Top quark physics at the LHC with the CMS detector

Sergey Senkin

School of Physics

University of Bristol

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Doctor of Philosophy in the Faculty of Science

Autumn 2013

Contents

1. Theoretical Background

1.1 The Standard Model of Particle Physics

1.1.1 Electroweak theory

1.1.2 Electroweak symmetry breaking

1.2 Top Quark Physics within the Standard Model

1.2.1 Top quark production at the LHC

1.2.2 Top quark decay

1.2.3 Background processes

1.2.4 Top quark mass

1.2.5 Importance of top quark physics

1.3 Physics Beyond The Standard Model

1.3.1 Supersymmetry

1.3.2 Extra dimensions

2. The LHC and the CMS detector

2.1 The Large Hadron Collider

The LHC [1] is currently the largest and the most powerful particle accelerator ever built. It is installed in the 26.7 km tunnel that was originally constructed for the LEP accelerator in the 1980s. The tunnel lies at a depth of 45 m to 170 m underground between the Jura mountain and Lake Geneva, being the main part of the CERN accelerator complex.

The machine is designed to accelerate proton beams and provide collisions at a centre of mass energy of $\sqrt{s} = 14$ TeV. Unlike particle-antiparticle colliders, the LHC requires two rings with opposite magnetic dipole fields in order to maintain and collide two counter-rotating proton beams. Since the tunnel was originally designed for the electron-positron LEP, it has an internal diameter of 3.7 m which is not enough to install two separate independent rings. Therefore, a twin-bore magnet design was adopted [2], which resulted in substantial cost savings.

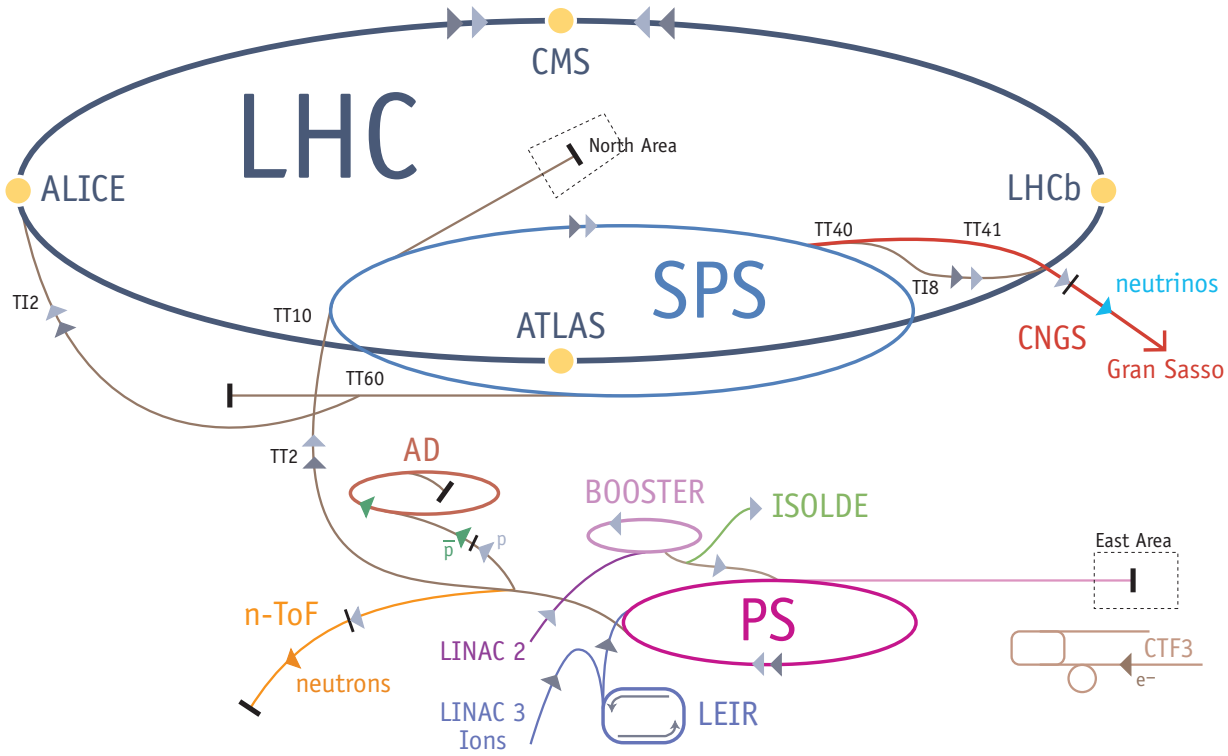


Figure 2.1: CERN accelerator complex.

A schematic view of the LHC accelerator chain is shown in Figure 2.1. Initially, the protons are obtained by stripping orbiting electrons from hydrogen atoms. Then they are injected into the linear accelerator LINAC2 to reach the energy of 50 MeV and enter the Proton Synchrotron Booster (PSB). The booster accelerates them to 1.4 GeV and passes the beam to the Proton Synchrotron (PS) where the energy rises to 25 GeV. In the next step, protons enter the Super Proton Synchrotron (SPS) where they are accelerated to 450 GeV. Finally, the beam is transferred to the LHC in both clockwise and anti-clockwise directions where it takes about 20 minutes to reach the design 7 TeV energy (per beam).

[add the number of magnets, total energy stored, number of bunches, bunch spacing etc.?]

The LHC has four interaction points, providing collisions to four major experiments. Two of them, CMS and ATLAS, are multi-purpose high-luminosity experiments with a peak luminosity of $L = 10^{34}$ cm^{-2}s^{-1}. The other two experiments operate at low luminosities and have more specific physics goals: LHCb studies b-meson decays, and Alice is a dedicated heavy ion experiment.

The instantaneous luminosity of a collider can be calculated as

$$L = \frac{n_1 n_2 f}{4\pi\sigma_x\sigma_y}, \tag{2.1}$$

where n_1 and n_2 are the numbers of particles in each of the colliding bunches, f is the revolution frequency, σ_x and σ_y are the horizontal and vertical beam sizes, assuming the two beams have the same size.

The number of events generated in the collisions per second is given by

$$N_{events} = L \times \sigma, \tag{2.2}$$

where σ is the cross section of the process under study.

[add the plots with cross sections and production rates?]

The LHC started operating on the 10th of September 2008, with the first beams fully circulating in both rings. However, only 9 days later a magnet quench occurred in two sectors of the tunnel, which was caused by an electrical fault due to a bad connection between two magnets. A consequent liquid helium explosion damaged a total of 53 superconducting magnets. Over a year was spent on repairs and tests, and the first collisions were recorded on the 23rd of November 2009 at a centre of mass energy of 0.9 TeV. The following few months showed the continuous ramp up of the beam energies up to 3.5 TeV per beam which was achieved on the 30rd of March 2010 when the LHC physics programme started.

Throughout the rest of 2010, the two general-purpose LHC experiments (CMS and ATLAS) recorded approximately 40 pb^{-1} of data, which resulted in the first measurements of various physics processes at the LHC. The following year became the main 7 TeV data-taking period, with about 5 fb^{-1} of data recorded by ATLAS and CMS. On the 5th of April 2012 the centre of mass energy was increased to 8 TeV, and July of 2012 marked the first major discovery of a new boson which was later shown to be consistent with the Standard Model Higgs boson, according to approximately 21.8 fb^{-1} of data recorded until early 2013. A long shut-down is planned for the following two years with various upgrades scheduled. The next physics run is expected in 2015 with the beam energy increased up to 6 or 7 TeV.

[add any upgrade details and distant future plans, like SLHC?]

2.2 The CMS Detector

The Compact Muon Solenoid [3] is a general-purpose detector designed to carry out precise measurements of the Standard Model and searches for physics beyond it. The primary design requirement was the ability to discover the nature of electroweak symmetry breaking, and the first observation of a Higgs boson was obtained in the Summer of 2012 [4].

The detector is installed at one of the LHC interaction points (Point 5) at about 100 m underground near the French village of Cessy, between the Jura mountains and Lake Geneva. The overall dimensions of the CMS detector are a length of 21.6 m, a diameter of 14.6 m and a total weight of 12 500 t.

The sectional view of CMS is shown in Figure 2.2. In the centre of the detector, tracking and calorimetry systems are surrounded by the superconducting solenoid. On the outermost part of it the magnetic flux is returned through the iron yoke in which the muon system is also integrated. All the sub-systems are discussed in the following sections in more detail.

The cylindrical shape of the CMS detector dictates using a cylindrical coordinate system, with the origin centred at the interaction point, the x-axis pointing towards the centre of the LHC ring, the y-axis pointing upwards and the z-axis pointing along the beamline in the anti-clockwise direction. The azimuthal angle ϕ is measured from the x-axis in the transverse $(x-y)$ plane and the polar angle θ is measured from the z-axis. The radial distance to the beamline is denoted by r. Pseudorapidity is defined as:

$$\eta = -\ln \tan \frac{\theta}{2}. \tag{2.3}$$

This implies that the particles moving in the transverse plane (perpendicular to the beamline) have a pseudorapidity of 0, whereas the beam direction has an infinite pseudorapidity.

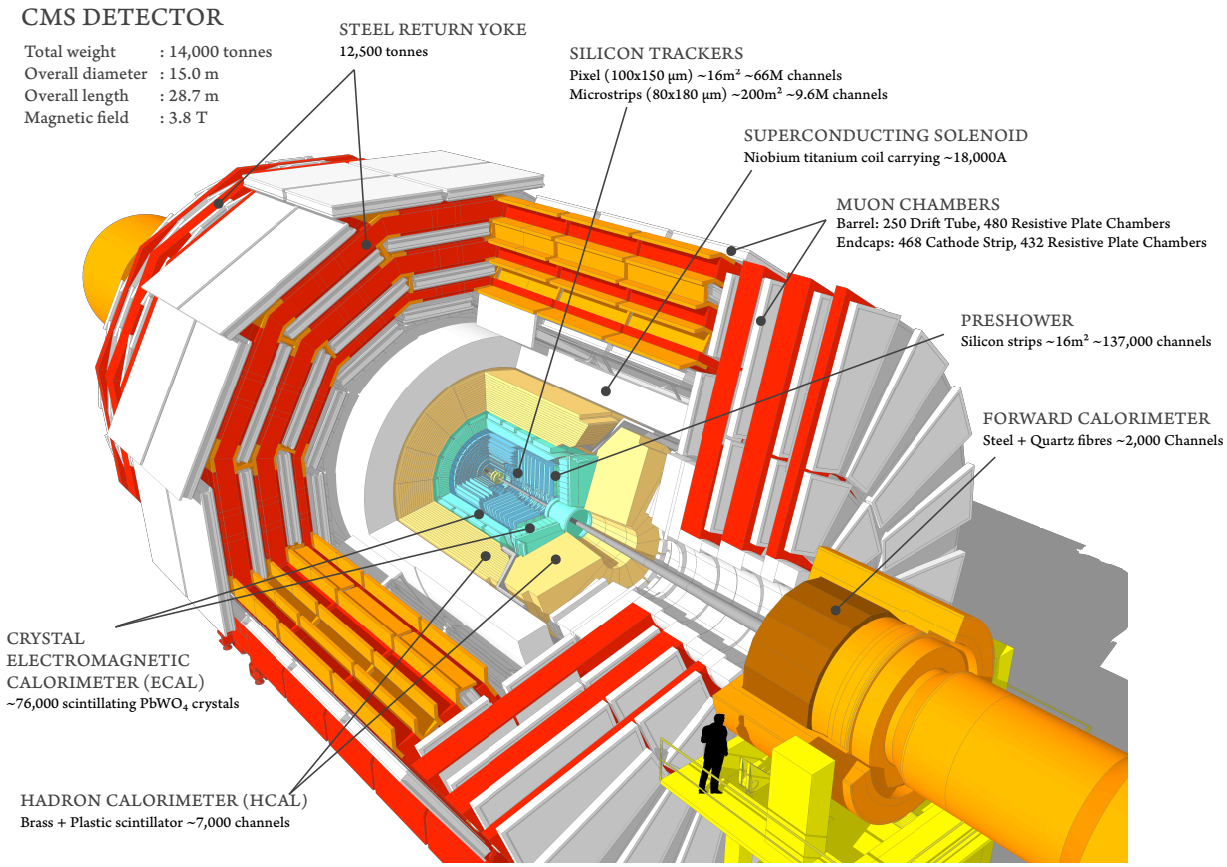


Figure 2.2: Sectional view of the CMS detector.

Considering the cylindrical shape of the detector, it has barrel and endcap regions, with the transition occurring at $\eta \sim 1.4$. The momentum and energy transverse to the beamline are denoted by p_T and E_T respectively; the imbalance of the energy measured in the transverse plane, called missing transverse energy, is denoted by $E_{\mathrm{T}}^{\mathrm{miss}}$.

2.2.1 Inner Tracking System

The tracking system lies in the heart of the CMS detector and is the closest to the interaction point where the particle flux has the highest value. This imposes demanding requirements on the configuration of the system. At design luminosity of $L = 10^{34}$ cm^{-2}s^{-1} with the bunch spacing of 25 ns, an average of 1000 particles from about 25 proton-proton interactions (pile-up vertices) is expected to traverse the tracker for each bunch crossing. However, up until the long shutdown a bunch spacing of 50 ns was used, which meant a higher number of protons in each bunch leading to approximately twice the number of pile-up vertices. Therefore, in order for the particle tracks to be identified reliably and separately for each bunch crossing, the tracker requires very fine granularity and fast response parameters. Another complication caused by the intense particle flux is the severe radiation damage, so the tracker has to be highly resilient in operating in the harsh environment for a reasonable lifetime.

To meet these requirements on granularity, response time and radiation resilience, the tracker design was chosen to be based on silicon detector technology. Although capable of meeting such conditions, this technology has a disadvantage of a high power density of on-detector electronics. This implies the necessity of an efficient cooling system. Moreover, a large amount of dense material interacting with the particles leads to higher multiple scattering, bremsstrahlung, photon conversions and nuclear interactions. Therefore, there are complications in the reconstruction of the tracks, meaning some loss of efficiency and precision. This will be discussed in detail later on in the object reconstruction section.

Figure 2.3 shows the overall layout of the tracking system. It consists of the inner pixel detector, located in the vicinity of the interaction point, and silicon strip tracker detectors: inner barrel and disks (TIB and TID), outer barrel (TOB) and endcaps (TEC). The geometrical acceptance of the tracker system goes up to $|\eta| < 2.5$. The outer radius of the CMS tracker reaches approximately 110 cm, and its total length is about 540 cm.

The pixel detector consists of three layers of pixel sensors at radii of 4.4 cm, 7.3 cm and 10.2 cm from the beamline in the barrel region. In addition there are two endcap disks on each side at $|z| = 34.5$ cm and 46.5 cm. The pixel size equals 100×150 µm^2 in $r\phi \times z$ coordinates. The pixel detector has 66 million pixels and the total area of about 1 m^2.

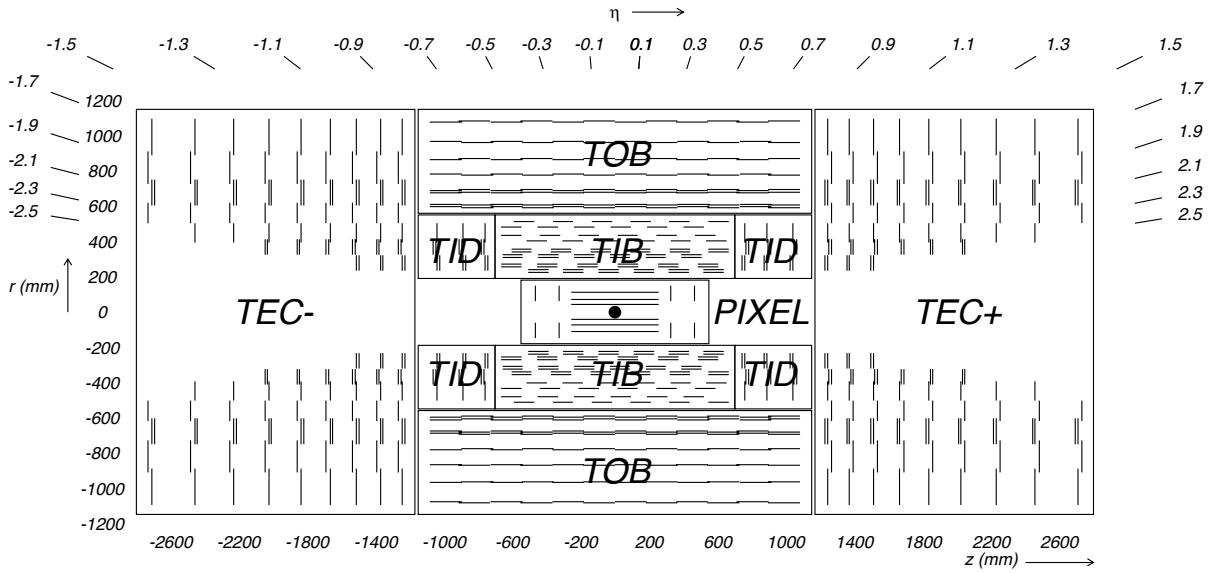


Figure 2.3: Cross-section of the CMS tracker system [3].

The silicon strip tracker consists of several layers of silicon microstrip detectors. It covers the region between 20 cm to 110 cm in radius and extends up to ±280 cm in the z direction. The Tracker Inner Barrel (TIB) is made out of 4 layers and the Tracker Outer Barrel (TOB) has 6 layers in it. The tracker endcaps (TEC) comprise 9 disks, and there are also the tracker inner disks (TID) that consist of 3 disks filling the gap between TIB and TEC as shown in Figure 2.3. There are 9.3 million silicon strips covering the area of about 200 m^2. The silicon sensors' thickness varies between 320 and 500 µm and the strip pitch varies from 80 µm in the TIB to 180 µm in TOB and TEC.

The silicon detectors of the tracker, the readout electronics and support structure form a considerable amount of material for the particles traversing from the interaction point. Figure 2.4 [3] shows the material budget of the CMS tracker in units of radiation lengths[1] (X_0). It grows from about 0.4 X_0 to 1.8 X_0 in the barrel region, and then decreases to about 1 X_0 in the endcaps. This causes a substantial conversion rate for photons and electrons in the tracker material; it also will be discussed in more detail in the electron reconstruction section.

[perhaps need to add the p_T resolution plots]

[1]A material's radiation length is the mean distance over which a high-energy electron loses all but $1/e$ of its energy by bremsstrahlung; this is equal to 7/9 of the mean free path for pair production by a high-energy photon.

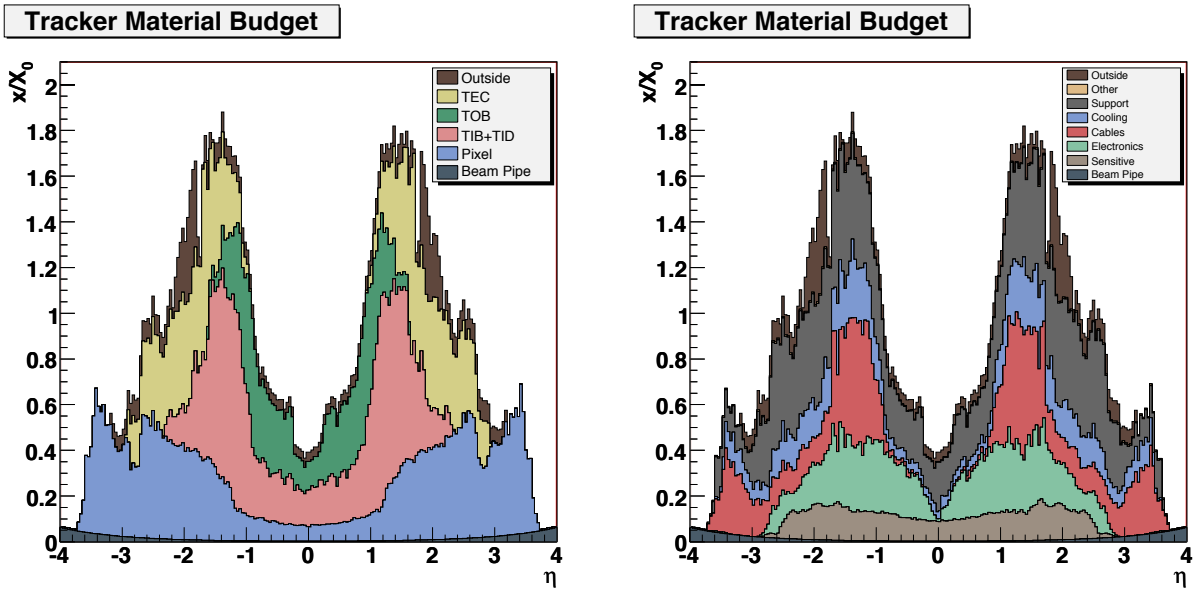


Figure 2.4: Material budget as a function of pseudorapidity η for the different sub-detectors of the tracker (left) and broken down into the functional contributions (right), in units of radiation length [3].

2.2.2 Electromagnetic Calorimeter

The next detector subsystem which is surrounding the tracker is the electromagnetic calorimeter, or ECAL. It is of a primary importance for the analyses described in this thesis, as it provides information for the electron and positron reconstruction. Combination of this information with that from the tracking system must ensure a precise measurement of electron position and momentum, and also sufficient background removal. It has to effectively distinguish the energy deposit shape of an electromagnetic particle from the one of a hadronic particle, which requires good segmentation and high resolution.

ECAL is a hermetic, high-granularity, high-resolution scintillating crystal calorimeter consisting of 61 200 lead tungstate ($PbWO_4$) crystals located in the central barrel region ($|\eta|< 1.479$), and 7324 crystals in each of the two endcaps ($1.479 <|\eta|< 3.0$). All crystals are followed by photodetectors reading and amplifying their scintillation: avalanche photodiodes (APD) are used in the barrel, and vacuum phototriodes (VPTs) are used in the endcaps. These different choices were caused by the configuration of the magnetic field and the expected level of radiation.

The layout of the ECAL sub-detector is shown in Figure 2.5. An additional preshower detector is used in the endcap region to lower the required detector depth. Its principal

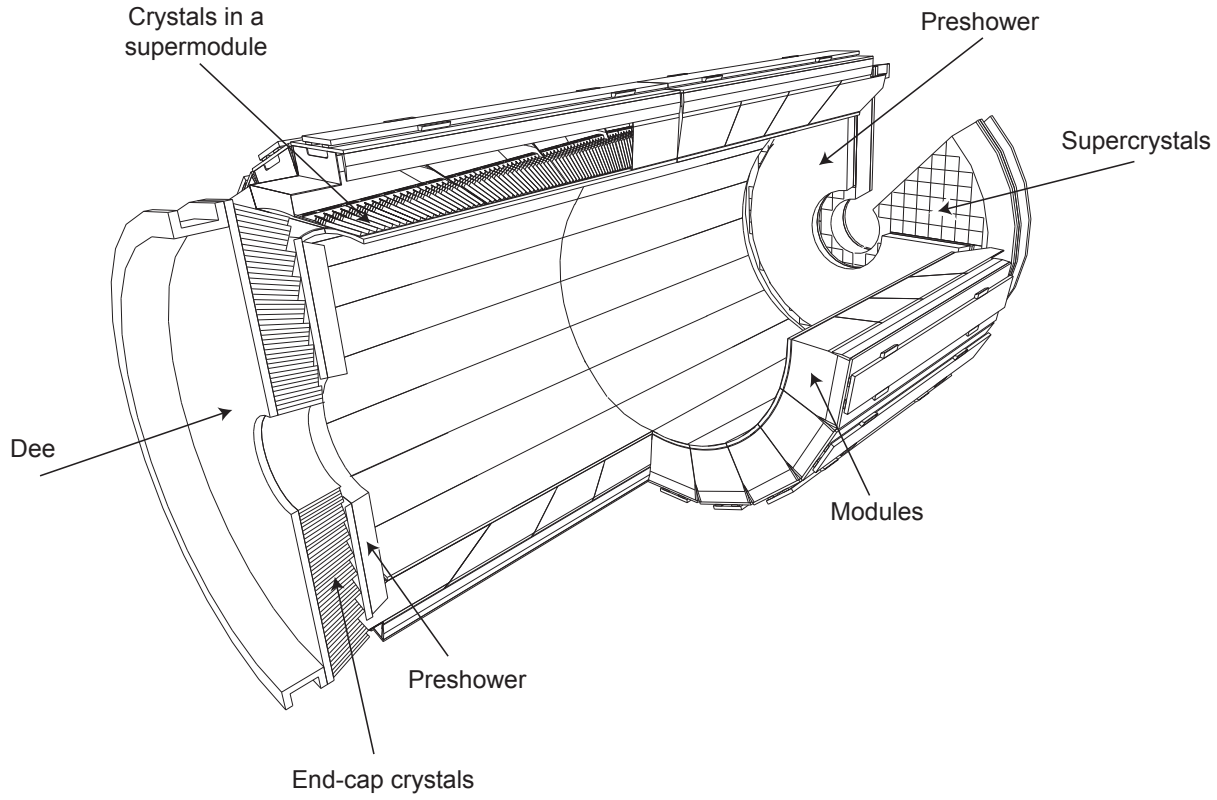


Figure 2.5: Layout of the CMS electromagnetic calorimeter [3].

aim is to identify neutral pions in the endcaps, but it also helps to distinguish neutral pions and electrons from minimum ionising particles and improves the position determination of electrons and photons with high granularity.

Table 2.1: ECAL crystal characteristics

	Barrel	Endcaps
number of crystals	61 200	14 648
crystal cross-section in (η, ϕ)	0.0174×0.0174	varies
crystal cross-section at the front	22×22 mm^2	28.62×28.62 mm^2
crystal cross-section at the rear	26×26 mm^2	30×30 mm^2
crystal length	230 mm ($25.8X_0$)	220 mm ($24.7X_0$)

The main geometrical characteristics of the ECAL crystals are shown in Table 2.1. The choice of lead tungstate was driven by the constraints of the CMS design. It is a very dense material (8.28 g/cm^3) with a short radiation length of $X_0 = 0.89$ cm, which allows the calorimeter to fit inside the compact magnet. Lead tungstate also has a small Molière radius[1] of 2.2 cm, which allows a calorimeter with fine granularity. Finally, the crystals emit 80 % of their scintillation light in just 25 ns, however the light yield is relatively low. At 18 °C, about 4.5 photoelectrons per MeV are collected. The dependence of the light yield on temperature requires a cooling system capable of keeping the crystal temperature stable within ±0.05 °C to preserve energy resolution [5].

The energy-dependent resolution of the calorimeter can be parameterised as follows [3]:

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{S}{\sqrt{E}}\right)^2 + \left(\frac{N}{E}\right)^2 + C^2. \tag{2.4}$$

where S is the stochastic term, N is the noise term, and C is the constant term. Figure 2.6 shows the energy resolution measured using incident electrons, during the beam tests in 2004.

2.2.3 Hadron Calorimeter

The hadron calorimeter (HCAL) is the next sub-detector located mostly inside the solenoid and completing the CMS calorimetry system. It is essential for the measurement of

[1]The Molière radius R_μ is a characteristic constant of a material giving the scale of the transverse dimension of the fully contained electromagnetic showers initiated by an incident high energy electron or photon. It is defined as the radius of a cylinder containing an average of 90 % of the shower's energy deposition.

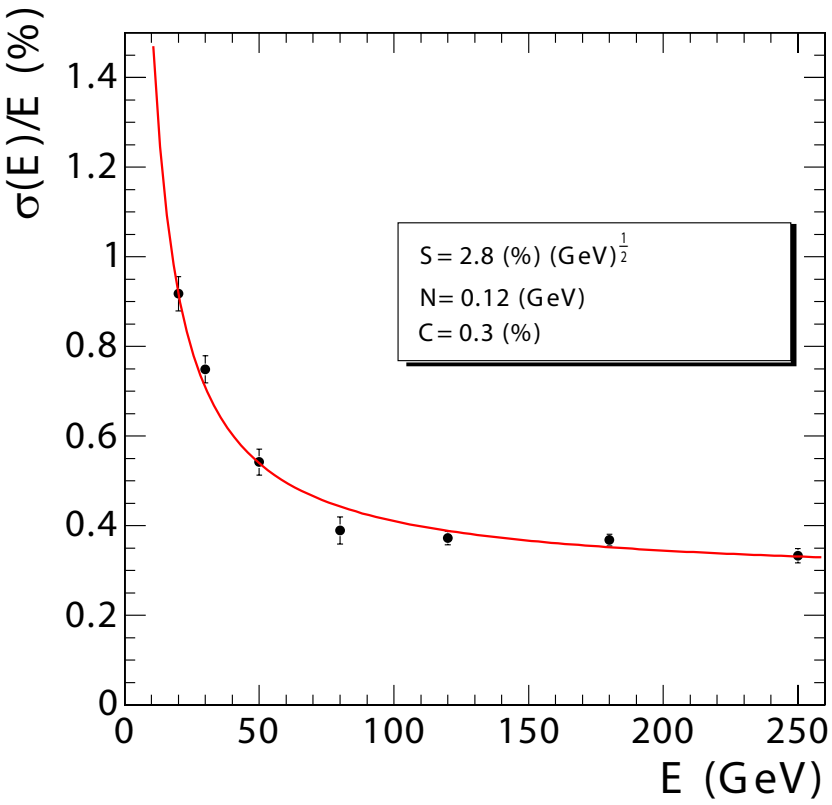


Figure 2.6: ECAL energy resolution derived from the test beam measurements as a function of deposited energy. The stochastic, noise, and constant contributions are shown [3].

hadron jets and missing transverse energy.

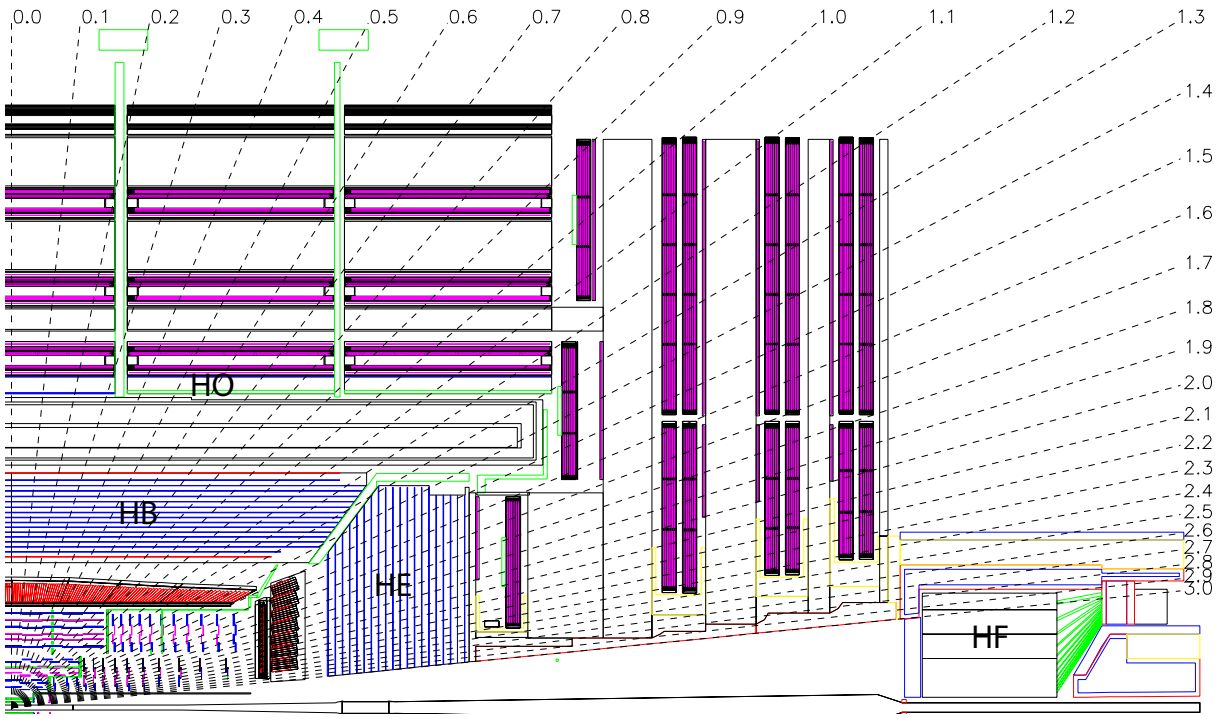


Figure 2.7: Longitudinal view of the CMS detector showing the locations of the hadron barrel (HB), endcap (HE), outer (HO) and forward (HF) calorimeters [3].

As shown in Figure 2.7, HCAL consists of four subsystems: the hadron barrel calorimeter (HB), the hadron endcap calorimeter (HE), the hadron outer calorimeter (HO) and the hadron forward calorimeter (HF). The barrel and endcap parts (HB, HE) cover the pseudorapidity range up to $|\eta|< 3.0$, and the forward part (HF) extends it to a total coverage of $|\eta|< 5.0$. HCAL surrounds ECAL from its outer limit of 1.77 m from the beamline, to the inner limit of the magnet coil at 2.95 m from the beamline. However, due to space limitations the barrel calorimeters do not contain complete hadronic showers, therefore an outer calorimeter (HO) was designed to measure the energy leakage. It is placed in the muon system just outside of the solenoid in the barrel region.

HCAL is a sampling calorimeter consisting of alternating layers of brass and stainless steel absorbers, and plastic scintillators as active elements. The choice of the absorber material was caused by its short hadronic interaction length and its property of being non-magnetic, which is crucial in the strong magnetic field of the CMS magnet. The scintillation light is guided by embedded wavelength-shifting (WLS) fibres. The light from the WLS is then

transmitted via a network of clear fibres, arranged in read-out towers, to hybrid photodiodes (HPDs) [3].

Both HB and HE scintillators have a granularity of $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$ for $|\eta| < 1.6$, and $\Delta\eta \times \Delta\phi = 0.17 \times 0.17$ for $|\eta| \geq 1.6$. The tower segmentation of the forward calorimeter (HF) varies from $\Delta\eta \times \Delta\phi = 0.175 \times 0.175$ at $|\eta| = 3.0$ to $\Delta\eta \times \Delta\phi = 0.3 \times 0.35$ at at $|\eta| = 5.0$. The HF is placed at about 11 m from the interaction point, and is essential to reconstruct very forward hadron jets. Together with HO, it provides the hermeticity of the calorimetry system, making it possible to measure the transverse missing energy to a reasonable precision.

[perhaps need to add the energy resolution plots]

2.2.4 Superconducting Magnet

The superconducting solenoid is a central feature of the CMS apparatus, essentially giving it its name. The magnet has a length of 12.5 m, diameter of 6.3 m and mass of 220 t. Although it was initially designed to sustain a uniform magnetic field of 4 T within the 5.9 m diameter free bore, operation at 3.8 T was chosen in order to increase the lifetime. The magnetic field is returned by a massive iron yoke. The main parameters of the CMS magnet are shown in Table 2.2.

Table 2.2: Parameters of the CMS superconducting solenoid [5] [6].

Field	3.8 T
Inner Bore	5.9 m
Length	12.5 m
Number of Turns	2168
Current	18 160 kA
Stored energy	2.3 GJ

The large bending power of the solenoid is required to bend the tracks of high energy charged particles to an extent where good momentum resolution is achieved. The design requirement for the strength of the magnetic field was the ability to unambiguously determine the sign of the electric charge for muons with a momentum of ≈ 1 TeV/c [5].

The solenoid coil is constructed from four layers of superconducting high-purity niobium-titanium cable co-extruded with pure aluminium, which acts as a thermal stabiliser. The cold mass is cooled down to 4.5 K by liquid helium. If a fast discharge happens (e.g. caused by a magnet quench), about 3 days are necessary to re-cool the coil.

2.2.5 Muon System

The last sub-detector placed on the outermost part of CMS is the muon system. Since the muons are the most penetrating particles detectable by CMS, they have the cleanest signature and play an important role in many physics analyses. Due to their ability to travel through the many layers of the calorimeters, muons are relatively easy to identify and separate from the background.

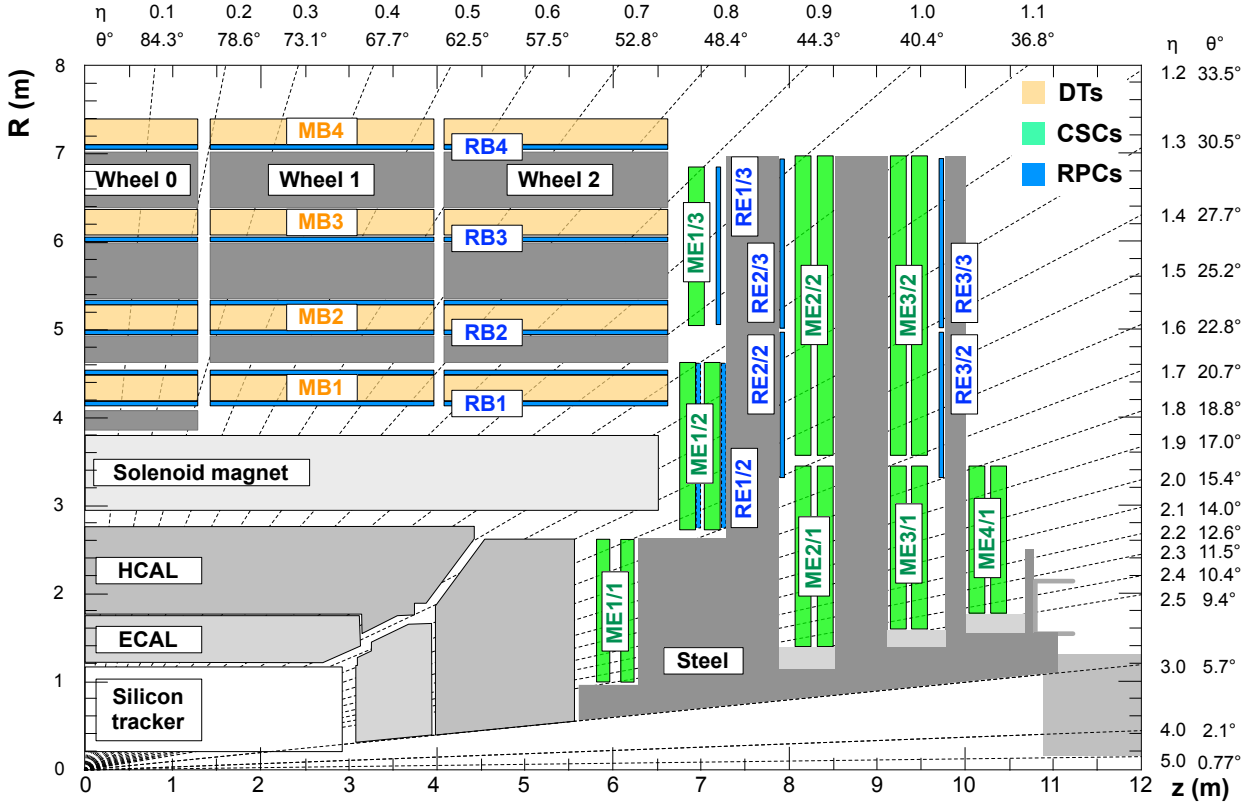


Figure 2.8: Layout of one quarter of the CMS muon system. Four drift tube (DT, in light orange) stations are labeled MB (muon barrel) and the cathode strip chambers (CSC, in green) are labeled ME (muon endcap). Resistive plate chambers (RPC, in blue) are in both the barrel and the endcaps of CMS, where they are labeled RB and RE, respectively.

The layout of the CMS muon system is shown in Figure 2.8. It consists of the drift tubes (DT), cathode strip chambers (CSC) and resistive plate chambers (RPC). The entire system surrounds the solenoid and covers the pseudorapidity region of $|\eta| < 2.4$.

The drift tubes are located in the barrel region ($|\eta| < 1.2$). Consisting of four stations, they form concentric cylinders around the beam line; there are 250 drift chambers with about 172 000 sensitive wires in total. When a muon passes through the volume, it knocks

electrons off the atoms of the gas, which then follow the electric field and reach the positively-charged wires, providing information on the muon's position. The chambers are filled with the gas mixture of 85 % Ar and 15 % CO_2, where the muon drift time does not exceed 380 ns. Although this value is bigger than the typical bunch crossing time (25 or 50 ns), it is sufficient because of the small muon rate in this region.

In the endcaps, the cathode strip chambers cover the pseudorapidity region of $0.9 < |\eta| < 2.4$. Each of 468 CSCs is a trapezoidal multi-wire proportional chamber consisting of 6 gas gaps with a plane of radial cathode strips and a plane of anode wires which are roughly perpendicular. A charged muon traversing each plane of a chamber causes gas ionisation and a subsequent electron avalanche which produces a charge on the anode wire and an image charge on the cathode strips. The gas used in CSCs is a mixture of Ar, CO_2 and CF_4.

The resistive plate chambers system is complementary to both DT and CSC systems, and is located in both barrel and endcap regions ($|\eta| < 2.1$). RPCs also operate in avalanche mode with a gas mixture of $C_2H_2F_4$, C_4H_{10} and SF_6, and due to an excellent time resolution of about 1 ns they provide fast information for triggering. The spacial resolution is, however, quite limited ($\approx$ 1 cm, compared to $\approx$ 100 µm for DTs and CSCs).

The muon momentum is measured in both the tracker and the muon system. As it can be seen on Figure 2.9, both sub-systems contribute to the momentum resolution at different p_T values. This happens due to the difference in the magnetic field and detector technology. For low-p_T muons, the best momentum resolution is obtained in the tracker, whereas in the high-p_T region the muon system provides a significant improvement. Therefore, by using information from both the silicon tracker and the muon chambers (i.e. reconstructing the "global muon"), the momentum resolution is improved in the whole p_T region up to a $\approx$ 1 TeV/c level.

2.2.6 Trigger and Data Acquisition

At design LHC luminosity of $L = 10^{34}$ cm^{-2}s^{-1}, approximately 25 collisions are expected to occur at each crossing of the proton bunches. The bunch spacing of 25 ns corresponds to a crossing rate of 40 MHz. Since every event produces $\sim$ 1 MB of raw data, it corresponds to a total data production of 40 TB s^{-1}. Attempting to store all of this data is clearly beyond the available technology. Moreover, only a fraction of events contain hard scattering processes that are of interest, therefore an effective trigger system had to be implemented.

The CMS trigger is a two-level system, consisting of two independent parts: the Level-1 (L1) trigger and the High-Level Trigger (HLT). The L1 trigger is a hardware system implemented in programmable electronics residing partly on detector, and partly in the underground control room located at approximately 90 m from the experimental cavern.

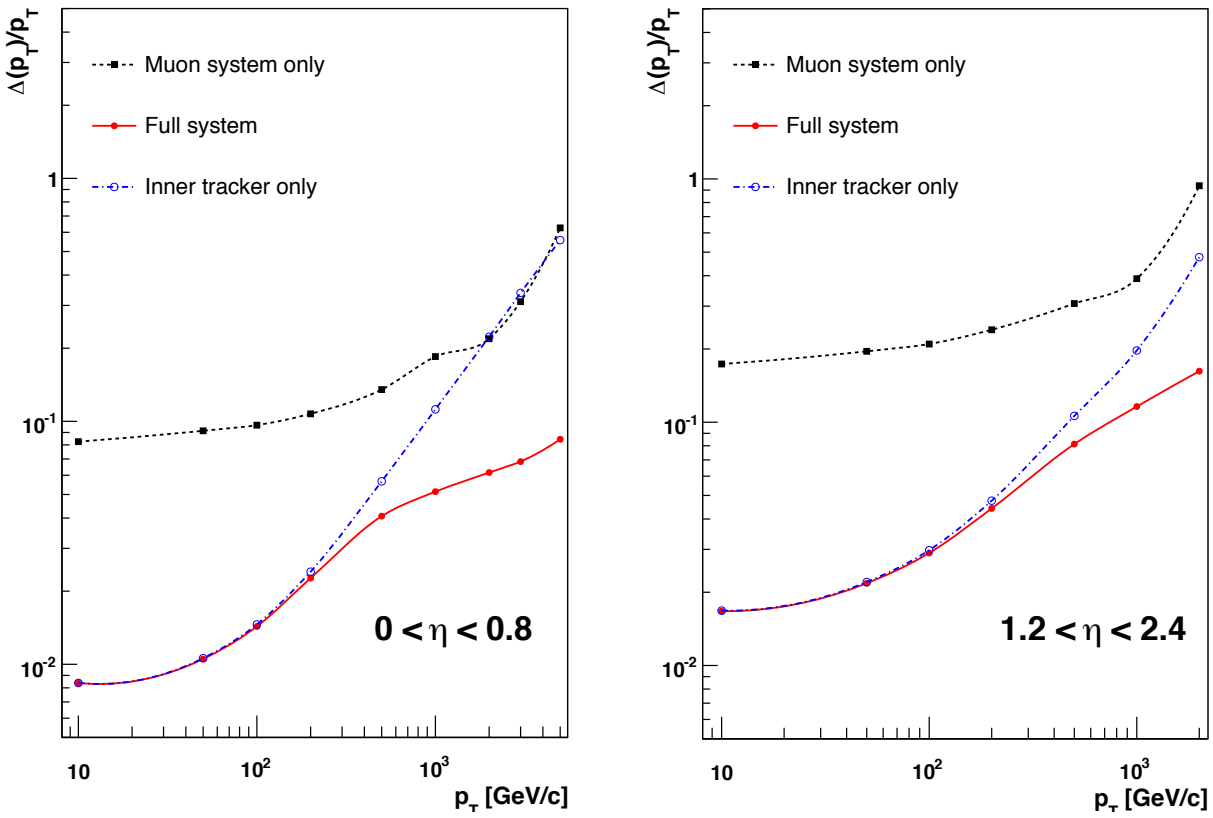


Figure 2.9: The muon transverse momentum resolution as a function of the transverse momentum (p_{T}) using the muon system only (black), the inner tracking only (blue), and both (red), in regions of $|\eta|< 0.8$ (left) and $1.2 <|\eta|< 2.4$ (right) [3].

The maximum latency between the collision and the L1 accept decision received by front-end electronics is 3.2 µs. During this amount of time, the complete event information is buffered in pipelined memories on the detector. The only information used for the L1 trigger decision is that from the muon system and the calorimetry. Since the reconstruction of tracks exceeds the time scale required for the L1 decision, the tracker information can't be used. The L1 trigger reduces the event rate from ~ 40 MHz to ~ 100 kHz, corresponding to a data flow of about 100 GB s^{-1}. These events are fed into the HLT system.

The High-Level Trigger is a software system implemented in a single CPU farm, sometimes referred to as the "Event Filter Farm". Having access to the full event information, customised algorithms of increasing complexity are used which results in a highly flexible trigger system. The event rate is reduced down to ~ 300 Hz, with the final data rate of approximately 300 MB s^{-1} being stored on a large disk cache at the experimental site (the Storage Manager) and later on transferred to CERN Tier 0 for further processing (see Section 2.3).

Since the start of the LHC running, the operating conditions have been changing drastically. During the start-up year of 2010, the instantaneous luminosity went up from about 10^{27} cm^{-2}s^{-1} to approximately 0.2×10^{32} cm^{-2}s^{-1}. In 2011 the luminosity ramped up to a factor of 20 above that of 2010, reaching approximately 4×10^{33} cm^{-2}s^{-1}. This required a lot of continuous effort to control the trigger rates at reasonable level, whilst also keeping its efficiency acceptable. In 2012 the luminosity was more stable, peaking at $\approx 7.6 \times 10^{33}$ cm^{-2}s^{-1} which is just a factor of 2 above the 2011 values. However, it still came as a challenge because of the impact of pile-up. At a bunch spacing of 50 ns and increased centre-of-mass energy of 8 TeV, the average number of pile-up vertices nearly doubled comparing to that in 2011, which required a major CPU extension and implementation of sophisticated PU mitigation techniques at the HLT level. The author's contribution to the HLT development of the trigger paths important for top physics is described in Chapter 3.

2.3 Computing

The vast amounts of data delivered by the CMS detector impose high requirements on the offline computing system. During 2010–2012 operation, CMS collected ~ 10 PB of raw data per year. Including Monte Carlo simulations, reconstructed data and analysis skims, the total annual amount of data essentially doubles. To handle the distributed storage and processing of this data, not just for CMS but for the entire high energy physics community using the LHC, a worldwide LHC computing grid (WLCG) has been put in place.

WLCG is a global collaboration of more than 150 computing centres in about 40 countries.

The grid has a tiered architecture, comprising 4 tiers with different resources and services. The first one, Tier 0, is based at CERN and is responsible for data-taking. It accepts raw data from the data acquisition system and repacks it into primary datasets according to the trigger information. The raw data is archived to tape, and is also prompt-reconstructed (within 48 hours) before being distributed to the Tier 1 (T1) centres around the world. There are 8 T1 sites based at large national laboratories in collaborating countries (e.g. RAL in the UK and FNAL in the US). Each of the T1 centres is used for large-scale centrally organised data-processing activities. The data is then distributed in the reduced format (see Section 2.3.1) to a more numerous set of Tier 2 centres, typically located at collaborating universities. Each of these centres is used for the grid-based analysis and Monte Carlo simulation for the whole experiment, as well as local services for groups maintaining them. The last stage of computing system, Tier 3, is meant solely for the local institution's user analysis.

2.3.1 Event Data Model

In the basis of the CMS Event Data Model lies the concept of an event, which is physically a result of a single collision in the LHC. From a software point of view, the event is a C++ object container storing raw data from a single readout of detector electronics (e.g. hits in various sub-detectors), as well as reconstructed data which is based on this information, such as tracks, clusters and physics objects. All these C++ objects are stored in ROOT format [7].

The EDM makes use of three main data formats, based on different levels of detail and precision:

- RAW format, containing full information from the detector as well as L1 and HLT trigger decisions, with the event size of ~ 1.5 MB.
- RECO (reconstructed data) format, which is obtained from raw data by application of pattern recognition and compression algorithms. This data includes reconstructed detector hits, clusters and physics objects (electrons, muons, etc.). The typical event size is ~ 250 kB.
- AOD (Analysis Oriented Data) format, produced by filtering the RECO data from the reconstructed detector objects, leaving just the high-level physics objects required for analysis. The event size is reduced down to ~ 50 kB.

The RECO and AOD data are analysis-ready data formats, produced centrally and used by many physics analysis groups. However, further simplification of the data is also a common

practice. By transforming the C++ objects produced by CMS software into plain basic types or vectors of them, only including the analysis-specific content, the event size can be reduced down to ~ 3 kB level depending on the needs of a particular analysis. This data format is often referred to as private "ntuples", and it requires specific analysis software capable of restructuring the data into user-defined classes. By following this approach, the analysis can be run locally and generally much faster than processing the RECO or AOD data. However, it requires "ntuplising" this data every time when new centrally-recommended physics objects or corrections are produced.

2.3.2 Analysis Software

Both of the analyses described in this thesis use the CMS software framework (CMSSW [8]), as well as Bristol Analysis Tools (BAT [9]). The differential cross section analysis also uses an additional level of python scripts for post-processing [10].

CMSSW is the key CMS software framework built around the Event Data Model (see Section 2.3.1). The framework is essential for purposes of Monte Carlo simulation, detector calibration and alignment, as well as data reconstruction and analysis. CMSSW has a modular architecture, consisting of one configurable executable (cmsRun) and a large set of plug-in modules that contain all the code needed for event processing (reconstruction algorithms, calibration, etc.). Different versions of CMSSW were used for different analyses:

- `CMSSW_4_2_8` for the top mass analysis on 2011 data;
- `CMSSW_4_4_4` for the missing transverse energy analysis on 2011 data;
- `CMSSW_5_3_9` for the top cross pair cross section analysis on 2012 data.

Corresponding versions were used to produce ntuples for processing by BAT, which was used to read the data, apply selections, calculate high-level variables and to create various histograms of distributions. BAT was originally started in 2010 by Dr. Lukasz Kreczko for the needs of the Bristol top group, later on also developed by the author and other researchers from Bristol and affiliated top groups. Like CMSSW, this framework has a modular structure, with its classes falling in four main categories:

- readers, for translating plain data types from ROOT files into C++ objects;
- RECO objects, i.e. output of the readers (physical objects like leptons, jets and its collections);

- selections, for application of event selections;
- analysers for creating histograms, applying selections, algorithms, and filling histograms.

All analysers are independent from each other, making the analysis chain stable and reliable. The final set of python scripts is used to prepare the histograms, perform fitting and unfolding procedures (in case of cross section analysis), and producing final tables and plots. Rootpy package [11] was used to access ROOT libraries in python interface, and matplotlib [12] was used to create plots.

2.4 Object Reconstruction

Most CMS analyses, including the ones described in this thesis, adopt a reconstruction technique called Particle Flow (PF) [13]. This algorithm is used to obtain a global event description at level of individually reconstructed particles by means of combining information coming from all sub-detector systems. The ultimate goal is to determine type, energy and momentum of all the particles in the event with highest possible precision and in the most optimal way. The types of these particles include electrons, muons, charged hadrons, neutral hadrons and photons. All these particles are then used to reconstruct jets (Section 2.4.3), missing transverse energy (Section 2.4.1) and tau leptons from their decay products.

2.4.1 Electron Reconstruction

The reconstruction of the $t\bar{t}$ pair with an electron in the final state imposes high requirements on the electron identification and its energy-momentum measurement, precision of which is of major importance for both top mass and $t\bar{t}$ cross section measurements.

Although the CMS detector is equipped with highly accurate ECAL and tracker systems, electron identification and reconstruction is still a challenging task due to the large amount of tracker material (see Section 2.2.1). This results in a significant Bremsstrahlung photon emission, which often causes an ECAL energy deposit to be widely spread in azimuthal direction because of the high magnetic field. Therefore, dedicated algorithms were developed in order to collect all Bremsstrahlung energy deposits in the calorimeter (Bremsstrahlung recovery), and also to take into account the kinks in the electron trajectory caused by photon emissions.

Electron reconstruction in CMS has following distinct stages: seeding, track finding, pre-identification, Bremsstrahlung recovery, track-cluster linking and final identification. Historically, the original seeding algorithm was designed and optimised for isolated high-p_{T} electrons. This approach starts from ECAL clusters, and therefore is called the 'ECAL-driven' seeding. It is based on the property of the ECAL energy deposits to have narrow

width in the η coordinate, and to be widely spread in ϕ (azimuthal direction) like it was mentioned above. The electron and all the associated Bremsstrahlung energy deposits form a single "super-cluster", and the ability to correctly determine it affects the overall performance of this method. Only super-clusters with transverse energy above 4 GeV are taken into account. Super-clusters are then matched to pairs or triplets of hits in the inner tracker layers, forming the track seeds on which the electron tracks are built upon.

The performance of the ECAL-driven method is not very well suited for non-isolated and low-p_{T} electrons. This occurs mainly due to the fact that the super-cluster position and energy can be highly biased by the impact of overlapping particles, especially if the electron happens to be within a jet and therefore non-isolated. Also, high track multiplicity complicates the backward propagation from a super-cluster, because it can be consistent with a number of track seeds corresponding to other particles. To minimise the number of these fake seeds, the ratio between the HCAL and ECAL energy deposits (H/E) is required to be smaller than 0.15. The HCAL towers used in the calculation of this ratio are taken within a cone of $\Delta R = 0.3$ behind the super-cluster position. Although this helps to keep the fake seed rate under control, the efficiency for non-isolated electrons becomes rather limited. As for the low-p_{T} electrons, the wider azimuthal spread of Bremsstrahlung photons leads to poorer reconstruction of the super-cluster, biasing its position and therefore preventing the efficient matching with a track seed.

Within the particle flow method, efficient reconstruction of non-isolated and low-p_{T} electrons is particularly important since it affects the reconstruction of jets and missing transverse energy. Therefore, a different ('tracker-driven') seeding algorithm is used, which starts from reconstruction of tracks. The baseline of the CMS track reconstruction is the Kalman filter (KF) [14], which is a linear least-squares estimator based solely on Gaussian probability density functions. It is particularly suitable for muon reconstruction since it is dominated by multiple Coulomb scattering and its impact is well modelled by Gaussian fluctuations. However, this approach usually fails for electrons because Bremsstrahlung photon emission is highly non-Gaussian. To accommodate for the resulting kinks in the electron trajectory, the Gaussian-Sum Filter (GSF) [15] is used, which is essentially a non-linear generalisation of the Kalman Filter. In this method, Bremsstrahlung energy loss is modelled by a Gaussian mixture, therefore GSF track fit provides a better estimate for the inner and outer track momentum comparing to the KF algorithm. The downside of this approach is its high CPU usage, which means it can be run on a limited number of seeds.

The GSF tracks are reconstructed upon all ECAL-driven seeds. In case of the 'tracker-driven' seeds, a pre-identification based on high-purity KF tracks has been adopted. This

procedure starts with the tracks reconstructed with very tight criteria, thus decreasing the fake rate yet compromising on tracking efficiency. Then an iterative-tracking strategy is carried out by means of removing hits unambiguously assigned to tracks from the previous iteration, and also progressively relaxing track seeding criteria. This approach leads to both high efficiency and low fake rate, which is crucial for low-p_{T} and non-isolated electrons.

In the next step, track-cluster matching has to be performed. In case if electron has negligible Bremsstrahlung emission, the track is well reconstructed with the KF algorithm all the way to the ECAL internal surface, where the closest cluster is matched to the track. The corresponding cluster energy is compared with the track momentum, and if the ratio (E/p) is close to unity, the track is selected. On the contrary, if the electron experiences a significant Bremsstrahlung emission, other track characteristics have to be exploited. In this case a selection based on the number of hits in the tracker and the χ^2_{KF} of the KF fit is applied before running a GSF refit. Finally, the number of hits, the GSF refit χ^2_{GSF}, $\chi^2_{\mathrm{KF}}/\chi^2_{\mathrm{GSF}}$ ratio, the energy loss measured by the track and the quality of the ECAL cluster-track matching are fed into a multivariate analysis using a Boosted Decision Trees (BDT) estimator.

Both tracker-driven and ECAL-driven seeds are used to obtain the GSF track collection of electron candidates. In the particle flow algorithm, it is necessary to link both electron and Bremsstrahlung energy deposits to the GSF track. A super-cluster is linked to a track if the extrapolated position from the outermost tracker measurement is within the boundaries of one of the ECAL cells at the expected depth of the electron shower maximum. The preshower-ECAL and ECAL-HCAL links are made in a similar way.

Another important particle flow procedure, also driven by GSF tracks, is Bremsstrahlung recovery. In order to reconstruct an electron with correctly assigned energy and momentum, it is crucial to identify all energy deposits from Bremsstrahlung photons, thus forming a super-cluster. This procedure is carried out for each tracker layer by computing a straight-line extrapolation tangent to the track, up to the calorimeter. To determine a Bremsstrahlung photon, track-cluster linking is performed as described above. To limit the charged hadron contamination, clusters already assigned to KF tracks are not included in the calculation. Also, the distance in η coordinate between the extrapolation and the cluster is required to be smaller than 0.015, which helps to reduce the neutral particles background.

2.4.1.1 Electron Identification

Electron reconstruction in CMS is based on a characteristic signature that electrons leave in the tracker and calorimetry systems. However, other objects like charged hadrons, jets or photon conversions can produce very similar signatures and therefore may be reconstructed as electrons. Therefore, in order to distinguish these “fake” electrons from “real” ones, a

further selection has to be applied. This procedure is referred to as electron identification (or electron ID).

Initially the electron identification is performed at the final stage of the electron reconstruction process. The working points of the cuts applied are selected to be loose enough in order to satisfy most CMS analyses requirements. Afterwards, more specific (tighter) ID cuts are applied for each individual analysis, defining the working point in the trade-off between selection efficiency and fakes contamination. This will be discussed separately in the selection description for each analysis in corresponding chapters.

There are several electron identification algorithms used by various CMS analyses, and four of them are used in the analyses described in this thesis: simple cut-based (SCB ID), cuts in categories (CiC ID), particle flow (PF ID) and multi-variate analysis (MVA ID).

Simple cut-based identification is used in the High-Level Trigger, and therefore has to be as simple, fast and robust as possible. Cuts are applied on the following variables:

- H/E, i.e. the ratio of hadronic energy of the HCAL towers centred at the super-cluster position of the electron in a cone of radius $\Delta R = 0.15$, and the super-cluster electromagnetic energy;

- $\Delta\eta_{\text{in}}$, i.e. the difference in η between the extrapolated track position and η of the super-cluster;

- $\Delta\phi_{\text{in}}$, i.e. the difference in ϕ between the extrapolated track position and ϕ of the super-cluster;

- $\sigma_{i\eta,i\eta}$, i.e. cluster shape of the electron energy in a 5×5 block of crystals around the seed crystal (the one with the highest energy) [16]:

$$\sigma_{i\eta,i\eta} = \sum_{5\times5 \text{ crystals}} \left(\eta_i - \eta_{\text{seed cluster}}\right)^2 \frac{E_i}{E_{\text{seed cluster}}} \tag{2.5}$$

Although this identification method has an advantage of its simplicity, it does not show the best signal efficiency and background rejection. One of the more complex methods is the CiC ID, which exploits the categorisation of electrons. It is optimised to select electrons from different sources (W, Z and J/ψ decays) and reject fakes from jets or conversions. In order to achieve higher efficiency, electron candidates are split into categories with different signal to background ratio, allowing to better tune the working points of the cuts. The first step in categorisation is done by division between barrel and endcap regions of the detector.

Since the properties of both the tracker and calorimetry systems differ significantly in these two regions, different cuts are applied. The second step is based on the following observables:

- f_{brem}, or measured bremsstrahlung fraction, defined as:

$$f_{\text{brem}} = \frac{p_{\text{in}} - p_{\text{out}}}{p_{\text{in}}} \tag{2.6}$$

  where p_{in} is the initial track momentum at the vertex and p_{out} is the track momentum at the last hit;

- E/p, which is the ratio of the super-cluster energy and the initial track momentum.

Three categories of electron candidates are distinguished [17]:

- "Low-Brem": $0.9 < E/p < 1.2 - f_{\text{brem}} < 0.12$ (barrel), $0.82 < E/p < 1.22 - f_{\text{brem}} < 0.2$ (endcap), the fake-like region with high number of both real and fake electrons;
- "Bremming": $0.9 < E/p < 1.2 - f_{\text{brem}} > 0.12$ (barrel), $0.82 < E/p < 1.22 - f_{\text{brem}} > 0.2$ (endcap), the electrons-like region with a little contamination from fakes;
- "Bad-Track": remaining regions with a low number of real electrons.

The third and the final step in categorisation is based simply on the electron transverse energy, with the lower threshold taken as 10 GeV. In each category, the selection is performed on the basic ID variables already mentioned above for the simple cut-based identification. On top of these variables, isolation and conversion rejection variables are also used, they are discussed in Section 2.4.1.2 and Section 2.4.1.3, respectively.

The cuts within CiC ID are applied in order to maximise the signal to background ratio. Depending on the needs of different analyses, nine levels of cut severity are implemented: *VeryLoose, Loose, Medium, Tight, SuperTight, HyperTight(1-4)*. Each step decreases the fake rate by about a factor of two for electrons with $E_{\text{T}} > 20$ GeV [17]. The CiC ID is used as a primary electron identification method in the Top Mass analysis (Chapter 4).

Particle flow ID is the final step of the particle flow electron reconstruction. Following the particle flow concept, is uses information from all the CMS sub-detectors obtained in previous reconstruction steps to build new observables for electron identification. The complete list of PF ID observables is given below:

- p_{T} and η of the GSF track;

- GSF $\sigma_{p_\mathrm{T}}/p_\mathrm{T}$, transverse momentum resolution of the GSF track;
- $\#\mathrm{hits}_\mathrm{KF}$, number of reconstructed KF track hits;
- χ^2_GSF and χ^2_KF, GSF and KF goodness-of-fits;
- $\Delta\eta$: distance in η between the position of the cluster and the extrapolated position of the GSF track;
- $\sigma_{\eta\eta}$, cluster shape (Equation 2.5) of the ECAL cluster linked to the GSF track;
- $H/(H+E_\mathrm{e})$, hadron fraction of the shower, where H is the energy of the hadron cluster linked to the GSF track;
- $(E_\mathrm{e}+\sum E_\gamma)/p_\mathrm{in}$, ratio between the super-cluster energy and the inner track momentum;
- $E_\mathrm{e}/p_\mathrm{out}$, ratio between the electron cluster energy and the track outer-momentum;
- $\sum E_\gamma/(p_\mathrm{in}-p_\mathrm{out})$, the ratio between the Bremsstrahlung photon energy as measured by ECAL and by the tracker;
- $EarlyBrem$, flag of $(E_\mathrm{e}+\sum E_\gamma) > p_\mathrm{in}$ inequality, corresponding to an electron emitting an "early" Bremsstrahlung photon, i.e. before it has crossed at least three tracker layers;
- $LateBrem$, flag of $E_\mathrm{e} > p_\mathrm{out}$ inequality, corresponding to an electron emitting a "late" Bremsstrahlung electron, when the ECAL clustering is not able to disentangle the overlapping electron and photon showers;
- $f_\mathrm{brem} = (p_\mathrm{in}-p_\mathrm{out})/p_\mathrm{in}$, Bremsstrahlung fraction (Equation 2.6).

All these variables are combined into a single discriminator by a multivariate analysis technique (BDT method), which has been trained on signal and background Monte Carlo samples. PF ID is a relatively loose identification method, since it has to satisfy the needs of all analyses using particle flow collections.

Finally, MVA electron identification is used in top cross sections analysis on 2012 data. It is another multivariate analysis technique, optimised to select isolated electrons from W and Z decays. The variables used in MVA ID are also combined into a single discriminator, they are largely similar to the ones used in PF ID:

- p_T and η of the GSF track;

- $\#\text{hits}_{\text{KF}}$, number of reconstructed KF track hits;
- χ^2_{GSF} and χ^2_{KF}, GSF and KF goodness-of-fits;
- $\Delta\eta$: distance in η between the position of the cluster and the extrapolated position of the GSF track;
- $\Delta\phi$: distance in ϕ between the position of the cluster and the extrapolated position of the track;
- $\Delta\eta_{vtx}$: distance in η between the position of the cluster and the position of the GSF track at vertex;
- $\Delta\phi_{vtx}$: distance in ϕ between the position of the cluster and the position of the GSF track at vertex;
- $\sigma_{\eta\eta}$, cluster shape in η;
- $\sigma_{\phi\phi}$, cluster shape in ϕ;
- η width of the super-cluster;
- ϕ width of the super-cluster;
- H/E, ratio of HCAL and ECAL cluster energy;
- $E_{\text{super-cluster}}/p_{\text{T}}$, ratio between the super-cluster energy and the track momentum;
- $1/E_{\text{super-cluster}} - 1/p_{\text{T}}$, difference between inverse super-cluster energy and inverse track momentum;
- $E_{\text{e}}/p_{\text{out}}$, ratio between the electron cluster energy and the track outer-momentum;
- $1 - E_{1\times5}/E_{5\times5}$, where $E_{1\times5}$ is the energy in the central 1×5 strip of the 5×5 electron cluster and $E_{5\times5}$ its total energy;
- $E_{3\times3}/E_{\text{super-cluster, raw}}$, ratio of the energy of a cluster of 3×3 and the uncorrected (raw) energy of the super-cluster;
- $E_{\text{PS}}/E_{\text{super-cluster, raw}}$, ratio of the energy in the preshower detector and the raw super-cluster energy (only in the endcap region).

2.4.1.2 Electron Isolation

Isolation is an observable that allows to distinguish prompt electrons (i.e. the ones from W and Z decays) from jets faking electrons and electrons within jets. It is essentially a measure of activity around the particle. In CMS, there are two different ways of quantifying isolation: detector-based and particle-based. The detector-based isolation is defined separately for each detector sub-system (tracker, ECAL and HCAL):

- Tracker isolation, calculated as the sum of transverse momenta of all track within a cone of $\Delta R = 0.3$ around the electron, excluding the electron momentum itself;
- ECAL isolation, i.e. the sum of transverse energy of all ECAL clusters within a cone of $\Delta R = 0.3$ around the super-cluster position. The footprint of the original electron is also removed;
- HCAL isolation, defined as the sum of transverse energy of all HCAL towers within a cone of $\Delta R = 0.3$ centred at the super-cluster position.

Particle-based isolation exploits the particle flow information: it is calculated as the sum of transverse energy of all PF particles in the cone of $\Delta R = 0.3$ around the electron. Both detector-based and particle-based isolation definitions are often normalised to the electron transverse momentum (or energy in case of calorimeter isolation) in order to improve signal efficiency. The normalised sum of the tracker, ECAL and HCAL isolation variables is referred to as detector-based relative isolation (or reliso), similarly normalised particle-based isolation is called PF reliso. These quantities are crucial in various methods to estimate the QCD background contribution for many analyses, and will be used for both electrons and muons throughout this thesis.

2.4.1.3 Identification of photon conversions

Interacting with detector material, photons can convert into electron-positron pairs. Due to the large amount of material budget in the tracker (Figure 2.4), especially in the end-cap region, there is a high chance of conversions to happen. The resulting electrons can successfully fake prompt signal electrons, passing all the identification criteria and appearing isolated if the initial photon was isolated, too. Therefore conversions constitute a large proportion of the QCD background to top signal. The electron-positron pair may not be symmetrical in transverse momenta, therefore a veto on a second electron is not sufficient to reject such electrons and other conversion identification criteria are necessary.

The simplest method is based on counting the number of missing hits in the tracker. Conversions are most likely to happen at some distance from the interaction point, essentially

anywhere between the point of photon production and the end of the tracker. Therefore electrons produced in such conversions are likely not to traverse through all the pixel layers. To separate these electrons from the prompt ones, a cut on the number of missing layers can be used. However, this approach fails if conversion occurs in the beam pipe, or if the reconstructed electron track is paired with unrelated hits in the tracker, making it look like a prompt electron.

A slightly more sophisticated method is a partner track method. It is based on geometrical cuts on *dist* and *dcot* variables, which refer to the distances between tangent points of any two tracks in the $r-\phi$-plane and $r-z$-plane, respectively:

$$dist = ||\vec{r}_1 - \vec{r}_2| - r_1 - r_2| \tag{2.7}$$

$$dcot = \left| \frac{1}{\tan\theta_1} - \frac{1}{\tan\theta_2} \right| \tag{2.8}$$

Here $\vec{r}_{1,2}$ are radial vectors of the two tracks, $r_{1,2}$ are the track radii and $\theta_{1,2}$ – angles between the tracks and the beam pipe. Apart from the cuts on *dist* and *dcot* variables, tracks should have opposite charge in order to be considered to come from a photon conversion.

These two methods for conversion identification were used in the top mass analysis on 2011 data. For the top cross sections analysis on 2012 data, along with the number of hits method, a more advanced technique called vertex fit was used. It is essentially a full vertex fit of all pairs of tracks, with the selection being made on the the fit probability. The method benefits from the full use of track uncertainties and covariances, combining all information into a single discriminator. However, increased complexity of this technique makes it more CPU-intensive than the partner track method.

2.4.2 Muon Reconstruction

A good muon reconstruction and identification was one of the main CMS design requirements. Initially, muons are reconstructed independently in the tracker ("tracker tracks") and in the muon system ("standalone muon tracks") using the Kalman Filter technique [14]. Based on these objects, two different reconstruction methods are used: [18]

- Global muon reconstruction. Each standalone muon track reconstructed in the muon system is matched with the tracker track by propagating onto a common surface. A global fit is performed on the combined collection of hits from both tracks, and the resulting muon is referred to as a *global muon*.

- Tracker muon reconstruction. All tracker tracks above certain threshold ($p_{\mathrm{T}} > 0.5$ GeV,

$p > 0.5$ GeV) are extrapolated to the muon system, taking into account possible energy losses, magnetic field and multiple Coulomb scattering. If the track matches at least one muon segment in the muon system, i.e. a short track stub made of DT or CSC hits, it is considered a *tracker muon*.

Global muon reconstruction has a high efficiency for high-p_{T} muons, penetrating through more than one muon station. On the contrary, tracker muon reconstruction is more efficient for low-p_{T} muons ($p_{\mathrm{T}} < 5$ GeV), as it requires just a single muon segment. Since the $\mathrm{t\bar{t}}$ analyses described in this thesis have a semileptonic signature with exactly one energetic muon in the final state (in case of the muon channel), only the global muon reconstruction method is used, as its momentum resolution at high transverse momenta benefits from both the tracker and the muon system (Figure 2.9).

Following the reconstruction, the quality of the muon objects is verified by applying identification criteria. This is done in order to suppress hadronic punch-through, muons from decays in flight and cosmic muons. Selection is applied on the following observables:

- normalised χ^2 (χ^2/number of degrees of freedom) of the global muon fit;
- number of muon chamber hits in the global muon fit;
- number of muon stations with muon segments;
- transverse impact parameter d_{xy} (closest approach of the track to the primary vertex);
- longitudinal distance d_z of the tracker track w.r.t. the primary vertex;
- number of hits in the pixel detector;
- number of hits in the tracker layers.

The muon identification also exploits the CMS particle flow event reconstruction, making use of information coming from all sub-detectors. Particle flow muons (PF muons) are identified by imposing selection on all muon candidates reconstructed with the global muon method. This selection was optimised for identification of muons in jets, minimising the fake rate from misidentified charged hadrons, which is crucial for correct reconstruction of jets and missing transverse energy (described in more detail in the following sections). Depending on the analysis specifics, isolation requirements can also be applied; isolation definitions closely follow the ones for electrons (Section 2.4.1.2).

2.4.3 Jet Reconstruction

Hadronisation of quarks and gluons leads to production of narrow cones of particles moving in approximately one direction, called jets. This happens due to colour confinement, as particles carrying a colour charge cannot exist in free form, they have to fragment into hadrons before they can be detected directly. Therefore, to measure the initial parton's momentum and energy, all these particles must be combined into jets.

The CMS particle flow algorithm implies reconstruction of individual particles (charged and neutral hadrons, electrons, muons, photons) before combining (or clustering) them into jets. A few different jet clustering techniques exist, but the one used predominantly in CMS and exclusively in this work is anti-k_{t} algorithm [19], which defines the distance between constituent particles as:

$$d_{ij} = \min\left(\frac{1}{k_{t,i}^2}, \frac{1}{k_{t,j}^2}\right)\frac{\Delta_{i,j}^2}{R^2} \tag{2.9}$$

where $\Delta_{i,j}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$, $k_{t,i/j}$, $y_{i/j}$ and $\phi_{i/j}$ are respectively transverse momenta, rapidities and azimuth angles of particles i/j, and R is the radius parameter.

The clustering proceeds by identifying the smallest distances between particles and recombining them until all jets are formed and no particles are left. An event typically has a few well-separated hard (high-p_{T}) particles and a large amount of soft (low-p_{T}) particles. The distance d_{1i} between a hard particle 1 and a soft particle i will be fully determined by the transverse momentum of the hard particle and the Δ_{1i} separation. On the other hand, distance between soft particles with similar separation will be substantially larger. Therefore, soft particles tend to cluster around the hard ones before they cluster amongst themselves. On the output, the anti-k_{t} algorithm forms conical jets with boundaries resilient to soft radiation.

2.4.3.1 Jet Energy Corrections

Due to non-linear and non-uniform response of the calorimetry systems, jets reconstructed using detector inputs typically have energies different to the ones of corresponding Monte Carlo particle jets (or generator jets), reconstructed by clustering the four-momenta of all stable particles generated in Monte Carlo simulation. Therefore some mapping procedure is necessary. Corrections applied to reconstructed jets in order to translate the measured energy to the true particle (or parton) energy are referred to as jet energy corrections, or JEC.

CMS has adopted a factorised approach of applying jet energy corrections, meaning that each correction level takes care of a different effect. The set of corrections is applied sequentially, i.e. the output of each step is the input to the next one. Essentially, each level

of correction is a scaling of a jet four-momentum, with a scale factor depending on various parameters of the jet, typically pseudorapidity and transverse momentum. Currently, the correction levels go as follows:

- L1 Offset correction;
- L2 Relative (η) correction;
- L3 Absolute (p_{T}) correction;
- L4 EMF (electromagnetic energy fraction) correction;
- L5 Flavour correction;
- L6 UE (underlying event) correction;
- L7 Parton correction.

The goal of the L1 correction is subtraction of pile-up and electronic noise contributions from the jet energy. The energy from pile-up vertices can be deposited in calorimeters since the additional proton-proton collisions occur close enough in time to the hard scattering process. Electronic noise in calorimeter readouts also creates additional energy offset which needs to be corrected for. The scale factors for L1 correction are derived using data-driven methods (zero-bias collisions).

The relative L2 correction is designed to flatten the jet response in pseudorapidity. These corrections are extracted with respect to the barrel region in bins of p_{T}, and therefore are uncorrelated with the following L3 corrections. Both Monte Carlo and data-driven (dijet balance) methods are used to derive the L2 scale factors. Once a jet is corrected for η dependence, it is corrected back to the particle level by applying absolute L3 correction. The goal of this correction is to flatten the jet response in transverse momentum. Derivation of L3 scale factors is done by using Monte Carlo truth information, or data-driven Z/γ+jet balance techniques.

The L4 EMF correction takes care of variations in jet response with respect to electromagnetic energy fraction, i.e. the fraction of energy deposited in ECAL by hadrons. Although it has been shown that EMF-dependent correction in three parameters (p_{T}, η, EMF) can improve the observed jet resolution, this correction is optional for most CMS analyses and hasn't been used in this work.

Another optional correction is the L5 flavour correction, which is intended to correct the jets depending on the flavour of initiating partons, i.e. light quarks, heavy (b and c quarks),

or gluons. Jets originating from heavy quarks are different from light quark jets in a few aspects. The average number of charged hadrons within b-jets is higher than the one for the light jets, although the average momentum of charge hadrons is smaller. Moreover, b-hadrons have a larger branching fraction into semileptonic decays with neutrinos in the final state that cannot be detected. All these factors may lead to a substantially smaller charged hadron energy fraction comparing to the light jets, therefore the average calorimeter energy response is lower for the b-jets. These effects are meant to be taken care of by the L5 correction, which is derived either from Monte Carlo or data-driven $\mathrm{t\bar{t}}$ events.

Underlying event [20] refers to all the activity in the proton-proton collision apart from the process of interest, i.e. the hard scattering process. It includes particles coming from additional parton interactions (or multiple scatterings) and beam remnants. Depending on the analysis goals, underlying event may also include initial and final state radiation (ISR and FSR) which represent the soft gluon radiation before and after the hard scattering process, respectively. Underlying event activity results in production of additional soft jets which can bias the jet energy measurement. Optional L6 UE correction attempts to mitigate such bias. However, it was not used in the analyses described in this thesis, since underlying event is an intrinsic part of proton-proton interactions effectively modelled by Monte Carlo simulation. In particular, the top cross sections analysis is designed to be very sensitive to such higher-order processes, and not taking them into account may result in compromising the discriminating power between Monte Carlo generators.

Finally, L7 parton correction attempts to correct the jet energies back to the parton level. It is also optional and was only used in the top mass analysis. The correction factors were derived from Monte Carlo simulations by comparing the generator jets momenta to their matched partons.

Due to the fact that CMS simulation is not perfectly tuned to the data yet, additional residual L2L3 corrections are applied in order to achieve better agreement between data and simulation. It is essentially a small residual η- and p_{T}-dependent calibration applied exclusively to data, which will remain in place until CMS develops a perfectly tuned simulation reproducing the data features out of the box.

2.4.3.2 Particle Flow Jet Identification

To ensure the quality of the jets, a final identification criteria are applied to all jet objects. Particle flow jet identification (or PF jet ID) is used to reduce the noise and rate of electrons reconstructed as jets. The cuts are applied on the following observables :

- number of constituent particles;

- NHF (neutral hadron energy fraction);
- CHF (charged hadron energy fraction);
- NEF (neutral electromagnetic energy fraction);
- CEF (charged electromagnetic energy fraction).

A rather loose identification cuts (referred to as "loose PF jet ID") were used in this work: a requirement of more than one constituent particle in the jet, NHF$<$ 0.99, CHF$<$ 0.99, NEF$>$ 0 and CEF$>$ 0 in pseudorapidity region of $\eta < 2.4$.

2.4.3.3 b-tagging

The identification of jets originating from b-quarks, or b-tagging, is one of the most important tools in top quark physics, capable of significantly decreasing the background contamination of signal processes. A variety of b-tagging algorithms have been developed by CMS [21], and the one that was used in this work is called combined secondary vertex (CSV) algorithm. It is based on the fact that b-hadrons have a significant lifetime ($\sim 10^{-12}$ s) and can travel a distance of a few centimetres before decaying. Therefore jets originating from b-quarks are likely to have a secondary vertex located at a considerable distance from the primary vertex, which can be used as an efficient discriminator between light jets and b-jets. In order to maximise its efficiency, apart from the secondary vertex information the CSV algorithm also exploits the track-based lifetime information. The following variables are used in the algorithm:

- number of tracks in the jet;
- number of tracks at the secondary vertex;
- secondary vertex category;
- secondary vertex invariant mass;
- ratio of the total track energy at secondary vertex with respect to all tracks in the jet;
- pseudorapidities of tracks at secondary vertex with respect to the jet axis;
- impact parameter significance (i.e. ratio of IP to its uncertainty) of the first track that increases the invariant mass above the charm threshold of 1.5 GeV (tracks are ordered by IP significance; the mass of the system is recalculated after adding each track);

- impact parameter significances of each track in the jet.

All these observables are combined into a single discriminator, the "medium" working point of which is used in this work. It provides ~ 70 % b-tagging efficiency for mis-tag rate of approximately 1 % [21].

2.4.4 Missing Transverse Energy

Due to the energy and momentum conservation, the sum of transverse momenta and energy of all particles in the final state of proton-proton collisions is expected to be zero. However, some particles can escape the detector without being reconstructed, therefore creating the imbalance in transverse momentum which is referred to as missing transverse energy, defined as:

$$\vec{E}_{\mathrm{T}}^{\mathrm{miss}} = -\sum_{i} \vec{p^{i}}_{\mathrm{T}} \tag{2.10}$$

where $\vec{p^{i}}_{\mathrm{T}}$ are the transverse momentum vectors of all reconstructed particles. The modulus of $\vec{E}_{\mathrm{T}}^{\mathrm{miss}}$ vector is denoted by $E_{\mathrm{T}}^{\mathrm{miss}}$.

Accurate reconstruction of missing transverse energy is crucial for precise measurements of Standard Model processes with neutrinos in the final state. Top quark pair semileptonic decay is one of such processes, therefore analyses covered in this thesis implicitly (top mass) or explicitly ($\mathrm{t\bar{t}}$ cross section with respect to $E_{\mathrm{T}}^{\mathrm{miss}}$-related variables) rely on efficient reconstruction of $E_{\mathrm{T}}^{\mathrm{miss}}$. Misidentification and misreconstruction of any visible particles in the event contribute to $E_{\mathrm{T}}^{\mathrm{miss}}$ measurement, therefore it is a rather demanding task.

Just like in the case of leptons and jets, particle flow was used for $E_{\mathrm{T}}^{\mathrm{miss}}$ reconstruction: in equation 2.10 the transverse momentum vectors are of the particles reconstructed using particle flow algorithm. This procedure gives so-called raw $E_{\mathrm{T}}^{\mathrm{miss}}$ on the output. The raw $E_{\mathrm{T}}^{\mathrm{miss}}$ is systematically different from true $E_{\mathrm{T}}^{\mathrm{miss}}$, which denotes the transverse momentum carried by invisible particles. This happens mainly due to non-compensating nature of the calorimeters, effects of pile-up, noise, etc. Therefore, a set of corrections is applied:

- Type-0, which corrects $E_{\mathrm{T}}^{\mathrm{miss}}$ for pile-up;
- Type-I, a propagation of jet energy corrections (Section 2.4.3.1) to $E_{\mathrm{T}}^{\mathrm{miss}}$;
- xy-shift correction, reducing the $E_{\mathrm{T}}^{\mathrm{miss}}$ ϕ modulation.

The causes of systematic $E_{\mathrm{T}}^{\mathrm{miss}}$ ϕ modulation include detector misalignment, beam spot displacement, inactive calorimeter cells and anisotropic detector response. xy-shift correction mitigates these effects, making the measured $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution closer to true $E_{\mathrm{T}}^{\mathrm{miss}}$ distribution which is flat in ϕ because of the rotational symmetry of the collisions.

All these corrections were applied to missing transverse energy in the analyses described in this work. However, the xy-shift correction was not applied in the top mass analysis, since it was not available at the time. As this analysis is not particularly sensitive to $E_{\mathrm{T}}^{\mathrm{miss}}$, the ϕ modulation is not expected to affect the top mass measurement and its resolution.

2.5 Summary

In this chapter, the Large Hadron Collider (LHC) has been introduced to the reader. The CMS experiment including all its subsystems has been described in detail, the overview of the CMS computing model and analysis software have been shown. Reconstruction and identification methods of various analysis objects including particle flow algorithm have been discussed in detail.

3. High level triggers for Top Physics

The LHC is often referred to as a top quark factory, producing a $t\bar{t}$ pair nearly each second of its nominal operation. While the production rate of ≈ 1 Hz seems manageable in terms of recording the data, it is significantly complicated by background processes with similar signatures occurring at much higher rates.

The trigger is the starting point of any physics event selection process, and therefore is clearly important for any physics analysis. As it was mentioned in Section 2.2.6, the CMS L1 trigger rate is limited to ~ 100 kHz. In order to meet the data recording constraints of approximately 300 MB s^{-1}, this rate is further reduced down to ~ 300 Hz, which is done by the HLT system. The total rate budget has to be shared between various physics analysis groups (e.g. Top, Higgs, Exotica, etc). Corresponding allocations are determined by CMS trigger coordination according to CMS physics goals, and can be a matter of serious debate.

During the LHC operation in 2011 and 2012 under conditions of gradual increase of instantaneous luminosity and pile-up, but very limited rate budget, trigger developers constantly tackled the challenge of finding the best compromise between growing rates and maintaining reasonable signal acceptance. While the simplest approach is tightening the cuts on physical quantities like lepton or jet transverse momenta, it is not favourable since it lowers the number of stored signal events and decreases the phase space which is crucial for new physics searches as well as Standard Model precision measurements. Therefore, development of more efficient algorithms allowing to keep high level of acceptance for signal events, whilst effectively rejecting background events, is the most preferable solution. This can often be achieved by increasing the level of approximation of the online (HLT) object reconstruction, making the algorithms closer to their sophisticated offline counterparts. However, it leads to a higher execution time, which is limited by computing resources available for HLT reconstruction. Hence, the CPU timing is another major constraint faced by the HLT developers.

This chapter covers the author's contribution to development and verification of High-Level Triggers for top physics with semileptonic signature, where one of the W bosons decays into an electron and a neutrino. The description of top triggers, efficiency measurement, validation of jet energy corrections and pile-up subtraction applied at the HLT level are discussed in relevant sections of this chapter.

3.1 Level-1 triggers

The CMS L1 trigger [22] is built of custom electronics processing the data from the calorimeters and the muon system. It is the first filtering stage of any physics selection, therefore it has high requirements on efficiency and phase space. The input rate of events, i.e. the beam crossing frequency of $\sim 5 \times 10^8$ Hz must be reduced down to ~ 100 kHz, being a major challenge considering a gradual increase of instantaneous luminosity throughout the LHC operation. The rate budget of ~ 100 kHz is shared between several L1 triggers corresponding to different physics objects being present in the event. For top physics signatures with a single electron in the final state, the following triggers have been used:

- L1_SingleEG_18
- L1_SingleEG_20
- L1_SingleEG_22

These triggers are referred to as electron/photon (e/γ, or EG) triggers. They scan the ECAL for groups of crystal cells (4×4) with a cumulative energy above the threshold that is suggested by the name of the trigger (18 GeV, 20 GeV and 22 GeV). The choice of threshold is determined by sustainability of the trigger rate at a given instantaneous luminosity.

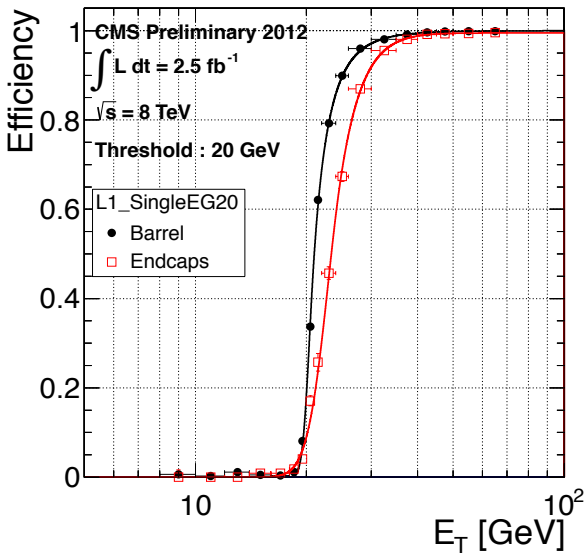


Figure 3.1: Efficiency of the 20 GeV e/γ trigger as a function of offline E_{T}, shown separately for barrel and endcap regions [23].

The efficiency of the L1_SingleEG_20 trigger as a function of the offline-reconstructed electron transverse momentum is shown on Figure 3.1. The L1 trigger decisions are then used as an input "seeds" to the HLT system, which is described in the following section.

3.2 High-level triggers for top physics

The High-Level Trigger [24] is the crucial part of CMS event selection process. As it was mentioned in Section 2.2.6, it is based on software algorithms running on the Event Filter Farm, i.e. a large cluster of commercial CPUs. The HLT reconstruction, often referred to as online reconstruction, is implemented in the same software framework (CMSSW) which is used for offline reconstruction, and the algorithms can be very similar. However, the key difference between online and offline reconstruction is the running time. Since the HLT selection has to be performed in real time, it imposes a significant constraint on computing resources, enforcing to compromise on robustness and efficiency of online algorithms. With an exception of small samples for performance monitoring, data rejected by the HLT is lost irrevocably. Therefore, correct and efficient operation of the HLT is of major importance for CMS physics programme.

Modular structure of CMSSW provides high flexibility in use of selection and reconstruction algorithms, allowing their continuous optimisation according to changes in physics needs and data-taking conditions. Various modules account for reconstruction of different physics objects and their matching with L1 objects, filtering, logging, monitoring, etc. All these modules are grouped into so-called trigger "paths", ultimately giving trigger decisions on whether to accept an event or not. A typical example of a trigger path used for top physics is an electron-plus-jets trigger, which requires the presence of isolated electron and at least three energetic jets in the event.

There are two types of trigger paths: prescaled and unprescaled. Prescaling is a method of reducing the trigger rate by only recording a fraction of events that pass the trigger selection. For example, if the prescale value of a particular trigger is 10, only one out of ten events that fire the trigger will be actually recorded. Usually prescaled triggers are used for performance monitoring and background estimation as they can impose a much looser selection criteria yet still have manageable rate. However, such triggers are unfavourable for signal selection, as they greatly reduce signal acceptance.

A set of trigger paths and their prescale factors is combined in the HLT configuration, referred to as the HLT menu or table [25]. Since the start of data-taking in 2010, the CMS trigger coordination adopted a "trigger train" model, implying the schedule of regular deadlines for updates in the trigger menu. These deadlines are usually imposed fortnightly,

according to changes in instantaneous luminosity or other alterations in data-taking conditions. In order for a trigger path to be included in the trigger menu, it has to be implemented in the HLT configurations database, validated by measuring the trigger rate and signal efficiency, tested for CPU timing and finally approved by the trigger studies group.

As it was mentioned before, the top quark pair decay covered in this work has a semileptonic signature, containing an electron, at least four jets and a neutrino in form of $E_{\mathrm{T}}^{\mathrm{miss}}$. Due to the nature of the decay, all these objects are highly energetic and can be triggered on with rather high thresholds on transverse energy. During the start-up year of 2010 when the instantaneous luminosity went up from $\sim 10^{27}$ cm^{-2}s^{-1} to $\sim 10^{31}$ cm^{-2}s^{-1}, top analyses with the signature of interest used the single electron trigger, requiring just one electron with certain isolation and transverse momentum criteria. However, by extrapolating the trigger rates on higher luminosities foreseen in the following years it became obvious that the single electron trigger rate would quickly become too high for the rate budget restrictions at the time. Therefore, alternative ways of adding other objects from the $\mathrm{t\bar{t}}$ signature were explored, naturally leading to electron-plus-jets trigger solution.

A typical electron-plus-jets trigger consists of a standard electron module, a jet module, a cleaning module to remove overlap between jet and electron collections, and a jet multiplicity filter. The electron module is constructed from the following sub-modules:

- L1 object and ECAL super-cluster matching;
- ECAL transverse energy filter;
- ECAL electron ID and isolation filter;
- HCAL electron ID and isolation filter;
- tracker electron ID and isolation filter.

To minimise the total running time, all the sub-modules are ordered by speed, starting from the fastest. The matching module finds the ECAL super-cluster closest to the L1 object in η and ϕ dimensions. If the super-cluster is located within a certain η and ϕ range, the event is passed to the next module, otherwise it is rejected.

The following (ECAL) module calculates the super-cluster transverse energy (E_{T}), and in case if it is above 25 GeV, the electron ID and isolation criteria are imposed for both ECAL and HCAL (see Section 2.4.1). The working points of identification and isolation and their naming conventions are shown in Table 3.1.

In the main trigger path used for signal rather than background estimation, Very Tight (VT) working point for calorimeter identification (CaloID) and Tight (T) working point for calorimeter isolation (CaloIso) were used. This was motivated by consistency with similar criteria for the single electron trigger used in 2010.

For the events that pass all described criteria in the electron module, a simplified algorithm for offline iterative tracking is applied. This algorithm has a faster running time at the expense of worse resolution. In order to decrease the CPU usage, the Combinatorial Track Finder (CTF) [26] is used instead of the GSF one, and the tracking is confined to a small region around the electron. The events are required to pass the tight criteria of the tracker electron ID (TrkId) which is calculated at this stage. Finally, the tight working point of the tracker isolation (TrkIso) is applied, where the track p_{T} is summed over all tracks within $\Delta R = 0.3$ cone around the electron track, excluding the electron itself. Provided that event passes all the electron sub-modules, the jet module is then executed.

The jet reconstruction is performed with the anti-k_{t} algorithm (see Section 2.4.3) with a radius parameter of $R = 0.5$. Historically, at the early stages of LHC operation, CMS analyses mostly used calorimeter jets, i.e. jets reconstructed using calorimeter information exclusively. Therefore, initially the trigger jet module used calorimeter jet collection, which provided fast reconstruction yet worse resolution hence lower trigger efficiency comparing to PF jets which were introduced later on during 2011 data-taking. In order to work online, PF jet reconstruction was modified to be compatible with the simplified tracking algorithm. Despite the fact that running the jet module with PF jet reconstruction is highly CPU-intensive and therefore can not be used as an unprescaled stand-alone trigger, it works effectively together with the electron module since it reduces the input event rate considerably. Both calorimeter and PF jet versions of the module impose the 30 GeV cut on transverse momentum on the jets as well as the pseudorapidity cut of $|\eta| < 2.6$. Subsequently, the jet

Working point	**CaloID**	**CaloIso**	**TrkId**	**TrkIso**
VeryLoose (VL)	$H/E < 0.15$ (0.10)	ECAL iso/$E_{\mathrm{T}} < 0.2$ (0.2)	$\Delta\eta < 0.01$ (0.01)	track iso/$p_{\mathrm{T}} < 0.2$ (0.2)
	$\sigma_{i\eta i\eta} < 0.024$ (0.040)	HCAL iso/$E_{\mathrm{T}} < 0.2$ (0.2)	$\Delta\phi < 0.15$ (0.10)	
Loose (L)	$H/E < 0.15$ (0.10)			
	$\sigma_{i\eta i\eta} < 0.014$ (0.035)			
Tight (T)	$H/E < 0.10$ (0.075)	ECAL iso/$E_{\mathrm{T}} < 0.125$ (0.075)	$\Delta\eta < 0.008$ (0.008)	track iso/$p_{\mathrm{T}} < 0.125$ (0.125)
	$\sigma_{i\eta i\eta} < 0.011$ (0.031)	HCAL iso/$E_{\mathrm{T}} < 0.125$ (0.125)	$\Delta\phi < 0.07$ (0.05)	
VeryTight (VT)	$H/E < 0.05$ (0.05)			
	$\sigma_{i\eta i\eta} < 0.011$ (0.031)			
WP80	$H/E < 0.10$ (0.05)	ECAL iso/$E_{\mathrm{T}} < 0.15$ (0.1)	$\Delta\eta < 0.007$ (0.007)	track iso/$p_{\mathrm{T}} < 0.05$ (0.05)
	$\sigma_{i\eta i\eta} < 0.01$ (0.03)	HCAL iso/$E_{\mathrm{T}} < 0.1$ (0.1)	$\Delta\phi < 0.06$ (0.03)	

Table 3.1: Naming conventions for electron trigger working points. The given values are for the barrel region of the detector and the values in brackets are for the endcap region.

collection is passed to the cleaning module.

Within CMS, electron and jet collections are reconstructed independently, and since they often leave similar footprints in the detector, these collections can overlap, causing potential double-counting. The electron-jet cleaning module removes electron candidates found by the electron module from the collection of jet candidates found by the jet module. A jet is removed from the jet collection if there is an electron within a cone of $\Delta R = 0.3$ around the jet axis. However, occasionally a genuine jet can be incorrectly identified as an electron. In this case the fake electron can be removed from the jet collection, biasing the performance of the jet multiplicity filter. To tackle this issue, separate jet collections are created for all electrons in the event, where jets are only cleaned from additional (non-signal) electrons. All these jet collections are passed to the following jet multiplicity filter, which accepts and event if at least one collection contains a desired number of cleaned jets.

Electron-plus-jets triggers successfully functioned throughout 2011 and 2012 years of data taking. During 2011, they were used in the top mass analysis and differential cross section analysis with respect to missing transverse energy, described in this thesis. The 2012 cross section analysis used a single electron trigger with a lepton p_{T} threshold of 27 GeV, reasonably tight (WP80) lepton isolation requirements, but no specific jet requirements. In contrast to 2011 running, this trigger was decided to be unprescaled in the trigger menu for the whole period of data-taking in 2012, regardless of having a significant rate. It is favourable for many analyses as it is much more straightforward in terms of calculating efficiency and acceptance scale factors, also reducing the complexity of estimating the systematic errors associated with the triggering.

3.3 Trigger efficiency measurement

One of the most important characteristics of a trigger path is its efficiency. In a broad sense, selection efficiency can be defined as a conditional probability that a single event passes the selection, given all other conditions (detector configuration, preselection, etc). In the context of the trigger, efficiency highly depends on the offline selection it is measured with respect to. As a matter of fact, different studies can apply different offline selections, therefore the same trigger may have different efficiencies for each of them.

The trigger efficiency can be measured in both data and Monte Carlo simulation using various methods. For all of them, the study has to be performed on some initial preselected dataset. Ideally, this preselection needs to be unbiased with respect to selection imposed by the trigger and offline selection. However, in reality it implies running on vast amounts of data, with very little number of events satisfying the trigger selection, leading to insufficient

statistics or non-feasible computing resources needed for the study. Therefore, some other trigger is used for preselecting the initial dataset, somewhat correlated to the trigger under study in order to obtain reasonable statistics. To correctly measure the efficiency of a trigger path, this correlation has to be taken into account.

The efficiency of a trigger A can be written as the following conditional probability [27]:

$$\epsilon_A = P(A|\vec{x}, T, D) \tag{3.1}$$

where $\vec{x}$ are reference quantities used for triggering (e.g. lepton p_{T}), and T is an offline selection, and D is a combination of all other factors like detector effects. Assuming that these factors, along with the offline selection and quantities used for triggering are fixed, we can denote $\epsilon_A = P(A)$ for brevity.

If B is the trigger used for preselection, then

$$P(A, B) = P(A|B) \cdot P(B) = P(B|A) \cdot P(A) \tag{3.2}$$

Therefore, as it immediately follows from Bayes' theorem,

$$P(A) = \frac{P(A|B) \cdot P(B)}{P(B|A)} \tag{3.3}$$

Here the conditional probability $P(A|B)$ can be estimated by taking the ratio of the numbers of events that pass the triggers A and B, given that they all pass offline selection. Efficiency of the auxiliary trigger $\epsilon_B = P(B)$ can be estimated from data by using a dataset obtained with looser ("minimum bias") selection. Unfortunately, conditional probability $P(B|A)$ can not be measured using real data, as by the nature of the study the trigger A only considers events preselected by the trigger B. However, the trigger B is usually chosen in such way that it is safe to assume that $P(B|A) = 1$ with negligible uncertainty. This is the case if the preselection trigger criteria are considerably looser than those of the trigger under study. To summarise, the trigger efficiency can be measured as:

$$\epsilon_A = P(A|B) \cdot P(B) = \frac{N_A}{N_B} \cdot \epsilon_B \tag{3.4}$$

where $N_{A(B)}$ is the number of events that pass the offline selection and fire the trigger A (B).

In case of electron plus jets triggers, the efficiency can be factorised in contributions of

leptonic and hadronic parts of the trigger:

$$\epsilon_{ele+jets} = \epsilon_{ele} \cdot \epsilon_{jets} \tag{3.5}$$

This method can be used under assumption that the probability of finding an electron in the event is independent of the presence of the jets in the event, i.e. leptonic and hadronic "legs" of the trigger are uncorrelated. Although such correlations do exist, it has been shown that most of them are negligible and factorisation method provides a meaningful estimate of the trigger efficiency [28].

Trigger efficiency can be parametrised by various physical quantities of the triggered objects, e.g. reconstructed jet pseudorapidity and transverse momentum. As the background spectrum and trigger rate fall down as a function of jet p_{T}, trigger efficiency can be described by a turn-on curve approaching a plateau region, where the efficiency is maximum and constant within the statistical uncertainty. It is important for the trigger to have a sharp turn-on, with the nominal 95 % efficiency point being below or approximately at the cut value of offline reconstructed jet p_{T}. This can be achieved by adjusting the online cuts to be lower than offline ones: if these thresholds are close, then the trigger would bias the analysis distributions. In this case, non-flat p_{T}-dependent scale factors have to be applied to the total number of signal events, which also complicates estimation of the systematic uncertainty attributed to the trigger. As for the efficiency dependency on the jet angular distribution, it is expected to be flat – however, as it can be seen from the following section, this is not always the case.

3.3.1 Validation of jet energy corrections and charged hadron subtraction for top triggers

In the end of 2011, to match the jet reconstruction method used by most CMS analyses, the calorimeter-based jet module in electron-plus-jets was replaced with PF-based one. This was done in an attempt to improve the p_{T} resolution of online jet reconstruction, reduce the rate and obtain sharper turn-on curves of the trigger efficiency. In the first iteration of PF-based trigger paths, no jet energy corrections were applied to the jets online, which resulted in observation of a non-flat trigger efficiency with respect to jet pseudorapidity. After a few months of work, responsible physics object group (known as "JetMET") provided a series of jet energy corrections to be applied at the HLT level, which were tested by the author.

Pile-up turned out to be another major issue particularly specific to 2012 data-taking period. Increased centre-of-mass energy and instantaneous luminosity at a bunch spacing of 50 ns effectively doubled the average number of pile-up vertices comparing to 2011 conditions.

This led to a significant increase in rate of PF-jets triggers, requiring implementation of pile-up mitigation techniques at the HLT level. As a result, charged hadron subtraction (CHS) for PF jets, also known as "PFnoPU", was introduced online. As the name suggests, it essentially removes the charged hadrons associated with secondary vertices from PF candidates, which helped to reduce the trigger rate dramatically. As top triggers operate with PF jets, they were greatly affected by pile-up, therefore this correction also had to be validated for electron-plus-jets triggers.

In this work, an impact of both jet energy corrections and pile-up subtraction on electron-plus-jets trigger efficiency was investigated. A rather complicated name of the main signal trigger under study (HLT_Ele25_CaloIdVT_CaloIsoT_TrkIdT_TrkIsoT_TriCentralPFJet30) suggests working points of tracker and calorimeter ID and isolation criteria for the signal electron (see Table 3.1), as well as transverse momenta requirements for both electron (25 GeV) and each of the three particle flow jets (30 GeV). Since only hadronic part of the trigger was of interest for this particular study, the efficiency was measured by using a "Single Electron" primary dataset of raw 2012 data. This primary dataset was constructed by accepting events passing at least one of the several single electron triggers with various p_{T} and isolation requirements. One of the loosest triggers, referred to as HLT_Ele27_WP80, operates with working point "WP80", also shown in Table 3.1, and electron p_{T} threshold of 27 GeV.

Offline selection used to obtain a clean $\mathrm{t\bar{t}}$ sample is outlined in Section 4.2, with an exception of the jet multiplicity requirement: here at least three jets were required. A tighter requirement of at least four jets was also of interest, as described below. The trigger efficiency is then calculated as:

$$\epsilon = \frac{N_{\mathrm{fired}}}{N_{\mathrm{selected}}} \tag{3.6}$$

where N_{selected} is the number of events from the primary dataset passing the described offline selection, and N_{fired} is the number of events which in addition fired the electron plus three jet trigger.

All the efficiencies were measured in bins of jet p_{T} and η. For these variables, different fit functions were used. The efficiency in p_{T} was fitted with the default function describing the turn-on curve:

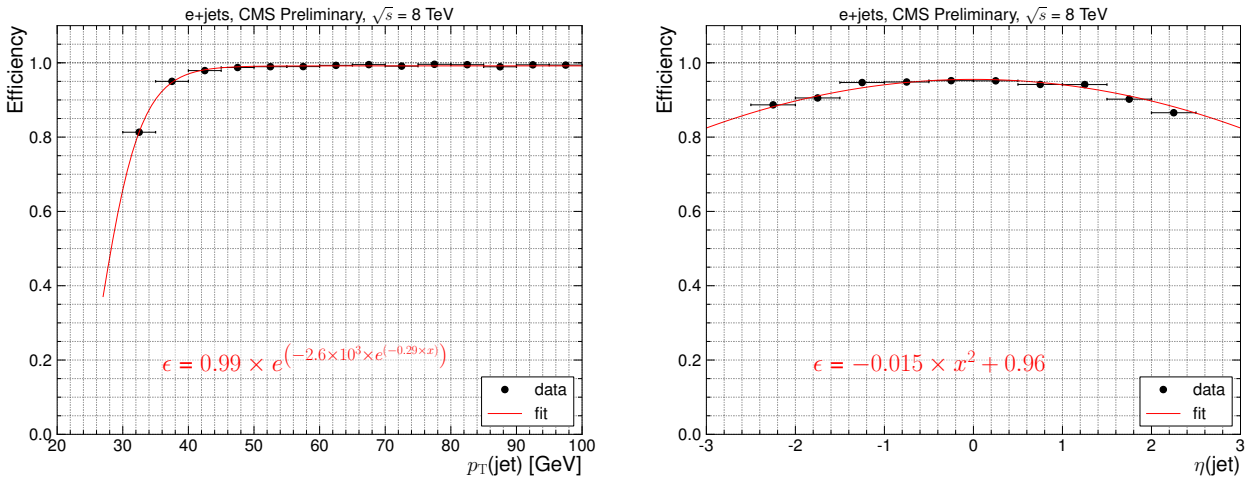
$$\epsilon(p_{\mathrm{T}}) = A \times e^{\left(B \times e^{C \times p_{\mathrm{T}}}\right)} \tag{3.7}$$

while the efficiency in η was fitted with

$$\epsilon(\eta) = A \times \eta^2 + B \times \eta + C \tag{3.8}$$

where A, B and C are the fit parameters.

Although the η dependency was expected to be flat, the observed structure in the η distribution of the PF jet triggers suggested the use of a parabolic fit function. In fact, this drop of the efficiency in the endcap region was the main motivation for the study, as it was believed to be due to the fact that the p_{T}- and η-dependent jet energy corrections were not applied online, causing inconsistency between HLT and offline reconstructed jets. Efficiency of the main electron-plus-jets trigger with no jet energy corrections applied is shown on Figure 3.2. Clearly, parabolic shape is seen on the efficiency plot with respect to pseudorapidity of a third reconstructed jet (all the jets were sorted by the value of transverse momentum).

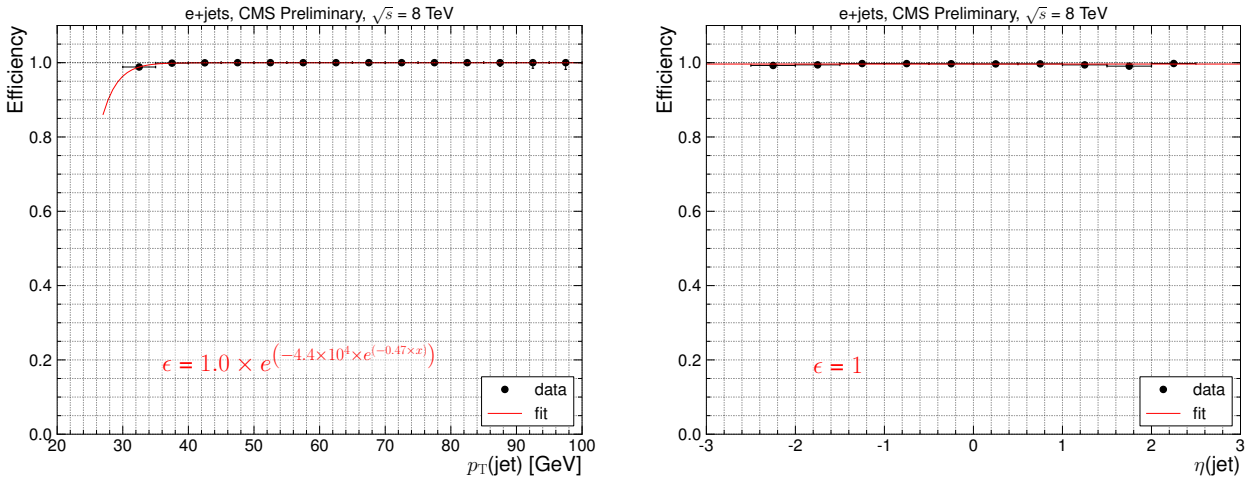


(a) (b)

Figure 3.2: Trigger efficiency for HLT_Ele25_CaloIdVT_CaloIsoT_TrkIdT_TrkIsoT_TriCentralPFJet30 (uncorrected) as a function of jet p_{T} (a) and η (b) for events with at least three jets

It has to be emphasised that only efficiencies with respect to offline selection with at least three jets were affected. Requiring a fourth jet in the event increases the trigger efficiency to nearly 100 %, as it can be seen on Figure 3.3. This implies that observed inefficiency in the endcap region was not critical for analyses requiring at least four jets in the event (including the ones in this thesis), however, it was still important to investigate this effect in order to better understand any possible systematic errors due to the triggering.

For the purpose of the study, a modified HLT menu was created, containing new versions of the main signal trigger with jet energy corrections and charged hadron subtraction applied. This menu was run on a $\mathrm{t\bar{t}}$ skim of a RAW dataset mentioned above, and resulting trigger



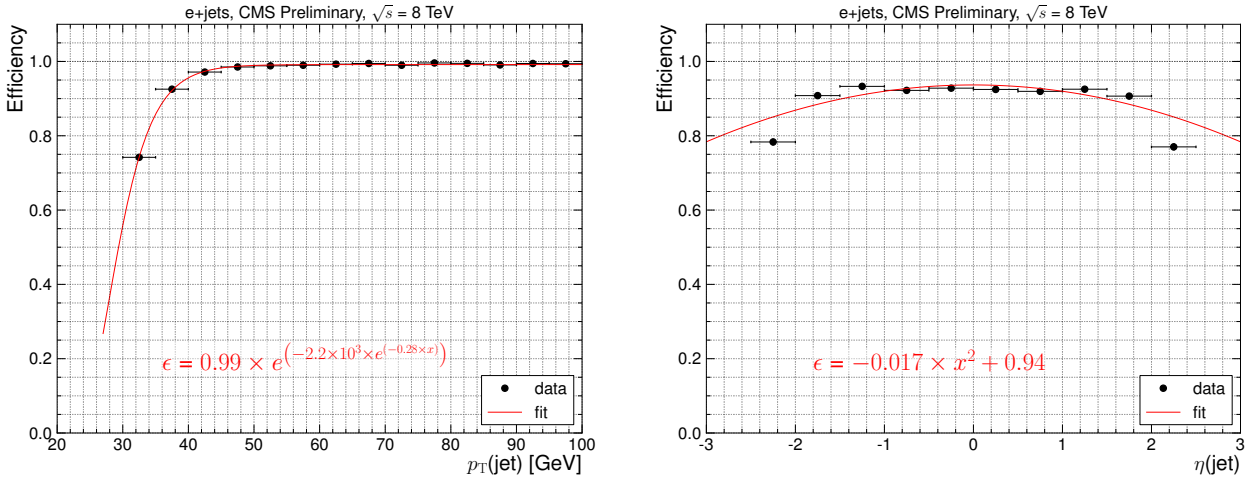
(a) (b)

Figure 3.3: Trigger efficiency for HLT_Ele25_CaloIdVT_CaloIsoT_TrkIdT_TrkIsoT_TriCentralPFJet30 (uncorrected) as a function of jet p_T (a) and η (b) for events with at least four jets

flags as well as jet collections were used to produce the final plots.

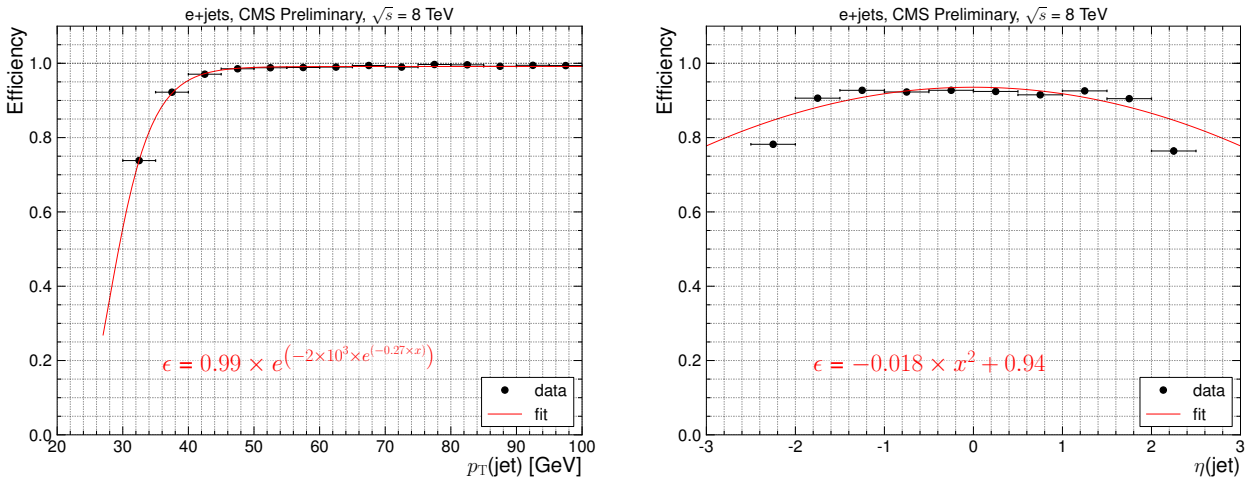
Figure 3.4 presents the results for the aforementioned trigger path but with jet energy corrections applied, and Figure 3.5 shows it for the same path with both charged hadron subtraction and jet corrections applied. It is obvious that all these corrections haven't solved the inefficiency problem in the endcaps, which caused some confusion in the group that provided them and triggered additional scrutiny from the trigger and jet/E_T^{miss} experts.

In order to gain further understanding, jet response distributions were investigated. Jet response is a ratio of the transverse momenta of offline-reconstructed and HLT jets, ideally produced from the same objects in the detector. To find this ratio, all the HLT jets were matched with offline jets found within a cone of $\Delta R = 0.3$ for each online jet. Matching efficiency proved to be close to unity, and following distributions were produced only for the successfully matched jets. Figure 3.6 shows the response plots for trigger paths with different corrections: uncorrected, only jet energy corrected (JEC) and with both jet energy corrections and charged hadron subtraction (JEC+CHS) applied. It can be seen that corrections work reasonably well in the barrel region, but substantially underestimate the energy of online jets in the endcap region. The response plot shown in slices of offline jet transverse momenta (c) suggests that corrections work better for jets with higher energy, which can also be seen from the distribution with respect to jet p_T, however, the overall response in the endcap region is still far from being perfect.



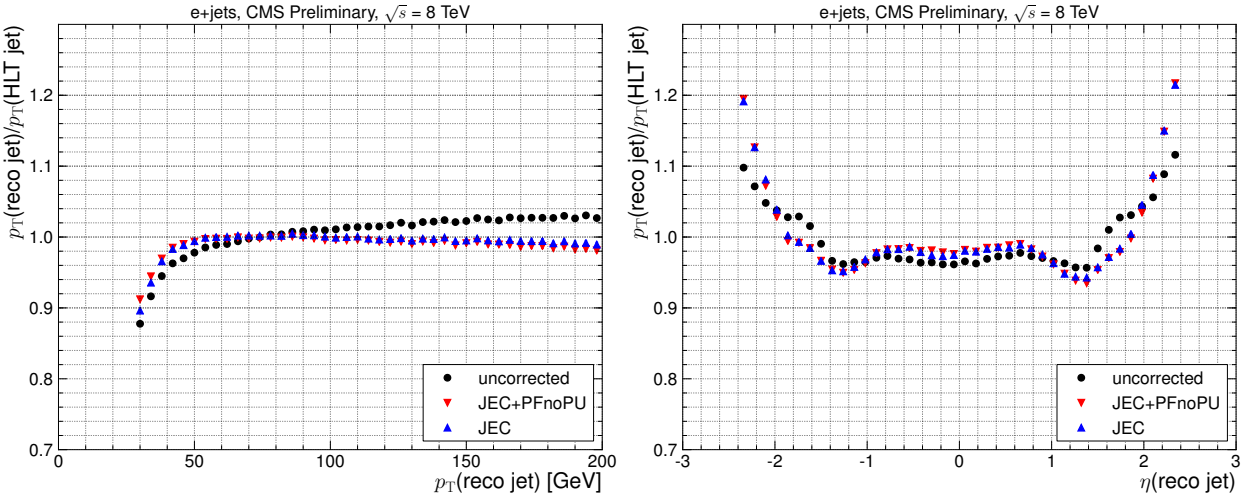
(a) (b)

Figure 3.4: Trigger efficiency for HLT_Ele25_CaloIdVT_CaloIsoT_TrkIdT_TrkIsoT_TriCentralPFJet30 with jet energy corrections applied online, as a function of jet p_T (a) and η (b) for events with at least three jets

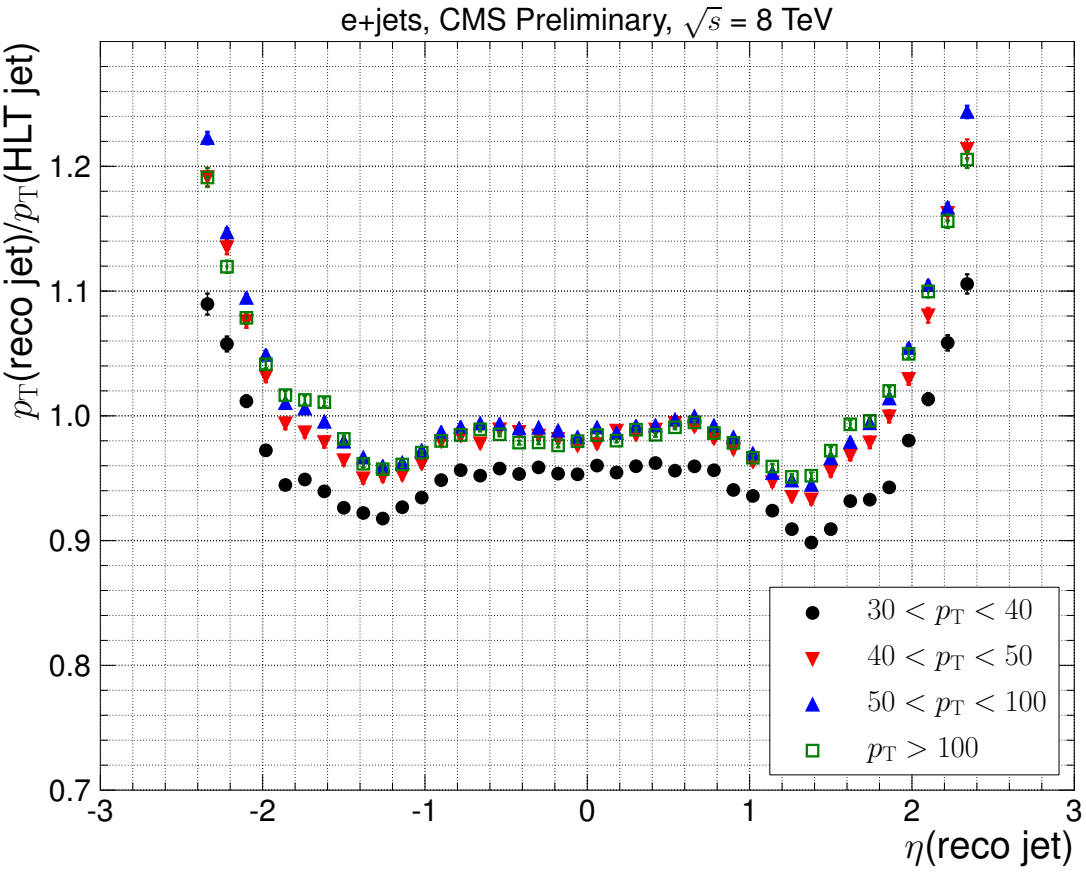


(a) (b)

Figure 3.5: Trigger efficiency for HLT_Ele25_CaloIdVT_CaloIsoT_TrkIdT_TrkIsoT_TriCentralPFJet30 with jet energy corrections and charged hadron subtraction applied online, as a function of jet p_T (a) and η (b) for events with at least three jets



(a) (b)



(c)

Figure 3.6: Ratio of matched offline and online reconstructed jet transverse momenta as a function of offline jet p_T (a) and η (b) for different jet corrections and in slices of offline jet p_T (c)

Shorty after these results were presented to the trigger coordination, they were confirmed independently by other experts within the JetMET group. Because of the time constraints associated with the luminosity increase schedule, a following workaround solution was suggested. In order to re-gain the trigger efficiency, the threshold for the third jet was dropped from 30 GeV to 20 GeV, which allowed to apply jet energy corrections and charged hadron subtraction with a minimum loss in the efficiency in the endcap region. However, this was a subject of a different study.

3.4 Summary

This chapter covered the service work performed by the author for CMS collaboration, namely development, validation and maintenance of high-level triggers for top quark physics, specifically for the electron channel of semi-leptonic $t\bar{t}$ decay signature. The importance and typical challenges of trigger development has been discussed. The modular structure of high level triggers has been described in detail for electron-plus-jets paths. A core of author's contribution: the investigation of the effect of jet energy corrections applied on the HLT level to top trigger paths has been presented, showing how it affects trigger efficiency and jet response distributions. An impact of charged hadron subtraction applied online has also been studied and demonstrated.

4. Top Quark Mass Measurement

The top quark mass measurement using 2011 LHC data recorded by the CMS detector at a centre of mass energy of $\sqrt{s} = 7$ TeV is presented in this chapter. The analysis was a cross-check to the official CMS top quark mass measurement at 7 TeV in lepton plus jets channel published in 2012 [29] in Journal of High Energy Physics. Only electron plus jets channel is described in this thesis, as the muon side of the analysis was performed by a different group at CERN, although in close collaboration.

The mass extraction technique used in this analysis, referred to as the ideogram method, is essentially the same to one used in the CMS measurement of the mass difference between top and antitop quarks [30] and the top quark mass measurement in 2010 [31]. In this chapter, this technique is described in detail. Additionally, data and Monte Carlo samples, event selection process, the kinematic fit used for $t\bar{t}$ reconstruction, as well as the calibration of the ideogram method, systematic uncertainties and the final result are presented in relevant sections of the chapter.

4.1 Data and Simulation

4.1.1 Data

The data used in this work is the full 2011 dataset recorded by the CMS detector with a total integrated luminosity of 5.0 ± 0.1 fb^{-1}. As only the electron channel is covered by this particular analysis, all data was pre-selected by single electron plus jets high level triggers, described in detail in Chapter 3.

4.1.2 Simulation of signal and background processes

To develop and test any analysis technique in particle physics, simulated events from Monte Carlo (MC) generators as well as a working simulated model of the detector are inevitably needed. In this analysis, $t\bar{t}$ signal decays and background processes were generated with a set of different MC generators including MadGraph [32], PYTHIA [33, 34] and POWHEG [35]. Supplementary package (TAUOLA [36]) was used for tau lepton decay description. In this section, all the event generators as well as different MC tunes used for evaluation of systematic uncertainties are briefly described. Additionally, Geant4-based [37] CMS detector simulation is mentioned, and eventually Monte Carlo samples for all processes are presented.

4.1.2.1 Monte Carlo event generators

Simulation of $t\bar{t}$ signal and background processes, described in Section 1.2, was performed using several Monte Carlo event generators. The process of simulating a hadron collision event is rather complicated, and usually includes generation of the following steps:

- incoming hadrons (protons);
- hard scattering of partons;
- boson radiation (Z, γ, g);
- parton shower;
- hadronisation of partons.

The performance of MC generators is usually optimised for one or more of these steps, therefore they are often used in combination, e.g. $t\bar{t}$ signal is modelled with MadGraph and PYTHIA. Standalone generators can also be used for modelling the theory systematic effects, which is also exploited in this thesis.

MadGraph [32] is a matrix-element Monte Carlo event generator. For any renormalisable Lagrangian-based model it produces all possible Feynman diagrams and automatically generates its matrix elements at the tree-level, performing the integration over all phase-space. This calculation is then used to produce the cross section of various processes and subprocesses as well as partons and kinematics of an event, including decays that are described using spin correlations. Partons from matrix element calculations are then matched to parton showers from hadronisation of quarks and gluons, which is simulated in PYTHIA (see below) along with the fragmentation of initial protons and the soft scattering of underlying event (mentioned in Section 2.4.3.1). The matching is done according to so-called MLM prescription [38] if a parton-jet pair satisfies a certain ΔR separation criteria. If no or more than one matched jets are found, events are rejected. There is also a certain transverse energy threshold requirement for the partons to be considered in the matching. Default values as well as variations used to estimate the systematic effect of the matching threshold are shown in Table 4.1.

PYTHIA [33, 34] is one of the most widely used MC generator in high energy physics. It is a standard tool used for simulation of quark and gluon hadronisation, multi-particle production, beam remnants, initial protons fragmentation, etc. In this work, PYTHIA is

used to generate QCD multi-jet production and underlying event on top of other generators providing partons from hard processes.

MC@NLO [39] is a parton shower MC generator providing next-to-leading (NLO) order corrections for a simulated process. It represents a major improvement in precision and accuracy of physics simulation comparing to leading order (LO) generators like PYTHIA. NLO implementation can handle additional partons in the final state coming from the hard scattering, which is not possible with LO calculations. Therefore it includes additional dynamic features, giving a major advantage for heavy flavour physics.

POWHEG (Positive Weight Hardest Emission Generator) [35] also works at NLO precision. Its main difference to MC@NLO generator lies in the order of process generation: POWHEG initially generates the hardest radiation. This technique allows to avoid double-counting coming from subsequent radiation which may happen with MC@NLO method when matching QCD NLO calculations with parton showers.

4.1.2.2 Matching Threshold and Factorisation Scale

As described above with the example of MadGraph generator, there is a certain threshold for partons transverse momenta from matrix element calculations to be matched with parton showers, often referred to as ME-PS threshold. To estimate the systematic uncertainty arising from it, additional MC samples were generated with either halved or doubled value of the threshold. This procedure was performed for signal $t\bar{t}$ + jets and background W + jets and Z + jets samples. Similarly, the systematic effect of the choice if factorisation scale Q^2 (described in Section 1.2.1) is determined. A summary of central values and variations used to estimate these systematic uncertainties is shown in Table 4.1, whereas the list of systematic MC samples is shown in Table 4.4.

4.1.2.3 CMS detector simulation

Detector simulation is an essential final step on top of the full physics simulation of proton-proton collisions. All generated particles resulting from these collisions are propagated to the simulated layers of the detector so that their interaction with the detector material and detector response can also be modelled. The CMS simulation is based on the Geant4 (GEometry ANd Tracking [37]) package, which is implemented in CMS software CMSSW [8]. This package fully recreates the geometry of the detector, including all its subsystems as well as a detailed map of the magnetic field.

In general, the simulation procedure includes modelling of the following steps:

- interaction region;

parameter	factorisation scale	matching threshold
$t\bar{t}$ events		
central value	$Q^2 = m_t^2 + \sum_{\text{jets}} p_T^2$	20 GeV
variations	$(0.5 \cdot Q)^2, (2 \cdot Q)^2$	10 GeV and 40 GeV
W + jets and Z + jets events		
central value	$Q^2 = m_{W/Z}^2 + \sum_{\text{jets}} p_T^2$	10 GeV
variations	$(0.5 \cdot Q)^2, (2 \cdot Q)^2$	5 GeV and 20 GeV

Table 4.1: Central values and variations used for systematic uncertainties of factorisation/renormalisation scale (Q^2) and matrix element – parton shower (ME-PS) matching in MadGraph MC generator.

- particles traversing all the layers of the detector and the interaction processes;
- multiple interactions per beam crossing (pile-up) and event overlay effects;
- digital readout of the detector (digitisation).

The data stream generated with this process is very similar to the output of the real detector, although naturally in addition to reconstructed objects, generated events also include collections of generated objects allowing to access the MC truth, which is essential for any particle physics analysis.

4.1.2.4 Monte Carlo samples

The nominal signal $t\bar{t}$ MC sample used in this analysis had been generated using MadGraph generator with a top quark mass of $m_t = 172.5$ GeV. To calibrate the mass extraction technique as it will be explained in Section 4.5, 8 additional $t\bar{t}$ samples were used with different top quark masses ranged between 161.5 GeV and 184.5 GeV. The W + jets and Z + jets background processes (see Section 1.2.3) were also generated with MadGraph, whereas single top background was generated with POWHEG generator. Full list of signal $t\bar{t}$ and background MC samples with values of cross section, numbers of generated events and corresponding integrated luminosities is shown in Table 4.2.

The QCD background samples are available in two different sets, one of which is pre-filtered to contain jets with high electromagnetic content (e/γ enriched), and the second one is enriched with electrons coming from decays of heavy flavour quarks ($b/c \to e\nu$). This pre-selection is performed on generator level to ensure reasonable statistics after the

$t\bar{t}$ selection imposed in top quark analyses with electrons in the final state. Both sets of QCD samples are generated with PYTHIA generator in different exclusive bins of $\hat{p}_T$ (i.e. transverse momentum of the outgoing partons defined in the rest frame of the hard process). The γ+jets MC samples were produced with MadGraph in different bins of H_T at generator level, which denotes the sum of transverse momenta of all high-p_T objects. Table 4.3 gives a list of QCD and γ+jets samples with values of production cross sections, pre-selection efficiencies, numbers of generated events and corresponding integrated luminosities.

Process	Generator	σ (pb)	# events	$\int \mathcal{L}dt$ (fb^{-1})
$t\bar{t}$ + jets	MadGraph			
$m_t = 161.5$ GeV		157.5	1620072	10.3
$m_t = 163.5$ GeV		157.5	1633197	10.4
$m_t = 166.5$ GeV		157.5	1669034	10.6
$m_t = 169.5$ GeV		157.5	1606570	10.2
$m_t = 172.5$ GeV		157.5	7490162	47.6
$m_t = 175.5$ GeV		157.5	1538301	9.8
$m_t = 178.5$ GeV		157.5	1648519	10.5
$m_t = 181.5$ GeV		157.5	1665350	10.6
$m_t = 184.5$ GeV		157.5	1671859	10.6
W + jets (W $\to l\nu$)	MadGraph	31314	81345381	2.6
Z/$\gamma^* \to l^+l^-$ + jets, $m(ll) > 50$ GeV	MadGraph	3048	36222153	11.9
Single top	POWHEG			
top t-channel		42.6	3814228	89.5
anti-top t-channel		22.0	1944822	88.4
top s-channel		2.72	259971	92.6
anti-top s-channel		1.49	137980	92.6
top tW-channel		5.3	814390	153.7
anti-top tW-channel		5.3	809984	152.8

Table 4.2: Signal and background Monte Carlo samples with cross sections at $\sqrt{s} = 7$ TeV, numbers of generated events and corresponding integrated luminosities.

Process	Generator	σ (pb)	filter efficiency	# events	$\int \mathcal{L} dt$ (fb^{-1})
QCD (e/γ enriched)	PYTHIA				
20 GeV $< \hat{p}_{\mathrm{T}} <$ 30 GeV		2.355×10^8	7.3×10^{-3}	34720808	2.0×10^{-2}
30 GeV $< \hat{p}_{\mathrm{T}} <$ 80 GeV		5.93×10^7	0.059	70375915	2.0×10^{-2}
80 GeV $< \hat{p}_{\mathrm{T}} <$ 170 GeV		9.06×10^5	0.148	8150669	6.1×10^{-2}
QCD (b/c $\rightarrow e\nu$)	PYTHIA				
20 GeV $< \hat{p}_{\mathrm{T}} <$ 30 GeV		2.355×10^8	4.6×10^{-4}	2002588	1.8×10^{-2}
30 GeV $< \hat{p}_{\mathrm{T}} <$ 80 GeV		5.93×10^7	2.34×10^{-3}	2030030	1.5×10^{-2}
80 GeV $< \hat{p}_{\mathrm{T}} <$ 170 GeV		9.06×10^5	0.0104	1082690	0.1
γ + jets	MadGraph				
40 GeV $< H_{\mathrm{T}} <$ 100 GeV		23 620	1	12730863	0.5
100 GeV $< H_{\mathrm{T}} <$ 200 GeV		3476	1	1536287	0.4
$H_{\mathrm{T}} >$ 200 GeV		485	1	9377168	19.3

Table 4.3: QCD multi-jet background and γ + jets MC samples with cross sections at $\sqrt{s} = 7$ TeV, numbers of generated events and corresponding integrated luminosities.

Process	Generator	σ (pb)	# events	$\int \mathcal{L} dt$ (fb^{-1})
$\mathrm{t\bar{t}}$ + jets	MadGraph			
0.5 $\times$ matching threshold		157.5	1607808	10.2
2 $\times$ matching threshold		157.5	4029823	25.5
$0.5 \times Q$		157.5	4004587	25.4
$2 \times Q$		157.5	3696269	23.4
W + jets (W $\rightarrow l\nu$)	MadGraph			
0.5 $\times$ matching threshold		29690	9956679	0.3
2 $\times$ matching threshold		30290	10461655	0.3
$0.5 \times Q$		33300	10092532	0.3
$2 \times Q$		32000	9784907	0.3
Z + jets (Z $\rightarrow ll$)	MadGraph			
0.5 $\times$ matching threshold		2888	1614808	0.6
2 $\times$ matching threshold		2915	1641121	0.6
$0.5 \times Q$		3312	1658855	0.5
$2 \times Q$		2954	1592742	0.5

Table 4.4: Systematic MC samples with cross sections at $\sqrt{s} = 7$ TeV, numbers of generated events and corresponding integrated luminosities. Factorisation scale Q and matching threshold systematic uncertainties (see Section 4.1.2.2) are estimated with variations of $\mathrm{t\bar{t}}$ + jets, W + jets and Z + jets samples.

4.2 Event Selection

Event selection applied in this analysis essentially follows the top quark group reference selection that is recommended as a standard CMS top analyses selection optimised for SM $t\bar{t}$ production. It is designed to select the semi-leptonic signature of a $t\bar{t}$ decay with an isolated lepton, four jets (including two jets from b-quarks) and a neutrino in the form of missing transverse energy. All objects are reconstructed using the particle flow method as described in Section 2.4.

As this analysis focuses on the electron channel, the event selection proceeds as follows:

1. preselection;
2. trigger;
3. electron candidate selection;
4. dilepton veto;
5. conversion veto;
6. muon veto;
7. jet selection;
8. b-tagging.

Preselection

Preselection is performed in order to reduce the number of events stored for local analysis. All events are required to have at least one electron with a p_{T} above 30 GeV or at least one muon with a p_{T} above 20 GeV. Additional event cleaning is performed at this stage to decrease the number of events with substantial detector noise. A small fraction of events with anomalous HCAL noise is removed by HCAL noise filter. Furthermore, beam scraping is mitigated by the appropriate filter requiring at least 25 % tracks with high purity if there are more than 10 tracks in the event. Finally, preselection stage includes the requirement of at least one good primary vertex in the event which has to be located within 24 cm in z-direction from the centre of CMS and within a radial distance of 2 cm. The primary vertex is also required to have at least four degrees of freedom which is determined from the vertex fit and is strongly correlated with the number of tracks compatible with the primary interaction region [40].

Trigger

The trigger requirement is imposed in a way it has been described in Chapter 3. Essentially, all the events are required to fire the electron-plus-three-jets trigger in its available version. A full list of triggers can be found in the appendix A.1.

Electron candidate selection

Exactly one electron candidate satisfying all following criteria is required to be present in the event. As soon as the electron-plus-jets trigger has the lepton p_{T} threshold of 25 GeV, the electron candidate transverse momentum is required to be above 30 GeV (in order to be in the trigger efficiency plateau region). Furthermore, the electron has to be within the tracker region of $|\eta| < 2.5$ excluding the ECAL barrel-endcap transition regions of $1.4442 < |\eta| < 1.566$. The CiC electron ID (see Section 2.4.1) with the tightest working point ("HyperTight") is used for electron identification purposes. This working point provides the misidentification rate of less than 1 %, and overall identification efficiency of 75 % [17].

In order to only select electrons originating at the interaction vertex, the x-y-distance (d_{xy}) between the electron track and the interaction point (2D impact parameter) is required to be less than 0.02 cm. Contribution of electrons inside the jets from QCD background events, as well as fake electrons is reduced by imposing the PF-based relative isolation cut of reliso < 0.1. A ΔR-cone with size of 0.3 is used in calculation of the reliso variable (Section 2.4.1.2).

Dilepton veto

A veto requirement on a second electron candidate (or dilepton veto) is used to reject events with any additional loose electrons. Namely, event is rejected if there is a second electron satisfying looser electron candidate criteria: a lower E_{T} threshold of 15 GeV and a looser cut on reliso (< 0.2).

Additionally, in order to reduce Z + jets background, the following veto is applied. If the event contains a second electron with $E_{\mathrm{T}} > 30$ GeV and reliso < 1.0 that forms such an invariant mass with the signal electron candidate that it is close to the Z mass peak, it is also rejected. The invariant mass window around the Z peak is chosen to be 76 GeV $< m_{ee} < 106$ GeV.

Conversion veto

As it was mentioned in Section 2.4.1.3, two methods are used in this analysis to remove the electrons coming from photon conversions: missing pixel layers method and partner track matching method. If missing pixel layer hits or a matching partner track are present in the event, it is rejected.

Muon veto

Other $\mathrm{t\bar{t}}$ channels, including muon and dilepton decay modes, can contaminate events passing the electron plus jets selection. To reduce this contamination, events containing a global (see Section 2.4.2) isolated muon are rejected. The muon is required to have a p_{T} above 10 GeV, $|\eta| < 2.5$ and reliso < 0.2 with a ΔR cone of 0.4.

Jet selection and b-tagging

Final steps in the selection help to further reduce the background by imposing constraints on the number of jets. This is particularly effective to mitigate the W + jets contamination, as the number of W + jets events decreases exponentially with increasing number of jets. In this analysis, at least four jets with $p_{\mathrm{T}} > 30$ GeV and $|\eta| < 2.4$ are required to be present in the event. These jets have to pass the loose PF jet ID (see Section 2.4.3). Finally, the CSV b-tagging algorithm with medium working point (Section 2.4.3.3) is used to identify the two b-quarks from the $\mathrm{t\bar{t}}$ decay.

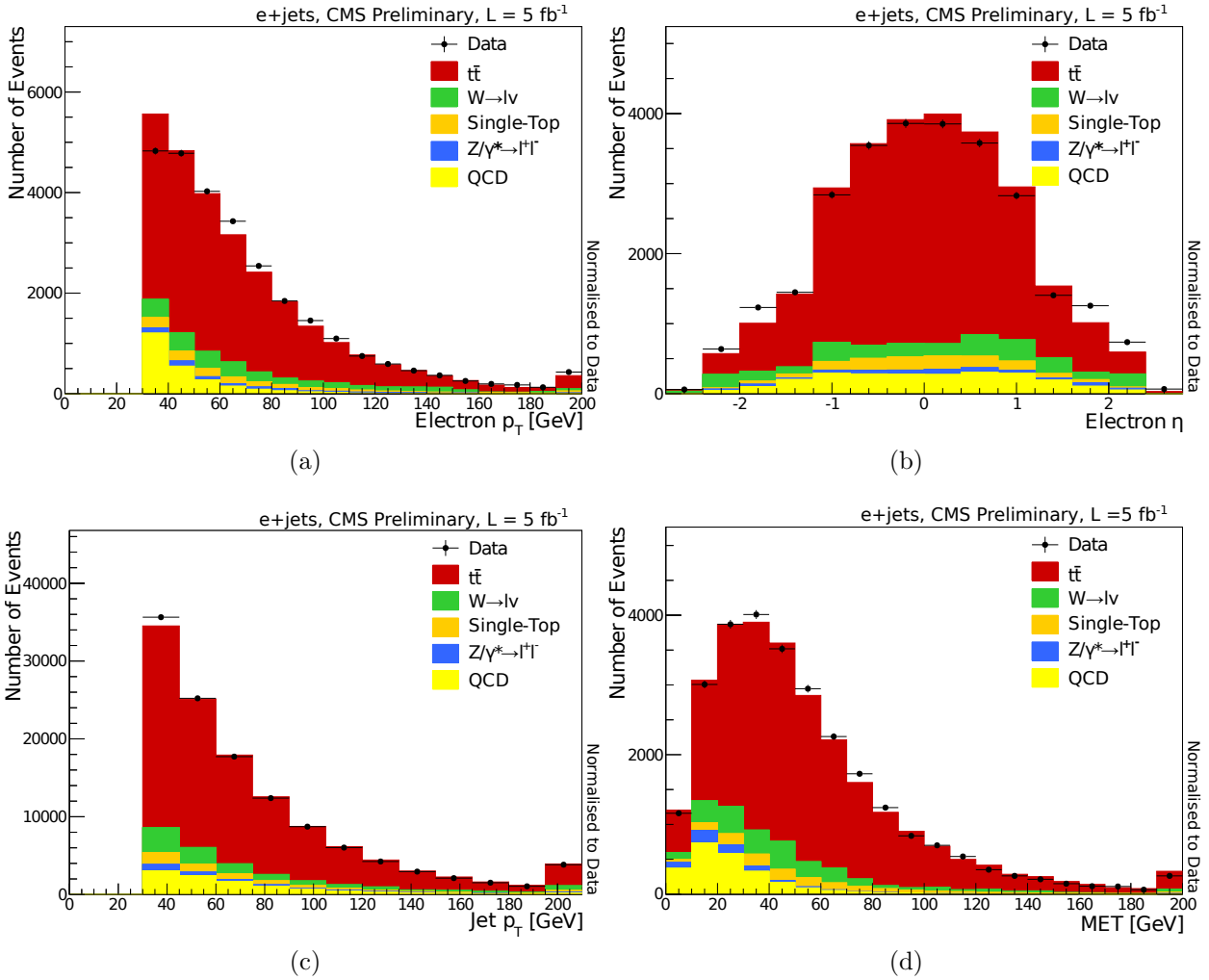


Figure 4.1: Kinematic variable distributions after all selection cuts. (a) electron p_{T}, (b) electron η, (c) jet p_{T} for all jets passing the selection, (d) $E_{\mathrm{T}}^{\mathrm{miss}}$.

4.3 Kinematic Fit

In order to precisely measure the top quark mass, objects in the final state need to be associated with originating partons of the top quark pair decay. In this analysis, a kinematic fit is employed to fully reconstruct the event kinematics under the $t\bar{t}$ hypothesis, thus improving the resolution of measured quantities by exploiting the knowledge of decay process.

The HitFit fitting package [41] used in this work originates from DØ collaboration. Based on the SQUAW algorithm [42] developed in Lawrence Berkeley National Laboratory, the original HitFit code was written by Scott Stuart Snyder in Fortran for Run I of the Tevatron. It was successfully used in the first direct measurement of the top quark mass by DØ for lepton plus jets $t\bar{t}$ events [43]. The package was then migrated to C++ for Run II of the Tevatron, and exploited in a series of $t\bar{t}$ analyses including the updated top quark mass measurement using the ideogram method [44], first measurement of the forward-backward charge asymmetry in $t\bar{t}$ production [45] and some others.

One of the key features of the HitFit package is its independence of specific experiments and detector geometries. The only external dependence is on CLHEP [46], a C++ library that provides utility classes for linear algebra, geometry, vector arithmetic, etc., which was specifically developed for high energy physics simulation and analysis software and is notably used by Geant4 package. In 2011 HitFit was ported to CMSSW and later on successfully used in a number of CMS top quark analyses, including the top mass measurement in lepton plus jets channel [29].

HitFit is designed to fit lepton plus jets events under the $t\bar{t}$ hypothesis. Specifically, it strives to constrain the event to a hypothesis of production of two heavy particles, each decaying into a W boson and b quark. According to the semi-leptonic signature, one of the W boson decays into an electron-neutrino pair, while the other decays into a light quark-antiquark pair.

The input to the fit includes the four-momenta of the lepton, four leading jets and missing transverse energy, as well as their respective resolutions that were calculated using Monte Carlo samples. Since only four leading jets are used, there are $4! = 24$ ways to associate these four reconstructed jets with the partons from the $t\bar{t}$ decay. The number of permutations is reduced to $4!/2 = 12$ since the light jets from the hadronic W decay are interchangeable. However, for each hypothesis the z component of neutrino four-momentum needs to be determined. This is done by requiring the two heavy particles to have equal mass ($m_{t} = m_{\bar{t}}$), which yields to a quadratic equation which has either two real or two complex solutions that share the real part. If case of the two real solutions the fit is performed for each of them. If

there are two complex solutions, the fit is performed once for the common real part of them, but the fit result is included in the event content twice. Therefore, each hypothesis has two possible values of the neutrino four-momentum, which doubles the number of permutations back to $4! = 24$.

The kinematic fit is performed by minimising the χ^2 function defined as

$$\chi^2 = (\mathbf{x} - \mathbf{x}^m)^T \mathbf{G} (\mathbf{x} - \mathbf{x}^m) \tag{4.1}$$

where $\mathbf{x}^m$ is the vector or measured observables, $\mathbf{x}$ is the vector of fitted variables and $\mathbf{G}$ is the inverse of the error matrix given by the resolutions of the observables.

The following kinematic constraints are imposed:

1. Reconstructed top quarks from both hadronic and leptonic legs are required to have equal masses:
$$m(\mathrm{t} \to \ell\nu\mathrm{b}) = m(\bar{\mathrm{t}} \to \mathrm{q}\bar{\mathrm{q}}\bar{\mathrm{b}}), \tag{4.2}$$
2. The W boson masses reconstructed in both hadronic and leptonic legs are fixed:
$$m(\ell\nu) = m(\mathrm{q}\bar{\mathrm{q}}) = m_{\mathrm{W}} = 80.4 \text{ GeV}. \tag{4.3}$$

Apart from exploiting the knowledge of the W mass, the b quark mass of 4.7 GeV is also used. Light quarks and leptons are considered to be massless.

The full description of the fitting algorithm can be found in the appendix of the HitFit author's PhD thesis [47]. Essentially, since the constraints are non-linear, an iterative technique is used. Starting with the measured observables, the constraint equations are linearised by expansion in a power series around the starting point. The minimisation is then solved with these linearised constraints, providing the new starting value for the next minimisation step which is repeated until the χ^2 converges or a maximum number of iterations is reached. Events are rejected if there is no permutation resulting in a converging fit. Some kinematic variables for permutations giving the best (smallest) χ^2 values are shown in Figure 4.2.

Typically, an event has a few permutations resulting in a converged fit. The simplest approach of picking a permutation with the best χ^2 value is disfavoured, since the Monte Carlo studies have shown that the correct permutation does not necessarily have the best χ^2. This value can change drastically with relatively small variations of the input parameters, as different permutations may obtain the smallest value of χ^2. On the other hand, some solutions yield very large χ^2 value, which typically happens for badly reconstructed or background events. Therefore, a loose cut of $\chi^2 = 20$ is applied for all permutations on

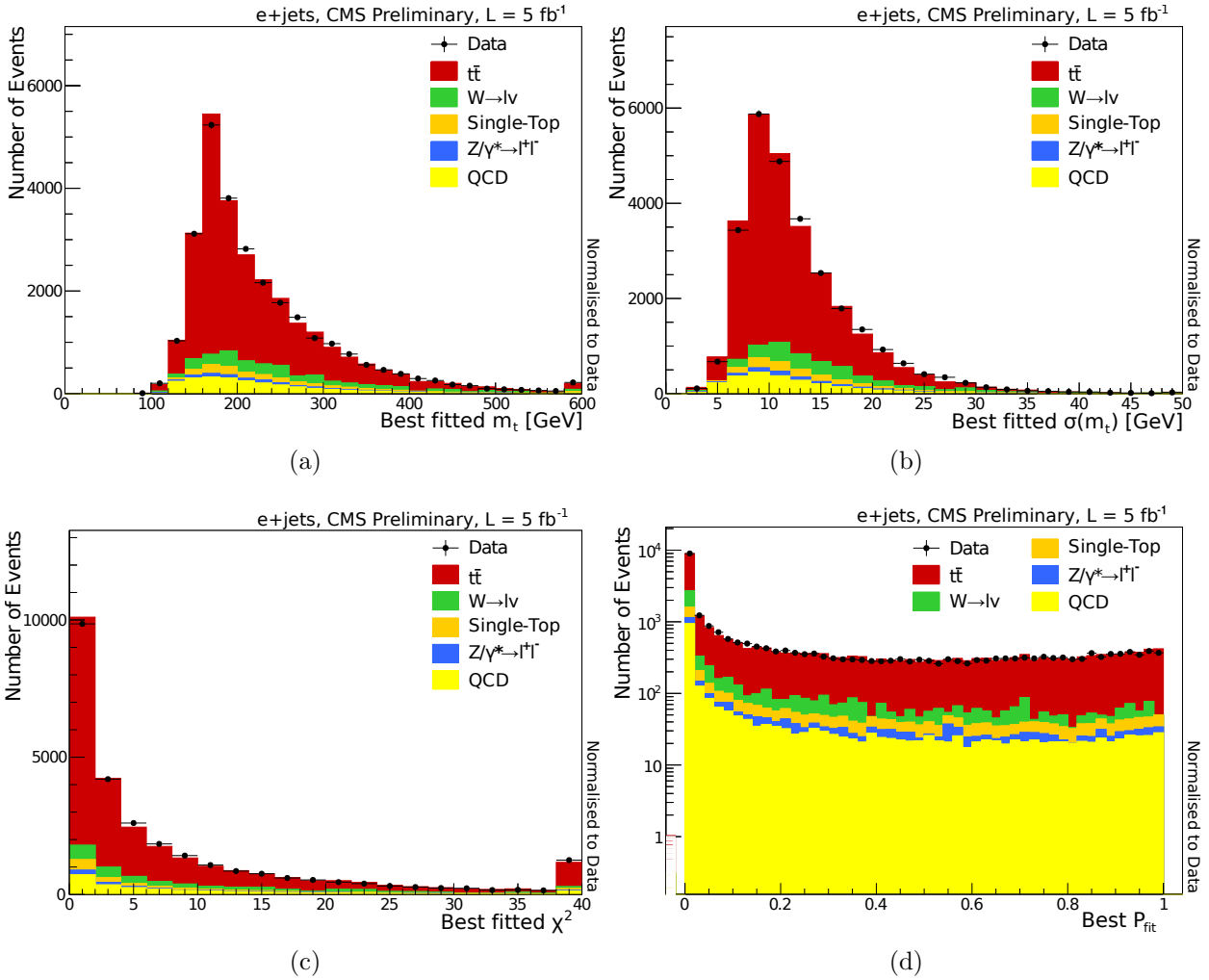


Figure 4.2: Distribution of (a) fitted top quark mass, (b) its standard deviation, (c) χ^2 and (d) fit probability for the best fitted hypotheses.

the output of the converged fit. This is the final selection criteria on top of the reference selection mentioned in Section 4.2.

The compatibility of a permutation with the $t\bar{t}$ hypothesis is quantified by the following fit probability:

$$P_{\text{fit}} = \exp(-\frac{1}{2}\chi^2). \tag{4.4}$$

This probability is used in the next step of the analysis, referred to as the ideogram method.

4.4 The Ideogram Method

The ideogram method is a powerful mass extraction technique. Its main idea lies in finding the likelihood distribution of the particle mass given the particular data sample, which is performed by evaluating the likelihood of measured observables given a generated particle mass for each event in the sample. The signal likelihood is evaluated from the theoretically expected Breit-Wigner distribution convoluted with the Gaussian experimental resolution on event-by-event basis. The whole procedure is calibrated using Monte Carlo simulation of a series of particle masses in the expected region.

One of the first applications of the ideogram method tracks back to the W boson mass measurement by the DELPHI Collaboration at CERN LEP collider [48, 49]. It was later on used in a series of top quark mass measurements, including the DØ measurement in lepton plus jets channel [44] and CDF measurement in all-hadronic channel [50]. The main CMS top quark mass measurement in lepton plus jets channel at 7 TeV [29] which is cross-checked by this analysis is also based on the ideogram method, however, with *in situ* JES (jet energy scale) measurement in a joint likelihood fit. As JES is a dominant systematics in this analysis, this approach proved to be more precise, but let us not neglect the importance of a cross-check.

4.4.1 Event likelihood

In the basis of the ideogram method lies the event likelihood, also referred to as the ideogram. It is calculated for each event in the sample as a function of the hypothesised top quark mass m_{t} in the following way:

$$\mathcal{L}_{\text{event}}\left(\mathbf{x}|m_{\text{t}}, f_{\text{t}\bar{\text{t}}}\right) = f_{\text{t}\bar{\text{t}}} P_{\text{t}\bar{\text{t}}}\left(\mathbf{x}|m_{\text{t}}\right) + \left(1 - f_{\text{t}\bar{\text{t}}}\right) P_{\text{bkg}}\left(\mathbf{x}\right). \tag{4.5}$$

Here $\mathbf{x}$ is the set of event observables with the top mass information from the kinematic fit, $f_{\text{t}\bar{\text{t}}}$ is the fraction of $t\bar{t}$ events in the data sample, and $P_{\text{t}\bar{\text{t}}}$ and P_{bkg} are probability densities for signal and background events, respectively. Essentially, by construction of the

likelihood, the first term corresponds to the probability for the event to be signal, whilst the second term is its probability to be background. The kinematic information $\mathbf{x}$ includes the fitted mass m_i, the estimated standard deviation for each fitted mass $\sigma(m_i)$ and the goodness-of-fit χ^2_i for all permutations, i.e. all possible jet-parton assignments and neutrino solutions surviving the selection criteria on the output of the fit.

Jet-parton assignment weights

The signal and background probabilities forming the event likelihood are calculated as a weighted sum over jet-parton assignments and neutrino solutions coming from the kinematic fit. Each permutation is weighted by the probability that it is correct, which is essentially the fit probability P_{fit} introduced in Equation 4.4. However, this particular implementation of the ideogram method also exploits the b-tagging information, thus the fit probability is multiplied by an additional term p_{btags}, which is the probability that the particular jet-parton assignment is compatible with the observed b-tags:

$$p_{\text{btags}} = \prod_{j=1}^{4} p^j \tag{4.6}$$

Here the index j runs over all four jets in the kinematic fit, and the probability p^j can be either ε_l, $(1-\varepsilon_l)$, ε_b, or $(1-\varepsilon_b)$, depending on the flavour of the jet as derived by the fit (light jet or b-jet), and whether the jet is b-tagged or not. As a recap from Section 2.4.3.3, for the medium working point of CSV b-tagging algorithm used in this analysis, b-tagging efficiency is approximately ~ 70 %, whilst the mis-tag rate (or $(1-\varepsilon_l)$) is approximately 1 %.

Therefore, for each such permutation the relative weight is w_i is calculated as:

$$w_i = p_{\text{fit}} \cdot p_{\text{btags}} = \exp(-\frac{1}{2}\chi^2) \cdot \prod_{j=1}^{4} p^j \tag{4.7}$$

These weights are used in the calculation of signal and background probability densities, described hereafter.

Signal probability

The $t\bar{t}$ signal probability in the event likelihood shown in Equation 4.5 is calculated as a weighted sum over all permutations:

$$P_{t\bar{t}}\left(\mathbf{x}|m_t\right) = \sum_{i}^{24} w_i \left(f_{\mathrm{cp}} \cdot \int_{m_{\mathrm{min}}}^{m_{\mathrm{max}}} \mathrm{G}\left(m'|m_i,\sigma_i\right) \mathrm{RBW}\left(m'|m_t,\Gamma_t\right) \mathrm{d}m' + (1-f_{\mathrm{cp}})\mathrm{WP}(m_i|m_t) \right) \tag{4.8}$$

Here f_{cp} is the fraction of events in which the maximum weight is assigned to the correct permutation, estimated from simulated $t\bar{t}$ events. For each permutation, the first term expresses the probability that it has a correct jet-parton assignment (correct permutation). It is calculated as a convolution of a relativistic Breit-Wigner distribution $\mathrm{RBW}(m'|m_t,\Gamma_t)$ and a Gaussian $\mathrm{G}(m'|m_i,\sigma_i)$.

The relativistic Breit-Wigner distribution describes the theoretical top quark mass distribution for a given hypothesised top quark mass value m_t with a top quark width Γ_t set to 2 GeV, according to the latest experimental results [51]. As suggested by Standard Model electroweak corrections and $s-$dependence of the phase space [52], the following form of RBW is used:

$$\mathrm{RBW}(m'|m_t,\Gamma_t) \propto \frac{m'^2}{\left(m'^2 - {m_t}^2\right)^2 + m'^4 \frac{\Gamma_t}{{m_t}^2}} \tag{4.9}$$

The Gaussian in the convolution represents the detector resolution. It is centred at the fitted mass value m_i with a width of the fitted mass uncertainty σ_i as estimated by the fit:

$$\mathrm{G}(m'|m_i,\sigma_i) = \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left(\frac{(m'-m_i)^2}{2\sigma_i^2}\right) \tag{4.10}$$

The resulting probability density for $t\bar{t}$ events with correct permutations is shown in Figure 4.3. The Monte Carlo sample with top mass of 166.5 GeV was used, for illustration.

The second term shows the probability of a wrong permutation in the signal event. This probability density, denoted $\mathrm{WP}(m_i|m_t)$, is calculated by fitting the Crystal Ball analytical function to the fitted top quark mass distribution for permutations with wrong jet-parton assignment, estimated purely from Monte Carlo $t\bar{t}$ sample whilst taking the permutation weights (Equation 4.7) into account. The wrong permutations fall in two categories: the first one corresponds to events where the four leading jets come from the $t\bar{t}$ decay, but are not assigned to the partons correctly; the second one are events with one ore more of the four leading jets coming from the soft gluon radiation (ISR or FSR). It was found that the fitted top quark mass distributions have similar shapes for both categories, therefore they

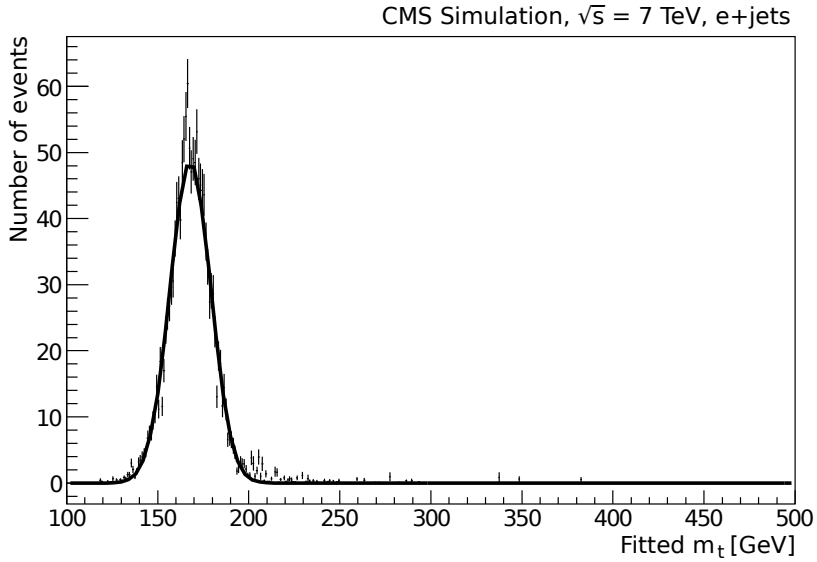


Figure 4.3: The fitted top quark mass distribution for $t\bar{t}$ events with correct permutations (MC truth, $m_t = 166.5$ GeV).

were combined for the purpose of fitting. The Crystal Ball function has the form of:

$$\mathrm{CB}(x;\mu,\sigma,\alpha,n) = N \times \begin{cases} \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{(x-\mu)^2}{2\sigma^2}) & \text{if } \frac{x-\mu}{\sigma} > -\alpha, \\ \frac{1}{\sqrt{2\pi}\sigma}(\frac{n}{|\alpha|})^n \exp(-\frac{|\alpha|^2}{2})(\frac{n}{|\alpha|} - |\alpha| - \frac{x-\mu}{\sigma})^{-n} & \text{if } \frac{x-\mu}{\sigma} \leq -\alpha. \end{cases} \quad (4.11)$$

The exponential order of $n = 5$ was used, and parameters μ, σ, and α were parametrised as functions of the top quark mass. The fitted distributions of wrong permutations spectra for three different top masses (166.5 GeV, 172.5 GeV and 178.5 GeV) are shown in Figure 4.4.

Background probability

The shape of the background probability density in Equation 4.5 is assumed not to depend on the top quark mass, and is estimated from the background Monte Carlo samples. The procedure is done in a similar way to the wrong permutations probability, and the jet-parton assignment weights (Equation 4.7) are also taken into account. Since the dominant background is W + jets, it was used exclusively for extracting the fitted top quark mass distribution by running the kinematic fit. The contribution from single top, Z+jets and QCD was expected to be small, and it was confirmed that their probability density distributions have similar shapes to that of the W + jets sample. Moreover, the Monte Carlo calibration

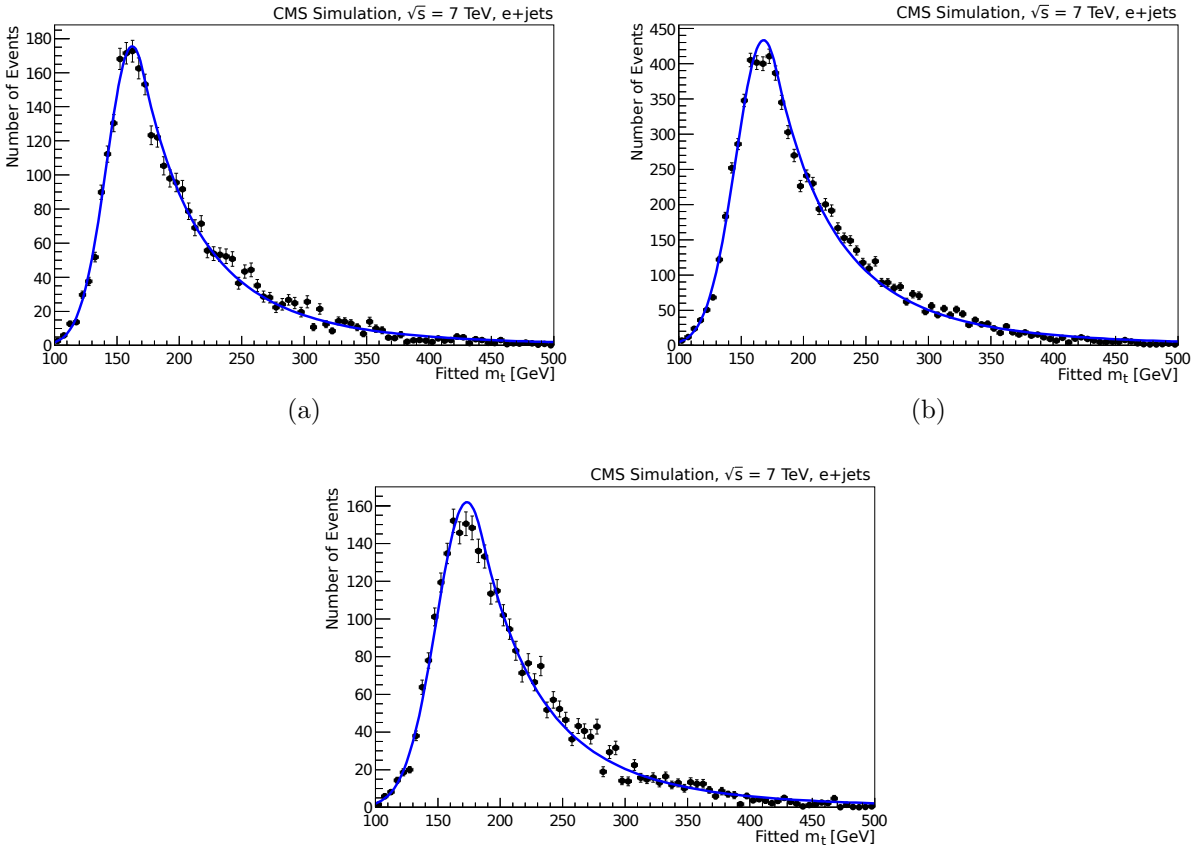


Figure 4.4: The fitted top quark mass distribution for $t\bar{t}$ events with wrong permutations (MC truth) for (a) $m_t = 166.5$ GeV, (b) $m_t = 172.5$ GeV, (c) $m_t = 178.5$ GeV.

described later on in Section 4.5 takes all mentioned backgrounds sources into account, therefore any possible bias due to this approximation is corrected for.

The fitted top quark mass spectrum for background events as estimated from W + jets MC sample is shown in Figure 4.5. This distribution is described by the Landau function, fitted by the ROOT package [53].

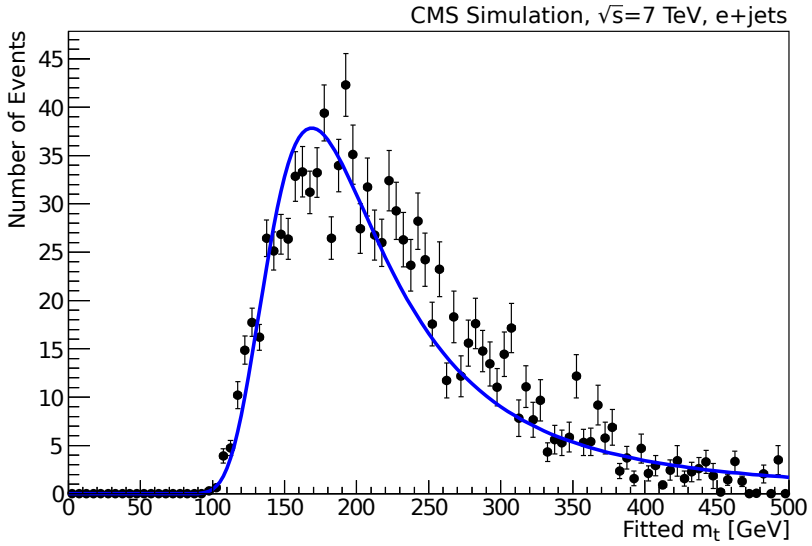


Figure 4.5: The fitted top quark mass distribution for $t\bar{t}$ events in W + jets Monte Carlo sample.

4.4.2 Combined likelihood fit

Finally, for the extraction of the top quark mass the total likelihood for the entire data sample is calculated as the product of all the event likelihoods:

$$\mathcal{L}_{\text{sample}} = \prod_{\text{all events}} \mathcal{L}_{\text{event}}\left(\mathbf{x}|m_{\text{t}}\right) \tag{4.12}$$

To extract the most probable top quark mass, this likelihood needs to be maximised. In practice, it is simplified by performing the minimisation of the summed negative log-likelihoods:

$$-2\log\mathcal{L}_{\text{sample}} = -2\sum_{\text{all events}} \log\mathcal{L}_{\text{event}}\left(\mathbf{x}|m_{\text{t}}\right) \tag{4.13}$$

The result with the full dataset is shown in Figure 4.6. The top quark mass and its statistical uncertainty are calculated by using the parabolic interpolation with three lowest

points in the log-likelihood curve. However, this result can not be regarded as the final measurement, as it does not take into account the possible mass bias of the ideogram method. The mass calibration technique is described in the next section.

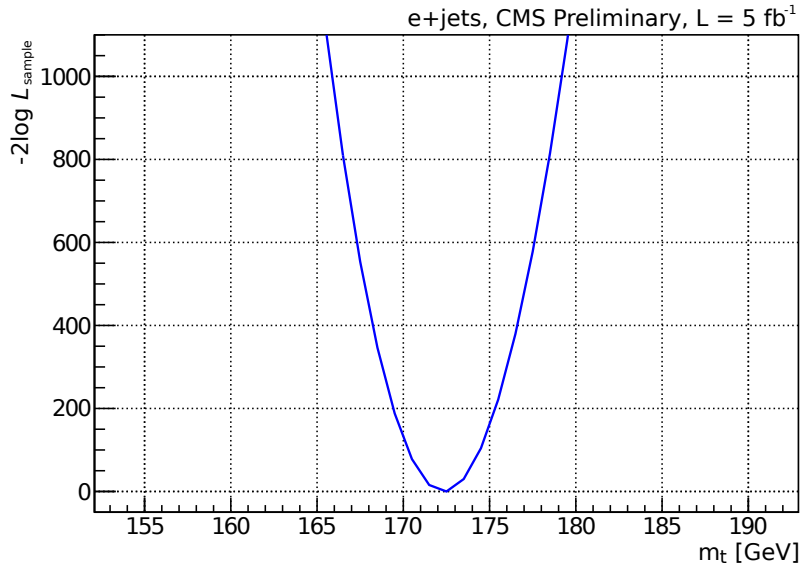


Figure 4.6: Log-likelihood curve as a function of top quark mass for e+jets dataset before calibration.

4.5 Mass calibration

By construction of the ideogram method, it has approximations which can lead to possible bias in the top quark mass measurement. To account for this bias, a calibration is performed by running the measurement on Monte Carlo simulation with known true top quark mass. Pseudo-experiments are used to measure the bias in the fitted mass, as well as to calibrate the statistical uncertainty of the measurement.

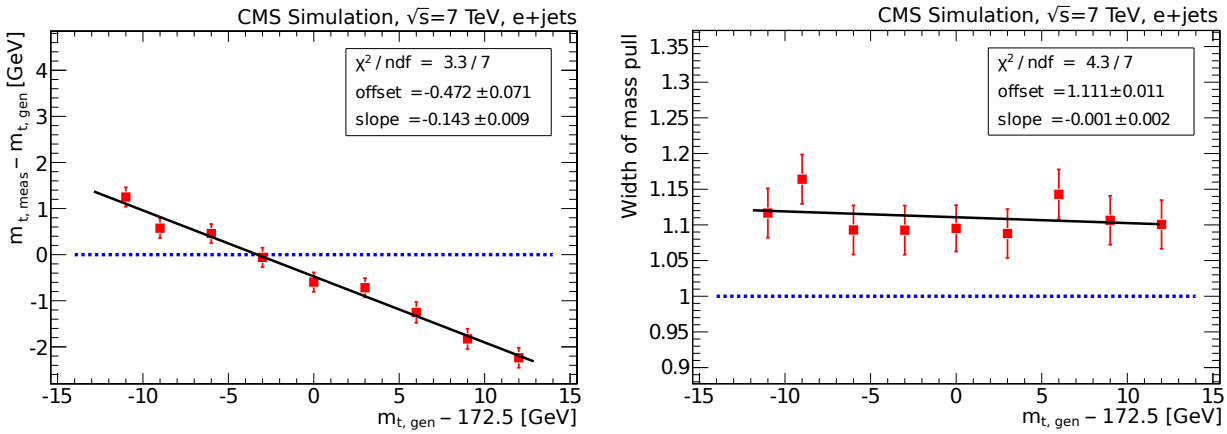
For each of the nine generated $t\bar{t}$ samples with mass points between 161.5 GeV and 184.5 GeV, 5000 pseudo-experiments were performed with events picked from Monte Carlo samples representing different sources of events, so that the numbers of events follow a Poisson distribution centred around the corresponding numbers observed in real data. The observed value of 23 727 electron plus jets events with an average signal fraction of 76 % was used in each pseudo-experiment. The background was estimated by using events from W + jets MC sample, whereas the impact of other background sources was accounted for separately as a systematic uncertainty (Section 4.6). For every pseudo-experiment all the event likelihoods are calculated, the total log-likelihood is formed and the top quark mass

(m_i) with its statistical uncertainty (σ_i) are extracted in the same way as in the real measurement described above. These results are used to derive the bias and pull distributions, defined as follows:

$$\text{bias} = m_{\text{t}}^{\text{meas}} - m_{\text{t}}^{\text{gen}} \quad \text{and} \quad \text{pull}_i = \frac{m_i - m_{\text{t}}^{\text{meas}}}{\sigma_i}, \tag{4.14}$$

where $m_{\text{t}}^{\text{meas}}$ is the mean top quark mass over all pseudo-experiments, for a given generated mass point. The width of the pull is defined as the standard deviation (σ) of a Gaussian fitted to the pull distribution.

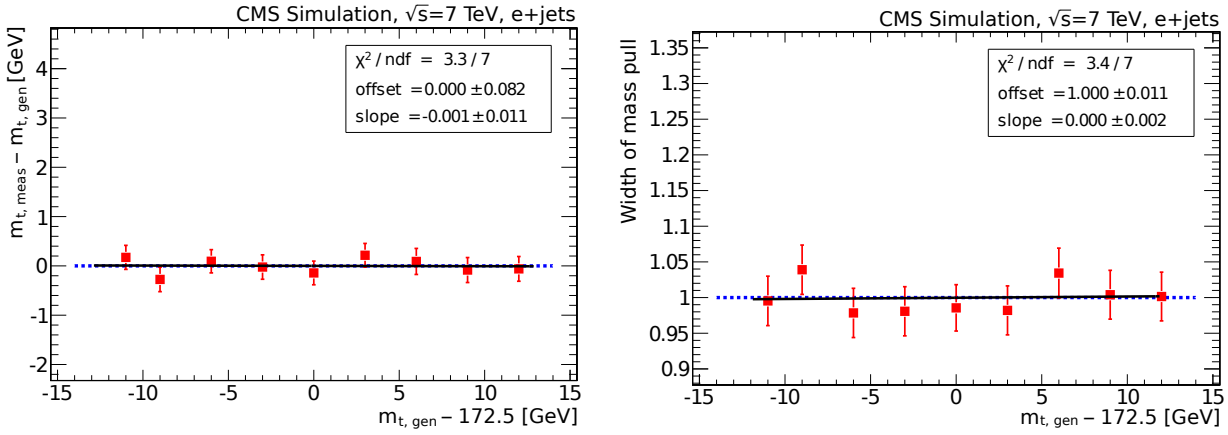
The mass bias and the width of the pull as a function of the generated top quark mass are shown in Figure 4.7. Clearly, using the ideogram method out of the box reveals a significant bias which needs to be corrected for. It can also be seen that the width of the pull distribution is approximately 11 % larger than unity, suggesting that the ideogram method underestimates the statistical uncertainty.



(a) (b)

Figure 4.7: Average mass bias (a) and width of the pull (b) as a function of true top quark mass.

Due to the apparent bias, the fitted linear calibration curves in both distributions are used to correct the final measurement of the top quark mass. As a closure test, the mass bias and pull width as a function of true top quark mass are calculated after applying the full calibration. The distributions are shown in Figure 4.8. Evidently, the closure tests confirm that the top quark mass and its statistical uncertainty are unbiased after calibration.



(a) (b)

Figure 4.8: Closure test, showing the average mass bias (a) and width of the pull (b) as a function of true top quark mass after calibration.

4.6 Systematic Uncertainties

Like any measurement, the top quark mass measurement is prone to various systematic errors. In attempt to evaluate them, several sources of systematic effects are considered in this analysis. These include theoretical uncertainties on Monte Carlo modelling, assumptions made within the ideogram method, as well as imperfect understanding of detector effects and reconstruction algorithms.

Systematic uncertainties are calculated by running the analysis with tuned MC samples where systematic effects are varied by ± 1 standard deviation, and comparing the result with the nominal one. The obtained difference in fitted Δm_{t} is regarded as an uncertainty due to a given systematic effect. The total systematic error is calculated as a sum of all components in quadrature, under assumption that all the effects are uncorrelated. Summary of all systematic uncertainties is given in Table 4.5, whereas all corresponding sources are described below.

Fit calibration statistics. To estimate the systematic effect of the calibration with pseudo-experiments described in Section 4.5, the statistical uncertainty of this calibration is propagated to the final measurement by regarding it as a systematic error. This uncertainty was evaluated with the central $\mathrm{t\bar{t}}$ MC sample with the top quark mass of 172.5 GeV.

b-tagging. The nominal ("medium") working point of the CSVM algorithm is varied in order to reflect the uncertainty of the b-tagging efficiency of ~ 4 % [21]. The average effect

Source of the systematic uncertainty	Δm_{t} (GeV)
Fit calibration statistics	0.12
b-tagging	0.10
Jet Energy Scale (overall data/MC)	1.21
Factorisation scale	1.38
ME-PS matching threshold	0.93
Colour Reconnection	0.03
Underlying event	0.08
Non-$\mathrm{t\bar{t}}$ background	0.61
Pile up	0.22
PDF	0.15
Total	2.17

Table 4.5: Overview of the systematic uncertainties and their effects on the top quark mass measurement.

of this change on the mass measurement (approximately 0.1 GeV) is quoted as a systematic uncertainty.

Jet energy scale. Since the top quark mass measurement is based on knowledge about at least four energetic jets in semi-leptonic $\mathrm{t\bar{t}}$ decay, it is very sensitive to any uncertainties in jet energies. Hence, the jet energy scale is a dominant systematic in this analysis, particularly depending on our understanding of the CMS detector. To quantify this systematic effect, the jet energies are varied within $\pm 1\sigma$ uncertainty of the over jet energy correction, obtained from a designated jet energy calibration and resolution study [54]. This p_{T}- and η-dependent uncertainty incorporates various contributions, including an offset due to noise and pile-up, absolute and relative discrepancies between data and MC, flavour-dependent corrections, etc. After applying asymmetric $\pm 1\sigma$ uncertainties, an average systematic effect of 1.21 GeV on the fitted top mass was observed.

Factorisation scale. As mentioned in Section 4.1.2.2, the factorisation scale is varied by factors of 0.5 and 2 with respect to the nominal value used in the central generated samples. The procedure is performed for $\mathrm{t\bar{t}}$ + jets, W + jets and Z + jets Monte Carlo samples independently as the factorisation scales are considered uncorrelated. To combine the systematic uncertainty, the average absolute size of the up and down effects are added in quadrature. Due to the lack of statistics available in variations of W + jets and Z + jets samples, this systematic uncertainty is one of the largest in the top quark mass measurement.

PDF variation eigenvector	Effect on the top quark mass (MeV)			
	up	down	(|up|+|down|)/2	max(|up|, |down|)
1	18.4	-18.8	18.6	18.8
2	-2.9	3.1	3.0	3.1
3	2.3	-2.4	2.4	2.4
4	5.5	-5.9	5.7	5.9
5	22.1	-23.0	22.5	23.0
6	34.2	-31.7	33.0	34.2
7	-6.7	7.6	7.2	7.6
8	2.9	-3.0	2.9	3.0
9	-11.2	10.5	10.8	11.2
10	4.5	-5.1	4.8	5.1
11	77.0	-66.7	71.9	77.0
12	19.7	-14.5	17.1	19.7
13	-23.6	28.4	26.0	28.4
14	-2.6	2.7	2.7	2.7
15	1.7	-6.1	3.9	6.1
16	61.3	-43.9	52.6	61.3
17	15.9	-14.8	15.3	15.9
18	-2.9	-7.3	5.1	7.3
19	-2.5	4.0	3.3	4.0
20	-3.6	-1.0	2.3	3.6
21	-4.5	-2.3	3.4	4.5
22	-17.3	-18.3	17.8	18.3
Total unc. (MeV)	116.1	101.1	106.6	118.1

Table 4.6: Estimated systematic effect of PDF variations on the top quark mass meaurement.

ME-PS matching threshold. The systematic effect of choice of ME-PS threshold, also mentioned in Section 4.1.2.2, is estimated by varying the default value of 20 GeV by factors of 0.5 and 2. Similarly to factorisation scale systematic uncertainty, the matching thresholds are varied independently for $t\bar{t}$ + jets, W + jets and Z + jets Monte Carlo samples and then added in quadrature. The same issue with the lack of statistics appears here, although to a lesser extent.

Colour reconnection. Different modelling of colour reconnections [55] result in a systematic uncertainty that is estimated by varying the underlying event tune in simulation. The observed shifts in top quark mass are quoted as a systematic uncertainty.

Underlying event. Non-perturbative QCD effects are taken into account by tuning PYTHIA to measurements of the underlying event [**Chatrchyan:2011id**]. The uncertainties are estimated by comparing in simulation two tunes with increased and decreased underlying event activity to a central tune (the Perugia 2011 tune to the Perugia 2011 mpiHi and Perugia 2011 Tevatron tunes described in [**Skands:2010ak**]).

The individual variarions are shown in Table **??**. The largest difference in the mass offset, 0.08 GeV, is quoted as systematic uncertainty.

Background modelling. Thanks to the use of b-tagging information in the likelihood, the effect of most non-$t\bar{t}$ backgrounds is reduced. To study the remaining uncertainty due to the exact sample composition, we did various checks. Changing the signal fraction in the pseudo experiments from 0.8 to 0.85 or 0.75 has a small effect on the calibration curve: the average variations is 0.28 GeV.

The difference between the calibration using only WJets to model the backgrounds and using a precise mix of background processes as listed in Table **??** is 0.27 GeV. Completely removing QCD from the mix yields a shift of 0.152 GeV. When removing 30% of the other components one-by-one from the mix, the effects are 0.08 GeV for WJets, 0.012 GeV for ZJets and 0.009 GeV for Single Top. To estimate the possible effect of uncertainties in the BG shape, all these numbers are added in quadrature, given a total BG uncertainty of 0.32 GeV

Pileup. The effect of pile up has been evaluated by scaling the pile up weight of the signal simulated events up and down, corresponding to a shift of the average vertex multiplicities by ± 0.6. The largest shifts in top mass is quoted as systematic uncertainties.

PDF. The Monte Carlo samples were generated using the CTEQ 6.6 parton distribution functions (PDFs). The uncertainties of these PDFs can be described by 22 orthogonal parameters. Up and down variations of these parameters result in $2 \cdot 22$ additional PDFs, which we employ to evaluate the impact of the PDF uncertainties on our measurement, re-weighting the events according to the deviation of the respective PDF from the original one.

The resulting changes in the reconstructed top mass offset are determined using the $t\bar{t}$ sample at the central value of the top mass. The observed shift in mass (in the lepton+jets channel) for each of the PDF variations is listed in Table 4.6. In general the shifts are very small and in many cases the "up" and "down" variation lead to an opposite shift in the mass, approximately equal is size. Taking the sum in quadrature of the average absolute

value of the shift ("up" or "down") of each variation yields an estimated combined PDF uncertainty of 140 MeV on the calibration offset. Correcting for a calibration slope of 0.87 gives an effect of 161 MeV, and we thus quote 0.16 GeV as potential systematic effect due to the PDF uncertainties. The corresponding results for e+jets and mu+jets are and 160 MeV respectively.

The overall systematic uncertainty is taken to be the quadratic sum of all its components.

4.7 Results

After extracting the top quark mass using combined likelihood fit, performing mass calibration and estimating systematic uncertainties as shown in previous sections, the following result was obtained in the electron plus jets channel:

$$m_{\mathrm{t}} = 172.87 \pm 0.27 \text{ (stat.)} \pm 2.17 \text{ (syst.) GeV}. \tag{4.15}$$

Being a cross-check result, this measurement is consistent with the official CMS top quark mass measurement at 7 TeV in lepton plus jets channel [29]:

$$m_{\mathrm{t}} = 173.49 \pm 0.43 \text{ (stat.+JES)} \pm 0.98 \text{ (syst.) GeV}. \tag{4.16}$$

This result expectedly has a significantly lower uncertainty due to *in situ* measurement of the jet energy scale (JES) in a joint likelihood fit. It was obtained using the combined fit in the electrons plus jets and muon plus jets channels, whereas separate fits in each channel yield following results:

$$\begin{aligned} \text{e+jets: } m_{\mathrm{t}} &= 173.72 \pm 0.66 \text{ (stat.+JES)} \pm 1.00 \text{ (syst.) GeV}, \\ \mu\text{+jets: } m_{\mathrm{t}} &= 173.22 \pm 0.56 \text{ (stat.+JES)} \pm 1.06 \text{ (syst.) GeV}. \end{aligned} \tag{4.17}$$

All results are consistent with the latest Tevatron combination result of $m_{\mathrm{t}} = 173.18 \pm 0.56$ (stat.) ± 0.75 (syst.) GeV [56] and the latest LHC combination result of $m_{\mathrm{t}} = 173.29 \pm 0.23$ (stat.) ± 0.92 (syst.) GeV [57].

4.8 Summary

In this chapter, a top quark mass measurement in the electron plus jets channel was presented using full 2011 dataset recorded by the CMS detector with a total integrated luminosity of 5.0 fb^{-1}. The mass extraction technique called ideogram method was described in detail, as well other technicalities of the analysis. The result is consistent with the official CMS top quark mass measurement in the lepton plus jets channel at 7 TeV, as well as latest combination results from the Tevatron and the LHC experiments.

A. High level triggers

A.1 Electron triggers

A full list of electron triggers including their version numbers used in the $E_{\mathrm{T}}^{\mathrm{miss}}$ analysis (Chapters...):

- HLT-Ele25-CaloIdVT-TrkIdT-CentralTriJet30-v1
  for run number $\leq$ 161216
- HLT-Ele25-CaloIdVT-TrkIdT-CentralTriJet30-v2
  for run number $\leq$ 163269
- HLT-Ele25-CaloIdVT-TrkIdT-TriCentralJet30-v3
  for run number $\leq$ 165969
- HLT-Ele25-CaloIdVT-CaloIsoT-TrkIdT-TrkIsoT-TriCentralJet30-v1
  for run number $\leq$ 166967
- HLT-Ele25-CaloIdVT-CaloIsoT-TrkIdT-TrkIsoT-TriCentralJet30-v2
  for run number $\leq$ 167913
- HLT-Ele25-CaloIdVT-CaloIsoT-TrkIdT-TrkIsoT-TriCentralJet30-v4
  for run number $\leq$ 173235
- HLT-Ele25-CaloIdVT-CaloIsoT-TrkIdT-TrkIsoT-TriCentralJet30-v5
  for run number $\leq$ 178380
- HLT-Ele25-CaloIdVT-CaloIsoT-TrkIdT-TrkIsoT-TriCentralPFJet30-v2
  for run number $\leq$ 179889
- HLT-Ele25-CaloIdVT-CaloIsoT-TrkIdT-TrkIsoT-TriCentralPFJet30-v3
  for run number $\leq$ 180252

A.2 Muon triggers

A full list of muon triggers including their version numbers used in the $E_{\mathrm{T}}^{\mathrm{miss}}$ analysis (Chapter...):

- HLT-IsoMu17-v5 for run number $\leq$ 163269

- HLT-IsoMu17-v6 for run number $\leq$ 163870
- HLT-IsoMu17-v8 for run number $\leq$ 165969
- HLT-IsoMu17-v9 for run number $\leq$ 166345
- HLT-IsoMu17-v10 for run number $\leq$ 166373
- HLT-IsoMu17-v9 for run number $\leq$ 167077
- HLT-IsoMu17-v11 for run number $\leq$ 167914
- HLT-IsoMu17-TriCentralJet30-v5 for run number $\leq$ 172619
- HLT-IsoMu17-DiCentralJet30-v5 for run number $\leq$ 173198
- HLT-IsoMu17-eta2p1-TriCentralJet30-v1 for run number $\leq$ 173656
- HLT-IsoMu17-eta2p1-DiCentralJet30-v1 for run number $\leq$ 173692
- HLT-IsoMu17-eta2p1-TriCentralJet30-v1 for run number $\leq$ 178419
- HLT-IsoMu17-eta2p1-TriCentralPFJet30-v2 for run number $\leq$ 179958
- HLT-IsoMu17-eta2p1-TriCentralPFJet30-v3 for run number $\leq$ 180252

In total 24 % of the data were taken with the single muon trigger, 11 % with the muon-plus- two-jet trigger, 48 % with the muon-plus-three-jet triggers and the last 16 % were taken with the muon-plus-three-jet trigger where the jets are reconstructed with a modified particle flow algorithm (see Chapter 3).

A full list of muon triggers including their version numbers used in the $E_{\mathrm{T}}^{\mathrm{miss}}$ analysis (Chapter...):

- HLT-IsoMu17-v5 for run number $\leq$ 163269
- HLT-IsoMu17-v6 for run number $\leq$ 163870
- HLT-IsoMu17-v8 for run number $\leq$ 165969
- HLT-IsoMu17-v9 for run number $\leq$ 166345
- HLT-IsoMu17-v10 for run number $\leq$ 166373
- HLT-IsoMu17-v9 for run number $\leq$ 167077

- HLT-IsoMu17-v11 for run number $\leq$ 167914
- HLT-IsoMu24-v8 for run number $\leq$ 170249
- HLT-IsoMu24-v9 for run number $\leq$ 173236
- HLT-IsoMu24-eta2p1-v6 for run number $\leq$ 178420
- HLT-IsoMu24-eta2p1-v9 for run number $\leq$ 180252

B. Top mass analysis control plots

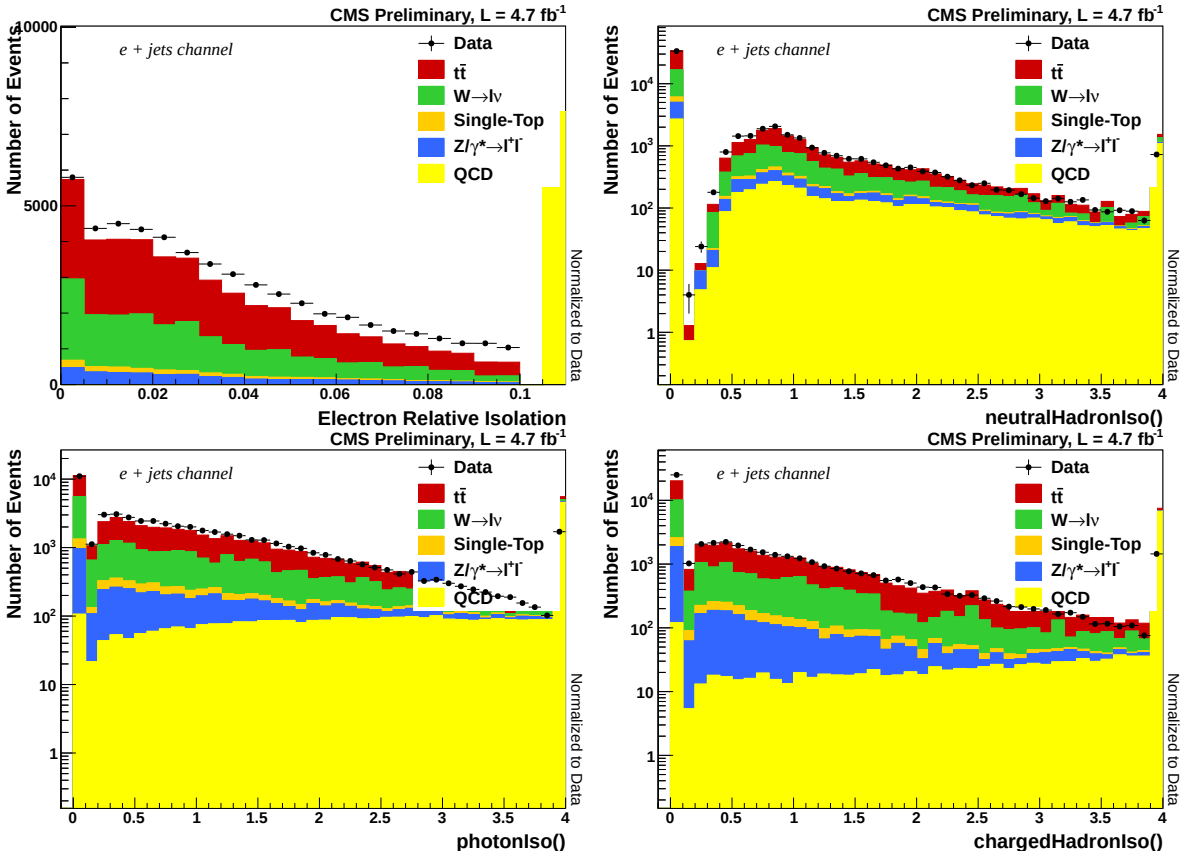


Figure B.1: Control plots for lepton isolation quantities. Left are for μ+jets, right are for e+jets.

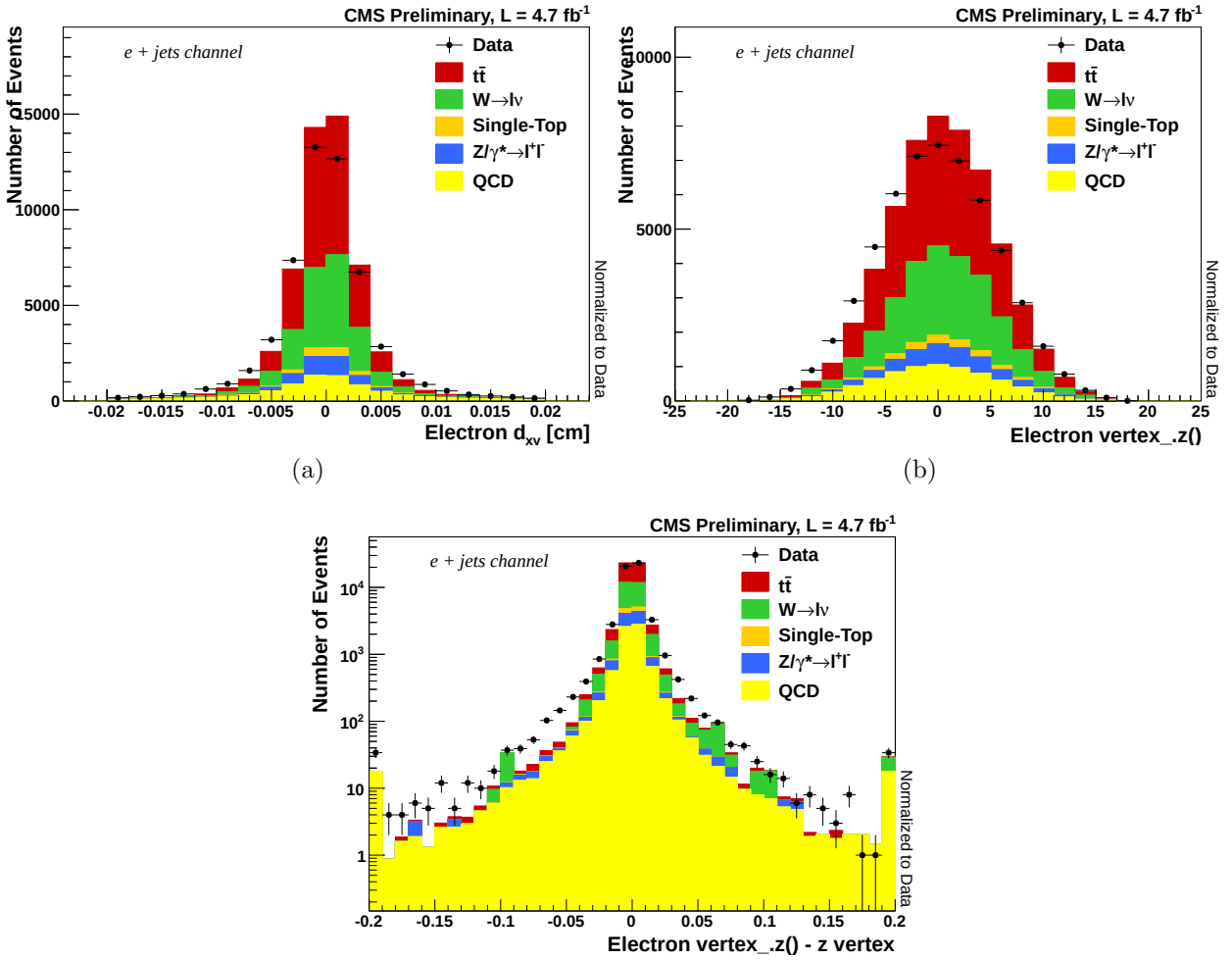


Figure B.2: Control plots for lepton kinematic quantities. Left are for μ+jets, right are for e+jets.

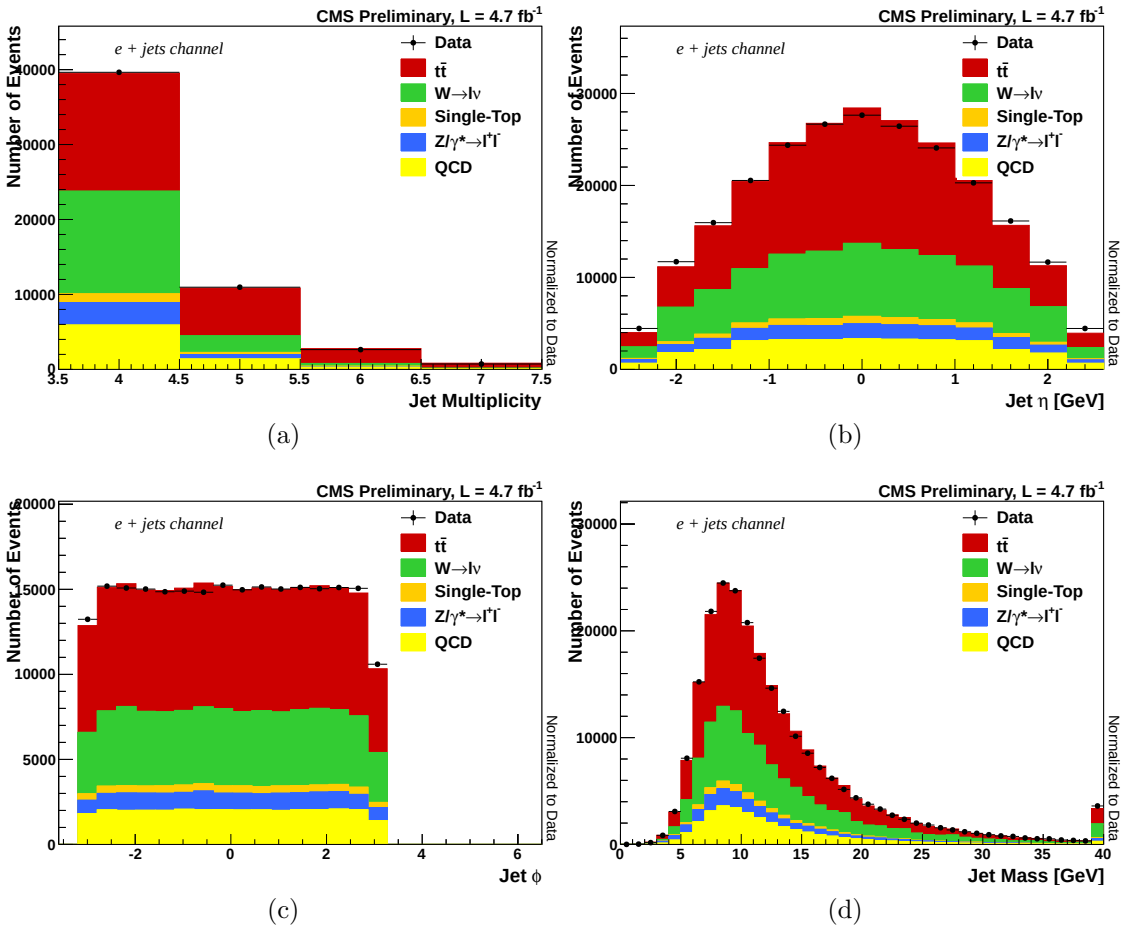


Figure B.3: Control plots for jet kinematic quantities. Left are for μ+jets, right are for e+jets.

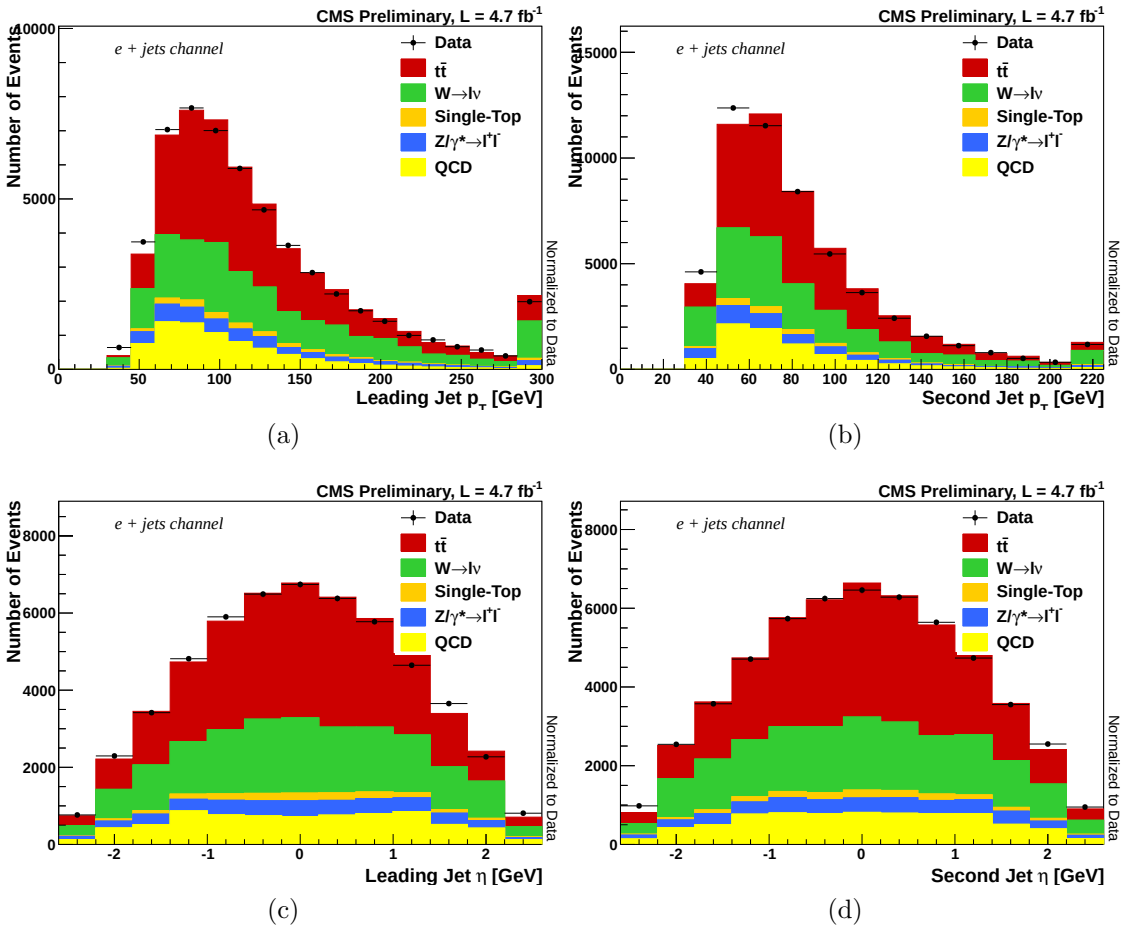


Figure B.4: Control plots for the jet pseudorapidity and transverse momemtum for the 2 leading jets. Left are for μ+jets, right are for e+jets.

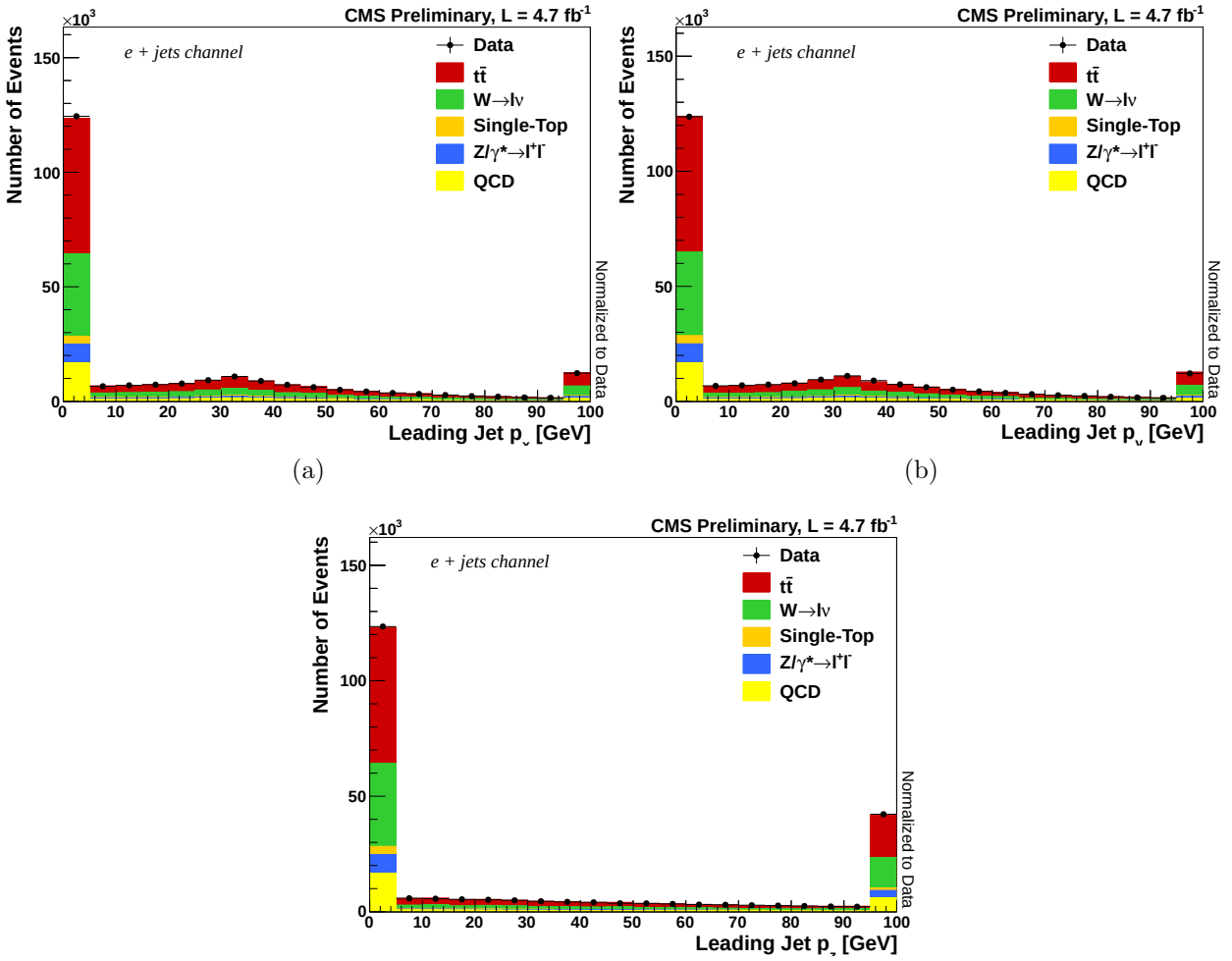


Figure B.5: Control plots for the jet momentum components. Left are for μ+jets, right are for e+jets.

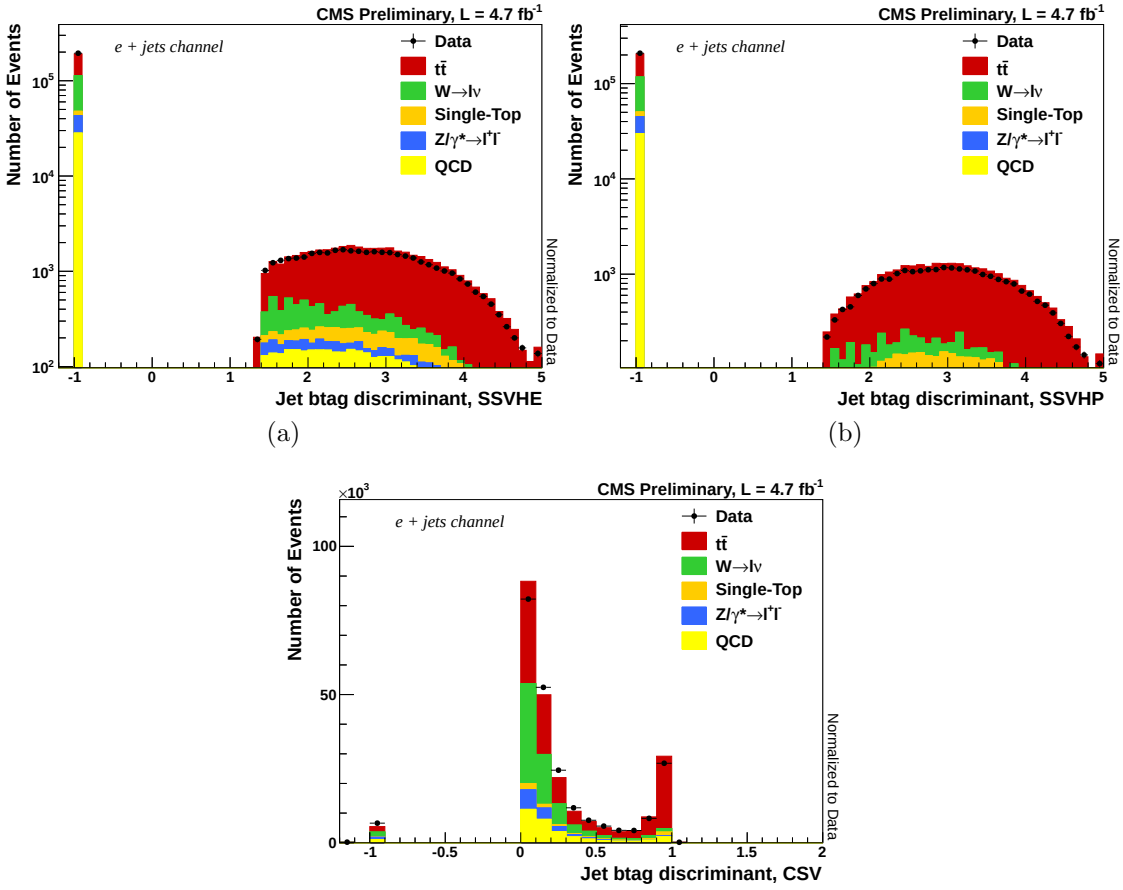


Figure B.6: Control plots for the different b-tagging discriminants for e+jets.

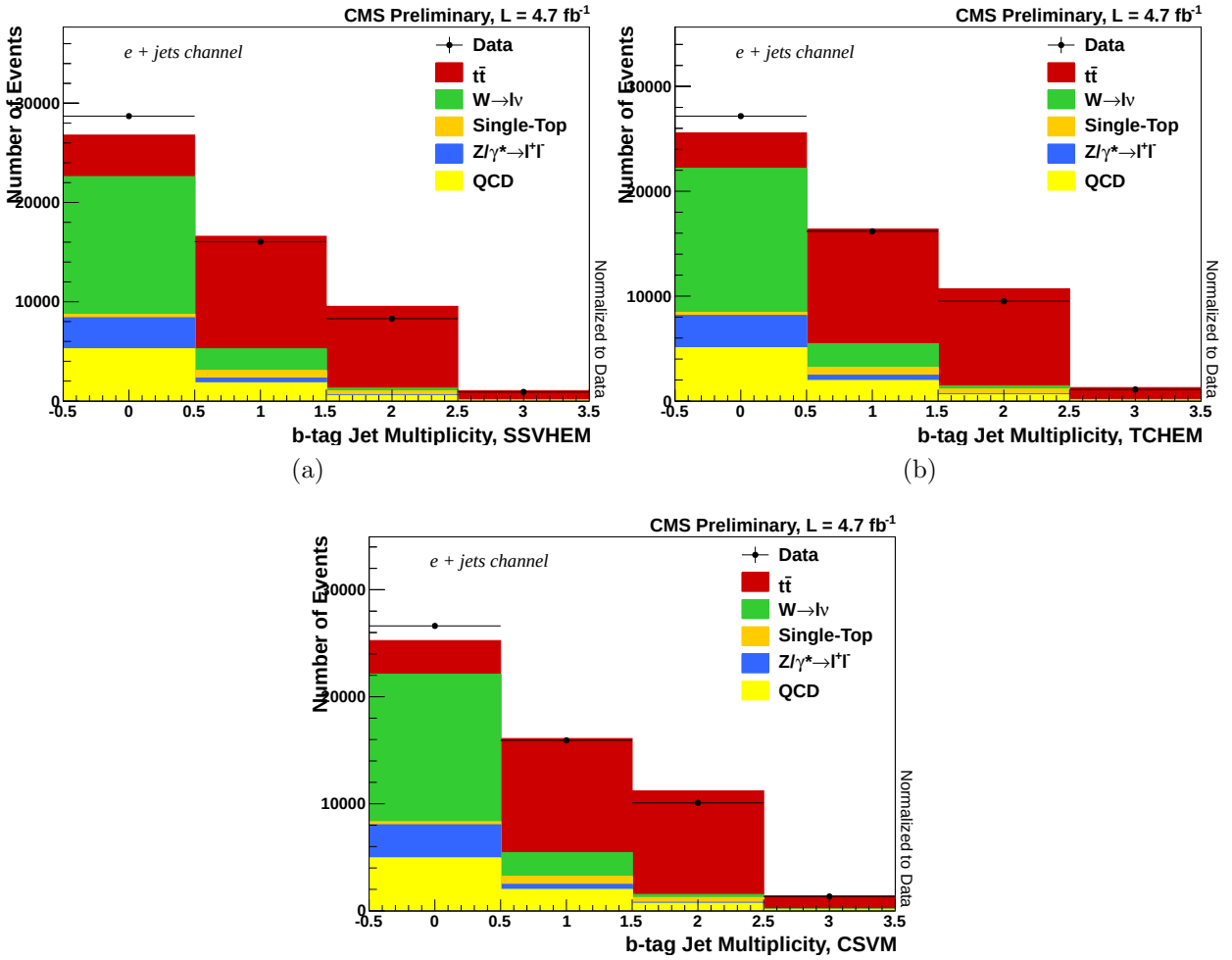


Figure B.7: Control plots for the different b-tagging discriminants. Left are for μ+jets, right are for e+jets.

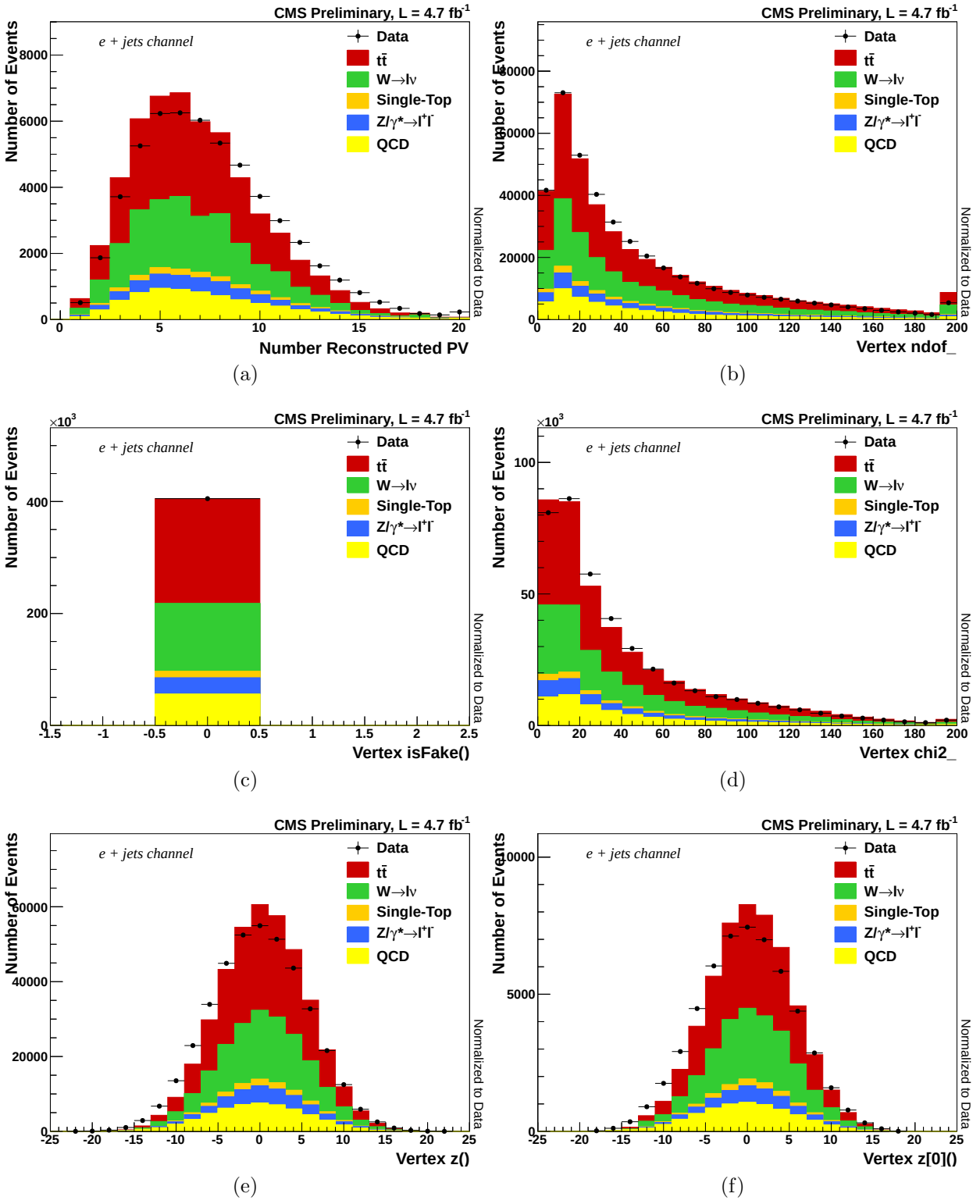
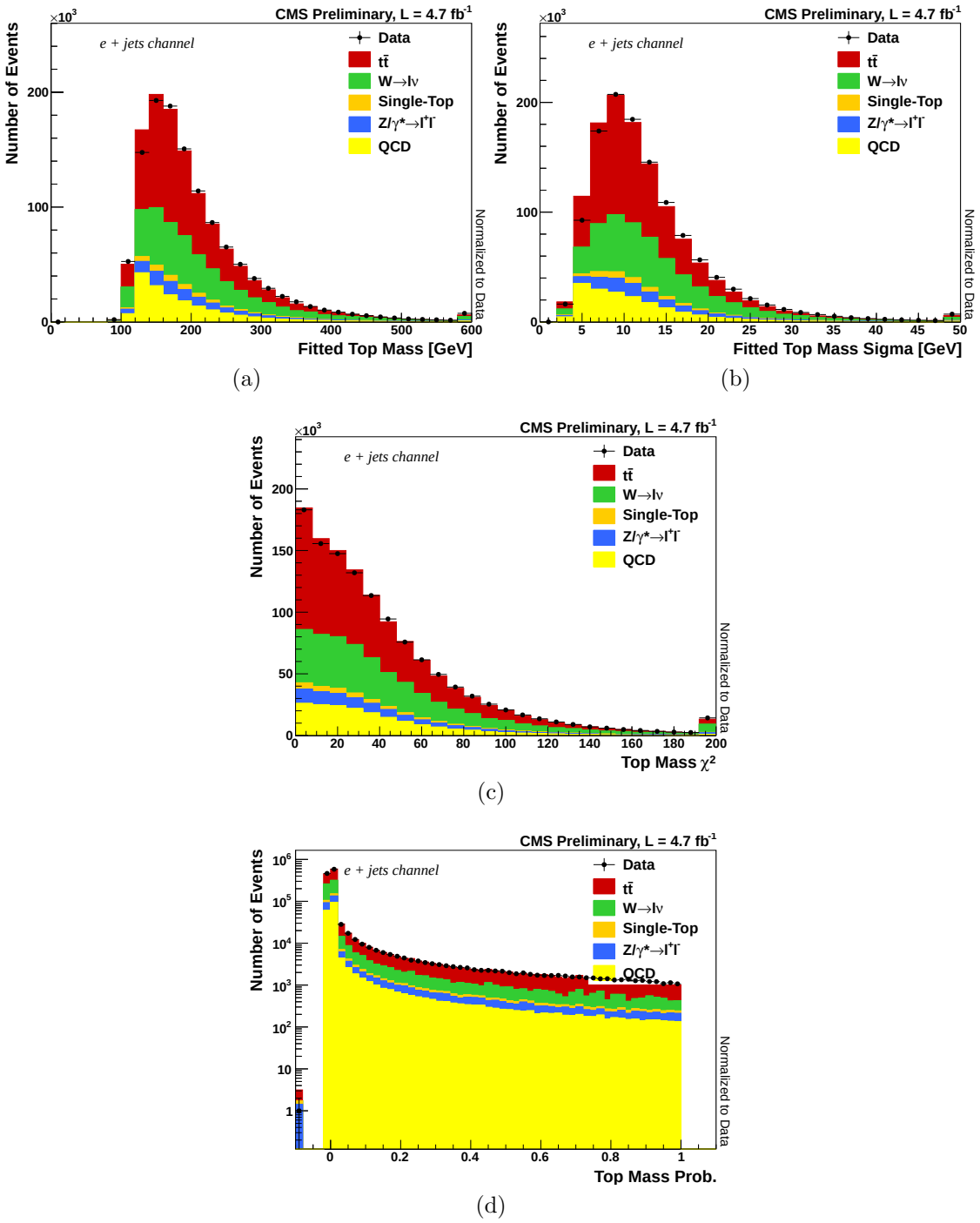


Figure B.8: Control plots for different primary vertex quantities in the e+jets channel.



(a) (b) (c) (d)

Figure B.9: Control plots for a few fitted quantities (all combinations are shown). Left are for μ+jets, right are for e+jets.

References

[1] L. Evans and P. Bryant. "LHC Machine". *JINST* 3.08 (2008), S08001. DOI: 10.1088/1748-0221/3/08/S08001.

[2] J.P. Blewett. "200 GeV intersecting storage accelerators". *Proceedings of The 8th International Conference on High-Energy Accelerators*. CERN. Geneva, Switzerland, 1971, p. 501.

[3] The CMS Collaboration. "The CMS experiment at the CERN LHC". *JINST* 3.08 (2008), S08004. DOI: 10.1088/1748-0221/3/08/S08004.

[4] The CMS Collaboration. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". *Physics Letters B* 716.1 (2012), pp. 30–61. DOI: 10.1016/j.physletb.2012.08.021.

[5] The CMS Collaboration. *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software*. Technical Design Report CMS. Geneva, Switzerland: CERN, 2006.

[6] The CMS Collaboration. "Precise mapping of the magnetic field in the CMS barrel yoke using cosmic rays". *Journal of Instrumentation* 5.03 (2010). DOI: 10.1088/1748-0221/5/03/T03021.

[7] Rene Brun and Fons Rademakers. "ROOT – An object oriented data analysis framework". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 389.12 (1997). New Computing Techniques in Physics Research V, pp. 81 –86. DOI: 10.1016/S0168-9002(97)00048-X.

[8] *CMSSW software framework*. URL: http://cms-sw.github.io/cmssw.

[9] *Bristol Analysis Tools*. URL: https://github.com/BristolTopGroup/AnalysisSoftware.

[10] *Daily Python Scripts*. URL: https://github.com/BristolTopGroup/DailyPythonScripts.

[11] *rootpy: Pythonic ROOT package*. URL: http://rootpy.org.

[12] *matplotlib plotting library*. URL: http://matplotlib.org.

[13] The CMS Collaboration. *Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET*. Tech. rep. CMS-PAS-PFT-09-001. Geneva, Switzerland: CERN, 2009.

[14] R. Frühwirth. "Application of Kalman filtering to track and vertex fitting". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 262.23 (1987), pp. 444 –450. DOI: 10.1016/0168-9002(87)90887-4.

[15] W Adam et al. "Reconstruction of electrons with the Gaussian-sum filter in the CMS tracker at the LHC". *Journal of Physics G: Nuclear and Particle Physics* 31.9 (2005), N9. DOI: 10.1088/0954-3899/31/9/N01.

[16] S. Baffioni et al. "Electron reconstruction in CMS". *The European Physical Journal C* 49.4 (2007), pp. 1099–1116. DOI: 10.1140/epjc/s10052-006-0175-5.

[17] The CMS Collaboration. *Cuts in Categories (CiC) Electron Identification*. Online; accessed 04-September-2013. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideCategoryBasedElectronID.

[18] The CMS collaboration. "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV". *Journal of Instrumentation* 7.10 (2012), P10002. DOI: 10.1088/1748-0221/7/10/P10002.

[19] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. "The anti-k_t jet clustering algorithm". *Journal of High Energy Physics* 2008.04 (2008), p. 063. DOI: 10.1088/1126-6708/2008/04/063.

[20] Rick D. Field. "The underlying event in hard scattering processes". *eConf* C010630 (2001), P501. arXiv: hep-ph/0201192.

[21] The CMS collaboration. "Identification of b-quark jets with the CMS experiment". *Journal of Instrumentation* 8.04 (2013), P04013. DOI: 10.1088/1748-0221/8/04/P04013.

[22] The CMS Collaboration. *CMS TriDAS project: Technical Design Report, Volume 1: The Trigger Systems*. CERN-LHCC-2000-038; CMS-TDR-6-1. CERN/LHCC, 2000.

[23] J. Brooke. "Performance of the CMS Level-1 Trigger". *Proceedings of Science* ICHEP2012 (2013), p. 508. arXiv: 1302.2469 [hep-ex].

[24] The CMS collaboration. "The CMS High Level Trigger System". *2007 15th IEEE-NPSS Real-Time Conference*. Institute of Electrical and Electronics Engineers, 2007. DOI: 10.1109/RTC.2007.4382773.

[25] The CMS collaboration. "Commissioning of the CMS High Level Trigger". *Journal of Instrumentation* 4.10 (2009), P10005. DOI: 10.1088/1748-0221/4/10/P10005.

[26] Wolfgang Adam et al. *Track Reconstruction in the CMS tracker*. Tech. rep. CMS-NOTE-2006-041. Geneva, Swtizerland: CERN, 2006.

[27] Diego Casadei. "Estimating the selection efficiency". *Journal of Instrumentation* 7.08 (2012), P08021. DOI: 10.1088/1748-0221/7/08/P08021.

[28] M. Agelou et al. *Top Trigger Efficiency Measurements and the top trigger package*. Tech. rep. DØ Note 4512. 2004.

[29] The CMS Collaboration. "Measurement of the top-quark mass in $t\bar{t}$ events with lepton+jets final states in pp collisions at $\sqrt{s} = 7$ TeV". *Journal of High Energy Physics* 2012.12 (2012), pp. 1–37. DOI: 10.1007/JHEP12(2012)105.

[30] The CMS Collaboration. "Measurement of the mass difference between top and anti-top quarks". *Journal of High Energy Physics* 2012.6 (2012), pp. 1–36. DOI: 10.1007/JHEP06(2012)109.

[31] The CMS Collaboration. *Measurement of the top quark mass in the l+jets channel*. Tech. rep. CMS-PAS-TOP-10-009. Geneva, Switzerland: CERN, 2011.

[32] Johan Alwall et al. "MadGraph 5: going beyond". *Journal of High Energy Physics* 2011.6 (2011), pp. 1–40. DOI: 10.1007/JHEP06(2011)128.

[33] Torbjörn Sjöstrand et al. "High-energy-physics event generation with Pythia 6.1". *Computer Physics Communications* 135.2 (2001), pp. 238–259. DOI: 10.1016/S0010-4655(00)00236-8.

[34] Torbjörn Sjöstrand, Stephen Mrenna, and Peter Z. Skands. "PYTHIA 6.4 Physics and Manual". *JHEP* 05 (2006). DOI: 10.1088/1126-6708/2006/05/026. arXiv: hep-ph/0603175.

[35] Stefano Frixione, Paolo Nason, and Carlo Oleari. "Matching NLO QCD computations with parton shower simulations: the POWHEG method". *Journal of High Energy Physics* 11 (2007). DOI: 10.1088/1126-6708/2007/11/070.

[36] Z. Was. "TAUOLA the library for lepton decay, and KKMC/KORALB/KORALZ/ status report". *Nuclear Physics B - Proceedings Supplements* 98.13 (2001), pp. 96 –102. DOI: 10.1016/S0920-5632(01)01200-2.

[37] GEANT4 Collaboration. "GEANT4 - a simulation toolkit". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250 –303. DOI: 10.1016/S0168-9002(03)01368-8.

[38] Stefan Höche et al. "Matching parton showers and matrix elements" (2006). Proceedings of the "HERA and the LHC" workshop, CERN/DESY 2004/2005". arXiv: hep-ph/0602031.

[39] Stefano Frixione and Bryan R. Webber. "Matching NLO QCD computations and parton shower simulations". *Journal of High Energy Physics* 06 (2002). DOI: 10.1088/1126-6708/2002/06/029.

[40] The CMS Collaboration. *Tracking and Primary Vertex Results in First 7 TeV Collisions*. Tech. rep. CMS-PAS-TRK-10-005. Geneva, Switzerland: CERN, 2010.

[41] Suharyo Sumowidagdo. *Programming documentation of HitFit: A Kinematic Fitter for top quark-antiquark pair lepton+jets event*. CMS AN-2011/171. 2011.

[42] Orin I. Dahl, T.B. Day, and Frank T. Solmitz. *SQUAW Kinematic Fitting Program*. Tech. rep. P-126. Berkeley, California: Lawrence Radiation Laboratory, 1968.

[43] DØ Collaboration. "Direct measurement of the top quark mass by the DØ Collaboration". *Phys. Rev. D* 58 (5 1998). DOI: 10.1103/PhysRevD.58.052001.

[44] DØ Collaboration. "Measurement of the top quark mass in the lepton+jets channel using the ideogram method". *Phys. Rev. D* 75 (9 2007). DOI: 10.1103/PhysRevD.75.092001.

[45] DØ Collaboration. "Measurement of the Forward-Backward Charge Asymmetry in Top-Quark Pair Production". *Phys. Rev. Lett.* 100 (14 2008). DOI: 10.1103/PhysRevLett.100.142002.

[46] *CLHEP - A Class Library for High Energy Physics*. URL: http://cern.ch/CLHEP.

[47] Scott Stuart Snyder. "Measurement of the Top Quark Mass at DØ". FERMILAB-THESIS-1995-27. PhD thesis. Stony Brook University, 1995.

[48] DELPHI Collaboration. "Measurement of the W-pair cross-section and of the W mass in e^+e^- interactions at 172 GeV". *The European Physical Journal C - Particles and Fields* 2.4 (1998), pp. 581–595. DOI: 10.1007/s100529800857.

[49] DELPHI Collaboration. "Measurement of the mass and width of the W boson in e^+e^- collisions at $\sqrt{s}$ = 161-209 GeV". *The European Physical Journal C* 55.1 (2008), pp. 1–38. DOI: 10.1140/epjc/s10052-008-0585-7.

[50] CDF Collaboration. "Measurement of the Top-Quark Mass in All-Hadronic Decays in $p\overline{p}$ Collisions at CDF II". *Phys. Rev. Lett.* 98 (14 2007). DOI: 10.1103/PhysRevLett.98.142001.

[51] DØ Collaboration. "Improved determination of the width of the top quark". *Phys. Rev. D* 85 (9 2012), p. 091104. DOI: 10.1103/PhysRevD.85.091104.

[52] The ALEPH Collaboration et al. "Precision electroweak measurements on the Z resonance". *Physics Reports* 427.56 (2006), pp. 257 –454. DOI: 10.1016/j.physrep.2005.12.006.

[53] K.S. Kölbig and B. Schorr. "A program package for the Landau distribution". *Computer Physics Communications* 31.1 (1984), pp. 97 –111. DOI: 10.1016/0010-4655(84)90085-7.

[54] The CMS collaboration. "Determination of jet energy calibration and transverse momentum resolution in CMS". *Journal of Instrumentation* 6.11 (2011). DOI: 10.1088/1748-0221/6/11/P11002.

[55] Daniel Wicke and Peter Z. Skands. "Non-perturbative QCD Effects and the Top Mass at the Tevatron". *Nuovo Cim.* B123 (2008), S1. arXiv: 0807.3248.

[56] CDF and DØ Collaborations. "Combination of the top-quark mass measurements from the Tevatron collider". *Phys. Rev. D* 86 (9 2012), p. 092003. DOI: 10.1103/PhysRevD.86.092003.

[57] ATLAS and CMS Collaborations. *Combination of ATLAS and CMS results on the mass of the top-quark using up to 4.9 fb^{-1} of $\sqrt{s}$ = 7 TeV LHC data.* Tech. rep. ATLAS-CONF-2013-102. Geneva: CERN, 2013.