Summer 7-25-2013

TTbar Production Cross Section Measurement in the Muon Plus Jets Channel using Soft Electron Tagging in *pp* Collisions at sqrt(*s*) = 8 TeV

Jason Keller
University of Nebraska-Lincoln, jdkeller7@gmail.com

TTBAR PRODUCTION CROSS SECTION MEASUREMENT IN THE MUON PLUS JETS CHANNEL USING SOFT ELECTRON TAGGING IN PP COLLISIONS AT $\sqrt{S} = 8$ TEV

by

Jason Keller

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Doctor of Philosophy

Major: Physics & Astronomy

Under the Supervision of Professor Daniel Claes

Lincoln, Nebraska

August, 2013

TTBAR PRODUCTION CROSS SECTION MEASUREMENT IN THE MUON PLUS JETS CHANNEL USING SOFT ELECTRON TAGGING IN PP COLLISIONS AT $\sqrt{S} = 8$ TEV

Jason Keller, Ph.D.

University of Nebraska, 2013

Adviser: Daniel Claes

A measurement is made of the top quark pair production cross section through the decay channel $t\bar{t} \to \mu\nu_\mu +$ jets, carried out using the Compact Muon Solenoid (CMS) detector at the Large Hadron Collider (LHC), located at CERN in Geneva, Switzerland. The top quark pair events were produced in proton-proton collisions at a center-of-mass energy of 8 TeV. To facilitate the detection of $t\bar{t}$ events, jets produced from bottom quarks, which are decay products of the top quarks, are identified using electrons with low transverse momentum with respect to the beam axis (soft electrons). The dominant background, W + jets production, is estimated using the preferential production of W^+ over W^- at the LHC. Techniques to measure or estimate the remaining backgrounds, and the identification efficiency of the soft electron technique are also discussed. The final result of this study is

$$\sigma_{t\bar{t}} = 228.9 \pm 4.8\,(\text{stat})\,^{+27.7}_{-27.8}\,(\text{syst.}) \pm 10.1\,(\text{lumi})\,\text{pb}.$$

ACKNOWLEDGMENTS

This dissertation would not have been possible without the help and support of so many people in so many ways. The work represented by this dissertation required a lot of effort not only on my part, but from those around me who offered me endless support, and a foundation of knowledge and hard work. These acknowledgements express gratitude to those who have greatly assisted me in this endeavor, but the list of people is by no means complete. To my friends and teachers who have helped me get to where I am today, I thank you greatly.

First and foremost, I would like to thank my advisor, Dan Claes, who is a good role model of what a scientist and professor should be. He always went out of his way to make time for me, no matter how busy he was, and would always offer his advice, and help me analyze issues in a systematic way. He would offer multiple solutions to every problem, and always listened to all my crazy ideas. He even agreed to a few of them. He also shares my love of comic books, which means be must be a great person.

Additionally, I would like to thank Prof. Greg Snow, who (along with Prof. Claes), took a chance on me as an undergraduate and hired me as a student research assistant. I learned many things under their guidance in the CROP lab, and they even let me tinker around with a few things so I could learn even more. It was this experience that cultivated my interest in high energy physics, and made me want to pursue it more in graduate school.

I would also like to thank Professors Ken Bloom, Aaron Dominguez, and Ilya Kravchenko for their help and support. They were always willing to give any advice that they could, give helpful, constructive criticism that helped me improve as an experimentalist, and taught me the importance of critical thinking and working

systematically. They are an example of the success that comes with hard work and perseverance.

A special thanks to Professors Herman Batelaan, Anthony Starace, and Jennifer McKitrick for agreeing to serve on my supervisory committee, and for providing an alternative point of view and set of questions. Profs. Batelaan and Starace also taught me in many physics courses throughout my career, and helped me prepare to become a better student and researcher.

Last, but certainly not least, I would like to thank my wife Stacie, my father Bob, and my brother Corey, who were a constant source of support throughout my life. Their support, both in my graduate career and in life, has meant a lot to me, and I will forever be endebted to them. Their effect on me as a person is much more than can be expressed on paper.

Contents

List of Figures

List of Tables

Chapter 1

Physics of the Standard Model

1.1 The Standard Model

Particle physics is the study of the fundamental particles of the Universe and their interactions. While interactions have been studied for centuries, the investigation of particles is relatively new, dating to the early 20th century. Quantum field theory, also developed in the 20th century, recognized that forces could be explained through the use of particles as force carriers. The observations of particles, and how their interactions work, has led to a collection of theories known as the Standard Model. This collection of theories has been tested many times, and has been found to quite accurately describe the matter that we see, the interactions they experience, and has made additional predictions which have also proved fruitful [1]. This chapter will describe the Standard Model, and will end with the description of its most massive known quark to date, the top quark.

1.1.1 Fundamental Particles

The fundamental particles of matter can be grouped into three *generations* (Tab. 1.1). The three generations are divided into quarks and leptons. Quarks can experience all three interactions of the Standard Model (strong, electromagnetic, weak), while charged leptons (e, μ, τ) only experience the electromagnetic and weak interactions. Neutrinos, being neutral, only experience the weak interaction. All the particles in Table 1.1 are fermions, and thus have half-integer spin. The term *fundamental* implies these particles are not known to have substructure. Thus, these particles are the elementary building blocks of matter.

Table 1.1: The fundamental particles of matter within the Standard Model.

	1st Generation	2nd Generation	3rd Generation	Charge (e)
Quarks	up (u) down (d)	charm (c) strange (s)	top (t) bottom (b)	+2/3 -1/3
Leptons	electron (e) e neutrino (ν_e)	muon (μ) μ neutrino (ν_μ)	tau (τ) τ neutrino (ν_τ)	-1 0

The first generation makes up the visible matter in the Universe. The up and down quarks are the valence quarks for the proton and neutron. These form the nuclei of elemental atoms, with electrons occupying orbital shells outside the nucleus. The other quarks may also form bound states similar to the proton and neutron. A particle containing three bound quarks is called a *baryon*. The proton and neutron are examples of baryons in nature. Two-quark bound states may also be formed, which are called *mesons*. Mesons are integer-spin particles and obey Bose-Einstein statistics. Collectively, baryons and mesons are called hadrons. Despite the non-integer charge of quarks, all observed hadrons have integer charge.

Interactions between the fundamental particles occur through the exchange of *gauge bosons* (Tab. 1.2). All force carriers are bosons, which means they have integer

spin. The two most familiar forces in the list are the electromagnetic force and gravity. The electromagnetic force is mediated by the neutral and massless photon. The neutrality and masslessness of the photon implies that the electromagnetic force has infinite range.

Table 1.2: The gauge bosons of the Standard Model [2].

Force	Boson	Mass (GeV)	Relative Strength
Strong	gluon (g)	0	1
Electromagnetic	photon (γ)	0	10^{-2}
Weak	$W^{\pm}, Z^0$	80.4, 91.2	10^{-6}
Gravity	graviton (G)	0	10^{-30}

The force carrier for the strong force is the *gluon*. Strong force interactions only occur for particles with *color charge*, an innate property of quarks that is analogous to electric charge, but comes in three values (red, blue, green) and their opposites (anti-red, anti-blue, anti-green). Thus, the strong force is responsible for the formation of hadrons, which only exist in color-neutral states, also known as white states. This explains why hadrons have integer charge, since they must have three quarks with three different colors (similarly for anti-quarks with anti-colors), or a quark-antiquark pair with color and anti-color. This means baryons must have electric charge of $0, \pm 1e, \pm 2e$, while mesons must have electric charge of $0, \pm 1e$. The gluon is massless, but is not color-neutral, implying the strong force is a short range force. Another unique result of the gluon color charge is the related concepts of asymptotic freedom and quark confinement [3]. The coupling constant of the strong force decreases with energy and increases with distance, leading to the requirement of nature that quarks must be bound. Bare quarks have never been observed in nature.

The weak force is mediated by two massive bosons called the W and Z. The W has electric charge, while the Z is neutral. The weak force is responsible for both nuclear

β decay and neutrino interactions with matter. Due to the large mass of both weak bosons, the weak force is incredibly short range and very weak.

The Standard Model describes the interactions of the strong, electromagnetic, and weak forces. Gravity, lacking an acceptable quantum description, is not included and is not discussed here further. Though the Standard Model describes multiple theories, it is not a unified theory, since not all interactions are governed by a single force or single set of equations. While there is ongoing work to create such a theory that withstands experimental tests, the Standard Model has shown to be self-consistent and very robust when compared to observation [1].

1.1.2 Mathematical Formulation

The Standard Model is a *quantum field theory*, in which every particle is described by a field ψ in four-dimensional space-time. The kinematics of particles, and any interactions between them, are represented by Lagrangian density functions, $\mathcal{L}$ (often just called the Lagrangian). Such functions are formed by enforcing a postulated set of symmetries that must be obeyed for a given field. Equations of state for the Lagrangian are found using the Euler-Lagrange equation:

$$\frac{\partial}{\partial x_\mu}\left(\frac{\partial \mathcal{L}}{\partial(\partial\psi/\partial x_\mu)}\right) - \frac{\partial \mathcal{L}}{\partial \psi} = 0. \tag{1.1}$$

An example of Equation (1.1) in practice is using the free Dirac field,

$$\mathcal{L} = i\overline{\psi}\gamma_\mu\partial^\mu\psi - m\overline{\psi}\psi, \tag{1.2}$$

to derive the free Dirac equation,

$$(i\gamma_\mu \partial^\mu - m)\psi = 0, \tag{1.3}$$

which describes the motion of a free fermion. Solving Equation (1.3) gives the four Dirac spinors, four-component vectors that describe spin-up and spin-down particles and antiparticles.

A fundamental postulate of the Standard Model is that the dynamics and interactions of particles must be invariant under local gauge transformations. The general form of such a transformation on a field ψ is of the form

$$\psi(x) \to U(x)\psi(x) \equiv e^{i\alpha_a(x)T_a}\psi(x), \tag{1.4}$$

where U is an arbitrary $n \times n$ matrix shown parametrized in its general form. T_a are a complete set of linearly independent, unitary $n \times n$ matrices, known as the *generators* of the group. Additionally, α_a are the group parameters, which are physically interpreted as the coupling strengths. A summation over the suffix a is implied. When requiring local gauge invariance, additional fields need to be introduced which describe particles mediating the interactions.

An example of such a concept is introducing local gauge invariance to *quantum electrodynamics* (QED), which is the quantum field description of the electromagnetic interaction. QED has $U(1)$ group symmetry, which reduces Equation (1.4) to

$$\psi(x) \to e^{i\alpha(x)}\psi(x).$$

However, the Lagrangian given in Equation (1.2) is not invariant under this transformation since the derivative term will now include an additional derivative on the

local gauge,

$$\partial_\mu \psi \rightarrow e^{i\alpha(x)} \partial_\mu \psi + i e^{i\alpha(x)} \psi \, \partial_\mu \alpha.$$

To account for this, the covariant derivative, ∂_μ, must be modified so it transforms covariantly under gauge transformations:

$$D_\mu \psi \rightarrow e^{i\alpha(x)} D_\mu \psi.$$

The modified derivative, D_μ, takes the form

$$D_\mu \equiv \partial_\mu - ieA_\mu, \tag{1.5}$$

with A_μ being a vector field which transforms as

$$A_\mu \rightarrow A_\mu + \frac{1}{e} \partial_\mu \alpha$$

to cancel out the unwanted term in the gauge transformation of the Lagrangian. Introducing the modified covariant derivative to Equation (1.2) gives

$$\mathcal{L} = \overline{\psi} \left(i\gamma^\mu \partial_\mu - m \right) \psi + e\overline{\psi} \gamma^\mu \psi A_\mu, \tag{1.6}$$

which is now invariant under local gauge transformation. A natural result of the invariance requirement is an interaction of a Dirac particle with charge $-e$ with the vector field A_μ (the second term of Equation (1.6)), which is the same interaction an electron experiences with a photon field. Thus, the requirement of local gauge invariance for a system with $U(1)$ group symmetry necessitates the QED interaction. A kinetic energy term for the photon field may be introduced into the Lagrangian

using the electromagnetic field tensor,

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu, \tag{1.7}$$

which is local gauge invariant, and results in a kinetic energy term of

$$\mathcal{L}_{Kin} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu}.$$

Note that an attempt to include a mass term for the photon field of the form

$$\mathcal{L}_{mass} = \frac{1}{2} m A_\mu A^\mu$$

would break the gauge invariance of the Lagrangian, implying that the photon field must be massless. This is in agreement with expectations from relativity, since the photon field moves at the speed of light.

Further interesting results can be seen when requiring *quantum chromodynamics* (QCD), the quantum description of the strong force, must also be locally gauge invariant. The theory of QCD is described by the $SU(3)_C$ symmetry group, where C denotes the three possible color charges. This causes the T_a in Equation (1.4) to become 3×3 matrices, and for the summation over a to run from $a = 1, \ldots, 8$. The QCD symmetry group is also non-Abelian because the commutation relationships for the T_a matrices is given as

$$[T_a, T_b] = i f_{abc} T_c, \tag{1.8}$$

with f_{abc} being the structure constants of the group.

Because the free Lagrangian for quarks will have a similar form to that given in Equation (1.2), enforcing the requirement of local gauge invariance will lead to

a similar mishap that occurred for QED. Following a similar path to the one given above requires the introduction of eight gauge fields, G^a_μ which must transform as

$$G^a_\mu \to G^a_\mu - \frac{1}{g}\partial_\mu \alpha_a - f_{abc}\alpha_b G^c_\mu, \tag{1.9}$$

taking Equation (1.8) into account. Redefining the covariant derivative as

$$D_\mu = \partial_\mu + igT_a G^a_\mu$$

gives

$$\mathcal{L} = \overline{q}(i\gamma^\mu \partial_\mu - m)q - g(\overline{q}\gamma^\mu T_a q)G^a_\mu - \frac{1}{4}G^a_{\mu\nu}G^{\mu\nu}_a, \tag{1.10}$$

with

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu - g f_{abc} G^b_\mu G^c_\nu. \tag{1.11}$$

The final term of Equation (1.11) introduces the interesting property of self-interactions among the eight gauge bosons of QCD, which are the gluons. This implies that gluons must also have color charge as a direct result of the non-Abelian nature of the symmetry group.

A major hallmark of the Standard Model is the unification of the electromagnetic and weak interactions into a single theory, the *electroweak* interaction, by Weinberg [4] and Salam [5]. The weak interaction is a parity-violating interaction with two charged and one neutral mediating bosons. The two charged currents have the form (using the first generation of leptons)

$$\begin{aligned} J^+_\mu &= \bar{\nu}\gamma_\mu \frac{1}{2}(1-\gamma^5)e = \nu_L \gamma_\mu e_L, \\ J^-_\mu &= \bar{e}\gamma_\mu \frac{1}{2}(1-\gamma^5)\nu = \bar{e}_L \gamma_\mu \nu_L. \end{aligned} \tag{1.12}$$

If the doublet

$$\chi_L = \begin{pmatrix} \nu \\ e^- \end{pmatrix} \tag{1.13}$$

is introduced along with the traditional "step-up" and "step-down" operators of $SU(2)$ groups, $\tau_\pm = \frac{1}{2}(\tau_1 \pm i\tau_2)$, Equation (1.12) can be rewritten as

$$J_\mu^\pm = \bar{\chi}_L \gamma_\mu \tau_\pm \chi_L, \tag{1.14}$$

with $\tau_{1,2,3}$ being the Pauli matrices. This suggests a possible $SU(2)_L$ symmetry, with the L subscript denoting coupling only with left-handed fermions. Thus, there is a "weak isospin" triplet of weak currents, written as

$$J_\mu^i = \bar{\chi}_L \gamma_\mu \frac{1}{2} \tau_i \chi_L, \tag{1.15}$$

with $i = 1, 2, 3$. Each of these currents is associated with a field W_μ^i related to the observed charged fields by

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp i W_\mu^2). \tag{1.16}$$

A Feynman diagram for such a current is shown in Figure 1.1.

The remaining field W_μ^3 represents a neutral weak current. However, it cannot be attributed to the known Z^0 boson since Z interactions contain a component of right-handed interactions. Another neutral current which has both right-handed and left-handed components is the electromagnetic interaction. In an attempt to save the $SU(2)_L$ symmetry found above, the Z boson and the photon are used to form two orthogonal fields which have definite transformation properties under $SU(2)_L$. These

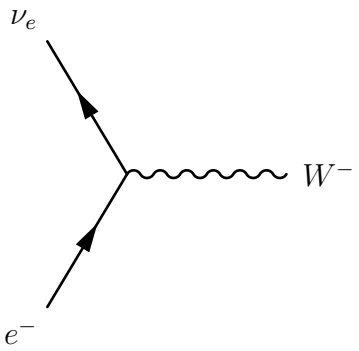

Figure 1.1: Leading order Feynman diagram for the weak charged current.

fields have the form

$$\begin{aligned} A_\mu &= W_\mu^3 \sin\theta_W + B_\mu \cos\theta_W, \\ Z_\mu &= W_\mu^3 \cos\theta_W - B_\mu \sin\theta_W, \end{aligned} \tag{1.17}$$

with θ_W being the Weinberg angle, which must be determined experimentally. The field W_μ^3 is the neutral field that is part of the weak isospin triplet. The newly introduced field B_μ corresponds to the weak hypercharge current, with hypercharge (Y) defined as

$$Q = T^3 + \frac{Y}{2}, \tag{1.18}$$

where Q is the electric charge of the fermion, and T^3 is the "z-component" of the weak isospin. The hypercharge field, similarly to the photon field, has $U(1)_Y$ symmetry. So, the electroweak interaction has $SU(2)_L \otimes U(1)_Y$ symmetry. The Lagrangian for the electroweak interaction is given by

$$\mathcal{L}_{EW} = \bar{\chi}_L \gamma^\mu \left(i\partial_\mu - g\frac{1}{2}\tau \cdot W_\mu - g'\frac{Y}{2}B_\mu \right) \chi_L + \bar{\chi}_R \gamma^\mu \left(i\partial_\mu - g'\frac{Y}{2}B_\mu \right) \chi_R, \tag{1.19}$$

with $\chi_{L,R}$ being the left-handed doublets and right-handed singlets. The coupling

constants g and g' for the respective W_μ and B_μ fields are related by

$$g \sin\theta_W = g' \cos\theta_W = e, \tag{1.20}$$

with e being the absolute value of the electric charge of an electron.

The resulting electroweak symmetry forms left-handed $SU(2)$ doublets exactly given by the quark and lepton generations of the Standard Model (Tab. 1.1). Right-handed singlets are also formed from the individual particles (excluding neutrinos, which are only left-handed [6]), and do not transform under $SU(2)$. An interesting feature of the Standard Model is the weak isospin quark doublets do not exactly correspond to the strong eigenstates. The symmetry of the electroweak interaction implies that flavor-changing currents only work within generations. However, quark decays between generations are observed. This implies that the weak quark doublet states must be somehow represented as superpositions of the strong eigenstates. If the quark doublets are written as

$$\begin{pmatrix} u \\ \tilde{d} \end{pmatrix}, \begin{pmatrix} c \\ \tilde{s} \end{pmatrix}, \begin{pmatrix} t \\ \tilde{b} \end{pmatrix}, \tag{1.21}$$

then the weak states $\tilde{d}, \tilde{s}, \tilde{b}$ correspond to the strong eigenstates by

$$\begin{pmatrix} \tilde{d} \\ \tilde{s} \\ \tilde{b} \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}. \tag{1.22}$$

This is the CKM matrix, developed by Cabibbo [7], and Kobayashi and Maskawa [8]. It is a unitary matrix with a complex phase, allowing for CP violation in the Standard

Model. Magnitudes of the CKM matrix elements are determined experimentally, and indicate that weak flavor transitions within generations (diagonal terms) are favored, while transitions across generations (off-diagonal terms) are suppressed.

An additional component needs to be included in the Standard Model Lagrangian to incorporate the mass of the $W^{\pm}$ and Z^0 bosons, as well as the fermions. Simply introducing mass breaks the coveted local gauge invariance of the theory. To circumvent this problem, the Higgs mechanism [9] is used to introduce mass through the process of spontaneous symmetry breaking. The Higgs mechanism starts by introducing a scalar field ϕ which interacts with the electroweak fields, represented by

$$\mathcal{L}_H = \left| \left(i\partial_\mu - g\frac{1}{2}\tau \cdot W_\mu - g'\frac{Y}{2}B_\mu \right) \phi \right|^2 - V(\phi), \tag{1.23}$$

where $V(\phi)$ is the Higgs potential

$$V(\phi) = \mu^2 \phi^\dagger \phi + \lambda(\phi^\dagger \phi)^2. \tag{1.24}$$

The field ϕ is described by a complex isopsin doublet

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}. \tag{1.25}$$

For $\mu^2 < 0$ and $\lambda > 0$, the Higgs potential of Equation (1.24) has a local maximum at $\phi = 0$ and local minima at

$$\frac{1}{2}(\phi_1^2 + \phi_2^2 + \phi_3^2 + \phi_4^2) = -\frac{\mu^2}{2\lambda} \equiv \frac{v^2}{2}. \tag{1.26}$$

To introduce mass in a way that preserves local gauge invariance, the symmetry of

the Higgs potential must be broken by choosing a particular minimum:

$$\phi_1 = \phi_2 = \phi_4 = 0,\ \phi_3 = v,$$

resulting in

$$\phi_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}. \tag{1.27}$$

Applying Equation (1.27) to the relevant terms of Equation (1.23), and using Equations (1.16), (1.17), and (1.20), gives

$$\left| \left(-ig\frac{1}{2}\tau \cdot W_\mu - i\frac{g'}{2}B_\mu \right) \phi \right|^2 = \left(\frac{1}{2}vg \right)^2 W_\mu^+ W^{-\mu} + \frac{1}{2}\left(\frac{1}{2}v\sqrt{g^2+g'^2} \right)^2 Z_\mu Z^\mu + 0 A_\mu A^\mu. \tag{1.28}$$

From this, $M_W = \frac{1}{2}vg$, $M_Z = \frac{1}{2}v\sqrt{g^2+g'^2}$, and $M_A = 0$. The massless photon is a direct result of the choice of the Higgs potential minimum.

The Higgs field can also be used to generate masses for the quarks and charged leptons through Yukawa interactions between the scalar Higgs field (ϕ) and the particle field (ψ). This has the form of

$$\mathcal{L}_Y = -G_Y \bar{\psi} \phi \psi. \tag{1.29}$$

The factor G_Y is the Yukawa coupling of the particle field to the scalar field. Performing the calculation similar to that of Equation (1.23) gives $G_Y \propto m_\psi$, or the strength of the particle coupling to the Higgs field is proportional to the particle mass. With the top quark being the most massive known elementary particle, this means the study of the top quark has major indirect importance to the study of the Higgs sector.

1.1.3 Precision Tests of the Standard Model

The Standard Model can be used to predict the values of many physical quantities. Some of these quantities may be calculated directly, such as the masses and decay widths of the weak bosons. Many of these quantities, however, are calculated through their effects on radiative corrections to the theory. These quantities include the top quark and Higgs boson masses, as well as many other quantities outside the scope of this thesis. These predictions can be tested against experimental measurements to determine the quality of the Standard Model as a theory to describe direct observation. Figure 1.2 shows the comparison of the theoretical to actual values using a quantity called the pull, which is the difference between the measured and theoretical fit values, divided by the measured uncertainty. There is very good agreement between theory and observation, with many theoretical calculations agreeing with experimental measurements within one standard deviation. Quantities with discrepancies beyond one standard deviation point to potential deficiencies in the Standard Model, and possible new physics. However, none of the discrepancies reach the level of three standard deviations required for evidence.

1.2 The Top Quark

The study of the top quark is relatively new. Its existence was first theorized by Kobayashi and Maskawa [8] in 1973, and was first observed by the CDF [10] and D0 [11] experiments at the Tevatron at Fermilab in 1995. The study of the top quark has become important because it provides a window to possible new physics. The top quark provides an indirect method for constraining theoretical models (such as the Higgs boson mass), a direct method for searching for new particles using top quark resonances, and is an irreducible background for many new physics searches. It is

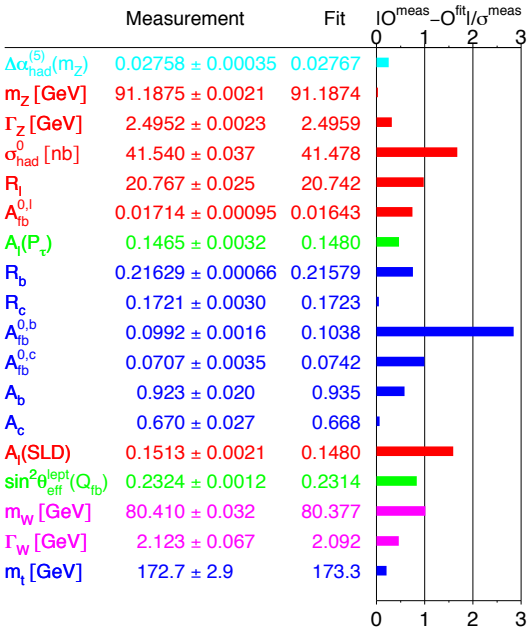

Figure 1.2: Comparison of the theoretical predictions of the Standard Model to experimental observation [1].

necessary to further study top quark physics at the LHC to aid in the study of new physics. This section will discuss the properties of the top quark, its production, and its physics signatures through decay.

1.2.1 Properties of the Top Quark

Table 1.3 shows the currently known properties of the top quark. The top quark mass is measured at the 0.6% level, and is by far the most precisely measured quark mass. The quick decay time implies that top quarks will decay before the hadronization process occurs.

Table 1.3: Properties of the top quark (t) [2].

Property	Value
Mass	173.5 ± 0.6 (stat.) ± 0.8 (syst.) GeV
Width	$2.0 + 0.7 - 0.6$ GeV
Charge	$+\frac{2}{3}e$
$\lvert V_{td}\rvert$	$(8.4 \pm 0.6) \times 10^{-3}$
$\lvert V_{ts}\rvert$	$(42.9 \pm 2.6) \times 10^{-3}$
$\lvert V_{tb}\rvert$	0.89 ± 0.07

1.2.2 $t\bar{t}$ Production

At hadron colliders, like the LHC, top quark pair production occurs predominantly via the strong interaction. The leading order (LO) processes are gluon fusion and quark-antiquark annihilation (Fig. 1.3). Since the LHC collides protons, the antiquarks will come from the sea quarks of the proton. Electroweak interactions may also produce $t\bar{t}$ pairs, but these contributions will be negligible compared to the strong processes at the LHC.

The theoretical calculation for the $t\bar{t}$ production cross section (Tab. 1.4) may be done using the parton model, in which each proton contains quasi-free partons that share the proton longitudinal momentum p_A. So, parton i with longitudinal momentum p_i carries the fraction $x_i = p_i/p_A$. Using the factorization theorem [12], which allows the cross section calculation to be performed using perturbation theory, the production cross section may be written as [13]

$$\sigma(pp \to t\bar{t}) = \sum_{i,j} \int dx_i \, dx_j f_{i,p}(x_i, \mu^2) f_{j,p}(x_j, \mu^2) \cdot \hat{\sigma}_{ij}(ij \to t\bar{t}; \hat{s}, \mu^2). \tag{1.30}$$

The parton distribution functions $f_{i,p}$ represent the probability density for finding a parton i inside the proton carrying longitudinal momentum fraction x_i. The parton distribution functions and the parton-parton cross section $\hat{\sigma}_{ij}$ depends on the

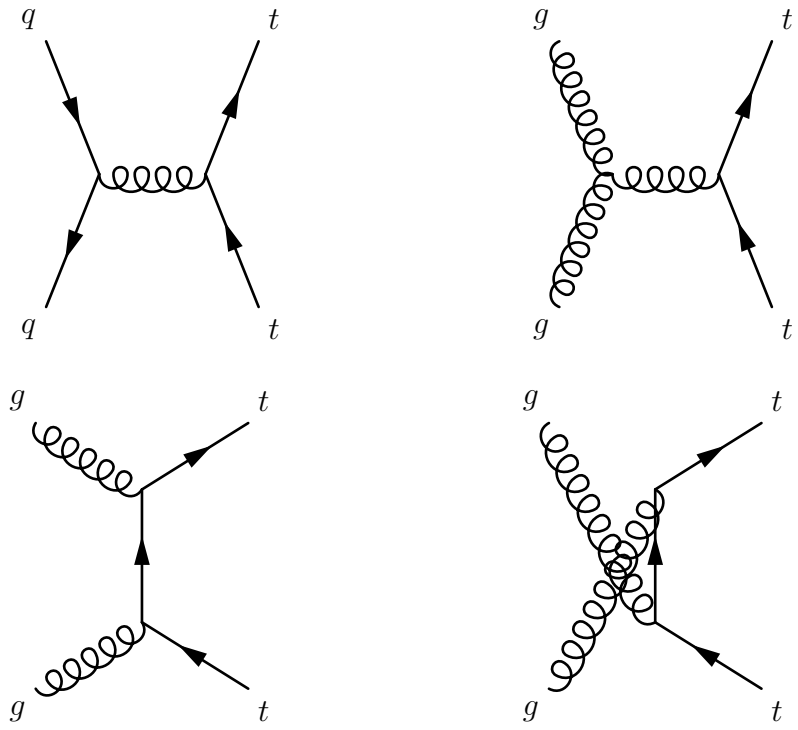

Figure 1.3: Leading order Feynman diagrams for $t\bar{t}$ production, showing quark-antiquark annihilation (top left) and gluon-gluon fusion (top right and bottom).

factorization and renormalization scale μ, an unphysical quantity which allows the calculation of the cross section to finite order. For calculating the production of heavy quark pairs, μ is typically set to m_t, the largest quark mass. Cross section predictions are generally calculated to finite order, which means the resulting calculation is dependent on the choice of μ. The effect of this dependence is tested by varying the parameter $\mu \to 2\mu, \mu/2$ and reperforming the calculation.

The parton-parton cross section $\hat{\sigma}$ can be calculated using perturbative QCD. The

differential cross section for the gluon fusion process is

$$\frac{d\hat{\sigma}}{d\hat{t}}(g_1 g_2 \to t\bar{t}) = \frac{\pi\alpha_s^2}{8\hat{s}^2}\left[\frac{6(m_t^2-\hat{t})(m_t^2-\hat{u})}{\hat{s}^2} - \frac{m_t^2(\hat{s}^2-4m_t^2)}{3(m_t^2-\hat{t})(m_t^2-\hat{u})} + \frac{4}{3}\frac{(m_t^2-\hat{t})(m_t^2-\hat{u})-2m_t^2(m_t^2+\hat{t})}{(m_t^2-\hat{t})^2} + \frac{4}{3}\frac{(m_t^2-\hat{t})(m_t^2-\hat{u})-2m_t^2(m_t^2+\hat{u})}{(m_t^2-\hat{u})^2} - 3\frac{(m_t^2-\hat{t})(m_t^2-\hat{u})-m_t^2(\hat{u}-\hat{t})}{\hat{s}(m_t^2-\hat{t})^2} - 3\frac{(m_t^2-\hat{t})(m_t^2-\hat{u})-m_t^2(\hat{t}-\hat{u})}{\hat{s}(m_t^2-\hat{u})^2}\right], \tag{1.31}$$

where $\hat{s}, \hat{t}, \hat{u}$ are the traditional invariant Mandelstam variables. The differential cross section for the quark-antiquark annihilation process is

$$\frac{d\hat{\sigma}}{d\hat{t}}(q\bar{q} \to t\bar{t}) = \frac{4\pi\alpha_s^2}{9\hat{s}^4}[(m_t^2-\hat{t})^2 + (m_t^2-\hat{u})^2 + 2m_t^2\hat{s}]. \tag{1.32}$$

At energies near the top pair kinematic threshold, $s = 4M_{top}^2$, the $q\bar{q}$ annihilation mechanism becomes the dominant source when the incident quarks are the valence quarks. At the Tevatron, a $p\bar{p}$ collider located at FNAL in Batavia, IL, 80-90% of the $t\bar{t}$ production was due to $q\bar{q}$ annihilation. At energies much higher than the kinematic threshold, the gluon-gluon fusion process dominates for both $p\bar{p}$ and pp collisions. Therefore, the gluon fusion process is the predominant source of top quark pairs at the LHC, a pp collider [13].

1.2.3 Top Quark Decay

In the Standard Model, the top quark decays predominantly by $t \to W^+ b$, with weak decays to the s and d quarks being CKM-suppressed by factors of $|V_{ts}|^2$ and $|V_{td}|^2$.

Table 1.4: Approximate NNLO $t\bar{t}$ production cross section versus center of mass energy for pp collisions, with scale and pdf uncertainties [14].

Energy	$\sigma\pm$ (scale) $\pm$ (pdf) (pb)
7 TeV	$163^{+7}_{-5} \pm 9$
8 TeV	$234^{+10}_{-7} \pm 12$
14 TeV	920^{+50+33}_{-39-35}

The top quark decay width, including first order QCD corrections, is given by [13]

$$\Gamma_t = \frac{G_F m_t^3}{8\pi\sqrt{2}} |V_{tb}|^2 \left(1 - \frac{m_W^2}{m_t^2}\right)^2 \left(1 + 2\frac{m_W^2}{m_t^2}\right)^2 \left(1 - \frac{2\alpha_s}{3\pi} f(y)\right), \tag{1.33}$$

with $y = (m_W/m_t)^2$ and $f(y) = 2\pi^2/3 - 2.5 - 3y + 4.5y^2 - 3y^2 \ln y$. Using the world averages for all masses and constants gives $\Gamma_t = 1.76$ GeV, which corresponds to $\tau_t \approx 5 \times 10^{-25}$ s. The extremely short decay time of the top quark means the decay occurs before hadronization can take place. Thus, the top quark has never been observed in a bound state. This is one reason why the top quark mass can be measured with good precision.

For top quark pairs, the decay $t\bar{t} \to W^+W^-b\bar{b}$ is the predominant decay channel, due to the relative size of $|V_{tb}|$ to $|V_{ts,d}|$. The W boson will decay leptonically approximately one-third of the time and hadronically the remaining two-thirds (Figure 1.4). This leads to three distinct categories of $t\bar{t}$ events.

- All hadronic (≈ 45 %): both W bosons decay hadronically, producing all jets in the final state. This state provides all the kinematic information of the $t\bar{t}$ pair, but has a large background, and is heavily dependent on the quality of the jet reconstruction and energy measurement.

- Semileptonic (≈ 44 %): one W boson decays to a lepton (e, μ, τ) and a neutrino, while the other decays hadronically. This state provides nearly as much

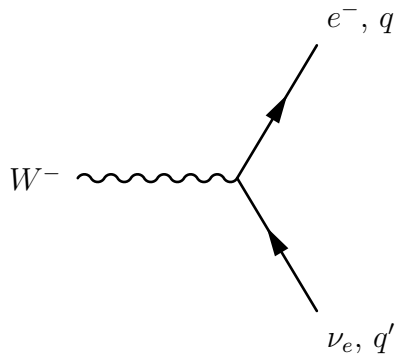

Figure 1.4: Leading order Feynman diagram for W boson decay.

statistics as the all hadronic channel, but loses kinematic information due to the neutrino from the W decay. Reconstruction of the τ channel can also be difficult, since it decays hadronically most of the time, and mimics the e and μ channels when it decays leptonically.

- Dileptonic (≈ 11 %): both W bosons decay to leptons. This state has the least statistics compared to the others, and has two neutrinos in the final state, but has the highest purity compared to the higher statistics channels.

This thesis describes the search for top pair events using the semileptonic channel, specifically where the lepton in the event is a muon (Fig. 1.5). Approximately 14 % of $t\bar{t}$ events decay in this manner. This particular channel was chosen due to the larger available statistics, the better quality of the muon reconstruction (Section 3.3), and the reduced Drell-Yan background, compared to the $e+$ jets channel, when a soft electron within a jet is required (See Section 3.5 for definition and discussion of jets).

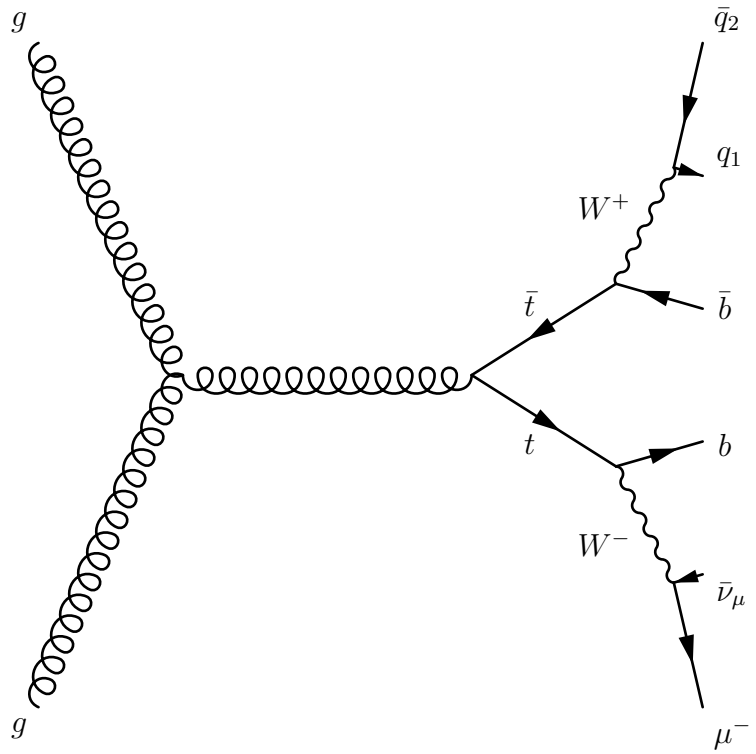

Figure 1.5: Example leading order Feynman diagram for $t\bar{t} \to \mu^- \bar{\nu}_\mu b \bar{b} q_1 \bar{q}_2$.

Chapter 2

Experimental Apparatus

2.1 The Large Hadron Collider

The Large Hadron Collider (LHC) [15] is a two-ring superconducting hadron accelerator and collider installed within a tunnel 26.7 km in circumference at CERN, located in Geneva, Switzerland (Fig. 2.1). It is designed to collide protons with a center-of-mass energy of 14 TeV, with an instantaneous luminosity of 10^{34} cm^{-2} s^{-1}. As of the time of this writing, the center-of-mass energy for proton collisions is 8 TeV with an instantaneous luminosity of 8×10^{33} cm^{-2} s^{-1}. The accelerator also accommodates collisions with lead ions. The LHC has two rings to accommodate two counter-circulating beams of hadrons. Due to space limitations, the two rings are coupled by 1,232 two-in-one dipole magnets (Fig. 2.2) within the same cryostat [15].

The rings cross at four locations, which are the collision points for the four experiments at the LHC: CMS, ATLAS, ALICE, and LHCb (Fig. 2.3). CMS and ATLAS are general purpose experiments while ALICE and LHCb are specialized experiments studying heavy ion collisions and bottom physics, respectively. To achieve full collision energy, protons are accelerated to 25 GeV within the Proton Synchrotron (PS).

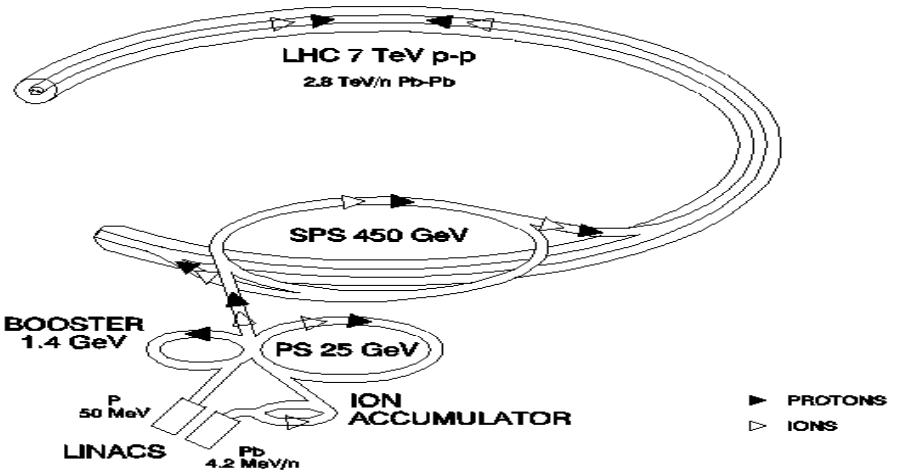

Figure 2.1: Schematic view of the LHC accelerator complex.

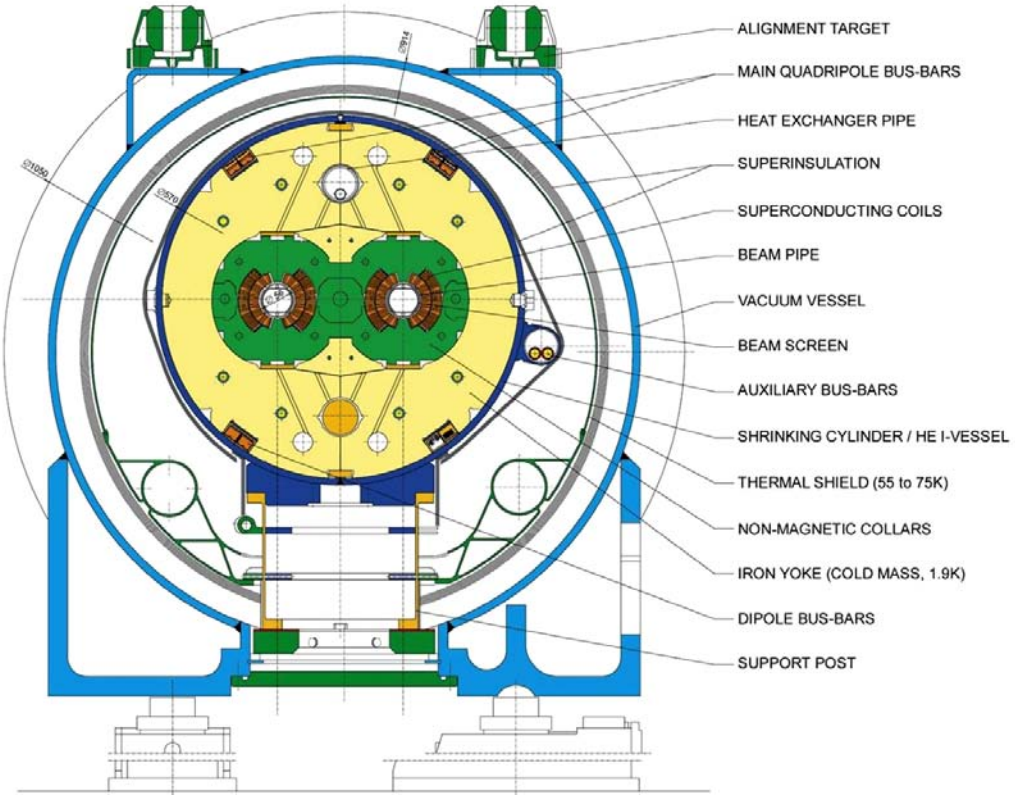

Figure 2.2: The cross-sectional view of an LHC dipole magnet.

Then, they are injected into the Super Proton Synchrotron and accelerated up to 450 GeV, at which point the beams are injected into the LHC rings to achieve the operating center-of-mass energy.

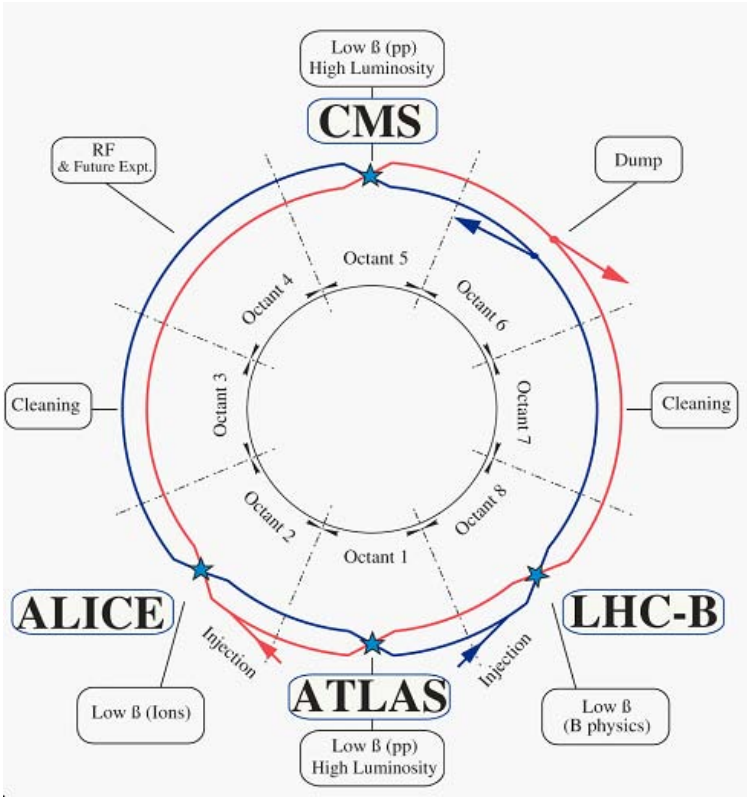

Figure 2.3: Schematic view of the LHC accelerator complex with LHC experiments.

The goal of the LHC is to examine physics which lies beyond the Standard Model by supplying collisions at energies much larger than previously obtained at other accelerators. The number of events per second for a given interaction produced at the collider is given by

$$N_{event} = L\sigma_{event}, \tag{2.1}$$

where σ_{event} is the cross section for the interaction under study, and L is the instan-

taneous luminosity of collisions, given by

$$L = \frac{N_b^2 n_b f_{rev} \gamma_r}{4\pi \epsilon_n \beta^*} F, \tag{2.2}$$

with N_b being the number of particles per bunch, n_b the number of bunches per beam, f_{rev} the revolution frequency, γ_r the relativistic gamma factor, ϵ_n the normalized transverse beam emittance, β^* the size of the beam envelope, and F the geometric reduction factor due to the crossing angle of the bunches at the interaction point:

$$F = \left(1 + \left(\frac{\theta_c \sigma_z}{2\sigma^*}\right)^2\right)^{-1/2} \tag{2.3}$$

with θ_c being the crossing angle, σ_z the RMS bunch length, and σ^* the transverse RMS beam size.

The goal of the LHC is to maximize the luminosity available for the general purpose experiments, CMS and ATLAS. To do this, the LHC has focused on having a large N_b and f_{rev} (often represented using the time between bunch crossings), while reducing ϵ_n and β^*. The major challenges to this are the ability to handle beam-beam interactions during bunch crossings, the mechanical aperture of the LHC arcs, the maximally produceable dipole field before quenching, the ability to handle both the beam energy and heat load should the beam need to be dumped, electromagnetic interactions within the beam, and the turnaround time needed to produce new bunches [15]. This has led to the need to reduce some operating parameters away from design to meet more realistic goals. Table 2.1 shows the values of the operating parameters listed above during a period of 2012 data taking versus design values.

Table 2.1: LHC operating parameters during a period of 2012 data taking compared to LHC design parameters [16].

Parameter	2012	Design
N_b	1.48×10^{11}	1.15×10^{11}
Time between bunches	50 ns	25 ns
ϵ_n	2.6 μm	3.75 μm
β^*	0.6 m	0.55 m

2.2 The CMS Detector

The Compact Muon Solenoid (CMS) [17] is a multi-purpose detector located at the LHC (Fig. 2.4). The primary goal of CMS is to perform discovery searches, notably for the Higgs boson. Located 100 meters underground near the French village of Cessy, CMS is assembled from components produced and tested by scientists all over the globe. At LHC design energies of 14 TeV, the total proton-proton cross section is expected to be 100 mb. Given the design luminosity of 10^{34} cm^{-2} s^{-1}, this leads to an event rate of approximately 10^9 inelastic events per second. Additionally, 20 inelastic pileup collisions, secondary pp interactions within the same bunch crossing, will be superimposed on top of the event of interest, leading to the production of approximately 1000 charged particles from the interaction region at each collision. This requires a detector which can detect individual particles with great resolution and precision at a very rapid pace. The large amount of radiation from each interaction also requires the use of radiation-hard materials and electronics.

The CMS experiment is designed with the following principles in mind to meet the analysis goals of the LHC.

- Good muon reconstruction, with particle identification and momentum resolution over a wide geometric and kinematic range, good dimuon mass resolution ($\approx 1\%$ at 100 GeV), and ability to determine the muon charge with $p < 1$ TeV.

- Good charged-particle momentum resolution and track reconstruction efficiency. Efficient tagging and triggering of b jets and taus, requiring the use of a fine grained pixel detector near the interaction region.

- Good electromagnetic energy resolution, good diphoton and dielectron mass resolution ($\approx 1\%$ at 100 GeV), with wide geometric coverage, rejection of neutral pions, and efficient isolation at high luminosities for electromagnetic particles.

- Good missing transverse energy ($\not\!\!E_T$) and dijet mass resolution, using hadron calorimeters with large geometric coverage and fine lateral segmentation.

The CMS detector design, discussed in the subsequent sections, meets these demands through the use of a high magnetic field solenoid, a fully silicon-based tracking system, and a homogenous crystal-based electromagnetic calorimeter.

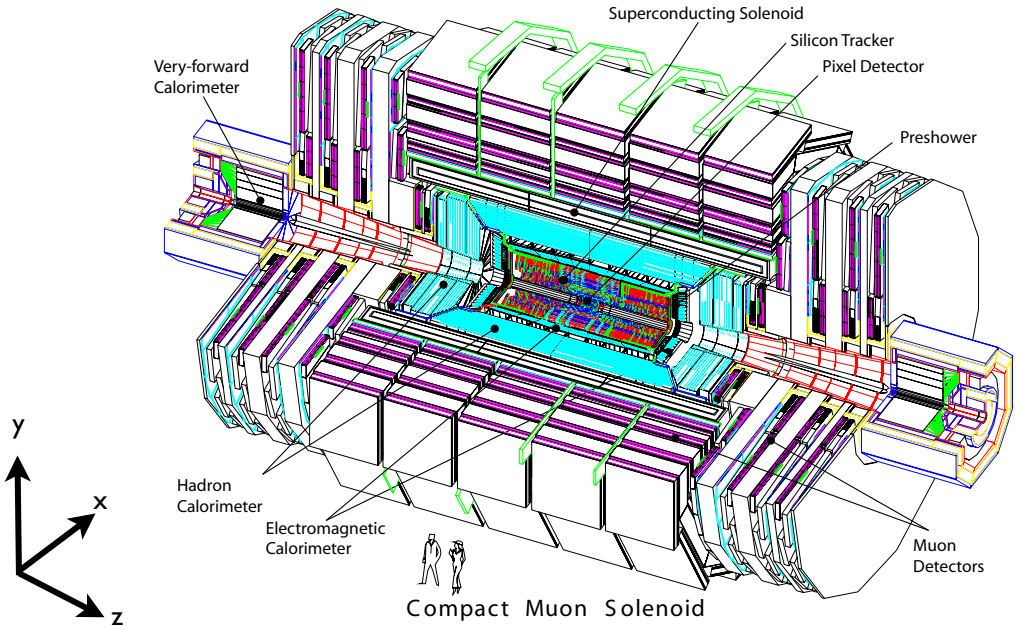

Figure 2.4: Schematic view of the CMS detector.

CMS has adopted a coordinate system with the origin centered at the nominal collision point, the y-axis pointing vertically upward, the x-axis pointing radially inward toward the LHC center, and the z-axis pointing along the beam direction from LHC Point 5 toward the Jura mountains. The azimuthal angle (ϕ) is measured from the x-axis in the $x - y$ plane. The radial coordinate in this plane is denoted as r. The polar angle (θ) is measured from the z-axis. A quantity related to the polar angle, called the pseudorapidity (η), is defined as

$$\eta = -\ln(\tan(\theta/2)). \tag{2.4}$$

Values of η relative to the detector are shown in Figure 2.5. From this, the momentum (p_T) and energy transverse (E_T) to the beam direction are computed from x and y components. The quantity ($\not\!\!E_T$) measures the imbalance in energy measured in the transverse plane, hypothetically due to neutral, non-interacting particles, by vectorially summing the component of energy from reconstructed objects which is transverse to the beam axis and multiplying by a minus sign. This is so the vector points in the opposite direction of the total reconstructed transverse energy, representing any particle produced in the event which was not reconstructed. Since there is no activity from the incident particles that is transverse to the beam, ($\not\!\!E_T$) should ideally be zero, requiring a transverse energy balance.

2.2.1 The Superconducting Solenoid

The solenoid for CMS has been designed to produce a uniform magnetic field of 4 T, though it is limited to 3.8 T during operation to prolong its lifetime. The dimensions of the solenoid are 6 m in diameter and 12.5 m in length, with a full-current stored energy of 2.6 GJ. The magnetic flux is returned through a 10,000 metric ton yoke

comprised of 5 wheels and 2 endcaps, which themselves contain three disks each. The cold mass of the solenoid alone is 220 metric tons in mass, containing four winding layers of a stabilized NbTi conductor. The use of four windings are unique to CMS when compared to magnets used at previous experiments, which use only one winding. The greater number of windings are required to produce such a high magnetic field, which requires 4.2×10^7 Amperes / turn to produce. As a result of the size and high field of the magnet, the ratio between the stored energy and the mass is high (11.6 KJ/kg), causing a large mechanical deformation (0.15 %) while energizing the solenoid. This is much larger than values obtained by previous experiments.

2.2.2 The Tracking System

The tracking system is designed to meet two important goals: provide a precise and efficient trajectory measurement for charged particles from LHC collisions, and provide precise reconstruction of resulting secondary vertices. The tracker surrounds the interaction region, with a length of 5.8 m and diameter of 2.5 m. There are estimated to be approximately 1000 particles passing through the detector at design LHC luminosity, requiring the tracker to have high granularity to provide accurate trajectories with a fast response time. The high flux of particles also requires the use of a radiation-hard material. All these requirements lead to the use of silicon as the basis for the tracking detector.

The tracker is composed of an inner pixel detector and an outer strip detector (Fig. 2.5). The pixel detector (Figure 2.6) is the innermost part of CMS and is nearest to the interaction region, with three barrel layers at mean radii of 4.4, 7.3, and 10.2 cm, and two forward disks at $z = \pm 34.5$ and $z = \pm 46.5$ cm. The barrel cylinder (example layer Figure 2.7) is 53 cm long, and the forward disks (example half-disk Figure 2.8)

extend from 6 to 15 cm in radius. There are a total of 66 million pixels (48 million barrel + 18 million forward) in total covering an area of 1.06 m^2 (0.78 m^2 barrel + 0.28 m^2 forward). Each pixel has dimensions of 100 x 150 x 285 μm^3, providing the fine granularity required for precision trajectory and vertex reconstruction. The pixel detector covers a pseudorapidity range of $-2.5 < \eta < 2.5$, with the barrel and forward components arranged as to provide three tracking points over the most of the pseudorapidity range to provide maximum measurement precision.

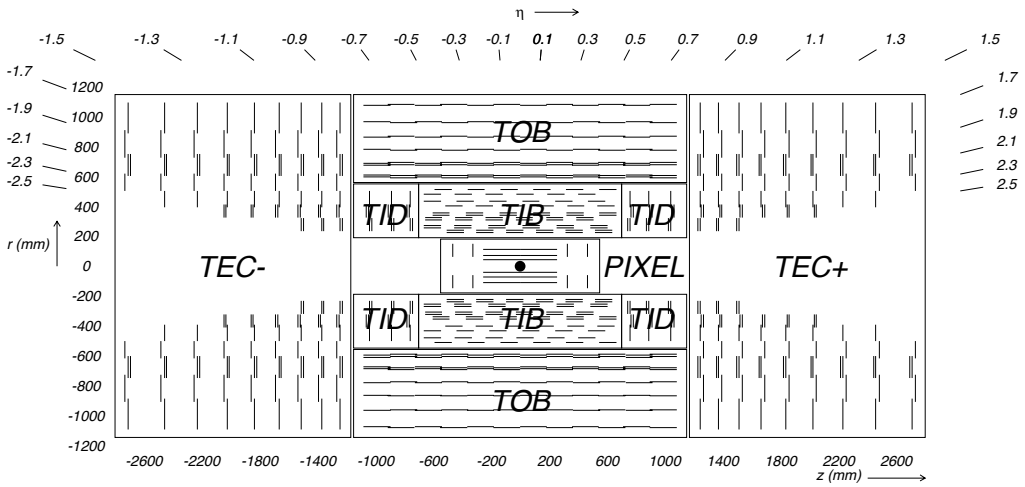

Figure 2.5: A schematic view of the CMS inner tracker, showing both pixel and strip components.

The strip tracker occupies the radial region of the CMS detector between 20 and 116 cm. It is composed of three subsystems: the tracker inner barrel/disks (TIB/TID, Figure 2.9), the tracker outer barrel (TOB, FIgure 2.10), and the tracker end caps (TEC, Figure 2.11). The TIB/TID extend toward 55 cm in radius and is comprised of 4 barrel layers, and three inner disks located at both ends of the barrel. The TIB/TID delivers up to 4 measurements in the $r - \phi$ plane used to contribute in a trajectory calculation using 320 μm thick silicon micro-strip sensors parallel to the beam axis in the barrel, and radial to the beam axis in the disks. In the TIB, the

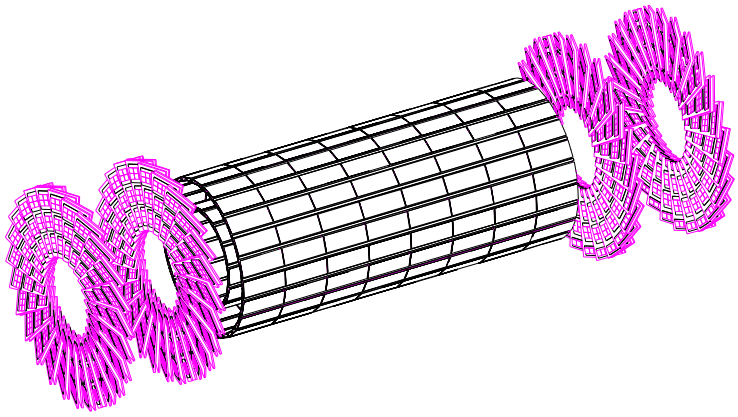

Figure 2.6: The CMS pixel detector in its default configuration [18].

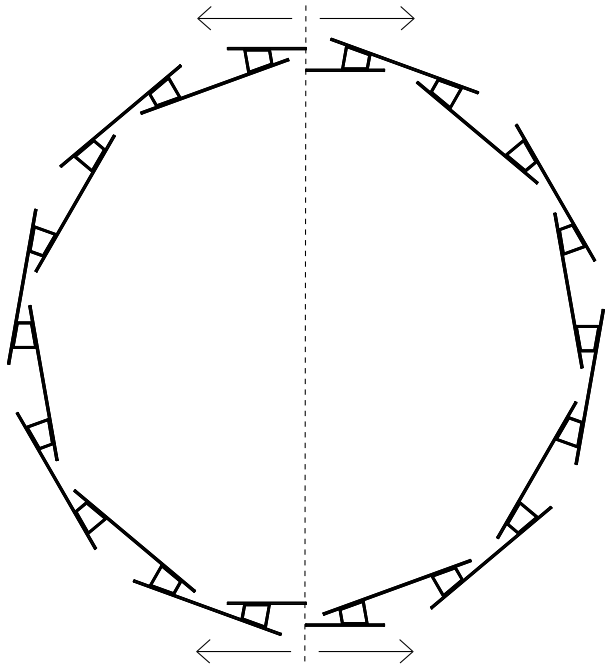

Figure 2.7: Example layer of the CMS pixel barrel detector [18].

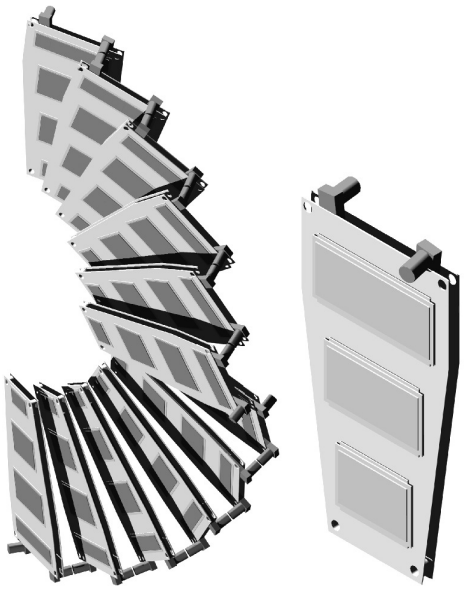

Figure 2.8: Example CMS forward pixel half-disk with module [18].

strips in the first two layers have a pitch of 80 μm, while the strips in the subsequent layers have a pitch of 120 μm. This results in single point resolution of 23 and 35 μm in the respective layers. The pitch in the TID varies from 100 to 141 μm.

The TOB surrounds the TIB/TID system. It has an outer radius of 116 cm and has 6 layers comprised of 500 μm tick micro-strip sensors. The pitches of these strips are 183 μm in the first four layers, and 122 μm in the last two. The additional 6 $r-\phi$ measurements provided by the TOB have single point resolution of 53 μm and 35 μm, respectively. Beyond the 236 cm range of the TOB is the TEC, which are 18 total disks split evenly on either side of the barrel, located at 124 cm $< |z| <$ 282 cm with 22.5 cm $< |r| <$ 113.5 cm. Each TEC disk carries up to 7 rings of micro-strip detectors oriented radially with respect to the beam pipe, having 97 μm to 184 μm average pitch. The micro-strips are 320 μm thick in the inner four rings, and 500 μm thick on the outer rings. The TEC provides up to 9 ϕ measurements for trajectory.

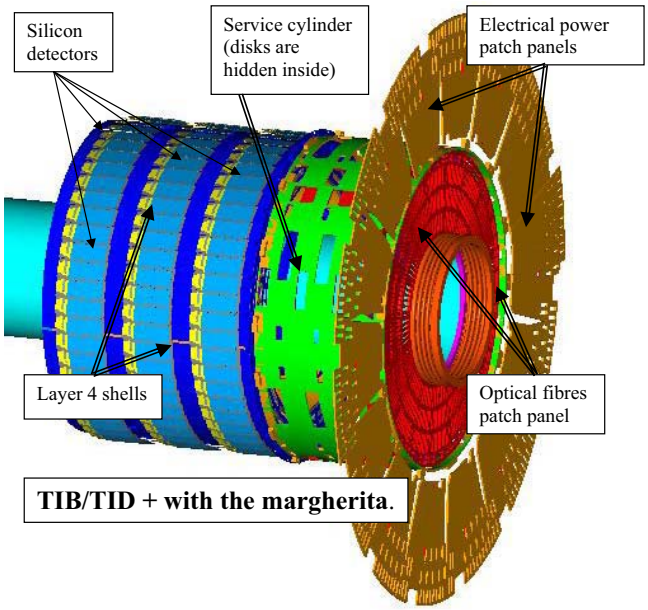

Figure 2.9: The Silicon Tracker Inner Barrel and Inner Disk detectors with service cylinder (margherita).

2.2.3 The Electromagnetic Calorimeter

The electromagnetic calorimeter (ECAL) is a detector designed to measure the energy of particles produced in electromagnetic interactions, specifically electrons and photons (Fig. 2.12). It is designed to be as hermetic and homogeneous as possible, and contains 61,200 lead tungstate ($PbWO_4$) crystals (Figure 2.13) in the barrel, and 7,324 crystals in each of two end caps on either end of the barrel. To facilitate the discrimination between neutral pions and photons in the endcap region, a preshower detector is placed in front of the endcaps. Particles striking the detector produce photons in the crystals that are collected using avalanche photodiodes in the barrel and vacuum photodiodes in the endcaps. Using high density crystals allows for a detector which is fast, has fine granularity in the $\eta - \phi$ plane, and is radiation hard.

The ECAL barrel (Figure 2.14) covers a pseudorapidity range of $|\eta| < 1.479$. The

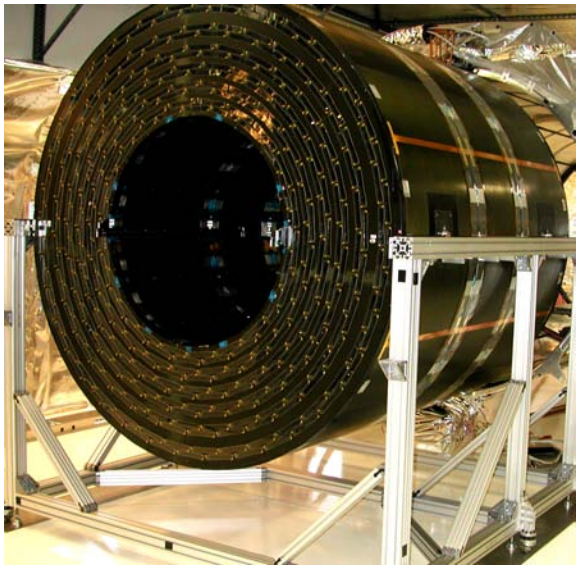

Figure 2.10: The Silicon Tracker Outer Barrel detector.

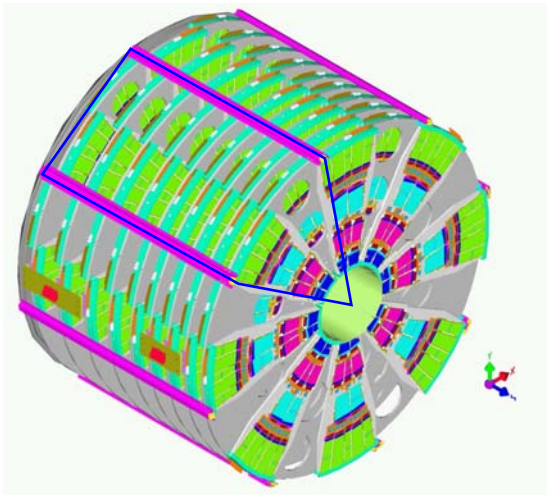

Figure 2.11: The Silicon Tracker End Caps.

center of the front faces of the crystals are at a radius of 1.29 m. The 61,200 crystals in the barrel result in a granularity which is 360-fold in the ϕ direction and (2×85)-fold in the η direction. The crystals have a tapered shape, which varies slightly with pseudorapidity. The crystals are mounted such that their axes make an angle of 3°

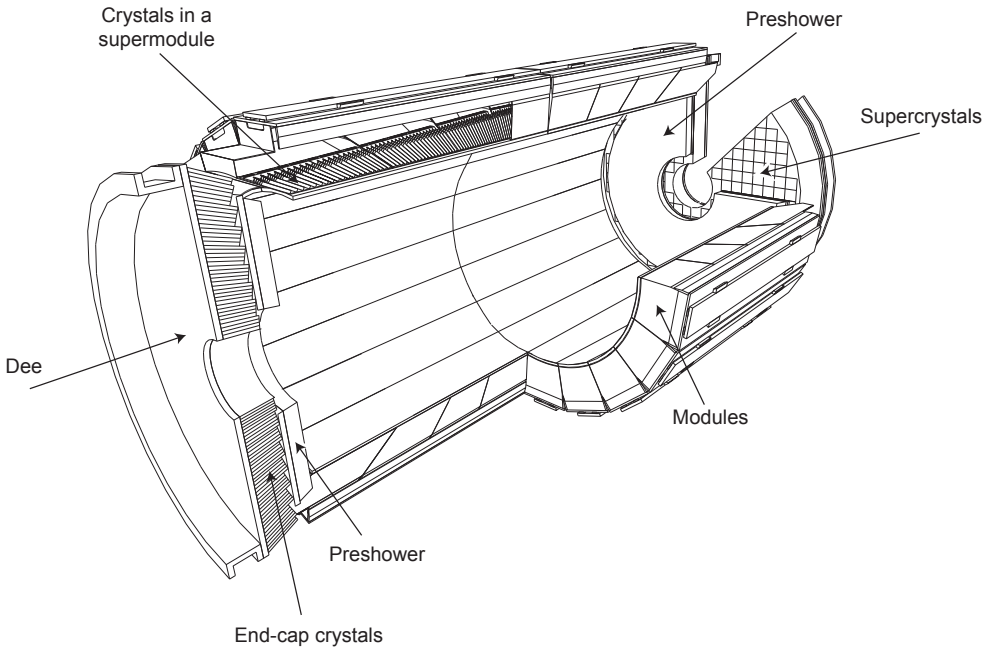

Figure 2.12: A schematic view of the CMS electromagnetic calorimeter (ECAL).

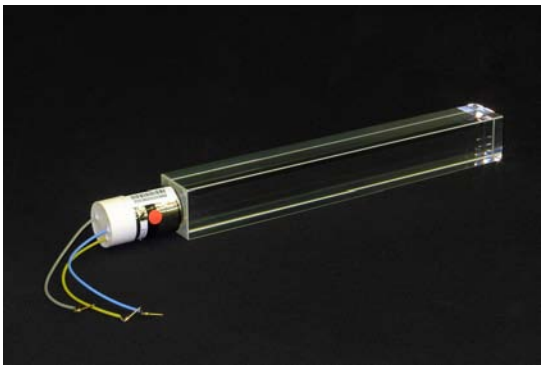

Figure 2.13: A lead tungstate ($PbWO_4$) crystal.

with respect to the nominal interaction vertex in both η and ϕ. This is done to avoid cracks in the detector which allow particles to pass through undetected. The angular cross section of each crystal is approximately 0.0174×0.0174 in $\eta - \phi$ space. The crystal length is 230 mm, corresponding to 25.8 X_0. Given all these dimensions, the total barrel volume is 8.14 m^3 with a total weight of 67.4 metric tons.

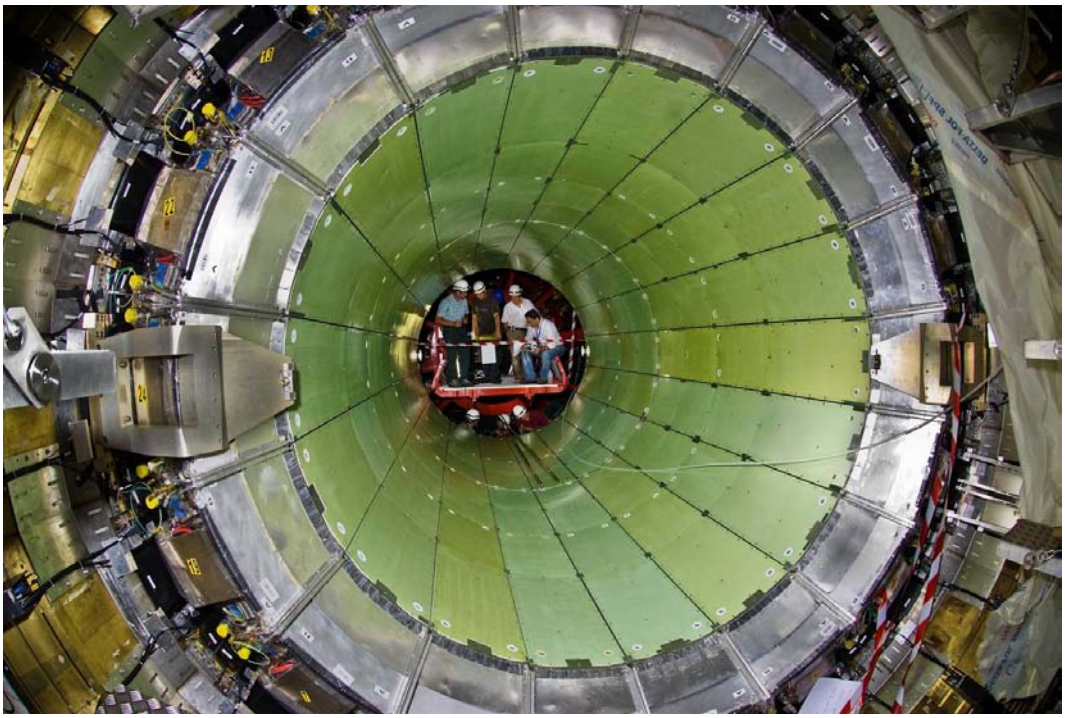

Figure 2.14: The CMS ECAL Barrel.

The ECAL endcap covers the pseudorapidity range of $1.479 < |\eta| < 3.0$. The longitudinal distance between the nominal interaction point and the endcap is 315.4 cm, which takes into account a shift of the endcap toward the interaction point when the magnetic field is turned on. The endcap contains identically shaped crystals grouped mechanically into 5×5 blocks called supercrystals. Each endcap is divided into two halves called dees (Figure 2.15), due to their shape being like the letter "D".

Each dee has 3,662 crystals, which are contained in 138 supercrystals and 18 partial supercrystals located on the inner and outer circumference of the dee. The crystals are arranged in a rectangular grid in the $x - y$ plane. Each crystal has a front face cross section of 28.62×28.62 mm^2, a rear face cross section of 30×30 mm^2, and a length of 220 mm (24.7 X_0). Given these dimensions, the total volume of the end caps is 2.90 m^3 with a total weight of 24.0 metric tons.

Figure 2.15: An example ECAL dee structure.

The final component of the ECAL is the preshower detector, which is used to discriminate between neutral pions and photons in the endcap region by using the energy deposition profile of these particles in the detector. The preshower is a sampling calorimeter with two layers: lead radiators initiate electromagnetic showers from incoming particles, while silicon strip sensors placed behind each radiator measure the

deposited energy and shower profiles. The total preshower thickness is 20 cm. The material thickness of the preshower is 2 X_0 prior to the first sensor plane, and 1 X_0 prior to the second sensor plane. This corresponds to 95% of single incident photons showering before the second sensor plane. The sensor planes are oriented orthogonally with respect to each other. Each silicon sensor has an active area of 61 mm^2 divided into 32 strips. The thickness of the silicon in each sensor is 320 μm.

2.2.4 The Hadronic Calorimeter

The hadronic calorimeter (HCAL) is designed to measure the energy of particles produced in hadronic interactions, specifically long-lived mesons and baryons (Fig. 2.18). The HCAL sits beyond the ECAL in both the barrel and endcaps, and sits within the solenoid. Due to the space limitations between the ECAL barrel and the solenoid (1.77 m $< R <$ 2.95 m), an additional detector called the outer hadronic calorimeter (HO) acts as a tail catcher to catch energetic hadrons which manage to pass through the HCAL and solenoid material. At $|\eta| > 3$, an additional detector called the forward hadron calorimeter (HF) resides at 11.2 m beyond the nominal interaction point, and extend hadronic detection to $|\eta| < 5.2$.

The HCAL barrel (HB, Figure 2.16) is a sampling calorimeter which covers the range of $|\eta| < 1.3$. The HB contains 36 identical azimuthal wedges constructed out of flat brass absorber plates (8 50.5 mm inner plates and 6 56.5 mm outer plates) aligned parallel to the beam axis. Each wedge is divided into four sectors in ϕ. The plates are assembled such that there is no projective dead material for the full radial extent of the wedge. To provide structural strength, the inner and outer plates are made of stainless steel (one 40 mm front plate and one 75 mm back plate). Contained within the layers of absorber plates are plastic scintillators (70,000 total) divided in to 16 η

sectors, resulting in a segmentation of 0.087×0.087 in the $\eta - \phi$ plane. The total absorber thickness at $\theta = 90°$ is 5.82 λ_I. The ECAL barrel provides an additional 1.1 λ_I of material.

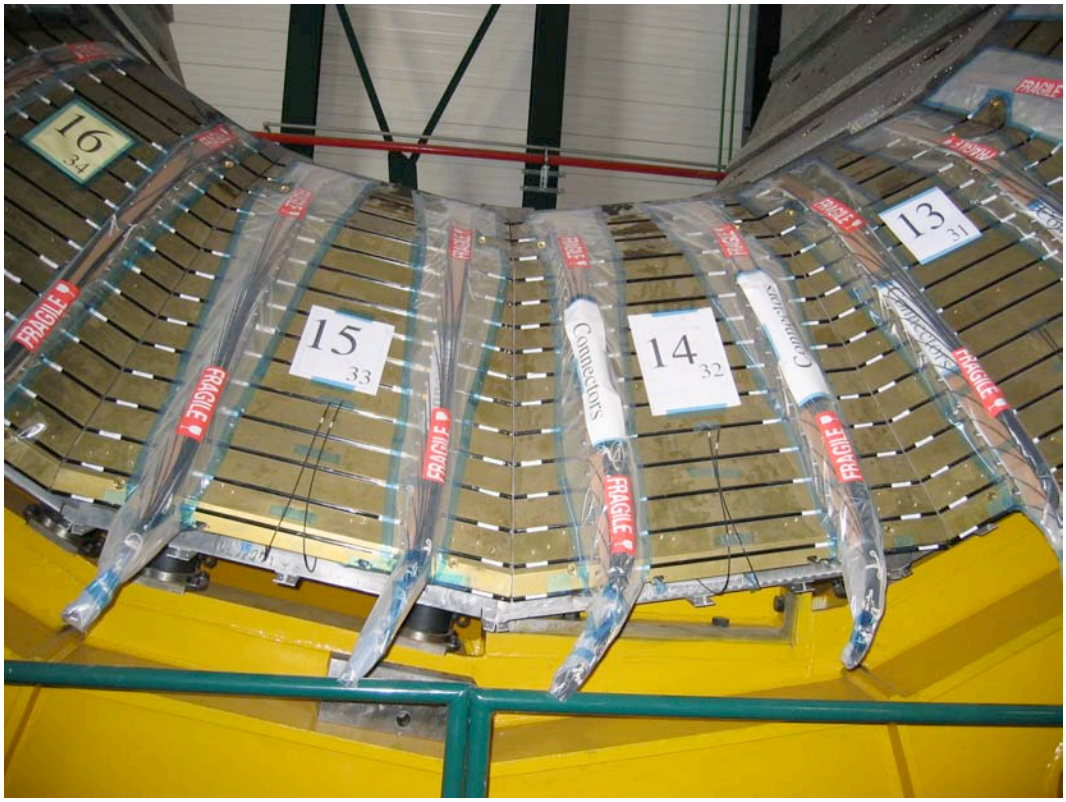

Figure 2.16: A closeup view of the wedges of the HCAL Barrel detector.

The HCAL endcap (HE, Figure 2.17) cover the pseudorapidity range of $1.3 < |\eta| < 3$, a region expected to contain approximately 34% of particles produced in the final state. The HE is sandwiched between the ECAL endcaps in front and the muon endcap yoke in back. Like the HB, the HE also contains brass absorbers, which are 79 mm thick. A spacing of 9 mm between the absorbers is used to accommodate the scintilators (20,916 total). The granularity in $\eta - \phi$ space provided by the HE is 0.087×0.087 for $|\eta| < 1.6$ and 0.17×0.17 for $|\eta| \geq 1.6$. The total length of the

endcap is about 10 λ_I, which includes the ECAL endcap crystals.

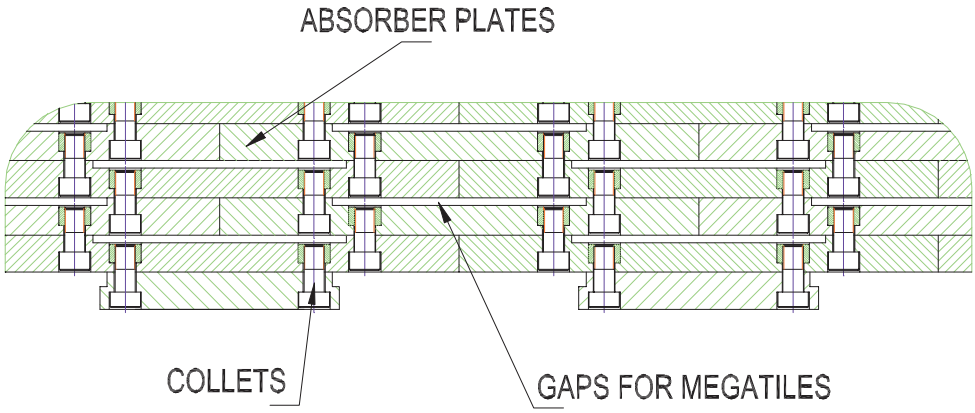

Figure 2.17: A schematic view of the HCAL Endcap detector.

The HO is an additional hadronic detector in the barrel of CMS to detect the energy from hadrons too energetic to be contained within the HB alone. Using the solenoid as additional absorber material, the HO resides outside the magnet to sample the energy of the remaining particles. The HO resides within the iron return yoke, being the first sensitive layer in the 5 rings of the yoke at $r = 4.07$ m. Due to minimal absorber depth at the central yoke ring, an additional HO layer is placed in front of the central ring at $r = 3.82$ m. The HO is comprised of scintillation tiles which provide a granularity equivalent to that of the HB by design.

The HF is the final component of the HCAL, which is located in the pseudorapidity range of $3 < |\eta| < 5.2$. This pseudorapidity range is a very hostile environment compared to the rest of CMS. At the design energy and luminosity of the LHC, 760 GeV of the proton-proton interaction will be deposited in the HF on average, compared to 100 GeV for the rest of the detector. The charged hadron rates are also extremely high, requiring a different detector design than what is used for the rest of the HCAL. Quartz fibers (fused-silica core and polymer hard cladding) are used as the

active medium of the HF due to its fast response and radiation hardness. Cherenkov radiation is used to generate the signal for the hadronic shower when the showering particles are above the Cherenkov threshold ($E \geq 190$ keV for electrons). The quartz fibers have a diameter of $600 \pm 10\,\mu$m for the silica core. In total, over 1000 km of quartz fibers are used in the HF. The HF is a cylindrical steel structure with an outer radius of 130.0 cm. The steel acts as the absorber material, and has the form of 5 mm thick grooved plates. The quartz fibers are inserted within these grooves, and run parallel to the beam line. The total granularity of the HF is 0.175×0.175 in the $\eta - \phi$ plane.

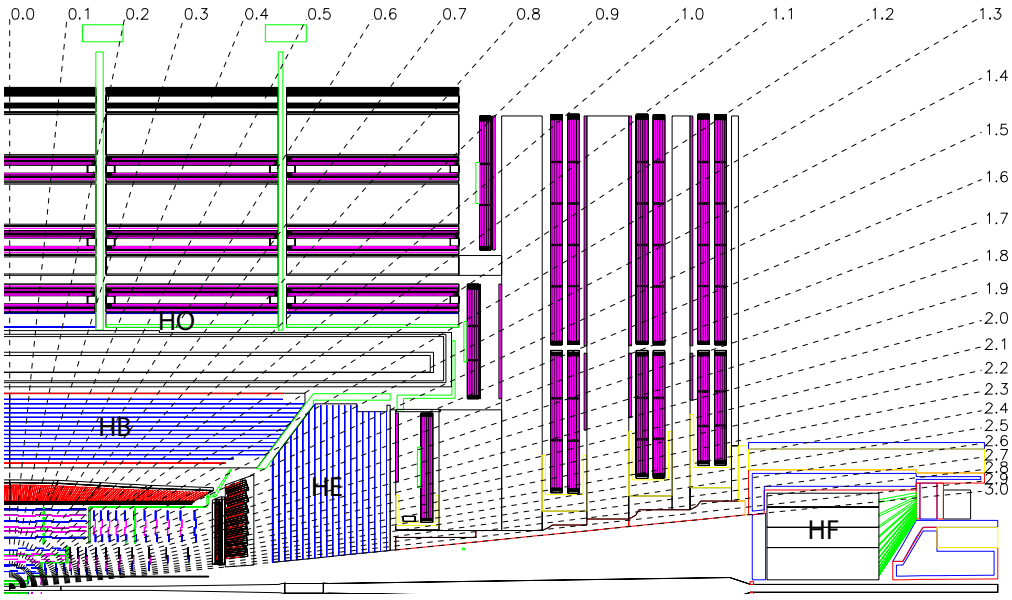

Figure 2.18: A schematic view of the CMS hadronic calorimeter (HCAL), showing the HCAL barrel (HB), HCAL endcap (HE), outer hadronic calorimeter (HO), and the forward hadronic calorimeter (HF).

2.2.5 The Muon System

The muon system, as the name implies, is responsible for the measurement, identification, and triggering of muons. The muon system is the outermost detector at CMS, with the idea being that muons, which are minimum ionizing particles, will pass through the bulk of inner material minimally perturbed, while other particles will be contained within their respective detectors.

As with other detectors, the muon system uses both a barrel and endcap design. The barrel, covering a pseudorapidity range of $|\eta| < 1.2$, contains drift tube chambers (Figure 2.19) organized into four stations. These stations are interspersed among the layers of the flux return plates, meant to act as absorbent material for muon identification. The first three stations contain eight chambers (divided into two groups of four), oriented parallel to the beam axis to perform measurements in the $r - \phi$ bending plane, and an additional four chambers oriented orthogonally to the beam axis, which provide measurements along the z direction. The fourth station is without the z-measuring chambers. The chambers are separated as much as possible to provide the best angular resolution.

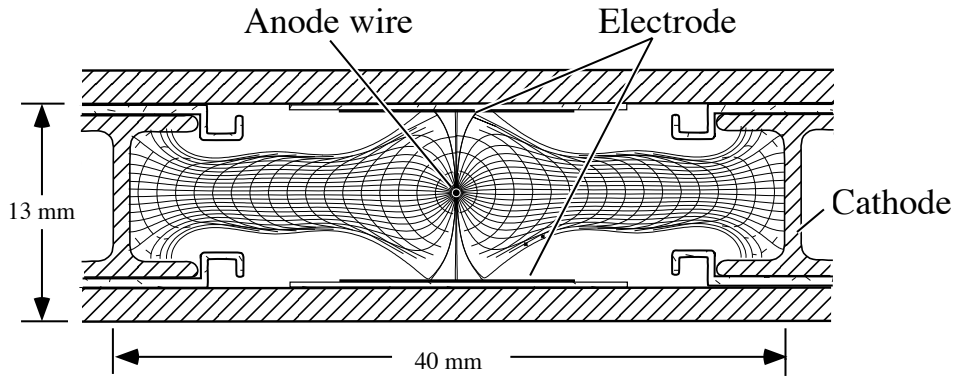

Figure 2.19: A schematic of a drift tube chamber with drift lines and isochrones [19].

The endcaps of the muon system ($0.9 < |\eta| < 2.4$) use cathode strip chambers (CSC, Figure 2.20) to perform measurements. This is due to the higher expected rates of both muons and background, and the non-uniform magnetic field in this region. The CSCs provide fast response times with fine segmentation and resistance to radiation. Each endcap contains four stations of CSCs aligned perpendicularly to the beam line and interspersed among the flux return plates. The CSCs provide precision measurements in the $r - \phi$ bending plane. The anode wires of the CSCs run approximately perpendicular to the strips, and are read out in order to provide measurements of both η and beam-crossing time for the muon.

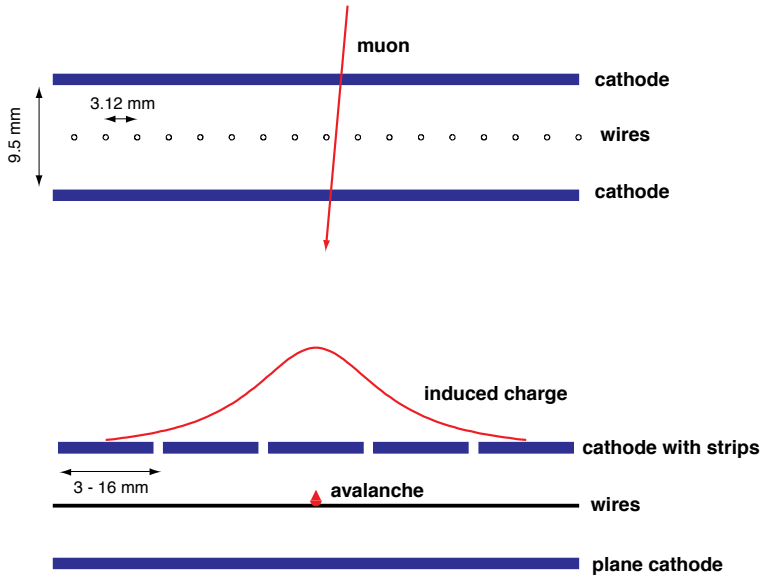

Figure 2.20: A schematic of a cathode strip chamber [19].

An additional system utilizing resistive plate chambers (RPC, Figure 2.21) is also installed in the muon system which acts as a dedicated muon trigger system. The RPCs are useful for triggering as they provide fast, highly-segmented measurements with a sharp p_T threshold over $|\eta| < 1.6$. RPCs are double-gap chambers which

ensure good operation even at high rates. A total of six barrel layers contain RPCs, two in each of the first two stations, and one in each of the last two. The redundancy in the first two station allows for the ability to trigger on low-p_T muons that may stop before reaching the outer stations. In the endcap region, a plane of RPCs is placed in each of the first three stations to allow the trigger to use coincidences between the RPCs and CSCs in order to reduce background and improve time resolution, as well as p_T resolution.

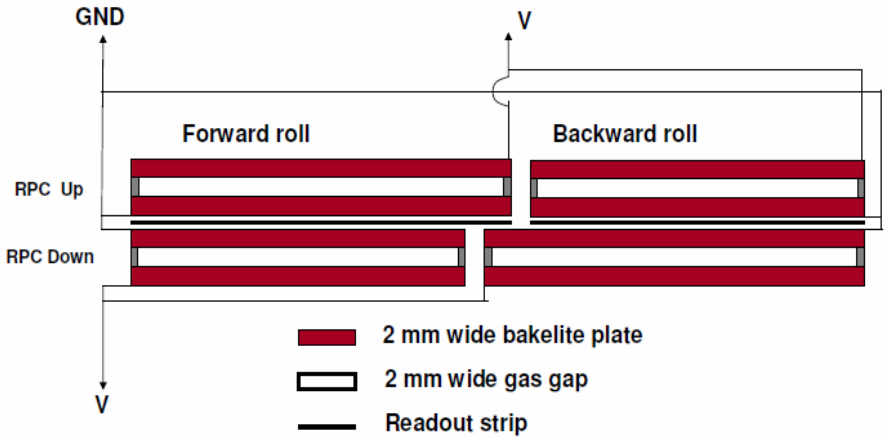

Figure 2.21: A schematic of a resistive plate chamber [20].

2.2.6 Data Acquisition and Triggering

At LHC design energy and luminosity, the total event rate is expected to be too large to be both reconstructed and stored for analysis. Additionally, not all events produced may be classified as interesting for the sake of analysis. To reduce the event rate, and to particularly select those events useful for interesting analysis, a triggering system is needed to immediately analyze data as it comes from the detector. The triggering system at CMS utilizes a two step approach:

- A set of custom-designed, programmable electronics called the Level-1 (L1) trigger, which reduces the event rate to tens of kHz (Fig. 2.22).
- A software system operated on a filter farm containing a huge number of processors, called the High Level Trigger (HLT), which reduces the event rate to approximately 100 Hz.

The L1 trigger uses coarsely segmented data from the muon and calorimeter systems, holding the high resolution data in front end electronics through the use of local, regional, and global components. The local triggers, also called Trigger Primitive Generators (TPG), are based on energy measurements in the calorimeter trigger system along with hit patterns or track segments in the muon trigger system. The regional triggers combine the information from the TPGs using pattern logic to reconstruct ranked and sorted trigger objects, such as electron or muon candidates in limited spatial regions. The ranking includes consideration of energy/momentum calculation and quality determination, which represents the confidence in parameter measurements given knowledge of the detector elements used along with trigger electronics. The Global Calorimeter and Muon Triggers determine the highest rank objects across the whole experiment and transfer this information to the Global Trigger. Ultimately, it is the Global Trigger which decides whether to accept or reject an event at the L1 level, given the information. Upon a positive event evaluation, the L1 information is then sent to the HLT, which uses a set of executables called *trigger paths* to determine the object content (electrons, muons, etc.) of the event at the trigger level. These trigger paths are used to divide the data into datasets which can be used for physics analyses.

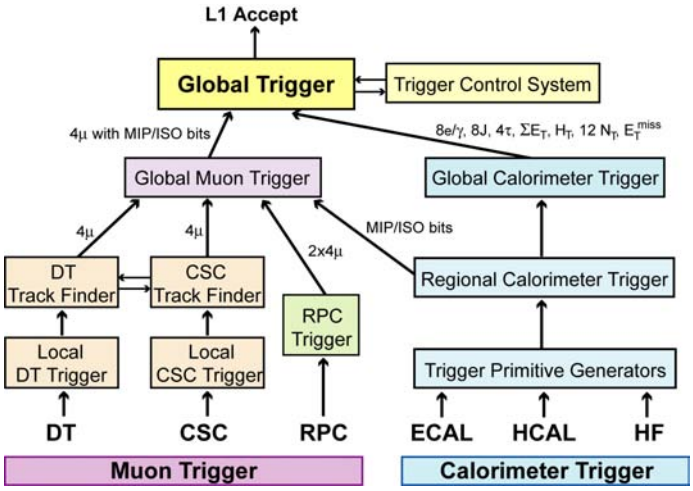

Figure 2.22: The L1 trigger system at CMS.

2.2.7 Computing at CMS

The computing structure for CMS is what allows physics analyses to be performed, and must allow for the storage, transfer, and evaluation of data for the lifetime of the experiment. The computing system must allow for the processing of real-time detector information, provide safe storage of raw data, perform reliable pattern recognition, event filtering, data reduction, and additionally support the production and distribution of simulated samples. This requires a system of large scale that supports efficient operation of data transfer and pattern recognition, flexibility to adapt to the increasing needs and demands of the experiment and its members, but must also be manageable in both construction and maintenance.

To accomplish these goals, CMS uses a combination of a software framework, data formats, and computing. The software framework, known as CMSSW [21], uses a set of compiled plugins called modules, which performs needed tasks on stored data. These modules come in the form of producers, filters, and analyzers. As the names

imply, producers add new information to data that may be needed by future users, filters apply selection criteria to constrain data to be used, and analyzers read in data and perform calculations without any modification of the source. Additionally, information is read in and written out through the use of input and output modules. These framework modules are designed to be insulated from the computing environment on which they run, execute independently of one another, and communicate only through writing to or retrieving from data, rather than to one another.

Data from CMS is stored in a format developed using the ROOT framework [22], developed by physicists and programmers located at CERN. Further, CMS makes use of several different event formats:

- **RAW format**: contains fully recorded detector readout information along with trigger information. The RAW format is used for offline reconstruction to convert detector information into physics objects, such as electrons, muons, photons, and tracks. The RAW data is permanently archived in safe storage, designed to occupy 1.5 MB / event. An extension of the RAW format is also used for simulated datasets, which occupies 2 MB / event, due to additional Monte Carlo truth information.

- **RECO format**: Reconstructed data produced by applying pattern recognition modules to RAW data. The result is physics objects which may be used in physics analyses in addition to reconstructed inputs (detector hits, energy clusters) to these objects.

- **AOD format**: Analysis Object Data, which is a compact form of the RECO format. It is meant to be easily transportable and storable, while providing all the necessary information for a typical physics analysis. An extension of this

format, AODSIM, is used for simulated datasets, and provides pertinent Monte Carlo truth information needed by the average user.

- **DQM format**: Data Quality Monitoring, which is information used to determine the quality of both simulated events and real collision data by comparing the information, in histogram form, to some centrally defined reference.

To facilitate the computing needs of CMS, in terms of data collection, reconstruction, storage, and analysis, a computing paradigm (Fig. 2.23) is developed to allow for the large scale needs of the experiment and its users:

- **Tier-0 center**: the only Tier-0 center resides at CERN. It is used to accept data from the online system and copy it to permanent mass storage. Along with this, the Tier-0 performs prompt reconstruction on the RAW data to produce RECO primary datasets, designed with physics signatures in mind. Tier-0 also transmits copies of both RAW and RECO data to Tier-1 centers.

- **Tier-1 centers**: Seven Tier-1 centers reside at various national labs and computing centers around the world, including one at FNAL in the US. The emphasis on Tier-1 centers is to provide reliable delivery of data-intensive services, requiring access to large CPU facilities, mass storage system with tape archive, and high speed international network connections. Tier-1's also provide long-term safe storage of RAW data, taking custodial responsibility for a fraction of CMS data. They must also store and serve both reconstructed data and simulated datasets to Tier-2 centers, which must be provided with rapid access to the Tier-1's. Tier-1 centers are also used to carry out re-reconstruction of RAW data as new techniques or detector conditions become available.

- **Tier-2 centers**: 50 Tier-2 centers are located throughout the world at various institutes affiliated with CMS, including eight at US institutions, with one at the University of Nebraska-Lincoln. Tier-2 centers typically divide resources up between those used by CMS for additional data reprocessing/simulation sample production, and those resources dedicated to users who examine the data for physics purposes. Due to the analysis requirements on Tier-2's, they must provide local mass storage to contain both real and simulated datasets copied from Tier-1 centers, as well as mass storage for users to contain personal event files. Tier-2's also provide support for specialized activities of analysis groups, such as tracking or b tagging work.

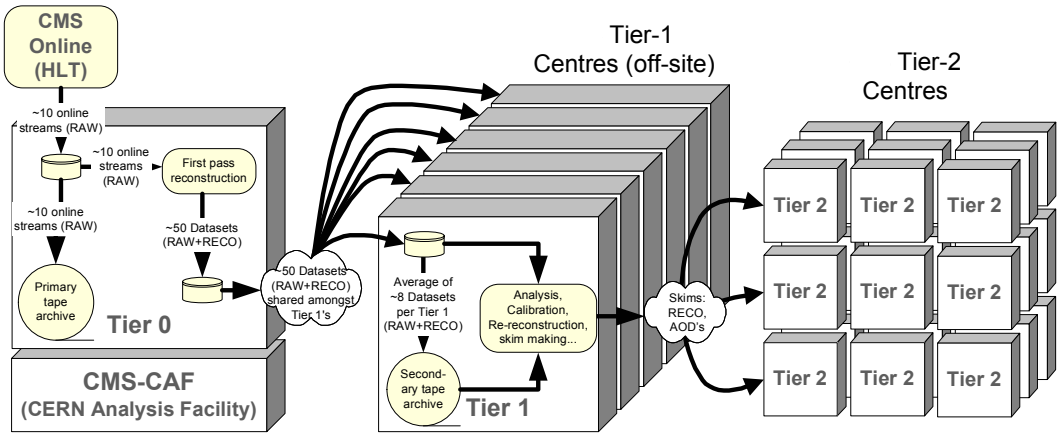

Figure 2.23: A schematic of the CMS computing hierarchy.

Chapter 3

Trigger Selection and Event Reconstruction

With the components of the CMS detector in place, it is now important to describe how these components are used to detect particles. This chapter discusses the complex algorithms used to reconstruct the particles or objects used in this analysis based on the detector information collected during data taking.

3.1 Trigger Selection and Datasets

11.6 fb^{-1} of data was used for this analysis, which was contained within the following datasets.

- /SingleMu/Run2012A-recover-06Aug2012-v1/AOD
- /SingleMu/Run2012A-13Jul2012-v1/AOD
- /SingleMu/Run2012B-13Jul2012-v1/AOD
- /SingleMu/Run2012C-PromptReco-v2/AOD

The phrase "SingleMu" denotes that the dataset is composed of events which have been accepted due to a single muon HLT trigger. The various run ranges in which the data was collected are described by Run2012A,B,C. A date contained within the dataset name describes when the data was rereconstructed from prompt reconstruction to account for any deficiencies in reconstruction, or to use new information to improve object finding.

For selection of events, the HLT trigger path "HLT_IsoMu24_eta2p1" was used, which requires a trigger muon with $p_T > 24$ GeV and $|\eta| < 2.1$. Additionally, the trigger muon must be isolated, which means that the total transverse momentum/energy of tracks/calorimeter clusters within a cone of $\Delta R < 0.24$ around the muon, divided by the muon's transverse momentum, must be less than 0.1. This particular trigger is selected since it matches the profile of what $t\bar{t}$ production provides in the muon plus jets channel, an isolated muon with high transverse momentum centrally located in the detector, while reducing the multijet background. The trigger also uses the very pure and efficient muon reconstruction for selecting events (Section 3.3).

Simulation samples (Tab. 3.1) are used to model both signal and background to either determine agreement with data, or to ascertain the level of agreement between estimation techniques and theoretical distributions. All simulation samples used contain the suffix "/Summer12_DR53X-PU_S10_START53_V7A-v1/AODSIM," which describes when the samples were produced, what pileup profile was used to simulate the effect of additional particles in the event through multiple pp interactions per bunch crossing, and other conditions used, such as the alignment of detector components and the location of dead detector channels, to name a few. For these samples, the additional events from pileup were mixed into the event, with the number of additional events coming from a random distribution provided by the "mix_2012_Summer_50ns_PoissonOOTPU_cfi.py" configuration.

Table 3.1: Simulation samples used in the analysis, with theoretical production cross sections and number of events.

Sample Name	Cross section (pb)	Number of Events
/TTJets_MassiveBinDECAY_TuneZ2star_8TeV-madgraph-tauola	225.2	6917468
/WJetsToLNu_TuneZ2Star_8TeV-madgraph-tarball	36257.2	18393090
/DYJetsToLL_M-50_TuneZ2Star_8TeV-madgraph-tarball	3503.7	30459503
/T_s-channel_TuneZ2star_8TeV-powheg-tauola	3.79	259961
/T_t-channel_TuneZ2star_8TeV-powheg-tauola	56.4	3749876
/T_tW-channel-DR_TuneZ2star_8TeV-powheg-tauola	11.1	496952
/Tbar_s-channel_TuneZ2star_8TeV-powheg-tauola	1.76	139974
/Tbar_t-channel_TuneZ2star_8TeV-powheg-tauola	30.7	1904972
/Tbar_tW-channel-DR_TuneZ2star_8TeV-powheg-tauola	11.1	493460
/WW_TuneZ2star_8TeV_pythia6_tauola	33.6	10000431
/WZ_TuneZ2star_8TeV_pythia6_tauola	12.6	10000283
/ZZ_TuneZ2star_8TeV_pythia6_tauola	5.2	9799908
/QCD_Pt_20_MuEnrichedPt_15_TuneZ2star_8TeV_pythia6	134680	21484602

Additional simulation samples (Tab. 3.2) are used to estimate systematic uncertainties due to variations in the factorization scale and in the matrix element to parton shower matching threshold.

Datasets are also used for the measurement of the soft electron reconstruction efficiency, discussed in Section 5.2. These are listed below.

- /MultiJet/Run2012A-13Jul2012-v1/AOD
- /BJetPlusX/Run2012B-13Jul2012-v1/AOD
- /BJetPlusX/Run2012C-24Aug2012-v2/AOD
- /BJetPlusX/Run2012C-PromptReco-v2/AOD
- /BJetPlusX/Run2012D-PromptReco-v1/AOD

The triggers used for selecting events require high p_T, multijet events with at least two jets having medium to high b-tag discriminator selection criteria (Tab. 3.3). The principles and discriminators used for b-tagging are discussed in Chapter 4.

3.2 Track Reconstruction

A track is a helical path with a unique set of helix parameters reconstructed using the tracker discussed in Section 2.2.2. The track represents the path a charged particle takes through the detector as it moves through the magnetic field. The track not only provides the path information of the particle, but also its momentum by using the curvature of the path in the presence of the magnetic field. The paths are formed from hits, which are locations where a charged particle passed through the tracker, leaving behind a detectable signal (see Section 2.2.2 for details). The hits are then combined to form the path of travel through the detector. Along the way, each newly added hit

Table 3.2: Systematic samples used in the analysis, with theoretical production cross sections and number of events. All systematic samples contain the "/Summer12_DR53X-PU_S10_START53_V7A-v1/AODSIM" suffix.

Sample Name	Cross section (pb)	Number of Events
/TTJets_scaleup_TuneZ2star_8TeV-madgraph-tauola	225.2	5008626
/TTJets_scaledown_TuneZ2star_8TeV-madgraph-tauola	225.2	5353628
/TTJets_matchingup_TuneZ2star_8TeV-madgraph-tauola	225.2	5415010
/TTJets_matchingdown_TuneZ2star_8TeV-madgraph-tauola	225.2	5476728
/DYJetsToLL_M-50_scaleup_TuneZ2star_8TeV-madgraph-tauola	3503.7	2170270
/DYJetsToLL_M-50_scaledown_TuneZ2star_8TeV-madgraph-tauola	3503.7	1934901
/DYJetsToLL_M-50_matchingup_TuneZ2star_8TeV-madgraph-tauola	3503.7	1985529
/DYJetsToLL_M-50_matchingdown_TuneZ2star_8TeV-madgraph-tauola	3503.7	2112387

Table 3.3: Triggers per sample used for the soft electron reconstruction efficiency measurement.

Trigger	Sample(s)
HLT_Jet160Eta2p4_Jet120Eta2p4_DiBTagIP3DFastPVLoose	MultiJet, BJetPlusX
HLT_Jet160Eta2p4_Jet120Eta2p4_DiBTagIP3DLoose	MultiJet, BJetPlusX
HLT_Jet60Eta1p7_Jet53Eta1p7_DiBTagIP3DFastPV	MultiJet, BJetPlusX
HLT_Jet60Eta1p7_Jet53Eta1p7_DiBTagIP3D	MultiJet, BJetPlusX
HLT_Jet80Eta1p7_Jet70Eta1p7_DiBTagIP3DFastPV	MultiJet, BJetPlusX
HLT_Jet80Eta1p7_Jet70Eta1p7_DiBTagIP3D	MultiJet, BJetPlusX
HLT_QuadJet75_55_35_20_BTagIP_VBF	MultiJet, BJetPlusX
HLT_QuadJet75_55_38_20_BTagIP_VBF	MultiJet, BJetPlusX
HLT_QuadPFJet75_55_35_20_BTagCSV_VBF	MultiJet, BJetPlusX
HLT_QuadPFJet75_55_38_20_BTagCSV_VBF	MultiJet, BJetPlusX
HLT_DiPFJet80_DiPFJet30_BTagCSVd07d05	BJetPlusX
HLT_DiPFJet80_DiPFJet30_BTagCSVd07d05d03_PFDiJetPt120	BJetPlusX
HLT_DiPFJet80_DiPFJet30_BTagCSVd07d05d03	BJetPlusX
HLT_DiPFJet80_DiPFJet30_BTagCSVd07d05d05	BJetPlusX

is used to update the helix parameters of the track. The methods for reconstructing these detector hits and the formation of trajectories are discussed in this section.

3.2.1 Hit Reconstruction

For the purposes of hit reconstruction, each silicon sensor of the tracker can be thought of as being a plane with a local coordinate system (u, v). The signals collected within this layer of the tracker are reconstructed into hits in a process called local hit reconstruction [23]. The coordinates of the plane run parallel/perpendicular to the pixels and strips in the detector, with u being defined as the coordinate perpendicular to the magnetic field, except in the stereo strip tracker, where it is the coordinate at the largest angle from the magnetic field.

In the pixel detector, zero suppression is applied to the sensor readout chips, with adjustable thresholds for each pixel. Pixel clusters are then formed from adjacent pixels reporting collected charge above the zero suppression. Any residual charge miscalibration, due to pixel-to-pixel variations, are estimated from laboratory measurements and included in simulation. With these pixel clusters formed, two methods for calculating the hit position of the charged particle are used: a fast algorithm used during pattern recognition, and a more precise algorithm used during the final track reconstruction [23].

For the pattern recognition step, the fast position finding algorithm uses the sum of the charge collected in the cluster projected along the local coordinates u and v. The hit position is then determined using the equations [23]

$$u_{rec} = u_C + \frac{Q^u_{last} - Q^u_{first}}{2(Q^u_{last} + Q^u_{first})}|W^u - W^u_{inner}| - \frac{L_u}{2}, \tag{3.1}$$

$$v_{rec} = v_C + \frac{Q^v_{last} - Q^v_{first}}{2(Q^v_{last} + Q^v_{first})}|W^v - W^v_{inner}| - \frac{L_v}{2}, \tag{3.2}$$

with Q_{first} and Q_{last} are the charges collected in the first and last pixels along each projection, (u_C, v_C) is the geometrical center of the cluster (excluding the first and last pixel). $L_{u,v} = D \tan \Theta_L^{u,v}$ is the Lorentz shift due to the motion of the sensor charges in the magnetic field. D is the sensor thickness and $\Theta_L^{u,v}$ is the Lorentz angle. $W_{inner}^{u,v}$ is the inner width of the cluster projections, excluding the first and last pixels. The total width $W_{u,v}$ is given by

$$W^{u,v} = D \tan(\alpha^{u,v} - \pi/2) + D \tan \Theta_L^{u,v} \tag{3.3}$$

The angles $\alpha^{u,v}$ are the particle impact angles with respect to the respective coordinate. For clusters of only one pixel, the center of the pixel is used after adjusting for the Lorentz drift.

For the final track hit, hit position in the pixel detector is estimated using the template method. The observed cluster charge distribution is compared to cluster charge shapes generated in simulation. The simulated cluster position which best matches the observed cluster charge distribution is used to find the reconstructed hit position [23]. To do the matching, the observed cluster charge can be described by charge P_i in each pixel i. This observed charge distribution is compared to each expected cluster shape S_{ij} for each hit position bin j. The best match is then found by minimizing the χ^2, given by

$$\chi^2 = \sum_i \left(\frac{P_i - N_j S_{ij}}{\Delta P_i} \right)^2, \tag{3.4}$$

$$N_j = \frac{\sum_I P_i/(\Delta P_i)^2}{\sum_i S_{ij}/(\Delta P_i)^2}, \tag{3.5}$$

with ΔP_i being the expected RMS of charge P_i, derived from simulation [23]. This χ^2 value is also used to reject outlier hits during the track if the cluster shape is

incompatible with the track trajectory.

For the strip tracker, algorithms are run in the front-end driver of the strip module to perform pedestal subtraction, common mode subtraction, and zero suppression to reject noise. A strip is accepted as being hit if it has charge exceeding five times the expected channel noise, or if the strip and one of its neighbors have a charge exceeding two times the channel noise [23]. Strips passing these requirements are called digis. Cluster formation starts with the use of seed digis, which are digis that have charge at least three times larger than noise. For each seed, adjacent strips are added if their charge exceeds twice the noise. The total cluster is kept if its total charge exceeds five times the cluster noise, defined as $\sigma_{\text{cluster}} = \sqrt{\sum_i \sigma_i^2}$, with σ_i begin the strip noise for strip i. The cluster position is then found using the charge weighted average of the strips, adjusting for the Lorentz shift.

3.2.2 Track Fitting

The CMS track finding software used to find tracks produced by most particles is known as the Combinatorial Track Finder (CTF), and tracks found from this algorithms as collectively known as CTFTracks. The CTF algorithm works by making multiple iterations over the collection of reconstructed tracker hits. The earlier iterations attempt to find the tracks that are easier to find (tracks with relatively high transverse momentum nominally from the interaction point), while the later iterations attempt to reconstruct both tracks with low transverse momentum, or highly displaced tracks. Each iteration removes hits successfully associated to tracks from consideration for the next iteration, reducing potential issues from combinatorics [23].

For all iterations, the track finding follows four steps [23]:

- Seed generation: The reconstruction of rudimentary track candidates using a

few (2 - 3) tracker hits. These seeds are used as an initial estimate for trajectories and their uncertainties.

- Track finding: the seed trajectories are extrapolated in search of additional tracker hits.

- Track fitting: estimates for track parameters are calculated based on the trajectory found.

- Track selection: tracks failing certain criteria are removed from consideration.

The seed generation defines starting trajectory parameters, called a trajectory seed, and the associated uncertainties for track candidates. Because of the quasi-uniform magnetic field surrounding the detector, charged particles follow helical paths. Thus, trajectory seeds contain the initial values for five parameters [24]:

- q/p: The trajectory curvature of the helix, derived from the equation of the momentum of a charged particle in a magnetic field, p = qBR,

- $\cot\theta$: the cotangent of the polar angle between the momentum vector and the z-axis,

- ϕ_0, the azimuthal angle at the point of closest approach with respect to the beam axis,

- d_0: the radial component of the point of closest approach with respect to the beam axis,

- z_0: the z-component of the point of closest approach with respect to the beam axis.

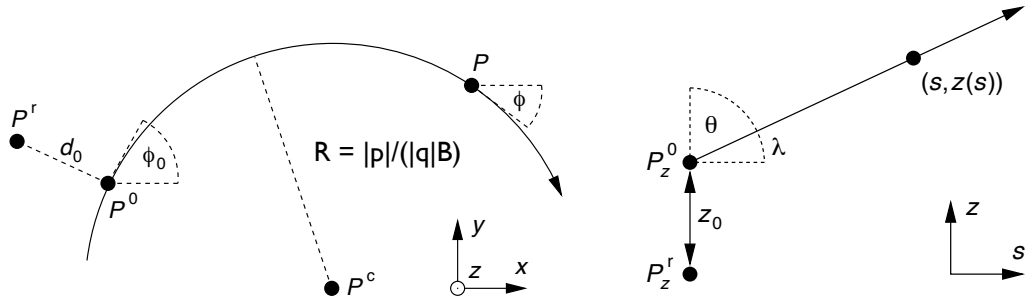

Figure 3.1: The track helix parameters in the $x-y$ (left) and $s-z$ (right) cylindrical planes, with $\mathbf{P}^r$ representing a reference point and $\mathbf{P}^0$ the point of closest approach of the track [25].

Figure 3.1 shows these parameters in both the $x-y$ and $s-z$ cylindrical planes, with s being the radial direction of the detector.

Seed generation begins with hits from the inner layers of the detector, since the granularity of the pixel detector allows for better parameter estimation. This will reduce trajectory degradation due to interaction with the detector material, and to allow for the reconstruction of low momentum tracks, which may not make it through the bulk of the detector. Seeds are formed from hits in pairs or triplets of the detector layers. Different regions of the detector have different requirements for acceptable trajectory parameters. For pairs of hits, the trajectory parameters found must be acceptable given the trajectory requirements of the given region. Hit triplets are formed from pairs of hits by requiring that the trajectory parameters at the outer hit be compatible with the parameters from the pair. For pairs of hits, an additional constraint is applied requiring that the seed originates from the beam spot to improve the parameter estimation.

The track finding step uses an implementation of the Kalman filter method, which moves from successive layers one by one, and attempts to add a new hit within the next layer to the old trajectory. The successive movement through the detector layers,

rather than an attempt at a global fit, is used to account for multiple scattering of the charged particle as it passes through the detector, in addition to accounting for energy lost due to moving through detector material [26]. The track fitting is divided into four steps [23]. The first step is the navigation step, which uses trajectory parameters evaluated at the current layer to find which adjacent detector layers are compatible with the current trajectory. An analytical propagator, assuming charged particle motion in a uniform magnetic field and no detector effects, is used for the navigation. The second step is the searching step, which finds compatible detectors in the layers returned by the previous step. A detector is compatible if it is within three standard deviations of the extrapolated trajectory position on that layer. The third step is the grouping step, which groups all hits from these detectors. A χ^2 test is used to determine if a hit is compatible with a trajectory measurement at the same layer. A $\chi^2 < 30$ is required. The final step is the updating step, which updates the trajectory based on the new hit information. If multiple, compatible hits exist, multiple track candidates emerge. For the second, third, and fourth steps, a propagator is used which takes both multiple scattering and energy loss into account. This process is repeated for each additional detector layer. The material propagator accomplishes this by inflating the uncertainty on the trajectory parameters according to the predicted rms scattering angle in the detector material, and by adjusting the momentum of the trajectory, taking the mean energy loss into account. To prevent the number of track candidates from growing exponentially, only the five best candidates are taken, which are those candidates having the lowest normalized χ^2 values.

Once the hit finding reaches a minimum configurable number of hits, the hit search reverses course, moving from outside-in, instead of inside-out. To begin this process, a new seed is formed using all the hits found in the previous step minus the hits of the previous seed. The inner hit finder then moves from layer to layer as described

above to find inner hits. This is done to account for additional hits that may be in the seeding layers due to overlapping sensors, and to find hits closer to the interaction region if the seed is formed from outer layers of the detector

Due to the number of possible combinations of seed generation and hit finding, it is possible for a track produced from one particle to be found more than once during the reconstruction. It is important to resolve the multiple reproductions so they do not all appear in the final track collection. This is done by using a trajectory cleaner that compares the fraction of shared hits for all track candidates, according to the equation $f_{shared} = \frac{N_{shared}^{hits}}{\min(N_1^{hits}, N_2^{hits})}$, with $N_{1,2}^{hits}$ being the number of total hits for each track being considered. If the fraction exceeds some given value, currently 0.19 [23], the track with the fewest hits is removed, or the track with the largest χ^2 value is removed, if both tracks have the same number of hits.

The track fitting step is necessary to remove any biases introduced in the trajectory calculation due to constraints applied during the seed generation stage. Track fitting uses the Kalman filter in two iterations. First, the fitting step starts with the innermost hit, and moves sequentially through each associated hit, updating the trajectory parameters along the way. Upon reaching the final, outermost hit, the process is reversed for the smoothing step, where the Kalman filter is run again going from the outermost to the inner most hit. The input for the smoothing step is the final trajectory result from the fitting step. The final track parameters are then the average of the fitting and smoothing steps [23].

To achieve better precision during the fitting/smoothing step, a Runge-Kutta propagator is used to move along the collection of associated hits. The Runge-Kutta propagator is able to account for both material effects and an inhomogeneous magnetic field, thus allowing for tracks which are not perfect helices. Upon completion of both fitting and smoothing steps, outlier hits are removed from the track, using either

a χ^2 matching requirement, or by using the template compatibility method discussed previously in this section.

The final step in track reconstruction is the selection step. Due to the large multiplicity of particles in LHC collisions, it is very possible for fake tracks, or tracks that are not due to a real charged particles, may be reconstructed. The selection step seeks to remove as many of these fake tracks as possible, while keeping practically all of the real tracks. The selection step places requirements on the number of total hits for the track, the total number of tracker layers where hits are expected, but not found, and the normalized χ^2 of the track fit. Requirements are also placed on the track to ensure that the track is compatible with originating from the interaction region. This will not be a common feature for fake tracks, which will be due to random hit combinations, and will not necessarily point back to the interaction region.

3.3 Muon Reconstruction

At CMS, muons are reconstructed by both the tracker and the muon detector, described in Section 2.2.5. As such, muons can be reconstructed using three strategies: stand-alone muons, global muons, and tracker muons. Each of these muon categories will be further described below. In each case, the muons are reconstructed using strategies similar to those discussed in Section 3.2, though the seed finding and propagation strategies may be different to take different detector designs into account.

3.3.1 Stand-alone Muon Reconstruction

Stand-alone muons are reconstructed using only hits in the Muon detector. Seeds are formed in stations using patterns of segments which meet geometrical criteria. The

p_T of the seed candidate is estimated using parameterizations of the form [27]

$$p_T = A - \frac{B}{\Delta\phi}. \tag{3.6}$$

The definition of $\Delta\phi$ depends on the hits contained within the seed. For seed candidates within the DT portion of the muon detector, $\Delta\phi$ is the bending angle of the segment with respect to the primary vertex direction. For seeds in the CSC region or for any seeds which overlap regions, $\Delta\phi$ is the difference between the ϕ positions of the two segments. If none of these conditions are met, then the direction of the highest quality segment is used [27].

These trajectory seeds are then used to build trajectories and propagate through the muon detector, similarly to what is discussed in Section 3.2. For the barrel region, the seed trajectory is used to propagate from the current layer to the next layer. The trajectory is used to find the chambers which are compatible, based on the direction ϕ. For the endcap region, the seed is used to find the chambers compatible within the disk, based on the radius r. Following the hit propagation, the trajectory is built using the in-out and out-in approach similar to what is discussed in Section 3.2. Stand-alone muon candidates are formed from the trajectories built from muon detector seeds by determining the trajectory parameters by propagating the stand-alone track to the point of closest approach to the beam line. To improve the momentum resolution, an interaction point constraint is imposed [27].

3.3.2 Global Muon Reconstruction

Global muon reconstruction combines the muon track in the silicon detector with the track in the muon detector. The goal is to gain increased resolution at lower momenta through the use of the tracker track, while gaining the fake muon discrimination and

resolution at higher momenta through the stand-alone track [27]. The global muon reconstruction proceeds by combining reconstructed tracker tracks with stand-alone muon tracks through a process called track matching. Due to the large multiplicity of tracker tracks, the stand-alone muon tracks are used to select tracker candidates for matching. First, tracker candidates are selected using a rectangular η-ϕ window around the stand-alone track. To reduce the fake rate and the reconstruction time, while keeping efficiency high, the rectangular window is defined with seven parameters [27]:

- The origin position of the tracking region, generally the interaction point.
- The allowed z spread of the region origin.
- The allowed r spread of the region origin.
- The direction from the origin to the stand-alone track.
- The ϕ size of the tracking region.
- The η size of the tracking region.
- The minimum p_T of tracks in the tracking region to determine the required curvature of tracks.

The next phase is to create a one-to-one match between a tracker track and a stand-alone track. This match is made by comparing the five trajectory parameters of the tracker and stand-alone tracks when propagated to a common surface. The surface chosen is one which both minimizes the total uncertainty of the trajectory parameters and reduces the number of potential silicon track candidates. Natural choices for surfaces are the outer layer of the silicon tracker, the inner surface of the muon detector, the cylinder containing the last silicon track hit, or the cylinder

containing the first muon chamber hit. Once the tracks are propagated to a common surface, a set of four variables are used to discriminate good matches from bad [27]:

- Compare the trajectory parameters, $\vec{p}_i$, of the tracks using a χ^2 similarity test with the trajectory covariance matrices, C_i:

$$\chi^2 = (\vec{p}_1 - \vec{p}_2)^T [C_1 + C_2]^{-1} (\vec{p}_1 - \vec{p}_2). \tag{3.7}$$

- Compare the track positions on the propagated surface in local coordinates (x, y):

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{3.8}$$

- Compare the track positions in (η, ϕ) space, using an equation similar to (3.8).

- Compare the track *directions* in (η, ϕ) space, using the momentum vectors.

For each stand-alone track, the silicon track candidates are looped over to calculate the discriminators listed above. A successful match is found when discriminator values for the pair fulfill certain tight requirements. If no match is found, then the best match is chosen to be the pair which has the most compatible directions at the interaction point, with less stringent cuts.

For the final global muon track fit, hits from the stand-alone muon track and the matched silicon track hit are combined and the track fit is performed similarly to what is described in Section 3.2. If multiple global track candidates are found for a given stand-alone track, the global track that has the lowest χ^2 is chosen, guaranteeing that there is one global track per stand-alone track.

3.3.3 Tracker Muon Reconstruction

The final method of muon reconstruction is the tracker muon reconstruction. Tracker muons are meant to address instances where muons may not leave enough hits in the muon detector, either due to being low momentum, or due to the configuration of the detector. No refit is attempted with the matching hits. Consequently, tracker muon reconstruction considers all tracker tracks as muon candidates, and attempts to match these tracks with hits in the muon detector. Additionally, the calorimeter signatures of these tracks can be used to further identify tracker muons, with the benefit of gaining muons that would otherwise be lost comes the cost of a rise in the fake rate. Therefore, great care must be taking in selecting tracker muons for analysis.

Tracker muon reconstruction begins by propagating silicon tracks to the calorimeter, and the energy of reconstructed hits passed in the propagation is stored in the tracker muon object, which allows for additional muon identification, since charged hadrons will leave more energy in the calorimeter than muons. The silicon tracks are propagated further into the muon detector. Any chambers that are crossed or nearly crossed (to account for multiple scattering) during the propagation are stored in the tracker muon object. The tracker muon algorithm then uses the crossed/nearly crossed chambers to calculate quantities useful for muon identification [27]:

- The distance between the propagated track and the nearest chamber edge in local (x, y) coordinates. The distance is given a negative sign if the propagated track position is inside the active volume, and is positive otherwise.

- The one-sigma uncertainty on the quantity calculated above, taking the covariance of the propagation into account.

- The position of the extrapolated track inside the chamber in local coordinates.
- The slope of the extrapolated track (dX/dZ) and (dY/dZ) where (X, Y, Z) are the local chamber coordinates.
- The one sigma uncertainties in the positions and slopes.
- The detector ID of the chamber.
- A vector of associated segments.

Segments are associated to tracks through a process called arbitration, which attempts to assign the correct segments to the correct track in cases where two tracks cross or nearly cross the same segment. Arbitration is performed using the distance between propagated track and chamber in local coordinates.

Figure 3.2 shows the reconstruction efficiency in data and simulation for both tracker and global muons in the barrel and endcap region of the CMS detector as a function of muon p_T. Figures 3.3 and 3.4 shows the muon misidentification probability, the fraction of times a hadron is misidentified as a muon, as a function of hadron momentum and eta. These plots show the muon reconstruction is highly efficient, but also very pure, making it ideal for physics analyses involving leptons.

3.4 Electron Reconstruction

Electron reconstruction is a major challenge at CMS due the large amount of tracker material and intense magnetic field, which lead to severe radiation loss and scattering as an electron moves through the detector. This leads to very poor reconstruction efficiency and resolution, especially for low momentum electrons, unless steps are taken to acquire as much electron information as possible. This section will discuss

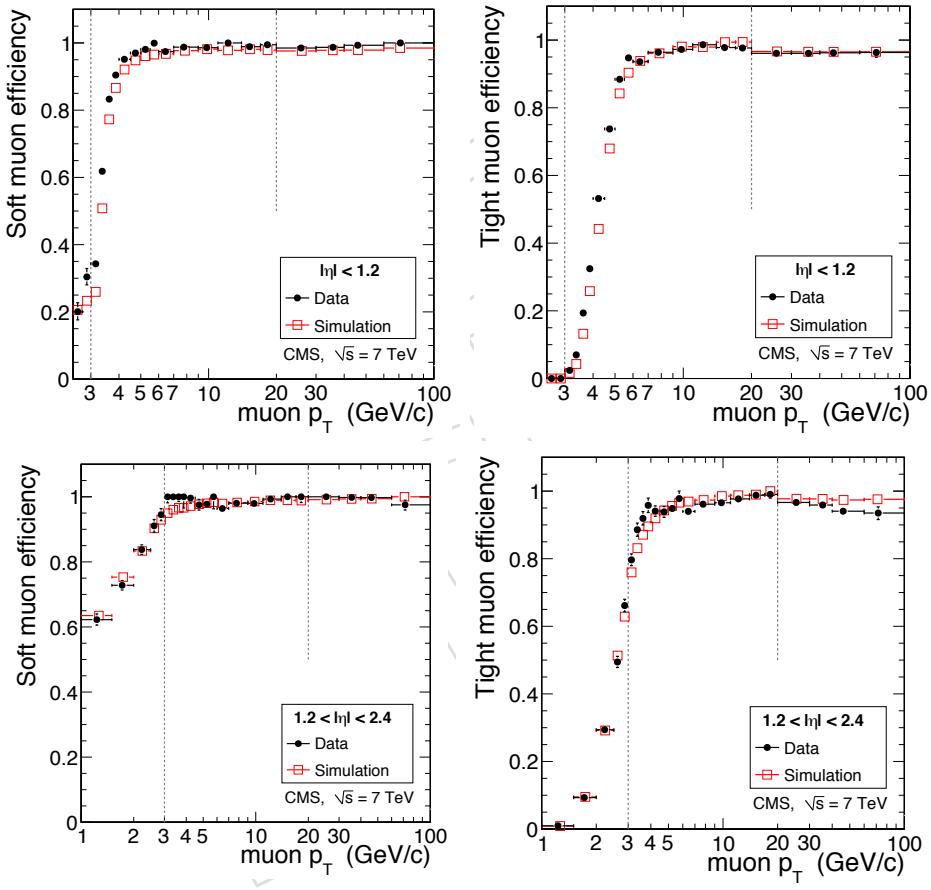

Figure 3.2: The muon reconstruction efficiency in data (black points) and simulation (hollow red boxes) for tracker muons (soft muons, left) and global muons (tight muons, right) as a function of muon p_T for the barrel ($|\eta| < 1.2$, top) and endcap ($1.2 < |\eta| < 2.4$, bottom) of the CMS detector [28].

the reconstruction of electrons through a process called particle flow, which seeks to use the combined detector information, rather than using the detector reconstruction sequentially, to reconstruct particles. Particle flow takes advantage of the precision performance of the tracker to improve the resolution of the calorimeter for reconstruction purposes. The reconstruction of electrons requires looking for both the seed cluster, or impact point of the electron with the calorimeter, and the bremsstrahlung clusters, clusters of energy in the calorimeter due to radiation emitted from the elec-

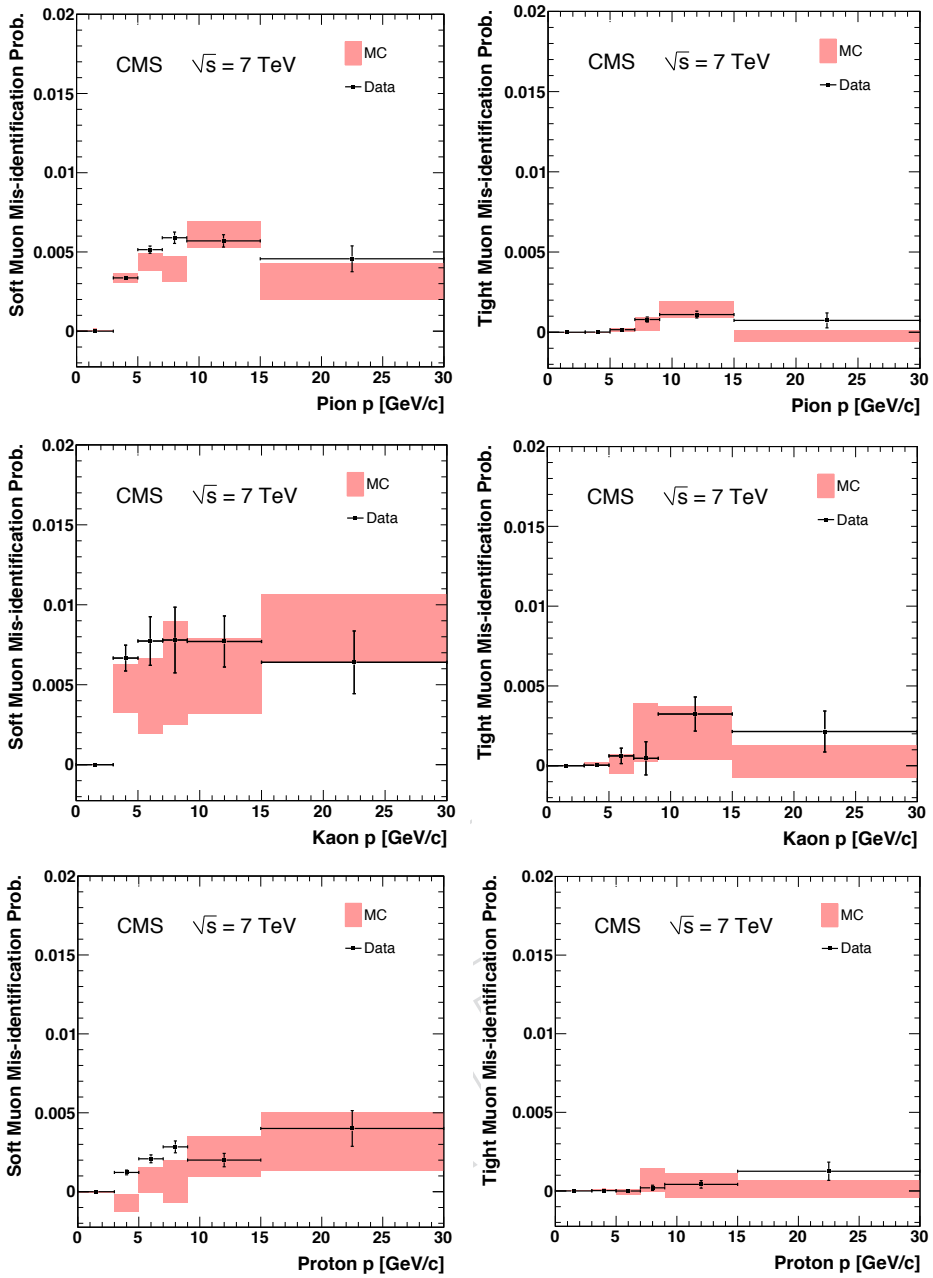

Figure 3.3: The muon misidentification probability in data (black points) and simulation (red shaded) for tracker muons (soft muons, left) and global muons (tight muons, right) as a function of hadron momentum for pions (top), kaons (middle), and protons (bottom) [28].

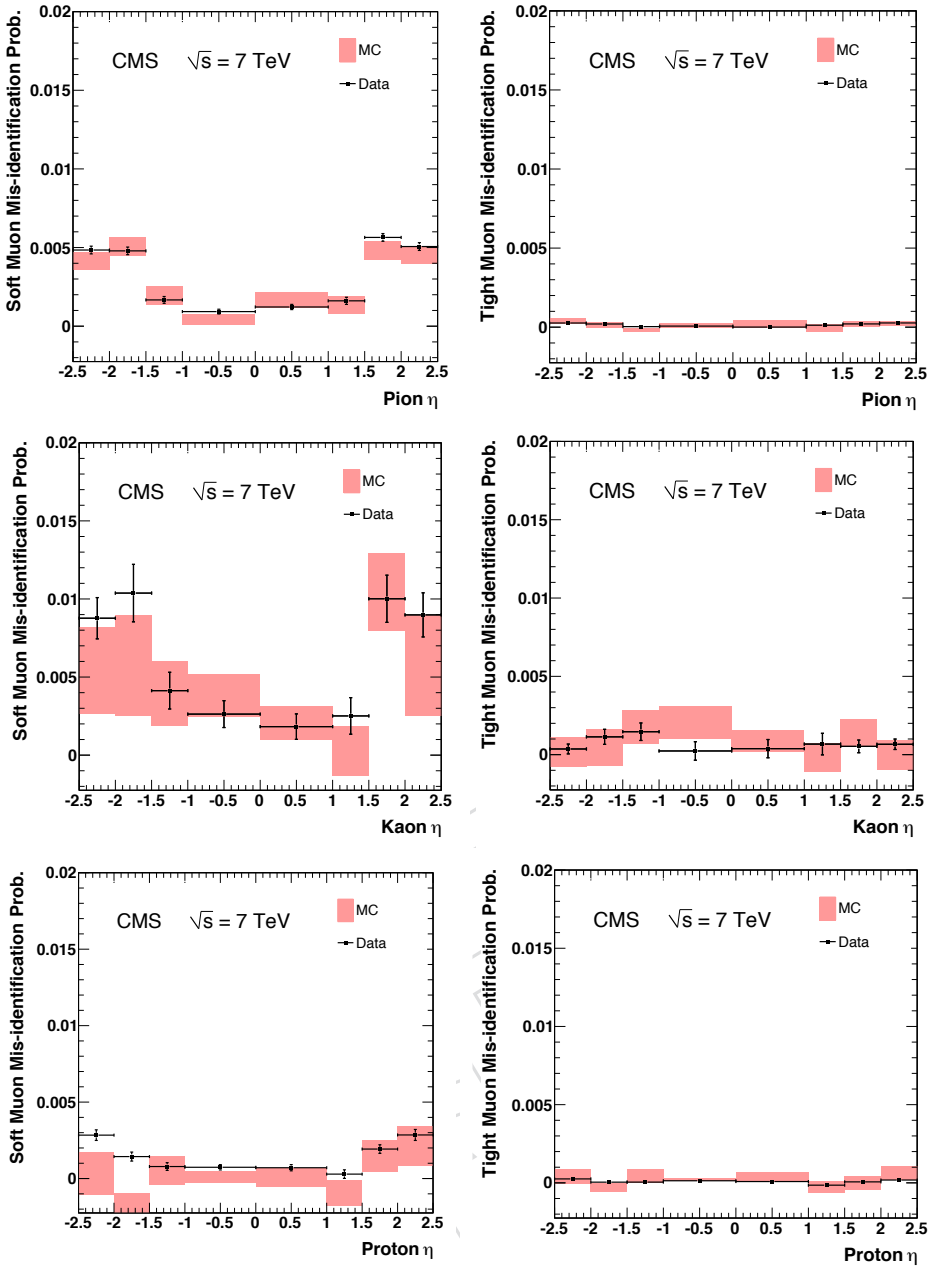

Figure 3.4: The muon misidentification probability in data (black points) and simulation (red shaded) for tracker muons (soft muons, left) and global muons (tight muons, right) as a function of hadron eta for pions (top), kaons (middle), and protons (bottom) [28].

tron during its flight through the detector. These clusters must be combined with a track reconstructed with a different algorithm than is discussed in Section 3.2 to account for the severe scattering and radiation loss. These two components combined can give a wealth of information about the electron itself, and can be used to remove fake electron candidates.

The process begins with the clustering of energy deposits in the electromagnetic calorimeter. The clustering algorithm used in particle flow is designed to be both highly efficient for particles with low transverse momentum, and to separate clusters from very close particles [29]. The clustering is performed separately for each calorimeter sub detector. The first step of the process identifies locally high energy deposits in the cells of the calorimeter (above 230 MeV in the barrel and 600 MeV in the endcap), which are used as seeds for the clustering. Subsequently, additional cells are added to the cluster if they are share a side with cells already added to the cluster, and have an energy above a configurable threshold, which is currently set to two standard deviations above the cell noise. This process is called topological clustering. Finally, these topological clusters are combined to form particle flow clusters. One particle flow cluster is meant to represent the total energy of a particle deposited in the calorimeter, such as the seed cluster and brem clusters of an electron.

Electron reconstruction includes its own unique electron track reconstruction, since the algorithm discussed in Section 3.2 does not adequately account for the radiation loss of electrons moving through the tracker. Electron energy loss is modeled by the Bethe-Heitler formula, which is non-Gaussian, and therefore unsuitable for use in a Kalman track finding algorithm [23]. To account for this, electron track reconstruction is performed using the Gaussian Sum Filter (GSF) algorithm. The GSF algorithm approximates the Bethe-Heitler energy loss for electrons as a sum of Gaussians, with each Gaussian being associated with its own Kalman filter [30]. Thus,

each GSF track is actually composed of a weighted sum of Kalman tracks, depending on the quality of each Kalman track compared to the hits in the sub detector. This allows for GSF track parameters to be calculated in two different ways, using either the weighted mean of the Kalman track parameters, or using the track parameters from the Kalman track with the highest weight.

At CMS, GSF tracks are currently seeded using two approaches: an ECAL-seeded approach and a track-seeded approach. The ECAL-seeded approach utilizes the fact that any electron track must be accompanied by a cluster in the electromagnetic calorimeter. Starting with reconstructed electromagnetic clusters, the GSF ECAL seeding algorithm forms a trajectory window from the cluster position down to the inner layers of the pixel detector. Curved trajectory windows are used to account for the electron charge, and two trajectory windows are formed for both positive and negative tracks. These trajectory windows are used to find hits in the pixel detector that are compatible with having come from the beam spot. The compatible pixel hits are then used to form the trajectory seed used for the next steps of the GSF track reconstruction.

The track-seeded approach utilizes the CTF tracks described in Section 3.2 to seed the GSF track reconstruction. Though the CTF algorithm is not well suited for reconstructing electron tracks, it will still find some component of the electron track, or the entire track itself, in trajectory segments where the electron does not radiate too much energy. Since GSF track finding is a time-consuming process, containing multiple Kalman track components, not all CTF tracks are used for GSF track seeding. A two-part process is needed to preselect the CTF tracks for GSF track seeding. The first step uses CTF tracks which successfully propagate from the interaction point to the calorimeter, and is used for electrons which do not radiate much energy. An attempt is made to match tracks with clusters in the ECAL using two variables: χ^2_{geom}

and the ratio between the cluster energy and the track momentum at the outermost layer. The variable χ^2_{geom} is a variable that describes the geometric compatibility between the calorimeter cluster and the track propagated to the calorimeter, given by

$$\chi^2_{geom} = \left(\frac{\Delta\phi}{\sigma_\phi}\right)^2 + \left(\frac{\Delta\eta}{\sigma_\eta}\right)^2, \tag{3.9}$$

which the sum of the differences of (η, ϕ) of the cluster position and the extrapolated track position divided by their uncertainties squared. CTF tracks must have $\chi^2_{geom} < 10$ and $E/p > 0.65, 0.75$ for tracks with p_T less than or greater than 6 GeV.

CTF tracks failing the first selection step, either because they fail the cuts of the first step or fail the ECAL propagation, are preselected by the second step, which uses the number of CTF track hits and the CTF track normalized fit χ^2 as selection variables. The CTF tracks preselected by a second step are assumed to be poorly fitted tracks, otherwise they would have passed the first step, and are thus expected to have a lower number of hits or a higher normalized fit χ^2 than tracks from other sources. CTF tracks passing either step one or step two are used as trajectory seeds for GSF track fitting.

The trajectory building follows a similar strategy as given in Section 3.2. One difference involves the modeling of the energy loss for electrons, which is discussed above in this section. Another key difference lies in the reduction in trajectory candidates when moving from one layer to the next. When moving to a compatible layer that includes many hits, many candidate trajectories are tried in parallel, as with the standard tracking algorithm. To choose which of these many candidates to move on to the subsequent layers, rather than using a χ^2 cut, the best two candidates, using their χ^2 value, are kept to propagate on. This is done to keep the efficiency to a higher level than would be maintained with a cut on the fit χ^2 [30].

With both the calorimeter clusters and electron tracks reconstructed, the particle flow algorithm combines them using a linking algorithm. To begin, the reconstructed GSF track is propagated into the ECAL at a depth corresponding to the expected electron shower profile. The GSF track is then linked to an ECAL cluster if the extrapolated track position lies within the boundaries of a reconstructed cluster. This cluster is designated as the seed cluster, defined as the cluster of energy due to the electron impact with the calorimeter. Additionally, bremsstrahlung clusters are linked to the track by using tangent lines from the track to the ECAL. Tangent lines are used to mimic photons emitted from the track, which will move in a straight line from emission point to the calorimeter. If the tangent line lies within a reconstructed calorimeter cluster, the cluster is linked to the track as a brem cluster [29].

In spite of these implementations meant to improve the electron reconstruction, the signal-to-background ratio is rather poor since charged hadrons can mimic the electron signature in the ECAL, and are also accompanied by a reconstructed track. In order to reduce this background, a preselection step must be implemented. In spite of the similar signatures of electrons and charged hadrons, there are characteristic variables, which provide separation between the electron signal and the reconstructed fakes. These characteristic variables are given below [31].

- E_{ECAL}/p_{Trk}: the ratio of the seed cluster energy in the ECAL to the outer track momentum. Since electrons lose all of their energy in the ECAL, the ratio should peak around 1, while for charged hadrons, which should only lose part of their energy, the ratio will be closer to 0.

- χ^2_{geom}: the cluster-track position matching variable, given by equation (3.9).

- Number of hits per track.

- $|p_T^{in} - p_T^{out}|/p_T^{in}$: Also known as the brem fraction, this is the difference between the momentum at the interaction region and the momentum at the calorimeter divided by the inner momentum. This variable describes the amount of energy radiated by the track as it moves through the detector. Charged hadrons should emit almost no energy, so the distribution should be close to 0.

- χ^2_{GSF}: the GSF track fit χ^2.

- $\chi^2_{CTF}/\chi^2_{GSF}$: The ratio of the CTF track χ^2 fit to the GSF track χ^2 fit for the given particle. Charged hadrons should see almost no improvement from the GSF track fit, while electrons will generally have worse CTF χ^2 values.

These variables are combined using a multivariate tool known as a Boosted Decision Tree (BDT) in the TMVA framework [32]. To be considered an electron in the final reconstruction, a candidate must have a BDT value greater than -0.1, though the user is able to make tighter cuts to achieve greater purity. Figure 3.5 shows the particle flow BDT distribution for different particles [33] .

3.5 Jet Reconstruction

Jets are defined as a cluster of particles moving in a common direction. Jets are typically composed of hadrons, and are due to strong interactions that occur within the event. Thus, jets generally move in the direction of the original parton that created them. At CMS, the jet reconstruction is performed using particles as reconstructed in the particle flow framework. The previous sections in this chapter describe the reconstruction of muons and electrons in this framework, but jet reconstruction requires the reconstruction of additional particles: charged hadrons, neutral hadrons, and photons (which come from the decay of π^0 mesons). This section will briefly dis-

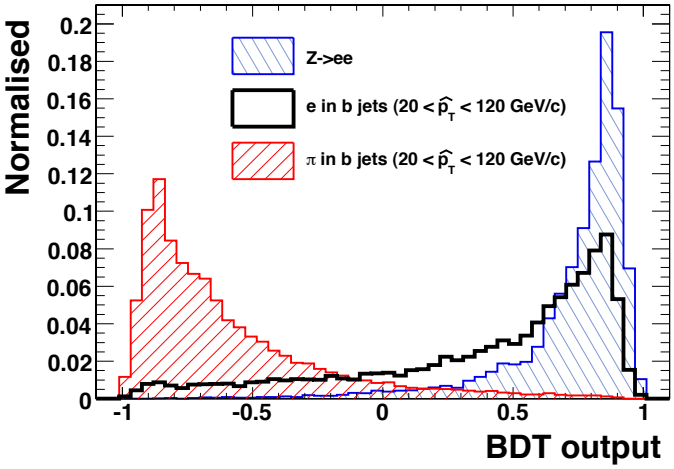

Figure 3.5: The particle flow BDT distribution for electrons from Z boson decay (blue hashed), electrons from b quark decay (black), and background pions in b jets (red hashed) from simulation.

cuss the reconstruction of charged and neutral hadrons, and photons in the particle flow framework, and how all particles are clustered to form jets at CMS.

3.5.1 The Reconstruction of Charged and Neutral Hadrons, and Photons

The reconstruction of photons and charged and neutral hadrons requires the use of silicon tracks and calorimeter clusters, and knowledge of how each of these particles interacts with the calorimeter. The particle flow reconstruction works to find these three objects simultaneously with the idea being that any energy in the HCAL associated with a track must due to a charged hadron, while any remaining energy in the HCAL (after accounting for charged hadrons) is due to a neutral hadron, and likewise, any excess energy in the ECAL (after accounting for charged and neutral hadrons) is due to a photon. This requires knowing how a charged hadron deposits energy in both the ECAL and HCAL, how neutral hadrons deposit energy in both

calorimeter components (and how that relates to the initial energy), and similarly for photons in the ECAL. To obtain this information requires calibrating the ECAL and HCAL response for charged hadrons, and calibrating separately how the excess energy relates to both neutral hadrons and photons [29].

To begin the reconstruction, tracks that are not used for the muon and electron reconstruction are propagated to the HCAL to be linked with a HCAL cluster. Several tracks can be linked to the same hadronic cluster, which requires comparing the total momenta of the linked tracks to the total energy of the cluster. An electromagnetic cluster may be linked to the track to account for energy deposited into the ECAL. Successful linkage of tracks to calorimeter clusters leads to charged hadron reconstruction. The track-calorimeter calibration is used to remove the total energy contribution from charged hadrons. Next, excess unlinked HCAL clusters are linked to nearby ECAL clusters whose energy distribution does not fit the profile of having come from a photon. Unlinked HCAL clusters or HCAL-ECAL linked clusters form neutral hadrons. The remaining ECAL clusters with a photon distribution profile are finally used to reconstruct photons.

To perform the energy calibration, calorimeter test beam data is used, and is refined with measurements from real data. The calibrated energy is found using the function [29]

$$E_{\text{calib}} = a + b(E, \eta)E_{\text{ECAL}} + c(E, \eta)E_{\text{HCAL}}, \tag{3.10}$$

with a, b, c being coefficients depending on the pseudorapidity of the particle, and E, an estimate of its true energy, taken from the momentum measurement from the track, or the total calorimeter energy, whichever is highest. For a given value of a, b

and c are found using a χ^2 fit of the form

$$\chi^2 = \sum_{i=1}^{N} \frac{(E^i_{\text{calib}} - E^i)^2}{\sigma_i^2(E^i_{\text{calib}})}, \tag{3.11}$$

with E^i and σ_i being the estimated true energy and expected energy resolution for the i^{th} charged hadron. The sum runs over all events and is separated for those particles in the barrel versus endcap, and for particles which deposit energy only in the HCAL, only in the ECAL, or in both.

3.5.2 The Reconstruction of Jets at CMS

As stated previously, jets are clusters of particles which move in a common direction. It is generally helpful to visualize these jets as a set of particles contained within a cone whose point is located at the interaction point. Jets are often reconstructed with this cone convention in mind, with the center of the face of the cone being the most energetic component of the jet (particle, calorimeter cell), and the direction of the jet being the energy or momentum weighted average of the jet constituents within the cone. There are many ways to reconstruct jets, adding new constituents iteratively, or using fixed cones centered around energetic objects present in an event. Each method comes with its own set of advantages and disadvantages. This section will discuss how jets at CMS are reconstructed using the anti-kT algorithm.

Particles reconstructed in the particle flow framework (discussed in the previous sections of this chapter) are used as inputs for the jet finding. In the anti-kT algorithm, the reconstructed particles are clustered using a distance parameter, called d_{ij}. The indices i and j can refer to an unclustered particle or a set of clustered particles, called pseudojets. The anti-kT algorithm compares d_{ij} of object i and object j to the distance between i and the beam, whose distance is d_{iB}. If d_{ij} is smaller than

d_{iB}, then object i is clustered with object j, and the process continues with the next object in the set. Otherwise, object i is removed from consideration for the current cluster, but is saved for the formation of other clusters. The process continues until no more objects are left for consideration, or no more clustering can be performed [34]. The distance parameters are defined by

$$d_{ij} = \min(k_{ti}^{-2}, k_{tj}^{-2})\frac{\Delta_{ij}^2}{R^2}, \tag{3.12}$$

$$d_{iB} = k_{ti}^{-2}, \tag{3.13}$$

$$\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2, \tag{3.14}$$

with k_t, y, and ϕ being the transverse momentum, rapidity, and azimuth of the object. The parameter R is a user input to the algorithm which describes the radius of the cluster cone [34].

Due to the nature of its operation, the anti-kT algorithm will produce conical jets in which soft particles will preferentially cluster around hard particles. Thus, the addition of soft radiation does not affect the formation of jets. The anti-kT algorithm also avoids the need for having a splitting and merging step for jets (a common pitfall for other algorithms) by having the splitting and merging built into the algorithm itself. This removes the ambiguity of which particle belongs to which jet.

Chapter 4

B-Tagging Techniques at CMS

The term "b-tagging" is defined as techniques used to find (or "tag") jets produced from bottom quarks. Such jets are typically expected in new physics signatures. Thus, the ability to find jets from bottom quarks, and suppress jets from other sources, is a powerful tool at collider experiments. This chapter describes what distinguishes bottom quark jets from other jets, the techniques used to tag these jets, and the tools used to validate the performance of the algorithms.

4.1 Principles of b-Tagging

B-Tagging techniques utilize the physics of bottom quarks (b quarks) versus other quarks. Bottom quarks are the second heaviest quark, with a mass of roughly 4 GeV[2]. Hadrons produced from bottom quarks decay via the weak process, through the emission of a virtual W boson. Since the W boson emitted is off-shell, and the decay of the bottom quark to lighter quark is CKM-suppressed (due to a factor of $|V_{xb}|^2$), bottom hadrons have relatively long lifetimes compared to other decaying hadrons, traveling a measurable distance from the interaction region. Thus, particles

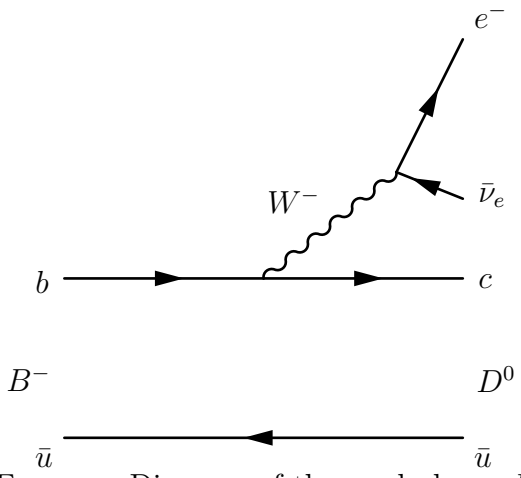

Figure 4.1: Feynman Diagram of the weak decay $B^- \to D^0 e^- \bar{\nu}_e$.

appearing from bottom hadron decay will have higher impact parameters with respect to the primary vertex than those particles which decay near, or are produced from, the primary vertex. If enough charged particles are produced in the decay, it is also possible to produce a secondary vertex, which has a significant decay length with respect to the primary vertex. Another consequence of the weak decay is the presence of semileptonic decays in bottom quark jets in much larger proportions compared to their lighter counterparts (Fig. 4.1) [2].

Knowing the difference between b jets and light jets, what does it mean to "tag" a jet? To tag a jet, a variable needs to be defined which utilizes some property that discriminates b jets from other jets. This variable can be related to the impact parameter of tracks, the distance between a secondary and primary vertex, or any other suitable quantity. The value of this variable, called a discriminator, is calculated for each jet, using the tracks associated to the jet, or some property of the jet itself. Typically, the discriminator is defined such that a larger cut on the discriminator value tends to be associated with jets that are b-like. The actual act of "tagging" is making a cut on the discriminator value of jets in an event. Jets with discriminator values that pass an applied cut are assumed to be b jets, and are said to be "b-

tagged." The analysis can then proceed with some level of confidence that the tagged jets are b jets. This level of confidence depends on the tagging efficiency (the fraction of actual b jets that pass a given discriminator cut) and the mistag rate (the fraction of non-b jets passing a given discriminator cut). The goal of any b-tagger is to have a much higher tagging efficiency than mistag rate. Both quantities depend on the separation in the discriminator distributions between b and non-b jets, and the value of the cut made on the discriminator itself.

4.2 Primary Vertex Reconstruction

An important ingredient of tagging b jets is knowing where the initial proton-proton interaction took place. This is the purpose of primary vertex reconstruction. The reconstruction uses charged particle tracks reconstructed from the silicon tracker. These tracks must fulfill quality requirements to ensure a good reconstructed primary vertex:

- Total number of pixel tracker layers with hits > 2,
- Total number of silicon tracker layers with hits > 5,
- Normalized χ^2 for the track fit < 20.0,
- Impact parameter in the transverse plane with respect to the beamspot < 5.0 cm.

Selected tracks are then clustered based on their z-coordinate value for the point of closest approach with respect to the beam line, using a method known as *deterministic annealing*, a minimization optimization procedure which is an analog of annealing in thermodynamics. For typical clustering by z-coordinate, the goal would be to find

the cluster configuration which minimizes the overall χ^2 of the system, given by

$$\chi^2 = \sum_{ik} c_{ik} \frac{(z_i - z_k)^2}{\sigma_i^2}, \tag{4.1}$$

with z_i, z_k being the z-coordinate for track i and vertex k, σ_i being the measurement uncertainty on z_i, and c_{ik} being a matrix element which describes the track-to-vertex association ($c_{ik} = 1$ when track i belongs to vertex k, $c_{ik} = 0$ otherwise). Deterministic annealing replaces the matrix element c_{ik} with a probability p_{ik}, which can take values between zero and one. The annealing then finds the most likely distribution of assignment of tracks to vertices, using the principle of maximum entropy. The total χ^2 of the system is decreased gradually, while keeping the system in the highest probability state, analogous to the cooling of an ensemble of particles in statistical mechanics.

For deterministic annealing, the "energy" of track i attached to vertex k is defined as

$$E_{ik} = \frac{(z_i - z_k)^2}{\sigma_i^2}, \tag{4.2}$$

and the mean "energy" for the ensemble of tracks and vertices is

$$E = \sum_k \sum_i p_i \rho_k p_{ik} E_{ik}, \tag{4.3}$$

with ρ_k being the cluster weight (defined below), and p_{ik} being the probability of having track i belong to cluster k at "temperature" T. p_i is a constant weight which is used to describe the impact parameter of the track with respect to the beamspot, given by

$$p_i = \frac{1}{1 + e^{\frac{IP^2}{\sigma_{IP}^2} - d_0^2}}. \tag{4.4}$$

d_0 is a user-defined cutoff for the weight. If d_0 is set to 0, then p_i is hard-coded to 1 by default [35].

The annealing occurs in two phases, the splitting phase and the assignment phase. For the splitting phase, the annealing variables take on the following definitions and constraints.

$$\sum_k \rho_k p_{ik} = 1, \tag{4.5}$$

$$\sum_k \rho_k = 1, \tag{4.6}$$

$$p_{ik} = \frac{e^{-E_{ik}/T}}{\sum_{k'} \rho_{k'} e^{E_{ik'}/T}}, \tag{4.7}$$

$$z_k = \frac{\sum_i p_i p_{ik} z_i/\sigma_i^2}{\sum_i p_i p_{ik}/\sigma_i^2}, \tag{4.8}$$

$$\rho_k = \frac{\rho_k \sum_i p_i p_{ik}}{\sum_i p_i}. \tag{4.9}$$

During the splitting phase, the "temperature" T is cooled and the clusters begin to split along the z-coordinate. Thus, the positions of the clusters are determined along z. The splitting phase continues until T reaches a user-defined value of T_{min} [35]. Then, the assignment phase begins, which assigns tracks to the variously formed clusters, following equations similar to those defined above, but with $\rho_k = 1$. The assignment process proceeds until $T = 1$ [35].

Once the clusters of tracks are formed, vertices are formed from the clusters using the adaptive vertex finder, a vertexing algorithm that uses annealing procedures very similar to deterministic annealing. The adaptive vertex finder is a least-squares vertex fitter at heart, but it assigns a weight to each track associated to the vertex, defined as

$$w_i(\chi_i^2) = \frac{e^{\chi_i^2/2T}}{e^{\chi_i^2/2T} + e^{\sigma_{cut}^2/2T}}. \tag{4.10}$$

σ_{cut} is a user-defined parameter for which $w_i = 0.5$. The vertex weight may be interpreted as a track-to-vertex association probability, and the vertex fit becomes a minimization of a weighted least-squares fitter. As a result, the vertex fitter gives a higher weight to tracks most likely to have come from a common vertex, and reduces the weight of potential outlier tracks [36].

The final step in primary vertex reconstruction is the ordering step. The vertices are ordered according to the sum of the square of the transverse momentum ($\sum p_T^2$) of the tracks associated with them. It is assumed that the vertex with the highest $\sum p_T^2$ is the primary interaction vertex, and it is this vertex which is referred to throughout the remaining sections as the primary vertex. Figure 4.2 shows the resulting primary vertex reconstruction efficiency versus number of vertex tracks, as measured in 7 TeV data and simulation samples. Figure 4.3 shows the resulting primary vertex position resolution versus number of tracks. The resulting efficiency and resolution are quire good, especially compared to the decay lengths of long lived particles, such as those discussed in this section.

4.3 B-Tagging with Displaced Tracks

Having found the primary vertex, b-tagging can proceed through the use of displaced tracks and the reconstruction of secondary vertices. To begin the process, tracks are associated to reconstructed jets using the quantity $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$, with η and ϕ for both track and jet defined with respect to the primary vertex. For the default reconstruction configuration, it is required that $\Delta R < 0.5$. The tracks then go through a selection process to ensure quality b-tagging. The criteria for selected tracks are [37]

- Track transverse momentum > 1.0 GeV,

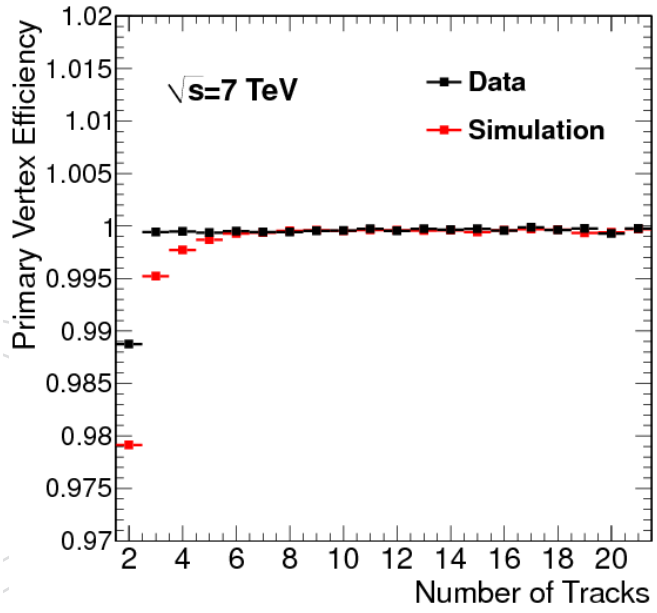

Figure 4.2: The primary vertex reconstruction efficiency vs number of vertex tracks for 7 TeV data and simulation [23].

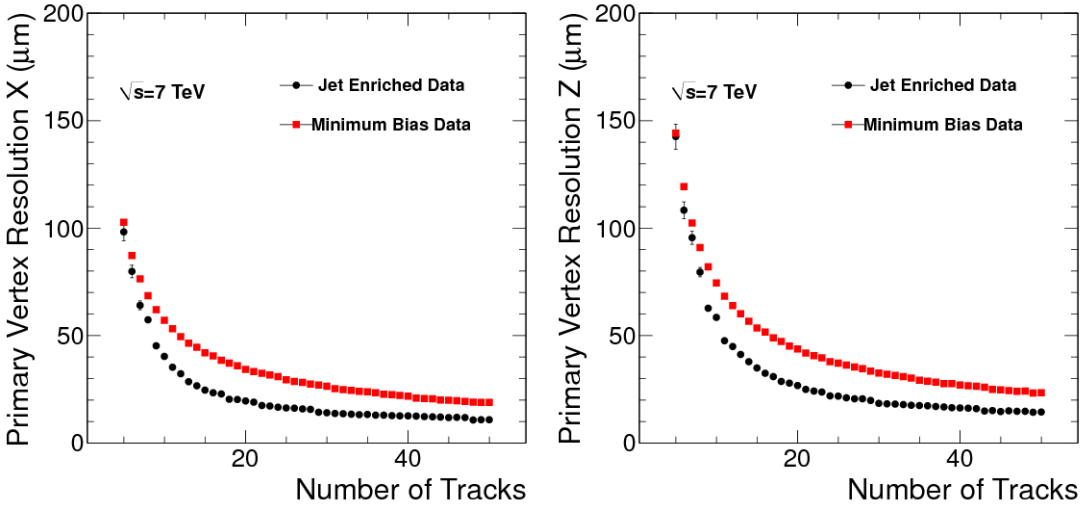

Figure 4.3: The primary vertex position resolution vs number of vertex tracks for both the x (left) and z (right) positions for 7 TeV data and simulation [23].

- Track fit normalized $\chi^2 < 5.0$,
- Number of track pixel hits > 1,
- Number of track silicon hits > 7,
- Track impact parameter in the transverse plane with respect to the primary vertex < 0.2 cm,
- Track impact parameter in the longitudinal plane with respect to the primary vertex < 17.0 cm.

Once tracks are selected, a quantity called the signed impact parameter significance is calculated. The impact parameter for a track with respect to the primary vertex is defined as the distance between the point of closest approach (pca) of the track to the primary vertex (pv) and the vertex itself:

$$\text{IP} = |\vec{pca} - \vec{pv}|, \tag{4.11}$$

with both $\vec{pca}$ and $\vec{pv}$ being position vectors for each point with respect to the detector origin.

For b-tagging, the impact parameter is calculated in the transverse plane, or in three dimensions. Due to the long lifetime of b hadrons, tracks from their decay are expected to have larger impact parameters than tracks from lighter jets. However, the impact parameter quantity itself does not necessarily provide a good handle for whether or not a track came from a displaced decay, since large impact parameters can also be a result of a poorly resolved track or primary vertex. A better descriptor

is the impact parameter significance, defined as

$$\mathrm{IPS} = \frac{IP}{\sigma_{\mathrm{IP}}}, \tag{4.12}$$

with IP being the impact parameter value, and σ_{IP} being the uncertainty of the IP value. The IP significance describes how significant the distance from the track pca to the primary vertex really is. b Jets will contain more significantly distant tracks than their lighter counterparts. Finally, a sign can be added to the impact parameter significance, which describes whether or not the track in question came from a decay which was upstream (in the direction of) the jet direction, or downstream of that direction. From physical principles, tracks from b decay should not only have tracks significantly distant from the interaction point, they should also come from a decay moving in the direction of the jet itself, since the jet would not exist without the presence of the bottom quark. However, lighter jets, which will have tracks from the interaction vertex primarily associated to it, will have tracks with random orientations. Thus, the inclusion of the sign becomes a powerful discriminator between b jets and light quark jets. To obtain this sign, the dot product between the jet direction and a vector pointing from the primary vertex to the track pca is used.

$$\mathrm{SIPS} = \frac{\vec{\mathrm{pca}} \cdot \vec{\mathrm{jet}}}{|\vec{\mathrm{pca}} \cdot \vec{\mathrm{jet}}|} \frac{IP}{\sigma_{\mathrm{IP}}}. \tag{4.13}$$

The distribution for the signed impact parameter significance (Fig. 4.4) will be symmetric about zero for light quark jets, but will be asymmetric toward the positive region for b jets. Figure 4.5 shows a sample signed impact parameter distribution for the third highest SIPS track associated to a jet, which shows a clear distinction between b jets and light (udsg) jets.

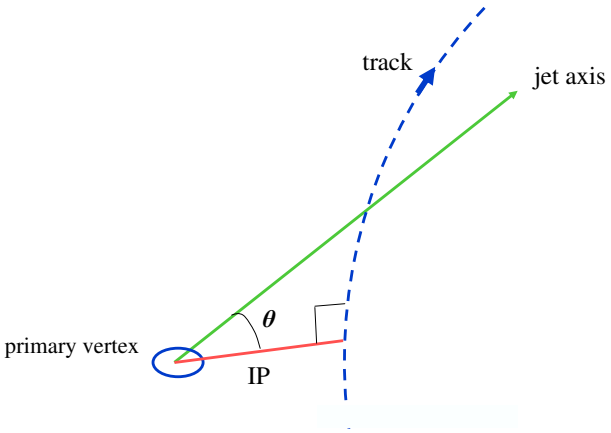

Figure 4.4: The signed impact parameter significance for a track associated to a jet. The track shown here has positive impact parameter significance. Negative signed impact parameter significance values are obtained for tracks with $\theta > \pi/2$.

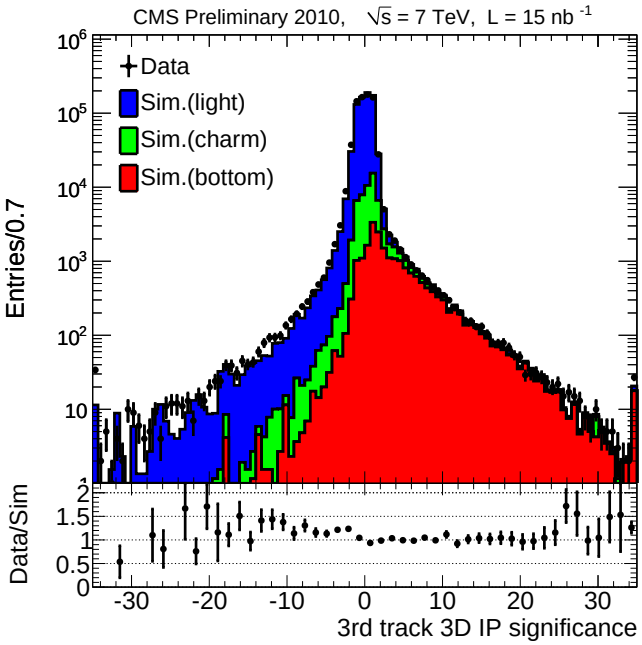

Figure 4.5: The signed impact parameter significance distribution for the third highest SIPS track associated to a jet. The points represent data, while the filled regions represent simulation (red = bottom jets, green = charm jets, blue = udsg jets) [37].

Using the signed impact parameter significance, a simplistic b-tagger, called the Track Counting tagger, can be defined. For this tagger, the jet-associated tracks are placed in descending order with respect to their signed impact parameter significance. Thus, the most significant tracks in the jet are first, and the least significant are last. The Track Counting tagger then makes a cut on the signed impact parameter significance of the *nth* track, the implication being that by making such a cut, one is counting how many tracks within the jet have an impact parameter significance above some value. Within CMS, two versions of this tagger are defined (Figure 4.6): the TrackCountingHighEfficiency tagger, which cuts on the impact parameter significance of the second track ($n = 2$), and the TrackCountingHighPurity tagger, which cuts on the third track ($n = 3$).

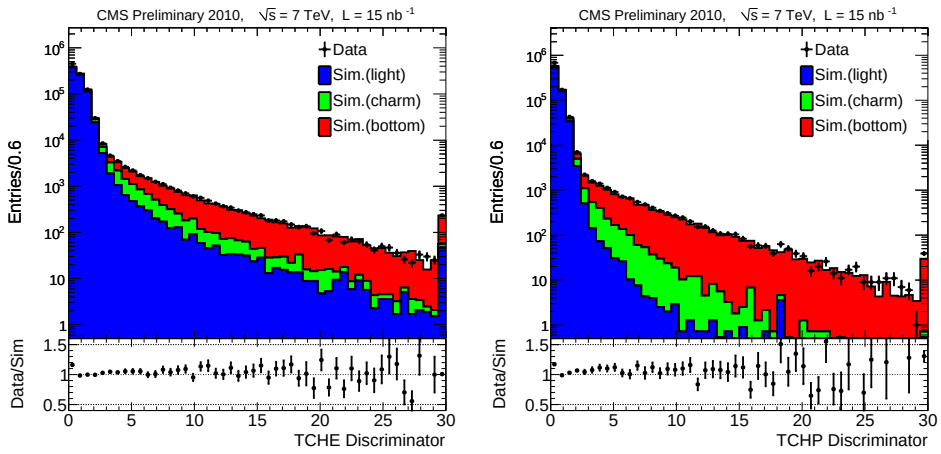

Figure 4.6: The TrackCountingHighEfficiency (left) and TrackCountingHighPurity (right) discriminator distributions [37].

The notion of the signed impact parameter significance of a track can be reconceptualized into a variable called track probability. Track probability indicates with what probability a track came from the primary vertex, given its signed impact parameter significance. The probability distribution functions (PDFs) are computed

using a sample of tracks in data likely to have come from the primary vertex, which is obtained by using tracks with negative signed impact parameter significance. The distributions are assumed to be the same for positive-signed tracks. To account for differences in PDF shape due to track quality, different PDFs are computed for tracks assigned to different quality classes. These classes are formed using the following variables [38].

- Number of pixel detector hits,
- $|\eta|$ of the track,
- Momentum p of the track,
- Normalized χ^2 of the track fit.

The track probability is then defined as [38]

$$P_{tr}(S) = \int_{|S|}^{\infty} \mathrm{PDF}(x)\, dx, \tag{4.14}$$

with S being the signed impact parameter significance of the track.

Track probability is important because it allows for the development of a tagger using information from all tracks in a jet, rather than just a single track. At CMS, these taggers are called the Jet Probability taggers. Two forms of these taggers exist at CMS: the JetProbability tagger, and the JetBProbability Tagger (Fig. 4.7). Both taggers follow the same principles, but the JetBProbability tagger gives more weight to the four highest signed impact parameter significant tracks. For both taggers, the track probability is converted to a jet probability, using the equation [38]

$$P_{jet} = \Pi \cdot \sum_{j=0}^{N-1} \frac{(-\ln \Pi)^j}{j!}, \tag{4.15}$$

with

$$\Pi = \prod_{i=1}^{N} \max(P_{tr}(i), 0.005) \tag{4.16}$$

being the product of probabilities of tracks associated to the jet. The cutoff value of 0.005 is introduced to avoid having tracks with probabilities close to zero spoiling the jet probability, though this leads to saturation peaks in the final discriminator [38]. Finally, the JetProbability discriminator is defined at [38]

$$D_{JP} = -\log(P_{jet})/4, \tag{4.17}$$

and the JetBProability discriminator is defined as

$$D_{JBP} = -\log(P_{jet}^{all})/4 - \log(P_{jet}^{4Trks})/4, \tag{4.18}$$

with P_{jet}^{4Trks} being the jet probability using only the four tracks with highest probability. If less than four jet tracks exist, then all the jet tracks are used.

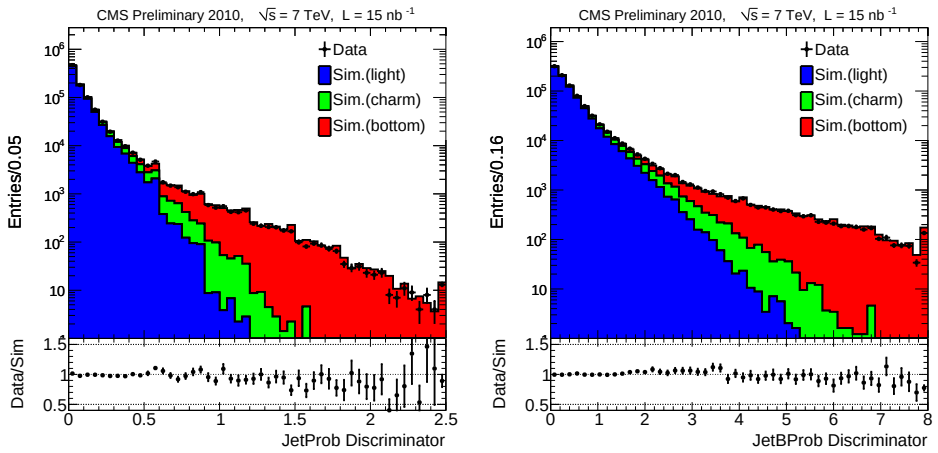

Figure 4.7: The JetProbability (left) and JetBProbability (right) discriminator distributions [37].

4.4 B-Tagging with Secondary Vertices

With tracks associated to a jet, it is also possible to reconstruct secondary vertices within the jet. Secondary vertices are defined as any vertices reconstructed outside of, or away from, the primary vertex. The presence of secondary vertices are a sign of particles with significant lifetime. Thus, secondary vertices within jets are a sign of b jets. Another powerful reason for the use of secondary vertices in b tagging is due to a side effect of b decays. From CKM theory, b hadrons will predominately decay to c hadrons, which themselves also have significant lifetimes. Thus, b jets have the potential to have two secondary vertices, and a great chance that at least one of them will be found in the reconstruction, making secondary vertex reconstruction very useful for b tagging.

To begin secondary vertex reconstruction, tracks are associated to jets, and are selected, as described in Section 4.3. To improve the quality of secondary vertices that are reconstructed, and reduce the fake rate, additional cuts are applied to the selected tracks [38]:

- ΔR between track and jet < 0.3,
- Distance of closest approach between track and jet < 0.07 cm.

The vertex reconstruction is performed using the adaptive vertex fitter, which is described in Section 4.2, but in an iterative manner. All selected jet-associated tracks are used to reconstruct a vertex. Those tracks which have a weight of 0.5 with respect to the reconstructed vertex (see equation (4.10)) are removed from the list, and the remainder of tracks are used to find a new vertex. The process continues until no more tracks remain, or a valid vertex cannot be found with the remaining tracks. The first attempt at reconstructing a vertex from jet-associated tracks is assumed

to be rediscovering the primary vertex (where many of the tracks will come from), and $\sigma_{cut} = 1.8$ in equation (4.10), and subsequent vertex reconstruction attempts use $\sigma_{cut} = 6.0$. The initial vertex fit also includes the beam spot position as a constraint to avoid, as much as possible, the association of secondary vertex tracks to the primary vertex. These cuts are selected to increase the b decay vertex finding efficiency, and immediately remove tracks highly compatible with the primary vertex [38].

Having reconstructed secondary vertices, further selection cuts are applied [38]:

- Fraction of secondary vertex tracks shared with the primary vertex < 0.65,
- Distance between secondary vertex and beam spot in the transverse plane < 2.5 cm,
- ΔR between the secondary vertex flight direction, and the jet axis < 0.5,
- For two-track vertices, the invariant mass of the vertex is not within the K_s mass window: $m_{K_s} = 0.5 \pm 0.05$ GeV,
- 2D decay length significance ($D_{xy}/\sigma_{D_{xy}}$) with respect to the primary vertex > 3.0,
- $D_{xy} > 0.1$ mm.

These cuts are defined to select significantly distant secondary vertices, while removing vertices composed of primary vertex tracks which failed the initial vertex fit, and long-lived decays that commonly occur in light jets: K_s and Λ hadrons, or other two-track processes, such as photon conversion to an e^+e^- pair.

CMS has two simple secondary vertex taggers which utilize the three-dimensional decay length significance of the secondary vertex relative to the primary vertex as

input for their discriminators, given by [38]

$$D_{SSV} = \log\left(1 + \frac{D_{3D}}{\sigma_{3D}}\right). \tag{4.19}$$

What distinguishes these two taggers is the number of tracks associated with the secondary vertex. The more efficient tagger, called the SimpleSecondaryVertexHighEfficiency (SSVHE) tagger, requires secondary vertices to have at least two tracks, while the other, the SimpleSecondaryVertexHighPurity (SSVHP) tagger, requires three-track vertices (Fig. 4.8). One can note the relative fraction of b jets is much larger for the high purity case than the high efficiency.

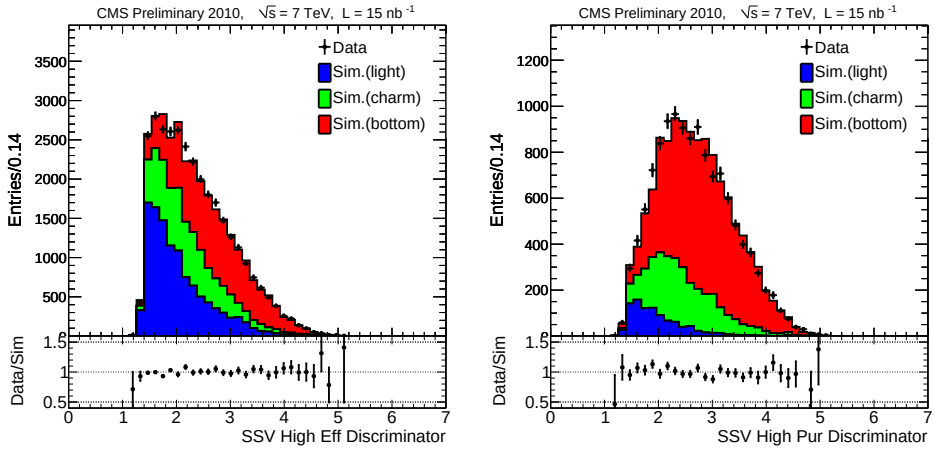

Figure 4.8: The SimpleSecondaryVertexHighEfficiency (left) and SimpleSecondaryVertexHighPurity (right) discriminator distributions [37].

While very powerful, the simple secondary vertex taggers suffer from an inherent limitation of not being fully efficient in finding b jets, due to decays which lack the requisite number of tracks for vertex reconstruction, tracks from decay which fail the track selection cuts, or vertices which fail the vertex selection cuts. This is not ideal, as it would be useful to have a tagger which is capable of finding b jets with 100%

efficiency. In addition, it is useful to develop a tagger that reduces the mistag rate, while still being highly efficient. This led to the development of the combined secondary vertex tagger. At its heart, it is a secondary vertex tagger, but it utilizes more information obtained from the vertex to increase the discrimination power between b jets and light jets. To make the tagger fully efficient, it uses information from the jet-associated tracks alone, when a reconstructed vertex is not present.

The combined secondary vertex tagger places a given jet into one of three categories [37]:

- RecoVertex: The jet contains a valid, reconstructed secondary vertex.
- PseudoVertex: The jet contains no reconstructed secondary vertex, but contains enough high impact parameter significant tracks (2D signed impact parameter significance > 2.0 cm) that can be treated as though they form a vertex.
- NoVertex: The jet contains no reconstructed vertex, or not enough high impact parameter tracks.

Based on whichever category the jet falls, different variables (Fig. 4.9) may be calculated for use with the combined tagger [37]:

- 2D flight distance significance (RecoVertex).
- ΔR between the vertex flight direction and jet axis (RecoVertex): real secondary vertices will tend to be collinear with the jet axis, whereas fake or bad vertices will tend to be more random.
- Vertex mass (RecoVertex, PseudoVertex): The invariant mass of the vertex assuming all of the tracks are pions. A correction is applied to recover momentum

lost due to invisible particles by using the transverse momentum p_T of the momentum sum vector of the vertex tracks with respect to the vertex fight vector. The final equation for the mass is given by

$$M_{corr} = \sqrt{M_{raw}^2 + p_T^2} + |p_T|. \tag{4.20}$$

- Number of vertex tracks (RecoVertex, PseudoVertex).
- Vertex energy ratio (RecoVertex, PseudoVertex): the energy ratio of the vertex tracks with respect to all tracks associated to the jet.
- $|\eta_{rel}|$ of all vertex tracks (RecoVertex, PseudoVertex): the pseudorapidity of each vertex track with respect to the jet axis.
- 2D signed IP significance of the first track above the charm threshold (RecoVertex, PseudoVertex): The jet-associated tracks are ordered by 2D IP significance. The invariant mass of the tracks are computed, adding one track at a time. 2D IP significance of the first track to lift the invariant mass above 1.5 GeV (the charm threshold) is used.
- Number of selected tracks in the jet (all categories).
- 3D signed impact significance of all selected tracks (all categories).

PDFs are formed for each variable, for each category where the variable applies, in bins of p_T and $|\eta|$ defined by divisions of $p_T = 40, 60, 90, 150$ GeV and $|\eta| = 1.2, 2.1$. The pdfs are then combined into a likelihood ratio, using the function [37]

$$\mathrm{LR}_{b,c,udsg} = \frac{\mathcal{L}_b}{\mathcal{L}_b + \mathcal{L}_{c,udsg}} \tag{4.21}$$

$$\mathcal{L} = \prod_i p_{b,c,udsg}(x_i) \tag{4.22}$$

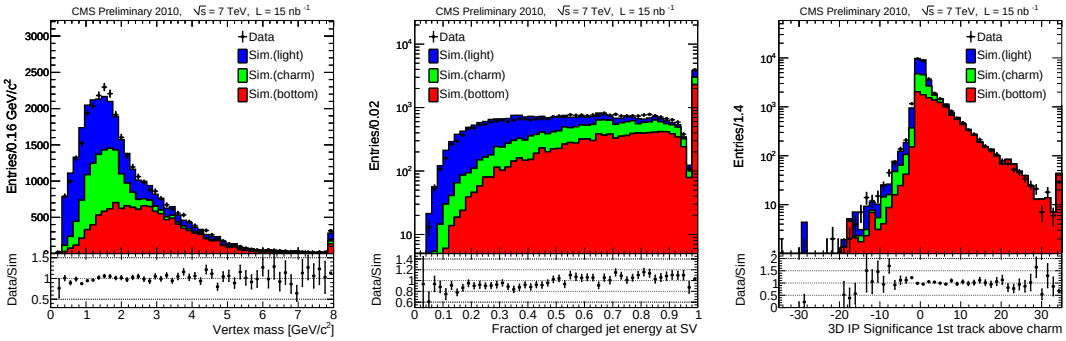

Figure 4.9: The vertex mass (left), vertex energy ratio (middle), and IP significance above charm threshold (right) distributions [37].

with $p_{b,c,udsg}(x)$ being the PDF value at x that the jet is a b, c, or light (udsg) jet. The likelihood ratio is computed twice to determine the likelihood the jet is b vs c, and b vs udsg. These two values are then combined to form the combined secondary vertex discriminator [37]:

$$D_{CSV} = 0.75L_{b,udsg} + 0.25L_{b,c}. \tag{4.23}$$

The values of 0.75 and 0.25 for the flavor composition of the discriminator comes from the flavor composition of hadronic $t\bar{t}$ decays. The use of categories allows for all jet information to be used in the discrimination, and the use of a likelihood allows for better separation between the jet flavors, so that the mistag rate is reduced (Fig. 4.10).

Figure 4.11 shows the b-tagging performance versus the non-b mistag rate for the impact parameter and secondary vertex-based b-taggers.

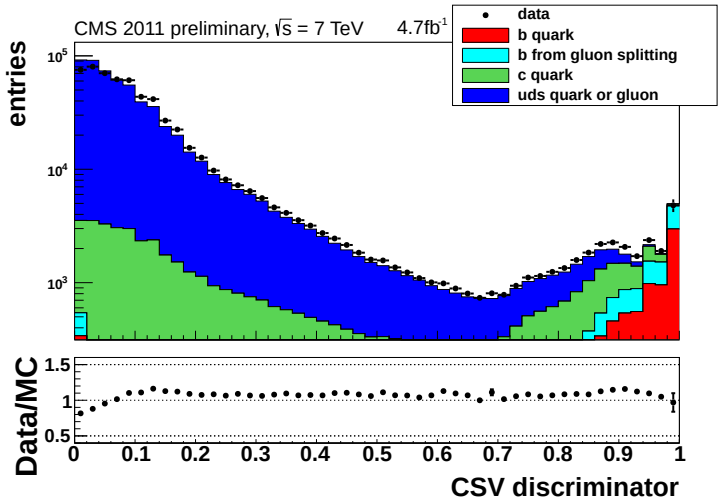

Figure 4.10: The CSV discriminator distribution [39].

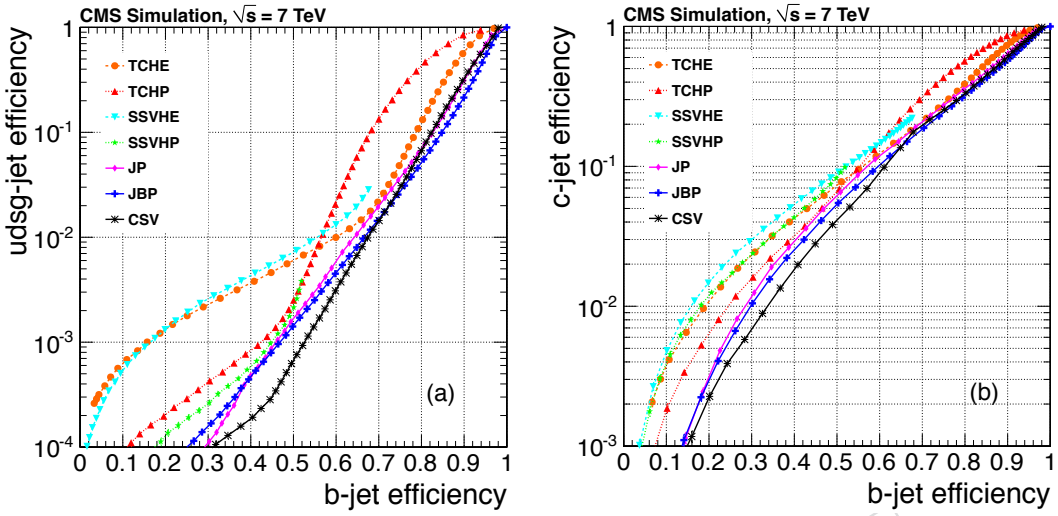

Figure 4.11: The b-tagging efficiency versus non-b mistag rate for both light quark (a) and c quark (b) jets for the impact parameter and secondary vertex-based taggers [40].

4.5 B-Tagging with Soft Leptons

Due to the weak nature of b hadron decays, it is also possible to tag b jets using soft (low p_T) leptons (e, μ) within jets. The branching ratios of b quarks to leptons gives an a priori estimate of the maximal b-tagging efficiency. The branching ratio for the process $b \to l\nu X$ is $10.69 \pm 0.22\%$ [2]. There is an additional contribution from the semileptonic cascade decay $b \to c \to l\nu X$, which is $9.6 \pm 0.4\%$ [2], for a total maximum tagging efficiency of $20.29 \pm 0.45\%$ per b jet.

However, these are not the only sources of reconstructed leptons in jets. Pions in jets can fake muons by punching through to the muon detector and leaving hits associated with a track, thus being reconstructed as a muon. Pions can also leave a signature in the electromagnetic calorimeter similar to electrons, and be incorrectly reconstructed as such. Real decays from non-b sources also contribute to the soft lepton tagging background. Hadrons from c quarks also decay weakly, and will also provide a natural source of leptons. Also, neutral pions can provide real electrons through the decay $\pi^0 \to ee\gamma$, and the much more dominant decay of $\pi^0 \to \gamma\gamma$, which can contribute electrons through photon conversion.

To reduce the background, and make a more effective b-tagger with leptons, it is necessary to develop a discriminator variable, as with the case of tracks and secondary vertices. A natural discriminator for leptons is the signed impact parameter significance (Fig. 4.12), since both electron and muon are a track produced from a b decay.

A second discriminator for soft lepton b-tagging is a quantity called $p_{T,Rel}$, defined as the transverse momentum of the lepton with respect to the original b hadron direction (Fig. 4.13). Since the b hadronization carries with it a large fraction of the original parton momentum, the reconstructed jet direction may be used to represent

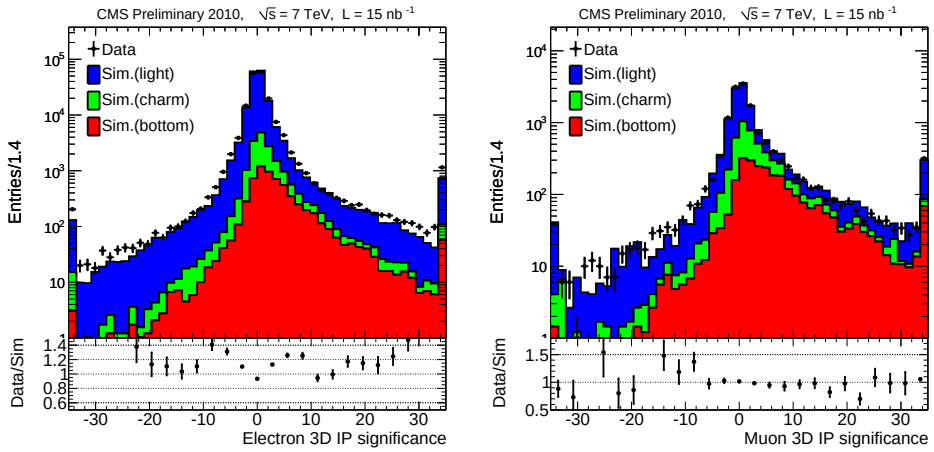

Figure 4.12: The 3D signed impact parameter significance for electrons (left) and muons (right) [37].

the b hadron direction to good approximation. Due to the much larger mass of b hadrons compared to hadrons from other quarks ($\sim$ 5 GeV for b hadrons vs. $\sim$ 2 GeV for c hadrons and $\sim$ 0.1 GeV for udsg hadrons), the $p_{T,Rel}$ from a b hadron decay should be larger than for lighter hadron decays.

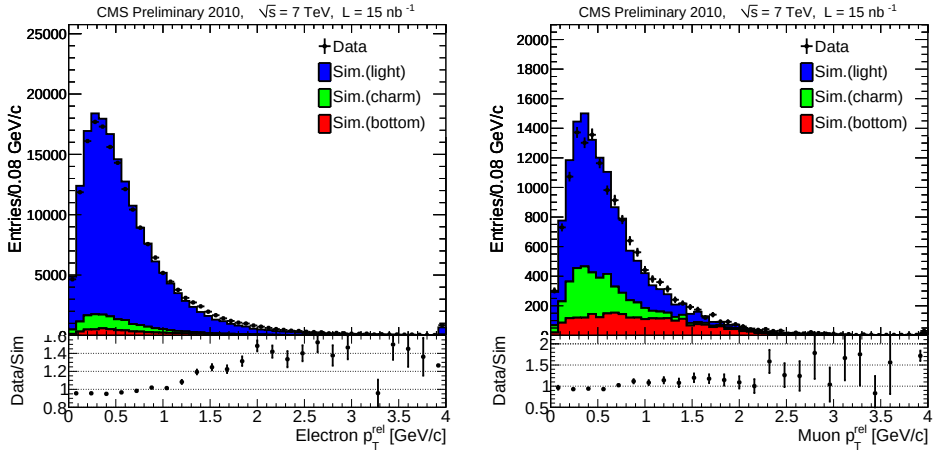

Figure 4.13: The $p_{T,Rel}$ distribution for electrons (left) and muons (right) [37].

CMSSW currently implements both lepton discriminators as taggers in its recon-

struction software. For both leptons, $\Delta R < 0.4$ is required between the lepton and jet directions [37]. Due to the very low background, muons may fulfill very lax selection criteria, requiring only that the muon is reconstructed using the global muon reconstruction algorithm. Due to higher background from fake sources, electrons must withstand more stringent selection. These values were chose to maximize the signal significance of b-jets tagged from electrons compared to jets tagged from fake electron sources. The concept of signal significance is discussed in greater detail in Section 5.1.

- For soft electrons in the barrel:
  - electron $p_T > 2.0$ GeV.
  - ΔR between electron super cluster and track < 0.017.
  - $0.05 <$ Electron supercluster energy / electron track momentum < 2.28
  - ParticleFlow mva discriminator > -0.1
- For soft electrons in the endcap:
  - electron $p_T > 2.0$ GeV.
  - ΔR between electron super cluster and track < 0.006.
  - Inverse brem fraction < 7.0 (see Section 3.4).
  - ParticleFlow mva discriminator > -0.24.

As stated earlier in this section, the maximum tagging efficiency using soft leptons is $\sim 20\%$, which can make the method appear unattractive compared to the taggers discussed in sections 4.3 and 4.4. However, there are instances in which tagging with soft leptons can be very useful. Soft lepton tagging provides information about the charge or flavor of the original b hadron, depending on whether the hadron is charged

or neutral. This can be useful in measuring the charge of the top quark [41], or in analyzing CP violation in neutral B mesons [42]. Soft lepton tagging can also provide a highly orthogonal sample to an analysis primarily using secondary vertex tagging since soft lepton decays will not produce a reconstructed secondary vertex in many cases. Analyses using soft lepton tagging are also orthogonal to each other, provided they do not use the same lepton, since the probability of having two soft lepton tags between two b jets is low.

4.6 Validation of B-Tagging Algorithms

The previous sections in this chapter discuss the expected behavior of b-tagging algorithms implemented at CMS, both in how these algorithms work and their discriminator distributions per jet flavor. An analyzer wishing to use one of these b-taggers needs to have a high level of confidence the tagger is working as expected. To ensure this confidence is justified, the CMS b-tagging physics object group (POG) has implemented a set of software packages which reproduces the histograms shown above (along with many more), and allows for the comparison of these distributions between CMS software releases to determine if any unexpected differences exist, and by what degree any known software changes implemented affect the b-tagging. This section describes the process of validating the b-tagging algorithms.

The b-tagging validation software utilizes a CMS software framework known as data quality monitoring (DQM). This DQM framework provides the ability to produce histograms for reconstructed objects on the fly (Section 2.2.7). DQM provides tools for producing histograms for online and offline settings. Histograms are produced online during central shifts that monitor detector performance during data taking. The online histograms are produced using output from various components of the

detector itself. By looking at the histograms as they are produced per run, a shifter can make the decision of whether a run, or sections of the run, need to be flagged as bad based on detector performance. Histograms are produced offline once the offline reconstruction has run. The offline histograms are then used by a validator to determine if the reconstruction is producing reasonable output compared to some standard expectation. The b-tag validation falls into the offline DQM category. If differences appear in the b-tag histograms, the validator must then determine the cause of these difference. Since b-tagging takes output from many pieces of the reconstruction chain as input, it is important for the validator to be aware of any changes made in the reconstruction code, and how these changes may impact b-tagging performance.

The b-tagging validation package is capable of running on both data and simulation samples, and produces the same set of histograms for both cases. However, the validation package can produce histograms for the different jet flavors (b,c, and udsg), since the generator level information is available. The list of histograms produced by the validation suite is given below [43].

- non-b jet efficiency vs b jet efficiency (simulation): The fraction of non-b jets which pass a given discriminator at a given cut are plotted vs the fraction of b jets which pass the same cut. This plot is made separately for c and udsg jets, and for each discriminator listed in the previous sections of this chapter.

- efficiency vs discriminator cut (simulation): The fraction of jets of a given flavor that pass a given cut for a given discriminator. The plot is made separately for b, c, and udsg jets, and for each discriminator listed in the previous sections of this chapter.

- discriminator (data, simulation): The discriminator histogram for the b-taggers

discussed previously.

- input variables (data, simulation): Each tagger takes given variables (signed impact parameter significance, secondary vertex decay length, secondary vertex mass, etc.) as input. Each of these input variables are plotted.

- tag correlation plots (data, simulation): The tag discriminators are plotted against each other, two at a time, to look for correlations between them. If requested, profile histograms are also produced.

The validation package also allows for these histograms to be produced in bins of jet p_T and $|\eta|$ to examine b-tagging dependence in different regions of detector and phase space (Fig. 4.14).

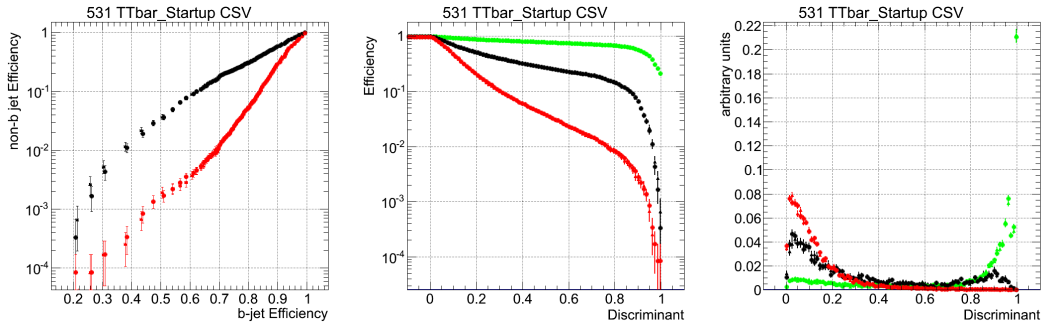

Figure 4.14: The non-b jet vs b jet efficiency (left), efficiency vs discriminator cut (middle), and discriminator (right) plots for the combined secondary vertex tagger. Black dots represent c jets, red dots represent udsg jets, and green dots represent b jets.

For the purpose of release validation, histograms are produced on two types of datasets. For validation with data, there are the ValSkim datasets, which are produced from a very small subset of a single run of data taking. To be used in the ValSkim dataset, the data must have been taken during a period where all sub detectors are marked as "good," and when there are good collisions taking place. For

validation with simulation, release validation (relval) samples are produced. These samples contain a small number of events (~ 9000), so they may be produced quickly. Multiple event processes are produced to accommodate the many POGs at CMS. For validating b-tagging software, the relval $t\bar{t}$ and QCD samples are used, since they both contain b jets, but are expected to have jets at different multiplicities.

Chapter 5

Soft Electron Tagging

The previous chapter describes what b-tagging means at CMS, and demonstrates that there are many ways to tag b jets. This analysis focuses on tagging b jets using the presence of soft electrons within jets. This alone does not fully describe how the tagging process works, since it has not yet been stated what the quality of the electrons is, what the kinematic requirements are, or how the discriminators discussed in Section 4.5 are used. This chapter will define the soft electron tagger used in this analysis, describe the process used to develop this definition, and how the resulting tagging efficiency is measured.

5.1 Tagger Definition

The electrons used for tagging come from the Particle Flow reconstruction, as described in Section 3.4. These electrons are chosen because of the demonstrated ability of the algorithm to find low momentum electrons in non-isolated environments, as is required for tagging b jets. To increase the likelihood the electron comes from a hadron decay, the tagging electron is required to be within $\Delta R < 0.4$ of a jet fulfilling

the selection criteria given in Section 6.2.1. Additionally, the larger mass of B hadrons compared to those from lighter quarks means electrons from b decay will have higher transverse momentum than those produced from lighter sources. To accommodate this, a cut of $p_T > 5.0$ GeV is applied to tagging electrons. This p_T requirement is low in comparison to many analyses looking for isolated electrons from prompt sources, which generally require $p_T > 20$ GeV or more.

Even with the selection cuts previously discussed, there is still some allowance for more stringent selection to increase the purity of $t\bar{t}$ events over the background. Three variables which can be used to achieve this goal are

- Particle Flow electron mva (see Section 3.4).
- electron $p_{T,Rel}$ (see Section 4.5).
- electron signed impact parameter significance (see Section 4.5).

The utility of these variables can be determined by calculating how cuts on these variables (or any combination of these variables) affects the signal significance, defined as

$$S(x) = \frac{N_S(x)}{\sqrt{N_S(x) + N_B(x)}}, \tag{5.1}$$

with $N_{S,B}(x)$ being the number of signal and background events passing a cut at x of a given variable. x can also represent a set of variables, in which case x becomes $\vec{x}$ and the significance is calculated for a given cut in the $\vec{x}$ parameter space. The goal is to find the cut, or set of cuts $\vec{p}$, which maximize S:

$$\left.\frac{\partial S}{\partial \vec{x}}\right|_{\vec{x}=\vec{p}} = 0. \tag{5.2}$$

Since the functional forms of $N_{S,B}(\vec{x})$ are often not known a priori, the typical method for finding the maximum significance is to count the number of signal and background events passing an applied cut in simulation, and plotting S as a function of $\vec{x}$.

Figure 5.1 shows the distributions for the electron mva, $p_{T,Rel}$ and signed impact parameter significance for both data and simulation (simulation weighted to data luminosity) and the resulting signal significance as a function of discriminator value. The event selection for these plots is given in Section 6.2.1. Given the tightness of the event selection cuts already in place, there is no significance gain by using either of the three discriminators, and thus, no additional cut is applied.

5.2 Tagging Efficiency

Given the selection found in the previous section, the next step is to calculate the tagging efficiency. For this analysis, the tagging efficiency is calculated using a $t\bar{t}$ simulation sample. To account for differences between data and simulation, a data-to-simulation scale factor is applied. The scale factor is determined by calculating the soft electron reconstruction efficiency in both data and Monte Carlo samples.

The reconstruction efficiency is estimated by comparing the number of B meson decays that result in soft electrons versus soft muons. This method works by using the semileptonic branching fraction of B mesons. For any decay, the branching fraction is defined empirically as

$$\Gamma_{X\to Y} = \frac{N^{Total}_{X\to Y}}{N^{Total}_{X}} = \frac{N^{Obs}_{X\to Y}}{N^{Total}_{X} A_{X\to Y} \epsilon_X \epsilon_Y}, \tag{5.3}$$

with $N^{Total/Obs}$ being the total/observed number of decays, $A_{X\to Y}$ being the acceptance of the $X \to Y$ process, and $\epsilon_{X/Y}$ being the reconstruction efficiencies of X/Y.

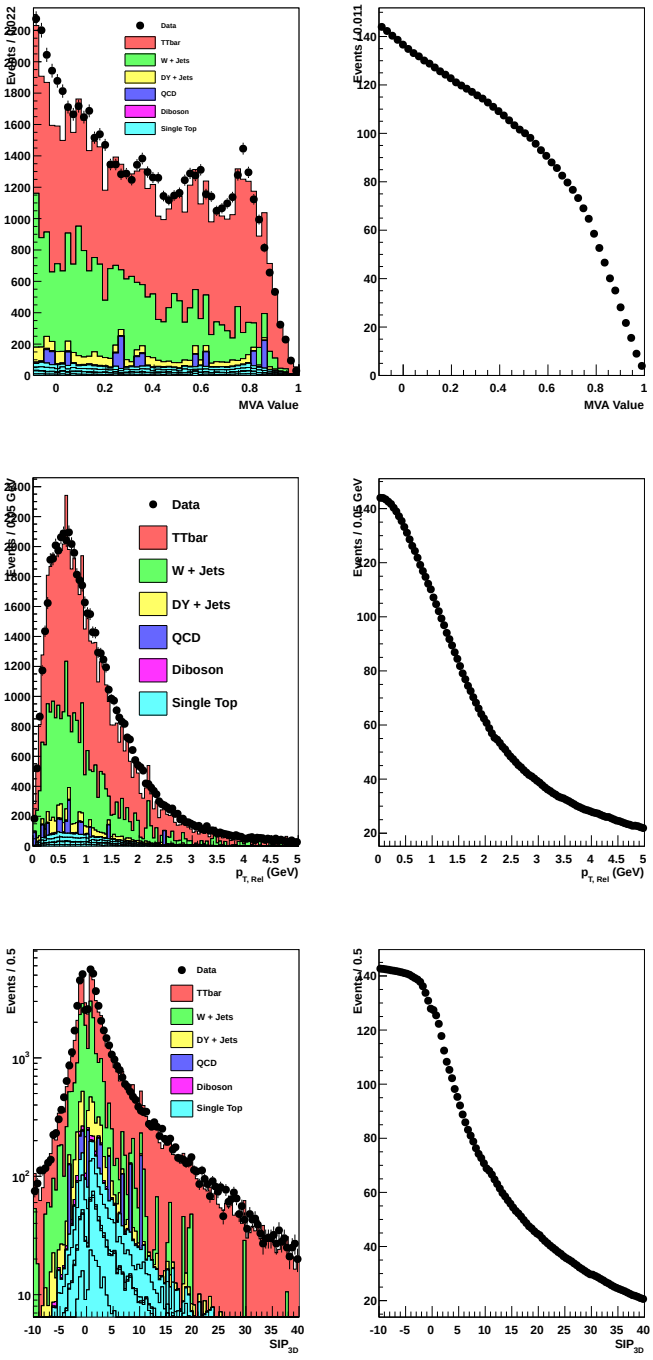

Figure 5.1: Discriminator distribution (left) and signal significance plots (right) for soft electrons in events selected for a $t\bar{t}$ analysis (Section 6.2.1). Simulation is normalized to data luminosity.

Comparing the branching ratio of $X \to Y$ and $X \to Z$, using equation (5.3), then

$$\frac{\Gamma_{X\to Y}}{\Gamma_{X\to Z}} = \frac{N^{Obs}_{X\to Y}}{N^{Obs}_{X\to Z}} \frac{A_{X\to Z}}{A_{X\to Y}} \frac{\epsilon_Z}{\epsilon_Y} \tag{5.4}$$

is obtained. Notice that ϵ_X and N^{Total}_X cancels out when taking the ratio. Using the decay $B \to Dl\nu$, with $l = e, \mu$, and knowing the ratio of the branching fractions for the $B \to Dl\nu$ process is 1 [2], equation (5.4) becomes

$$\epsilon_e = \epsilon_\mu \frac{N^{Obs}_{B\to De\nu}}{N^{Obs}_{B\to D\mu\nu}} \frac{A_{B\to D\mu\nu}}{A_{B\to De\nu}}. \tag{5.5}$$

Thus, the reconstruction efficiency for electrons from B meson decay can be estimated by using the muon reconstruction efficiency, and the relative ratios of the observed B decays. The acceptance quantities are calculated using simulation. The muon reconstruction efficiency, ϵ_μ, and the $N^{Obs}_{B\to Dl\nu}$ quantities are calculated using data, and the methods used are discussed below. The data samples used to calculate these quantities are listed in Section 3.1.

To find the number of observed B decays for each lepton, it is necessary to reconstruct the B decay vertices. A specific decay channel is selected for the reconstruction, which is

$$\begin{aligned} B^0 &\to D^*(2010)^- l^+ \nu_l, \\ D^*(2010)^- &\to \bar{D}^0 \pi^-, \\ \bar{D}^0 &\to K^+ \pi^-. \end{aligned}$$

This decay leaves a lepton and three charged hadrons in the final state. The branching ratio for the B meson decay above is $4.95 \pm 0.110\%$. The branching ratios for the decay

of the excited charged and neutral D mesons are $67.7 \pm 0.500\%$ and $3.88 \pm 0.0500\%$, giving a branching ratio for the overall process of $0.130 \pm 0.00350\%$.

To reconstruct the B meson, each of the decays listed above are reconstructed sequentially. First, the neutral D meson is found by forming a vertex from the tracks of two charged hadrons. The mass of the D meson is used as a kinematic constraint in the vertex fit [26], which is 1.86 ± 0.0013 GeV [2]. Next, the excited D meson is reconstructed using the resulting neutral D meson vertex and a third charged hadron, using the mass difference between the excited and neutral D mesons as a kinematic constraint, which is 0.145 ± 0.0000100 GeV [2]. Finally, the B meson vertex is fit with the excited D meson and a lepton. Since the neutrino from the B decay is undetectable, the mass of the B vertex will not be the true B mass, so no constraint is applied.

Each of the reconstructed objects used in the above reconstruction must pass the following selection cuts.

- Charged Hadrons:
  - The charged hadrons must be PFChargedHadrons (Section 3.5.1).
  - $p_T > 0.6$ GeV and $|\eta| < 2.5$.
  - The number of track hits for the charged hadron > 7, the number of pixel hits > 1, and the normalized $\chi^2 < 5.0$.
  - $|d_0| < 0.2$ cm and $|d_z| < 0.5$ cm, with both measured with respect to the primary vertex.
  - The three charged hadrons must pass clusterization using the same technique as the Iterative Vertex Finder [44].

  - The charge sum of the three tracks is required to be the opposite of the lepton.
  - The kaon is identified as the sole track from the decay which has the same sign as the lepton.
  - The pion from the D^* decay is required to have a p_T less than the pion from the D^0 decay (slow pion).
  - The primary vertex weight (equation (4.10)) for the kaon and pion < 0.5.

- Muons
  - The muon must be a PFMuon, GlobalMuon, and TrackerMuon (Section 3.3).
  - $p_T > 5.0$ GeV and $|\eta| < 2.5$.
  - The global track normalized $\chi^2 < 10.0$ and the number of muon detector hits > 0.
  - The inner track must have at least 5 layers with hits and at least 1 pixel hit.
  - There must be at least 2 matched stations.

Electrons must pass the selection given in Section 5.1. Finally, the D meson vertices must have a vertex χ^2 fit probability of > 0.01, while the B vertex must have a χ^2 fit probability > 0.001.

Figures 5.2 and 5.3 show the resulting D^0 and $D^* - D^0$ mass difference between data and simulation. The simulation sample represented in these plots is the $t\bar{t}$ sample listed in Section 3.1. The background represented in the plots indicates reconstructed B vertices in which the reconstructed lepton or charged hadrons are not appropriately

identified using Monte Carlo truth. Figure 5.4 shows the B mass plots divided into lepton p_T bins of $[5.0, 9.0]$, $[9.0, 14.0]$, and $[14.0, 20.0]$ GeV. These plots show good agreement for the reconstructed B mass. The data and simulation histograms in these plots are normalized to unit area, with the relative signal and background contributions coming from simulation. Data and simulation agree reasonably well, though slight shift in the central mass values between between the two point to small discrepancies in how the simulation models the data.

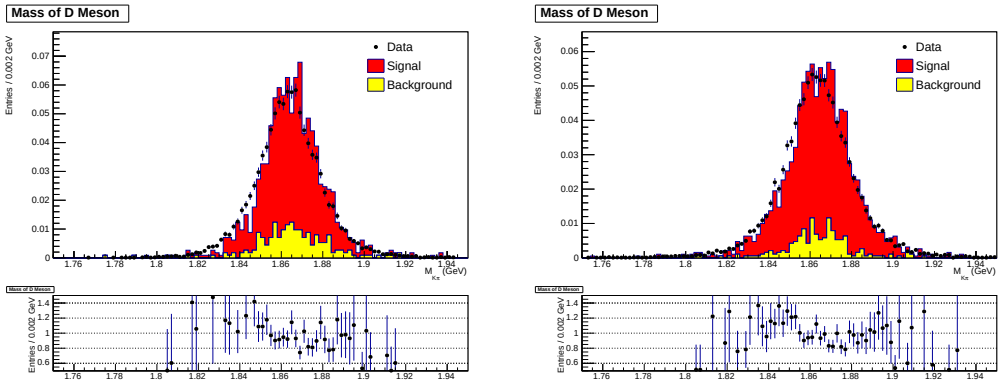

Figure 5.2: The D^0 meson mass for electrons (left) and muons (right).

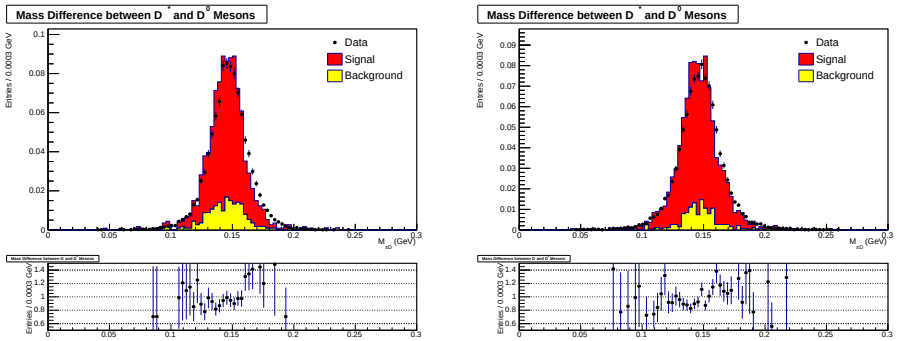

Figure 5.3: The $D^* - D^0$ meson mass difference for electrons (left) and muons (right).

To determine the number of observed B decays in data, RooFit [45] is used to find the relative signal and background contributions to the B mass distribution.

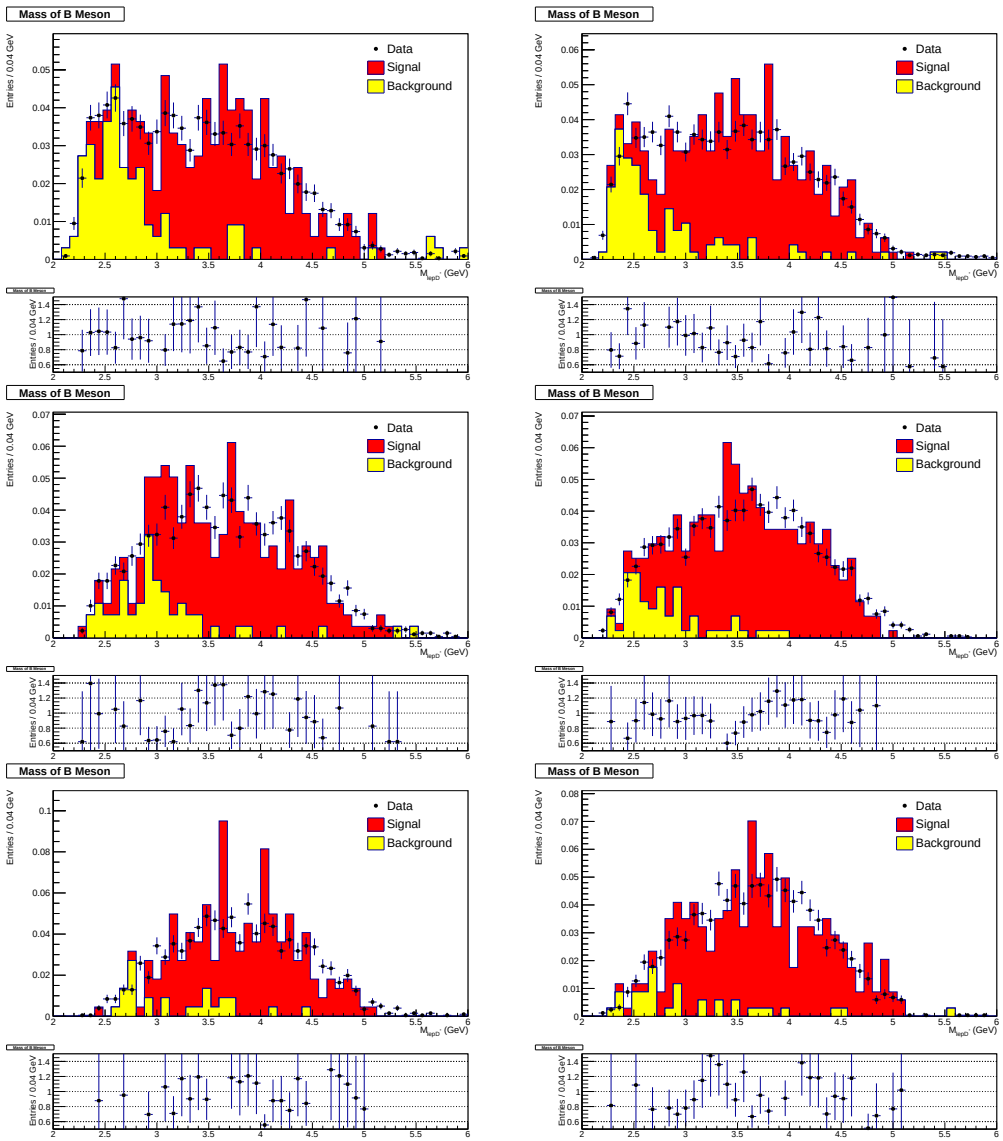

Figure 5.4: The B meson mass for electrons (left) and muons (right) in lepton p_T bins of [5.0, 9.0] (top), [9.0, 14.0] (middle), and [14.0, 20.0] GeV (bottom).

The signal is formed by using the RooKeysPdf class, which constructs the shape by using kernel estimation [46]. The signal mass shape is obtained from simulation. The background mass distribution is modeled using a gamma distribution, which has the functional form

$$f(x;\mu,\gamma,\beta) = \frac{1}{\Gamma(\gamma)\beta^{\gamma}}(x-\mu)^{\gamma-1}\exp\left(-\frac{x-\mu}{\beta}\right), \tag{5.6}$$

with β being the scale parameter, γ the shape parameter, μ the location parameter, and $\Gamma(\gamma)$ is the gamma function

$$\Gamma(\gamma) = \int_0^{\infty} t^{\gamma-1}\exp\left(-t\right)\,dt. \tag{5.7}$$

This distribution was chosen because it characterizes the low mass peak, but long tail, of the background quite well. The fits are performed in lepton p_T bins of [5.0, 9.0], [9.0, 14.0], and [14.0, 20.0] GeV using an unbinned maximum likelihood fit. The resulting fits are shown in Figures 5.5 - 5.7. Table 5.1 shows the outcome of the fit, and the number of observed B decays for each lepton in each p_T bin.

As a test to determine how well the soft electron reconstruction efficiency can be estimated, the method is applied to the $t\bar{t}$ simulation sample listed in Table 3.1. The simulation sample contains the Monte Carlo truth information regarding what particles are produced and how, allowing for the calculation of the reconstruction efficiency of electrons and muons from B decay from simulation. The muon reconstruction efficiency and the B meson decay fits can be used in Equation (5.5) to estimate the electron reconstruction efficiency and see how it compares to the actual value. The result is shown in Figure 5.8. The agreement between the estimated and actual electron reconstruction efficiencies is quite good, showing that the method can

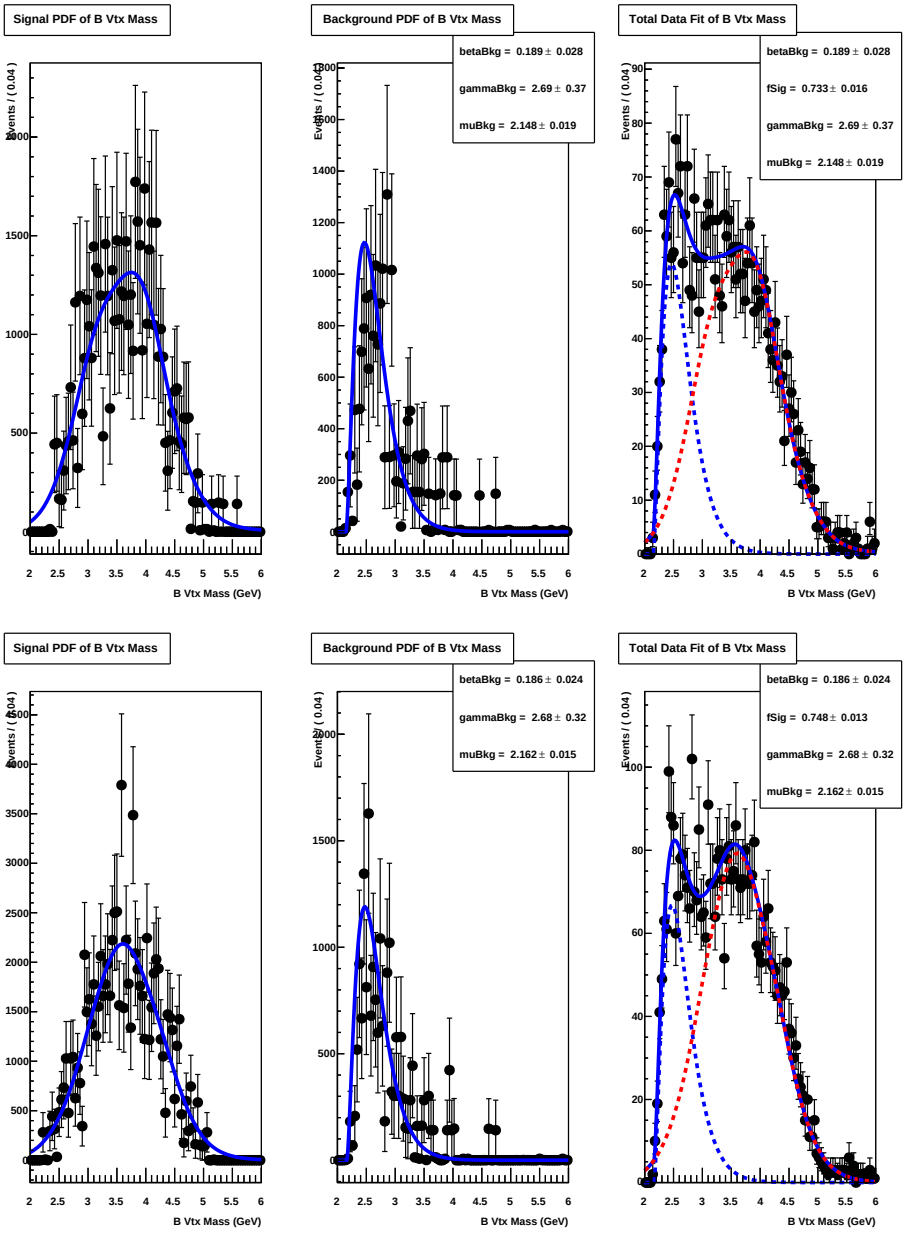

Figure 5.5: The B meson mass fit for electrons (top) and muons (bottom) in lepton p_T bins of [5.0, 9.0] GeV. Each subplot shows the kernel estimated signal pdf compared to simulation signal (left), the resulting gamma distribution fit from data compared to simulation background (middle), and the total fit in data (right). The total fit is shown as the solid blue line, the background contribution as the dotted blue line, and the signal contribution as the dotted red line.

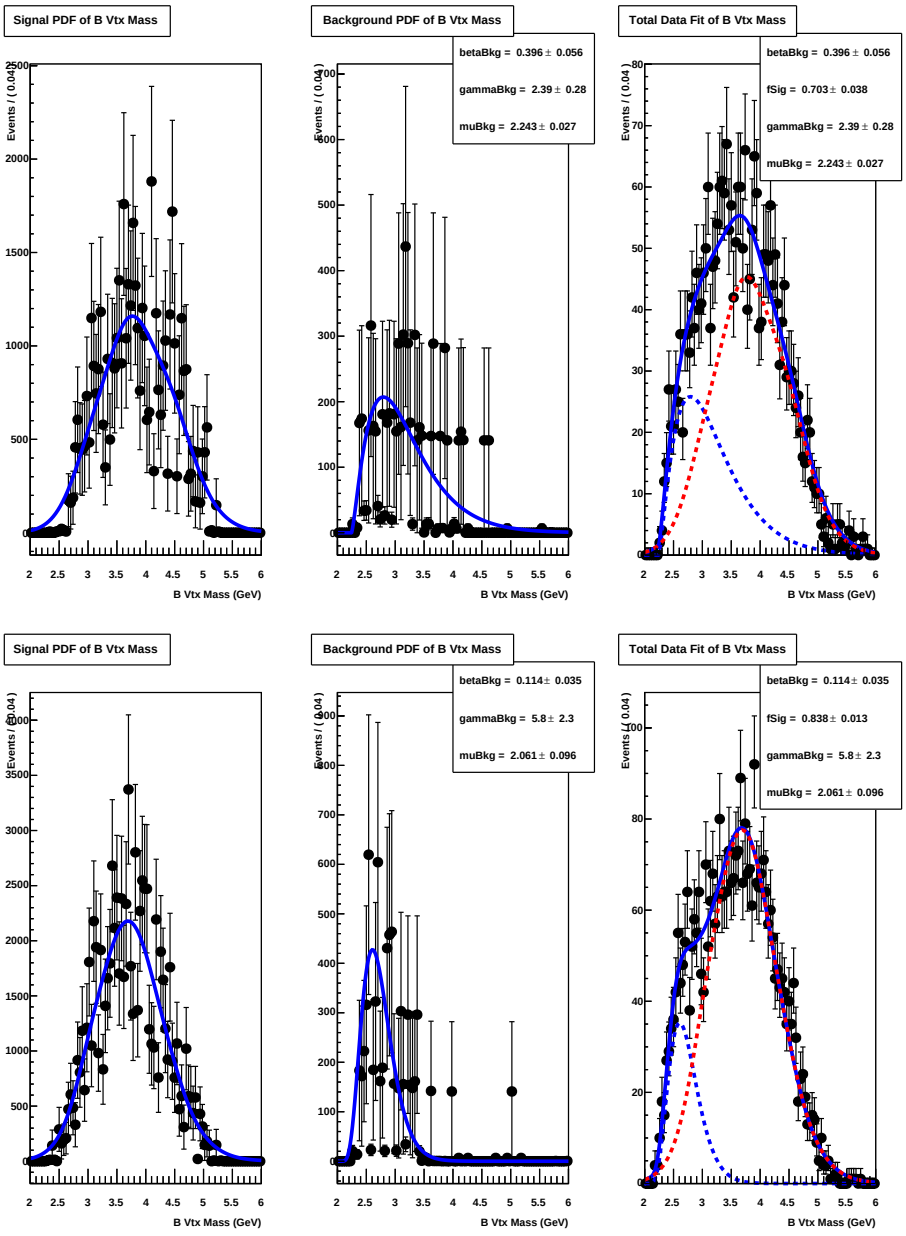

Figure 5.6: The B meson mass fit for electrons (top) and muons (bottom) in lepton p_T bins of [9.0, 14.0] GeV. Each subplot shows the kernel estimated signal pdf compared to simulation signal (left), the resulting gamma distribution fit from data compared to simulation background (middle), and the total fit in data (right). The total fit is shown as the solid blue line, the background contribution as the dotted blue line, and the signal contribution as the dotted red line.

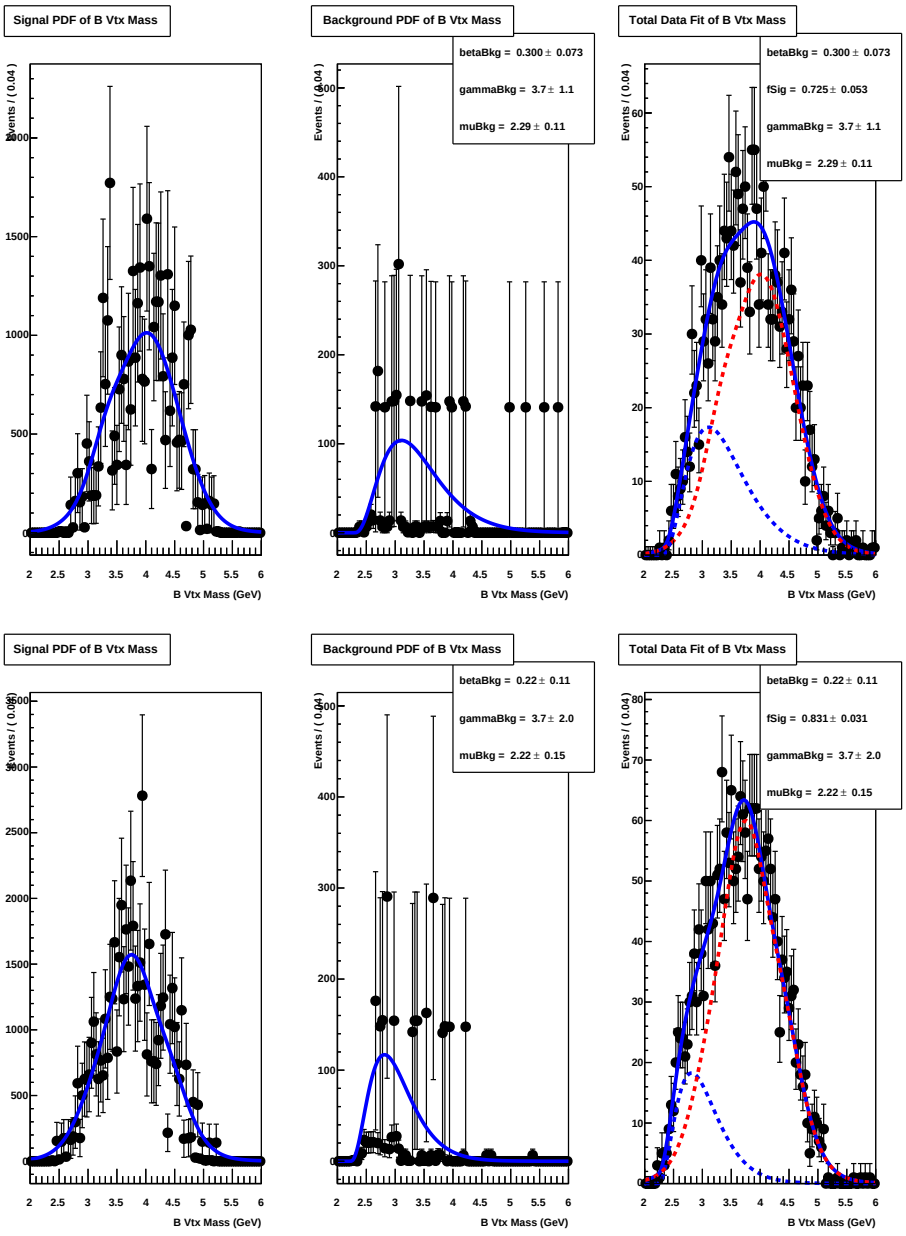

Figure 5.7: The B meson mass fit for electrons (top) and muons (bottom) in lepton p_T bins of [14.0, 20.0] GeV. Each subplot shows the kernel estimated signal pdf compared to simulation signal (left), the resulting gamma distribution fit from data compared to simulation background (middle), and the total fit in data (right). The total fit is shown as the solid blue line, the background contribution as the dotted blue line, and the signal contribution as the dotted red line.

Table 5.1: The total number of events passing the B vertex selection, the fractional signal contribution from the mass fit, and the number of observed B decays per lepton.

	Electron			Muon		
p_T Bin (GeV)	Total	Signal Fraction	Observed	Total	Signal Fraction	Observed
5.0 - 9.0	3265 ± 57	0.733 ± 0.016	2393 ± 67	4198 ± 65	0.748 ± 0.013	3140 ± 73
9.0 - 14.0	2691 ± 52	0.703 ± 0.038	1892 ± 109	3456 ± 59	0.838 ± 0.013	2896 ± 67
14.0 - 20.0	2013 ± 49	0.725 ± 0.053	1459 ± 112	2519 ± 50	0.831 ± 0.031	2093 ± 88

be a useful tool for determining the soft electron reconstruction efficiency.

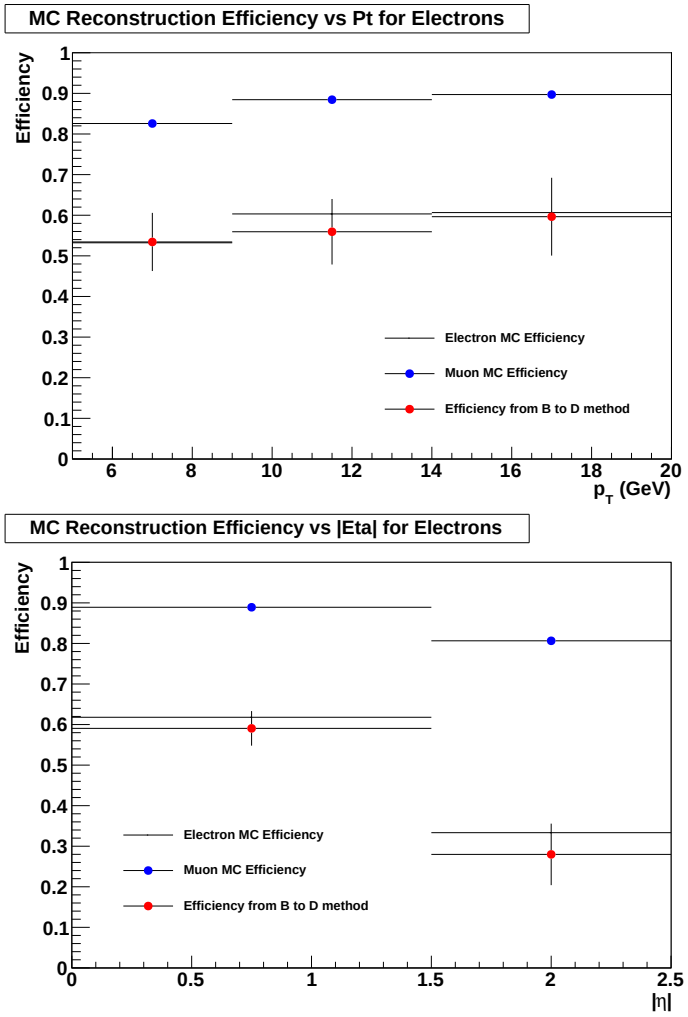

Figure 5.8: The muon (blue points) and electron reconstruction efficiency (both estimated (red points) and actual (black points)), in the $t\bar{t}$ simulation sample, as a function of lepton p_T (top) and $|\eta|$ (bottom).

Next, the muon reconstruction efficiency is measured in both data and simulation. To do this, the tag and probe method is used, which pairs a loosely selected object (probe) with a tightly selected muon (tag). The tag/probe pair is required to form a mass resonance to ensure that the probe is a muon. A set of selection cuts are then

applied to the probe, and the selection efficiency is calculated using

$$\epsilon_{TP} = \frac{\text{Number of probes passing selection}}{\text{Total number of probes}} \tag{5.8}$$

The official CMS tag and probe package [47] is used. The efficiency is measured in three components:

- The muon track reconstruction efficiency (standalone muon is the probe).
- The track-to-GlobalMuon reconstruction efficiency.
- The muon ID efficiency.

For each component, the J/ψ resonance ($m_{J/\psi} = 3.096$ GeV [2]) is used. To count the number of events in resonance against those that are background, RooFit [45] is used to perform the unbinned maximum likelihood fits. The signal resonance is modeled using the single-sided Crystal Ball function, which has the functional form

$$\begin{aligned} f(x;\mu,\sigma,\alpha,n) &= N \begin{cases} \frac{A(\alpha,n)}{(B(\alpha,n)-u(x,\mu,\sigma))^n}, & u(x,\mu,\sigma) < -\alpha, \\ \exp\left(-\frac{u^2(x,\mu,\sigma)}{2}\right), & u(x,\mu,\sigma) \geq -\alpha, \end{cases} \\ A(\alpha,n) &= \left(\frac{n}{\alpha}\right)^n \exp\left(-\frac{\alpha^2}{2}\right), \\ B(\alpha,n) &= \frac{n}{\alpha} - \alpha, \\ u(x,\mu,\sigma) &= \frac{x-\mu}{\sigma}. \end{aligned} \tag{5.9}$$

The background is fit with a second order Chebyshev polynomial:

$$f(x;c_0,c_1,c_2) = 1 + c_0 + c_1 x + c_2(2x^2 - 1). \tag{5.10}$$

The Crystal Ball function is useful for describing peaks affected by energy loss in the

reconstruction [48], while Chebyshev polynomials are useful in modeling continuous background distributions [45].The fits are performed in probe p_T bins equivalent to those used for the B mass fits. Figures 5.9 - 5.19 show the resulting fits and efficiencies for each of the three efficiency components for data and simulation. Figure 5.20 shows the total efficiency, which is the product of the efficiency components, and the data/simulation scale factor. The scale factor is fit with a constant term, with the result being 0.9974 ± 0.0283. This is used to correct the ratio calculation in simulation to ensure that differences between data and simulation are due solely to the electron reconstruction.

Finally to find the data/simulation factor for electrons, the electron/muon efficiency ratio is calculated by using the two fractions in equation (5.5) and moving the muon efficiency to the left hand side of the equation. The simulation calculation is corrected using the data/simulation scale factor previously found above for muons. The resulting ratios and scale factor, fitted with a constant term, are shown in Figure 5.21. The electron data/mc scale factor is found to be 1.028 ± 0.081, which is very comparable to scale factors obtained for the other b-taggers discussed in Sections 4.3 and 4.4, which range from $0.87 - 1.01$ in value and $0.01 - 0.07$ in uncertainty [49]. This scale factor is used to correct the soft electron tagging efficiency found using the $t\bar{t}$ simulation sample.

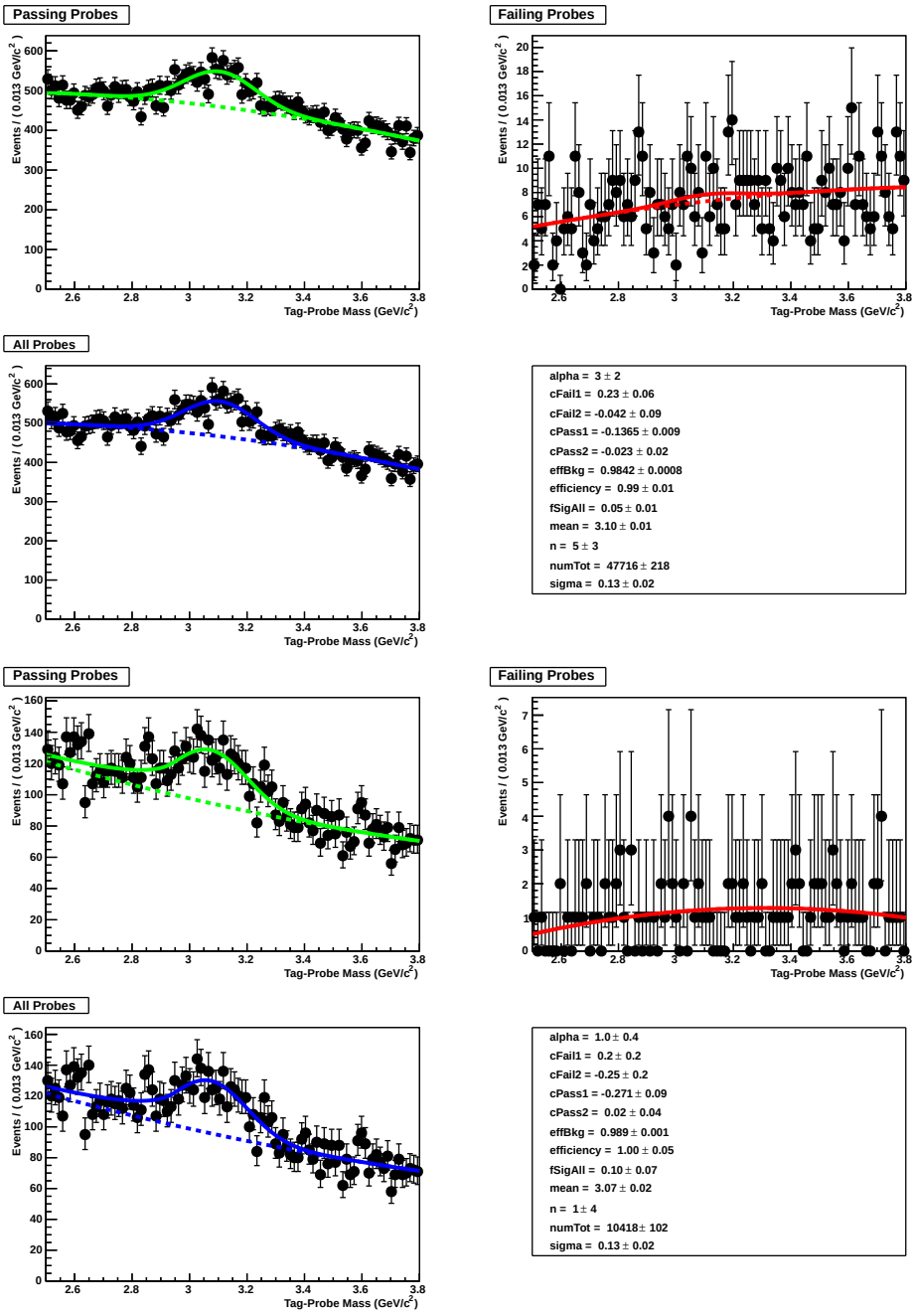

Figure 5.9: Resulting fits and efficiencies for the muon tracking efficiency in data (top) and simulation (bottom) for the probe p_T of [5.0, 9.0] GeV. Solid lines show the total fit, while dashed lines show the background contribution.

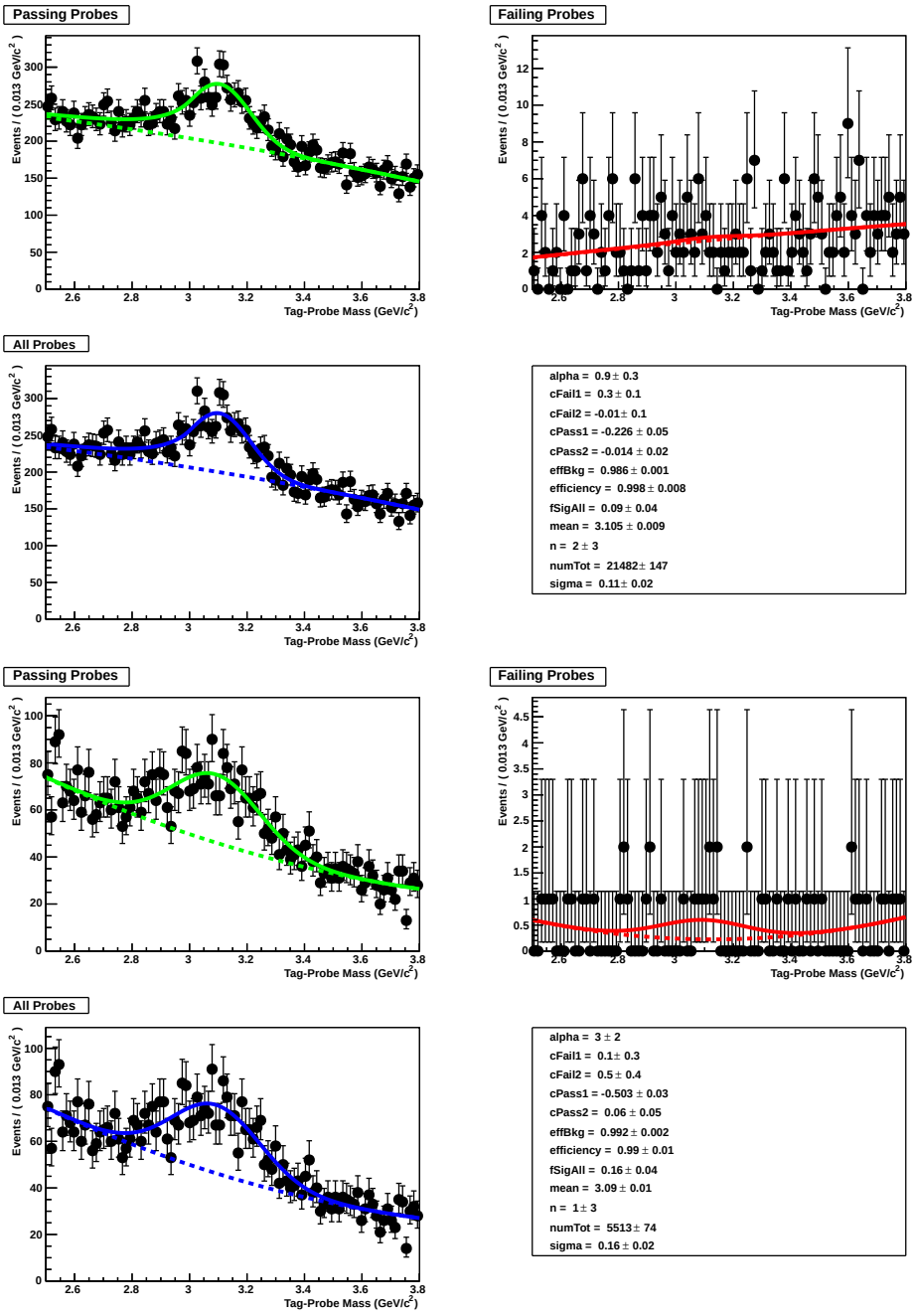

Figure 5.10: Resulting fits and efficiencies for the muon tracking efficiency in data (top) and (simulation) for the probe p_T of [9.0, 14.0] GeV. Solid lines show the total fit, while dashed lines show the background contribution.

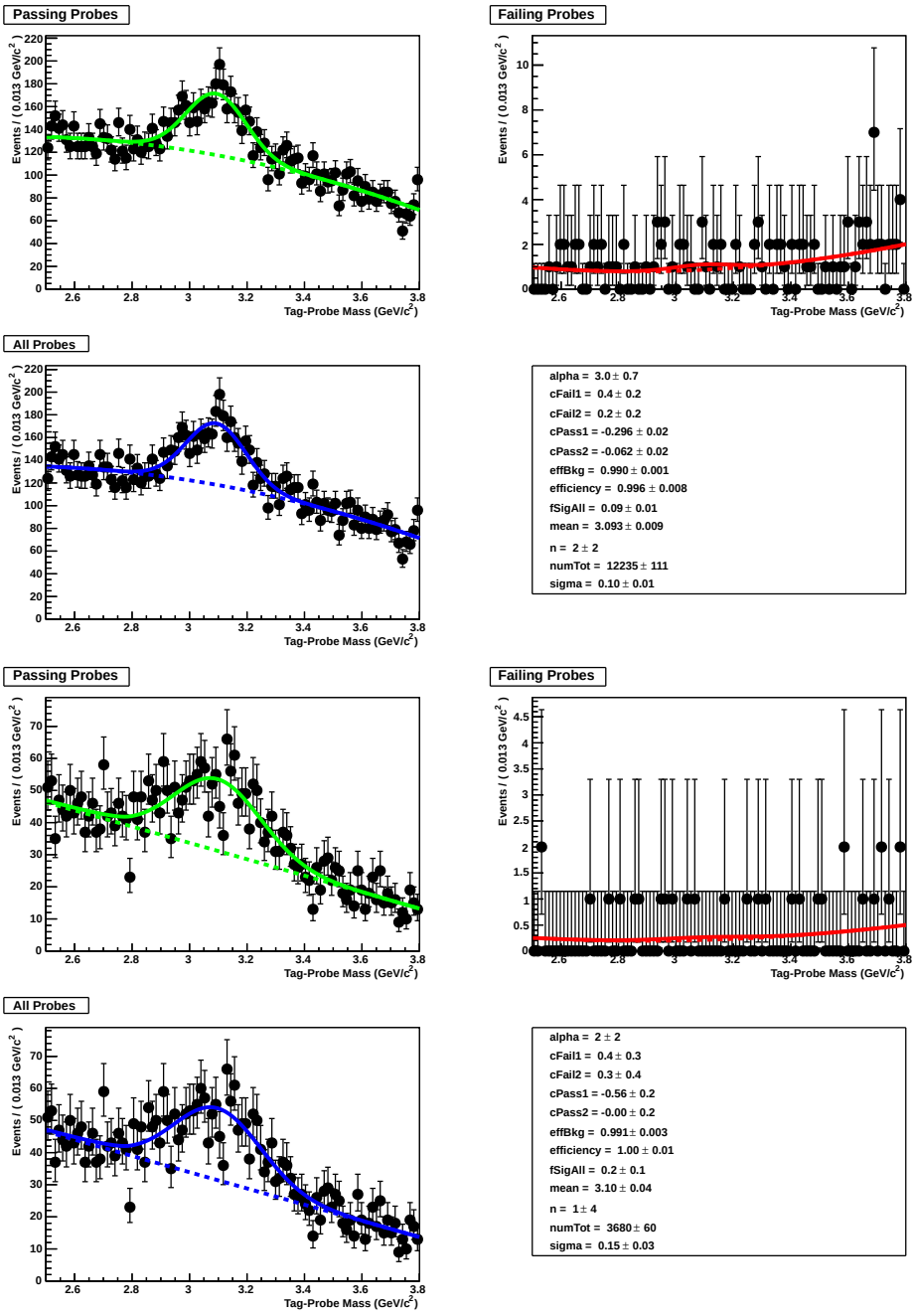

Figure 5.11: Resulting fits and efficiencies for the muon tracking efficiency in data (top) and simulation (bottom) for the probe p_T of [14.0, 20.0] GeV. Solid lines show the total fit, while dashed lines show the background contribution.

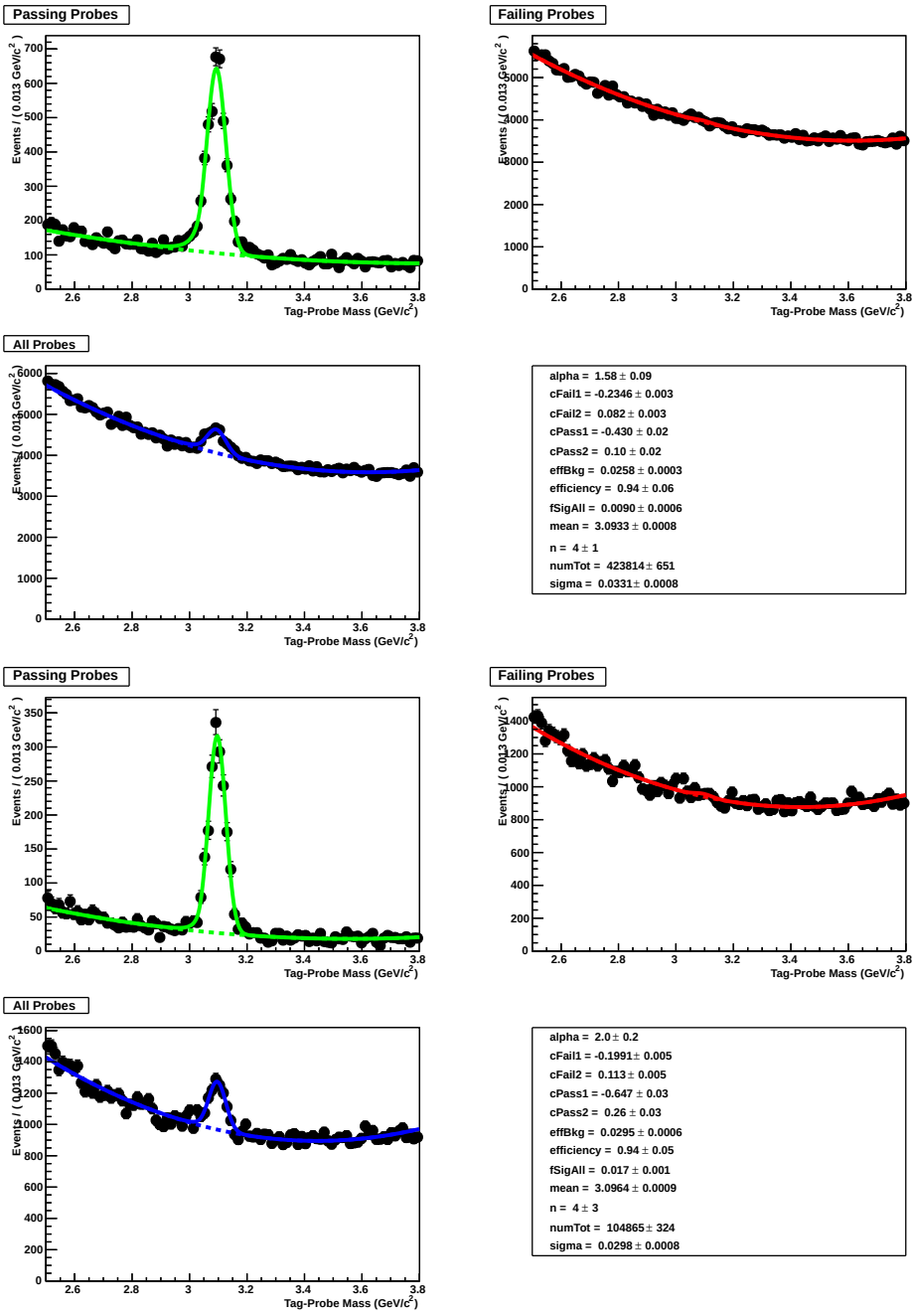

Figure 5.12: Resulting fits and efficiencies for the track-to-GlobalMuon efficiency in data (top) and simulation (bottom) for the probe p_T of [5.0, 9.0] GeV. Solid lines show the total fit, while dashed lines show the background contribution.

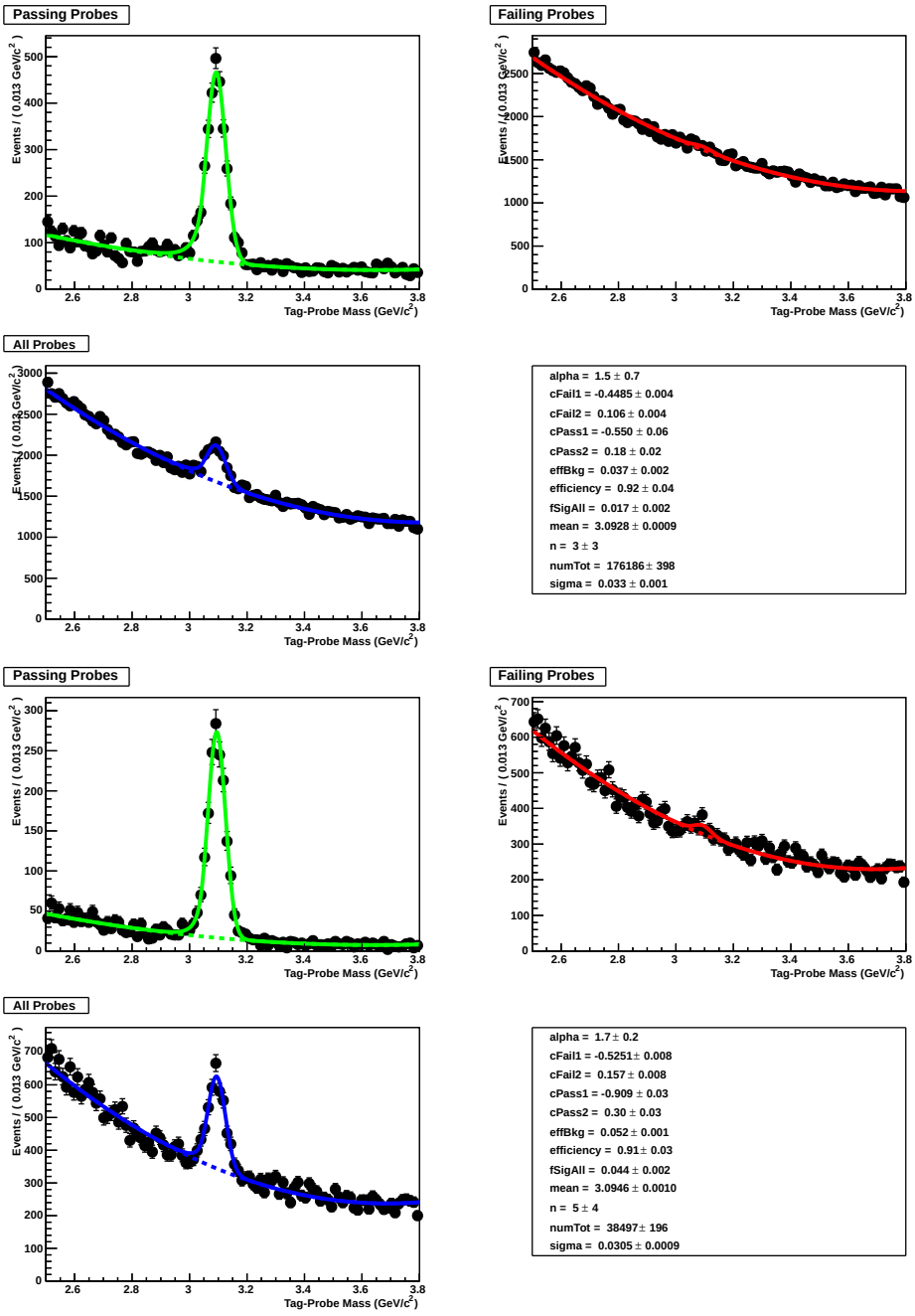

Figure 5.13: Resulting fits and efficiencies for the track-to-GlobalMuon efficiency in data (top) and simulation (bottom) for the probe p_T of [9.0, 14.0] GeV. Solid lines show the total fit, while dashed lines show the background contribution.

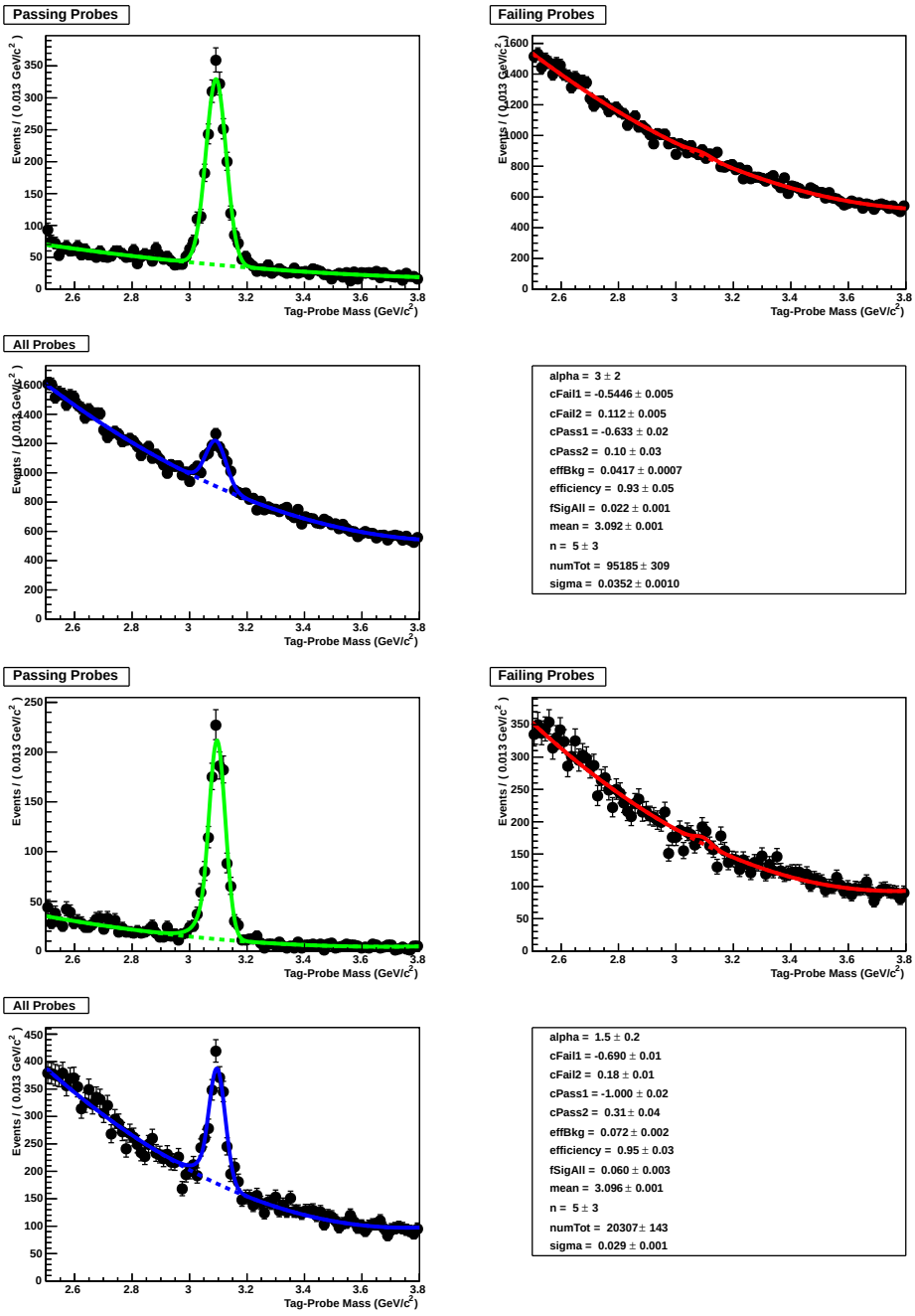

Figure 5.14: Resulting fits and efficiencies for the track-to-GlobalMuon efficiency in data (top) and simulation (bottom) for the probe p_T of [14.0, 20.0] GeV. Solid lines show the total fit, while dashed lines show the background contribution.

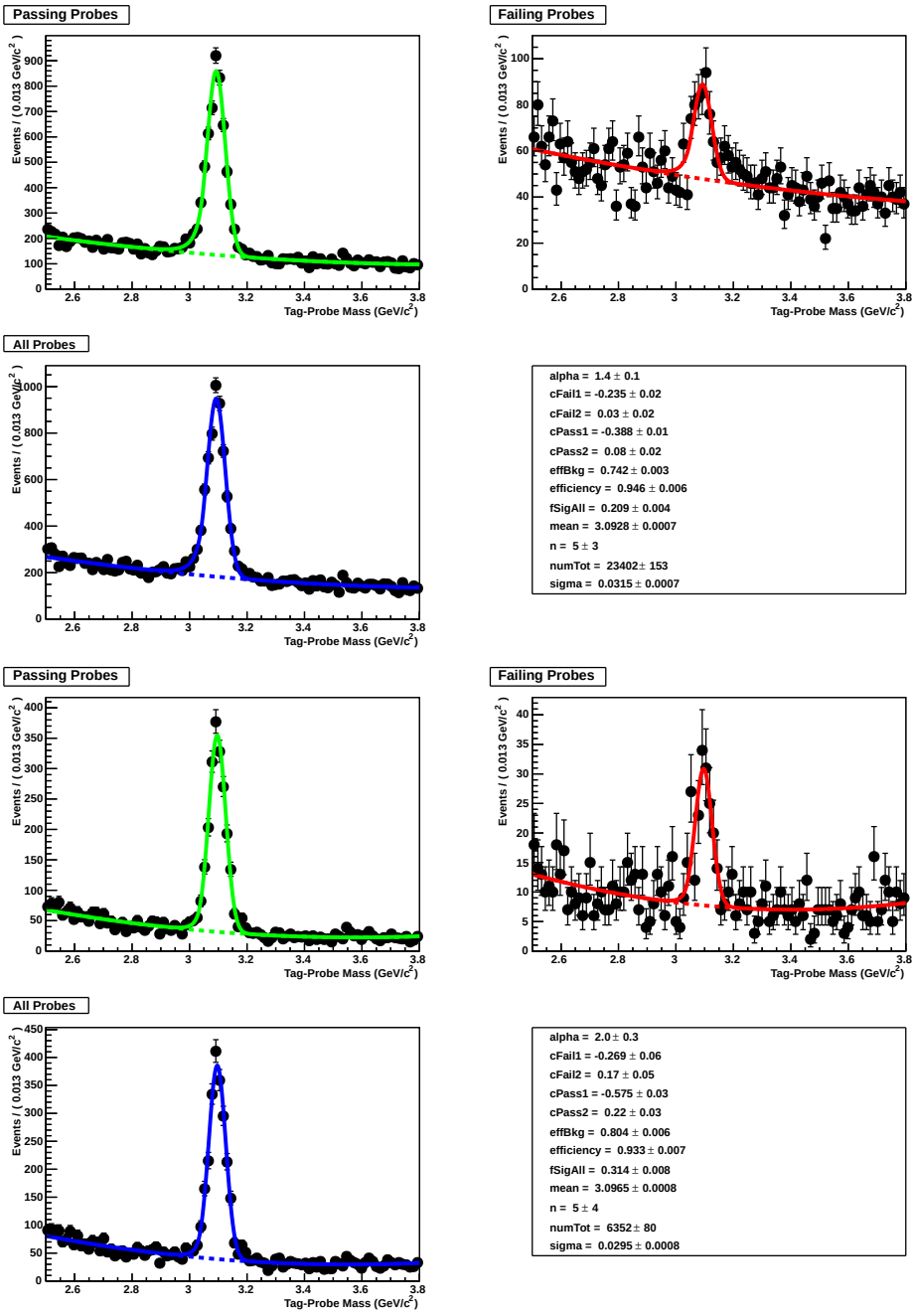

Figure 5.15: Resulting fits and efficiencies for the muon ID efficiency in data (top) and simulation (bottom) for the probe p_T of [5.0, 9.0] GeV. Solid lines show the total fit, while dashed lines show the background contribution.

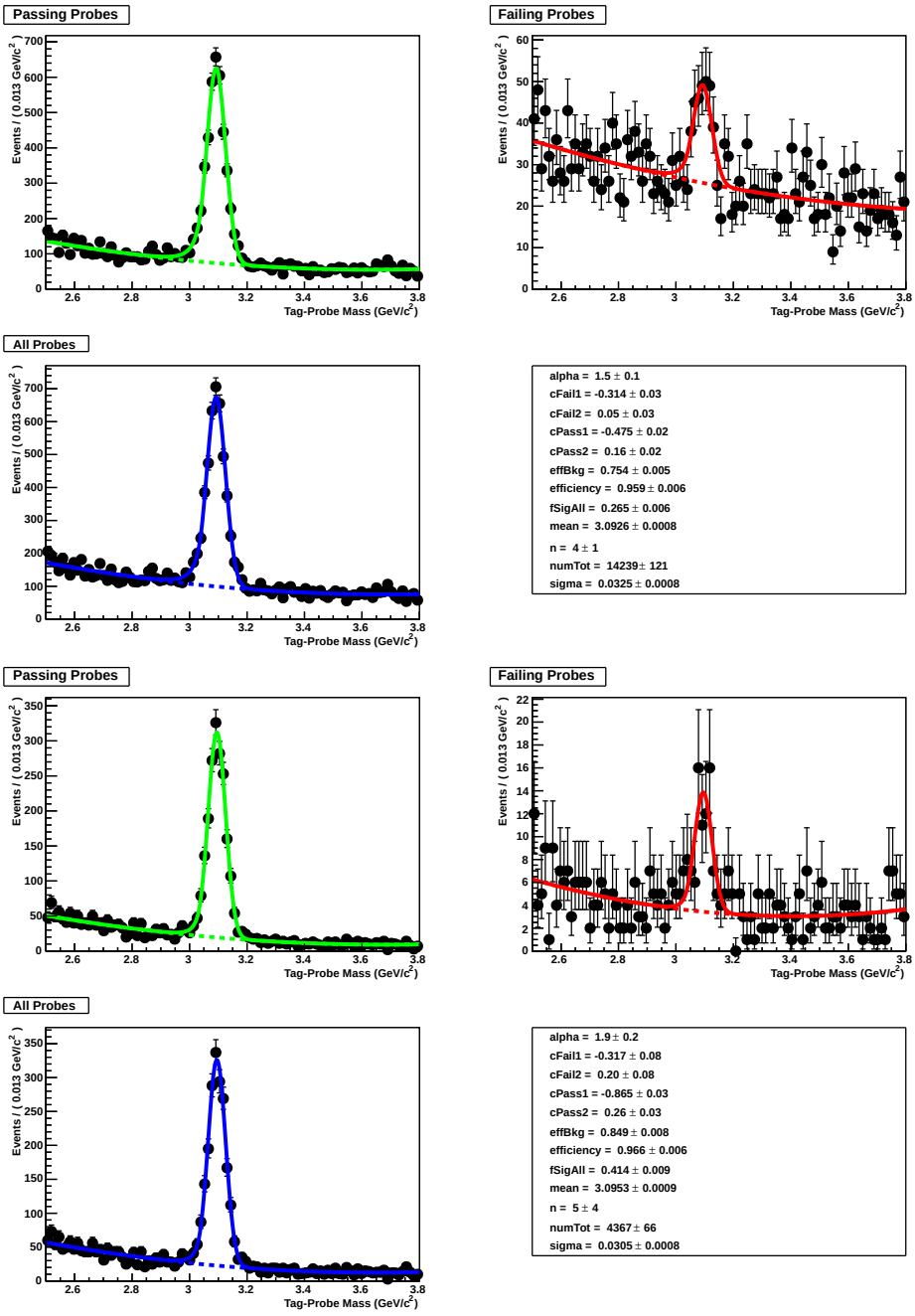

Figure 5.16: Resulting fits and efficiencies for the muon ID efficiency in data (top) and simulation (bottom) for the probe p_T of [9.0, 14.0] GeV. Solid lines show the total fit, while dashed lines show the background contribution.

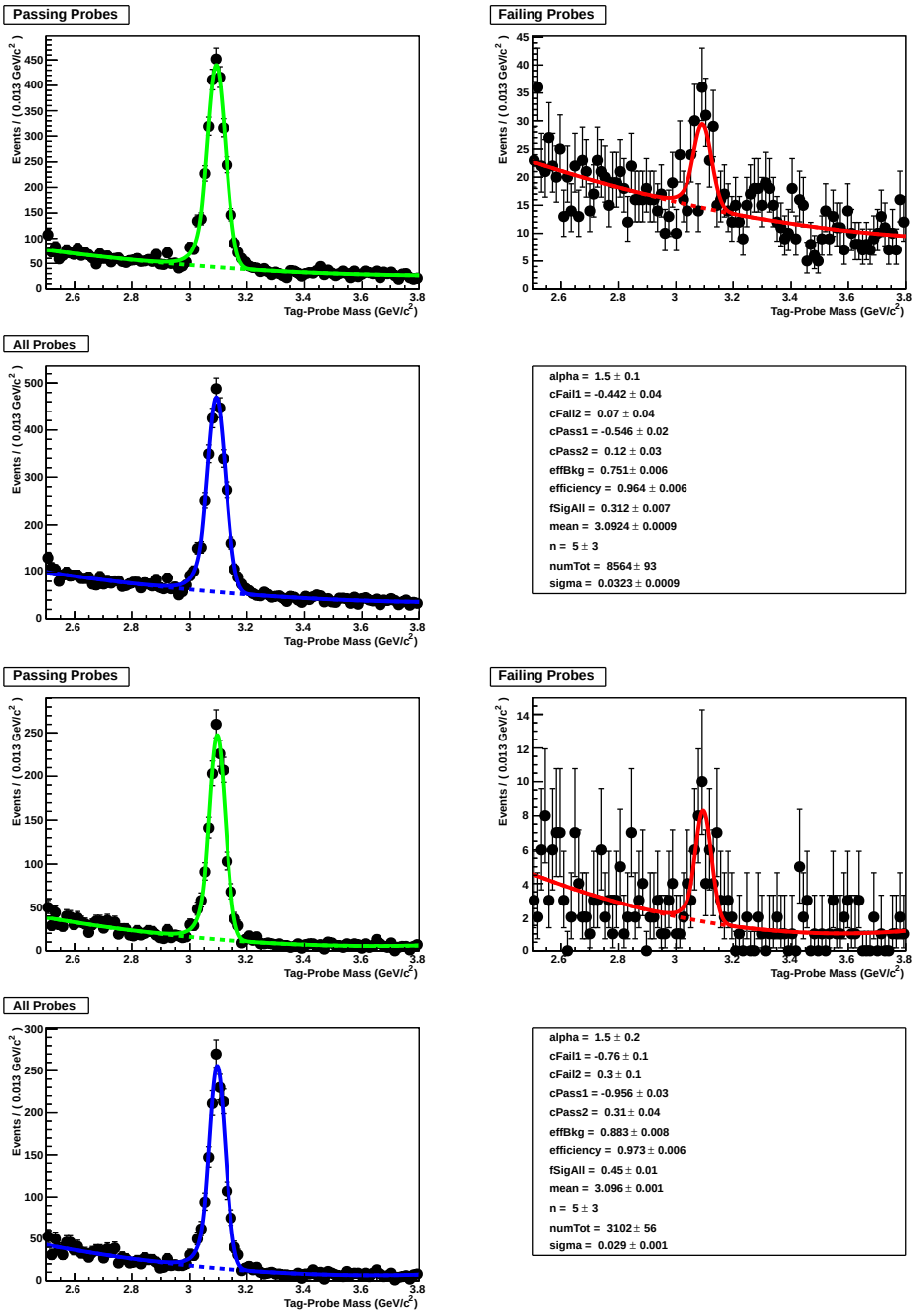

Figure 5.17: Resulting fits and efficiencies for the muon ID efficiency in data (top) and simulation (bottom) for the probe p_T of [14.0, 20.0] GeV. Solid lines show the total fit, while dashed lines show the background contribution.

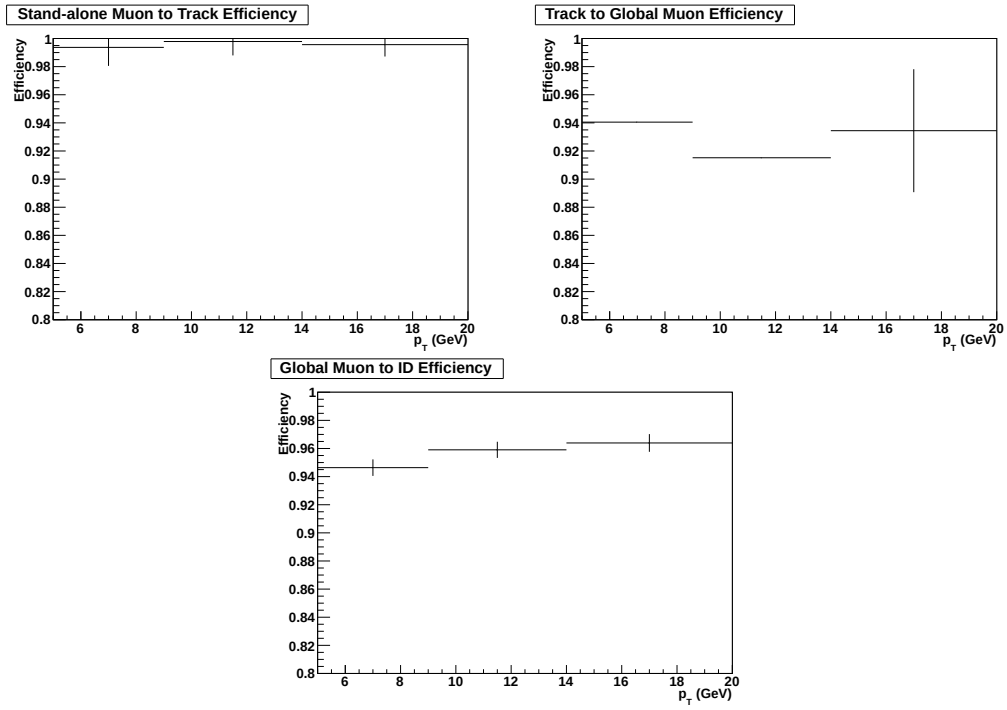

Figure 5.18: The efficiency results in data for the muon tracking efficiency (top left), track-to-Global Muon efficiency (top right), and Global muon ID efficiency (bottom)

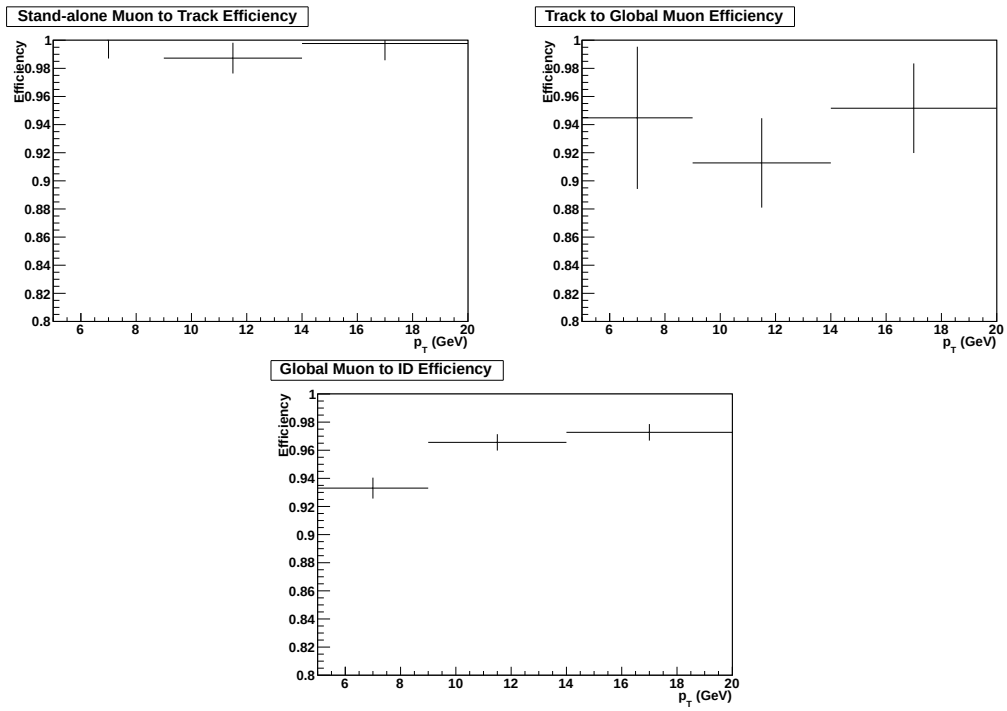

Figure 5.19: The efficiency results in simulation for the muon tracking efficiency (top left), track-to-Global Muon efficiency (top right), and Global muon ID efficiency (bottom)

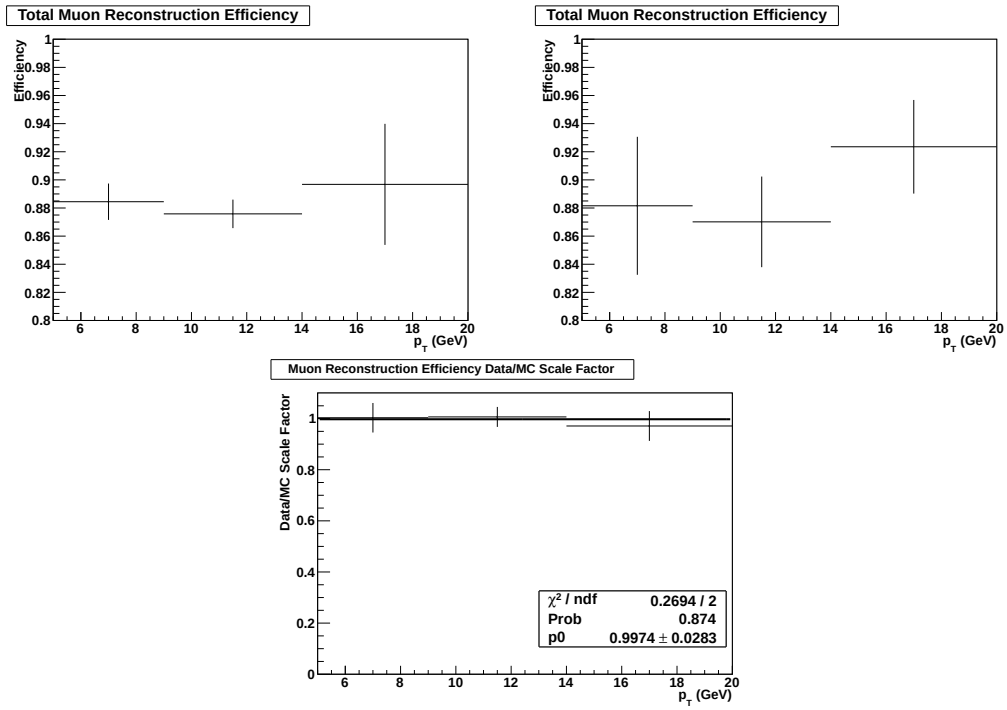

Figure 5.20: The total muon reconstruction and ID efficiency for data (top left) and simulation (top right), and the resulting data/simulation scale factor with constant fit (bottom).

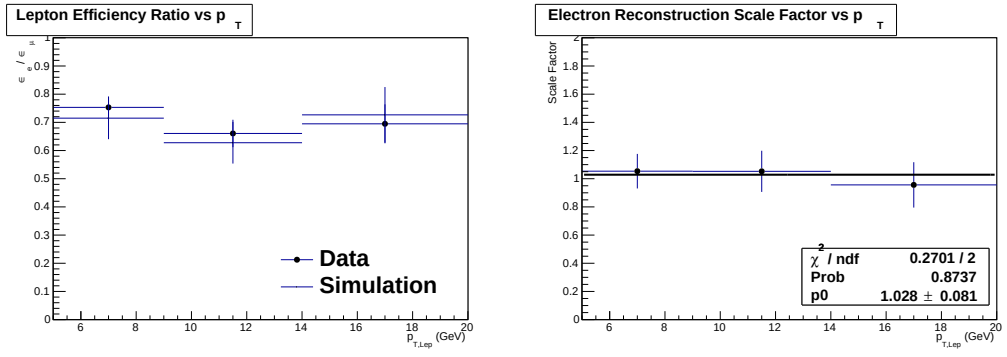

Figure 5.21: The lepton efficiency ratios for both data and simulation (left) and the resulting scale factor, fitted with a constant term (right).

Chapter 6

Cross Section

In particle physics, the cross section is the probability for a given interaction to occur. The number of events produced for a process is the cross section times the luminosity (Equation (2.1)). Cross sections are useful in determining the relative rates of interactions and for tests of the Standard Model, since any significant disagreement from the theoretical Standard Model cross section points to new physics sources, or can point to new physics by providing a measurement of couplings to other particles. Thus, the measurement of cross sections is important in testing the Standard Model. Cross section measurements are also necessary for determining the background contribution of a process to new physics sources. This chapter will outline the components necessary for measuring a cross section, how these components are determined, and the resulting systematic uncertainties.

6.1 Equation and Inputs to Cross Section

The equation to empirically calculate the cross section for a given process is

$$\sigma = \frac{N_{Total} - N_{Bkg}}{A * \epsilon * L_{Int}}. \tag{6.1}$$

N_{Total} is the total number of events which survive selection cuts, and will be the total of signal and background events. N_{Bkg} is the estimated number of background events which survive selection cuts. A is the acceptance, which is the fraction of signal events that one expects to observe, given the geometric and kinematic constraints of the detector. Acceptance is calculated completely from simulation. The quantity ϵ is the selection efficiency of those events which fall within the acceptance. Finally, L_{Int} is the integrated luminosity over which the N_{Total} number of events was collected.

6.2 Event Selection

This analysis reexamines the top quark pair production cross section in the muon + jets final state, with the novel use of soft electrons to tag bottom quark jets at CMS. Since events containing a pair of top quarks will contain two jets from bottom quarks in the final state, one of the final state jets is required to have an electron. This section will outline the event selection criteria and the resulting selection efficiency.

6.2.1 Selection Criteria

Each event must come from a centrally certified list of good runs and luminosity sections, in which all detector systems were running, and the CMS detector was operating correctly. In addition, the HLT_IsoMu24_eta2p1 trigger must be passed, which requires at least one isolated muon with $p_T > 24$ GeV and $|\eta| < 2.1$. The

event must also have a valid primary vertex with at least four degrees of freedom (a measure of the number of tracks compatible with having come from the primary vertex), and both a longitudinal distance less than 24 cm and a transverse distance less than 2 cm, both measured with respect to the detector origin. In addition, each event must pass the following filters to ensure it is of good quality.

- HCAL Noise filter to ensure the event is not triggered due to HCAL noise and that no object is reconstructed with this noise.
- A beam halo filter to ensure the particle contribution due to beam halo is not excessive.
- A filter which checks if the HCAL calibration laser was spuriously firing during the event.
- A filter which removes events due to beam scraping by ensuring 25% of reconstructed tracks are high purity tracks.

Requirements are also placed on the reconstructed objects within the event. There must be exactly one muon which passes the following criteria.

- It must be a muon which is reconstructed from the Global Muon method and is selected from the particle flow reconstruction (Section 3.3).
- It must have $p_T > 26$ GeV and $|\eta| < 2.1$.
- The normalized χ^2 for the global track must be less than ten.
- The global track must have at least one valid hit in the muon detector and have at least one matched station.

- The track must have hits in more than five silicon tracker layers and have at least one pixel detector hit.
- The d_0 of the track with respect to the primary vertex must be less than 0.02 cm.
- The d_z of the track with respect to the primary vertex must be less than 0.5 cm.
- The total p_T (corrected for pileup) of particles within $\Delta R < 0.4$ of the muon, divided by the muon p_T, must be less than 0.12.

To reduce the Drell-Yan background (Section 6.3.2), there must be no additional muon which satisfies the following requirements.

- A muon which is reconstructed using the Global Muon or Tracker Muon reconstruction methods and is selected from the particle flow reconstruction (Section 3.3).
- The muon has $p_T > 10$ GeV and $|\eta| < 2.5$.
- The total p_T (corrected for pileup) of particles within $\Delta R < 0.4$ of the muon, divided by the muon p_T, is less than 0.2.

To reduce contributions from various dilepton background sources, the event must not have an isolated electron which fulfills the given criteria.

- It is an electron which is reconstructed using the particle flow method (Section 3.4).
- The electron has $p_T > 20$ GeV and $|\eta| < 2.5$.

- The total p_T (corrected for pileup) of particles within $\Delta R < 0.3$ of the electron, divided by the muon p_T, is less than 0.2.

Jets in the event are reconstructed as specified in Section 3.5. Isolated muons and electrons are removed from the set of particles used to reconstruct jets, as well as particles from pileup sources. Additionally, jet energy corrections are applied to account for detector response as a function of both jet p_T and η. There must be at least three jets in the event which pass the following selection.

- Jet $p_T > 35$ GeV and $|\eta| < 2.5$.
- The jet must be composed of more than one particle and at least one of these particles must be charged, if the jet is within $|\eta| < 2.4$.
- The fraction of jet energy due to neutral hadrons must be < 0.99.
- The fraction of jet energy due to neutral EM particles must be < 0.99.
- The fraction of jet energy due to charged EM particles must be < 0.99, if the jet is within $|\eta| < 2.4$.
- The fraction of jet energy due to charged hadrons must be > 0, if the jet is within $|\eta| < 2.4$.

Finally, at least one of the jets selected with the above criteria must have a soft electron, selected as described in Section 5.1, within $\Delta R < 0.4$.

6.2.2 Acceptance and Selection Efficiency

The acceptance and selection efficiency are calculated using Monte Carlo truth information from the $t\bar{t}$ sample listed in Section 3.1. The selection efficiency is factorized

into the pre-tag and post-tag selection efficiencies. The acceptance times pre-tag selection efficiency is calculated by dividing the number of events which pass the pre-tag selection by the total number of events generated in the $t\bar{t}$ sample. This comes to $5.45 \pm 0.00858\%$. This value is further corrected by a data-to-simulation scale factor to account for differences in the muon reconstruction, id, and HLT selection efficiencies, which is 0.998 ± 0.0014 (stat) ± 0.056 (syst) [50]. The post-tag selection efficiency is found in simulation by taking the number of $t\bar{t}$ events passing the soft electron tagging cut and dividing by the number of events which pass the pre-tag selection. This number is corrected by the scale factor which is found in Section 5.2.

6.3 Background Estimation

6.3.1 W Boson + Jets Background

W boson + jet production (Fig. 6.1) is the largest background to top pair production due to its relatively large cross section and both final states contain a W boson decaying leptonically with jets. Due to the size of the background and the difficulty to estimate it from theory, it is preferable to have a data-driven manner to estimate the contribution of the background to the signal region, rather than having to rely on simulation.

Fortunately, it is possible to estimate the W + jet background due to how W bosons are produced in collision versus how they are produced in top decay. The LHC is a proton-proton collider, with a proton having the valence quark configuration of two up quarks and one down quark. Due to the asymmetry in the number of up quarks to down quarks in protons, collisions will produce more W^+ bosons than W^-, since $u\bar{d} \to W^+$, while $\bar{u}d \to W^-$ [51]. Top quark pair production creates W^+ and

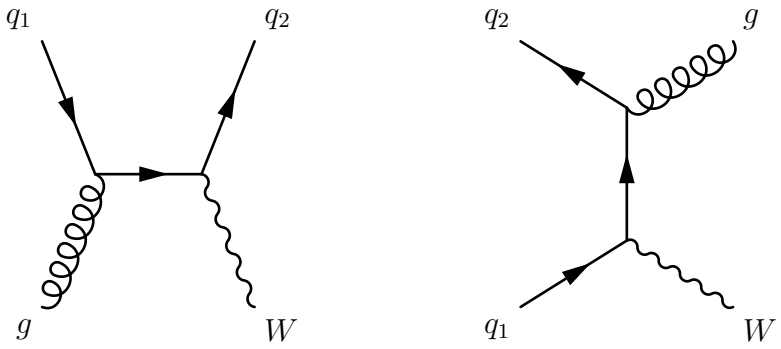

Figure 6.1: Example Feynman diagrams for W + Jet production.

W^- in equal numbers. Therefore, this W boson production asymmetry can be used to estimate the W + jets background.

The asymmetry can be calculated by counting the muons produced in the final state. W + jets will have more μ^+ than μ^-, while top pair events will produce μ^+ and μ^- in equal numbers. Muons from other background sources will also produce the positive and negative variety in equal numbers, hence the only source of the asymmetry is due to W production. Using the number of μ^+ and μ^- produced in W + jet events, the asymmetry is represented as

$$A_W = \frac{N_{\mu^+}^{\mathrm{W+Jets}} - N_{\mu^-}^{\mathrm{W+Jets}}}{N_{\mu^+}^{\mathrm{W+Jets}} + N_{\mu^-}^{\mathrm{W+Jets}}}. \tag{6.2}$$

Since W + jets is the only source of muon charge asymmetry, $N_{\mu^+}^{\mathrm{W+Jets}} - N_{\mu^-}^{\mathrm{W+Jets}} = N_{\mu^+}^{\mathrm{Total}} - N_{\mu^-}^{\mathrm{Total}}$, while the denominator $N_{\mu^+}^{\mathrm{W+Jets}} + N_{\mu^-}^{\mathrm{W+Jets}} = N_{\mathrm{Total}}^{\mathrm{W+Jets}}$. Thus, Equation (6.2) can be rewritten as

$$N_{\mathrm{Total}}^{\mathrm{W+Jets}} = \frac{N_{\mu^+}^{\mathrm{Total}} - N_{\mu^-}^{\mathrm{Total}}}{A_W}, \tag{6.3}$$

$$= R_W(N_{\mu^+}^{\mathrm{Total}} - N_{\mu^-}^{\mathrm{Total}}), \tag{6.4}$$

with $R_W = A_W^{-1}$. Estimating the number of W + jet events requires knowing the

total number of μ^+ and μ^- events in data and the W boson production asymmetry.

Lacking an experimental determination of the W boson production asymmetry, it may be estimated using a mix of simulation and data. If one assumes the ratio of asymmetries for different jet bins of W production is the same for data and simulation, the relationship

$$\frac{R_{MC}(N_{Jets}=i)}{R_{MC}(N_{Jets}=0)} = \frac{R_{Data}(N_{Jets}=i)}{R_{Data}(N_{Jets}=0)} \tag{6.5}$$

may be written. With some algebraic manipulation, this becomes

$$R_{Data}(N_{Jets}=i) = \frac{R_{MC}(N_{Jets}=i)}{R_{MC}(N_{Jets}=0)} R_{Data}(N_{Jets}=0). \tag{6.6}$$

Thus, determining the asymmetry in the higher jet bins in data requires finding the asymmetry in the 0−jet bin. To do this, contributions from other sources must be subtracted from the 0−jet bin in data using the methods described in the subsequent sections. The contribution from top pair production is estimated using simulation. Once the non-W production contributions are removed, R may be calculated for the 0−jet bin by taking the reciprocal of Equation (6.2) and assuming the remainder of events is completely W + jets. Equation (6.6) is then used to propagate the asymmetry estimate to the higher jet bins in data. These propagated asymmetry values are then used with Equation (6.4) and the difference in number of μ^+ and μ^- events in data to estimate the total number of W + jets events in the corresponding jet multiplicity bin. Figure 6.2 shows the value of R versus the number of jets in data. The trend in R versus the jet multiplicity is due to relative changes in the W + jet production subprocess contribution (quark-quark vs quark-gluon vs gluon-gluon interactions) for different jet bins [51]. Figure 6.3 shows how the pre-tag estimate in data compares to Monte Carlo truth.

To estimate the post-tagged W + jets contribution, the charge asymmetry is again

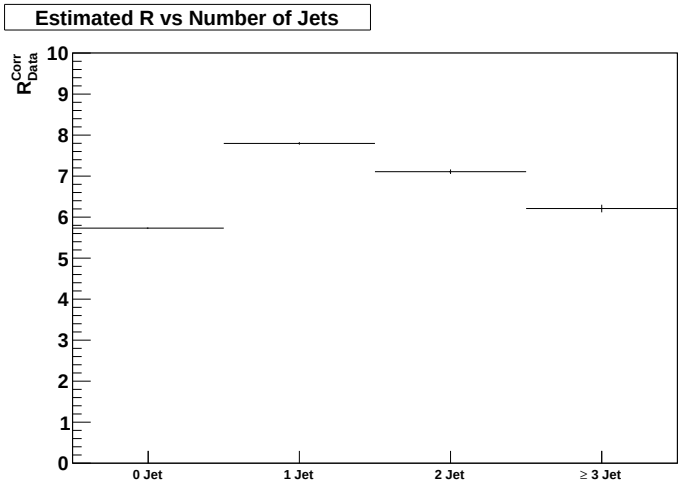

Figure 6.2: R versus jet multiplicity in data

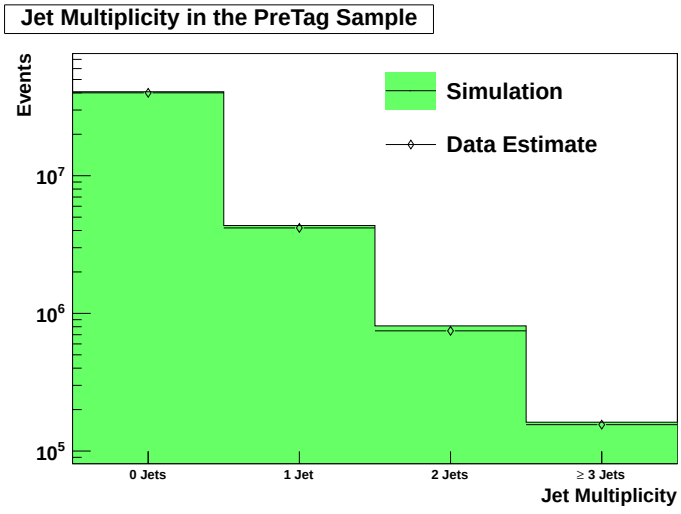

Figure 6.3: W + Jets pre-tag estimate vs Monte Carlo truth normalized to 11.6 fb^{-1} .

used with the assumption that the asymmetry does not affect the tag rate. Once the tag requirement has been applied to the dataset, the number of tagged events with μ^- is subtracted from the number of tagged events with μ^+. If the asymmetry does not affect the W + jets tag rate, then the remaining events after subtraction should be pure tagged W + jet events. The tag rate, τ, is then estimated using the equation

$$\tau = \frac{N_{\mu^+}^{post-tag} - N_{\mu^-}^{post-tag}}{N_{\mu^+}^{pre-tag} - N_{\mu^-}^{pre-tag}}. \tag{6.7}$$

This provides a purely data-driven method of estimating the W + jets post-tag estimate. The tag rate versus jet multiplicity is shown in Figure 6.4.

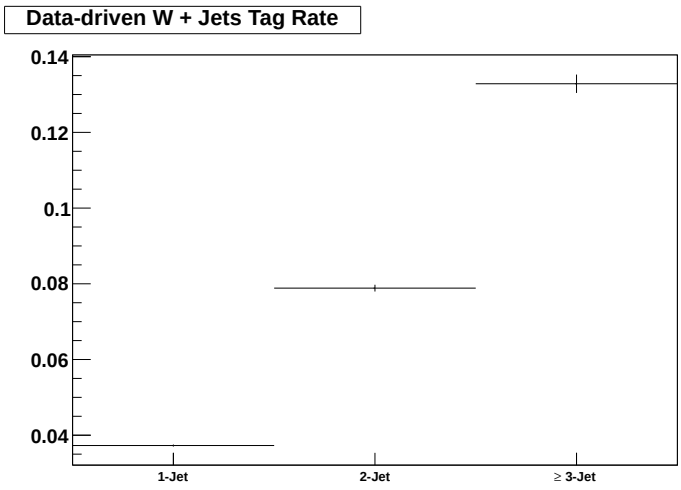

Figure 6.4: W + Jet tag rate versus jet multiplicity.

6.3.2 Drell-Yan Production + Jets Background

Drell-Yan + Jets (Fig. 6.5) are the class of events which have the hadronic production of dileptons through a real or virtual Z boson or a virtual photon and jets [6]. This is a potential background through both dimuon production, where one of the muons

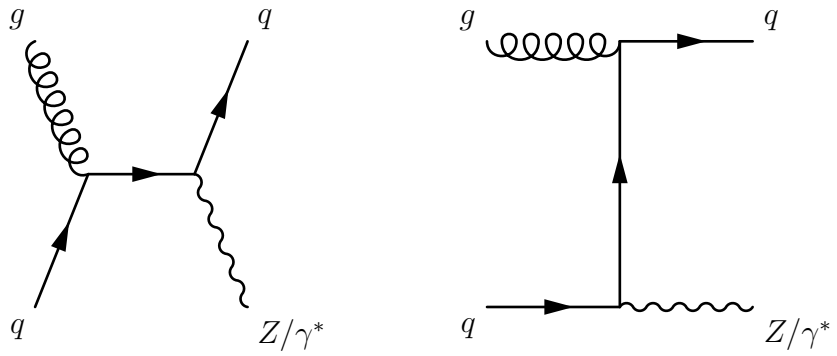

Figure 6.5: Example Feynman diagrams for Drell-Yan + Jets production.

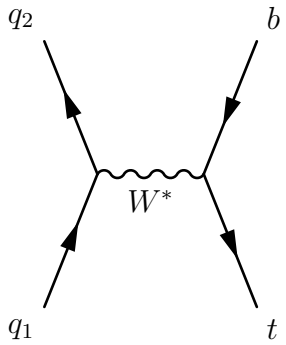

Figure 6.6: Example Feynman diagram for single top s-channel production.

is high p_T and highly isolated, and through ditau production, where the taus decay to a muon and an electron, which can fake the electron tag if it is near a jet. The contribution of this process to the background is estimated using simulation.

6.3.3 Single Top Background

Single Top production may occur at hadron colliders in three ways [52]:

- s-channel: the production of a time-like W boson which decays to a top-bottom quark pair (Fig. 6.6).
- t-channel: the fusing of a space-like W boson with a bottom quark to produce a single top quark (Fig. 6.7).

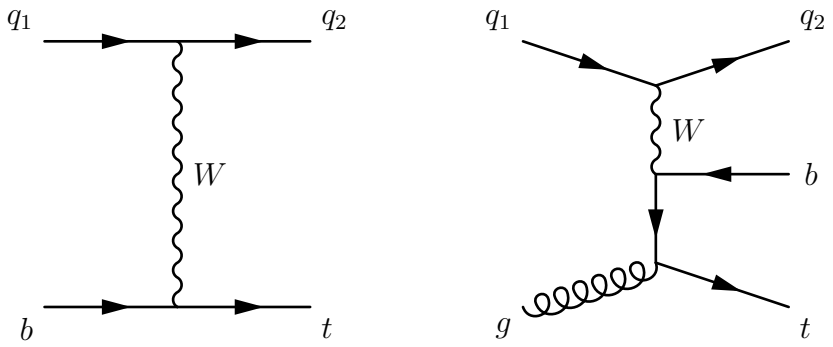

Figure 6.7: Example Feynman diagrams for single top t-channel production.

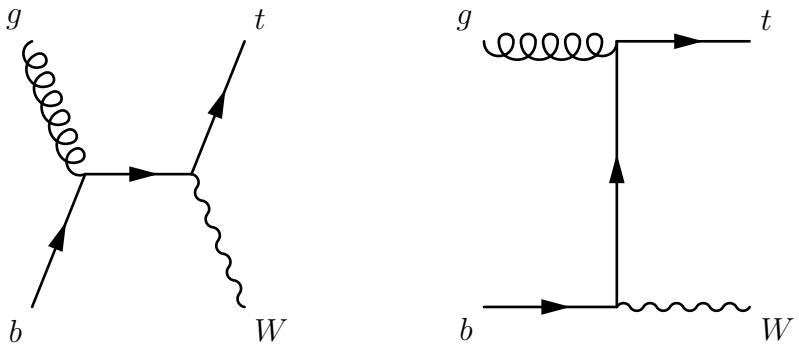

Figure 6.8: Example Feynman diagrams for single top tW-channel production.

- tW-channel: a bottom quark radiates a real W boson and produces a top quark as a result (Fig. 6.8).

Contributions from each channel are expected to be quite small compared to signal, and are estimated using simulation.

6.3.4 Diboson Background

Diboson background (Fig. 6.9) involves the production of a diboson (WW, WZ, or ZZ) pair with one of the bosons decaying leptonically and the other hadronically. The contribution of this background is expected to be quite small compared to signal, and is estimated using simulation.

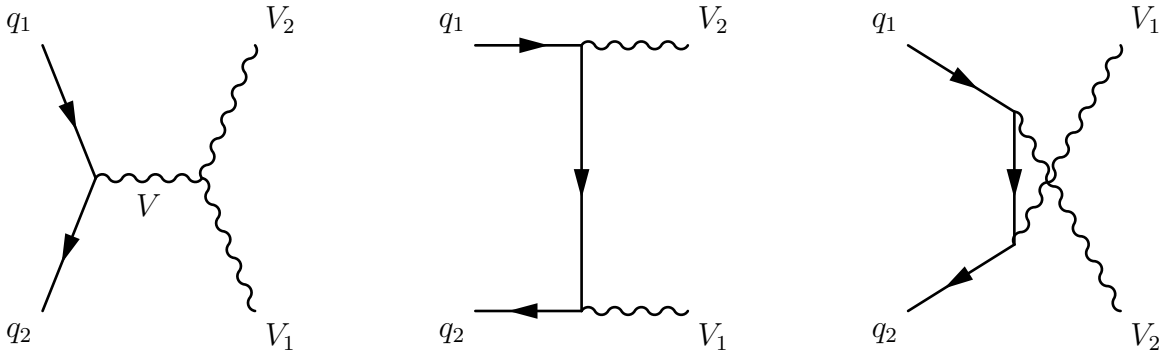

Figure 6.9: Leading order Feynman diagrams for diboson production ($V, V_{1,2} = W, Z, \gamma$).

6.3.5 QCD Multijet Background

QCD background is the multijet background which is not classified under the previously discussed categories and is typified by the production of jets through QCD processes. The required presence of a muon in the final state implies these events are due to heavy flavor jet production, with at least one jet producing a high p_T muon. Though this muon will typically be non-isolated, there will still be some contribution to the signal region that must be accounted for. Fortunately, the QCD background can be estimated in a completely data-driven manner.

The method used in this analysis is called the ABCD method. This method estimates the number of QCD events in the signal region by using event counts in QCD-dominated regions. The method requires creating a two-dimensional space using uncorrelated variables, v_1 and v_2, which characterize the signal and QCD background events in such a way to allow for the splitting of the (v_1, v_2) space into four distinct regions called signal region A and QCD-dominated regions B, C, and D. Ideally, regions B, C and D will contain purely QCD, while the signal and non-QCD background processes will reside only in region A (Fig. 6.10). The method works by assuming region C is a template for the behavior in region A, and that regions B and D provide

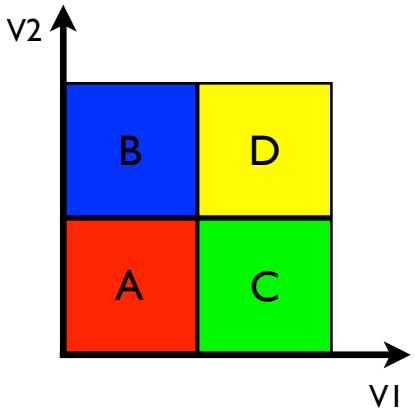

Figure 6.10: A representation of regions A, B, C, and D in $(v1, v2)$ space.

the appropriate scaling to go from C to A. If the variables v_1 and v_2 are uncorrelated for QCD events, it must be true that

$$\frac{N_A^{QCD}}{N_C^{QCD}} = \frac{N_B^{QCD}}{N_D^{QCD}} \text{ or } \frac{N_A^{QCD}}{N_B^{QCD}} = \frac{N_C^{QCD}}{N_D^{QCD}}. \tag{6.8}$$

In other words, the ratio of the number of events in adjacent regions should be equivalent when moving along the v_1 or v_2 axis. Combining these two notions gives the equation for estimating the number of QCD events in region A:

$$N_A^{QCD} = \frac{N_B}{N_D} N_C, \tag{6.9}$$

with the N_i representing the number of QCD events in regions A, B, C, and D. Since B, C, and D are assumed to be pure QCD, the QCD superscript is dropped. Thus, estimating the QCD contribution to region A simply requires counting the number of events in the non-signal regions.

This method can be modified slightly to allow for the calculation of systematic uncertainties. This is done by introducing additional, QCD-dominated regions, E

and F (Fig. 6.11). The relationship between regions E and F is treated in the same manner as regions A and C. Modifying equation (6.9) appropriately, the relation

$$1 = \frac{N_B N_F}{N_D N_E} \tag{6.10}$$

is obtained. Any deviation from 1 in calculating the result of equation (6.10) is treated as a systematic uncertainty. The subject of QCD background systematics is discussed again in Section 6.4.1.

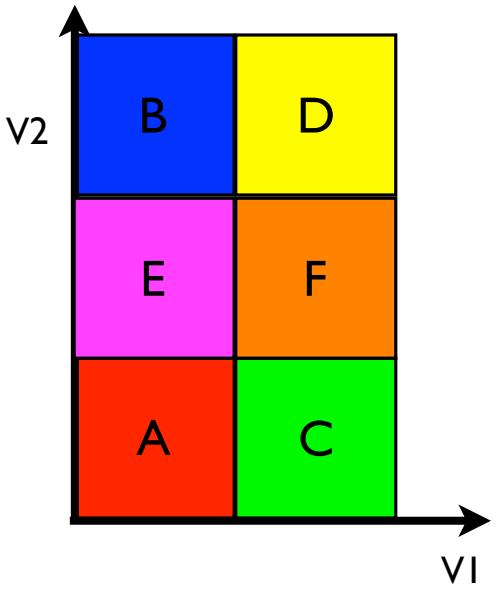

Figure 6.11: A representation of regions A, B, C, D, E, and F in $(v1, v2)$ space.

For this analysis, the relative isolation of the primary muon, defined as the total p_T (corrected for pileup) of particles within $\Delta R < 0.4$ of the muon divided by the muon p_T, and the primary muon $|d_0|$ with respect to the beamspot are used to create the two-dimensional space. The signal is expected to have very isolated muons close to the beamspot, while QCD may have non-isolated muons or muons some distance away from the beamspot due to long-lived decays. Using the QCD simulation sample

mentioned in Section 3.1, the boundaries were chosen for the ABCD regions such that the value obtained using equation (6.9) was closest to the true value of QCD events in the signal region (Fig. 6.12).

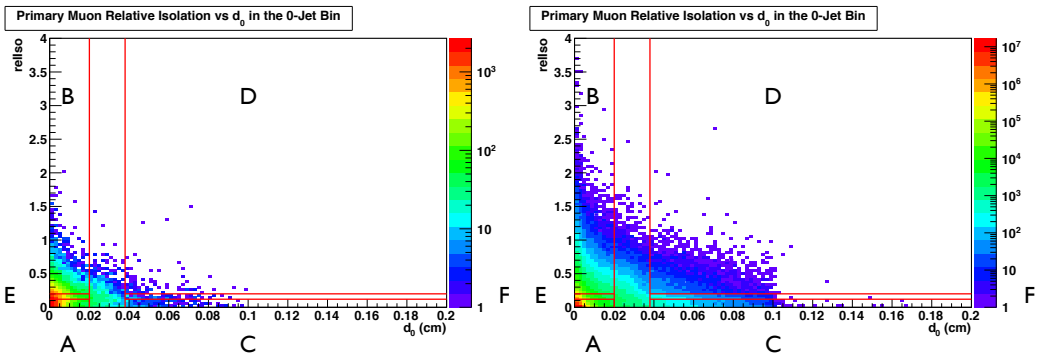

Figure 6.12: The ABCD regions, with systematic regions E and F, shown for the 0-Jet bin for both simulation (left) and data (right).

Estimates for the QCD background in the signal region are shown in Table 6.1 for the datasets listed in Section 3.1 with 11.6 fb^{-1} in the pre-tag case. For comparison, a similar table is shown using the simulation sample in Table 6.2.

Table 6.1: QCD background estimate for the pre-tag case in the signal region for 11.6 fb^{-1} of data. Numbers listed under $N_{A,\mathrm{MC}}$ are taken from the QCD simulation sample and scaled to the data luminosity. Percentage errors listed are statistical uncertainties. Uncertainties for $N_{B,C,D}$ are Poisson uncertainties (not shown).

Bin	N_B	N_C	N_D	$N_{A,\mathrm{Calc}}$	$N_{A,\mathrm{MC}}$
0-Jet	719052	20684	6882	$2161130 \pm 1.4\%$	$1253340 \pm 1.1\%$
1-Jet	1456420	22774	51761	$640800 \pm 1.0\%$	$486983 \pm 0.7\%$
2-Jet	918251	4991	68215	$67185 \pm 1.5\%$	$70389 \pm 1.5\%$
$\geq$ 3-Jet	167521	888	13169	$11296 \pm 3.5\%$	$10762 \pm 3.4\%$

To obtain the QCD estimate in the post-tag case, region C is used to estimate the tag rate in region A by dividing the number of tagged events by the total number of events in region C. The number of QCD events in the post-tag signal region is taken

Table 6.2: QCD background estimate for the pre-tag case in the QCD simulation sample. Numbers listed under $N_{A,\text{True}}$ are the actual number of QCD events in the signal region. Percentage errors listed are statistical uncertainties. Uncertainties for $N_{B,C,D}$ are Poisson uncertainties (not shown).

Bin	N_B	N_C	N_D	$N_{A,\text{Calc}}$	$N_{A,\text{True}}$
0-Jet	9196	350	187	$17236 \pm 9.1\%$	17254
1-Jet	27357	344	1405	$6697 \pm 6\%$	8159
2-Jet	20502	76	1617	$968 \pm 12\%$	1027
$\geq$ 3-Jet	3382	12	282	$148 \pm 29\%$	162

to be the product of the number of pre-tag QCD events in region A and the tag rate in region C. Table 6.3 shows the tag rate in region C and the estimated number of QCD events in the signal region in data, compared with the actual number of tags in QCD simulation.

Table 6.3: The post-tag QCD estimate, compared with the true number of tags in QCD simulation. Percent errors given are the statistical errors.

Bin	Region C Tag Rate	$N_{A,\text{Calc}}$	$N_{A,\text{MC}}$
1-Jet	0.043 ± 0.001	$27554 \pm 3\%$	$22496 \pm 0.6\%$
2-Jet	0.083 ± 0.004	$5576 \pm 5\%$	$5599 \pm 1.3\%$
$\geq$ 3-Jet	0.15 ± 0.01	$1694 \pm 8\%$	$1600 \pm 2.5\%$

6.4 Systematic Uncertainties

Systematic uncertainties are distinct from statistical uncertainties, which occur due to the random nature of the process being measured. Systematic uncertainties are due to biases in how a quantity is measured. These biases can be introduced through the use of calculated quantities which carry some uncertainty themselves, assumptions that are made in making calculations, uncertainties in simulation models used, among

other sources. This section describes the sources of systematic uncertainty in this analysis and how their effect is estimated.

6.4.1 Background Systematics

To estimate the pre-tag systematic uncertainty for the W + Jets estimate, Equation (6.6) can be rewritten as

$$\Delta N_{+,-}^{\mathrm{Total,Data}}(N_{Jets}=i) = \frac{\Delta N_{+,-}^{\mathrm{Total,MC}}(N_{Jets}=i)}{\Delta N_{+,-}^{\mathrm{Total,MC}}(N_{Jets}=0)} \Delta N_{+,-}^{\mathrm{Total,Data}}(N_{Jets}=0), \quad (6.11)$$

with $\Delta N_{+,-}^{\mathrm{Total}} = N_{\mu^+}^{\mathrm{Total}} - N_{\mu^-}^{\mathrm{Total}}$. The calculated value for $\Delta N_{+,-}^{\mathrm{Total}}$ in data can be compared to the actual value, since the only source of muon charge asymmetry in data should be due to W + Jets production. The largest difference between the calculated and actual values of $\Delta N_{+,-}^{\mathrm{Total}}$ is taken as a systematic uncertainty, which is 2.2 %. Figure 6.13 shows the comparison between the calculated and actual values of $\Delta N_{+,-}^{\mathrm{Total}}$. Systematic uncertainties due to QCD and other background estimates in the 0-jet bin must also be included in the pre-tag uncertainty.

To obtain the post-tag W + Jet systematic uncertainty, the method for determining the tag rate in data, discussed in Section 6.3.1, is applied in simulation and compared to the true tag rate. The largest difference between the estimated and Monte Carlo truth tag rate is taken as a systematic uncertainty, which is 6.7 %. Figure 6.14 shows the comparison between the calculated and Monte Carlo truth tag rate.

The systematic uncertainty for the pre-tag QCD estimate is calculated as described in Section 6.3.5. For the post-tag estimate, an additional systematic uncertainty is taken on the tag rate estimate in region C. The tag rates for regions E and F are calculated and compared and the difference is taken as a systematic uncertainty,

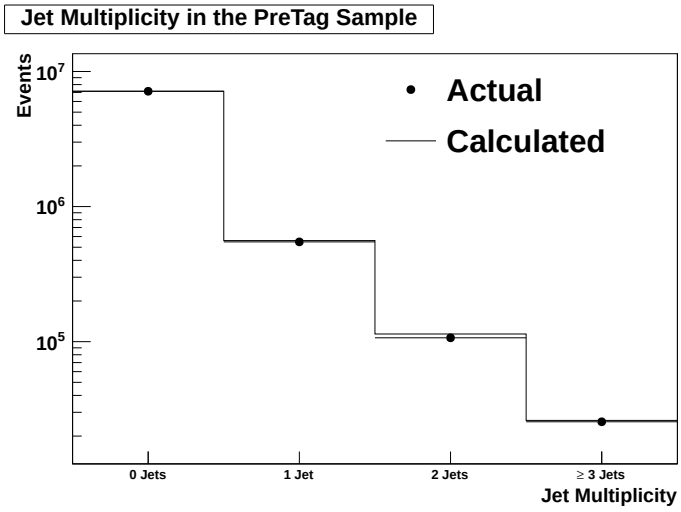

Figure 6.13: Comparison of calculated to actual $\Delta N_{+,-}^{\mathrm{Total}}$ in data.

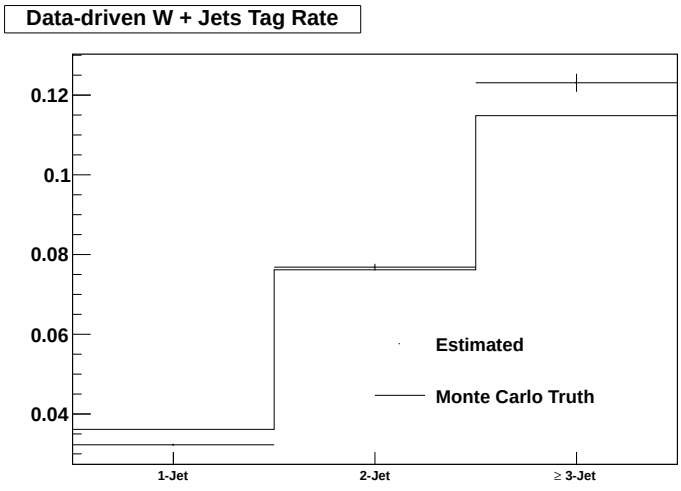

Figure 6.14: Comparison of the estimated and Monte Carlo truth tag rates.

assuming the same level of discrepancy for regions A and C. The total systematic uncertainty for the QCD estimate in the post-tag case is then calculated using propagation of errors. Table 6.4 lists all systematic uncertainties related to the QCD estimate. Though these uncertainties appear large, the effect on the cross section is small, due to the relatively small size of the background, as shown in Table 6.5.

Table 6.4: Systematic uncertainties in the QCD background estimate.

Bin	Pre-Tag Sys Unc.	Tag Rate Sys. Unc.	Post-Tag Sys. Unc.
0-Jet	60 %	0 %	60 %
1-Jet	48 %	21 %	52 %
2-Jet	38 %	4.5 %	38 %
$\geq$ 3-Jet	58 %	16 %	60 %

Uncertainties for the single top and diboson samples are estimated by scaling their respective normalization by $\pm 30\%$. Uncertainty in the Drell-Yan + Jets samples are estimated by using the Q^2 and matrix element to parton shower matching threshold samples listed in Section 3.1.

6.4.2 Luminosity

The uncertainty on the luminosity calculation is provided by the CMS Luminosity group, and is taken to be 4.5 % [50].

6.4.3 Soft Electron Tagging Efficiency

The uncertainty on the data-to-simulation scale factor used to corrected the post-tag selection efficiency is taken as a systematic uncertainty, which is 7.88 %. See Chapter 5.2 for details.

6.4.4 Event Generation

Event generation systematics cover uncertainties due to the Q^2 factorization scale, matrix element to parton shower matching threshold, and parton distribution function (pdf) uncertainties. To estimate the effects of the Q^2 scale and matching thresholds, special samples mentioned in Section 3.1 are used to determine the effect on the acceptance and selection efficiency.

The pdf uncertainties require individually varying the 22 input parameters of the CTEQ 6.6 pdfs used in the event generator up and down to determine their effect on the cross section [53]. This is done by determining 44 weights (2 per pdf parameter) to use for reweighing events used in the analysis. The effects of each weight (up or down) are added in quadrature to get the total pdf uncertainty.

6.4.5 Jet Systematics

Additional uncertainties must be applied for both the jet energy scale (JES) and resolution (JER). JES is the correction applied to jet energies to account for detector response as a function of p_T and η. The correction is thought to move the jet's measured energy to its "true" value. However, since the correction factors must be measured themselves, there is some uncertainty in these factors that must be accounted for in analyses which use them. To do this, the correction factor per jet is scaled up and down by the respective uncertainty. The resulting differences in the final cross section are taken as a systematic uncertainty.

JER accounts for differences in energy resolution between data and simulation. To give simulation the same energy resolution as data, jet energies in Monte Carlo are smeared using factors that depend on the jet $|\eta|$. To determine the JER uncertainty, the jets are resmeared by scaling the $|\eta|$-dependent factors up or down by their

respective uncertainties, and the cross section is recalculated. Any difference is taken as a systematic uncertainty.

6.4.6 Pileup

To account for uncertainty in the pileup reweighing of simulated events to match the pileup profile of data, the minimum bias cross section used to calculate the data profile is shifted up and down by ±5%. The effect of the up/down pileup reweightings on the $t\bar{t}$ calculated cross section is taken as a systematic uncertainty.

6.4.7 Total

The itemized list of systematic uncertainties and the total uncertainty on the $t\bar{t}$ cross section is listed in Table 6.5.

Table 6.5: Systematic uncertainties on the measured $t\bar{t}$ cross section.

Quantity	$\delta\sigma_{t\bar{t}}$ Uncertainty (pb)		$\delta\sigma_{t\bar{t}}$ Uncertainty (%)	
QCD Estimate	+4.3	-6.4	+1.9	-2.8
QCD Estimate (W + Jets)	+3.7	-3.4	+1.6	-1.5
W + Jets	+9.4	-9.2	+4.1	-4.0
DY + Jets	+4.4	-7.8	+1.9	-3.4
Other Bkg	±5.6		±2.4	
Tagging Efficiency	±18.0		±7.9	
Muon Scale Factor	±11.7		±5.1	
Q^2 Scale	+5.2	-6.5	+2.3	-2.9
ME-PS Matching	+2.5	-3.0	+1.1	-1.3
PDF	±0.94		±0.41	
JES	+9.6	-4.5	+4.2	-2.0
JER	+2.5	-0.8	+1.1	-0.4
Pileup	+1.2	-1.8	+0.5	-0.8
Total	+27.7	-27.8	+12.1	-12.2

6.5 Result

Table 6.6 shows the cut flow results for the data and simulation samples when all but the jet multiplicity and tagging cuts in Section 6.2.1 are applied. Data-driven estimates are not used in Table 6.6, since these rely on the preselection cuts to be applied. Table 6.7 shows the cut flow results for the pre-tag and post-tag datasets. Data-driven estimates are used where available. The agreement between data and simulation is quite good, and any discrepancy is covered by systematic uncertainties.

Figure 6.15 shows the jet multiplicity for the pre-tag and post-tag events after all selection cuts are applied. The QCD and W + Jet contributions represent the calculated estimates discussed in Sections 6.3.1 and 6.3.5. All other contributions are the Monte Carlo predicted contributions, when normalized using their respective cross sections and the total data luminosity. The hashed region in each bin reflects the total systematic uncertainty due to the background estimates. There is very good agreement in the signal-free bins, showing that the background estimation methods used are quite good at predicting their respective contributions to the total data. The signal bin shows clear evidence of $t\bar{t}$ events, as does Figure 5.1, which was made with the signal selection.

The measured $t\bar{t}$ cross section using soft electron tagging at $\sqrt{s} = 8$ TeV is

$$\sigma_{t\bar{t}} = 228.9 \pm 4.8\,(\text{stat})\,^{+27.7}_{-27.8}\,(\text{syst.}) \pm 10.1\,(\text{lumi})\,\text{pb},$$

which is in good agreement with the theoretical prediction given in Table 1.4 [14].

Table 6.6: The preselection cut flow table. All numbers where the physics process is specifically stated come from simulation samples, and are scaled using their theoretical cross-sections to match the data luminosity.

Cut	$t\bar{t}$	W + Jets	DY + Jets	QCD	Single Top	Diboson	Total MC	Data
Trigger	349062	53952829	8246847	14789538	101381	57074	77496731	79187558
Vertex	349060	53952669	8246835	14789538	101381	57074	77496557	79186496
Tight Muon	285506	45134704	4117287	1740036	86018	45829	52009380	53275159
Loose Muon Veto	276968	45134109	2590517	1732328	85139	43156	49862217	51334304
Loose Elec. Veto	250206	45129856	2577416	1732037	82381	39826	49811722	51280381

Table 6.7: The cut flow for the pre-tag and post-tag datasets. Estimates obtained from data-driven methods are denoted with (data). Estimates obtained from simulation are scaled using their theoretical cross-sections to match the data luminosity.

Cut	$t\bar{t}$	W + Jets (data)	DY + Jets	QCD (data)	Single Top	Diboson	Total MC	Data
0-Jet pre-tag	4370	40008747	1950595	2161130	11567	15532	44151941	44151941
1-Jet pre-tag	29831	4175461	490805	640800	33011	16530	5386438	5702891
2-jet pre-tag	74253	746700	78155	67185	26268	6577	999138	1073367
$\geq$ 3-jet pre-tag	144774	155445	18600	11296	12833	1735	344683	352182
1-Jet post-tag	2941	155669	11828	27554	3094	703	201789	221485
2-jet post-tag	12765	58628	5366	5576	3961	533	86829	96627
$\geq$ 3-jet post-tag	35317	20437	2090	1694	2735	224	62497	63611

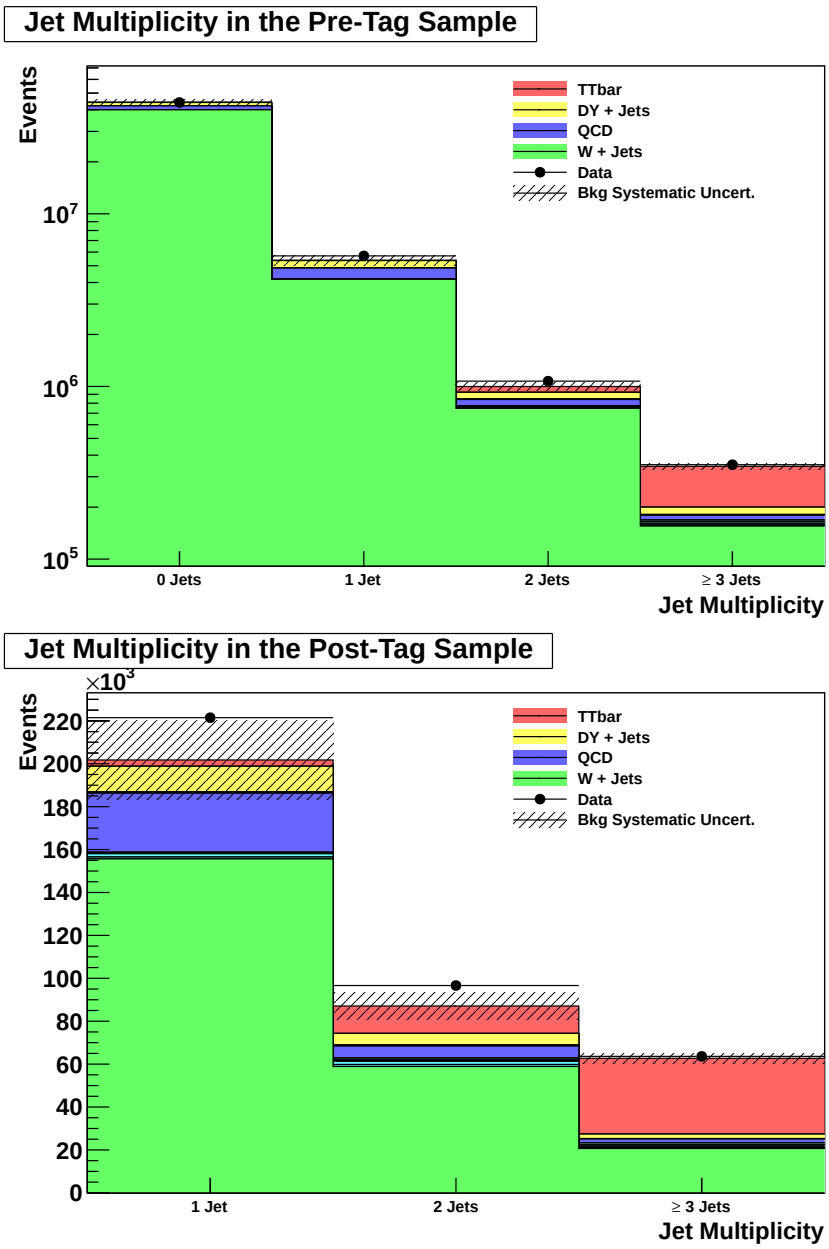

Figure 6.15: Jet Multiplicity for pre-tag (top) and post-tag (bottom) events after selection. Contributions are represented by calculated estimates where applicable. The hashed region is the total background estimate systematic uncertainty.

Chapter 7

Conclusion

This document describes the measurement of the $t\bar{t}$ production cross section at $\sqrt{s} = 8$ TeV at CMS using soft electron tagging. The measurement was performed in the muon + jet final state with 11.6 fb^{-1} of data. To estimate the W + Jets background, which is the most dominant background, the asymmetry in the production of W^+ over W^- at the LHC was used. The asymmetry method shows good agreement to simulation. The QCD background was estimated using the ABCD method. All other background estimates are obtained with Monte Carlo samples and theoretical cross sections. The measured cross section is in good agreement with two different theoretical predictions.

This analysis is the first instance in which soft electron tagging has been successfully used at CMS. Despite the many obstacles which were overcome, the resulting good agreement of the final measurement shows soft electron tagging has viability in future analyses. The demonstrated performance allows for the use of soft electrons to repeat analyses which used soft muons, such as the measurement of the top quark charge [41]. The use of both soft muons and soft electrons in an analysis would also provide increased statistics, and a built-in cross check of results, since tagging with

soft muons and soft electrons would provide two greatly orthogonal datasets. Most importantly, this analysis shows that soft electrons are an additional b-tagging tool, which can be used or included in other more robust b-taggers, such as multivariate taggers.

Bibliography

[1] Grunewald, M. W. Precision tests of the standard model URL http://arxiv.org/abs/hep-ex/0511018. hep-ex/0511018.

[2] Beringer, J. *et al.* The review of particle physics. *Phys. Rev. D86* **010001** (2012).

[3] Bethke, S. Experimental tests of asymptotic freedom. *Progress in Particle and Nuclear Physics* **58**, 351–386 (2007). URL http://www.sciencedirect.com/science/article/pii/S0146641006000615.

[4] Weinberg, S. A model of leptons. *Physical Review Letters* **19**, 1264–1266 (1967). URL http://link.aps.org/doi/10.1103/PhysRevLett.19.1264.

[5] Salam, A. & Ward, J. C. Electromagnetic and weak interactions. *Physics Letters* **13**, 168–171 (1964). URL http://www.sciencedirect.com/science/article/pii/0031916364907115.

[6] Halzen, F. & Martin, A. D. *Quarks & Leptons: An Introductory Course in Modern Particle Physics* (Wiley, 1984).

[7] Cabibbo, N. Unitary symmetry and leptonic decays. *Physical Review Letters* **10**, 531–533 (1963). URL http://link.aps.org/doi/10.1103/PhysRevLett.10.531.

[8] Kobayashi, M. & Maskawa, T. CP-violation in the renormalizable theory of weak interaction. *Progress of Theoretical Physics* **49**, 652–657 (1973). URL http://ptp.ipap.jp/link?PTP/49/652/.

[9] Higgs, P. W. Broken symmetries and the masses of gauge bosons. *Physical Review Letters* **13**, 508–509 (1964). URL http://link.aps.org/doi/10.1103/PhysRevLett.13.508.

[10] The CDF Collaboration. Observation of top quark production in $p\bar{p}$ collisions with the collider detector at fermilab. *Physical Review Letters* **74**, 2626–2631 (1995). URL http://link.aps.org/doi/10.1103/PhysRevLett.74.2626.

[11] The D0 Collaboration. Search for high mass top quark production in $p\bar{p}$ collisions at sqrt[s]= 1.8 tev. *Physical Review Letters* **74**, 2422–2426 (1995). URL http://link.aps.org/doi/10.1103/PhysRevLett.74.2422.

[12] Collins, J. C., Soper, D. E. & Sterman, G. Factorization of hard processes in qcd URL http://arxiv.org/abs/hep-ph/0409313. hep-ph/0409313.

[13] Wagner, W. Top quark physics in hadron collisions. *Reports on Progress in Physics* **68**, 2409 (2005). URL http://stacks.iop.org/0034-4885/68/i=10/a=R03.

[14] Kidonakis, N. Differential and total cross sections for top pair and single top production (2012). URL http://arxiv.org/abs/1205.3453. 1205.3453.

[15] (ed), L. E. & (ed), P. B. LHC machine. *JINST* **3** (2008).

[16] Zimmermann, F. LHC: The machine. URL http://www-conf.slac.stanford.edu/ssi/2012/Presentations/Zimmermann.pdf.

[17] The CMS Collaboration. The CMS experiment at the CERN LHC. *JINST* **3** (2008).

[18] The CMS Collaboration. The tracker project technical design report. *CERN/LHCC 98-6* (1998).

[19] *The CMS muon project: Technical Design Report*. Technical Design Report CMS (CERN, Geneva, 1997).

[20] CMS Collaboration. Performance study of the cms barrel resistive plate chambers with cosmic rays (2009). URL `http://arxiv.org/abs/0911.4045`. `0911.4045`.

[21] CMSSW LXR. URL `http://cmslxr.fnal.gov/lxr/`.

[22] Brun, R. & Rademakers, F. ROOT —an object oriented data analysis framework. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **389**, 81–86 (1997). URL `http://www.sciencedirect.com/science/article/pii/S016890029700048X`.

[23] The CMS Collaboration. Description and performance of the CMS track and primary vertex reconstruction. *CMS PAPER TRK-11-001* (2011).

[24] T.Miao, Wenzel, H. & Yumiceva, F. Beam position determination using tracks. *CMS NOTE 2007/021* (2007).

[25] Krämer, T. Track parameters in LCIO. *LC-DET-2006-004* (2006).

[26] Früwirth, R., Regler, M., Bock, R. K., Grote, H. & Notz, D. *Data Analysis Techniques for High-Energy Physics* (Cambridge University Press, 2000), 2nd ed. edn.

[27] Abbiendi, G. *et al.* Muon reconstruction in the CMS detector. *CMS AN-2008/097* (2008).

[28] The CMS Collaboration. Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV. *MUO-10-004* (2010).

[29] The CMS Collaboration. Particle-flow event reconstruction in CMS and performance for jets, taus, and e_T^{miss}. *CMS PAS PFT-09-001* (2009).

[30] Baffioni, S. *et al.* Electron reconstruction in CMS. *CMS NOTE 2006/40* (2006).

[31] M.Pioppi. Electron pre-identification in the particle flow framework. *CMS AN-2008/032* (2008).

[32] Hoecker, A. *et al.* TMVA - toolkit for multivariate data analysis. *PoS ACAT:040,2007* URL `http://arxiv.org/abs/physics/0703039`. `physics/0703039`.

[33] Beaudette, F., Benedetti, D., Janot, P. & Pioppi, M. Electron reconstruction within the particle flow algorithm. *CMS AN-2010/034* (2010).

[34] Cacciari, M., Salam, G. P. & Soyez, G. The anti-k_t jet clustering algorithm (2008). URL `http://arxiv.org/abs/0802.1189v2`. `0802.1189v2`.

[35] Erdman, W. Offline primary vertex reconstruction with deterministic annealing clustering. *CMS IN-2011/014* .

[36] Waltenberger, W. Adaptive vertex reconstruction. *CMS NOTE-2008/033* .

[37] The CMS Collaboration. Commissioning of b-jet identification with pp collisions at $\sqrt{s} = 7$ TeV. *CMS PAS BTV-10-001* (2010).

[38] T.Speer *et al.* Algorithms for b jet identification in CMS. *CMS AN-2009/085* (2009).

[39] The CMS Collaboration. b-jet identification in the CMS experiment. *CMS PAS BTV-11-004* (2011).

[40] The CMS Collaboration. Identification of b-quark jets at the CMS experiment. *BTV-12-001* (2012).

[41] The CDF Collaboration. Exclusion of an exotic top quark with -4/3 electric charge using soft lepton tagging. *arXiv:1006.4597v1* (2010).

[42] The CDF Collaboration. Measurement of $B^0 - \bar{B}^0$ flavor oscillations using jet-charge and lepton flavor tagging in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV. *arXiv:hep-ex/9903011v1* (1999).

[43] Btag validation software. URL `https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideBtagValidation`.

[44] Wehrli, L., Dissertori, G. & Grab, C. *Measurement of BB Angular Correlations based on Secondary Vertex Reconstruction in Proton-Proton Collisions at 7 TeV*. Ph.D. thesis, Zurich, ETH (2012).

[45] Verkerke, W. & Kirkby, D. The RooFit toolkit for data modeling. *arXiv:physics/0306116v1* (2003).

[46] Cranmer, K. S. Kernel estimation in high-energy physics. *Comput.Phys.Commun.* **136**, 198–207 (2001). URL `http://arxiv.org/abs/hep-ex/0011057`. `hep-ex/0011057`.

[47] Adam, N., Berryhill, J., Halyo, V., Hunt, A. & Mishra, K. Generic tag and probe tool for measuring efficiency at CMS with early data. *CMS AN-2009/111* (2009).

[48] Gaiser, J. E. Charmonium spectroscopy from radiative decays of the j/psi and psi-prime. *SLAC-R-255* (1982).

[49] The CMS Collaboration. Identification of b-quark jets with the cms experiment (2012). URL `http://arxiv.org/abs/1211.4462`. `1211.4462`.

[50] The CMS Collaboration. Top pair cross section in e/mu + jets at 8 TeV. *CMS PAS TOP-12-006* (2012).

[51] Kom, C. H. & Stirling, W. J. Charge asymmetry in W + jets production at the LHC (2010). URL `http://arxiv.org/abs/1004.3404`. `1004.3404`.

[52] Najafabadi, M. M. Single top production at LHC URL `http://arxiv.org/abs/hep-ex/0605034`. `hep-ex/0605034`.

[53] Biallass, P., Hebbeker, T., Hof, C., Meyer, A. & Pieta, H. Parton distribution uncertainty determination within CMSSW. *CMS AN-2009/048* (2009).