ABSTRACT

Search for New Physics Using Top Quark Pairs Produced in Association with a Boosted Z or Higgs Boson in Effective Field Theory

Bryan David Caraway, Ph.D.

Advisor: Kenichi Hatakeyama, Ph.D.

A data sample containing top quark pairs ($t\bar{t}$) produced in association with a boosted Z or Higgs boson is used to search for signs of new physics within the framework of effective field theory. The data correspond to an integrated luminosity of 138 fb^{-1} of proton-proton collisions produced at a center-of-mass energy of 13 TeV at the LHC and collected by the CMS experiment. Selected collision events contain a single lepton and hadronic jets, including two identified with the decay of bottom quarks, plus an additional large-radius jet with high transverse momentum identified as a Z or Higgs boson decaying to a bottom quark pair. Machine learning techniques are employed to discriminate $t\bar{t}Z$ and $t\bar{t}H$ events from background processes, which are dominated by $t\bar{t}$ + jets production. The signal strengths of boosted $t\bar{t}Z$ and $t\bar{t}H$ processes are measured, and upper limits are placed on the $t\bar{t}Z$ and $t\bar{t}H$ differential cross sections as a function of the Z or Higgs boson transverse momentum. In addition, effects of physics beyond the standard model are probed using a framework in which

the standard model is considered to be the low-energy effective field theory of a higher-scale theory. Eight possible dimension-six operators are added to the standard model Lagrangian and their corresponding coefficients are constrained via a fit to the data.

Search for New Physics Using Top Quark Pairs Produced in Association with a Boosted Z or Higgs Boson in Effective Field Theory

by

Bryan David Caraway, B.S., M.A.

A Dissertation

Approved by the Department of Physics

Lorin S. Matthews, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

Kenichi Hatakeyama, Ph.D., Chairperson

Andrew Brinkerhoff, Ph.D.

Jay R. Dittmann, Ph.D.

Walter Wilcox, Ph.D.

Gerald B. Cleaver, Ph.D.

Liang Dong, Ph.D.

Accepted by the Graduate School
May 2022

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF ACRONYMS

ALICE	A Large Ion Collider Experiment
APD	Avalanche Photodiode
ATLAS	A Toroidal LHC Apparatus
AUC	Area Under the Curve
BPIX	Barrel Pixel
BSM	Beyond the Standard Model
CERN	European Organization for Nuclear Research
CKM	Cabibbo-Kobayashi-Maskawa
CL	Confidence Level
CMS	Compact Muon Solenoid
CSC	Cathode Strip Chamber
DNN	Deep Neural Network
DT	Drift Tube
DY	Drell-Yan
ECAL	Electromagnetic Calorimeter
EFT	Effective Field Theory
EWSB	Electroweak Symmetry Breaking
FPIX	Forward Pixel
FPR	False Positive Rate
FSR	Final State Radiation
GSF	Gaussian Sum Filter
HB	Hadron Calorimeter Barrel
HCAL	Hadron Calorimeter
HE	Hadron Calorimeter Endcap
HF	Hadron Calorimeter Forward

HL-LHC	High Luminosity Large Hadron Collider
HLT	High Level Trigger
HO	Hadron Calorimeter Outer
HPD	Hybrid Photodiode
ISR	Initial State Radiation
JER	Jet Energy Resolution
JES	Jet Energy Scale
JMR	Jet Mass Resolution
JMS	Jet Mass Scale
L1	Level 1
LEP	Large Electron-Positron
LHCb	Large Hadron Collider Beauty
LHC	Large Hadron Collider
LINAC	Linear Accelerator
LO	Leading Order
MC	Monte Carlo
ML	Machine Learning
MVA	Multi-variate Algorithm
NLO	Next-to-Leading Order
NN	Neural Network
NNLL	Next-to-Next-to-Leading Logarithmic
NNLO	Next-to-Next-to-Leading Order
NP	Nuisance Parameter
PDF	Parton Distribution Function
PF	Particle Flow
PMT	Photomultiplier Tube
POG	Physics Object Group
POI	Parameter Of Interest

P5	Point 5
PS	Parton Shower
PU	Pileup
PV	Primary Vertex
QCD	Quantum Chromodynamics
QED	Quantum Electrodynamics
QFT	Quantum Field Theory
ReLU	Rectified Linear Unit
ROC	Receiver Operating Characteristic
RPC	Resistive Plate Chamber
SF	Scale Factor
SiPM	Silicon Photomultiplier
SM	Standard Model
SPS	Super Proton Synchrotron
STXS	Simplified Template Cross Section
SUSY	Supersymmetry
SV	Secondary Vertex
TEC	Tracker Endcap
TIB	Tracker Inner Barrel
TID	Tracker Inner Disks
TOB	Tracker Outer Barrel
TPR	True Positive Rate
VEV	Vacuum Expectation Value
VPT	Vacuum Phototriode
WC	Wilson Coefficients
WP	Working Point
4FS	4-Flavor Scheme
5FS	5-Flavor Scheme

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

During my time at Baylor University, several people played a significant role in helping me complete this work. These individuals provided the necessary knowledge and support for me to succeed as a graduate student. I would like to take this opportunity to express my immense gratitude. First, I would like to thank my advisor, Dr. Kenichi Hatakeyama, for your time, patience, and guidance over the years. Under your tutelage, I learned a great deal not only about particle physics, but also how to work relentlessly to achieve my goals. Next, I want to express my thanks to Dr. Jay Dittmann, for introducing me to hardware work and for always being available when I needed advice. I would also like to thank Dr. Andrew Brinkerhoff, for your knowledge and perspective on the numerous challenges we encountered during this work. Additionally, I would like to thank the extremely knowledgable Dr. Joe Pastika and Dr. Jon Wilson. I appreciate the lessons I learned from our interactions and your contributions to this work. I also want to thank Dr. Chris Madrid, for the engaging conversations and for your friendship. Working with you was truly a pleasure. I want to express my thanks to Dr. Caleb Smith, Ankush Kanuganti, Brooks McMaster, and all others who helped me along the way. To everyone I mentioned, you played an instrumental role in my growth and I will always be grateful. Lastly, I am very grateful for the financial support I received from the Department of Energy (DE-SC0007861), which has made my education and research possible.

This work is dedicated to my parents, Karen and David, and to Monica, for your unwavering support and love during this journey.

CHAPTER ONE

Introduction

The standard model (SM) of particle physics is the best known theoretical explanation for the interactions of matter at the smallest space and time scales. Over the period of its development starting in the latter half of the twentieth century, the predictive power of the SM theory has been validated through the discovery of the elementary particles such as the Z and W bosons in 1983 at CERN's Super Proton-Antiproton Synchrotron, and the top quark in 1995 at the Tevatron located at the Fermi National Accelerator Laboratory. The most notable and last confirmation of the SM was the joint discovery of the Higgs boson in 2012 at CERN's Large Hadron Collider (LHC) by the CMS and ATLAS Collaborations. Additionally, precision measurements of the physical properties and interactions of the fundamental particles have strengthened the soundness of the SM.

However, several observed phenomena cannot be explained through the SM alone. The existence of dark matter, massive neutrinos, and the matter-antimatter asymmetry all indicate a more comprehensive theory beyond the standard model (BSM). To date, experiments searching for specific BSM models have yet to discover any evidence for their validity.

Many of these searches have taken place at the world's largest and most energetic particle accelerator, the LHC, which began operation in 2009. The LHC collides two proton beams at a center-of-mass energy of $\sqrt{s} = 13$ TeV at several points along the collider. Two of the interaction points are at the center of large general purpose

particle detectors, CMS and ATLAS. The collision energies achieved by the LHC allow for the production of the heaviest, second and third heaviest elementary particles: the top quark, Higgs boson and Z boson respectively. The amount data collected at the LHC is sufficient to observe extremely rare processes involving top quark pair associated with a heavy boson.

Effective field theory (EFT) treats the SM as a low-energy approximation to a more fundamental higher-energy theory with new particles at an arbitrary, unknown mass scale Λ. One advantage of EFT is that it introduces the effects of new physics on the current SM interactions in a model-independent way with the assumption that new particles are unable to be produced at the current energies achieved at the LHC. These are modeled as high-order operators in the SM Lagrangian with an unknown coupling strength known as the Wilson coefficient (WC). A careful analysis of observed SM processes may yield observations consistent with the effects of operators in EFT. The downside of this framework is the inability to measure the mass of new particles. However, signs of new physics signatures may assist theorists to refine BSM models which can guide experimentalists in their search.

The focus of this thesis is a search for new physics using EFT in a data set containing top quark pairs associated with a Z ($t\bar{t}Z$) or Higgs boson ($t\bar{t}H$). The data set, which is collected by the CMS detector during the Run 2 data-taking period (2016–2018) requires a single charged lepton from the decay of a top quark pair, plus a Z or Higgs boson with a large transverse momentum decaying to a bottom quark pair ($b\bar{b}$). Phenomenological studies motivate this choice by indicating that $t\bar{t}Z$ or $t\bar{t}H$ with a Lorentz boosted heavy boson demonstrate an enhanced sensitivity

from the effects of the EFT operators. Data consistent with this signature are used to construct a model to constrain the Wilson coefficients associated with these operators.

This thesis is organized as follows. Chapter Two provides an overview of the theoretical framework of the standard model theory, open questions regarding this theory, and an introduction to a model-independent extension of the SM using effective field theory. Next, Chapter Three describes in detail the Large Hadron Collider and the Compact Muon Solenoid experiment, and how data are collected. The concepts of data simulation and data reconstruction are introduced in Chapter Four, with a focus on details directly concerning this thesis. Chapter Five presents an analysis performed on data with the purpose of searching for signs of new physics using the effective field theory framework. Lastly, Chapter Six summarizes this work.

CHAPTER TWO

Theory

The current best description of the known fundamental particles and the interactions between them is the standard model theory of particle physics. The SM is a gauge theory that satisfies $\mathrm{SU(3)_C} \otimes \mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$ gauge invariance. It is a largely successful theory, making numerous predictions which closely agree with experimental observations. However, there are a few observations which the SM fails to describe, and predicates an extension to the theory. Section 2.1 provides an overview of the SM including a few open questions, and is mostly based on Refs. [1–6]. Section 2.2 describes a method to extend the standard model using effective field theory.

2.1 The Standard Model

The standard model of particle physics describes the known elementary particles and the fundamental interactions between them. As depicted in Fig. 2.1, the list of observed particles includes the six quarks, six leptons, four gauge bosons, and one scalar boson. They are classified into two major categories: fermions and bosons.

The fermions make up all of the observed matter in the known universe. They have a spin of one-half and obey Fermi-Dirac statistics, and thereby the Pauli exclusion principle. Additionally, every fermion in the SM has an antiparticle partner which is nearly identical except it has the opposite quantum numbers. For example, the electric charge changes from $-e$ to $+e$ where e is the elementary charge. The elementary particles of the fermion family are subdivided into the quarks and leptons, with each grouped into three "generations". Quarks and leptons each have six types

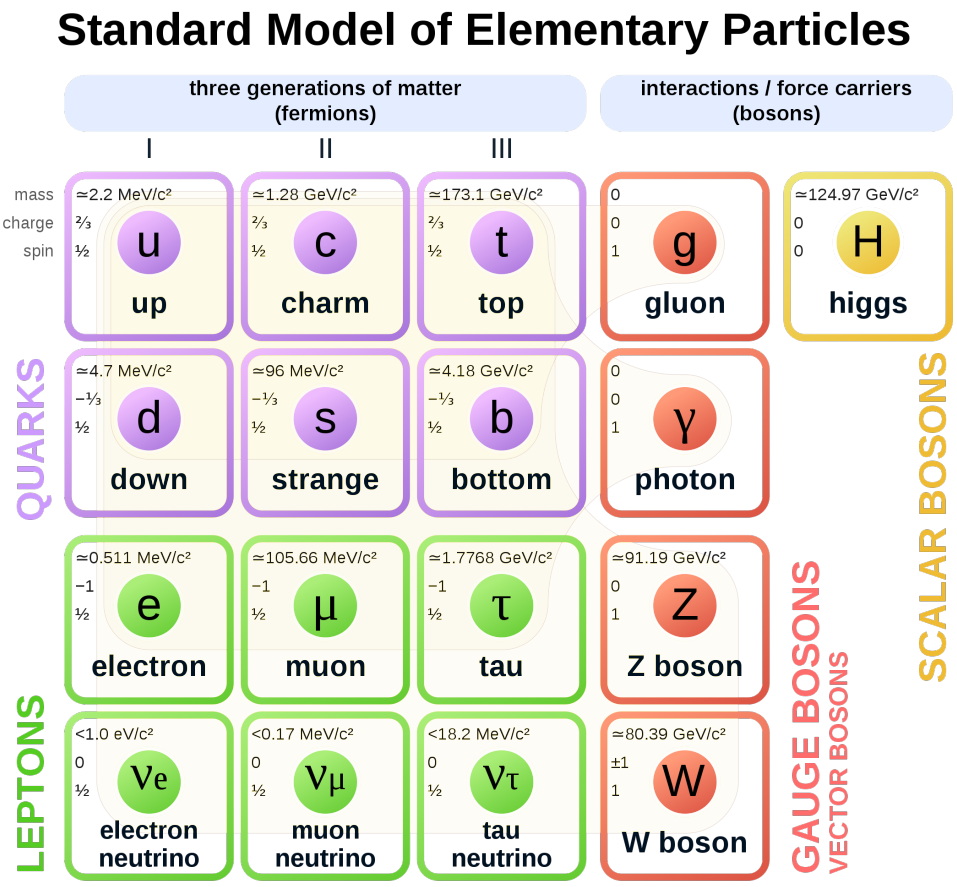

Figure 2.1: An outline of the SM of elementary particles which is grouped according to the characteristics of the particles. The observed particles included in the SM are the six quarks, six leptons, four gauge bosons, and the Higgs scalar boson. Figure source [7].

of particles, or so-called "flavors". The flavors belonging to the third generation have relatively large mass, are short-lived, and decay to lower generation particles until reaching the stable first generation flavors. It is for this reason that all ordinary matter comprises electrons, up quarks, and down quarks.

The electron (e), muon (μ), and tau (τ) are the charged leptons ($-e$ electric charge) of the first, second, and third generations respectively. Each charged lepton has a corresponding neutrino (ν_e, ν_μ, ν_τ), which is electrically neutral, has near-zero mass, and an extremely low rate of interaction. In fact, neutrinos are known to often pass through the entire Earth without stopping. Charged leptons may interact with other particles through the electromagnetic force which is mediated by photons, and both charged leptons and neutrinos interact via the weak force which is mediated by the W and Z bosons.

Quarks also have three generations of elementary particles. The up (u) and down (d) quarks compose the first generation, followed by the charm (c) and strange (s) quarks for the second, and lastly the third generation consists of the top (t) and bottom (b) quarks. Unlike the charged leptons, quarks have fractional electric charge. The up, charm, and top flavors all have $+\frac{2}{3}e$ charge, while the down, strange, and bottom quarks have a charge of $-\frac{1}{3}e$. In addition to the weak and electromagnetic force, quarks interact with the strong force. Because of the unique nature of quarks and their connection with the strong force, they cannot exist freely and must bind with other quarks to form composite particles, known as hadrons. The class of integer spin hadrons made up of a quark and antiquark pair are referred to as mesons. Baryons are hadrons with half-integer spin which are made up three quarks such as protons (uud) and neutrons (udd).

The top quark in particular has some noteworthy characteristics that are worth mentioning. The predominant property of the top quark is its large mass of approximately 173 GeV, as measured by the CMS and ATLAS experiments at the Large Hadron Collider [8]. Not only is it the most massive quark, but it is also the most massive SM particle. Due to the top quark having a mass on the same order as the Higgs boson, the top quark has a strong coupling to the Higgs interaction as compared to the other quarks. Additionally, the top quark has an extremely short lifetime, and will decay almost immediately without forming a bound state with other quarks. The decay products are almost always a bottom quark and a W boson due to a low mixing probability of the top quark with other first and second generation quarks.

In contrast to fermions, bosons have integer spin, obey Bose-Einstein statistics, and serve as the mediator of the interactions between particles. The fundamental forces observed in nature arise from interactions with the SM bosons, which is why the vector gauge bosons are also referred to as force-carrier particles. Thus far, three of the four fundamental forces are described in the SM. The gluon mediates the strong force, the photon is the force carrier for the electromagnetic force, and the Z and $W^{\pm}$ bosons relay the weak force. All vector bosons have a spin of magnitude one. The only force left out of the SM description is the gravitational force, however at the elementary particle scale its magnitude is so small compared to the others that it can be safely be ignored. Last, but certainly not least, the Higgs boson is the only scalar boson in the SM. It is unique compared to the vector bosons in that its spin is zero and does not mediate a force. Rather, the Higgs particle is associated with how elementary particles gain their property of mass.

The theoretical framework of the standard model is based on quantum field theory and the Lagrangian density formalism. The SM Lagrangian $\mathcal{L}_{\mathrm{SM}}$ serves as a mathematical description of all the known particles and interactions at the fundamental level. Additionally, the $\mathcal{L}_{\mathrm{SM}}$ is built satisfying $\mathrm{SU(3)_C} \otimes \mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$ gauge symmetry. The short form of the $\mathcal{L}_{\mathrm{SM}}$ may be written as:

$$\begin{aligned}\mathcal{L}_{\mathrm{SM}} = \quad & \underbrace{-\frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a - \frac{1}{4}W^I_{\mu\nu}W^{\mu\nu}_I - \frac{1}{4}B_{\mu\nu}B^{\mu\nu}}_{\text{kinetic term for the respective } \mathrm{SU(3)_C},\ \mathrm{SU(2)_L},\ \text{and } \mathrm{U(1)_Y} \text{ gauge fields}} \\ & \underbrace{+i\bar{\psi}\gamma^\mu D_\mu \psi}_{\text{interaction gauge bosons / fermions}} \quad \underbrace{-\lambda_f(\bar{\psi}_L\varphi\psi_R + \bar{\psi}_R\varphi\psi_L)}_{\text{interaction Higgs / fermions}} \\ & \underbrace{+(D^\mu\varphi)^\dagger(D_\mu\varphi)}_{\text{Higgs kinetic term}} \quad \underbrace{-V(\varphi)}_{\text{Higgs scalar potential}} \quad ,\end{aligned} \tag{2.1}$$

where $D_\mu = \partial_\mu - ig_s T^a G^a_\mu - ig\tau^I W^I_\mu - \frac{i}{2}g'YB_\mu$. The following sections will provide further details on the theoretical framework of the SM as a quantum field theory and introduce the Lagrangian density formalism. Additionally, an introduction to quantum electrodynamics, quantum chromodynamics, electroweak unification, and electroweak symmetry breaking will be presented to contextualize the components of Eq. (2.1). Lastly, this section will conclude with a few select open questions left unanswered by the SM theory.

2.1.1 Introduction to Quantum Field Theory

Quantum field theory (QFT) serves as the theoretical foundation of the SM and is centered around the hypothesis that particles and waves may be expressed as continuous fields that exist throughout all spacetime. In this theory, an elementary particle is the result of an excitation of a field. The mathematical formalism utilized to dictate the dynamics of the fields and interactions among them is the Lagrangian

density, $\mathcal{L}$. To begin the process of writing down the entire SM Lagrangian, the Dirac Lagrangian $\mathcal{L}_{\text{Dirac}}$ of a free spinor field, ψ, is formulated as:

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

where γ^{μ} are the Dirac matrices and $\bar{\psi} \equiv \psi^{\dagger}\gamma^{0}$ is the adjoint Dirac spinor with $\psi^{\dagger}$ denoting the hermitian conjugate of ψ. From the Lagrangian, the action S is written as:

$$S = \int \mathcal{L}\ dt\ dx, \tag{2.3}$$

and the equations of motion of the system follows from the principle of least action:

$$\partial S = \frac{\partial \mathcal{L}}{\partial \psi} - \partial_{\mu}\frac{\partial \mathcal{L}}{\partial(\partial_{\mu}\psi)} = 0. \tag{2.4}$$

The solution to the equation above is a set of 4-component Dirac spinors, which may be interpreted as the kinematics of fermions.

A set of properties of the SM Lagrangian, which has important physical implications, is the observed symmetries that exist under certain transformations. According to Noether's theorem, each differential symmetry of the Lagrangian implies a corresponding conservation law. For example, it is trivial to see that the Dirac Lagrangian is invariant under the global U(1) transformation, which is expressed as:

$$\begin{aligned} \psi &\rightarrow e^{i\alpha}\psi, \\ \bar{\psi} &\rightarrow e^{-i\alpha}\bar{\psi}. \end{aligned} \tag{2.5}$$

The physical interpretation of this symmetry is a conserved current, and also the conservation of electric charge. However, it is more difficult, but more meaningful, to show that the Dirac Lagrangian is invariant under a local transformation. In this

example, the $\mathcal{L}_{\text{Dirac}}$ is not invariant under the following local transformation:

$$\begin{aligned} \psi &\to e^{iq\alpha(x)}\psi, \\ \bar{\psi} &\to e^{-iq\alpha(x)}\bar{\psi}. \end{aligned} \tag{2.6}$$

Compared to the global transformation, the derivative will act on the $\alpha(x)$ term and add an extra term to the transformed $\mathcal{L}_{\text{Dirac}}$. However, the Lagrangian formalism can be altered with a gauge field term to make it invariant under such local transformation.

A gauge field is a special vector field derived from gauge theory. The term gauge refers to a redundancy in the mathematical formalism of the degrees of freedom of a physical system. By taking advantage of the redundancies, any gauge transformation of the system will result in no net change even if it is a local transformation. Under such conditions, the system is said to be gauge invariant which implies a gauge symmetry, and leads to a conservation law. For example, we may reconsider the local transformation of Eq. (2.2) and replace the derivative ∂_μ with the covariant derivative D_μ which is defined as:

$$D_\mu = \partial_\mu - iqA_\mu(x), \tag{2.7}$$

where $A_\mu(x)$ is a gauge field which is required to transform as:

$$A_\mu(x) \to A_\mu(x) + \frac{1}{q}\partial_\mu\alpha(x). \tag{2.8}$$

Here, q is constant and describes the coupling strength of $A_\mu(x)$ with the spinor field ψ. With these additions, local transformation invariance of $\mathcal{L}_{\text{Dirac}}$ is preserved. This example represents a simple unitary local transformation. Moreover, this concept can be extended to higher orders with special unitary groups developed by Yang and Mills: SU(n) of order n. The group includes $n^2 - 1$ gauge generators T^a, where $a \in \{1, 2, \cdots, n^2-1\}$, and each generator is associated with a gauge field A^a_μ. Likewise,

the covariant derivative includes the generator terms and takes the form:

$$D_\mu = \partial_\mu - iqT^a A^a_\mu(x). \tag{2.9}$$

The physical manifestations of the gauge field A^a_μ are the spin-1 gauge bosons, and for each field an additional gauge invariant kinetic term is added to the Lagrangian. Additionally, if the gauge group associated with the gauge field is non-abelian, meaning the generators do not commute, then the associated gauge bosons are allowed to participate in self-interactions. Following these principles, the gauge bosons corresponding to the electromagnetic force, the strong force, and the weak force are added to the formalism.

2.1.2 Quantum Electrodynamics

The electromagnetic force is described by the gauge theory, quantum electrodynamics (QED). Thus far, the concepts discussed in Section 2.1.1 and the introduction of a simple unitary gauge group U(1) are congruent with QED. In fact, the gauge field A_μ is one and the same with respect to the photon field. As mentioned before, an additional gauge invariant kinematic term of the photon field is added to the Lagrangian and takes the form:

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial^\mu A^\nu, \tag{2.10}$$

where $F_{\mu\nu}$ is the field strength tensor. Now building upon the $\mathcal{L}_{\text{Dirac}}$, the QED Lagrangian, $\mathcal{L}_{\text{QED}}$, is written as:

$$\begin{aligned} \mathcal{L}_{\text{QED}} &= \underbrace{-\frac{1}{4}F_{\mu\nu}F^{\mu\nu}}_{\text{kinetic } A_\mu} \underbrace{- q\bar{\psi}\gamma^\mu A_\mu\psi}_{\text{interaction } \psi/A_\mu} \underbrace{+ i\bar{\psi}\gamma^\mu\partial_\mu\psi}_{\text{kinetic } \psi} \underbrace{- m\bar{\psi}\psi}_{\text{mass } \psi} \\ &= -\frac{1}{4}F_{\mu\nu}F^{\mu\nu} + \bar{\psi}(i\gamma^\mu D_\mu - m)\psi, \end{aligned} \tag{2.11}$$

where q is the QED coupling strength constant which is also the fine-structure constant $q = \frac{e^2}{4\pi}$. Using the same conditions mentioned in Section 2.1.1, the $\mathcal{L}_{\text{QED}}$ possess $\text{U}(1)_{\text{EM}}$ gauge invariance and leads to the conservation law of electric charge. The gauge boson associated with A_μ is none other than the photon. From $\mathcal{L}_{\text{QED}}$, several properties of the photon can be inferred. For example, the addition of a photon mass term in the Lagrangian would destroy the local invariance, thus it follows that the photon is massless. Additionally, the $\text{U}(1)_{\text{EM}}$ gauge group is abelian which means the photons cannot self-interact. Therefore photons do not possess an electric charge.

2.1.3 Quantum Chromodynamics

Quantum chromodynamics (QCD) is a $\text{SU}(3)_{\text{C}}$ gauge theory which describes the strong force. This symmetry group possesses $3^2 - 1 = 8$ generators which are the 3×3 Gell-Mann matrices defined as $T^a \equiv \frac{1}{2}\lambda^a$. By construction, T^a matrices are Hermitian, traceless, linearly independent, and satisfy the commutation relation:

$$[T^a, T^b] = if^{abc}T^c, \tag{2.12}$$

where f^{abc} are the structure constants of the $\text{SU}(3)_{\text{C}}$ group. The T^a generators associate with eight gauge fields; the particle manifestation of which are the eight gluons. The conservation law of QCD following the $\text{SU}(3)_{\text{C}}$ gauge symmetry is the conservation of the so-called color current and color charge. The conventions of color charge are especially unique, yet they serve as a helpful tool for understanding the behavior of quarks and gluons. Specifically, quarks possessing different color charges (red, green, blue) will bind together to form colorless composite hadrons. For example, a quark and antiquark with red and antired color charges, respectively, will form a colorless meson. Additionally, a baryon such as a proton will have a red, green, and

blue quark bound together. The strong force, which binds the quarks together, is facilitated by the gluon. The gluon, which has two different color charges, is able to swap the color charge between quarks and gluons.

In a similar approach to formulating the $\mathcal{L}_{\text{QED}}$, the imposition of $\text{SU}(3)_{\text{C}}$ gauge symmetry on the Dirac Lagrangian and the addition of a kinetic term for the gluon field yield the QCD Lagrangian, $\mathcal{L}_{\text{QCD}}$. The $\mathcal{L}_{\text{QCD}}$ is formulated as:

$$\mathcal{L}_{\text{QCD}} = -\frac{1}{4} G^a_{\mu\nu} G^{\mu\nu}_a + \bar{\psi}(i\gamma^\mu D_\mu - m)\psi, \tag{2.13}$$

where ψ represents the Dirac quark spinors, and $G^a_{\mu\nu}$ is the gluon field strength tensor. In this formulation, the covariant derivative is defined as

$$D_\mu = \partial_\mu - i g_s T^a G^a_\mu, \tag{2.14}$$

where g_s is the coupling strength constant between the quarks and gluon field G^a_μ. In all actuality, g_s is not strictly a constant; rather its value depends on the energy scale at which the strong interaction is probed. Similar to the photon, the addition of a gluon mass term in Eq. (2.13) would destroy the $\text{SU}(3)_{\text{C}}$ gauge invariance, thus the gluon is massless. However in contrast to the photon and the QED $\text{U}(1)_{\text{EM}}$ gauge theory, the symmetry in QCD is non-abelian which means the gluon may interact with other gluons. This property is reflected in the gluon field strength tensor:

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu + g_s f^{abc} G^b_\mu G^c_\nu. \tag{2.15}$$

The self-interaction of the gluon is ultimately responsible for the phenomenon of asymptotic freedom where the coupling strength decreases at higher energies and as the distance between interacting colored particles decreases. Likewise, the coupling strength diverges at lower energies and as the distance increases between quarks.

This phenomenon is described as the principle of color confinement which states that quarks only exist in bound colorless states with other quarks and cannot be isolated.

2.1.4 The Weak Force and Electroweak Unification

The study of the dynamics of nuclear decay, specifically β-decay where $n \rightarrow p + e^- + \bar{\nu_e}$, led to the postulation of the weak force. Later, more advanced studies of nuclear decay discovered that these interactions violated parity. This means that the weak force demonstrates bias based on the chirality, left-handedness or right-handedness, of a particle. As it turns out, the charged weak force only couples to left-handed particles while the neutral weak force couples to both left and right-handed particles. In order to write down the dynamics of the weak force, first the fermionic field ψ needs to be split according to chirality:

$$\begin{aligned} \psi_L &= (\frac{1-\gamma^5}{2})\psi, \\ \psi_R &= (\frac{1+\gamma^5}{2})\psi, \\ \psi &= \psi_L + \psi_R, \end{aligned} \tag{2.16}$$

where ψ_L and ψ_R are the left-handed and right-handed fermionic fields, respectively. The weak interaction also satisfies the SU(2) gauge group invariance, and the conservation law associated with this is isospin I_3. Additionally, left-handed fermions are grouped as doublets and have $I_3 = \pm\frac{1}{2}$, while right-handed fermions form singlets with $I_3 = 0$.

As hinted by the apparent charge current within β-decay, a form of the weak interaction carries electric charge which implies some relation to the electromagnetic force. This idea led to the theoretical development of a single unified electroweak force consisting of weak and electromagnetic forces. By unifying the two gauge theories, the

gauge symmetry that the electroweak force possesses is denoted as $\mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$. The conserved quantity of $\mathrm{U(1)_Y}$ is hypercharge which is a function of the isospin and electric charge:

$$Y = 2(Q - I_3). \tag{2.17}$$

The $\mathrm{SU(2)_L}$ gauge group has $2^2 - 1 = 3$ generators $\tau^a \equiv \frac{1}{2}\sigma^a$ where σ^a are 2×2 Pauli matrices and τ^a satisfies the relation:

$$[\tau^a, \tau^b] = i\varepsilon^{abc}\tau^c, \tag{2.18}$$

where ε^{abc} are the structure constants of the $\mathrm{SU(2)_L}$ group. The three boson gauge fields which correspond to the $\mathrm{SU(2)_L}$ generators are denoted as W^a_μ where $a \in \{1, 2, 3\}$. The W^0 boson is electrically neutral with $I_3 = 0$, and the W^1 and W^2 bosons are electrically charged with non-null isospin: $Q = \pm e$ and $I_3 = \pm 1$, respectively. As one might expect, the $\mathrm{U(1)_Y}$ gauge group with gauge field B_μ shares many similarities with the $\mathrm{U(1)_{EM}}$ group discussed in Section 2.1.2. Here, the boson B has isospin $I_3 = 0$.

Utilizing a similar convention to that of Eqs. (2.11) and (2.13), the Lagrangian for the electroweak interaction $\mathcal{L}_{\mathrm{EWK}}$ is written as:

$$\mathcal{L}_{\mathrm{EWK}} = -\frac{1}{4}W^a_{\mu\nu}W^{\mu\nu}_a - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} + \bar{\psi}_L(i\gamma^\mu D^L_\mu)\psi_L + \bar{\psi}_R(i\gamma^\mu D^R_\mu)\psi_R, \tag{2.19}$$

where the respective left and right-handed covariant derivatives are D^L_μ and D^R_μ, and the field strength tensors are $B^{\mu\nu}$ and $W^{\mu\nu}_a$. The covariant derivatives are written as:

$$\begin{aligned} D^L_\mu &= \partial_\mu - ig\tau^a W^a_\mu - \frac{i}{2}g'YB_\mu, \\ D^R_\mu &= \partial_\mu - ig'YB_\mu, \end{aligned} \tag{2.20}$$

where g and g' are the coupling strength terms for the gauge fields W^a_μ and B_μ respectively, and Y is the hypercharge. Additionally, the field strength tensors for

the $\mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$ gauge group are formulated as:

$$\begin{aligned} W^a_{\mu\nu} &= \partial_\mu W^a_\nu - \partial_\nu W^a_\mu + g\varepsilon^{abc} W^b_\mu W^c_\nu, \\ B_{\mu\nu} &= \partial_\mu B_\nu - \partial_\nu B_\mu. \end{aligned} \tag{2.21}$$

From Eqs. (2.19)–(2.21), there a few important things to interpret from the electroweak framework. First, there is no mass term for the W^a or B bosons in Eq. (2.19) as this would break the gauge symmetry for the $\mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$ gauge group. Additionally, there is no longer a mass term for the fermionic field as this would violate $\mathrm{SU(2)_L}$ gauge invariance due to the asymmetric behavior of the weak interaction with respect to particle chirality as evident in Eq. (2.20). Because the weak force can only interact with left-handed fermions and neutrinos have no electrical charge, the right-handed neutrino cannot participate in any of the electroweak interactions. Lastly, the $W^a_{\mu\nu}$ field strength tensor contains a term for self-interaction while $B_{\mu\nu}$ does not.

The physically observed particles of the $\mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$ gauge group are the $\mathrm{W}^\pm$ and Z bosons, and the photon which are linear superpositions of the gauge fields:

$$\begin{aligned} \mathrm{W}^\pm &= \frac{1}{\sqrt{2}}(W^1 \mp iW^2), \\ \begin{pmatrix} \gamma \\ \mathrm{Z} \end{pmatrix} &= \begin{pmatrix} \cos(\theta_W) & \sin(\theta_W) \\ -\sin(\theta_W) & \cos(\theta_W) \end{pmatrix} \begin{pmatrix} B \\ W^0 \end{pmatrix}, \end{aligned} \tag{2.22}$$

where θ_W is the weak mixing angle (Weinberg angle).

2.1.5 The Higgs Boson and Electroweak Symmetry Breaking

In Section 2.1.4, it was mentioned that under the imposed $\mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$ gauge invariance the electroweak gauge bosons and fermions are prohibited from having mass. However, the $\mathrm{W}^\pm$ and Z bosons and the quarks have been observed to be

massive. This tension between the electroweak theory and experimental observation is resolved with the Higgs mechanism developed by Brout, Englert, and Higgs [9, 10]. The Higgs mechanism introduces a complex scalar field φ to the $\mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$ gauge group, and is represented as a doublet comprising charged and neutral components with the form:

$$\varphi = \begin{pmatrix} \varphi^+ \\ \varphi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \varphi^1 + i\varphi^2 \\ \varphi^3 + i\varphi^4 \end{pmatrix}, \tag{2.23}$$

and a scalar potential $V(\varphi)$ which is written as:

$$V(\varphi) = \mu^2 \varphi^\dagger \varphi + \lambda (\varphi^\dagger \varphi)^2, \tag{2.24}$$

where μ and λ are constants. Given these conditions, the Lagrangian of the Higgs mechanism $\mathcal{L}_{\mathrm{Higgs}}$ is formulated as:

$$\mathcal{L}_{\mathrm{Higgs}} = (D^\mu \varphi)^\dagger (D_\mu \varphi) - V(\varphi). \tag{2.25}$$

From Eq. (2.24) , the constants dictate the shape of the Higgs scalar potential. First, the coupling constant λ should be positive for the potential to be bounded by zero from below. The choice $\mu^2 > 0$ yields a potential with a single minimum vacuum potential state $V(\varphi_0) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ where φ_0 is the ground state of the scalar field. On the other hand, the choice $\mu^2 < 0$ gives the "Mexican hat" potential, depicted in Fig. 2.2, which has an infinite amount degenerate minima and a non-zero vacuum expectation value (VEV) v expressed as:

$$\varphi_0^\dagger \varphi_0 = \frac{-\mu^2}{2\lambda} \equiv \frac{v}{2}. \tag{2.26}$$

While there are a number of solutions for the ground state scalar field, an interesting choice assigns the neutral component of the scalar field the VEV, thereby retaining

$\mathrm{U(1)_{EM}}$ charge symmetry but breaking electroweak symmetry (EWSB):

$$\varphi_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}. \tag{2.27}$$

Under this assumption, the bosons associated with $\mathrm{SU(2)_L}$ gauge group, known as Goldstone bosons, provide movement between the degenerate minima; that is around of crown of the "Mexican hat". As pointed out by Englert, Brout, and Higgs, a local unitary gauge transformation exploits this degeneracy and causes these bosons to vanish from the Lagrangian yielding an unique minimum and the remaining neutral scalar boson φ^0. This is none other than the famous Higgs boson (H). With these conditions, the scalar field is extended from the ground state with the physical Higgs field $\mathrm{H}(x)$:

$$\varphi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + \mathrm{H}(x) \end{pmatrix}. \tag{2.28}$$

Inserting Eq. (2.28) into the $\mathcal{L}_{\mathrm{Higgs}}$ combined with the covariant derivatives of $\mathrm{SU(2)_L} \otimes \mathrm{U(1)_Y}$, the Higgs field mixes with the weak interaction fields and generates mass terms for the $\mathrm{W}^{\pm}$ and Z bosons in terms of the VEV and the coupling strengths:

$$\begin{aligned} m_{\mathrm{W}} &= \frac{gv}{2}, \\ m_{\mathrm{Z}} &= \frac{v\sqrt{g^2 + g'^2}}{2}, \\ m_{\mathrm{H}} &= v\sqrt{2\lambda}, \end{aligned} \tag{2.29}$$

where λ is the same self-coupling strength term from before and the means by which the Higgs boson obtains its mass. Under the assumptions of the Higgs mechanism, the photon does not interact with the Higgs field and therefore retains its property of masslessness. Similarly, the massless gluon is unable to interact with the Higgs boson.

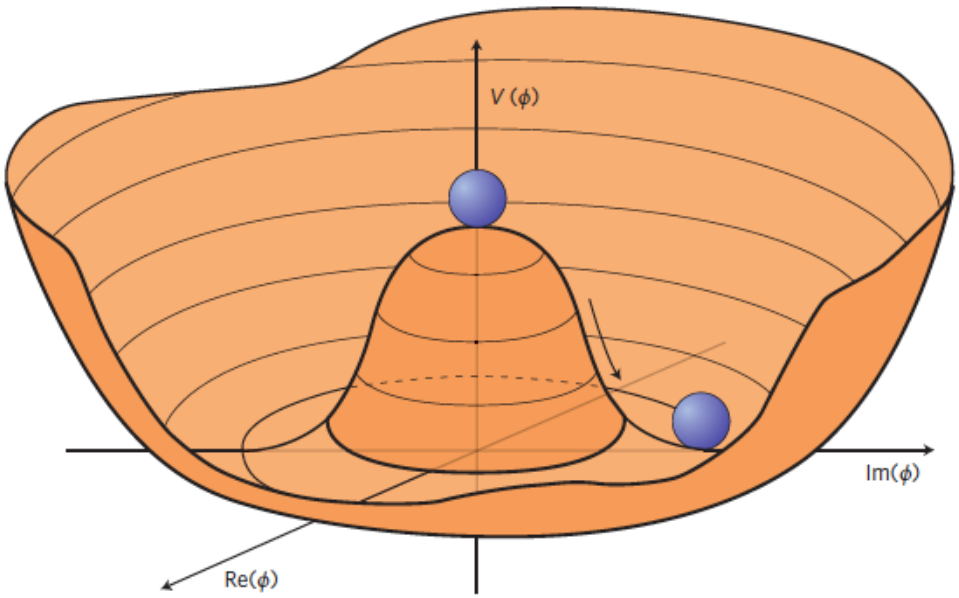

Figure 2.2: An illustration of the shape of the "Mexican hat" Higgs scalar potential. Figure source [11].

Fermions, specifically charged leptons and quarks, also acquire mass by their interactions with the Higgs field. This is accomplished through the Yukawa interactions where the Higgs field mixes with the fermionic field as seen in the Yukawa Lagrangian, $\mathcal{L}_{\text{Yukawa}}$:

$$\begin{aligned} \mathcal{L}_{\text{Yukawa}} &= -\lambda_f(\bar{\psi}_L\varphi\psi_R + \bar{\psi}_R\varphi\psi_L), \\ &= -\lambda_{f_{ij}}(\bar{q}_i\varphi d_j + \bar{q}_i\tilde{\varphi}d_j + \bar{l}_i\varphi e_j) + \text{h.c.}\ , \end{aligned} \tag{2.30}$$

where λ_f is the Yukawa coupling strength of a particular fermion to the Higgs field, $\tilde{\varphi} = i\sigma_2\varphi^*$ where σ_2 is the second Pauli matrix, h.c. is the hermitian conjugate of the terms that come before, and q_i and l_i are the respective left-handed quark and lepton doublets. Additionally, e_i, and u_i and d_i are the right-handed charged lepton and quark singlets, respectively. The Yukawa coupling is not known a priori, and is inferred through the measurement of fermion mass m_f:

$$\lambda_f = m_f\frac{\sqrt{2}}{2}. \tag{2.31}$$

Importantly, because the SM does not allow for right-handed neutrinos, neutrinos cannot interact with the Higgs field in the Yukawa Lagrangian. Therefore, they are considered massless in the SM.

With the discovery of the Higgs boson in 2012 by the CMS and ATLAS experiments at the Large Hadron Collider, the Higgs field was finally validated as a way to generate mass for the SM particles. Since then, the mass of the Higgs boson has been measured to be approximately 125 GeV, as shown in Fig. 2.3, making it the second most massive particle in the SM.

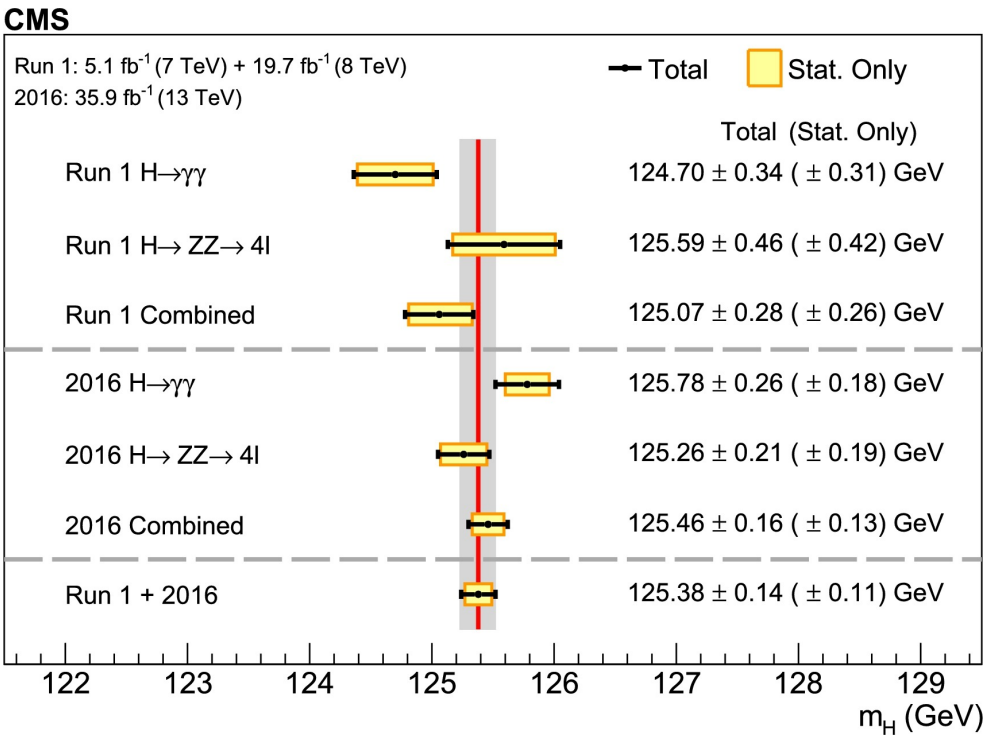

Figure 2.3: Measurements of the Higgs boson mass performed by the CMS experiment at the Large Hadron Collider. Figure source [12].

2.1.6 *Open Questions of the Standard Model*

The SM has been a largely successful theory, providing an accurate description of observable matter and the fundamental interactions between them. However, there are many open questions regarding the SM. For instance, there are several experimental observations which the SM cannot explain, and secondly certain free parameters of the $\mathcal{L}_{\text{SM}}$ do not have a priori motivation. A short summary of some of the major shortcomings is provided below:

- *Gravity:* The current SM theory includes three out of the four known forces in the universe, and leaves out a description for gravity. While the contribution of gravity is negligible at the quantum scale, this poses a problem in that the SM is not a complete description of all the known interactions. The attempt of adding a boson responsible for mediating the gravitational force, the so-called graviton, has been shown to cause the SM theory to diverge resulting in an unphysical theory.
- *Unification of the Forces:* For a long time, the unification of the electromagnetic, weak, and strong forces has been sought after by theorists. Such a description may be relevant for some extreme environments at high energy scales such as immediately following the Big Bang. Unfortunately, the SM does not converge to a unified description of the forces.
- *Dark Matter and Dark Energy:* The existence of dark matter is supported by measurements of the cosmological constant performed by the Planck satellite [13] as well as other astronomical observations [14, 15]. According to cosmology, dark matter should make up about 26% of all energy in the universe. The SM explains approximately 5% of the energy in the universe, but

it does not provide a suitable particle candidate matching the characteristics of dark matter. The measurements from the Planck satellite also support the existence of dark energy which is hypothesized to be responsible for the accelerated expansion of the universe. Dark energy is thought to comprise roughly 69% of all energy in the universe. Again, the SM provides no explanation consistent with dark energy.

- *Matter-Antimatter Asymmetry:* The observable matter that exists in the universe appears to be entirely made up of normal matter. This points to a mechanism with asymmetric behavior towards matter and antimatter, with a positive bias towards matter. Currently, the SM does not explain the observed matter-antimatter asymmetry.
- *Neutrino Mass:* Neutrinos have been observed to oscillate between generations. This observations implies that neutrinos have a non-zero mass, which is not included in the current SM description. While the mass term for the neutrino can be added to the SM, there is no mechanism to handle right-handed neutrinos. Also, there is an open question about how neutrinos should be theoretically described. They could be modeled as Dirac fermions or as Majorana fermions. If the latter, then the neutrino would be its own antiparticle. Under the Majorana description, neutrino mass can be added to the SM Lagrangian without violating gauge invariance.
- *The Hierarchy Problem:* The Higgs boson mass has been measured to be about 125 GeV. This can be expressed as its bare mass term plus additional terms describing the loop corrections: ${m_{\mathrm{H}}}^2 = (m_{\mathrm{H}}^0)^2 - \frac{|\lambda_{\mathrm{UV}f}|^2}{8\pi^2}\Lambda_{\mathrm{UV}}^2 + \cdots$. These corrections scale with the Yukawa coupling λ_f and a high energy cut-off scale

Λ_{UV} on the order of 10^{19} GeV, where quantum gravity is no longer negligible. For the Higgs mass to match what is observed, the bare mass term must be a highly tuned parameter in order to cancel the loop corrections. The cancellation seems artificial to the extent that theorists consider this to be a problem with the naturalness of the SM theory.

2.2 Effective Field Theory

As discussed in Section 2.1.6, the SM in its current form is incomplete. This fact motivates theorists to develop theories that go beyond the SM to help explain certain observed phenomena. One such BSM model, and perhaps the most promising, is Supersymmetry (SUSY) [16], which hypothesizes a supersymmetric partner particle for every current SM particle. With the SUSY model, several issues with the SM are conveniently addressed such as the hierarchy problem and the dark matter candidate. However, direct searches for SUSY particles at the Large Hadron Collider by the CMS and ATLAS experiments have, thus far, been unsuccessful. It may be that new particles exist at an energy scale which exceeds the capabilities of the Large Hadron Collider. However even if this is the case, effective field theory (EFT) [17–19] can be used to indirectly search for new particles by looking for novel interactions of the SM particles.

An EFT is a low-energy approximation of a more fundamental theory that exists at an energy scale Λ, with hypothetical particles assumed to have mass much larger than the existing SM particles. Thus, interactions mediated by new particles will be very off-shell and can be modeled as a point-like interaction as depicted in Fig. 2.4. In this way, the EFT formulation does not need to factor in a new gauge field

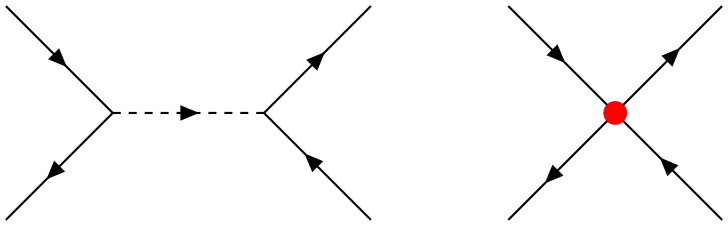

Figure 2.4: Feynman diagrams depicting an interaction mediated by a new particle produced on-shell (left) versus the low-energy approximated interaction modeled with EFT (right) represented by the red dot.

or scalar field when considering new physics effects in SM processes. Rather, higher-order operators defined as the product of SM gauge and scalar fields, and suppressed by the energy scale term $1/\Lambda$, are added to the dimension-four SM Lagrangian density:

$$\mathcal{L}_{EFT} = \mathcal{L}_{SM}^{(4)} + \frac{1}{\Lambda}\sum_i c_i^{(5)} O_i^{(5)} + \frac{1}{\Lambda^2}\sum_i c_i^{(6)} O_i^{(6)} + \cdots , \tag{2.32}$$

where $O_i^{(d)}$ are the effective operators of dimension d, and $c_i^{(d)}$ are the associated dimensionless coupling strengths otherwise known as the Wilson coefficients (WCs). In principle, effective field theory can extend the SM up to an arbitrary number of dimensions, however the energy scale term will suppress the contribution from higher-order operators. Therefore, the dimension-five and six operators will most likely have the largest measurable effect and are considered first when testing the EFT hypothesis. The new effective operators are required to satisfy the gauge invariance of the SM, but they do not necessarily conserve "coincidental" symmetries of the SM Lagrangian such as baryon and lepton numbers. The dimension-five operators violate lepton number conservation [20, 21], so only the 59 dimension-six operators constituting the Warsaw basis [18] remain to be considered. Of these, eight operators involve the interaction of at least one heavy boson field with two heavy quark fields as defined in Table 2.1. The physics analysis in Chapter Five probes the coupling

strength of these operators within events containing a top quark pair associated with a Z or Higgs boson. Example diagrams of $t\bar{t}H$ and $t\bar{t}Z$ constructed with EFT operators may be found in Appendix A.

Table 2.1: A set of dimension-six operators involving two quarks and at least one heavy boson. Additionally, the couplings are restricted to involve only third-generation quarks. The quantity $\sigma^{\mu\nu}$ is defined as $\frac{i}{2}(\gamma^\mu\gamma^\nu - \gamma^\nu\gamma^\mu)$ where γ^μ denotes the Dirac matrices. The third-generation quark doublet is represented by q, and u and d represent the right-handed third-generation quark singlets. Furthermore, $(\varphi^\dagger i\overleftrightarrow{D}_\mu\varphi) \equiv \varphi^\dagger(iD_\mu\varphi) - (iD_\mu\varphi^\dagger)\varphi$ and $(\varphi^\dagger i\overleftrightarrow{D}^I_\mu\varphi) \equiv \varphi^\dagger\tau^I(iD_\mu\varphi) - (iD_\mu\varphi^\dagger)\tau^I\varphi$. The abbreviations S_W and C_W denote the sine and cosine of the weak mixing angle in the unitary gauge, respectively.

Operator	Definition	WC
‡$O^{(ij)}_{u\varphi}$	$\bar{q}_i u_j\tilde{\varphi}\,(\varphi^\dagger\varphi)$	$c_{t\varphi} + ic^I_{t\varphi}$
$O^{1(ij)}_{\varphi q}$	$(\varphi^\dagger i\overleftrightarrow{D}_\mu\varphi)(\bar{q}_i\gamma^\mu q_j)$	$c^-_{\varphi Q} + c^3_{\varphi Q}$
$O^{3(ij)}_{\varphi q}$	$(\varphi^\dagger i\overleftrightarrow{D}^I_\mu\varphi)(\bar{q}_i\gamma^\mu\tau^I q_j)$	$c^3_{\varphi Q}$
$O^{(ij)}_{\varphi u}$	$(\varphi^\dagger i\overleftrightarrow{D}_\mu\varphi)(\bar{u}_i\gamma^\mu u_j)$	$c_{\varphi t}$
‡$O^{(ij)}_{\varphi ud}$	$(\tilde{\varphi}^\dagger iD_\mu\varphi)(\bar{u}_i\gamma^\mu d_j)$	$c_{\varphi tb} + ic^I_{\varphi tb}$
‡$O^{(ij)}_{uW}$	$(\bar{q}_i\sigma^{\mu\nu}\tau^I u_j)\,\tilde{\varphi}W^I_{\mu\nu}$	$c_{tW} + ic^I_{tW}$
‡$O^{(ij)}_{dW}$	$(\bar{q}_i\sigma^{\mu\nu}\tau^I d_j)\,\varphi W^I_{\mu\nu}$	$c_{bW} + ic^I_{bW}$
‡$O^{(ij)}_{uB}$	$(\bar{q}_i\sigma^{\mu\nu}u_j)\,\tilde{\varphi}B_{\mu\nu}$	$(C_W c_{tW} - c_{tZ})/S_W + i(C_W c^I_{tW} - c^I_{tZ})/S_W$

CHAPTER THREE

The LHC and the CMS Experiment

The particle collider has been the apparatus of choice to test the predictions of the standard model theory. Within the experimental sector of high energy physics, colliders have progressed in complexity and scale in order to accelerate particles and collide them at high energies. The culmination of this progression is the Large Hadron Collider. Particle detectors are built along the LHC to detect the products of the colliding particles. While there are several detectors located on the LHC, the one relevant to this thesis is the Compact Muon Solenoid (CMS) detector. This chapter provides details of the LHC and a description of the CMS detector.

3.1 The Large Hadron Collider

The European Organization for Nuclear Research, known as CERN, is the world's largest particle physics laboratory. It was founded in 1954 and is located on the French-Swiss border near Geneva, Switzerland. Since its inception, the facility has hosted a number of state-of-the-art experiments including the largest, most powerful particle accelerator, the LHC. The LHC was built in the pre-existing tunnel system constructed for its predecessor, the Large Electron-Positron (LEP) accelerator. The main tunnel which houses the beam infrastructure is underground at a depth between 50 and 175 m, and circular in shape with a circumference of about 26.7 km. The LHC is designed to collide proton-proton (pp) beams at a center-of-mass energy of up to 14 TeV.

3.1.1 *The CERN Accelerator Complex*

The proton bunches supplied to the LHC originate from a system of pre-accelerators. This complex hosts several experiments as illustrated in Fig. 3.1. The

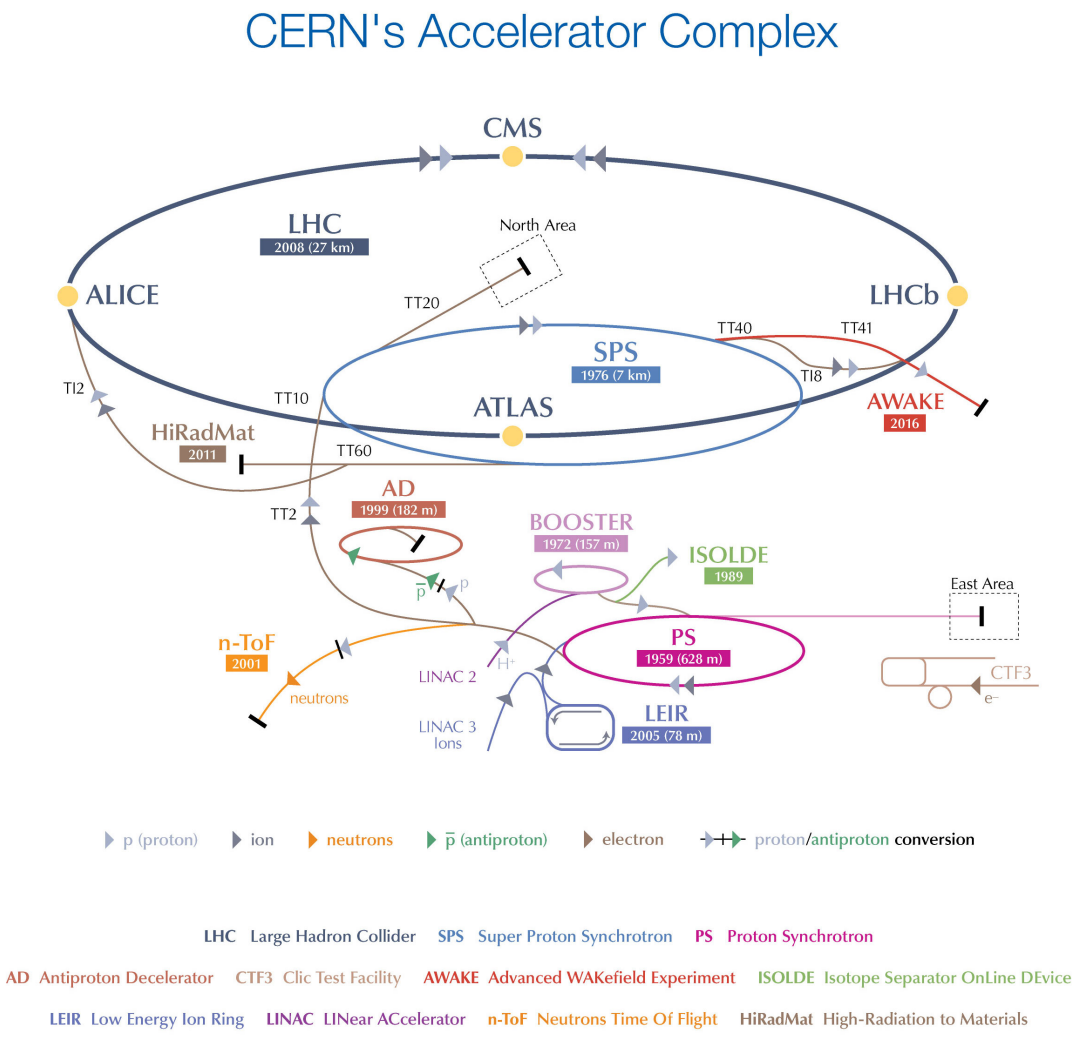

Figure 3.1: Overview of the CERN accelerator complex. Image source [22].

protons begin their journey from a bottle of hydrogen. The hydrogen atoms are striped of their electrons by applying an electromagnetic field. Shortly after, they pass through a linear accelerator (LINAC 2). At this stage, the protons are accelerated,

increasing their energy to 50 MeV. They are injected into the proton synchrotron and accelerated to attain an energy of 25 GeV, and then passed to the super proton synchrotron (SPS) which boosts the energy to 450 GeV. Finally, the beams are transferred to the LHC in clockwise and anti-clockwise directions in separate beam pipes where they are boosted to a maximum energy of 7 TeV. At four points along the LHC ring, the protons are made to collide at the center of particle detectors which are designed to capture the products of a collision and reconstruct fundamental physics phenomena. The experiments located at these four collision points are:

- *ALICE (A Large Ion Collider Experiment):* The ALICE experiment analyzes heavy-ion (lead) collisions in order to study the nature of the quark-gluon plasma.
- *ATLAS (A Toroidal LHC Apparatus):* ATLAS is one of the main general purpose detectors on the LHC and is the sister experiment of CMS. The main design feature is a large toroidal magnet. It is also the largest particle detector ever constructed.
- *CMS (Compact Muon Solenoid):* The CMS detector is the counterpart of the ATLAS detector, and also serves as a general purpose detector. It is heavier despite its smaller volume when compared to ATLAS and features a 3.8 T superconducting solenoid magnet. The CMS experiment is described further in Section 3.2.
- *LHCb (The Large Hadron Collider Beauty):* The LHCb experiment specializes in studying physics involving the b quark in order to investigate matter-antimatter asymmetry in the universe. The design is uniquely asymmetric and focuses on the detection of forward particles.

3.1.2 The Design of the LHC

The technical specifications of the LHC design [23] incorporate updated and novel technologies in order to meet the goal of $\sqrt{s} = 14$ TeV pp collisions, and an instantaneous luminosity of approximately 10^{34} cm^{-2} s^{-1}. The main components of the LHC are the superconducting magnets and the radiofrequency cavities.

The bending of the beam is accomplished by massive superconducting dipole magnets made of niobium-titanium coils, and cooled to 1.9 K. In this superconducting state, the LHC magnets generate powerful 8.4 T magnetic fields which are necessary to steer the beam. Due to space limitations within the LHC tunnel, the dipole magnets utilize a so-called twin-bore design which surrounds both the anti-clockwise and clockwise beam pipes as illustrated in Fig. 3.2. Additionally, there are quadrupole magnets and various multipole magnets which squeeze, correct, and direct the beam towards the interaction points. An astonishing 120 tonnes of liquid helium are required to cool the 1232 dipole and 392 quadrupole magnets along the 27 km long tunnel.

As protons travel through the LHC, they pass through a series of radiofrequency (RF) cavities. The LHC RF cavities are constructed from niobium and copper, and cooled to a superconducting temperature of 4.5 K. An electromagnetic field is applied over the RF cavities, designed to resonate the EM waves, and oscillates at a precise frequency to accelerate the charged proton bunches. After several turns along the LHC, the protons will reach their desired energy.

At capacity, the LHC will hold 2808 proton bunches with each containing approximately 10^{11} protons. Each proton bunch is spaced out by 25 ns which means the collision frequency is 40 MHz. The rate of proton collisions at the LHC and across

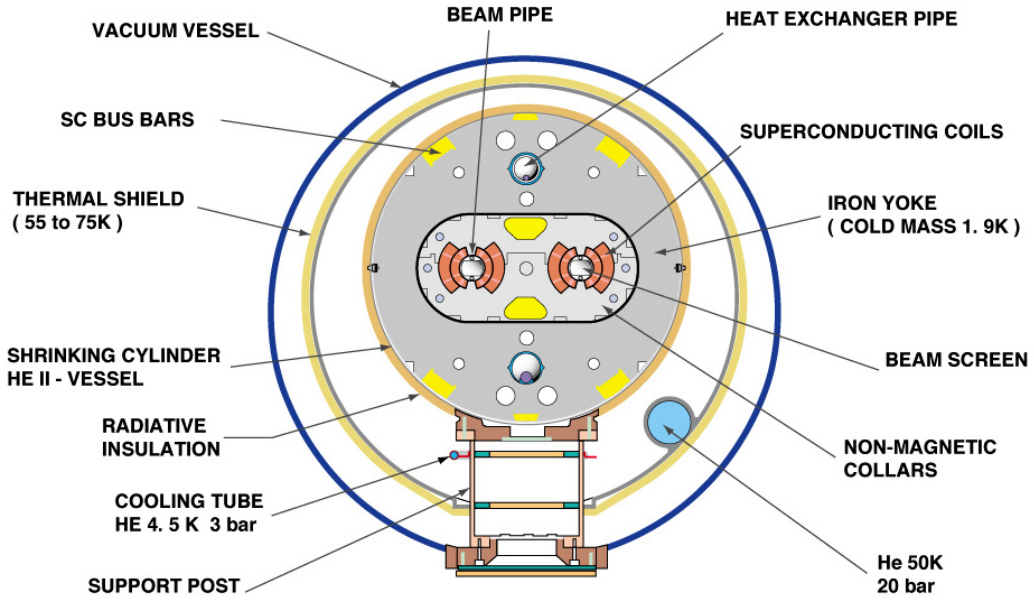

Figure 3.2: Cross section of an LHC dipole magnet. Image source [24].

many high-energy experiments is described by the instantaneous luminosity $\mathcal{L}$ [25]

$$\mathcal{L} = \frac{N_b^2 f n_b}{4\pi} F, \tag{3.1}$$

where N_b is the number of protons per bunch, n_b is the number of proton bunches, f is the revolution frequency of the beam, and F is a geometrical factor (with units cm^{-2}) which is a function of the transverse trajectory and the crossing angle of the collision. Given a particular process α, the number of events expected over a time interval is formulated in terms of the luminosity

$$\frac{dN_\alpha}{dt} = \mathcal{L}\sigma_\alpha, \tag{3.2}$$

where σ_α is the cross section of the process. High-energy physicists often express the quantity of events collected during a data-taking run as the integrated luminosity $\mathcal{L}_{\text{int}} = \int \mathcal{L}\ dt$ with units of inverse femtobarns fb^{-1}. Figure 3.3 summarizes the cumulative integrated luminosity delivered to the CMS detector for several LHC run

periods. From these terms, the total number of events expected at the end of a data-taking period is written as

$$N_\alpha = \mathcal{L}_{\text{int}} \sigma_\alpha. \tag{3.3}$$

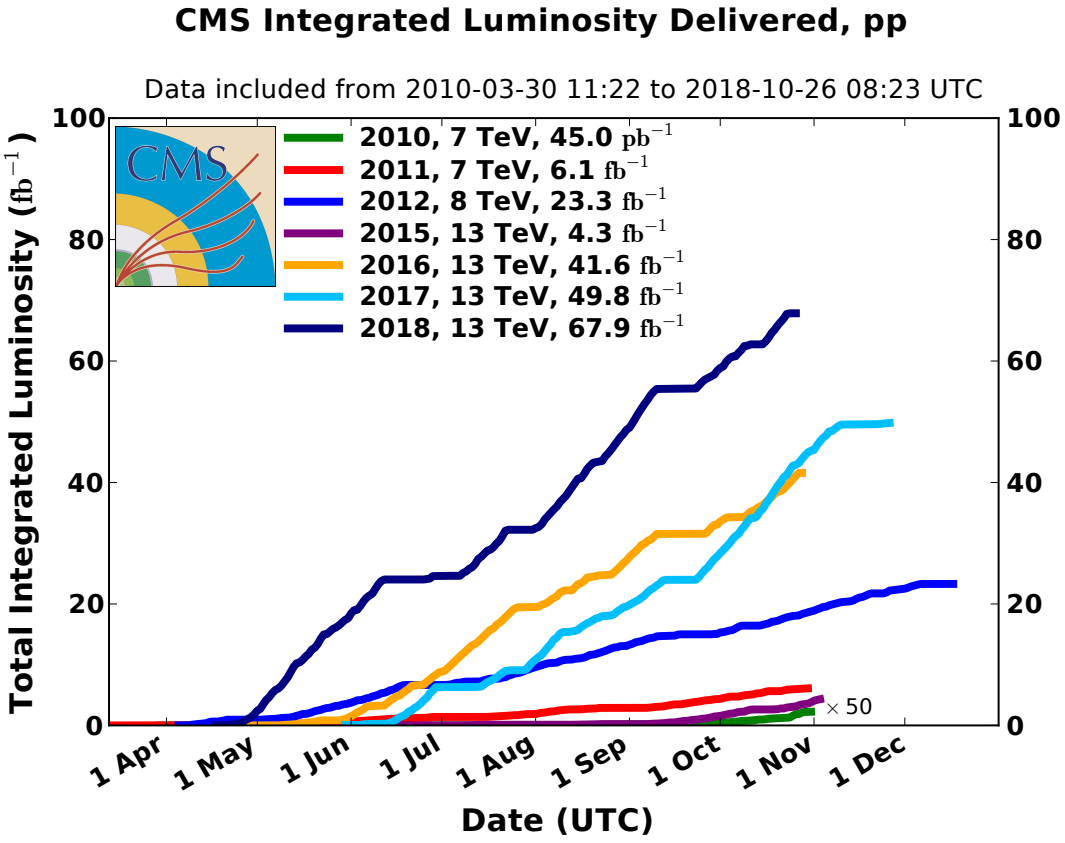

Figure 3.3: Cumulative integrated luminosity delivered to the CMS detector for several different LHC run periods. Figure source [26].

During a proton-proton bunch crossing, multiple proton on proton collisions can occur. These additional collisions other than the one of our main interest are referred to as in-time pileup (PU). A similar phenomena called out-of-time pileup refers to the additional proton-proton collisions which occur just before and after the collision of interest. The mean number of pileup interactions can be tuned by altering

the bunch spacing, the number of protons in a bunch, and the intensity which the proton bunches are squeezed prior to the collision. Pileup may seem undesirable as it can obscure interesting physics of the main pp collision, however the benefit is an increased chance of capturing an interesting event per bunch crossing. During the initial data-taking runs at the LHC, the mean number of pileup interactions was relatively low. However as the luminosity and collision energy increased, the pileup also increased, as illustrated in Fig. 3.4. High energy physicists are able to handle increasing levels of pileup through better calibrated electronics and more efficient pileup-suppression algorithms.

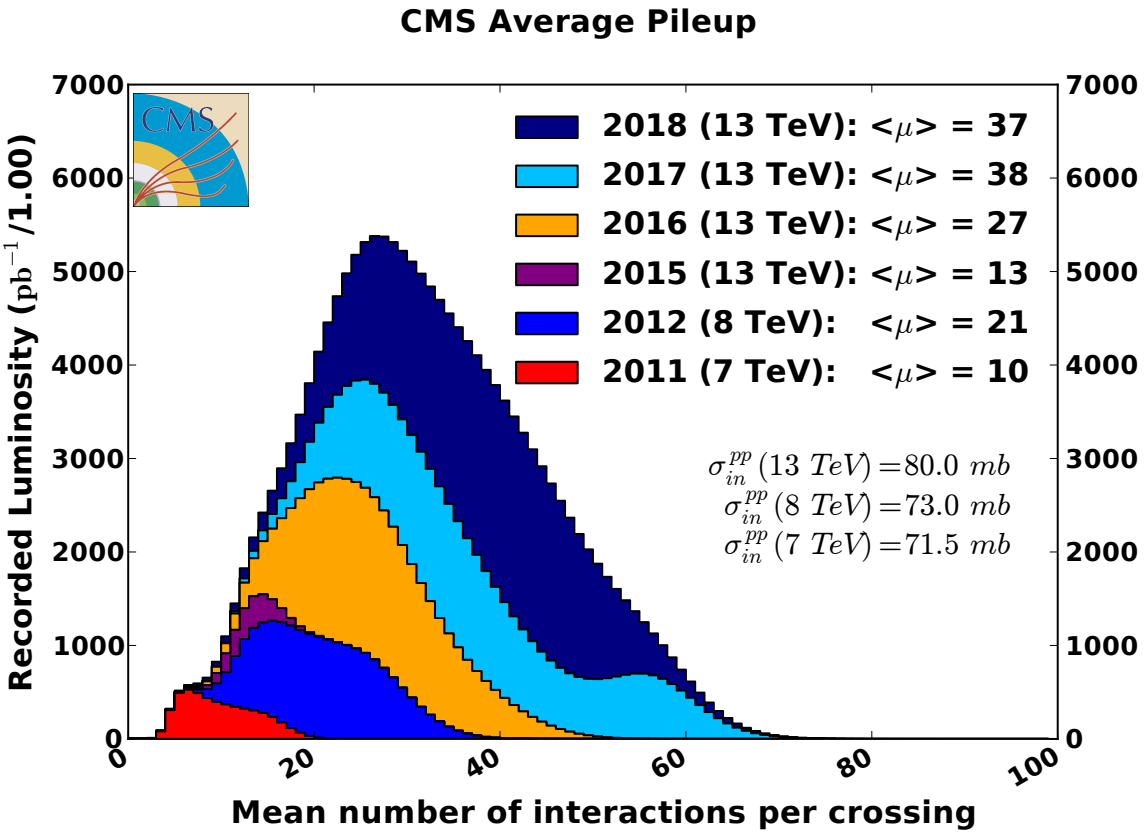

Figure 3.4: Distribution of the average number of interactions per pp collision for various data-taking years. The mean pileup per year is also displayed. Figure source [26].

The conditions for the data analyzed in this work correspond to those of the 2016, 2017, and 2018 LHC operational periods. This data-taking time frame is collectively referred to as "Run 2". The average number of pileup interactions ranged from 27 to 38, and pp collisions of about 159 fb^{-1} of integrated luminosity were delivered to the CMS detector at a center-of-mass energy of 13 TeV.

3.2 The Compact Muon Solenoid Experiment

The Compact Muon Solenoid (CMS) [27–29] detector is one of the two general purpose detectors at the LHC. It is located on the opposite side of the LHC ring from the ATLAS detector, about 100 m underground, at interaction point 5 (P5). In terms of local geographical landmarks, P5 is situated between the Jura mountains and Lac Léman (Lake Geneva), just outside of the small French township of Cessy. The CMS detector is massive, yet dense, hence the usage of "compact" in its name. It weighs approximately $14,000$ tonnes, and measures 28.7 m in length and 15.0 m in diameter. The CMS detector is cylindrical in shape, with concentric subsystems about the interaction point as depicted in Fig. 3.5. The subdetectors are geometrically divided into a barrel component and two endcap components. Multiple subsystems are specially designed to detect certain particles produced in pp collisions. The innermost subsystem is the silicon tracker, followed by the electromagnetic calorimeter (ECAL) and the hadron calorimeter (HCAL). The main feature of the CMS detector is the superconducting solenoid which generates a powerful 3.8 T magnetic field and surrounds the tracker, ECAL, and HCAL. The magnetic field bends the trajectory of charged particles traveling through the detector and is crucial for precisely measuring their momenta. The muon detectors are embedded in a iron return yoke outside the

solenoid. Each of these subsystems and their components are discussed in detail in the following sections.

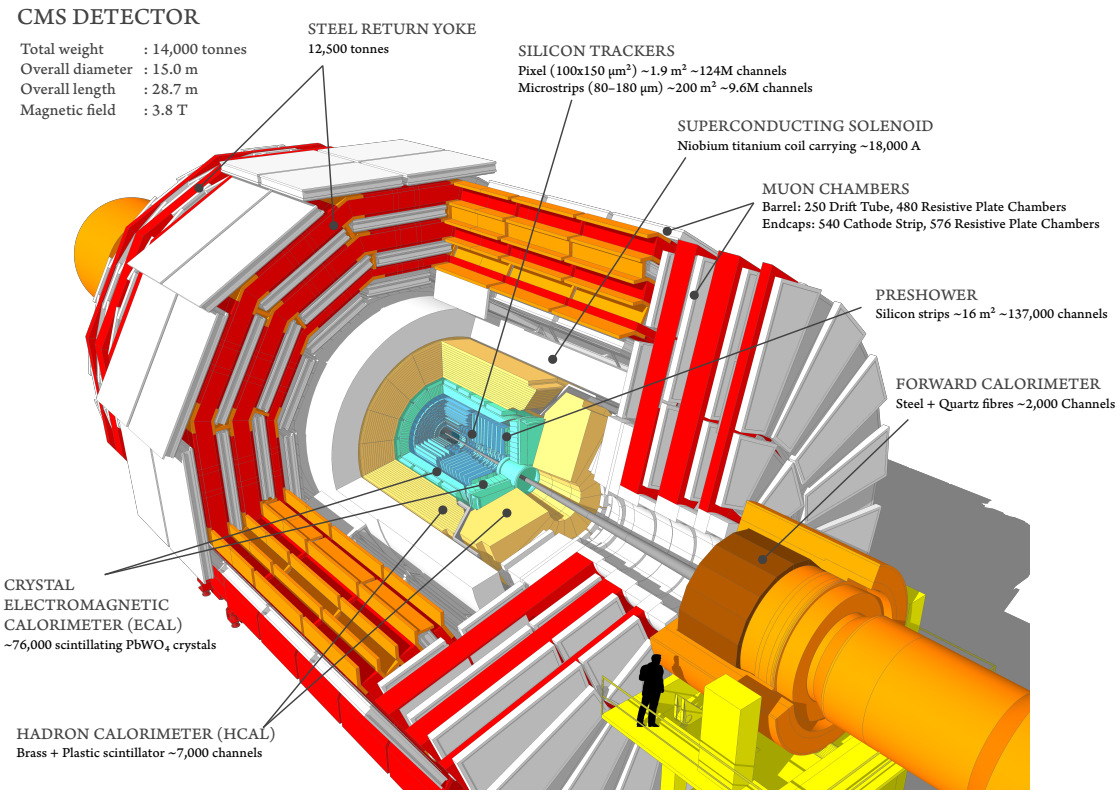

Figure 3.5: An overview of the CMS detector. The subsystems are labeled in the diagram, and the silhouette of a person is added for scale. Figure source [30].

3.2.1 The CMS Coordinate System

The coordinate system adopted by the CMS Collaboration defines the nominal collision point as the origin. The x-axis points towards the center of the LHC ring, the y-axis points vertically upward towards the surface, and the z-axis points towards the Jura mountains along the counterclockwise beam direction. Due to the cylindrical shape of the CMS detector, positions are given in terms of the azimuthal angle ϕ, the polar angle θ, and the radial distance r. The rapidity y (not to be confused with the

y-axis) is a useful quantity in hadron collider physics. This is defined as:

$$y = \frac{1}{2} \ln \left(\frac{E + p_z}{E - p_z} \right), \tag{3.4}$$

where E is the energy of the particle and p_z is the beam-direction component of the momentum. When a Lorentz boost is applied along the beam-direction, differences in rapidity Δy are Lorentz invariant. However, the quantities E and p_z can be difficult to precisely measure especially when the particle's mass is not known. Additionally, the z-components of the momentum of the interacting partons are unknown since they individually carry a varying fraction of the proton's total momentum. For these reasons, the rapidity is approximated as the pseudorapidity assuming $E \gg m$ or $p_{\mathrm{T}} \gg m$. Compared to the rapidity, pseudorapidity is more easily quantified. The pseudorapidity η is a function of θ and is written as:

$$\eta = -\ln \left(\tan \left(\frac{\theta}{2} \right) \right). \tag{3.5}$$

The angular separation ΔR between particles is written in terms of the azimuthal angle and the pseudorapidity:

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}. \tag{3.6}$$

Lastly, the collision products' momenta are often measured in the transverse direction (the x-y plane), denoted as p_{T}. This is due to the fact that the colliding partons p_z is unknown; however, their initial momentum in the x-y plane is nearly zero

3.2.2 The Solenoid Magnet

The CMS magnet enables the precise measurement of charged particles' momenta as well as the sign of their electric charge. The 3.8 T magnetic field generated

by the CMS solenoid forces muons, electrons, and charged hadrons to move in a helical motion which is tracked to determine the particle's momentum. As depicted in Fig. 3.6, the outside solenoid is surrounded by three layers of an iron return yoke. This provides integral support to the whole CMS detector, and also has the effect of confining the magnetic field lines to create a homogeneous field within the solenoid. During operation, the solenoid is cooled down to about 4.6 K which brings the $12,000$ tonnes magnet to a superconducting state. A 19.5 kA current through the solenoid, composed of $2,168$ turns of high-purity aluminium wire, generates the 3.8 T magnetic field. The magnet has a inner diameter of 5.9 m and a length of 12.9 m, which is enough to enclose the barrel region of the inner-tracker, ECAL, and HCAL subsystems. The solenoid enables a momentum resolution of $\Delta p/p \approx 10\%$ for particles with 1 TeV of momentum, measured by the tracker.

3.2.3 The Inner Tracker

The subdetector which is closest to the interaction point is the inner tracker detector. The purpose of the tracker is to determine the trajectories and momenta of charged particles with $p_{\mathrm{T}} > 1$ GeV as they come under the influence of the solenoid's magnetic field and move through the innermost part of the CMS detector. This is accomplished by detecting charged particles and connecting the hits assuming a helical trajectory to form a track. Particles can be traced back to an originating vertex, whether it be the primary interaction vertex, a pileup interaction vertex, or a secondary vertex. Secondary vertices are indicative of b quark decays and their identification is very important in the data analysis discussed in Chapter Five.

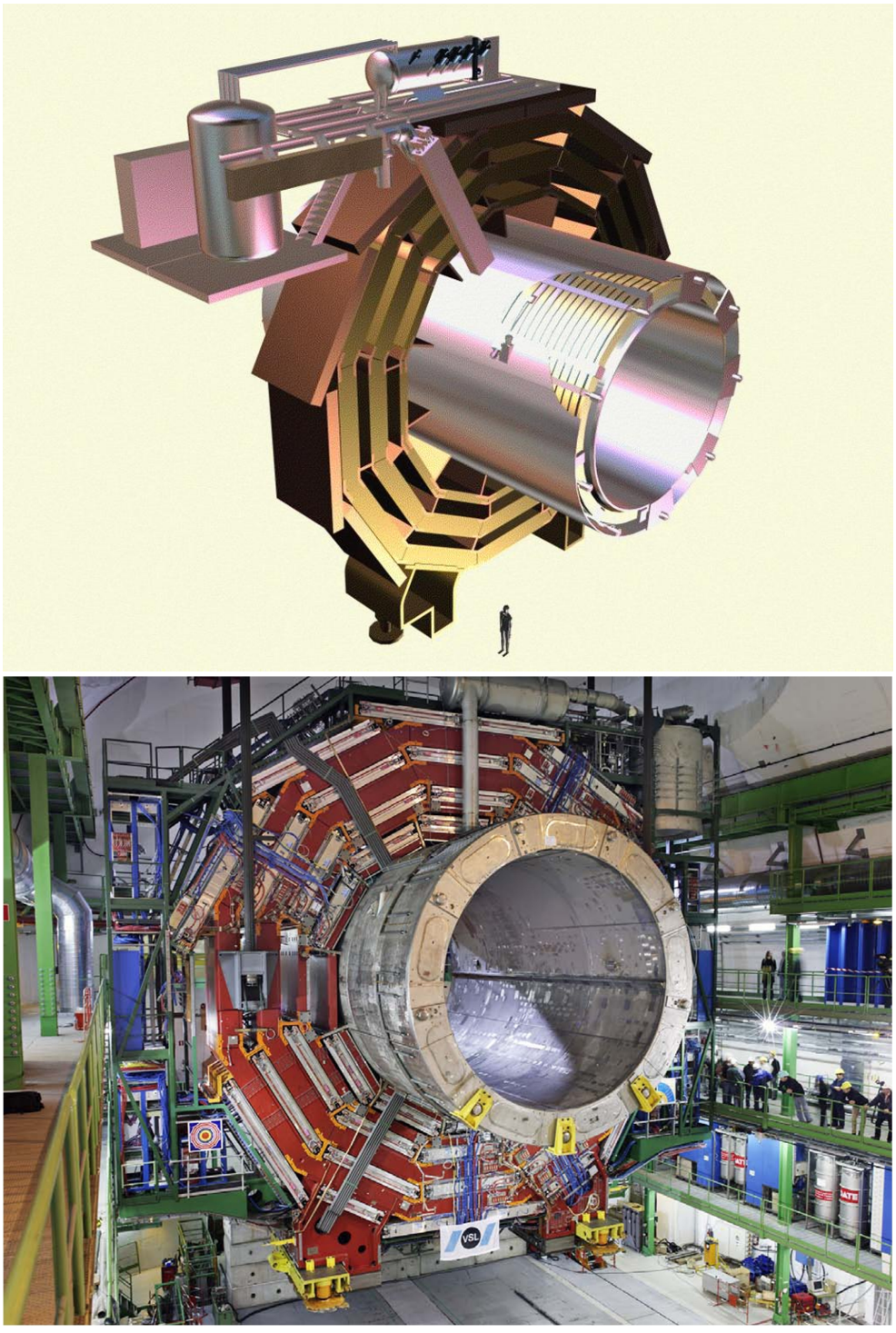

Figure 3.6: (Top) Artistic view of the CMS solenoid, cryostat, and support structures. (Bottom) Picture of the solenoid taken during the assembly of the CMS detector. Image source [31].

As displayed in Fig. 3.7, the tracker is made up of concentric layers of silicon modules. In the endcap region, the layers are arranged to be transverse to the beam pipe, while the layers in the barrel are parallel to the beam. The use of silicon in the tracker is beneficial for a couple of reasons. Silicon is resistant to radiation which is prevalent in this region of the CMS detector when the LHC is delivering particle collisions. Secondly, the thin silicon modules interact minimally, reducing the energy loss of a charged particle as it travels through the tracker.

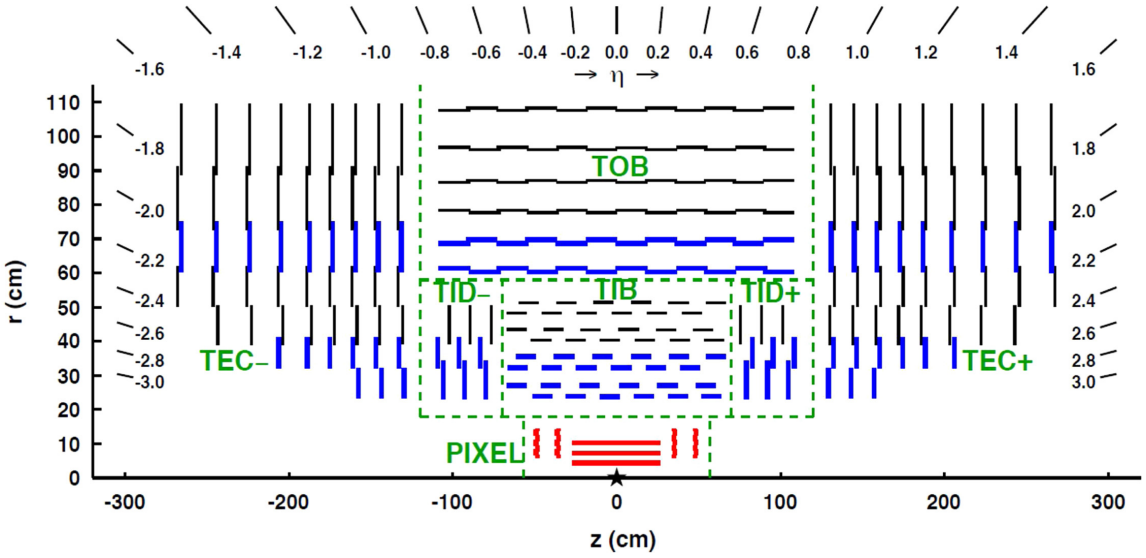

Figure 3.7: Schematic of the original tracker system. The pixel detector is closest to the interaction point, followed by the inner and outer silicon strip tracker. Figure source [32].

The part of the tracker which is closest to the beam pipe, the pixel detector, is made up of tiny silicon modules with a pixel size of 100×150 μm^2. Due to the pixel detector's proximity to the beam pipe, it is the most critical component of the inner tracker in determining precise track parameters. However, this also means that the

pixel detector experiences the largest flux of particles, and thus the most radiation damage out of all the CMS subsystems.

During a break in LHC operation between the 2016 and 2017 data-taking periods, the pixel detector was deemed insufficient for future LHC conditions and was completely replaced with an upgraded pixel detector. This upgrade is part of larger initiative to replace several CMS components known as the "Phase 1" upgrades. In addition to replacing modules with radiation damage, there are a few improvements with the new pixel detector which are worth mentioning. As shown in Fig 3.8, the new pixel detector has more silicon modules, $1,440$ in total, and layers compared to the original. In the barrel and endcap regions, there is an additional layer and both are placed closer to the interaction point. These changes improve momentum resolution and identification of vertices.

The silicon strip detector is the outermost part of the inner tracker and surrounds the pixel layers. This part of the tracker is made up of silicon modules with a larger surface area and thickness. The inner barrel tracker (TIB) has four layers of modules with a strip pitch varying from 80 μm to 205 μm and thickness 320 μm, and the outer barrel tracker (TOB) has six layers with the same surface area but a larger thickness of 500 μm. In the endcap, the inner disc tracker (TID) has three layers of silicon strips arranged on a disc followed by the endcap tracker (TEC) which has nine more layers of discs. In total, the strip detector is made up of $15,000$ modules.

3.2.4 The Electromagnetic Calorimeter

The next CMS subsystem, in terms of proximity to the interaction point, is the electromagnetic calorimeter (ECAL) [34]. The main purpose of the ECAL is

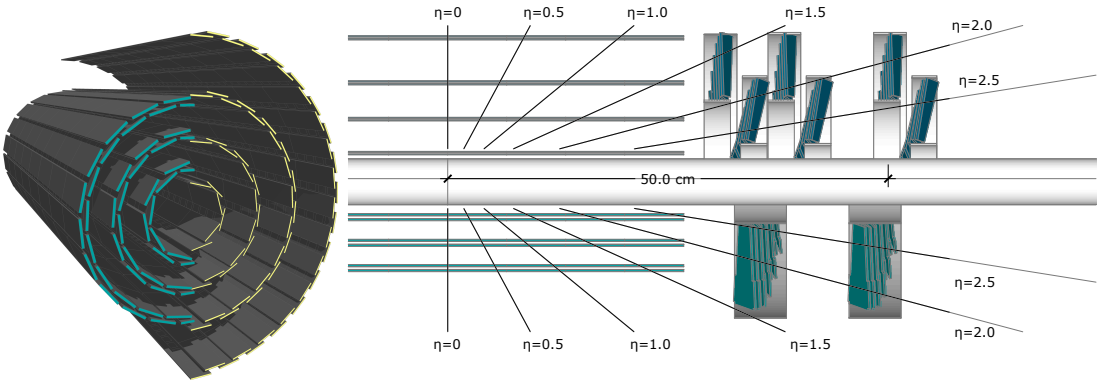

Figure 3.8: In 2017, the pixel detector was upgraded to increase tracker performance. (Left) Side-by-side comparison between the 3-dimensional geometry of the legacy and upgraded barrel pixel detector, on the left and right respectively. (Right) The barrel portion (BPIX) was upgraded to have four layers, up from three. The endcap region (FPIX) now has three layers, up from two. Figure source [33].

to measure the energy deposits of electromagnetic showers originating from photons and electrons. Instead of passively detecting a particle like the inner tracker, the ECAL will fully absorb the particle's energy. The ECAL is homogeneous and is primarily comprised of lead tungstate ($PbWO_4$) crystals. In the barrel region of ECAL (EB), there are $61,200$ crystals, and $7,324$ crystals in the each ECAL endcap (EE). The EB crystals have a front-face cross section of 22×22 mm^2 and a length of 230 mm, providing coverage up to $|\eta| < 1.479$. The crystals in the EE extend the pseudorapidity coverage $1.479 < |\eta| < 3.0$, and have a front-face cross section of 28.6×28.6 mm^2 and a length of 220 mm. Images of the lead tungstate crystals and a diagram of the ECAL layout are shown in Figs 3.9 and 3.10. In front of each ECAL endcap, there is a preshower (ES) detector which has two layers of lead absorbers interleaved with silicon strip detectors. The ES helps distinguish between two closely-spaced photons originating from neutral pions and high-energy photons from the

hard scattering process. Additionally, the preshower detector improves positional resolution of electrons and photons in the region $1.653 < |\eta| < 2.6$.

As electrons and photons travel through the lead layers of the preshower and the crystals, the material causes the particles to shower until all their energy is absorbed. Additionally, the crystals will scintillate, producing blue light as the electromagnetic shower passes through. Photodetectors, located at the base of the crystal, capture this light and amplify the signal based on a calibrated gain. In the ECAL barrel, these are avalanche photodiodes (APDs), while the ECAL endcap has vacuum phototriodes (VPTs) which perform better in the higher particle flux of the endcap.

3.2.5 The Hadron Calorimeter

The hadron calorimeter (HCAL) [38] is located, mostly, in-between the ECAL detector and the solenoid. The main objective of the HCAL detector is to detect neutral and charged hadrons, which pass through the tracker and ECAL without losing a significant fraction of their energy. Similar to the ECAL, the HCAL is constructed of materials designed to fully absorb and thereby measure the energy of incident particles. However unlike the ECAL, the HCAL is heterogeneous having alternating layers of absorber and scintillator. The absorber causes the hadrons to shower, producing a spray of particles. Some portion of the shower will be absorbed and some will pass through into the plastic scintillator. The particles cause the scintillator to produce light proportional to the energy of the shower. The shower passes onto the next layer of absorbing material and so-on until all energy has been fully absorbed.

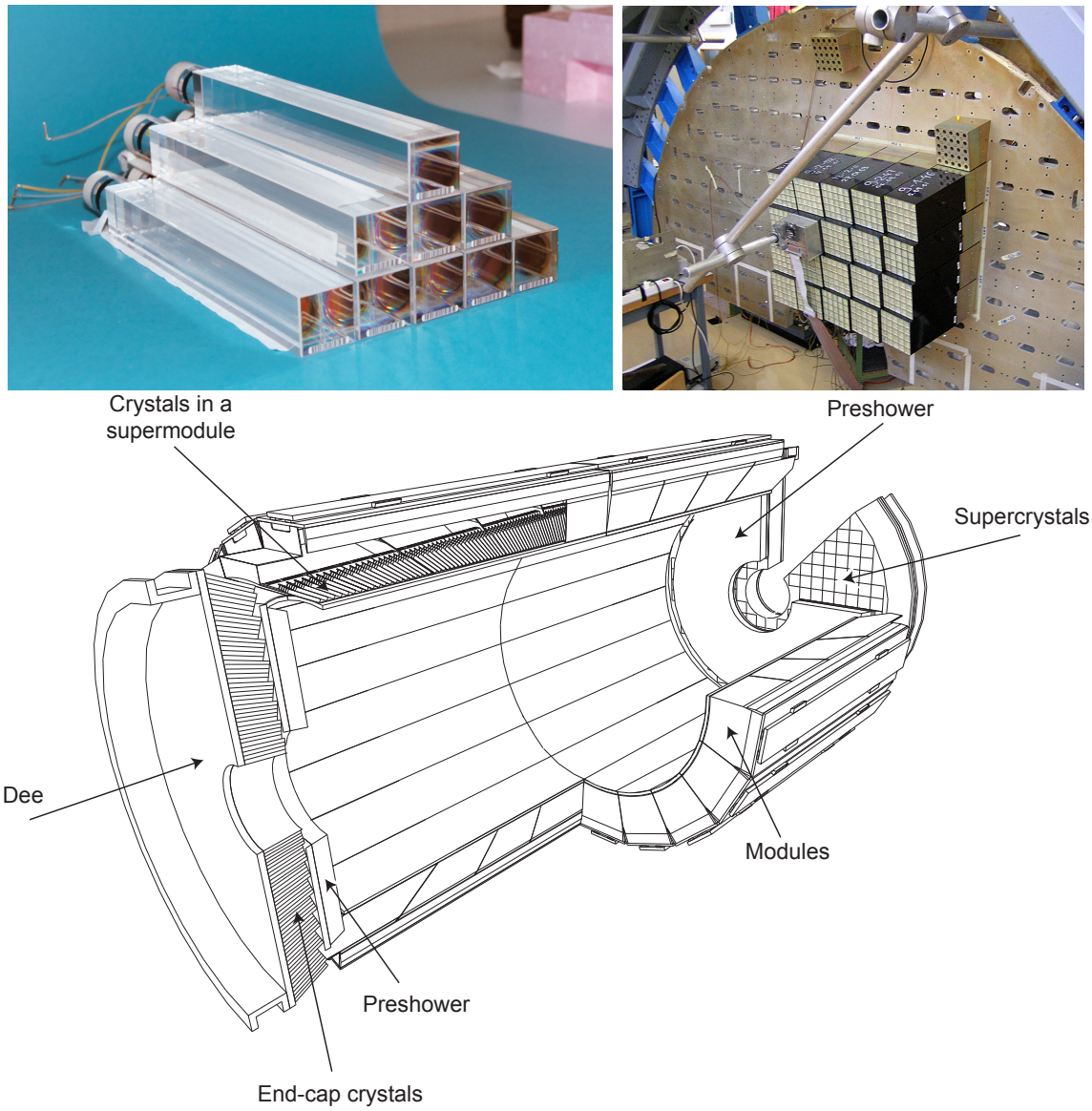

Figure 3.9: (Top) Photographs of the lead-tungstate ECAL crystals before (left) and during (right) their installation. (Bottom) Diagram of the ECAL crystal arrangement within the CMS detector. Image and figure sources [35–37].

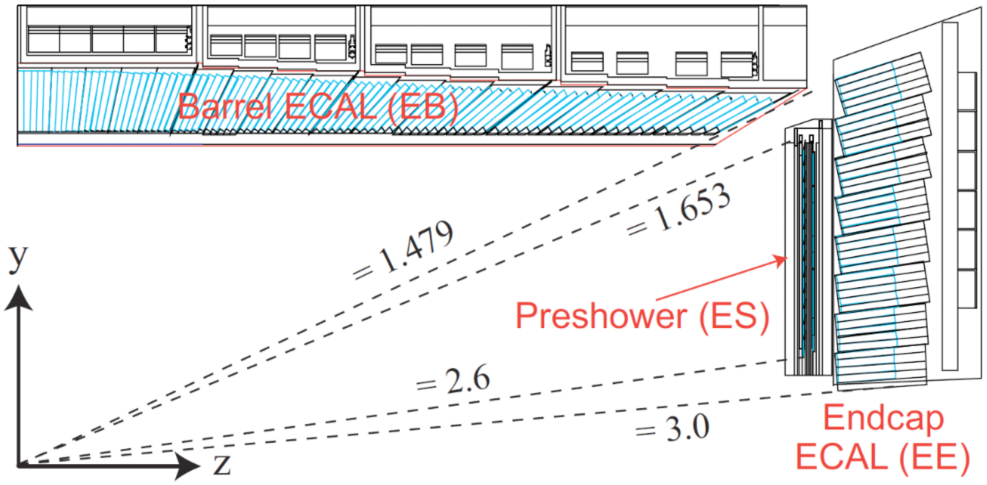

Figure 3.10: Schematic of the η coverage of the CMS ECAL subsystems including the barrel ECAL (EB), endcap ECAL (EE), and the preshower detector (ES). Figure source [28].

The CMS HCAL consists of four subdetectors, namely the barrel (HB), outer barrel (HO), endcap (HE), and forward region (HF). Images of the installation of the HB, HE, and HF subsystems are shown in Fig. 3.11. The HB and HE subdetectors contain brass absorber and plastic scintillator, and come together to form a hermetic seal which maximizes coverage. The design of the HB and HE detectors was motivated by the need to fully capture oncoming hadrons, while being compact enough to fit within the dimensions of the solenoid. The HCAL outer barrel detector is positioned just outside of the solenoid. It consists of layers of scintillator and is designed to detect and stop hadrons that occasionally leave the HB and punch through the magnet coil. For HB, HE, and HO, the mechanism to readout the signal is the same, i.e. the plastic scintillator scintillates blue light which is captured and directed by optical fibers that shift the wavelength to green light. Green light is routed to hybrid photodiodes (HPDs) located within readout modules which ultimately convert the optical signal to a digital signal.

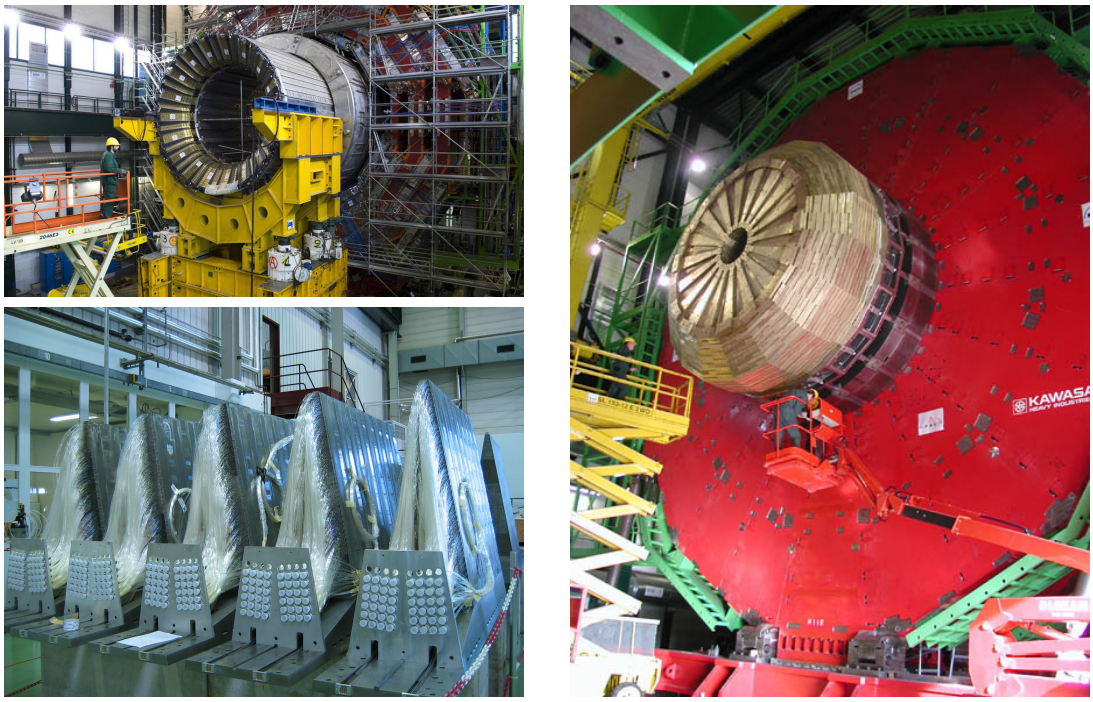

Figure 3.11: Photographs of various stages of the installation of the HCAL barrel (top-left), HCAL endcap (right), and HCAL forward (bottom-left) onto the CMS detector. Image sources [39–41].

The last component of HCAL, HF, resides in the forward area of the CMS detector, $|\eta| > 3.0$. The HF is unique compared to the other HCAL subsystems in terms of its material make up. It has a large block of steel absorber to resist the large amount of radiation, and over 1000 km of quartz crystal fiber embedded in the steel. As particles travel through HF, Cherenkov light is produced in the fibers. The HF is able to distinguish between shallow showers created by electrons and photons versus hadron showers, using two different length quartz fibers. The optical signal passes through the fibers into photomultiplier tubes (PMTs) which convert the light to an electrical signal.

During the Run 2 period of LHC operation, the HCAL subsystem was a beneficiary of the Phase 1 upgrade initiative [42]. The main purpose of this upgrade was

to replace antiquated readout electronics with electronics capable of operating in the high-pileup conditions of future LHC data-taking periods. One crucial component was the replacement of the HPDs, which had incurred radiation damage, with radiation resistant silicon photomultipliers (SiPMs). Additionally, the new electronics increased depth segmentation information for the HE and HB signal readout as depicted in Fig. 3.12. The HO detector was upgraded first prior to the 2016 data-taking period. After LHC operation in 2016 concluded, the upgrades to the HF detector followed. Next, the HE detector Phase 1 upgrade was completed prior to the 2018 data-taking period. Lastly, the HB upgrades took place after Run 2.

3.2.6 The Muon Detector

Unlike the particles discussed previously, the properties of the muon enable it to travel through several meters of material with little to no interaction. Thus, the CMS muon detectors [43] are located outside of the solenoid where muons are the only expected detectable particles. The muon subsystem consists of several layers of detectors embedded in the return yoke. The placement of these systems is designed to complement the inner tracker detector and be able to extend the detection of muon tracks to the boundaries of the CMS detector. There are three muon detector types utilized in the CMS detector. As shown in Fig. 3.13, the drift tubes (DTs) are located in the barrel region and cathode strip chambers (CSCs) are positioned in the endcap. Additionally, resistive plate chambers (RPCs) supplement both detector technologies and are located in the barrel and endcap regions.

The DT chambers provide muon detection coverage in the barrel region $|\eta| < 1.2$. There are about 250 DTs spread out over four concentric layers running parallel to

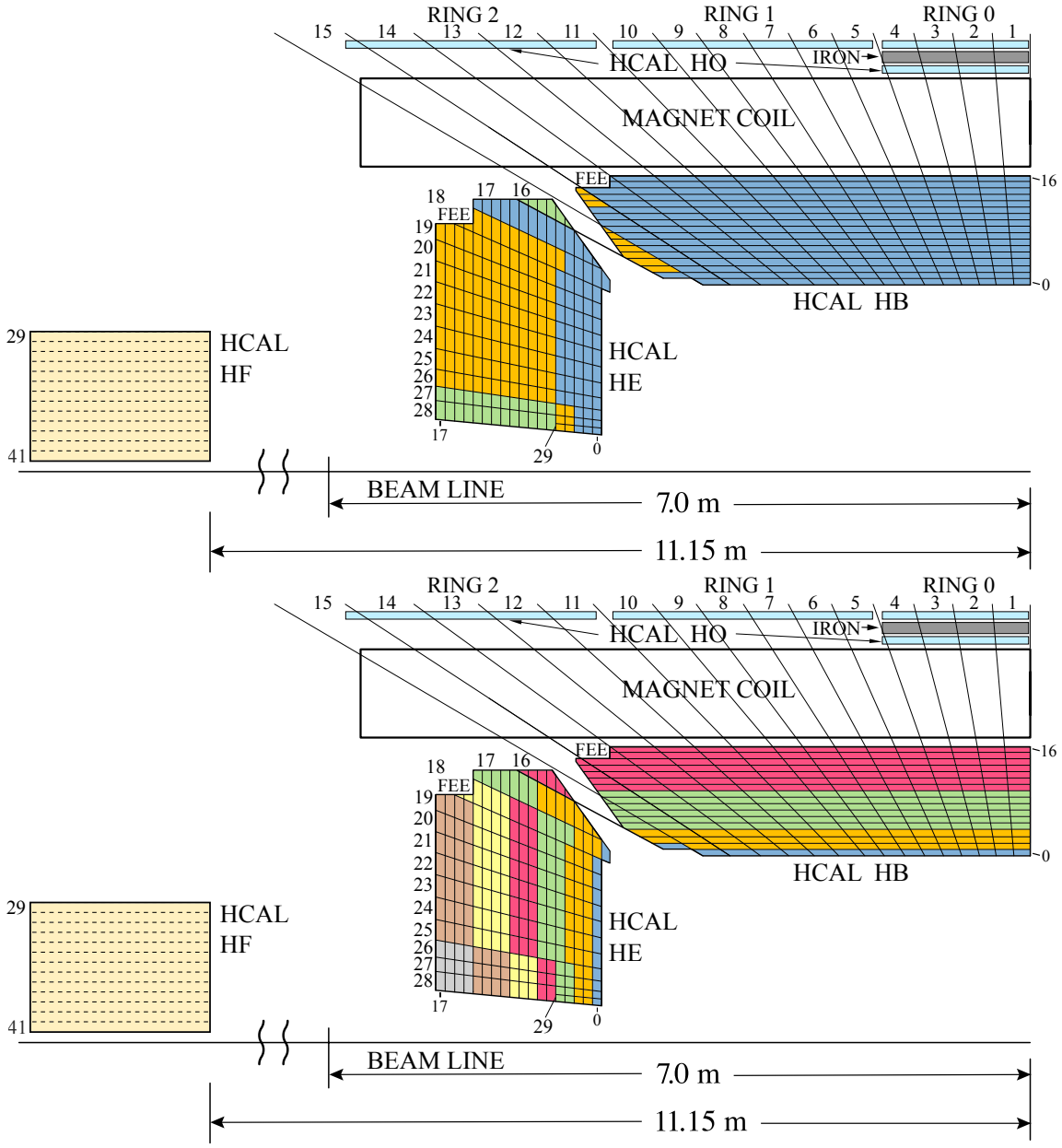

Figure 3.12: Diagram of the depth segmentation of the legacy HCAL (top) and the Phase 1 HCAL (bottom). The different colors represent the distinguishable depth information readout from the HB and HE. Figure source [42].

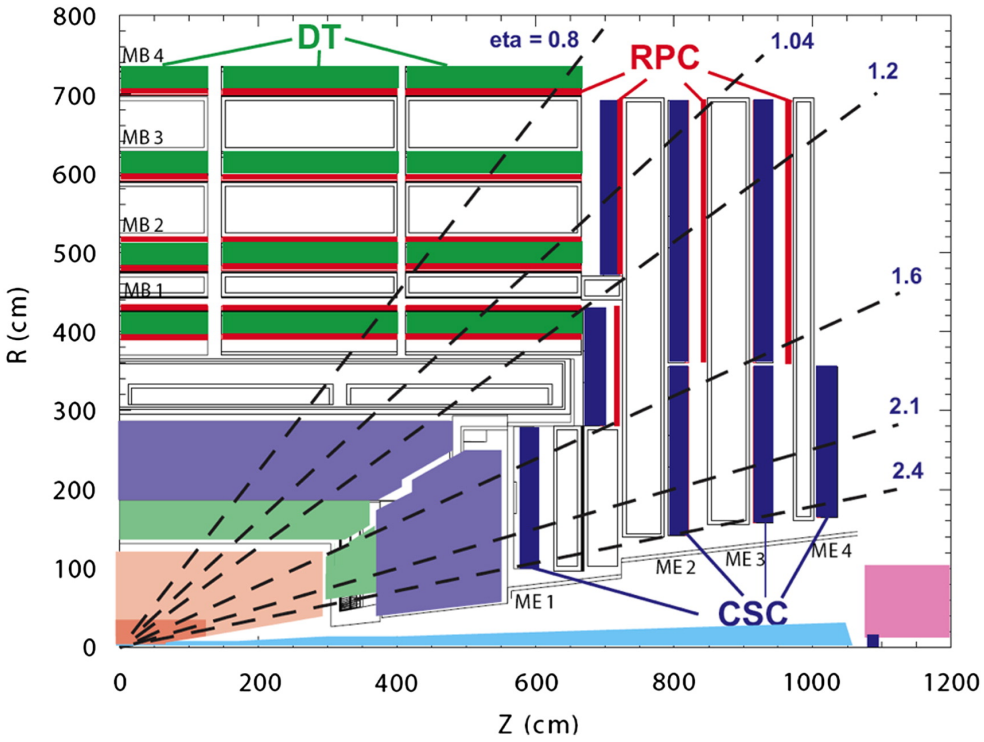

Figure 3.13: An illustration depicting the layout and position of the CMS muon DT, CSC, and RPC subdetectors. Figure source [28].

the beam pipe. The tubes which compose the DTs have a cross section of 1.3×4.2 cm^2 and a length ranging from 2 to 2.5 m. Each DT contains many cells which are filled with an ionizable gas and have a positively-charged wire running through the center. When a muon passes through a cell, it ionizes the gas causing free electrons to drift towards the wire. The electron coming into contact with the wire generates a readout signal which contains information on the position of the detection.

The CSC modules are located in the endcap region of the CMS detector and provide detection coverage in $0.9 < |\eta| < 2.4$. They are filled with gas and contain anode wires oriented perpendicular to the cathode strips. Instead of DTs, the utilization of CSCs was based on their property of being radiation hard which is relevant in

the high-particle flux environment of the endcap. In total, there are 468 CSC modules located in the CMS endcap in 4 layers, oriented perpendicular to the beam pipe.

The barrel and endcap have 6 and 3 layers of RPCs, respectively. The RPCs provide supplemental timing information to the positional readout of the DT and CSC modules. They utilize two parallel plates with high resistivity with a strong voltage applied across the plates. A signal is produced when a muon passes through the plates causing an avalanche of free electrons and thus a measurable current. This type of technology has a fast readout and provides excellent timing information of the order of 1 ns making it possible to precisely assign muons with bunch crossings.

3.2.7 The Trigger System and the Worldwide Computing Grid

The LHC produces collisions at a rate of 40 MHz, the vast majority of which are dominated by uninteresting QCD processes. In fact, the expected rate of interesting physics such as the Higgs boson production is of the order of 1 Hz or one Higgs boson produced per second. While experimentalists would like to store and analyze every pp collision, there is simply not enough storage space or readout bandwidth to do so. Therefore, the CMS Collaboration created a two-tier trigger system [44, 45] to identify and store any event that may contain interesting physics.

The first tier is the hardware-based Level-1 (L1) trigger. The L1 trigger uses computationally fast, custom hardware designed to make a "keep" or "discard" decision every 3.2 μs, with an output rate of keep decisions below 100 kHz. The hardware is located as close as possible to the CMS detector in the service cavern to reduce latency. In order to comply with this time constraint, the L1 trigger is only able to view primitive objects, quickly constructed from a subset of the readout from the

muon systems, the ECAL, and the HCAL. During this time, the event's full raw detector readout is stored in memory buffers. As illustrated in Fig. 3.14, the L1 trigger combines the various subdetector primitives into a global trigger which offers a final decision to allow an event to be further analyzed or to be discarded.

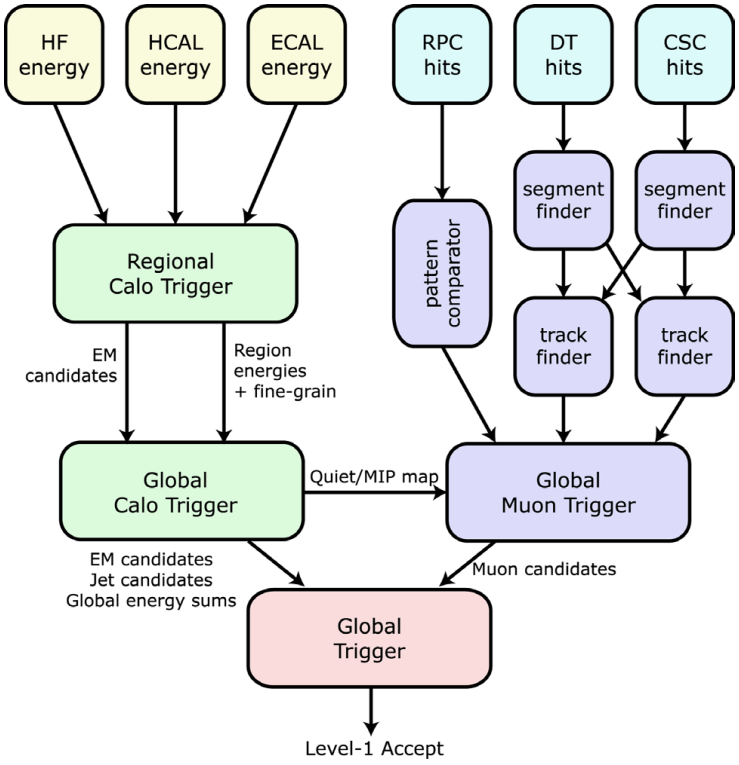

Figure 3.14: Information flow chart of the L1 trigger system. The global trigger is constructed using limited information from the calorimeters and muon subsystems. Image source [46].

The high level trigger (HLT) is the second phase of the CMS trigger system. Events which pass the L1 trigger are sent to a computer farm located at P5 nearby the CMS control room. At this stage, the full event information is available for a nearly offline-equivalent reconstruction, after which the events are categorized based on a predetermined set of HLT paths. The HLT paths were designed by the CMS

Collaboration to accommodate a wide range of potential data analyses. Events which successfully pass any one of these trigger paths are saved to storage, however around 99% of events fail to pass any of the trigger paths and are subsequently discarded. This corresponds to about 1000 events saved every second, or 1000 Hz.

Events that pass the L1 trigger and HLT are stored and undergo full event reconstruction (discussed in Chapter Four). Additionally, most data analyses typically need a large amount of simulated data in order to build and test theoretical models. This requires an immense amount of processing power and storage capacity. For the needs of CMS and other CERN projects, the Worldwide LHC Computing Grid [47] was established to connect computing centers from all around the world. The grid enables users to access data securely and utilize the computing resources of participating institutions and universities.

CHAPTER FOUR

Event Simulation and Reconstruction

Before a physics analysis can take place, the products of the proton-proton collision are reconstructed from the readout of detector electronics. The reconstruction process is accomplished by a series of algorithms built from empirical and theoretical principles. At the outset, the particle flow algorithm pieces together information from various CMS subdetectors to form a physical representation of the collision event. Initially, the event description consists of electrons, photons, muons, neutral hadrons and charged hadrons, and their respective positions and momenta. This information is refined using a variety of cut-based or machine learning algorithms into more analyzable physics objects such as jets, heavy-flavor jets, missing transverse momentum, and others. During each stage of reconstruction, various forms of noise suppression, corrections, and event cleaning are also applied. Additionally, kinematic criteria are imposed on the physics objects to reduced the likelihood of background contamination. After the collision events have been fully reconstructed, a physics analysis of those events may take place.

The model of any CMS physics analysis is developed from a collection of simulated proton-proton collisions. The simulation is created by a suite of software which synthesizes the event description in a sequential way. First, the hard scattering process is simulated, followed by the immediate decay products of the collision. The software also simulates the chain of decay into relatively long-lived particles and their

interactions with the CMS detector. At this stage, the data comprises simulated detector readout. Just like real collision events, the simulated events are reconstructed. However at the end of this procedure, there are systematic differences between the two. To rectify this, corrections are applied to enhance the agreement between the simulation and real data.

This chapter will provide a more detailed description of the steps involved in simulating collision events. Additionally, the algorithms utilized by the CMS Collaboration to reconstruct both real and simulated events are covered. Finally, this chapter provides definitions of physics objects relevant to the work of this thesis.

4.1 Event Simulation

Producing accurate simulations of the proton-proton collisions in the CMS detector is essential for testing theoretical predictions. Given the complexity of these collisions and the particle interactions, events are simulated using Monte Carlo (MC) sampling techniques [48]. A diagram of a typical proton-proton collision including all of the individual components of an event is shown in Fig. 4.1. The following subsections will provide a brief explanation of how the simulated MC events are produced.

- *Hard Scattering:* In a proton-proton collision, two partons (quarks or gluons) interact with each other. The probability of these interactions as a function of interaction energy depends on the parton density functions (PDFs). From the type of partons, the matrix elements are calculated to determine the cross section of the process and to produce simulated events.

- *Parton Shower:* Additional gluons, quarks, or photons may radiate off of partons that go into or come out of the hard scattering in simulated events. These are called initial state radiation (ISR) and final state radiation (FSR), respectively. The parton shower (PS) is responsible for providing a description of ISR and FSR up until a certain energy scale has been reached.
- *Hadronization:* Due to color confinement, sets of colored partons hadronize to become mesons and baryons. At the cut-off energy scale, perturbative QCD diverges so this process is described by phenomenological models. These stable particles eventually interact with the detector material.
- *Underlying Event:* Partons that are not part of the hard interaction may still undergo soft scattering. These interactions result in the underlying event (UE) which is modeled separately.
- *Detector Simulation:* This step simulates the interaction between stable particles and the detector. This includes the full CMS detector description, and how certain particles interact with different detector materials.

4.1.1 The Hard Scattering of Partons

The first step in event generation is simulating the hard scattering of the proton-proton collisions and its resulting products. This requires a physical description of the proton. Protons are composed of two up quarks and one down quark. However, this description of the proton is too simplistic and is only relevant in describing the valence quarks of the proton. In fact, a proton's internal structure is more complicated as it is also made up of gluons and quarks which are constantly splitting and annihilating. The probability of probing a proton at a certain energy

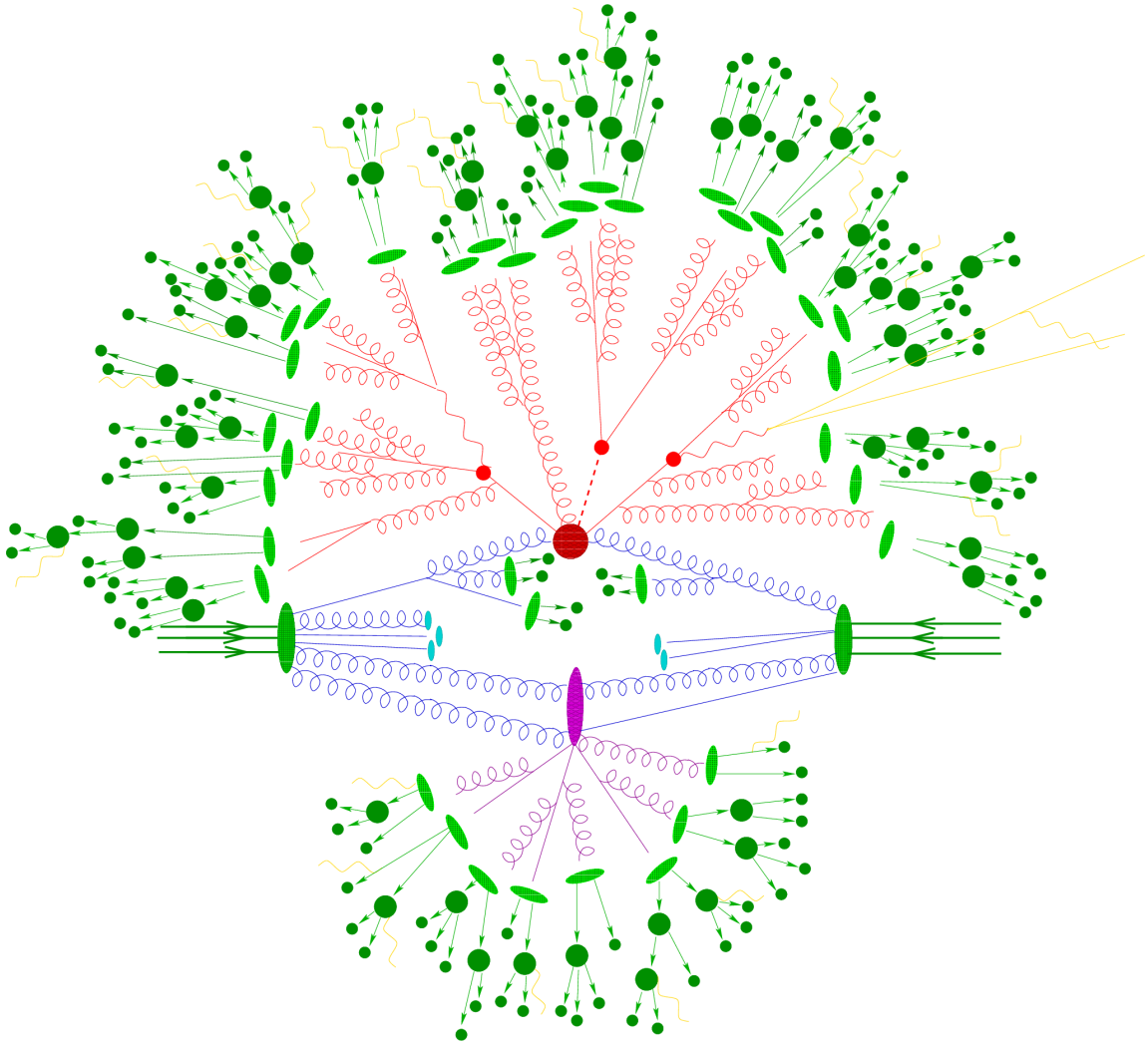

Figure 4.1: An illustration of a pp collision. The partons from the incoming protons are shown as blue lines. The hard scattering is indicated by the red blob. Red lines representing the immediate products of the hard process eventually hadronize which are shown as light green blobs. The purple blob and its products is an example of the underlying event. Figure source [49].

scale Q^2, and interacting with a given parton is determined by the parton distribution functions of the proton. Collaborations such as NNPDF [50, 51] form these functions based on observations from a wide variety of processes involving protons, e.g. the deep inelastic scattering process as well as the Drell-Yan (DY) and multijet processes at hadron colliders. By deriving the PDFs in a process agnostic way, they can be applied to all types of simulated processes. An example of a PDF set for two different energy scales is displayed in Fig. 4.2. With this information and the matrix element (ME) generators, the cross section of a SM process can be calculated. ME generators require information about the target process including a list of final-state particles, couplings, and other settings determined by the user. There are several parameter settings, such as the strong coupling constant α_S which influence the simulated cross section. Two more noteworthy arbitrary parameters also influence the generator: the renormalization scale μ_R and factorization scale μ_F. The μ_R value regulates the divergences in perturbative QCD that appear when calculating the cross section, and μ_F sets the boundary between short and long-range particle interaction. With perturbative QCD, generators can include expansion terms in the calculation for increased accuracy. The default accuracy without additional perturbations is referred to as "leading-order" (LO). Increasing the accuracy requires the calculation of additional perturbative QCD terms. Each expansion prepends a "next-to" to LO, with the first expansion being referred to as "next-to-leading order" (NLO), the second being "next-to-next-to-leading order" (NNLO), and so on. The ME generators which are common in the CMS Collaboration are POWHEG [52–55] (NLO), MADGRAPH (LO), and MADGRAPH5_aMC@NLO (NLO) [56]. With these tools, individual events are generated using MC sampling to populate a kinematic phase space which is also

specified by the user. In addition to SM, BSM processes can be generated by altering the settings of the ME calculation. In this work, the UNIVERSAL FEYNRULES OUTPUT (UFO) model [57] automizes the generator settings of new physics and is used to implement EFT operators into the ME calculation.

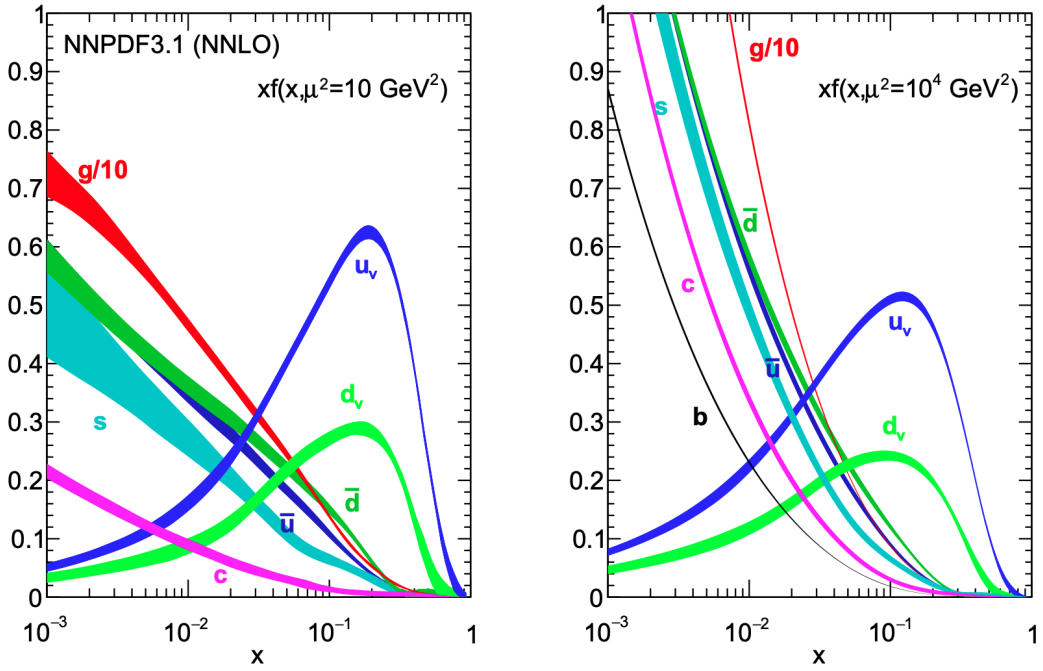

Figure 4.2: The NNPDF3.1 (NNLO) PDFs as a function of the fractional parton momentum with respect to the proton momentum, and computed at a scale $\mu^2 = 10$ GeV^2 (left) and $\mu^2 = 10^4$ GeV^2 (right). Here, μ is the same as Q which is the primary notation in the main text. Image source [51].

4.1.2 Parton Shower

During the hard scattering process, additional particles may radiate from the colliding partons or off of hard-scattered particles. While the ME generator can simulate such radiation, it does not necessarily provide the best description of radiation especially at a lower energy scale. Therefore, the CMS Collaboration has widely

adopted the PYTHIA [58] program to simulate these extra radiated particles as parton showers. This treatment includes the possibility for soft radiation where a gluon radiates from a colored quark, for collinear gluon splitting where a gluon splits into either quark-antiquark pair or gluons, or radiated photons. There are two major categories of parton showers. Initial-state radiation (ISR) and final-state radiation (FSR) occur when radiation takes place before or after the hard scatter of partons, respectively. The parton shower is simulated up until an energy scale threshold Λ_{QCD} is reached, after which hadronization will need to be considered. Using a ME generator with PS may lead to ambiguities in the sense that the same part of an event may be redundantly generated by both programs. The MADGRAPH program uses a matching scheme to avoid the double counting of phenomena in the event. For the MADGRAPH5_aMC@NLO program, the FxFx [59] matrix-element matching procedure is used while the leading-order MADGRAPH program utilizes the MLM [60] scheme. The POWHEG program regulates ME-PS merging by scaling the cross section for real emissions by a damping function ${h_{\mathrm{damp}}}^2/({p_{\mathrm{T}}}^2) + {h_{\mathrm{damp}}}^2$, which is a function of radiation p_{T} and the damping parameter h_{damp}. The value for h_{damp} is dependent on the underlying event tune which will be discussed further in Section 4.1.4. For generators with the underlying event tune TUNECUETP8M1, the h_{damp} variable is set to equal the approximate mass of the top quark $m_{\mathrm{t}} = 172.5$ GeV. However the tune for the POWHEG program was updated to TUNECP5 for which the value of h_{damp} was determined empirically [58] to be $1.379\, m_{\mathrm{t}}$.

4.1.3 Hadronization

So far, the simulation steps discussed in previous sections are responsible for simulating colored quarks and gluons up to a certain energy scale. After this threshold, perturbative QCD starts to diverge and cannot be used to describe what happens next. From first principles, color confinement stipulates that these particles will pair with other particles to form stable, color neutral hadrons. This process is called hadronization. At this energy scale, particle physicists must rely on a phenomenological models to simulate these non-perturbative processes up until they interact with the detector. Within the CMS Collaboration, the PYTHIA program is most commonly used for the simulation of hadronization.

4.1.4 Underlying Event

As discussed previously, the hard scattering of proton-proton collisions involves the interaction of one parton per participating proton. However, the entirety of each of the proton's energy does not participate in this interaction. The underlying event (UE) [58, 61] describes the process in which the proton-proton interaction provides additional hadrons unrelated to the hard scattering. The mechanism for UE involves the proton remnants with color undergoing parton showering or hadronization. Soft scattering from non-hard-scattering partons may also occur and contribute towards UE. For UE, the low-energy products cannot be modeled with perturbative QCD. Instead a phenomenological approach is used to construct a model with UE parameters tuned to data. The UE tune TUNECP5 [58] is adopted for most of simulation samples used in this work, with the exception of some background samples for 2016

which the tune TuneCUETP8M1 [61] is utilized. In this work, the PYTHIA program is responsible for UE when simulating this step of the event generation.

4.1.5 Detector Simulation

The final step in simulation involves creating a virtual description of the CMS detector for the stable final-state particles to interact with. Using the GEANT4 [62] toolkit, each detector and subdetector with specifications of its material is used to build a simulation model of the CMS detector. Everything from structural support structures, cooling infrastructure, wiring, dead channels, the magnetic field, and electronics is included in the detector simulation. With a precise model of the CMS detector, the GEANT4 program takes the incident particles and infers the various interactions that take place with the material and the magnetic field. At this step, the simulation of pileup is added to the event description. Many proton bunch-crossings occur before and after the hard process as well. This type of pileup is called out-of-time pileup. These effects are simulated by overlaying particle interactions from additional bunch-crossings and shifting their timing to mimic the expected level of pileup.

Lastly, the GEANT4 toolkit uses the location and the amount of energy captured by the detector to simulate the digitization, including electronic noise, of the signal. As with real data, the thresholds for the L1 and HLT triggers are included in the simulation and the pass or fail status of the event per trigger is recorded. At this point, the simulation and real data should be fairly close, however this form of the data is not easily analyzable. The reconstruction process transforms the data into readable objects which may be further analyzed.

4.2 Event Reconstruction

The purpose of the CMS event reconstruction is to translate the detector readout into the information of particles produced in pp collisions. This is accomplished by utilizing information from all subdetectors and building an event description comprised of the aforementioned particles. For this, the CMS Collaboration implemented the particle flow (PF) algorithm [63], which has grown to be an invaluable, sophisticated software thanks to the efforts of many algorithm developers.

Several topics concerning the reconstruction of data events are discussed in the following subsections. Particle flow is able to link tracks from the silicon tracker, deposited energy clusters in the calorimeters, and muon detector activity, and amalgamate this information to form of set of identified particles. A more in-depth description of PF is covered in Section 4.2.1. The particle flow particle candidates can be further refined into reconstructed physics objects which are suitable for analysis. This refinement is specific to the type of particle or set of particles and involves passing selective identification criteria, algorithmic clustering, and machine learning based classification. The list of physics objects and their criteria, as utilized in the analysis discussed in Chapter Five, are enumerated in Section 4.2.2. One aspect of reconstruction is identifying events which are likely corrupted by anomalous detector signals or reconstruction issues. Affected events are filtered out. Additionally, there are some observed differences between reconstructed simulated data and real data. To remedy this, corrections are applied to simulated events in the form of scale factors which bring the simulated data closer to the real collision data. Both the event filters and the simulation-to-data corrections are discussed in Section 4.2.3.

4.2.1 Particle Flow

The particle flow algorithm combines the information from all subsystems of the CMS detector to reconstruct muons, electrons, photons, neutral hadrons, and charged hadrons. Collectively, these particles are known as PF candidates. Particle flow utilizes an order of operations and several techniques specific to the type of candidate to eventually produce an output. However before this step can be executed, particle flow needs to construct higher-level information from the raw data. This is divided into three categories which include calorimeter clustering, track construction, and the linkage of this information.

- *Track Reconstruction:* Hits are associated according to an iterative tracking algorithm, based on the Kalman filter [32], to reconstruct the path of a charged particle, also called a track, as it moves through the magnetic field. Likewise, muons are charged particles and are able to be traced in the inner tracker and in the outer muon detectors. Individual muon detector hits are pieced together to form muon tracks. Using the track information, the CMS Collaboration is able achieve high levels of momentum resolution for charged particles.
- *Calorimeter Clustering:* As electrons, photons, and hadrons move through the calorimeters, they will interact with the absorptive material and create a spray of particles in a recurrent fashion until all energy has dissipated. This has the effect of distributing the products of the original particle across multiple channels in the ECAL and HCAL. Particle flow creates clusters from the calorimeter cells based on the magnitude of the deposited energy. At this step, cells belonging to ECAL and HCAL are processed separately.

- *Linking Algorithm:* Particle flow implements a linking algorithm to connect tracks and clusters across several subsystems. In this way, tracks can be associated with ECAL clusters or HCAL clusters, ECAL clusters can combine with HCAL clusters, and muon tracks can be connected with compatible tracks from the inner-tracker. This level of information is referred to as PF blocks.

Particle flow pieces this information together to identify the type of particle that interacted with the detector. For each PF candidate, an example of their interaction with the detector is shown in Fig. 4.3. For the PF algorithm, the order of operations matter. PF candidates with the cleanest signature are identified first, followed by candidates with the next-cleanest signature and so on. The PF blocks related to the identified PF candidate are removed from the algorithm when considering the next potential candidate. The PF algorithm is a complex algorithm. Therefore, it is difficult to describe the whole algorithm chain in detail, but generally, muon candidates are reconstructed first, then electrons, charges hadrons, photons, and neutral hadrons.

(1) Muons will pass through the whole of the CMS detector with detectable signatures in the inner-tracker and muon systems. Muon tracks and compatible inner-tracker tracks are reconstructed as PF muons.

(2) Electrons passively interact with the inner-tracker and are absorbed by the ECAL subdetector. Electron tracks are identified using a Gaussian Sum Filter which accounts for energy loss due to braking radiation, bremsstrahlung. ECAL clusters are matched to the electron track according to track trajectory and track energy. From this information, PF electrons are reconstructed.

(3) Charged hadrons will interact with the inner-tracker, deposit some energy in the ECAL, and deposit the remaining energy in the HCAL. Because electrons and muons have been identified, the remaining tracks in the inner-tracker should come from charged hadrons. The track trajectory is matched to the closest HCAL and ECAL clusters. If the track energy is higher than the sum of the energy in the calorimeter clusters, then nearby ECAL clusters which are compatible with the track trajectory are added gradually until parity is achieved. After this condition is met, charged hadrons are reconstructed.

(4) Photons pass through the inner-tracker without leaving a trace and will deposit their energy in the ECAL. Out of the remaining ECAL clusters, those clusters which are not associated with HCAL clusters will be reconstructed as photons.

(5) Neutral hadrons will pass through the inner-tracker undetected and will deposit energy in the ECAL and HCAL. The remaining ECAL and HCAL clusters which are geographically connected will be reconstructed as neutral hadrons.

After proceeding through the particle flow algorithm, the event description has been transformed from raw information of detector hits to a list of particles that likely interacted with the detector. However, there is still room to refine this further and reconstruct the higher-level physics objects that were produced directly from the hard process.

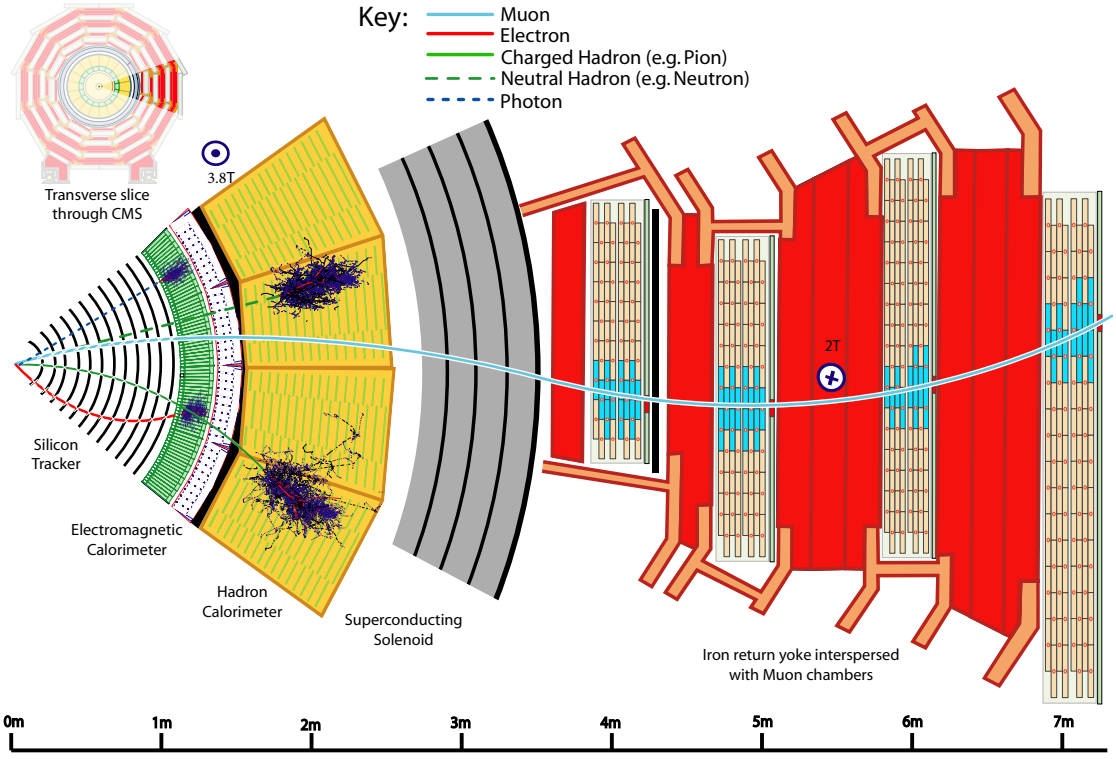

Figure 4.3: A depiction of a transverse slice of the CMS detector, with examples of how different particles interact with the detector. Figure source [63].

4.2.2 Reconstruction of Physics Objects

Following particle flow, the event includes a list of reconstructed PF particle candidates. While this information is useful, a full physics analysis requires higher-level physics objects which better represent the main physics production of the hard process. For example, an analysis interested in the $t\bar{t}$ process will seek to reconstruct top quarks or the immediate products of top quark decay in order to discriminate $t\bar{t}$ events from everything else. In this way, a data set made up of $t\bar{t}$ events is probed to determine the physical nature of this process. Many physics objects are common across analyses within the CMS Collaboration. For this reason, physics object groups (POGs) were formed to develop and standardize object definitions. Each POG consists of experts who are responsible for providing recommendations for object identification and improving upon object-related algorithms among other things. The physics analysis discussed in Chapter Five relies on a number of these objects which will be discussed further in the following subsections.

4.2.2.1 Primary Vertex Reconstruction In a reconstructed event, several tracks can be traced back to the same interaction vertex using the adaptive vertex filter [64] fitting algorithm. The event description includes many vertices due to the hard scatter and pileup, and the vertex associated with the hard process is considered as the primary vertex (PV). Out of all the vertices in the event, the primary vertex is the one with the largest value of $\sum p_{\mathrm{T}}^2$ of the jets and missing transverse momentum that are formed from tracks associated with that vertex. Both jets and

missing transverse momentum are discussed in Sections 4.2.2.5 and 4.2.2.2, respectively. Reconstruction of the PV is critical for pileup mitigation and reconstructing other physics objects.

4.2.2.2 Missing transverse momentum The initial conditions of the pp collisions dictate that the protons are travelling along the beam axis with zero net momentum in the transverse plane. Following the principles of conservation of momentum, the vector sum of the p_{T} of the collision products should equate to zero provided that those products are visible to the detector. However, neutrinos pass through the detector undetected and reconstruction of low-p_{T} particles may fail. The magnitude and direction of this phenomena is determined by measuring the missing transverse momentum ($\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}$). These are defined as

$$\vec{p}_{\mathrm{T}}^{\,\mathrm{miss}} = -\sum_{i=1}^{N} \vec{p}_{\mathrm{T}}^{\,(i)}, \tag{4.1}$$

where the summation is over all PF candidate particles. Its magnitude is denoted by $p_{\mathrm{T}}^{\mathrm{miss}}$.

4.2.2.3 Muons Muon candidates are reconstructed as tracks in the tracking system consistent with measurements in the muon system, and associated with calorimeter deposits compatible with the muon hypothesis. In addition to the criteria set by the particle flow algorithm, reconstructed muons pass more stringent identification and isolation requirements. For identification, muons are identified according to the medium working point (WP) following the recommendations of the muon POG [65]. Also, the significance of the three-dimensional impact parameter (SIP3D) of the track of the muon candidate is calculated as the distance of closest approach to the PV divided by the uncertainty. A requirement for muon candidates

that SIP3D$^{\mu} < 4$ standard deviations ensures that the muon in question is strongly associated with the hard process. Reconstructed muons satisfy the mini-isolation (I^{μ}_{mini}) criterion. The I^{μ}_{mini} is calculated as the scalar sum of the p_{T} of all charged hadron, neutral hadron, and photon PF candidates within a cone around the muon, where the charged hadrons are associated with the PV and the neutral hadron and photon contributions are corrected for pileup. The radius of the cone (R) is a function of the muon p_{T}:

$$R = \begin{cases} 0.2, & p_{T,\text{lep}} \leq 50 \text{ GeV} \\ \frac{10 \text{ GeV}}{p_{T,\text{lep}}}, & p_{T,\text{lep}} \in (50 \text{ GeV}, 200 \text{ GeV}) \\ 0.05, & p_{T,\text{lep}} \geq 200 \text{ GeV} \end{cases} \tag{4.2}$$

Reconstructed muons are required to have a $I^{\mu}_{\text{mini}}/p^{\mu}_{\text{T}} < 0.2$ and $p^{\mu}_{\text{T}} > 30$ GeV. Lastly due to limitations in the tracker coverage, they also have $|\eta^{\mu}| < 2.4$.

So far, this set of criteria is motivated by the effort to reconstruct muons which originate from the hard process, however some consideration is given to muons which are produced as the result of hadron decay. These are dubbed as soft muons since they are typically lower in p_{T}. These objects are reconstructed from high purity, connecting tracks within the inner tracker [65].

4.2.2.4 Electrons Electrons can be challenging to reconstruct due to their tendency to radiate a significant amount of energy through bremsstrahlung. The current strategy for their reconstruction is to combine ECAL clusters from suspected radiated photons with the electron cluster, using a more optimal Gaussian Sum Filter (GSF) [66] track fitting algorithm. Further guidelines for electron reconstruction have been set in place by the Egamma POG. For example, electrons are identified

with a cut-based approach at the "tight" working point [67, 68]. For the purposes of the analysis discussed in Chapter Five, the selection related to the isolation of the electron is omitted from the identification criteria and replaced with I^e_{mini}. The reconstruction requirements of electrons are similar to muons except more strict due to the prevalence of fake electron signatures and non-prompt electron production from hadron decay. Electrons are required to pass $I^e_{\text{mini}}/p^e_{\text{T}} < 0.1$ [69]. Additionally, reconstructed electrons have $|\eta^e| < 2.5$ and $p^e_{\text{T}} > 30$ GeV for the 2016 run period, and $p^e_{\text{T}} > 35$ GeV for the 2017 and 2018 run periods. A higher p^e_{T} threshold for 2017 and 2018 is due to changes in the electron trigger. Lastly, the significance of an electron's 3-dimensional track impact parameter must be less than 4 standard deviations away from the primary vertex (SIP3D$^e < 4$). Soft electrons originating from hadron decay are also reconstructed with the cut-based identification at the "loose" working point.

4.2.2.5 Jets As a byproduct of the pp collision, quarks will be produced and will undergo hadronization. Those hadrons will decay or radiate particles, and those products will do the same and so on. The effect is a spray of particles in a cone-like shape which is referred to as a jet. In terms of detector interaction, the jet's energy will mostly be in the form charged and neutral hadrons, and thus deposit their energy in the HCAL and some in the ECAL. Jets are reconstructed by clustering PF candidates according to the anti-k_T algorithm [70].

This anti-k_T clustering is a sequential recombination algorithm which proceeds by first calculating

$$\begin{aligned} d_{ij} &= \min(\frac{1}{p^2_{T,i}}, \frac{1}{p^2_{T,j}})\frac{\Delta R^2_{ij}}{R^2}, \\ d_{iB} &= \frac{1}{p^2_{T,i}} \end{aligned} \tag{4.3}$$

for particles i and j, and where R is the distance parameter and ΔR_{ij} is the spatial separation. This is formulated in terms of the azimuthal angle ϕ and rapidity y:

$$\Delta R_{ij} = \sqrt{(y_i - y_j)^2 + (\phi_i - \phi_j)^2}. \tag{4.4}$$

Next, the algorithm compares the smallest d_{ij} out of possible particle pairs and d_{iB}. If $d_{ij} < d_{iB}$ then particles i and j are merged and the process repeats until $d_{ij} \geq d_{iB}$. At the end of the algorithm, anti-k_T will produce a list of jets. The jet definition most common among CMS analyses is computed with the anti-k_T algorithm where the distance parameter is set to $R = 0.4$. With this configuration, the anti-k_T algorithm reconstructs so-called AK4 jets. Likewise, the AK8 jets are reconstructed when the distance parameter is set to $R = 0.8$. The AK8 jet is the preferred way to reconstruct hadronic decays of boosted W, Z, and Higgs bosons.

Before the clustering step, the charged hadron subtraction (CHS) or the pileup per particle identification (PUPPI) technique is used to mitigate the contribution of pileup to the event description [71]. The CHS algorithm is used to remove PF charged particles which are traced back to a pileup vertex (a non-primary vertex) and is the most commonly used pileup mitigation technique for reconstructing AK4 jets. The preferred method for AK8 jets removes pileup contamination from the PF candidates in the anti-k_T algorithm following the PUPPI protocol. The PUPPI algorithm applies more constraints to charged particles and rescales the energy of neutral particles based on their probability to originate from the primary vertex. This method also provides a better reconstruction of the jet substructure.

Following the clustering step, there are some observed differences between the four-momentum of the reconstructed jets and the generator-level particle jets. It

is understood that this is caused by the persistent contamination of pileup, non-inclusion of low-energy particles in the reconstruction, or non-linear effects in the detector response. The CMS Collaboration mitigates these issues using a series of sequential corrections to the reconstructed jet as displayed in Fig. 4.4. Each step in the jet energy scale (JES) corrections [72] is specific to the type of issue being accounted for and simply translates the p_{T} and mass of the jet closer to their true values. Also, the sequence has separate calibrations for AK4 and AK8 jets. The steps correct for pileup, detector response, and residual differences between the simulation and data.

- *Pileup:* The pileup offset corrections are based on the observations of the simulation of dijet events processed with and without pileup. Additionally for data events, a residual correction is derived based on differences between data and simulation using the random cone (RC) method [72].
- *Response:* After the pileup offset corrections are applied, a simulated sample of QCD multijet events is used to correct for poor performance or non-linear effects in the detector response. This process derives a correction for the reconstructed jet energy as a function of p_{T} and η due to differences in detector behavior based on the magnitude of these observables. Corrections related to the detector response will bring the energy of the reconstructed jet closer to the generator-level particle jet energy.
- *Residuals:* The final two steps in the sequence use a data driven method to correct jet energy only in data events. Both steps exploit the expected p_{T} balance of the event. The first residual applies an η-dependent correction determined with dijet events, relative to a jet of similar p_{T} in the barrel

reference region $|\eta| < 1.3$. The second residual is a p_{T}-dependent correction, which is measured in Drell-Yan plus jets or photon plus jets events based on the leptons or photons that recoil off of one or more jets. The derived residual corrections for the reconstructed jet energies restores p_{T} balance, on average, to the event.

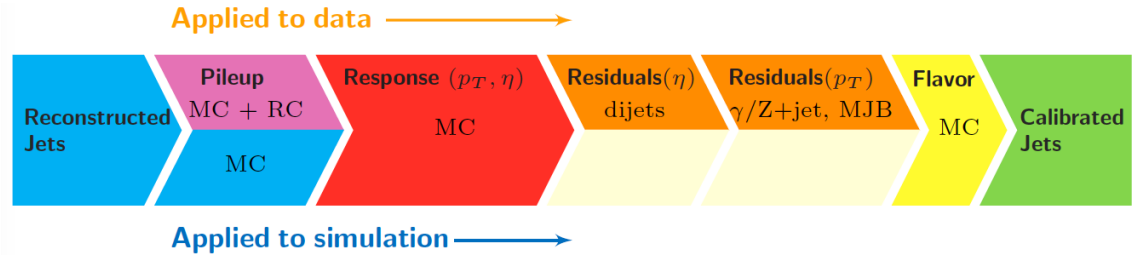

Figure 4.4: The sequence of corrections for reconstructed jets in data and simulation. The corrections marked with MC are derived from studies of simulation, RC corresponds to random cone, and MJB signifies the analysis of multijet events The flavor step is optional and not utilized for either AK4 or AK8 jets in this thesis. Figure source [72].

After the JES corrections, there exist differences in the observed jet energy resolution (JER) between jets in simulation and data. Generally, simulated jets tend to have better resolution compared to data. Therefore the JER corrections [72] apply a smearing factor c_{JER} to worsen the energy resolution in simulated reconstructed jets. This method depends on whether or not the jet can be matched to a generator-level jet. If the reconstructed jet is matched, then c_{JER} is defined as

$$c_{\mathrm{JER}} = 1 + (s_{\mathrm{JER}} - 1)\frac{p_{\mathrm{T}} - {p_{\mathrm{T}}}^{\mathrm{ptcl}}}{p_{\mathrm{T}}} \tag{4.5}$$

where $p_{\mathrm{T}}^{\mathrm{ptcl}}$ is the p_{T} of the generator-level jet and s_{JER} is the JER scale factor. The JER scale factor is determined by data and simulation differences observed in dijet

and photon+jet events. A jet is matched to a generator-level particle when

$$\begin{aligned} \Delta R(\text{jet, ptcl}) &< R/2, \\ |p_{\mathrm{T}} - p_{\mathrm{T}}^{\mathrm{ptcl}}| &< 3\sigma_{\mathrm{JER}}(p_{\mathrm{T}}) \end{aligned} \tag{4.6}$$

where R is the distance parameter of the jet and σ_{JER} is the relative p_{T} resolution as measured in simulation. If the reconstructed jet is not matched to a generator-level particle, then the resolution is smeared stochastically according to

$$c_{\mathrm{JER}} = 1 + \mathcal{N}(0, \sigma_{\mathrm{JER}})\sqrt{\max(0, s_{\mathrm{JER}}^2 - 1)} \tag{4.7}$$

where $\mathcal{N}(0, \sigma_{\mathrm{JER}})$ is a random number sampled from a normal distribution with mean 0 and standard deviation σ_{JER}. After obtaining c_{JER}, the jet four-momentum is scaled by the smearing factor.

The collection of AK4 jets are refined by passing additional selections. AK4 jets are required to have $p_{\mathrm{T}} > 30$ GeV and $|\eta| < 2.4$. Additionally, they are required to pass the tight working point criteria of the "jet ID" [71]. AK4 jets with $p_{\mathrm{T}} < 50$ GeV, are required to pass the loose working point of "pileup ID" [71] in order to further minimize the contamination of jets originating from pileup. Lastly, AK4 jets must be spatially separated from electrons and muons with $\Delta R_{\mathrm{jet,\,lep}} > 0.4$.

The final list of AK8 jets undergo some additional reconstruction and pass a series of requirements. The AK8 jet mass is difficult to reconstruct due to soft wide-angle and collinear radiation contaminating the large cone-like area of the jet. The so-called "soft drop" algorithm [73] de-clusters the AK8 jet and prunes unwanted energy from the mass calculation. The procedure, using the "tagging mode", is as follows:

(1) The AK8 jet is reclustered following the Cambridge-Aachen (C/A) algorithm [74–76] into a new jet j. The next-to-final step of the clustering produces two subjets, j_1 and j_2.

(2) If the soft-drop condition is satisfied, $\frac{\min(p_{\mathrm{T}1},p_{\mathrm{T}2})}{p_{\mathrm{T}1}+p_{\mathrm{T}2}} > z_{\mathrm{cut}}(\frac{\Delta R_{12}}{R_0})^{\beta}$, then j is the final soft-drop jet.

(3) Else, j is redefined as the subjet with the larger p_{T} and we repeat through the procedure.

(4) If j can no longer be de-clustered, then remove j from consideration.

This algorithm with $z_{\mathrm{cut}} = 0.1$ and $\beta = 0$ will produce a soft-drop jet whose mass is defined as the soft-drop mass m_{SD}. Motivated by the strategy outlined in Chapter Five, AK8 jets are required to have a 50 GeV $< m_{\mathrm{SD}} <$ 200 GeV. AK8 jets must also satisfy the criteria $p_{\mathrm{T}} > 200$ GeV, $|\eta| < 2.4$, and pass the tight "jet ID" working point. Similar to AK4 jets, AK8 jets must be spatially separated from electrons and muons with $\Delta R_{\mathrm{jet,\,lep}} > 0.8$.

Additional corrections for the m_{SD} of the AK8 jet are applied to rectify differences between the reconstructed mass and the mass of the generator-level particle, as well as differences in the jet mass resolution in simulation and data. The jet mass scale "W-JMS" correction falls into the first category and uses simulated events containing a graviton decaying to a W boson. The calibration factor is determined as a function of jet p_{T} and η utilizing the W-boson mass peak, with the goal of scaling the m_{SD} to the mass of the generator-level W boson. In practice, the m_{SD} is corrected by simply multiplying the m_{SD} to the JMS correction factor $c_{\mathrm{JMS}}(p_{\mathrm{T}}, \eta)$.

$$m_{\mathrm{SD}}^{(\mathrm{corr})} = m_{\mathrm{SD}}^{(\mathrm{uncorr})} \times c_{\mathrm{JMS}}(p_{\mathrm{T}}, \eta) \tag{4.8}$$

While the JMS correction is based on the W boson mass, it is still applicable to the Z and Higgs bosons mass resonances. Another correction, based on the W boson studies, is applied to the m_{SD} to worsen the mass resolution of the reconstructed AK8 jets. The methodology and motivation are almost identical to the JER corrections. The jet mass resolution "W-JMR" corrections, applies a smearing factor c_{JMR} to the reconstruction m_{SD} of the AK8 jet. The value for the JMR correction c_{JMR} depends on whether of not the AK8 subjets are matched to generator-level subjets. This is satisfied if the following conditions are met:

$$\begin{gathered}\Delta R(\mathrm{subjet}_1,\, \mathrm{gen.\,subjet}_i) < 0.4, \\ \Delta R(\mathrm{subjet}_2,\, \mathrm{gen.\,subjet}_j) < 0.4, \\ \text{where } \; i \neq j.\end{gathered} \tag{4.9}$$

If the subjets are matched successfully then the generator-level m_{SD} is the calculated invariant mass of the two generator-level subjets and the mass smearing factor is defined as

$$c_{\mathrm{JMR}} = 1 + (s_{\mathrm{JMR}} - 1)\frac{m_{\mathrm{SD}} - {m_{\mathrm{SD}}}^{\mathrm{gen}}}{m_{\mathrm{SD}}} \tag{4.10}$$

where s_{JMR} is the JMR scale factor based on simulation and data differences observed in the resolution of the W boson mass peak. If the subjets are not matched then the smearing factor is defined as

$$c_{\mathrm{JMR}} = 1 + \mathcal{N}(0, \sigma_{\mathrm{JMR}})\sqrt{\max(0, s^2_{\mathrm{JMR}} - 1)} \tag{4.11}$$

where $\mathcal{N}(0, \sigma_{\mathrm{JMR}})$ is a random number sampled from a normal distribution with mean 0 and standard deviation σ_{JMR}. Figure 4.5 demonstrates the positive impact of applying JMS and JMR corrections to the reconstructed AK8 jet m_{SD}; bringing

the simulation and data closer in agreement with regards to the mean and standard deviation of the AK8 jet m_{SD} distribution.

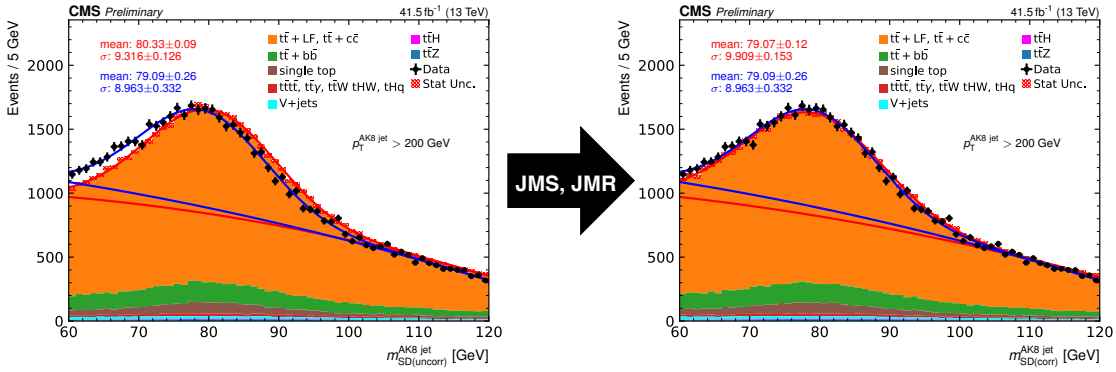

Figure 4.5: The effect of applying the JMS and JMR corrections to the reconstructed AK8 jet m_{SD}. (Left) Before corrections are applied and (right) after. Simulation (red) and data (blue) are fit to the linear combination of a Gaussian function and second-degree polynomial function. The best-fit values for the Gaussian mean and standard deviation are inscribed.

4.2.2.6 b-jets Many important standard model physics processes are associated with b quarks. For example, the top quark will almost exclusively decay to a b quark and a W boson. Also, the branching fraction of Higgs boson to b quarks is the highest out of all the Higgs boson decay channels. Likewise, the production of b quark pairs constitutes a relatively large fraction of hadronic Z boson decays. Thus, the reconstruction of b quark objects is one key component in identifying events with $\mathrm{t\bar{t}}$, a Z boson, or a Higgs boson. In the analysis covered in Chapter Five, all of these physics processes are relevant, making the identification of b quarks an essential part of the signal extraction.

Jets originating from b quarks have a distinguishing feature. When b quarks undergo hadronization, they produce B hadrons which have a relatively long lifetime

on the order of 1.5 picoseconds in their own rest frame. Because of this property, B hadrons will travel a short distance from the PV and decay thereafter. The observed products of B hadrons will produce tracks which can be traced back to a secondary vertex (SV) which is slightly displaced on the order of millimeters with respect to the PV. The B hadron has other notable properties such as its relatively large mass of 5 GeV and a 20% likelihood for an electron or muon to be a product of its decay. All of the above characteristics, as displayed in Fig. 4.6, are exploited to identify b-jets. However, it is important to note that C hadrons have similar characteristics of B hadrons with some distinguishable differences. For this reason, jets originating from b or c quarks are referred to as heavy-flavor (HF) jets. On the other hand, jets originating from u, d, or s quarks, or gluons are called light-flavor (LF) jets.

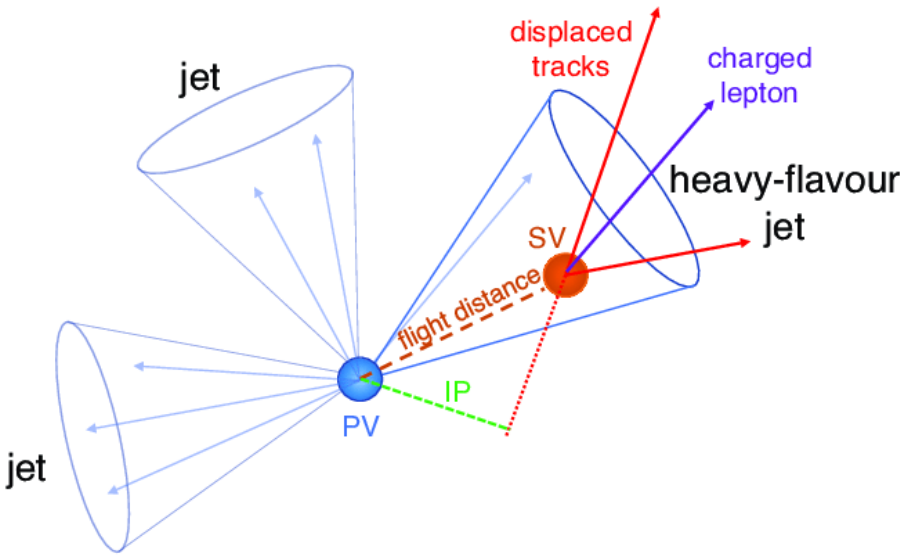

Figure 4.6: Diagram of a heavy-flavor jet with a secondary vertex. The decay products includes jets and a lepton, and have tracks which are displaced with respect to the primary vertex. Figure source [77].

The CMS Collaboration has developed several algorithms to identify AK4 b-jets based on the properties of the jet. While previous methods were efficient at identification, the DeepCSV [77] algorithm was developed to further improve upon the tagging efficiency of AK4 jets originating from b quarks. The DeepCSV discriminator is a type of machine learning algorithm which is trained to distinguish multiple categories of jet flavor classification based on 19 reconstructed properties of the AK4 jet. The data set used to train and validate the DeepCSV model is made up of simulated events of $t\bar{t}$ or QCD multijet production. In practice, DeepCSV will provide a confidence score, a value between zero and one, that the jet is a b-jet. The confidence threshold is dependent on the data-taking year and the desired tagging efficiency. However, a higher tagging efficiency will result in a larger misidentification rate. In order to balance the two, the "medium" working point (WP) is chosen as the threshold for AK4 b-jet classification and corresponds to a tagging efficiency of 68%. This also corresponds to a 1% misidentification probability jets arising from gluons and up, down, and strange quarks, and 12% for charm quark jets. The DeepCSV medium working point for all Run 2 data-taking years is summarized in Table 4.1. Additionally, Fig. 4.7 illustrates the relative performance of DeepCSV with respect to other b-jet identification methods.

Table 4.1: The DeepCSV medium working point for every year.

Year	2016	2017	2018
DeepCSV WP	0.6321	0.4941	0.4148

Because the DeepCSV machine learning algorithm is trained and evaluated exclusively using simulated data, the observed discriminator output in data events

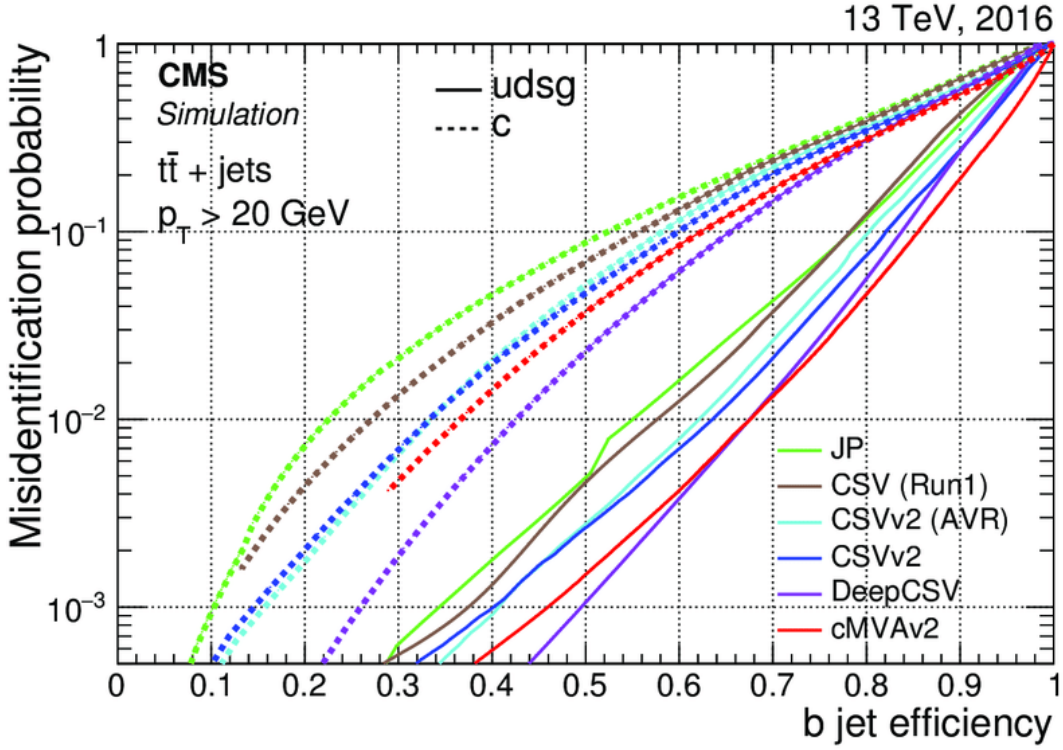

Figure 4.7: Misidentification probability versus b-jet tagging efficiency for several b tagging algorithms. The dashed curve is the misidentification rate with respect to c quarks, and the non-dashed curves to gluons and u, d, and s quarks. Figure source [77].

is slightly different in comparison to MC events. For this reason, a correction factor is derived for simulated events based on the flavor, the reconstructed p_{T} and η of the AK4 jet, and also the DeepCSV score [77]. With the scale factor (SF), an event reweighting is calculated by the following expression

$$W^{(\mathrm{b-tag})} = \prod_{i}^{N_{\mathrm{jets}}} \mathrm{SF}_{f_i, d_i, p_{\mathrm{T}i}, \eta_i} \tag{4.12}$$

where $W^{(\mathrm{b-tag})}$ is the event reweighting, and $\mathrm{SF}_{f_i, d_i, p_{\mathrm{T},i}, \eta_i}$ is the scale factor correction parameterized by jet-flavor f, DeepCSV discriminator score d, p_{T}, and η. Finally, the reweighting is normalized to prevent it from changing the overall yield of the simulation according to the AK4 jet multiplicity.

For AK8 jets, the DeepAK8 [78] machine learning algorithm has been widely adopted by the CMS Collaboration to classify reconstructed large-size jets. This method was designed to identify the originating particle that produced the AK8 jet based on the properties of the PF candidate particle constituents and SV information. In total, 42 variables for each reconstructed particle (up to 100 particles), and 15 variables for SV (up to 7 SVs) serve as inputs to the algorithm. It is trained and validated with simulated reconstructed AK8 jets with $p_{\mathrm{T}} > 200$ GeV. An alternate version of DeepAK8 is the mass-decorrelated (MD) model meaning the algorithm is trained to classify jets and punished for any bias towards the jet mass. This is very useful for identifying common decay modes across different parent particles that may have different fundamental masses, such as the Z and Higgs boson decaying to a pair of b quarks. Figure 4.8 illustrates a comparison in performance of DeepAK8 and DeepAK8-MD to other previous identification algorithms for AK8 jets originating from the Z or Higgs boson. With this configuration, the DeepAK8-MD $\mathrm{b\bar{b}}$ versus

light flavor (bbvL) discriminator identifies AK8 jets which originate from collimated b quarks or a Lorentz boosted b quark regardless of the parent process. Similar to

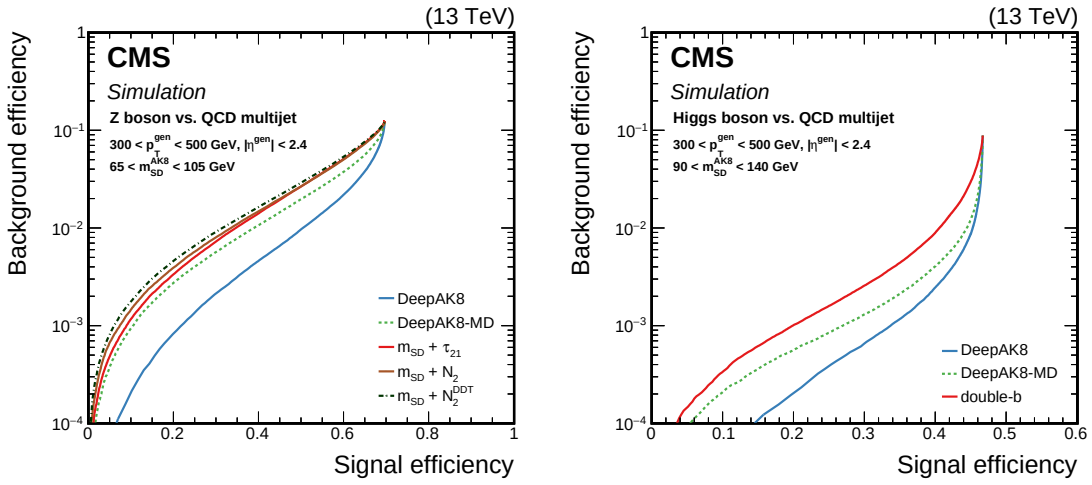

Figure 4.8: Background efficiency versus signal efficiency of the various AK8 jet identification algorithms including the DeepAK8 MVA. The plots display the performance for the hadronically decaying Z boson (left) and Higgs boson decaying to b quarks (right). Figure source [78].

the DeepCSV algorithm, the DeepAK8-MD bbvL performance is slightly different between simulated and real reconstructed AK8 jets. For this reason, a scale factor is derived to correct for the observed differences in the tagging efficiency and the mistag rate.

4.2.3 Event Filters and Corrections

The CMS detector is a very complex detector, and consequently the algorithm to reconstruct its readout information is also complicated. Unfortunately, issues with subdetector performance or occasional poor anomalous reconstruction may impact the quality of data or simulation events. For this reason, several event filters are

adopted by this work to discard events altogether, and event corrections to reweight the simulation to better match the data.

4.2.3.1 Event Filters The first set of filters will remove simulated or data events based on anomalies in the reconstruction process which cause spurious $p_{\mathrm{T}}^{\mathrm{miss}}$. The complete list of $p_{\mathrm{T}}^{\mathrm{miss}}$ filters applied to events are:

- *Primary Vertex Filter:* This filter removes events which fail to have a good primary vertex.
- *Beam Halo Filter:* This filter removes events contaminated by beam halo.
- *HBHE Noise Filter:* This filter removes events with abnormal levels of noise in the hadron barrel and hadron endcap calorimeters.
- *HBHE Isolation Noise Filter:* This topological filter removes events with clusters of noise in the hadron calorimeter.
- *ECAL Trigger Primitive Filter:* This filter removes events where a significant amount of energy appears to be lost in dead readout cells.
- *Bad PF Muon Filter:* This filter removes events with a low quality PF muon with large p_{T}.
- *EE Bad Scintillator Filter:* This filter only removes real data events with anomalous pulses due to faulty ECAL endcap crystals.
- *ECAL Bad Calibration Filter:* This filter removes events corresponding to the 2017 and 2018 data-taking years with anomalous pulses due to faulty calibrations in the ECAL.

4.2.3.2 Event Reweighting Several conditions that were present during data taking at the CMS detector are not included in the default MC simulation. Additionally, simulation with lower precision event generators may be scaled up to a high level of precision in perturbative QCD. For both scenarios, an ad hoc reweighting is applied to simulated events to correct for discrepancies between data and simulation.

- *Pileup Reweighting:* The distribution of pileup interactions is not exactly the same in simulation and data. We apply a set of pileup weights [79] which reshape the underlying luminosity distribution in MC to better match the data. Because the distribution of instantaneous luminosity in data is unknown until the completion of a data-taking year, these weights are applied on the simulation instead of being a part of the default event generation.
- *ECAL L1 Prefiring Issue:* In 2016 and 2017, the so-called "ECAL L1 prefiring issue" causes events with a significant amount of energy in the ECAL endcap, $2 < |\eta| < 3$, to suffer from a reduced trigger efficiency. This is due to a gradual timing shift in ECAL caused by radiation damage and a gradual decay of the response in the crystals. The simulation is corrected based on the pre-firing inefficiency in the trigger [80] to account for this effect in the data.
- *2018 HEM Failure:* During the 2018 data-taking period, sectors of the HCAL endcap subdetector on the minus side, "HEM", became inoperable. The impacted region corresponds to $-3.0 < \eta < -1.4$, and $-1.57 < \phi < -0.87$. The particle flow algorithm may generate additional misidentified electrons or jets with reduced energies in the disabled sector, due to energy measured in ECAL but no corresponding energy measured in HCAL. Simulated jets in

this sector have an additional uncertainty on the jet energy to cover for the observed disagreement between data and simulation. Data events containing reconstructed electrons in the affected region with $p_{\mathrm{T}} > 20$ GeV are vetoed. MC events with an electron, with the same criteria, are re-weighted according to the ratio of the luminosity before the HEM issue divided by the total luminosity.

- *Top p_{T} Reweighting:* Differences in the top p_{T} spectra between data and simulation generated at NLO QCD accuracy have been observed. For this reason, theorists have calculated $\mathrm{t\bar{t}}$ production at the LHC at NNLO QCD and including NLO EW corrections (NNLO QCD+NLO EW) [81]. The corrections are applied to simulated $\mathrm{t\bar{t}}$ production using event weights. The additional weight can be written as

$$W^{(\mathrm{topp_T})} = \sqrt{\mathrm{SF}(\mathrm{t_1\, p_T^{gen}}) \times \mathrm{SF}(\mathrm{t_2\, p_T^{gen}})} \tag{4.13}$$

where $\mathrm{SF}(\mathrm{t\, p_T^{gen}})$ is the scale factor as a function of the generator-level top quark p_{T} with the form

$$\mathrm{SF}(\mathrm{p_T}) = a\mathrm{e}^{bp_{\mathrm{T}}} - cp_{\mathrm{T}} + d \tag{4.14}$$

where $a = 0.103$, $b = -0.0118$, $c = -0.000134$ and $d = 0.973$. As shown in Fig. 4.9, the scale factor parameters are derived from a parametric fit of the ratio of the top p_{T} spectra between NLO QCD and NNLO QCD+EW $\mathrm{t\bar{t}}$.

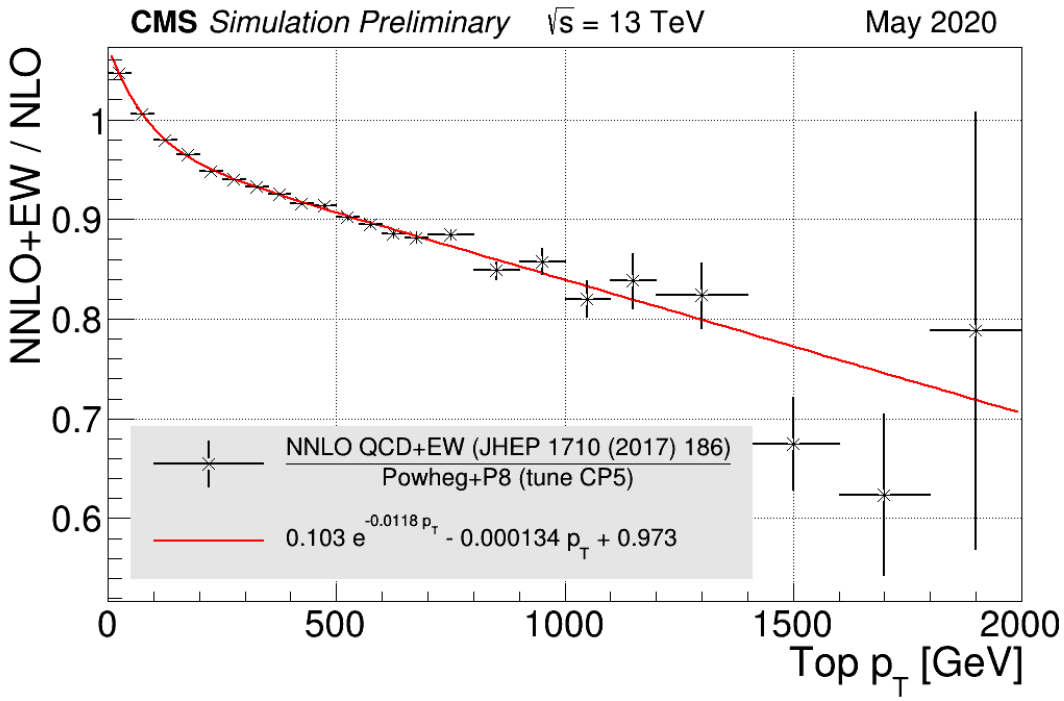

Figure 4.9: Parametric fit of the ratio between NNLO QCD+NLO EW and POWHEG +PYTHIA 8 NLO QCD of the generated top quark p_{T} spectra in $\mathrm{t\bar{t}}$ production. Figure adopted from the information in [81].

CHAPTER FIVE

Search for New Physics in $t\bar{t}Z$ and $t\bar{t}H$ in Effective Field Theory

5.1 Introduction

Searches for new physics at the LHC targeting exotic Higgs, supersymmetry, and other BSM theories have yet to observe significant deviations from the standard model. Null results from direct searches for new particles suggest that such new particles may be too massive ($\gg 1$ TeV) to be detected directly at the LHC. Therefore, effective field theory, as outlined in Chapter Two, can be used as a model-agnostic approach to probe new physics in the form of a higher-order extension of the SM theory. The extension contains a set of EFT operators which add new couplings between known SM particles. In particular, the large coupling between the top quark and the Higgs and Z boson is prone to BSM effects for relevant processes with a large Lorentz boost such as $t\bar{t}$ associated with a Z or Higgs boson [82–85].

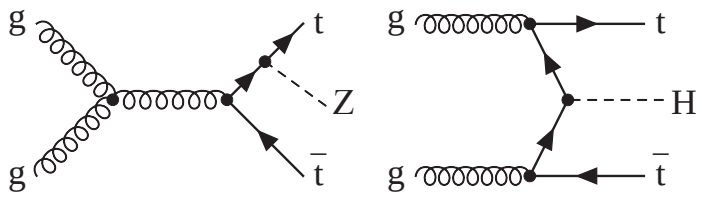

Figure 5.1: Example tree-level Feynman diagrams for the $t\bar{t}Z$ and $t\bar{t}H$ production processes.

Both the $t\bar{t}Z$ and $t\bar{t}H$ processes, shown in Fig. 5.1, have a relatively low production rate at the LHC. This, combined with the high production rate of backgrounds such as the production of top-antitop quark pair, has made it a challenge to measure

the $t\bar{t}Z$ and $t\bar{t}H$ cross sections with high precision. To date, analyses have been able to achieve a precision of 8% for the $t\bar{t}Z$ cross section [86–88] and 20% for the $t\bar{t}H$ cross section [89, 90]. Many of these measurements, especially of $t\bar{t}Z$, have been performed in final states containing multiple charged leptons. The advantage of the multi-lepton final state, despite the low branching fraction of W and Z bosons decaying to leptons, is the large reduction of prominent LHC background processes and a signature with a relatively low fake-rate. Additionally, CMS analyses have probed EFT effects in the multi-lepton final states of the $t\bar{t}Z$ and $t\bar{t}H$ processes [82, 85].

This thesis focuses on pp collisions producing $t\bar{t}Z$ and $t\bar{t}H$ with decays containing a single lepton. Notably in $t\bar{t}Z$ and $t\bar{t}H$ production, the branching fraction of the single lepton decay is much larger than the multi-lepton decay; however, the $t\bar{t}$ background with additional jets ($t\bar{t}$ + jets) lowers the level of signal purity within the single lepton phase space. A large amount of the $t\bar{t}$ + jets background can be suppressed using boosted jet tagging techniques to identify the signature of a high-p_T Z or the Higgs boson decaying into quarks. Particularly, the DeepAK8 tagging algorithm aimed at identifying bosons decaying to $b\bar{b}$ pair has an excellent tagging efficiency [78]. The mass of the reconstructed boson, as determined by the soft-drop algorithm [91], categorizes the object as a Z or Higgs boson. The amount of background contamination stemming from $t\bar{t}$ + jets events is reduced by employing a custom machine learning algorithm which is trained to distinguish $t\bar{t}Z$ and $t\bar{t}H$ events from background. This neural network (NN) classifies an event based on inputs containing information about the reconstructed $t\bar{t}$ system and the Z or Higgs boson candidate, and the overall topology of the event.

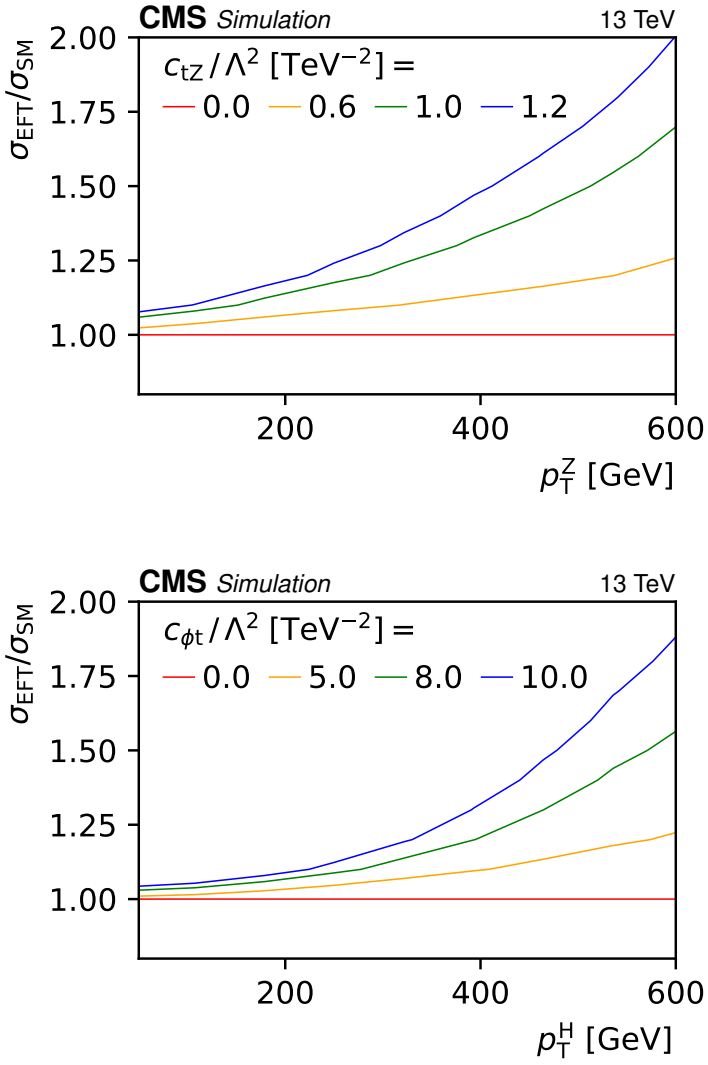

Figure 5.2: A contour plot illustrating the ratio of the production rate in EFT over the SM theory given the boson p_T and Wilson coefficient setting. The top plot displays the cross section ratio σ_{EFT}/σ_{SM} of $t\bar{t}Z$ for several values of c_{tZ}/Λ^2 versus the Z boson p_T. The bottom plot illustrates the impact $c_{\varphi t}/\Lambda^2$ has on the relative production rate of $t\bar{t}H$ versus Higgs boson p_T. In both cases, a difference between EFT and the SM allow for tighter constraints on the WCs at high boson p_T.

When probing deviations from the SM, the effects from non-zero WCs generally become more pronounced at high Z or Higgs boson p_T as demonstrated in Fig. 5.2. Thus, analyzing events with a Z or Higgs boson with a large Lorentz boost should allow tighter constraints on the EFT model. Moreover, measuring the p_T of the Z

or Higgs boson differentially may further increase the potential of providing tighter constraints. This strategy has been adopted in a couple CMS analyses which use the differential measurements of $t\bar{t}Z$ and $t\bar{t}\gamma$ [83, 84] to set limits on WCs. In this work, the p_T of the Z or Higgs boson candidate along with its mass and the neural network output form three discriminating variables. The expected signal and background yields in each bin in this three-dimensional space are determined from MC simulated events and are a function of nuisance parameters, which model sources of systematic uncertainty, as well as parameters of interest (POIs). The expected yields are fit to the real data, thereby measuring what values of the POIs, such as the WCs, are compatible with data observed at the LHC. In summary, deviations from the SM as a result of EFT operators are probed in $t\bar{t}Z$ and $t\bar{t}H$ events containing a single lepton and a boosted Z or Higgs boson decaying into a $b\bar{b}$ pair.

5.2 Data and Simulation Samples

This analysis uses the full data set of pp collisions collected in 2016, 2017, and 2018 with the CMS detector. The data set corresponds to a collision center-of-mass energy of 13 TeV and a total integrated luminosity of 138 fb^{-1}. By year, the total integrated luminosities are 36.3, 41.5, and 59.7 fb^{-1} for 2016, 2017, and 2018, respectively. The data set comprises events passing a predetermined combination of triggers which identifies events with an electron or a muon. The set of triggers, outlined in Table 5.1, are made up of HLT triggers connected with a logical "or" and vary due to differences in run conditions between years at the CMS detector and the LHC. The MC simulation used to model the signal and background events are made

Table 5.1: The trigger paths used for the collection of the single-electron and the single-muon data set. For multiple triggers in the same run era, the triggers are combined with a logical "or".

Trigger paths single-e channel	Run era
HLT_Ele27_WPTight_Gsf_v*	2016 B–H
HLT_Photon175_v*	2016 B–H
HLT_Ele115_CaloIdVT_GsfTrkIdT_v*	2016 B–H
HLT_Ele45_CaloIdVT_GsfTrkIdT_PFJet200_PFJet50_v*	2016 B–H
HLT_Ele50_CaloIdVT_GsfTrkIdT_PFJet165_v*	2016 B–H
HLT_Ele32_WPTight_Gsf_L1DoubleEG_v*	2017 B–F
HLT_Ele35_WPTight_Gsf_v*	2017 B–F
HLT_Photon200_v*	2017 B–F
HLT_Ele115_CaloIdVT_GsfTrkIdT_v*	2017 C–F
HLT_Ele50_CaloIdVT_GsfTrkIdT_PFJet165_v*	2017 C–F
HLT_Ele32_WPTight_Gsf_v*	2018 A–D
HLT_Photon200_v*	2018 A–D
HLT_Ele115_CaloIdVT_GsfTrkIdT_v*	2018 A–D
HLT_Ele50_CaloIdVT_GsfTrkIdT_PFJet165_v*	2018 A–D
Trigger paths single-μ channel	**Run era**
HLT_IsoMu24_v*	2016 B–H
HLT_IsoTkMu24_v*	2016 B–H
HLT_Mu50_v*	2016 B–H
HLT_TkMu50_v*	2016 B†–H
HLT_IsoMu27_v*	2017 B–F
HLT_Mu50_v*	2017 B–F
HLT_OldMu100_v*	2017 C–F
HLT_TkMu100_v*	2017 C–F
HLT_IsoMu24_v*	2018 A–D
HLT_Mu50_v*	2018 A–D
HLT_OldMu100_v*	2018 A–D
HLT_TkMu100_v*	2018 A–D

†: Trigger partially available for data taking period

using several generators. A list of the samples including information about the MC generator and cross section is displayed in Table 5.2.

The signal samples used to model the expected SM rate of $t\bar{t}H$ and $t\bar{t}Z$ are generated at NLO accuracy with the POWHEG and MADGRAPH5_aMC@NLO v2 program, respectively. The $t\bar{t}Z$ and $t\bar{t}H$ SM signal samples are normalized to 0.86 pb, which is computed at NLO with resummation to next-to-next-to-leading-logarithmic (NNLL) accuracy in QCD [92], and 0.507 pb, which is computed at NLO in QCD [93], respectively. The simulation samples corresponding to the main background in this analysis, $t\bar{t}$ + jets, is generated with the POWHEG program at NLO accuracy. The total yield of the $t\bar{t}$ + jets events is scaled to the inclusive cross section of 831.76 pb, which corresponds to NNLO+NNLL accuracy [94–100]. The POWHEG and MADGRAPH5_aMC@NLO v2 program are used to generate the t-channel and s-channel single top samples at NLO accuracy, respectively. Background processes containing a top quark and a variety of other SM particles, $t\bar{t}W$, $t\bar{t}\gamma$, $t\bar{t}t\bar{t}$, and tZq, are made with MADGRAPH5_aMC@NLO v2 at NLO in QCD. Also, MADGRAPH5_aMC@NLO v2 is used to create tHq and tZW samples at LO accuracy. Events containing a vector boson and additional jets, referred to as V+jets, include W+jets and DY+jets which are generated at LO with the MADGRAPH5_aMC@NLO v2 program. Simulated backgrounds other than $t\bar{t}$ + jets are normalized to their predicted cross sections, which are taken from theoretical calculations at NLO or NNLO in QCD [92, 101–105].

For the $t\bar{t}$ + jets process, the background is subdivided based on the flavor of the additional jets. The motivation for this is to provide a more accurate model of background events which closely resemble the targeted signal signature. Events with $t\bar{t}$ + LF and $t\bar{t}$ + $c\bar{c}$ processes are derived from the same MC sample, and are

generated with the five-flavor scheme (5FS). The most critical background process, $t\bar{t} + b\bar{b}$, is generated with special settings [106, 107] in the four-flavor scheme (4FS). These settings provide a more robust simulation of $t\bar{t} + b\bar{b}$ events. However, the total expected $t\bar{t} + b\bar{b}$ yield is taken from the 5FS $t\bar{t}$ + jets sample. The difference between the 4FS and 5FS is that in the former, b quarks only appear in the final states. In the 5FS, b quarks are included in the initial states and in the PDF. The $t\bar{t} + b\bar{b}$ events are removed from the 5FS sample; leaving the remaining "other $t\bar{t}$ + jets" processes, i.e. $t\bar{t}$ + LF and $t\bar{t} + c\bar{c}$.

The PYTHIA 8.226 (8.230) program is utilized to simulate the parton showering and hadronization for the 2016 (2017 and 2018) MC samples. The procedure which matches partons from the MADGRAPH5_aMC@NLO generator and those from parton showers corresponds to the FxFx and MLM schemes for NLO and LO samples respectively. The underlying event component is simulated with the TUNECP5 for the majority of MC samples. For the $t\bar{t}W$, single top s-channel and lepton tW-channel, $t\bar{t}\gamma$, $t\bar{t}t\bar{t}$, W+jets, and DY + jets samples corresponding to the 2016 year, the underlying event is simulated with TUNECUETP8M1. MC samples with the CP5 tune use the NNPDF3.1 NNLO [51] PDFs, while those generated with the CUETP8M1 tune use the NNPDF3.0 LO [50] PDFs.

An effective field theory interpretation of the process yields are incorporated into the analysis using additional $t\bar{t}Z$, $t\bar{t}H$, and $t\bar{t} + b\bar{b}$ samples generated with the "dim6top" EFT model program [108]. The EFT model utilized in this analysis is consistent with the following:

- The degrees of freedom implemented in the dim6top EFT model are derived from the Warsaw basis of dimension-6 operators [18].

Table 5.2: A list of the MC samples used in the analysis tabulated in terms of the channel, ME generator, the cross section (σ) which the channel is normalized to, and the accuracy of the cross section calculation. Additionally, W+jets and DY + jets comprise subsamples binned according to the generator-level H_{T} in intervals starting from 400 GeV to infinity.

Sample	Channel	MC generator	σ [pb]	σ accuracy
$t\bar{t}H$	$H \to b\bar{b}$	POWHEG	0.2934	NLO
	$H \to \text{non} - b\bar{b}$	POWHEG	0.2150	NLO
$t\bar{t}Z$	$Z \to q\bar{q}$	MadGraph5_aMC@NLO	0.6012	NLO+NNLL
	$Z \to b\bar{b}$	MadGraph5_aMC@NLO	0.1157	NLO+NNLL
	$Z \to ll/\nu\nu$	MadGraph5_aMC@NLO	0.2589	NLO+NNLL
$t\bar{t}$ + jets	fully hadronic	POWHEG	377.96	NNLO+NNLL
	single lepton	POWHEG	365.46	NNLO+NNLL
	dilepton	POWHEG	88.34	NNLO+NNLL
Single top	s-channel (lepton)	MadGraph5_aMC@NLO	3.36	NLO
	t-channel (t)	POWHEG	136.02	NLO
	t-channel ($\bar{t}$)	POWHEG	80.95	NLO
	tW-channel (t)	POWHEG	35.85	NNLO
	tW-channel ($\bar{t}$)	POWHEG	35.85	NNLO
	tW-channel (t, lepton)	POWHEG	19.12	NNLO
	tW-channel ($\bar{t}$, lepton)	POWHEG	19.12	NNLO
$t\bar{t}W$	$W \to l\nu$	MadGraph5_aMC@NLO	0.1792	NLO
	$W \to q\bar{q}$	MadGraph5_aMC@NLO	0.3708	NLO
$t\bar{t}\gamma$	inclusive	MadGraph5_aMC@NLO	3.697	NLO
$t\bar{t}t\bar{t}$	inclusive	MadGraph5_aMC@NLO	0.0091	NLO
tZq	inclusive	MadGraph5_aMC@NLO	0.0758	NLO
tHW	inclusive	MadGraph5_aMC@NLO	0.0152	NLO
tHq	inclusive	MadGraph5_aMC@NLO	0.0743	NLO
W+jets	$W \to l\nu$	MadGraph5_aMC@NLO	—	NNLO
DY + jets	$Z \to ll$ ($m_{ll} > 50$ GeV)	MadGraph5_aMC@NLO	—	NNLO

- Λ is conventionally set to 1 TeV.
- The Cabibbo-Kobayashi-Maskawa (CKM) matrix is approximated as a unit matrix.

- The masses of u, d, s, c, e, μ fermions are set to zero by default.
- Baryon and lepton number violating operators are not included.
- Only tree-level simulation is possible.
- The operators involve a heavy boson and a third generation quark and are displayed in Table 2.1.
- The imaginary WCs lead to CP violation and are excluded as they are generally constrained [108].

The MadGraph5_aMC@NLO v2 program is utilized to generate the EFT samples at LO accuracy in QCD and to implement the EFT operators via the event reweighting feature. The $t\bar{t}Z$ and $t\bar{t}H$ samples are generated with up to one extra parton in the final state in order to bring the precision closer to NLO in QCD [82, 109]. The variables that dictate the magnitude of the coupling strengths of the EFT operators are the eight Wilson coefficients: c_{tW}, c_{tZ}, $c_{t\varphi}$, $c_{\varphi t}$, $c^{3}_{\varphi Q}$, $c^{-}_{\varphi Q}$, c_{bW}, and $c_{\varphi tb}$. For each event, 184 alternative event weights are created corresponding to unique combinations of values of the WCs. Each weight is a point in the 8-dimensional EFT model space. These discrete points are parameterized into a quadratic function which computes the EFT weight w_{EFT} of that event for any combinations of WC settings. The quadratic relation is constructed using the normal equation

$$S = (X^{\mathrm{T}} \times X)^{-1} \times X^{\mathrm{T}} \times y \tag{5.1}$$

where X is a 45×184 matrix with a shape corresponding to the 45 possible combinations of WC values including the SM term and the 184 reweighting, y contains the value of the 184 alternative weights, and S are the 45 quadratic relation parameters. S includes interference parameters between coefficients and self-interference

parameters (s_{2ij}), interference between WC and SM (s_{1i}), and the SM (s_0). Finally, S parameters are used to build a quadratic function for computing the w_{EFT} for any coefficient value

$$w_{\mathrm{EFT}}\left(\frac{\vec{c}}{\Lambda^2}\right) = s_0 + \sum_i s_{1i}\frac{c_i}{\Lambda^2} + \sum_{i,j} s_{2ij}\frac{c_i}{\Lambda^2}\frac{c_j}{\Lambda^2}, \tag{5.2}$$

where subscripts i and j are indices for WCs under consideration. Notably, $w_{\mathrm{EFT}} = w_{\mathrm{SM}}$ or s_0 when all Wilson coefficients are set equal to zero.

5.3 Event Selection

The trigger and baseline event selection narrow down the data set into a manageable size and reduce the amount of background contamination in the data set. The targeted event signature comprises the single lepton decay of $\mathrm{t\bar{t}}$ in addition to a heavy boson decaying to two b quarks. The $\mathrm{t\bar{t}}$ decay products will have a single electron or muon from one of the two W bosons, missing transverse momentum from the leptonically decaying W boson, one b quark from each top quark, and two quarks from the other W boson. The physical characteristics of this signature are used to construct a set of criteria to select events consistent with this description. An outline of the trigger and baseline criteria is provided in the following subsections.

5.3.1 Trigger

The data collected at the CMS detector are required to pass the single lepton trigger as enumerated in Table 5.1. Likewise for the MC data set, a simulated version of the single lepton trigger is applied to MC events to emulate the trigger performance observed in real data. Because of residual differences in trigger performance between the simulated and real data set, a correction factor is applied to MC simulation

to bring it closer to the real data. More information about the derivation of the correction factor is presented in Appendix C.

5.3.2 Baseline Selection

The baseline selection requirements, summarized in Table 5.3, are engineered to reduce contamination from the QCD multijet background and to identify the decay signature of the single lepton $t\bar{t}$ decay, as well as the Z or Higgs boson $b\bar{b}$ decay. The criteria are as follows:

- *Exactly One Reconstructed Muon or Electron:* Events are required to contain one lepton, passing the lepton object criteria. This is limited to one electron or one muon passing the relevant object criteria as discussed in Section 4.2.2.4 and Section 4.2.2.3, respectively. This selection complements the single lepton trigger and reduces the QCD multijet background. For signal events, one lepton is produced from semi-leptonic $t\bar{t}$ decay.
- *Number of AK4 Jets* ≥ 5*:* Analysis events require at least five AK4 jets passing the object criteria defined in Section 4.2.2.5. Ideally, signal $t\bar{t}Z$ or $t\bar{t}H$ events contain two b-jets from the $t\bar{t}$ system, two b-jets from the decay of a Z or Higgs boson, and two additional jets from the $t\bar{t}$ system stemming from a hadronically decaying W boson. The decay products of a Z or Higgs boson with a sufficiently large Lorentz boost may include jets which are collimated such that they are reconstructed as a single jet. Taking this into account, a minimum of five reconstructed jets within acceptance are required.
- *Exactly One* Z *or Higgs Boson Candidate:* The decay products of a Z or Higgs boson, with a sufficiently large Lorentz boost, will be collimated enough to

be reconstructed as an AK8 jet. A machine learning algorithm is used to identify such jets which are consistent with Z or Higgs boson decaying to a $b\bar{b}$ pair. The algorithm is based on a deep neural network, the DeepAK8 $b\bar{b}$ tagger [78], which is developed in such a way to be unbiased with respect to the jet mass. The strategy of decorrelating the mass from the tagger allows for the construction of mass regions corresponding to the mass of the Z or Higgs boson, and side-band regions for constraining the background estimation. Therefore, the reconstruction criteria for the Z or Higgs boson candidate is an AK8 jet with $p_{\mathrm{T}} > 200$ GeV, $50 < m_{\mathrm{SD}} < 200$ GeV, and a tagger score > 0.8. This working point combined with a m_{SD} mass requirement has a tagging efficiency of 65–85% for $\mathrm{Z} \to b\bar{b}$ decays and 40–75% for $\mathrm{H} \to b\bar{b}$ decays, depending on the p_{T} of the AK8 jet. The misidentification rate is approximately 2.5% for QCD hadronic jets. If more than one AK8 jet fulfill the requirements, the jet with the highest $b\bar{b}$ tagger score is defined as the Z or Higgs boson candidate.

- $p_{\mathrm{T}}^{\mathrm{miss}} \geq 20$ GeV *:* A moderate amount of $p_{\mathrm{T}}^{\mathrm{miss}}$, as defined in Section 4.2.2.2, is expected from the decay of the $t\bar{t}$ pair, where a W boson decays to a lepton and neutrino. This criteria also helps eliminate the QCD multijet background events which characteristically have near zero $p_{\mathrm{T}}^{\mathrm{miss}}$.
- *At Least Two AK4 b-jets Separated from the* Z *or Higgs Boson Candidate:* The production of $t\bar{t}Z$ or $t\bar{t}H$ will contain four b quarks at tree level: two from the Higgs or Z boson decay and two from $t\bar{t}$ decay. The b quarks from the $t\bar{t}$ decay are identified using a minimum distance separation requirement from the Higgs or Z boson candidate of 0.8 in η-ϕ coordinates.

Additionally, events which are identified to likely contain non-prompt leptons from J/Ψ or Upsilon decay are vetoed due to insufficient simulation of this phenomena. Following the object definitions in Sections 4.2.2.3 and 4.2.2.4, events with a lepton and soft lepton pair of the same flavor with an invariant mass of less than 12 GeV are discarded. After the baseline selection has been applied, the agreement between data and MC simulation is checked in distributions of several kinematic properties of the event description. Overall, there are no obtuse differences as illustrated in Figs. 5.3–5.5.

Table 5.3: Summary table of analysis event selections

Pass single lepton trigger	`True`
N Electrons or muons	1
N AK4 jets	≥ 5
N AK8 Z/H candidates	1
$p_{\mathrm{T}}^{\mathrm{miss}}$	> 20 GeV
N AK4 b-jets, with ΔR(AK4 b-jet, Z/H candidate)> 0.8	≥ 2

5.4 Signal Enhancement with a Neural Network

The $\mathrm{t\bar{t}}$+jets background matches most of the expected signal's event signature, especially when $\mathrm{t\bar{t}}$ production includes extra radiated bquarks, $\mathrm{t\bar{t}} + \mathrm{b\bar{b}}$. At the LHC, $\mathrm{t\bar{t}}+$jets events are produced at a rate roughly 1000 times higher than the production of $\mathrm{t\bar{t}Z}$ or $\mathrm{t\bar{t}H}$. This means, in order to have any chance of observing signal with an acceptable degree of confidence, an algorithm needs to be able to identify signal events based on non-trivial and often nuanced information. In order to improve the separation of signal from background events, instead of a traditional cut-based

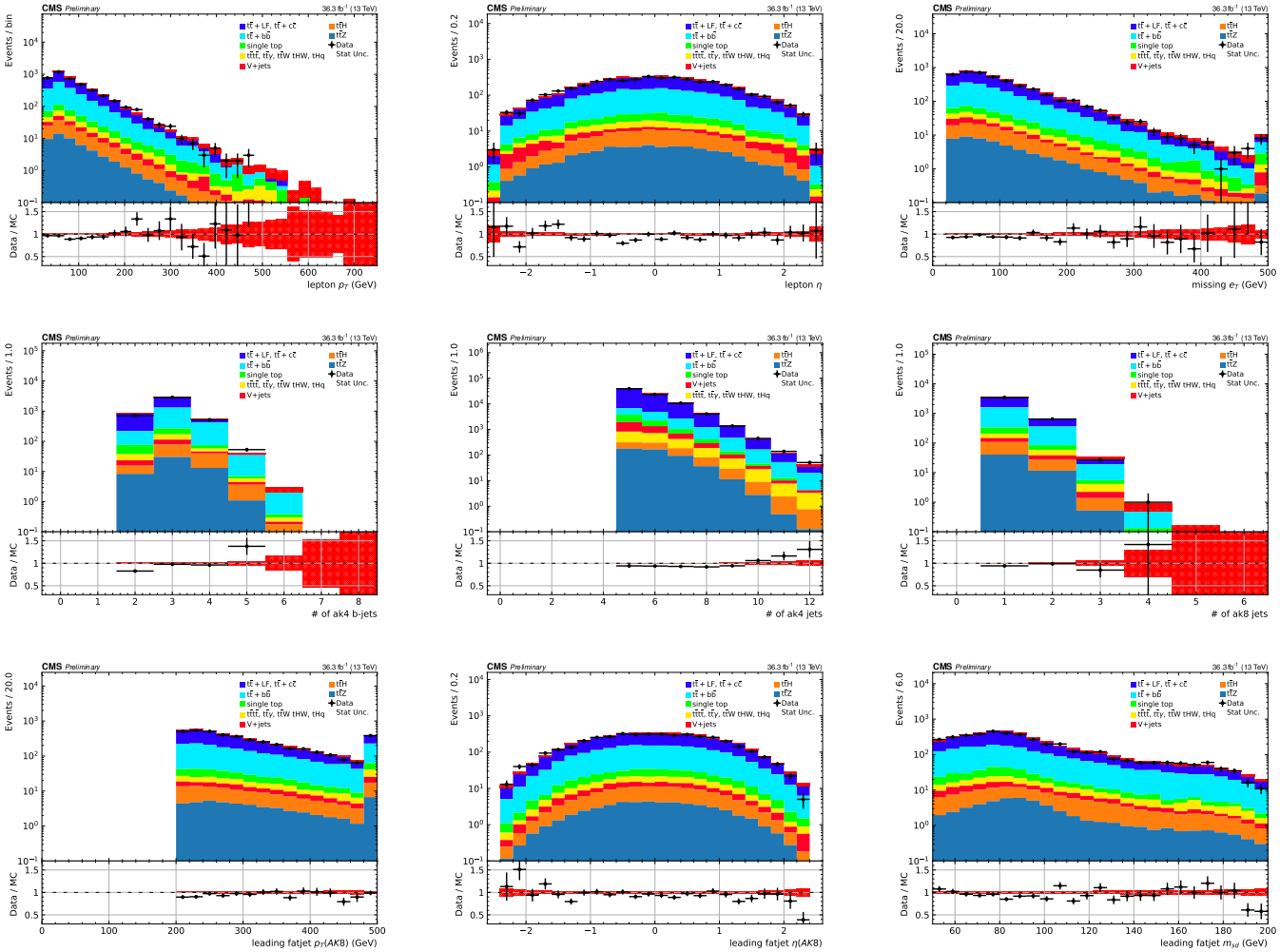

Figure 5.3: Data versus MC simulation in various kinematic properties made from events passing the baseline selection requirements corresponding to the 2016 data-taking year. The distributions are for lepton p_{T} (top left), lepton η (top center), $p_{\mathrm{T}}^{\mathrm{miss}}$ (top right), AK4 b-jet multiplicity (middle left), AK4 jet multiplicity (middle center), AK8 jet multiplicity (middle right), leading AK8 jet p_{T} (bottom left), leading AK8 jet η (bottom center), and leading AK8 jet m_{SD} (bottom right).

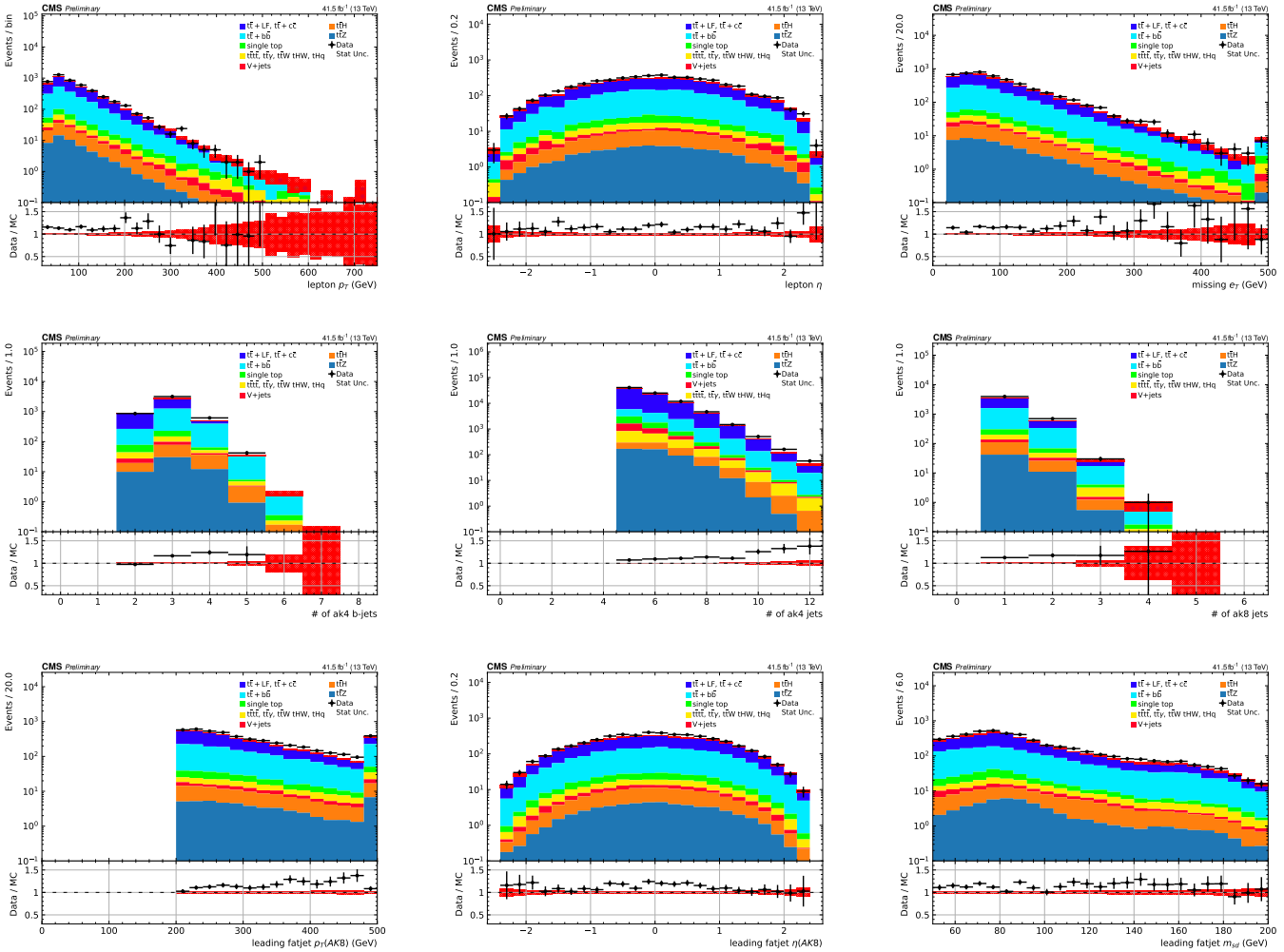

Figure 5.4: Data versus MC simulation in various kinematic properties made from events passing the baseline selection requirements corresponding to the 2017 data-taking year. The distributions are for lepton p_{T} (top left), lepton η (top center), $p_{\mathrm{T}}^{\mathrm{miss}}$ (top right), AK4 b-jet multiplicity (middle left), AK4 jet multiplicity (middle center), AK8 jet multiplicity (middle right), leading AK8 jet p_{T} (bottom left), leading AK8 jet η (bottom center), and leading AK8 jet m_{SD} (bottom right).

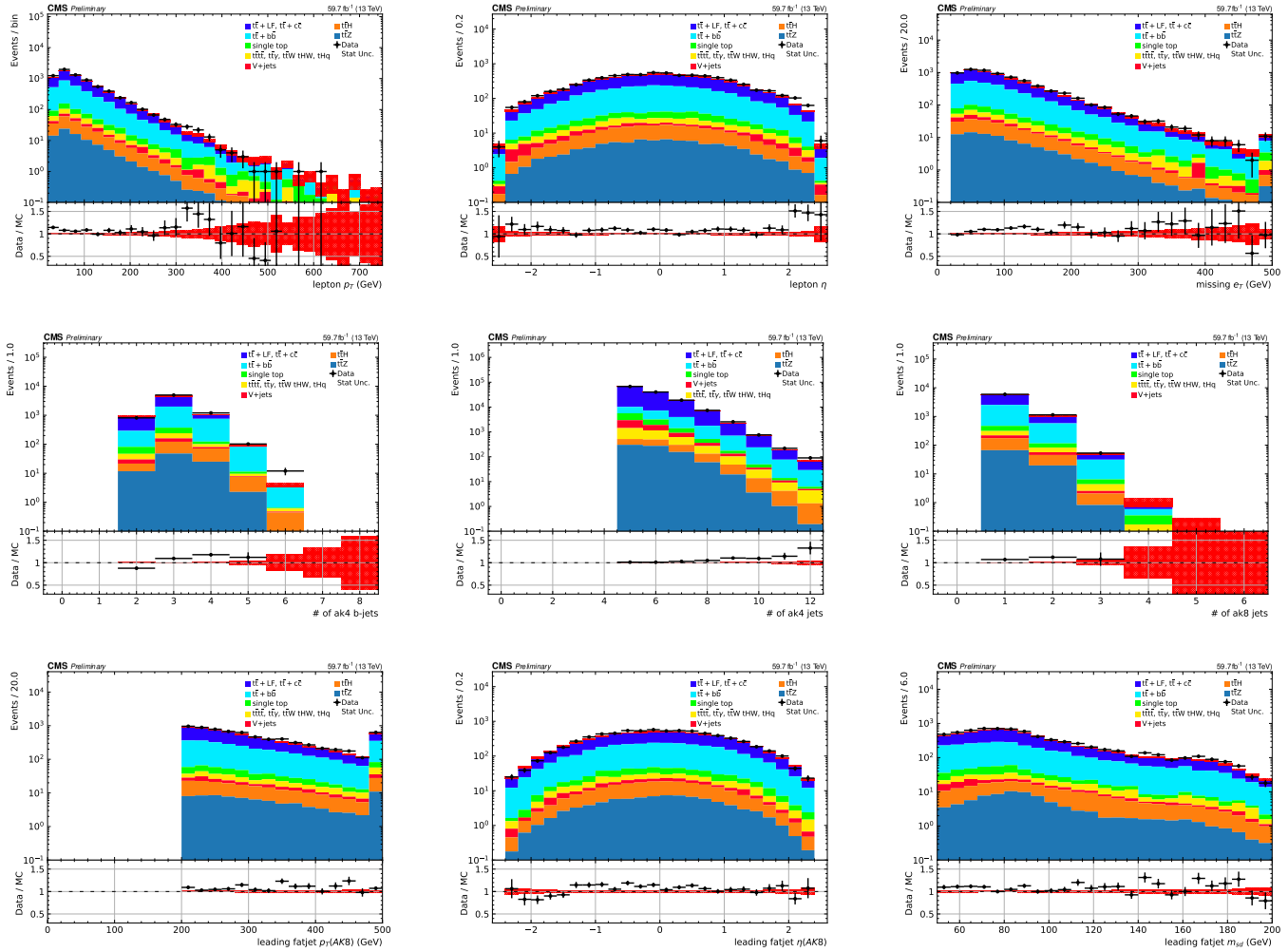

Figure 5.5: Data versus MC simulation in various kinematic properties made from events passing the baseline selection requirements corresponding to the 2018 data-taking year. The distributions are for lepton p_{T} (top left), lepton η (top center), $p_{\mathrm{T}}^{\mathrm{miss}}$ (top right), AK4 b-jet multiplicity (middle left), AK4 jet multiplicity (middle center), AK8 jet multiplicity (middle right), leading AK8 jet p_{T} (bottom left), leading AK8 jet η (bottom center), and leading AK8 jet m_{SD} (bottom right).

algorithm, the analysis utilizes a custom built neural network (NN) which is trained to distinguish "well-reconstructed" $t\bar{t}Z$ or $t\bar{t}H$ from $t\bar{t}$ and $t\bar{t}+b\bar{b}$ events. In the context of this work, signal events which are "well-reconstructed" are defined as simulated $t\bar{t}H$ or $t\bar{t}Z$ events where the generated Higgs or Z boson and its $b\bar{b}$ decay products spatially match the reconstructed Z or Higgs boson candidate. Many concepts in this chapter are based on the Refs. [110–112].

5.4.1 Principles of Neural Networks

Over the last decade, neural networks have gained in popularity out of the many available machine learning (ML) algorithms. This is due in part to the ability of a neural network to solve complicated and often non-linear problems in a diverse range of applications such as computer vision, natural language processing, regression, clustering, and classification. Often, people assume that because neural networks are used to solve complex issues, the inner-workings must also be intricate. However, this section will demonstrate the principles of neural networks are relatively straightforward.

Generally, there are two ways of constructing a neural network: supervised and unsupervised learning. The latter, while useful, is not utilized in this thesis and is not discussed further. A supervised neural network is trained on labeled data to infer a desired output, known a priori, based on a given set of input variables. The basic throughput of a neural network consists of taking a laminated batch of inputs, passing this information to a sequence of layers which performs non-linear transformations, and passing that result to a final layer which produces an interpretable output. The middle layers following the input layer and preceding the final layer are called hidden

layers. A neural network with one or more hidden layers is considered a deep neural network (DNN). The network's layers are made up of nodes that are fully-connected with the other neighboring layers' nodes. The strength of each connection is encoded as a weight. Additionally, each layer has a bias parameter. Figure 5.6 shows an example of a fully-connected deep neural network.

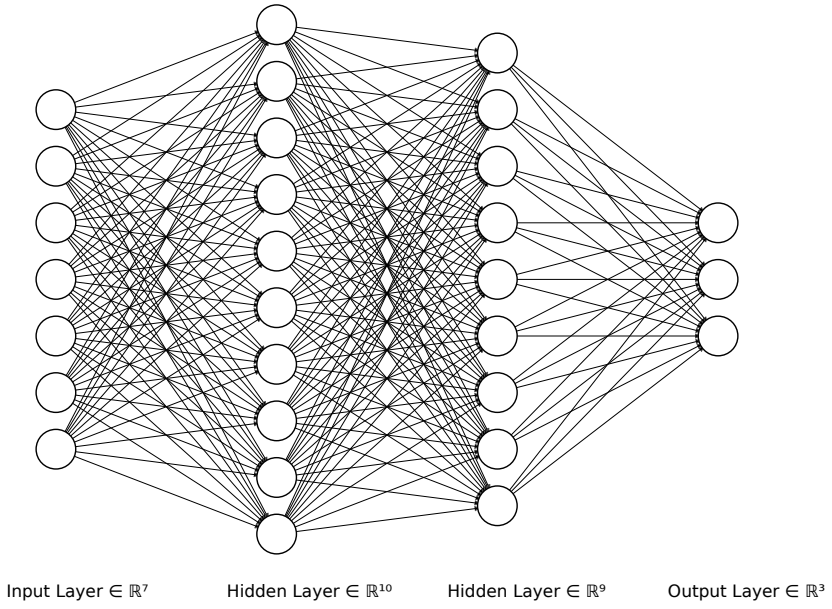

Figure 5.6: Example of a fully-connected deep neural network. The flow of information is from left to right. Each line connecting the nodes represent a weight in the model [113].

For every node, the linear combination of the previous layers output plus the bias is fed into an activation function. This is formulated as

$$a^{(i)}(z) = a\left(\sum_k w_{jk}^{(i)} x_j^{(i)} + b^{(i)}\right) \tag{5.3}$$

where a is the activation function for layer i, w_{jk} is the weight for node j, x_k is the k^{th} input from the previous layer, and b is the bias. For a sequence of fully

connected hidden layers, each iteration gradually transforms the initial inputs into non-trivial higher order features. This is the hallmark of neural networks and allows it to solve very complicated, non-linear problems. For the hidden layer nodes, a common activation function, and the one used in this work, is the rectified linear unit (ReLU) function which is given by

$$\text{ReLU}(z) = \max(0, z). \tag{5.4}$$

The function has a simple derivative which significantly reduces the complexity of optimizing the network's weights when the network is periodically updated.

The choice of activation function for the output layer depends on the problem the network is designed to solved. For multi-classification, neural networks will have a softmax activation function

$$\text{softmax}(z)_j = \frac{e^{z_j}}{\sum_l^K e^{z_l}}, \tag{5.5}$$

where each node $j = 1, ..., K$ corresponds to a supervised class, and the denominator is a normalizing term from all nodes in the output layer. The result is a real number ranging from zero to one and is interpreted as the level of confidence the event belongs to a particular class. A batch of outputs from the final layer proceeds to the loss function $J(w)$, which will be covered in more detail in a later section.

A technique called stochastic gradient descent is designed to optimize the model weights by minimizing the loss function. From the name, it may be obvious that this is accomplished by calculating the gradient of the loss function with respect to the weights. This number is subtracted from the current value of the weight which yields the updated weight

$$w_j \leftarrow w_j - \frac{\partial J}{\partial w_j}(w_1, w_2, ..., w_j), \tag{5.6}$$

where w_j is the j^{th} model weight. During training, several iterations of the full data set are repeated until the loss reaches an apparent minimum. Every iteration of the full data set is called an epoch.

However, producing an optimum model is notoriously computationally expensive. Researchers have streamlined the training process in several ways, all of which are adopted in this work. With respect to the previously described weight updating method, experts have improved upon the procedure. Stochastic gradient descent is known to suffer from the "vanishing gradient" problem which exponentially increases the computation time to minimize the neural network loss. The Adam, shorthand for adaptive moment estimation, weight optimizer circumvents this problem [114], albeit with a more complicated procedure. The Adam procedure updates the weights by iterating through the following steps:

(1) $m_j \leftarrow \beta_1 m_j - (1-\beta_1)\frac{\partial J}{\partial w_j}(w_1, w_2, ..., w_j)$

(2) $s_j \leftarrow \beta_2 s_j + (1-\beta_2)\frac{\partial J}{\partial w_j}(w_1, w_2, ..., w_j)^2$

(3) $\hat{m}_j \leftarrow \frac{m_j}{1-\beta_1}$

(4) $\hat{s}_j \leftarrow \frac{s_j}{1-\beta_2}$

(5) $w_j \leftarrow w_j + \eta\frac{\hat{m}_j}{\sqrt{\hat{s}_j+\epsilon}}$

where β_1, β_2, and ϵ are constants with values of 0.9, 0.999, and 10^{-7} respectively, and η is the learning rate defined by the user.

Also, the implementation of batch normalization and dropout layers help with the training efficiency. Batch normalization standardizes the previous layer outputs via trainable mean and variance parameters [115]. While this does increase the time to train over the first few epochs, it decreases the overall number of iterations to reach an optimum model. Dropout layers operate by excluding a percentage of the

incoming node outputs by setting their value to zero. The outputs which are dropped are randomly selected for each iteration of training. Adding dropout to a neural network provides a method to sample many neural network architectures with the benefit of only having to train a single model [116].

5.4.2 Selection of Input Features

The process of selecting input variables is vitally important in building a successful neural network. The input variables, also referred to as features, are chosen based on the following criteria. They have an adequate ability to distinguish signal and background, are not highly correlated or anti-correlated with other features, and are reasonably modeled in the analysis phase space by the simulation. This section will list all input variables used in the training as well as the initial motivation for selecting them, and describe how this set of inputs passes the inclusion threshold.

The input features fall into one of three categories. They are "$t\bar{t}$ system", "Z or Higgs boson candidate substructure", and "event topology". The full list of input variables and the category to which they belong is displayed on Tables 5.4 and 5.5. The first set of variables use objects not overlapping the Z or Higgs boson candidate and are designed to describe the kinematics of the $t\bar{t}$ system within signal. In this context, "not overlapping the Z or Higgs boson candidate" translates to a ΔR separation no less than 0.8 between the Z or Higgs boson candidate and the physics object in question. Given that the Z or Higgs boson candidate is chosen correctly, all physics objects and their kinematic combinatorics should look $t\bar{t}$-like, specifically where $t\bar{t}$ decays semi-leptonically. The next set of variables provide substructure information of the reconstructed Z or Higgs boson candidate. Well-reconstructed

signal events should have a reconstructed Z or Higgs boson candidate consistent with $b\bar{b}$ decays. Lastly, the remaining variables describe the overall event topology of the $t\bar{t}Z$ and $t\bar{t}H$ processes. In addition, it is within the neural network's ability to engineer higher-level features, in the hidden layers, by combining information belonging to variables from or across these categories. However, these will not be listed as it is non-trivial to extract a physical interpretation of these features from the neural network.

The discriminatory power of each variable is evaluated using "mutual information" [118]. This metric quantifies the dependence, or amount of shared information, an input variable has with the event's signal or background classification. For every pair of input features, their absolute linear dependence is determined by Pearson correlation. Additionally, the modeling of each input variable is evaluated using the p-score of the χ^2 goodness-of-fit. This is accomplished by creating a binned distribution of a variable in question populated with simulated events, and modeling relevant sources of systematic uncertainty. Then, the χ^2 statistic is computed for real data and thousands of randomly sampled toy data sets, thus providing a distribution of χ^2 values. The p-score is the ratio of MC simulation χ^2 values above the data χ^2 value with respect to the total amount of toy data sets generated. A p-score of > 0.05 is the standard threshold when determining if a variable is well-modeled, and is compatible with the null hypothesis i.e. the disagreement between the simulated model and data are due to random fluctuations. However, this assertion is statistical in nature and should be regarded with the quantity of variables tested as to not unnecessarily discard inputs that fall below the threshold due to chance. A comprehensive summary of p-scores for input variables is shown in Fig. 5.7. In this plot, two of the

Table 5.4: A list of the neural network input variables pertaining to the $\mathrm{t\bar{t}}$ system. The "+" represents the relativistic four-momentum sum. Some variables are calculated for both the highest p_T (leading) and second-highest p_T (subleading) jet as indicated.

Name	Description
$\mathrm{t\bar{t}}$ **system**	
b p_T	p_T of the leading (subleading) b-jet
b score	`DeepCSV` score of the leading (subleading) b-jet
q p_T	p_T of the leading (subleading) non-b-jet
q score	`DeepCSV` score of the leading (subleading) non-b-jet
$\Delta R(\mathrm{b},\mathrm{q})$	minimum ΔR between the leading (subleading) b-jet and any non-b-jet
$\Delta R(\mathrm{q},\mathrm{q})$	ΔR between the non-b-jets closest and next-to-closest to the leading (subleading) b-jet
$m(\mathrm{q}+\mathrm{q})$	invariant mass of the non-b-jets closest and next-to-closest to the leading (subleading) b-jet
$\Delta R(\mathrm{b},\mathrm{q}+\mathrm{q})$	ΔR between the leading (subleading) b-jet and the sum of the nearest and next-to-nearest non-b-jets
$m(\mathrm{b}+\mathrm{q}+\mathrm{q})$	invariant mass of the leading (subleading) b-jet and the nearest and next-to-nearest non-b-jets
$\Delta R(\mathrm{Z/H},\mathrm{b}+\mathrm{q}+\mathrm{q})$	ΔR between the Z/H boson candidate and the sum of the leading (subleading) b-jet and the non-b-jets nearest and next-to-nearest to the leading (subleading) b-jet
$\Delta R(\mathrm{Z/H},\mathrm{b}+\mathrm{b}+\mathrm{q}+\mathrm{q}+\ell)$	ΔR between the Z/H boson candidate and the sum of the leading and subleading b-jets, the non-b-jets nearest and next-to-nearest to the leading (subleading) b-jet, and the lepton
$m_\mathrm{T}(\mathrm{b}+\ell+\vec{p}_\mathrm{T}^{\,\mathrm{miss}})$	transverse mass of the subleading b-jet, the lepton, and $\vec{p}_\mathrm{T}^{\,\mathrm{miss}}$
$m(\mathrm{Z/H}+\mathrm{b})$	invariant mass of the Z/H boson candidate and the nearest b-jet
$m(\mathrm{b}+\mathrm{b})$	invariant mass of the leading and subleading b-jets
$\Delta R(\mathrm{b},\mathrm{b})$	ΔR between the leading and subleading b-jets
$\Delta R(\mathrm{Z/H},\mathrm{q})$	ΔR between the Z/H boson candidate and the leading non-b-jet
$\Delta R(\mathrm{Z/H},\mathrm{b})$	ΔR between the Z/H boson candidate and the leading b-jet
$\Delta R(\mathrm{Z/H},\ell)$	ΔR between Z/H boson candidate and the lepton
$m(\mathrm{Z/H}+\ell)$	invariant mass of the Z/H boson candidate and the lepton
$\Delta R(\mathrm{b},\ell)$	ΔR between the leading (subleading) b-jet and the lepton
$m(\mathrm{b}+\ell)$	invariant mass of the leading (subleading) b-jet and the lepton
$N(\mathrm{b_{out}})$	number of b-jets outside the Z/H boson candidate cone ($\Delta R > 0.8$)
$N(\mathrm{q_{out}})$	number of non-b-jets outside the Z/H boson candidate cone

Table 5.5: A list of the neural network input variables pertaining to the Z or Higgs boson candidate substructure and the event topology. The "+" represents the relativistic four-momentum sum. Some variables are calculated for both the highest p_{T} (leading) and second-highest p_{T} (subleading) jet as indicated.

Name	Description
Z/H **boson candidate substructure**	
b_{in} score	maximum (minimum) `DeepCSV` score of AK4 jets within the Z/H boson candidate cone ($\Delta R \leq 0.8$)
$\Delta R(\mathrm{b}_{\mathrm{in}}, \mathrm{b}_{\mathrm{out}})$	ΔR between a b-jet within the Z/H boson candidate cone and the leading b-jet outside of the Z/H boson candidate cone
$N(\mathrm{b}_{\mathrm{in}})$	number of b-jets within the Z/H boson candidate cone
$N(\mathrm{q}_{\mathrm{in}})$	number of non-b-jets within the Z/H boson candidate cone
Z/H $\mathrm{b\bar{b}}$ score	AK8 $\mathrm{b\bar{b}}$ tagger score of the Z/H boson candidate
Event topology	
N(AK8 jets)	number of AK8 jets including the Z/H boson candidate
N(AK4 jets)	number of AK4 jets
N(Z/H)	number of AK8 jets with a minimum AK8 $\mathrm{b\bar{b}}$ tagger score of 0.8
AK8 m_{SD}	maximum m_{SD} of AK8 jets excluding the Z/H boson candidate
$H_{\mathrm{T}}(\mathrm{b}_{\mathrm{out}})$	H_{T} of the b-jets outside the Z/H boson candidate cone
$H_{\mathrm{T}}(\mathrm{b}_{\mathrm{out}}, \mathrm{q}_{\mathrm{out}}, \ell)$	H_{T} of all AK4 jets outside the Z/H boson candidate cone and the lepton
sphericity	sphericity calculated from the AK4 jets and the lepton [117]
aplanarity	aplanarity calculated from the AK4 jets and the lepton [117]

fifty input variables have a p-score below 0.05. The probability of this occurrence is $\binom{50}{2} \times (0.05^2) \times (0.95^{48}) = 26\%$, and is not significant enough to alter the list of inputs.

5.4.3 Neural Network Architecture and Training

The multi-classifier neural network was built and trained using the open-source Python library, Keras [119], which interfaces with Tensorflow [120]. The output of the network is the level of confidence an event belongs to each of the three supervised classes: signal, $\mathrm{t\bar{t}} + \mathrm{LF}$, and $\mathrm{t\bar{t}} + \mathrm{b\bar{b}}$. The prominent $\mathrm{t\bar{t}} + \mathrm{LF}$ and $\mathrm{t\bar{t}} + \mathrm{b\bar{b}}$ backgrounds are supervised separately due to their distinct way of impersonating signal.

Many have described the process of designing the architecture of a neural network as an artform. This is due in part because of the complex nature of neural networks, but also the determination of the best architecture is often ill-defined and

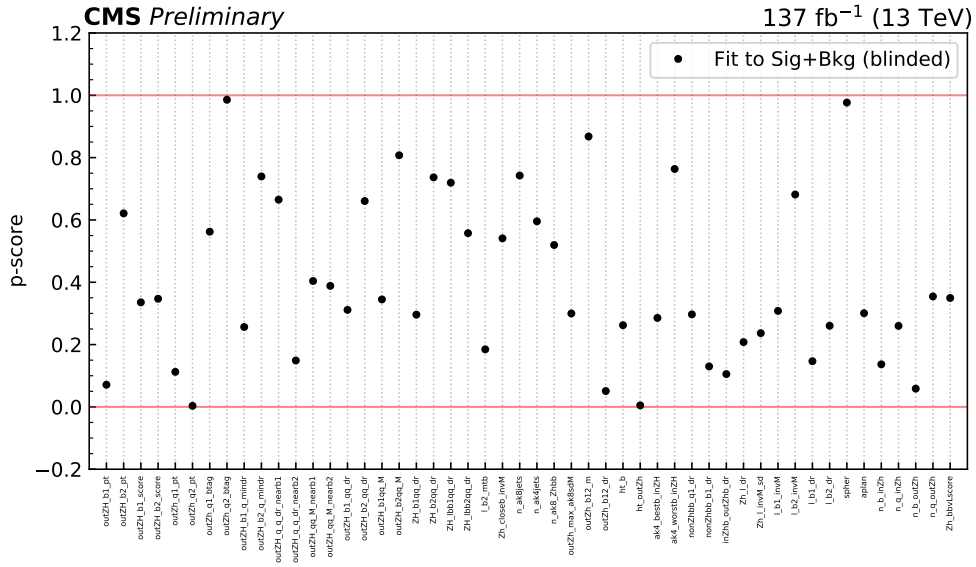

Figure 5.7: Summary of p-score values calculated per neural network input.

nuanced. Relatively small alterations in a single hyperparameter can result in large changes in the neural network performance. For this analysis, hyperparameters are optimized using a strategy called grid search. Grid search is an semi-exhaustive search of combinations of possible hyperparameter values. The selection of a set of hyperparameters is determined by the training performance for that set. The hyperparameters which are optimized in this way include the number of hidden layers, the number of perceptrons in each hidden layer, the fraction setting in the dropout layer, the learning rate, and the focus and recall parameters of the cost function. A summary of the layer-by-layer architecture is presented in Table 5.6.

One unique characteristic of this neural network is its custom built loss function, categorical focal loss [121]. Focal loss is implemented in order to circumvent a common issue in the field of machine learning called the "imbalance problem". This occurs when there are unequal amounts of examples between the supervised

Table 5.6: Architecture of the dense neural network by layer. Layer throughput is sequential, from top to bottom.

Input (50 nodes)
↓
Batch Normalization
↓
Dense (128 nodes, ReLu)
↓
Dense (64 nodes, ReLu)
↓
Dropout (0.5)
↓
Output (3 nodes, Softmax)

categories. This results in a performance bias towards the category with the highest number of examples. Focal loss works against this by treating the cost separately for each class. For example, assuming there are less signal events than background, incorrectly classifying events labeled as signal will penalize the training more than incorrectly classifying background events. This will effectively cancel out the imbalance in the training data set. The related hyperparameters which are tuned to achieve this effect are the focus and recall parameters. The formulation of the categorical focal loss function is

$$J(y, \hat{y}) = \sum_{i=1}^{K} -\alpha_i (1 - y_i)^{\gamma} \times \hat{y}_i \log(y_i) \tag{5.7}$$

where K denotes the number of supervised classes, $\hat{y}_i$ is the ground truth value or the event label, y_i is the prediction from the DNN, α is the recall parameter, and γ is the focus parameter.

The neural network is trained on simulated data consisting of well-matched $t\bar{t}Z$ and $t\bar{t}H$, as well as $t\bar{t}$+LF and $t\bar{t}+b\bar{b}$ events corresponding to all years. This data set is

split to form the test, training, and validation data sets, each having an important role in the training of the neural network. During the training phase, the network learns from the training data set and minimizes the loss per epoch. Simultaneously, the loss is calculated for the validation data; however this information is not used to train the model. If the validation loss starts to increase between epochs, while the training data set loss is decreasing, then the DNN training is halted. This strategy known is as "early-stopping". By exiting the training early, the neural network is prevented from becoming overtrained or biased towards the training data. Additionally, the validation data set is used to determine the model's optimal hyperparameters via the grid search method. Once the model is trained and its hyperparameters are chosen, the test data set is used to assess the overall performance of the network and its ability to generalize on never-seen-before data.

5.4.4 Performance and Validation

Although the neural network provides three outputs, the signal-class node corresponding to the confidence of an event being signal-like is the only output explicitly utilized from the neural network in this analysis. Henceforth, every instance in reference to the neural network score or output should be interpreted as such. Following the training, the neural network is deployed on the entire analysis data set. The performance of the network is evaluated by plotting the "receiver operating characteristic" (ROC) curve and computing the "area under the curve" (AUC). The ROC curve is traced by the sensitivity and the fall-out for various score thresholds. The sensitivity is otherwise known as the true positive rate (TPR) and the fall-out is

known as the false positive rate (FPR):

$$\text{sensitivity} = \frac{\text{number of true positives}}{\text{number of true positives} + \text{number of false negatives}},$$

$$\text{fall-out} = \frac{\text{number of false positives}}{\text{number of false positives} + \text{number of true negatives}}$$

Figure 5.8 illustrates the resulting ROC curve and the relative scores in signal and $t\bar{t}$ + jets background events, and demonstrates reasonable differentiation between signal and background events.

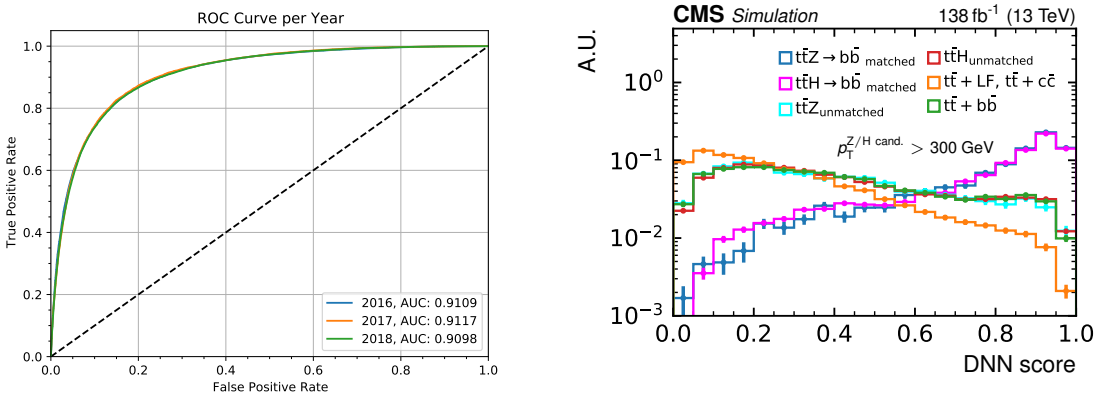

Figure 5.8: (Left) ROC curve quantifying neural network performance for all years. The AUC, displayed in the legend, is the magnitude of the area underneath the curve. (Right) The DNN performance for Run 2 simulation signal and background normalized to one, where signal events are distinguished by whether or not it is ΔR-matched (0.6) to both of the generated bquarks from the Z/H boson decay. Plots are made from events that pass the analysis baseline event selection. Error bars only account for the statistical uncertainties.

Despite efforts to regularize the neural network during training, it is still reasonably possible that the final model may have bias to some phenomena only present in simulated events. To show that the network is able to generalize to data, the p-score was calculated in a way similar to the neural network input variables for the sixty-four node outputs of the final hidden layer of the trained model. A summary

of p-scores is displayed in Fig. 5.9. Out of the sixty-four outputs, four nodes have a p-score below 0.05. This outcome has a probability $\binom{64}{4} \times (0.05^4) \times (0.95^{60}) = 18\%$, and is not significant enough to overhaul the neural network model.

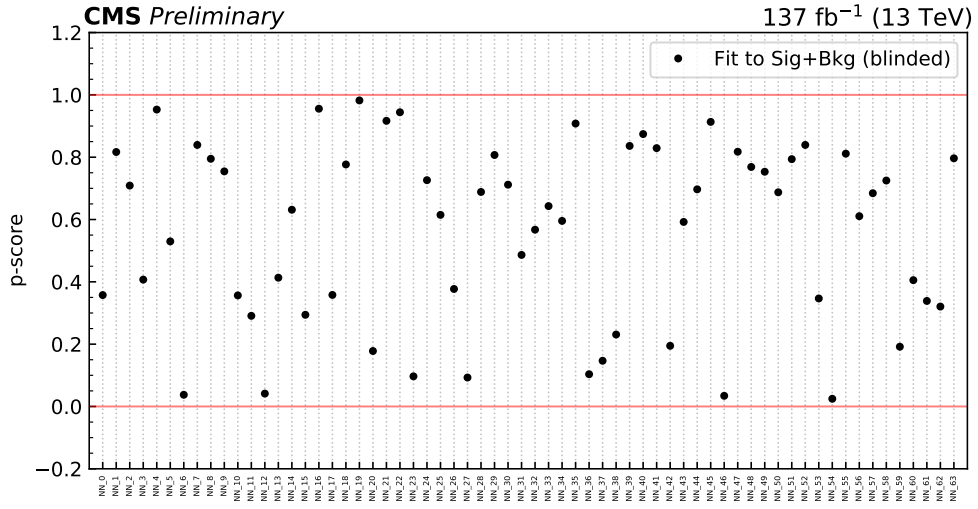

Figure 5.9: Summary of p-score values calculated for each final hidden layer node output.

5.5 Signal Extraction

A set of observables is used in order to measure physical attributes of $t\bar{t}H$ and $t\bar{t}Z$, as well as to place limits on new physics within the EFT framework. These observables are in the form of contiguous bins whose boundaries are determined by event-level kinematics. Events are placed in bins according to their reconstructed Higgs or Z bosons p_T, reconstructed Higgs or Z boson mass, and neural network score. Bin intervals are strategically engineered to form groupings of events which are enriched in signal and, inversely, groupings which are dominated by background.

Once all events have been organized according to their attributes, the bins are used to construct templates which profile the expectation from simulated events and the observation from data collected with the CMS detector. Then, a statistical interpretation of the analysis templates produces a likelihood function which quantifies the compatibility between the prediction and the observed data. The prediction is a function of several model parameters including the parameters of interest (POI), whose best-fit values are obtained by maximizing the likelihood function. Additionally, several methods within the statistical formalism are used to obtain the confidence level (CL) intervals of the POIs, or the upper limits on the POIs. The following section provide a description of how the template bins are defined, and also an overview of the statistical model used for signal extraction.

5.5.1 Analysis Bins

Events are organized into bins based on three dimensions of event-level attributes. The distributions of the attributes are illustrated individually for every data-taking year in Fig. 5.10.

- Z *or Higgs Boson* p_{T}*:* The spectrum of reconstructed Z or Higgs boson p_{T} is divided into three intervals: $200 < p_{\mathrm{T}} < 300$ GeV, $300 < p_{\mathrm{T}} < 450$ GeV, and $p_{\mathrm{T}} > 450$ GeV. The interval definitions are motivated by relative performance differences in the reconstruction of the Z or Higgs boson, the simplified template cross section (STXS) binning definition for $\mathrm{t\bar{t}H}$ [93], and according to the phenomenology that EFT operators will generally have a greater impact on events with a high boson p_{T}.

- *Deep Neural Network:* The DNN bin edges are determined by quantiles of "well-reconstructed" simulated signal events. Signal events are considered to be "well-reconstructed" when the reconstructed Higgs or Z boson candidate is ΔR-matched (0.6) to both of the generator-level b quarks from the Z/H boson decay. Six quantiles are defined based on percentiles of the neural network output distribution. The percentiles that determine the quantiles are 0%, 5%, 25%, 35%, 50%, 70%, and 100% of the total expected, well-reconstructed signal yield. Bin edges are computed according to the event's Z/Higgs boson candidate p_{T} and the data-taking year, and are tabulated in Table 5.7.
- Z *or Higgs Boson* m_{SD}*:* Lastly, within each p_{T} and NN score interval, events are divided according to their reconstructed Z or Higgs boson candidate soft-drop mass. Figure 5.11 shows the m_{SD} distributions of reconstructed Z or Higgs boson candidates in three p_{T} ranges for simulated signal and background events. Given this information, the mass intervals are chosen as $50 < m_{\mathrm{SD}} < 75$ GeV, $75 < m_{\mathrm{SD}} < 105$ GeV, $105 < m_{\mathrm{SD}} < 145$ GeV and $145 < m_{\mathrm{SD}} < 200$ GeV. However due to the lack of simulated signal and background events, the two highest mass intervals are merged, corresponding to the interval $105 < m_{\mathrm{SD}} < 200$ GeV for events with a Z or Higgs boson candidate within the interval $200 < p_{\mathrm{T}} < 300$ GeV. The two middle mass intervals contain the mass of the Z and Higgs boson, 91.2 GeV and 125.1 GeV respectively [122]. These mass intervals de-correlate $\mathrm{t\bar{t}Z}$ and $\mathrm{t\bar{t}H}$ events and are essential in measuring their respective cross sections separately.

There are 66 analysis bins per template and one template per data-taking year, meaning the analysis consists of 198 analysis bins in total. The bins which contain the

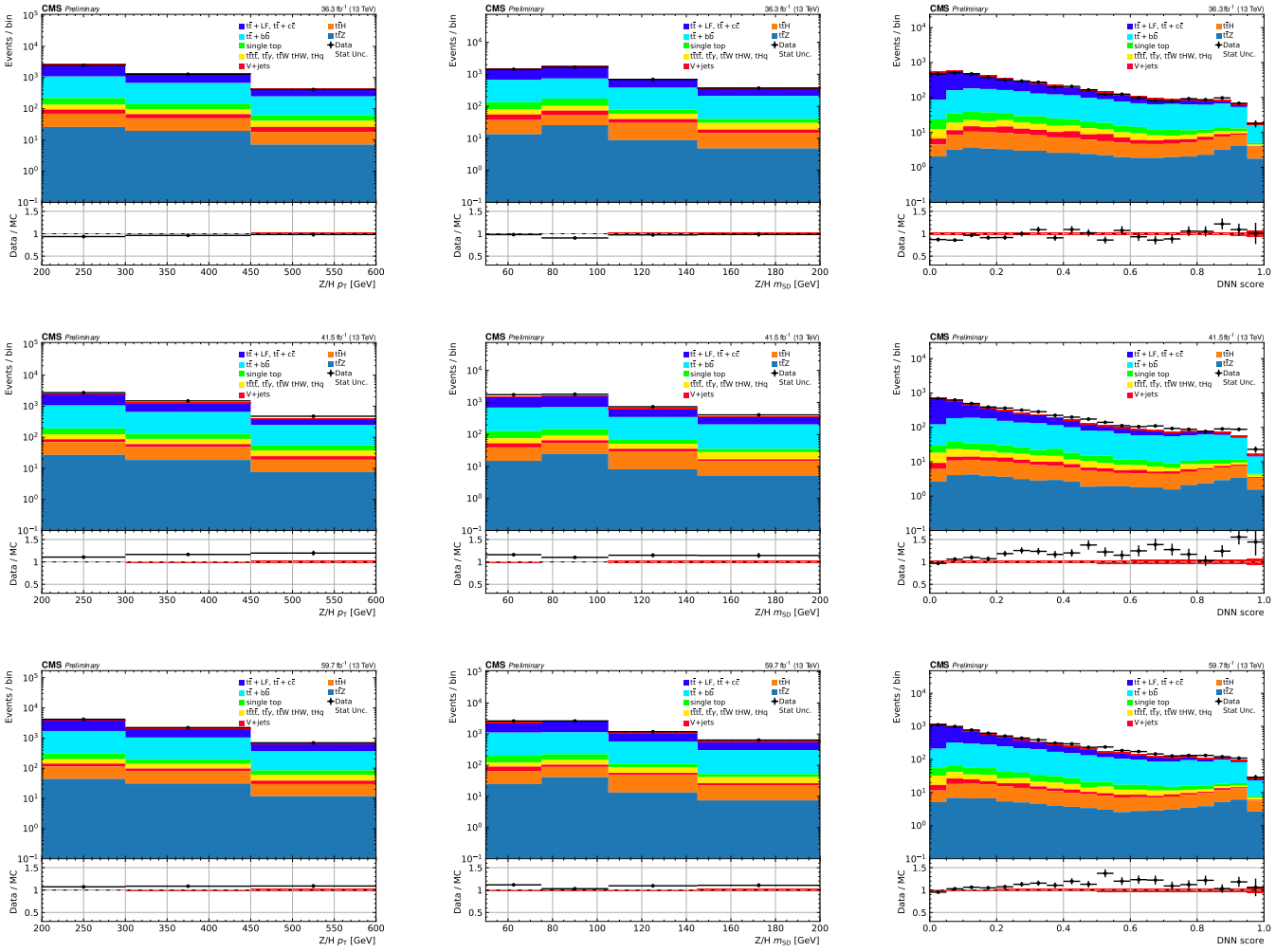

Figure 5.10: Data versus MC simulation in the distributions of the Z or Higgs boson candidate p_{T} (left column), the Z or Higgs boson candidate m_{SD} (middle column), and the DNN score (right column). Events passing the baseline selection requirements corresponding to the 2016 (top row), 2017 (middle row), and 2018 (bottom row) data-taking years.

Table 5.7: DNN right-most bin edges calibrated for the Z/Higgs boson candidate p_{T} (GeV) for 2016, 2017, and 2018.

Z/Higgs boson cand. p_{T} (GeV)	NN_1	NN_2	NN_3	NN_4	NN_5	NN_6
	2016					
$200 < p_{\mathrm{T}} < 300$	0.21	0.59	0.70	0.82	0.90	1.00
$300 < p_{\mathrm{T}} < 450$	0.32	0.67	0.78	0.86	0.92	1.00
$p_{\mathrm{T}} > 450$	0.40	0.72	0.80	0.87	0.92	1.00
	2017					
$200 < p_{\mathrm{T}} < 300$	0.18	0.53	0.65	0.80	0.88	1.00
$300 < p_{\mathrm{T}} < 450$	0.23	0.62	0.72	0.83	0.91	1.00
$p_{\mathrm{T}} > 450$	0.34	0.71	0.79	0.87	0.92	1.00
	2018					
$200 < p_{\mathrm{T}} < 300$	0.17	0.51	0.63	0.78	0.88	1.00
$300 < p_{\mathrm{T}} < 450$	0.21	0.58	0.70	0.82	0.90	1.00
$p_{\mathrm{T}} > 450$	0.31	0.69	0.78	0.87	0.92	1.00

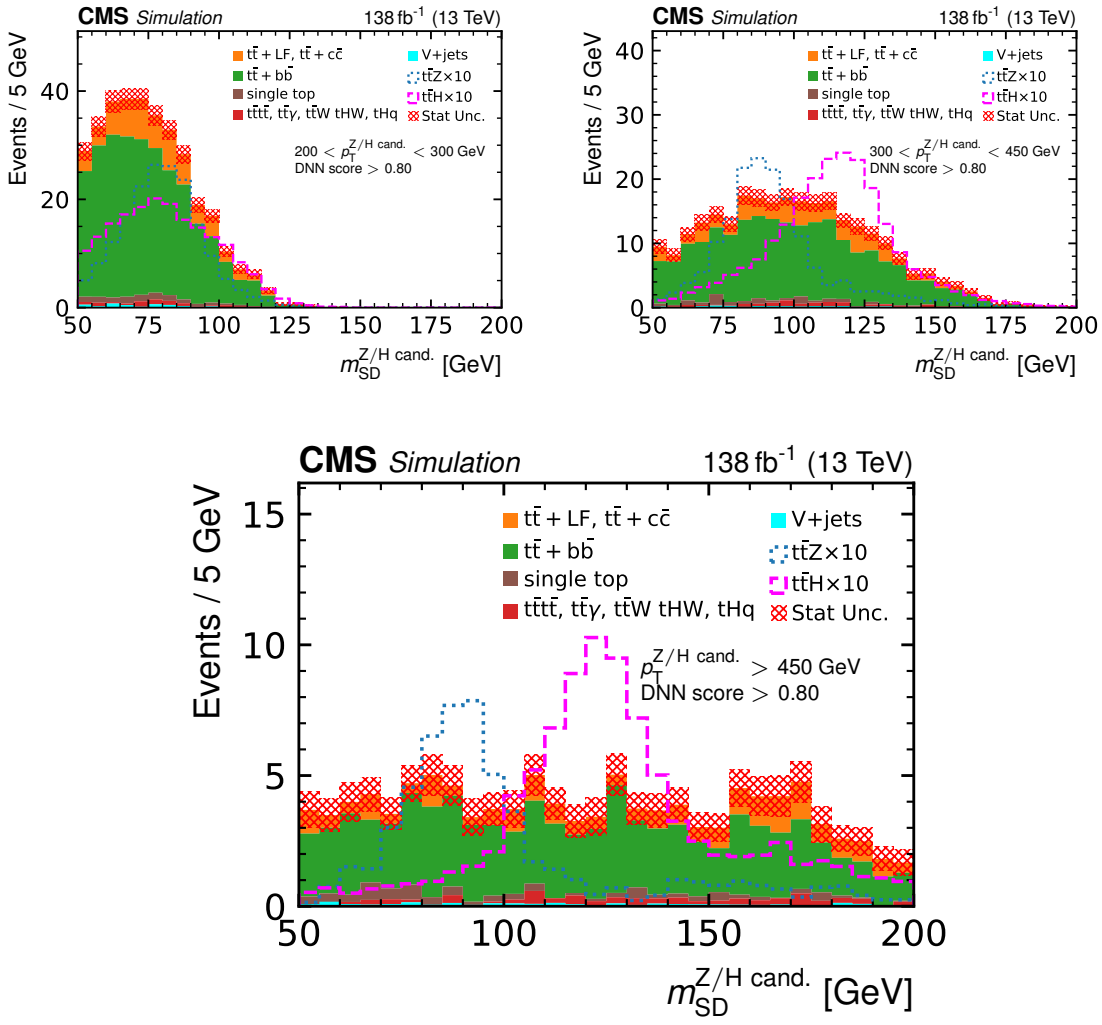

Figure 5.11: Soft-drop mass distributions of the Z or Higgs boson candidate in three p_{T} ranges of 200–300 GeV (upper), 300–450 GeV (middle), and above 450 GeV (lower) for Run 2 simulated data set with DNN score > 0.8. The $\mathrm{t\bar{t}Z}$ and $\mathrm{t\bar{t}H}$ distributions are scaled by 10 for visibility.

highest levels of signal purity correspond to the p_{T} intervals > 300 GeV, the center mass intervals $75 < m_{\mathrm{SD}} < 145$ GeV, and the top two NN score intervals. For technical reasons, the analysis templates are flattened from a three dimensional histogram to a one dimensional histogram. A cartoon depicting this procedure for each Z or Higgs boson candidate p_{T} interval is shown in Fig. 5.12. Figures 5.13–5.15 display the templates constructed for this analysis.

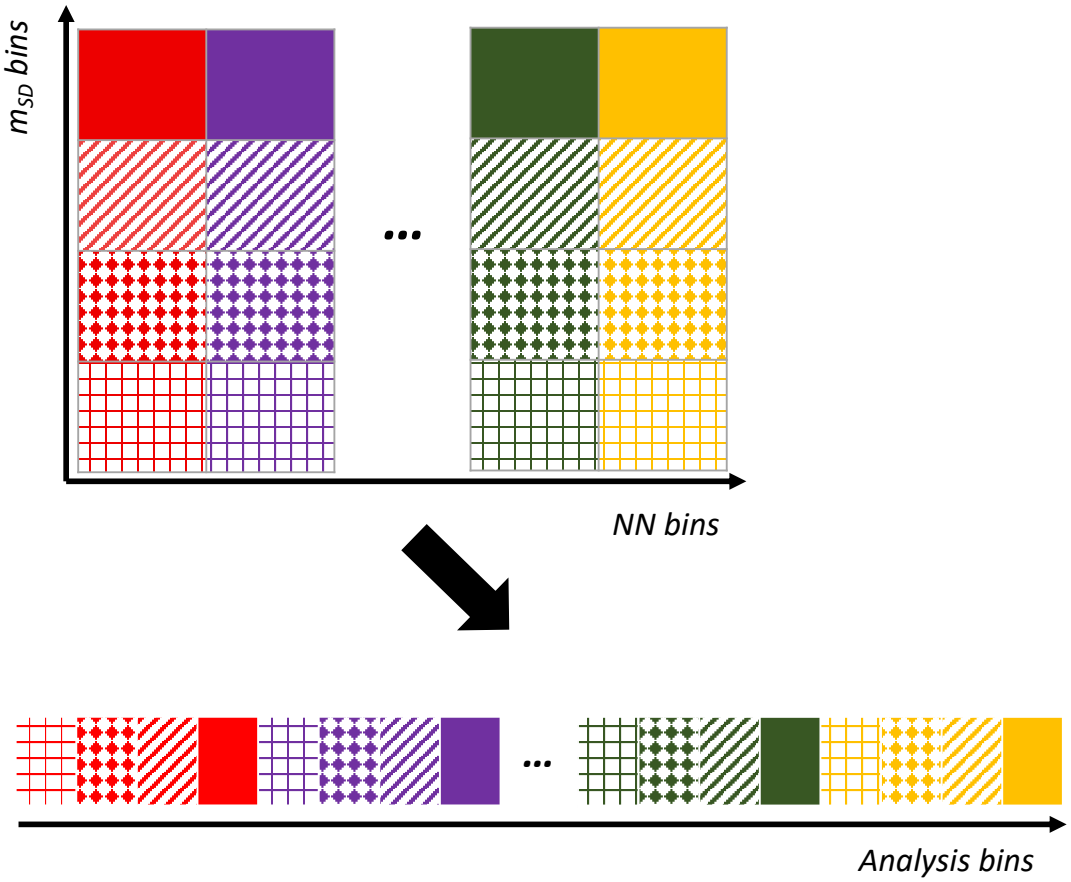

Figure 5.12: Cartoon illustrating the construction of the 1D analysis templates for each Z or Higgs boson candidate p_{T} interval from a 2D template histogram of m_{SD} vs. NN score.

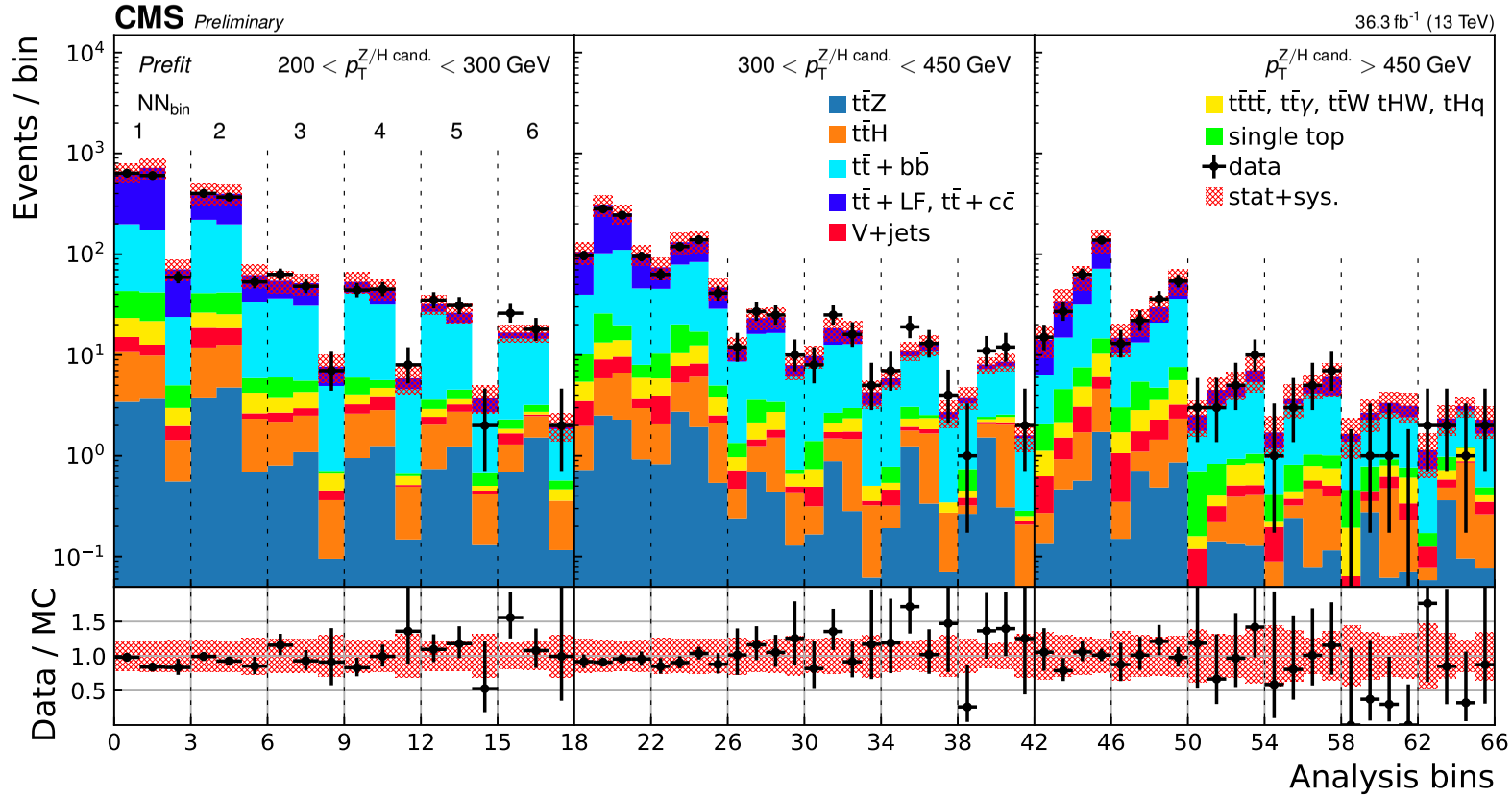

Figure 5.13: Analysis template with pre-fit expected and observed yields for the 2016 data-taking period. The analysis bins are defined as functions of the DNN score, and the p_{T} and m_{SD} of the Z/H boson candidate AK8 jet. The red hatch marks correspond to the total statistical and systematic uncertainty on the expected yield. The data points have a vertical error bar denoting the Poisson error.

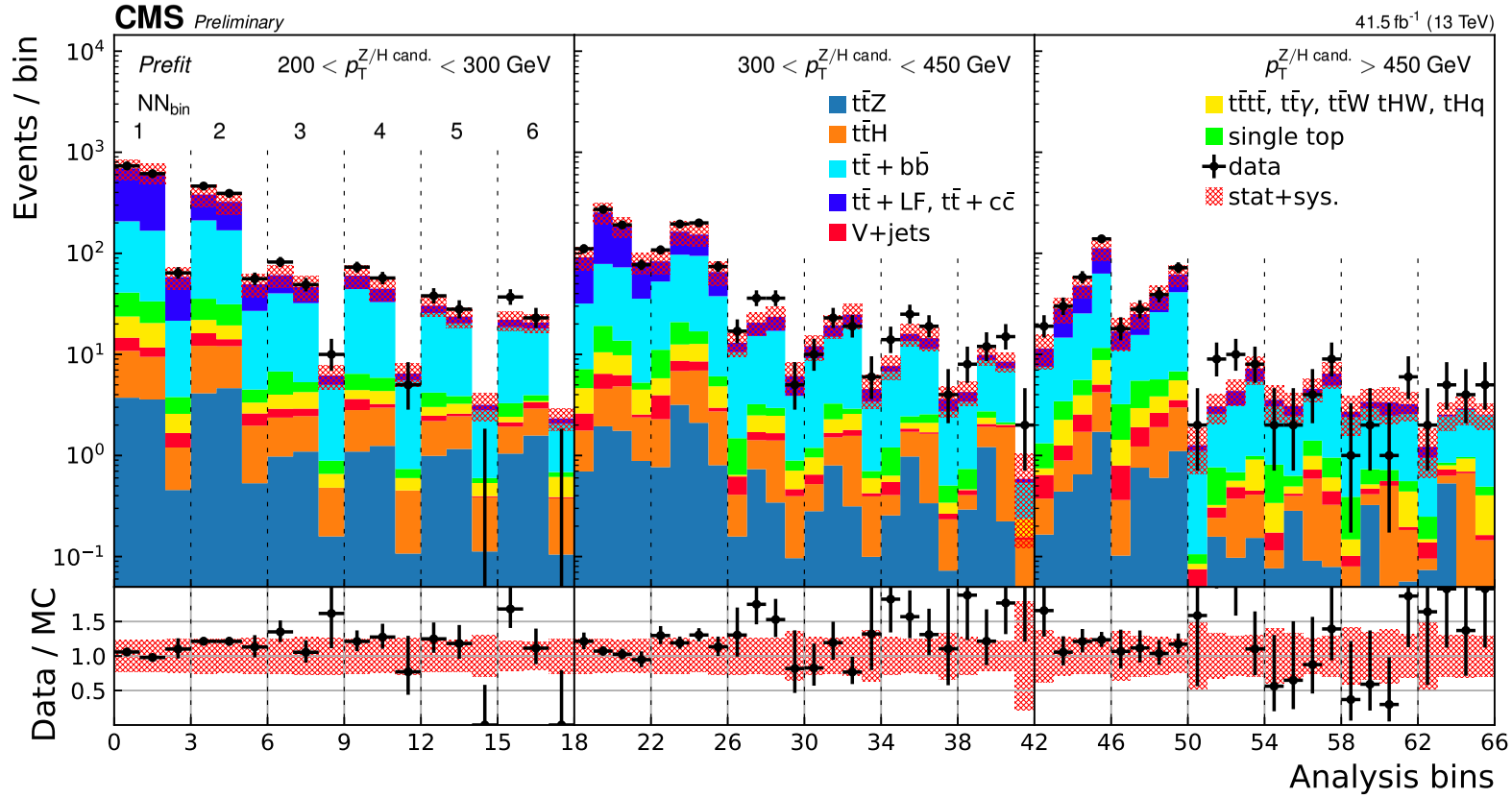

Figure 5.14: Analysis template with pre-fit expected and observed yields for the 2017 data-taking period. The analysis bins are defined as functions of the DNN score, and the p_{T} and m_{SD} of the Z/H boson candidate AK8 jet. The red hatch marks correspond to the total statistical and systematic uncertainty on the expected yield. The data points have a vertical error bar denoting the Poisson error.

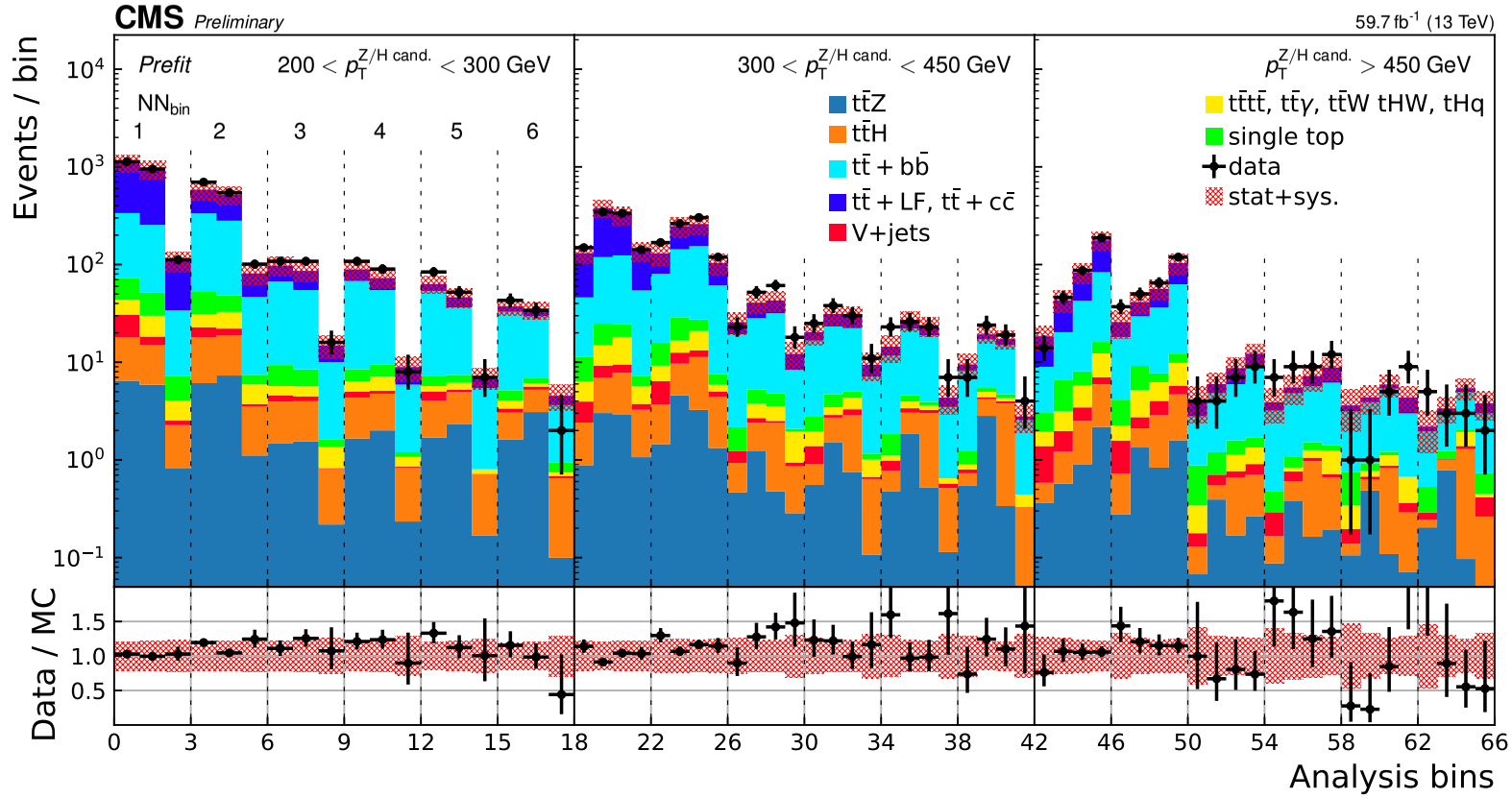

Figure 5.15: Analysis template with pre-fit expected and observed yields for the 2018 data-taking period. The analysis bins are defined as functions of the DNN score, and the p_{T} and m_{SD} of the Z/H boson candidate AK8 jet. The red hatch marks correspond to the total statistical and systematic uncertainty on the expected yield. The data points have a vertical error bar denoting the Poisson error.

5.5.2 *The Parameters of Interest*

The parameters of interest (POIs) are, simply put, the parameters the analysis is seeking to measure in the observed data. The POIs are free parameters which affect the predicted yields of certain SM processes such as $t\bar{t}H$ and $t\bar{t}Z$. In this way, the POIs may increase or decrease the production yield in the template bins to better fit the data. Their best-fit values and uncertainties are measured by maximizing the likelihood function of the data and simulation templates.

The signal strength modifier is the first type of POI the analysis measures. It corresponds to the ratio of the observed cross section of a signal process divided by the SM prediction. For this analysis, two signal strength modifiers are measured simultaneously. They are $\mu_{t\bar{t}H}$ and $\mu_{t\bar{t}Z}$, which are defined as

$$\mu_{t\bar{t}H} = \frac{\sigma(pp \to t\bar{t}H, p_T^H > 200 \text{ GeV})}{\sigma_{SM}(pp \to t\bar{t}H, p_T^H > 200 \text{ GeV})}, \tag{5.8}$$

$$\mu_{t\bar{t}Z} = \frac{\sigma(pp \to t\bar{t}Z, p_T^Z > 200 \text{ GeV})}{\sigma_{SM}(pp \to t\bar{t}Z, p_T^Z > 200 \text{ GeV})}, \tag{5.9}$$

where σ is the cross section. Here, signal is defined as $t\bar{t}Z$ or $t\bar{t}H$ where the respective Z or Higgs boson p_T is greater than 200 GeV in order to align with the reconstruction criteria of the Lorentz boosted boson. From the definition, the signal strength modifier can be used to measure the cross section of a process. Furthermore, the signal may be subdivided into contiguous intervals in the boson p_T to measure the cross section differentially. Using this approach, additional POIs are defined for the $t\bar{t}H$ and $t\bar{t}Z$ processes in the Z of Higgs boson p_T intervals corresponding to $200 < p_T^{Z/H} \leq 300$, $300 < p_T^{Z/H} \leq 450$, and $p_T^{Z/H} > 450$ GeV.

The other POIs this thesis considers are the Wilson coefficients introduced in Section 2.2. As outlined in Eq. (5.2), a parametrization of the EFT weight is

determined for every simulated event. Additionally, the magnitude of the w_{EFT} is a function of the WCs which are not known a priori. The events populate the analysis template bins, where the bin yields N_{EFT} are the sum of the event weights and are formulated as:

$$\begin{aligned} N_{\mathrm{EFT}}\left(\frac{\vec{c}}{\Lambda^2}\right) &= \sum_k w_{\mathrm{EFT},k}\left(\frac{\vec{c}}{\Lambda^2}\right) = \sum_k \left(s_{0k} + \sum_i s_{1ik}\frac{c_i}{\Lambda^2} + \sum_{i,j} s_{2ijk}\frac{c_i}{\Lambda^2}\frac{c_j}{\Lambda^2}\right) \\ &= S_0 + \sum_i S_{1i}\frac{c_i}{\Lambda^2} + \sum_{i,j} S_{2ij}\frac{c_i}{\Lambda^2}\frac{c_j}{\Lambda^2}, \end{aligned} \tag{5.10}$$

where $S_0 = \sum_k s_{0k}$, $S_{1i} = \sum_k s_{1ik}$, and $S_{2ij} = \sum_k s_{2ijk}$, and summing over k events. The bin yields N_{EFT}, which are modeled with LO precision, are normalized to the LO SM yield $N_{\mathrm{SM}} = N_{\mathrm{EFT}}\left(\frac{\vec{c}}{\Lambda^2} = 0\right)$ resulting in a factor which quantifies the multiplicative divergence from the SM prediction. This procedure allows EFT to be modeled in the existing analysis templates with signal generated at NLO precision by simply multiplying the expected signal yield by this factor:

$$N_{\mathrm{exp}}\left(\frac{\vec{c}}{\Lambda^2}\right) = N_{\mathrm{NLO\ SM}} \times \frac{N_{\mathrm{EFT}}\left(\frac{\vec{c}}{\Lambda^2}\right)}{N_{\mathrm{LO\ SM}}}. \tag{5.11}$$

In this way, the analysis templates are morphed according to the WC values which best fit the data. Importantly, this method of modeling the EFT effects is viable if and only if the NLO SM and LO SM signal yields demonstrate reasonable agreement in the template bins, and if the impacts on the signal process cross section from EFT are similar for simulation with NLO and LO precisions. A validation study was performed to confirm this, and more details are presented in Appendix B.

5.5.3 Maximum Likelihood Estimation

The binned likelihood function $L\left(\mu, \theta \mid D_i\right)$ is used to estimate the value of the POI and the CL interval of the estimation. The likelihood is constructed as the

product of Poisson probabilities associated with the bins from the analysis templates, outlined in Section 5.5.1. There are three main components of the probability function for i bins: the expected signal yield $S_i(\mu, \theta)$, the background estimation $B_i(\theta)$, and observed data D_i. The signal is a function of the POI μ, and both the signal and background estimations are a function of the so-called nuisance parameters θ (NPs). In the context of this work, μ represents the signal strength modifiers or the Wilson coefficients. These are associated with sources of systematic uncertainty in the background and signal modeling, and are generally encoded as a log-normal probability density function. More details about the systematic uncertainties considered in this thesis are covered in Section 5.6. The binned likelihood function is formulated as

$$L(\mu, \theta \mid D_i) = \prod_i \frac{\lambda_i(\mu, \theta)^{D_i} \, e^{-\lambda_i(\mu,\theta)}}{D_i!}, \tag{5.12}$$

$$\lambda_i(\mu, \theta) = S_i(\mu, \theta) + B_i(\theta). \tag{5.13}$$

The model parameter values which best fit the data, $\hat{\mu}$ and $\hat{\theta}$, are those which maximize the likelihood function. However this practice is computationally expensive, therefore the minimization of two times the negative log-likelihood $-2 \ln L$ is solved numerically with the Minuit tool [123].

5.5.4 Profiled Likelihood

The CL intervals of the POIs are determined by profiling the likelihood function. This technique calculates the ratio of the maximized likelihood and another maximized likelihood but computed at a fixed value for one or more POIs. Like before, it is computationally advantageous to solve the negative log of the likelihood ratio or equivalently the difference in the negative log likelihoods. This is formulated

as

$$q_\mu = -2 \ln \frac{L\left(\mu, \hat{\theta}_\mu \mid D_i\right)}{L\left(\hat{\mu}, \hat{\theta} \mid D_i\right)}, \tag{5.14}$$

where q_μ is the test statistic for a given value μ, and $\hat{\theta}_\mu$ is the nuisance parameter which maximizes the likelihood function for a fixed value of μ. Using this method, the test statistic is calculated for multiple values of the POI, and the resulting profile is utilized to determine the CL intervals.

5.5.5 Asymptotic Limits

When the POI is a signal strength of a SM process, it is unphysical for it to be negative. This is relevant when the experimental model is not sensitive enough to detect the signal process in question. In this scenario, the uncertainty of the signal strength modifier is modeled asymptotically in order to extract the 95% CL upper limit of the POI. The methodology for determining the upper limit of a POI compares the p-scores of the background only hypothesis p_b, against the signal plus background hypothesis p_μ. Both are defined as

$$p_\mu = \int_{q_\mu^{\text{obs}}}^{\infty} f\left(q_\mu \mid \mu, \hat{\theta}_\mu\right) dq_\mu, \tag{5.15}$$

$$1 - p_b = \int_{q_\mu^{\text{obs}}}^{\infty} f\left(q_\mu \mid 0, \hat{\theta}_0\right) dq_\mu, \tag{5.16}$$

where $f\left(q_\mu \mid \mu, \hat{\theta}_\mu\right)$ is the probability density function for the test statistic q_μ given μ and $\hat{\theta}_\mu$, and q_μ^{obs} is the observed test statistic. The upper limit of μ is determined by solving the inequality

$$\frac{p_\mu}{1 - p_b} \leq 1 - \alpha, \tag{5.17}$$

where α is the confidence level interval, e.g. solving for $\alpha = .95$ will result in the 95% CL interval upper limit of μ.

5.6 Systematic Uncertainties

Several sources of systematic uncertainty are included as nuisance parameters with a log-normal prior in the modeling of the signal and background predictions. These may affect the shape of the predictions in the analysis templates, the rate of the signal and background yields, or both. The sources of uncertainty belong to one of two major categories: experimental and theoretical. A brief outline of the sources including a description of the affected background or signal process, correlations across data-taking years, and the effect on the shape or rate of the prediction is provided in the following subsections and in Tables 5.8 and 5.9.

Additionally, statistical fluctuations due to the limited size of the MC samples are implemented as nuisance parameters in the likelihood function using the method described in Refs. [124, 125]. This approach will introduce a nuisance parameter with a Gaussian prior for every bin in the templates with an amount of unweighted events greater than a specified threshold. For this work, a threshold of 10 is chosen. For bins with a yield of less than or equal to 10, a set of nuisances for every signal and background process is included and modeled with a Poisson prior. This evaluation of the overall number of unweighted events in a bin does not include the contribution from expected signal events.

5.6.1 Theoretical Uncertainties

- *Cross Section (QCD Scales and PDF+α_s):* The signal and background processes are scaled according to theoretical predictions of the cross section of at least NLO accuracy. The uncertainties in the cross section are derived from the QCD scale and the choice of renormalization μ_{R} and factorization μ_{F}

Table 5.8: Theoretical sources of uncertainty considered in the analysis. "Type" refers to rate (R), shape (S), or both (R+S) uncertainties. "Corr." indicates whether the uncertainty is treated as correlated (C), partially correlated (P), or uncorrelated (U) across the years 2016–2018.

Source	Type	Corr.	Description
QCD scales	R	C	Scale uncertainty of (N)NLO prediction, independent for $t\bar{t}H$, $t\bar{t}Z$, $t\bar{t}$ + jets, single top, $t(\bar{t})$ + X, V + jets
PDF+α_s (gg)	R	C	PDF uncertainty for gg initiated processes, independent for $t\bar{t}H$, $t\bar{t}Z$
PDF+α_s ($q\bar{q}$)	R	C	PDF uncertainty of $q\bar{q}$ initiated processes
PDF+α_s (qg)	R	C	PDF uncertainty of qg initiated processes
μ_R scale	S	C	Renormalization scale uncertainty of the ME generator, independent for $t\bar{t}H$, $t\bar{t}Z$, $t\bar{t} + b\bar{b}$ (4FS), other $t\bar{t}$ + jets (5FS)
μ_F scale	S	C	Factorization scale uncertainty of the ME generator, independent for $t\bar{t}H$, $t\bar{t}Z$, $t\bar{t} + b\bar{b}$ (4FS), other $t\bar{t}$ + jets (5FS)
PDF+α_s shape	S	C	From NNPDF variations, independent for $t\bar{t} + b\bar{b}$ (4FS), and $t\bar{t}Z$, $t\bar{t}H$, and other $t\bar{t}$ + jets (5FS)
PS scale ISR	S	C	Initial state radiation uncertainty of the PS (PYTHIA), independent for $t\bar{t}Z$, $t\bar{t}H$, $t\bar{t} + b\bar{b}$ (4FS) and other $t\bar{t}$ + jets (5FS)
PS scale FSR	S	C	Final state radiation uncertainty of the PS (PYTHIA), independent for $t\bar{t}Z$, $t\bar{t}H$, $t\bar{t} + b\bar{b}$ (4FS), and other $t\bar{t}$ + jets (5FS)
ME-PS matching ($t\bar{t}$)	R	C	NLO ME to PS matching (for $t\bar{t}$ + jets events), independent for $t\bar{t} + b\bar{b}$ (4FS), and other $t\bar{t}$ + jets (5FS)
Underlying event ($t\bar{t}$)	R	C	Underlying event (for all $t\bar{t}$ + jets events)
Top p_T reweighting	R+S	C	Correction to top quark p_T spectra in $t\bar{t}$+ jets simulation
Cross section of $t\bar{t} + c\bar{c}$	R	C	A 50% uncertainty on the rate of $t\bar{t} + c\bar{c}$ within $t\bar{t}$ + jets (5FS)
Cross section of $t\bar{t} + b\bar{b}$	R	C	A freely floating rate for the $t\bar{t}+b\bar{b}$ (4FS) cross section
Collinear gluon splitting	R	C	Additional 50% rate uncertainty on $t\bar{t}$ + 2b (a subprocess of $t\bar{t} + b\bar{b}$ events)

Table 5.9: Experimental sources of uncertainty considered in the analysis. "Type" refers to rate (R) or shape (S) uncertainties. "Corr." indicates whether the uncertainty is treated as correlated (C), partially correlated (P), or uncorrelated (U) across the years 2016–2018.

Source	Type	Corr.	Description
Integrated luminosity	R	P	Signal and all backgrounds
Lepton ID/Iso_{mini}	R+S	U	Signal and all backgrounds
Trigger efficiency	R+S	U	Signal and all backgrounds
Pileup	R+S	U	Signal and all backgrounds
Jet energy scale	R+S	P	$t\bar{t}Z$, $t\bar{t}H$, $t\bar{t}$ + jets and single top
Jet energy resolution	R+S	U	$t\bar{t}Z$, $t\bar{t}H$, $t\bar{t}$ + jets and single top
Ak8 jet mass scale	R+S	U	$t\bar{t}Z$, $t\bar{t}H$, $t\bar{t}$ + jets and single top
Ak8 jet mass resolution	R+S	U	$t\bar{t}Z$, $t\bar{t}H$, $t\bar{t}$ + jets and single top
b jet tag HF fraction	R+S	C	Signal and all backgrounds
b jet tag LF fraction	R+S	C	Signal and all backgrounds
b jet tag HF/LF stat (linear)	R+S	U	Signal and all backgrounds
b jet tag HF/LF stat (quadratic)	R+S	U	Signal and all backgrounds
b jet tag charm (linear)	R+S	C	Signal and all backgrounds
b jet tag charm (quadratic)	R+S	C	Signal and all backgrounds
$b\bar{b}$ jet tag	R+S	U	Signal and all backgrounds

scales in the matrix-element calculation, and is also derived from the choice of PDF set and α_s. Furthermore, the PDF+α_s uncertainty is split up according to participating partons in the hard scattering. For background processes which share a common set of initiating partons, their PDF+α_s uncertainty is treated as fully correlated. Otherwise, a set of independent nuisance parameters are assigned according to signal and background processes for both the QCD scale and PDF+α_s uncertainties. Additionally, the nuisance parameters are treated as correlated between the years, and are summarized in Tables 5.10 and 5.11.

- μ_R, μ_F *Scales:* The choice of renormalization scale μ_R and the factorization scale μ_F can affect the kinematic properties in the signal, $t\bar{t} + b\bar{b}$, and other

Table 5.10: The inclusive cross section uncertainties due to the QCD scale and the choice of renormalization and factorization scales in the matrix-element calculation on the signal and the background processes.

Process	QCD scale
$t\bar{t}H$	$+5.8\%/-9.2\%$
$t\bar{t}Z$	$+8.1\%/-9.3\%$
$t\bar{t}$ + jets	$+2.4\%/-3.5\%$
$t(\bar{t})$ + X $^{(\dagger)}$	$+18.1\%/-12.5\%$
V + jets	$+0.8\%/-0.4\%$
single top	$+3.1\%/-2.1\%$

$^{(\dagger)}$: Estimated from the envelope of the μ_R and μ_F scales in the MC simulation

Table 5.11: The uncertainties on the cross section of signal and background processes due to the choice of PDF+α_s and associated with the interacting partons. The uncertainties for the same initial state (same column) and for different processes (different rows) are treated as fully correlated.

	$gg_{t\bar{t}H}$	$gg_{t\bar{t}Z}$
$t\bar{t}H$	$\pm3.6\%$	
$t\bar{t}Z$		$\pm3.5\%$

	gg	$q\bar{q}$	qg
$t\bar{t}$ + jets	$\pm4.2\%$		
$t(\bar{t})$ + X $^{(\dagger)}$		$\pm4.5\%$	
V + jets		$\pm3.8\%$	
single top			$\pm2.8\%$

$^{(\dagger)}$: Estimated from the PDF+α_s set in the MC simulation

$t\bar{t}$ + jets processes. These are modeled as uncertainties by varying the scales independently by a factor of 0.5 or 2, and propagating the variations of the shape to the templates in the fit. The respective scale variations are accessed via a set of alternative weights in the MC simulation. Since the normalization uncertainties of the matrix-element generator are covered by the (N)NLO

cross section uncertainties (Table 5.10), only the shape or acceptance variation of the shape distributions is considered here, i.e. the variations are scaled to retain the overall normalization in the inclusive phase-space. Additionally, changes in the relative fraction of $t\bar{t} + b\bar{b}$ events in the inclusive $t\bar{t}$ + jets phase-space are considered pre-fit based on the composition of events in the $t\bar{t}$+jets (5FS) MC sample for these variations. The set of nuisance parameters derived from the variations on the shape are treated as uncorrelated among the $t\bar{t}H$, $t\bar{t}Z$, $t\bar{t}+b\bar{b}$, and other $t\bar{t}$+jets (5FS) processes, and correlated across the data-taking years.

- *PDF+α_s Shape:* The shape variation of the signal and background processes due to the choice of PDF set and α_s is treated as a set of nuisance parameters in the fit. The uncertainty on the PDF is estimated as the RMS of all residuals for the `NNPDF31_nnlo_as_0118_nf_4` PDF set, and as the quadratic sum of all residuals for the `NNPDF31_nnlo_hessian_pdfas` PDF set, following the definition of the PDF variations in each case. The `NNPDF31_nnlo_as_0118_nf_4` set is used for the $t\bar{t} + b\bar{b}$ (4FS) sample and the `NNPDF31_nnlo_hessian_pdfas` set is used for the $t\bar{t}H$, $t\bar{t}Z$, and other $t\bar{t}$ + jets (5FS) processes. The PDF set uncertainties are included in the fit as two independent nuisances parameters based on the PDF set used. Additionally, both NNPDF sets include variations on the α_s scale. Since the overall normalization uncertainties of the PDF and α_s are covered by the (N)NLO cross section uncertainties (Table 5.11), only the shape or acceptance variation on the simulation is considered.
- *PS Scales (ISR/FSR):* The simulation of the parton shower is varied by changing the scales which govern the ISR and FSR by a factor of 0.5 and 2. The

variations are stored in the MC simulation as alternative weights and are treated as uncertainties in the fit. Since the normalization uncertainties of the matrix-element generator are covered by the (N)NLO cross section uncertainties (Table 5.10), only the shape or acceptance variation is considered here. Variations in the ISR and FSR will alter the relative fraction of $t\bar{t}+b\bar{b}$ events within $t\bar{t}+$jets. This is considered pre-fit by propagating this effect from $t\bar{t}+b\bar{b}$ within the $t\bar{t}+$jets (5FS) sample to the $t\bar{t}+b\bar{b}$ (4FS) sample. The ISR and FSR uncertainties are treated as uncorrelated among the $t\bar{t}+b\bar{b}$ (4FS) and other $t\bar{t}+$jets (5FS) processes.

- *ME-PS Matching:* For the POWHEG $t\bar{t}+$jets (5FS) and $t\bar{t}+b\bar{b}$ (4FS) MC samples, the uncertainty of the matching between the matrix-element generator and the parton shower are estimated by varying the $h_{\text{damp}} = 1.379^{+0.926}_{-0.5052}\, m_t$ parameter. Two dedicated MC samples for $t\bar{t}+b\bar{b}$ and $t\bar{t}+$jets with $h_{\text{damp}} = 2.305\, m_t$ and $h_{\text{damp}} = 0.874\, m_t$ are propagated to the analysis templates to determine the uncertainty. However, the limited number of events in the MC samples leads to large statistical fluctuations in the bins of the templates. Instead, the overall rate variation on the distributions is used to model the uncertainty due to variations in the h_{damp} setting. There are two independent nuisance parameters for the ME-PS matching uncertainty included in the fit, one for the $t\bar{t}+b\bar{b}$ (4FS) process and another for the other $t\bar{t}+$jets (5FS) processes.
- *Underlying Event:* Variations in the modeling of the underlying event tune are treated as uncertainties in the fit. Similar to the ME-PS uncertainty, dedicated MC samples with alternative tune are propagated to the analysis

templates to estimate the uncertainty. Since there are no dedicated samples available for the $t\bar{t} + b\bar{b}$ (4FS) samples, the relative effect of the UE tune is derived from the fraction of $t\bar{t} + b\bar{b}$ events in the $t\bar{t} +$ jets (5FS) sample. This treatment is reasonable since the variations of the UE tune affect the soft-particle regime and should not depend on the simulation of the hard process, substantially. The MC samples with varied UE tune are limited in the number of events, so only the overall rate effect on the $t\bar{t} +$ jets process is used to determine the uncertainty. A single nuisance parameter is included in the fit for the $t\bar{t} +$ jets processes.

- *Top p_T Reweighting:* Differences in the p_T spectra of the top quarks between data and simulation have been observed [126]. Therefore, a theory-based correction is applied as an event reweighting, as detailed in Section 4.2.3.2. The magnitude of this correction is taken as the uncertainty and only affects the $t\bar{t} +$ jets background.
- *Cross Section of the $t\bar{t} + c\bar{c}$ Process:* The cross section of $t\bar{t}$ with at least two additional jets coming from a charm quark is measured with an uncertainty of approximately 20% [127]. However, we assign a larger cross section uncertainty of 50% to account for differences in our categorization of $t\bar{t} + c\bar{c}$ production which includes one or more charm jets, and also differences in phase space with respect to Ref. [127]. This uncertainty is treated as fully correlated among the years.
- *Cross Section and Modeling of $t\bar{t} + b\bar{b}$ Process:* Due to the uncertain cross section of the $t\bar{t} + b\bar{b}$ process, the normalization of $t\bar{t} + b\bar{b}$ is allowed to freely float in the fit. Also, discrepancies between data and simulation in

the frequency of the production $t\bar{t} + b\bar{b}$ events where the gluon splits into very collinear b-jets, $t\bar{t} + 2b$, have been studied [128]. The conclusions in Ref. [128] motivate a 50% uncertainty on the rate of the subset of $t\bar{t} + b\bar{b}$ events consistent with $t\bar{t} + 2b$. This uncertainty is treated as fully correlated among the years.

5.6.2 Experimental Uncertainties

- *Luminosity:* The expected yields in the analysis templates are scaled using the integrated luminosity measured each data-taking year. There are several sources of uncertainty associated with the luminosity estimate with varying degrees of magnitude and different correlations across the years [129–131]. A summary of the uncertainties and correlations for every data-taking year is summarized in Table 5.12. A set of nuisance parameters are introduced in the fit and affect the overall yield of the analysis templates.

Table 5.12: The uncertainty on the luminosity estimate per data-taking year as a percentage. Some uncertainties are correlated or not correlated between years as indicated below.

Year	2016	2017	2018
Uncorrelated 2016	1.0	0.0	0.0
Uncorrelated 2017	0.0	2.0	0.0
Uncorrelated 2018	0.0	0.0	1.5
Correlated 2016-2018	0.6	0.9	2.0
Correlated 2017-2018	0.0	0.6	0.2

- *Lepton Identification and Isolation:* The corrections related to the identification and isolation requirements of the muons and electrons have an associated p_{T} and η dependent uncertainty. This affects both the rate and the shape of

the simulated signal and background. Two sets of nuisance parameters, one set for each lepton flavor, are included in the fit and are uncorrelated across the data-taking years due to changes in the detector configuration and the reconstruction.

- *Trigger Efficiency:* The uncertainty in the calculation of the trigger efficiency correction factor, discussed in Section 5.3.1 and Appendix C, is included in the fit. This is implemented as two sets of nuisance parameters for electrons and muons, which are uncorrelated across data-taking years due to the changes in the trigger definition. An additional nuisance parameter is associated to the uncertainty of the "ECAL L1 Prefiring Issue" correction mentioned in Section 4.2.3.2. This uncertainty only affects the 2016 and 2017 data-taking years. The uncertainties relating to the trigger efficiency affect the rate and shape of the simulated processes.
- *Pileup:* Effects due to the uncertainty in the distribution of the number of pileup interactions are evaluated by varying the cross section used to predict the number of pileup interactions in the MC simulation by 4.6% from its nominal value. The variations affect the rate and shape, and are propagated to the analysis templates. The set of nuisance parameters representing this uncertainty are uncorrelated across data-taking years in order to account for differences in data and simulation in the event vertex multiplicity distribution.
- *Jet Energy Scale (JES):* Introduced in Section 4.2.2.5, the uncertainty in the JES is evaluated by shifting the jet energy scale applied to the reconstructed jets in the MC simulation. These impact the kinematic properties of both the collection of AK4 and AK8 jets, and affect the rate and shape of the

signal and background simulation. Several additional nuisance parameters are implemented in the fit based on the 11 independent subsources of the JES uncertainty with varying degrees of correlation across the years. An additional source of uncertainty stemming from the HEM issue in 2018 is applied. This uncertainty carries a 20% energy variation for jets with $-2.5 < \eta < -1.3$, and $-1.57 < \phi < -0.87$, and a 35% variation for jets with $-3.0 < \eta < -2.5$, and $-1.57 < \phi < -0.87$.

- *Jet Energy Resolution (JER):* Introduced in Section 4.2.2.5, observed differences between the energy resolution of AK4 and AK8 jets between the data and simulation necessitates a smearing of the jet energy. The smearing is varied within its uncertainty independently for AK4 and AK8 jets, and the rate and shape effects are propagated to the analysis templates. A set of nuisance parameters, independent by year, are included in the fit.
- *Jet Mass Corrections (JMS, JMR):* Similar to the JES and JER corrections, the scale (JMS) and resolution (JMR) of the softdrop mass of the AK8 jets are corrected based on observed differences in data and simulation. An independent set of nuisance parameters, independent by year and association to the JMS and JMR corrections, impact the rate and shape of the simulated signal and background.
- *b-tag Efficiency:* The efficiency of the DeepCSV tagger, used to identify b-jets as covered in Section 4.2.2.6, is different in simulation and in data. Because the DeepCSV tagger score is used as an input to the DNN, the shape of the score distribution is corrected based on the flavor properties and kinematics of the AK4 jet. Several uncertainties associated to the correction is included

in the fit as nuisance parameters. Two are related to varying the purity of light-flavor (g, u, d, s) and heavy-flavor (c, b) jets in the data set used to derive the correction. These are correlated across the data-taking years. Four additional sets of nuisance parameters account for the statistical uncertainty associated with the size of the MC samples utilized in the correction factor derivation. For these, linear and quadratic distortions due to the finite size of the samples for both light-flavor and heavy-flavor simulated data sets are accounted for. All four uncertainties are estimated independently for each year and are therefore uncorrelated across data-taking years. Lastly, two additional nuisance parameters related to the purity of the heavy-flavor data set with respect to charms quarks are included in the fit and are correlated across data-taking years. All nuisance parameters related to the b-tag efficiency affect the rate and the shape of the simulated processes.

- $b\bar{b}$-*tag Efficiency:* The efficiency of the DeepAK8 $b\bar{b}$ tagging algorithm, used to identify a Z or Higgs boson candidate (Section 5.3.2), is corrected in the simulation to better match the data. The uncertainty of this procedure is included as a nuisance parameter in the fit which affects both the rate and the shape of the simulated processes. Additionally, the set of nuisance parameters related to the $b\bar{b}$ tagging efficiency are uncorrelated across the data-taking years.

5.7 Results

Utilizing the signal extraction techniques discussed in Section 5.5, and accounting for the systematic uncertainties enumerated in Section 5.6, the signal strengths

of the boosted $t\bar{t}Z$ and $t\bar{t}H$ processes are measured, 95% CL upper limits are placed on the differential $t\bar{t}Z$ and $t\bar{t}H$ cross sections, and the values of WCs associated with dimension-six EFT operators are constrained. This is accomplished by maximizing the binned likelihood function in Eq. (5.12), and determining the model parameters' best-fit values. The analysis templates after the fit to data, with signal fixed to the SM prediction, are illustrated in Figs. 5.16–5.18.

5.7.1 Signal Strengths and Upper Limits on the Differential Cross Sections

While the main focus of this analysis is to constrain new physics in EFT, a fit is performed on the SM hypothesis of the templates. This is done for a few reasons. First and foremost, the boosted regime of $t\bar{t}H$ and $t\bar{t}Z$ production has been largely unexplored in the past due to limited sensitivities. Therefore, a topic of interest is to quantify the sensitivity that can be achieved with the full Run 2 data set. Secondly, the observed limits on the signal provides some additional interpretability of the constraints on the WCs. Lastly, the SM assumption is used for validation studies of the model and the fit. More information regarding this is found in Appendix D.

The signal strength modifiers measured in this work correspond to the $t\bar{t}Z$ and $t\bar{t}H$ processes with a rapidity requirement $|y_{Z/H}| < 2.5$ in accordance with the STXS definition [93]. Also, the heavy boson at generator-level has a sufficient Lorentz boost $p_T^{Z/H} > 200$ GeV. The remaining $t\bar{t}Z$ and $t\bar{t}H$ is fixed to the SM prediction because it is not feasible to reconstruct the Z or Higgs boson as a boosted object. The results are shown in Table 5.13 and Fig. 5.19, and the impacts from major sources of systematic uncertainty on $\mu_{t\bar{t}Z}$ and $\mu_{t\bar{t}H}$ are listed in Table 5.14. Additionally, the correlations between the signal strengths and the normalization of the $t\bar{t}$ + jets background are

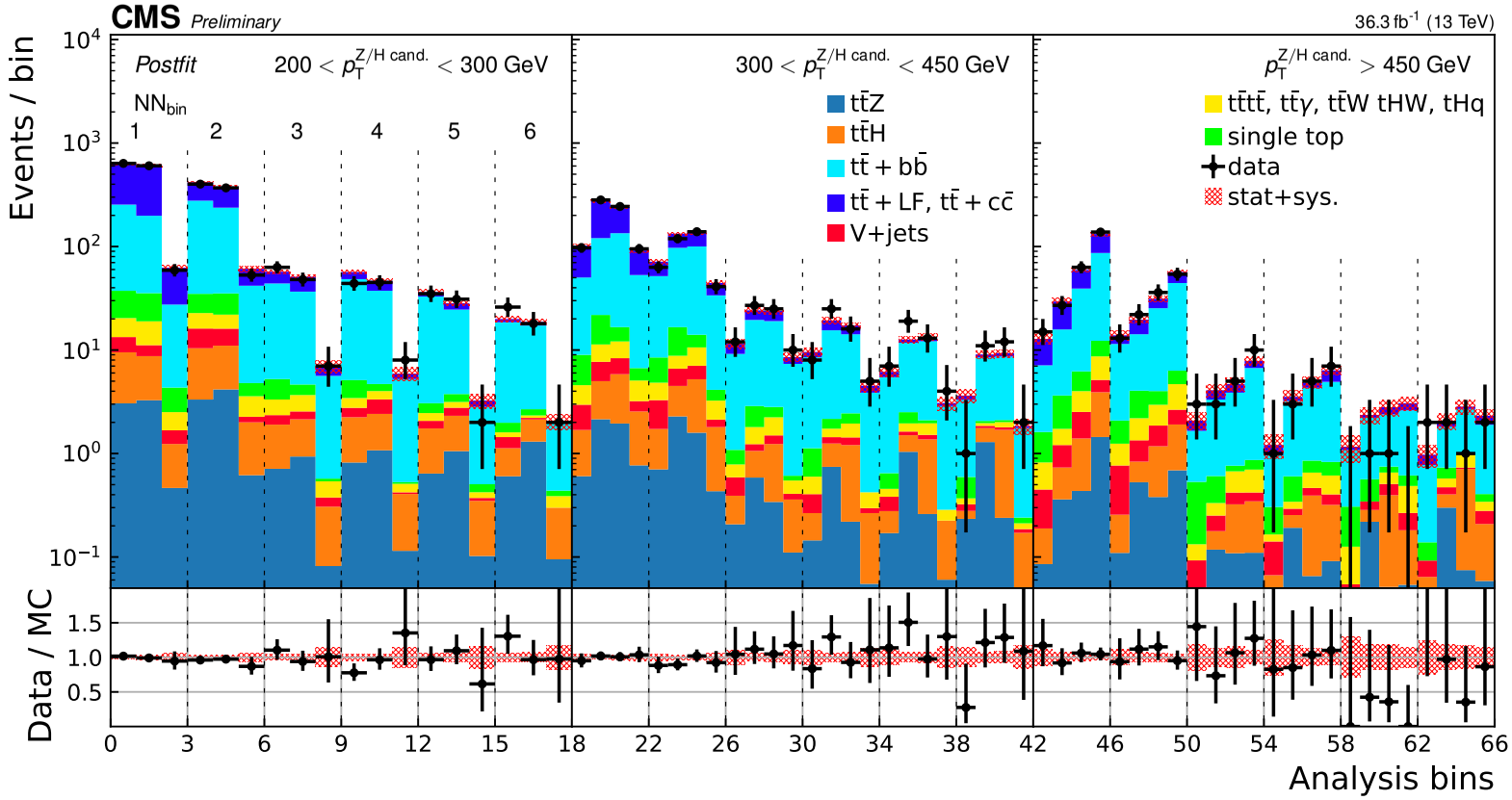

Figure 5.16: Post-fit expected and observed yields for the 2016 data-taking period in each analysis bin. In the fit, the $t\bar{t}Z$ and $t\bar{t}H$ signal cross sections are fixed to the SM predictions. The analysis bins are defined as functions of the DNN score, and the p_T and m_{SD} of the boson candidate AK8 jet. The largest groupings of bins are defined by the AK8 jet p_T. The smaller groups are defined by the DNN score, and the smallest groups, 3 or 4 bins with the same p_T and DNN score, correspond to the AK8 jet m_{SD}.

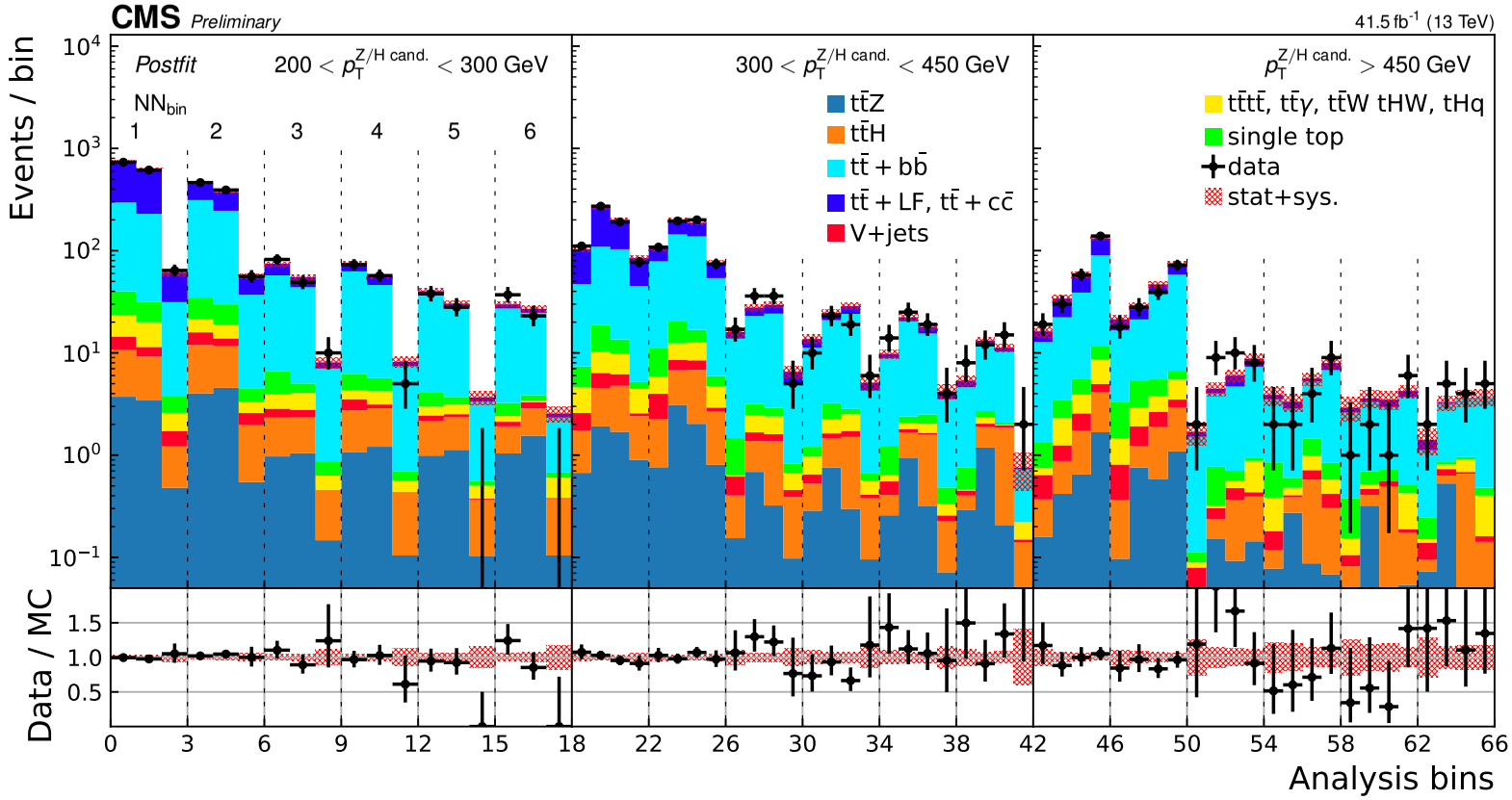

Figure 5.17: Post-fit expected and observed yields for the 2017 data-taking period in each analysis bin. In the fit, the $t\bar{t}Z$ and $t\bar{t}H$ signal cross sections are fixed to the SM predictions. The analysis bins are defined as functions of the DNN score, and the p_T and m_{SD} of the boson candidate AK8 jet. The largest groupings of bins are defined by the AK8 jet p_T. The smaller groups are defined by the DNN score, and the smallest groups, 3 or 4 bins with the same p_T and DNN score, correspond to the AK8 jet m_{SD}.

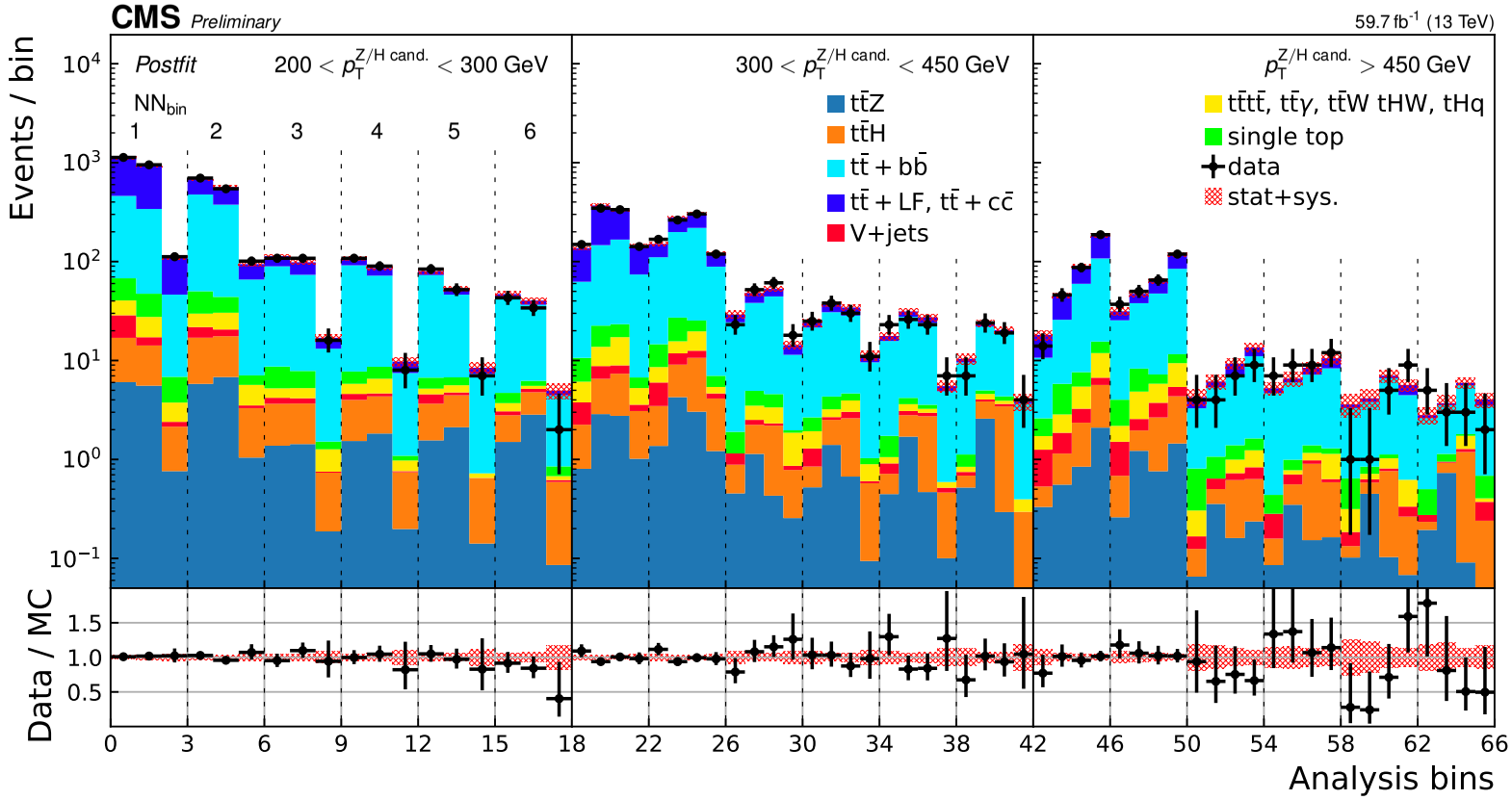

Figure 5.18: Post-fit expected and observed yields for the 2018 data-taking period in each analysis bin. In the fit, the $t\bar{t}Z$ and $t\bar{t}H$ signal cross sections are fixed to the SM predictions. The analysis bins are defined as functions of the DNN score, and the p_T and m_{SD} of the boson candidate AK8 jet. The largest groupings of bins are defined by the AK8 jet p_T. The smaller groups are defined by the DNN score, and the smallest groups, 3 or 4 bins with the same p_T and DNN score, correspond to the AK8 jet m_{SD}.

shown in Table 5.15. Both of the $t\bar{t}Z$ and $t\bar{t}H$ signal strengths are fit lower than what is expected, and $\mu_{t\bar{t}H}$ has a best-fit value below zero which is unphysical. The negative $t\bar{t}H$ signal strength is due to lower event yields in data than what is predicted in the bins which are most sensitive to the $t\bar{t}H$ signal. However as indicated by Table 5.13, the uncertainty of this result is dominated by the statistical limitations of the data. As more data are collected and analyzed, the sensitivity to observing $t\bar{t}Z$ and $t\bar{t}H$ will increase. The observed yield for the $t\bar{t} + b\bar{b}$ background is higher than what is expected, however this is consistent with the findings in other analysis measurements performed by the CMS Collaboration [127, 132, 133].

Table 5.13: The expected and observed best-fit signal strength modifiers $\mu_{t\bar{t}Z}$ and $\mu_{t\bar{t}H}$ for simulated Z or Higgs boson $p_T > 200$ GeV. The observed uncertainties are broken down into the components arising from the limited size of the data, the limited size of the simulation samples, experimental uncertainties, and theoretical uncertainties.

Signal strength	Observed	Stat.	MC Stat.	Experiment	Theory	Expected
$\mu_{t\bar{t}Z}$	$0.65^{+1.05}_{-0.98}$	$^{+0.80}_{-0.76}$	$^{+0.37}_{-0.38}$	$^{+0.38}_{-0.31}$	$^{+0.42}_{-0.38}$	$1.00^{+0.92}_{-0.84}$
$\mu_{t\bar{t}H}$	$-0.33^{+0.87}_{-0.85}$	$^{+0.72}_{-0.65}$	$^{+0.32}_{-0.34}$	$^{+0.19}_{-0.17}$	$^{+0.30}_{-0.38}$	$1.00^{+0.79}_{-0.73}$

In addition to the signal strength modifier, the 95% CL upper limits are placed on the differential cross sections for the production of $t\bar{t}H$ and $t\bar{t}Z$ as a function of the Higgs or Z boson p_T. The upper limits are extracted from the analysis templates utilizing a maximum-likelihood unfolding technique as described in Ref. [134]. The same rapidity requirements that are used in the measurement of $\mu_{t\bar{t}H}$ and $\mu_{t\bar{t}Z}$ are imposed here, while also fixing the subset of signal with $p_T^{Z/H} \leq 200$ GeV to the SM prediction during the fit to data. The $t\bar{t}Z$ and $t\bar{t}H$ signal events passing these criteria

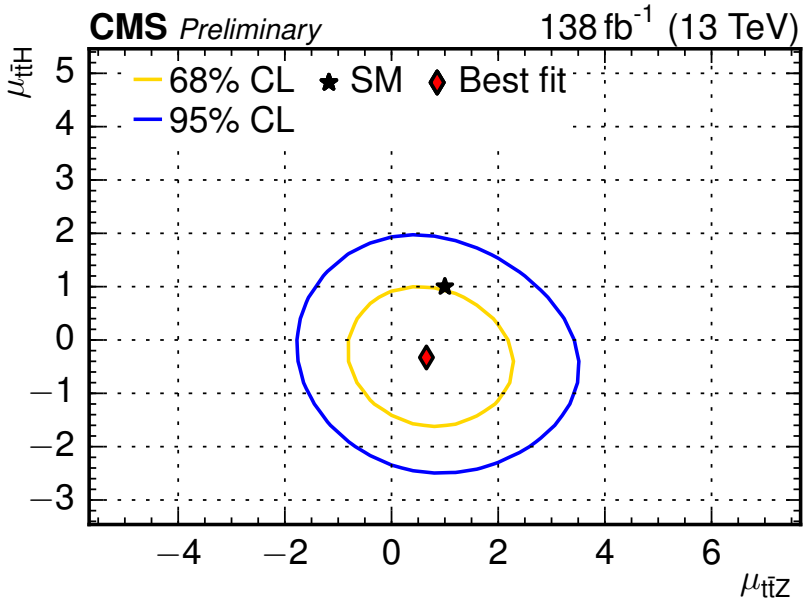

Figure 5.19: The observed best-fit signal strength modifiers $\mu_{t\bar{t}H}$ versus $\mu_{t\bar{t}Z}$ for simulated Higgs or Z boson $p_T > 200$ GeV. The contours show the 68% and 95% confidence level regions.

Table 5.14: Major sources of uncertainty in the measurement of the signal strength modifiers $\mu_{t\bar{t}Z}$ and $\mu_{t\bar{t}H}$ for simulated Z or Higgs boson $p_T > 200$ GeV.

Source of uncertainty	$\Delta\mu_{t\bar{t}Z}$	$\Delta\mu_{t\bar{t}H}$
$t\bar{t} + c\bar{c}$ cross section	$^{+0.24}_{-0.22}$	$^{+0.17}_{-0.16}$
$t\bar{t} + b\bar{b}$ cross section	$^{+0.17}_{-0.23}$	$^{+0.15}_{-0.22}$
$t\bar{t} + 2b$ cross section	$^{+0.03}_{-0.03}$	$^{+0.10}_{-0.10}$
μ_R and μ_F scales	$^{+0.19}_{-0.14}$	$^{+0.10}_{-0.16}$
Parton shower	$^{+0.15}_{-0.16}$	$^{+0.06}_{-0.05}$
Top quark p_T modeling in $t\bar{t}$	$^{+0.01}_{-0.01}$	$^{+0.11}_{-0.13}$
b-tag efficiency	$^{+0.25}_{-0.13}$	$^{+0.10}_{-0.11}$
$b\bar{b}$-tag efficiency	$^{+0.17}_{-0.12}$	$^{+0.04}_{-0.03}$
Jet energy scale and resolution	$^{+0.11}_{-0.10}$	$^{+0.11}_{-0.12}$
Jet mass scale and resolution	$^{+0.10}_{-0.11}$	$^{+0.08}_{-0.08}$

Table 5.15: Observed (expected) correlations between the signal strength modifiers $\mu_{t\bar{t}H}$ and $\mu_{t\bar{t}Z}$, and the theoretical normalization nuisance parameters of $t\bar{t}$ + jets and $t\bar{t} + b\bar{b}$. These are extracted from the covariance matrix of the fit to the full Run 2 data set (Asimov data set).

NP	$\mu_{t\bar{t}H}$	$\mu_{t\bar{t}Z}$	$\sigma_{t\bar{t}+\text{jets}}$	$\sigma_{t\bar{t}+b\bar{b}}$
$\mu_{t\bar{t}H}$	1.00	−0.04 (−0.09)	−0.02 (0.03)	−0.25 (−0.27)
$\mu_{t\bar{t}Z}$	−0.04 (−0.09)	1.00	−0.02 (0.02)	−0.29 (−0.27)
$\sigma_{t\bar{t}+\text{jets}}$	−0.02 (0.03)	−0.02 (0.02)	1.00	−0.08 (−0.04)
$\sigma_{t\bar{t}+b\bar{b}}$	−0.25 (−0.27)	−0.29 (−0.27)	−0.08 (−0.04)	1.00

are divided into subsamples based on the generator-level p_{T} of the heavy boson. Each subsample is assigned a unique POI. Those intervals are: $200 < p_{\text{T}}^{\text{Z/H}} \leq 300$ GeV, $300 < p_{\text{T}}^{\text{Z/H}} \leq 450$ GeV, $p_{\text{T}}^{\text{Z/H}} > 450$ GeV. The POIs corresponding to the subsamples are profiled simultaneously, and the 95% upper limits of the differential cross sections are obtained from the fit results. These are summarized in Table 5.16, and illustrated individually for $t\bar{t}Z$ and $t\bar{t}H$ in Fig. 5.20 and 5.21 respectively.

Table 5.16: Observed (median expected ±1 standard deviation) 95% CL upper limits for $t\bar{t}Z$ and $t\bar{t}H$ differential cross sections.

Signal	$p_{\text{T}}^{\text{Z/H}}$ (GeV) interval	95% CL upper limit (fb)	95% CL upper limit/SM
$t\bar{t}Z$	(200, 300]	359 (492^{+216}_{-143})	3.42 ($4.69^{+2.06}_{-1.36}$)
	(300, 450]	208 (135^{+58}_{-39})	4.88 ($3.17^{+1.37}_{-0.91}$)
	(450, ∞)	49.1 ($50.7^{+23.0}_{-15.4}$)	4.02 ($4.16^{+1.89}_{-1.26}$)
$t\bar{t}H$	(200, 300]	418 (736^{+296}_{-210})	8.02 ($14.1^{+5.7}_{-4.0}$)
	(300, 450]	59.9 ($47.3^{+20.5}_{-13.9}$)	3.24 ($2.55^{+1.11}_{-0.75}$)
	(450, ∞)	9.78 ($16.5^{+7.4}_{-4.9}$)	1.96 ($3.30^{+1.49}_{-0.98}$)

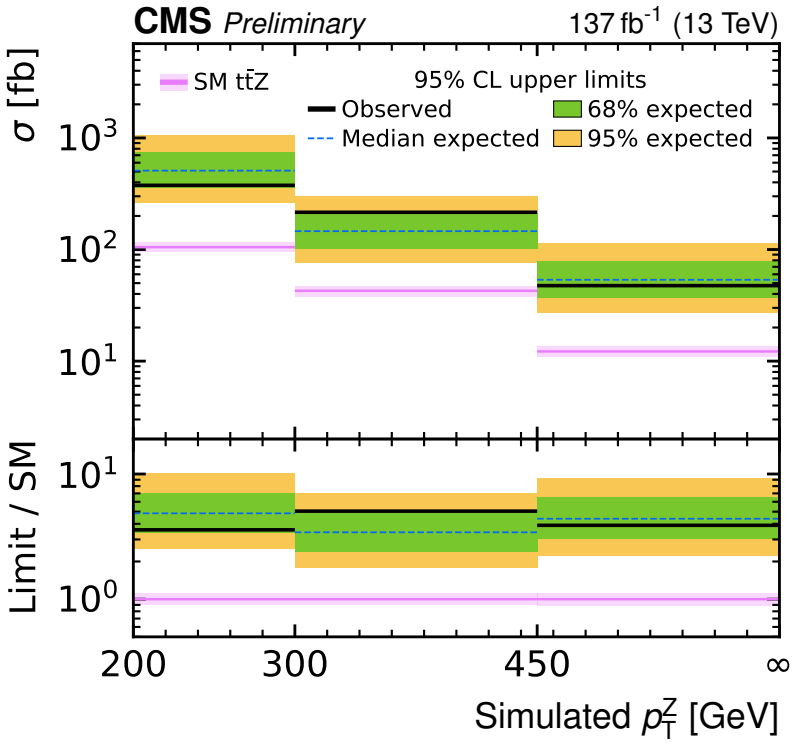

Figure 5.20: Observed and expected 95% CL upper limits on the $t\bar{t}Z$ differential cross sections as a function of Z p_T. The green and yellow bands show the expected 95% CL upper limits while the black lines represent the observed 95% CL upper limits. The magenta lines show the SM predicted differential cross sections with PDF + α_S and QCD scale uncertainties. The lower panel shows the ratio of the expected and observed upper limits on the differential cross sections to the SM differential cross sections. The last bin is unbounded, extending to large values in p_T.

5.7.2 Effective Field Theory Constraints

A fit of the templates to the data is repeated as before but with the addition of potential effects from dimension-six EFT operators and the WCs which dictate the magnitude of those effects. The implementation of WCs as POIs in the likelihood function is discussed in Section 5.5.2. As before, a rapidity requirement on the $t\bar{t}Z$ and $t\bar{t}H$ signal events is applied, i.e. $|y_{Z/H}| < 2.5$. The goal of this measurement is to constrain the WCs within a range of uncertainty and probe potential deviations from the SM theory. This is accomplished in a few ways. First, the likelihood is profiled

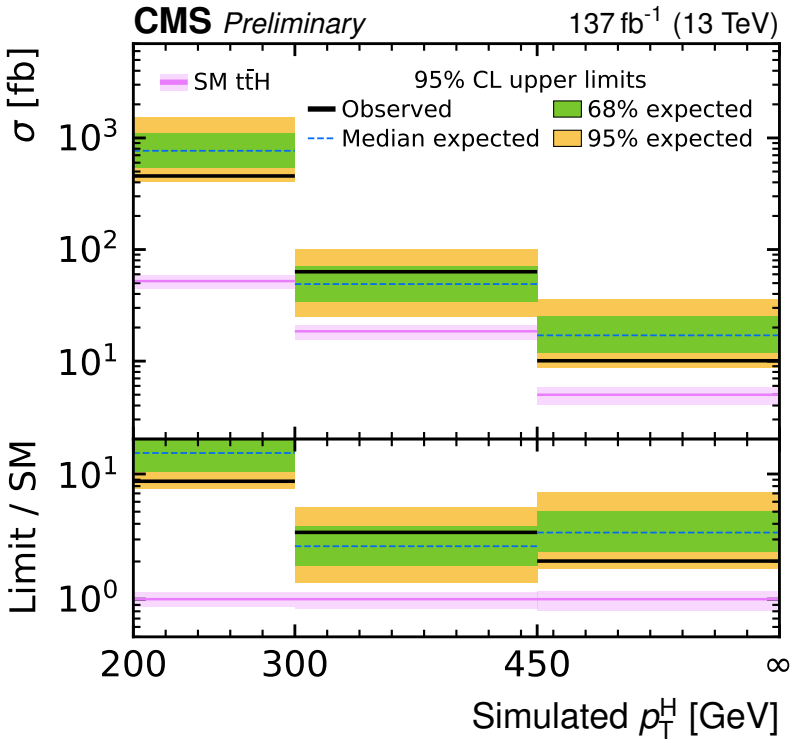

Figure 5.21: Observed and expected 95% CL upper limits on the $t\bar{t}H$ differential cross sections as a function of Higgs boson p_T. The green and yellow bands show the expected 95% CL upper limits while the black lines represent the observed 95% CL upper limits. The magenta lines show the SM predicted differential cross sections with PDF + α_S and QCD scale uncertainties. The lower panel shows the ratio of the expected and observed upper limits on the differential cross sections to the SM differential cross sections. The last bin is unbounded, extending to large values in p_T.

for each WC while fixing the seven other WCs to their SM value, zero, the results of which are displayed in Fig. 5.22. Next, the likelihood is profiled again for each WC, however this time the other WCs are also profiled simultaneously. The likelihood profiling for the WCs where the others are allowed to float is shown in Fig. 5.23. In both cases, the 68% and 95% CL intervals are extracted and illustrated in Fig. 5.24, and the 95% CL intervals are summarized in Table 5.17. Also, the impact EFT has on the predicted bin yields, given the values of Wilson coefficients corresponding to

their best-fit and their 95% CL interval upper bound for the scenario where all other WCs are fixed to their SM value, are shown in Figs. 5.25–5.32.

Table 5.17: Observed 95% CL intervals on the eight WCs in EFT. The intervals are determined by scanning over a single WC while either treating the other seven as profiled, or fixing the other seven to the SM value of zero.

WC/Λ^2[TeV^{-2}]	95% CL interval (profiled)	95% CL interval (fixed)
$c_{\mathrm{t}\varphi}$	[0.55, 29.50]	[0.25, 30.04]
$c^{-}_{\varphi Q}$	[−8.27, 9.93]	[−6.55, 8.72]
$c^{3}_{\varphi Q}$	[−4.44, 3.91]	[−4.13, 3.04]
$c_{\varphi\mathrm{t}}$	[−12.72, 7.91]	[−12.02, 6.32]
$c_{\varphi\mathrm{tb}}$	[−10.19, 11.65]	[−9.88, 10.75]
c_{tW}	[−1.63, 1.57]	[−1.05, 0.96]
c_{bW}	[−4.54, 4.54]	[−4.46, 4.41]
c_{tZ}	[−1.68, 1.67]	[−1.05, 1.11]

For some WCs, the constraints loosen as the others are profiled due to correlations in the way they affect the signal yield. Further analysis of these correlations is accomplished by profiling pairs of WCs simultaneously while fixing the other WCs to zero. The 68%, 95%, and 99.7% CL intervals from the profilings are shown in Fig. 5.33. Two of these WC pairs ($c^{3}_{\varphi Q}$, $c^{-}_{\varphi Q}$) and (c_{tW}, c_{tZ}) are associated with common operators $O^{1(ij)}_{\varphi\mathrm{q}}$ and ${}^{\ddagger}O^{(ij)}_{uB}$, respectively, and the other WC pair ($c_{\varphi\mathrm{t}}$, $c^{-}_{\varphi Q}$) corresponds to the operators $O^{(ij)}_{\varphi\mathrm{u}}$ and $O^{1(ij)}_{\varphi\mathrm{q}}$ of the similar structure, leading to the observed correlations. The likelihood scans for the WCs are sometimes bimodal, as can be seen in Fig. 5.23, especially for $c_{\mathrm{t}\varphi}$ and c_{bW}. This bimodality arises because the expected yields in each analysis bin are quadratic as functions of the WCs, so the observed yields may be most consistent with the expectations at two distinct

WC values. All of the WC constraints are in close agreement with the SM with the exception of $c_{t\varphi}$, which shows a weak tension with the SM. This tension is dominated by the yield of $t\bar{t}H$, which is smaller than expected, as seen in Figs. 5.19 and 5.21. These results complement or strengthen existing constraints on $c_{t\varphi}$, $c_{\varphi tb}$, c_{bW}, and c_{tW} in particular, in a phase space unexplored by other analyses [82–85, 135].

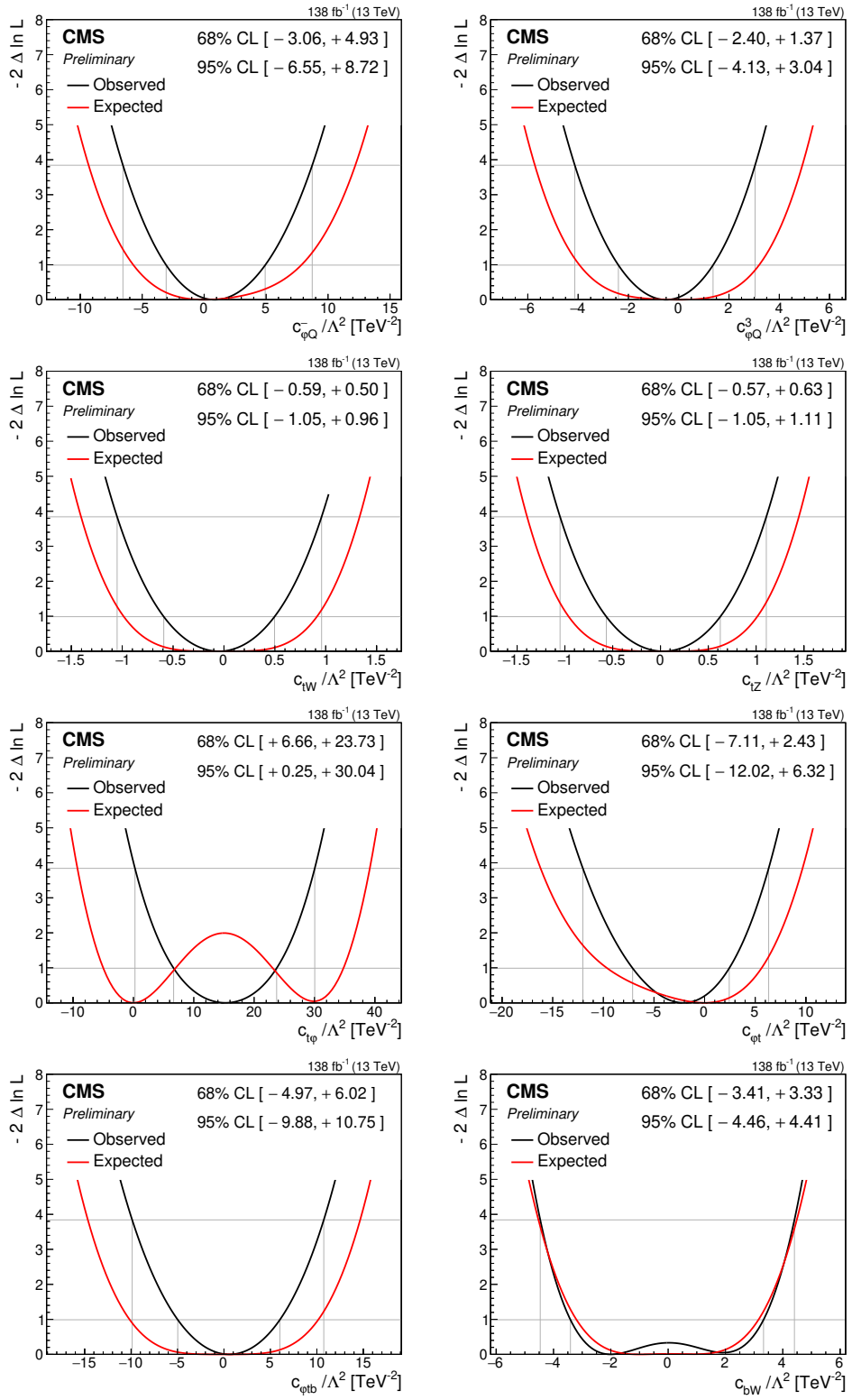

Figure 5.22: Observed (black) and expected (red) one-dimensional scans of the negative log-likelihood as a function of each of the eight WCs when the seven other WCs are fixed to their SM values. The 68% and 95% CL intervals are indicated by thin gray lines.

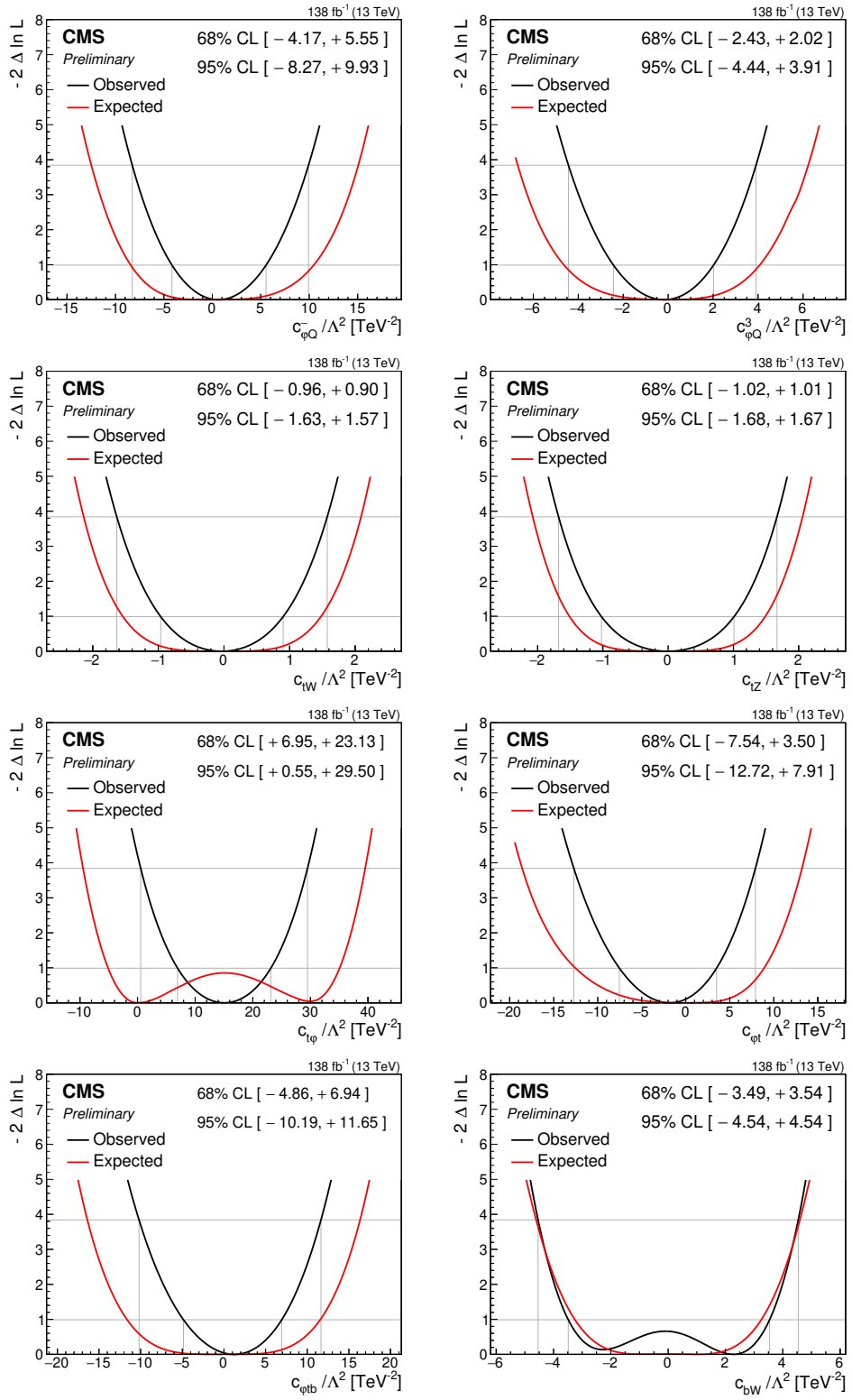

Figure 5.23: Observed (black) and expected (red) one-dimensional scans of the negative log-likelihood as a function of each of the eight WCs where all other other WCs are simultaneously profiled. The 68% and 95% CL intervals are indicated by thin gray lines.

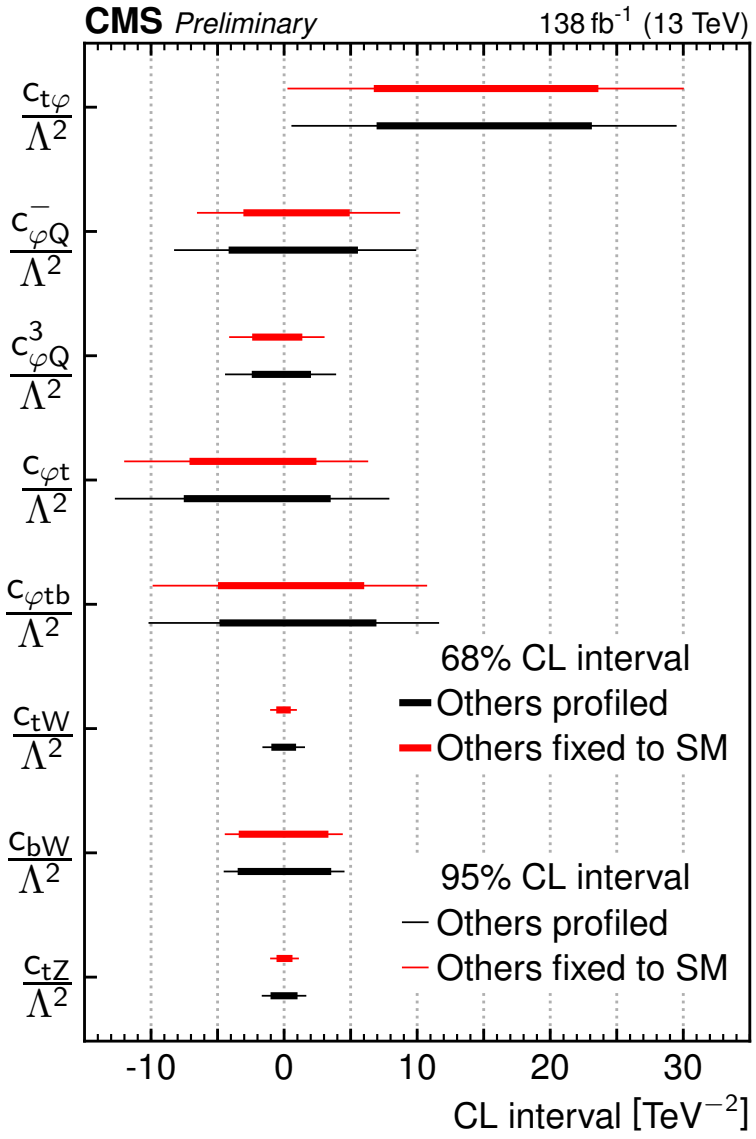

Figure 5.24: The observed 68% and 95% CL intervals for the WCs. The intervals are found by scanning over a single WC while either treating the other seven as profiled, or fixing the other seven to the SM value of zero.

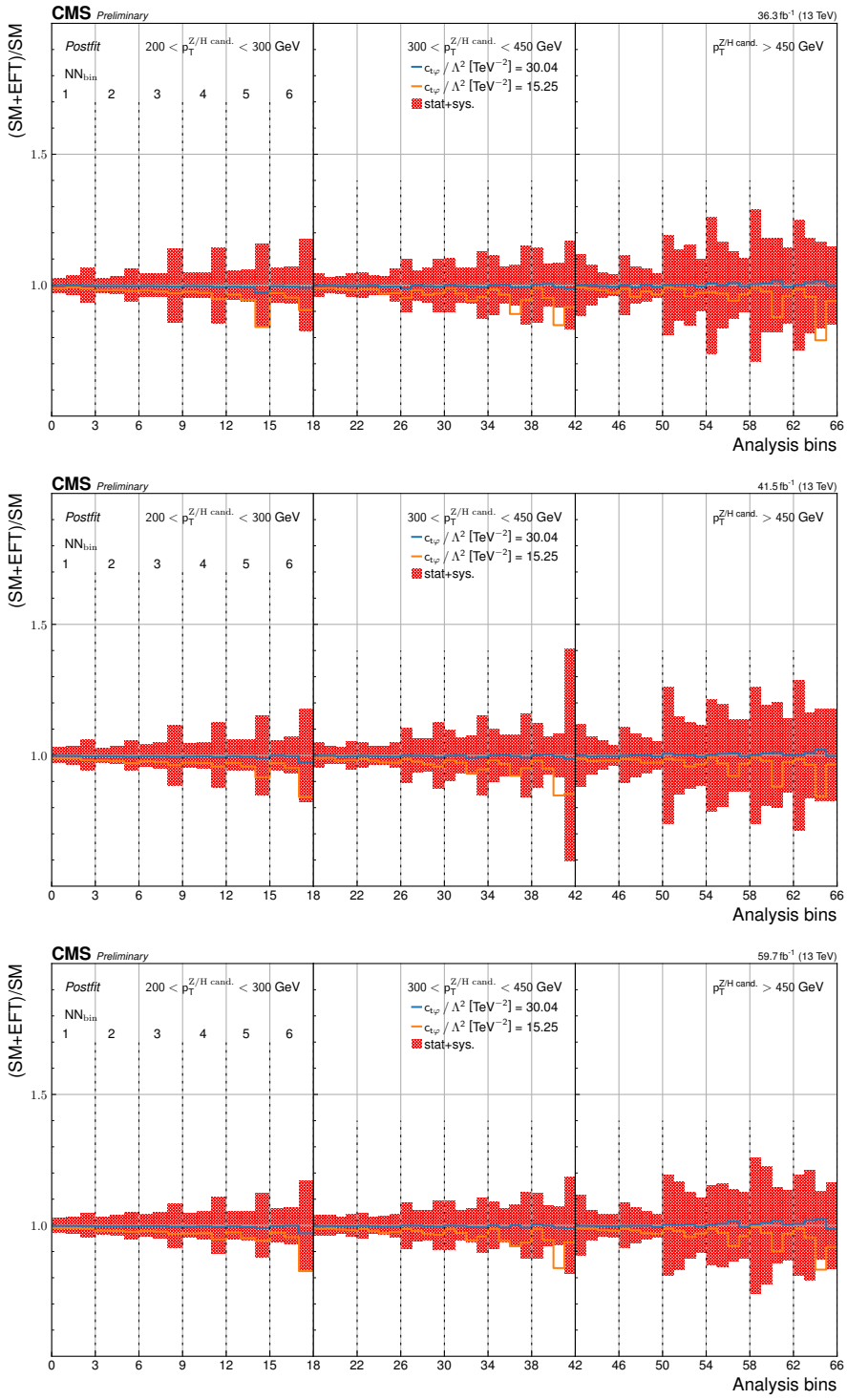

Figure 5.25: The impact that $c_{t\varphi}$ has on the predicted bin yields in the analysis templates for the 2016 (top), 2017 (middle), and 2018 (bottom) data-taking years. The plotted lines are the best-fit value of $c_{t\varphi}$ (orange) and the upper bound of the observed 95% CL interval of $c_{t\varphi}$ (blue) where all other WCs are fixed to the SM. The red bands correspond to the SM post-fit uncertainties.

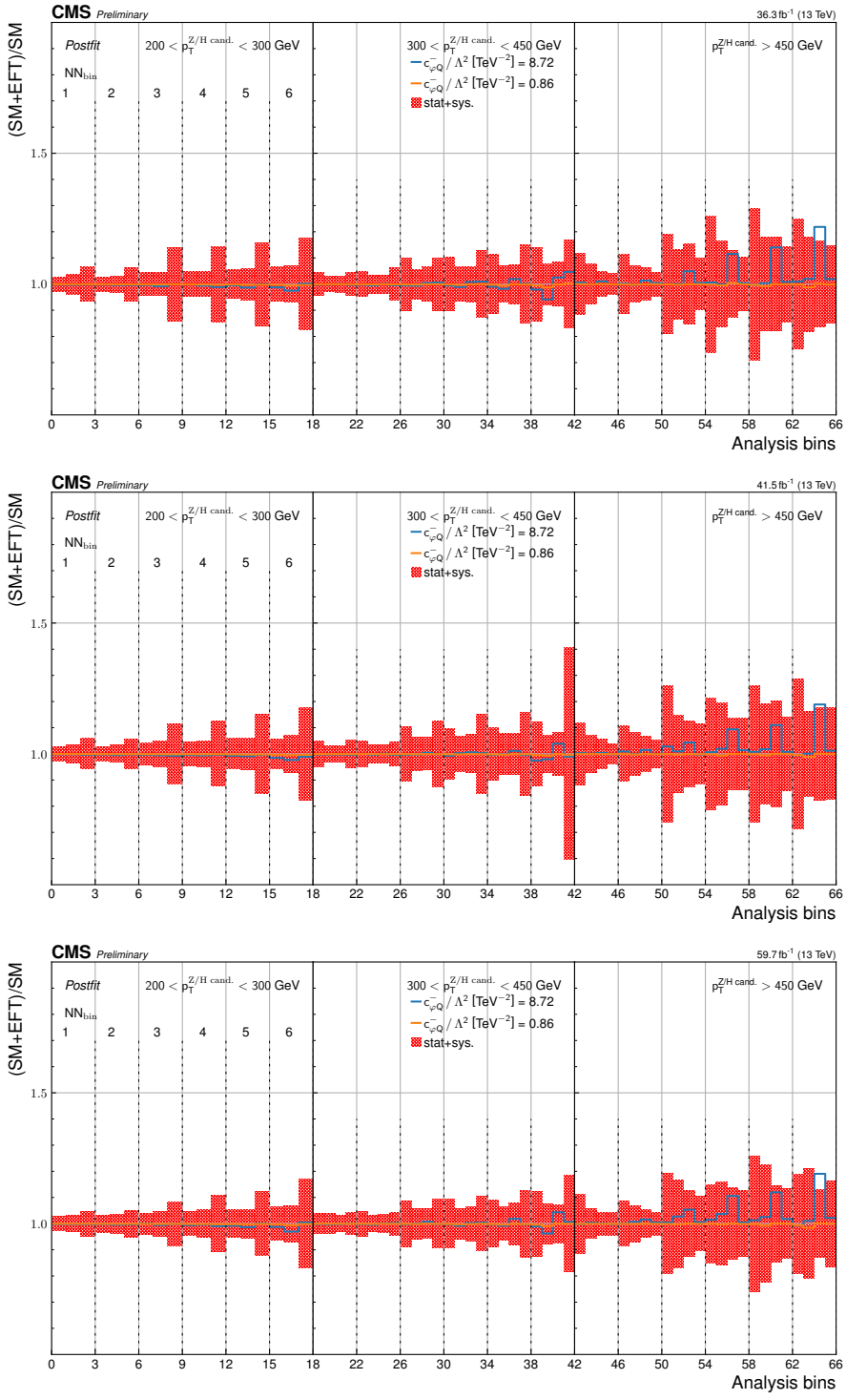

Figure 5.26: The impact that $c_{\varphi Q}^{-}$ has on the predicted bin yields in the analysis templates for the 2016 (top), 2017 (middle), and 2018 (bottom) data-taking years. The plotted lines are the best-fit value of $c_{\varphi Q}^{-}$ (orange) and the upper bound of the observed 95% CL interval of $c_{\varphi Q}^{-}$ (blue) where all other WCs are fixed to the SM. The red bands correspond to the SM post-fit uncertainties.

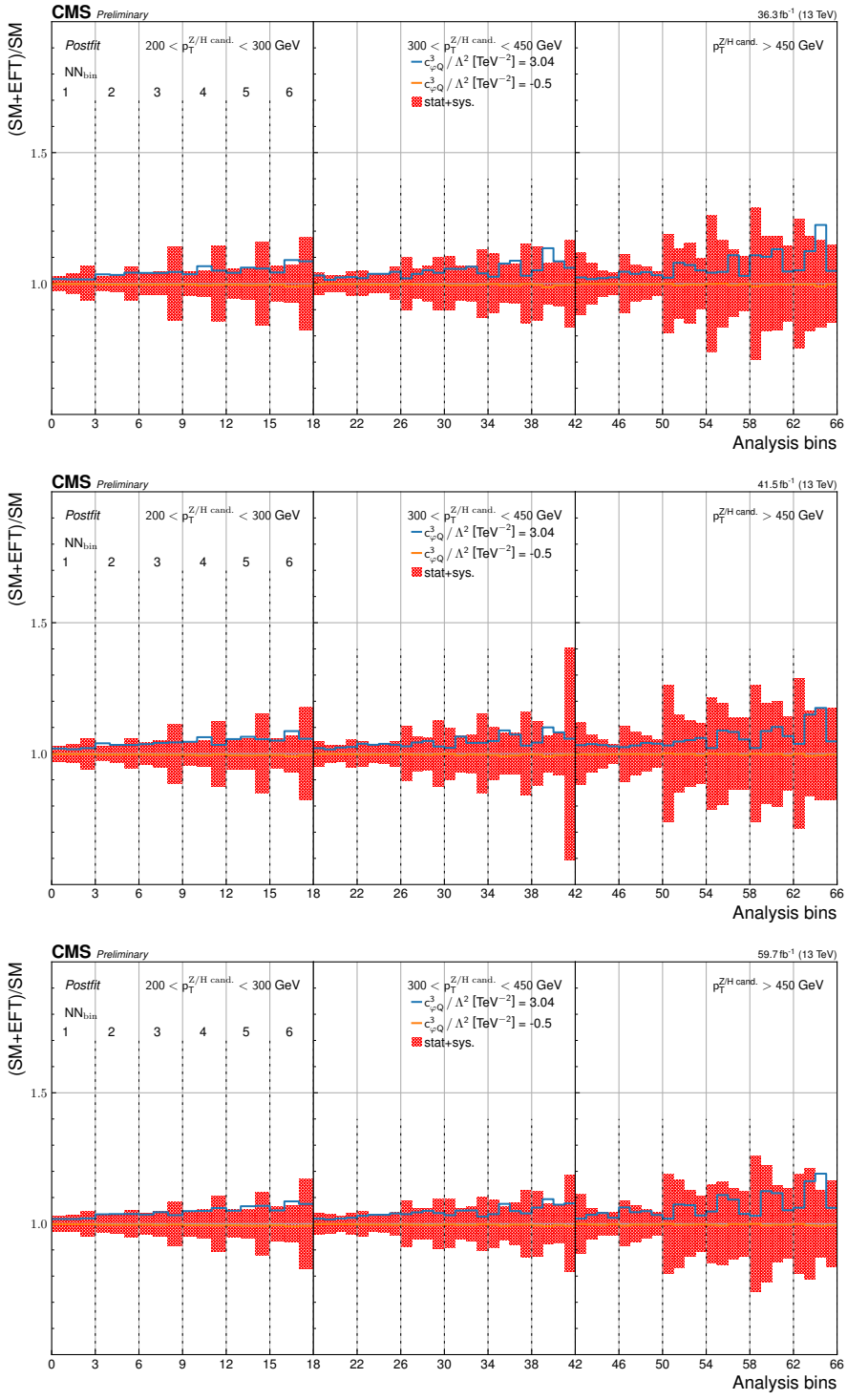

Figure 5.27: The impact that $c^3_{\varphi Q}$ has on the predicted bin yields in the analysis templates for the 2016 (top), 2017 (middle), and 2018 (bottom) data-taking years. The plotted lines are the best-fit value of $c^3_{\varphi Q}$ (orange) and the upper bound of the observed 95% CL interval of $c^3_{\varphi Q}$ (blue) where all other WCs are fixed to the SM. The red bands correspond to the SM post-fit uncertainties.

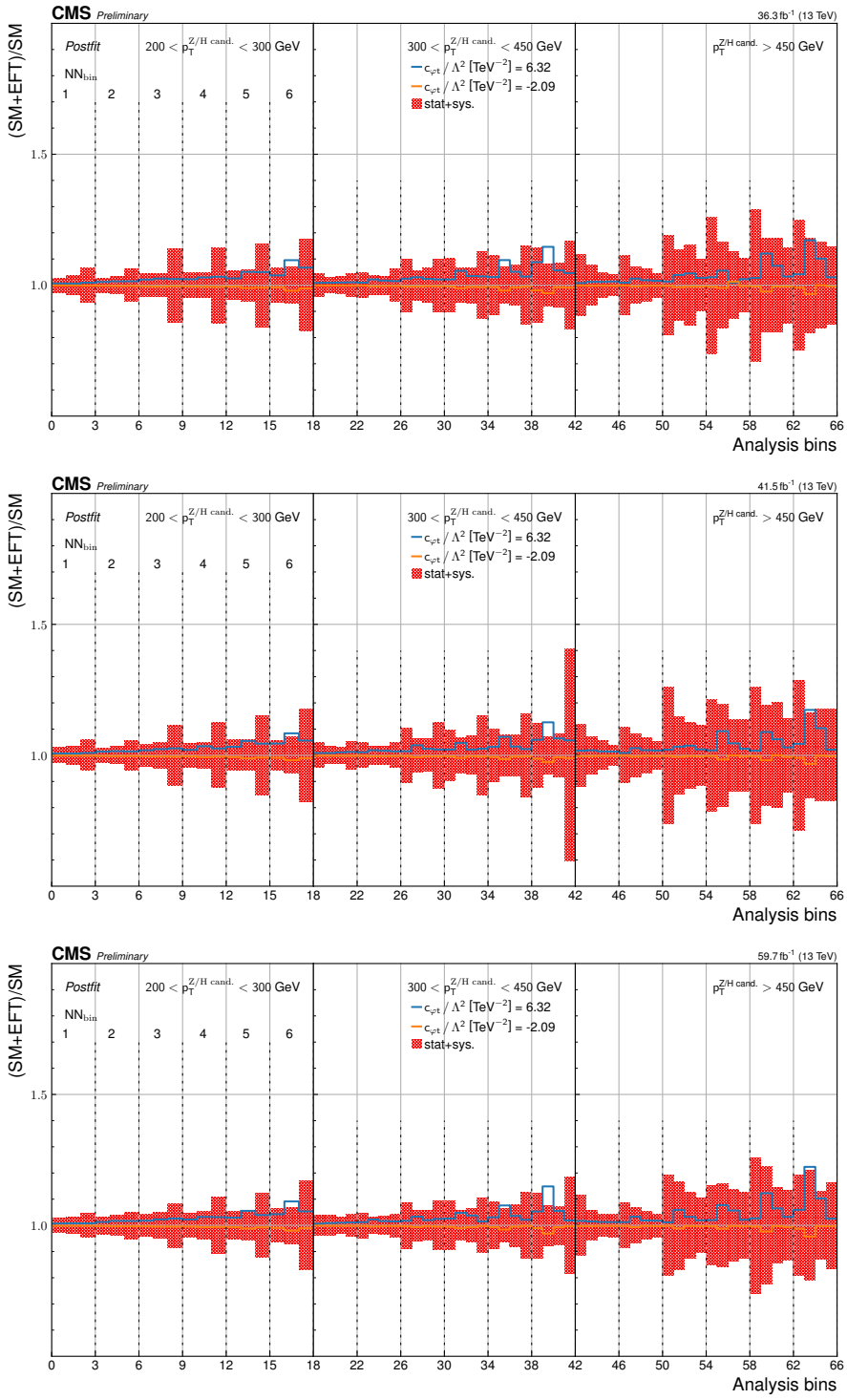

Figure 5.28: The impact that $c_{\varphi t}$ has on the predicted bin yields in the analysis templates for the 2016 (top), 2017 (middle), and 2018 (bottom) data-taking years. The plotted lines are the best-fit value of $c_{\varphi t}$ (orange) and the upper bound of the observed 95% CL interval of $c_{\varphi t}$ (blue) where all other WCs are fixed to the SM. The red bands correspond to the SM post-fit uncertainties.

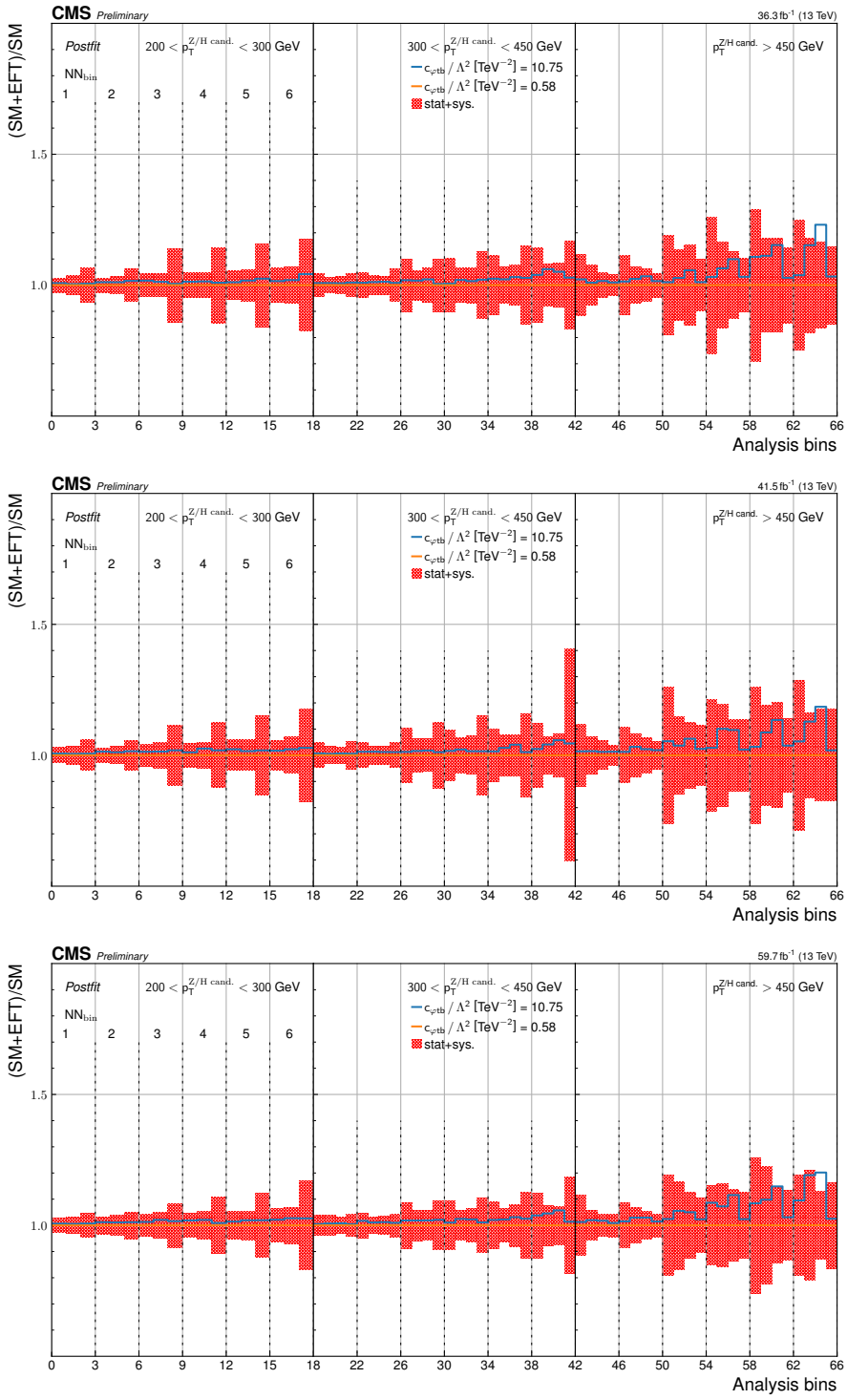

Figure 5.29: The impact that $c_{\varphi \mathrm{tb}}$ has on the predicted bin yields in the analysis templates for the 2016 (top), 2017 (middle), and 2018 (bottom) data-taking years. The plotted lines are the best-fit value of $c_{\varphi \mathrm{tb}}$ (orange) and the upper bound of the observed 95% CL interval of $c_{\varphi \mathrm{tb}}$ (blue) where all other WCs are fixed to the SM. The red bands correspond to the SM post-fit uncertainties.

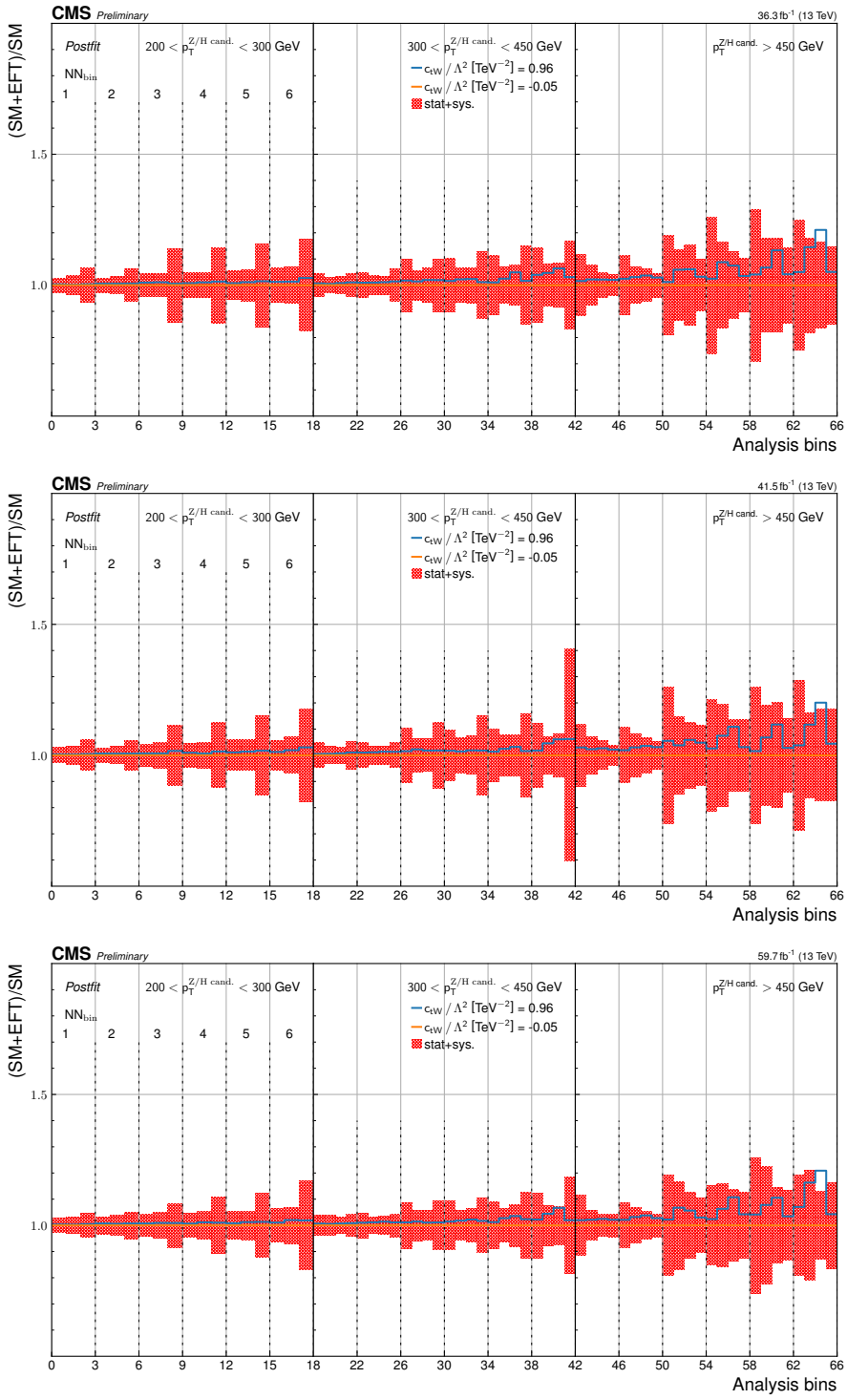

Figure 5.30: The impact that c_{tW} has on the predicted bin yields in the analysis templates for the 2016 (top), 2017 (middle), and 2018 (bottom) data-taking years. The plotted lines are the best-fit value of c_{tW} (orange) and the upper bound of the observed 95% CL interval of c_{tW} (blue) where all other WCs are fixed to the SM. The red bands correspond to the SM post-fit uncertainties.

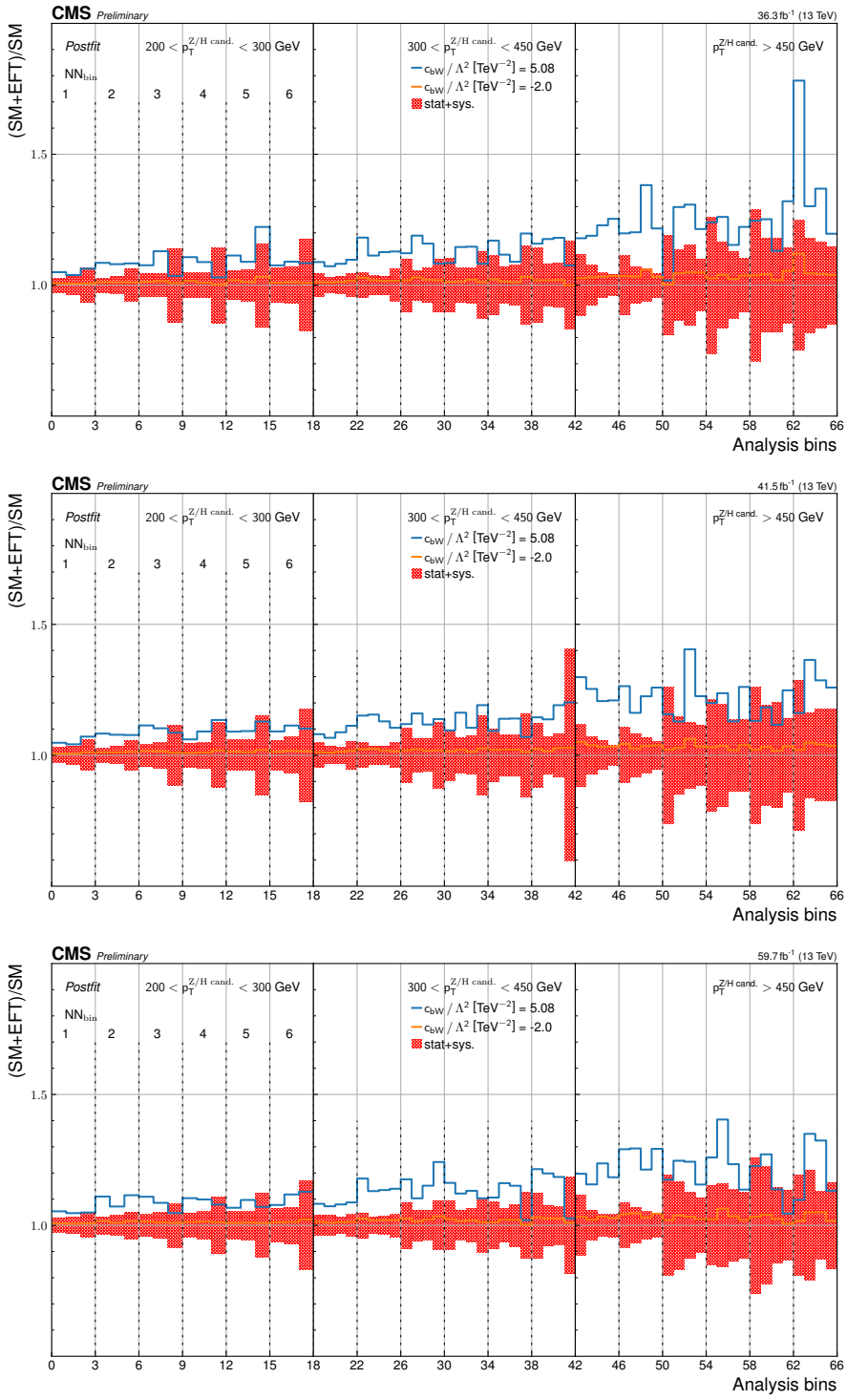

Figure 5.31: The impact that c_{bW} has on the predicted bin yields in the analysis templates for the 2016 (top), 2017 (middle), and 2018 (bottom) data-taking years. The plotted lines are the best-fit value of c_{bW} (orange) and the upper bound of the observed 95% CL interval of c_{bW} (blue) where all other WCs are fixed to the SM. The red bands correspond to the SM post-fit uncertainties.

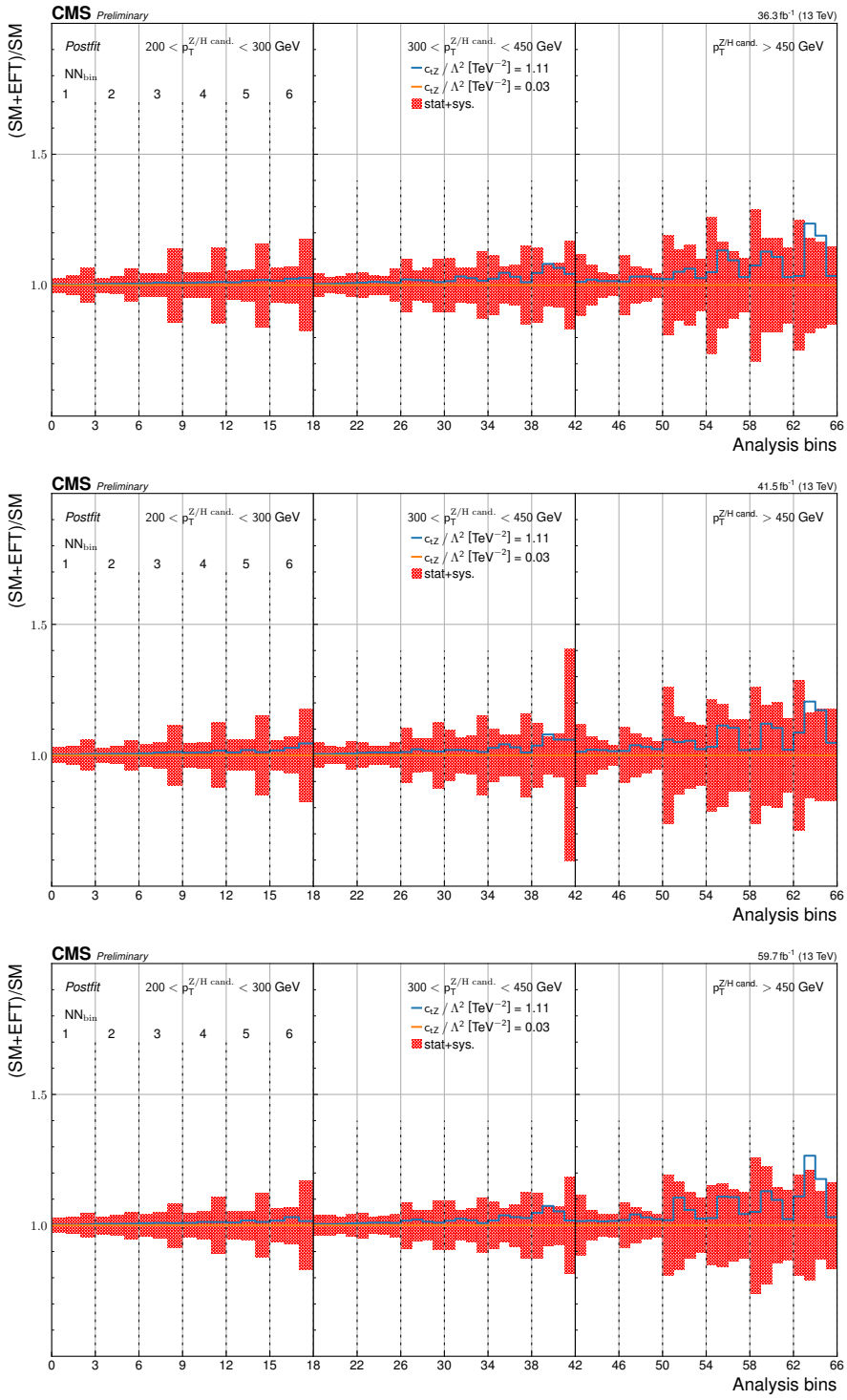

Figure 5.32: The impact that c_{tZ} has on the predicted bin yields in the analysis templates for the 2016 (top), 2017 (middle), and 2018 (bottom) data-taking years. The plotted lines are the best-fit value of c_{tZ} (orange) and the upper bound of the observed 95% CL interval of c_{tZ} (blue) where all other WCs are fixed to the SM. The red bands correspond to the SM post-fit uncertainties.

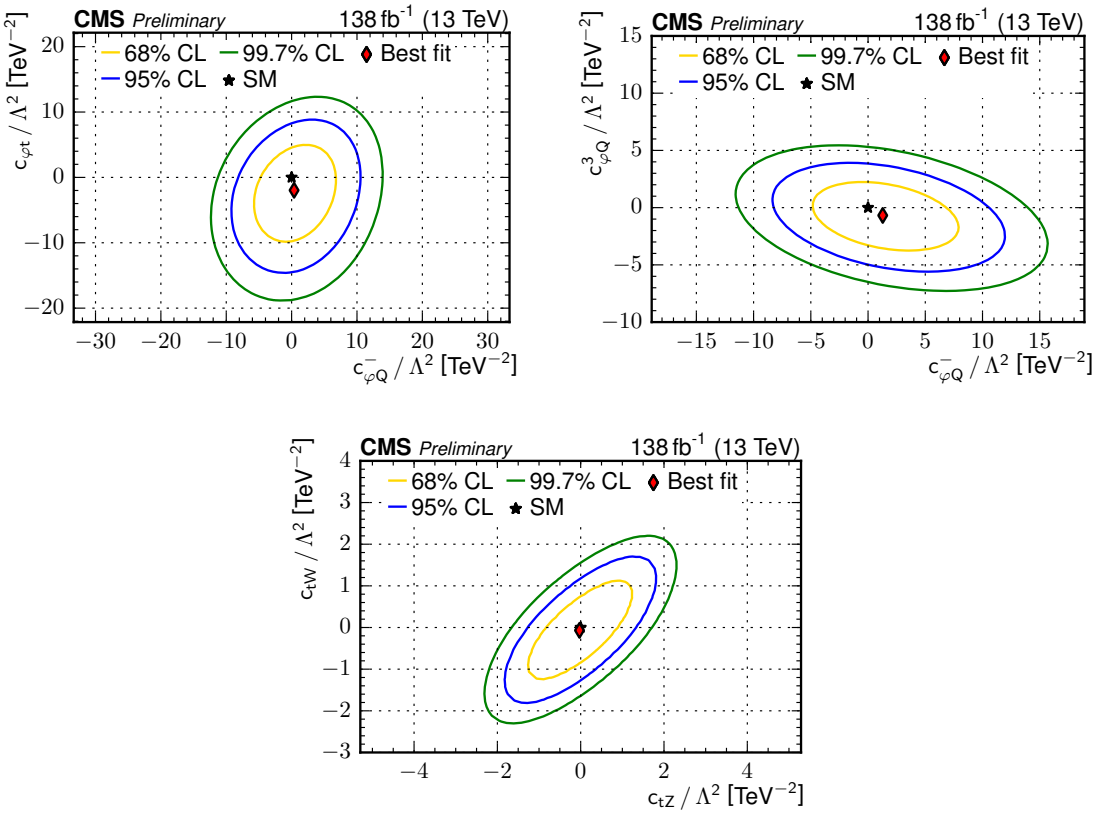

Figure 5.33: Observed two-dimensional scans of the negative log-likelihood as a function of two of the eight WCs when all other WCs are fixed to their SM values. The pair of WCs scanned correspond to the top three highest observed correlation coefficients out of all pairs. They are $c_{\varphi t}$ versus $c^{-}_{\varphi Q}$ (upper left), $c^{3}_{\varphi Q}$ versus $c^{-}_{\varphi Q}$ (upper right), and c_{tW} versus c_{tZ} (lower). The 68%, 95%, and 99.7% CL intervals are indicated by the yellow, blue, and green lines respectively.

CHAPTER SIX

Summary

A search for new physics is performed in events containing a top quark pair associated with a Z or Higgs boson in the effective field theory (EFT) framework. Additionally, the signal strength modifiers for the production of $t\bar{t}Z$ and $t\bar{t}H$ are measured, and 95% confidence level (CL) upper limits are placed on the differential cross section as a function of the Z or Higgs boson transverse momentum p_T. The analysis utilizes the full Run 2 data set of the proton-proton collisions with center-of-mass energy $\sqrt{s} = 13$ TeV. The data, corresponding to an integrated luminosity of 138 fb^{-1}, were recorded by the Compact Muon Solenoid (CMS) detector at the Large Hadron Collider (LHC). The measurements are found to be consistent with the standard model (SM) of particles physics.

The results are obtained by analyzing data containing a $t\bar{t}$ pair, whose decay produces a single lepton, plus a Z or Higgs boson decaying to $b\bar{b}$ pair where the Z or Higgs bosons is boosted, with $p_T^{Z/H} > 200$ GeV. The analysis employs a deep neural network (DNN) to discriminate the $t\bar{t}Z$ and $t\bar{t}H$ signal events from the $t\bar{t}+$jets dominated background based on the general event description as well as the physical properties of the reconstructed objects. Events are binned according to quantiles of the DNN output, the p_T of the Z or Higgs boson candidate, and also the mass of the Z or Higgs boson candidate. The templates are formed from these bins and are fit to the data by maximizing the likelihood function. From this procedure, the best-fit

values and their uncertainties are extracted related to the $t\bar{t}Z$ and $t\bar{t}H$ cross sections and EFT effects.

The signal strength modifiers for boosted $t\bar{t}Z$ and $t\bar{t}H$ production, relative to their SM predicted cross sections, are measured to be $\mu_{t\bar{t}Z} = 0.65^{+1.05}_{-0.98}$ and $\mu_{t\bar{t}H} = -0.33^{+0.87}_{-0.85}$ at the 68% CL. The 95% CL upper limits on the differential $t\bar{t}Z$ and $t\bar{t}H$ cross sections are placed in a range from 2 to 5 times the SM predicted cross sections when the Z or Higgs boson has $p_T > 300$ GeV. The results of this analysis in the SM framework represent the most stringent limits to date on the cross sections for the production of $t\bar{t}Z$ and $t\bar{t}H$ with Z or Higgs boson $p_T > 450$ GeV.

Eight Wilson coefficients (WCs) associated with eight dimension-six operators which involve a top quark and heavy boson in the leading order EFT framework are measured and constrained within a 95% CL interval. Multiple EFT scenarios are presented including where one WC is measured at a time while the others are fixed to their SM value of zero, where all WCs are profiled simultaneously, and where pairs of WCs are measured while the others are fixed to zero. In general, the WCs are consistent with the SM. However, there is an observed tension of approximately 2 standard deviations with respect to the SM theory in the profile of the $c_{t\varphi}$ WC. This deviation from the SM is due to low observed event yields in bins sensitive to the $t\bar{t}H$ signal, and is not significant enough to draw conclusions on the physicality of the ${}^{\ddagger}O^{(ij)}_{u\varphi}$ EFT operator, nor to reject the SM hypothesis. This is the first analysis to place constraints on the WCs in the dimension-six EFT framework using $t\bar{t}Z$ and $t\bar{t}H$ events with a boosted high p_T Z or Higgs boson decaying to a bottom quark pair.

The sensitivity of the results are mostly hindered by the size of the current LHC data set. Future endeavors emulating the strategy of this analysis will benefit

from a larger data set collected during Run 3, which is planned to take place from 2022 through 2026. The current forecast in the amount of data collected is roughly 2 to 3 times the amount of data collected during Run 2. After Run 3, substantial upgrades to the LHC tunnel and injection system as well as upgrades to the CMS detector are scheduled to take place. Subsequent data-taking runs in the high luminosity LHC environment (HL-LHC) are projected to collect data sets much larger than the preceding runs.

Beyond waiting for more data to be collected, there are a few more aspects which similar future analyses can improve upon. Advancements in reconstruction and identification techniques will in principle be able to enhance the signal better while rejecting background. In conjunction, novel neural network architectures and training regimens may assist in determining phase spaces which are sensitive to beyond the standard model (BSM) effects in the EFT framework. Lastly, improvements in the $t\bar{t}$ + jets simulation modeling will help reduce the theoretical uncertainty in future work.

There is still a large amount of data to be collected and analyzed over the course of the lifetime of the LHC experiment. On the horizon, there are plans for other exciting experiments in the field of high energy physics which may guide the focus of the search for new phenomena. As long as evidence for model-specific BSM theories remains to be observed, the EFT framework will be an indispensable tool when searching for new physics at the LHC.

APPENDICES

APPENDIX A

Diagrams of $t\bar{t}Z$ and $t\bar{t}H$ with EFT Vertices

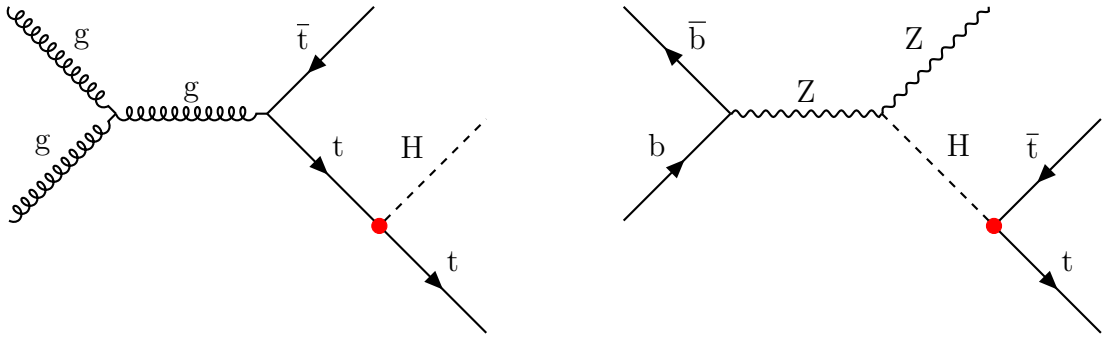

Figure A.1: Example diagrams of $t\bar{t}H$ (left) and $t\bar{t}Z$ (right) production with the inclusion of dimension-six EFT operators. The red-dot vertex corresponds to the operator ${}^{\ddagger}O_{u\varphi}^{(ij)}$ and WC $c_{t\varphi}$.

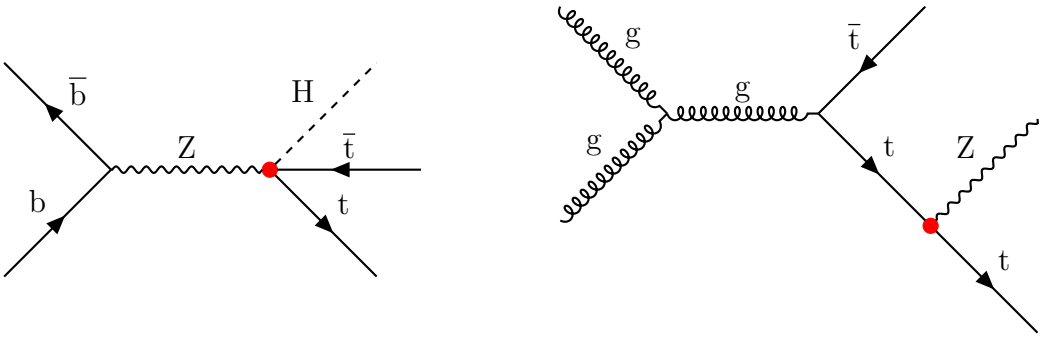

Figure A.2: Example diagrams of $t\bar{t}H$ (left) and $t\bar{t}Z$ (right) production with the inclusion of dimension-six EFT operators. The red-dot vertex corresponds to the operator $O_{\varphi q}^{1(ij)}$ and WC $c_{\varphi Q}^{-}$.

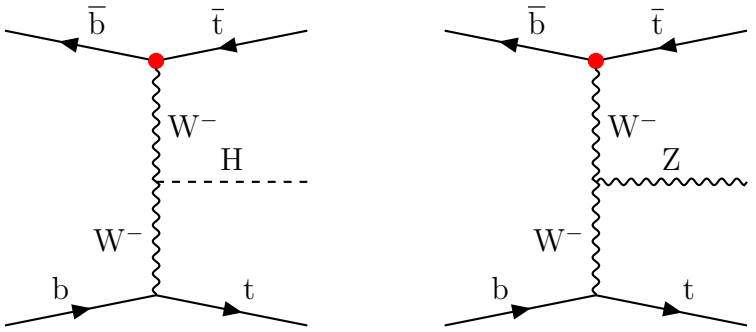

Figure A.3: Example diagrams of $t\bar{t}H$ (left) and $t\bar{t}Z$ (right) production with the inclusion of dimension-six EFT operators. The red-dot vertex corresponds to the operator $O_{\varphi q}^{3(ij)}$ and WC $c_{\varphi Q}^{3}$.

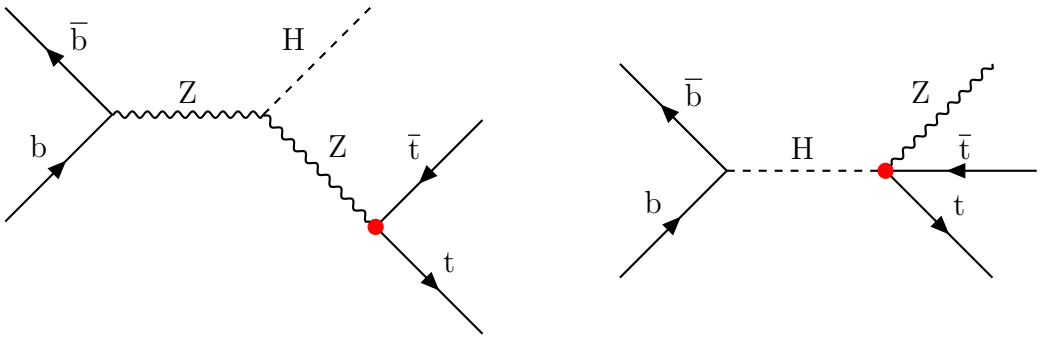

Figure A.4: Example diagrams of $t\bar{t}H$ (left) and $t\bar{t}Z$ (right) production with the inclusion of dimension-six EFT operators. The red-dot vertex corresponds to the operator $O_{\varphi u}^{(ij)}$ and WC $c_{\varphi t}$.

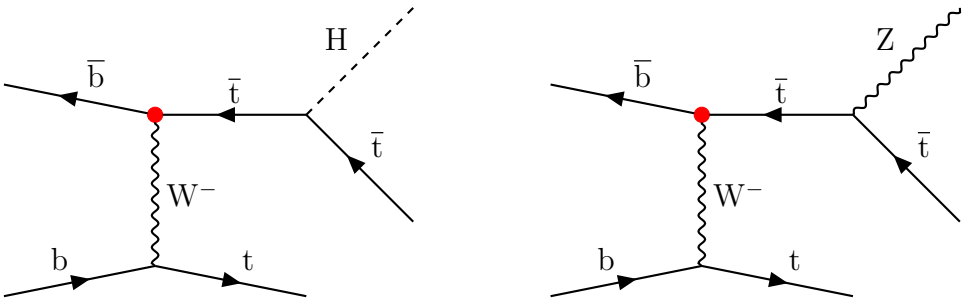

Figure A.5: Example diagrams of $t\bar{t}H$ (left) and $t\bar{t}Z$ (right) production with the inclusion of dimension-six EFT operators. The red-dot vertex corresponds to the operator ${}^{\ddagger}O_{\varphi ud}^{(ij)}$ and WC $c_{\varphi tb}$.

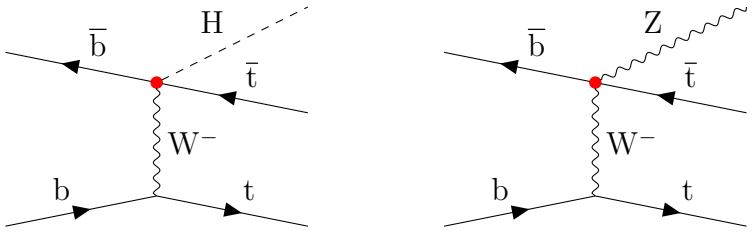

Figure A.6: Example diagrams of $t\bar{t}H$ (left) and $t\bar{t}Z$ (right) production with the inclusion of dimension-six EFT operators. The red-dot vertex corresponds to the operator ${}^{\ddagger}O_{uW}^{(ij)}$ and WC c_{tW}.

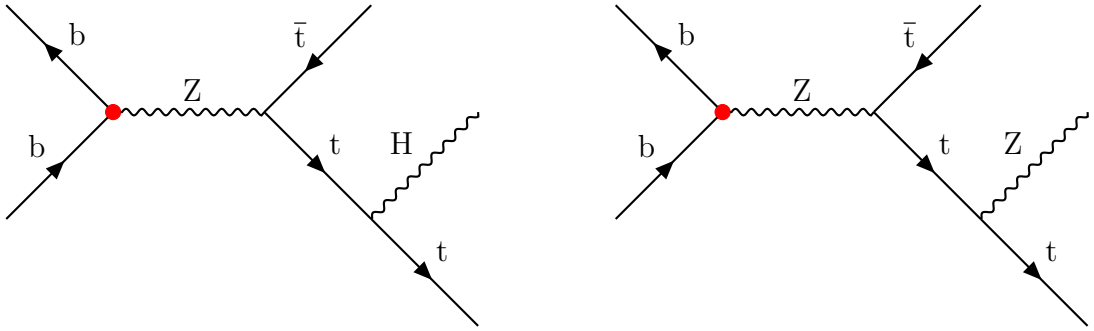

Figure A.7: Example diagrams of $t\bar{t}H$ (left) and $t\bar{t}Z$ (right) production with the inclusion of dimension-six EFT operators. The red-dot vertex corresponds to the operator ${}^{\ddagger}O_{dW}^{(ij)}$ and WC c_{bW}.

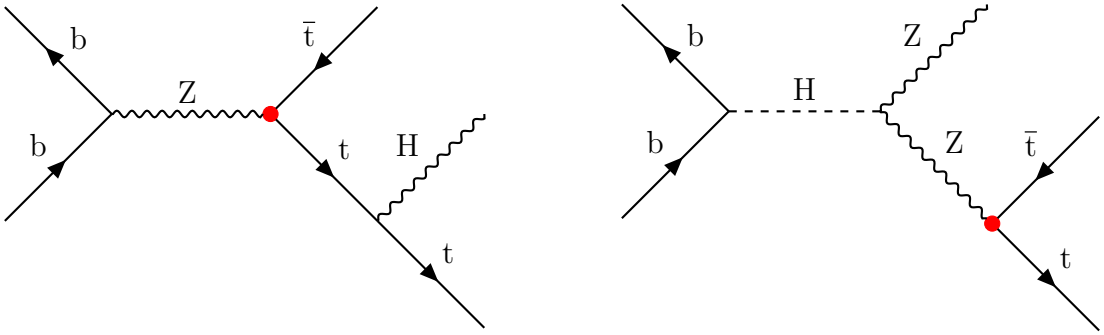

Figure A.8: Example diagrams of $t\bar{t}H$ (left) and $t\bar{t}Z$ (right) production with the inclusion of dimension-six EFT operators. The red-dot vertex corresponds to the operator ${}^{\ddagger}O_{uB}^{(ij)}$ and WC c_{tZ}.

APPENDIX B

Validation of the EFT MC Samples

The viability of using the LO EFT MC samples to model the impact of EFT on the signal samples is based on two criteria. First, there must be reasonable agreement in the shapes of the DNN score, reconstructed Z or Higgs boson p_{T}, and reconstructed Z or Higgs boson softdrop mass distributions between the LO EFT MC samples, where the WCs are set to their SM values, and the NLO MC samples. This agreement is shown for the simulated $\mathrm{t\bar{t}Z}$, $\mathrm{t\bar{t}H}$, and $\mathrm{t\bar{t}} + \mathrm{b\bar{b}}$ processes in Figs. B.1 and B.2.

Second, the impacts from EFT effects in the LO MC samples must also be compatible with impacts in NLO MC samples generated with non-zero WC values. The NLO MC samples are generated with EFT effects using the SMEFT@NLO framework [136]. Because of the computation cost of running this program, only the cross section is computed corresponding to a single WC value for comparison. Additionally, the SMEFT@NLO framework does not include the dimension-six operator corresponding to the WC $c_{\varphi \mathrm{tb}}$, and effects from certain WCs on the $\mathrm{t\bar{t}Z}$ or $\mathrm{t\bar{t}H}$ process cannot be simulated. These are $c^3_{\varphi Q}$, c_{tW}, and $c_{\mathrm{t}\varphi}$ for the $\mathrm{t\bar{t}Z}$ production, and $c^3_{\varphi Q}$, $c^-_{\varphi Q}$, $c_{\varphi \mathrm{t}}$, c_{tW}, and c_{tZ} for the $\mathrm{t\bar{t}H}$ production. Where applicable, the NLO and LO simulation of EFT effects on the production rates of $\mathrm{t\bar{t}Z}$ and $\mathrm{t\bar{t}H}$ agree reasonably well as shown in Fig. B.4.

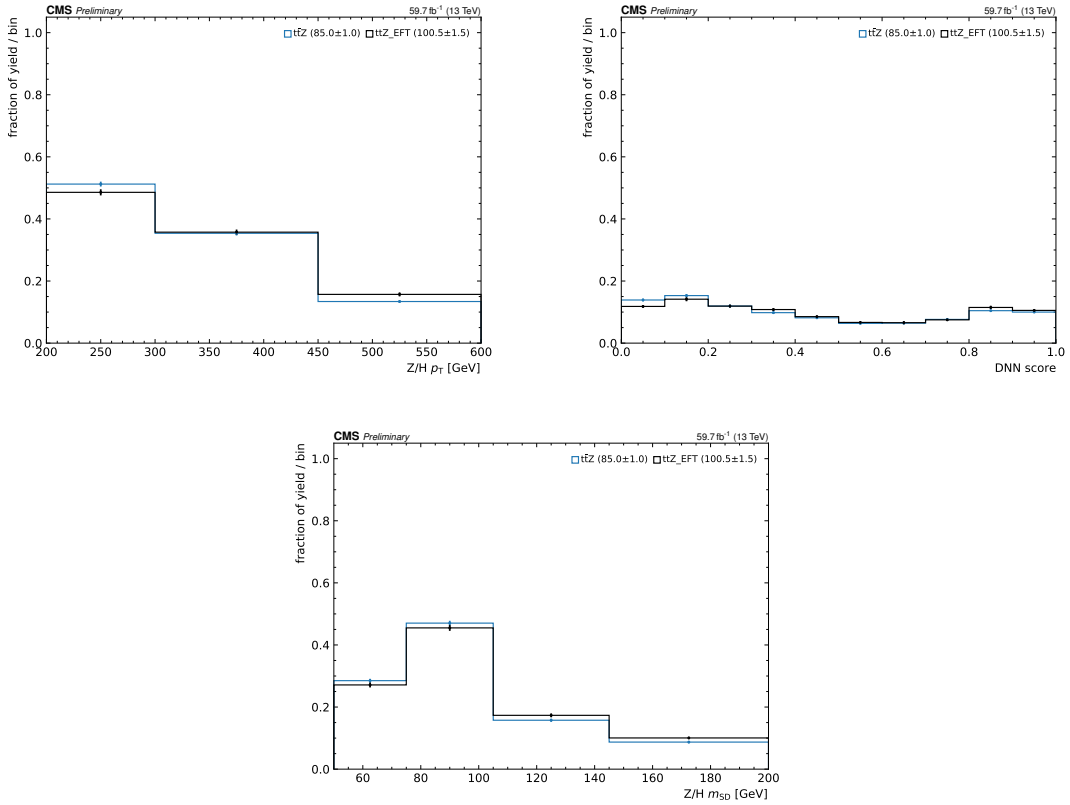

Figure B.1: Comparison in the reconstructed Z/Higgs boson candidate p_{T}, DNN output, and the Z/Higgs boson candidate m_{SD} between the privately produced EFT $\mathrm{t\bar{t}Z}$ samples simulated at LO with up to an extra parton versus the centrally produced NLO $\mathrm{t\bar{t}Z}$ samples. A k-factor is applied to the EFT samples to bring the overall normalization to NLO precision.

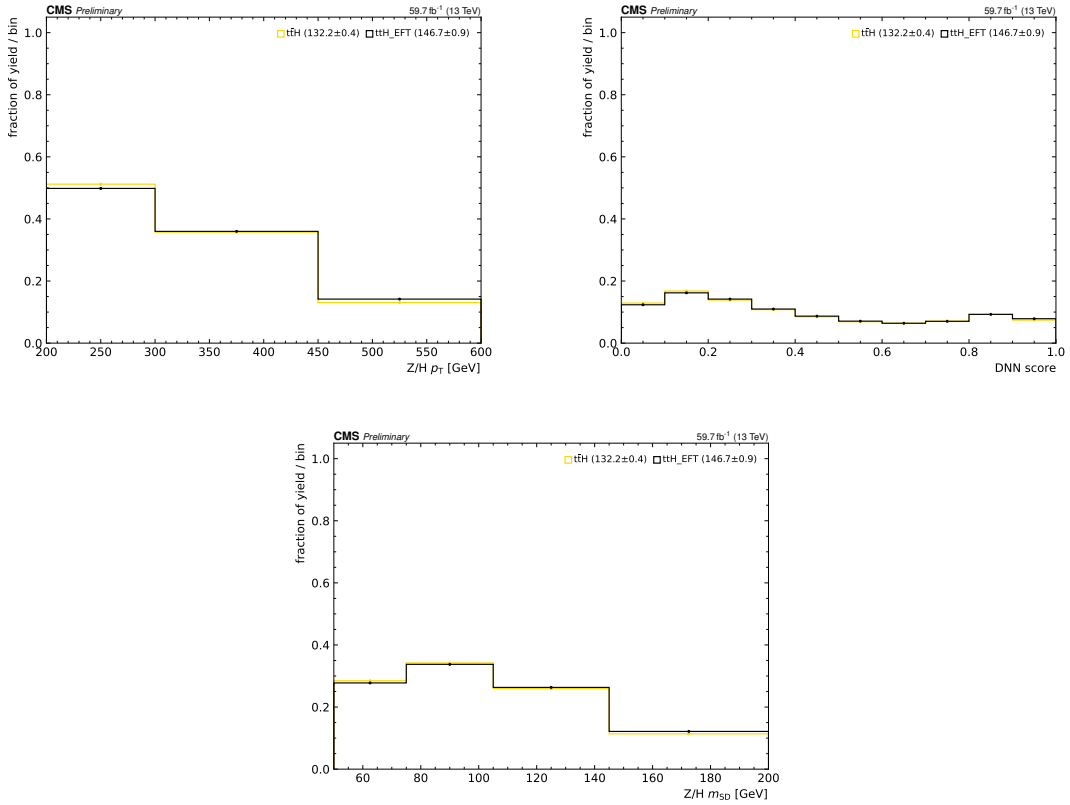

Figure B.2: Comparison in the reconstructed Z/Higgs boson candidate p_{T}, DNN output, and the Z/Higgs boson candidate m_{SD} between the privately produced EFT $\mathrm{t\bar{t}H}$ samples simulated at LO with up to an extra parton versus the centrally produced NLO $\mathrm{t\bar{t}H}$ samples. A k-factor is applied to the EFT samples to bring the overall normalization to NLO precision.

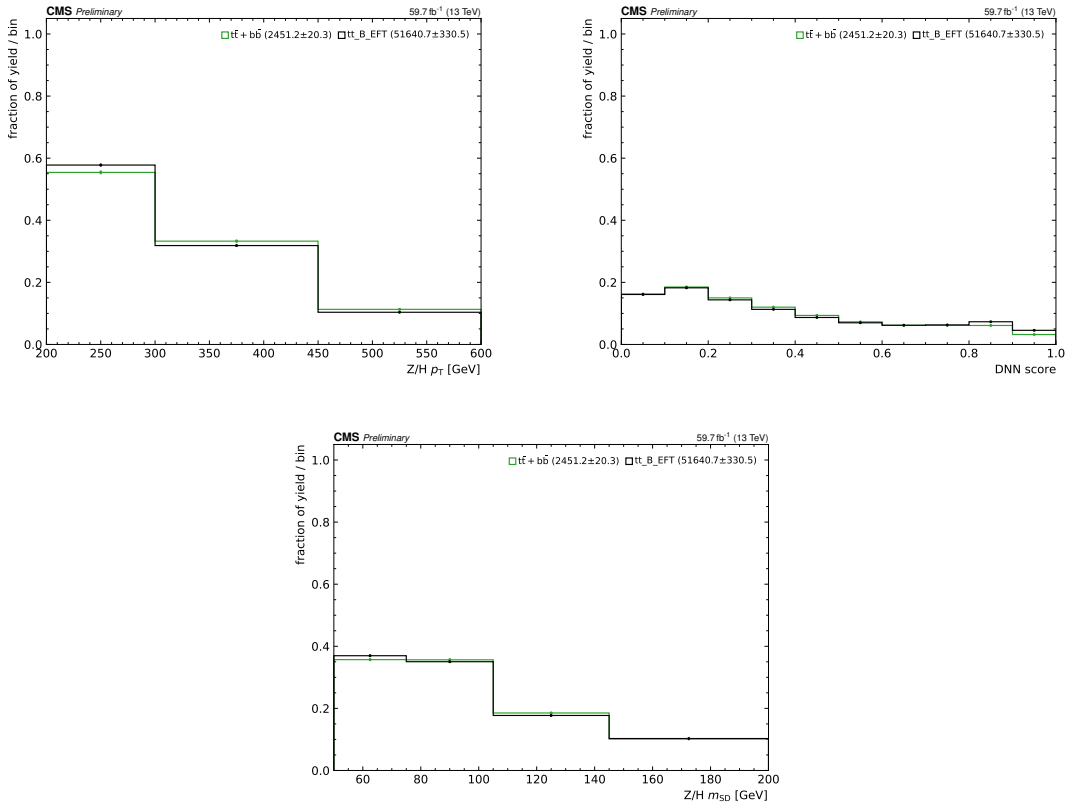

Figure B.3: Comparison in the reconstructed Z/Higgs boson candidate p_{T}, DNN output, and the Z/Higgs boson candidate m_{SD} between the privately produced EFT $\mathrm{t\bar{t}} + \mathrm{b\bar{b}}$ samples simulated at LO versus the central NLO $\mathrm{t\bar{t}} + \mathrm{b\bar{b}}$ samples. A k-factor is applied to the EFT samples to bring the overall normalization to NLO precision.

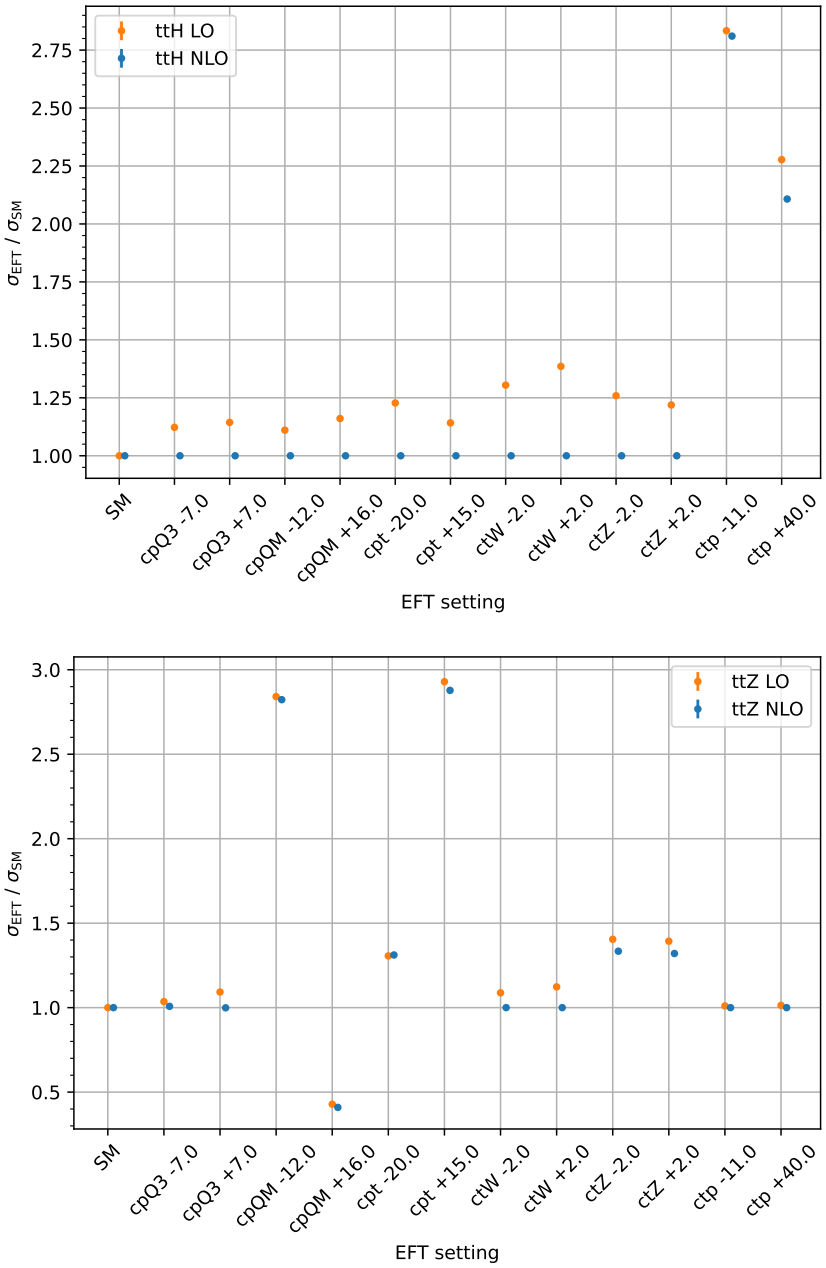

Figure B.4: Comparison in EFT effects relative to the SM cross section between the privately produced EFT cross sections simulated at LO with up to an extra parton versus privately generated SMEFT@NLO cross sections. Some WCs are not accounted for in SMEFT@NLO and their resulting cross sections are exactly equal to the SM.

APPENDIX C

Trigger Efficiency Scale Factors

A correction factor for the simulated trigger is derived in a phase space which is similar yet orthogonal to the one defined by the baseline event selection in Table 5.3:

- number of AK4 jets ≥ 4
- number of AK8 jets ≥ 1
- 1 or 2 b-tagged AK4 jets
- 1 electron
- 1 muon
- $p_{\mathrm{T}}^{\mathrm{miss}} > 20$ GeV

This phase space is dominated by $\mathrm{t\bar{t}}+\mathrm{jets}$ production with the $\mathrm{t\bar{t}}$ decaying into two leptons. The data are required to pass the reference trigger in order to preserve orthogonality. When measuring the single electron trigger efficiency, the single muon trigger is utilized as the reference trigger and visa versa for the single muon trigger efficiency. The efficiency is calculated for each year as a function of the lepton p_{T} and η, and is defined as the ratio of events passing the trigger over the total. The single electron trigger efficiency measurements for simulation and data are displayed in Figs. C.1 and C.2, respectively. Figures C.3 and C.4 contain the single muon trigger efficiency measurements for simulation and data, respectively. The trigger efficiency scale factors, as shown in Fig. C.5 for the single electron data and Fig. C.6 for the single muon data, are calculated as

$$\mathrm{SF}_{\mathrm{trigger}} = \frac{\epsilon_{\mathrm{Data}}}{\epsilon_{\mathrm{MC}}}, \tag{C.1}$$

where ϵ is the trigger efficiency. The correction is applied to the simulation by multiplying the MC event weight by the scale factor. The sources of uncertainty in the calculation of the trigger efficiency scale factor include the statistical uncertainty as determined by the binomial confidence interval [137], the correlation between the reference and the target triggers, and the potential contamination from events with a fake muon or electron such as the $t\bar{t}$ + jets process where only a single lepton is produced from the decay of $t\bar{t}$. The correlation coefficient between the target and the reference triggers is defined as

$$\alpha = \frac{\epsilon^{\mathrm{MC}}_{\mathrm{ref.\,trigger}} \times \epsilon^{\mathrm{MC}}_{\mathrm{target\,trigger}}}{\epsilon^{\mathrm{MC}}_{\mathrm{both\,triggers}}}, \tag{C.2}$$

where $\epsilon^{\mathrm{MC}}_{\mathrm{both\,triggers}}$ is the trigger efficiency when both reference and target triggers are required, and α is the correlation coefficient, e.g. $\alpha = 1$ means the reference and the target triggers are fully uncorrelated. The correlation between the two is treated as a systematic uncertainty, and is given by

$$\text{Sys. uncertainty} = (1 - \alpha) \times \text{Nominal SF.} \tag{C.3}$$

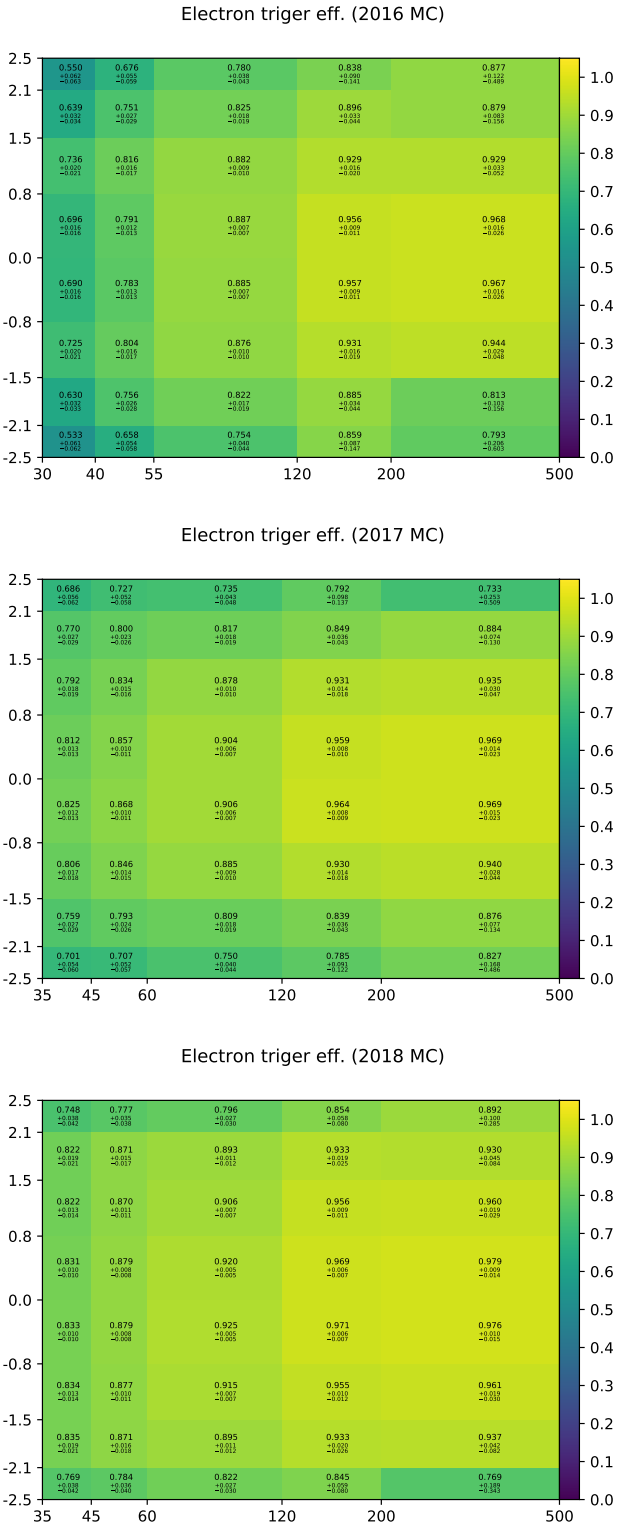

Figure C.1: Single-electron trigger efficiency measured with simulated events as a function of reconstructed electron p_{T} and η for the 2016 (top), 2017 (middle), and 2018 (bottom) data-taking years. The efficiency and error are annotated within each cell of the plot.

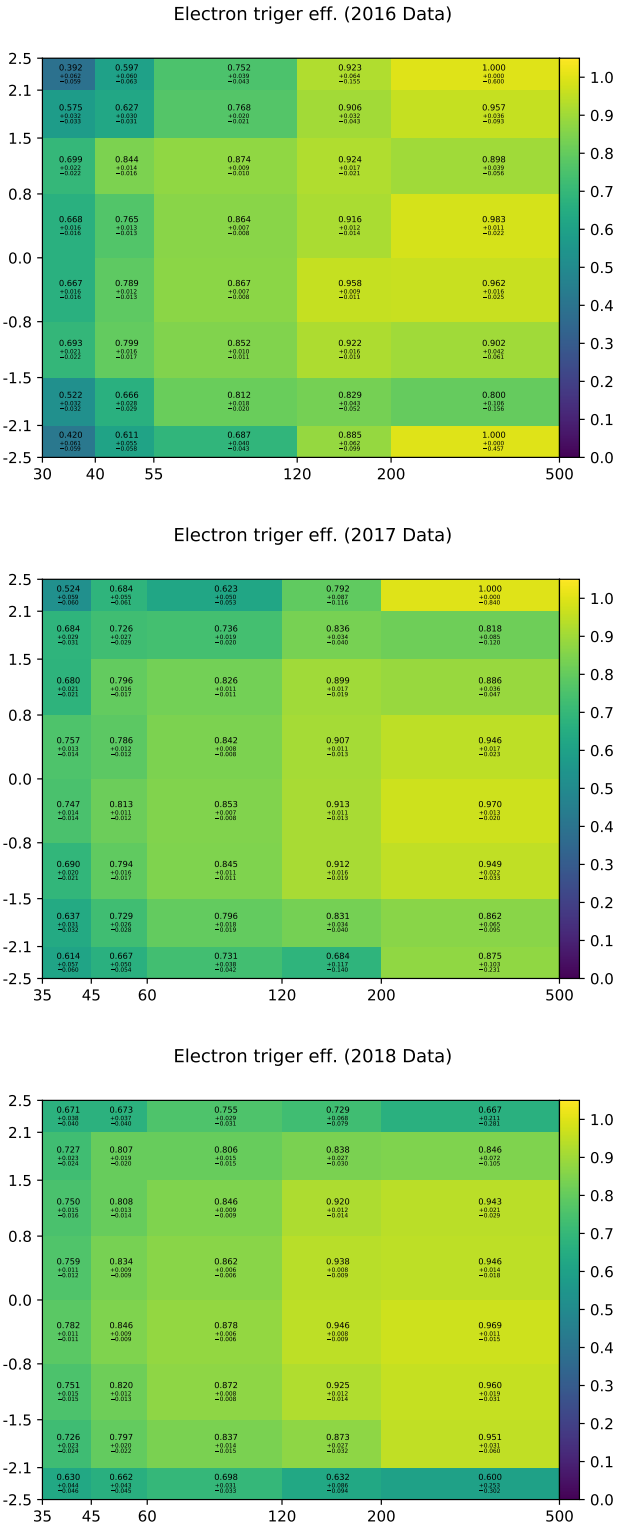

Figure C.2: Single-electron trigger efficiency measured with data events as a function of reconstructed electron p_{T} and η for the 2016 (top), 2017 (middle), and 2018 (bottom) data-taking years. The efficiency and error are annotated within each cell of the plot.

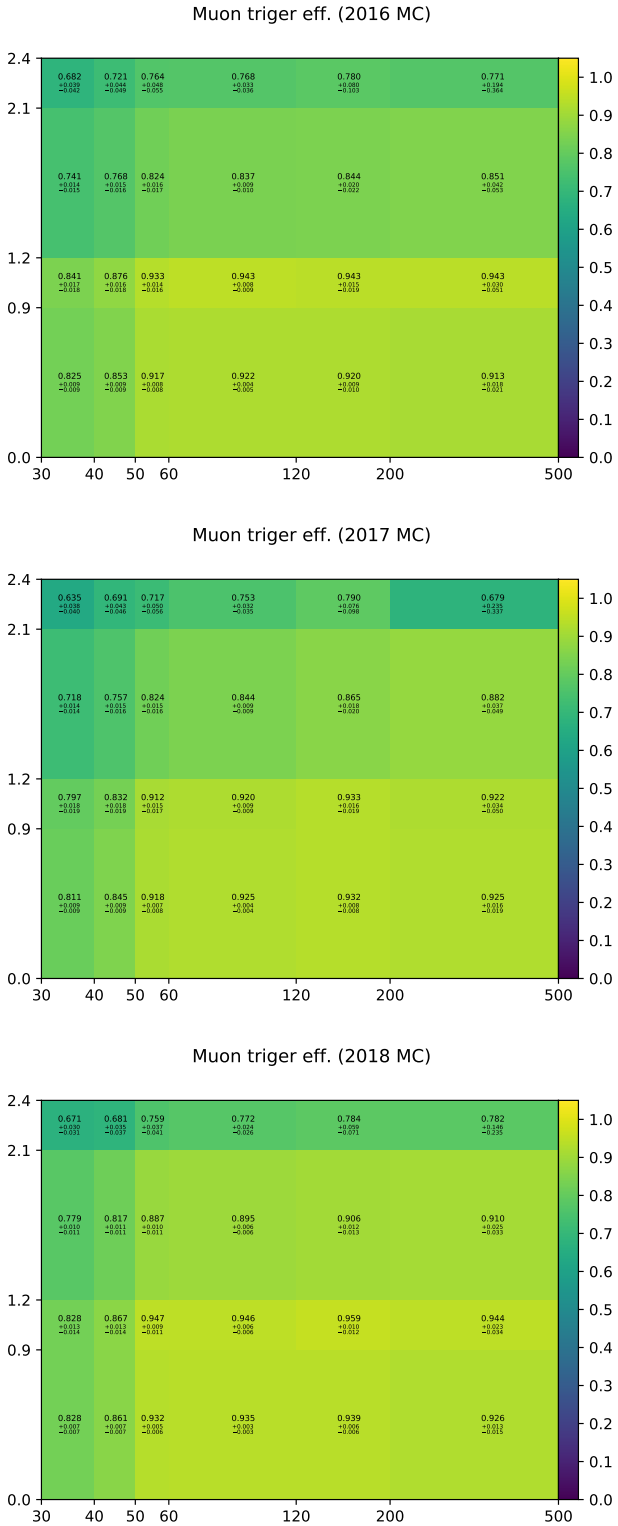

Figure C.3: Single-muon trigger efficiency measured with simulated events as a function of reconstructed muon p_{T} and η for the 2016 (top), 2017 (middle), and 2018 (bottom) data-taking years. The efficiency and error are annotated within each cell of the plot.

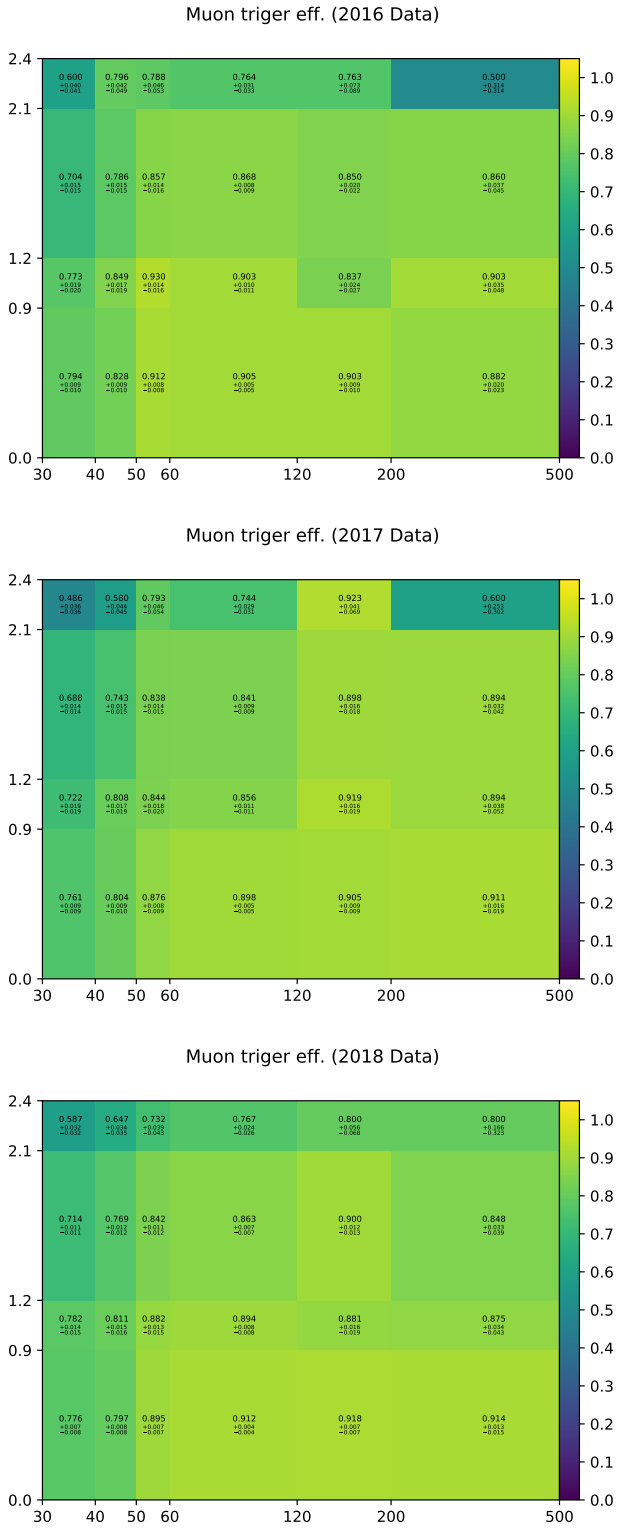

Figure C.4: Single-muon trigger efficiency measured with data events as a function of reconstructed muon p_{T} and η for the 2016 (top), 2017 (middle), and 2018 (bottom) data-taking years. The efficiency and error are annotated within each cell of the plot.

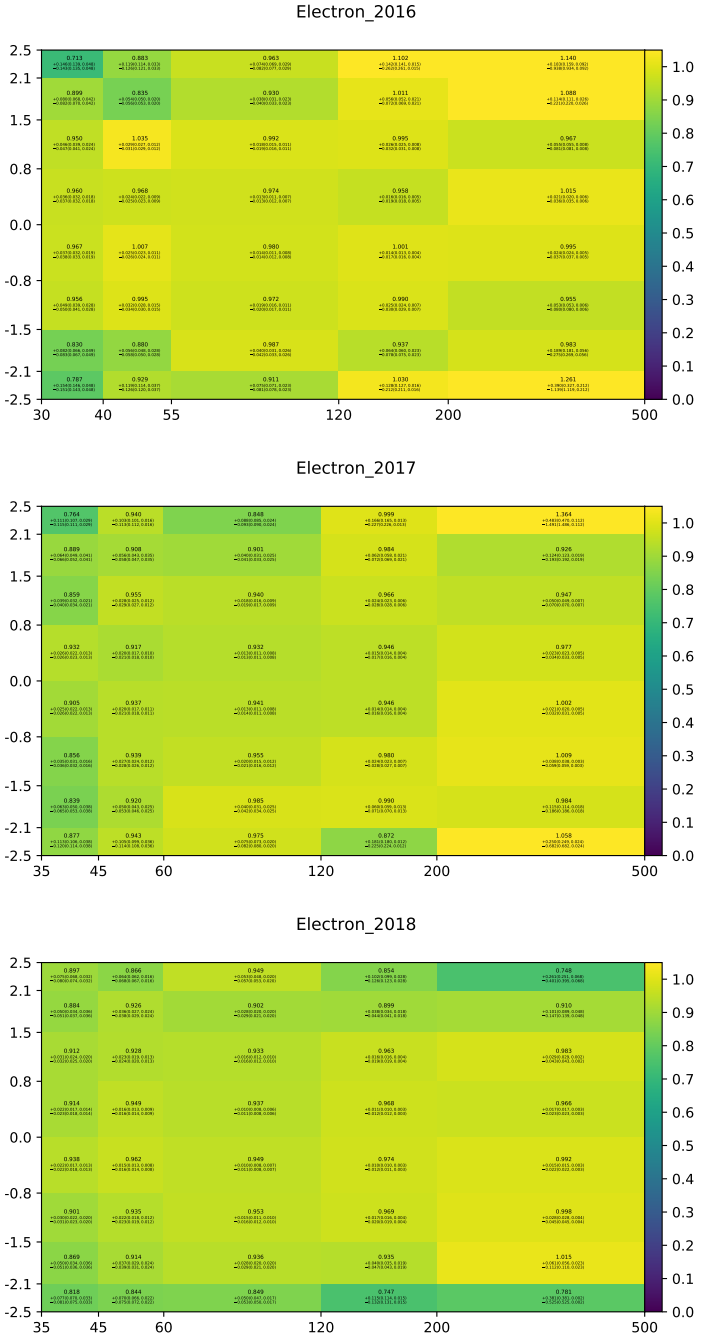

Figure C.5: Single-electron trigger efficiency scale factor measured as a function of reconstructed electron p_{T} and η for the 2016 (top), 2017 (middle), and 2018 (bottom) data-taking years. The efficiency and error are annotated within each cell of the plot.

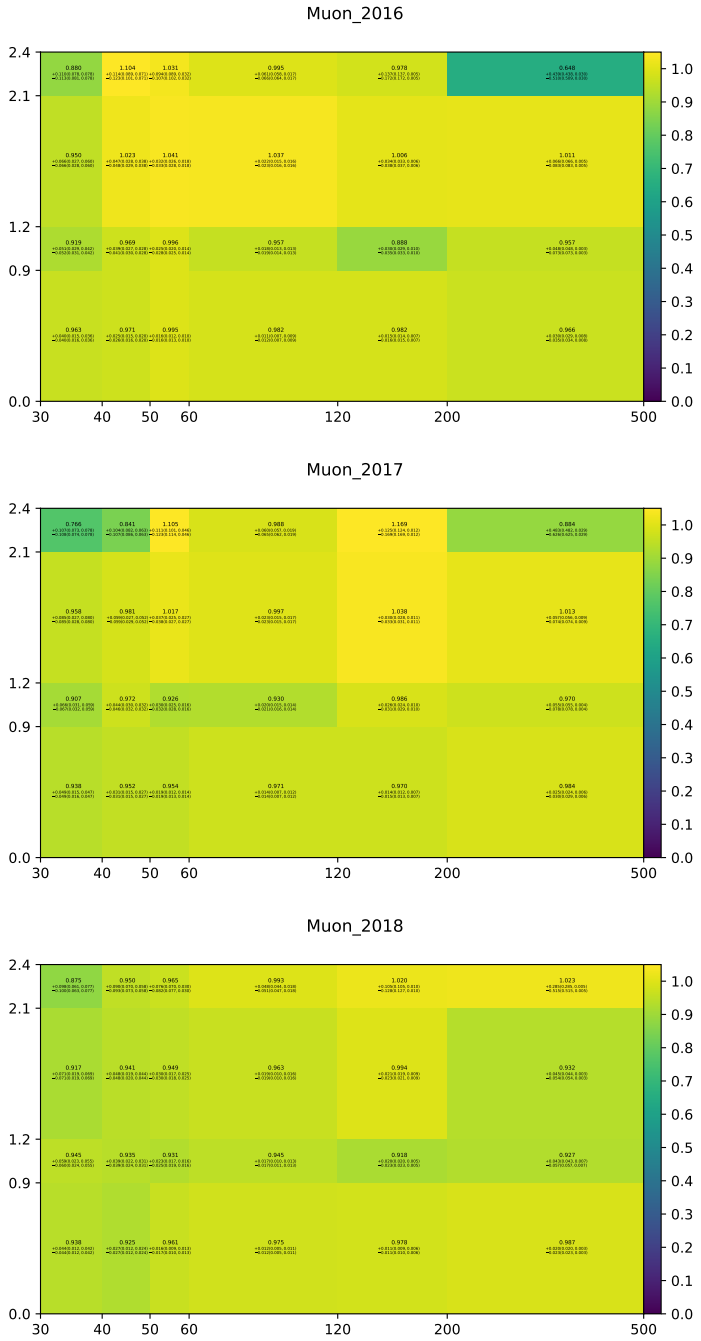

Figure C.6: Single-muon trigger efficiency scale factor measured as a function of reconstructed muon p_{T} and η for the 2016 (top), 2017 (middle), and 2018 (bottom) data-taking years. The efficiency and error are annotated within each cell of the plot.

APPENDIX D

Validation of the Fit to the Data

The quality of the fit to the data is assessed by examining the nuisance parameters of the model. In general, if the model is an excellent representation of what to expect in data then the parameters of the likelihood function should not change drastically when the likelihood is maximized. To quantify how much the model has to shift to better fit the data, the pulls $(\hat{\theta} - \theta_0)/\Delta\theta$ for each nuisance parameter are calculated. When the model exactly agrees with the observed data, the pull will have a value of 0 and a standard deviation equal to 1. Otherwise those values will likely shift away from zero as seen in the left columns in Figs. D.1 and D.2, and also in Figs. D.3 and D.4. A shifted pull is assessed on a case by case basis to determine whether or not the result makes sense, e.g. "tt2bxsec" and "ttCxsec" are pulled away from zero and are constrained tighter than a standard deviation equal to one. This is not alarming because the uncertainties associated with these nuisances are not well established a priori. Overall, the pulls on the nuisance parameters behave reasonably.

Additionally, the impact that each source of systematic uncertainty has on the signal strength modifiers $\mu_{\mathrm{t\bar{t}H}}$ and $\mu_{\mathrm{t\bar{t}Z}}$ is compared between what is expected and what is observed. Illustrations of the top thirty impacts are shown in Figs. D.1 and D.2. Overall, the behavior of the nuisance parameters does not vary unreasonably between scenarios.

Lastly, a goodness of fit test is performed to determine whether of not to reject the null hypothesis, i.e. the model is statistically compatible with the data.

The result of the test is a p-score of 0.977 which means the null hypothesis should not be rejected.

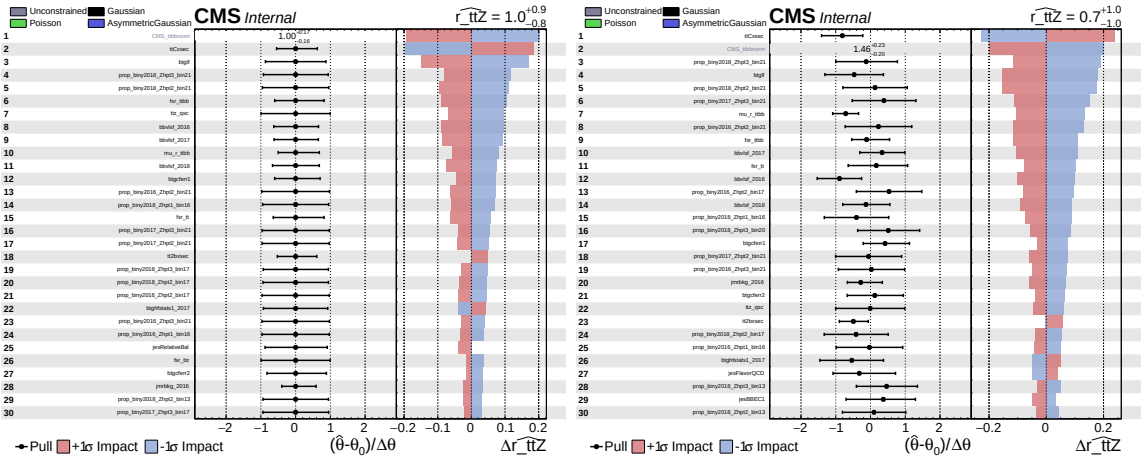

Figure D.1: The impacts and pulls of the top thirty nuisance parameters with respect to the best-fit and uncertainty of the signal strength modifier $\mu_{t\bar{t}Z}$. The expected impacts and pulls are shown on the left and the observed impacts and pulls are shown on the right.

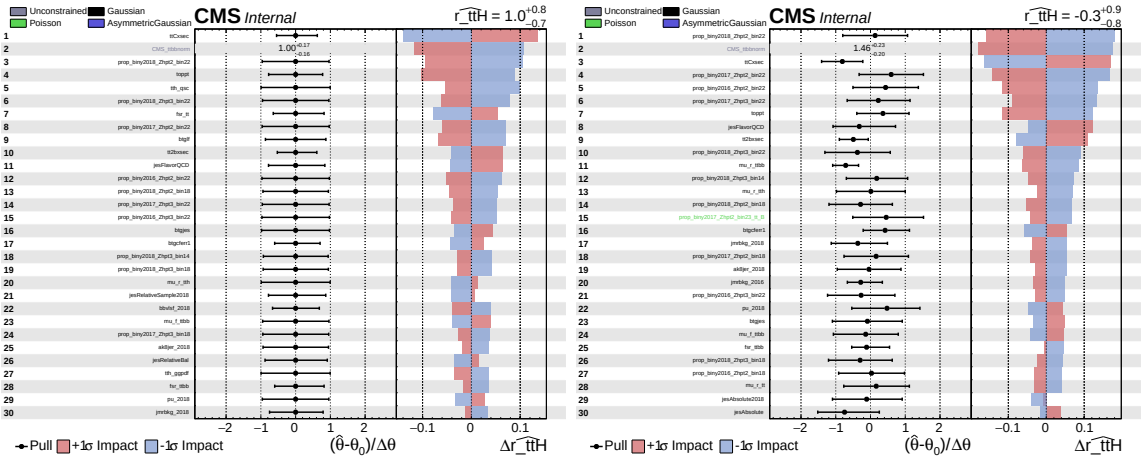

Figure D.2: The impacts and pulls of the top thirty nuisance parameters with respect to the best-fit and uncertainty of the signal strength modifier $\mu_{t\bar{t}H}$. The expected impacts and pulls are shown on the left and the observed impacts and pulls are shown on the right.

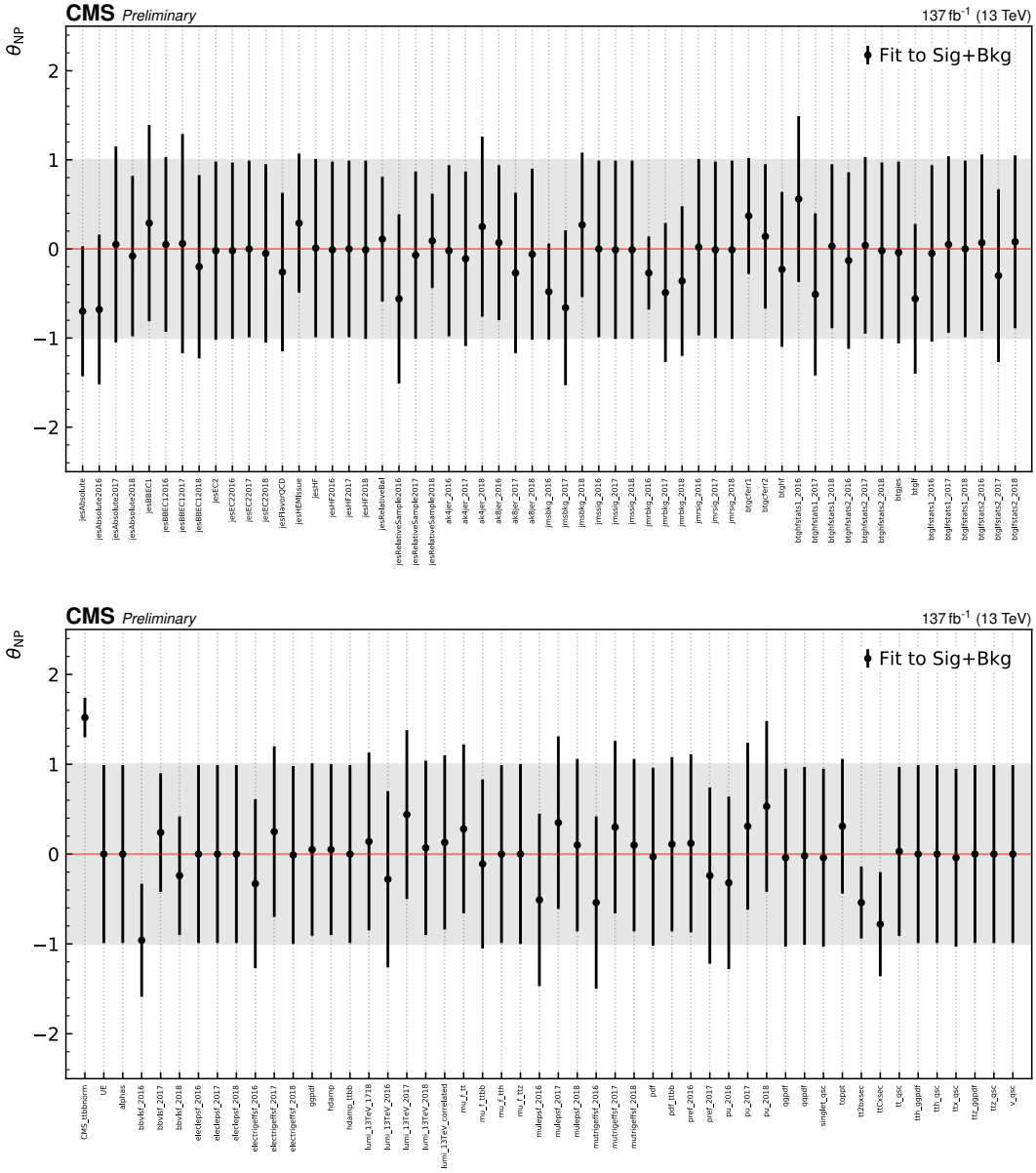

Figure D.3: The change in the best-fit values and constraints of the model's nuisance parameters after fitting to the full Run 2 data set. The parameters in the figures are arranged by the source of the nuisance, i.e. JES, JER, JMS, JMR, and b-tag efficiency (top) and the remaining non-MC-statistics related nuisances (bottom).

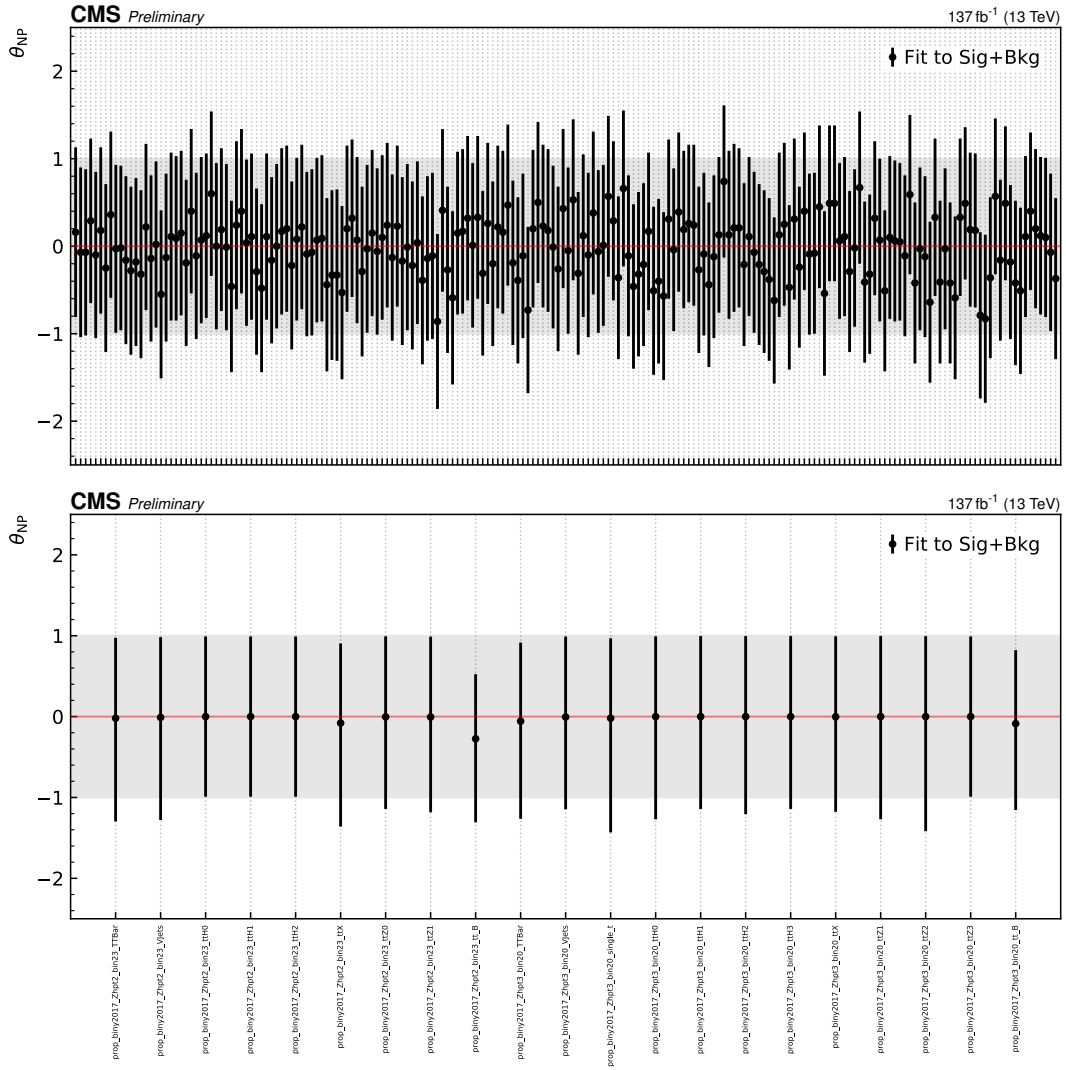

Figure D.4: The change in the best-fit values and constraints of the model's nuisance parameters after fitting to the full Run 2 data set. The parameters in the figures are arranged by the source of the nuisance, i.e. the MC-statistics nuisances with a Gaussian prior (top) and the MC-statistics with a Poisson prior (bottom).

BIBLIOGRAPHY

[1] C. P. Burgess and G. D. Moore, "The standard model: A primer". Cambridge University Press, 12, 2006. ISBN 978-0-511-25485-7, 978-1-107-40426-7, 978-0-521-86036-9.

[2] A. J. Larkoski, "Elementary Particle Physics: An Intuitive Introduction". Cambridge University Press, 2019. doi:10.1017/9781108633758.

[3] M. E. Peskin and D. V. Schroeder, "An introduction to quantum field theory". Westview, Boulder, CO, 1995.

[4] M. Thomson, "Modern Particle Physics". Cambridge University Press, 2013. doi:10.1017/CBO9781139525367.

[5] T. Lancaster and S. J. Blundell, "Quantum field theory for the gifted amateur". Oxford University Press, Oxford, 2014. doi:10.1093/acprof:oso/9780199699322.001.0001, ISBN 9780199699322.

[6] D. Griffiths, "Introduction to Elementary Particles: Second Revised Edition". Wiley VCH, 2008. ISBN 9783527406012.

[7] W. Commons, "File:standard model of elementary particles.svg — wikimedia commons, the free media repository", 2021. [Online; accessed 15-November-2021]. https://commons.wikimedia.org/w/index.php?title=File:Standard_Model_of_Elementary_Particles.svg&oldid=585729208.

[8] LHC Top Working Group, "LHCTopWG Summary Plots". TWiki: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCTopWGSummaryPlots r72.

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

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

[11] J. Ellis, M. K. Gaillard, and D. V. Nanopoulos, "An updated historical profile of the higgs boson", doi:10.1142/9789814733519_0014, arXiv:1504.07217.

[12] CMS Collaboration, "A measurement of the Higgs boson mass in the diphoton decay channel", *Phys. Lett. B* **805** (2020) 135425, doi:10.1016/j.physletb.2020.135425, arXiv:2002.06398.

[13] Planck Collaboration, "Planck 2018 results. VI. cosmological parameters", *Astron. Astrophys.* **641** (2020) A6, doi:10.1051/0004-6361/201833910, arXiv:1807.06209.

[14] V. C. Rubin and W. K. Ford, Jr., "Rotation of the Andromeda nebula from a spectroscopic survey of emission regions", *Astrophys. J.* **159** (1970) 379, doi:10.1086/150317.

[15] D. Clowe, A. Gonzalez, and M. Markevitch, "Weak lensing mass reconstruction of the interacting cluster 1E0657-558: Direct evidence for the existence of dark matter", *Astrophys. J.* **604** (2004) 596, doi:10.1086/381970, arXiv:astro-ph/0312273.

[16] Y. Shadmi, "Introduction to Supersymmetry", in *2014 European School of High-Energy Physics*, p. 95. 2016. arXiv:1708.00772. doi:10.5170/CERN-2016-003.95.

[17] W. Buchmuller and D. Wyler, "Effective Lagrangian analysis of new interactions and flavor conservation", *Nucl. Phys. B* **268** (1986) 621, doi:10.1016/0550-3213(86)90262-2.

[18] B. Grzadkowski, M. Iskrzynski, M. Misiak, and J. Rosiek, "Dimension-six terms in the standard model Lagrangian", *JHEP* **10** (2010) 085, doi:10.1007/JHEP10(2010)085, arXiv:1008.4884.

[19] A. Falkowski and R. Rattazzi, "Which EFT", *JHEP* **10** (2019) 255, doi:10.1007/JHEP10(2019)255, arXiv:1902.05936.

[20] C. Degrande et al., "Effective field theory: A modern approach to anomalous couplings", *Annals Phys.* **335** (2013) 21, doi:10.1016/j.aop.2013.04.016, arXiv:1205.4231.

[21] A. Kobach, "Baryon number, lepton number, and operator dimension in the standard model", *Phys. Lett. B* **758** (2016) 455, doi:10.1016/j.physletb.2016.05.050, arXiv:1604.05726.

[22] F. Marcastel, "CERN's Accelerator Complex. La chane des acclrateurs du CERN.", (2013). General Photo.

[23] O. S. Brning et al., "LHC Design Report". CERN Yellow Reports. CERN, 2004. doi:10.5170/CERN-2004-003-V-1.

[24] J.-L. Caron, "Cross section of LHC dipole. Dipole LHC: coupe transversale.", (May, 1998). AC Collection. Legacy of AC. Pictures from 1992 to 2002.

[25] W. Herr and B. Muratori, "Concept of luminosity", doi:10.5170/CERN-2006-002.361.

[26] CMS Collaboration, "CMS Luminosity Public Results". TWiki: https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults r164.

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

[28] CMS Collaboration, "CMS Physics: Technical Design Report Volume 1: Detector Performance and Software", Technical Design Report CERN-LHCC-2006-001, CMS-TDR-8-1, 2006.

[29] CMS Collaboration, "CMS Physics: Technical Design Report Volume 2: Physics Performance", *J. Phys. G* **34** (2007) 995, doi:10.1088/0954-3899/34/6/S01.

[30] CMS Collaboration, "Cutaway diagrams of CMS detector.", (2019).

[31] F. Kircher et al., "Magnetic tests of the CMS superconducting magnet", *IEEE Transactions on Applied Superconductivity* **18** (2008) 356, doi:10.1109/TASC.2008.920571.

[32] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", *JINST* **9** (2014), no. 10, P10009, doi:10.1088/1748-0221/9/10/p10009.

[33] CMS Collaboration, "CMS Technical Design Report for the Pixel Detector Upgrade", Technical Design Report CERN-LHCC-2012-016, CMS-TDR-11, 2012.

[34] CMS Collaboration, "The CMS electromagnetic calorimeter project", Technical Design Report CERN-LHCC-97-033, CMS-TDR-4, 1997.

[35] J. Williams, Jun, 2008. https://hepwww.pp.rl.ac.uk/groups/CMSvpt/bestphotos/index.htm.

[36] "Last crystals for the CMS chandelier. Arrive des derniers cristaux de CMS", 2008. http://cds.cern.ch/record/1101276.

[37] CMS Collaboration, P. K. Siddireddy, "The CMS ECAL trigger and DAQ system: electronics auto-recovery and monitoring", 2018. arXiv:1806.09136.

[38] CMS Collaboration, "The CMS hadron calorimeter project", Technical Design Report CERN-LHCC-97-031, CMS-TDR-2, 1997.

[39] M. Brice, L. Vaillet, and L. Lazic, "Images of the CMS HCAL Barrel (HB)", (2008). CMS Collection.

[40] T. Virdee, M. Brice, and L. Veillet, "Images of CMS HCAL Endcap (HE)", (2008). CMS Collection.

[41] M. Brice, T. Virdee, and T. Camporesi, "Images of CMS HCAL Forward Calorimeter (HF)", (2008). CMS Collection.

[42] CMS Collaboration, "CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter", Technical Design Report CERN-LHCC-2012-015, CMS-TDR-10, 2012.

[43] CMS Collaboration, "The CMS muon project", Technical Design Report CERN-LHCC-97-031, CMS-TDR-2, 1997.

[44] CMS Collaboration, "The CMS trigger system", *JINST* **12** (2017) P01020, doi:10.1088/1748-0221/12/01/P01020, arXiv:1609.02366.

[45] V. Gori, "The CMS high level trigger", *International Journal of Modern Physics: Conference Series* **31** (2014) 1460297, doi:10.1142/s201019451460297x.

[46] J. Brooke et al., "Hardware and firmware for the CMS global calorimeter trigger", in *9th Workshop on Electronics for LHC Expriments*, p. 226. 2003. doi:10.5170/CERN-2003-006.226.

[47] I. Bird et al., "Update of the computing models of the WLCG and the LHC experiments", Technical Design Report CERN-LHCC-2014-014, LCG-TDR-002, 2014.

[48] A. Adekitan, "Monte Carlo simulation", doi:10.13140/RG.2.2.15207.16806.

[49] S. Hche, "Introduction to parton-shower event generators", in *Theoretical Advanced Study Institute in Elementary Particle Physics: Journeys Through the Precision Frontier: Amplitudes for Colliders*, p. 235. 2015. arXiv:1411.4085. doi:10.1142/9789814678766_0005.

[50] NNPDF Collaboration, "Parton distributions for the LHC Run II", *JHEP* **04** (2015) 040, doi:10.1007/JHEP04(2015)040, arXiv:1410.8849.

[51] NNPDF Collaboration, "Parton distributions from high-precision collider data", *Eur. Phys. J. C* **77** (2017) 663, doi:10.1140/epjc/s10052-017-5199-5, arXiv:1706.00428.

[52] S. Alioli, P. Nason, C. Oleari, and E. Re, "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX", *JHEP* **06** (2010) 043, doi:10.1007/JHEP06(2010)043, arXiv:1002.2581.

[53] H. B. Hartanto, B. Jäger, L. Reina, and D. Wackeroth, "Higgs boson production in association with top quarks in the POWHEG BOX", *Phys. Rev. D* **91** (2015) 094003, doi:10.1103/PhysRevD.91.094003, arXiv:1501.04498.

[54] S. Alioli, P. Nason, C. Oleari, and E. Re, "NLO single-top production matched with shower in POWHEG: s- and t-channel contributions", *JHEP* **09** (2009) 111, doi:10.1088/1126-6708/2009/09/111, arXiv:0907.4076.

[55] E. Re, "Single-top Wt-channel production matched with parton showers using the POWHEG method", *Eur. Phys. J. C* **71** (2011) 1547, doi:10.1140/epjc/s10052-011-1547-z, arXiv:1009.2450.

[56] J. Alwall et al., "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations", *JHEP* **07** (2014) 079, doi:10.1007/JHEP07(2014)079, arXiv:1405.0301.

[57] C. Degrande et al., "UFO - The Universal FeynRules Output", *Comput. Phys. Commun.* **183** (2012) 1201, doi:10.1016/j.cpc.2012.01.022, arXiv:1108.2040.

[58] CMS Collaboration, "Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements", *Eur. Phys. J. C* **80** (2020) 4, doi:10.1140/epjc/s10052-019-7499-4, arXiv:1903.12179.

[59] R. Frederix and S. Frixione, "Merging meets matching in MC@NLO", *JHEP* **12** (2012) 061, doi:10.1007/JHEP12(2012)061, arXiv:1209.6215.

[60] M. L. Mangano, M. Moretti, F. Piccinini, and M. Treccani, "Matching matrix elements and shower evolution for top-pair production in hadronic collisions", *JHEP* **01** (2007) 013, doi:10.1088/1126-6708/2007/01/013, arXiv:hep-ph/0611129.

[61] CMS Collaboration, "Event generator tunes obtained from underlying event and multiparton scattering measurements", *Eur. Phys. J. C* **76** (2016) 155, doi:10.1140/epjc/s10052-016-3988-x, arXiv:1512.00815.

[62] GEANT4 Collaboration, "GEANT4—a simulation toolkit", *Nucl. Instrum. Meth. A* **506** (2003) 250, doi:10.1016/S0168-9002(03)01368-8.

[63] CMS Collaboration, "Particle-flow reconstruction and global event description with the CMS detector", *JINST* **12** (2017) P10003, doi:10.1088/1748-0221/12/10/P10003, arXiv:1706.04965.

[64] R. Frhwirth, W. Waltenberger, and P. Vanlaer, "Adaptive vertex fitting", CMS Note CMS-NOTE-2007-008, 2007.

[65] CMS Collaboration, "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV", *JINST* **13** (2018) P06015, doi:10.1088/1748-0221/13/06/P06015, arXiv:1804.04528.

[66] W. Adam, R. Frhwirth, A. Strandlie, and T. Todorov, "Reconstruction of electrons with the Gaussian-sum filter in the CMS tracker at the LHC", *J. Phys. G* **31** (2005), no. 9, N9, doi:10.1088/0954-3899/31/9/n01.

[67] CMS Collaboration, "Electron Identification Based on Simple Cuts". TWiki: https://twiki.cern.ch/twiki/bin/view/CMSPublic/EgammaPublicData#Implementing_the_Simple_Cut_Base r5.

[68] CMS Collaboration, "Electron Cut Based ID for 94X samples". Indico: https://indico.cern.ch/event/732971/contributions/3022843/attachments/1658685/2656462/eleIdTuning.pdf.

[69] CMS Collaboration, "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV", *JINST* **10** (2015) P06005, doi:10.1088/1748-0221/10/06/P06005, arXiv:1502.02701.

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

[71] CMS Collaboration, "Pileup mitigation at CMS in 13 TeV data", *JINST* **15** (2020), no. 09, P09018, doi:10.1088/1748-0221/15/09/P09018, arXiv:2003.00503.

[72] CMS Collaboration, "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV", *JINST* **12** (2017) P02014, doi:10.1088/1748-0221/12/02/P02014, arXiv:1607.03663.

[73] M. Dasgupta, A. Fregoso, S. Marzani, and G. P. Salam, "Towards an understanding of jet substructure", *JHEP* **09** (2013) 029, doi:10.1007/JHEP09(2013)029, arXiv:1307.0007.

[74] CMS Collaboration, "A Cambridge-Aachen (C-A) based Jet Algorithm for boosted top-jet tagging", CMS Physics Analysis Summary CMS-PAS-JME-09-001, 2009.

[75] Y. L. Dokshitzer, G. D. Leder, S. Moretti, and B. R. Webber, "Better jet clustering algorithms", *JHEP* **08** (1997) 001, doi:10.1088/1126-6708/1997/08/001, arXiv:hep-ph/9707323.

[76] M. Wobisch and T. Wengler, "Hadronization corrections to jet cross-sections in deep inelastic scattering", in *Workshop on Monte Carlo Generators for HERA Physics*, p. 270. 1998. arXiv:hep-ph/9907280.

[77] CMS Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV", *JINST* **13** (2018) P05011, doi:10.1088/1748-0221/13/05/P05011, arXiv:1712.07158.

[78] CMS Collaboration, "Identification of heavy, energetic, hadronically decaying particles using machine-learning techniques", *JINST* **15** (2020) P06005, doi:10.1088/1748-0221/15/06/P06005, arXiv:2004.08262.

[79] CMS Collaboration, "Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 13$ TeV", *JHEP* **07** (2018) 161, doi:10.1007/JHEP07(2018)161, arXiv:1802.02613.

[80] CMS Collaboration, "Performance of the CMS level-1 trigger in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JINST* **15** (2020), no. 10, P10017, doi:10.1088/1748-0221/15/10/P10017, arXiv:2006.10165.

[81] M. Czakon et al., "Top-pair production at the LHC through NNLO QCD and NLO EW", *JHEP* **10** (2017) 186, doi:10.1007/JHEP10(2017)186, arXiv:1705.04105.

[82] CMS Collaboration, "Search for new physics in top quark production with additional leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV using effective field theory", *JHEP* **03** (2021) 095, doi:10.1007/JHEP03(2021)095, arXiv:2012.04120.

[83] CMS Collaboration, "Measurement of top quark pair production in association with a Z boson in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **56** (2020) doi:10.1007/JHEP03(2020)056, arXiv:1907.11270.

[84] CMS Collaboration, "Measurement of the inclusive and differential $\mathrm{t\bar{t}}\gamma$ cross sections in the single-lepton channel and EFT interpretation at $\sqrt{s} = 13$ TeV", *JHEP* **12** (2021) 180, doi:10.1007/JHEP12(2021)180, arXiv:2107.01508.

[85] CMS Collaboration, "Probing effective field theory operators in the associated production of top quarks with a Z boson in multilepton final states at $\sqrt{s} = 13$ TeV", *JHEP* **12** (2021) 083, doi:10.1007/JHEP12(2021)083, arXiv:2107.13896.

[86] CMS Collaboration, "Measurement of the cross section for top quark pair production in association with a W or Z boson in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **08** (2018) 011, doi:10.1007/JHEP08(2018)011, arXiv:1711.02547.

[87] CMS Collaboration, "Observation of single top quark production in association with a Z boson in proton-proton collisions at $\sqrt{s} = 13$ TeV", *Phys. Rev. Lett.* **122** (2019) 132003, doi:10.1103/PhysRevLett.122.132003, arXiv:1812.05900.

[88] ATLAS Collaboration, "Measurement of the $t\bar{t}Z$ and $t\bar{t}W$ cross sections in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", *Phys. Rev. D* **99** (2019) 072009, doi:10.1103/PhysRevD.99.072009, arXiv:1901.03584.

[89] CMS Collaboration, "Observation of $\mathrm{t\bar{t}H}$ production", *Phys. Rev. Lett.* **120** (2018) 231801, doi:10.1103/PhysRevLett.120.231801, arXiv:1804.02610.

[90] ATLAS Collaboration, "Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector", *Phys. Lett. B* **784** (2018) 173, doi:10.1016/j.physletb.2018.07.035, arXiv:1806.00425.

[91] A. J. Larkoski, S. Marzani, G. Soyez, and J. Thaler, "Soft drop", *JHEP* **05** (2014) 146, doi:10.1007/JHEP05(2014)146, arXiv:1402.2657.

[92] A. Kulesza et al., "Associated production of a top quark pair with a heavy electroweak gauge boson at NLO+NNLL accuracy", *Eur. Phys. J. C* **79** (2019) 249, doi:10.1140/epjc/s10052-019-6746-z, arXiv:1812.08622.

[93] LHC Higgs Cross Section Working Group Collaboration, "Handbook of LHC Higgs cross sections: 4. deciphering the nature of the Higgs sector", technical report, 10, 2016. doi:10.23731/CYRM-2017-002, arXiv:1610.07922.

[94] M. Beneke, P. Falgari, S. Klein, and C. Schwinn, "Hadronic top-quark pair production with NNLL threshold resummation", *Nucl. Phys. B* **855** (2012) 695, doi:10.1016/j.nuclphysb.2011.10.021, arXiv:1109.1536.

[95] M. Cacciari et al., "Top-pair production at hadron colliders with next-to-next-to-leading logarithmic soft-gluon resummation", *Phys. Lett. B* **710** (2012) 612, doi:10.1016/j.physletb.2012.03.013, arXiv:1111.5869.

[96] P. Bärnreuther, M. Czakon, and A. Mitov, "Percent-level-precision physics at the Tevatron: next-to-next-to-leading order QCD corrections to $q\bar{q} \to t\bar{t} + X$", *Phys. Rev. Lett.* **109** (2012) 132001, doi:10.1103/PhysRevLett.109.132001, arXiv:1204.5201.

[97] M. Czakon and A. Mitov, "NNLO corrections to top-pair production at hadron colliders: the all-fermionic scattering channels", *JHEP* **12** (2012) 054, doi:10.1007/JHEP12(2012)054, arXiv:1207.0236.

[98] M. Czakon and A. Mitov, "NNLO corrections to top pair production at hadron colliders: the quark-gluon reaction", *JHEP* **01** (2013) 080, doi:10.1007/JHEP01(2013)080, arXiv:1210.6832.

[99] M. Czakon, P. Fiedler, and A. Mitov, "Total top-quark pair-production cross section at hadron colliders through $O(\alpha_S^4)$", *Phys. Rev. Lett.* **110** (2013) 252004, doi:10.1103/PhysRevLett.110.252004, arXiv:1303.6254.

[100] M. Czakon and A. Mitov, "Top++: a program for the calculation of the top-pair cross-section at hadron colliders", *Comput. Phys. Commun.* **185** (2014) 2930, doi:10.1016/j.cpc.2014.06.021, arXiv:1112.5675.

[101] S. Quackenbush, R. Gavin, Y. Li, and F. Petriello, "W physics at the LHC with FEWZ 2.1", *Comput. Phys. Commun.* **184** (2013) 209, doi:10.1016/j.cpc.2012.09.005, arXiv:1201.5896.

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

[103] N. Kidonakis, "Two-loop soft anomalous dimensions for single top quark associated production with a W^- or H^-", *Phys. Rev. D* **82** (2010) 054018, doi:10.1103/PhysRevD.82.054018, arXiv:1005.4451.

[104] M. Aliev et al., "HATHOR: hadronic top and heavy quarks cross section calculator", *Comput. Phys. Commun.* **182** (2011) 1034, doi:10.1016/j.cpc.2010.12.040, arXiv:1007.1327.

[105] P. Kant et al., "HatHor for single top-quark production: updated predictions and uncertainty estimates for single top-quark production in hadronic collisions", *Comput. Phys. Commun.* **191** (2015) 74, doi:10.1016/j.cpc.2015.02.001, arXiv:1406.4403.

[106] T. Ježo, J. M. Lindert, N. Moretti, and S. Pozzorini, "New NLOPS predictions for $t\bar{t}+b$-jet production at the LHC", *Eur. Phys. J. C* **78** (2018) 502, doi:10.1140/epjc/s10052-018-5956-0, arXiv:1802.00426.

[107] F. Buccioni, S. Kallweit, S. Pozzorini, and M. F. Zoller, "NLO QCD predictions for $t\bar{t}b\bar{b}$ production in association with a light jet at the LHC", *JHEP* **12** (2019) 015, doi:10.1007/JHEP12(2019)015, arXiv:1907.13624.

[108] J. A. Aguilar-Saavedra et al., "Interpreting top-quark LHC measurements in the standard-model effective field theory", LHC TOP WG note CERN-LPCC-2018-01, 2018. arXiv:1802.07237.

[109] R. Goldouzian et al., "Matching in $pp \to t\bar{t}W/Z/h+$ jet SMEFT studies", *JHEP* **06** (2021) 151, doi:10.1007/JHEP06(2021)151, arXiv:2012.06872.

[110] A. Gron, "Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition". O'Reilly Media, Inc., 2019. ISBN 9781492032649.

[111] H. Brink, J. W. Richards, and M. Fetherolf, "Real-World Machine Learning". Manning Publication Co., 2016. ISBN 9781617291920.

[112] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, "Learning From Data". AMLBook, 2012. ISBN 9781600490064.

[113] A. LeNail, "Nn-svg: Publication-ready neural network architecture schematics", *Journal of Open Source Software* **4** (2019), no. 33, 747, doi:10.21105/joss.00747.

[114] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", (2017). arXiv:1412.6980.

[115] S. Ioffe and C. Szegedy, "Batch normalization: accelerating deep network training by reducing internal covariate shift", (2015). arXiv:1502.03167.

[116] G. E. Hinton et al., "Improving neural networks by preventing co-adaptation of feature detectors", (2012). arXiv:1207.0580.

[117] ATLAS Collaboration, "Measurement of event shapes at large momentum transfer with the ATLAS detector in pp collisions at $\sqrt{s} = 7$ TeV", *Eur. Phys. J. C* **72** (2012) 2211, doi:10.1140/epjc/s10052-012-2211-y, arXiv:1206.2135.

[118] P. E. Latham and Y. Roudi, "Mutual information", *Scholarpedia* **4** (2009), no. 1, 1658, doi:10.4249/scholarpedia.1658.

[119] F. Chollet et al., "Keras", 2015. https://github.com/fchollet/keras.

[120] M. Abadi, A. Agarwal, P. Barham et al., "TensorFlow: Large-scale machine learning on heterogeneous systems", 2015. Software available from tensorflow.org. https://www.tensorflow.org.

[121] T.-Y. Lin et al., "Focal loss for dense object detection", (2018). arXiv:1708.02002.

[122] Particle Data Group, M. Tanabashi et al., "Review of particle physics", *Phys. Rev. D* **98** (2018) 030001, doi:10.1103/PhysRevD.98.030001.

[123] F. James, "MINUIT: Function minimization and error analysis reference manual", 1998. CERN Program Library Long Writeups.

[124] R. J. Barlow and C. Beeston, "Fitting using finite Monte Carlo samples", *Comput. Phys. Commun.* **77** (1993) 219, doi:10.1016/0010-4655(93)90005-W.

[125] J. S. Conway, "Incorporating nuisance parameters in likelihoods for multisource spectra", in *PHYSTAT 2011*, p. 115. 2011. arXiv:1103.0354. doi:10.5170/CERN-2011-006.115.

[126] CMS Collaboration, "Measurement of differential cross sections for top quark pair production using the lepton+jets final state in proton-proton collisions at 13 TeV", *Phys. Rev. D* **95** (2017) 092001, doi:10.1103/PhysRevD.95.092001, arXiv:1610.04191.

[127] CMS Collaboration, "First measurement of the cross section for top quark pair production with additional charm jets using dileptonic final states in pp collisions at $\sqrt{s} = 13$ TeV", *Phys. Lett. B* **820** (2021) 136565, doi:10.1016/j.physletb.2021.136565, arXiv:2012.09225.

[128] G. Ridolfi, M. Ubiali, and M. Zaro, "A fragmentation-based study of heavy quark production", *JHEP* **01** (2020) 196, doi:10.1007/JHEP01(2020)196, arXiv:1911.01975.

[129] CMS Collaboration, "Precision luminosity measurement in proton-proton collisions at $\sqrt{s} = 13$ TeV in 2015 and 2016 at CMS", (2021). arXiv:2104.01927.

[130] CMS Collaboration, "CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV", CMS Physics Analysis Summary CMS-PAS-LUM-17-004, 2018.

[131] CMS Collaboration, "CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV", CMS Physics Analysis Summary CMS-PAS-LUM-18-002, 2019.

[132] CMS Collaboration, "Measurement of the cross section for $\mathrm{t\bar{t}}$ production with additional jets and b jets in pp collisions at $\sqrt{s} = 13$ TeV", *JHEP* **07** (2020) 125, doi:10.1007/JHEP07(2020)125, arXiv:2003.06467.

[133] CMS Collaboration, "Measurement of the $\mathrm{t\bar{t}b\bar{b}}$ production cross section in the all-jet final state in pp collisions at $\sqrt{s} = 13$ TeV", *Phys. Lett. B* **803** (2020) 135285, doi:10.1016/j.physletb.2020.135285, arXiv:1909.05306.

[134] CMS Collaboration, "Measurement and interpretation of differential cross sections for Higgs boson production at $\sqrt{s} = 13$ TeV", *Phys. Lett. B* **792** (2019) 369, doi:10.1016/j.physletb.2019.03.059, arXiv:1812.06504.

[135] CMS Collaboration, "Measurement of the inclusive and differential $\mathrm{t\bar{t}}\gamma$ cross sections in the dilepton channel and effective field theory interpretation in proton-proton collisions at $\sqrt{s}$ =13 TeV", (2022). arXiv:2201.07301. Submitted to *JHEP*.

[136] C. Degrande et al., "Automated one-loop computations in the standard model effective field theory", *Phys. Rev. D* **103** (2021) 096024, doi:10.1103/PhysRevD.103.096024, arXiv:2008.11743.

[137] C. J. Clopper and E. S. Pearson, "The use of confidence or fiducial limits illustrated in the case of the binomial", *Biometrika* **26** (12, 1934) 404–413, doi:10.1093/biomet/26.4.404.