University of California
Santa Barbara

Search for $t\bar{t}H(H\rightarrow c\bar{c})$ in the single-lepton channel using the full Run II data sample

A dissertation submitted in partial satisfaction
of the requirements for the degree

Doctor of Philosophy
in
Physics

by

Phillip Tyler Masterson

Committee in charge:

Professor Joseph Incandela, Chair
Professor Claudio Campagnari
Professor Nathaniel Craig

March 2024

The Dissertation of Phillip Tyler Masterson is approved.

Professor Claudio Campagnari

Professor Nathaniel Craig

Professor Joseph Incandela, Committee Chair

December 2023

Search for $t\bar{t}H(H\rightarrow c\bar{c})$ in the single-lepton channel using the full Run II data sample

by

Phillip Tyler Masterson

To Karen Masterson, who taught me to follow my dreams. To Daniel Masterson, who showed me how to persevere on the journey there. And to Ilse Denchfield, who left footprints I hope I will someday be standing in.

Acknowledgements

This dissertation could not have been written without a tremendous amount of support, both direct and indirect. To everyone who offered help along the way, regardless of its form, I owe a huge debt of gratitude.

Firstly, I would like to thank everyone who I've had the privilege of working alongside during the past five years. Loukas, for for the advice you offered throughout my stay at CERN and for your guidance on how to grow as a physicist. Huilin, for your invaluable expertise and unfailing willingness to give feedback. Joe, for the small fraction of your experience that you've managed to pass on to me. Ioannis, Maarten, Jan, Sebastian, and the rest of my analysis group, for everything that you've done for both your own pieces of the project and my own channel. Melissa, for offering an experienced hand throughout my entire stay at CERN. Valentina, Umesh, Alethea, Susanne, and the rest of the CMS . Watching this analysis come together has been a hugely rewarding experience, and I couldn't have gotten this far without you.

I would also like to thank the enormous number of people I've met and friends I've made along the way. Raghav, for the countless late nights in Geneva and for the occasional shove out of my comfort zone. Farouk and Amina, for the many relaxed days in the office and for the many friends I've met through you. Rea, Naomi, Sanje, Daniel, Sofia, and the rest of the lunch group for keeping my afternoons interesting and for the amazing home-cooked meals. Emily, Mason, and the seemingly-infinite number of people I met through Melissa, for the fond memories of skiing, trivia and floating down the Arve. The rest of the trivia crew, for the great accomplishment of not placing dead last every time. Amy and Leo, for being awesome housemates. Ryan, Erin, Michael, and everyone else from UCSB, for the fond memories in Santa Barbara. The regulars in the CERN games club and our DND campaign, for more than a fair amount of hijinks. The last

five years wouldn't have been anywhere near as incredible as they were without you all.

Finally, I owe more than I can express to my family. Mom and dad, for helping me become the person that I am. Thomas, for keeping me grounded no matter how high up in the atmosphere you happen to be. Emily, for adding color to everything that you're involved in. Mackenzie, for always offering a listening ear and for tying the four of us together. And lastly, Omi, for your quiet strength and inspirational generosity. Thank you all so much.

Curriculum Vitæ

Phillip Tyler Masterson

Education

2024	Ph.D. in Physics (Expected), University of California, Santa Barbara.
2021	M.A. in Physics, University of California, Santa Barbara.
2018	B.S. in Physics, University of California, Santa Barbara.
2018	B.S. in Computing, University of California, Santa Barbara.

Professional appointments

2020 - 2024	Graduate Student Researcher, Department of Physics, University of California, Santa Barbara.
2018 - 2020	Teaching Assistant, Department of Physics, University of California, Santa Barbara.

Publications

The LDMX collaboration, Group, C. et al. Photon-rejection Power of the Light Dark Matter eXperiment in an 8 GeV Beam. J. High Energ. Phys. 2020, 3 (2020).

The CMS collaboration, Acar B. et al. Performance of the CMS High Granularity Calorimeter prototype to charged pion beams of 20−300 GeV/c. Journal of Instrumentation, Volume 18 (2023).

The CMS collaboration, Acar B. et al. Response of a CMS HGCAL silicon-pad electromagnetic calorimeter prototype to 20-300 GeV positrons. Journal of Instrumentation, Volume 17 (2022).

The CMS collaboration, Acar B. et al. Construction and commissioning of CMS CE prototype silicon modules. Journal of Instrumentation, Volume 16 (2021).

The CMS collaboration, Acar B. et al. The DAQ system of the 12,000 Channel CMS High Granularity Calorimeter Prototype. Journal of Instrumentation, Volume 16 (2021).

The LDMX collaboration, Akesson, T., Blinov, N. et al. A high efficiency photon veto for the Light Dark Matter eXperiment. J. High Energ. Phys. 2020, 3 (2020).

Fields of study

Major field: High-energy particle physics

Studies in experimental Higgs physics with Professor Joseph Incandela

Studies in the CMS HGCAL ugprade with Professor Joseph Incandela

Studies in Light Dark Matter eXperiment design with Professor Joseph Incandela

Abstract

Search for $t\bar{t}H(H\rightarrow c\bar{c})$ in the single-lepton channel using the full Run II data sample

by

Phillip Tyler Masterson

Higgs decay to a pair of charm quarks is a process that has potential to shed light on the standard model Higgs-charm Yukawa coupling and BSM physics in the form of two Higgs doublet models (2HDMs). However, it has not yet been observed experimentally. TThis thesis presents the current status of a search for the titular $t\bar{t}H(h\rightarrow c\bar{c})$ process. The primary focus in this work is the single-lepton channel, which will utilize the full CMS Run II data set of 138 fb^{-1}. This analysis takes advantage of recent advances in machine learning to obtain high jet-tagging and event classification performance, and reports a preliminary new constraint on the $t\bar{t}H(h\rightarrow c\bar{c})$ cross-section.

Contents

Chapter 1

Introduction

Despite being only slightly over fifty years old, the standard model of particle physics (SM) is one of the most successful scientific theories in history. The SM encompasses three of the four fundamental forces—electromagnetism, the weak force, and the strong force, with gravity remaining the exception—within the framework of quantum field theory, and in doing so produces an impressive variety of predictions that span many orders of magnitude in energy and interaction strengths. Research in particle physics over the past several decades has corroborated the SM's explanatory power, from extremely high-precision measurements of the electron's magnetic moment[1] to the discovery of the Higgs boson in 2012[2][3].

However, the SM is far from a complete theory of everything. In addition to its famous incompatibility with our current best theory of gravity, general relativity, the SM struggles to explain observations such as dark matter, the matter-antimatter asymmetry of the universe, and the strong force's preservation of CP symmetry. Recent astrophysical observations and theoretical considerations have provided tantalizing hints about the nature of beyond-the-standard-model (BSM) physics, but direct experimental detection of BSM phenomena, e.g. particulate dark matter, has so far evaded physicists. Con-

sequently, a key priority of modern particle physics is to precision-test the SM. Both possible outcomes—extending the range of physics that we know the SM describes, or direct experimental access to BSM physics—will help advance the field.

Due to its relatively recent discovery and critical role in the standard model, the Higgs boson is a natural starting point for BSM physics searches. In particular, couplings between the Higgs and other SM particles could be sensitive to yet-undiscovered physics. Over the past several years, the CMS experiment has measured couplings between the Higgs and several SM particles to good precision. See Figure 1.1 for a visualization. The next-easiest coupling to measure is the coupling between the Higgs and the charm quark. If the Higgs-charm Yukawa coupling y_c matches the SM value, the decay process $H \to c\bar{c}$ will likely be too difficult to observe at the LHC in the immediate future. However, BSM physics could enhance the branching fraction of this decay mode to the point where it could be observed with previously-recorded data.

This thesis describes a search for the process $gg \to t\bar{t}H$ ($H \to c\bar{c}$) using Run 2 CMS data. Due to the small SM value of y_c, the difficulty of distinguishing charm jets from other jet flavors, and the presence of large QCD backgrounds, a constraint within an order of magnitude or so of $y_{c(\mathrm{SM})}$ can only be achieved by making full use of cutting-edge analysis techniques. The crucial task of jet tagging in this analysis is handled by ParticleNet, a point cloud-based neural net architecture that achieves state-of-the-art performance on a wide variety of jet tagging benchmarks. After splitting events into three independent channels—dilepton, single-lepton, and fully-hadronic—events are classified by a transformer-based neural net, Particle Transformer, into seven different categories. Five categories correspond to key background processes and are used as control regions. To better handle the irreducible background from the similar process $t\bar{t}H$ ($H \to b\bar{b}$), the remaining two categories are used as signal regions, one for $t\bar{t}H$ ($H \to b\bar{b}$) and one for the process of interest $t\bar{t}H$ ($H \to c\bar{c}$). The observed limits on the $t\bar{t}H$ ($H \to c\bar{c}$) cross-section

are subsequently determined by standard fitting methods.

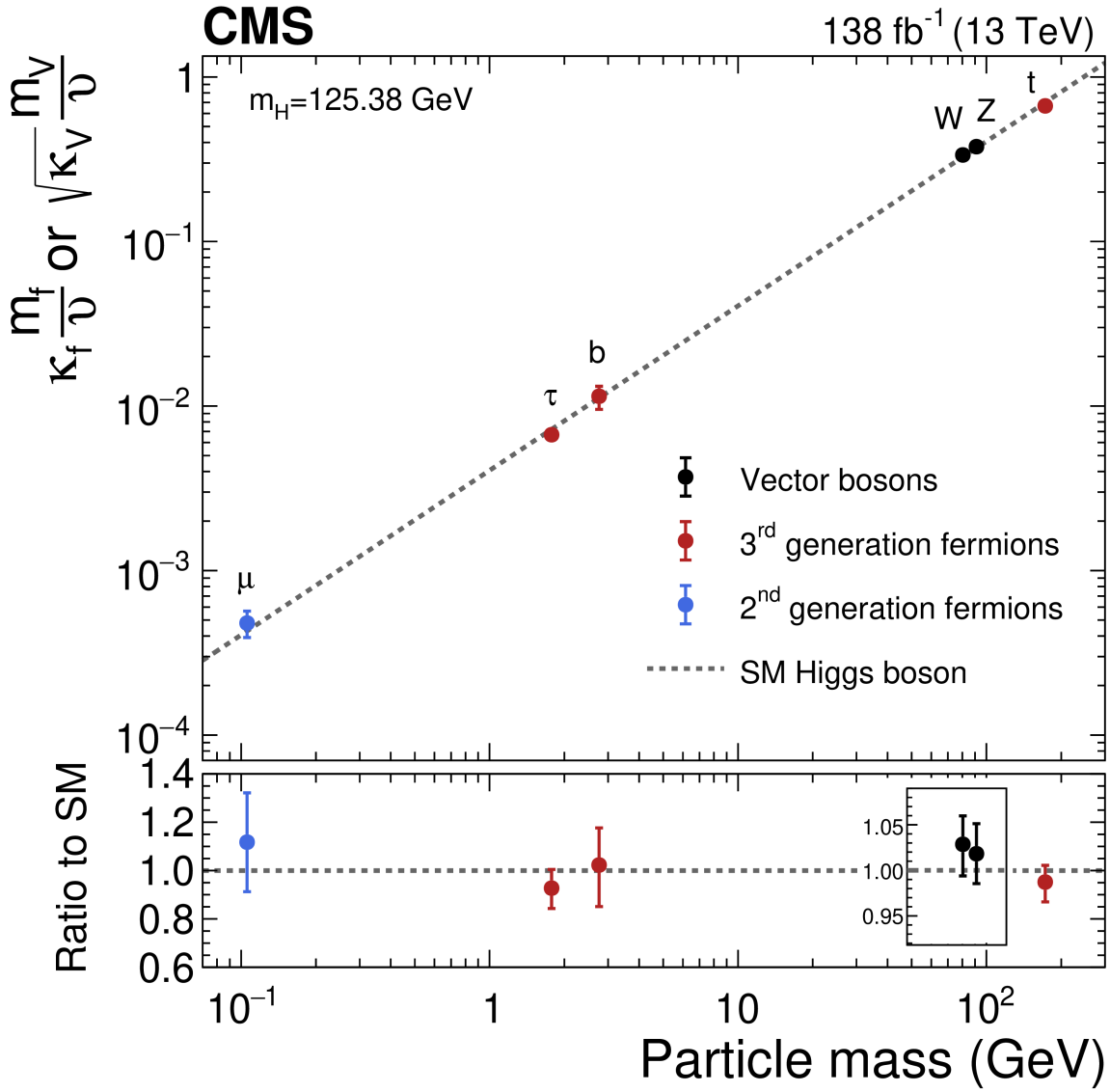


Figure 1.1: Figure of all currently observed couplings between the Higgs and standard model particles. So far, all have values consistent with standard model predictions. The most recently measured was the b coupling in 2018[4]; the c quark is the easiest of the remaining couplings to observe. From [5].

Chapter 2

The Standard Model and Beyond

The standard model is a renormalizable quantum field theory with nineteen free parameters. It encompasses seventeen types of fundamental particles, each of which are excited states of their corresponding quantum fields, and characterizes the interactions between them with three forces of nature. Using the formalism of quantum field theory, a theoretical framework that unifies special relativity and quantum mechanics, the SM makes a diverse list of predictions that accurately describe most phenomena that experimental physicists have observed to date. However, the SM is widely expected to be a low-energy approximation of a more fundamental theory due to several experimental and theoretical shortcomings, and many of its proposed extensions are testable in the near future.

The following chapter summarizes the most important aspects of the standard model. Section 2.1 provides an overview of the SM's particle content. Section 2.2 goes into detail on each of the three fundamental forces that the SM describes, as well as the Higgs mechanism and the role it plays in mass generation. Lastly, section 2.3 highlights a handful of problems with the SM and proposed solutions that will motivate the search described in this thesis.

2.1 Particle content of the standard model

The particles of the standard model fall into two overarching categories, distinguished by their spin: Fermions, with half-integer spin, and bosons, with integer spin. Fermions obey a type of quantum statistics known as Fermi-Dirac statistics, which describes systems of identical particles where no two particles may occupy the same quantum state (the Pauli exclusion principle). Conversely, bosons are described by Bose-Einstein statistics, which follows from the assumption that two identical particles may occupy the same quantum state. Fermions are massive and constitute all known matter in the universe, while bosons are responsible for mediating the interactions between fermions, and in the case of the Higgs, explaining the origin of their masses.

The standard model has twelve fermions, which can be further subdivided into two categories: quarks and leptons. The six quarks are distinguished by their color charges of red, green, and blue, and their ability to interact with the strong force, as well as their fractional electric charges of $+\frac{2}{3}$ or $-\frac{1}{3}$ for "up-type" and "down-type" quarks, respectively. Quarks may also interact and change flavor via the weak force. Like all fermions, quarks come in three generations; each successive generation of quarks is physically identical apart from their successively greater masses. Up (u) and down (d) quarks comprise the first generation, and their low masses of a few MeV make them the lightest quarks. The second-generation quarks, charm (c) and strange (s), have intermediate masses, while the third-generation of top (t) and bottom (b) quarks are the heaviest fundamental fermions. The unique nature of the strong force—specifically, the fact that it does not grow weaker with distance, discussed later in this section—keeps quarks confined within composite particles called hadrons. When produced in high-energy collisions, quarks will create new hadrons around themselves in order to maintain color confinement; these clusters of particles are known as jets. The top quark is an exception: As the heaviest

particle in the standard model, its high mass makes its lifetime shorter than the time it takes to hadronize, and consequently it will always decay into jets instead of forming hadrons.

Unlike quarks, the six leptons do not interact with the strong force and are not confined within composite particles. The first three, the electron (e), muon (μ), and tau (τ), have an electric charge of -1 and may therefore interact either electromagnetically or via the weak force. In contrast, the electron neutrino (ν_e), the muon neutrino (ν_μ), and the tau neutrino (ν_τ) interact only through the weak force. In the standard model, neutrinos are massless, although the discovery of neutrino oscillations in more recent years has proven that they are actually extremely light but massive.

Interactions between fermions are mediated by bosons, integer-spin particles that carry fundamental forces. The electromagnetic force is carried by photons (γ), massless spin-1 particles. The weak force is carried by three massive spin-1 particles, the electrically charged W^+ and W^- and the Z. Unlike in the case of electromagnetism, the high mass of the $W^\pm$ and Z—high enough that both decay before they can propagate over any significant distance—restricts the weak force to short ranges. Gluons mediate the strong force. Although usually regarded as a single particle, there are eight types of gluons; each is a superposition of states of color and anticolor. Gluons are massless, but generally confined within hadrons due to the unique properties of the strong force. Finally, the Higgs boson is a massive spin-0 particle that the standard model requires in order to explain how the W and Z bosons acquire mass. This Higgs mechanism will be discussed in-depth later in this thesis.

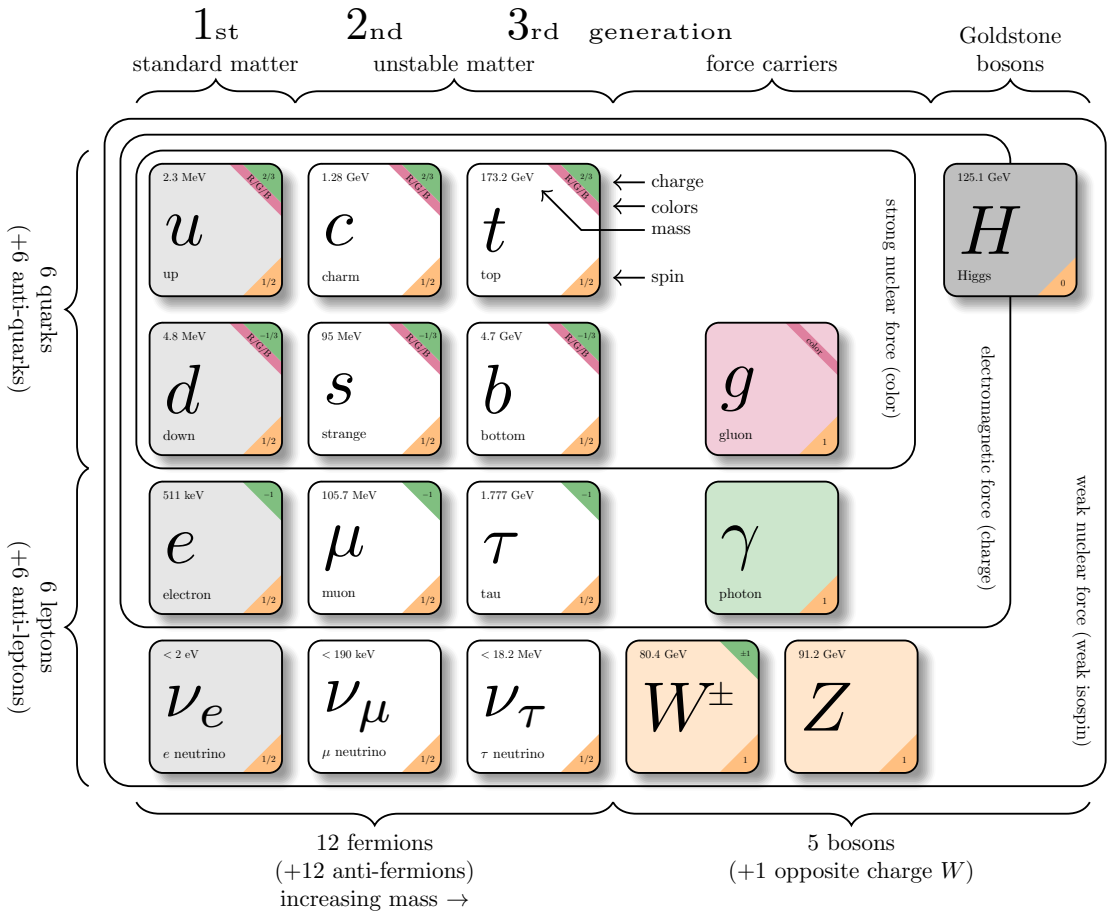


Figure 2.1: Table of all particles in the standard model, with the three generations of fermions occupying the leftmost three columns and bosons taking up the remainder. Boxes are used to indicate which forces a particle may interact with: Neutrinos inside the "weak nuclear force" box interact only via the weak force, leptons inside the "electromagnetic force" box interact via the weak force and the electromagnetic force, and so on for quarks. The electric charge of each particle is indicated by the numbers in green corners, while the numbers in orange corners denote the spin. Additionally, masses (or rough constraints if the mass is unknown) are given by the number above each particle's symbol; particles with no labeled mass are massless. (Every fermion in the first three columns has a corresponding antiparticle with opposite charge, but these are not listed.) Template modified from 2.1.

2.2 The standard model as a gauge field theory

As a quantum field theory, the standard model is represented by a Lagrangian density $\mathcal{L}_{SM}$. The dynamics of the standard model are derived using the principle of least action: Given the action

$$S = \int \mathcal{L} d^4x, \tag{2.1}$$

where the integral is performed over all spatial and temporal dimensions, the equations of motion are such that S is stationary. The form of the Lagrangian density itself is largely determined by its gauge symmetries. That is, the Lagrangian density (and hence the dynamics of the system) is invariant under certain local transformations, and this condition heavily constrains what form it may take.

To illustrate this, we may begin with the Lagrangian of a free fermion of mass m:

$$\mathcal{L}_f = \bar{\psi}(i\gamma^\mu \partial_\mu - m)\psi = i\bar{\psi}\not{\partial}\psi - m\bar{\psi}) \tag{2.2}$$

Here, ψ is a Dirac spinor, γ^μ denote the four Dirac matrices, and $\not{\partial}$ uses the standard Dirac notation

$$\not{\partial} = \gamma^\mu \partial_\mu. \tag{2.3}$$

We will see in the following sections that by modifying $\mathcal{L}_f$ to obey various local symmetries, we can begin to piece together the various components of the standard model.

2.2.1 Quantum electrodynamics

Although the electromagnetic force must be unified with the weak force to properly fit into the standard model, it is illustrative to first consider the case of quantum electrody-

namics (QED), as it emerges from a simple U(1) symmetry. Under a U(1) transformation, ψ transforms as

$$\psi \to e^{iq\chi(x)}\psi, \tag{2.4}$$

where q is an integer multiple of the electron charge e and $\chi(x)$ is a scalar function of the spacetime coordinates. Similarly,

$$\bar{\psi} \to e^{-iq\chi(x)}\bar{\psi}. \tag{2.5}$$

By default, $\mathcal{L}_f$ is not invariant under this transformation, producing an extra term $-q\bar{\psi}\psi\partial\!\!\!/\chi(x)$. However, we may solve this by introducing a new term into the Lagrangian:

$$\begin{aligned} \mathcal{L}_{\text{f+A}} &= \mathcal{L}_f - q\bar{\psi}A\!\!\!/\psi \\ &= i\bar{\psi}D\!\!\!/\psi - m\bar{\psi}) \end{aligned} \tag{2.6}$$

Here A_μ is a new vector field that transforms as $A_\mu \to A_\mu - \partial\chi(x)$, and $D_\mu = \partial_\mu + iqA_\mu$ is the gauge-covariant derivative. It can now readily be seen that imposing a U(1) gauge symmetry leads to the addition of a new boson, and a new force that it mediates, to our theory. However, the new terms added only determine the interactions between ψ and A. To make this Lagrangian into a proper quantum field theory, we must add a final term—the electromagnetic field tensor—that allows the new force carriers described by the field A_μ to propagate as free particles:

$$\mathcal{L}_{\text{QED}} = i\bar{\psi}D\!\!\!/\psi - m\bar{\psi}\psi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu}. \tag{2.7}$$

Here $F_{\mu\nu} = \partial_{mu}A_\nu - \partial_\nu A_\mu$. With some effort, Maxwell's equations may be derived from

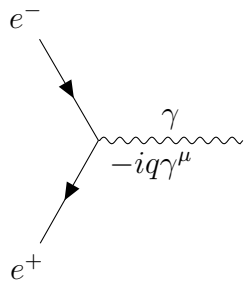


Figure 2.2: Feynman diagram of the fundamental QED vertex. Here the vertex itself is labeled with the corresponding vertex factor $-iq\gamma^\mu$, which will appear once in the final cross-section calculation for every vertex in the diagram. The vertex can be rotated freely so long as fermions and antifermions are exchanged in accordance with crossing symmetry.

the electromagnetic field tensor using the principle of least action.[6]

$\mathcal{L}_{\text{QED}}$ is the Lagrangian of quantum electrodynamics. It is sufficient to derive Feynman rules and diagrams for more complicated processes, starting from the fundamental QED vertex in figure 2.2. However, $\mathcal{L}_{\text{QED}}$ is a simplified model. To see how electrodynamics and the strong interaction are incorporated into the standard model, one must look beyond U(1) into non-abelian symmetry groups.

2.2.2 Quantum chromodynamics

Because electroweak interactions introduce a significant complication in the form of massive force carriers, we will focus on quantum chromodynamics (QCD) first. Like QED, QCD is invariant under transformations by a symmetry group, this time SU(3). Fermi spinors must be replaced by quark spinors, which transform as triplets under SU(3). However, while U(1) has a single generator, SU(3) has eight generators $T_a = \lambda_a/2$, where λ_a are the Gell-Mann matrices. We now now need eight vector fields to fix L_{QCD} instead of one. Additionally, SU(3) is non-abelian, as its generators T_a do not commute. The nonzero commutators lead to new terms in the field tensor:

$$\mathcal{L}_{\text{QCD}} = \bar{Q}(i\not{D} - m)Q - \frac{1}{4}G^{a\mu\nu}G^{a}_{\mu\nu} \tag{2.8}$$

$$G^{a}_{\mu\nu} = \partial_\mu G^{a}_{\nu} - \partial_\nu G^{a}_{\mu} + g_S f^{abc} G^{b}_{\mu} G^{c}_{\nu} \tag{2.9}$$

Here G^{a}_{μ} are newly-introduced vector fields, $D_\mu = \partial_\mu - ig_S T^a G^{a}_{\mu}$ is the new gauge-covariant derivative, and a is summed from 1 to 8. This is the QCD Lagrangian for a single fermion species. Although it appears similar to the QED Lagrangian at a glance, apart from the larger number of vector fields, there is a key difference: The new term in equation 2.9, which will produce terms with three and four fields in the field strength tensor when 2.8 is fully expanded out. This means that the gluon fields G^{a}_{μ} can self-interact, producing two new fundamental vertices as shown in Figure 2.3. These self-interacting terms make QCD far more complex than QED in practice, allowing for exotic states such as "glueballs" and introducing a variety of mathematical difficulties.

A larger complication arises when couplings are taken into account. When measuring couplings in the lab, higher-order Feynman diagrams will contribute to the process of interest. Like most processes in QFT, these contributions will depend on the energy scale k of the interaction, so the measured coupling will depend on k as well. In the case of QED, the coupling $\alpha(k^2)$ is relatively constant and of order 10^{-2} at low energies (<TeV). $\alpha(k^2)$'s smallness is mathematically convenient: Because each QED vertex contributes an additional factor of α when computing cross-sections and vertex corrections, higher-order Feynman diagrams will contribute exponentially less, and perturbation theory can be used to solve problems in QED with high accuracy by only considering the simplest few diagrams.

However, QCD is less cooperative. Its coupling constant $\alpha_S(k^2)$ is approximately:

$$\alpha_S(k^2) = \frac{\alpha_S(\mu^2)}{1 + B\alpha_S(\mu^2)\ln(\frac{k^2}{\mu^2})} \tag{2.10}$$

where μ is a reference energy known as the renormalization scale and B is a constant that depends on the number of low mass quark flavors ($m_q \ll \mu$). Although $\alpha_S(k^2)$ decreases with k, making high-energy calculations accessible, it increases to O(1) at energies below a few GeV. This makes perturbation theory useless at that scale, since higher-order Feynman diagrams will contribute heavily—a major problem, since most visible matter is made of low-energy quarks! QCD remains extremely difficult to handle mathematically, although modern techniques such as lattice QCD have found some success.

An interesting consequence of the running of α_S is color confinement. At extremely high energies or extremely short distances, $\alpha_S(k^2)$ shrinks, and the interaction strength between quarks in close proximity vanishes. Conversely, quarks at lower energies or larger distances interact much more strongly. This behavior is called asymptotic freedom. As a consequence, quarks in nature are always confined into colorless composite particles called hadrons. Free quarks cannot exist in nature—try to tear one out of a hadron, and the energy required will increase to the point where new quark-antiquark pairs will be created spontaneously, keeping the newly-separated quark confined within new hadrons. In experiments such as CMS, this process of hadronization shows up as condensed showers of particles known as jets.

The Lagrangian density of the standard model $\mathcal{L}_{SM}$ is invariant under the symmetry group SU(3) $\times$ SU(2) $\times$ SU(1). In particular, the dynamics of the strong force are invariant under SU(3), while those of the electroweak interaction are invariant under SU(2) $\times$ SU(1).

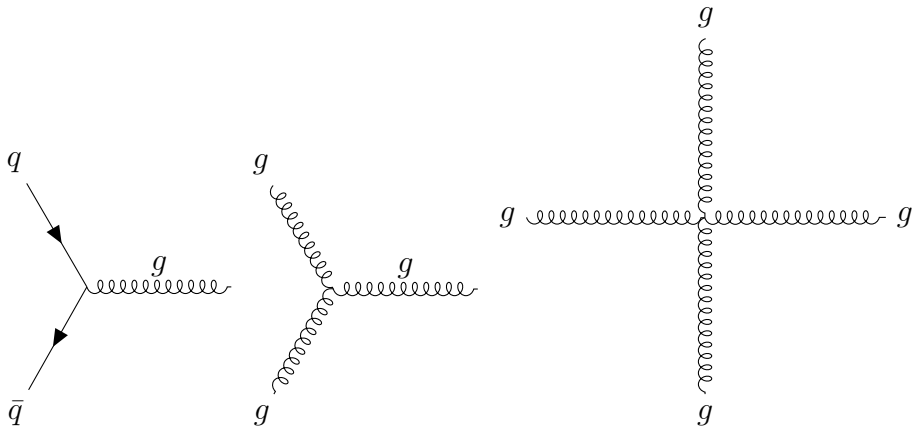


Figure 2.3: Feynman diagrams of the three allowed vertices in QCD. The first vertex is the only means by which fermions may interact with gluons; the second two are self-interaction vertices for gluons that have no analogue in QED.

2.2.3 Electroweak theory and spontaneous symmetry breaking

In QED and QCD, all gauge bosons are massless, while the W and Z bosons that mediate the weak force are massive. This poses a problem: Mass terms for bosons are not gauge-invariant, so they cannot be simply added onto the Lagrangian. The solution is to take a new approach that unifies electromagnetism with the weak force.

To accomplish this, we must construct a Lagrangian that is invariant under SU(2) $\times$ U(1) transformations. This group has four generators, denoted by Y (for U(1)), T_1, T_2, and T_3 (for SU(2)). Before taking the usual approach for creating a gauge-invariant Lagrangian, however, a key experimental constraint must be taken into account: The weak force maximally violates parity. To accomplish this while preserving Lorentz invariance, the weak interaction must only interact with the left-handed chiral components of fermions. Ignoring mass for the time being, the first part of the gauge-invariant Lagrangian for the electroweak interaction is

$$\mathcal{L}_{\mathrm{EW}'} = i\bar{\psi}_L \not{D}_L \psi_L + i\bar{\psi}_R \not{D}_R \psi_R - \frac{1}{4} W^{i\mu\nu} W^i_{\mu\nu} - \frac{1}{4} B^{\mu\nu} B_{\mu\nu}, \tag{2.11}$$

where W^i and B are new gauge fields and $\not{D}_{R/L}$ are gauge-covariant derivatives that act only on right- and left-handed fermions. (The definition of $W^i_{\mu\nu}$ mirrors equation 2.9, although with the SU(3) structure constants f^{abc} replaced with SU(2)'s ϵ^{abc}.) $D\!\!\!/_{R/L}$ were previously identical, but since W couples only to left-handed fermions, $D_{L\mu}$ and $D_{R\mu}$ now differ:

$$D_{L\mu} = \partial_\mu + igW^i_\mu T^i + i\frac{g'}{2}Y_L B_\mu \tag{2.12}$$

$$D_{R\mu} = \partial_\mu + i\frac{g'}{2}Y_R B_\mu \tag{2.13}$$

Y is the weak hypercharge operator, which may take on different values for left- and right-handed components and is accordingly split into Y_L and Y_R. With some effort,[7] it can be shown that neither B nor any of W^i have the couplings that one would expect to see from the photon. It turns out that the physical fields of the electroweak theory are:

$$A_\mu = B_\mu \cos\theta_W + W^3_\mu \sin\theta_W \tag{2.14}$$

$$Z_\mu = -B_\mu \sin\theta_W + W^3_\mu \cos\theta_W \tag{2.15}$$

$$W^\pm_\mu = \frac{1}{\sqrt{2}}(W^1_\mu \pm iW^2_\mu) \tag{2.16}$$

θ_W is the weak mixing angle, which also relates the weak isospin and weak hypercharge couplings g and g' via $\sin\theta_W = \frac{g'}{\sqrt{g^2+g'^2}}$.

2.2.4 The Higgs mechanism

We return now to the problem of mass. Although $\mathcal{L}_{\text{EW}'}$ describes many key features of the electroweak interaction, it is not possible to introduce mass terms for $W^\pm$ and

Z without violating the gauge symmetry. Instead, we can introduce two new complex scalar fields as a weak isospin doublet:

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} \tag{2.17}$$

Retaining the gauge-covariant derivative D_μ from above, we can now define a gauge-invariant Lagrangian

$$\mathcal{L}_H = (D_\mu \phi)^\dagger (D^\mu \phi) - V(\phi). \tag{2.18}$$

$V(\psi)$ is the Higgs potential

$$V(\phi) = \mu^2 \phi^\dagger \phi + \lambda (\phi^\dagger \phi)^2, \tag{2.19}$$

If $\mu^2<0$, $\phi = (0,0)^T$ is an unstable local maximum, and $V(\psi)$ is minimized when $\phi^\dagger\phi = -\mu^2/2\lambda$. Consequently, the principle of least action implies that ϕ will spontaneously "fall" into these states of minimum potential. A simplified example is shown in figure 2.4.

We can arbitrarily define the chosen state as $\phi = (0, v/\sqrt{2})^T$, where v is real. It can be shown that perturbations around this state $(0, (v + H(x))/\sqrt{2})^T$ are the only states that turn out to be physical in practice, corresponding to a new scalar particle H: The Higgs boson.[8] (Perturbations around the remaining three degrees of freedom are called Goldstone bosons; they are nonphysical and and can be absorbed into longitudinal polarization components of the W and Z bosons.) Upon expanding out the first term in the Lagrangian $(D_\mu\phi)^\dagger(D^\mu\phi)$, terms containing $W^1_\mu W^{1\mu}$, $W^2_\mu W^{2\mu}$, and $Z_\mu Z^\mu$ will appear. In short, mass terms for the Z and W bosons emerge from their interactions with the Higgs field, which has taken on a nonzero value through spontaneous symmetry breaking. We

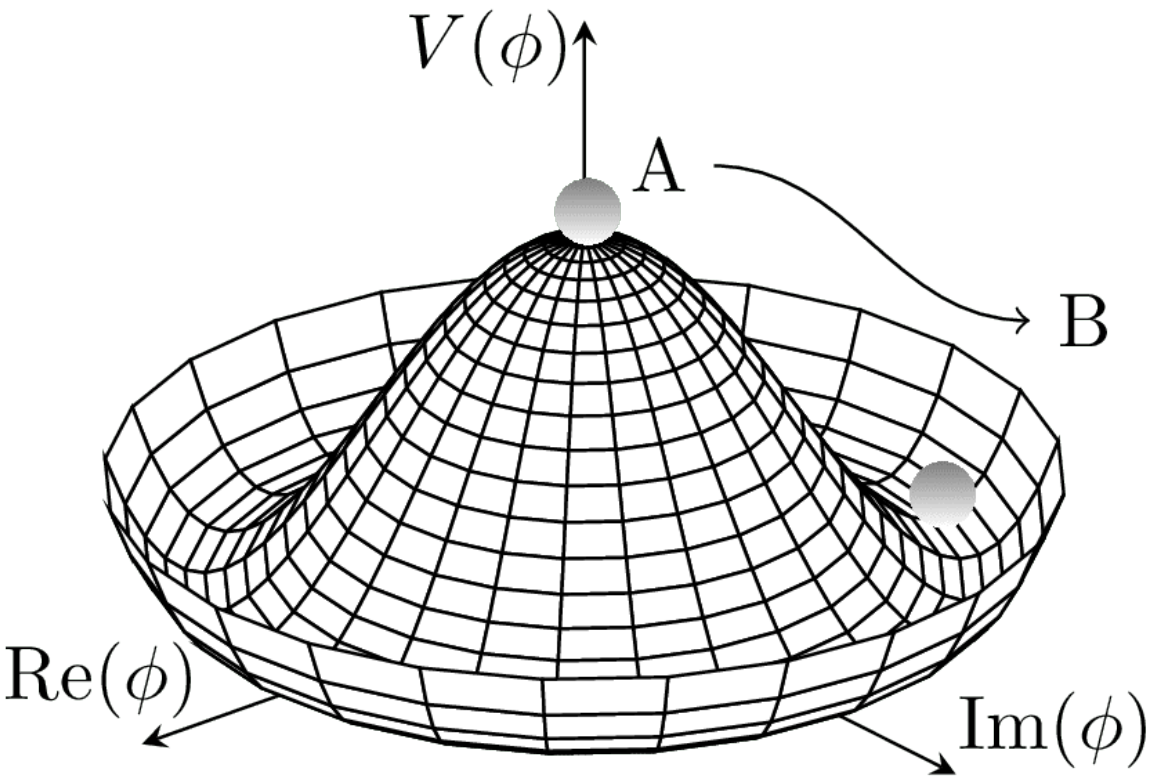


Figure 2.4: Illustrative example of a Higgs-like potential. Here, the Lagrangian contains a scalar singlet ϕ instead of a doublet, making it possible to plot both the real and imaginary components in three dimensions. (The equivalent plot for the SM Higgs potential would be five-dimensional.) The principle of least action dictates that ϕ will fall from an unstable maximum at $\phi = 0$(A) to a degenerate minimum state in the "trough" (B). From [5]

can therefore identify $m_W = \frac{1}{2}gv$, $m_Z = \frac{1}{2}v\sqrt{g^2 + g'^2}$, and $m_H = \sqrt{2\lambda v^2}$.

In a similar vein, the Higgs mechanism can also be used to generate masses for the standard model fermions. Using a Lagrangian of the form

$$\mathcal{L} = -\sqrt{2}\frac{m_f}{v}[\bar{L}\phi R + (\bar{L}\phi R)^\dagger], \tag{2.20}$$

where L and R are left- and right-handed SU(2) doublets, mass terms for the charged leptons and up-type quarks can be generated. A slightly modified version of 2.20 yields masses for down-type quarks as well. Both Lagrangians also yield interaction terms of the form $\frac{m_f}{v}\bar{u}uH$, revealing that the Yukawa coupling between the SM fermions (neutrinos excepted) and the Higgs is directly proportional to their mass. Measuring the Higgs branching fraction to various fermions is therefore a good way for experimentalists to probe the SM. (Recall Figure 2.4 from the introduction for a brief summary of the currently observed couplings and their compatibility with the SM.)

One final piece is necessary to bring the standard model together: The quark states that transform as SU(2) doublets are not quark flavor eigenstates. Instead, they are superpositions of them. Experimentally, the states can be related by the Cabibbo–Kobayashi–Maskawa (CKM) matrix

$$\begin{bmatrix} d' \\ s' \\ b' \end{bmatrix} = \begin{bmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{bmatrix} \begin{bmatrix} d \\ s \\ b \end{bmatrix}. \tag{2.21}$$

It can be shown that $|V_{ij}|^2$ is the probability of a j quark decaying into a i quark. V_{ij} are in general complex, and the presence of a complex phase in the CKM matrix leads to the observed violation of CP symmetry by the weak interaction.

2.3 Beyond the standard model

Although the standard model is both wide-reaching in scope and precise in many of its predictions, it is far from a complete theory of nature. Physicists have discovered several phenomena that the SM does not predict and cannot be easily modified to explain. Moreover, a number of theoretical issues remain: The SM has a few mathematical quirks that lack a clear explanation and seem like they could be consequences of a more fundamental theory. Research into all of these areas is ongoing, but despite receiving a tremendous amount of attention and resources over the past few decades, the way forward is still unclear.

Although the ttH(H→cc) process considered in this thesis is predicted by the standard model, it is quite sensitive to a particular type of new physics: 2 Higgs doublet models (2HDM). The following section will explain some issues with the SM that motivate these models, then finish with a quick overview of a simple 2HDM.

2.3.1 The strong CP problem

Although the Lagrangian for the SM formulation of the strong force was laid out in Equation 2.8, it is not the most general formulation of QCD. One new term can be added and one tweak can be made to the $\bar{Q}i\not{D}Q$ term while keeping the theory renormalizable and compatible with SU(3) symmetry. The most general QCD Lagrangian is

$$\mathcal{L}_{\text{QCD'}} = \bar{Q}(i\not{D} - me^{i\theta'\gamma_5})Q - \frac{1}{4}G^{a\mu\nu}G^a_{\mu\nu} + \theta\frac{g^2}{32\pi^2}G^{a\mu\nu}\tilde{G}^a_{\mu\nu} \tag{2.22}$$

Notably, both of these modifications violate CP symmetry. Things can be made slightly cleaner by (without loss of generality) redefining the mass phase of the quark fields via $\phi \to e^{i\theta'\gamma_5/2}\phi$. This leaves the physics of QCD unchanged, but conveniently moves the

entire CP-violating term into the last part of Equation 2.22 (called the θ term).

In practice, however, no CP violation has ever been observed in strong interactions. The strongest constraints come from measurements of the electric dipole moment of the neutron,[9] which impose a limit of roughly $\theta < 10^{-10}$.[10] Since θ can theoretically take on any value between 0 and 2π, it appears to be fine-tuned to have a value near 0, a coincidence that invites an explanation.

A promising solution is to introduce a new pseudoscalar particle, the axion. It can be shown that if the axion couples to QCD via the term

$$\mathcal{L} \subset \frac{a}{f_a} \frac{g^2}{32\pi^2} G^{a\mu\nu} \tilde{G}^a_{\mu\nu}, \tag{2.23}$$

where f_a is a new coupling constant, θ will spontaneously relax to 0, and the measured electric dipole moment of the neutron will be 0.[11] The simplicity of this approach is appealing, but no axions have been discovered so far; all physicists have been able to do is place weak constraints on their existence.[12]

2.3.2 Baryogenesis

Another unexpected property of the universe is the observed asymmetry between matter and antimatter. Nearly all observable matter is commposed of baryons; hardly any is antibaryonic. This suggests the existence of physical laws that favored the production of baryons over antibaryons in the early universe. In order for this to occur, three criteria known as the Sakharov conditions must be met:[13]

- The process must involve baryon number violation.
- The process must violate CP symmetry.

- The process must have occurred quickly relative to the rate of expansion of the universe (or the universe would have reached thermal equilibrium).

The last criterion is fairly easy to meet within inflationary models. However, the first two are thornier: The standard model forbids baryon number violation, and although the weak interaction does violate CP symmetry, this is not sufficient to explain the observed baryon asymmetry. A large amount of experimental and theoretical work has been invested in searching for differences between the physics of baryons and antibaryons, but the cause of baryogenesis remains a mystery.

2.3.3 Supersymmetry

Supersymmetry, or SUSY, is one of the most famous concepts in BSM physics. In short, supersymmetry proposes extending the SM by introducing a new spacetime symmetry between fermions and bosons. For every known particle, this symmetry would imply the existence of a new particle whose spin differs by 1/2. In a theory with unbroken supersymmetry, particles and their supersymmetric counterparts (superpartners) would be identical apart from their spin. In practice, however, spontaneous breaking of this symmetry allows SM particles and their superpartners to take on different masses and charges. This leads to a rich phenomenology, with supersymmetric theories predicting the existence of many yet-undiscovered particles that could be produced by modern colliders.

Notably, SUSY has the potential to solve many open problems in physics. The gauge hierarchy problem is the most direct application: Naively, one-loop corrections to the Higgs mass are expected to be extremely large (proportional to Λ^2, where Λ is the energy scale at which the SM beaks down), and must cancel with the bare Higgs mass to one part in 10^{30} to reproduce the observed Higgs mass of 125 GeV. SUSY resolves this apparent

fine-tuning problem, turning the Λ^2 dependence into a logarithmic one and bringing the one-loop correction down to a more modest value.[14] Moreover, SUSY provides a possible explanation for dark matter in the form of particles like neutralinos (mixed eigenstates of the photon, W, Z, and Higgs superpartners). In many SUSY models, neutralinos are the lightest supersymmetric particles, and thus are stable products of decaying superpartners. Their large predicted mass of O(100-1000 GeV) and weak interactions with ordinary matter make them promising dark matter candidates. Beyond this, SUSY opens up an avenue for unifying the strong force with the electroweak force at high energies, a phenomena called grand unification, as well as providing some support for more ambitious theories such as string theory.

At present, many versions of SUSY face a significant problem: They are constrained by experimental results. The minimal supersymmetric standard model (MSSM), for instance, is the simplest theory of supersymmetry that remains compatible with the SM, introducing the smallest possible number of new fields and interactions. It is currently in severe tension with observed Higgs mass of 125 GeV, as this relatively high Higgs mass is hard to explain without raising the predicted stop mass to a point where supersymmetry has trouble solving the hierarchy problem.[15] Furthermore, a large number of superpartner searches have been carried out at the LHC, ruling out many SUSY models that predict superpartners below the TeV scale.[16] Many untested possibilities remain at higher energies, but they often have to take a a more convoluted approach to avoid the existing constraints.

2.3.4 2 Higgs doublet models

Modern electroweak theory includes a single Higgs doublet. In principle, however, there is no reason why the standard model must be limited to a single Higgs. As

mentioned previously, 2 Higgs doublet models (2HDMs) are a viable alternative that, although subject to significant constraints, are still worth investigating further.

In the case of the SM, there are four scalar degrees of freedom in the Higgs sector, three corresponding to Goldstone bosons that become longitudinal modes of the W and Z and one physical degree of freedom that becomes the Higgs. In the simplest version of the SM with supersymmetry, however, the scalars are part of chiral doublets, meaning that their complex conjugates have opposite chirality. The only way for up- and down-type quarks to acquire mass simultaneously, then, is for there to be a second Higgs doublet; this is also required to cancel gauge anomalies in one-loop Feynman diagrams. Consequently, this 2HDM has eight degrees of freedom, two for each component of the doublets. This leads to a total of five physical degrees of freedom: The SM Higgs, one additional neutral scalar, one neutral pseudoscalar, and two electrically-charged scalars. However, more possibilities exist. The most fully general 2HDM has fourteen degrees of freedom instead of eight, and is much more phenomenologically rich.

In 2HDMs, not all fermions acquire their mass through coupling to the same Higgses. The first and second generations of fermions may gain their masses through a different mechanism than the third generation—for instance, the first Higgs doublet may only provide mass to the third generation fermions, while the second doublet provides mass to the remainder. In these scenarios, the Higgs coupling to first- and second-generation fermions may be unexpectedly high, leading to excesses in related processes. The charm quark is the heaviest quark in the first two generations (and therefore has a relatively large coupling to the Higgs), making processes involving the decay $H \to c\bar{c}$ a natural place to look for signatures of 2HDMs.[17][18][19][20]

Chapter 3

The LHC and the CMS Detector

The best available modern tool for probing the high energies of the electroweak scale is the Large Hadron Collider (LHC) and its four main experiments. In particular, this thesis focuses on data produced by the CMS experiment.

This chapter opens with a brief overview of the fundamentals of particle colliders in section 3.1, followed by a description of the LHC and its capabilities. The CMS experiment and its subdetectors will be explained in depth in section 3.3.

3.1 The physics of particle colliders

In particle collision experiments, a key kinematic quantity is the invariant mass (or center-of-mass energy) $\sqrt{s} = (\sum_i p_i)^2$, where p_i is the 4-momentum of the i'th particle in the system. $\sqrt{s}$ effectively sets the maximum energy scale of the physics. For instance, Higgs boson production will be heavily suppressed if $\sqrt{s}$ drops too far below the electroweak scale of 200 GeV; the production rate rises as $\sqrt{s}$ increases.

Since the outcome of a given process in QFT is nondeterministic (or at least unpredictable), the only way to gather information on rare processes is to collect data

from a massive number of events. It is therefore useful to define a quantity tied to the probability of producing a given process independently of the event rate, as well as a process-independent quantity tied to the rate at which events are produced at an experiment. The former is known as the interaction cross-section σ, which has units of area; it is a quantum mechanical quantity that is unique to each process. The process for deriving the cross-section of a process from the Lagrangian of the underlying quantum field theory is lengthy, but well-established; see [21] for a rigorous treatment. The latter is called the instantaneous luminosity $\mathfrak{L}$. It has units of inverse area per second and depends on the event rate, $\sqrt{s}$, and beam characteristics of the experiment in question. Given a rate of event production dN/dt, the two are related as follows:

$$\frac{dN}{dt} = \sigma\mathfrak{L}(t) \tag{3.1}$$

In physics analyses working with a fixed amount of data, it is often more useful to use the integrated luminosity $\mathfrak{L}$, which gives the expected event yield for a process when multiplied by the process' cross-section:

$$N = \sigma \int \mathfrak{L}(t)dt \tag{3.2}$$

By convention, this is simply referred to as the luminosity. Many key processes in the standard model (most notably, Higgs production) have extremely small cross-sections, often on the order of picobarns (10^{-40} m^{-2}) or less.[22] Studying these processes requires extremely high-luminosity particle colliders, with $\sqrt{s}$ far above the electroweak scale and high collision rates.

Modern particle accelerator design is limited by several factors. Firstly, the critical kinematic role of $\sqrt{s}$ limits the reach of fixed-target experiments: A beam of energy E impacting a fixed target has one-fourth the center-of-mass energy as a head-on collision

between two identical beams. Additionally, our ability to accelerate particles quickly is limited. Charged particles can be accelerated with strong electric fields in radiofrequency (RF) cavities; however, extremely high voltage gradients can lead to RF breakdown and cause damage. Consequently, modern RF cavities for electron and positron accelerators can only reach gradients of around tens of MeV/m, while proton and ion accelerators are limited to a few MeV/m.[23] A simple way to get around this limit is to pass a particle beam through the same RF cavities multiple times by bending them into a circular path with magnetic fields. Unfortunately, this introduces two additional problems. First, higher beam energies require either stronger superconducting magnets or higher quantities of them, which can impose significant engineering and financial constraints. Second, accelerating electric charges in a circle causes them to emit synchrotron radiation, the intensity of which is roughly proportional to $1/m^4$ for a fixed beam energy. Since protons are roughly 2,000 times more massive than electrons, this means that massive amounts of synchrotron radiation make circular electron colliders impractical at the TeV scale given current technology, so the highest-energy modern accelerators must be circular proton colliders. Protons do bring some additional downsides—as composite particles, they produce larger and messier backgrounds, and the nature of partons makes it impossible to precisely know the CM energy of each collision—but although this gives linear colliders a greater ceiling for precise measurements, the higher achievable beam energies of circular colliders have paid large dividends in terms of access to new physics.

3.2 The Large Hadron Collider

The Large Hadron Collider (LHC) is the highest-energy particle accelerator ever built. A 10-kilometer-wide circular accelerator, the LHC consists of two counterrotating beams with a maximum energy of 6.8 TeV each, for a total $\sqrt{s}$ of 13.6 TeV. Located on the border

of Switzerland and France, the LHC was constructed between 1998 and 2008 by the European Organization for Nuclear Research (CERN), and has been in regular operation since its first data-taking run in 2010. Inside its 27-km-long tunnel, the LHC's beam is guided by 1232 superconducting dipole magnets, focused by 392 quadruple magnets, and "kicked" up to higher energies by 16 RF cavity generators. Each beam is comprised of 2808 bunches of particles; a bunch structure is required because of the pulsed nature of RF cavities. Proton beams have nearly 100 billion protons per bunch, although the LHC is also capable of colliding heavy ions such as lead nuclei. Prior to entering the LHC, the beams are first boosted up to 450 GeV by a linear accelerator and a series of booster rings. See Figure 3.1 for details. After the beams have been injected, the magnet current and beam energy are then gradually ramped up to full strength over the course of 25 minutes, at which point the LHC is ready to begin making collisions.

At the LHC's four crossing points, both beams are temporarily focused and overlapped to produce collisions. Each crossing point has a dedicated detector built around it, each with a different design and goal. At LHC Points 1 and 5, ATLAS and CMS are the two largest experiments at the LHC; both are general-purpose detectors designed to gather data for a wide variety of physics analyses. The remaining two experiments are more specialized: ALICE, at Point 2, is focused on studying quark-gluon plasmas produced by heavy ion collisions, while LHCb's goal is to better understand the physics of b-hadrons and CP violation.

3.3 The CMS detector

The analysis discussed in this thesis is conducted using data from the Compact Muon Solenoid experiment, or CMS. At four stories high and weighing in at fourteen thousand tons, the CMS detector is the second-largest at the LHC after ATLAS. It is a general-

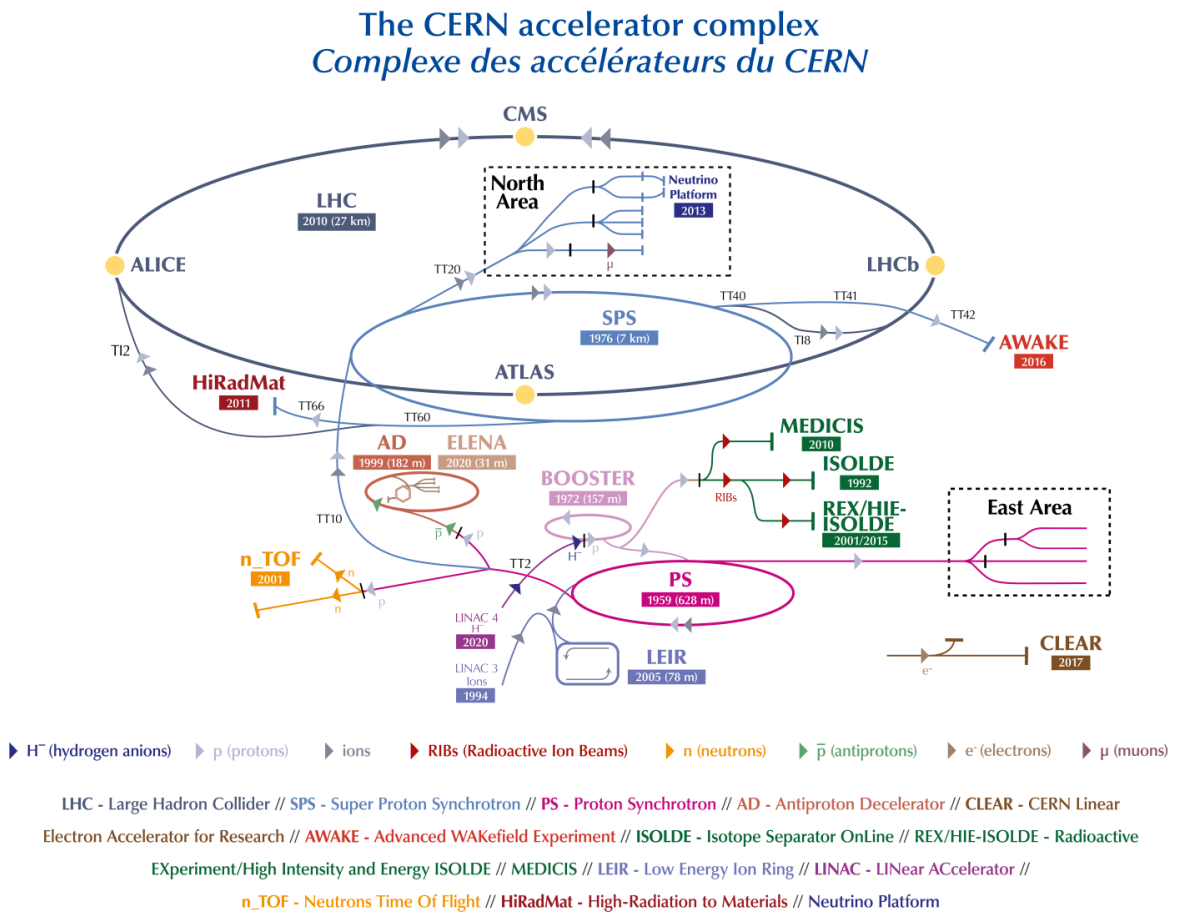


Figure 3.1: A schematic of the CERN accelerator complex. The light gray arrows indicate the path that protons take. Protons are initially accelerated by LINAC 4 and are ramped up to higher energies at three booster rings (BOOSTER, the Proton Synchrotron (PS), and the Super Proton Synchrotron (SPS)) before entering the LHC. Many are eventually funneled into the LHC, but some are used for other CERN experiments and test facilities. From [24].

purpose detector with the goal of advancing the knowledge frontier of high energy physics. To ensure the maximum reach and precision of its physics program, CMS is designed to identify and measure nearly every particle produced in proton-proton collisions produced by the LHC. Moreover, CMS is tailored to have especially good sensitivity to the main channels in which the Higgs boson was expected to be observed, $H \to ZZ^* \to 4\ell$ and $H \to \gamma\gamma$. This led to three key subgoals:[25]

1. To have a powerful and redundant muon detection system. Muons are unlikely to be produced in background collisions due to their high mass and status as second-generation leptons, so they can be useful to filter for events that involve new physics.

2. To have a high-resolution calorimeter for detecting photons and electrons. This has direct applications to both Higgs discovery channels.

3. To have high tracking resolution. This is required to enable 1 and 2.[25][26]

The CMS detector has successfully met these goals in practice, as evidenced by its and ATLAS' joint discovery of the Higgs in 2012. The only known particles that CMS cannot detect with good resolution are neutrinos; essentially everything else beyond ten degrees of the beamline is covered. (Setting aside the difficulty of placing equipment so close to the beam, the slightly-imperfect coverage is not an issue in practice, since particles deflected at small angles usually involve processes with low momentum transfer and thus uninteresting physics.)

To attain this degree of coverage, CMS consists of multiple concentric subdetectors, each tailored for a specific purpose. See Figure 3.5 for an illustration. Since its first data-taking beam fill in 2010, CMS has collected nearly 250 $\text{fb}^{-}1$ of data over the course of three runs. Due to the low luminosity of Run 1 (2010-2012) data and the current lack of familiarity with the Run 3 dataset (2022-present), the analysis described in this thesis

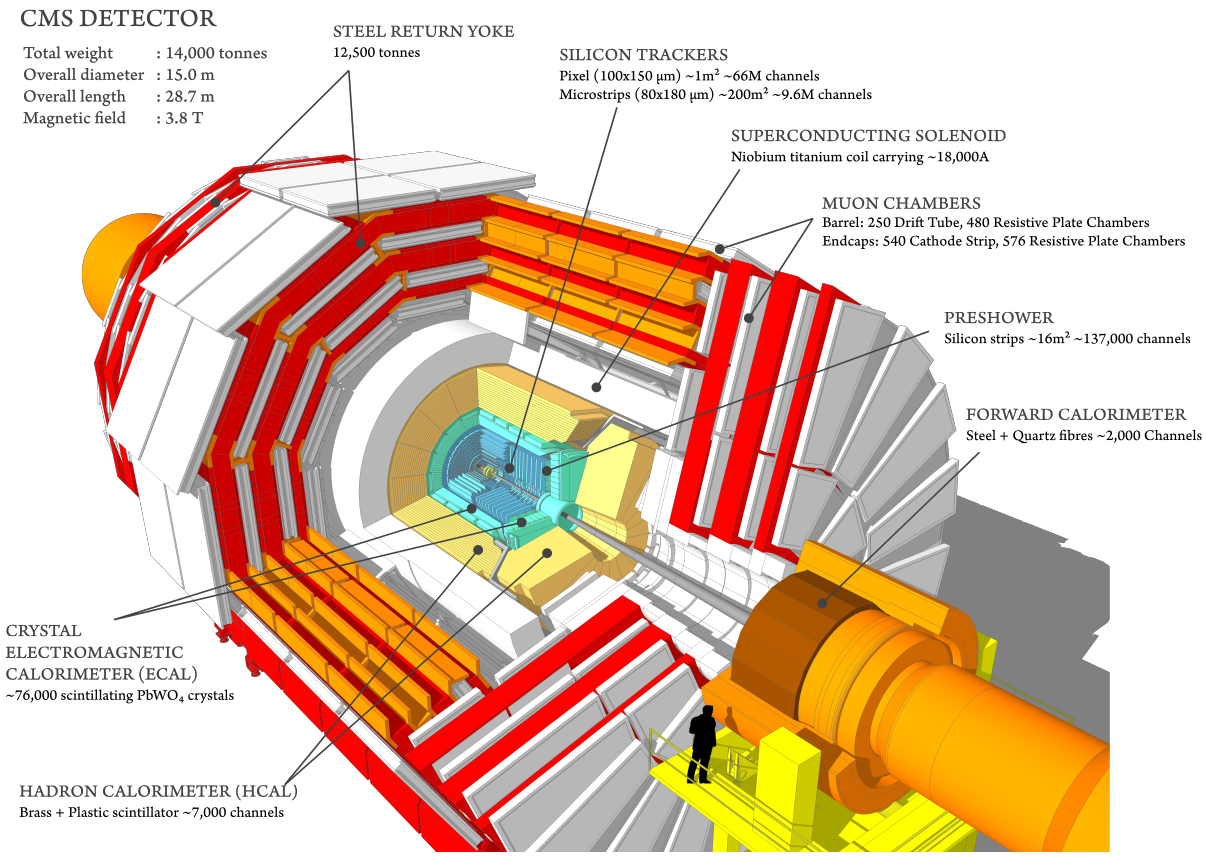


Figure 3.2: Cutaway illustration of the CMS detector. Each of the subdetectors are labeled. From [28]

focuses primarily on Run 2 data (2015-2018). The luminosity of each run is illustrated in Figure 3.4.

An overview of the CMS subsystems follows, drawing primarily from the CMS Technical Design Report.[27]

3.3.1 The CMS coordinate system

The origin of the CMS coordinate system is simply taken to be the interaction point at the exact middle of the detector. The z-axis lies along the beamline (pointing west toward the Jura mountains), the y-axis points upwards, and the x-axis points radially inwards toward the center of the LHC. In polar coordinates, the z-axis is unchanged; the

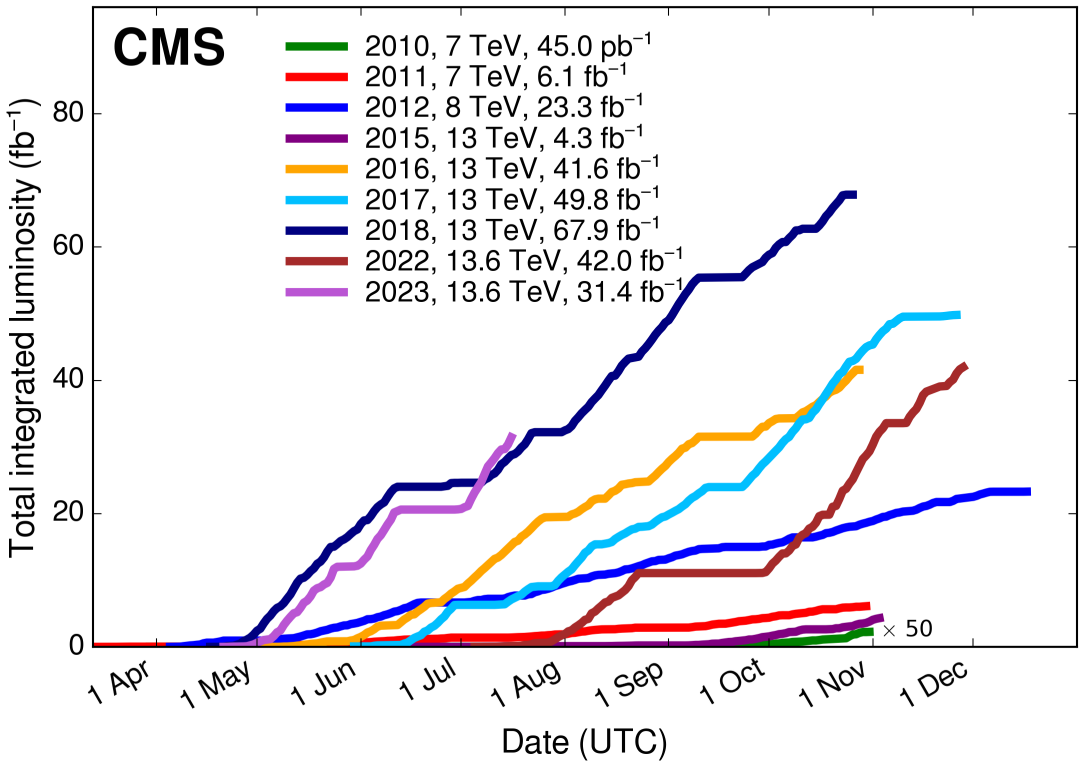


Figure 3.3: Plot of the integrated luminosity per year as a function of time. The early runs in 2010 took very little data, and are magnfied by a factor of 50 in this plot for visibility. The curves gradually get steeper over time as a result of upgrades to the LHC that have increased its collision rate and luminosity. From [29].

polar angle θ is measured from the z-axis, while the azimuthal angle ϕ is measured from the x-axis along the $x - y$ plane.

However, θ is inconvenient to use in practice. Protons are composite particles made of quarks and gluons, and the fraction of the total momentum carried by each parton—speaking roughly, each component of the proton—is not equal. As a result, the center of mass frame of the collision is not necessarily the rest frame of CMS, and it will often be boosted along the z-axis. It is therefore worth defining a θ-like coordinate that easily transforms under boosts. We can define the rapidity of a particle:

$$y = \frac{1}{2} \ln \left(\frac{E + p_z}{E - p_z} \right), \tag{3.3}$$

Here E and p_z are the measured energy and momentum along the z-axis, respectively. The advantage of y is that differences in rapidity are identical in all frames, which can be seen by applying a Lorentz transformation to y:

$$y' = y + \frac{1}{2} \ln \left(\frac{1 - \beta}{1 + \beta} \right) \tag{3.4}$$

When considering the difference in rapidity between two arbitrary events $\Delta y = y_2 - y_1$, the ln terms will cancel out, yielding the same interval in all frames. Another advantage of this choice of coordinate is that particle flux remains roughly constant as η changes.[8]

Since virtually all particles observed at the LHC are in the ultra-relativistic regime, we can make the high-energy approximation $p_z \approx E \cos\theta$, producing a quantity called the pseudorapidity:

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

η is exclusively a function of θ (which is measured by the tracker) and is much easier to

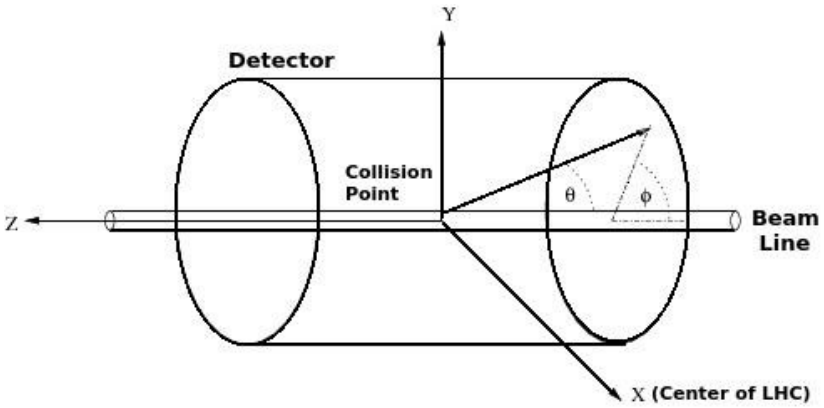


Figure 3.4: Schematic of the (right-handed) CMS coordinate system. From [30].

compute than y as a result.

The last relevant quantity at CMS is the transverse component of momentum $p_T \approx E\cos(\theta)$. Although the total p_z of a collision at the LHC is usually nonzero, the total transverse momentum is almost exactly zero. A nonzero measured p_T indicates that some products of the collision may not have been detected—likely neutrinos, but possibly something more exotic like dark matter.

3.3.2 The solenoid magnet

The most noteworthy feature of CMS, and the origin of its name, is the superconducting solenoid magnet that it is built around. Because charged particles curve in magnetic fields, with the radius of curvature and direction of the path determined by the particle's charge and momentum, pairing a powerful magnetic field with high-resolution tracking software is a powerful way to collect identifying information.

The CMS design specifications call for high momentum resolution for muons. The key benchmark is high-energy muons (E 1 TeV), which curve only slightly in a magnetic field; a field strong enough to bend muons by $\Delta p/p \approx 10\%$ at 1 TeV is necessary to reliably measure their charge. To meet this goal, engineers designed a six-meter-diameter, 13-

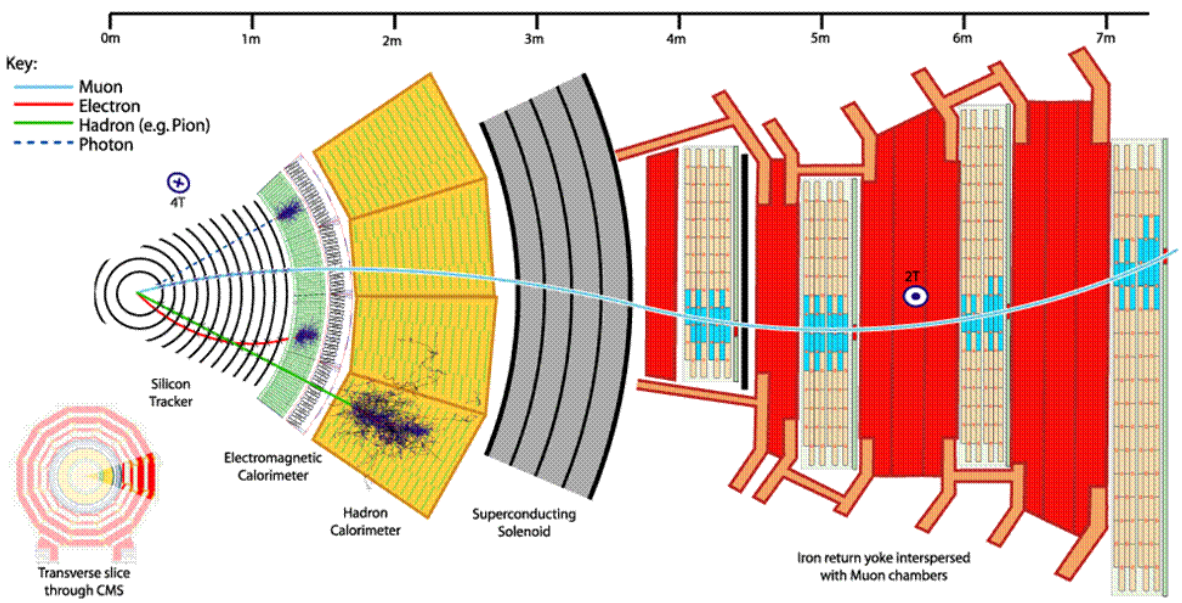


Figure 3.5: From [31].

meter-long superconducting solenoid that produces an axial magnetic field of 3.8 Tesla inside of it. The niobium-titanium coil magnet is kept at at a temperature of 4.7K to ensure that it stays superconducting whenever current is flowing through it. Additionally, the return magnetic field outside the solenoid is condensed and directed by a large iron return yoke that surrounds it. Because of the return yoke, the magnetic field in the muon system is as high as 2 Tesla (pointing in the opposite direction of the main magnetic field). Figure 3.5 gives an example of how the magnetic field affects the trajectory of a muon traveling through the CMS barrel.

Since the solenoid occupies the space between the muon chambers and the inner layers of the detector, it is a major constraint for both the original design and future upgrades: All inner parts of the detector must fit inside it.

3.3.3 The Inner Tracking System

The CMS tracker is the innermost subdetector. Immersed in the solenoid's 3.8 T magnetic field, the inner tracker's main job is to precisely measure the paths of charged particles produced in collisions, making it possible to reconstruct their momenta with high accuracy later. Additionally, the tracker is responsible for identifying both primary vertices (points where a hard collision occurs) and secondary vertices (produced when a heavy particle such as a tau travels a short distance from the primary vertex before decaying).

Both the pixel and strip tracker operate based on the same principles. When high-energy particles pass through a semiconductor, they leave a large number of electron-hole pairs in their wake. By setting up a high electric potential across a thin (O(100 μm)) silicon sensor, these currents can be collected, and when paired with a sufficiently powerful amplifier, produce a detectable pulse. On top of the need for extremely high-resolution sensors, especially in the innermost layers of the tracker, an additional requirement is for the subdetector to be radiation hard. Due to its close proximity to the beam, the inner tracker experiences unprecedented radiation doses over the lifetime of the experiment (see Figure 3.6), which degrade the electrical properties of silicon over time. Recent advances in silicon detectors have made it possible to build electronics that can withstand these doses, and by keeping the tracker cooled at -20 C the effects of radiation damage can be minimized.

The tracker is comprised of two distinct sections, the pixel tracker and the strip tracker. The pixel tracker has an inner radius of 3 cm and an outer radius of 16 cm, bringing it extremely close to the interaction point. It consists of four cylindrical layers of high-resolution silicon modules, plus six partially-overlapping discs at each end. Each module contains over 66 thousand 100×150 μm pixels for a total of 124 million readout

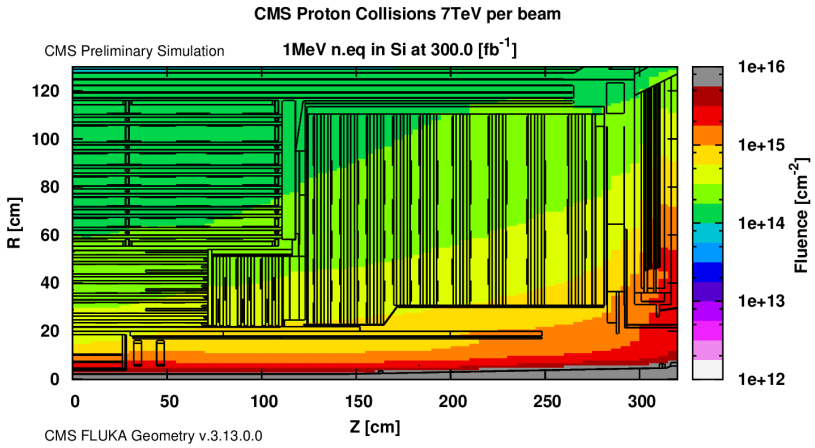


Figure 3.6: Simulation of the total fluence expected in the tracker by the end of Run 3. The fluence is expressed in terms of the number of 1 MeV neutrons per square centimeter it would take to cause an equivalent amount of damage. From [34].

channels.[32] The strip tracker is much larger, extending from a barrel radius of 20 cm to 116 cm, and is comprised of silicon microstrip detectors instead of pixels. The barrel section that surrounds the pixel tracker has ten concentric layers, while the endcaps consist of twelve discs of sensors. Although the greater distance from the beampipe lessens the resolution requirements for the strip tracker, it still has a total of 9.3 million detector channels.[33] As a whole, the tracker has comprehensive coverage up to $|\eta| < 3.0$, as well as the < 25 ns temporal resolution necessary to handle the event rates of the LHC. Furthermore, to avoid interfering with the outer detectors, the tracker accomplishes all this with less than two radiation lengths' worth of material (and less than one radiation length for $|\eta| < 1$).

3.3.4 The Electromagnetic Calorimeter

As discussed previously, CMS's electromagnetic calorimeter (ECAL) is tasked with measuring electron and photon energies with high precision. It successfully meets its design goals of high photon and lepton energy resolution by making use of scintillating

lead tungstate (PbWO4) crystals. When a high-energy photon from the interaction point (or photons produced by electron bremmsstrahlung) travels through the crystal, electron-positron pairs will be created via pair production. These pairs will continue forming a shower in the crystal until they eventually reach an energy low enough to excite atoms in the scintillating crystal, which will then emit light upon returning to their ground state. These small pulses of light are then picked up by sensitive photodiodes, and the amount of light emitted is proportional to the energy deposited in the ECAL. Beyond their scintillating properties, PbWO4 crystals were chosen for two key reasons: They produce light very quickly after absorbing a particle shower (on the order of the 25 ns spacing between events at the LHC), and they have a short radiation length of 0.89 cm due to their density (leading to a compact and high-resolution calorimeter).

The ECAL consists of two sections, a barrel region and two endcaps. Each is made of a single layer of PbWO4 crystals and photodiodes. The 76,000 or so crystals are slightly over 20 cm long and cover an area of roughly 2.5 $\times$ 2.5 cm each. (The barrel and endcap use slightly different crystal dimensions.) Additionally, the endcaps contain an extra preshower layer made of alternating lead absorber layers and silicon sensor strips; this is primarily to better distinguish lone photons from boosted pions that decay to a pair of overlapping photons. Overall, the ECAL is highly effective: It has an energy resolution of around 2% in the barrel and up to 5% in the endcaps, as well as good η coverage of $|\eta| < 2.6$. (A small crack in the calorimeter with poor resolution exists between the barrel and the endcaps, but this does not degrade performance significantly.)[35]

3.3.5 The Hadronic Calorimeter

Because the ECAL is not dense enough to absorb most neutral and charged hadrons, it is surrounded by a hadronic calorimeter (HCAL) designed to capture them. Like

the ECAL, the HCAL uses scintillators to detect high-energy particles, but it takes a sampling calorimeter approach: Layers of brass absorber are interleaved with layers of plastic scintillator, and the showers created by high-energy particles passing through the absorber are detected by the scintillators. At roughly 23 radiation lengths deep, the HCAL absorbs most of the remaining particles produced by collisions, so that only muons, undetectable neutrinos, and a small number of high-energy hadron showers reach the solenoid magnet.

The HCAL is divided into four sections. The HCAL barrel (HB) occupies nearly all remaining space between the ECAL and the solenoid, from a radius of 1.77 m to 2.95 m. Two endcaps (HE) enclose the cylinder on either side, fully covering the interaction point up to $|\eta|3$. Since the HB is relatively thin near the middle compared to the endcaps (5.82 interaction lengths at $\eta = 0$, compared to 10 for the HE), some high-energy jets manage to slip past the HCAL and the solenoid, leading to the inclusion of a "tail catcher" outer HCAL section (HO) consisting of two scintillator layers and an iron plate outside the solenoid. Finally, an extremely radiation-hard forward calorimeter provides additional coverage up to $|\eta| < 5$. Compared to the ECAL, the HCAL features less energy resolution, ranging from 5-10% for 300 GeV jets to 50% for 20 GeV jets in the barrel, but due to the more complex nature of jets and the lower relevance to Higgs physics, this is both difficult to avoid and not a major obstacle to physics analyses.[36]

3.3.6 The Muon System

The outermost subdetector is the CMS muon system. Largely integrated into the return yoke, the muon system attains the high level of muon momentum resolution described in the introduction. Like the ECAL, the muon system is divided into barrel and endcap regions; unlike the ECAL, different technologies are used for each. The

barrel has a relatively low, fairly constant magnetic field and a smaller muon flux, so lower-resolution drift tubes (DT) are acceptable. DTs are gas-filled chambers with a wire cathode passing through the middle. As a muon passes through the gas, the electric field triggers an avalanche of electrons which is picked up by the cathode. The amount of time the electrons take to reach the cathode is known based on the physics of the gas, and when multiple perpendicular layers of drift tubes are overlapped, this allows for precise determination of the muon's position. For $1.2 \leq \eta \leq 2.5$, the higher particle flux and more complicated magnetic field necessitate the use of higher-resolution cathode strip chambers (CSCs). CSCs are composed of a series of wire anodes and perpendicular copper cathodes suspended in a chamber of gas. When high-energy muons ionize atoms in the gas, the electrons are collected by the cathode strips and the ions are collected by the anode, producing two separate signals that allow for precise determination of the muon's position. Finally, a set of resistive plate chambers (RPCs) is installed between the solenoid and the DTs. While lacking in spacial resolution compared to the DTs, the RPCs have the advantage of high timing resolution, and are used to match muons with the bunch crossing that produced them.

The full muon system must be paired with the inner tracker to achieve its full performance. While worse than the tracker at handling low-energy muons, it significantly augments the tracker's capabilities at energies above 500 GeV, as shown in Figure 3.7.

(CMS is in the process of adding an additional set of gas electron multipliers (GEMs) between the endcaps and the HCAL to improve reconstruction and trigger efficiencies. However, the first GEMs were only installed after Run 2, so they do not affect the data in this analysis.[37])

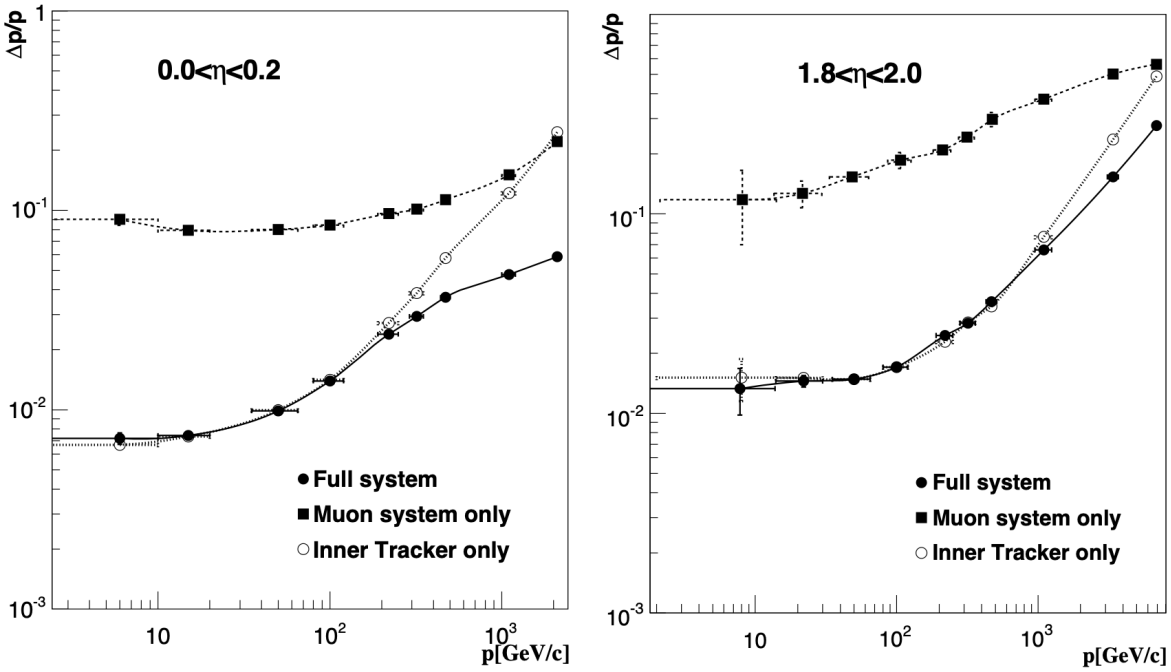


Figure 3.7: Sample plots of CMS's muon momentum resolution using the inner tracker, the muon systems, and both subdetectors combined. Since the resolution varies with η, two narrow pseudorapidity ranges are considered, one near the middle of the barrel with $0 < \eta < 0.2$ (left) and one inside the endcaps with $1.8 < \eta < 2.0$ (right). At high energy, the momentum resolution is much better than it would be if the two subsystems reconstructed the momentum independently of each other. From [37].

3.3.7 The Trigger System

During normal LHC operation, bunch crossings at the CMS interaction point occur at a rate of approximately 40 MHz. However, event data can only be written to long-term storage at a rate of roughly 1 kHz. CMS therefore requires a high-rate trigger system capable of rejecting almost 10^5 events for every accepted event.

The trigger system is divided into two stages. The first, the Level 1 (L1) trigger, must achieve a rejection factor of 10^3 while spending no longer than 1μs to reach a decision on each event. To achieve this speed, the L1 algorithm is implemented using custom-built hardware. Drawing on data from the calorimeters and muon systems, the L1 trigger checks for the presence of muons, jets, photons, and other objects that exceed certain p_T and E thresholds; the missing transverse energy E_T^{miss} is also considered. Event data is stored in designated buffers while the calculation is performed and is immediately dumped if the event fails to pass the trigger.

The high-level trigger (HLT) reduces an event rate of nearly 100 kHz down to the required 100 Hz. Unlike the L1 trigger, the HLT is implemented by a designated processor farm, and can be quickly reconfigured if better algorithms and computing methods are developed. The HLT performs a partial reconstruction of each event, considering only the minimum amount of data necessary to make a decision. To maximize speed, the HLT is broken up into a series of steps—calorimeter and muon data, followed by tracker data, followed by full track reconstruction—each of which will immediately discard the event if it fails a requirement.[27]

Chapter 4

Simulation and Reconstruction

Over the course of the last several decades, modern particle physics experiments have grown in both ambition and complexity. CMS is no exception: In terms of the size of its collaboration, the complexity of the detector, and the sheer quantity of data produced, CMS is one of the most complicated experiments ever conducted. As a consequence, predicting what the detector will see from the perspective of read-out and interpreting the recorded data are highly non-trivial. Major advances in computing technology in recent years have proven to be critical for managing both of these challenges. Most notably, CMS uses cutting-edge software to both generate simulated events and translate raw detector data into physical objects.

This chapter contains a brief overview of software methods used to simulate and reconstruct data from CMS. Section 4.1 describes the CMS simulation framework, from event generation to the simulated detector response. Section 4.2 describes the process of reconstructing CMS data into higher-level objects via the Particle Flow algorithm.

4.1 Monte Carlo Simulations

In principle, one could calculate the probability distribution function for all possible observations at CMS using only the Lagrangian of the standard model. However, since this is computationally impossible, a Monte Carlo (MC) strategy is used instead. Rather than compute loop-level corrections and other adjustments for every event, which is currently infeasible (especially for QCD processes), the CMS software framework CMSSW[38] employs random sampling of known probability distribution functions to generate a large number of events.

Event generation with CMSSW involves two steps. In the first, simulated particle 4-momenta from the collision point are created using dedicated software libraries called generators. Generators account for a wide variety of phenomena across several steps, from the initial hard collision itself to initial- and final-state processes to hadronization and jet formation. In practice, several different approaches to event generation that yield events with subtly different kinematics are available. For instance, one scheme may treat up, down, and charm quarks as having the same flavor, while another treats charms individually. Often, picking one is impossible: Different schemes may do a better job simulating different aspects of the physics. Since these differences are sometimes relevant to the overall performance of an analysis, many analyses must ensure that their search strategy is versatile enough to handle MC events produced via several different event generation methods.

In the second step, the generated particles are propagated through the CMS detector. This step takes into account not only the gradual energy loss as high-energy particles travel through the material of the detector, but also more complicated effects such as showering and secondary vertices produced by nuclear interactions. CMSSW then translates the energy loss in each part of the detector into "hits", applying an additional layer

called digitization that simulates the electronics' response to the deposited energy. The final results are largely identical to real data samples produced by CMS, although with one key exception: Generator-level information about the particles produced in each event is available, while in data, the only information that one can use is the electronics' response. This lets experimentalists perform detailed checks for each type of event to ensure that their analysis techniques work properly on MC samples.

Because of the high complexity and approximate nature of the simulations, MC samples must be extensively validated against data. Public samples made available for the CMS collaboration are checked in a variety of ways, but this is not sufficient: Every analysis considers only a small subset of MC events, so differences may slip through the cracks, and MC simulations are known to be imperfect in a variety of ways. Consequently, every CMS analysis must take care to independently validate their MC samples in their region of consideration.

4.2 Object Reconstruction

After raw detector output has been produced—either via simulation or real data collection—it must be reconstructed prior to being used in an analysis. In practice, most CMS analyses begin at the level of physics objects, combinations of particle species information and 4-momentum. In theory, the physics objects contain all possible information that can be extracted from the collision point, although complications like CMS's inability to detect neutrinos and the challenges of hadron identification mean that some additional processing may be required. Reconstruction, the process of arriving at these physics objects from raw detector output, is a crucial process that depends not only on the comprehensiveness of the detector but also on the quality of the reconstruction algorithms.

In older experiments, reconstruction was often performed at the subdetector level: Muon objects were formed using only data from the muon systems, jets were formed using only data from the calorimeters, and so on. CMS, however, uses a more holistic approach that begins with individual subdetectors but eventually makes use of all detector layers. The specialized algorithm developed for this purpose is called Particle Flow (PF). PF first saw use at the ALEPH experiment at the Large Electron-Positron Collider (the accelerator that previously occupied the LHC's circular tunnel), which dealt with the easier environment of electron-positron collisions.[39] Adapting the algorithm to the much messier case of proton-proton collisions required a great deal of refinement, but because of the excellent performance and granularity of CMS, PF has been very successful as a framework for reconstruction.

Broadly, PF is a two-step process. In the first step, low-level objects—PF elements—are reconstructed from raw detector output. In the second, particle species, momentum, and energy information are reconstructed using the PF elements, yielding the final PF objects.

4.2.1 PF Element Reconstruction

PF elements can be divided into three categories: Tracks, vertices, and calorimeter clusters. The first two are computed together; calorimeter clusters are created independently.

At CMS, charged particle tracks are reconstructed using the inner tracker and the muon system. This relies on the Kalman filtering algorithm, which takes an iterative approach to track-finding: At each step, the easiest-to-identify tracks are found and removed from further consideration. In each iteration, "seed" hits are initially chosen in the pixel tracker, and a curve is fit to them to begin the track. Hits from the strips

tracker are considered next and added to the track if the fit is sufficiently good; the fit result is adjusted each time. At the end of the iteration, all hits corresponding to several best-fit tracks are excluded, and the algorithm repeats a set number of times. The HLT, which uses a similar approach for its partial reconstruction, only goes through two iterations; offline reconstruction performs a maximum of six. After tracks have been identified, primary and secondary vertices can be found by calculating impact parameters for each track and tracing them back to common origins. (Sometimes, "kinked" tracks are produced by nuclear interactions with the tracker. These are handled separately.) Thanks to the tracker's high resolution, vertices can be identified with a precision on the order of 10 μm in the three spacial dimensions.[40] Tracks in the muon system can be identified in a similar fashion, although the fitting algorithm is modified to take the changing magnetic field into account.

The basic PF elements for the ECAL and HCAL are calorimeter clusters. CMS uses a custom algorithm for assembling clusters: Calorimeter hits that are both local maxima and above a certain energy threshold are designated as seeds, and neighboring hits above a noise threshold are added repeatedly to yield topological clusters. Next, each topological cluster is assumed to be the result of N separate Gaussian energy deposits, where N is the number of seeds. A best-fit result for the N Gaussians is obtained via an iterative algorithm, and the final Gaussian parameters are used as the parameters of the PF elements. The results are then calibrated to produce better estimates of the deposited energy; this is critical for CMS's ability to distinguish between overlapping charged and neutral particles.

4.2.2 PF Object Reconstruction

One all PF elements have been created, the Particle Flow algorithm begins its second step: The linking algorithm. Linking is the process of connecting elements from different subdetectors together to reconstruct the final particle candidates. The algorithm works on a pair-wise basis, comparing two elements together to determine whether they could be produced by the same particle. (To avoid performing n^2 comparisons, which would be computationally costly, the linking algorithm only considers elements with a low separation in η and ϕ.)

Like the procedure for PF element reconstruction, the linking algorithm searches for the easiest-to-find candidates first, then excludes them from further consideration and moves on. The first reconstructed particle candidates are muons, which have a distinct signature of tracks in the tracker and muon system and nothing in the calorimeters. Muons that are isolated from other particles are removed first; a second step is then performed to catch non-isolated muons that overlap with jets.

Next, electrons and isolated photons are handled simultaneously. Electrons usually emit bremsstrahlung while passing through the tracker, so photon and electron reconstruction involve similar considerations. The key difference is the presence of a track in the inner tracker linked to the ECAL cluster. After various checks to ensure that the candidate is isolated from other elements in the event, electron candidates are additionally examined with boosted decision trees (BDTs) and must pass a number of other electron identification criteria to avoid misidentification. For photon candidates, the key verification mechanism is comparing the ratio of energy deposited in the ECAL to energy deposited in the HCAL: Photons will deposit most of their energy in the ECAL, while the reverse is true for neutral hadrons. As with muons, all PF elements linked to electrons and photons are masked from later steps by default.

Third, hadrons and non-isolated photons (e.g. near-overlapping photons from π^0 decay) are reconstructed together. As usual, photons are primarily distinguished by the absence of tracks in the inner tracker, while hadrons deposit a conspicuous amount of energy in the HCAL. If the candidate set of elements also passes a variety of other criteria, the particle species is then determined by comparing the energy deposited in the calorimeters against the momentum as estimated by the tracker. An excess of energy is a sign of extra photons or neutral hadrons, a deficit prompts a muon search with loosened identification criteria, and rough compatibility implies a charged hadron.

Finally, a handful of special cases are handled separately. Hadrons will often interact with nuclei in the tracker, producing an average of one extra secondary vertex per $\mathrm{t\bar{t}}$ event.[41] These are identified with a dedicated algorithm. A subsequent post-processing step handles any remaining ambiguities; for example, an unphysically high p_T^{miss} suggests the presence of a missed muon.

Chapter 5

Analysis Techniques

The fundamental goal of modern analysis methods in particle physics is too compare the theoretical predictions of the standard model with experimental data. As touched on in the previous chapter, one of the major challenges is that there are several degrees of separation between data taken by modern detectors and the underlying physics that produced it: Reconstruction involves several steps of assembling progressively higher-order objects from detector-level information, and none of them are perfectly accurate. Another challenge is theoretical uncertainties caused by the difficulty of e.g. describing lower-energy interactions in QCD. Moreover, final states involving QCD are heavily complicated by jets, the condensed sprays of hadrons that quarks produce. To handle the uncertainties and complicated event topologies, most CMS analyses make use of a broad library of techniques; this thesis is no exception.

This chapter summarizes several statistical and analytic methods used in the analysis described by this thesis. Section 5.1 provides an overview of maximum likelihood estimation, the main statistical technique used for data analysis here. Section 5.2 goes into detail on jet tagging. Finally, section 5.3 introduces a method for event classification using transformer neural networks.

5.1 Statistical methods

Modern physics analyses need powerful statistical methods that can handle large quantities of multidimensional data and a large number of associated uncertainties. Furthermore, since data-MC differences are hard to avoid yet impactful for high-precision measurements, a truly reliable approach must be capable of adjusting MC-based predictions to account for real data—without looking at any signal data.

5.1.1 Likelihood Functions

To begin, let us introduce the concept of a likelihood function. The likelihood function is the probability of observing the data in an experiment as a function of some (often unknown) parameters describing the experiment. For A-dimensional data x_a and B parameters θ_b, it can be written as

$$L = f(x_1, x_2, ..., x_A; \theta_1, \theta_2, ..., \theta_B). \tag{5.1}$$

The parameters θ_b are very general, and can include theoretical parameters like coupling constants, experimental effects like reconstruction uncertainties, and more. Parameters with unknown values that are nonetheless required to estimate the parameters of interest are called nuisance parameters. Note that f is equivalent the probability density function of x_a if θ_b are assumed to be fixed instead of x_a; the opposite is true for the likelihood function.

If a large number of data points $\vec{x}^i$ are collected instead of a single measurement, the total likelihood function is simply the product of the likelihood functions for each individual measurement. Assuming a total of N events, and using vectors $\vec{x}^i$ and $\vec{\theta}$ as shorthand for $x_1^i, x_2^i, ...$ and $\theta_1, \theta_2, ...$, the likelihood function is

$$L(\vec{x}^1, \vec{x}^2, ..., \vec{x}^N; \vec{\theta}) = \prod_{i=1}^{N} f(\vec{x}^i; \vec{\theta}). \tag{5.2}$$

In practice, the number of measurements N is often random as well, as is the case when recording events at CMS. Here, N obeys a Poisson distribution:

$$\text{Pois}(N, \lambda) = \frac{\lambda^N e^{-\lambda}}{N!} \tag{5.3}$$

where λ is a function of all parameters $\vec{\theta}$. We can adjust for this by multiplying L by the PDF for N to yield

$$L = \frac{\lambda^N e^{-\lambda}}{N!} \prod_{i=1}^{N} f(\vec{x}^i; \vec{\theta}). \tag{5.4}$$

This is the extended likelihood function for a probability distribution function $f(\vec{x}^i; \vec{\theta})$.

In principle, f can be any arbitrary PDF. However, f is often infeasible or impossible to define analytically. In these cases, an easier approach is to divide f into n_{bin} bins, which turns the likelihood function into the product of n_{bin} Poisson distributions. If we let $\mu_i(\vec{\theta})$ be the expected number of entries in the ith bin and denote the measured number of entries per bin be n_i, we obtain:

$$L(\vec{n}, \vec{\theta}) = \prod_{i=1}^{n_{bin}} \text{Pois} f(n_i; \mu_i(\vec{\theta})) \tag{5.5}$$

We now have a way to express the likelihood of $\vec{\theta}$ given experimental data $\vec{n}$. (Note that the case of multiple observable variables with different binning is handled trivially; each bin contributes one Poisson term to the product regardless of which quantity is being measured.)

5.1.2 Maximum Likelihood Estimation

Typically, we are interested in the most likely "true" values of $\vec{\theta}$ in light of observations (in a frequentist sense). This corresponds to the values of $\vec{\theta}$ that maximize L. However, calculating the product of a large number of terms can pose computational problems, since this multiplication tends to produce extremely small values that may cause floating-point underflows and other issues. To remedy this, one can note that since lnx is a monotonically increasing function, maximizing L is equivalent to maximizing $\ln L$. The ln turns the product into a sum, which is much more convenient to work with. In practice, it is standard to instead minimize $-2\ln L$; this turns out to be slightly cleaner mathematically. We find:

$$-2\ln L(\vec{n},\vec{\theta}) = 2\sum_{i}^{N} \mu_i(\theta) - \vec{n_i \ln \mu_i(\theta)} + \ln n_i! \tag{5.6}$$

This procedure is called a maximum likelihood estimate.

For observations with a large number of entries, the distribution of the expected number of entries per bin is approximately Gaussian, with variance n_i. In light of the second term in the above equation, one can see that maximizing L is equivalent to minimizing

$$\chi^2 = \sum_{i=1}^{n_{bins}} \frac{(n_i - \mu(x_i,\vec{\theta}))^2}{n_i} \tag{5.7}$$

This is recognizably a chi-squared distribution, which has the well-known PDF

$$\mathrm{P}(\chi^2; n_{df}) = \frac{2^{-n_{df}/2}}{\Gamma n_{df}/2}\chi^{n_{df}-2}e^{-\chi^2/2}, \tag{5.8}$$

where $n_{df} = n_{bins} - \dim(\vec{\theta})$ is the number of degrees of freedom. A key advantage of this approximation is that knowing P allows one to estimate the goodness of fit: Low values

of χ^2 indicate a good fit, while high values imply that the assumed model may be wrong. This can be formally interpreted as a p-value, i.e. the probability of obtaining a fit result at least that extreme if the model is correct, via the cumulative distribution function of the χ^2 distribution.[42]

5.1.3 Physics Applications

At modern experiments like CMS, one must go beyond the basic considerations discussed above. First, observed data consists of events produced by a wide variety of processes. A particular final state can be produced in several ways, especially after taking into account complications like detector inefficiencies and mistags. Consequently, μ_i is typically the sum of several different histograms, each corresponding to the contribution from a particular process; each histogram is called a template. These templates are constructed from several sets of MC samples, with each event weighted to mimic the distribution of real data.

Moreover, since MC simulations are imperfect, a common approach is to estimate backgrounds with data-driven methods: Using data to fine-tune estimates of background distributions. It is now standard to divide the phase space of all selected events into signal regions (SRs) and control regions (CRs), with the selections chosen such that the SRs enriched in the signal process of interest and the CRs are enriched in particular backgrounds. This allows the normalization for each of the background processes to be estimated more precisely (as nuisance parameters in a maximum likelihood fit), at which point their values can be (somewhat) safely extrapolated into the signal region. A common background-driven approach is the ABCD method, but complicated analyses may rely on more sophisticated approaches.

Currently, CMS has dedicated software libraries for statistical analysis. The combine

tool, a RooFit-based library, incorporates a wide variety of statistical methods and is used to perform fits and set limits in this analysis.[43][44]

5.2 Jet Tagging at CMS

A key challenge in often encountered in high-energy physics experiments is jet tagging. As discussed previously, quarks produced by high-energy collisions cannot exist as free particles, and instead produce a condensed spray of hadrons called a jet. Jets at CMS are typically assembled via the anti-k_t clustering algorithm,[45] which groups neighboring hadrons in momentum space together. (This is in contrast to cone algorithms, which group jets together based solely on their separation in η and ϕ.) The anti-k_t algorithm depends only on a radius parameter that sets the angular scale of the jet sizes; a typical value is $R = 0.4$, meaning that most jets will be on the scale of a cone with $\sqrt{\eta^2 + \phi^2} \leq 0.4$.

After defining jets, the next step is determining the type of particle that produced the reconstructed jets, a process known as flavor tagging. Because each jet is composed of a large number of light hadrons, and because the mass of the original quark is typically small compared to its kinetic energy, extracting flavor information from the jet is challenging.

5.2.1 Fundamentals of Flavor Tagging

The key to jet tagging is secondary vertex identification. Although t quarks decay too quickly to hadronize and cannot form single jets, b quarks will form b hadrons, which typically have a lifetime on the order of 10^{-12} s. Since these b hadrons are boosted due to the energy of the collision, time dilation ensures that their lifetimes are long enough for them to travel a few hundred microns before decaying, which produces a secondary vertex

a short distance from the primary vertex (PV).[46] See figure 5.1 for an illustration. These secondary vertices can be identified by CMS's inner tracker, which has enough resolution to calculate the impact parameters for each track (their distance of closest approach to the PV) with sub-100 μm precision. Flavor tagging algorithms can therefore make use of this information to distinguish b jets from light-flavor u, d, s, and gluon jets, which do not produce secondary vertices.

c quarks also form short-lived hadrons, so c jets and b jets have similar kinematics. However, c hadrons have an even shorter lifetime than b hadrons, resulting in SVs with a smaller displacement from the beamline. This is problematic in two ways: First, it lowers the tracker's SV reconstruction efficiency, and second, it produces an intermediate case between b and light jets that is hard to distinguish from either. As a result, c tagging is significantly harder than b tagging, with cutting-edge algorithms only able to reach efficiencies on the order of 50%.

Historically, a wide variety of techniques have been used for jet tagging. Earlier approaches utilized custom-designed algorithms. An early "Jet Probability" algorithm, for instance, used the two-dimensional and three-dimensional impact parameters of each track in a jet to estimate the probability that the tracks originated from a PV; the log of this probability could be used as a b-tagging discriminator.[47] The main downside of these approaches is that custom algorithms are insensitive to more subtle differences in jet substructure; they will not use any kinematic information that they are not explicitly designed to account for. In contrast, state-of-the-art methods rely on deep learning. Because neural networks are capable of learning complex correlations between a large number of jet features, they are well-suited for the task of tagging, and are rapidly evolving as both software and hardware continue to improve.

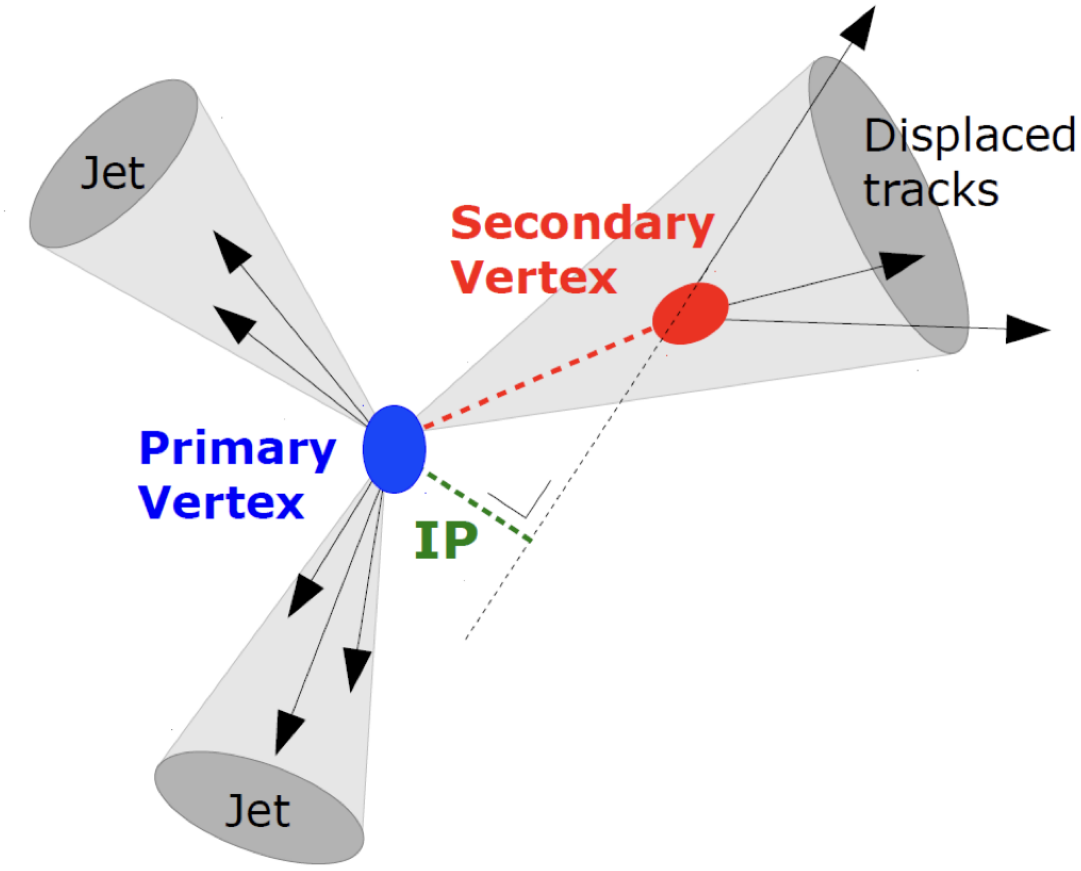


Figure 5.1: Illustration of a secondary vertex (red) produced when a hadron produced by a primary vertex (blue) decays. The impact parameter (IP) of the uppermost track is shown in green. From [46].

5.2.2 ParticleNet

[48]

The neural net architecture selected for jet tagging in this thesis is ParticleNet. Like many other recent jet taggers, ParticleNet is a convolutional neural network (CNN). CNNs are built around an operation called convolution, which skims over many small subsections ("patches") of the input data, applies a filter to each patch, and aggregates the results to yield an output matrix. CNNs take advantage of the fact that many forms of data—for instance, images—have large spatially-local correlations by grouping together information from small patches of input data. Convolution can filter large input objects for simple patterns such as edges, then repeatedly pass the results on to higher-level convolution blocks to identify higher-level features.

CNNs are a powerful tool for identifying higher-level features across a wide range of applications, from image recognition to natural language processing; it is natural to apply them to jet tagging. However, the nature of jets poses some problems for traditional CNN architectures. The simplest way to model a jet is as an image, which can be done by treating the calorimeter as a grid of pixels where the "intensity" of each pixel is the amount of energy deposited in that part of the detector. Unfortunately, the relatively small number of nonzero pixels in an average jet makes this computationally inefficient, with large amounts of processing power and memory wasted on pixels that contain 0s. This representation also has difficulty integrating other forms of data that cannot easily be translated into an image-like representation, such as particle IDs from the tracker. An alternative approach is to model a jet as a set of its constituent particles. This sidesteps both of the previous issues, but introduces a new problem: Sets are not permutation-invariant, so an ordering for the particles must be chosen. This can lead to inefficiencies via accidental choice of a suboptimal ordering.

ParticleNet has its roots in a third way to represent jets: Point clouds. A point cloud is an unordered, permutation-invariant set of points. Each point is defined by several coordinates, which can in principle be anything—simple (x, y, z) spatial coordinates, angular coordinates like ϕ and η, deposited energy, numbers representing particle specie, and so on. This representation has been successfully used in the past to analyze and classify images, and extending it to jets is a natural next step. The main challenge for a point cloud-based jet tagger is defining a permutation-invariant convolution operation that can be applied to a small "patch" of the input. ParticleNet defines each patch to be the k nearest neighbors of each point, with distance computed over the spatial coordinates. We may then define the following edge convolution (EdgeConv) operation for the ith particle in a jet x_i:

$$x_i' = \square_{j=1}^{k} h_\Theta(x_i, x_{i_j}) \tag{5.9}$$

Here $\square$ represents a permutation-invariant channel-aggregating operation, such as the mean or maximum of the inputs. $h_\Theta(x_i, x_{i_j})$ is a permutation-invariant function of a point's coordinates x_i and k nearest neighbors x_{i_j}; it is defined by a number of learnable parameters Θ that will be tuned as the neural network is trained. See Figure 5.2 for an illustration. In short, EdgeConv aggregates information about the relationships between the central point x_i and its k nearest neighbors into a single matrix. Because EdgeConv acts on and returns a matrix, stacking it repeatedly to form a deep neural network is trivial. All that remains is to define the dimensions of each EdgeConv block and to aggregate the results with pooling and fully connected layers. As shown in Figure 5.3, even three EdgeConv layers is sufficient to recognize complex features in three-dimensional images.

In ParticleNet, the number of nearest neighbors k is typically of order 10. $\square$ is

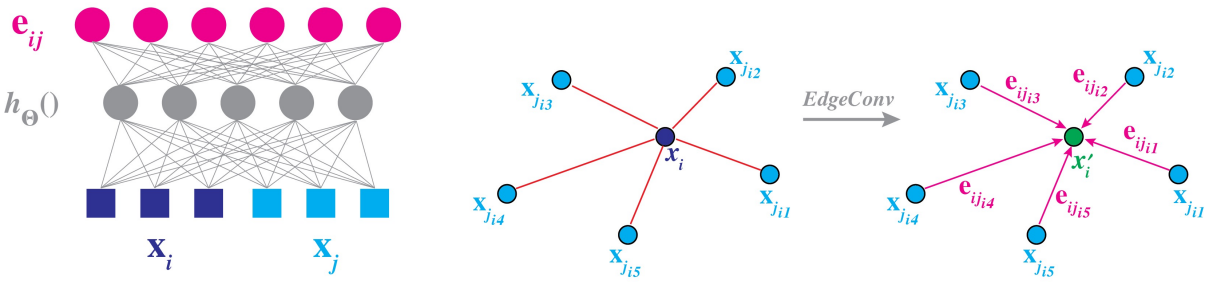


Figure 5.2: Left: Computing a single edge feature $h_\Theta(x_i, x_j)$ for two particles x_i and x_j. Right: A visual representation of the EdgeConv block's behavior. The k nearest neighbors to each particle x_i are calculated, and the edge features are aggregated together to produce the input for the next layer of the network. From [49].

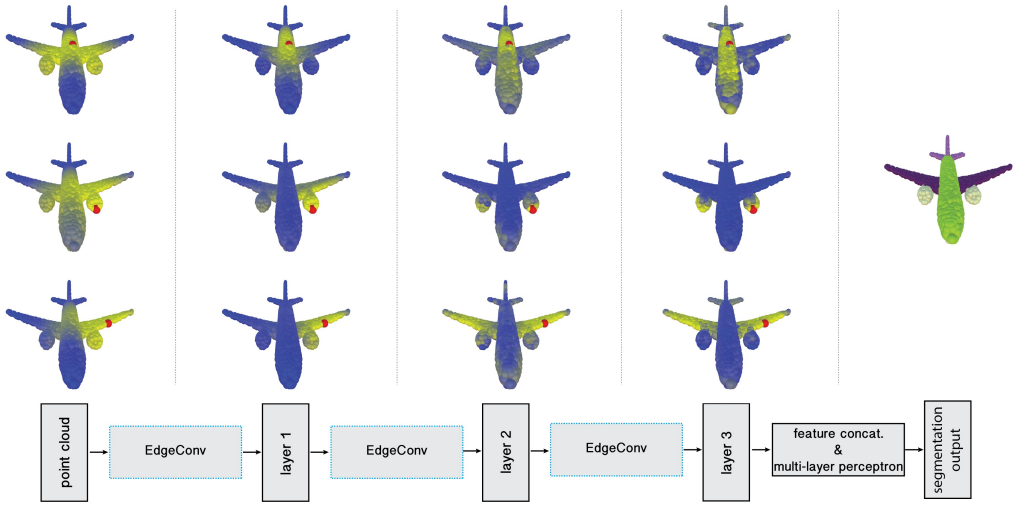


Figure 5.3: Example of EdgeConv identifying features in 3D images. The color gradient in this image represents the feature space distance between a red selected point and all other points on the object, with yellow meaning closer. Initially this is just the geometric distance between the points, but successive EdgeConv blocks will group points based on progressively higher-order features until the object can be categorized neatly. From [49].

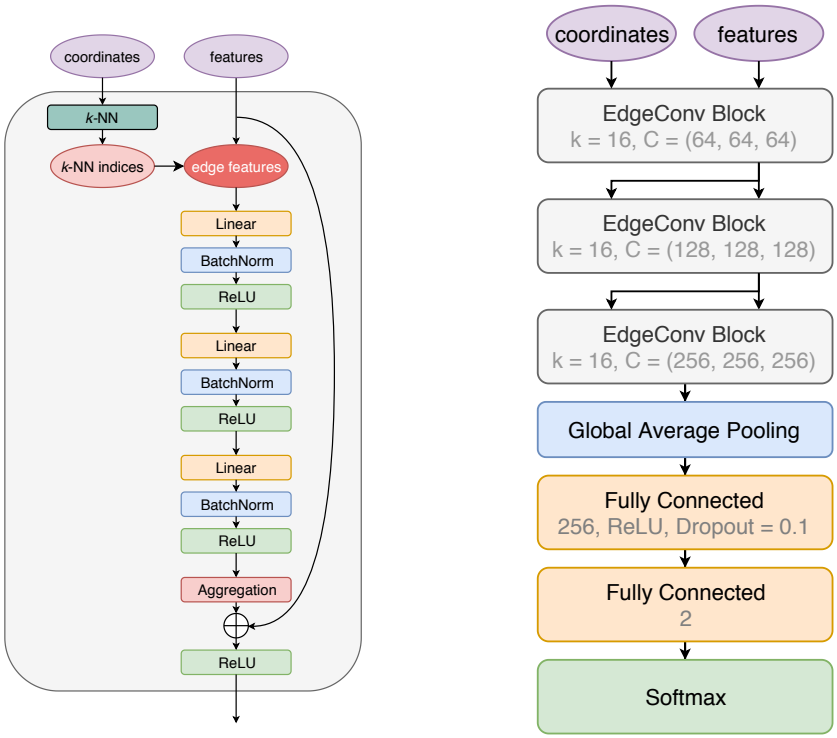


Figure 5.4: Left: Structure of a single EdgeConv block. Right: Schematic of ParticleNet's architecture. Text in gray lists the various layer parameters for the standard ParticleNet implementation, e.g. $k = 16$ nearest neighbors; see [48] for details. In this thesis, a reduced-size version of ParticleNet is used, e.g. with k=8 and C=(96,96,96) for the first layer.

chosen to be the mean operation, and h_Θ is a custom multi-layer perceptron. The full ParticleNet architecture is shown in Figure 5.4. In practice, ParticleNet significantly outperforms state-of-the-art jet tagging CNNs used by CMS such as DeepJet, as 5.5 makes clear.

5.3 Event Classification with Particle Transformer

Another challenging task at CMS is event classification. Even after reconstruction, jet tagging, and filters to reduce backgrounds, it is often difficult to determine which process caused a particular detector signature. This is a particularly pressing issue when

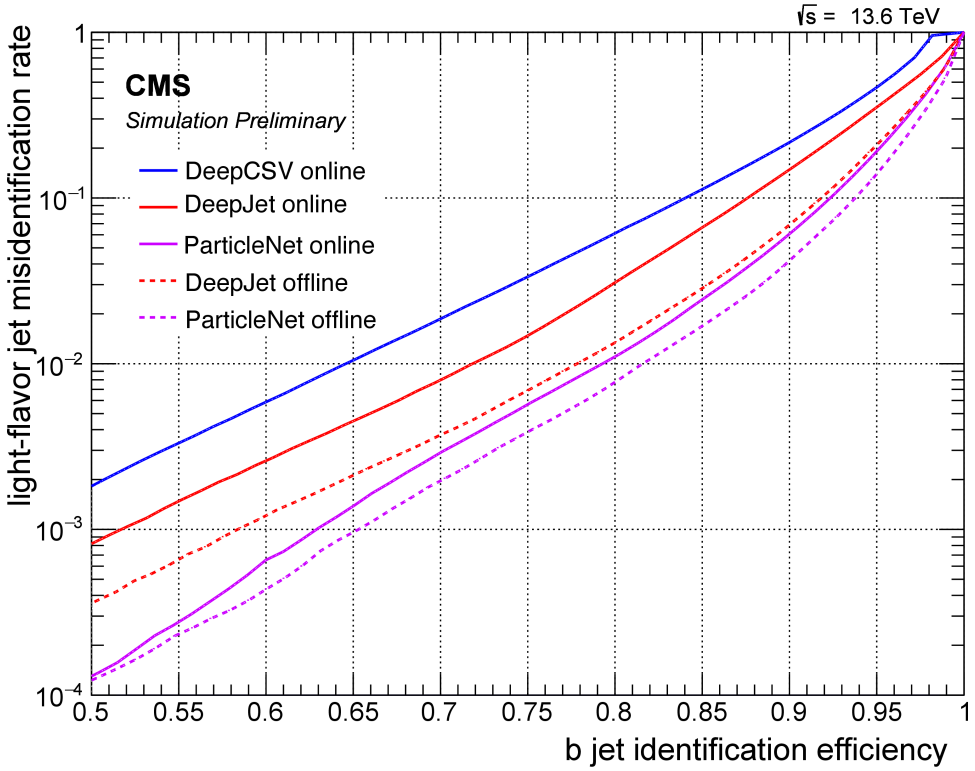


Figure 5.5: ParticleNet b-tagging efficiency on run 3 data at CMS. Notably, ParticleNet outperforms DeepJet, the previous state-of-the-art jet tagging model used by CMS, by a significant margin. (Although run 3 data, it is not sufficiently different from run 2 data to cause a difference in performance.) From [50].

searching for rare processes, since even small numbers of misclassified events can obscure a weak signal. Due to the complicated topologies of many final states at CMS, particularly those involving several jets in the final state, machine learning has also seen widespread use in this context. ML is especially good at exploiting subtle correlations between different features of an event that can't be separated out by simple cuts or selections.

As we have seen in the previous section, the particle cloud model is very successful at handling the various complexities present in jets. Given the generality of the concept, it is natural to use it as an event classifier by applying it to entire events instead of individual jets. This has been done before with ParticleNet in the context of the under-development Light Dark Matter Experiment (LDMX)—although it involves an environment very different from CMS's, ParticleNet has nonetheless outperformed alternative veto algorithms.[51]

Here we consider a recent machine learning model based on the same point cloud concept, Particle Transformer.[52] Like ParticleNet, Particle Transformer is a deep neural net designed to be invariant under permutations of its input particles. However, two key improvements have been made. First, Particle Transformer explicitly uses pairwise relationships between particles in its structure; ParticleNet does not and is less sensitive to these correlations. Second, Particle Transformer is a transformer-based architecture that makes use of the attention mechanism, while ParticleNet is a simpler convolutional neural net. A brief overview of the Particle Transformer architecture follows.

Particle Transformer accepts two inputs: C per-particle features, and C' interaction features for every pair of particles. Per-particle features include the usual kinematic variables such as energy, ϕ, η, and so on. For interaction features, the following variables

are calculated for each pair of particles a and b:

$$\Delta = \sqrt{(y_a - y_b)^2 + (\phi_a - \phi_b)^2} \tag{5.10}$$

$$k_T = \min(p_{T,a}, p_{T,b})\Delta \tag{5.11}$$

$$z = \frac{\min(p_{T,a}, p_{T,b})}{p_{T,a} + p_{T,b}} \tag{5.12}$$

$$m^2 = (E_a + E_b)^2 - ||\mathbf{p_a} + \mathbf{p_b}||^2 \tag{5.13}$$

These quantities tend to be long-tailed, so to improve stability and overall performance of the NN, the natural log of all of each quantity is used instead of the raw value.

As mentioned above, Particle Transformer is a transformer-based architecture. Transformer models are a recent development in deep learning that use an "attention" mechanism to focus on the most important parts of the input while paying less attention to extraneous information.[53] In general, transformer models consist of two sections, an encoder and a decoder. The encoder makes up the first few layers of the model; its purpose is to learn how to "encode" the input into a form that makes it easier to classify. The decoder then reads the output of the encoder and decides how to classify the input. See Figure 5.6 for a graphical representation of Particle Transformer's architecture.

Particle Transformer was initially tested for jet tagging, a task at which it outperforms other state-of-the-art taggers such as ParticleNet. In particular, Table 5.4 shows Particle Transformer's performance on a large jet tagging dataset; Particle Transformer achieves better results for all considered categories. As is the case with ParticleNet, Particle Transformer's generality means that it can also be used as an event classifier with no changes to the architecture. The work described in this thesis is the first analysis to use it for this purpose.

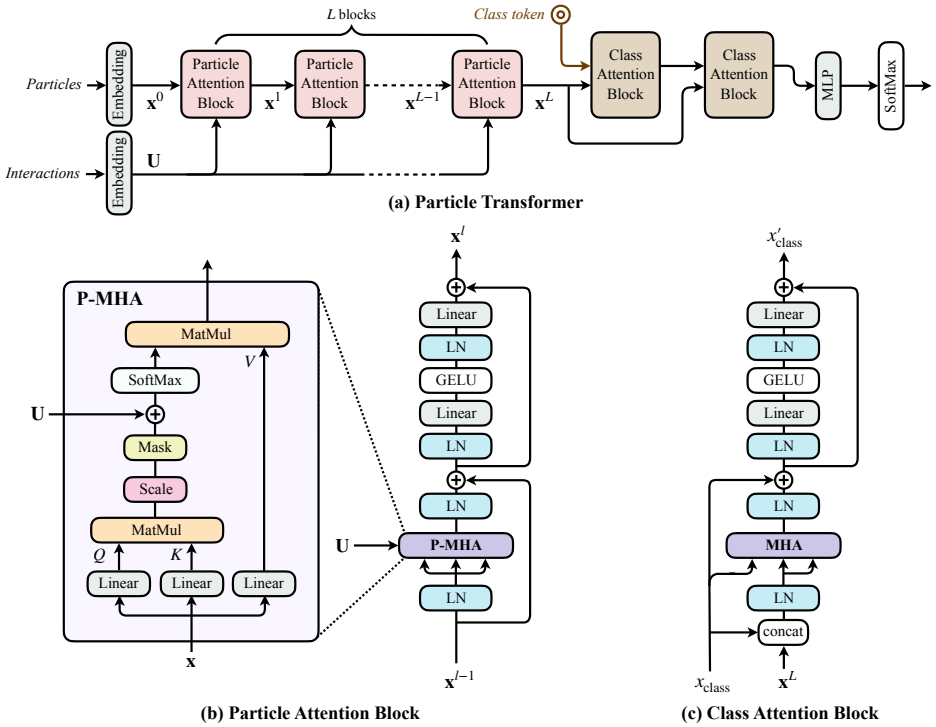


Figure 5.6: Top: Schematic of the Particle Transformer architecture. The first three "particle attention" blocks are the encoder, while the last two "class attention" blocks are the decoder. Bottom: Schematic of the particle attention (left) and class attention (right) blocks. Here P-MHA refers to a specialized multi-headed attention module designed to focus on pairwise features between particles; see [52] 2022 for details.

	Accuracy	AUC	$H \to b\bar{b}$ $\text{Rej}_{50\%}$	$H \to c\bar{c}$ $\text{Rej}_{50\%}$	$H \to gg$ $\text{Rej}_{50\%}$	$H \to 4q$ $\text{Rej}_{50\%}$	$H \to \ell\nu qq'$ $\text{Rej}_{99\%}$	$t \to bqq'$ $\text{Rej}_{50\%}$	$t \to b\ell\nu$ $\text{Rej}_{99.5\%}$	$W \to qq'$ $\text{Rej}_{50\%}$	$Z \to q\bar{q}$ $\text{Rej}_{50\%}$
PFN	0.772	0.9714	2924	841	5	198	265	797	721	189	159
P-CNN	0.809	0.9789	4890	1276	88	474	947	2907	2304	241	204
ParticleNet	0.844	0.9849	7634	2475	104	954	3339	10526	11173	347	283
ParT	**0.861**	**0.9877**	**10638**	**4149**	**123**	**1864**	**5479**	**32787**	**15873**	**543**	**402**

Table 5.1: Table comparing Particle Transformer's (ParT's) jet-tagging performance against that of other commonly used models. Here $\text{Rej}_{50\%}$ is the inverse of the false positive rate at a true positive rate of 50%. A higher value corresponds to a higher rejection rate. Evidently, Particle Transformer outperforms standard convolutional neural net models as well as ParticleNet by a significant margin.[52]

Chapter 6

Searching for $t\bar{t}Hc\bar{c}$

This chapter introduces the $t\bar{t}H(H\rightarrow c\bar{c})$ process, which is the focus of this thesis. As discussed in Chapters 1 and 2, the need for an accurate measurement of the Higgs-charm coupling is well-motivated by experimental and theoretical considerations: It is the easiest yet-unobserved Higgs coupling to measure, and BSM physics may manifest as a Yukawa coupling in excess of the SM value. To this end, an analysis to constrain the Higgs-charm coupling using a new method is currently in the final stages of development.

Section 6.1 provides some background and experimental context for Higgs-charm coupling measurements. Section 6.2 introduces the $t\bar{t}H(h\rightarrow c\bar{c})$ process, as well as the various challenges that a measurement of this final state faces. Finally, section 6.3 gives an overview of the analysis' structure in the context of previous searches for $t\bar{t}H(H\rightarrow c\bar{c})$.

6.1 Experimental context

Several previous attempts to constrain the Higgs-charm Yukawa coupling have been made. Before proceeding, we will establish the following definitions per convention: $\sigma(X)$ is the cross-section of the process X, $\mathcal{B}(H \rightarrow q\bar{q})$ is the branching fraction for the $H\rightarrow q\bar{q}$

decay mode, and κ_c^2 is defined as the ratio of the predicted decay width $\Gamma_{c\bar{c}}^{SM}$ to the measured decay width $\Gamma_{c\bar{c}}$.[54]

The current strongest limit comes from a search for a Higgs produced in conjunction with a leptonically-decaying V (W or Z) boson, where the Higgs decays to two charms. The authors established an observed (expected) limit of $\sigma(VH)\mathcal{B}(H \to c\bar{c}) < 0.94(0.50^{+0.22}_{-0.15})$ at the 95% confidence level. Equivalently, this sets the current upper bound of μ at $14(7.6^{+3.4}_{-2.3})$ times the SM value, as well as a constraint on the Yukawa coupling modifier of $1.1 < |\kappa_c| < 5.5$ ($|\kappa_c| < 3.4$). Like the analysis discussed in this thesis, the $VHc\bar{c}$ analysis used ParticleNet for jet tagging; however, their primary tool for event classification and background mitigation was a boosted decision tree (BDT).[55]

Another notable effort investigated an alternative channel: Higgs production through gluon-gluon fusion. A Higgs produced in this manner will be heavily boosted, so the experimental signature for this analysis was two overlapping high-p_T charm jets. While less sensitive than the $VHc\bar{c}$ approach, boosted ggH had the advantage of considering a highly orthogonal search region: A different Higgs production mode, a different p_T range, and a separate region of data. The analysis used DeepJet to reconstruct Higgs candidates, and reached an overall limit on μ of 45 (38) times the SM value.[56]

Finally, a search in the $VHc\bar{c}$ channel was conducted with the ATLAS Run II dataset. Using a multivariate algorithm for charm tagging as well as a separate b-tagging algorithm to veto remaining backgrounds, they obtained a limit of 26 (31) times the SM value at the 95% confidence level, or equivalently $|\kappa_c| < 8.5(12.4)$.[57]

Another approach that has been considered is searching for a Higgs produced in conjunction with a single charm quark.[58][59] This the advantage of a clear signature: Unlike in the other approaches, the Higgs can be reconstructed via standard decay channels, requiring only a single charm tag. A CMS analysis is ongoing; the result has not yet been published. Additionally, the possibility of using a scouting approach—using data

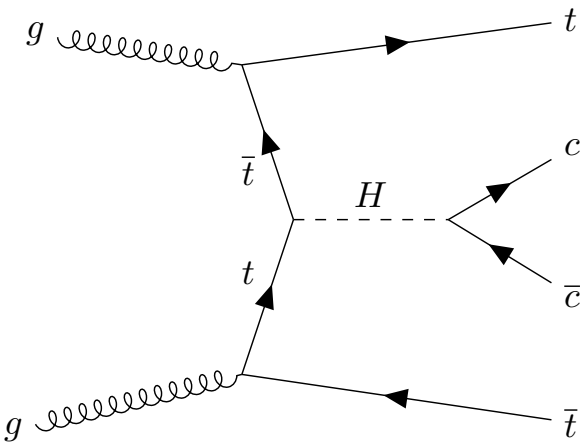


Figure 6.1: Feynman diagram of the ttH(H→ $c\bar{c}$ process.

from the online reconstruction, which is more inclusive than the standard CMS triggers and results in a much larger sample of events—to probe this production mode has also been considered. Unfortunately, preliminary estimates of the search sensitivity were not encouraging and this direction was ultimately not pursued by our research group.

6.2 Introduction to $t\bar{t}H(h\rightarrow c\bar{c})$

A production mode that has not yet been investigated is a Higgs produced in conjunction with two top quarks, with the Higgs decaying to two charms, $t\bar{t}H(h\rightarrow c\bar{c})$. Figure 6.1 contains a simple Feynman diagram of the process. Although the presence of two top quarks results in a complicated final state with multiple jets, this process is predicted to have a reasonably high yield and has the potential to be competitive with $VHc\bar{c}$ with a sufficiently powerful background rejection strategy.

This analysis faces numerous challenges. First, compared to other Higgs decay modes such as $H \rightarrow b\bar{b}$, the branching fraction $\mathcal{B}(H \rightarrow c\bar{c})$ is quite small, resulting in a weak signal. Second, significant backgrounds are expected in all channels, necessitating a robust background estimation and rejection strategy. A critical component of this strategy is a

powerful jet tagger: Two hard-to-tag charm quarks are present in the event signature, meaning that this process can only be taken advantage of with the assistance of a highly accurate jet tagger. Finally, the presence of numerous jets in the final state mandates a powerful method for event classification.

6.3 Towards a ttH(H$\rightarrow c\bar{c}$) measurement

Thanks to the machine learning tools established in Chapter 5, the time is ripe for a proper search for $t\bar{t}H(h\rightarrow c\bar{c})$. The analysis described in this thesis makes use of all three years of Run 2 data, for a total of 138 fb^{-1}. It is divided into three channels, fully-hadronic (FH), dilepton (DL), and single-lepton (SL). In the FH channel, both top quarks decay hadronically, almost always to a b quark in conjunction with a lighter quark-antiquark pair. The FH channel is complex: As is the norm with all-hadronic channels, large QCD backgrounds are present, resulting in a need for heavy reliance on machine learning. Despite this, the projected sensitivity is comparable to that of the SL channel. In the dilepton channel, both tops decay to either an electron or a muon, a corresponding neutrino, and a b quark. These leptons are easier to identify, but this is counterbalanced by the top's lower branching fraction to leptons, resulting in a less powerful channel. Finally, the single-lepton channel features one hadronic top decay and one leptonic top decay, as illustrated in Figure 6.2. The presence of a lepton in the final state helps reduce QCD backgrounds, making the SL channel a reasonably powerful one. This thesis will focus primarily on the SL channel, which can be further divided into two streams: Events with a single electron and events with a single muon.

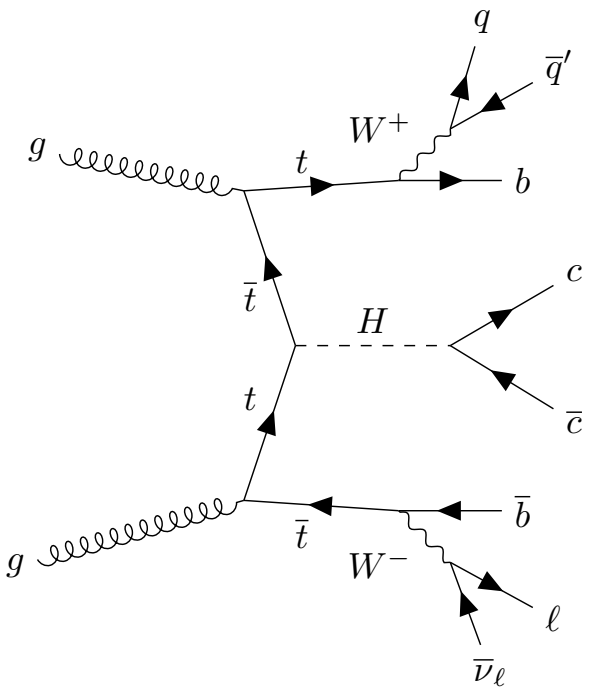


Figure 6.2: Feynman diagram of the single-lepton channel. q and q' are different quark flavors, with the branching ratio of each allowed combination being proportional to the corresponding CKM matrix element; e.g. $|V_{ud}|^2$ for an up and an anti-down.

Chapter 7

Event and Object Selection

Before proceeding with the analysis itself, a number of preliminary steps must be taken. All necessary data and Monte Carlo samples must be assembled, filtered for event quality, and pared down by the triggers chosen for the analysis. Additionally, because Particle-Flow is designed to be fairly loose in its object identification criteria to accommodate the wide range of analyses performed at CMS, an additional set of criteria for each physics object must be applied. Once this is done, the jets can then be flavor-tagged. Finally, a number of corrections must be made to account for data-MC differences in trigger efficiencies, a handful of problems with the detector, and several other important effects.

This chapter begins with an overview of all SL data and Monte Carlo samples in Section 7.1, followed by the choice of triggers in section 7.2. Physics object definitions and flavor tagging are covered in section 7.3. Finally, section 7.4 discusses various corrections made prior to the background estimation step.

7.1 Datasets

All data and Monte Carlo samples used in this analysis are stored using the `NanoAODv9` format. In the case of MC, some samples are only used indirectly for calculating systematics and comparing different event generation methods.

Before proceeding, it must be noted that all samples for 2016 are affected by a significant issue with the strip tracker. At the beginning of Run 2, the pre-amplifier of the APV25 readout chip for the silicon strip sensors was unexpectedly impacted by the tracker's low operating temperature, leading to an unusually slow discharge. This resulted in a low signal-to-noise ratio in the tracker at high event rates. The issue was fixed as of August 13th, 2016 after the pre-amplifier bias voltage (VFP) was adjusted; however, the problem had affected a large portion of the 2016 dataset by then.[60] Because of the significant difference in tracker performance before and after this issue was fixed, the 2016 data and MC samples have been split into two eras, denoted "2016 pre-VFP" and "2016 post-VFP". The detector response in the 2016 pre-VFP MC files has been altered to emulate the tracker's performance while the issue was present.

7.1.1 Data samples

To maximize its sensitivity, this analysis makes use of the full Run 2 dataset for a total integrated luminosity of 137.64 fb^{-1}. Data samples for 2016 pre-VFP, 2016 post-VFP, 2017, and 2018 are listed in Tables **??**, 7.1.1, 7.1.1, and 7.1.1, respectively. Table 7.5 lists the luminosity for each year, as well as the samples used for the single-electron and single-muon streams.

Table 7.1: 2016preVFP primary datasets used in the analysis.

Sample	Run range
/SingleMuon/Run2016B-ver2_HIPM_UL2016_MiniAODv2-v2/MINIAOD	273150-275376
/SingleMuon/Run2016C-HIPM_UL2016_MiniAODv2-v2/MINIAOD	275657-276283
/SingleMuon/Run2016D-HIPM_UL2016_MiniAODv2-v2/MINIAOD	276315-276811
/SingleMuon/Run2016E-HIPM_UL2016_MiniAODv2-v2/MINIAOD	276831-277420
/SingleMuon/Run2016F-HIPM_UL2016_MiniAODv2-v2/MINIAOD	277932-278807
/SingleElectron/Run2016B-ver2_HIPM_UL2016_MiniAODv2-v2/MINIAOD	273150-275376
/SingleElectron/Run2016C-HIPM_UL2016_MiniAODv2-v2/MINIAOD	275657-276283
/SingleElectron/Run2016D-HIPM_UL2016_MiniAODv2-v2/MINIAOD	276315-276811
/SingleElectron/Run2016E-HIPM_UL2016_MiniAODv2-v5/MINIAOD	276831-277420
/SingleElectron/Run2016F-HIPM_UL2016_MiniAODv2-v2/MINIAOD	277932-278807

Table 7.2: 2016postVFP primary datasets used in the analysis.

Sample	Run range
/SingleMuon/Run2016F-UL2016_MiniAODv2-v2/MINIAOD	278769-278808
/SingleMuon/Run2016G-UL2016_MiniAODv2-v2/MINIAOD	278820-280385
/SingleMuon/Run2016H-UL2016_MiniAODv2-v2/MINIAOD	280919-284044
/SingleElectron/Run2016F-UL2016_MiniAODv2-v2/MINIAOD	278769-278808
/SingleElectron/Run2016G-UL2016_MiniAODv2-v2/MINIAOD	278820-280385
/SingleElectron/Run2016H-UL2016_MiniAODv2-v2/MINIAOD	280919-284044

7.1.2 Monte Carlo samples

Monte Carlo samples for this analysis were selected to ensure coverage of all significant background processes. All samples use the CP5 tune,[61] and parton showering is modeled with pythia v8.240. Where relevant, samples are generated using a Higgs mass of 125 GeV and a top quark mass of 172.5 GeV.

All MC samples used for background modeling are listed in Table 7.6. The samples were generated using the following methods:

- **$t\bar{t}$ ($t\bar{t}b\bar{b}$ excepted), $t\bar{t}W$, single top (t-channel only), and $t\bar{t}H$:** Generated with POWHEG v2 at next-leading order (NLO).
- **$t\bar{t}b\bar{b}$:** Generated with POWHEG-BOX-RES with OpenLoops at NLO using a four-flavor scheme (see below).

Table 7.3: 2017 primary datasets used in the analysis.

Sample	Run range
/SingleMuon/Run2017B-UL2017_MiniAODv2-v1/MINIAOD	297046-299329
/SingleMuon/Run2017C-UL2017_MiniAODv2-v1/MINIAOD	299368-302029
/SingleMuon/Run2017D-UL2017_MiniAODv2-v1/MINIAOD	302030-303434
/SingleMuon/Run2017E-UL2017_MiniAODv2-v1/MINIAOD	303824-304797
/SingleMuon/Run2017F-UL2017_MiniAODv2-v1/MINIAOD	305040-306462
/SingleElectron/Run2017B-UL2017_MiniAODv2-v1/MINIAOD	297046-299329
/SingleElectron/Run2017C-UL2017_MiniAODv2-v1/MINIAOD	299368-302029
/SingleElectron/Run2017D-UL2017_MiniAODv2-v1/MINIAOD	302030-303434
/SingleElectron/Run2017E-UL2017_MiniAODv2-v1/MINIAOD	303824-304797
/SingleElectron/Run2017F-UL2017_MiniAODv2-v1/MINIAOD	305040-306462

Table 7.4: 2018 primary datasets used in the analysis.

Sample	Run range
/SingleMuon/Run2018A-UL2018_MiniAODv2_GT36-v2/MINIAOD	315252-316995
/SingleMuon/Run2018B-UL2018_MiniAODv2_GT36-v2/MINIAOD	317080-319310
/SingleMuon/Run2018C-UL2018_MiniAODv2_GT36-v3/MINIAOD	319337-320065
/SingleMuon/Run2018D-UL2018_MiniAODv2_GT36-v2/MINIAOD	320673-325175
/EGamma/Run2018A-UL2018_MiniAODv2_GT36-v1/MINIAOD	315252-316995
/EGamma/Run2018B-UL2018_MiniAODv2_GT36-v1/MINIAOD	317080-319310
/EGamma/Run2018C-UL2018_MiniAODv2_GT36-v1/MINIAOD	319337-320065
/EGamma/Run2018D-UL2018_MiniAODv2_GT36-v2/MINIAOD	320673-325175

- **$t\bar{t}Z$ and s-channel $t\bar{t}$:** Generated with MadGraph5_aMC@NLO v2.6.5.
- **$t\bar{t}W$:** Generated with MadGraph5_aMC@NLO v2.6.5. Additionally, at most one extra jet was simulated and merged into each event using FxFx.
- **Heavy particle decay for $t\bar{t}W$, $t\bar{t}Z$, and t-channel single top:** Modeled with MadSpin.

Of particular note are the four-flavor scheme samples. (Note: Flavor scheme refers to the number of quark flavors used in MC simulations. In the four-flavor scheme, u, d, and s quarks are simulated together as a single flavor; in the five-flavor scheme, s quarks are given their own flavor.) In this analysis, $t\bar{t}b\bar{b}$ events are an especially critical background.

Table 7.5: All data samples used in the single-lepton channel and their corresponding luminosities. Samples for the single-electron and single-muon streams are indicated separately.

Year	Luminosity	Muon Datasets	Electron Datasets
2016pre	19.52 fb^{-1}	`SingleMuon:B-F`	`SingleElectron:B-F`
2016post	16.81 fb^{-1}	`SingleMuon:F-H`	`SingleElectron:F-H`
2017	41.48 fb^{-1}	`SingleMuon:B-F`	`SingleElectron:B-F`
2018	59.83 fb^{-1}	`SingleMuon:A-D`	`EGamma:A-D`

As explained previously in 5.2.1, b jets are easily mistaken for c jets; this leads to small but significant mistag rates. Moreover, the $t\bar{t}b\bar{b}$ process has a relatively high branching fraction, resulting in a significant number of signal-like $t\bar{t}b\bar{b}$ events. (See the c-tagging section of Figure 7.1 later in this thesis for sample working points.) This background must therefore be modeled as precisely as possible. $t\bar{t}b\bar{b}$ events are conventionally modeled with $t\bar{t}$ events in the five-flavor scheme (5FS), but because the b quarks in these simulations are generated at the parton shower level, they come packaged with significant uncertainties and are impacted strongly by the choice of tune. Moreover, $t\bar{t}$ simulations typically overestimate the $t\bar{t}b\bar{b}$ cross-section by a factor of 1.2-1.4.

Here, we use an improved $t\bar{t}b\bar{b}$ model developed by an in-progress (currently unpublished) analysis searching for $t\bar{t}H(H\rightarrow b\bar{b})$. $t\bar{t}b\bar{b}$ events are generated at matrix-level by a dedicated 4FS simulation.[62] The kinematics of this $t\bar{t}b\bar{b}$ sample differ significantly from those of $t\bar{t}b\bar{b}$ events in $t\bar{t}$, as found in [63]; the lower-level nature of the simulation makes this method preferred even though existing $t\bar{t}b\bar{b}$ measurements are not precise enough to conclusively favor one approach over the other. These $t\bar{t}b\bar{b}$ events are then merged into the $t\bar{t}$ sample via a simple procedure: Every $t\bar{t}b\bar{b}$ event in the 5FS sample is replaced with an event from the 4FS sample. The normalization is kept free-floating in the eventual fit, but in pre-fit plots, $t\bar{t}b\bar{b}$ is normalized to the 5FS yield, which is known to be more accurate than the 4FS normalization.

Table 7.6: All Monte Carlo samples used in the single-lepton channel.

Channel	Sample	XS [pb]
$t\bar{t}H c\bar{c}$	ttHTocc_M125_TuneCP5_13TeV-powheg-pythia8	0.015
$t\bar{t}H b\bar{b}$	ttHTobb_M125_TuneCP5_13TeV-powheg-pythia8	0.295
$t\bar{t}$	TTToHadronic_TuneCP5_13TeV-powheg-pythia8	379.265
	TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8	366.226
	TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8	88.409
$t\bar{t}b\bar{b}$	TTbb_4f_TTToHadronic_TuneCP5-Powheg-Openloops-Pythia8	19.902
	TTbb_4f_TTToSemiLeptonic_TuneCP5-Powheg-Openloops-Pythia8	19.218
	TTbb_4f_TTTo2L2Nu_TuneCP5-Powheg-Openloops-Pythia8	4.639
Single top	ST_s-channel_4f_hadronicDecays_TuneCP5_13TeV-amcatnlo-pythia8	3.110
	ST_t-channel_antitop_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8	80.0
	ST_t-channel_top_4f_InclusiveDecays_TuneCP5_13TeV-powheg-madspin-pythia8	134.2
	ST_tW_antitop_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8	39.65
	ST_tW_top_5f_inclusiveDecays_TuneCP5_13TeV-powheg-pythia8	39.65
	ST_s-channel_4f_leptonDecays_TuneCP5_13TeV-amcatnlo-pythia8	3.729
	ST_tW_antitop_5f_NoFullyHadronicDecays_TuneCP5_13TeV-powheg-pythia8	21.617
	ST_tW_top_5f_NoFullyHadronicDecays_TuneCP5_13TeV-powheg-pythia8	21.617
$t\bar{t}W$	TTWJetsToLNu_TuneCP5_13TeV-amcatnloFXFX-madspin-pythia8	0.196
	TTWJetsToQQ_TuneCP5_13TeV-amcatnloFXFX-madspin-pythia8	0.405
$t\bar{t}Z$	TTZToLLNuNu_M-10_TuneCP5_13TeV-amcatnlo-pythia8	0.253
	TTZToQQ_TuneCP5_13TeV-amcatnlo-pythia8	0.586
QCD	QCD_HT300to500_TuneCP5_13TeV-madgraphMLM-pythia8	322600
	QCD_HT500to700_TuneCP5_13TeV-madgraphMLM-pythia8	29980
	QCD_HT700to1000_TuneCP5_13TeV-madgraphMLM-pythia8	6334
	QCD_HT1000to1500_TuneCP5_13TeV-madgraphMLM-pythia8	1088
	QCD_HT1500to2000_TuneCP5_13TeV-madgraphMLM-pythia8	99.11
	QCD_HT2000toInf_TuneCP5_13TeV-madgraphMLM-pythia8	20.23
W+jets	WJetsToQQ_HT-200to400_TuneCP5_13TeV-madgraphMLM-pythia8	2549.0
	WJetsToQQ_HT-400to600_TuneCP5_13TeV-madgraphMLM-pythia8	276.5
	WJetsToQQ_HT-600to800_TuneCP5_13TeV-madgraphMLM-pythia8	59.25
	WJetsToQQ_HT-800toInf_TuneCP5_13TeV-madgraphMLM-pythia8	28.75
	WJetsToLNu_0J_TuneCP5_13TeV-amcatnloFXFX-pythia8	48716.955
	WJetsToLNu_1J_TuneCP5_13TeV-amcatnloFXFX-pythia8	8107.312
	WJetsToLNu_2J_TuneCP5_13TeV-amcatnloFXFX-pythia8	3049.263
Z+jets	ZJetsToQQ_HT-200to400_TuneCP5_13TeV-madgraphMLM-pythia8	1012.0
	ZJetsToQQ_HT-400to600_TuneCP5_13TeV-madgraphMLM-pythia8	114.2
	ZJetsToQQ_HT-600to800_TuneCP5_13TeV-madgraphMLM-pythia8	25.34
	ZJetsToQQ_HT-800toInf_TuneCP5_13TeV-madgraphMLM-pythia8	12.99
	DYJetsToLL_M-10to50_TuneCP5_13TeV-madgraphMLM-pythia8	22635
	DYJetsToLL_M-50_TuneCP5_13TeV-amcatnloFXFX-pythia8	6077.22
$t\bar{t}H\tau\bar{\tau}$	ttHToTauTau_M125_TuneCP5_13TeV-powheg-pythia8	0.032
ttHNonbb	ttHToNonbb_M125_TuneCP5_13TeV-powheg-pythia8	0.212

Additionally, a number of alternative MC samples are used in the analysis for a variety of other purposes. They include:

- **HDAMP:** The emission cross-section in POWHEG is scaled by a damping function h_{damp}, which has a significant effect on initial- and final-state radiation at high p_T ranges. The correct value of this damping function depends on the process being simulated, and the uncertainty in this value must be taken into account by examining MC samples with a higher and lower value of h_{damp}. $h_{\text{damp}} = 1.379 m_t$, the nominal value for the 4FS $t\bar{t}b\bar{b}$ sample, is used in this analysis.

- **Herwig:** Herwig is an alternative to the standard pythia generator, and is known to model data more accurately in some situations. Some samples generated with Herwig are included to check for any differences with pythia that could affect the final results.

- **FxFx:** FxFx is an alternative to POWHEG for computing NLO corrections to parton showering. As with Herwig, FxFx samples are included for comparison with the standard POWHEG samples.

- **Five-flavor scheme:** 5FS samples were also generated for the purpose of comparison against the 4FS samples described above.

A full list of alternative MC samples may be found in Table 7.7.

7.2 Triggers

Two sets of triggers are required for the single-lepton channel, one for the single-electron stream and one for the single-muon stream. The triggers for each year are listed

Table 7.7: All remaining MC samples used in this analysis.

Channel	Sample
$t\bar{t}$, h_damp up	TTToHadronic_hdampUP_TuneCP5_13TeV-powheg-pythia8
	TTTo2L2Nu_hdampUP_TuneCP5_13TeV-powheg-pythia8
$t\bar{t}$, h_damp down	TTToHadronic_hdampDOWN_TuneCP5_13TeV-powheg-pythia8
	TTToSemiLeptonic_hdampDOWN_TuneCP5_13TeV-powheg-pythia8
	TTTo2L2Nu_hdampDOWN_TuneCP5_13TeV-powheg-pythia8
$t\bar{t}b\bar{b}$, h_damp up	TTbb_4f_TTToHadronic_hdampUP_TuneCP5-Powheg-Openloops-Pythia8
	TTbb_4f_TTToSemiLeptonic_hdampUP_TuneCP5-Powheg-Openloops-Pythia8
	TTbb_4f_TTTo2L2Nu_hdampUP_TuneCP5-Powheg-Openloops-Pythia8
$t\bar{t}b\bar{b}$, h_damp down	TTbb_4f_TTToHadronic_hdampDOWN_TuneCP5-Powheg-Openloops-Pythia8
	TTbb_4f_TTToSemiLeptonic_hdampDOWN_TuneCP5-Powheg-Openloops-Pythia8
	TTbb_4f_TTTo2L2Nu_hdampDOWN_TuneCP5-Powheg-Openloops-Pythia8
$t\bar{t}$+jets	TTJets_TuneCP5_13TeV-amcatnloFXFX-pythia8
$t\bar{t}$ herwig	TT_TuneCH3_13TeV-powheg-herwig7
$t\bar{t}Hb\bar{b}$ FxFx	ttHJetTobb_M125_TuneCP5_13TeV_amcatnloFXFX_madspin_pythia8
$t\bar{t}$HNonbb FxFx	ttHJetToNonbb_M125_TuneCP5_13TeV_amcatnloFXFX_madspin_pythia8

in Tables 7.2, 7.2, and 7.2, and follow standard CMS recommendations for selecting single electrons and muons.

Table 7.8: Triggers used for 2016 data.

Stream	Triggers	Run Era
e	HLT_Ele27_WPTight_Gsf	B-H
μ	HLT_IsoMu24	B-H
	HLT_IsoTkMu24	B-H

Table 7.9: Triggers used for 2017 data.

Stream	Triggers	Run Era
e	HLT_Ele32_WPTight_Gsf_L1DoubleEG_L1Flag	B-F
	HLT_Ele28_eta2p1_WPTight_Gsf_HT150	B-F
μ	HLT_IsoMu27	B-F

Table 7.10: Triggers used for 2018 data.

Stream	Triggers	Run Era
e	HLT_Ele32_WPTight_Gsf	A-D
	HLT_Ele28_eta2p1_WPTight_Gsf_HT150	A-D
μ	HLT_IsoMu24	A-D

7.3 Object Selection

As mentioned previously, all physics objects must be assembled from ParticleFlow candidates using an established set of selection criteria. This section describes the criteria used for objects in the SL channel.

7.3.1 Primary vertices

The primary vertex (PV) of each event is defined to be the vertex with the highest summed p_T^2 of all matching ParticleFlow candidates. The following criteria must be satisified:

- The PV is a vertex created from a fit to reconstructed tracks.
- The fit contains at least 5 degrees of freedom.
- The PV is no more than 24 cm from the origin of the detector ($|z| < 24$ cm).
- The PV is at most 2 cm away from the beamline (ρ<2 cm).

If no PV matching these criteria is found in an event, the event is discarded.

7.3.2 Leptons

The single-lepton channel uses "tight" identification criteria for both electrons and muons, following the standard CMS recommendations. (The FH and DL channels use looser criteria, since the former is more concerned with rejecting possible leptons and the latter needs higher signal efficiency.) Electron candidates must fall above a year-dependent p_T threshold (29 GeV for 2017 and 26 GeV for other years) and have a pseudorapidity in the range $|\eta| < 2.4$. Additionally, they must pass the MVA-based ID and isolation requirements set by the `mvaFall17V2Iso_WP80` working point, which

corresponds to an electron signal efficiency of 80%.[64] Like electrons, muons must also fall above a year-dependent p_T threshold (29 GeV for 2016 and 30 for 2017 and 2018) and have a pseudorapidity in the range $|\eta| < 2.4$. Furthermore, they must pass the `CutBasedIdTight` criteria and have a PF relative isolation of at most 0.06 ($\delta\beta$-corrected, $\Delta R < 0.4$), as well as impact parameters of $|d_{xy}| < 0.05$ cm and $|d_z| < 0.2$ cm relative to the primary vertex (`TightRelIso`).[65]

7.3.3 Jets

As mentioned previously, jets are reconstructed by the anti-k_T algorithm with $R = 0.4$. All jets must additionally have a p_T above 25 GeV, fall within the usual pseudorapidity range of $|\eta| < 2.4$, and pass the `tightLepVeto` requirement. Lastly, jets with p_T <50 GeV must also pass a tight pileup ID requirement. This is determined during reconstruction: Jets are evaluated by a boosted decision tree trained to distinguish good data from pileup and must exceed a certain discriminator threshold.[66]

Jets are tagged by a reduced-size version of ParticleNet, using k=8 nearest neighbors. An 8-category multiclass approach is used, with separate categories for c quarks, b quarks, light (uds) quarks, and gluons as well as two additional categories for pileup and undefined jets. The truth-level definitions of each category used in the training are given in Table 7.11. A wide variety of ParticleFlow candidate and SV variables are used as input features, including η and ϕ, 2- and 3-dimensional impact parameters, reconstructed energy in the case of PF candidates, and more. See [67] for a full list. Lastly, the training dataset was weighted to match the p_T and η distributions of the combined b+bb category; bins were chosen to be mostly flat in p_T and η. The model was trained and exported using the weaver framework.[68] As Figures 7.1 demonstrates, ParticleNet significantly outperforms the standard DeepJet neural net on both b and c tagging, reaching

50% charm tagging efficiency against light and gluon jets at a mistag rate of 10%. The performance improvement is particularly large at high p_T.

Table 7.11: Definition of truth-level jet categories used for ParticleNet training.

Label	Pileup	Flavor	Light/Gluon
label_b	genjet_pt > 0	nBHadrons == 1	-
label_bb	""	nBHadrons > 1	-
label_c	""	nBHadrons == 0 && nCHadrons == 1	-
label_cc	""	nBHadrons == 0 && nCHadrons >= 1	-
label_uds	""	hadronFlavour == 0	|partonFlavour| == 1,2,3
label_g	""	hadronFlavour == 0	partonFlavour == 21
label_undef	""	hadronFlavour == 0	partonFlavour == 0
label_pu	genjet_pt $\leq$0	nBHadrons == 1	-

After jets have been identified and tagged, all events in the SL channel are required to have at least 5 jets. The DL channel requires at least 4 and the FH channel requires at least 6. Additionally, events must have at least 3 b or c jets that pass a "medium" working point. Any events failing these criteria are discarded; however, events with 4 jets are used for a validation region later.

7.3.4 Missing transverse energy

The raw missing transverse energy (MET) is defined as the negative vectorial sum of the p_T of all ParticleFlow candidates in the event. This value is then corrected per the recommendations in [69] for detector misalignment, calorimeter inefficiencies, and other effects; the result is used as an estimate of the true MET.

7.3.5 Data quality filters

In addition to the triggers and requirements mentioned above, all events must pass a number of standard data quality filters.[70] See Table 7.3.5 for a full list. Note that the

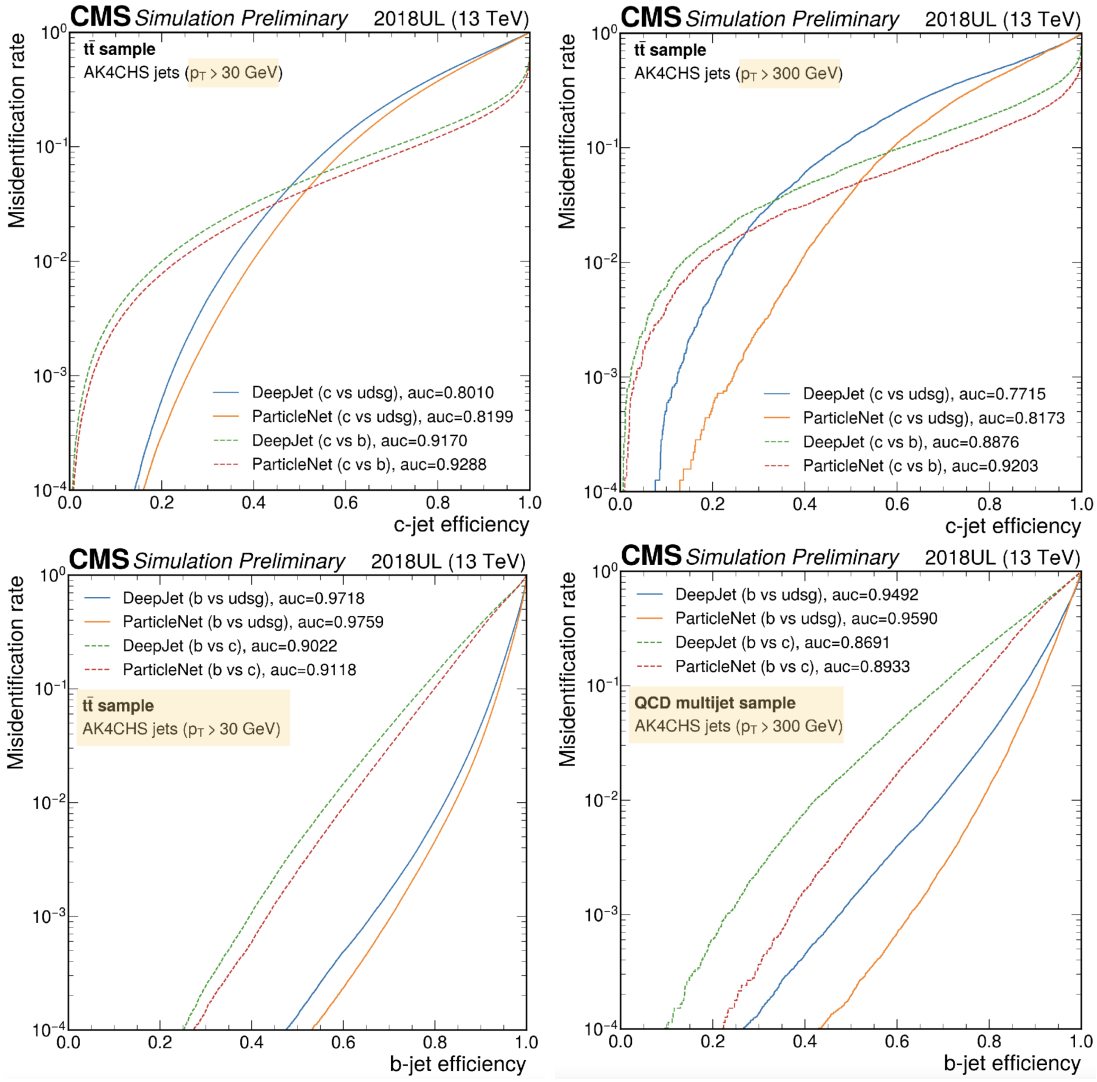


Figure 7.1: Top: ROC curves for ParticleNet's c-tagging performance above 30 GeV (left) and above 300 GeV (right). Bottom: ROC curves for ParticleNet's b-tagging performance. ParticleNet outperforms the DeepJet standard at all working points, particularly at high p_T. Plots by Huilin Qu.

same filters apply to all years with the exception of the EE bad calibration filter, which is only used for 2017 and 2018 samples.

Table 7.12: MET filters used for data quality.

Filter	NanoAOD branch
Primary vertex filter	`Flag_goodVertices`
Beam halo filter	`Flag_globalSuperTightHalo2016Filter`
HBHE noise filter	`Flag_HBHENoiseFilter`
HBHEiso noise filter	`Flab_HBHENoiseIsoFilter`
ECAL trigger primitive filter	`Flag_EcalDeadCellTriggerPrimitiveFilter`
Bad PF muon filter	`Flag_BadPFMuonFilter`
EE badSC noise filter (data only)	`Flag_eeBadScFilter`
EE bad calibration filter (2017-18 only)	`Flag_ecalBadCalibFilterV2`

7.3.6 The HEM15/16 issue

In mid-2018, the power supply unit for the HEM 15 and 16 regions of the HCAL Endcap encountered a rare, fatal failure mode. The problem resulted in voltages spikes that damaged the front-end readout, and from June 30th on (i.e. runs 2018C and 2018D), HEM 15 and 16 were rendered non-functional. This lowered the total coverage of the HCAL by 3%, resulting in erroneously low jet energies and a higher electron mistag rate in the affected region.[71] To mitigate the effects of this HEM15/16 issue, events in the single-lepton channel are vetoed if either the largest Higgs candidate jet or the electron falls in the affected HCAL region.

7.4 MC corrections

Even after ensuring that all events satisfy the above criteria, significant data-MC differences remain. Although some are due to the imperfect nature of MC simulations, many are caused by known issues with the CMS detector or inefficiencies in data analysis algorithms. Several corrections are applied to MC samples to compensate.

7.4.1 Top p_T reweighting

In Runs 1 and 2, both CMS and ATLAS discovered that the top p_T spectrum for $t\bar{t}$ events was softer than expected. Although part of this discrepancy was eliminated by applying NNLO and higher-order corrections to the event generators, a significant difference still remains and is not currently understood. Consequently, the TOP PAG group regularly calculates reweighting functions for $t\bar{t}$ that take this difference into account. These take the form of scale factors $\mathrm{SF}(p_T)$ that are a function of top p_T; the weight for each event used in this analysis is $\sqrt{\mathrm{SF(t)SF}(\bar{\mathrm{t}})}$.[72]

7.4.2 L1 prefiring adjustment

During 2016 and 2017, the ECAL underwent a gradual timing shift that was not taken into account by the L1 trigger. Portions of some events with a large amount of energy deposited in the ECAL region $2.0 < |\eta| < 3.0$ consequently spilled over into the following bunch crossing ("prefiring"). Because the L1 trigger is set up to reject events from consecutive crossings, these events were rejected. To account for this, weights corresponding to the probability of the prefiring issue occurring for a given event were added to all affected data.[73]

7.4.3 Trigger scale factors

Trigger efficiency in CMS data typically varies with a number of factors, including p_T, η, and detector conditions for each year. These effects can be quite large and are often not well-modeled by MC, so one must compute and apply trigger scale factors to compensate for this. For the SL channel, scale factors for the single-muon triggers are provided by the muon Physics Object Group (POG).[74] However, scale factors for the single-electron triggers were computed manually. For a full description of the electron

trigger scale factor calculation and results, see Appendix A.

7.4.4 Lepton scale factors

Additional scale factors are necessary to account for broad-ranging discrepancies in data and MC lepton behavior. They are provided centrally by the corresponding POGs.[75] As is the case with the trigger scale factors, the lepton scale factors are binned in p_T and η for each year of Run 2, and are computed independently for electrons and muons.

7.4.5 Flavor tagging scale factors

The final set of scale factors is used to correct the ParticleNet flavor tagging results, ensuring that the jet-tagging methods are accurate. To make calculating these SFs easier, all jets are split into 11 distinct categories based on their combined b and c scores s_{b+c} and their b vs c score $s_{b/c}$:

$$s_{b+c} = \frac{s_b + s_{bb} + s_c + s_{cc}}{s_b + s_{bb} + s_c + s_{cc} + s_{LF}} \tag{7.1}$$

$$s_{b/c} = \frac{s_b + s_{bb}}{s_b + s_{bb} + s_c + s_{cc}} \tag{7.2}$$

SFs for each of these 11 working points are derived independently of each other. See Figure 7.3 for sample cut efficiencies at each working point and for a graphical representation of the category definitions.

The final scale factors were derived using three data samples enriched in each of the desired jet flavors: $Z\ell\ell$+1 jet for light flavor-enriched, $W\ell\nu$ (requiring a soft muon in the jet) for charm-enriched, and $t\bar{t} \to e\mu$+2 jets for bottom-enriched. The final values

	WP	eff(b)	eff(c)	eff(light)
b-tagging	B4	**0.4126**	0.0007	0.0001
	B3	**0.1156**	0.0026	0.0002
	B2	**0.1041**	0.0087	0.0006
	B1	**0.0980**	0.0272	0.0020
	B0	**0.0861**	0.0757	0.0043
c-tagging	C4	0.0020	**0.0750**	0.0016
	C3	0.0125	**0.1419**	0.0037
	C2	0.0465	**0.1456**	0.0055
	C1	0.0563	**0.1651**	0.0712
	C0	0.0332	**0.1442**	0.1647
untagged	L0	0.0331	0.2134	0.7461

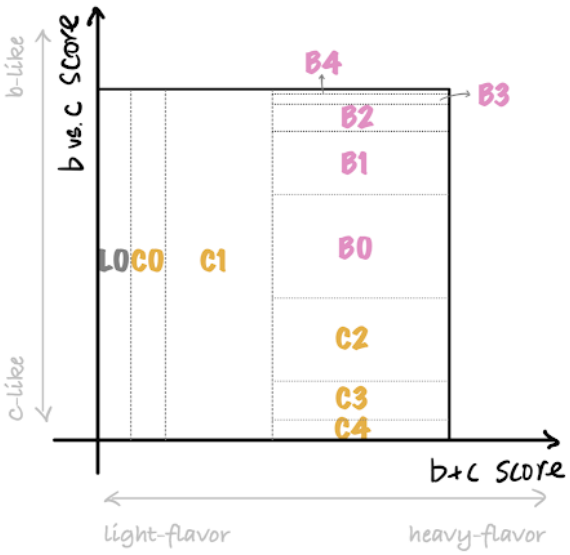


Figure 7.2: Left: All working points for the PN and their efficiencies in each jet category. Right: Definitions for each of the working points in graphical format. Plots by Huilin Qu.

were then determined as a function of jet p_T by a fit; any p_T bins where the SFs could not reliably be constrained by the fit were automatically assigned a value of 1 with a maximal uncertainty of (0.3,3). Figure 7.3 shows the results for each flavor category.

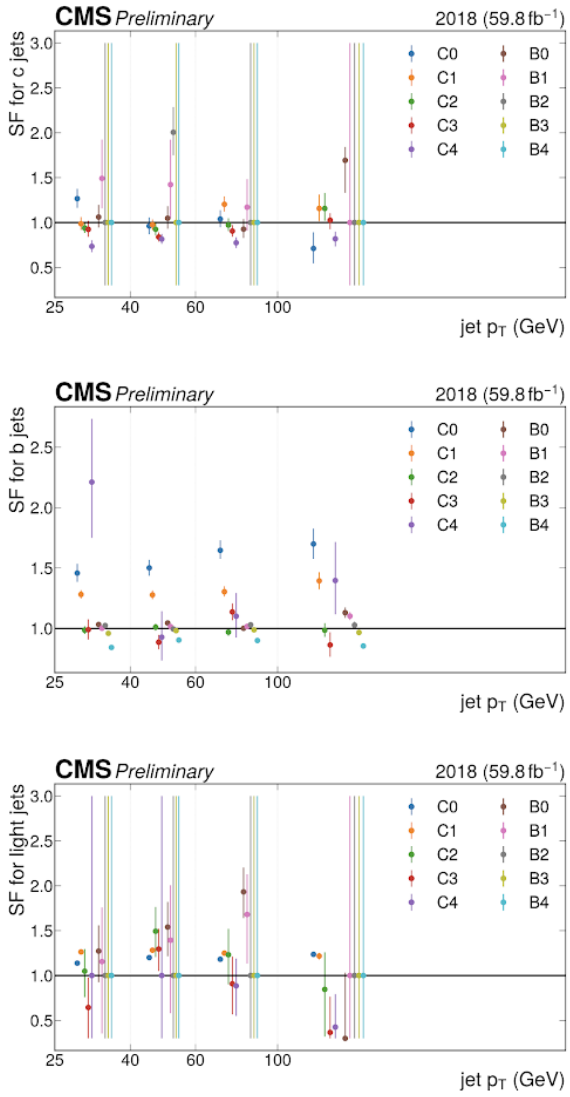


Figure 7.3: Flavor tagging scale factors for the trained ParticleNet model. Plot by Huilin Qu.

Chapter 8

Background Estimation

The extremely small cross-section of the $t\bar{t}Hc\bar{c}$ process necessitates a highly robust background estimation strategy. As discussed previously in chapter 4, this cannot be done using Monte Carlo simulations alone; the difficulty of modeling background processes involving QCD and the weak nature of the signal means that the background estimation method must be fundamentally data-driven.

This chapter begins with an overview of the major backgrounds in the single-lepton channel in section 8.1. The background estimation method is described in two parts: Event classification via Particle Transformer in section 8.2 and signal extraction via a maximum likelihood fit in 8.3. Finally, section 8.4 explains how the approach is validated by examining similar but orthogonal regions of phase space.

8.1 Background processes

As a relatively light quark, charms couple weakly to the Higgs, leading to a very low SM branching ratio of 2.86×10^{-2}. In comparison, the SM likelihood of a Higgs decaying to a pair of b quarks is 5.76×10^{-1}, over thirty times higher.[76] Moreover,

the SM gg→t$\bar{t}$H cross-section is also small, on the order of 0.085 pb.[77] This makes the t$\bar{t}$H(H→ c$\bar{c}$) signal easy for other processes to obscure: After the trigger and initial Particle Transformer selections have been applied, t$\bar{t}$Hc$\bar{c}$ make up only 0.08% and 0.09% of all surviving events in the single-electron and single-muon streams, respectively. See figure 8.4 for a full comparison with other backgrounds.

The largest backgrounds are processes of the form t$\bar{t}$+q$\bar{q}$, where q is any lighter flavor of quark. These can be split into five different categories:

1. **t$\bar{t}$ + light flavor (LF):** Here q is one of the light flavors (u, d, s) or a gluon. Before the Particle Transformer selections described below, this is by far the most common background process.

2. **t$\bar{t}$+bb:** In this case, a low-p_T b$\bar{b}$ pair is produced in conjunction with the tops, leading to two clearly resolved b jets.

3. **t$\bar{t}$+bj:** Similar to the above, except the produced pair of b quarks is heavily boosted. In this case, the two resulting jets overlap, producing a single merged jet. This event topology is different enough that t$\bar{t}$+b events are handled as a separate category from t$\bar{t}$+b$\bar{b}$.

4. **t$\bar{t}$+cc:** Similar to 2, except with two c quarks.

5. **t$\bar{t}$+cj:** Similar to 3, except with two boosted c quarks.

The sixth-most common background is single-top, i.e. a single top quark produced in conjunction with lighter quarks or a W boson. Another concern is t$\bar{t}$H(H→ b$\bar{b}$), which is identical to the signal process except with the charm quarks replaced with bottoms. The remaining backgrounds each make up a small fraction of the total; t$\bar{t}$Z(Z→ q$\bar{q}$) is the most notable due to its kinematic similarity t$\bar{t}$H(H→ c$\bar{c}$). Feynman diagrams for these key backgrounds are shown in Figure 8.1.

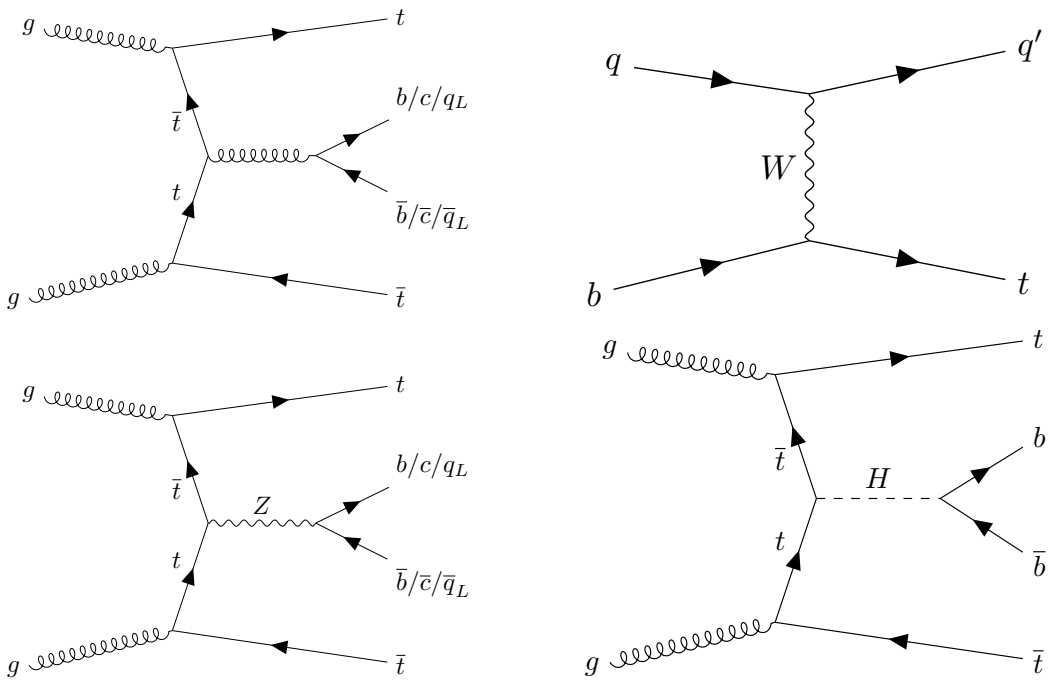


Figure 8.1: Feynman diagrams for the $t\bar{t}b\bar{b}/c\bar{c}$/LF (top left), single-top (top right), $t\bar{t}Zb\bar{b}/c\bar{c}$ (bottom left), and $t\bar{t}Hb\bar{b}$ (bottom right) background processes. Note that the single-top diagram is only one of five common production modes at the LHC.

It is worth noting that this is not necessarily a complete list of backgrounds. In future stages of the analysis, small effects from processes such as W+jets, Z+jets, tH+W, and tH+q may also be taken into account. However, these contributions to the overall background are expected to be insignificant, and they will not be investigated until the analysis is in a later stage of readiness.

8.2 Event classification with Particle Transformer

The large and complex backgrounds in this analysis make the need for a powerful event classifier clear, a role which is filled by a Particle Transformer model as described in section 5.3. In addition to the standard coordinates, the Particle Transformer model

takes $\ln p_T$, $\ln E$, and η for the single lepton and each jet as input, as well as $\ln E_{\text{MET}}$. The lepton is assigned an extra "isEl/isMu" flag to differentiate the two different streams. Event with at most eight jets are used during training; each jet is paired with eight flags indicating which ParticleNet b- and c-tagging score thresholds the jet passes. Events in the training sample are weighted by cross-section, then renormalized such that the sum of the weights in each category is equal. A total of ten event categories are used during the training, one for each of the following processes: The five $t\bar{t}$ backgrounds listed in the previous section; the three Z backgrounds $t\bar{t}Z(Z\rightarrow c\bar{c})$, $t\bar{t}Z(Z\rightarrow b\bar{b})$, and $t\bar{t}Z(Z\rightarrow q\bar{q})$; the $t\bar{t}H(H\rightarrow b\bar{b})$ background, and finally the $t\bar{t}H(H\rightarrow c\bar{c})$ signal process. The trained model assigns ten scores that reflect the probability of an event falling into each category, but since the scores must sum to 1, the result is nine independent Particle Transformer scores.

The natural next step is to categorize all events based on their Particle Transformer scores. However, the simplest method of doing this—sorting each event into the category corresponding to its largest score—runs into a complication: The normalization of tt+bb in each category differs between the 4FS $t\bar{t}b\bar{b}$ model and the unmodified 5FS $t\bar{t}$ model. As shown in Figure 8.2, the 5FS model predicts significantly fewer tt+bb events in the ttH and ttZ categories. Since the preference for the merged 4FS sample is only a general *a priori* preference, and since any difference in normalization could potentially bias the event count in the signal region, the safest approach is to take steps to minimize the differences between the 4FS and 5FS predictions. Moreover, Figure 8.3 illustrates that these normalization differences between the samples can be largely attributed to non-signal-like events. This suggests that imposing a cut on the combined signal score and defining all control and signal regions above this cut will bring the discrepancy down to more acceptable levels. A cut of `score_ttH+score_ttZ` > 0.6 was ultimately chosen to solve the problem while keeping statistics in the signal regions high.

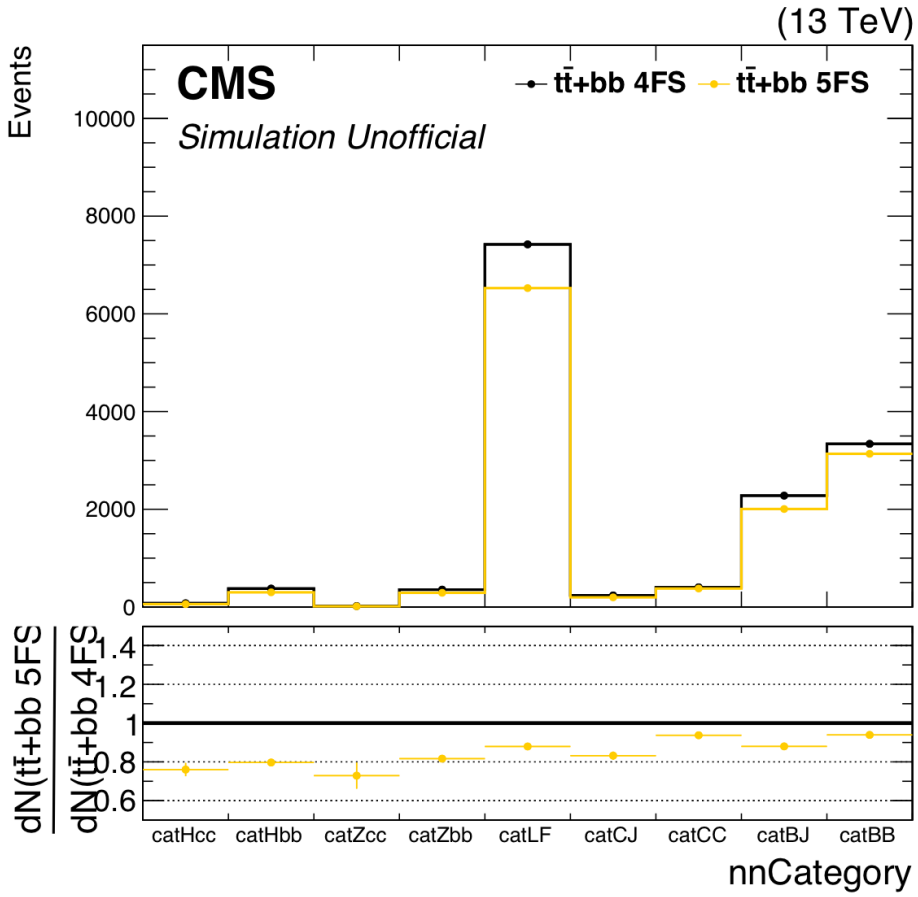


Figure 8.2: Event counts per category for the 2018 4FS and 5FS tt+bb samples in the SL channel, using the simple categorization scheme. A significant difference in normalization is visible between the four leftmost signal categories and the remaining background categories. (The same effect is present in other years; these are not shown for the sake of brevity.)

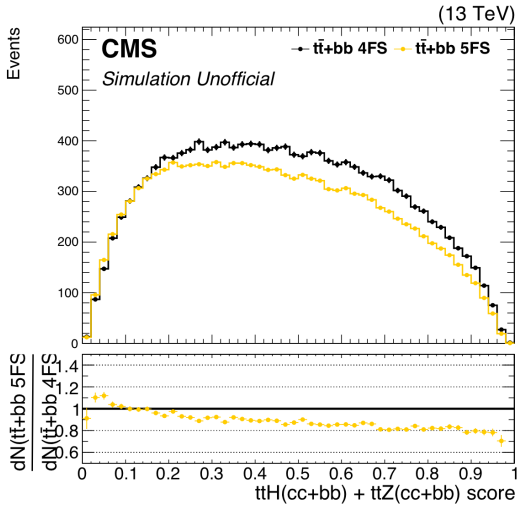


Figure 8.3: Plot comparing the combined signal score `score_ttHcc+score_ttHbb+score_ttZcc+score_ttZbb` for 2018 tt+bb events between the 4FS model and the 5FS model in the SL channel. A significant shape difference is present at low scores; cutting at 0.6 removes it at low cost in signal.

Another obstacle encountered by the simple categorization method is low purities in each category. This problem is solved in two steps. First, due to an excess of light flavor events in all categories, an additional cut of `score_ttLF` is introduced. The post-selection event distribution in each stream is shown in Figure 8.4; the choice of cuts removes most of the light flavor events, leaving reasonably large quantities of each of the major $t\bar{t}$ backgrounds. Second, the category definitions are refined: All signal events must pass a score cut of (`score_ttHcc+score_ttHbb+score_ttZcc+score_ttZbb`)/(1 − `score_ttZqq`) > 0.85, and the control region scores are weighted to filter out $t\bar{t}$+c events more strongly. The full categorization scheme is summarized schematically in Figure 8.5.

The 2018 MC event counts per category, as well as the purities of each category, are shown in Figure 8.7. The final category definitions reach a compromise between high purity and high statistics. Additionally, the 4FS-5FS normalization difference is greatly reduced, as shown in Figure 8.6. After categorization, the Particle Transformer scores for each category are renormalized to fall between 0 and 1; these renormalized scores

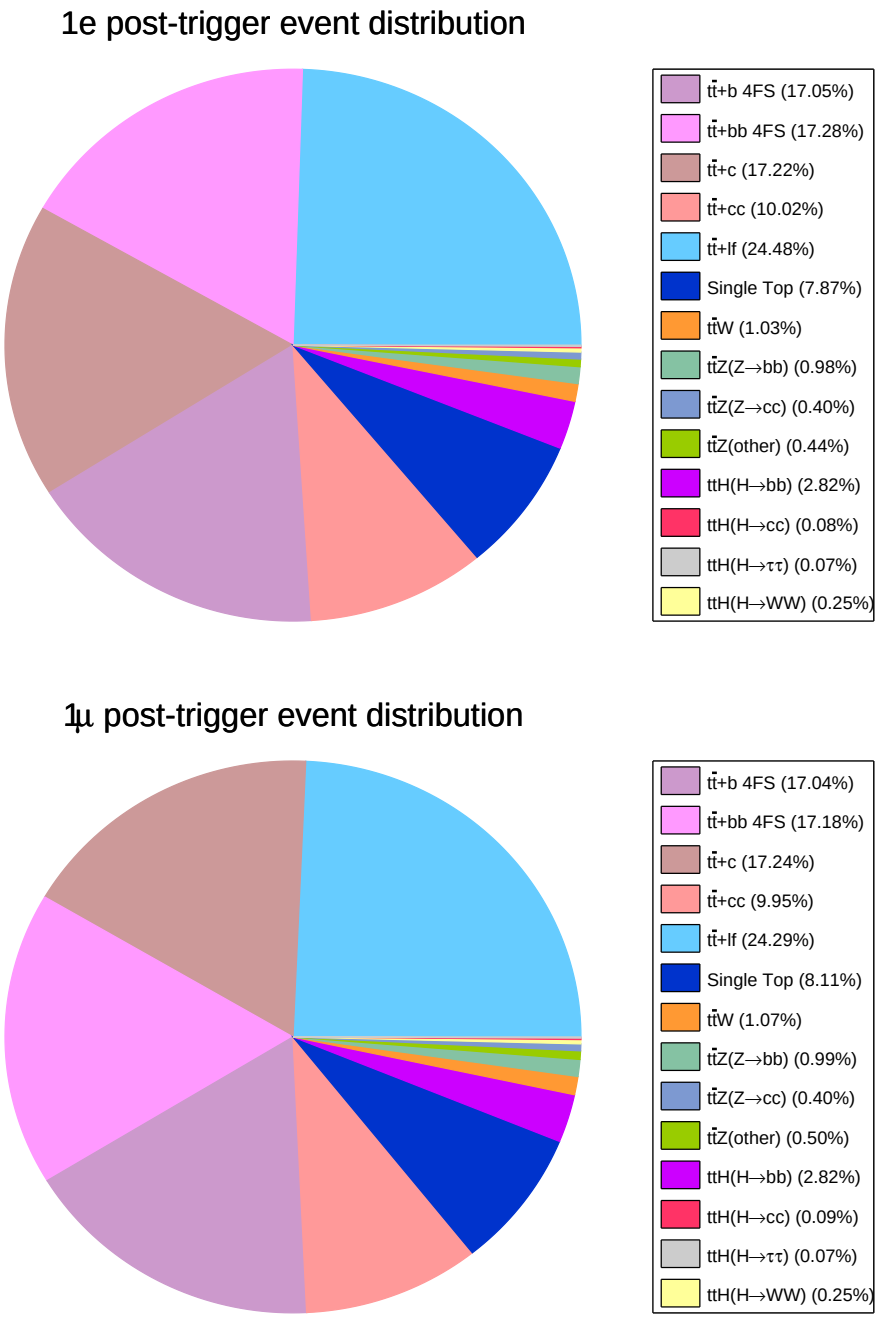


Figure 8.4: Plots showing the 2018 MC post-trigger, post-Particle Transformer-selection fractions of each event type in the single-electron (left) and single-muon (right) streams.

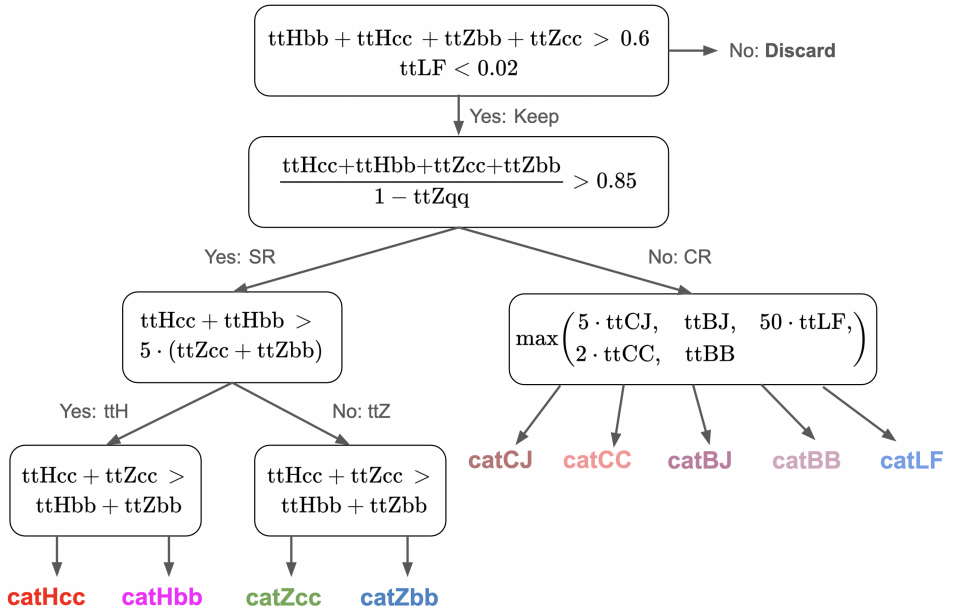


Figure 8.5: Flow diagram showing the categorization scheme used in this analysis. Process names are used as a shorthand for their corresponding scores, e.g. "ttHcc" for the $t\bar{t}Hc\bar{c}$ score. The merged-jet processes tt+b and tt+c are referred to as "ttBJ" and "ttCJ" to keep them distinct.

are used in the final fit. MC plots of the adjusted scores for the single-electron and single-muon streams are displayed in Figures 8.6 and 8.5, respectively.

8.3 Background estimation and signal extraction

After all surviving events have been categorized, a binned maximum likelihood fit is used to extract rate parameters for each process. The categories described above are split into signal and control regions: The catHcc, catHbb, catZcc, and catZbb categories are designated signal regions, and all remaining categories (all $t\bar{t}$ backgrounds) are designated control regions. Template histograms for each background process are generated using Monte Carlo samples following the procedure in 5.1.3, and their Particle Transformer score distributions for each category are used as the model for the binned maximum likelihood fit. The fit is performed using the combine tool. The normalization parameters for

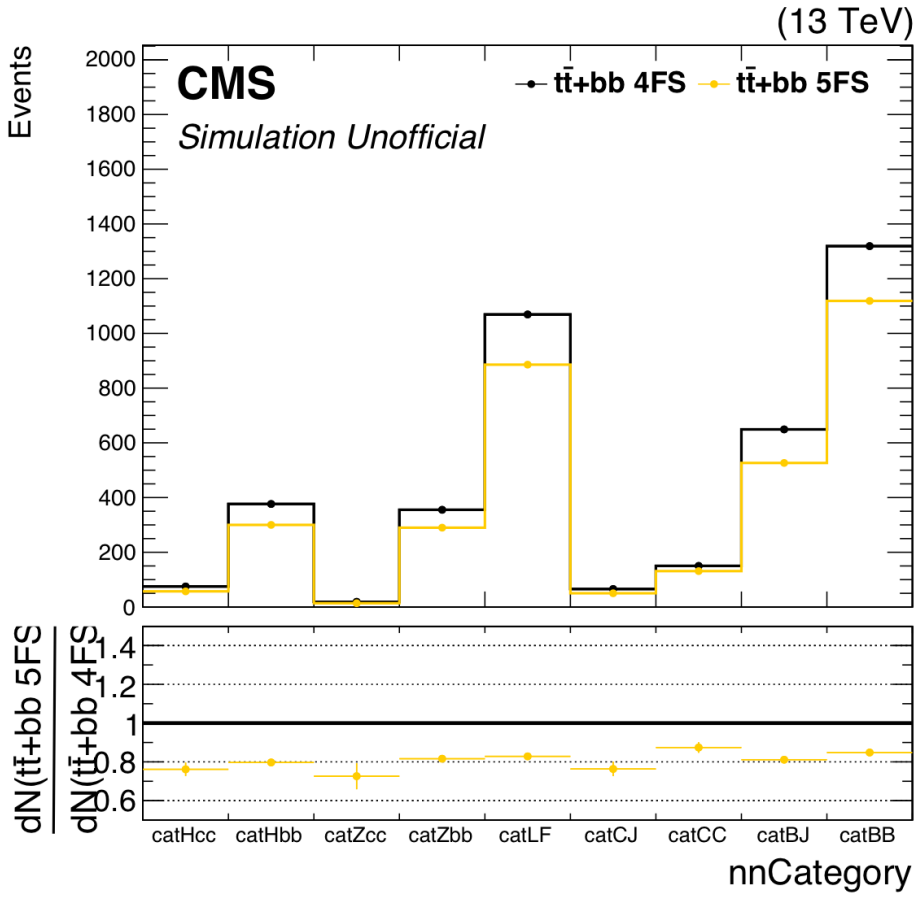


Figure 8.6: Event counts per category for the 2018 4FS and 5FS tt+bb samples, using the finalized categorization scheme. The signal and control regions are now much closer in terms of normalization.

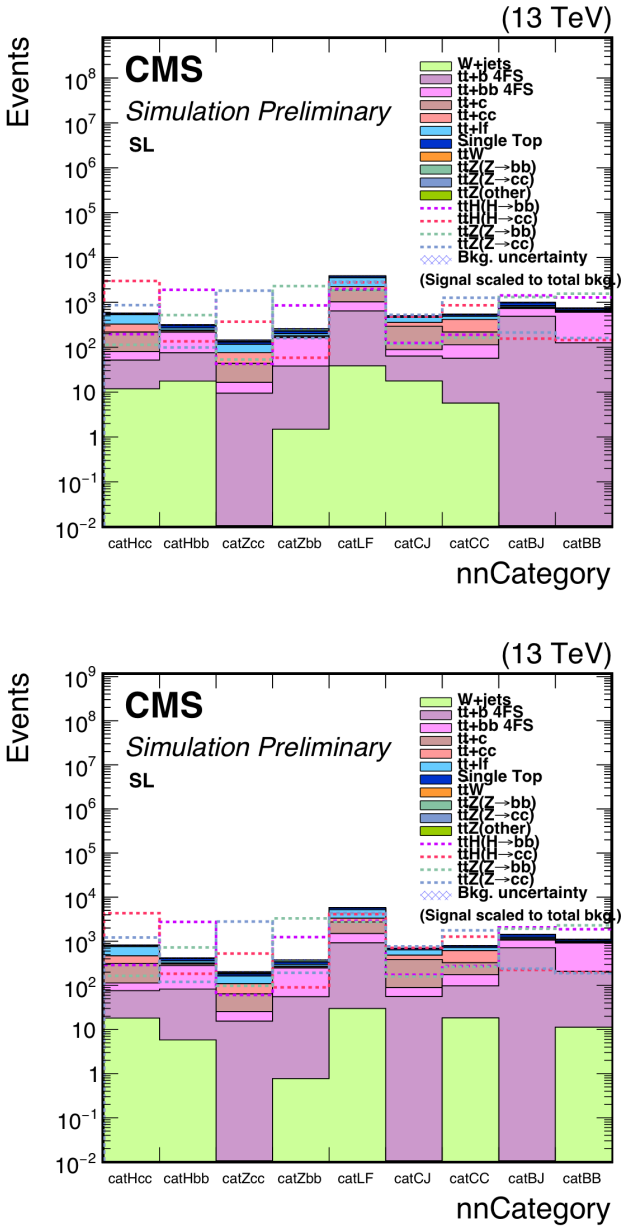
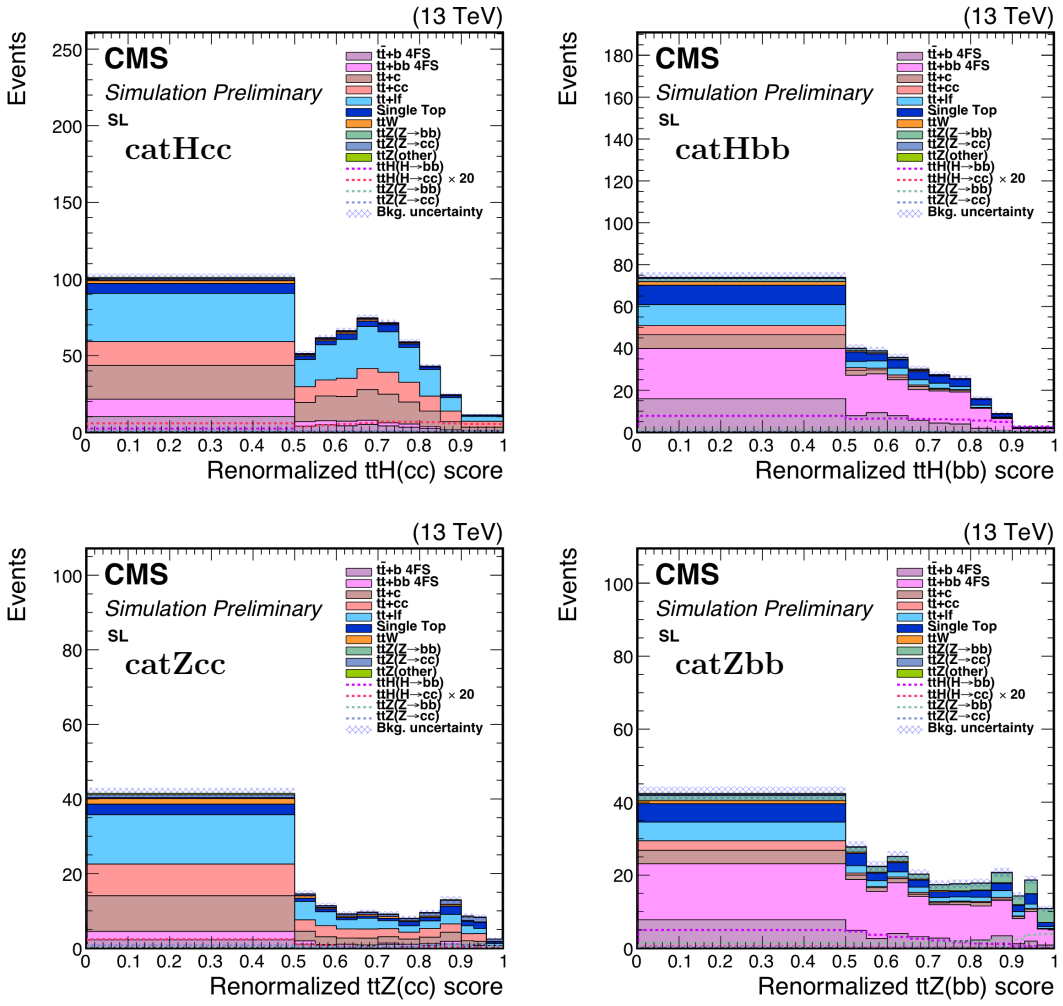
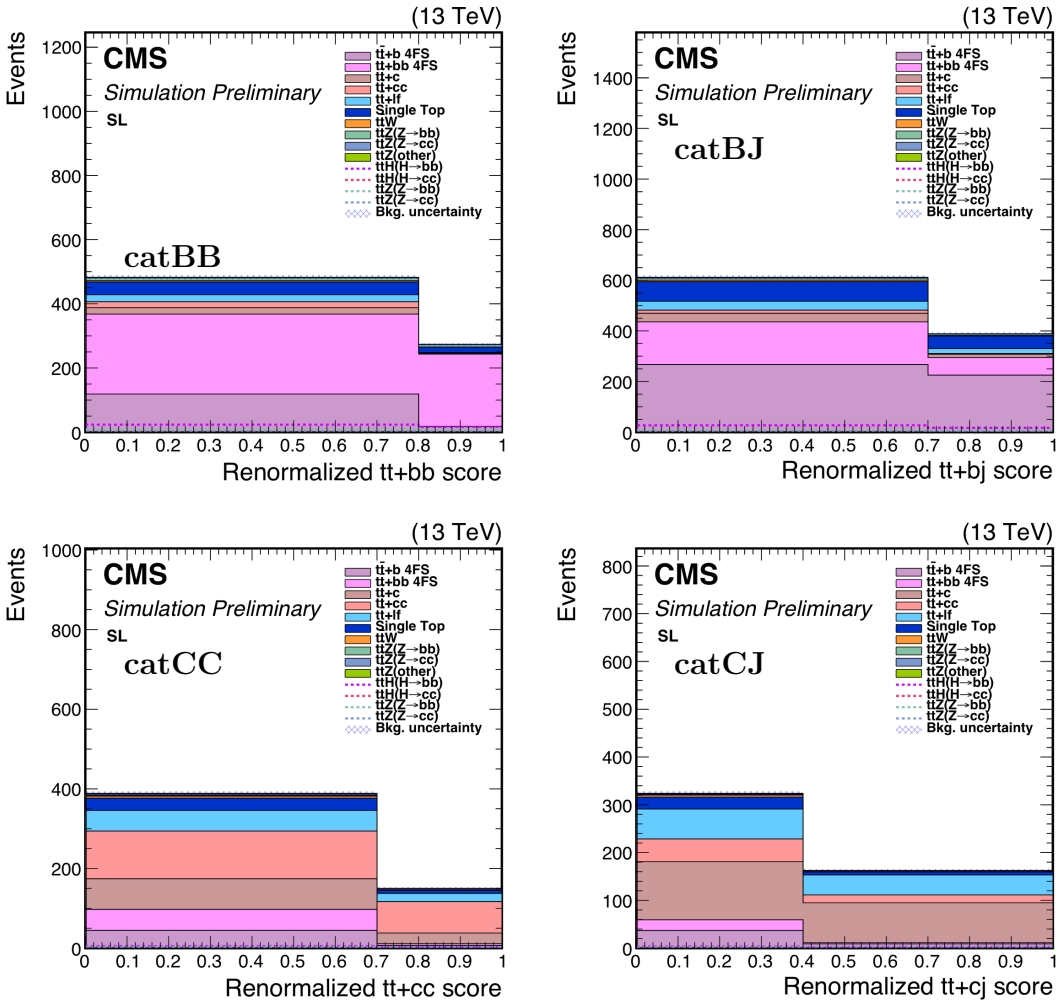


Figure 8.7: Histograms showing the composition of each event category for the single-electron (left) and single-muon (right) streams.


(13 TeV)
CMS
Simulation Preliminary
SL
catHcc
Events
Renormalized ttH(cc) score
(13 TeV)
CMS
Simulation Preliminary
SL
catHbb
Events
Renormalized ttH(bb) score
(13 TeV)
CMS
Simulation Preliminary
SL
catZcc
Events
Renormalized ttZ(cc) score
(13 TeV)
CMS
Simulation Preliminary
SL
catZbb
Events
Renormalized ttZ(bb) score
tt+b 4FS
tt+bb 4FS
tt+c
tt+cc
tt+lf
Single Top
ttW
ttZ(Z→bb)
ttZ(Z→cc)
ttZ(other)
ttH(H→bb)
ttH(H→cc) × 20
ttZ(Z→bb)
ttZ(Z→cc)
Bkg. uncertainty


(13 TeV)
CMS
Simulation Preliminary
SL
catBB
Events
Renormalized tt+bb score
catBJ
Renormalized tt+bj score
catCC
Renormalized tt+cc score
catCJ
Renormalized tt+cj score
tt+b 4FS
tt+bb 4FS
tt+c
tt+cc
tt+lf
Single Top
ttW
ttZ(Z→bb)
ttZ(Z→cc)
ttZ(other)
ttH(H→bb)
ttH(H→cc)
ttZ(Z→bb)
ttZ(Z→cc)
Bkg. uncertainty

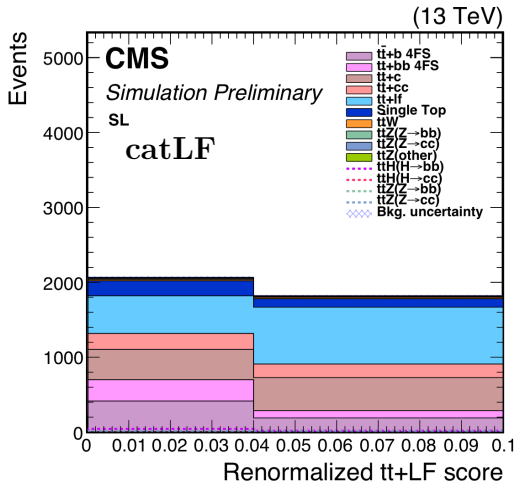


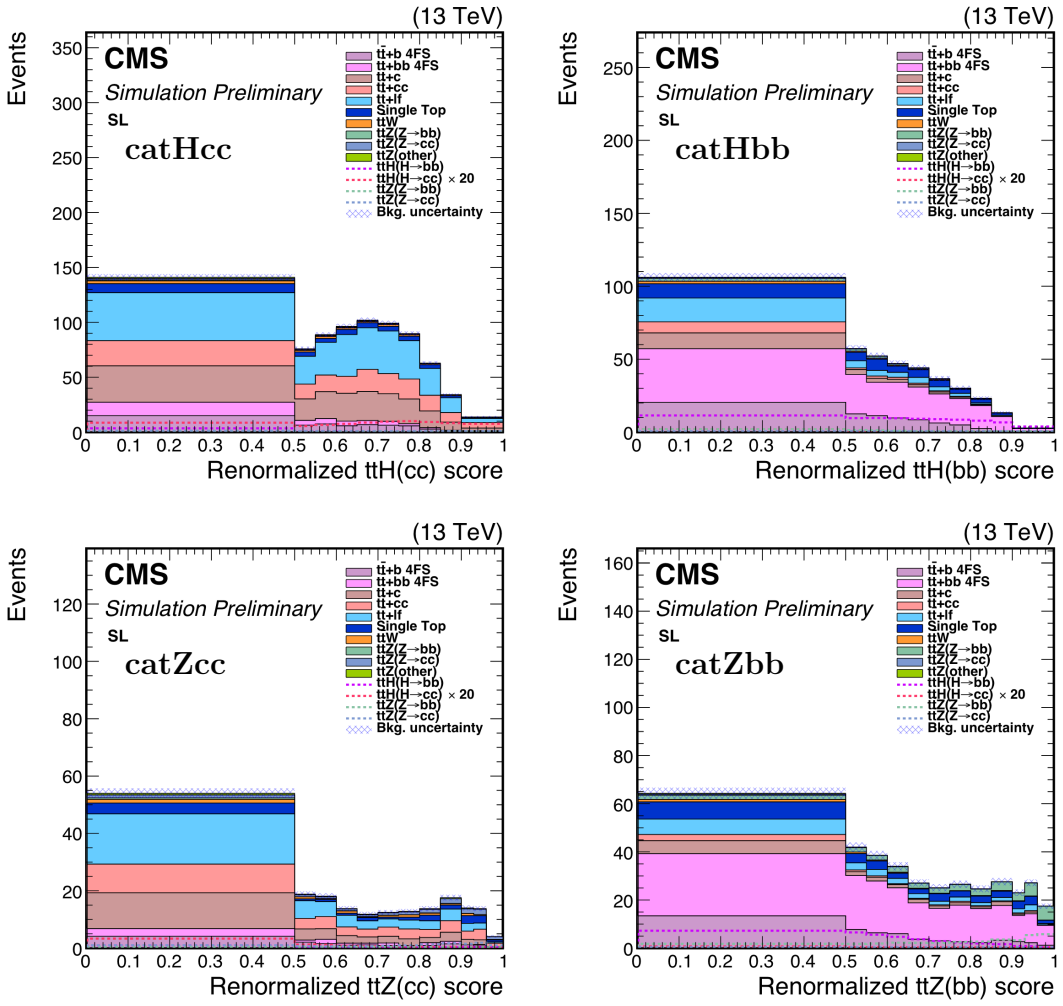
Figure 8.6: Histograms showing the distribution of renormalized Particle Transformer scores for 2018 single-electron MC events. Only events in the category corresponding to each score are included; e.g. only events in the ttH(cc) category are shown in the renormalized ttH(cc) score plot.

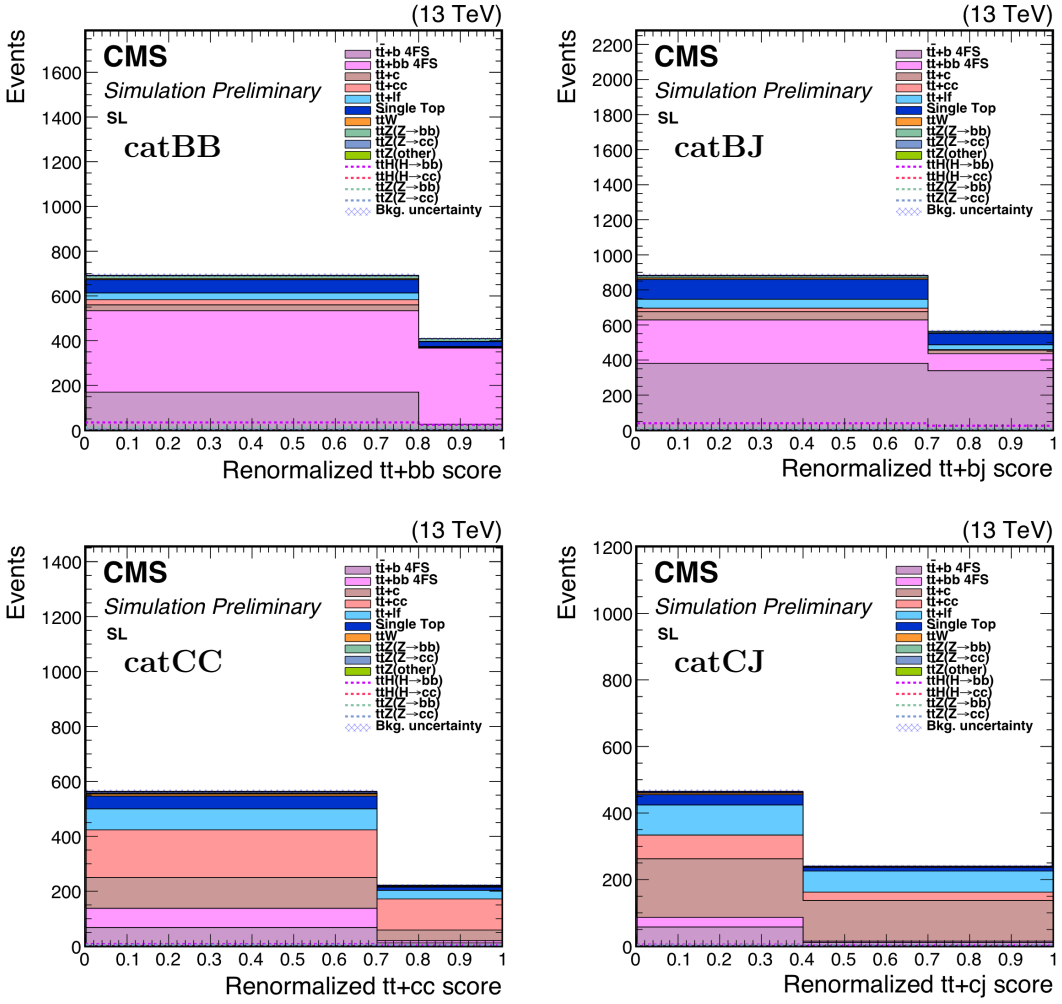
the nine key processes are left floating; the key parameter of interest is the normalization of the $t\bar{t}Hc\bar{c}$ signal, and the remaining normalization parameters are nuisances.

8.4 Validating the background estimation approach

To validate our methods in the single-lepton channel, a validation region is defined as close to the signal and control regions as possible. Events in the validation region must pass the usual criteria of $\geq$5 jets and $\geq$3 b or c jets. They must also pass a renormalized LF score cut of 0.02, and have a renormalized signal score between 0.4 and 0.6 (compared with a raw signal score of >0.6 for the control and signal regions). The precise definition is listed in table 8.4 In this validation region, a binned maximum-likelihood fit is performed following the method described in the previous section.

A 4-jet validation region was considered, but ultimately discarded for two reasons. First, the 4-jet phase space was deemed too kinematically different from the standard


(13 TeV)
CMS
Simulation Preliminary
SL
catHcc
Events
Renormalized ttH(cc) score
(13 TeV)
CMS
Simulation Preliminary
SL
catHbb
Events
Renormalized ttH(bb) score
(13 TeV)
CMS
Simulation Preliminary
SL
catZcc
Events
Renormalized ttZ(cc) score
(13 TeV)
CMS
Simulation Preliminary
SL
catZbb
Events
Renormalized ttZ(bb) score
tt̄+b 4FS
tt+bb 4FS
tt+c
tt+cc
tt+lf
Single Top
ttW
ttZ(Z→bb)
ttZ(Z→cc)
ttZ(other)
ttH(H→bb)
ttH(H→cc) × 20
ttZ(Z→bb)
ttZ(Z→cc)
Bkg. uncertainty


(13 TeV)
CMS
Simulation Preliminary
SL
catBB
Events
Renormalized tt+bb score
(13 TeV)
CMS
Simulation Preliminary
SL
catBJ
Events
Renormalized tt+bj score
(13 TeV)
CMS
Simulation Preliminary
SL
catCC
Events
Renormalized tt+cc score
(13 TeV)
CMS
Simulation Preliminary
SL
catCJ
Events
Renormalized tt+cj score
tt+b 4FS
tt+bb 4FS
tt+c
tt+cc
tt+lf
Single Top
ttW
ttZ(Z→bb)
ttZ(Z→cc)
ttZ(other)
ttH(H→bb)
ttH(H→cc)
ttZ(Z→bb)
ttZ(Z→cc)
Bkg. uncertainty

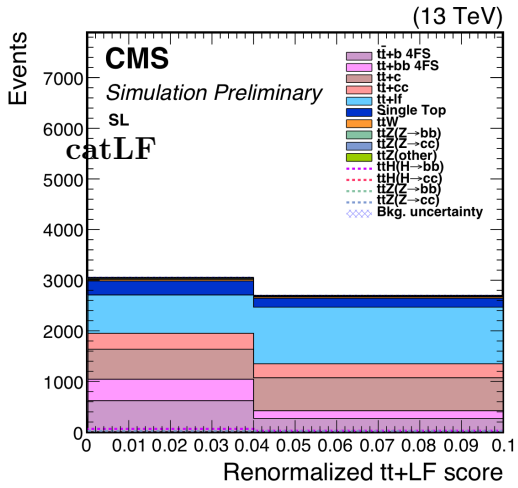


Figure 8.5: Histograms showing the distribution of renormalized Particle Transformer scores for 2018 single-muon MC events, using the same selections as in Figure 8.5.

≥ 5 jet selection; for instance, at least one jet must have been missed by the calorimeters for the majority of these events. Second, researchers working on a very similar, near-publication ttH(H→bb) analysis (HIG-19-011) discovered that $t\bar{t}$ backgrounds in the 4-jet region were especially poorly modeled, and had significantly different rate parameters for each background process in comparison to the $\geq$ 5-jet validation and signal regions. In line with their recommendations, we have opted to focus only on the 5-jet validation region for the time being.

Figure 8.6 and plots the 2018 single-electron MC shape predictions against data in the validation region. Although the normalization of some of the background processes differs significantly from MC predictions, such as in the catBB and catBJ regions, this is not unexpected given the the difficulty of simulating $t\bar{t}b\bar{b}$ backgrounds. See Figure 8.7 for the distributions after the fit has been performed and all MC event categories have been rescaled to match the best-fit rate parameters. No significant shape or normalization differences are found in any of the validation regions post-fit, corroborating the soundness of the background estimation method. A possibly-significant data-MC discrepancy

Table 8.1: Selection criteria used for the 5-jet and 4-jet validation regions, as well as the signal and control region criteria for comparison. All raw Particle Transformer scores are abbreviated by the name of their corresponding category (e.g. "ttHcc" for the catHcc score), and $N_{j,b|c}$ denotes the number of jets in an event that reach a medium b- or c-tag score threshold. All selection criteria not listed here apply to all three regions.

Region	Jet selection	Particle Transformer selection
Signal and control	$N_j \geq 5$, $N_{j,b\|c} \geq 3$	$(\text{ttHcc} + \text{ttHbb} + \text{ttZcc} + \text{ttZbb})/(1 - \text{ttZqq}) < 0.6$, $(\text{ttLF}/(1 - \text{ttZqq}) < 0.05$
5-jet validation	$N_j \geq 5$, $N_{j,b\|c} \geq 3$	$(\text{ttHcc} + \text{ttHbb} + \text{ttZcc} + \text{ttZbb})/(1 - \text{ttZqq}) > 0.4$, $(\text{ttHcc} + \text{ttHbb} + \text{ttZcc} + \text{ttZbb})/(1 - \text{ttZqq}) < 0.6$, $(ttLF/(1 - \text{ttZqq}) < 0.02$,
4-jet validation	$N_j = 4$, $N_{j,b\|c} \geq 3$	$(\text{ttHcc} + \text{ttHbb} + \text{ttZcc} + \text{ttZbb})/(1 - \text{ttZqq}) > 0.4$, $(\text{ttLF}/(1 - \text{ttZqq}) < 0.05$ $1.5 * \text{ttHbb}/(1 - \text{ttZqq} + 2 * \text{ttZcc} + 2 * \text{ttZbb}) > 0.95$, $6 * \text{ttZbb}/(1 - \text{ttZqq} + 2 * \text{ttZcc} + 2 * \text{ttZbb}) > 0.95$

is visible as an excess in the catBB category. Although it is possibly a non-issue–tt+bb backgrounds are known to be poorly modeled, and the aforementioned ttH(H→bb) analysis HIG-19-011 made a similar observation in their validation region–investigating this excess further is currently a high priority for this analysis, and it will be looked into thoroughly before proceeding to the pre-approval stage. (For the sake of readability, plots for other years and the muon stream have been omitted. The muon stream is very similar to the electron stream, and other years have no major anomalies compared to 2018.)

Please note that these results are preliminary and unofficial, and that the definitions of the validation regions may change in the final version of the analysis.

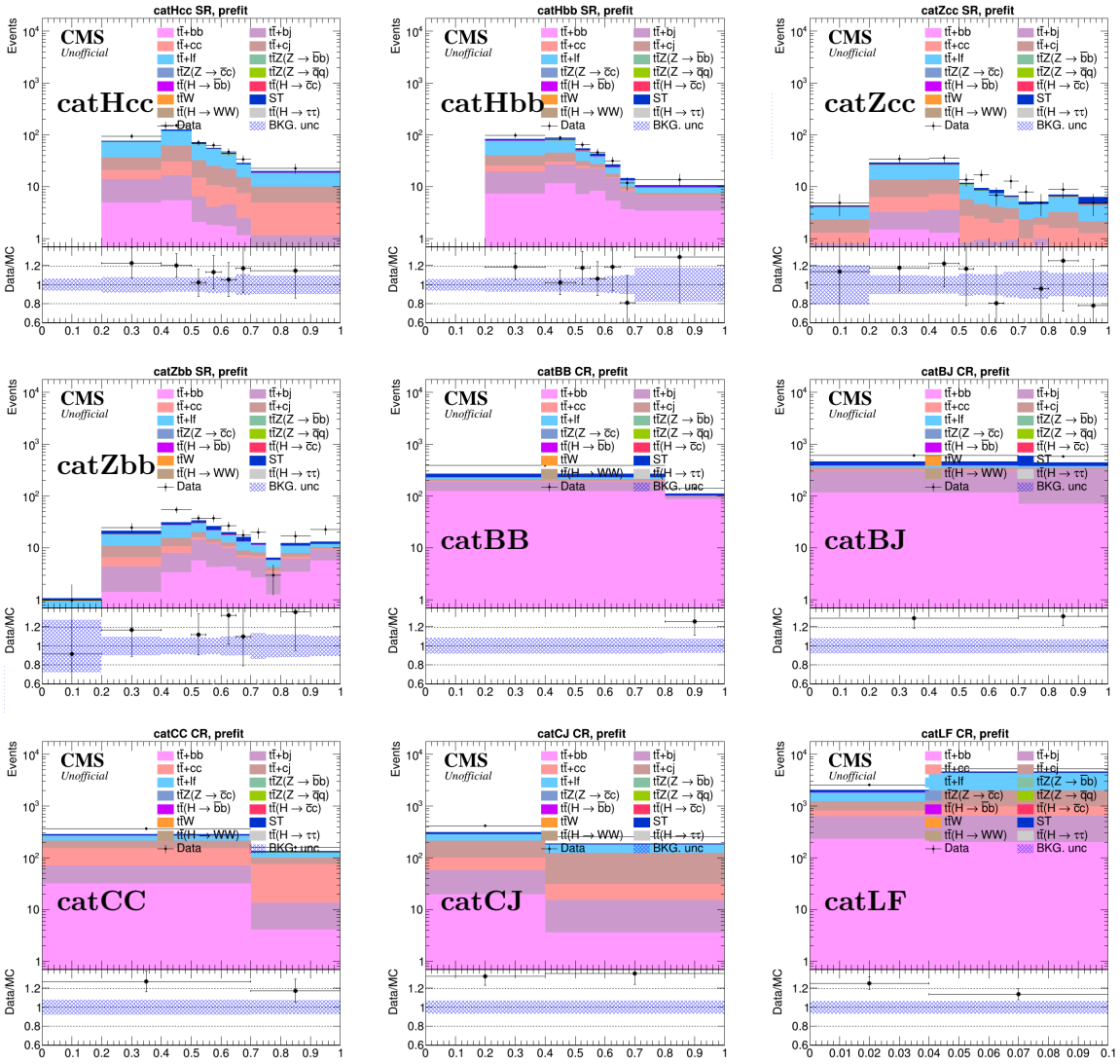


Figure 8.6: Data versus MC plots for 2018 single-electron events in the validation region, pre-fit.

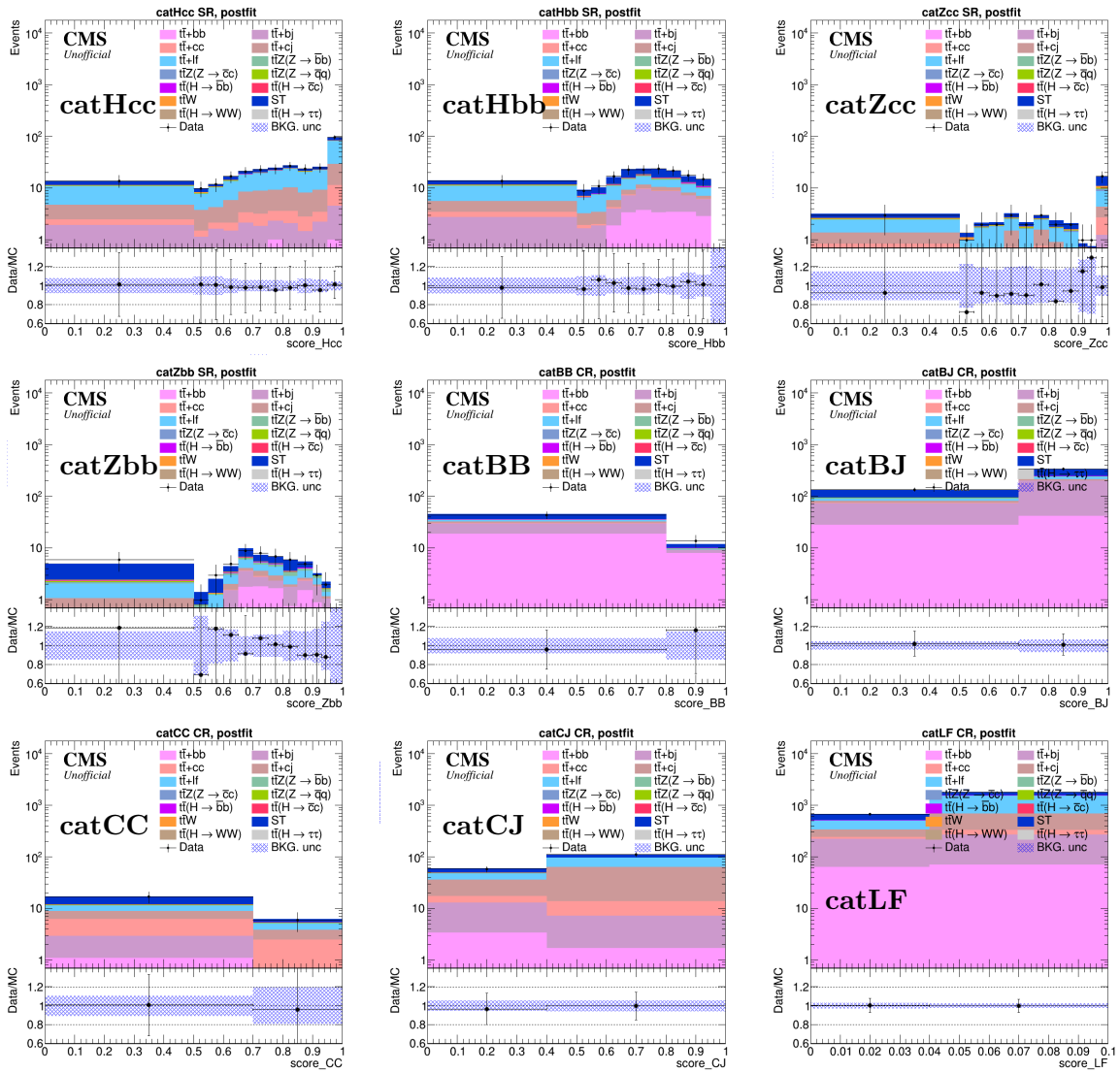


Figure 8.7: Data versus MC plots for 2018 single-electron events in the validation region, post-fit.

Chapter 9

Systematics

The $t\bar{t}H(h\rightarrow c\bar{c})$ analysis described in this thesis involves many steps, and each step introduces new sources of uncertainty. Many of these uncertainties affect theoretical predictions, such as variations in key parameters used in QCD calculations. Others are experimental in nature, such as imprecision in CMS luminosity measurements and issues caused by malfunctioning parts of the detector.

The following uncertainties are actively being accounted for in this analysis. With two exceptions (flavor tagging and JES systematics), all systematics in this section are in their final form.

- **Statistical:** The statistical uncertainties of each Particle Transformer score bin in data.

- **QCD scale:** Uncertainties in the QCD scale chosen for matrix element calculation in MC event generators. These are treated as rate uncertainties in the final fit, since they affect the overall normalization of each process (as opposed to shape uncertainties, which affect the shapes of variable distributions). These uncertainties are treated as correlated between processes that share the same uncertainty source,

as well as correlated between the three data-taking years.

- **Renormalization and factorization scale:** When performing calculations involving QCD, it is necessary to introduce two energy scales that affect the kinematics of MC simulations. The former, the renormalization scale μ_R, is needed to cancel out divergences in Feynman loop diagrams. It directly affects the strong coupling constant α_S. The latter, the factorization scale μ_F, draws a line between perturbative and non-perturbative calculations for hadron cross-sections; it impacts parton distribution functions. This analysis follows the standard recommendation to vary both parameters up and down by a factor of 2 (using different MC samples for each) and account for any changes with systematics.[78] These systematics are treated as uncorrelated among (5FS) $t\bar{t}H$ processes, (4FS) $t\bar{t}b\bar{b}$ processes, and other $t\bar{t}$ samples.

- **ISR and FSR:** Due to their dependence on α_S, QCD initial-state and final-state radiative processes are also affected by the choice of renormalization scale. As before, additional MC samples are generated after varying μ_R up and down by a factor of two, and they are used to derive systematic uncertainties.[79]

- **PDF shape uncertainty:** Although the PDF normalization uncertainties are covered by the above systematics, shape effects must also be taken int account. These uncertainties are computed by generating events with alternate PDF sets `NNPDF31_nnlo_as_0118_nf_4` (for the 4FS sample) and `NNPDF31_nnlo_hessian_pdfas` (for the 5FS samples) and comparing them against the baseline Particle Transformer score templates.[80] Systematics for 4FS and 5FS samples are treated as uncorrelated.

- **Luminosity:** The luminosity measured for each year of CMS data has an uncer-

tainty on the order of 1%. This is also included as a systematic.[81]

- **Lepton scale factors:** As mentioned in 7.4.4, p_T and η-dependent scale factors for lepton tracking and reconstruction efficiencies, as well as corresponding systematics, are provided by the MUO and EGM POGs.

- **Trigger scale factors:** The trigger scale factors discussed in section 7.4.3 are currently being used to correct event weights, and their uncertainties have been added to the workflow. See appendix A for details.

- $h_{\mathbf{damp}}$**:** The nominal value of h_{damp} for this analysis is 1.379. Systematics are calculated using up and down variations of 2.305 and 0.874. These systematics are treated as uncorrelated between the various $\mathrm{t\bar{t}}$ processes, and correlated per process between analysis categories and years.

- **L1 prefiring issue:** As discussed in section 7.4.2, this analysis uses CMS-provided scale factors to adjust for the L1 prefiring issue. These also come with uncertainties; they will be treated as uncorrelated between 2016 and 2017. (The prefiring was not present in 2018.)[73]

- **Jet energy corrections:** The jet energy scale and resolution are set by two parameters in the calorimeter reconstruction algorithms, JER and JES. For the former, samples with 1-sigma variations in JES are compared against the baseline templates to derive shape uncertainties.[82] For the latter, the total JES uncertainty—the envelope of all individual sources of uncertainty—is used as a systematic. While a total of 10 other sources of JES shape systematics exist and have not been added, the envelope systematic is the largest, and only fails to account for correlations between the remaining sources. The rest will only have a small effect on the overall uncertainties, and will be taken into account in the final version of the anslysis.[83]

- **MET corrections:** Uncertainties in the scale factors for the standard MET corrections are centrally provided, and accounted for via the recommended methods.[69]

- **Pileup:** The distribution of pileup interactions at CMS is also uncertain. The nominal cross-section used to predict this distribution is be varied by 4.6% up and down to yield fully-correlated rate and shape systematics.

- **Flavor tagging scale factors:** Finally, uncertainties in the flavor tagging scale factors discussed in section 7.4.5 must be included. These uncertainties are not yet finalized, and will be refined further in future versions of the analysis. In their present state, the flavor tagging systematics account for statistical uncertainties, variations in μ_F and μ_R, related variations in ISR and FSR, effects from the inclusion of W+jets and Z+jets samples, and pileup uncertainties.

Note that as this analysis is still in a preliminary state, the above systematics are subject to further refinement, and new sources of uncertainty may be accounted for as necessary.

Chapter 10

Preliminary Results

This chapter presents preliminary results for the $t\bar{t}H(H\to c\bar{c})$ search in the single-lepton channel. Data from the full CMS Run II dataset is reconstructed through the standard methods, and all jets are tagged using a ParticleNet model. All events that fail to pass the single-lepton triggers and baseline selection criteria are discarded. The remainder are then handed to the trained Particle Transformer model, which assigns a classification score for each of the ten event categories listed in 8.2 to each event. Based on their scores, events are then sorted into five control regions (catBB, catBJ, catCC, catCJ, and catLF) and four signal regions (catHcc, catHbb, catZcc, and catZbb). A binned maximum likelihood fit is then performed simultaneously in all regions to extract the best-fit rate parameters for each process and place constraints on the main $t\bar{t}H(H\to c\bar{c})$ signal process. All uncertainties described in 9 are applied to the fit.

Before proceeding, it must be noted that all results presented in this section are preliminary and not endorsed by CMS. Although the majority of the analysis workflow is settled, several key components are still a work in process. These include:

- Rare backgrounds: Processes such as W+jets, Z+jets, and QCD are expected to be largely handled by the baseline selections, but may still have a small effect on

the final score distributions. The final version of the analysis will include these samples in the background estimation.

- Validation: The signal-like processes $t\bar{t}Hb\bar{b}$, $t\bar{t}Zc\bar{c}$, and $t\bar{t}Zb\bar{b}$ will be used in the final version to further validate the background estimation method.

- Further study of $t\bar{t}$+$b\bar{b}$ backgrounds: As discussed in 8.4, a significant excess has been observed in the `catBB` control region. This will be investigated thoroughly before publication.

All of these will be implemented before the analysis group officially requests pre-approval from CMS. The current goal is for the analysis to reach this stage by early January.

For the preliminary unblinded result in this thesis, best-fit rate parameters for each of the five main background processes are shown in table 10.1. Expected and observed limits are shown in Table 10, and impact plots are available in Figure 10.0. (Expected and observed significances are not reported here, since they are all zero due to the smallness of the signal.) The overall expected and observed limits for Run II are 14.06 and 5.02 times the value of the standard model coupling, respectively.

A noteworthy takeaway is that the normalizations for most $t\bar{t}$ backgrounds are large relative to MC, suggesting potential effects from a mismodeling of $t\bar{t}$. The low observed limit suggests a similar possibility: An excess of $t\bar{t}$+$b\bar{b}$ and $t\bar{t}$+$c\bar{c}$ in their respective control regions could "squeeze" the $t\bar{t}Hc\bar{c}$ fit result down to lower values. Preliminary investigations in the primary validation region have suggested that a mismodelling could be caused by the Particle Transformer model picking up on subtle features in data that are not present in the MC samples used for training. However, more effort will be necessary before drawing firm conclusions. It is important to interpret this result as preliminary finding; further refinement of the background estimation techniques may affect the official

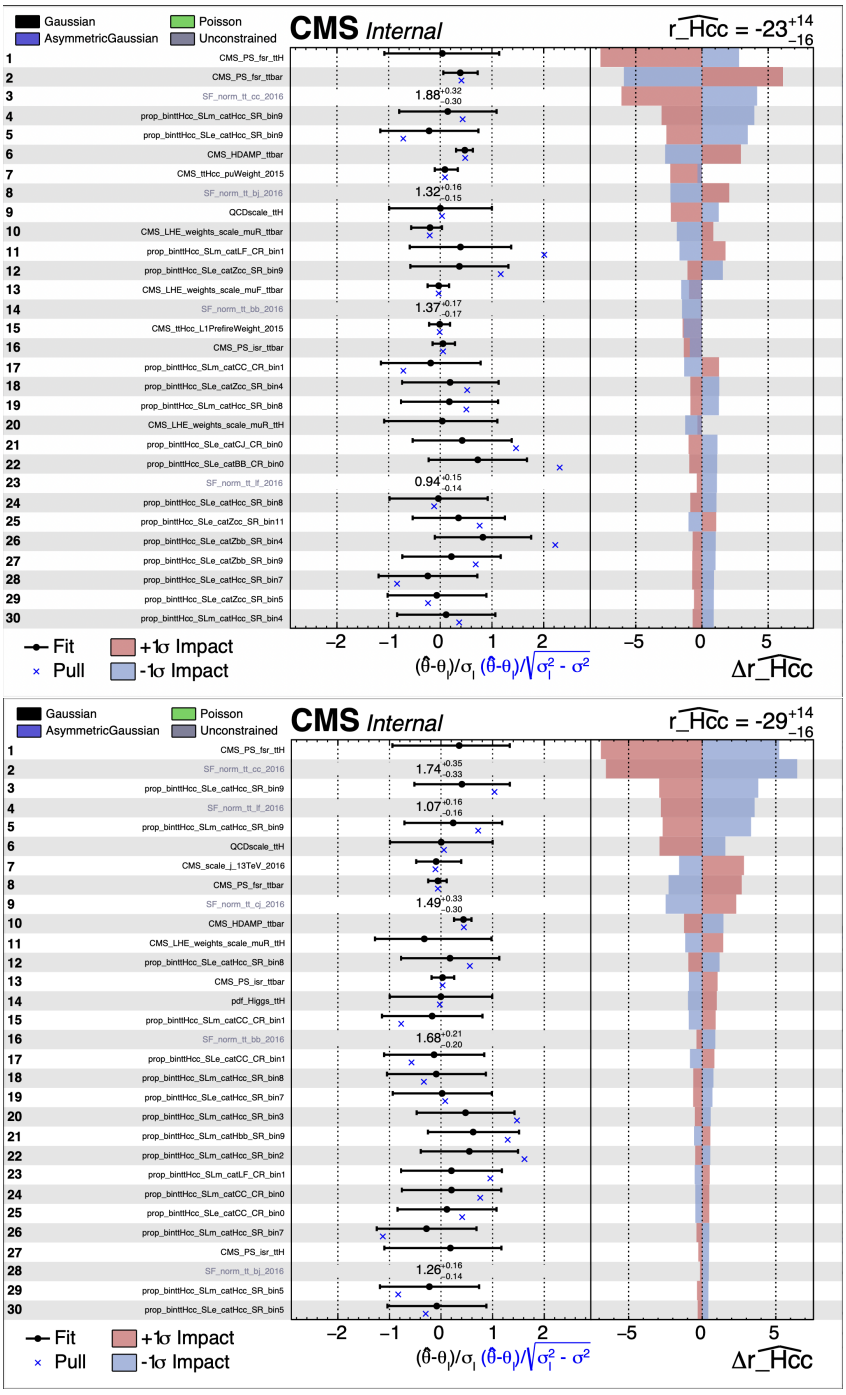
Table 10.1: Background normalization nuisance parameters obtained from the final fit.

	tt+bb	tt+bj	tt+cc	tt+cj	tt+LF
2016 pre-VFP	1.335±0.167	1.308±0.151	1.865±0.312	1.423±0.291	0.940±0.139
2016 post-VFP	1.427±0.175	1.255±0.152	1.660±0.351	1.502±0.318	1.034±0.166
2017	1.310±0.134	1.391±0.125	1.297±0.210	1.134±0.203	0.962±0.142
2018	1.419±0.120	1.225±0.109	1.286±0.198	1.534±0.191	0.678±0.127

result. A more thorough investigation of the $t\bar{t}$ backgrounds studied in this analysis is the current next objective of the $t\bar{t}Hc\bar{c}$ collaboration.

Table 10.2: Preliminary expected and observed limits for Run II in the single-lepton channel. Significances are not reported, since all expected and observed values are zero.

	2016 pre-VFP	2016 post-VFP	2017	2018	Run II
50% expected limit	29.12	33.87	20.75	17.13	14.06
$+1\sigma$ expected limit	43.40	50.07	30.42	25.11	20.34
$+2\sigma$ expected limit	62.98	72.20	43.44	35.96	28.66
Observed limit	18.00	15.74	14.35	12.75	5.42


Gaussian
Poisson
AsymmetricGaussian
Unconstrained
CMS Internal
r_Hcc = -23^{+14}_{-16}
1 CMS_PS_fsr_ttH
2 CMS_PS_fsr_ttbar
3 SF_norm_tt_cc_2016
4 prop_binttHcc_SLm_catHcc_SR_bin9
5 prop_binttHcc_SLe_catHcc_SR_bin9
6 CMS_HDAMP_ttbar
7 CMS_ttHcc_puWeight_2015
8 SF_norm_tt_bj_2016
9 QCDscale_ttH
10 CMS_LHE_weights_scale_muR_ttbar
11 prop_binttHcc_SLm_catLF_CR_bin1
12 prop_binttHcc_SLe_catZcc_SR_bin9
13 CMS_LHE_weights_scale_muF_ttbar
14 SF_norm_tt_bb_2016
15 CMS_ttHcc_L1PrefireWeight_2015
16 CMS_PS_isr_ttbar
17 prop_binttHcc_SLm_catCC_CR_bin1
18 prop_binttHcc_SLe_catZcc_SR_bin4
19 prop_binttHcc_SLm_catHcc_SR_bin8
20 CMS_LHE_weights_scale_muR_ttH
21 prop_binttHcc_SLe_catCJ_CR_bin0
22 prop_binttHcc_SLe_catBB_CR_bin0
23 SF_norm_tt_lf_2016
24 prop_binttHcc_SLe_catHcc_SR_bin8
25 prop_binttHcc_SLe_catZcc_SR_bin11
26 prop_binttHcc_SLe_catZbb_SR_bin4
27 prop_binttHcc_SLe_catZbb_SR_bin9
28 prop_binttHcc_SLe_catHcc_SR_bin7
29 prop_binttHcc_SLe_catZcc_SR_bin5
30 prop_binttHcc_SLm_catHcc_SR_bin4
1.88^{+0.32}_{-0.30}
1.32^{+0.16}_{-0.15}
1.37^{+0.17}_{-0.17}
0.94^{+0.15}_{-0.14}
Fit
Pull
+1σ Impact
-1σ Impact
−2 −1 0 1 2
(θ̂-θ_I)/σ_I
−5 0 5
(θ̂-θ_I)/√(σ_I² - σ²)
Δr_Hcc
Gaussian
Poisson
AsymmetricGaussian
Unconstrained
CMS Internal
r_Hcc = -29^{+14}_{-16}
1 CMS_PS_fsr_ttH
2 SF_norm_tt_cc_2016
3 prop_binttHcc_SLe_catHcc_SR_bin9
4 SF_norm_tt_lf_2016
5 prop_binttHcc_SLm_catHcc_SR_bin9
6 QCDscale_ttH
7 CMS_scale_j_13TeV_2016
8 CMS_PS_fsr_ttbar
9 SF_norm_tt_cj_2016
10 CMS_HDAMP_ttbar
11 CMS_LHE_weights_scale_muR_ttH
12 prop_binttHcc_SLe_catHcc_SR_bin8
13 CMS_PS_isr_ttbar
14 pdf_Higgs_ttH
15 prop_binttHcc_SLm_catCC_CR_bin1
16 SF_norm_tt_bb_2016
17 prop_binttHcc_SLe_catCC_CR_bin1
18 prop_binttHcc_SLm_catHcc_SR_bin8
19 prop_binttHcc_SLe_catHcc_SR_bin7
20 prop_binttHcc_SLm_catHcc_SR_bin3
21 prop_binttHcc_SLm_catHbb_SR_bin9
22 prop_binttHcc_SLm_catHcc_SR_bin2
23 prop_binttHcc_SLm_catLF_CR_bin1
24 prop_binttHcc_SLm_catCC_CR_bin0
25 prop_binttHcc_SLe_catCC_CR_bin0
26 prop_binttHcc_SLm_catHcc_SR_bin7
27 CMS_PS_isr_ttH
28 SF_norm_tt_bj_2016
29 prop_binttHcc_SLm_catHcc_SR_bin5
30 prop_binttHcc_SLe_catHcc_SR_bin5
1.74^{+0.35}_{-0.33}
1.07^{+0.16}_{-0.16}
1.49^{+0.33}_{-0.30}
1.68^{+0.21}_{-0.20}
1.26^{+0.16}_{-0.14}
Fit
Pull
+1σ Impact
-1σ Impact
−2 −1 0 1 2
(θ̂-θ_I)/σ_I
−5 0 5
(θ̂-θ_I)/√(σ_I² - σ²)
Δr_Hcc

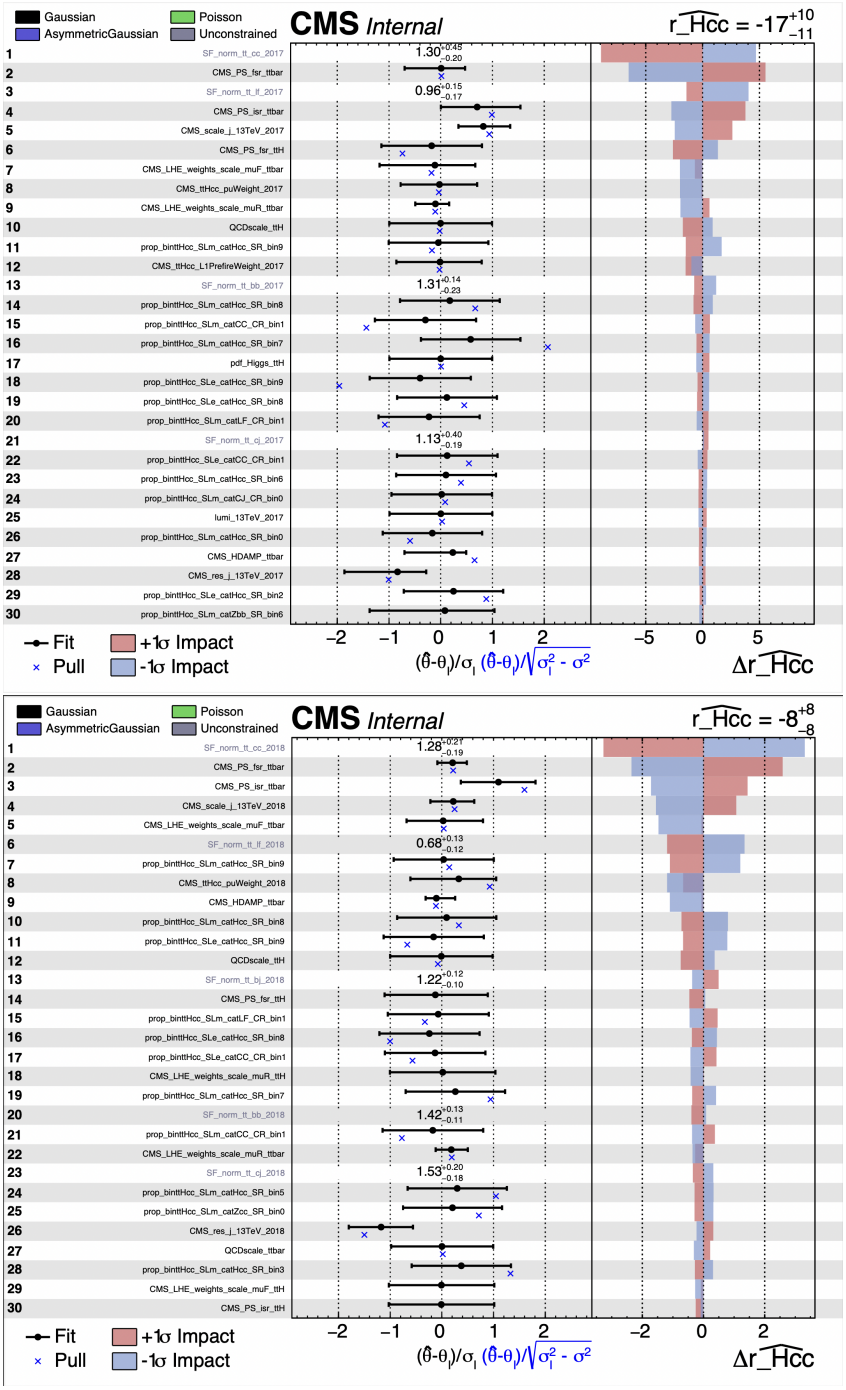


Figure 10.0: Impact plots for each year, in order of increasing year (treating 2016 Pre-VFP and 2016 Post-VFP separately as usual).

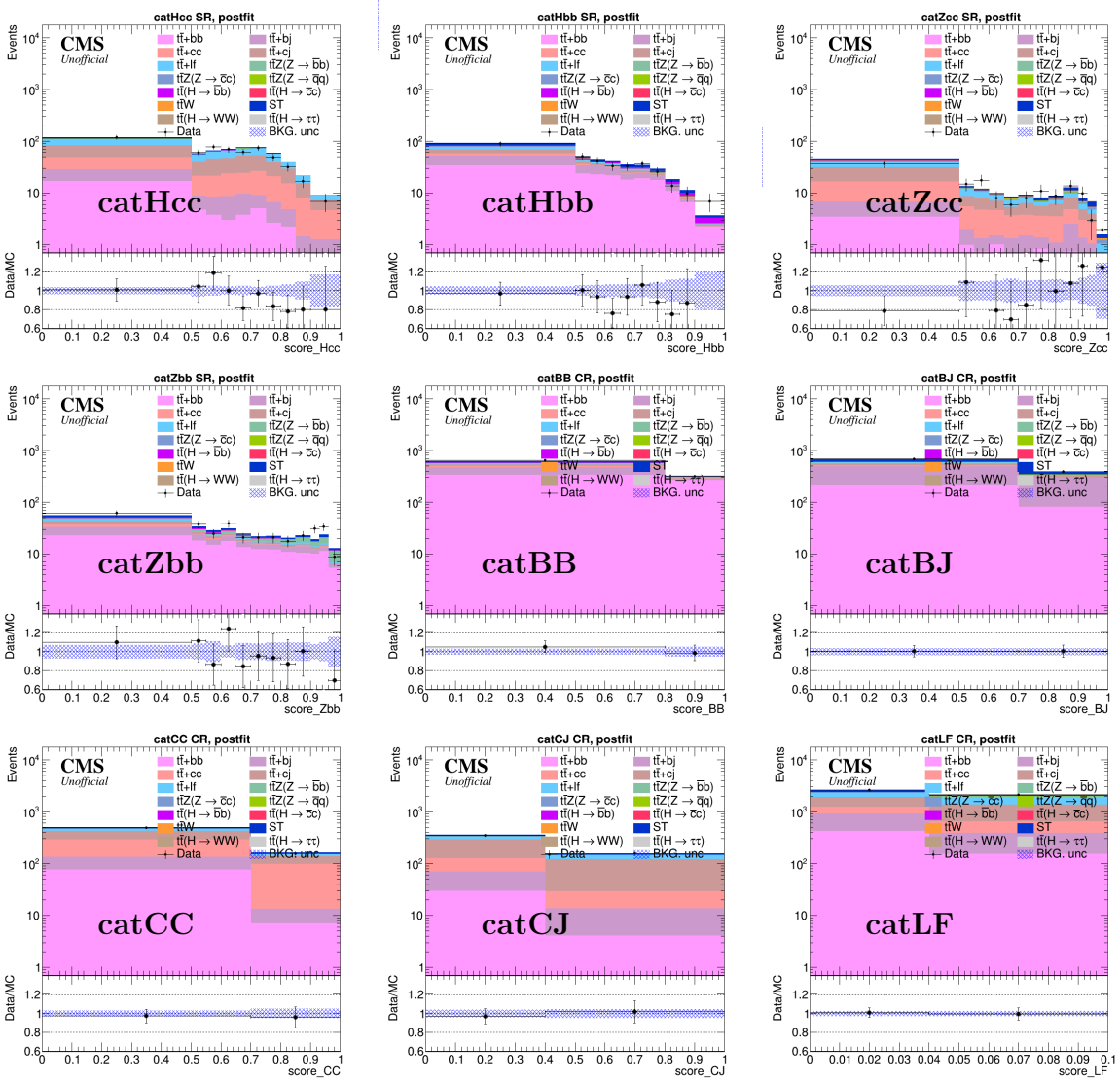


Figure 10.1: Final post-fit distributions for 2018 data in the single-electron channel.

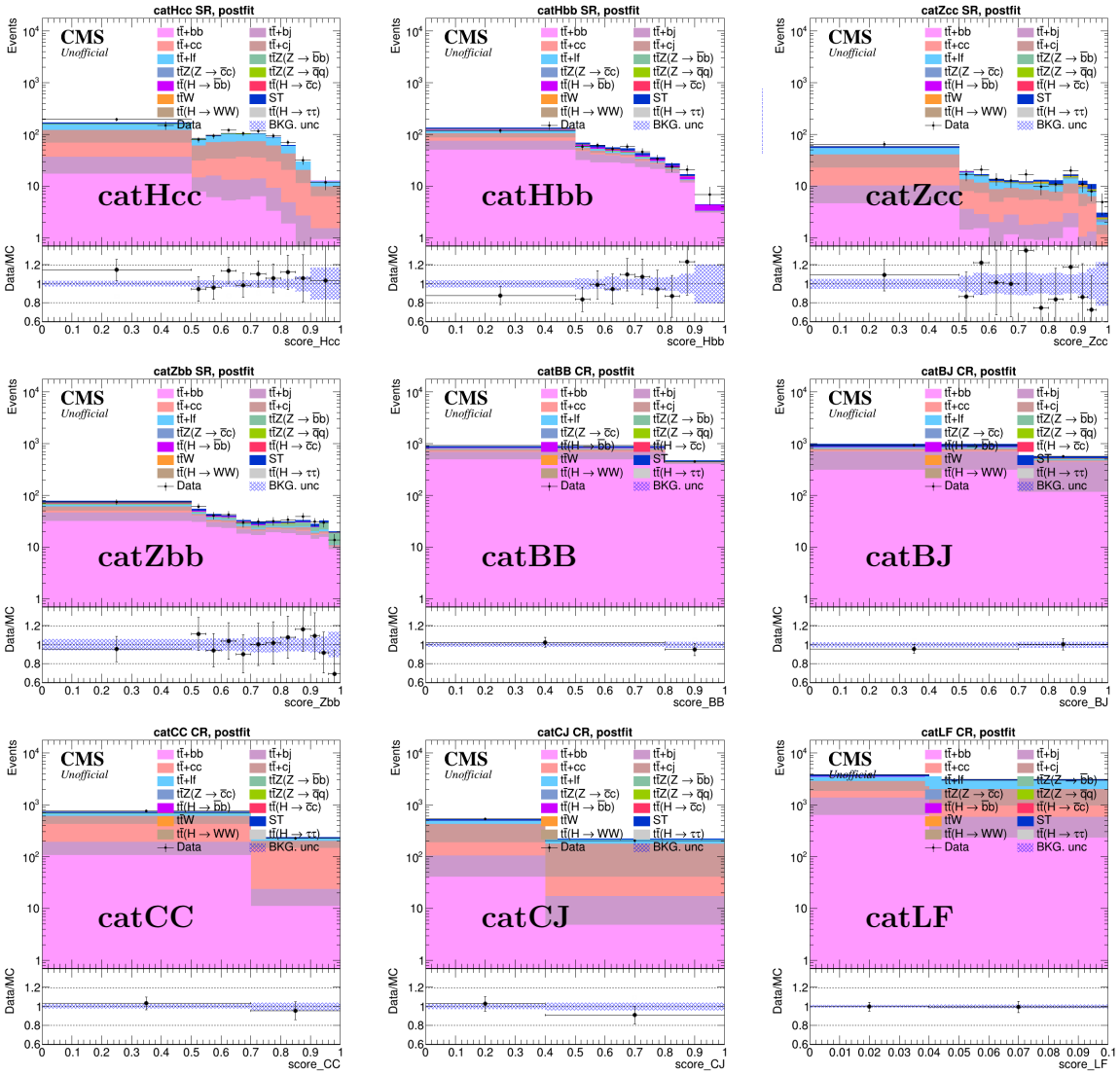


Figure 10.2: Final post-fit distributions for 2018 data in the single-muon channel.

Chapter 11

Conclusion

This thesis summarized the current status of an ongoing search for the $gg \to t\bar{t}H$ ($H \to c\bar{c}$) process. Thanks to state-of-the-art jet tagging methods and a novel approach to event classification using a transformer-based DNN, the expected limits from this analysis are highly competitive with the current best limits on the charm Yukawa coupling, particularly once the results from all channels have been combined. A preliminary observed (expected) result of 5.02 (14.06) times the standard model was obtained in the single-lepton channel using the full Run II dataset.

While it is premature to draw firm conclusions due to the preliminary nature of the result, a lack of an excess would further constrain BSM physics models that predict a magnified Higgs-charm Yukawa coupling, most notably 2HDMs where first- and second-generation quarks derive mass from a source other than the SM Higgs. This would provide further support for the standard model's validity across a broad range of measurements. In the longer term, this result motivates a search for $t\bar{t}H$ ($H \to c\bar{c}$) at the HL-LHC, since the much higher luminosity may allow future analyses to be sensitive to the SM value of the coupling. Regardless, the analysis techniques employed demonstrate the growing strength of modern machine learning techniques in particle physics, paving the way for

DNNs like Particle Transformer to play a major role in future background estimation strategies.

Appendix A

Electron Trigger Scale Factors

This appendix describes the technical details of the electron trigger scale factors for the single-lepton channel. Section A.1 explains the process of deriving the scale factors, while section A.1 covers the corresponding systematic uncertainties.

A.1 Calculation of SL electron trigger scale factors

As discussed briefly in Chapter 7, trigger scale factors are necessary to account for significant differences in trigger efficiency between data and MC samples. Figure A.1 contains several plots of key kinematic variables for 2018 data and MC with and without the signal trigger applied. Significant shape differences are visible in many of the plots, indicating that the trigger behaves differently on data and MC events; this difference must be compensated for with scale factors.

The signal triggers used in the single-electron channel are listed in A.1. As usual, the scale factors are defined as the ratio of the trigger efficiency in data to the trigger efficiency in MC:

Year	Triggers
2016	HLT_Ele27_WPTight_Gsf
2017	HLT_Ele32_WPTight_Gsf_L1DoubleEG_plusL1 HLT_Ele28_eta2p1_WPTight_Gsf_HT150
2018	HLT_Ele32_WPTight_Gsf HLT_Ele28_eta2p1_WPTight_Gsf_HT150

Table A.1: Signal triggers per year. All triggers are combined with a logical OR.

year	reference triggers
2016	HLT_IsoMu24 HLT_IsoTkMu24
2017	HLT_IsoMu27
2018	HLT_IsoMu24

Table A.2: Reference triggers per year. All triggers are combined with a logical OR.

$$\mathrm{SF}(\mathrm{bin}) = \frac{\epsilon(\mathrm{bin}, \mathrm{data})}{\epsilon(\mathrm{bin}, \mathrm{MC})} \tag{A.1}$$

The data and MC trigger efficiencies $\epsilon(\mathrm{bin}, \mathrm{data})$ and $\epsilon(\mathrm{bin}, \mathrm{MC})$ must be computed using a sample of events orthogonal to the main signal and control regions. Single-muon triggers are known to be nearly orthogonal to the single-electron triggers (apart from small correlations accounted for later), and are consequently used to select this sample. See Table A.2 for the full list. Since trigger efficiencies are known to depend strongly on both the electron transverse momentum (p_T) and electron supercluster η (η_{SC}), SFs are computed in independent bins of p_T and η_{SC}.

All data and MC samples used to compute the scale factors are listed in A.4 and A.3, respectively. Both dilepton and single-lepton samples were used, as false positives in the single-lepton sample will result in a small number of single-lepton events surviving the standard selections. However, since over 97% of all events passing the selections come from the dilepton sample, the final scale factors are computed using only dilepton events,

Year	Samples
2015	/TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8/ RunIISummer20UL16NanoAODAPVv9-106X_mcRun2_asymptotic_preVFP_v11-v1
	/TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8/ RunIISummer20UL16NanoAODAPVv9-106X_mcRun2_asymptotic_preVFP_v11-v1
2016	/TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8/ RunIISummer20UL16NanoAODv9-106X_mcRun2_asymptotic_v17-v1
	/TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8/ RunIISummer20UL16NanoAODv9-106X_mcRun2_asymptotic_v17-v1
2017	/TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8/ RunIISummer20UL17MiniAODv2-106X_mc2017_realistic_v9-v1
	/TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8/ RunIISummer20UL17MiniAODv2-106X_mc2017_realistic_v9-v1
2018	/TTToSemiLeptonic_TuneCP5_13TeV-powheg-pythia8/ RunIISummer20UL18MiniAODv2-106X_upgrade2018_realistic_v16_L1v1-v2
	/TTTo2L2Nu_TuneCP5_13TeV-powheg-pythia8/ RunIISummer20UL18MiniAODv2-106X_upgrade2018_realistic_v16_L1v1-v1

Table A.3: MC samples used in the SL single electron channel.

with the single lepton events accounted for via an additional systematic uncertainty (see below). The baseline event selection is $N_{\text{jets}} \geq 4$, in addition to the other standard event cleaning and quality criteria.

A notable complication was the failure of the HEM15/16 HCAL regions during runs 2018C and 2018D. To account for this, two separate checks were performed: First, an additional uncertainty was added to the energies of all reconstructed objects in the affected

Year	Samples
2016B-F	/SingleMuon/Run2016*-ver2_HIPM_UL2016_MiniAODv2_NanoAODv9-v2
2016G-H	/SingleMuon/Run2016*-UL2016_MiniAODv2_NanoAODv9-v1
2017	/SingleMuon/Run2017*-UL2017_MiniAODv2_NanoAODv9_GT36-v1
2018	/SingleMuon/Run2018*-UL2018_MiniAODv2_NanoAODv9_GT36-v1

Table A.4: Data samples used in the SL single electron channel.

region, and second, to set an upper bound on the magnitude of the effect, all leptons in the affected region were removed. Upon recalculating the scale factors and systematics after each change, neither was found to have a significant effect on the results.

Calculation of systematic uncertainties

In addition to the statistical uncertainty for each bin, four systematic uncertainties were combined to calculate the final uncertainty.

First, because the EleHT cross trigger includes a cut on HT, the scale factors themselves are expected to have a dependence on HT. While small, this effect is visible in A.2. To account for it, scale factors for each year were computed separately for events with HT < 400 GeV and HT > 400 GeV, and a systematic equal to half of the difference was computed for each bin:

$$\sigma_{\mathrm{HT}}(\mathrm{bin}) = \frac{1}{2}|\mathrm{SF}(\mathrm{bin}, \mathrm{HT} > 400) - \mathrm{SF}(\mathrm{bin}, \mathrm{HT} < 400)| \tag{A.2}$$

Second, the impact of adding the single lepton samples was accounted for by a similar method. Both the dilepton and single lepton samples were combined, and the overall scale factors were recalculated for the combination. The corresponding systematic was taken to be the magnitude of the difference:

$$\sigma_{\mathrm{SL}}(\mathrm{bin}) = |\mathrm{SF}(\mathrm{bin}, \mathrm{DL}) - \mathrm{SF}(\mathrm{bin}, \mathrm{DL{+}SL})| \tag{A.3}$$

Third, because the conditions of the CMS detector varied slightly over time, the trigger efficiency was expected to have a year dependence. To investigate this, scale factors were first computed separately for each year, then the luminosity-weighted average of all years was taken:

$$\mathrm{SF}(\mathrm{bin}, \mathrm{weighted}) = \frac{1}{\mathrm{Lumi}_{total}} \sum_{y} \mathrm{SF}(\mathrm{bin}, y) * \mathrm{Lumi}_{y} \tag{A.4}$$

...with y summed over 2016 (B-F), 2016 (F-H), 2017, and 2018. The final systematics were the magnitude of the difference between the luminosity-weighted scale factors and the nominal scale factors:

$$\sigma_{\mathrm{run}}(\mathrm{bin}) = |\mathrm{SF}(\mathrm{bin}, \mathrm{weighted}) - \mathrm{SF}(\mathrm{bin}, \mathrm{all\ data})| \tag{A.5}$$

Fourth, although the reference muon trigger would ideally be orthogonal to the signal electron trigger, this is not the case in reality. One must account for small but nonzero correlations between the two. We can define a correlation factor α:

$$\alpha = \frac{\epsilon_e^{MC} \epsilon_\mu^{MC}}{\epsilon_{e,\mu}^{MC}} \tag{A.6}$$

...where ϵ_e^{MC} is the efficiency of the signal trigger (on events passing the baseline selections), ϵ_μ^{MC} is the efficiency of the reference trigger, and $\epsilon_{e,\mu}^{MC}$ is the efficiency of the signal and reference triggers in combination. We can subsequently define a systematic uncertainty:

$$\sigma_{\mathrm{trigger}}(\mathrm{bin}) = (1 - \alpha) * \mathrm{SF}(\mathrm{bin}) \tag{A.7}$$

For each bin, the statistical uncertainty and the four systematics were summed together in quadrature to obtain an overall uncertainty.

Final single-electron scale factors

The final scale factors for each year are presented in A.3 through A.6. Overall, the results were found to be in reasonably good agreement with the previous findings in

[84]. The statistical uncertainty contributes roughly half of the final uncertainty and the HT systematic systematic contributes around another quarter, while the remaining three systematics were relatively small (<10%) in most cases.

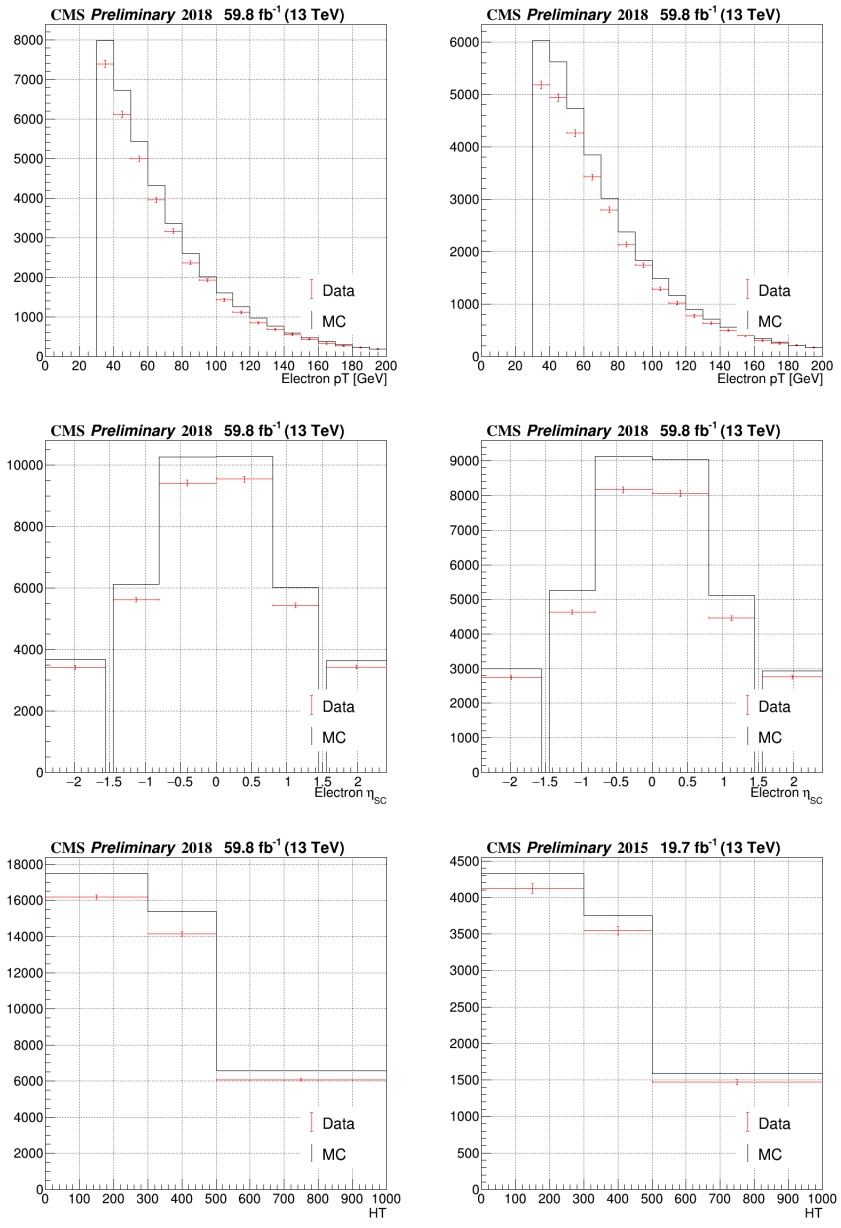


Figure A.1: 2018 distributions of several key kinematic variables for data and MC after the standard SL selections, with and without the signal trigger applied.

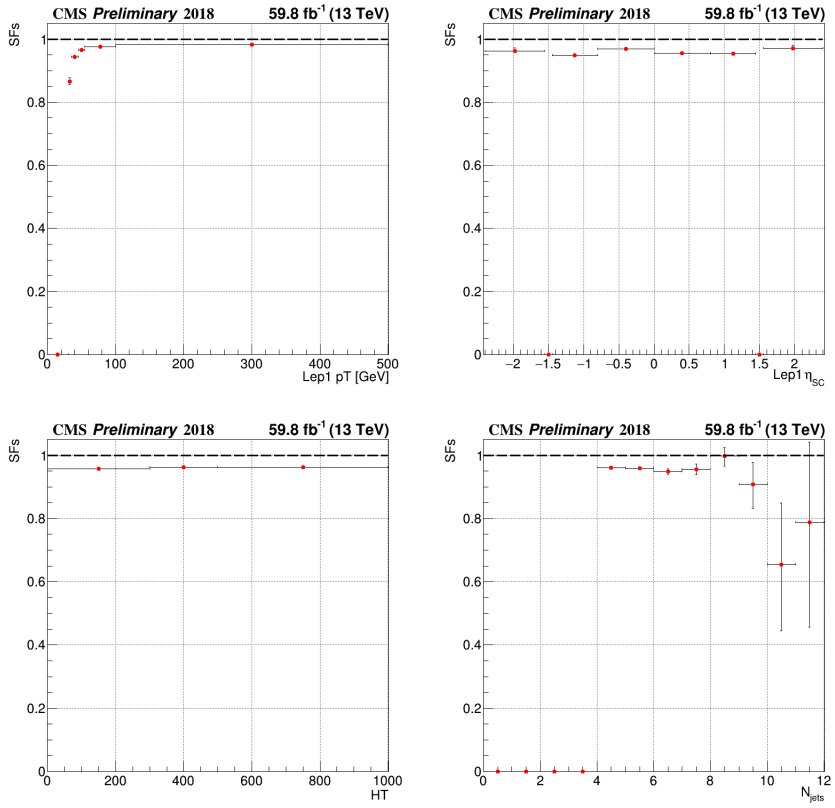


Figure A.2: One-dimensional 2018 scale factors for several key kinematic variables.

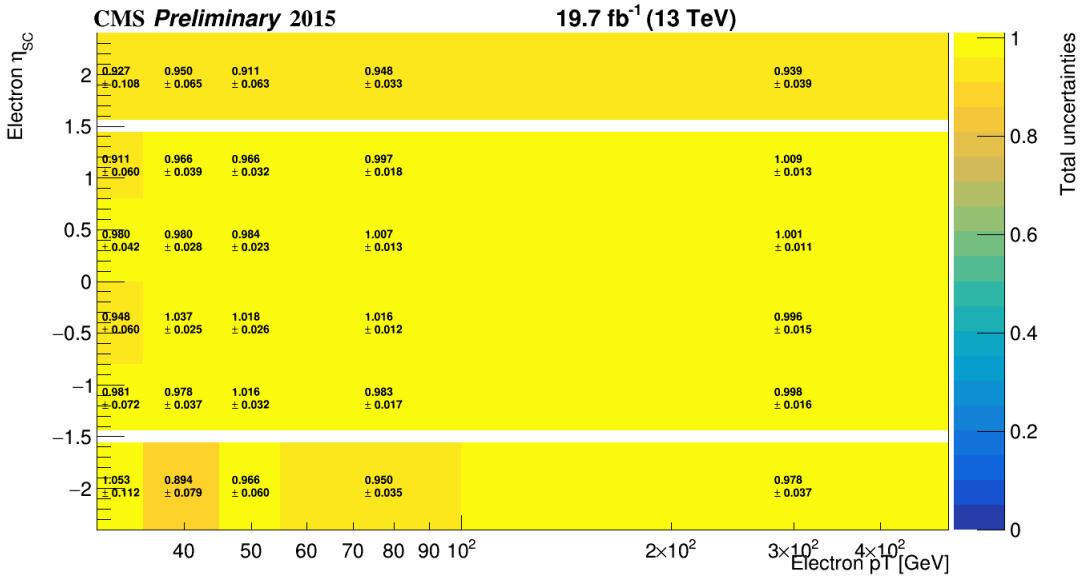


Figure A.3: Final 2-dimensional uncertainties for run 2016B-F.

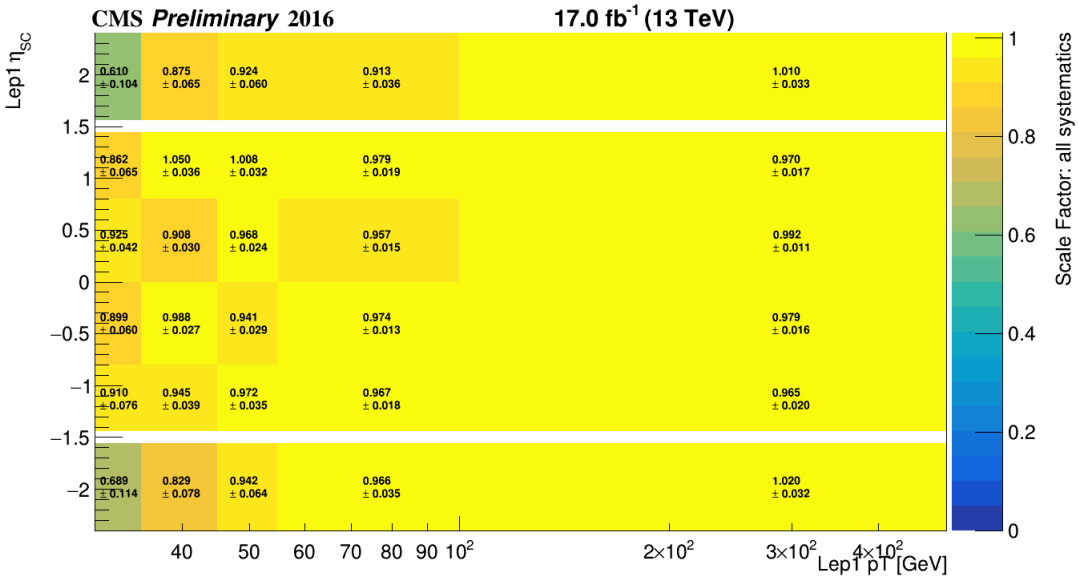


Figure A.4: Final 2-dimensional uncertainties for run 2016F-H.

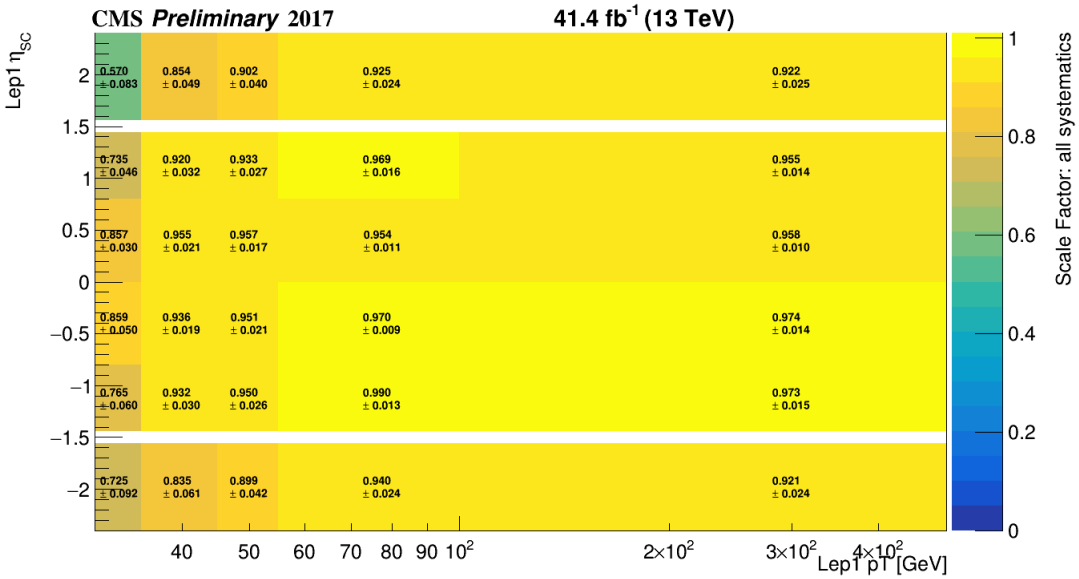


Figure A.5: Final 2-dimensional uncertainties for run 2017.

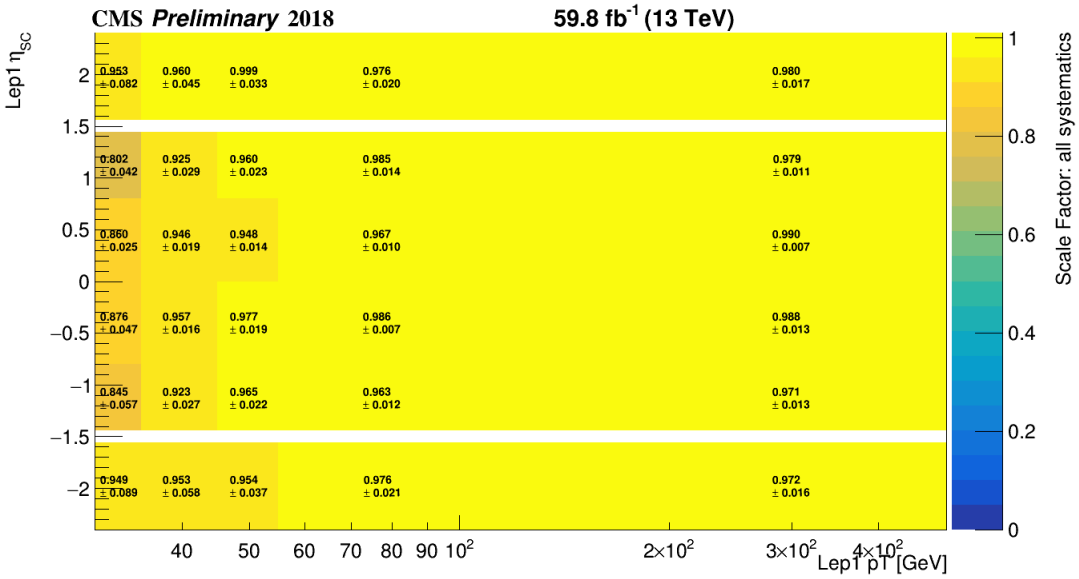


Figure A.6: Final 2-dimensional uncertainties for run 2018.

Bibliography

[1] X. Fan, T. Myers, B. Sukra, and G. Gabrielse, *Measurement of the electron magnetic moment*, *Physical Review Letters* **130** (feb, 2023).

[2] A. Collaboration, *Observation of a new particle in the search for the standard model higgs boson with the ATLAS detector at the LHC*, *Physics Letters B* **716** (sep, 2012) 1–29.

[3] C. Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, *Physics Letters B* **716** (sep, 2012) 30–61.

[4] **CMS** Collaboration, *Observation of Higgs boson decay to bottom quarks*, tech. rep., CERN, Geneva, 2018.

[5] "Example: Standard model of physics." `https://tikz.net/higgs-potential/`. Accessed: 2023-10-27.

[6] M. E. Peskin and D. V. Schroeder, *An Introduction to quantum field theory*. Addison-Wesley, Reading, USA, 1995.

[7] J. Horejsi, *Fundamentals of electroweak theory*, 2022.

[8] M. Thomson, *Modern particle physics*. Cambridge University Press, New York, 2013.

[9] G. Pignol and P. Schmidt-Wellenburg, *The search for the neutron electric dipole moment at psi*, 2021.

[10] C. Tamarit, W.-Y. Ai, J. S. Cruz, and B. Garbrecht, *The limits of the strong CP problem*, in *Proceedings of 7th Symposium on Prospects in the Physics of Discrete Symmetries, DISCRETE 2020-2021 — PoS(DISCRETE2020-2021)*, Sissa Medialab, may, 2022.

[11] A. Hook, *TASI Lectures on the Strong CP Problem and Axions*, *PoS* **TASI2018** (2019) 004.

[12] "Axion constraints 2020." https://indico.cern.ch/event/948186/contributions/4100241/attachments/2163557/3651055/axions2020.pdf. Accessed: 2023-11-2.

[13] A. D. Sakharov, *Violation of CP Invariance, C asymmetry, and baryon asymmetry of the universe*, *Pisma Zh. Eksp. Teor. Fiz.* **5** (1967) 32–35.

[14] J. L. Feng, *Naturalness and the status of supersymmetry*, *Annual Review of Nuclear and Particle Science* **63** (oct, 2013) 351–382.

[15] P. Draper, P. Meade, M. Reece, and D. Shih, *Implications of a 125 GeV higgs boson for the MSSM and low-scale supersymmetry breaking*, *Physical Review D* **85** (may, 2012).

[16] "Atlas run 2 searches for electroweak production of supersymmetric particles interpreted within the pmssm." https://cds.cern.ch/record/2870222. Accessed: 2023-11-4.

[17] D. Ghosh, R. S. Gupta, and G. Perez, *Is the higgs mechanism of fermion mass generation a fact? a yukawa-less first-two-generation model*, *Physics Letters B* **755** (apr, 2016) 504–508.

[18] F. Botella, G. Branco, M. Rebelo, and J. Silva-Marcos, *What if the masses of the first two quark families are not generated by the standard model higgs boson?*, *Physical Review D* **94** (dec, 2016).

[19] R. Harnik, J. Kopp, and J. Zupan, *Flavor violating higgs decays*, *Journal of High Energy Physics* **2013** (mar, 2013).

[20] W. Altmannshofer, J. Eby, S. Gori, M. Lotito, M. Martone, and D. Tuckler, *Collider signatures of flavorful higgs bosons*, *Physical Review D* **94** (dec, 2016).

[21] M. Srednicki, *Quantum Field Theory*. Cambridge Univ. Press, Cambridge, 2007.

[22] "Higgs cross sections for hl-lhc and he-lhc." https://twiki.cern.ch/twiki/bin/view/LHCPhysics/HiggsEuropeanStrategy. Accessed: 2023-11-4.

[23] J. e. a. Seeman, *Design and Principles of Linear Accelerators and Colliders*, pp. 295–336. Springer International Publishing, Cham, 2020.

[24] E. Lopienska, *The CERN accelerator complex, layout in 2022. Complexe des accélérateurs du CERN en janvier 2022*, . General Photo.

[25] **CMS** Collaboration, M. e. a. Della Negra, *CMS: letter of intent by the CMS Collaboration for a general purpose detector at LHC*, tech. rep., CERN, Geneva, 1992. Open presentation to the LHCC 5 November 1992, M. Della Negra/CERN, CMS Spokesman.

[26] "The cms detector: From design to discovery." `https://indico.cern.ch/event/1135177/contributions/4763030/attachments/2473381/4243770/CMS_Higgs_10y_v10.pdf`. Accessed: 2023-11-5.

[27] **CMS** Collaboration, G. L. e. a. Bayatian, *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*. Technical design report. CMS. CERN, Geneva, 2006. There is an error on cover due to a technical problem for some items.

[28] "The cms experiment at cern - detector." `https://cms.cern/detector`. Accessed: 2023-11-5.

[29] "Public cms luminosity information." `https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults`. Accessed: 2023-11-5.

[30] G. Perez, *Unitarization Models For Vector Boson Scattering at the LHC*. PhD thesis, KIT, 01, 2018.

[31] "Overview of cms goals and detector." `https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookCMSExperiment`. Accessed: 2023-11-5.

[32] **CMS** Collaboration, T. T. G. of the CMS Collaboration, *The CMS Phase-1 Pixel Detector Upgrade*, tech. rep., CERN, Geneva, 2020.

[33] C. Collaboration, *Commissioning and performance of the cms silicon strip tracker with cosmic ray muons*, *Journal of Instrumentation* **5** (mar, 2010) T03008.

[34] "Brilrselbahgc - plots for approval by members of bril rs." `https://twiki.cern.ch/twiki/bin/view/CMSPublic/BRILRSelbaHGC`. Accessed: 2023-11-6.

[35] **CMS** Collaboration, *The CMS electromagnetic calorimeter project: Technical Design Report*. Technical design report. CMS. CERN, Geneva, 1997.

[36] **CMS** Collaboration, *The CMS hadron calorimeter project: Technical Design Report*. Technical design report. CMS. CERN, Geneva, 1997.

[37] **CMS** Collaboration, J. G. Layter, *The CMS muon project: Technical Design Report*. Technical design report. CMS. CERN, Geneva, 1997.

[38] C. Collaboration, "Cmssw - github." `https://github.com/cms-sw/cmssw`.

[39] A. S. et al, *Particle-flow reconstruction and global event description with the CMS detector*, *Journal of Instrumentation* **12** (oct, 2017) P10003–P10003.

[40] T. C. Collaboration, *Description and performance of track and primary-vertex reconstruction with the CMS tracker*, *Journal of Instrumentation* **9** (oct, 2014) P10009–P10009.

[41] M. B. Andrews, *Search for exotic Higgs boson decays to merged photons employing a novel deep learning technique at CMS*, 2021.

[42] L. Lista, *Practical statistics for particle physicists*, *CERN Yellow Reports: School Proceedings* (2017) Vol 5 (2017): Proceedings of the 2016 European School of High–Energy Physics.

[43] **TheATLAS, TheCMS, TheLHCHiggsCombinationGroup** Collaboration, *Procedure for the LHC Higgs boson search combination in Summer 2011*, tech. rep., CERN, Geneva, 2011.

[44] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, *The European Physical Journal C* **71** (feb, 2011).

[45] M. Cacciari, G. P. Salam, and G. Soyez, *The anti-k_t jet clustering algorithm*, *Journal of High Energy Physics* **2008** (apr, 2008) 063–063.

[46] S. Moortgat, *Development of new charm-tagging methods for the search for Flavour Changing top-quark dark matter interactions at the LHC*, 2015.

[47] C. e. a. Bennett, "Workbook - b tagging." `https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookBTagging`.

[48] H. Qu and L. Gouskos, *Jet tagging via particle clouds*, *Physical Review D* **101** (Mar., 2020).

[49] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, *Dynamic graph cnn for learning on point clouds*, 2019.

[50] C. Collaboration, *Development of the cms detector for the cern lhc run 3*, 2023.

[51] L. Ostman, *Imaging using machine learning for the ldmx electromagnetic calorimeter*, Master's thesis, Lund University, 2020.

[52] H. Qu, C. Li, and S. Qian, *Particle transformer for jet tagging*, 2022.

[53] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, *Attention is all you need*, 2023.

[54] **LHCHiggsCrossSectionWorkingGroup** Collaboration, S. e. a. Heinemeyer, *Handbook of LHC Higgs Cross Sections: 3. Higgs Properties: Report of the LHC Higgs Cross Section Working Group*. CERN Yellow Reports: Monographs. 2013. Comments: 404 pages, 139 figures, to be submitted to CERN Report. Working Group web page: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CrossSections.

[55] **CMS** Collaboration, A. e. a. Tumasyan, *Search for Higgs boson decay to a charm quark-antiquark pair in proton-proton collisions at $\sqrt{s} = 13$ TeV*, *Phys. Rev. Lett.* **131** (2023) 061801, [arXiv:2205.0555]. All the figures and tables, including additional supplementary figures and tables, can be found at http://cms-results.web.cern.ch/cms-results/public-results/publications/HIG-21-008 (CMS Public Pages).

[56] **CMS** Collaboration, *Inclusive search for a boosted Higgs boson decaying to a charm quark pairs in proton-proton collisions at sqrt(s) = 13 TeV*, tech. rep., CERN, Geneva, 2022.

[57] **ATLAS** Collaboration, G. e. a. Aad, *Direct constraint on the Higgs–charm coupling from a search for Higgs boson decays into charm quarks with the ATLAS detector*, *Eur. Phys. J. C.* **82** (2022) 717.

[58] N. M. Coyle, C. E. M. Wagner, and V. Wei, *Bounding the charm yukawa coupling*, *Phys. Rev. D* **100** (Oct, 2019) 073013.

[59] I. Brivio, F. Goertz, and G. Isidori, *Probing the charm quark yukawa coupling in Higgs + Charm production*, *Phys. Rev. Lett.* **115** (Nov, 2015) 211801.

[60] **CMS** Collaboration, *Simulation of the Silicon Strip Tracker pre-amplifier in early 2016 data*, .

[61] A. M. e. a. Sirunyan, *Extraction and validation of a new set of cms pythia8 tunes from underlying-event measurements*, *The European Physical Journal C* **80** (Jan., 2020).

[62] T. Ježo, J. M. Lindert, N. Moretti, and S. Pozzorini, *New nlops predictions for $t\bar{t}$ + bjet production at the lhc*, *The European Physical Journal C* **78** (jun, 2018).

[63] **CMS** Collaboration, *Inclusive and differential cross section measurements of* $\mathrm{t\bar{t}b\bar{b}}$ *production in the lepton+jets channel at $\sqrt{s} = 13\,\mathrm{TeV}$ with the CMS detector*, tech. rep., CERN, Geneva, 2023.

[64] T. C. collaboration, *Electron and photon reconstruction and identification with the cms experiment at the cern lhc*, *Journal of Instrumentation* **16** (may, 2021) P05014.

[65] A. S. et al, *Performance of the cms muon detector and muon reconstruction with proton-proton collisions at √s=13 tev*, *Journal of Instrumentation* **13** (jun, 2018) P06015.

[66] N. e. a. Norjoharuddeen, *et identification in high pile-up environment (pileupjetid) for ultra legacy data*, 2023.

[67] H. Q. et al, *Particlenet for ak4 jet tagging*, 2020.

[68] H. Qu, *weaver-core: Pypi*, 2023.

[69] J. e. a. Rosenzweig, *Met analysis*, 2023.

[70] B. e. a. Maier, *Noise filter recommendations for run ii & run iii*, 2023.

[71] K. Borras, *Status of the cms experiment*, 2018.

[72] A. e. a. Jafari, *The modeling of the top quark p_t*, 2020.

[73] T. e. a. Laurent, *Reweighting recipe to emulate level 1 ecal and muon prefiring.*, 2023.

[74] F. e. a. Primavera, *Muon pog: Muon hlt*, 2023.

[75] C. P. O. Group, *Summary of common pog jsons*, 2023.

[76] L. H. working group (LHCHXWG), *Sm higgs branching ratios and total decay widths*, 2023.

[77] C. T. e. a. Potter, *Handbook of lhc higgs cross sections: 3. higgs properties: Report of the lhc higgs cross section working group*, 2013.

[78] A. e. a. Babaev, *Utilities for accessing pileup information for data*, 2023.

[79] S. Mrenna and P. Skands, *Automated parton-shower variations in pythia 8*, *Physical Review D* **94** (Oct., 2016).

[80] J. B. et al, *Pdf4lhc recommendations for lhc run ii*, *Journal of Physics G: Nuclear and Particle Physics* **43** (jan, 2016) 023001.

[81] A. e. a. Babaev, *Luminosity physics object group (lum pog)*, 2023.

[82] *Jet energy scale: Jet energy resolution*, 2023.

[83] *Jet energy scale uncertainty sources*, 2022.

[84] A. D. et al., "Trigger efficiencies and scale factors for the ttH(H→bb) measurement in single lepton and dilepton channels with the full run-2 data sample." AN-2019/008.