ABSTRACT

Title of dissertation: USING MACHINE LEARNING TO MEASURE THE CROSS SECTION OF TOP QUARK PAIRS IN THE MUON+JETS CHANNEL AT THE COMPACT MUON SOLENOID

Malina Kirn, Doctor of Philosophy, 2011

Dissertation directed by: Professor Nicholas Hadley
Department of Physics
Department of Applied Mathematics & Statistics, and Scientific Computation

The cross section for $pp \to t\bar{t}$ production at a center of mass energy of 7 TeV is measured using a data sample with integrated luminosity 36.1 pb^{-1} collected by the CMS detector at the LHC. The analysis is performed on a computing grid. Events with an isolated muon and three hadronic jets are analyzed using a multivariate machine learning algorithm. Kinematic variables and b tags are provided as input to the algorithm; output from the algorithm is used in a maximum likelihood fit to determine $t\bar{t}$ event yield. The measured cross section is $151 \pm 15(stat.)^{+35}_{-28}(syst.) \pm 6(lumi.)$ pb.

USING MACHINE LEARNING TO MEASURE THE CROSS SECTION OF TOP QUARK PAIRS IN THE MUON+JETS CHANNEL AT THE COMPACT MUON SOLENOID

by

Malina Aurelia Kirn

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2011

Advisory Committee:
Professor Nicholas Hadley
Professor Konstantina Travisa
Professor William Dorland
Research Professor Steven Wallace
Associate Professor Alan Sussman

Acknowledgments

Special thanks to Nick Hadley and Jeff Temple for many years of guidance. Additionally, the work in this dissertation would not be possible without tremendous assistance from many people:

- Daniel Bloch: b-tagging
- Brian Bockelman: Condor
- Kevin Burkett: track reconstruction
- Matthew Chadwick: b-tagging
- Dave Evans: ProdAgent
- Patrick Gartung: system administration
- Arvind Gopu: RSV
- Jasmin Gruschke: event selection
- Oliver Gutsche: CRAB, CMSSW, and grid services
- Burt Holzman: OSG
- Doug Johnson: system security

- Bockjoo Kim: CMSSW installation
- Nicolo Magini: PhEDEx
- Petar Maksimovic: fits
- Charles Plager: pseudo-experiments
- Salvatore Rappoccio: PAT and b-tagging
- Paul Rossman: PhEDEx
- Alain Roy: Condor
- Alex Sim: BeStMan
- Rob Snihur: Tier 3 clusters
- Bill Strossman: networks
- Michael Thomas: Rocks
- Anthony Tiradani: OSG
- Eric Vaandering: CRAB
- Jason Williams: cluster power
- Francisco Yumiceva: b-tagging

Table of Contents

List of Figures vi

List of Abbreviations x

1 Introduction 1

2 High energy physics 5
2.1 The Standard Model . . . 7
2.2 Top quark physics . . . 15
2.3 Chapter highlights . . . 18

3 Experimental apparatus 20
3.1 The Large Hadron Collider . . . 21
3.2 The Compact Muon Solenoid detector . . . 29
3.2.1 Superconducting solenoid . . . 34
3.2.2 Inner tracker . . . 36
3.2.3 Electromagnetic calorimeter . . . 37
3.2.4 Hadron calorimeter . . . 41
3.2.5 Muon chambers . . . 42
3.2.6 Luminosity monitors . . . 45
3.2.7 Event triggers . . . 46
3.3 Chapter highlights . . . 47

4 Grid computing 49
4.1 Open Science Grid . . . 49
4.1.1 Virtual organizations . . . 50
4.1.2 Authentication . . . 52
4.1.3 Compute and storage elements . . . 54
4.2 CMS resources, services, and software . . . 56
4.2.1 CMS global services . . . 57
4.2.2 PhEDEx . . . 59
4.2.3 CRAB . . . 62
4.2.4 ProdAgent . . . 63
4.2.5 Tiers . . . 65
4.3 The University of Maryland Tier 3 . . . 66
4.4 Chapter highlights . . . 70

5 Event simulation, reconstruction, and selection 71
5.1 Simulated data . . . 72
5.1.1 Event generation using MadGraph & Pythia . . . 72
5.1.2 Event simulation using GEANT . . . 75
5.2 Event digitization . . . 76
5.3 Event reconstruction . . . 76

5.3.1 Track, beamspot, and vertex reconstruction . . . 77
5.3.2 Muons . . . 80
5.3.3 Jets . . . 81
5.3.4 b-tagging . . . 84
5.4 Data and Monte Carlo simulated samples . . . 86
5.5 Event selection . . . 90
5.6 Chapter highlights . . . 94

6 Analysis techniques 96
6.1 Signal and background discrimination . . . 97
6.1.1 The backpropagation algorithm . . . 102
6.1.2 Input variables . . . 104
6.1.3 Discriminator . . . 106
6.2 Using data to correct or replace simulations . . . 108
6.2.1 $t\bar{t}$ and single top templates . . . 110
6.2.2 QCD template . . . 115
6.2.3 V+jets template . . . 117
6.3 Chapter highlights . . . 120

7 Measurement of the cross section and its uncertainty 123
7.1 Fitting . . . 123
7.2 Pseudo-experiments . . . 126
7.3 Fit statistical uncertainty . . . 127
7.4 Systematic uncertainties . . . 128
7.4.1 Jet energy scale . . . 130
7.4.2 b tagging . . . 131
7.4.3 V+heavy flavor . . . 132
7.4.4 Jet energy resolution . . . 133
7.4.5 MadGraph and Pythia event generation settings . . . 133
7.5 Chapter highlights . . . 135

8 Conclusion 136
8.1 The measured $t\bar{t}$ cross section . . . 136
8.2 Outlook . . . 138
8.3 Chapter highlights . . . 145

Bibliography 146

List of Figures

2.1 The elementary particles in the Standard Model are divided into leptons, quarks, and force carriers. Leptons and quarks come in three generations. Credit: [35]. 8

2.2 Possible leading order Feynman diagrams for the process $pp \to t\bar{t}$ (a) s-channel gluon fusion, (b) t-channel quark-gluon interaction, and (c) s-channel quark-antiquark annihilation. 12

2.3 Example modifications to the leading order diagrams in Figure 2.2 at (a) next to leading order (NLO) and (b) next-to-next to leading order (NNLO). 14

3.1 The SPS, the LHC, and the experimental caverns. The small silver ring is the SPS and the large blue ring is the LHC, both described in Section 3.1. The experimental cavern located opposite SPS on the LHC is for the CMS detector. Credit: [13]. 21

3.2 The CERN accelerator complex. Notable components are the Linac 2, the booster, the PS, the SPS, and the LHC. Credit: [13]. 25

3.3 The LHC dipole magnet. Credit: [13]. 26

3.4 The Compact Muon Solenoid. Credit: [16]. 30

3.5 Transverse view of CMS with different types of particles and the detectors with which they interact. Credit: [13]. 32

3.6 The $|\eta|$ distribution of a reconstructed muon originating from the massive top quark (red and pink) and lower mass W or Z bosons (green) and QCD (black). The distributions are normalized to show shape comparisons. 33

3.7 Quartered longitudinal view of CMS with values of $|\eta|$ shown as dashed lines, r values on the right edge, and z values on the lower edge. Credit: [16]. 34

3.8 End view of the solenoid and portions of the iron return yoke at CMS. Credit: [16]. 36

3.9 Quarter longitudinal view of the strip layers in the inner tracking detector at CMS. Blue layers are made of glued orthogonal strips while red layers are composed of single strips in one orientation. The marks drawn on the top and right edges are values of $|\eta|$. Credit: [16]. 38

3.10 Quartered longitudinal view of the electromagnetic calorimeter at CMS, composed of lead tungstante crystals. Dashed lines correspond to values of $|\eta|$. Credit: [16]. 39

3.11 Quartered longitudinal view of the hadronic calorimeter at CMS, a sampling calorimeter. Towers are numbered from 1 to 29 and are composed of 17 active readout layers in the barrel and 19 layers in the endcaps. Each readout layer is composed of plastic scintillators and a dense material, primarily brass. Credit: [16]. 42

3.12 Quartered longitudinal view of the muon chambers at CMS. The three types of gaseous cambers, drift tubes, resistive plate chambers, and cathode strip chambers are labeled. Boxed white regions correspond to the iron magnetic return yoke and the solenoid. The HCAL is shadowed in purple/blue, the ECAL in green, and the inner tracker in orange/red. Credit: [16]. 43

3.13 Delivered and recorded luminosity at CMS in 2010. Credit: [43]. . . . 46

4.1 Weekly number of consumed computing hours in OSG from July 2010-2011, shown by virtual organization. Credit: [53]. 51

4.2 Weekly PhEDEx transfer volume from July 2010-2011, shown by destination. Generated using a configured query on the PhEDEx web service. 60

5.1 Possible leading order Feynman diagrams for the process $gg \to t\bar{t}$+1 jet, where an additional gluon is produced from (a) initial state radiation and (b) final state radiation. 73

5.2 An event observed at CMS with a muon and four jets, one of which is identified ("tagged") as originating from a b quark. Dark green lines in the center are reconstructed tracks, blue and red histograms are energy deposits in the calorimeters, the red line traversing the muon chambers is the reconstructed muon, and the pale green segments are reconstructed jets. Tracks and calorimeter energy deposits with no corresponding jet or muon are interpreted as particles in the interaction that were not reconstructed, electronic noise, or the underlying event. Credit: [24]. 78

5.3 Distribution of b-tag discriminators for data (points) and light (blue), charm (green), and bottom (red) jets in simulated events. (a) TCHE, (b) TCHP, (c) SSVHE, (d) SSVHP. Credit: [18]. 87

5.4 Light flavor versus b flavor efficiency for different tagging algorithms. Jet probability (JP) and jet b probability (JBP) tagging algorithms were not considered for this analysis. Credit: [18]. 88

5.5 Leading order Feynman diagrams for single top production via (a) s-channel quark-antiquark annihilation and (b) t-channel quark-W interaction. 90

6.1 The output of a single neuron (Equation 6.2) is the value of a function (Equation 6.1) given a shifted weighted sum of the neuron inputs (Equation 6.3) as the function argument. 99

6.2 The MLP used for this analysis. Weights of form w^r_{j0} correspond to the neuron shift, while the remaining weights correspond to neuron-to-neuron edge weights. For example purposes, only some of the weights are explicitly labeled. Each output value y is given by Equation 6.2 and shown schematically in Figure 6.1. 101

6.3 Inputs to the neural network for simulated events: (a) b tag boolean, (b) ΔR_{12}, and (c) $|\eta^{\mu}|$. Simulated $t\bar{t}$ events are in red, single top events in pink, V+jets events in green, and QCD events in black. All distributions are normalized to unity to show shape comparisons. . . 107

6.4 Normalized simulated discriminator templates for (top left) signal $t\bar{t}$, (top right) W/Z+jets background, (bottom left) single top background, and (bottom right) QCD background. 109

6.5 Data- driven or corrected discriminator distributions for (top left) data-corrected $t\bar{t}$ (corrected tag rates and jet energy resolution), (bottom left) data-corrected single top (corrected tag rates and jet energy resolution), (top right) data-driven V+ jets (events with exactly two jets), and (bottom right) data-driven QCD (events with muon isolation greater than 0.1). The discriminator distributions of simulated events passing nominal selection are overlaid for reference, though the distributions are not expected to agree. 111

6.6 The $p_{\mathrm{T}}^{\mathrm{rel}}$ distribution of muons associated with b jets (red) and light or c jets (blue) for (a) jets tagged by the track counting high purity algorithm medium cut (TCHPM) and (b) jets not tagged by TCHPM. Credit: [18]. 113

6.7 The b tag distributions for negative and nominal algorithms. The data are shown as points and simulations as colored stacked histograms, with b jets shown in red, c jets in green, and light jets in blue. The negative tag discriminant values are indicated by lighter shading. Credit: [18]. 114

6.8 Distribution of muon isolation from all signal and background processes considered in this analysis for events passing all selection criteria except the requirement on muon isolation. The last bin contains events with values of RelIso greater than 1 (overflow). The number of predicted events is scaled by a luminosity of 36 pb^{-1}. 116

6.9 Normalized inputs to the neural network for simulated QCD events passing nominal event selection (solid red line), simulated events with muon isolation greater than 0.1 (dashed line), and data events with muon isolation greater than 0.1 (points) for (a) the b-tag boolean, (b) ΔR_{12}, and (c) $|\eta^{\mu}|$. 118

6.10 Number of jets with $p_{\mathrm{T}} > 30$ GeV from all signal and background processes considered in this analysis for events passing all selection criteria except the requirement of three or more jets. The number of predicted events is scaled by a luminosity of 36 pb^{-1}. 119

6.11 Normalized inputs to the neural network for simulated V+jets events passing nominal event selection (solid red line), simulated events with exactly two jets (dashed line), and data events with exactly two jets (points) for (a) the b-tag boolean, (b) ΔR_{12}, and (c) $|\eta^{\mu}|$. 121

7.1 Corrected $t\bar{t}$ event yields and pull distributions from 10,000 pseudo-experiments for an integrated luminosity of 36.1 pb^{-1}. The expected number of $t\bar{t}$ events is 387. 128

8.1 Results of template fit to the discriminant calculated from 36.1 pb^{-1} of data. 137

8.2 Inputs to the neural network for simulations and data, where event yields and statistical uncertainties are taken from the fit to the data: (a) the b tag boolean, (b) ΔR_{12}, and (c) $|\eta^{\mu}|$. 139

8.3 The sum of the transverse energy of all jets with $p_T > 30$ GeV (H_T^{30}) for various simulated interactions. All distributions are normalized to unity in order to show shape comparisons. 141

List of Abbreviations

b	bottom quark or bottom jet
c	charm quark or charm jet
g	gluon
μ	muon
ν	neutrino
q	quark
$t\bar{t}$	top and antitop quark pair
V	vector boson (W or Z boson)
W	W boson
Z	Z boson
b	barn
B	byte
eV	electron volts
ANN	artificial neural network
ATLAS	A Toroidal LHC Apparatus
BDII	Berkeley database information index
BeStMan	Berkeley storage manager
CDF	Collider Detector at Fermilab
CE	compute element
CEMon	CE monitor
CERN	European Organization for Nuclear Research
CKM	Cabibbo-Kobayashi-Maskawa
CMS	Compact Muon Solenoid
CMSSW	CMS software framework
CRAB	CMS remote analysis builder
CSC	cathode strip chamber
CTF	combinatorial track finder
DBS/DLS	data bookkeeping/location service
DT	drift tube
DN	distinguished name
ECAL	electromagnetic calorimeter
GEANT	geometry and tracking
HCAL	hadronic calorimeter
HDFS	Hadoop distributed file system
HLT	high level trigger
HS06	HEP-SPEC06
ISR/FSR	initial/final state radiation
JES	jet energy scale
L1	level 1 trigger
LHC	Large Hadron Collider

LO	leading order
ME	matrix element
MLP	multi layer perceptron
NFS	network filesystem
NLO	next to leading order
NNLO	next-to-next to leading order
OSG	Open Science Grid
PDF	parton distribution function
PhEDEx	physics experiment data export
ProdAgent	production agent
PS	proton synchrotron or parton shower
PV	primary vertex
QED	quantum electrodynamics
RAID	redundant array of independent disks
RF	radio frequency
RPC	resistive plate chamber
RSV	resource service validation
SE	storage element
SPS	super proton synchrotron
SRM	storage resource management
SSV	simple secondary vertex
SSVHE	simple secondary vertex high efficiency
SSVHEM	simple secondary vertex high efficiency medium
SSVHP	simple secondary vertex high purity
TC	track counting
TCHE	track counting high efficiency
TCHEM	track counting high efficiency medium
TCHP	track counting high purity
TMDB	transfer management database
VDT	virtual data toolkit
VO	virtual organization
WLCG	worldwide LHC computing grid
WM	workload management

Chapter 1

Introduction

Research in particle physics can be divided into three major thrusts: high energy physics, high intensity physics, and particle astrophysics. Particle physics research studies the interactions of particles via the fundamental forces: electromagnetism, the strong and weak nuclear forces, and gravity. The Large Hadron Collider (LHC) is located at the European Organization for Nuclear Research (CERN) in Geneva, Switzerland. The LHC is the highest energy particle accelerator in the world, currently colliding two beams of protons at 3.5 TeV per beam, or 7 TeV center of mass energy (labeled as $\sqrt{s}$). The LHC began collisions at 7 TeV on March 30, 2010. Experiments at the LHC include the Compact Muon Solenoid (CMS) and A Toroidal LHC ApparatuS (ATLAS), as well as two other smaller, focused experiments.

ATLAS and CMS are general purpose detectors, designed to discover any new particles that might be produced at the LHC. The Standard Model has been very successful at predicting all known particle behavior. The Tevatron, with a center of mass energy of 1.96 TeV, was the world's highest energy collider prior to the LHC and a large number of particle properties were measured at the Tevatron. Particle properties measured at the Tevatron and other colliders were extrapolated to the higher energy of the LHC using the Standard Model. The extrapolations need to be

confirmed at the LHC before discovery of any new particles can reliably be claimed.

The top quark was discovered in 1995 by the DØ [29] and Collider Detector at Fermilab (CDF) [14] experiments at the Tevatron. Its mass was unexpectedly large, 172 GeV, making it the heaviest known particle and stimulating theories beyond the Standard Model. Until the LHC began operations, the Tevatron was the only place on earth where top quarks could be produced and studied. A collision or interaction is referred to in particle physics as an event. At the LHC, an event is a collision between two protons. Events containing the expected experimental signature from top quarks are selected from the large set of all observed events. Background is anything that produces the same final state (signature) but is not produced by the physics of interest (signal). Determining if the selected collisions come from top quarks or background and measuring the rate at which top quarks are produced at the LHC are the primary goals of this dissertation.

Discrimination between top quark signal events and background events with the same signature is performed using a multivariate machine learning algorithm, specifically an artificial neural network (ANN). Machine learning algorithms are characterized by their goal, classification or regression (parameter estimation), and their learning method, supervised or unsupervised. We employ a supervised ANN for classification purposes, although we do not use the output of the ANN, termed the discriminant or discriminator, to classify individual events. Rather, the shape of the discriminator for signal and background events is fit to the discriminant observed in data. This determines the quantity of signal and background present in the selected events, yielding the rate of production of top-antitop quark pairs, or $t\bar{t}$, called the

production cross section.

The ANN is trained using simulated events. Monte Carlo simulations use the established rules of the Standard Model and previously measured parameters in the theory to simulate events as they would appear in the detectors of CMS. Some physical processes are especially well modeled by the simulations, while others are modeled less well. Data is used to correct or replace the simulations that are known to model certain types of particle interactions poorly.

The underlying assumptions in the Monte Carlo simulation and the experimental precision of measured particle characteristics introduce sources of uncertainty in the final measurement of the $t\bar{t}$ cross section. The cross section uncertainty is measured using pseudo-experiments (also known as Toy Monte Carlo), in which Monte Carlo simulations of events are changed to account for a particular type of uncertainty. The change in the measured $t\bar{t}$ cross section from each set of pseudo-experiments is then the uncertainty of the measured $t\bar{t}$ cross section from the uncertainty simulated in each set.

Event simulation, reconstruction, and analysis are performed on an international computing grid. CMS has 183 member institutions in 38 countries. Locating computing facilities in close proximity to the analysts significantly improves network latency, but requires underlying mechanisms to share data and compute cycles between individual computing sites. One such computing site is located at the University of Maryland. It is used to service the storage and analysis requirements of both local and remote users, including the analysis presented here.

We report a measurement of the $t\bar{t}$ production cross section at $\sqrt{s} = 7$ TeV in

proton-proton collisions with the CMS detector at the LHC. For this measurement, we provide particle characteristics as inputs to a machine learning algorithm in order to separate $t\bar{t}$ events from background. Data are used to correct or replace some Monte Carlo simulations. The Standard Model and top quark physics are discussed in Chapter 2; the experimental apparatus of the LHC and CMS in Chapter 3; the computing grid in Chapter 4; event simulation, reconstruction, and selection in Chapter 5. The multivariate machine learning algorithm and methods to correct or replace the Monte Carlo simulations with data are described in Chapter 6. The fitting procedure and sources of uncertainty are described in Chapter 7, with the measured $t\bar{t}$ cross section and outlook in Chapter 8. The contents of Section 4.3 and most of the material in Chapters 6, 7, and 8 are the unique work of the author.

Chapter 2

High energy physics

In 1913, Niels Bohr introduced the Bohr model of the hydrogen atom, which explained the origin of spectral emission lines from hydrogen. The Bohr model uses orbital shells with electrons orbiting the hydrogen nucleus at discrete radii. When an electron moves from one shell to another, it absorbs or emits photons. The photons have energy corresponding to the change in electron energy in the electromagnetic potential created by the nucleus. Since the energy of a photon is proportional to its frequency and the electrons orbit in discrete shells, the frequency of light emitted from a nucleus is quantized; these are called spectral emission lines. Eventually this was understood through quantum mechanics, in which particles possess certain quantum characteristics. The radius of a particle in orbit takes discrete values because it is related to the orbital angular momentum of the particle, which is quantized.

The nucleus was discovered by Ernest Rutherford in 1910 by firing particles at a thin foil of atoms. The scattering angles were only consistent with the existence of a dense atomic nucleus. Rutherford posited the nucleus was composed of particles with a net positive charge in order to electromagnetically attract the negative orbiting electrons. If the nucleus had multiple positive particles, termed protons, something had to prevent them from electromagnetically repelling each

other. Studies of nucleus spin and nuclear decay in the late 1920s and early 1930s indicated the nucleus also contained neutral particles, neutrons, which could not interact electromagnetically. Although several mechanisms had been proposed for the cause of nuclear structure, it was not until 1935 that Hideki Yukawa proposed the existence of a new nuclear force, the strong force. This force interacted with the particles in the nucleus, overcoming the electromagnetic repulsion between particles in the nucleus to bind them together. The force did not interact with the orbiting electrons. It was eventually discovered in the 1960s that the proton and neutron are themselves composite objects, formed from quarks bound together with gluons, the carrier of the strong force.

Rutherford also realized that radioactive decay was due to the disintegration of the nucleus. Enrico Fermi introduced the weak force as the mechanism causing most radioactive decays in 1933. Direct observation of the weak force carriers was not possible until 1983, when the charged W and neutral Z bosons were discovered at the Super Proton Synchrotron at CERN. The W and Z bosons are heavy, with mass of 80 and 91 GeV, respectively (the proton has mass 0.94 GeV) [50]. Due to their large masses, the W and Z lifetimes are short, making the weak force act over short distances. Radioactive decay also led to the discovery of the neutrino. Neutrinos were postulated in 1930 by Wolfgang Pauli because the products of radioactive decays appeared to violate conservation of energy and momentum. Since the neutrino is not charged and does not interact via the strong force, the neutrino did not interact with the detectors observing radioactive decay and carried unobserved energy and momentum from the decay.

2.1 The Standard Model

The studies of atomic structure and decay in the early 20th century led to the development of quantum mechanics and the Standard Model of particle physics. The Standard Model is the theory describing the interactions of all known particles. The known particles are categorized as leptons, quarks, and force carriers. See Figure 2.1. Leptons include the electron, the heavier muon and tau, and their accompanying neutrinos. Quarks are never observed in isolation, rather, they bond to produce hadrons such as the proton and neutron. Leptons and quarks are divided into three generations. For those particles whose mass has been measured, each generation is heavier than the previous, requiring increasing quantities of energy to produce. The photon is the electromagnetic force carrier, the gluon is the strong force carrier, and the W and Z bosons are the weak force carriers.

The theory of quantum electrodynamics (QED) describes electromagnetic particle interactions via quantum mechanics. In QED, the force of electromagnetism is exerted on particles possessing electric charge via the photon. QED is the mathematical basis for quantum field theory, earning Sin-Itiro Tomonaga, Julian Schwinger, and Richerd Feynman the 1965 Nobel prize in physics. Feynman developed diagrams and rules to describe all electromagnetic particle interactions, which have subsequently been extended for the weak and strong nuclear forces.

The weak force regulates nuclear decay and is exerted on particles possessing weak isospin via the W and Z bosons. W bosons possess weak isospin and electric charge, while Z bosons do not. Some quarks and leptons possess weak isospin and

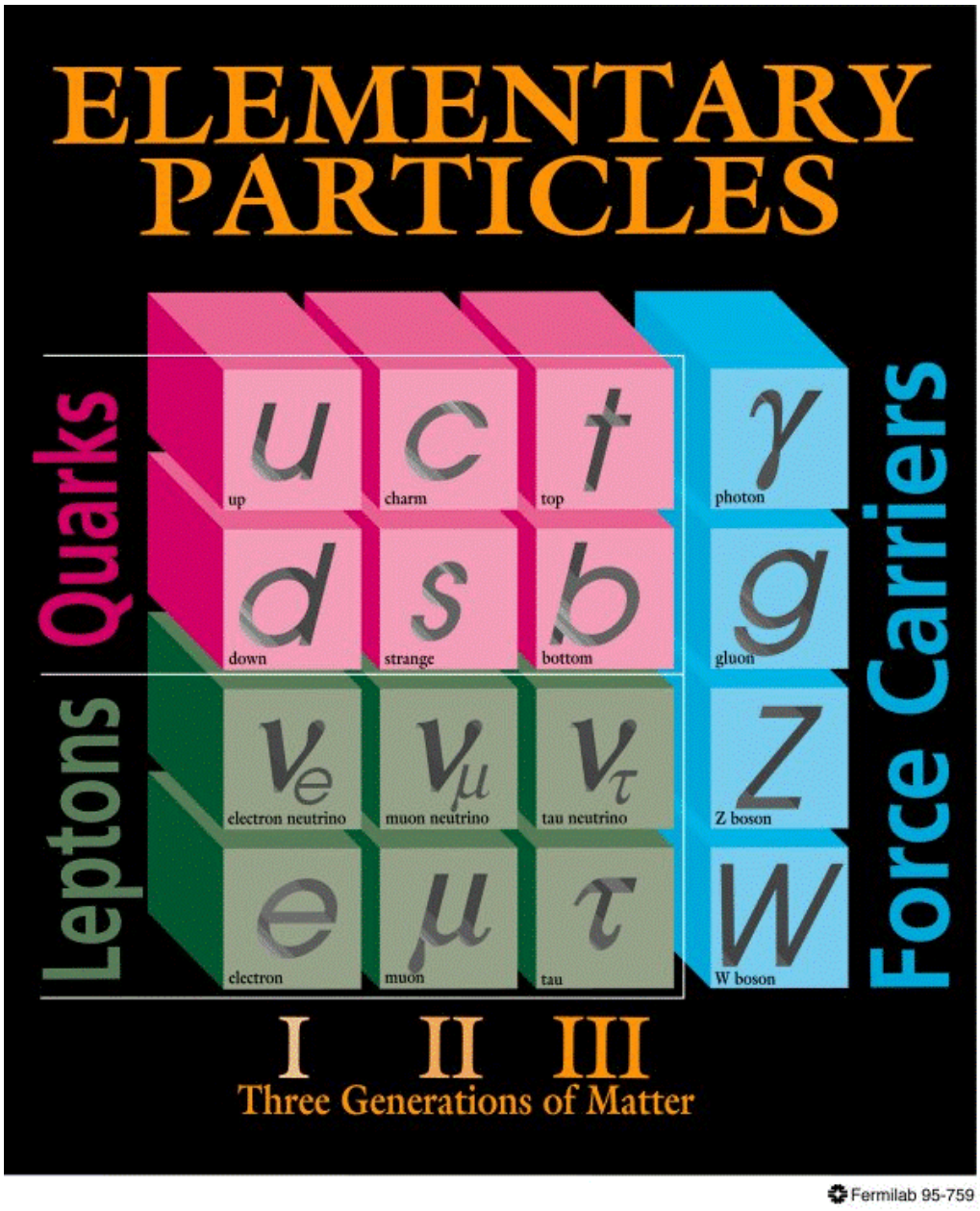


Figure 2.1: The elementary particles in the Standard Model are divided into leptons, quarks, and force carriers. Leptons and quarks come in three generations. Credit: [35].

therefore can interact via the weak force.

The strong force charge is termed color, which every quark possesses. Since individual quarks always have a color or anticolor, color neutral objects are formed by multiple quarks. The gluon carries both color and anticolor. The electric force exerted by a particle with electric charge decreases in strength over distance because the photon does not carry electric charge. However, since the gluon carries color charge, the strong force exerted by a colored object does not decrease over distance. Therefore, it is energetically favorable for colored quarks to bind into composite, color neutral particles (termed color confinement). Far away from a color neutral particle, the strong force is negligible.

Color confinement plays an important role in the process of hadronization. Individual quarks produced in a collision will propagate away from each other. As they get further apart, it becomes energetically favorable to create a new quark and antiquark near each original quark. The new particles bind with the original particles to produce color neutral objects, breaking the strong bond between the distant quarks. Due to the combinatorics of producing color neutral particles, the hadronization process continues until it is no longer energetically favorable to continue pulling new quark-antiquark pairs from the vacuum. Indeed, isolated quarks are not observed directly in particle detectors, rather, a jet of color neutral objects produced by hadronization of some originating quark or gluon is observed. The theory of quantum chromodynamics, or QCD, governs all colored interactions. Technically the term QCD includes every event containing a quark or gluon, but since the occurrence of quarks and gluons at the LHC is exceedingly common, QCD

is used in this dissertation to refer to events in which the produced particles are all hadrons.

The Standard Model also incorporates conservation laws for particle characteristics such as energy and momentum. Additional conservation laws especially important in interactions involving the top quark are electric charge and quark flavor. Quark flavor refers to the three generations of quarks shown in Figure 2.1. Each quark type has its own flavor, with particles having flavor of +1 and antiparticles having flavor of -1. Quark flavor is conserved by the strong force, but is not conserved by the weak force. This is especially significant in top quark physics, as weak force decay is the mechanism by which the heavy top quark eventually decays via intermediates to lighter quarks.

Notably missing from the Standard Model is the force carrier for gravity, which some theories beyond the Standard Model attempt to incorporate. At the energies of the LHC, the force of gravity is many orders of magnitude smaller than the other forces. Gravity impacts the behavior of particles at the LHC so little that predictions of particle interactions are formed without consideration of its effects. No theory beyond the Standard Model has yet been confirmed by experimental evidence, though the higher energies accessible by the LHC may allow discovery of particle behavior not consistent with the Standard Model.

An additional component of the Standard Model is the Higgs boson, which is the proposed mechanism by which particles acquire mass. Since the Higgs boson has not yet been discovered, it is not included in Figure 2.1, but it is predicted within the framework of the Standard Model. The discovery of the Higgs boson,

along with potential observation of particle behavior beyond that predicted by the Standard Model, are the two primary goals of the CMS and ATLAS experiments at the LHC.

Particle processes are typically written in text with arrows signifying the interaction parent(s) to the left and interaction product(s) to the right. For example, $pp \rightarrow t\bar{t} \rightarrow W^{+}bW^{-}\bar{b} \rightarrow q\bar{q}'b\mu\bar{\nu_{\mu}}\bar{b}$ refers to a process in which two protons interact to produce top and antitop quarks, which each decay to a W boson and bottom quark. The W bosons subsequently decay, one to a quark and antiquark of different flavor and one to a muon and a muon neutrino.

Interactions between particles are represented through the use of Feynman diagrams. Feynman diagrams add information such as mediating particles and specific particle interactions. Feynman diagrams use springs for gluons, waves for photons, dashed lines for the Z and W bosons, and directional lines for quarks and leptons. Feynman diagrams that represent the first interaction in the example above, two protons interacting to produce a top and antitop quark, can be drawn in three significantly contributing ways, as shown in Figure 2.2. Since the proton is a composite object, the proton constituents can interact in various ways to produce top and antitop quarks. Provided kinematic properties of the parent particles, a single Feynman diagram contains enough information to calculate the kinematic properties of all child particles. Conservation laws and the interaction rules of the Standard Model determine the allowed kinematic properties. The convention used for the Feynman diagrams in this dissertation is that time is on the horizontal axis, so initial incoming particles are drawn to the left while final outgoing particles are drawn to the right.

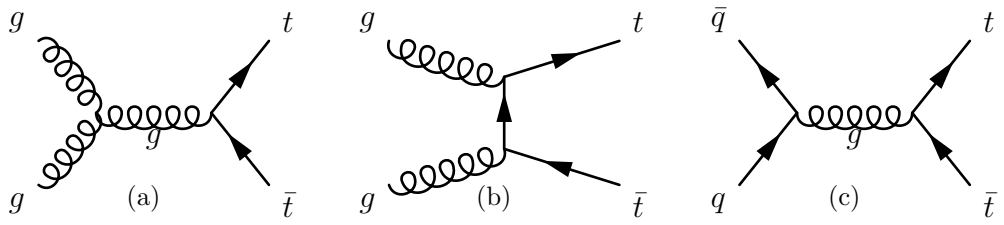


Figure 2.2: Possible leading order Feynman diagrams for the process $pp \to t\bar{t}$ (a) s-channel gluon fusion, (b) t-channel quark-gluon interaction, and (c) s-channel quark-antiquark annihilation.

The energy of a particle with mass and momentum is given by:

$$E^2 = p^2c^2 + m^2c^4 \tag{2.1}$$

where p is the momentum of the object and m is its mass. In high energy physics, $c = 1$ is chosen as a convention, which means energy, momentum, and mass can be quoted in units of electron volts (eV). An important feature of conservation laws is that they must hold in every frame of reference. In the rest frame of a particle, $p = 0$, therefore, the only energy available is the mass of the particle. Consider the rest frame of the top-antitop system in the Feynman diagram Figure 2.2c. The gluon decays to two top quarks, where the mass of each top quark is approximately 172 GeV [50]. Therefore, to produce two top quarks, the gluon must have at least 344 GeV of energy. When considering $t\bar{t}$ production from a gluon within the rest frame of the top-antitop system, one might think that energy conservation is violated, since the top quarks have non-zero rest mass, while the gluon is massless. However, Heisenberg's uncertainty principle plays an important role here. Heisenberg's uncertainty principle states that the uncertainty in a particle's energy is directly related to its lifetime: $\Delta E \Delta t \geq \hbar$. The gluon in Figure 2.2c is virtual, i.e., it exists

for a finite period of time. The less time the gluon exists Δt, the larger ΔE must get, i.e., a virtual gluon cannot have a definite energy. Though the mass energy of a gluon is zero when it is long lived, when the gluon is short lived, it is allowed to have a mass energy distribution that is not zero – the gluon is allowed to be "off mass shell". Conservation of energy is not violated in this process. Rather, the energy of short lived particles cannot be fixed with perfect precision, allowing this otherwise forbidden interaction to occur.

An amplitude describing the probability of going from the initial state to the final state can be formed using the information in the Feynman diagrams. By using the Lagrangian describing different types of particle interactions, a set of Feynman rules have been formulated that are applied to the particles and interaction vertices. The Feynman rules give the mathematical terms that are multiplied or added to create a probability amplitude. Like the kinematic solution of a Lagrangian, the probability amplitude describes the kinematics of the outgoing particles as a function of the kinematics of the incoming particles and vice versa. Unlike a classical Lagrangian, it is a statistical statement of the probability amplitude of the final state given the probability amplitude of the initial state and vice versa.

The complexity of a diagram indicates the mathematical order of the associated perturbative modification to the probability amplitude. The diagrams in Figure 2.2 are all leading order (LO), but more complicated interactions can occur in $pp \to t\bar{t}$. Example next to leading order (NLO) and next-to-next to leading order (NNLO) diagrams are shown in Figure 2.3 and generally contribute to the probability amplitude less significantly than the leading order processes.

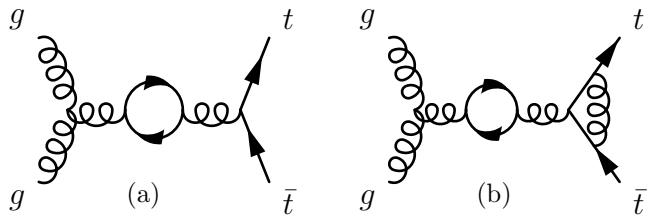


Figure 2.3: Example modifications to the leading order diagrams in Figure 2.2 at (a) next to leading order (NLO) and (b) next-to-next to leading order (NNLO).

Additionally, particle interactions with the same initial and final particles but different intermediate (virtual) particles interfere with each other. The principle of quantum interference is most clearly demonstrated by Young's "double slit" experiment, in which light passing through two small slits to illuminate a screen or plate shows a constructive and destructive pattern indicating the interference pattern expected from wave propagation. The double slit experiment was duplicated using a source emitting single photons [38], which also cumulatively landed in a pattern consistent with wave interference. When a particle is allowed access to multiple paths, it will follow all possible paths in a truly probabilistic way until the probability distribution is collapsed by observation. The Feynman diagrams in Figures 2.2a, 2.2b and Figure 2.3 all interfere because they contain exactly the same initial and final particles. Interference can add constructively or destructively to the probability amplitude, often as a function of the kinematics of the final particles. It is important to note that these terms contribute to the same probability amplitude; despite the particles taking different paths from initial to final state, all possible paths contribute to the probability amplitude describing the interaction.

The Feynman diagram in Figure 2.2c does not contribute to the same proba-

bility amplitude as the first two since it does not contain the same initial and final state particles. However, since proton interactions allow both gluons and quarks as initial state particles, the diagram in Figure 2.2c, as well as its corresponding higher order diagrams, still contributes to the overall production rate of $pp \to t\bar{t}$.

2.2 Top quark physics

The top quark occupies a unique position within the Standard Model. With a mass roughly equal to that of a tungsten atom, it is the only quark heavy enough to decay before it hadronizes. The large mass of the top quark has inspired numerous theories in which the top quark plays a special role in the generation of mass or in beyond Standard Model physics. For theories in which the top quark is not directly a part of the new physics, it is often an important background. Additionally, since the decay products of the top quark can involve every subdetector, top quarks can be used for cross-detector calibration.

The top quark decays via the weak force with a lifetime that is inversely proportional to its decay width:

$$\begin{aligned} \tau_t &= \hbar/\Gamma_t, \\ \Gamma_t^0 &= \frac{G_F m_t^3}{8\pi\sqrt{2}} \times |V_{tb}|^2 \end{aligned} \tag{2.2}$$

where Γ_t^0 is the leading order term in the top quark decay width, making the top quark lifetime proportional to $1/m_t^3$ [30]. Since the top quark is so massive, its lifetime is very short, $\mathcal{O}(10^{-25})$ seconds. The top quark nearly always decays to the next heaviest quark, a bottom quark (observed as a jet), and a W boson. W bosons

decay to a quark and an antiquark (two jets) with a branching ratio of roughly 2/3 and to a charged lepton and neutrino with a branching ratio of roughly 1/3. By searching for top quark pairs in which one W decays leptonically, background from QCD jet-only events is significantly reduced. Requiring one W to decay leptonically and one to decay hadronically yields a branching ratio of $1/3 \times 2/3 + 1/3 \times 2/3 = 4/9$ for $t\bar{t}$ events, i.e., 4/9 of $t\bar{t}$ events decay to a charged lepton, neutrino, and four jets (two b jets from the two top quarks, and two jets from a W). Requiring both W bosons to decay leptonically reduces the QCD background even further, but reduces the branching ratio and therefore the available statistics ($1/3 \times 1/3 = 1/9$). The decay of a top-antitop ($t\bar{t}$) pair is categorized by the decay of the W bosons produced by the pair. Thus, the channel in which both W bosons decay to leptons is referred to as the "dilepton" channel and the channel in which one W decays to leptons and the other to quark jets is referred to as the "lepton plus jets" channel.

Neutrinos only interact weakly with matter, so have a very low probability of interacting with any part of the detector, though they can be inferred by missing energy. Therefore, some searches for $t\bar{t}$ events require a lepton, jets, and missing energy. However, due to uncertainties in jet energy reconstruction, it was assumed that the uncertainty of reconstructing missing energy in early data taking would be large and we chose not to use missing energy in our event selection.

A further categorization of $t\bar{t}$ decays is made by specifying the flavor of the leptons produced by the W decays. The Standard Model contains three generations of leptons. From lightest to heaviest, they are: electron, muon, and tau (with their accompanying neutrinos). Muon reconstruction tends to be higher purity than

electron reconstruction because very few particles other than the muon penetrate to the muon chambers in CMS. Taus, as the heaviest leptons, decay to a variety of signatures and are difficult to discriminate from jets and other leptons. Therefore, due to the lower purity of reconstructed electrons and the variety of ways that taus decay, reconstructed muons are the best measured lepton flavor. We choose to consider $t\bar{t}$ events in the "muon+jets channel", in which one W decays to a muon (and muon neutrino) and the other W decays to two jets.

Since W bosons can decay leptonically, events that produce a leptonically decaying W plus a few additional jets are the largest source of background. The signature of a charged lepton, a neutrino, and four jets can be produced by top quark pairs, the signal, or by a W boson with extra jets, the largest background. Additional sources of background are leptonically decaying single top events produced with extra jets and QCD jet-only events in which one jet 'fakes' a charged lepton. Jets faking leptons will be discussed further in Section 5.3.3.

Bottom quarks hadronize into B hadrons, composite particles formed from at least one b quark. The B hadrons decay to a hadron containing only lighter quarks via the weak force. Calculating the decay lifetime of B hadrons is considerably more complicated than for the top quark because of the variety of B hadrons, the multiple mechanisms by which B hadrons decay, and the allowed transitions from one type of B hadron to another (B hadron mixing). However, in all decays, the lifetime of the B hadron is inversely proportional to the square of elements in the matrix V, as for the decay of top quarks in Equation 2.2. This is the Cabibbo-Kobayashi-Maskawa (CKM) matrix and entries in this matrix give the transition

amplitudes from one quark flavor to another via the weak force. Diagonal entries, corresponding to decays within a quark generation, have large values. Off diagonal entries, corresponding to decays between generations, are small. Since the b quark is the lightest quark in the third mass generation, the b can only decay to a different quark generation, the lifetime for which is given by off diagonal terms. The average lifetime of a B meson (a type of B hadron) is 1.6×10^{-12} seconds [50]. While the B hadron has a longer lifetime than the top, its lifetime is short enough that its decay occurs before it interacts with the CMS detector. Jets from b quarks can be identified (known as b tagging), which will be discussed in Section 5.3.4.

2.3 Chapter highlights

The Standard Model describes the interactions between particles and the probability with which the interactions occur. Particle interactions are fundamentally statistical. The known particles in the Standard Model are either quarks, leptons, or force carriers. Isolated quarks are never observed in detectors because of color confinement and are instead observed as jets. The heaviest quark, the top quark, is the only quark to decay before it hadronizes, but its lifetime is so short that only top quark decay products are observed in detectors. The top quark decays to a bottom quark and a W boson. The W boson decays to either two jets or a lepton and neutrino. We search for top quark pairs in the muon+jets channel, in which one W decays to a muon and muon neutrino while the other W decays to two jets. The bottom quark hadronizes to a B hadron, which has a long enough lifetime to

produce decay products that are displaced from the original interaction. Observing displacement of particles inside a jet allows identification of the jet as potentially originating from a b quark. The top quarks studied in this dissertation are observed by the CMS detector and produced by the collision of protons at the LHC.

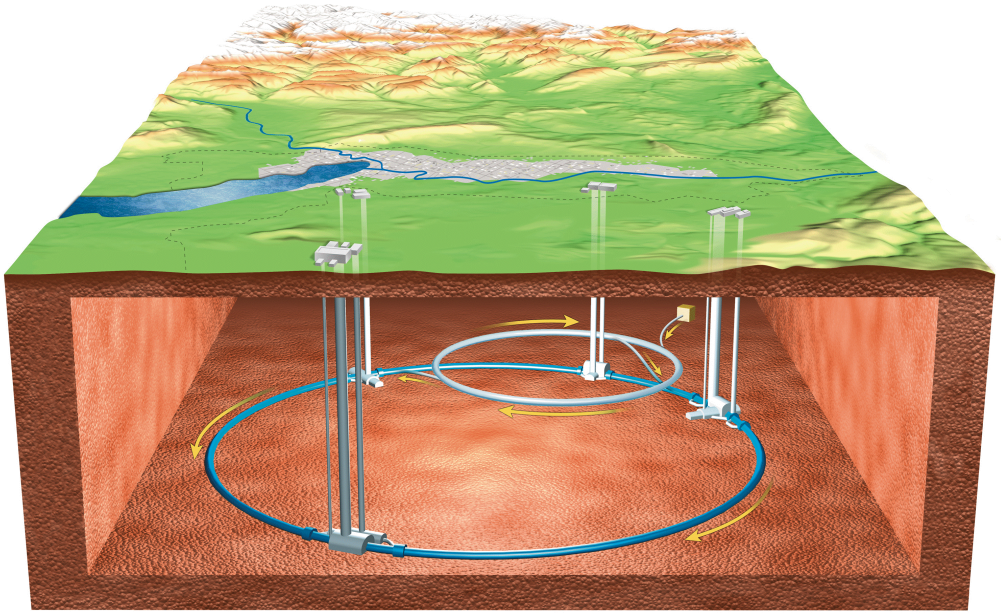
Chapter 3

Experimental apparatus

The CERN accelerator and detector complex is located in Geneva on the border between Switzerland and France. The LHC is 27 kilometers in circumference, buried with depth varying from 50 to 175 meters. Four detectors lie on the LHC ring including CMS. See Figure 3.1. The tunnel used by the LHC was originally constructed from 1983 to 1988 for the Large Electron-Positron Collider, which operated from 1989 to 2000. LHC construction began in 2000 and the LHC first circulated beam on September 10, 2008. The first 7 TeV collisions at the LHC occurred on March 30, 2010.

Construction of the surface buildings for CMS began in 1998; the CMS cavern was built from 2000 to 2005. Design and construction of the CMS subdetectors began in the late 1990s at various institutions around the world. CMS was assembled in slices at CERN on the surface, then lowered and integrated together in the cavern. Assembling CMS in slices allowed cavern and detector construction to occur in parallel. It also enabled ease of maintenance inside the cavern. The location of the CMS cavern required innovative civil engineering to prevent collapse or flooding [31]. The substrate was a mixture of sand and gravel with two water tables followed by a soft sandstone. While the shafts were dug, liquid nitrogen was injected to freeze the water. The cavern itself was incrementally excavated, where each small excavation

was reinforced with sprayed concrete. The cavern walls were designed as a large support structure of concrete up to 4 meters thick, embedded with steel bars.



(a)

Figure 3.1: The SPS, the LHC, and the experimental caverns. The small silver ring is the SPS and the large blue ring is the LHC, both described in Section 3.1. The experimental cavern located opposite SPS on the LHC is for the CMS detector. Credit: [13].

3.1 The Large Hadron Collider

The CERN accelerator complex is designed to accelerate protons to high energy and collide them as frequently as possible. To that end, components of the complex perform several primary functions: accelerating the protons in the direction of the beam axis, z, grouping the protons into bunches in z, steering the protons

in circular orbits, and focusing the protons in the plane transverse to z. Radio frequency (RF) cavities are used to accelerate and bunch the protons, while magnets are used to steer and focus the protons.

Accelerating protons to the energy of the LHC requires several intermediate accelerators at CERN, which are listed in Table 3.1. Since protons are charged, the Lorentz force equation, $\vec{F} = q\vec{v} \times \vec{B}$, can be used to determine the needed strength of the magnetic field to contain particles of a given velocity within the radius of each accelerator in the chain. As the kinetic energy of the protons increases within a single accelerator, the magnetic field produced by the magnets in the accelerator increases to match. The minimum and maximum energy of an accelerator is given by its radius and the minimum and maximum strengths of its magnets, where each accelerator in the chain has a larger radius than the previous. Once protons reach the maximum energy within one accelerator, they are injected into the next accelerator in the chain. The protons finally reach an energy of 3.5 TeV in the LHC. Two beams of protons are collided, yielding a center of mass energy of 7 TeV.

The various components of the accelerator chain are shown in Figure 3.2. The duoplasmatron uses an electron cathode to ionize hydrogen gas, producing isolated protons. The protons, a plasma at this stage, are accelerated through charged grids and focused into a beam using quadrupole magnets. Protons leave the duoplasmatron with an energy of 100 keV. The Linac 2 primarily increases the proton energy using voltage differentials applied via alternating current. Since the alternating current would ordinarily result in no net gain in the energy of the protons, the sections of the Linac 2 that would cause the protons to accelerate backwards

Accelerator	Maximum energy
Duoplasmatron	100 keV
Linac 2	50 MeV
Proton Synchrotron Booster	1.4 GeV
Proton Synchrotron	25 GeV
Super Proton Synchrotron	450 GeV
LHC	7 TeV

Table 3.1: The accelerator chain at CERN and the maximum designed beam energy of each accelerator. The LHC is designed for 7 TeV energy per beam, but is currently operating at 3.5 TeV per beam.

are shielded. Protons leave the Linac 2 with an energy of 50 MeV, an energy gain of three orders of magnitude.

The proton synchrotron booster, proton synchrotron (PS), and super proton synchrotron (SPS) accelerate and bunch the protons through the use of radio frequency standing waves. Similar to the Linac 2, these standing waves yield both positive and negative acceleration. Unlike the Linac 2, the protons are not shielded from the portion of the wave that causes negative acceleration, which induces the protons to bunch. Protons that are forward of the future bunch location are pushed back by negative potentials while protons behind the bunch location are pushed forward by positive potentials. The booster forms single bunches that are released at intervals matching the harmonics of the PS, which subsequently splits each sin-

gle bunch into smaller bunches. Ultimately, bunches of protons leave the PS at a maximum rate of one bunch every 25 ns, though 2010 operation was at intervals of 100 ns or more.

The energy of a single proton dictates its radius inside the beam pipe and, therefore, the time of arrival of the proton at each accelerating cavity. The waves in the cavities are also time varying, delivering large potential "kicks" when low energy particles arrive at the cavity and smaller potential kicks when high energy particles arrive. The process is termed phase stability, discovered by Edwin McMillan. The booster injects protons into the PS with 1.4 GeV of energy, the PS injects protons into the SPS with 25 GeV of energy, and the SPS injects protons into the LHC with 450 GeV of energy. Including warm up and calibration, total injection time into the LHC is about 16 minutes.

The LHC can ramp proton energy from 450 GeV to a design maximum of 7 TeV, though the LHC is currently operating at a maximum energy of 3.5 TeV per beam. LHC magnet ramp up and ramp down takes approximately 20 minutes, yielding a theoretical total turnaround time between periods of beam ("fills") of 70 minutes. The actual turnaround time in 2010 averaged roughly 7 hours. In the LHC, dipole magnets steer the protons while quadrupole magnets steer and focus the protons. As in classical optics, a single focusing quadrupole will focus the beam on one axis perpendicular to the beam motion while defocusing the beam on the orthogonal axis. However, also as in classical optics, a pair of quadrupoles placed at different orientations yields net focusing on both axes. Radiofrequency cavities placed near the dipoles increase the energy of the protons. A schematic of the LHC

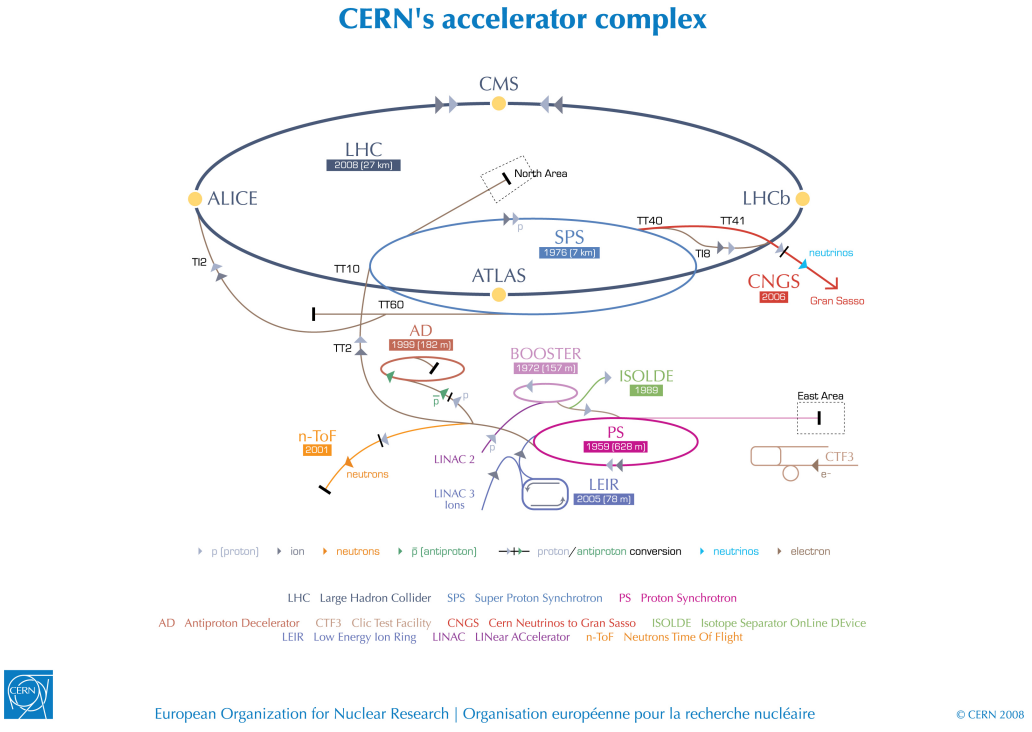


Figure 3.2: The CERN accelerator complex. Notable components are the Linac 2, the booster, the PS, the SPS, and the LHC. Credit: [13].

dipole magnets is shown in Figure 3.3. There are 1232 dipole magnets of length 15 meters and 392 quadrupole magnets of length 5 to 7 meters at the LHC. There are more than 9300 magnets at the LHC in total, though the dipole and quadrupole magnets constitute most of the length of the LHC, which is 27 km in circumference. Most of the magnets are held at low temperatures (less than 2 Kelvin) to enable superconductivity in the wires that create the magnetic field.

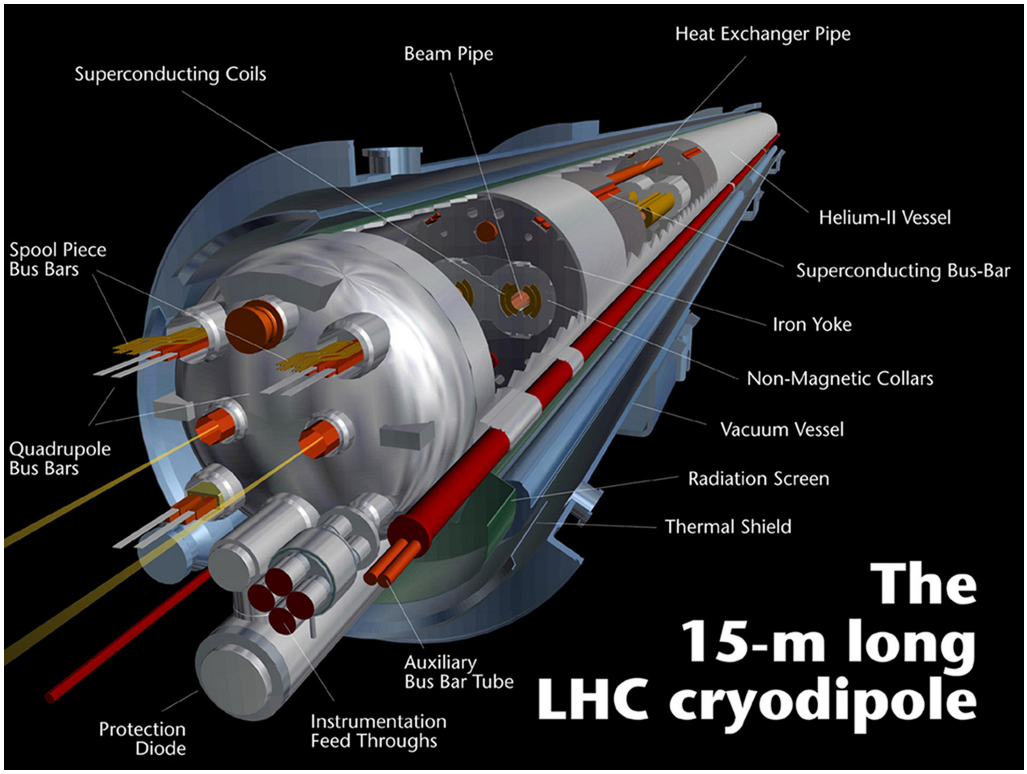


Figure 3.3: The LHC dipole magnet. Credit: [13].

The periods of time when a fixed number of bunches are circling in the LHC ring are termed fills. A fill ends when collisions cause the bunches to deform outside allowed variations, or when the number of protons in a bunch is depleted below an allowed level. At this time, the beam is intentionally "dumped" into an absorbing

material, the magnets are ramped down, and the injection process starts again. At peak operational energy and luminosity, the dump material must be capable of absorbing 200 MJ of energy, equivalent to stopping a train traveling at 200 km/hour in 80 millionths of a second [69]. Beam dumps are accomplished using a dilution magnet, which defocuses the beam by a factor of 100,000, and two cylindrical graphite blocks eight meters long and one meter in diameter, encased in concrete. During beam dump, portions of the graphite can reach 700° C.

Luminosity is a numerical measure that can be related to the expected number of collisions of protons. The rate at which collisions occur at an interaction point depends on the speed of the protons, the frequency of bunches circling the ring, the number of bunches, the number of protons in a bunch, the shape of the bunches, and the angle at which the beams cross. Specifically, the luminosity delivered by the LHC is given by:

$$\mathcal{L} = \frac{\gamma f k_B {N_p}^2}{4\pi\epsilon_n\beta^*} F \tag{3.1}$$

where γ is the Lorentz factor $(1/\sqrt{1-(v/c)^2})$, f is the revolution frequency, k_B is the number of bunches, and N_p is the number of protons per bunch [50]. The parameters ε_n, β^*, and F are related to the design of the accelerator (and, therefore, the shape of the bunches) and the angles of the beams at the interaction point at the detector.

Production cross section can be thought of as the probability of a particular type of event occurring, although cross section is a value with units of area. Since the probability of one object in motion striking another object (at rest or in motion)

is proportional to the cross sectional areas of each object, the probability of particle interactions can be characterized in units of area. Particle cross section is measured in barns (an interaction rate of a barn is very large, or "as big as a barn"), where 1 barn is equivalent to a cross sectional area of 10^{-24} cm^2. For comparison, the Bohr radius, which is the radius of an electron in the lowest energy state of a hydrogen atom, is 5.3×10^{-11} cm, which corresponds to a cross sectional area of 8.8×10^{-21} cm^2. The radius of the hydrogen nucleus, a proton, is 0.8×10^{-13} cm, yielding a cross sectional area of 2×10^{-26} cm^2.

Luminosity is measured in units of inverse barns per second, b^{-1}s^{-1}. The number of collisions from a particular type of particle interaction with production cross section σ is related to the luminosity by:

$$N = A\epsilon\sigma \int \mathcal{L}dt \tag{3.2}$$

where A refers to acceptance and ϵ to efficiency. Acceptance is a geometric property of the particle detector. While CMS is designed to be nearly hermetic, some particles go down the beam pipe or between cracks in the subdetectors. Efficiency is determined by the cuts used by the analyst, where the events with desired characteristics (signal) are selected from all available collisions (which contain both signal and background). Most selection cuts are designed to improve the purity of the final sample while retaining efficiency. For event selection, purity is defined as the number of signal events selected divided by the total number of events selected. Efficiency is defined as the number of signal events passing selection divided by the number of signal events prior to selection.

3.2 The Compact Muon Solenoid detector

CMS is a 12,500 metric ton instrument. It is 20 meters in length and 14 meters in height and width; it has of order 100 million digital channels. It can be likened to a 100 megapixel digital camera, four stories tall and wide by six stories long. CMS is composed of several subdetectors, each recording various properties of particles as they traverse the detector. The subdetectors come in two forms: tracking detectors and calorimeters. Tracking detectors record the passage of charged particles through their layers by measuring the ionization of the material in the detector as voltage peaks, termed "hits". Tracking detectors measure particle momentum via the radius of curvature of the particle as it propagates in the magnetic field produced by the solenoid. Calorimeters operate primarily on the principle of absorbing all the energy of the particle within the medium of the calorimeter, intentionally inducing the particle into many collisions that cascade into more collisions via a particle "shower". To induce these showers, calorimeters are composed of dense materials. Figure 3.4 is a diagram of CMS, which can roughly be described as concentric barrels with endcaps, where particles collide in the center of the detector. Going radially outward from the beam line, the subdetectors are the inner tracker, the electromagnetic calorimeter (ECAL), the hadronic calorimeter (HCAL), the solenoid, and the muon chambers (or outer tracker). A complete technical description of the CMS detector can be found in [16]. A summary is presented in this chapter.

Since different detectors are designed to interact with different particles, identification of particle type can be performed by determining the subdetectors with

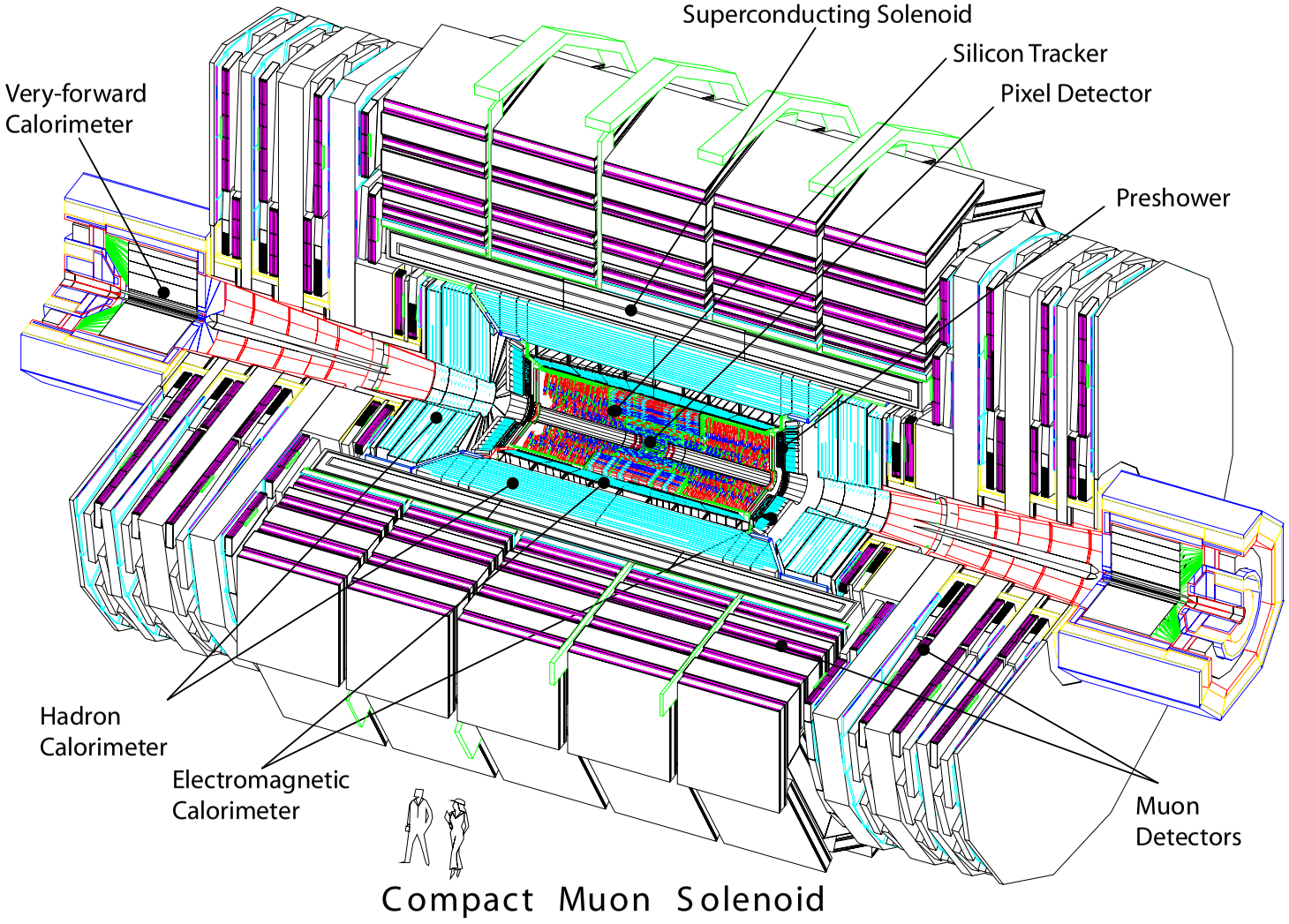


Figure 3.4: The Compact Muon Solenoid. Credit: [16].

which a particle interacts. Examples are shown in Figure 3.5. Electrons, as charged particles, leave hits in the inner tracker, then shower in the electromagnetic calorimeter. Photons are uncharged and thus do not leave hits in the inner tracker, but shower in the electromagnetic calorimeter. Of note here is that photons may interact in a process called pair production with the material in the inner tracker to produce an electron and positron (antielectron), which leave hits in the inner tracker, potentially leading to reconstructed particle tracks. However, if the pair production occurs outside the innermost layers of the tracker, particle tracks will not be reconstructed due to the lack of inner hits. Jets from quarks and gluons are visible in both the tracker and hadronic calorimeter (some particles in the jet are charged, so interact with the tracker). Electrically neutral hadrons, such as neutrons, do not leave tracks in the inner tracker, but do leave energy in HCAL. Finally, muons leave hits in the inner tracker and the outer tracker because they are charged, but they do not interact hadronically with HCAL. Because the muon is heavier than the electron, its electromagnetic radiation length is considerably longer, so the muon rarely leaves significant energy deposits in ECAL.

Cylindrical coordinates are useful to describe the geometry of CMS, with the z axis oriented with the beam line and the $r-\phi$ plane (the transverse plane) perpendicular to the z axis. High energy physicists use an additional coordinate, pseudorapidity, η, defined as $-ln(tan(\theta/2))$, where θ is the angle measured from the z axis in spherical coordinates. Rapidity, $\frac{1}{2}ln(\frac{E+p_z}{E-p_z})$, is defined such that the distribution of all particles in the lab reference frame is nearly flat in rapidity space. To achieve this, rapidity must be a function of the energy and momentum of the particle. While

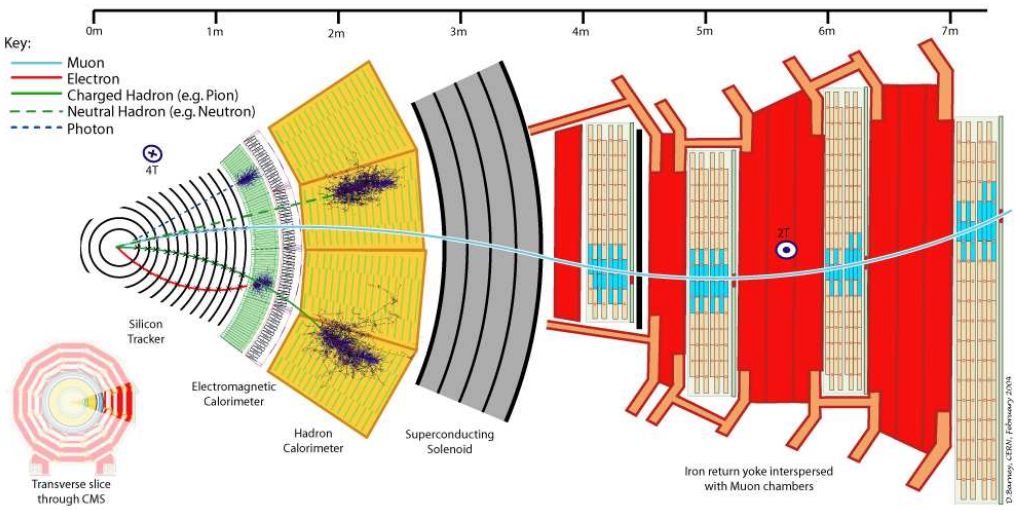


Figure 3.5: Transverse view of CMS with different types of particles and the detectors with which they interact. Credit: [13].

experiments directly measure particle momentum, extrapolation to particle energy requires knowledge of the particle mass, which is not directly measured by experiments. Rapidity is therefore not especially useful as a generalized coordinate in experiments. Pseudorapidity is defined by θ, making it an experimentally useful coordinate and, in the approximation that a particle is massless, is equal to rapidity. Thus, low mass particles have a nearly flat pseudorapidity distribution. For massive particles, a portion of the beam energy must be used to create the mass energy of the particle, reducing the boost in the z direction. Massive particles are produced with a non-uniform distribution in pseudorapidity, skewed in favor of smaller values, seen in Figure 3.6. The range of η values for CMS can be seen in Figure 3.7.

Another commonly used coordinate variable is ΔR, defined as $\sqrt{\Delta\phi^2 + \Delta\eta^2}$, which is similar to angular separation with an additional relativistic component

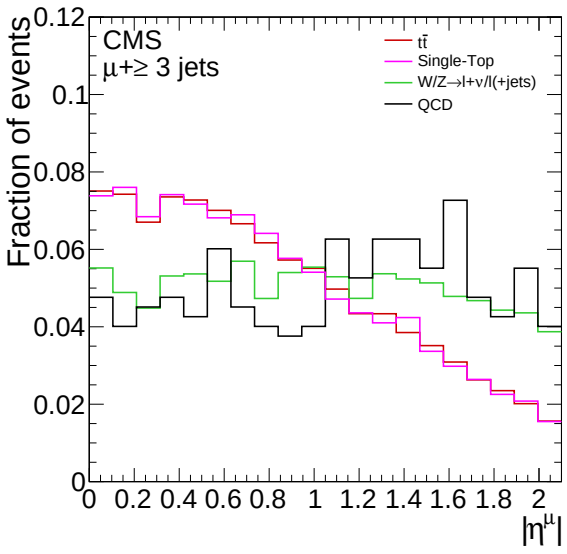


Figure 3.6: The $|\eta|$ distribution of a reconstructed muon originating from the massive top quark (red and pink) and lower mass W or Z bosons (green) and QCD (black). The distributions are normalized to show shape comparisons.

from η. Additionally, because particles deposit energy in measured locations in the calorimeters, energy is treated as a directional vector in particle physics. Specifically, the variable E_{T} refers to the energy deposited in the plane transverse to the beam axis.

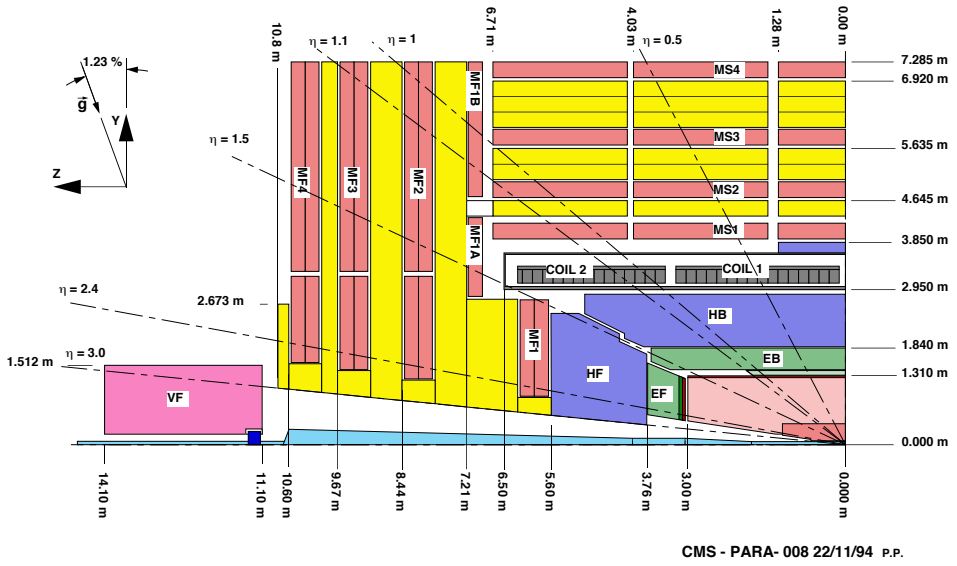


Figure 3.7: Quartered longitudinal view of CMS with values of $|\eta|$ shown as dashed lines, r values on the right edge, and z values on the lower edge. Credit: [16].

3.2.1 Superconducting solenoid

The superconducting solenoid produces a 3.8 Tesla field. By comparison, the Earth's magnetic field ranges in strength from $30 - 60$ μT. The magnetic field is needed to measure the momentum of charged particles as they traverse the inner

and outer tracking detectors.

The 13 meter long, 6 meter diameter solenoid is kept at superconducting temperatures using liquid helium. The solenoid is kept inside a vacuum tank, providing thermal insulation from the rest of the detector. Metal is conductive because the electrons in the outer orbits of a metal nucleus are energetic enough to move freely between the lattice of atoms in the metal. Conductivity, the ability of electrons to move freely, increases in a metal as it is cooled because the metal nuclei vibrate less as they get colder, reducing the rate at which electrons interact with the nuclei. However, even at absolute zero, the classical resistance of any imperfect metal cannot be zero due to interactions of the electrons with the impurities. Nonetheless, all metals possess a critical temperature, below which electrical resistance goes to zero. This quantum mechanical effect is still not fully understood, but is currently best modeled by the BCS theory [5]. The BCS theory posits that under certain conditions, it becomes energetically favorable for electrons to bond into Cooper pairs. While electrons repel due to their electrical charge, they also attract the positive nuclei forming the metal lattice. An electron distorts the nearby lattice and increases the density of positive charges, attracting other free electrons. The electrons bond in an energetically stable pair if the lattice does not have sufficient kinetic energy (temperature) to separate them. When this occurs, the Cooper pair does not experience any resistance from the lattice as the kinetic energy in the lattice is now bound internally to Cooper pairs; thus, the metal becomes superconducting.

Iron is interleaved with the muon chambers to shape the magnetic field from the solenoid. Magnetically permeable materials have low magnetic resistance, sim-

ilar to electrical resistance. Because iron is magnetically permeable, the magnetic field lines are in a lower energy state when inside the iron. Iron return yokes outside the solenoid are used to shape the magnetic field, making the fields lines parallel to the z axis with little divergence. The solenoid and the iron in the return yoke are the heaviest components of CMS. A picture of the solenoid and portions of the return yoke are shown in Figure 3.8.

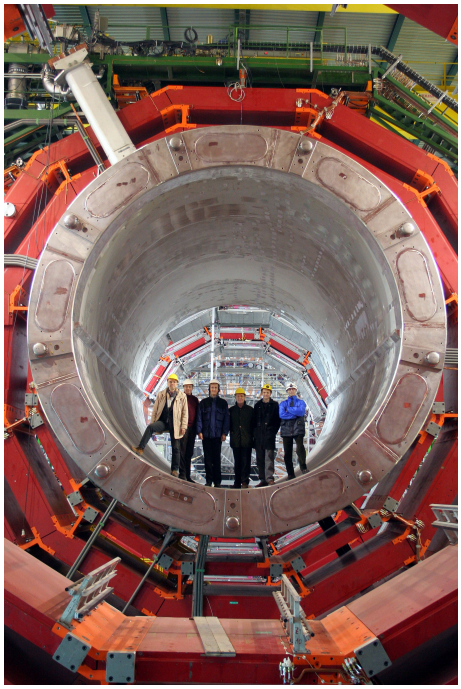


Figure 3.8: End view of the solenoid and portions of the iron return yoke at CMS. Credit: [16].

3.2.2 Inner tracker

Closest to the beam pipe is the inner tracking detector, which is composed of silicon pixel and strip sensors. The pixel sensors measure particle position with the

highest precision of all the sensors. They are placed close to the interaction region, which has a high particle density per unit area. The inner tracker measures the trajectory of all charged particles by the ionization of atoms in the silicon composing the tracker. The freed electrons in the silicon flow down the potential gradient from voltage applied to the end of each sensor. Each sensor then measures the change in voltage due to the arrival of the ionized electrons at the end of the sensor.

There are approximately 66 million pixel sensors and 9.6 million strip sensors. Pixels have detecting surfaces of dimension 150 by 100 μm, whereas strips have detecting surfaces with sizes that vary depending on position, from 80 to 180μm in one dimension and from 10 to 25 cm in the other dimension. The layout of the tracking detector is shown in Figure 3.9, with individual strip sensors drawn as blue or red lines (pixel sensors are not drawn), indicating single sided or glued layers. Glued layers are two strips in close proximity, one at zero angle with respect to the z-axis, the other at a slight angle, improving the resolution of measuring particle position.

3.2.3 Electromagnetic calorimeter

The electromagnetic calorimeter, shown in Figure 3.10, is designed to induce interactions with electromagnetic particles such as the electron and photon. Since ECAL is made of dense lead tungstate crystals, the electromagnetic showers are completely contained in the body of ECAL. The shower excites the atoms in the crystal, which release the stored energy as scintillation light, which is recorded by

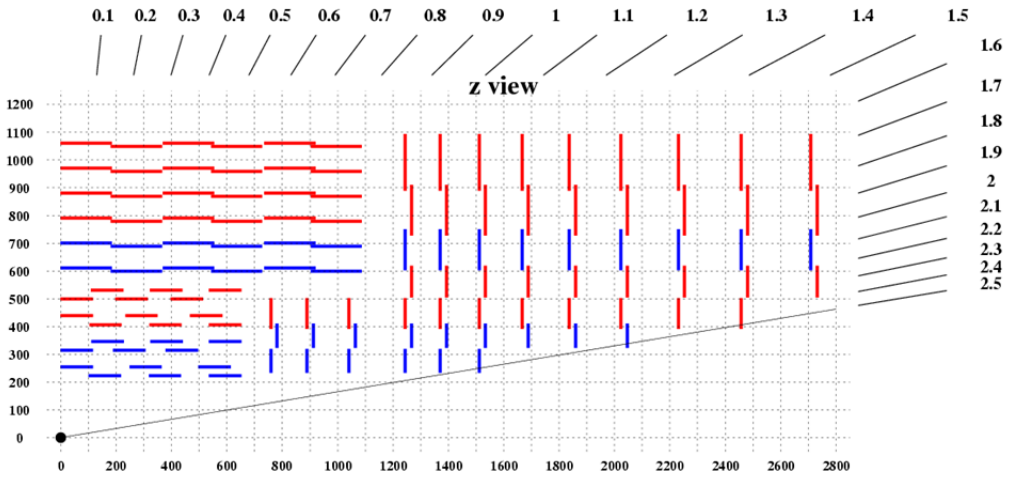


Figure 3.9: Quarter longitudinal view of the strip layers in the inner tracking detector at CMS. Blue layers are made of glued orthogonal strips while red layers are composed of single strips in one orientation. The marks drawn on the top and right edges are values of $|\eta|$. Credit: [16].

photodetectors.

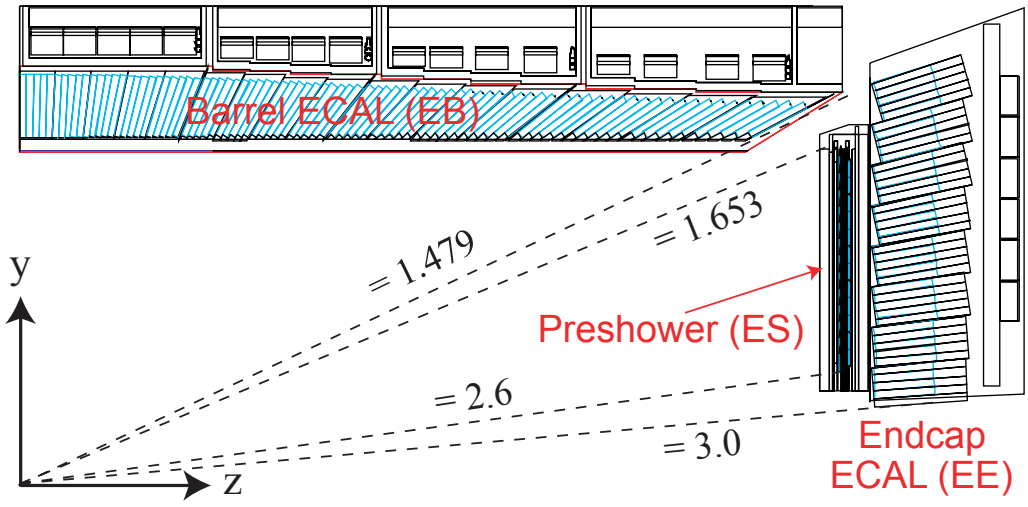


Figure 3.10: Quartered longitudinal view of the electromagnetic calorimeter at CMS, composed of lead tungstante crystals. Dashed lines correspond to values of $|\eta|$. Credit: [16].

In an electromagnetic shower, high energy electrons and photons interact with the atoms in the lead tungstate crystals, producing more electrons and photons. In the Bremsstrahlung process, a high energy electron interacts with the electromagnetic field of an atom and subsequently releases a free photon. In the process of pair production, a high energy photon produces an electron and positron that interact with the atom. In one radiation length, the particles lose $1/E$ of their energy. Note that radiation length describes the change in total energy of all the particles in the shower. The individual shower particles do not lose their energy smoothly, but in discrete amounts in every interaction. The electromagnetic shower continues until the individual electrons and photons no longer have sufficient energy to interact with

the atoms via the process of Bremsstrahlung radiation or pair production. Although muons interact electromagnetically, the radiation length in a material for a particle of mass m is proportional to m^2. The muon is 200 times as massive as the electron and has a radiation length 40,000 times larger than the electron. Since muons are $\mathcal{O}(10^4)$ times less likely to shower in ECAL, assuming muons do not shower in ECAL during event reconstruction is an excellent approximation.

During the formation of the electromagnetic shower, the electrons in the crystal atoms are excited to higher energy states. Because the atoms are not in their lowest energy state, the electrons eventually release their energy via scintillation photons. These photons have much lower energy than the photons in the shower and are released after the shower has passed, though lead tungstate releases about 80% of the scintillation light within the 25 ns of time between bunch crossings.

In the barrel, the scintillation light is directed using fiberoptic cables to avalanche photodiodes. The photons ionize the silicon atoms of the photodiode, converting light into electricity via the photoelectric effect. The electrons are then induced into an avalanche process of secondary emissions with a voltage potential accelerating the electrons towards the end of the photodiode. As each electron is accelerated, it strikes subsequent layers in the photodiode, ionizing the silicon, producing more electrons. The photodetectors used in the endcaps are vacuum phototriodes, which operate on the same principle as avalanche photodiodes. Because the magnetic field in the endcaps is stronger than the barrel, the phototriodes have fewer internal layers and are composed of different materials.

The photodetectors are designed to measure only the scintillation light from

the crystals in the calorimeter. Although the photodetectors are small and the shower energy is mostly dissipated in the volume of ECAL, shower components occasionally interact directly with the photodetectors, inducing a very energetic avalanche. Since this causes one photodetector to spike with a voltage change far in excess of its neighboring photodetectors, this occurrence can be readily detected and handled [19].

3.2.4 Hadron calorimeter

The hadronic calorimeter is designed to induce interactions with hadronic particles. The HCAL is primarily composed of brass interleaved with plastic scintillator. The innermost and outermost layers of the HCAL in the barrel are made of stainless steel for structural reinforcement. The high density brass and steel cause many interactions and induce a hadronic shower. The interleaved plastic scintillators have attached photodetectors to measure the energy of the shower as it develops. Hadronic showers develop similarly to electromagnetic showers, except the interaction of the shower hadronic particles with the atoms is primarily via the strong force. Since the interactions are via the gluon, hadronization also plays a role in hadronic shower development. While the crystals in ECAL act both as scintillators and as the dense interaction material, brass and steel are not transparent materials and must be interleaved with scintillating plastic. The HCAL is termed a sampling calorimeter, which tend to have reduced energy precision as only part of the shower is measured.

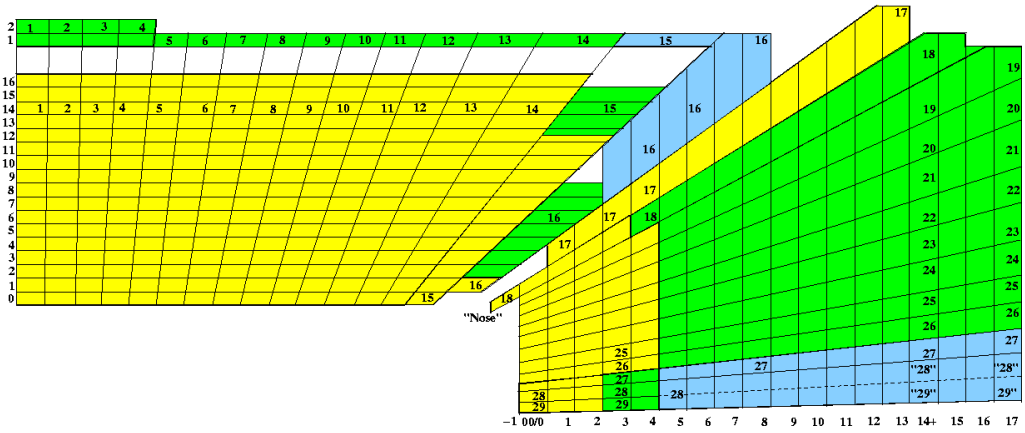


Figure 3.11: Quartered longitudinal view of the hadronic calorimeter at CMS, a sampling calorimeter. Towers are numbered from 1 to 29 and are composed of 17 active readout layers in the barrel and 19 layers in the endcaps. Each readout layer is composed of plastic scintillators and a dense material, primarily brass. Credit: [16].

3.2.5 Muon chambers

Nearly all products of the interaction are absorbed by ECAL or HCAL, leaving mostly muons remaining, earning the outer tracker the appellation "muon chambers". The muon chambers function similar to the inner tracker. However, since the muon chambers are farther from the interaction than the inner tracker, less precision is required when measuring the position of particles as they traverse this subdetector. Therefore, the muon chambers are composed of various types of gaseous chambers instead of silicon. As the muons travel through the gas, they ionize the gas atoms and the freed electrons are drawn towards wires via an applied potential.

There are three types of chambers in the muon chambers, shown in Figure 3.12: drift tubes, resistive plate chambers, and cathode strip chambers. Drift tubes are

used in the barrel, resistive plate chambers are used in both the barrel and endcaps, and cathode strip chambers are used in the endcaps. Different systems are used in different regions of η due to the varying radiation environments. Since low mass particles are distributed roughly equally in η and η is defined as $-ln(tan(\theta/2))$, detector components close to the beam are exposed to radiation that grows exponentially as $|\theta|$ decreases. The large muon chambers in Figure 3.12 are composed of brick-like layers of smaller gaseous chambers.

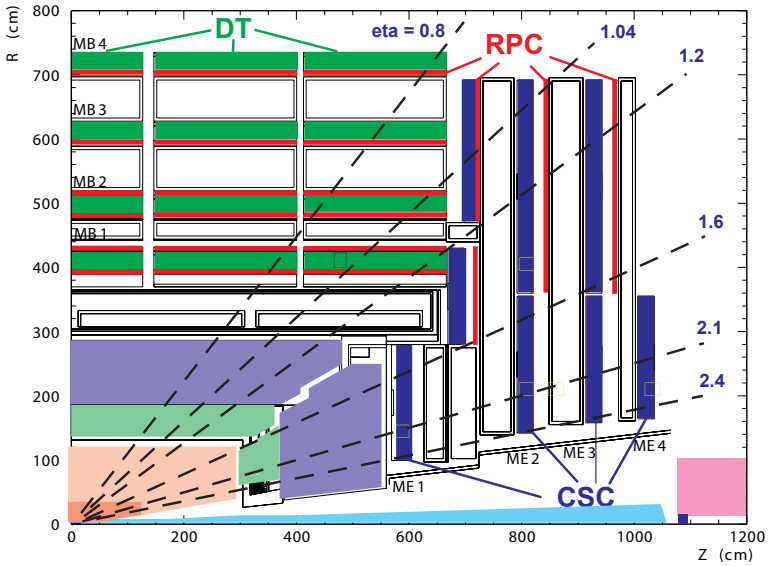


Figure 3.12: Quartered longitudinal view of the muon chambers at CMS. The three types of gaseous cambers, drift tubes, resistive plate chambers, and cathode strip chambers are labeled. Boxed white regions correspond to the iron magnetic return yoke and the solenoid. The HCAL is shadowed in purple/blue, the ECAL in green, and the inner tracker in orange/red. Credit: [16].

Drift tubes (DTs) measure the location where a charged particle passed through

the gas by measuring the elapsed time from the ionization to the time a voltage pulse is observed due to the arrival of the electrons at the wire. The ionized electrons drift at a known speed through the gas, giving the distance from the point of ionization to the wire. However, given the size of the individual chambers and the drift speed of the electrons, the elapsed time can be longer than the time between collisions. Therefore, the DTs cannot unambiguously determine the collision to which a voltage peak corresponds. Instead, the resistive plate chambers are used for this purpose.

Resistive plate chambers (RPCs) act much like avalanche photodiodes, where multiple layers of metal are surrounded by gas, developing an avalanche from the incident particle. Unlike photodiodes, the incident particle is not expected to be a photon, rather, the resistive plate chambers have a large surface area designed to directly interact with the muon. The muon ionizes an atom in the metal of the RPC and an electromagnetic avalanche develops. Electromagnetic avalanches develop quickly, so the resistive plate chambers determine the collision that created the muon. Due to their large surface area, the RPCs have poor position resolution, but the DTs can accurately measure position once the correct collision is provided by the RPCs.

The cathode strip chambers (CSCs) operate similarly to the drift tubes and also use the resistive plate chambers to determine the correct collisions. Due to the increased density of particles at high η, the CSCs have smaller gas chambers than the DTs, where each segment is composed of 7 panels of cathode strips with 6 gas gaps. Although the CSCs span an $|\eta|$ up to 2.4, the RPCs were not installed past $|\eta| = 1.6$ for initial low luminosity running. Therefore, the muon chambers are used

only to an $|\eta|$ of 1.6, but are capable of going up to an $|\eta|$ of 2.4.

3.2.6 Luminosity monitors

For limited ranges of η and transverse energy, there is a linear relationship between the average transverse energy deposited in HCAL and the number of collisions per unit time. A portion of the HCAL detector is located at very forward angles. Using pseudorapidity range $3.5 < |\eta| < 4.2$, the maximum transverse energy is kinematically limited to a few hundred GeV. By averaging the transverse energy deposited in sections of HCAL in this pseudorapidity range, the number of collisions can be approximated [21]. Some of the beam parameters in Equation 3.1 are known, giving the effective cross-sectional area of the bunches as they collide in CMS, $\sigma_{\text{eff}}(x)$ and $\sigma_{\text{eff}}(y)$. This is then related to the instantaneous luminosity, $\mathcal{L}_0$, i.e.:

$$\mathcal{L}_0 = \frac{\gamma f k_B {N_p}^2}{2\pi\sigma_{\text{eff}}(x)\sigma_{\text{eff}}(y)} \propto \langle E_{\text{T}} \rangle \tag{3.3}$$

The proportionality constant between $\langle E_{\text{T}} \rangle$ and $\mathcal{L}_0$ can be measured precisely during LHC runs specifically designed to measure luminosity. In normal data taking runs, collisions deposit energy in the forward HCAL detectors. The proportionality constant is then applied to the measured energy, yielding the instantaneous luminosity delivered by the LHC to CMS during periods of data taking. The measured luminosity has an uncertainty of 4% and is shown for the 2010 data taking period in Figure 3.13.

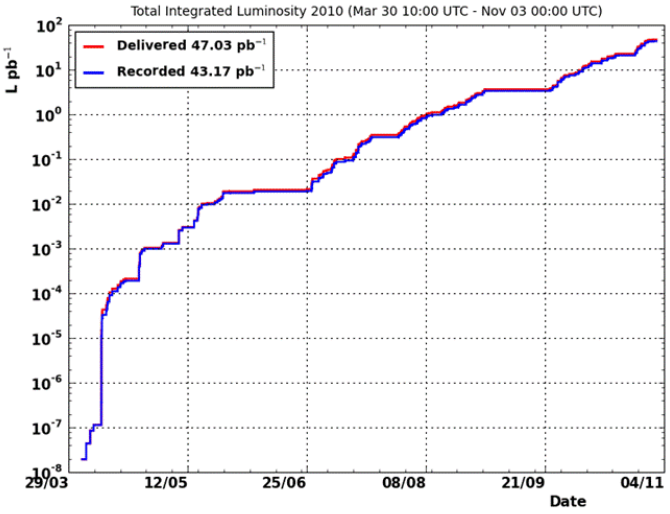


Figure 3.13: Delivered and recorded luminosity at CMS in 2010. Credit: [43].

3.2.7 Event triggers

Not all proton collisions involve unusual or physically 'interesting' interactions. Therefore, protons are collided at a much higher rate than they can be recorded. A trigger system is utilized which records an event if it appears to contain a desirable interaction. To record a baseline of typical collisions, some triggers are designed to accept events with minimal activity, termed minimum bias. Since the expected rate of these events is high, events passing only minimum bias criteria are accepted a fraction of the time ("prescaled"). The LHC is capable of delivering collisions at a rate near 10^7 Hz, or ten million collisions each second. The computing farm located at CERN is capable of reconstructing events at a rate of 150 Hz [15]. Therefore, triggers must reduce the rate of accepted events by five orders of magnitude. CMS uses a dual trigger system, termed level-1 (L1) and high level trigger (HLT). L1

electronics are usually co-located with the detector electronics while the HLT is encoded primarily in configurable software deployed on a computing farm.

Since this analysis searches for $t\bar{t}$ events in the muon+jets channel, we utilize events passing muon triggers. We use a single muon L1 trigger that requires a muon with p_{T} greater than 7 GeV. Because the instantaneous luminosity delivered by the LHC increased during 2010 operations, we used two muon high level triggers, first requiring a muon with p_{T} greater than 9 GeV, HLT_Mu9, then requiring a muon with p_{T} greater than 15 GeV, HLT_Mu15. By increasing the p_{T} threshold for the trigger, the rate of triggering was reduced, keeping the total rate of events handed to the CERN computing farm under 150 Hz.

The L1 muon trigger reconstructs muon candidates in the DTs, CSCs, and RPCs, then combines the candidates based on physical proximity. The L1 muon candidates are ranked based on detector type, p_{T}, quality, and η. The four best candidates with p_{T} greater than 7 GeV are passed to the muon HLT. The muon HLT uses the standalone muon reconstruction described in Section 5.3.2 to select muons.

3.3 Chapter highlights

The Large Hadron Collider currently accelerates and collides protons at a center of mass energy of 7 TeV. The number of proton collisions is characterized by luminosity. The expected number of collisions producing top quark pairs is given by $N = A\epsilon\sigma \int \mathcal{L}dt$, where σ is the production cross section of top quark pairs at the

LHC. Production cross section is given in units of barns, a unit of area related to the probability of interaction, while integrated luminosity is given in units of inverse barns.

Proton collisions from the LHC are observed by the CMS detector, amongst others. CMS is composed of the inner tracker, ECAL, HCAL, muon chambers, and a solenoid producing a large magnetic field. Tracking detectors record the passage of charged particles through tracking layers, measuring particle position and momentum (based on the radius of curvature inside the magnetic field). Calorimeters dissipate the energy of particles in electromagnetic or hadronic showers, measuring particle position and energy. Particles are identified based on the subdetectors with which they interact. Jets primarily interact with the inner tracker and HCAL while muons primarily interact with the inner tracker and muon chambers.

Proton collisions in CMS are triggered and recorded based on the preliminary identification of particles with desired characteristics. Data in CMS is divided into collections by sets of triggers. For example, all the events passing any muon trigger are contained in one dataset. The datasets are then distributed and analyzed around the world utilizing a computing grid described in the next chapter.

Chapter 4

Grid computing

CMS is an international collaboration that produces approximately 10 PB of data every year [15]. To achieve low latency access for users located around the world, the data are hosted at computing facilities of widely varying size. The distributed resources are interconnected using grid technologies, appearing as a single resource to the user. The distribution and analysis of LHC data is enacted through the use of several computational grids, collectively called the Worldwide LHC Computing Grid (WLCG) [68]. Most WLCG computing sites located in the North American continent participate in the Open Science Grid (OSG) [53].

4.1 Open Science Grid

OSG is composed of sites providing computing resources like CPU or storage, providers of central services such as site registration and monitoring, and a software stack enabling all parties to interact. Consumers are grouped into virtual organizations and sites are associated with certain VOs, though some sites allow opportunistic access to local resources by consumers in all VOs. The largest virtual organizations in OSG are ATLAS and CMS, sharing computing resources with each other and with other virtual organizations. Computing facilities participating in OSG run software in the Virtual Data Toolkit (VDT) [56], composed of numerous

subpackages. OSG itself provides site registration and monitoring services that can be used to determine the availability of computing resources.

4.1.1 Virtual organizations

A virtual organization (VO) is a group of individuals and/or institutions with shared rules or contracts governing resource use and provision. Physicists, computing facilities, and computing resource administrators related to the CMS experiment form the CMS VO. The CMS and ATLAS VOs are the largest consumers (and providers) of OSG resources, where the weekly number of hours spent computing for different VOs is shown in Figure 4.1. A computing grid benefits from the presence of multiple VOs because the resource usage patterns of each VO are often different. Each computing facility has contractual requirements regarding the proportion of computing services that must be available to members of each VO. Some computing facilities may be wholly owned by one VO, while others are owned by multiple VOs. If the user consumption of resources in one VO fluctuates low, members of other VOs are allowed opportunistic access to resources ordinarily owned by the low-consumption VO. Sharing resources amongst multiple VOs allows institutions to purchase fewer computing resources than the expected maximum demand of one VO. Instead, computing resources can be purchased as a function of the demand of all VOs, which tends to vary more smoothly in time than the demand of a single VO.

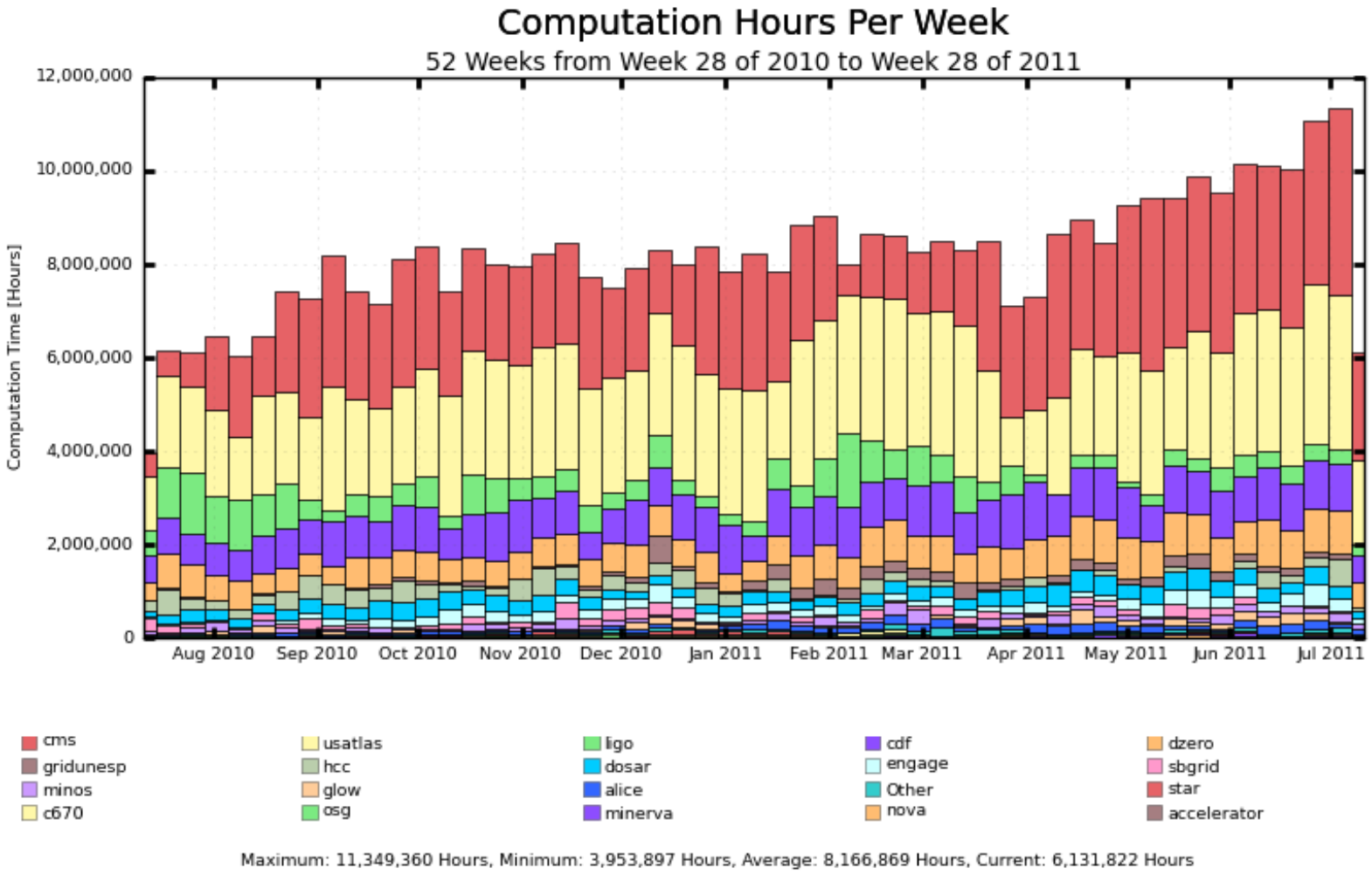


Figure 4.1: Weekly number of consumed computing hours in OSG from July 2010-2011, shown by virtual organization. Credit: [53].

4.1.2 Authentication

Authentication in OSG is performed using X.509 certificates [42]. The X.509 standard utilizes a public key infrastructure, in which a certificate contains a public key and private key. Public keys are widely distributed to any interested party, while private keys are held only by the owner of the certificate. In public key infrastructure, messages are encrypted with the recipient's public key and can only be decrypted using the recipient's private key. Conversely, signatures are encrypted using the signer's private key and are decrypted using the signer's public key. Authentication in a grid environment is done via the exchange of signatures, where the site must sign messages to the user and the user must sign messages to the site.

Running jobs usually need to communicate with grid services, but a running job may not have access to the private key of the job owner. Therefore, a grid job contains a proxy, which is a short-lived certificate containing its own private key. The proxy contains the needed information to allow the job to emulate the original certificate, without possessing the actual private key of the original certificate. Because the proxy is short lived and because a proxy is only sent after initial authentication has been successful, it is less dangerous to send the private key of a proxy over the grid than the private key of the certificate. WLCG requires VOMS-extended proxies [11], which are proxies issued by a Virtual Organization Management Service. Upon receiving their certificate, users can register the certificate with their VO, requesting to be a member of specific groups (such as site administrator or physicist) and to be granted certain roles (such as a high priority

user). Users create VOMS-extended proxies by contacting the VO Management Service(s) affiliated with the VO and optionally specifying the group and role they want the proxy to contain.

Grid certificates are issued to individual users and sites from certificate authorities. A given user or site can choose to honor certificates issued by particular certificate authorities. Certificate authorities issue certificate revocation lists, a compilation of previously issued certificates that have been revoked for any number of reasons, including the suspected compromise of the private key. A user or site accepting certificates from an authority must ensure that their copy of the revocation list from the authority is kept up to date.

Grid certificates and proxies contain a distinguished name (DN) that uniquely identifies the user. WLCG sites authorize users to access site resources by mapping the DN in a VOMS-extended proxy to a local login account, usually a UNIX username. By mapping proxies to local accounts, specific user accounts do not have to be replicated at every WLCG site, enabling site local management of user permissions and priorities. The groups and roles specified in the VOMS-extended proxy can be used to map the user to different login accounts, which may have different privileges or priorities at the local site. Additionally, VOMS-extended proxies allow WLCG sites to accept proxies only from users in their supported VOs. Similar to certificate revocation lists, sites must keep a list of VOMS banned users up to date.

4.1.3 Compute and storage elements

Computing facilities that provide services to the OSG do so via a compute element (CE), a storage element (SE), or both. A CE provides grid access to compute cycles while an SE provides grid access to storage. The CE and, in some cases, the SE, report information to OSG services. The OSG services federate the reports and make them available to administrators and users. OSG software is built on the Virtual Data Toolkit, which packages the software needed to provide CE and SE services. The VDT also includes the needed software to contact CE and SE services. OSG software also includes a few OSG specific utilities, especially a configuration utility. OSG software configures the software in the VDT according to site administrator specifications, as well as configuring services in the VDT to contact OSG services.

A CE is strictly defined as a service that wraps a computing facility's processing cycles, allowing users employing grid protocols to authenticate and request cycles. However, for a CE to successfully service grid jobs, the CE must run a local job manager as well as advertise and monitor CE availability. Therefore, the VDT for a CE includes Globus [36] to provide grid access, Condor [46] to manage jobs locally, CE Monitor (CEMon) [12] to advertise availability, and Resource Service Validation (RSV) [54] to monitor.

Globus provides the grid service that wraps site computing cycles, as well as the needed software to contact grid services and resources. Condor manages the compute cycles at the site, scheduling jobs based on resource availability, user

priority, and historical user consumption. Although the VDT includes Condor, OSG supports other local job schedulers such as PBS and SGE; installing Condor from the VDT is entirely optional. CEMon reports information about the CE and SE (if appropriate) to OSG's Berkeley Database Information Index (BDII) service [39]. BDII information can be either static, such as CE hostname and cluster hardware, or dynamic, such as the currently available number of job slots. Users wishing to discover available resources that meet their requirements contact the BDII service. RSV also runs on the CE, self-checking the status of CE and SE (if appropriate) services. The information from RSV is presented to administrators on a site-local website and is also reported to MyOSG [47]. Users and administrators can ascertain the status of OSG resources by consulting MyOSG web pages or by contacting the MyOSG service. RSV tests include the responsiveness of the CE and SE hosts to pings, directory permissions, the timestamps of the current certificate revocation lists, the ability to service jobs, and the ability to receive files.

Similar to a CE, an SE is a service that wraps a computing facility's storage, allowing users employing grid protocols to authenticate and request storage of or access to files. An SE also requires an underlying file system and may employ storage management solutions. The VDT for an SE includes Globus, Berkeley Storage Manager (BeStMan) [9], and dCache [33]. In the case where the SE has no affiliated CE, CEmon is also installed on the SE to report the SE data to BDII.

Globus, a toolkit itself, includes an implementation of a GridFTP [2] server, as well as software to interact with GridFTP services. GridFTP is a grid-extended file transfer protocol. The Globus GridFTP server wraps the underlying filesystem;

it does not provide storage management. Some SEs will host multiple GridFTP services to improve network bandwidth. OSG sites running an SE typically install either BeStMan or dCache. The BeStMan and dCache services contact the GridFTP service(s) and are themselves contacted via the storage resource management (SRM) protocol. In other words, BeStMan and dCache wrap GridFTP, which in turn wraps the underlying filesystem. Both BeStMan and dCache provide software to interact with SEs via the SRM protocol.

In addition to providing a grid service which communicates via the SRM protocol, BeStMan and dCache provide storage management services. dCache is designed as a distributed storage system. dCache handles data stored on multiple disks and tapes (if applicable) on multiple physical nodes. dCache stages data from tape to disk as appropriate and presents a unified filesystem interface to the user. BeStMan is a lightweight storage management service requiring minimal intervention by site administrators. BeStMan supports several transfer protocols and disk based file systems. The BeStMan service can be run in gateway mode, which does not provide any storage management, operating as a simple SRM wrapper of GridFTP. Using BeStMan in gateway mode allows sites to use their own internal storage management systems.

4.2 CMS resources, services, and software

The CMS VO uses its own high level global services, software, and resources to meet the specific grid computing needs of CMS. CMS global services monitor sites

and jobs, provide information on datasets, and store the conditions of the detector as a function of periods of operation. CMS uses several specialized software packages for (1) the production of simulated data at different grid sites, (2) job scheduling and management for user created analysis and production jobs, and (3) the distribution of real and simulated data from site to site. CMS organizes resource providers by tiers, where sites in different tiers are built, funded, and contractually obligated to perform specific tasks, i.e., sites in different tiers run different types of workflows. The tiered structure primarily governs the quantity and proportion of resources dedicated to data storage, distribution, processing, and analysis.

Individual sites have different hardware configurations and operating system software packages installed. Therefore, CMS uses a software platform (CMSSW) that makes event processing portable from site to site. CMSSW provides the framework for event simulation, reconstruction, and analysis; CMSSW is discussed in the next chapter. CMSSW jobs are sent to the grid sites hosting the dataset requested by the job or, for simulations, to any site with available resources.

4.2.1 CMS global services

CMS uses the Dashboard [45] to monitor user jobs and sites, which presents the information on a user friendly website. Dashboard can generate plots in real time to help users debug problems with their jobs. It also provides detailed information to site administrators regarding site tests. Site tests include a check of required CMSSW releases, the ability to run different types of CMSSW jobs, and file transfers

via SRM commands specifically utilized in CMS.

The conditions of the CMS detector during data taking periods are used by the event reconstruction algorithms in CMSSW. A conditions database tracks various properties of the CMS detector and the LHC beam, such as which channels in the detector are taking data at any given time. The raw data from CMS are processed immediately without utilizing full conditions information, termed online reconstruction. The full conditions database is not used during online reconstruction because some of the conditions can only be determined using information from online reconstruction, as discussed in Section 5.3. Additionally, the current conditions of the detector are often logged to the database after online reconstruction completes, so online reconstruction uses partial conditions that are not updated on the same intervals as the full conditions database. The data are then reconstructed a second time using the full conditions database; this step is termed offline reconstruction. Conditions are frequently needed for further event processing by users, so the conditions database is made available as a global service.

Datasets are tracked in CMS using the data bookkeeping and data location services (DBS/DLS) [52]. All real and officially simulated data are registered in DBS. Users can optionally register data they have produced. DBS maps dataset names to file blocks and blocks to files. Dataset registration also includes metadata, such as the type of data (e.g., raw, reconstructed, produced by users, etc.), the parent dataset (if applicable), and the configuration file and CMSSW release used to produce and/or process the data. DLS, a component of PhEDEx (described in the next section), maps file blocks to the grid sites storing replicas. A DBS/DLS

web portal allows users to access dataset metadata and storage locations.

4.2.2 PhEDEx

CMS uses Physics Experiment Data Export (PhEDEx) [6, 7] to manage data transfers between grid sites. PhEDEx utilizes blackboard and independent agent architecture [32], running as software agents that transmit and receive various types of tracked information to and from the PhEDEx blackboard. The PhEDEx blackboard is implemented as a transfer management database (TMDB) that stores routing tables, historical transfer performance, and current requests and transfers. DLS is one of the TMDB tables. The weekly transfer volume managed by PhEDEx from July 2010-2011 varied from approximately 200 TB to 1,100 TB, as shown in Figure 4.2. During this period, the CMS-cumulative transfer speed varied from a few hundred to a few thousand MB/sec.

Users request transfers of datasets or blocks to destination sites at a specified priority with the PhEDEx web service. The web service sends emails to the administrators of the destination sites and to global data managers, any of whom can approve the request. Once the request is approved, the web service creates the subscription in the TMDB. The web service also supplies monitoring information to users and administrators, including transfer status, transfer failures, and historical transfer rates. Figure 4.2 was generated using a configured query on the web service.

Several agents actively monitor the TMDB, translating subscriptions into transfer tasks. A limited number of files are scheduled for transfer per destina-

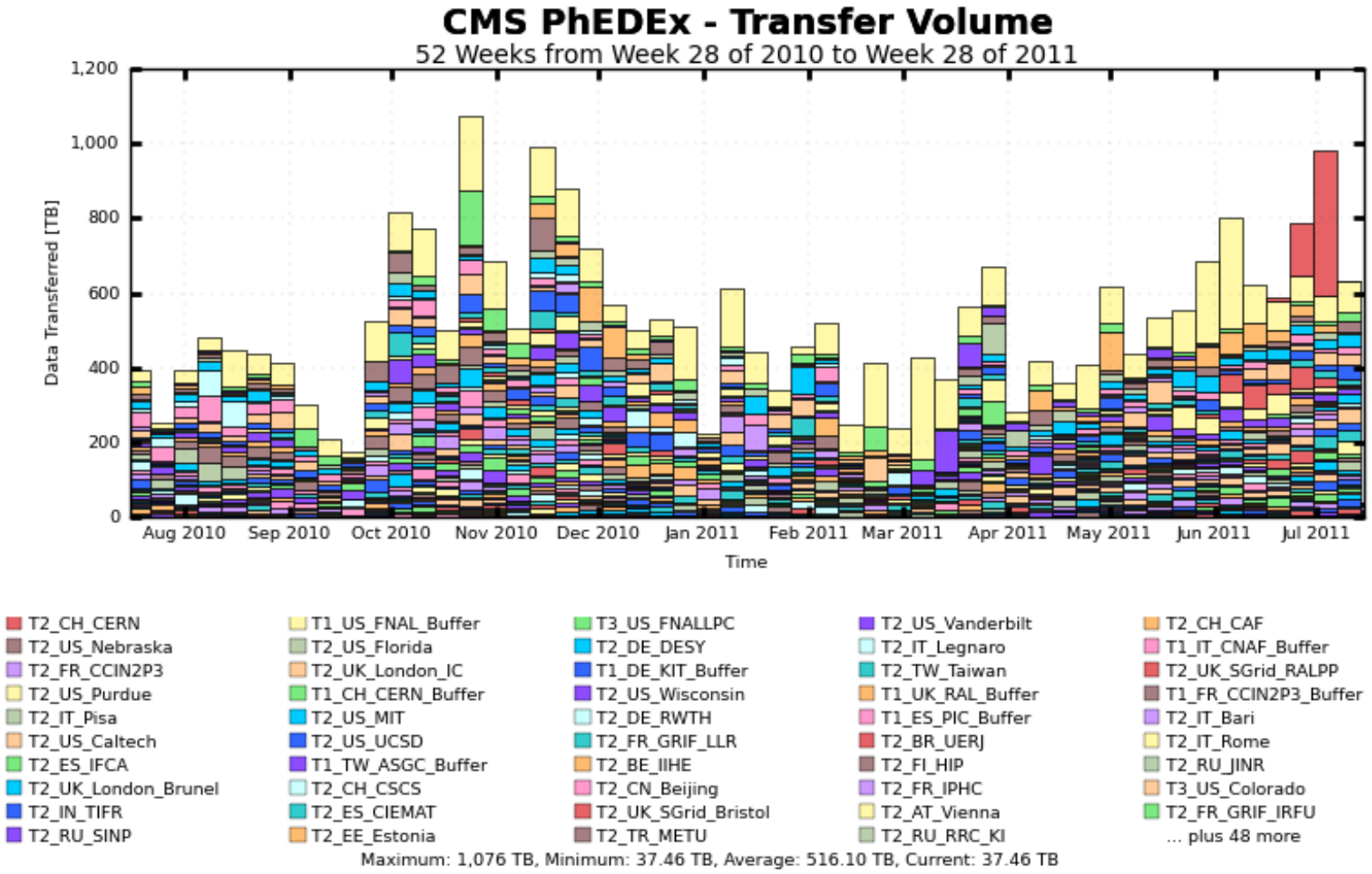


Figure 4.2: Weekly PhEDEx transfer volume from July 2010-2011, shown by destination. Generated using a configured query on the PhEDEx web service.

tion to keep the transfer table small and to avoid scheduling a transfer far in the future based on present network activity. The order of files chosen is determined by the priority of outstanding requests. Since multiple sites can host the same file block, the source host with the smallest expected transfer latency to the requested destination host is chosen. Expected latency is determined by summing (1) the latency of current PhEDEx requests on the transfer path and (2) the size of the scheduled files divided by the historical transfer rate from source to destination host. File transfer requests have expiration times (typically a few hours to a few days), automatically rescheduling the file transfer if it does not succeed within the allocated expiration time. The files to transfer, source host, and request expiration are then stored in the TMDB.

Site-local PhEDEx agents contact the TMDB at regular intervals. When the TMDB shows an outstanding transfer request to the site, the agent at the destination site pulls the data from the source host specified by the TMDB. The names of files are stored in DBS and TMDB using a unique logical file name that usually is not the same as the physical file name at the site. A physical file name might be written as a classic directory structure or as an SRM address and is different between sites. A site-local trivial file catalog is used to translate the logical to the physical file name. Sites capable of exporting data to other sites must make the files available in an SRM SE, while sites importing data can do so using any desired physical file name. PhEDEx and site network infrastructure settings are typically tuned to transfer files 2-4 GB in size. Once the transfer of a file block is complete, the new host is registered in DLS. Transfer rates are also recorded at regular intervals in the

TMDB. A number of other site-local agents are provided by PhEDEx, including agents for file deletion and block monitoring.

4.2.3 CRAB

Users submit custom data analysis and simulation jobs to the grid via the CMS Remote Analysis Builder (CRAB) [62]. CRAB primarily uses software from either Globus or gLite-UI to contact WLCG resources, though CRAB supports alternative submission mechanisms. gLite [37] is used by most of the European sites in the WLCG and is analogous to the VDT used by American sites. gLite-UI provides the gLite software needed to contact WLCG sites (the user interface). During job submission, CRAB contacts DBS/DLS to determine the sites hosting the requested dataset. BDII entries for each hosting site are then consulted to determine currently available cycles and the information needed to submit jobs to the site. When jobs are submitted, CRAB registers the jobs with the Dashboard. As CRAB jobs complete, they contact the Dashboard to update their state, including an exit code that conveys highly detailed information to the user. CRAB job submission to the CE or batch system of the servicing sites is either done directly by the user or via an intermediate CRABServer. Currently the submission mechanism is up to the user.

CRABServers are capable of submitting jobs to CEs via gLite or Condor glideIn, a particular type of grid-enabled Condor submission mechanism. Condor glideIn can yield lower job failure rates because glideIn utilizes 'pilot' jobs, jobs sent

to CEs before the actual job from the user. Once the pilot job contacts the CE, is scheduled by the CE job manager, and lands on a servicing compute node, it pulls the waiting user job from the CRABServer. Therefore, jobs submitted via the glideIn mechanism are guaranteed to land on a CE that is at least minimally functional. When glideIn jobs complete, the output is automatically sent back to the CRABServer. CRABServer monitors the state of jobs and as gLite jobs complete, CRABServer requests the output of the job from the servicing CE. This cleans up temporary space allocated to the job by the servicing site more rapidly than gLite jobs submitted directly by the user. CRABServer then hosts the output from either glideIn or gLite submitted jobs until the user requests it.

Users are constrained to produce small output files or to use CRAB in SE stageout mode. In this mode, CRAB sends the output directly from the servicing site to an SE where the user has access privileges. In SE stageout mode, users have the option to register their output to DBS/DLS. If users opt to register their dataset, it can be processed by CRAB again, which allows users to easily share custom datasets with one another. However, since user produced files can vary significantly in size, PhEDEx cannot be used to transfer the data produced by CRAB from one SE to another.

4.2.4 ProdAgent

Official data simulations are managed via ProdAgent [34], which is composed of a persistent relational database and application logic contained in python dae-

mons. The daemons communicate with each other via the database utilizing asynchronous encapsulated messages. Physicists create configuration files for CMSSW which produce simulated data, then request the production of a specified number of events from the configuration files. Requests are compiled and selected for production based on physics priorities and available resources. Because the requests go through a review process, simulations managed by ProdAgent are allocated more resources than custom user simulations submitted through CRAB.

Approved requests are handed to members of Data Operations who schedule and execute the jobs using several ProdAgent instances. Each ProdAgent instance manages production at a subset of CMS sites. For example, all US CMS sites participating in ProdAgent are managed by one ProdAgent instance. The ProdAgent software stack is capable of interfacing with a wide variety of job scheduling systems, including Globus and Condor. Different ProdAgents work independently, controlled and monitored by Data Operations staff. Failed jobs are resubmitted by operators as needed. As jobs complete, ProdAgent merges the output of multiple jobs at a single site, creating files 2-4 GB in size. ProdAgent registers the new files to DBS/DLS services running at the servicing site. The files are then copied or transferred from the site that produced the data to other sites designated as custodial hosts using PhEDEx, which registers the files to the CMS global DBS/DLS.

During the data taking period of this dissertation, ProdAgent was also used to manage offline reconstruction of the events taken by CMS. This has since been replaced by a new workload management (WM) system, which is implemented as a request manager, work queue, and independent WMAgents analagous to ProdA-

gents. The WM system automatically converts user requests into scheduled jobs at CMS sites with minimal intervention from staff in Data Operations.

4.2.5 Tiers

CMS organizes computing resources using a tiered structure from Tier 0 to Tier 3. The Tier 0 is located at CERN and stores a complete copy of the raw data. The Tier 0 also performs both online and offline reconstruction of the data, which is then distributed to Tier 1 sites via PhEDEx.

Tier 1 sites are located across the world and are designed primarily as national or regional bases of operation. The United States Tier 1 for CMS is located at Fermilab in Batavia, Illinois. A complete copy of all recorded and simulated data is stored between all Tier 1 sites. Tier 1 sites reprocess recorded and simulated data whenever new CMSSW releases become the CMS recommended standard. Data is transferred between Tier 1 sites to balance network loads and distributed from Tier 1 sites to Tier 2 and 3 sites based on user demand. Although Tier 1 sites are not contractually required to produce simulated data, they are used to produce simulated data via ProdAgent when either (1) no reprocessing is occurring, or (2) production demand exceeds the production resources supplied by Tier 2 sites.

Tier 2 sites allocate approximately half of their computing and storage resources to the production and storage of simulated data from ProdAgent and the other half to user driven analysis and simulation from CRAB. Simulated data are propagated back to the Tier 1 site affiliated with the production site. Tier 2 sites

can also distribute data directly to other Tier 2 and 3 sites.

Tier 3 sites are funded by individual university or institutional grants and thus are not contractually required by the CMS VO to provide any particular service. Tier 3 sites are highly flexible, sometimes filling short term gaps between user demand and officially available services from other tiers. Most users login interactively at Tier 3 sites, which serve as user access points to the larger grid. A number of Tier 0, 1, and 2 sites have collocated Tier 3 resources to meet the interactive needs of local analysts.

The CMS computing requirements for 2010 and 2011, organized by tier, are shown in Table 4.1. CPU resource pledges in the WLCG VO are in units of the HEP-SPEC06 (HS06) benchmark [40]. HEP-SPEC06 is an adaptation of the C++ benchmarks in CPU2006 [63] for high energy physics workflows. CPU2006 is designed to stress both the processor and memory. The total required computing resources by CMS for 2010 (2011) were 396 (550) kHS06 (kilo-HEP-SPEC06), 29 (40) PB of disk, and 37 (64) PB of tape. For comparison, the largest compute node at the University of Maryland Tier 3 provides 11 TB of usable disk and 24 job slots with a CINT2006 (CPU2006 integer tests) benchmark of 36.9 [63], which was evaluated using the Intel C++ professional compiler (CMS uses gcc).

4.3 The University of Maryland Tier 3

The University of Maryland runs a CMS Tier 3 providing both CE and SE services. The services are used by Maryland students and faculty and provide oppor-

Tier	CPU (kHS06)	Disk (PB)	Tape (PB)
2010			
0	96.6	4.0	14.3
1	103.5	11.9	23.1
2	196.2	13.3	N/A
2011			
0	106.1	4.4	21.1
1	131.4	15.8	43.1
2	312.0	19.9	N/A

Table 4.1: The CMS computing requirements for 2010 and 2011 summed over all sites in each tier. This table is taken from Reference [67].

tunistic computing cycles to the rest of the CMS VO. Local users typically submit CRAB jobs from the Maryland Tier 3 and use SE-stageout mode to send the output back to the Maryland Tier 3 SE. Local users may register their datasets to DBS/DLS, enabling access to the data by non-local users via the Maryland Tier 3 CE and SE. Remote users of the Maryland Tier 3 typically use the CE and SE to process locally stored datasets (from CRAB or PhEDEx) or use the CE to produce their own custom simulations.

The Maryland Tier 3 computing rack has two management nodes, two interactive nodes, and thirteen compute nodes. All the nodes run Scientific Linux [58], a derivative of Red Hat Enterprise Linux [55]. Condor is used for local job manage-

ment. The management and interactive nodes are on uninterruptible power supplies. The cluster internal network switch is a Dell PowerConnect 6248 with 48 gigabit ethernet ports. The hardware was purchased for order $100,000 over several years.

The cluster is managed by the head node. The head node is a Rocks [49] frontend, which serves the operating system and basic operating system settings to all other nodes, as well as some software. It also hosts user disk space using the network file system (NFS) and is the Condor pool manager. The server is a 2U Dell PowerEdge 2950 with two quad-core Intel Xeon E5440 processors with a clockspeed of 2.83 GHz, 8 GB of RAM, and 490 GB of disk. All software is installed on a 70 GB partition managed by RAID-1, while the user disk is on a 420 GB partition managed by RAID-5. Redundant Array of Independent Disks (RAID) [51] is a disk federation system, where different levels of RAID dictate how data is distributed across multiple disks, which yields redundancy under different types of disk failure. RAID-1 is a simple mirroring system, while RAID-5 employs data striping that loses no data when a single disk fails.

The grid management node runs an OSG CE and SE, CMSSW, and PhEDEx. CMSSW is network mounted and releases are installed and removed automatically by the CMS VO via the Maryland Tier 3 CE. The grid node hardware is identical to that of the head node. The operating system and PhEDEx are installed on the 70 GB RAID-1 partition and the network mounted VDT and CMSSW are on the 420 GB RAID-5 partition. The grid node also manages a 9 TB usable size disk volume with RAID-6. Like RAID-5, RAID-6 also employs data striping, but loses no data when two disks fail. This network mounted 9 TB volume serves as the Maryland

SE as well as local user disk space.

Users login to the interactive nodes, where they run local analysis jobs and submit jobs to the grid or the cluster using CRAB or Condor. Both gLite-UI and CRAB are installed by Rocks from the head node during operating system install. This enables rapid redeployment of the entire software stack should failure occur on one of the interactive nodes. The interactive nodes are 1U Dell PowerEdge 1950s with two quad core Intel Xeon E5440 processors with a clockspeed of 2.66 GHz, 16 GB of RAM, and two 146 GB disk drives. One local drive hosts the software while the other hosts temporary storage for users, which is garbage collected each week.

The cluster has two types of compute nodes. Eight of the compute nodes are PowerEdge 1950s with chips identical to the head node, 16 GB of RAM, and two 2 TB disk drives. CMSSW event simulation and reconstruction jobs are limited to consume less than 2 GB of memory, so the 1950 compute nodes have 8 Condor job slots. Five of the compute nodes are 2U Dell PowerEdge R510s with two six-core Intel Xeon X5650 processors with a clockspeed of 2.66 GHz, 48 GB of RAM, and twelve 2 TB disk drives. The chips are hyperthreaded to produce 24 virtual cores per R510, corresponding to 24 Condor job slots. The compute nodes have a total of 152 TB of disk space. This space will be federated over multiple disks and nodes using the Hadoop Distributed File System (HDFS) [4], yielding approximately 70 TB of usable space. With a larger disk volume, the Maryland Tier 3 will be capable of storing more official datasets, transferred via PhEDEx from Tier 1 and 2 sites.

4.4 Chapter highlights

CMS distributes and analyzes data using the WLCG. Sites contributing computing and storage resources to the WLCG provide them via CE and SE services. CMS resources are organized in Tiers. The Tier 0, located at CERN, performs online and offline reconstruction and hosts a complete copy of the data taken by CMS. Tier 1 sites reprocess and host a complete copy of real and simulated data, as well as produce simulated data via ProdAgent. Tier 2 sites produce simulated data via ProdAgent and service user jobs submitted via CRAB. Tier 3 sites provide interactive user access points to the rest of the grid. They are controlled by local users and provide no contractually obligated services to the WLCG, though Tier 3 sites frequently provide opportunistic access of computing and storage. PhEDEx transfers data taken by the CMS detector from the Tier 0 to Tier 1 sites and the simulated data made by ProdAgent from Tier 2 to Tier 1 sites. PhEDEx is also used to transfer data based on user demand.

CRAB, ProdAgent, and PhEDEx contact BDII to get information about sites in the WLCG and contact DBS/DLS to get or set CMS dataset metadata and location. Several monitoring services, such as RSV, MyOSG, and Dashboard, provide web services for users and administrators to discover site or job status. A conditions database tracks the state of the CMS detector, which is accessed by CMSSW jobs. CMSSW, installed at all CMS sites, is used to perform online and offline reconstruction, to simulate data via ProdAgent, to analyze or simulate data via CRAB, or to reconstruct, simulate, or analyze data interactively.

Chapter 5

Event simulation, reconstruction, and selection

All event simulation and reconstruction is performed within the CMS software framework, CMSSW [16]. CMSSW is a C++ and python based software package that wraps a number of other software packages, including MadGraph [3], Pythia [60], GEANT [1], and Root [10]. For this dissertation, MadGraph is used to create Monte Carlo simulations of events using matrix elements, where the matrix element is the probability amplitude of the fundamental interactions described by Feynman diagrams (described in Section 2.1). Pythia simulates the process of hadronization from the events generated by MadGraph, termed parton showering. The particles generated by MadGraph and Pythia are passed to GEANT, which simulates the response of the CMS detector.

Digitization and reconstruction are performed on both Monte Carlo simulated data and real data. During digitization, measurements by the CMS subdetectors (or GEANT simulations of subdetector measurements) are quantized into values of position and/or energy. Event reconstruction uses digitized position and energy information to identify the individual particles and their position and energy. In this context, purity is the number of true particles that were reconstructed divided by the total number of reconstructed particles (some reconstructed particles are erroneous). Reconstruction efficiency is the number of reconstructed particles divided by the

total number of true particles (some true particles are not reconstructed).

5.1 Simulated data

5.1.1 Event generation using MadGraph & Pythia

Four steps are used to computationally generate particle interactions: (1) a statement of the Feynman rules and particle properties like mass, (2) calculation of the matrix elements, (3) event generation with probability amplitude given by the matrix element, and (4) parton showering of hadronic particles. MadGraph performs the first three operations and Pythia the last. Matrix elements are implemented as functions that return the probability amplitude for an interaction given initial and final states as input to the function. The matrix element includes the effects of quantum interference and the integral of all possible intermediate states up to a specified perturbative order. MadGraph uses previously calculated probabilities of initial states to generate user specified final state particles via the intermediate state probability amplitudes given by the matrix elements. For collisions of composite particles like protons, the probability distribution of initial states is termed the parton distribution function, or PDF. The PDF describes the fraction of the total proton momentum carried by the initial state particles involved in the collision. Proton PDFs are phenomenological models with dozens of parameters; we use CTEQ66 [48]. MadGraph uses Monte Carlo simulation techniques to generate event final states.

Since the LHC collides high energy protons, production of additional jets is

highly probable. The additional jets can be radiated from the initial particles, termed initial state radiation (ISR), or from the final particles, termed final state radiation (FSR). Examples of both are shown in Figure 5.1. These jets can be radiated as final state particles themselves and can significantly modify the kinematics of the other final state particles. While additional virtual loops in the Feynman diagram (as in Figure 2.3) modify total event probability, they do not significantly modify the kinematics of the individual final state particles. Therefore, higher order perturbative calculations are needed for an accurate prediction of total cross section while leading order simulations with additional radiated jets are needed for an accurate prediction of particle kinematics. These simulations are typically done separately; MadGraph leading order simulations with additional jets are used in the full event processing chain and higher order simulations are used to predict the production cross section of the fully simulated lower order events.

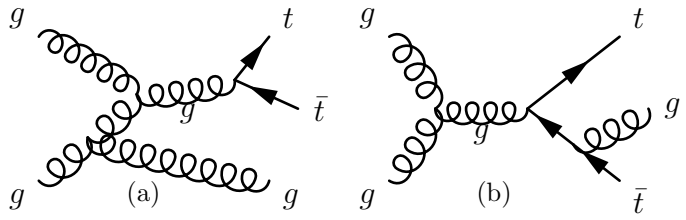


Figure 5.1: Possible leading order Feynman diagrams for the process $gg \to t\bar{t}$+1 jet, where an additional gluon is produced from (a) initial state radiation and (b) final state radiation.

Parton showering is the process of pulling particles from the vacuum due to color confinement, where the new particles eventually bond into hadrons during hadronization. The number of particles in parton showers grows exponentially as the

shower propagates. Therefore, parton showers are modeled using phenomenology, which leads to large uncertainties in computational models where hadronization is key. Pythia takes MadGraph events as input and generates the parton showers. However, a parton shower is composed of many quarks and gluons. It is therefore ambiguous as to whether the quarks and gluons generated by parton showers in Pythia are double counting the additional radiated jets already in MadGraph. To resolve the ambiguity, a matching threshold is used. Partons in Pythia with sufficient energy are treated as already included in the MadGraph event and lower energy partons are considered soft components of the shower, not already in the MadGraph event.

Pythia also models beam remnants using model parameters tuned by previous data measurements. Beam remnants are quarks and gluons in the proton that did not collide, but do hadronize after the internal bonds in the proton are broken by a hard scatter. Additionally, multiple interactions can occur between the proton constituents, where one interaction is a hard scatter and the other is a soft scatter. This is typically simulated by mixing real 'zero bias' events with simulations from Pythia, where 'zero bias' refers to detector activity measured by CMS that does not pass any trigger. While the particles from the beam remnant and multiple interactions are rarely triggered on or reconstructed by themselves, they can interact with the same pieces of the detector as the particles in the primary interaction. This reduces the precision of reconstructing particle trajectories and energies and therefore must be included in the simulations.

Pythia outputs a list of particles that have long enough lifetimes to travel

measurable distances before they decay. Usually these particles reach the material in the CMS detector. GEANT then takes the list of particles from Pythia and simulates the interactions of the particles with the CMS detector.

5.1.2 Event simulation using GEANT

GEANT (GEometry ANd Tracking) is a software package designed to simulate the passage of elementary particles through materials. It is used for a wide variety of purposes, including simulating output from medical scanning devices, interactions of cosmic radiation with spacecraft, and high energy, nuclear, and accelerator physics. In CMSSW, GEANT models shower development in the calorimeters and material ionization and particle scattering in the trackers. It takes as input the generated particles from Pythia. The location and types of materials in CMS (including passive elements like cables) are specified in the GEANT configuration. GEANT outputs simulated hits, which represent the energy loss from a single particle interaction with the calorimeters or trackers, or the ionization of the silicon or gas in the trackers as a particle traverses the material. GEANT particles are either generated particles from Pythia or secondary particles arising from interactions of particles with the material. Like MadGraph and Pythia, the rate of interactions and production of secondary particles is modeled using Monte Carlo simulation techniques. Additional utilities in CMSSW take the simulated hits from GEANT and convert them into the raw data format used for real data from the CMS detectors.

5.2 Event digitization

For trackers, the change in voltage measured in raw data is digitized into a hit. A hit refers to a location in the tracking chamber, which is determined by the layer of the tracking chamber, the time of the change in voltage and, for the inner tracker pixels and glued strips, the shape of the voltage change over multiple pixels or strips. Important considerations for hit digitization are efficiency and purity. Normal electrical noise will cause changes in the measured voltage. To improve purity, hit digitization requires a significant change in the measured voltage, which prevents the misidentification of random electronic noise as a particle traversing the tracking layer.

For calorimeters, the measured photon counts in raw data are digitized into energy. The light emitted by segments of the calorimeter (termed a cell) is channeled to specific photodetectors. Before CMS was completely assembled, the calorimeters were calibrated to determine the relationship between measured photon counts and incident energy. Additionally, scintillating materials lose transparency due to prolonged radiation. The number of photons corresponding to a unit of energy slowly decreases as the integrated radiation increases, so the energy response of the calorimeters require frequent recalibration.

5.3 Event reconstruction

Since we require a $t\bar{t}$ signature of a muon and jets, we describe below how muons and jets are reconstructed in some detail. To elucidate the task of event

reconstruction, an event containing a reconstructed muon and four jets is shown in Figure 5.2. Event reconstruction takes the digitized data from the CMS detectors and identifies individual particles and overall characteristics of the collision, such as where it occurred. Identification of particle type usually requires information from several detectors, as described in Section 3.2 and shown in Figure 3.5. Since particle identification uses knowledge of the location of the sensors in multiple layers and multiple subdetectors, physical alignment of the sensors with relation to each other is key. Additionally, knowledge of the conditions of the detector during the collision assists with the reconstruction of the particles in the event. For instance, a change in temperature or fluctuation in the magnetic field modifies the energy deposited by particles showering in the calorimeters or the radius of curvature as particles traverse the trackers. While objects reconstructed by the level 1 and high level triggers do not use a complete set of conditions, final event reconstruction does. A simulated conditions database, which represents the physical conditions during real collisions, is also used to simulate events in GEANT.

5.3.1 Track, beamspot, and vertex reconstruction

Tracks are intermediate physics objects that represent individual charged particles. This section describes tracks reconstructed in the inner tracker, while the next section describes tracks reconstructed in both the inner tracker and muon chambers. Track reconstruction consumes the majority of event reconstruction time due to the large number of combinations from matching thousands of hits to multiple tracks.

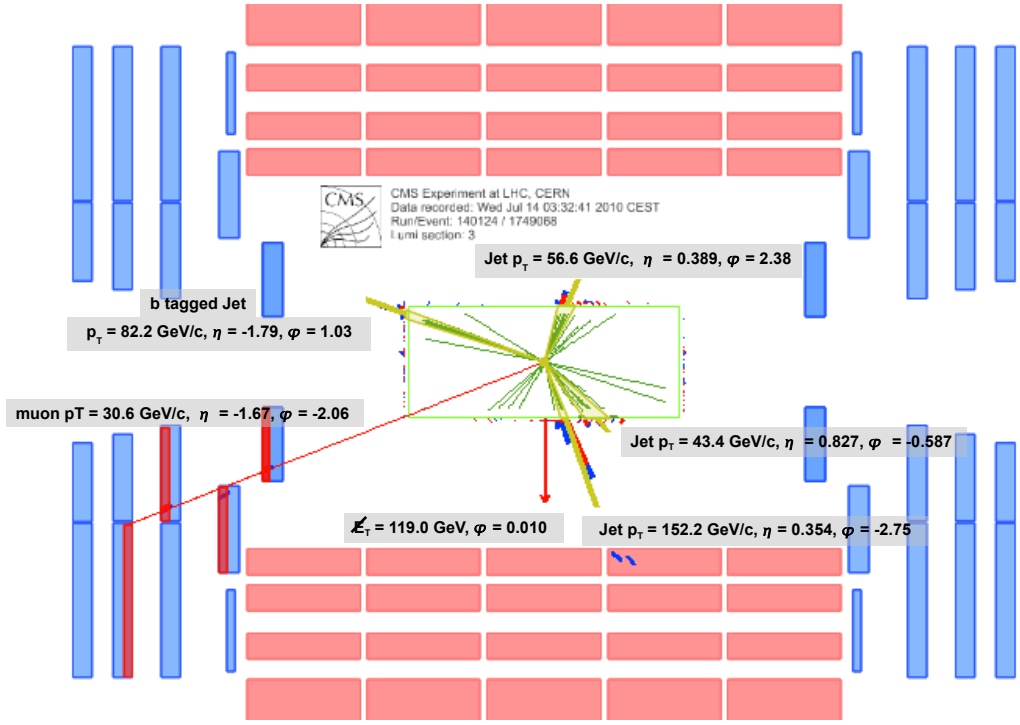


Figure 5.2: An event observed at CMS with a muon and four jets, one of which is identified ("tagged") as originating from a b quark. Dark green lines in the center are reconstructed tracks, blue and red histograms are energy deposits in the calorimeters, the red line traversing the muon chambers is the reconstructed muon, and the pale green segments are reconstructed jets. Tracks and calorimeter energy deposits with no corresponding jet or muon are interpreted as particles in the interaction that were not reconstructed, electronic noise, or the underlying event. Credit: [24].

The combinatorial track finder (CTF) [17] passes first guess tracks to a Kalman filter [44]. A charged particle track is a helix, where the particle orbits the magnetic field lines oriented in the z direction (along the beam axis). Three points are mathematically sufficient to define a helix, so CTF produces seeds (first guess tracks) by using either three pixel hits or two pixel hits and the expected beamspot (the intersection point of the proton beams). The seed, defined by the accurate pixel hits, is iteratively propagated outward through the layers of the tracker from the beamspot using a Kalman filter. Because a Kalman filter updates the hypothesis (the track) with the uncertainty of the new data (the hits), the uncertainty of the hypothesis grows. Therefore, the track hypothesis is propagated a second time from outer hits to inner hits to yield a relatively constant uncertainty along the length of the track ("error smoothing").

Between LHC fills, the beamspot may move. In 2009, the beamspot fill-to-fill variations were ± 0.5 mm in x and y and ± 2 cm in z. When the true beamspot is displaced from the beamspot assumed for track reconstruction, there is a correlation between the distance and angle of the closest approach of tracks to the beamspot in the transverse plane (where distance of closest approach in the transverse plane is termed impact parameter, or d_0). An iterative χ^2 fitter uses this correlation to determine the location of the beamspot for the duration of the fill and uploaded to the conditions database. Tracks are then reconstructed a second time using the new beamspot.

Although the interaction point of the bunches of protons in the LHC beams varies little within a single fill, the location of the individual interacting protons

within the bunch varies in every collision. The location of the actual proton interaction is termed the primary vertex. The primary vertex is displaced further from the beamspot in the z direction than in the transverse plane because the bunch size is larger along z, the direction of acceleration. Therefore, the location of the primary vertex is determined by searching for clusters of tracks that intersect at a common point in z. An adaptive vertex finder is used to identify the primary vertex. It is implemented as a generalized Kalman filter where the hypothesis used in the Kalman algorithm is the location of the vertex and the iteratively added data points in the Kalman algorithm are the reconstructed tracks. Unlike a standard Kalman filter, the adaptive vertex finder does not reject tracks inconsistent with the vertex hypothesis. Rather, the adaptive vertex finder weights tracks by their z distance from the vertex hypothesis and iteratively updates the vertex hypothesis and track weights until the hypothesis converges. The method is robust when more than one vertex is present in the event, which occurs whenever a particle decays a measurable distance away from the primary interaction. This is common in events containing b quarks, due to their relatively long decay lifetime. The primary vertex (PV) is assumed to be the vertex with the most tracks originating from it, while all other vertices are termed secondary vertices.

5.3.2 Muons

Muons are expected to leave tracks in both the inner tracker and the muon chambers. Standalone muon tracks are reconstructed with the Kalman filter using

hits in the muon chambers [8]. Standalone muon seeds are taken from (1) the online reconstructed tracks that caused the event to pass L1 muon triggers and (2) track segments within each muon chamber.

Two separate techniques are used to reconstruct muons from inner tracks and standalone muon tracks. "Global muons" are reconstructed starting with the standalone muon track and searching for a matching track in the inner track collection. The global muon track is determined with a combined fit using the hits in the standalone muon and inner tracks. "Tracker muons" are formed by extrapolating inner tracks to the muon chambers and searching for at least one matching track segment in the muon chambers. Global muon reconstruction is more efficient at higher momenta while tracker muon reconstruction is more efficient at lower momenta [23]. Requiring a muon to be reconstructed by both the global and tracker muon techniques yields the highest reconstruction purity.

5.3.3 Jets

For this dissertation, jet reconstruction is performed using the particle flow algorithm [66]. The algorithm matches inner tracks to clusters of energy in ECAL and HCAL. Charged hadrons are identified using tracks that match to energy clusters in HCAL, while photons and neutral hadrons are identified from energy clusters not associated with any tracks.

The particle flow algorithm takes previously reconstructed inner tracks and removes the hits associated with these tracks from the collection of hits. It then

iteratively loosens track seeding criteria to reconstruct additional tracks using the inner track reconstruction algorithm. Hits associated with tracks are removed after each iteration. This process improves track reconstruction efficiency while keeping the purity high because hits are removed, which reduces the number of combinatoric options at each step.

The particle flow algorithm clusters calorimeter cells in three steps. First, cluster seeds are chosen as cells with energy deposits above some minimum threshold. Then topological clusters are iteratively grown by including cells with an energy above a threshold and with at least one side in common with a cell already in the cluster. Finally, the energy and position of the clusters are adjusted by fractionally sharing the energy in a single calorimeter cell across multiple clusters according to the cell-cluster distance, which is recalculated after each adjustment.

Particle flow produces jets by using the inner tracks and calorimeter clusters in an anti-k_t algorithm [61]. Jets propagate with a cone-like shape, so jet reconstruction algorithms typically use cones with specified angular separation ΔR, which is usually scaled by jet energy to form the distance metric. CMS uses $R = 0.5$. Initially given a list of inner tracks and calorimeter clusters, the pseudo-code for the anti-k_t algorithm is:

WHILE (particles remain in the list):

- for each particle i, calculate $d_i = p_{\mathrm{T,i}}^{-2}$
- for each pair of particles i and j, calculate $d_{ij}^2 = min(p_{\mathrm{T,i}}^{-2}, p_{\mathrm{T,j}}^{-2})\Delta R_{ij}^2/R^2$
- for all calculated d_i and d_{ij}, find the minimum

- if the minimum is a d_{ij}, combine i and j into a new particle and add it to the list, removing the original i and j

- otherwise, declare i is a jet and remove it from the list

Some jet reconstruction algorithms incorrectly merge jets that have low energy radiation between them or are collinear, but the anti-k_t algorithm is infrared and collinear safe.

Inner tracks and calorimeter clusters are also linked by their location. The linked blocks within the jet then determine the charged, neutral, electromagnetic, and hadronic components of the jet. The linking algorithm considers every pair of elements in the event using geometrical distance as the link quality. Elements are created iteratively, where an element is a track or calorimeter cluster or a previously made pair of elements. Generally a link is kept when the element with larger geometrical uncertainty fully envelops the geometrical position of the element with smaller geometrical uncertainty. Charged hadron particle candidates are links between calorimeter clusters and at least one track. Neutral hadron and photon candidates are calorimeter clusters that are not linked with a track. Sometimes charged and neutral candidates overlap in the calorimeters. Charged and neutral particle overlap is identified by large energy excess in the calorimeter cluster with respect to the sum of the associated track momenta. Particle flow is applied in a similar fashion to reconstruct electrons and muons. For hadronic jet reconstruction purposes, particle flow electrons and muons are removed from the collection of particle flow tracks and calorimeter clusters.

Reconstructed muons originating from a jet are termed 'fake'. Jets are primarily composed of neutral and charged pions, which are the lightest hadrons containing two quarks. The charged pions leave tracks in the inner tracker. A jet shower is occasionally not fully absorbed within the calorimeter, leaving hits in the muon chambers. Therefore, the tracks in the inner tracker can geometrically match to hits in the muon chamber from the jet, leading to a reconstructed muon. Additionally, pions can decay to muons, also leaving hits in the muon chamber that geometrically match to tracks in the inner tracker. Fake reconstructed muons are removed from events by requiring the muon to be isolated; that is, energy near the muon track in the calorimeters and tracker must be below some threshold and the muon must be a minimum ΔR away from reconstructed jets. However, not all fake muons can be successfully removed without drastically reducing the efficiency of reconstructing muons from the primary interaction of the protons.

5.3.4 b-tagging

Due to the lifetime of the B meson, discussed in Section 2.2, B mesons decay inside the CMS detector, yielding additional tracks that do not intersect the primary vertex of the event. Additionally, since the b quark has a mass of 4.19 GeV [50], jets originating from b quarks often have additional tracks relative to jets from lighter quarks. The simple secondary vertex (SSV) b-tag algorithm [18] identifies b jets by requiring a reconstructed secondary vertex with N or more associated tracks. The SSV discriminator is the absolute value of the significance of the three dimensional

distance from the primary to secondary vertex. For the SSV high efficiency (SSVHE) b-tag, $N = 2$ and for SSV high purity (SSVHP), $N = 3$. The track counting (TC) b-tag algorithm [18] sorts all tracks associated with a jet in order of decreasing impact parameter significance. The impact parameter significance of the N^{th} track is used as discriminator value, where $N = 2$ for track counting high efficiency (TCHE) and $N = 3$ for track counting high purity (TCHP). A cut on the discriminator is used to achieve the desired b tag efficiency and purity.

Figure 5.3 shows the SSV and TC discriminators for QCD events in early data from CMS and simulated data. The jet flavor (b, c, or light) of jets in simulated events is known, where light jets refer to jets originating from a gluon or up, down, or strange quarks. Cutting on the value of the discriminator yields different efficiencies for different jet flavors, as shown in Figure 5.4. The charm quark has a mass of 1.3 GeV [50] and thus, like the bottom quark, jets from charm quarks often contain additional tracks that can increase the value of the tag discriminator relative to lighter quarks. Three separate cuts are defined for each discriminator, termed loose, medium, and tight, corresponding roughly to a 10% light fake rate in QCD events (the rate at which light jets are tagged as b), 1% fake rate, and 0.1% fake rate, respectively. Early studies indicated the SSVHE medium (SSVHEM) and TCHE medium (TCHEM) operating points provided the best performance in our analysis. SSVHEM corresponds to a cut value of 1.74 and TCHEM corresponds to a cut value of 3.3. For the jets in $t\bar{t}$ events passing the selection to be described in Section 5.5, the SSVHEM operating point had a b tag efficiency of 51% to 68% and light fake rate of 1% to 6% as a function of jet kinematics, while the TCHEM operating point

had a b tag efficiency of 55% to 74% and light fake rate of 1% to 6%. Since TCHEM had a higher efficiency and comparable fake rate as compared to SSVHEM, the TCHEM operating point was used in this analysis.

5.4 Data and Monte Carlo simulated samples

Measurement of the $t\bar{t}$ cross section is performed using data samples with an integrated luminosity of 36.1 pb^{-1} taken by CMS in 2010, summarized in Table 5.1. The datasets used are composed of events that pass at least one muon trigger. During early running, the instantaneous luminosity delivered by the LHC was low, so triggers with low thresholds could be used. Later runs at higher instantaneous luminosity required the use of triggers with higher thresholds.

Trigger name	Integrated luminosity (pb^{-1})
HLT_Mu9	8.3
HLT_Mu15	27.9
Total	36.1

Table 5.1: Trigger and integrated luminosities for data used in the analysis.

Monte Carlo simulations were generated for $t\bar{t}$ signal events as well as for those processes expected to act as significant backgrounds in a $t\bar{t}$ cross section measurement. The various simulated samples used in this analysis are listed in Table 5.2, along with the theoretical cross section for each process at $\sqrt{s} = 7$ TeV at the LHC. The simulated samples were generated at leading order with additional jets, while

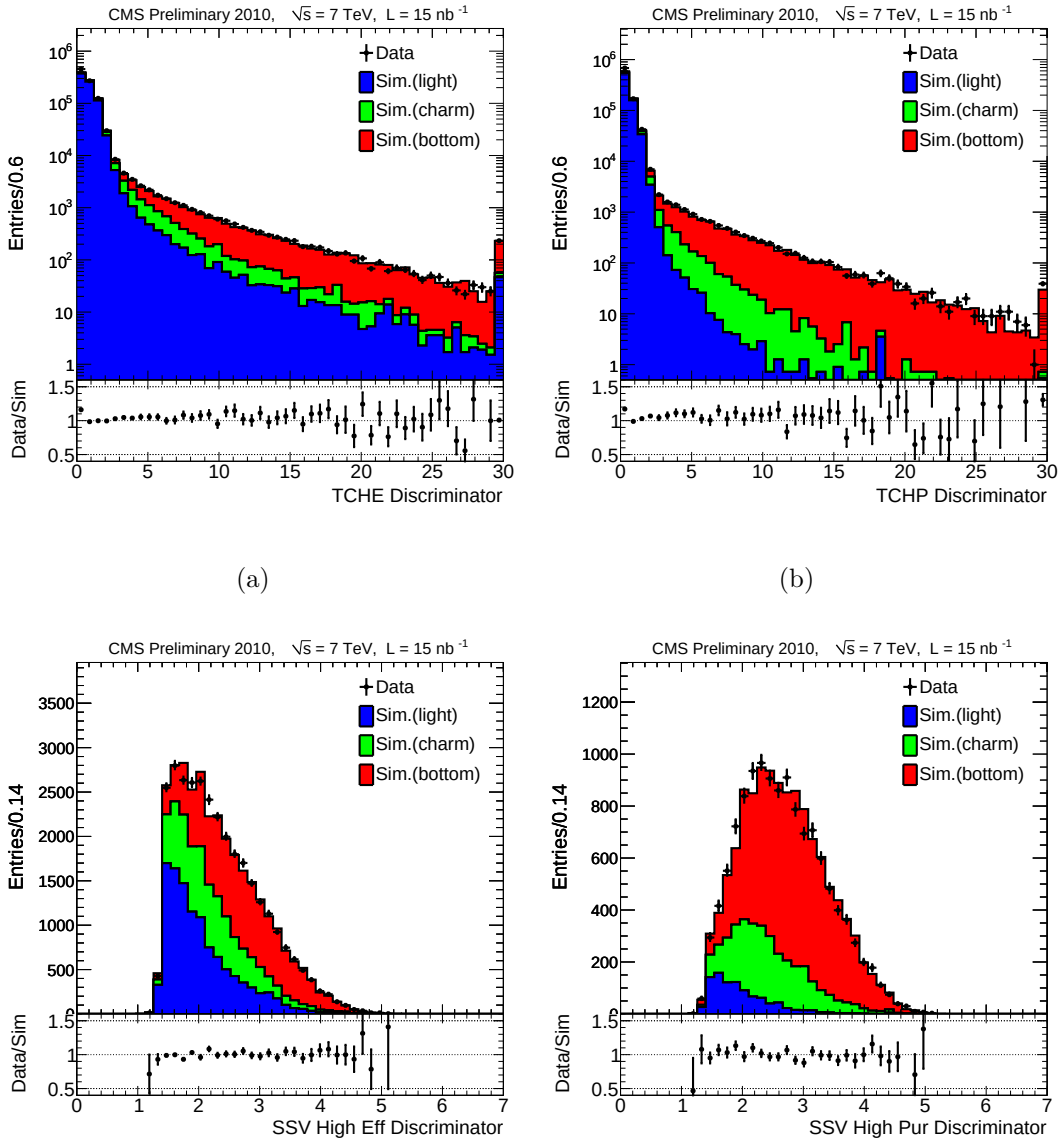


Figure 5.3: Distribution of b-tag discriminators for data (points) and light (blue), charm (green), and bottom (red) jets in simulated events. (a) TCHE, (b) TCHP, (c) SSVHE, (d) SSVHP. Credit: [18].

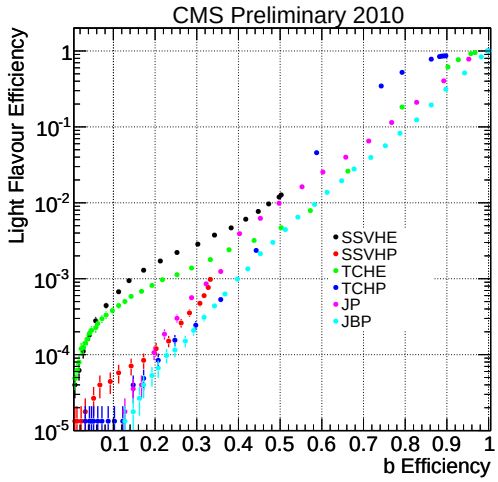


Figure 5.4: Light flavor versus b flavor efficiency for different tagging algorithms. Jet probability (JP) and jet b probability (JBP) tagging algorithms were not considered for this analysis. Credit: [18].

the theoretical cross sections were calculated at higher orders.

As discussed in Section 2.2, W bosons can decay leptonically, and events in which a leptonically decaying W boson is produced along with additional jets is the largest source of background in a semi-leptonic $t\bar{t}$ analysis. Z bosons and off-shell photons (γ^*) can also decay leptonically, but since they decay to two charged leptons, vetoing on the presence of a second lepton can reduce this background. The term "V+jets" used in this dissertation includes the simulated samples of W+jets and Z/γ^*+jets. QCD jet-only events do not contain a muon from the primary interaction, but have production cross section many orders of magnitude larger than the $t\bar{t}$signal. Because jets can fake muons, even a low muon fake rate coupled with the high QCD cross section yields a significant number of QCD background

Name	Cross section (pb)	Perturbative order
$t\bar{t}$+jets	157.5	NLO
W+jets	31,314	NNLO
Z/γ^*+jets	3,048	NNLO
QCD	84,679	LO
Top-s	1.4	NLO
Top-t	21.0	NLO
Top-tW	10.6	NLO

Table 5.2: Simulated datasets used in this analysis.

events with a reconstructed muon. Single top quarks produced with additional jets, where the top decays leptonically, are also a source of background. Single top events are listed as Top-s and Top-t in Table 5.2, referring to "s-channel" and "t-channel" production modes, shown in Figure 5.5. Additionally, single top events produced with a W boson (labeled Top-tW), where either one decays leptonically, can contain the muon+jets topology of $t\bar{t}$ semi-leptonic events. The term "single top" used in this dissertation includes the simulated samples of Top-s, Top-t, and Top-tW. All samples except for QCD were generated with MadGraph and Pythia, while the QCD sample was generated using only Pythia.

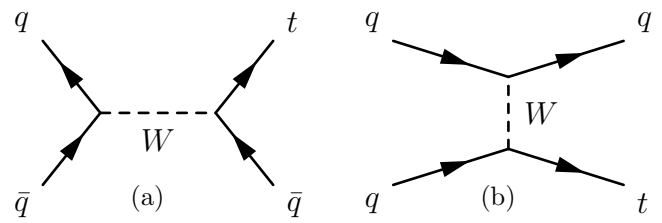


Figure 5.5: Leading order Feynman diagrams for single top production via (a) s-channel quark-antiquark annihilation and (b) t-channel quark-W interaction.

5.5 Event selection

Events containing the desired signature of a muon and jets are selected from simulated or real data samples. The event selection for $t\bar{t}$ semi-leptonic events follows the recommendations of the top physics group at CMS [25], which was optimized to reduce the rate at which non-$t\bar{t}$ background events pass selection. Although the expected event topology from the semileptonic decay of $t\bar{t}$ contains four high energy jets, only three jets are required in this analysis. It was found that requiring a fourth jet with $p_{\mathrm{T}} > 30$ GeV reduced the $t\bar{t}$ selection efficiency by roughly 50%, as shown in Table 5.3. Additionally, requiring a fourth jet greatly increases the systematic uncertainty in the cross section measurement arising from uncertainties in the jet energy scale (see Section 7.4.1). The selection requirements are listed below.

1. Trigger selection: All events must pass a single muon trigger. For early running, the HLT_Mu9 trigger was used for selection (requiring an HLT muon with p_{T} greater than 9 GeV). Eventually this trigger was prescaled, and selection was instead performed using the (unprescaled) HLT_Mu15 trigger (p_{T} greater than 15 GeV). All simulated events use HLT_Mu9.

2. Primary vertex: All events must have a good primary vertex. A primary vertex is considered "good" if each of the following conditions are met:

   - there are more than four degrees of freedom in the fit;
   - distance on the beam axis $|z|$ <24 cm;
   - radius in the transverse plane $r < 2.0$.

3. Exactly one good isolated muon is present in the event. A muon is considered to be "good" and "isolated" if each of the following conditions are met:

   - is reconstructed by both the global muon and tracker muon algorithms;
   - $p_{\mathrm{T}}^{\mu} > 20$ GeV;
   - $|\eta| < 2.1$;
   - RelIso<0.05, where RelIso is the relative isolation, defined as $\sum_{\Delta R(\mu)<0.3}(p_{\mathrm{T}}^{\mathrm{track}} + E_{\mathrm{T}}^{\mathrm{ECAL\ cells}} + E_{\mathrm{T}}^{\mathrm{HCAL\ cells}})/p_{\mathrm{T}}^{\mu}$, i.e., the relative transverse momentum and energy of tracks and calorimeter cells within $\Delta R < 0.3$ of the muon;
   - the global track has at least one hit in the muon chambers, muon chamber track segments formed from at least two muon stations, and the fit has normalized χ^2 <10;
   - the inner track has at least ten hits in the silicon tracker and at least one hit in the pixels;
   - $\Delta R(\mu, \mathrm{jet}) > 0.3$ for all jets passing the jet selection requirements in Step 6 below;

- $|d_0(\text{beamspot})| < 0.02$ cm;
- the z distance between the primary vertex and the μ inner track vertex is less than 1 cm.

4. No second muon ($p_{\text{T}} > 10$ GeV, $|\eta| < 2.5$, RelIso< 0.2) may be present in the event.

5. No electron ($E_{\text{T}} > 15$ GeV, $|\eta| < 2.5$, RelIso< 0.2) may be present in the event.

6. At least 3 particle flow jets must be present in the event, with each jet satisfying the following requirements:
   - $p_{\text{T}} > 30$ GeV;
   - $|\eta| < 2.4$;
   - composed of more than one particle candidate;
   - composed of at least one charged particle candidate;
   - charged electromagnetic energy fraction $< 99\%$;
   - neutral electromagnetic energy fraction $< 99\%$;
   - neutral hadronic energy fraction $< 99\%$;
   - charged hadronic energy fraction $> 0\%$.

The expected number of events as determined from simulated $t\bar{t}$ and background samples after each of the above selection cuts is shown in Table 5.4. The

expected event yields are scaled to an integrated luminosity of 36.1 pb^{-1} and compared with the actual number of events seen in data, where expected event yields are calculated using the predicted cross sections from theory. In this analysis, no other direct cuts beyond those outlined above are applied. Instead, to achieve additional discrimination between signal and background events, kinematic variables are fed into an artificial neural network, described in the following chapter.

Selection	MC Sample						
	$t\bar{t}$+jets	W+jets	Z/γ^*+jets	QCD	Top-s	Top-t	Top-tW
All	1306182	14805546	2543727	29504866	494967	484060	494961
HLT, PV	411062	3313802	79183	25023858	180293	168013	135302
Good μ	158071	2147221	201383	99224	91219	94514	63604
e veto	136895	2142659	196648	98377	89826	92681	54568
≥ 1 jets	135642	269838	34271	28984	81535	82358	52583
≥ 2 jets	124591	45438	5829	3047	49886	43829	41377
≥ 3 jets	90289	7395	1092	399	11945	13308	21158
≥ 4 jets	44100	1431	215	64	2026	3314	6282

Table 5.3: Simulated event counts for each selection cut. The HLT, PV selection cut includes cuts 1 and 2, requiring the event pass the high level trigger and contain a good primary vertex. The good μ selection cut includes cuts 3 and 4, requiring a good, isolated muon and no second loose muon.

Selection	MC Sample					Data
	$t\bar{t}$+jets	V+jets	QCD	Top	MC total	
HLT, PV	1792	287769	2595872	315	2885747	7718212
Good μ	672	172871	10293	173	184009	181911
e veto	584	172317	10205	168	183274	181270
≥ 1 jets	579	22113	3007	150	25848	28011
≥ 2 jets	532	3726	316	84	4658	5535
≥ 3 jets	387	613	41	27	1068	1271
≥ 4 jets	189	119	7	7	321	344

Table 5.4: Number of expected simulated and measured data events at 36.1 pb^{-1}. The HLT, PV selection cut includes cuts 1 and 2, requiring the event pass the high level trigger and contain a good primary vertex. The good μ selection cut includes cuts 3 and 4, requiring a good, isolated muon and no second loose muon.

5.6 Chapter highlights

The raw data taken by CMS is processed using digitization and reconstruction algorithms. Digitization algorithms turn voltages and photon counts from the trackers and calorimeters into hits and energy deposits, respectively. Event reconstruction uses the digitized hits and energy deposits in different CMS subdetectors to identify particles in the event, such as muons, jets, and jets originating from b quarks. Event reconstruction also estimates values for various particle characteristics, such as position, momentum, and energy.

Simulated data is generated from MadGraph matrix elements, which are squared to yield the probability of events with given initial and final states. Monte Carlo techniques are employed to generate datasets of simulated events with composition represented by the probabilities. Pythia takes the output from MadGraph and simulates parton showering and beam remnants. The interactions of the generated particles with the CMS detector are simulated by GEANT, producing simulated raw data. Simulated raw data is digitized and reconstructed by the same algorithms that process the real raw data.

Events containing an isolated, high-p_{T} reconstructed muon and three high-p_{T} reconstructed jets are accepted for further analysis. Simulations predict that 387 $t\bar{t}$, 613 V+jets, 41 QCD, and 27 single top events in 36.1 pb^{-1} of data are expected to pass selection by assuming Standard Model cross sections. 1,271 events passing selection are observed by CMS and will be used in a multivariate machine learning analysis designed to discriminate between events that originate from the $t\bar{t}$ signal or background, which will be described in the next chapter.

Chapter 6

Analysis techniques

We employ a machine learning algorithm to discriminate between signal and background events using event kinematics as input. The algorithm, an artificial neural network, uses parameterized underlying functions to approximate the multidimensional relationship between the inputs and output discriminant. Simulated data are used to iteratively calculate the ANN parameter values with the backpropagation algorithm, termed "supervised learning".

Due to uncertainties in the simulation, we use events in orthogonal control regions to model some background processes in the nominal signal region. Specifically, events with exactly two jets are used to model V+jets events passing nominal selection while events containing a muon with reversed isolation are used to model QCD events passing nominal selection.

We choose input variables to the ANN, which must provide signal and background discrimination, have distributions that are similar in the nominal and orthogonal event samples, and have small measurement uncertainty. Sources of measurement uncertainty are discussed in Section 7.4.

6.1 Signal and background discrimination

Machine learning algorithms can be compared to a maximum likelihood fit. Both methods tune internal parameters in a model by using the expected distribution of measurements predicted by the model. While maximum likelihood fits can be performed using multiple measured characteristics, they are typically done assuming the measurements are independent, thereby losing all correlation information. The measurements are assumed to be independent because a genuinely multidimensional maximum likelihood fit requires many simulations in order to yield the desired precision in the multidimensional space. Specifically, if using a binned likelihood fit with B bins in D dimensions, the number of simulations required to preserve the desired precision grows as B^D. Machine learning algorithms reduce the required number of simulations by approximating the multidimensional relationship between the inputs and outputs using a parameterized method. The required number of simulations scales with the number of parameters in the algorithm, which is typically only linearly coupled with the number of inputs and outputs.

Simulated $t\bar{t}$ events are given a training target of 1 while all other simulated events are given a training target of 0. A machine learning algorithm takes user specified inputs such as particle kinematics. The algorithm then "learns" or "trains" by using simulated events to iteratively tune internal parameters to yield discriminant values close to 1 for $t\bar{t}$ events and close to 0 for background events. The discriminant value for an event may be loosely interpreted as the probability that the event is signal, though the method employed does not strictly yield values between 0 and

1. After training, the discriminant is calculated for observed events. To test the validity of the method, the discriminant is also calculated for simulated events that were not used for training.

An artificial neural network is used to distinguish signal events from background events after the event selection is applied. An ANN approximates the multidimensional relationship between the inputs and outputs using a superposition of parameterized underlying functions. Historically, an ANN was thought to model the action of neurons in the brain, hence the origin of the name. For classification, a natural choice of underlying function is the nonlinear sigmoid function, shown in Equation 6.1, which is a continuous approximation of a step function.

$$f(x) = \frac{1}{1 + e^{-ax}} \tag{6.1}$$

Input neurons pass experimentally measured event characteristics to other neurons in the ANN. For binary classification, an output neuron is trained to calculate values close to either 0 or 1. Every other neuron in the network passes $f(x)$ to other neurons in the network, where x is a shifted sum of weighted outputs from other neurons.

A multi layer perceptron (MLP), a type of ANN, is constructed in layers, with the input neurons (or perceptrons) in the first layer, the output neurons in the last layer, and a specified number of hidden layers with a specified number of neurons in each. The input to each hidden and output neuron is a shifted sum of weighted outputs of the neurons in the previous network layer. Using the notation of [64], the network consists of $L+1$ layers, where the r^{th} layer contains k_r neurons. For each event i of N training events, with simulated event characteristics $x_j(i)$,

Equation 6.2 gives the neuron output values, $y_j^r(i)$, and Equation 6.3 gives the neuron input values, $v_j^r(i)$, where j is the index of the neuron in the layer and r is the layer index. The output values of the network are $y_j^L(i) = f(v_j^L(i))$. The values of the weights, w_{jk}^r, are iteratively calculated using the backpropagation algorithm. The action of a single neuron is represented by Figure 6.1.

$$y_j^0(i) = x_j(i), \qquad i = 1, ...N, j = 1, ..., k_0,$$

$$y_j^r(i) = f(v_j^r(i)), \qquad i = 1, ...N, r = 1, ..., L, j = 1, ..., k_r, \tag{6.2}$$

$$v_j^r(i) = w_{j0}^r + \sum_{k=1}^{k_{r-1}} w_{jk}^r y_k^{r-1}(i), \qquad i = 1, ...N, r = 1, ..., L, j = 1, ..., k_r \tag{6.3}$$

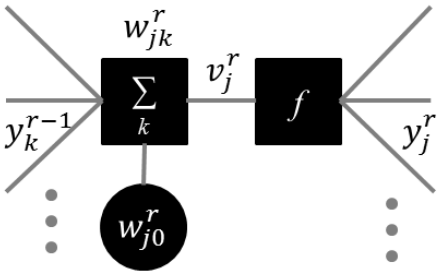


Figure 6.1: The output of a single neuron (Equation 6.2) is the value of a function (Equation 6.1) given a shifted weighted sum of the neuron inputs (Equation 6.3) as the function argument.

The MLP is implemented via the Root TMVA::MethodMLP package [41] and is wrapped by the MVAFramework utility within CMSSW [57]. A total of three kinematic variables are fed into the MLP, chosen from roughly forty considered. We chose to use an MLP structure with two hidden layers, with 5 neurons in one

layer and 4 neurons in the other, i.e., $k_0 = 3$, $k_1 = 5$, $k_2 = 4$, and $k_3 = 1$. A dual hidden layer architecture was chosen because one input is discontinuous (a boolean) and because dual hidden layer architectures are thought to be more stable than single hidden layer architectures with the same number of nodes [41]. The number of hidden nodes is determined by the number of simulated events using the rule of thumb of no more than one network parameter per 20-40 training events [59], though performance was seen to be relatively independent of small changes in the network topology. The visual representation of the network is shown in Figure 6.2.

Each signal and background sample is split into two independent subsamples for use with the MLP. One set of subsamples, termed the training sample, is fed into the neural network to train it. The second set, termed the testing sample, is used to generate test pseudo-experiments, described in Section 7.2, which make use of the trained network. Training is performed using 750 cycles to produce converged and stable discriminant values without overtraining, as recommended in TMVA documentation.

Training is normally performed with a mixed sample, where the event weights multiplied by the number of simulated events gives the event yields expected in data at an arbitrary luminosity. To test stability, we used a modified training sample containing only $t\bar{t}$ signal and V+jets background events (since V+jets events are the largest background), with event weights adjusted to provide a signal to background ratio of 1:1, which is significantly different from the expected $t\bar{t}$:V+jets ratio of 1:1.7, seen in Table 5.4. Performance of the MLP was not impacted by the change in relative numbers of training events.

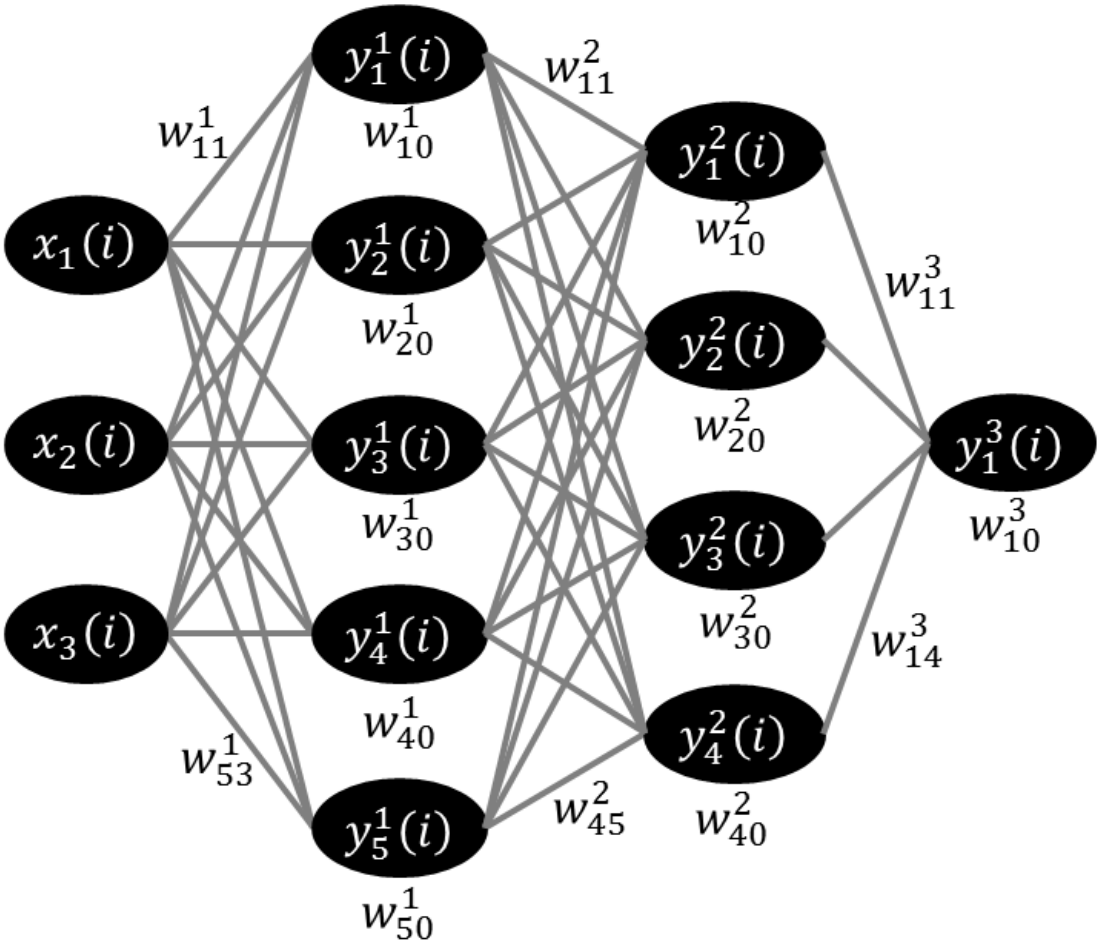


Figure 6.2: The MLP used for this analysis. Weights of form w_{j0}^r correspond to the neuron shift, while the remaining weights correspond to neuron-to-neuron edge weights. For example purposes, only some of the weights are explicitly labeled. Each output value y is given by Equation 6.2 and shown schematically in Figure 6.1.

6.1.1 The backpropagation algorithm

The network weights are determined during training using the backpropagation algorithm. The algorithm iteratively updates the network weights using steepest descent minimization of a cost function. If the training targets for event i are $y_j(i)$ and the network output is $y_j^L(i)$, then the error is given by the difference between the training target and the calculated network output, $y_j(i) - y_j^L(i)$. We use the cost function shown in Equation 6.4, a commonly used cost function for MLPs. Minimizing the cost function is equivalent to minimizing error accumulated over all events and outputs. Note that with one network output, $k_L = 1$.

$$C = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{k_L} (y_j(i) - y_j^L(i))^2 \tag{6.4}$$

The backpropagation algorithm iteratively propagates the cost function in the output layer back to previous layers by weighting the cost function with the network weights between the layers. The following pseudocode is taken with minor modifications from [64]:

- Initialization: Initialize all the ANN weights with small random values.
- Forward computations: For each of the training input vectors $\mathbf{x}(i) = [x_1(i), x_2(i), ..., x_{k_0}(i)]^T$, $i = 1, 2, ..., N$, compute all $v_j^r(i)$ and $y_j^r(i) = f(v_j^r(i))$, $r = 1, 2, ..., L$, $j = 1, 2, ..., k_r$ from Equation 6.3. Compute the cost function for the current estimate of weights from Equation 6.4.
- Backward computations: For each $i = 1, 2, ..., N$, $r = L, L-1, ..., 2$, and

$j = 1, 2, ..., k_r$, compute

$$\delta_j^L(i) = (f(v_j^L(i)) - y_j(i))f'(v_j^L(i))$$

$$\delta_j^{r-1}(i) = (\sum_{k=1}^{k_r} w_{jk}^r \delta_k^r(i))f'(v_j^{r-1}(i)).$$

- Update the weights: For learning rate μ, $\mathbf{w}_j^r = [w_{j0}^r, w_{j1}^r, ..., w_{jk_{r-1}}^r]^T$, $\mathbf{y}^{r-1}(i) = [1, y_1^{r-1}(i), ..., y_{k_{r-1}}^{r-1}(i)]^T$, $r = 1, 2, ..., L$, and $j = 1, 2, ..., k_r$, compute

$$\Delta \mathbf{w}_j^r = -\mu \sum_{i=1}^{N} \delta_j^r(i) \mathbf{y}^{r-1}(i)$$

$$\mathbf{w}_j^r(\text{new}) = \mathbf{w}_j^r(\text{old}) + \Delta \mathbf{w}_j^r.$$

The derivation of $\delta_j^r(i)$ is performed by differentiating the cost function with respect to the network weights. Steepest descent is only capable of finding local minima. To circumvent this issue, TMVA's implementation uses sequential backpropagation, which updates the network weights with each training event instead of using the cost function summed over all events. The weights are then updated according to $\mathbf{w}_j^r(i+1) = \mathbf{w}_j^r(i) - \mu \delta_j^r(i) \mathbf{y}^{r-1}(i)$. Events in the training set are also sampled in random order, where a single training epoch samples the entire training set once. TMVA recommends 500 or more training epochs with learning parameter $\mu = 0.02$, with exponential decay rate of 0.01.

Convergence of the network weights tends to be faster if the optimal values do not span many orders of magnitude. If the inputs have widely varying numerical ranges of values, the network weights must also widely vary. Additionally, if the inputs have a similar numerical range of values, then the network weights represent the relative importance of the inputs with respect to the output. To ensure a similar

numerical range of values between inputs, it is common practice to normalize the inputs, usually by their probability distribution function, so that the input values are constrained to be between 0 and 1. With the probability distribution function of input x denoted as $p(x)$, the input is transformed to $\tilde{x} = \int_{-\infty}^{x} p(x')dx' / \int_{-\infty}^{+\infty} p(x')dx'$. The normalization routine is implemented within the MVAFramework package.

6.1.2 Input variables

While MLPs are mathematically robust under irrelevant or duplicated inputs, they tend to converge more slowly in these cases. When using a fixed number of training epochs, it is safest to keep only those inputs that significantly impact the output. Therefore, it is common practice to progressively add or subtract network inputs, keeping only those inputs with relatively large network weights.

Approximately fourty kinematic variables were considered as inputs. To be deemed a viable input candidate, a variable must fulfill the obvious requirement of providing some discriminating power between $t\bar{t}$ signal and background processes. Additionally, some variables are measured with more uncertainty than others, impacting the uncertainty of the discriminator distribution. Finally, because this analysis uses events in orthogonal control regions to model some background processes in the nominal signal region, the candidate input variable distribution must be similar in the nominal and orthogonal event samples.

A subset of the approximately forty variables considered as inputs includes:

- the transverse momenta p_{T}^{j} for the three highest p_T jets in the event;

- ΔR_{ij} for various jet combinations i and j;
- the presence of at least one b tagged jet in the event;
- H_T^{30}, the scalar sum of transverse energies for all jets with $p_{\mathrm{T}} > 30$ GeV;
- p_{T}^{μ}, the transverse momentum of the muon;
- $|\eta^{\mu}|$, the normalized eta position of the muon;
- RelIso, muon relative isolation.

Each of the jet p_{T}^{i} and H_T^{30} variables provide some discrimination between $t\bar{t}$ signal and background processes. However, events with two jets have jet p_{T}^{i} distributions that are considerably different than V+jets events with three or more jets. Similarly, we consider b tags on only the two leading jets.

Muon combined relative isolation provides strong discrimination between $t\bar{t}$ and QCD processes; so strong, in fact, that it is used to define the orthogonal data region from which the QCD contribution is modeled. Since the variable defines the difference between the signal region and QCD control region, it cannot be used as an input to model QCD in the nominal selection region. Similarly, the muon transverse momentum differs significantly between the signal and QCD control region and thus is not an appropriate choice for an input variable.

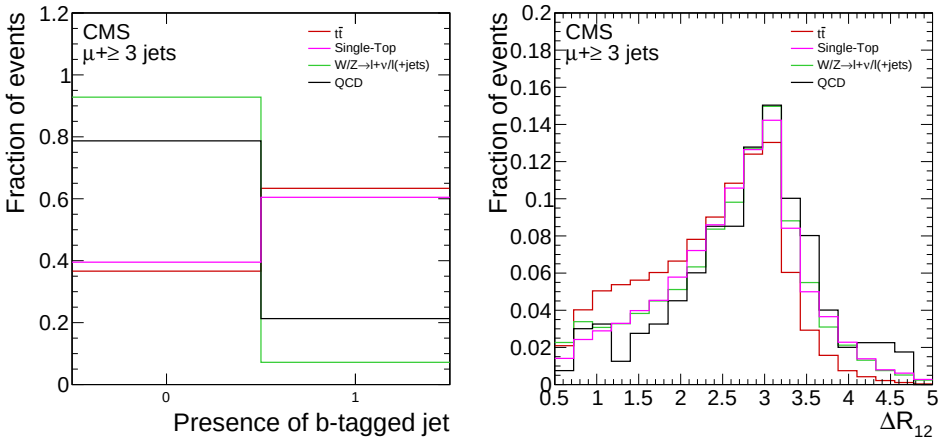
Other variables were considered and discarded for similar reasons. We have chosen a total of three variables that satisfy the necessary requirements for inputs to the neural network:

- A boolean variable indicating if at least one of the two leading jets in the event is tagged as a b.
- The angular separation ΔR between the two leading jets in the event, ΔR_{12}.
- The position $|\eta|$ of the reconstructed muon, $|\eta^{\mu}|$.

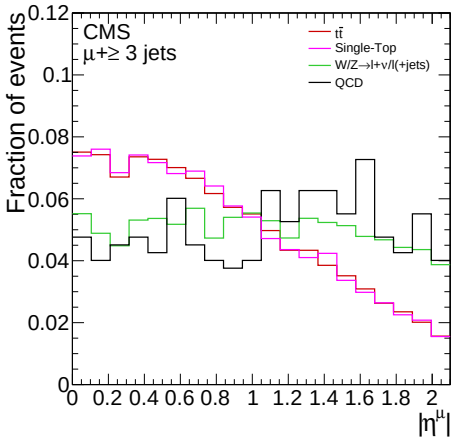
Since $t\bar{t}$ events contain two b jets, one might expect that using the number of tagged jets, rather than a simple boolean, might provide better discrimination between $t\bar{t}$ signal and background events containing a single b jet. However, it was found that using the number of b jets as a neural net input greatly increases the sensitivity of the output to the uncertainty of tagging efficiency, discussed in Section 7.4.2. Thus, the boolean tag variable is used instead. This boolean still provides strong discrimination between signal and background, as shown in Figure 6.3. The remaining variables, ΔR_{12} and $|\eta^{\mu}|$, each exhibit some difference in their distribution between signal and background events (Figure 6.3).

6.1.3 Discriminator

Discriminator output templates for the simulated training samples are shown in Figure 6.4. The $t\bar{t}$ template exhibits a dual-peak structure, with the bulk of events distributed in the region between [0.5-1.0], but a significant fraction of events in the region [0-0.5]. This structure arises primarily from the use of the b tag as a neural net input. Since the b tagging algorithm is not fully efficient at identifying real b jets and we do not consider all jets, some fraction of $t\bar{t}$ events report 0 b jets, resulting in the low values in the signal discriminator template.



(a) (b)



(c)

Figure 6.3: Inputs to the neural network for simulated events: (a) b tag boolean, (b) ΔR_{12}, and (c) $|\eta^{\mu}|$. Simulated $t\bar{t}$ events are in red, single top events in pink, V+jets events in green, and QCD events in black. All distributions are normalized to unity to show shape comparisons.

The single top background template in Figure 6.4 is similar in shape to the $t\bar{t}$ signal template. This is unsurprising, given the similarities in kinematics between the two event types. Single top events, unlike other backgrounds, are expected to always produce b jets and it has already been seen that the b tag input dominates the value of the discriminant output. Therefore single top events are not used during training since the labeling of single top events as background may cause the neural network to give $t\bar{t}$ events similar to single top events artificially low values. The expected number of single top events passing selection is an order of magnitude smaller than that of $t\bar{t}$ events.

6.2 Using data to correct or replace simulations

Due to the uncertainty of simulating colored interactions, techniques have been developed to use information from data to either correct or replace simulated events. While the decay of the B hadron itself is well understood, the jet environment in which the decay occurs is difficult to simulate. Therefore, the efficiency of identifying real b jets and the misidentification ("fake") rate of tagging non-b jets must be taken from data. Because of this, a tag boolean is used as input to the ANN rather than the tag value itself – the data-corrected values of efficiency and fake rate are calculated only at specific operating points. The efficiency and fake rate of tagging jets is measured from data and applied to the simulations of $t\bar{t}$ and single top events.

QCD (jet only) and V+jets events passing nominal event selection are modeled by events in real data passing orthogonal selection instead of by simulated data.

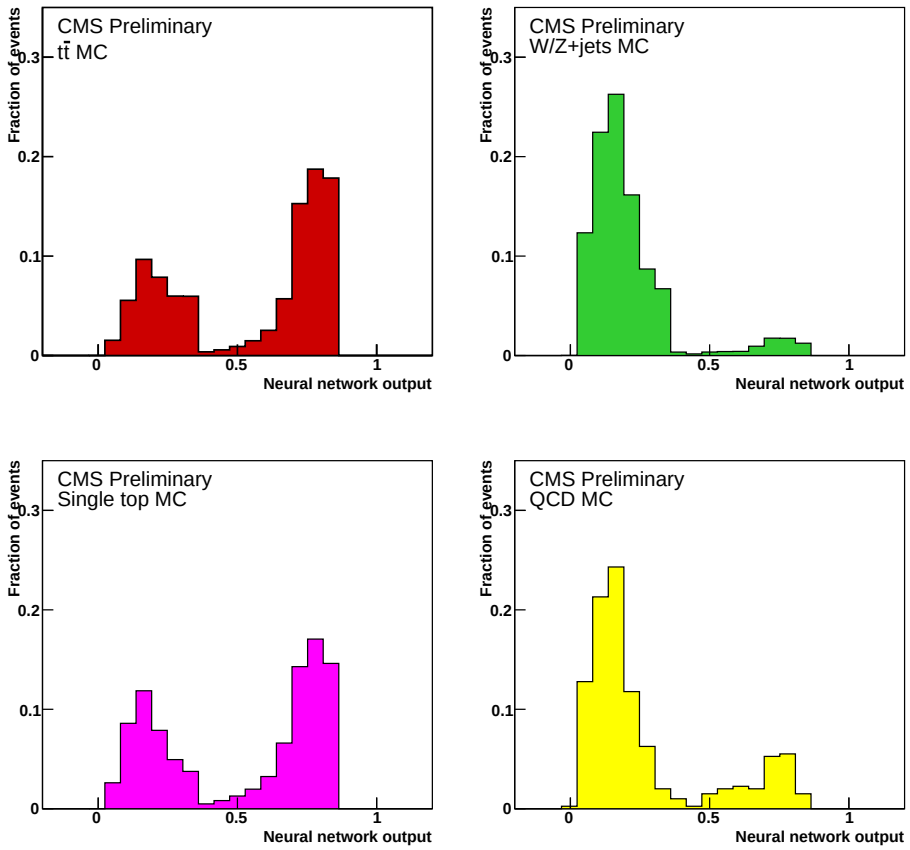


Figure 6.4: Normalized simulated discriminator templates for (top left) signal $t\bar{t}$, (top right) W/Z+jets background, (bottom left) single top background, and (bottom right) QCD background.

Using events from data provides a much larger sample of QCD-like events than can be produced from simulations, as well as data-driven modeling of the QCD ANN inputs. W events produced with additional jets are the largest source of background. Although the decays of the W and Z bosons are well understood, V events produced with additional heavy (b) jets are especially significant because the shape of the discriminator is strongly dependent on the b tag boolean input. Since the proportion of heavy and light jets in the V+jets simulated sample is subject to the same uncertainties present in QCD simulations, V+jets events passing nominal event selection are modeled by orthogonal events in real data.

The $t\bar{t}$ and single top discriminants are taken from simulations, but the simulated events are corrected by the b tag efficiency, b tag fake rate, and jet energy resolution measured in data ("data-corrected"). The discriminator templates for the V+jets and QCD backgrounds are produced directly from data ("data-driven"). Events passing orthogonal event selection of a muon with isolation greater than 0.1 (instead of less than 0.05) are used to form the QCD discriminant. Events passing orthogonal event selection of exactly two jets (instead of three or more) are used to form the V+jets discriminant. The data- driven or corrected ANN discriminator distributions are shown in Figure 6.5.

6.2.1 $t\bar{t}$ and single top templates

Tag rates are different for jets produced by b quarks, c quarks, light quarks (u, d, s), or gluons (g). Jets reconstructed in simulations are identified by the particle

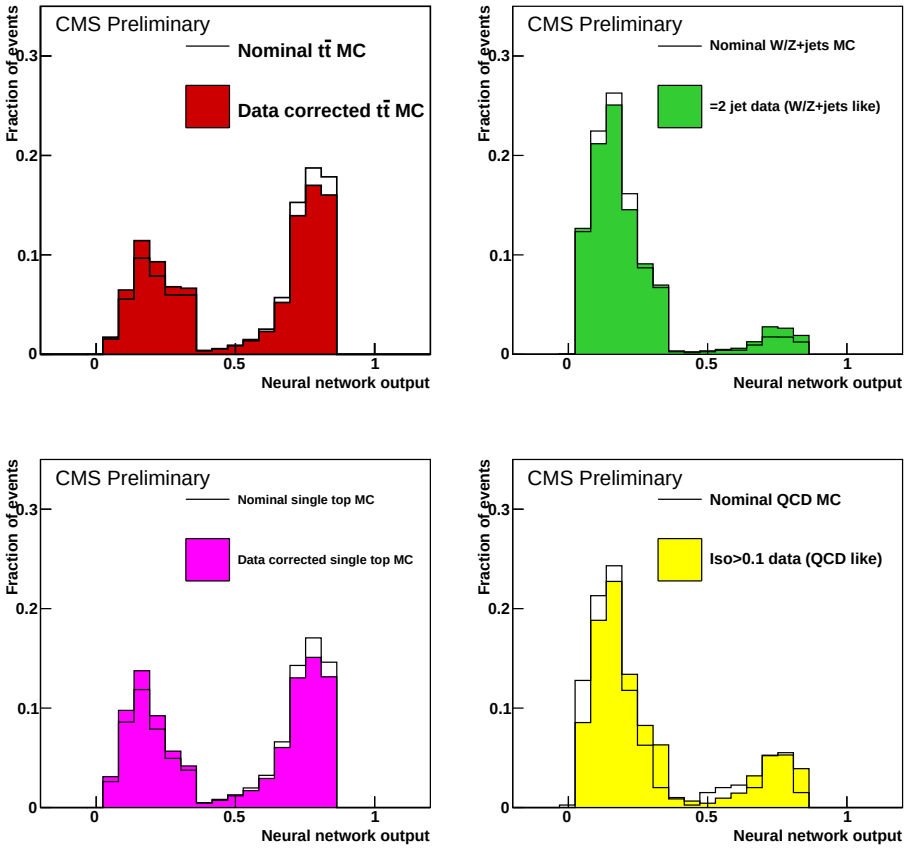
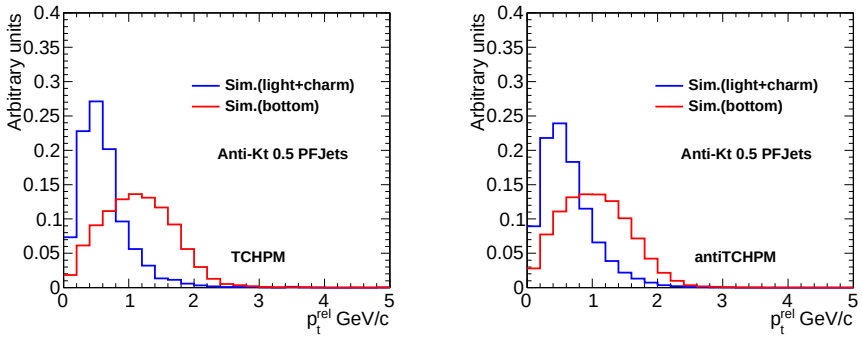


Figure 6.5: Data- driven or corrected discriminator distributions for (top left) data-corrected $t\bar{t}$ (corrected tag rates and jet energy resolution), (bottom left) data-corrected single top (corrected tag rates and jet energy resolution), (top right) data-driven $V+$ jets (events with exactly two jets), and (bottom right) data-driven QCD (events with muon isolation greater than 0.1). The discriminator distributions of simulated events passing nominal selection are overlaid for reference, though the distributions are not expected to agree.

that produced them [18]. Scale factors are calculated by measuring the tag rate of b-like and light-parton-like (u, d, s, or g) jets in data and dividing by the tag rate in simulations. The scale factors are then applied to simulated $t\bar{t}$ and single top events. The description of the calculation of the scale factors in the following paragraphs is taken from [18].

The B meson decay products include a muon in 11% of b jets, or in 20% of b jets including b to c cascades. Therefore, b jets can be identified in data by requiring a reconstructed muon inside the jet. Events are required to contain a muon that passes similar selection requirements as used in the nominal selection, but RelIso is ignored and the muon is required to be within a ΔR of 0.4 of a jet. Jets can fake muons by penetrating to the muon chambers or via pion decay, as discussed in Section 5.3.3. To determine the efficiency of tagging b jets from data, the fraction of tagged jets with muons must be adjusted to account for the fraction of jets with fake muons. The momentum of the muon in the plane transverse to the jet axis, $p_{\mathrm{T}}^{\mathrm{rel}}$, tends to be larger for b jets than for light or c jets, as shown in Figure 6.6. The fraction of jets with fake muons is determined using a maximum likelihood fit of the $p_{\mathrm{T}}^{\mathrm{rel}}$ distributions of tagged and untagged jets.

The scale factor for b tag efficiency, $\mathrm{SF}_b = e_b^{\mathrm{data}}/e_b^{\mathrm{MC}}$, is given by the tag efficiency measured in data, e_b^{data}, and the tag efficiency measured in simulations, e_b^{MC}, where e_b^{data} and e_b^{MC} are measured on events with a non-isolated muon. The scale factor can be applied to jets in different environments, such as the dense environment of a $t\bar{t}$ decay, specifically, $e_b^{t\bar{t}} = \mathrm{SF}_b e_b^{t\bar{t}\ \mathrm{MC}}$. The primary sources of uncertainty in SF_b are due to uncertainty in the fraction of jets with fake muons



(a) (b)

Figure 6.6: The $p_{\mathrm{T}}^{\mathrm{rel}}$ distribution of muons associated with b jets (red) and light or c jets (blue) for (a) jets tagged by the track counting high purity algorithm medium cut (TCHPM) and (b) jets not tagged by TCHPM. Credit: [18].

and in the shape of the $p_{\mathrm{T}}^{\mathrm{rel}}$ distribution. For the track counting high efficiency medium operating point, $SF_b = 0.900 \pm 0.135$.

The light jet fake rate is calculated by using tracks with negative impact parameters or secondary vertices with negative decay lengths. The track impact parameter is negative if $\mathbf{p}_{\mathrm{T}}^{\mathrm{track}} \cdot (\mathbf{d_0} - \mathbf{PV})$ is negative, e.g., if the dot product of the track transverse momentum vector and the vector pointing from the primary vertex to the point of closest approach in the transverse plane is negative. Likewise, the secondary vertex decay length is negative if $\mathbf{p}^{\mathrm{jet}} \cdot (\mathbf{SV} - \mathbf{PV})$ is negative. For the track counting b tag algorithms, the track ordering is inverted such that the tracks with the most negative impact parameter significance are used to calculate the tag discriminant. For the simple secondary vertex b tag algorithms, only those vertices with negative decay lengths are used. The negative and nominal tag discriminants

shown in Figure 6.7 demonstrate that most jets with negative impact parameter or secondary vertex decay length originate from light quarks or gluons.

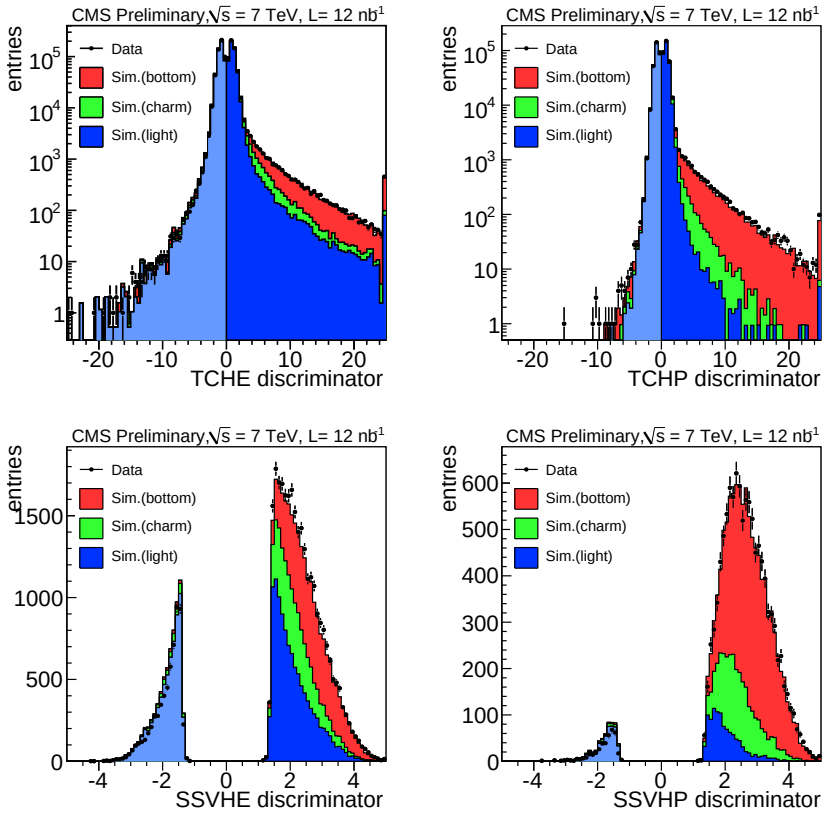


Figure 6.7: The b tag distributions for negative and nominal algorithms. The data are shown as points and simulations as colored stacked histograms, with b jets shown in red, c jets in green, and light jets in blue. The negative tag discriminant values are indicated by lighter shading. Credit: [18].

The negative tag rate in data, e_{-}^{data}, is the number of negative tagged jets divided by number of jets. The ratio between the tag rate of light jets and the negative tag rate of all jets, $R_l = e_l^{\text{MC}}/e_{-}^{\text{MC}}$, is determined from simulations. The

fake rate observed in data is then approximated by assuming $e_l^{\text{data}} = e_-^{\text{data}} R_l$ and the scale factor is $\text{SF}_l = e_l^{\text{data}}/e_l^{\text{MC}}$. The uncertainty of R_l is the primary source of uncertainty in e_l^{data} and SF_l. For TCHEM, SF_l is calculated as a function of jet p_{T} and η, with values ranging from 1.06 to 1.32 with uncertainty ± 0.13 to ± 0.28.

The scale factors are used to correct b tag rates in the $t\bar{t}$ and single top simulations. The simulated efficiencies for tagging b, c, light (u, d, or s), and g jets are binned in p_{T}-η space for the jets from $t\bar{t}$ and single top events that pass selection. For each simulated jet passing selection requirements, a random number between 0 and 1 is generated. If this random number is less than the scale factor multiplied by the efficiency given by the jet's flavor, p_{T}, and η, then the jet is treated as a b tagged jet. SF_b is used to correct b and c jet tag efficiencies, while SF_l is used to correct light and g jet tag efficiency. In the case where the flavor of a simulated jet cannot be determined (occurs in less than 0.2% of selected jets), the jet is treated as coming from a gluon.

Additionally, studies [20] indicate the jet energy resolution in the data is 10% worse than in the simulations. An additional correction is applied to the $t\bar{t}$ and single top simulated samples to account for this.

6.2.2 QCD template

The QCD background template is derived by reversing one of the nominal selection cuts in order to define a region that is dominated by QCD with minimal

contamination from $t\bar{t}$ signal or other backgrounds. The nominal muon isolation cut is reversed and made larger, i.e., only events in which the lead muon's RelIso value is greater than 0.1 (instead of less than 0.05) are selected and used to model QCD events in the signal region. The ΔR cut between the muon and all candidate jets remains at 0.3, as in the nominal selection. QCD events produce the bulk of events with isolation greater than 0.1, as shown in Figure 6.8. Simulations predict that 97% of events passing this selection are QCD.

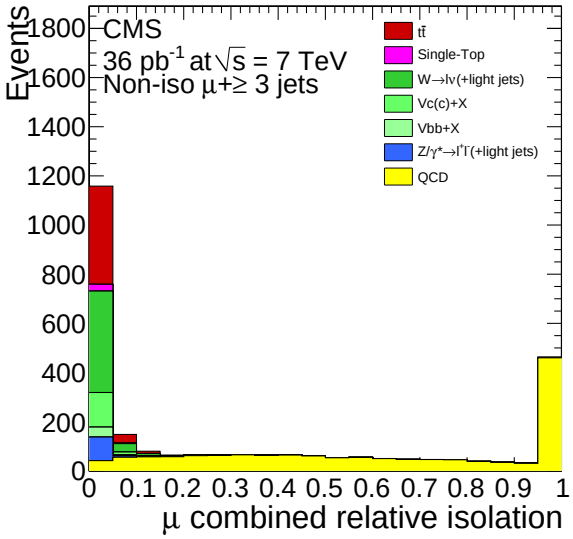


Figure 6.8: Distribution of muon isolation from all signal and background processes considered in this analysis for events passing all selection criteria except the requirement on muon isolation. The last bin contains events with values of RelIso greater than 1 (overflow). The number of predicted events is scaled by a luminosity of 36 pb^{-1}.

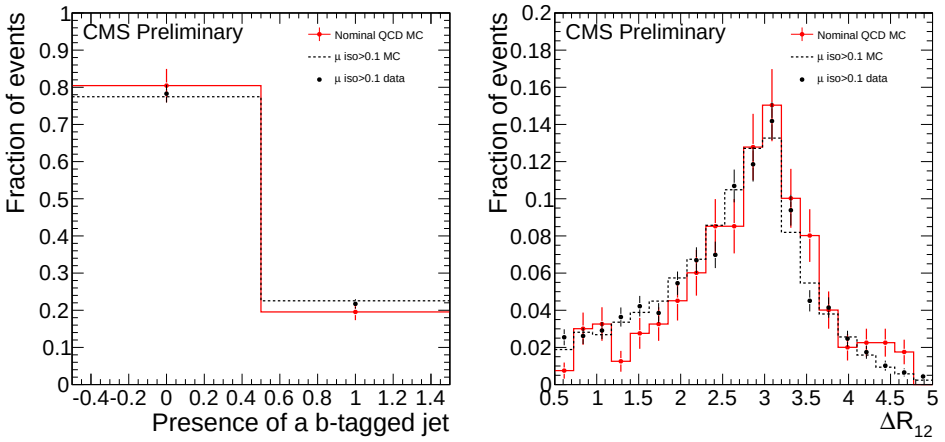
Data events passing the reversed isolation cut are then used to approximate the output of the neural network for QCD events. The neural net input distributions are

compared between the QCD events in the nominal selection region and events in the orthogonal data-driven region. These distributions are shown in Figure 6.9 and are found to be similar within the available statistics. For the purpose of comparing the simulated distributions to the data, the same data corrections applied to the $t\bar{t}$ and single top templates are also applied to simulated events in both the orthogonal and nominal regions. These corrections are not applied to the final neural network template used for QCD to determine the $t\bar{t}$ cross section, since this template is generated from inputs measured in data.

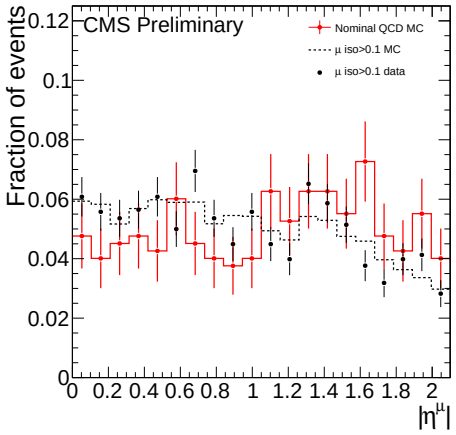
6.2.3 V+jets template

The fraction of events containing heavy flavor (b or c) quarks in V+jets events may not be well modeled by the simulations. To safely handle heavy flavor content in V+jets events, events taken from data are used to approximate V+jets events passing nominal selection, similar to the technique used for QCD. The nominal event selection is altered to require exactly two jets with $p_{\mathrm{T}} > 30$ GeV (instead of three or more). As can be seen in Figure 6.10, V+jets produce the bulk of events with exactly two jets. Simulations predict that 87% of two jet events are V+jets, 7% QCD, and 6% $t\bar{t}$ or single top.

Again, the neural net input distributions are compared for V+jets events passing the nominal selection and events in the orthogonal data-driven selection region. These distributions are shown in Figure 6.11 and are found to be similar within the available statistics. For the purpose of comparing the simulated distributions



(a) (b)



(c)

Figure 6.9: Normalized inputs to the neural network for simulated QCD events passing nominal event selection (solid red line), simulated events with muon isolation greater than 0.1 (dashed line), and data events with muon isolation greater than 0.1 (points) for (a) the b-tag boolean, (b) ΔR_{12}, and (c) $|\eta^{\mu}|$.

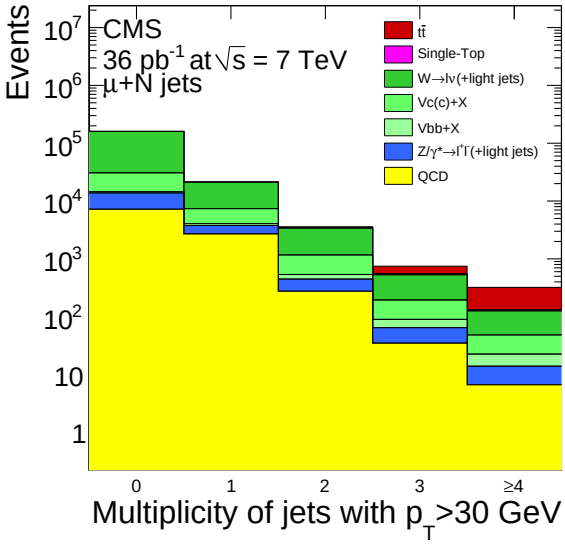


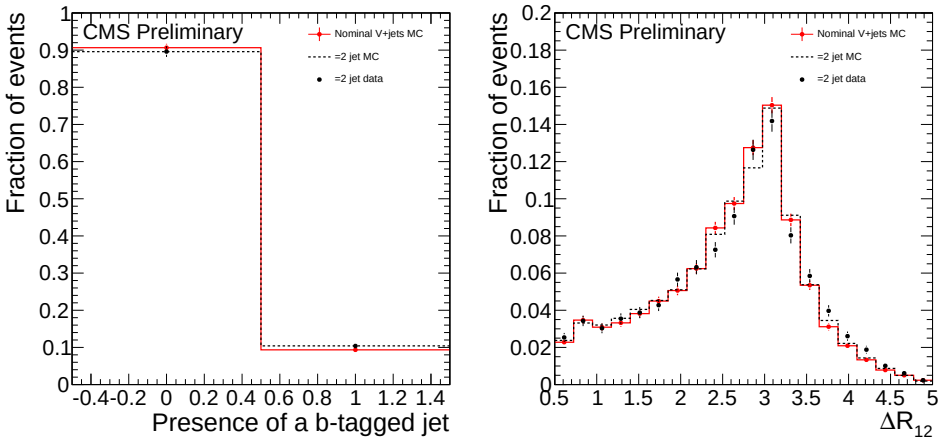
Figure 6.10: Number of jets with $p_{\mathrm{T}} > 30$ GeV from all signal and background processes considered in this analysis for events passing all selection criteria except the requirement of three or more jets. The number of predicted events is scaled by a luminosity of 36 pb^{-1}.

to the data, the same data corrections applied to the $t\bar{t}$ and single top templates are also applied to simulated events in both the orthogonal and nominal regions. Additionally, results from other $t\bar{t}$ μ+jets analyses utilizing a b tag [22] indicate that V+jets events containing a b jet occur at roughly twice the predicted rate (k_b=2), which is also applied to the simulated events. These corrections are not applied to the final neural network template used for V+jets to determine the $t\bar{t}$ cross section, since this template is generated from inputs measured in data.

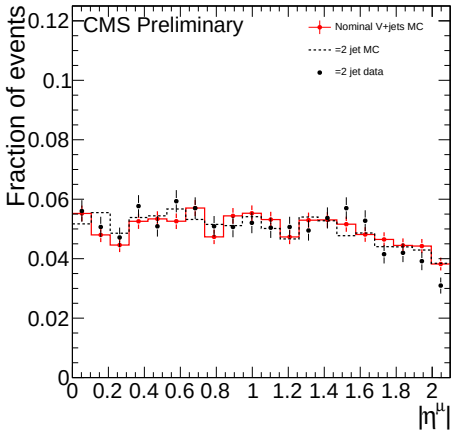
6.3 Chapter highlights

Discrimination between signal and background events is performed using an artificial neural network. The ANN is trained using simulations of $t\bar{t}$, V+jets, and QCD events. The ANN takes as input a b tag boolean indicating a jet in the event was tagged by TCHEM, the angular separation between the two leading jets, and the muon's position $|\eta|$.

Corrections to the jet energy resolution and b tag efficiency observed in data are applied to the simulated $t\bar{t}$ and single top samples. Simulations of QCD are replaced using events taken from data. Jets from b quarks in selected V+jets events generally do not come from the W or Z (V) boson, so are subject to the same uncertainties present in QCD simulations. Since the b tag boolean is an important ANN input, the relative amount of V+heavy flavor cannot be taken from simulations and must instead be modeled using data. Events with muon RelIso>0.1 are used to model QCD events in the signal region while events with two jets are used to model



(a) (b)



(c)

Figure 6.11: Normalized inputs to the neural network for simulated V+jets events passing nominal event selection (solid red line), simulated events with exactly two jets (dashed line), and data events with exactly two jets (points) for (a) the b-tag boolean, (b) ΔR_{12}, and (c) $|\eta^{\mu}|$.

V+jets events in the signal region. Events in these orthogonal selection regions have ANN input variable distributions similar to the QCD or V+jets events in the nominal signal region. The ANN discriminant templates for $t\bar{t}$, single top, V+jets, and QCD events are fit to the discriminant values observed in data to determine the event yield of each, as will be described in the next section.

Chapter 7

Measurement of the cross section and its uncertainty

To measure the cross section, the ANN discriminant for signal and background samples are fit to the data. The fit estimates the number of $t\bar{t}$ events present in the data, which can be converted to cross section using $N = A\epsilon\sigma \int \mathcal{L}dt$. The ANN discriminant fit templates for $t\bar{t}$ and single top are taken from their respective data-corrected simulated samples. Data passing orthogonal selection criteria are used for the V+jets and QCD templates. The statistical and systematic uncertainties of the measured cross section are evaluated using pseudo-experiments generated from simulations of $t\bar{t}$, single top, V+jets, and QCD.

7.1 Fitting

RooFit [65], a fitting package integrated with Root, is used to perform an extended binned likelihood fit of the ANN discriminant. Assuming the N_{data} selected data events are composed of $t\bar{t}$, single top, V+jets, and QCD events with probability distribution functions $P_{t\bar{t}}(x)$, $P_t(x)$, $P_V(x)$, and $P_{\text{QCD}}(x)$, the likelihood of the values of $N_{t\bar{t}}$, N_t, N_V, and N_{QCD} given observations $x_1, ..., x_{N_{\text{data}}}$ is:

$$L(N_{t\bar{t}}, N_t, N_V, N_{\mathrm{QCD}}|x_1, ..., x_{N_{\mathrm{data}}}) =$$
$$\prod_{i=1}^{N_{\mathrm{data}}} \frac{N_{t\bar{t}}P_{t\bar{t}}(x_i) + N_tP_t(x_i) + N_VP_V(x_i) + N_{\mathrm{QCD}}P_{\mathrm{QCD}}(x_i)}{N_{t\bar{t}} + N_t + N_V + N_{\mathrm{QCD}}} \tag{7.1}$$

The likelihood is maximized by iterating over possible values of $N_{t\bar{t}}$, N_t, N_V, and N_{QCD} subject to the constraints that $N_{\mathrm{data}} = N_{t\bar{t}} + N_t + N_V + N_{\mathrm{QCD}}$ and that each value is non-negative. Likelihood terms tend to span many orders of magnitude and maximization is usually done with inverted minimization routines. Therefore, maximum likelihood fits are typically coded as minimization of $-ln(L)$, which monotonically decreases as the likelihood increases.

We fit to the output of the ANN, so the x in the likelihood is the ANN discriminant, or the horizontal axis of the histograms in Figure 6.5. We interpret the unit normalized histograms of the ANN discriminants for $t\bar{t}$, single top, V+jets, and QCD as binned probability distribution functions. That is, the number of histogram entries in a given bin divided by the total number of histogram entries is taken as the integral of the probability density function over the range of x spanned by the bin. We assume the number of events in each bin is Poisson distributed. For bin j, the number of data events in the bin is denoted as $N_{j,\mathrm{data}}$. The number of $t\bar{t}$ histogram entries in bin j divided by the total number of $t\bar{t}$ histogram entries is denoted as $P_{j,t\bar{t}}$. $P_{j,t}$, $P_{j,V}$, and $P_{j,\mathrm{QCD}}$ are defined in a similar manner. The likelihood may thus be written:

$$N_j = N_{t\bar{t}} P_{j,t\bar{t}} + N_t P_{j,t} + N_V P_{j,V} + N_{\mathrm{QCD}} P_{j,\mathrm{QCD}}$$

$$L(N_{t\bar{t}}, N_t, N_V, N_{\mathrm{QCD}} | x_1, ..., x_{N_{\mathrm{data}}}) = \prod_{j=1}^{N_{\mathrm{bins}}} \frac{N_j^{N_{j,\mathrm{data}}} e^{-N_j}}{(N_{j,\mathrm{data}})!} \tag{7.2}$$

Due to the similarity of the QCD fit template to the V+jets fit template and the single top fit template to the $t\bar{t}$ fit template, there is a strong anti-correlation between N_{QCD} and N_V as well as $N_{t\bar{t}}$ and N_t. To improve the convergence of the fit in this scenario, the number of QCD and single top events are constrained. Gaussian (normal distribution) penalty functions are multiplied times the likelihood with means equal to the expected number of QCD and single top events of 41 and 27, respectively, shown in Table 5.4. A 100% width is assigned to the number of QCD events due to expected large uncertainties from the LO cross section. A 30% width is assigned to the number of single top events because, while the theory cross section has smaller uncertainty, it is inappropriate to set the single top uncertainty to be less than the expected precision of the measured $t\bar{t}$ cross section. The final form of the likelihood used is then:

$$L(N_{t\bar{t}}, N_t, N_V, N_{\mathrm{QCD}} | x_1, ..., x_{N_{\mathrm{data}}}) = \prod_{j=1}^{N_{\mathrm{bins}}} \frac{N_j^{N_{j,\mathrm{data}}} e^{-N_j}}{(N_{j,\mathrm{data}})!} \times \left(\frac{1}{\sqrt{2\pi}41} e^{\frac{-(N_{\mathrm{QCD}}-41)^2}{2\times 41^2}} \right) \times \left(\frac{1}{\sqrt{2\pi}(0.3\times 27)} e^{\frac{-(N_t-27)^2}{2\times(0.3\times 27)^2}} \right) \tag{7.3}$$

7.2 Pseudo-experiments

To test the validity of the analysis, 10,000 pseudo-experiments (also termed "Toy Monte Carlo") are generated from the simulated samples assuming an integrated luminosity of 36.1 pb^{-1}. The pseudo-data in each pseudo-experiment are composed of a Poisson-distributed number of events about the expected yields for each of the $t\bar{t}$, single top, V+jets, and QCD datasets. Pseudo-data are generated from the simulated datasets by randomly sampling appropriate ANN discriminant templates. The appropriate templates for creating the $t\bar{t}$ and single top ANN discriminants for each pseudo-experiment are the templates used in the final fit. Events in the orthogonal selection regions (RelIso>0.1 and two jet events) may not perfectly model QCD and V+jets events in the nominal region. To evaluate this effect, we must use simulated data. Therefore, pseudo-data are drawn from simulated V+jets and QCD events in the nominal region and the templates used to fit the pseudo-data are simulations of events in the orthogonal selection regions. For the purpose of having representative simulations in both the nominal and orthogonal selection regions, the data corrections applied to the $t\bar{t}$ and single top simulations are applied to the QCD and V+jets simulations. The V+jets $k_b = 2$ correction is also applied.

Because of the small expected QCD yield as well as its large uncertainty, it is possible for the fitter to give a QCD yield near the constraint of 0 events. This can introduce instability in the fit when the minimization routine attempts to determine the shape of the likelihood near the constraint. The fitted QCD yields are affected directly because of the constraint. The fitted V+jets yields are also affected because

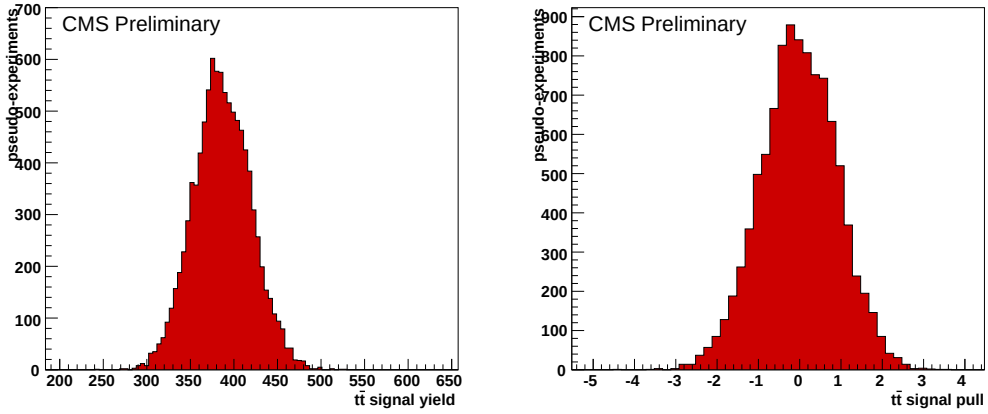
of the similarity of the QCD and V+jets templates, leading to an anti-correlation between fitted QCD and V+jets event yields. The $t\bar{t}$ and single top fitted event yields can also be slightly affected by this instability, since the QCD and V+jets templates, while similar, are not identical. To avoid the unstable region, pseudo-experiments with exceptionally low QCD yields are simply discarded. This was not an issue in the data, where the fitted number of QCD events was much larger than the 10 event cutoff that was used.

7.3 Fit statistical uncertainty

The pseudo-experiments indicate the presence of an intrinsic -3.1% bias in the measurement of the $t\bar{t}$ yield, where an average of 375 $t\bar{t}$ events are measured and 387 are expected. The bias is introduced by the use of data to form the V+jets and QCD templates. Differences between the input variables in the nominal and orthogonal selection regions, shown in Figures 6.9 and 6.11, cause the bias. The fit is subsequently corrected for the intrinsic bias.

The corrected $t\bar{t}$ yield and pull distribution from the pseudo-experiments is shown in Figure 7.1. Pull is defined as the difference between the corrected value from the fit in each pseudo-experiment and the expected number of events (387), divided by the corrected uncertainty of each fit, where RooFit calculates the fit uncertainty. A Gaussian fit to the corrected yield gives a mean of 386 and width 33 while a Gaussian fit to the corrected pull gives a mean of -0.02 and width 0.92. Because the pull width is less than 1, RooFit calculates a slightly larger fit uncertainty

than the average uncertainty over 10,000 pseudo-experiments. Since the uncertainty calculated by RooFit is quoted as the statistical uncertainty of the fit to the data, the quoted statistical uncertainty is conservative.



(a) (b)

Figure 7.1: Corrected $t\bar{t}$ event yields and pull distributions from 10,000 pseudo-experiments for an integrated luminosity of 36.1 pb^{-1}. The expected number of $t\bar{t}$ events is 387.

7.4 Systematic uncertainties

Systematic scenarios simulate the effect of changing some underlying assumption. A systematic scenario might change the efficiency of event selection or the shape of the ANN discriminant, or both. For example, we assume a value for the efficiency of b tagging jets, but if the efficiency in data is actually lower or higher, the ANN discriminant shape from data will be different than from simulation, as the shape is dependent on the fraction of events with b tagged jets. A system-

atic scenario increases or decreases the value of an underlying assumption by its uncertainty in order to determine if this systematically increases or decreases the measured $t\bar{t}$ cross section. All systematic uncertainties are evaluated by throwing 10,000 pseudo-experiments from datasets changed under various systematic scenarios and measuring the average change in the calculated $t\bar{t}$ cross section.

The fit templates for $t\bar{t}$ and single top are data-corrected simulations and do not change in systematic scenarios. Since the fit templates used for V+jets and QCD events are taken from data, the V+jets and QCD fit templates, as well as the pseudo-data, are varied under each systematic. Some systematics are from uncertainties in the simulation and the QCD and V+jets templates are taken from data. However, our measurement of intrinsic bias assumes the differences between the nominal and orthogonal events are modeled correctly by simulation. Therefore, the systematic pseudo-experiments must include changes to the simulations of V+jets and QCD events in both the nominal and orthogonal selection regions. This treats sources of additional bias that may only be present in systematic scenarios.

The dominant systematic uncertainties in the $t\bar{t}$ cross section measurement are from uncertainties in the jet energy scale (JES) and in the efficiency of the b tagging algorithm. Additional uncertainties from jet energy resolution, the heavy flavor content of V+jets events, the modeling of parton distribution functions within the simulation, variations of a cut on MadGraph to Pythia matrix element to parton shower matching, Q^2 simulation cut, and initial and final state radiation are also evaluated. The systematic uncertainties are summarized in Table 7.1 and described in the remainder of this chapter. Systematic uncertainties are measured with respect

to the expected yield, then scaled by the measured cross section.

Source	Uncertainty (%)
Jet energy scale	+9.7/-5.1
Jet energy resolution	±3.3
b tag efficiency	+16.1/-14.7
$V + b$ k factor	+5.2/-5.6
$V + c$ k factor	+4.4/-1.8
ISR/FSR	±5.0
Q^2	+6.8/-3.5
ME to PS matching	+6.0/-3.0
PDF	+0.6/-1.8
Combined	+22.8/-18.4

Table 7.1: List of systematic uncertainties.

7.4.1 Jet energy scale

The effect of consistently underestimated (overestimated) jet energy on the measured $t\bar{t}$ cross section is determined by shifting the energy of all jets in the simulated samples downward (upward) by the jet energy scale uncertainty prior to event selection. The change in jet energies affects both the ANN discriminant and the number of selected events, since the number of jets with $p_{\mathrm{T}} > 30$ GeV changes if jet energies are systematically shifted.

The estimate for the jet energy scale uncertainty of particle flow jets is measured to be less than 2.5%, with exact values parameterized as a function of jet p_{T} and η [66]. An additional flat uncertainty of 1.5% is added in quadrature with the jet dependent uncertainty to account for differences in software releases and calibration changes. A pileup uncertainty of (1.37 GeV)/p_{T} is also added in quadrature, where pileup uncertainty refers to the possibility that a reconstructed event may contain the products of multiple collisions. Finally, a b jet uncertainty of 2% for jets with $50 < p_{\mathrm{T}} < 200$ GeV and $|\eta| < 2.0$ or 3% for all other jets is also added in quadrature. The additional b jet uncertainty is applied to all jets to be conservative.

Pseudo-data are generated from the JES-shifted simulated events passing the nominal selection cuts. Since the $t\bar{t}$ and single top templates are data-corrected simulations, these fit templates are not changed in the pseudo-experiments. However, the QCD and V+jets templates are derived directly from data in the final fit. Therefore, the QCD and V+jets templates used in the JES pseudo-experiments are generated from the JES-shifted events that pass the orthogonal selection. The application of the JES uncertainty is found to alter the expected $t\bar{t}$ yields by +9.7% and -5.1%.

7.4.2 b tagging

The uncertainty due to systematic under- or over-estimates of the b tagging efficiency is estimated in a similar manner to the jet energy scale uncertainties. To estimate the effect from tag efficiency uncertainty, separate samples are generated

with the tag efficiency altered by $\pm 1\sigma$ from its expected value. In the case of c jets, where SF_b is used to correct the tag rate, the uncertainty from SF_b is doubled. For the track counting high efficiency medium operating point, the uncertainty of SF_l is calculated as a function of jet p_{T} and η and takes values ± 0.13 to ± 0.28. The uncertainty of SF_b is estimated as 15% of SF_b [18].

Just as with the jet energy scale uncertainties, appropriate templates are fit to pseudo-data to estimate the effect of this uncertainty on fitted $t\bar{t}$ event yield. Uncertainties on SF_l and SF_b are varied independently since their uncertainties are uncorrelated, then added in quadrature. Additionally, since the scale factors are applied to efficiencies of tagging b, c, uds, and g jets calculated from simulations, the statistical uncertainty of the efficiencies is varied coherently with their corresponding scale factor. The statistical uncertainty is considerably smaller ($< 5\%$) than uncertainties on SF_l and SF_b. A $\pm 1\sigma$ change in b tagging efficiency results in a change in fitted $t\bar{t}$ yield of +16.1% and -14.7%.

7.4.3 V+heavy flavor

Since the fit template for V+jets is taken from data, the heavy flavor content in V+jets is taken directly from data. However, because events in the orthogonal selection regions do not perfectly model events in the nominal selection region, the $t\bar{t}$ cross section measurement has an intrinsic bias of -3.1%. The intrinsic bias is measured using pseudo-experiments generated from simulated datasets, where a k factor of 2 is applied to simulated V+jets events containing a b jet. The uncertainty

on k_b is quite large, ± 1. Additionally, while a k factor of 1 is applied to V+jets events containing a c jet, it also has a large uncertainty of +1 and -0.5.

As is done for all sources of systematic uncertainty, simulated events in the orthogonal region and nominal regions are simultaneously varied under the systematic scenario while the fit templates for $t\bar{t}$ and single top, derived from simulations, are kept fixed. The average fitted $t\bar{t}$ yield changes by +5.2% and -5.6% due to uncertainty from k_b and by +4.4% and -1.8% due to uncertainty from k_c. These are additional systematic uncertainties due to uncertainty in the measurement of the intrinsic bias.

7.4.4 Jet energy resolution

Simulations were corrected to increase the jet energy resolution by 10%, which has an uncertainty of $\pm 10\%$. The upward and downward uncertainties of the jet energy resolution are evaluated by applying 20% and 0% smearings, respectively, and fitting the resulting distributions with the appropriate templates. The difference in $t\bar{t}$ selection efficiency between the three smeared samples is less than 0.02%. Shape differences in the ANN templates for the oversmeared and undersmeared samples produce a systematic uncertainty of $\pm 3.3\%$.

7.4.5 MadGraph and Pythia event generation settings

The simulation includes cuts on the squared momentum transfer, Q^2, and the quality of the MadGraph matrix element to Pythia parton shower match (ME to PS).

Both cuts are varied up and down in separate $t\bar{t}$ and V+jets samples. Additionally, the amount of initial and final state radiation (ISR/FSR) is varied up and down in the $t\bar{t}$ sample. In all cases, systematic uncertainties on the $t\bar{t}$ cross section due to these variations are calculated by throwing pseudo-data from the shifted samples and fitting with the appropriate templates. The systematic uncertainty from Q^2 is evaluated independently for the $t\bar{t}$ and V+jets samples, then added in quadrature, yielding an uncertainty on the $t\bar{t}$ cross section measurement of +6.8% and -3.5%. Similarly, the ME to PS matching cut is varied independently in $t\bar{t}$ and V+jets and added in quadrature, yielding an uncertainty of +6.0% and -3.0%. The systematic uncertainty introduced from varying the amount of ISR/FSR up and down in $t\bar{t}$ is $\pm$5.0%.

The simulations use the CTEQ66 parton distribution functions (PDF) [48], which is characterized by 22 parameters. In addition to this PDF, an additional set of 44 PDFs are also used, corresponding to positive and negative variations of each of the 22 parameters. The systematic uncertainty due to variations in these parameters is evaluated by weighting selected events by the varied PDFs, generating new samples from these weighted events, and then fitting the appropriate templates to pseudo-data generated from the reweighted events. The results of these pseudo-experiments show a negligible change in selection efficiency for the weighted samples, but a small change in ANN discriminant shape. Combining the uncertainties from each of the 22 different fit parameters results in a total systematic uncertainty of +0.6% and -1.8% due to variations in the PDF.

7.5 Chapter highlights

We measure $t\bar{t}$ cross section by fitting ANN discriminant templates for $t\bar{t}$, single top, V+jets, and QCD events to the ANN discriminant values observed in data. The fit is performed using a maximum likelihood method with Gaussian constraints on the single top and QCD event yields with 30% and 100% widths, respectively. Gaussian constraints are used because the fit templates for $t\bar{t}$ and single top events are similar to each other, as are the V+jets and QCD fit templates.

Pseudo-experiments are used to determine the performance of the fit. Pseudo-data are generated with a Poisson-varying number of events, with means given by the expected event yields. The pseudo-data are fit by simulated templates. This ensures the pseudo-experiments accurately represent any differences between the discriminant shapes in the nominal selection region and the orthogonal selection regions used to form the V+jets and QCD fit templates.

An intrinsic bias of -3.1% on the fitted $t\bar{t}$ cross section is measured, due to differences between the ANN inputs from events in the orthogonal and nominal regions. The pseudo-experiments also indicate the statistical uncertainty calculated by the fitter is conservative. The primary sources of systematic uncertainty are from the uncertainty of the b tag efficiency and the jet energy scale, with a combined systematic uncertainty of +22.8% and -18.4%. Changes to the intrinsic bias under sources of systematic uncertainty are treated by the pseudo-experiments, where systematically modified simulations are used to produce both the pseudo-data and the templates used for V+jets and QCD.

Chapter 8

Conclusion

8.1 The measured $t\bar{t}$ cross section

The neural net discriminant produced from the selected data events is shown in Figure 8.1. The result of the fit of the data-corrected $t\bar{t}$ and single top templates and data-driven V+jets and QCD templates to the data is also shown. The bias corrected result of the fit indicates a signal of $369 \pm 36(stat.)$ events; the remaining 902 events passing event selection are interpreted as background. This event yield corresponds to a cross section of $151 \pm 15(stat.)$ pb.

Table 8.1 shows the correlation table for the fit, where strong correlations between the QCD and V+jets fitted yields and between the $t\bar{t}$ and single top fitted yields are expected. These correlations are handled using Gaussian constraints on the amount of QCD and single top, as discussed in Section 7.1. The statistical uncertainty of the fit is 9.7% for the number of $t\bar{t}$ events, 6.0% for V+jets, 99.4% for QCD, and 29.8% for single top. As expected, the fit uncertainty of QCD and single top event yields is highly correlated with the Gaussian constraint width in the likelihood.

Figure 8.2 compares the data-corrected simulated distributions to the data distribution of the three variables used as inputs to the neural network. The event yields and statistical uncertainty of the simulated events are taken from the fit.

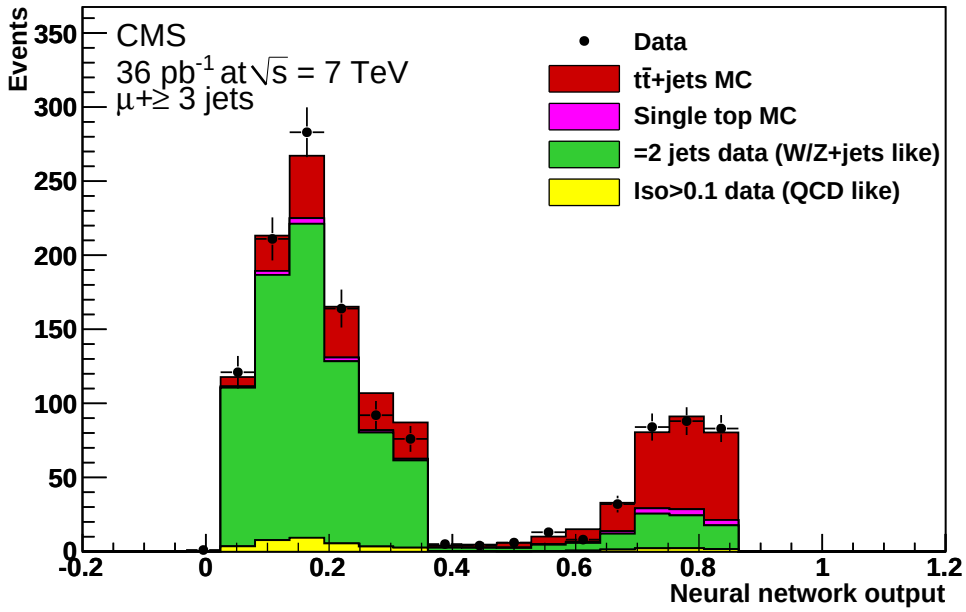


Figure 8.1: Results of template fit to the discriminant calculated from 36.1 pb^{-1} of data.

	QCD	single top	$t\bar{t}$	V+jets
QCD	1	-0.005	-0.257	-0.622
single top	-	1	-0.206	-0.016
$t\bar{t}$	-	-	1	-0.227
V+jets	-	-	-	1

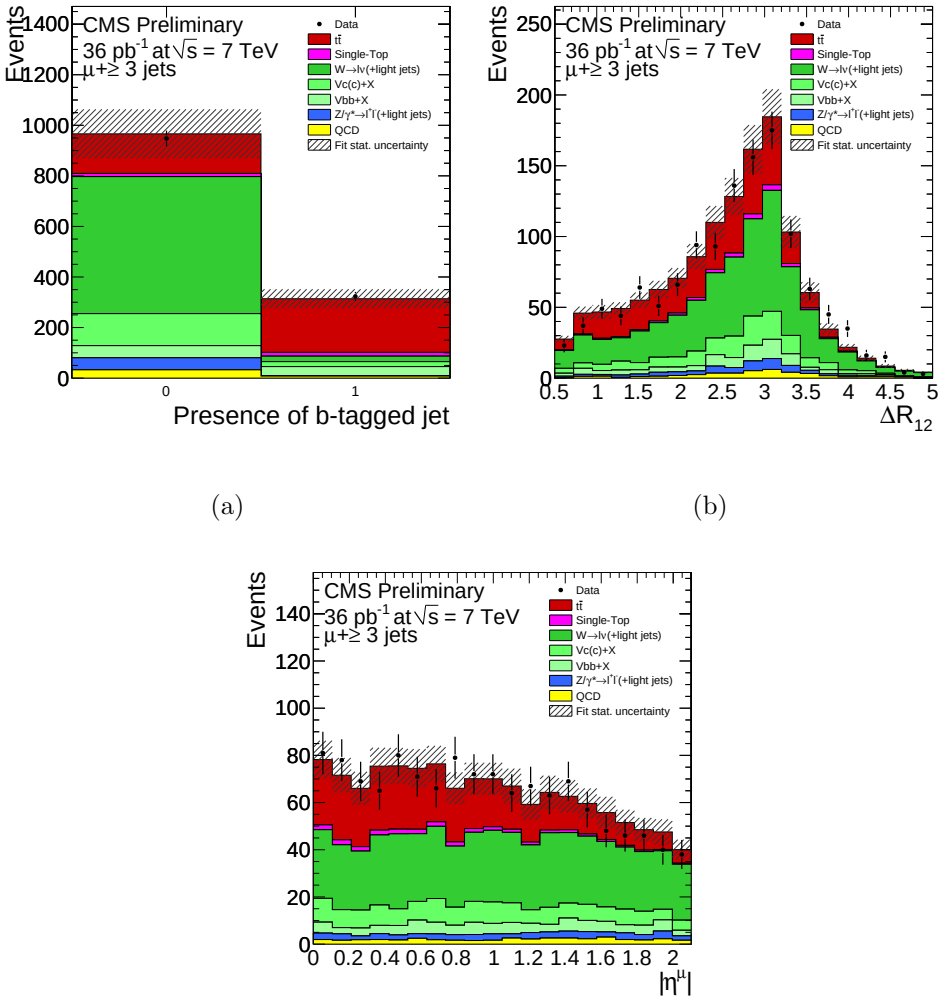
Table 8.1: Fit correlation table (symmetric).

By using data-corrected simulations, the input variable shapes can be compared between simulations of V+jets with different jet flavors, e.g., W+light, Z/γ^*+light, $V + c/cc$, and $V + bb$.

The combined systematic uncertainty is +35 pb and -28 pb. An additional 4% uncertainty due to the uncertainty in the luminosity measurement yields a cross section measurement of $151 \pm 15(stat.)^{+35}_{-28}(syst.) \pm 6(lumi.)$ pb, or 151^{+39}_{-32} pb. This result is consistent within one standard deviation with the theoretical NLO $t\bar{t}$ cross section.

8.2 Outlook

The $t\bar{t}$ cross section has also been measured by CMS in the dilepton [27] and electron+jets [22, 28] channels. These measurements were combined with the muon+jets measurement with the smallest uncertainty [22]. As of July 2011, the CMS measured $t\bar{t}$ cross section is 158 ± 19 pb, which has the smallest total un-



(a) (b)

(c)

Figure 8.2: Inputs to the neural network for simulations and data, where event yields and statistical uncertainties are taken from the fit to the data: (a) the b tag boolean, (b) ΔR_{12}, and (c) $|\eta^{\mu}|$.

certainty of all $t\bar{t}$ cross section measurements in pp collisions at a center of mass energy of 7 TeV. The cross section can also be measured in the hadronic channel; this measurement is currently being made by CMS. Finally, while $t\bar{t}$ can decay via tau+jets or di-tau, tau leptons are massive enough that they decay to jets or lighter leptons. These decays are not explicitly excluded by the current event selections used by the dilepton, lepton+jets, or hadronic analyses in CMS. The expected low statistical gain to be had by specifically searching for top to tau decays, combined with the difficulty of identifying taus due to their immediate decay, makes explicit tau channel measurements of the $t\bar{t}$ cross section a low priority at this time.

This analysis could be extended to include additional input variables to the ANN. Inputs can be chosen that would reduce both statistical and systematic uncertainties of the result. For example, jet energy is an excellent discriminator between $t\bar{t}$ and V+jets events because the jets in the lepton+jets $t\bar{t}$ decay channel can come directly from the top or W decays, while the jets in selected V+jets events come primarily from soft initial or final state radiation. The sum of the transverse jet energy of all jets with $p_T > 30$ GeV is shown in Figure 8.3. Using jet energy variables as inputs reduces statistical uncertainty because the ANN is better able to discriminate between signal and background, which reduces the uncertainty of the fit for number of signal events. While using jet energy variables as input will increase sensitivity of the analysis to the jet energy scale systematic uncertainty, it will reduce the sensitivity to the uncertainty of the b tag efficiencies, the largest systematic uncertainty in this analysis. Because the ANN shape would be less dependent on the b tag boolean input, changes to the discriminant shape due to changes in b tag

rates would be smaller.

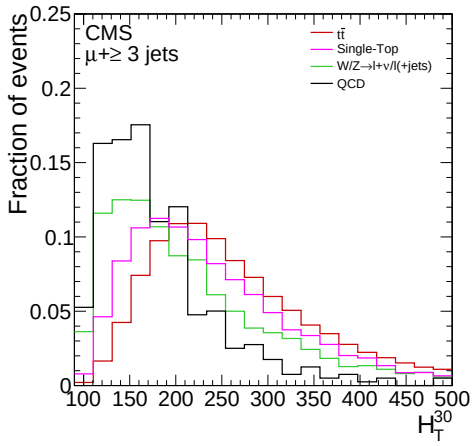


Figure 8.3: The sum of the transverse energy of all jets with $p_T > 30$ GeV (H_T^{30}) for various simulated interactions. All distributions are normalized to unity in order to show shape comparisons.

Jet energy variables were not used as inputs to the ANN in this analysis because jet energy distributions in two jet events do not match the jet energy distribution for V+jets events with three or more jets. By using an orthogonal selection region to model V+jets events, we were able to take heavy flavor content directly from data, but this restricts the use of variables related to jet energy. If a jet energy variable was used as an input, the ANN discriminator for two jet events would no longer accurately represent V+jets events in the nominal selection region. This was true of many considered and discarded variables, where the shape of the variables from two jet events did not match the shape from V+jets events with three or more jets. Therefore, in an attempt to use additional input variables, three strategies were considered and either discarded or postponed.

In the first strategy, an alternative orthogonal event selection was used to model V+jets events passing nominal selection. Di-muon events with a di-muon mass near 91 GeV, the mass of the Z boson, are produced almost entirely from Z boson decays; contamination from $t\bar{t}$, single top, and QCD processes is very small. Simulations predict that 94% of events with three or more jets and di-muon mass between 76 and 106 GeV are produced by V+jets decays. With three or more jets, the jet energies and heavy flavor content are nearly identical between events passing the nominal (single muon+three or more jets) and orthogonal (two muons+three or more jets) event selection. Input variables related to jet energy could be safely used from di-muon+three jet events to model V+jets events in the signal region without introducing additional sources of bias. Additionally, since two jet events do not perfectly model the flavor content of three jet V+jets events, the intrinsic bias changed under systematic scenarios that modified the b tag boolean input. Since di-muon+three jet events provide a better model of the flavor content of three jet V+jets events, sources of systematic uncertainty due to changes of the intrinsic bias from heavy flavor would be significantly reduced. In 36.1 pb^{-1} of data, only 69 events are expected to have three jets and di-muon mass between 76 and 106 GeV. The statistical occupancy of events with two muons and three jets used to form the V+jets fit template would be far too small to create a smooth fit template. However, as CMS accumulates more data, using di-muon events to model V+jets events will be an excellent strategy to consider.

In the second strategy, the two jet event ANN inputs could be corrected to match the ANN inputs of V+jets events passing nominal event selection using infor-

mation from simulations. Binned scale factors could be measured from simulations, defined as the ratio of number of events in the input variable bin between two jet events and nominal V+jets events. The scale factors could then be applied to the input variables taken from two jet events in data, applying a correction derived from simulations to the templates taken from data. Calculation of the scale factors would be greatly complicated by the fact that two jet events are contaminated by $t\bar{t}$ – the scale factors used would have to dynamically change during the fitting procedure to match the fitted event yields. Additionally, since the scale factors would be measured from simulations, they would be subject to the same systematic uncertainties impacting the data-corrected $t\bar{t}$ and single top templates. For example, a change to the jet energy scale could significantly change the binned scale factors used to correct any jet energy input from two jet events to the shape expected from V+jets events with three or more jets. Due to the difficulty of implementing scale factors as a function of fit yields and because of the potentially large uncertainty of the scale factors from systematic scenarios, this strategy was not pursued.

Alternatively, data-corrected simulations could be used for V+jets. The primary motivation for using events from data to model V+jets events was due to the uncertainty of the relative amount of heavy flavored jets in V+jets events, k_b and k_c. An obvious solution is to create independent fit templates for different V+jets flavors in simulated events passing nominal selection. This was performed: one template was made from the combination of V+light jet and $V + c$ jet events, since the discriminant shapes for both event types was similar, while a second template was made from $V + b$ events. The number of events corresponding to each template was

allowed to float freely during the fit so that no dependence on the heavy flavor in V+jets events was assumed. The $V + b$ fit result was strongly correlated with the fits to V + light/c and $t\bar{t}$, yielding a large uncertainty. This large uncertainty on $V+b$ content led to instability in systematic scenarios and therefore large systematic uncertainties on the $t\bar{t}$ cross section. There are alternative strategies to measure k_b and k_c from data, such as using the tag rates in 1, 2, 3, and 4 or more jet events. However, like the simulated scale factors for correcting templates derived from data, the contamination from $t\bar{t}$ events would require k_b and k_c to be dynamically fitted variables, making the implementation difficult and the benefits uncertain.

More traditional approaches to reduce systematic uncertainty are already being pursued in CMS. The single largest source of uncertainty in this analysis is due to uncertainty of the measured b tag scale factors, especially SF_b. Efforts are already underway to reduce the uncertainty of the b tag scale factors and preliminary results have measured SF_b with uncertainty less than half of the current uncertainty ($\pm$15%). We varied the uncertainty of SF_b in our analysis and found a roughly one to one correspondence between the uncertainty of SF_b and the resulting systematic uncertainty on the measured $t\bar{t}$ cross section. Therefore, the large +16.1% and -14.7% uncertainty of the measured $t\bar{t}$ cross section due to the uncertainty of the b tag efficiency may be reduced by half in the near future. This would make the systematic uncertainty due to b tag efficiency comparable to other sources of systematic uncertainty and significantly improve the final measured $t\bar{t}$ cross section.

8.3 Chapter highlights

Several strategies to reduce the statistical and systematic uncertainties of this analysis are available. Using events with a di-muon mass near the Z boson mass to model the V+jets fit template is especially promising once more data is accrued by CMS. As of July 18, 2011, CMS has recorded 1.23 fb^{-1} of luminosity, yielding approximately a factor of 30 more events than in this analysis, which uses data collected in 2010. Additionally, the single largest source of systematic uncertainty, the uncertainty of SF_b, is likely to decrease by approximately a factor of two in the near future.

Our measurement of the $t\bar{t}$ cross section in the muon+jets channel, as well as measurements made by CMS in other channels [22, 26, 27, 28], is consistent with the cross section predicted by the Standard Model. No evidence of physics beyond the Standard Model is indicated, although this analysis does not specifically rule out any predictions by theories beyond the Standard Model.

Bibliography

[1] S. Agostinelli et al. G4–a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A*, 506(3):250 – 303, 2003.

[2] W. Allcock, J. Bresnahan, R. Kettimuthu, and M. Link. The Globus Striped GridFTP Framework and Server. In *Proceedings of Super Computing (SC)*, Nov 2005.

[3] J. Alwall et al. MadGraph/MadEvent v4: the new web generation. *Journal of High Energy Physics*, 2007(09):028, 2007.

[4] Apache Hadoop. `http://hadoop.apache.org`.

[5] J. Bardeen, L. N. Cooper, and J. R. Schrieffer. Theory of superconductivity. *Phys. Rev.*, 108(5):1175–1204, Dec 1957.

[6] T. Barrass et al. Software agents in data and workflow management. In *Proceedings of Computing in High Energy Physics (CHEP)*, 2004.

[7] T. Barrass, J. Jernandez, and L. Tuura. Techniques for high-throughput, reliable transfer systems: break-down of PhEDEx design. In *Proceedings of Computing in High Energy and Nuclear Physics (CHEP)*, 2006.

[8] R. Bellan. Muon Reconstruction with the CMS Tracking System. *Nuclear Physics B - Proceedings Supplements*, 177-178:253 – 254, 2008. Proceedings of the Hadron Collider Physics Symposium 2007.

[9] Berkeley Storage Manager (BeStMan). `https://sdm.lbl.gov/bestman`.

[10] R. Brun and F. Rademakers. ROOT - An Object Oriented Data Analysis Framework. *Nucl. Inst. & Meth. in Phys. Res. A*, 389:81–86, 1997. See also `http://root.cern.ch`.

[11] A. Ceccanti, V. Ciaschini, M. Dimou, G. Garzoglio, T. Levshina, S. Traylen, and V. Venturi. Voms/vomrs utilization patterns and convergence plan. *Journal of Physics: Conference Series*, 219(6):062006, 2010.

[12] CEMon. `http://grid.pd.infn.it/cemon`.

[13] CERN document server (CDS). `http://cdsweb.cern.ch`.

[14] CDF Collaboration. Observation of Top Quark Production in $p\bar{p}$ Collisions with the Collider Detector at Fermilab. *Phys. Rev. Lett.*, 74(14):2626–2631, Apr 1995.

[15] CMS Collaboration. CMS: The Computing Project Technical Design Report. *CERN/LHCC*, 2005-023, 2005.

[16] CMS Collaboration. CMS technical design report, volume I: Detector performance and software. *CERN/LHCC*, 2006-001, 2006.

[17] CMS Collaboration. CMS tracking performance results from early LHC operation. *The European Physical Journal C*, 70:1165–1192, 2010. 10.1140/epjc/s10052-010-1491-3.

[18] CMS Collaboration. Commissioning of b-jet identification with pp collisions at $\sqrt{s} = 7$ TeV. *CMS Physics Analysis Summary*, PAS BTV-10-001, 2010.

[19] CMS Collaboration. Electromagnetic calorimeter commissioning and first results with 7 TeV data. *CERN/LHCC*, CMS-PAS-EGM-10-002, 2010.

[20] CMS Collaboration. Jet Energy Resolution in CMS at $\sqrt{s} = 7$ TeV. *CERN/LHCC*, CMS-PAS-JME-10-014, 2010.

[21] CMS Collaboration. Measurement of CMS Luminosity. *CERN/LHCC*, CMS-PAS-EWK-10-004, 2010.

[22] CMS Collaboration. Measurement of the $t\bar{t}$ Pair Production Cross Section at $\sqrt{s} = 7$ TeV using b-quark Jet Identification Techniques in Lepton+Jet Events. *CERN/LHCC*, CMS-PAS-TOP-10-003, 2010. `https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsTOP`.

[23] CMS Collaboration. Performance of muon identification in pp collisions at $\sqrt{s}$= 7 TeV. *CERN/LHCC*, CMS-PAS-MUO-10-002, 2010.

[24] CMS Collaboration. Selection of Top-Like Events in the Dilepton and Lepton-plus-Jets Channels in Early 7 TeV Data. *CERN/LHCC*, CMS-PAS-TOP-10-004, 2010.

[25] CMS Collaboration. Selection of Top-Like Events in the Dilepton and Lepton-plus-Jets Channels in Early 7 TeV Data. *CMS Physics Analysis Summary*, PAS TOP-10-004, 2010.

[26] CMS Collaboration. Combination of top pair production cross sections in pp collisions at 7 TeV and comparisons with theory. *CERN/LHCC*, CMS-PAS-TOP-11-001, 2011.

[27] CMS Collaboration. First measurement of the cross section for top-quark pair production in proton-proton collisions at $\sqrt{s} = 7$ tev. *Physics Letters B*, 695(5):424 – 443, 2011.

[28] CMS Collaboration. Measurement of the top-quark pair-production cross section in the lepton+jets channel without the use of b-tagging. *Accepted for publication by European Physical Journal C*, CMS-PAS-TOP-10-002, 2011. `https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsTOP`.

[29] DØ 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, March 1995.

[30] DØ Collaboration. Determination of the width of the top quark. *DØ Note*, 6034-CONF, 2010.

[31] Constructing a cavern for a giant. `http://cms.web.cern.ch/cms/Detector/CivilEngineering/Construction.html`.

[32] D. D. Corkill. Collaborating Software: Blackboard and Multi-Agent Systems and the Future. In *Proceedings of the International Lisp Conference*, 2003.

[33] M. Ernst, P. Fuhrmann, M. Gasthuber, T. Mkrtchyan, and C. Waldman. dCache, a distributed storage data caching system. In *Proceedings of Computing in High Energy and Nuclear Physics (CHEP)*, Sep 2001.

[34] D. Evans et al. Large Scale Job Management and Experience in Recent Data Challenges within the LHC CMS experiment. In *XII International Workshop on Advanced Computing and Analysis Techniques in Physics Research (ACAT)*, November 2008.

[35] Fermilab visual media services. `http://www-visualmedia.fnal.gov`.

[36] I. Foster and C. Kesselman. Globus: A Metacomputing Infrastructure Toolkit. *Intl J. Supercomputer Applications*, 11(2), 1997.

[37] gLite. `http://glite.cern.ch`.

[38] P. Grangier, G. Roger, and A. Aspect. Experimental Evidence for a Photon Anticorrelation Effect on a Beam Splitter: A New Light on Single-Photon Interferences. *EPL (Europhysics Letters)*, 1(4):173, 1986.

[39] Grid Information System. `https://svnweb.cern.ch/trac/gridinfo`.

[40] HEP-SPEC06 Benchmark. `http://hepix.caspur.it/benchmarks`.

[41] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, and H. Voss. TMVA: Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007.

[42] R. Housley, W. Ford, W. Polk, and D. Solo. Internet x.509 public key infrastructure certificate and CRL profile, 1999.

[43] iCMS. `http://cms.cern.ch/iCMS`.

[44] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45, 1960.

[45] E. Karavakis et al. CMS Dashboard Task Monitoring: A user-centric monitoring view. *Journal of Physics: Conference Series*, 219(7), 2010.

[46] M. Litzkow, M. Livny, and M. Mutka. Condor - a hunter of idle workstations. In *Proceedings of the 8th International Conference of Distributed Computing Systems*, June 1988.

[47] MyOSG. `http://myosg.grid.iu.edu`.

[48] P. M. Nadolsky et al. Implications of CTEQ global analysis for collider observables. *Phys. Rev.*, D78:013004, 2008.

[49] P. M. Papadopoulos, M. J. Katz, and G. Bruno. NPACI Rocks: Tools and Techniques for Easily Deploying Manageable Linux Clusters. In *"Proceedings of IEEE International Conference on Cluster Computing*, October 2001.

[50] Particle Data Group. The Review of Particle Physics. *J. Phys. G*, 37, 2010.

[51] D. A. Patterson, G. Gibson, and R. H. Katz. A case for redundant arrays of inexpensive disks (raid). In *Proceedings of the 1988 ACM SIGMOD international conference on Management of data*, SIGMOD '88, pages 109–116, New York, NY, USA, 1988. ACM.

[52] A. Delgado Peris et al. Data Location, Transfer and Bookkeeping in CMS. *Nuclear Physics B - Proceedings Supplements*, 177-178:279 – 280, 2008. Proceedings of the Hadron Collider Physics Symposium 2007, Proceedings of the Hadron Collider Physics Symposium 2007.

[53] R. Pordes et al. The open science grid. *Journal of Physics: Conference Series*, 78(1):012057, 2007.

[54] R. Quick, A. Gopu, and S. Hayashi. RSV: OSG Fabric Monitoring and Interoperation with WLCG Monitoring Systems. In *Computing in High Energy and Nuclear Physics (CHEP)*, 2009.

[55] Red Hat Enterprise Linux. `http://www.redhat.com/rhel`.

[56] A. Roy et al. Building and testing a production quality grid software distribution for the Open Science Grid. *Journal of Physics: Conference Series*, 180, 2009.

[57] C. Saout. A software framework for Multivariate Analysis Techniques for the CMS offline software. `https://twiki.cern.ch/twiki/pub/CMSPublic/SWGuideMVAFramework/saout_001_temp.pdf`, 2009.

[58] Scientific Linux. `http://www.scientificlinux.org`.

[59] H. Singh. A measurement of $t\bar{t}$ production cross section in $p\bar{p}$ collisions at $\sqrt{s} = 1.8$ TeV using neural networks. *PhD dissertation*, University of California Riverside, December 1999.

[60] T. Sjöstrand, S. Mrenna, and P. Skands. PYTHIA 6.4 physics and manual. *Journal of High Energy Physics*, 2006(05):026, 2006.

[61] G. Soyez. The SISCone and anti-kt jet algorithms. In *Proceedings of the International Workshop on Deep Inelastic Scattering and Related Subjects*, 2008.

[62] D. Spiga et al. The CMS Remote Analysis Builder (CRAB). In *Proceedings. 14th Int. Conf. on High Performance Computing*, volume 4873, 2007.

[63] SPEC CPU Subcommittee. SPEC CPU2006 Benchmark Descriptions. *Computer Architecture News*, 34(4), 2006.

[64] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*, pages 162–168. Elsevier, fourth edition, 2009.

[65] W. Verkerke and D. Kirkby. The RooFit toolkit for data modeling. *ArXiv e-prints*, physics/0306116, 2003.

[66] J. Weng. A Global Event Description using Particle Flow with the CMS Detector. *ArXiv e-prints*, 0810.3686, October 2008.

[67] WLCG Resource, Balance, and Usage. `http://gstat-wlcg.cern.ch/apps/pledges/resources`.

[68] Worldwide LHC Computing Grid. `http://lcg.web.cern.ch`.

[69] K. Yurkewicz. Protecting the LHC from itself. *Symmetry*, 04(10), 2007.