Vrije Universiteit Brussel

Faculteit Wetenschappen
Departement Natuurkunde

Measurement of the top quark pair production cross section at the LHC with the CMS experiment

Michael Maes

Promotor
Prof. Dr. Jorgen D'Hondt
Proefschrift ingediend met het oog op het behalen van
de academische graad Doctor in de Wetenschappen

September 2013

Doctoral examination commission

Chair: Prof. Dr. N. Van Eijndhoven (VUB)

Supervisor: Prof. Dr. J. D'Hondt (VUB)

Secretary: Prof. Dr. F. Blekman (VUB)

Prof. Dr. S. Bentvelsen (NIKHEF)

Prof. Dr. S. Lowette (VUB)

Prof. Dr. A. Sevrin (VUB)

Prof. Dr. R. Tenchini (INFN/Pisa)

Prof. Dr. ir. G. Vandersteen (VUB)

Contents

Introduction

All the particles that constitute the matter of the universe and the forces governing them, except gravity, are described by a theory called the Standard Model. This theory has been formulated in the 1970s and since then has been validated in numerous experiments. To date, the predictions made by this theory have been validated to high precision which is the basis of its success. However, the Standard Model has its flaws such as the inability to describe gravity. One of its longstanding issues was the inability to incorporate the mass of massive particles. This was solved by Brout, Englert and Higgs introducing a new mechanism to attribute mass to particles. For this method to be validated, a boson called the Brout-Englert-Higgs boson has to be discovered.

To search for new phenomena as well as the illusive BEH boson, large particle colliders have to be built where the universe can be probed at it's smallest scales. The Large Hadron Collider (LHC) that has been constructed at CERN came into operation in March 2010 and is the largest and most energetic collider in the world colliding proton beams at 7 and 8 TeV.

To study these proton-proton collisions provided by the LHC, large experiments have been constructed and among them is the Compact Muon Solenoid (CMS). The latest success of the Standard Model came when in 2012 the CMS and ATLAS collaborations claimed the discovery of a boson that is consistent with the long-sought Brout-Englert-Higgs boson. Nevertheless, the search for new phenomena continues at full force.

The top quark, the heaviest particle in the Standard Model, plays a very important role in the research at the LHC. The large sample of top quark events produced at the LHC provides a unique opportunity to measure this quarks properties with great precision. The production cross section of top quark pairs, for example, does not only benchmark the perturbative calculations in QCD but is also sensitive to various hypothetical new physics phenomena that would modify its value.

Additionally, the top quark also serves as a calibration tool. Since it almost exclusively decays into a W boson and a b quark, the large sample of b quarks can be used to calibrate b quark identification, or b-tagging, algorithms. These algorithms are used by various analyses to isolate signal processes containing b quarks in the final state from the usually abundant backgrounds.

In this thesis, a fully data-driven method is introduced to measure the b-tagging efficiency as well as its mis-tagging efficiency. Furthermore, b-tagging is used to isolate the top quarks from the background to measure the top quark pair production cross section.

In Chapter 1, the Standard model is introduced along with the current status of the

top quark research. Subsequently, the Large Hadron Collider and the CMS experiment are introduced in Chapter 2. To be able to develop and benchmark analysis techniques, simulated events are required. The event generation and detector simulation procedure is outlined in Chapter 3 together with the reconstruction of the detector signals to physics objects. Then, the reconstructed events can be passed along to the event selection step in Chapter 4 to separate signal and background events. In Chapter 5 the method to measure the b-tagging efficiency is introduced. This method is then extended in Chapter 6 to also measure the mis-tagging efficiency. The measurement of the top quark pair cross section is then detailed in Chapter 7. Finally, Chapter 8 concludes on the measurements and provides some future perspectives.

The results presented in Chapters 5, 6 and 7 are based on my own research. The measurement of the b-tagging efficiency has been published in [1–3] and the $t\bar{t}$ cross section in [4–6] with a paper publication for the 8 TeV result still in the pipeline. Finally, the measurement of the mis-tagging efficiency has not been made public so far.

Chapter 1

The top quark sector of the Standard Model

In our universe, all matter is built up by tiny invisible particles. These elementary particles[1] interact according to fundamental forces. Consequently, a theory describing the dynamics of our universe is required to contain a consistent description of all these particles and how they interact. The Standard Model of Particle Physics, or Standard Model (SM), effectively describes the known particles in our universe and is able to incorporate three out of four fundamental forces. Moreover, the Standard Model does not only excel in being the best attempt at a unified theory of the universe; it also excels in the high precision by which it can predict the outcome of numerous experiments carried out since its birth during the second part of the 20^{th} century.

One particle of the Standard Model is of particular interest in current research, namely the top quark. This quark was discovered in 1995 and to date is the heaviest known elementary particle. Its importance in the understanding of the Standard Model and the universe is two-fold. First, this quark provides a unique window to phenomena beyond the reach of the Standard Model as new hypothetical particle can decay in top quarks. Secondly, the top quark can be used as an experimental calibration tool to further improve precision measurements.

1.1 The Standard Model

The fundamental particles building up all matter in the universe are described by the Standard Model along with the electromagnetic, weak and strong forces that govern them [7–9]. Each of these three forces is accompanied by a force-carrying particle. In the first Section, the matter particles or fermions are introduced along with the force-carrying particles called bosons. Subsequently, in Section 1.1.2, these particles will be described as fields in a quantum field theory. In this framework, the fields can interact among each other by requiring the theory to be invariant under local gauge transformations. Furthermore, the gauge groups that make up the Standard Model will be discussed. Additionally, the mechanism of spontaneous symmetry breaking that

[1] An elementary particle cannot be further decomposed in other even smaller particles.

gives mass to all particles will be introduced. To end the discussion on the Standard Model, some of its problems and possible extensions will be outlined in Section 1.1.4.

1.1.1 Fermions and bosons: the building blocks of the universe

In the Standard Model two main categories of particles are distinguished: the bosons and the fermions. The twelve fermions, or matter particles, are considered to built up all known matter. Secondly the bosons, or force-carriers, are associated to the forces that govern the interaction between the particles.
The fermion group consists of 12 half-integer spin particles that are either labeled a lepton or a quark. The leptons group consists of the electron (e^-), muon (μ^-) and the tau (τ^-) particles. For each of these leptons an associated neutrino exists. These particles are charge neutral and as a consequence can only interact through the weak force. Because of their negligible mass ($<<$ 1 eV), neutrinos are considered massless throughout this thesis. Conversely, since the leptons carry a negative electrical charge they can interact both through the weak and the electromagnetic forces. The negative elementary charge of - 1 e for the leptons is further emphasised by the minus superscript in their symbol.

The second group of fermions consists of quarks. Each fermion generation distinguishes an up-type and a down-type quark. The up-type quark carries a +2/3 e charge while the down-quark has a -1/3 e electrical charge. As opposed to the leptons the quarks can also interact through the third fundamental force described by the Standard Model being the strong interaction.

Each of the fermions has an anti-particle with the same mass but an opposite electrical charge. The anti-particle of fermion f is denoted as $\bar{f}$. For the charged leptons, the anti-lepton is denoted with a "+" superscript rather than putting a bar on the symbol. This clearly states that the anti-lepton has a positive charge. Moreover, the anti-electron is called a positron.

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

Table 1.1: Overview of the fermions in the Standard Model along with their electrical charge

Within the fermion group, three generations can be identified as shown in Table 1.1 where each generation consists of a lepton, an associated neutrino and two quarks. The first generation is considered to built up all visible matter in the universe while all subsequent generations contain particles with identical quantum mechanical properties except for a higher mass. The proton consists of two up quarks and one down quark while a neutron is composed of two down quarks and one up quark. Protons and

	Boson	Mass (GeV)
Electromagnetic force	Photon (γ)	Massless
Weak force	$W^{\pm}$	80.385 ± 0.015
Weak force	Z^0	91.1876 ± 0.0021
Strong force	8 gluons (g)	Massless
Generating mass	H^0	125.7 ± 0.4

Table 1.2: Bosons of the Standard Model [10, 11]

neutrons along with electrons in their turn build up atoms which build up all known matter. The second and third generation do not occur as stable particles in nature, nevertheless they are present in cosmic rays and can be produced under laboratory conditions as well.

The forces that allow the fermions to interact are carried by integer-spin particles called bosons. First the electromagnetic force is governed by the massless photon. Next the massive $W^{\pm}$ and Z^0 bosons carry the weak interaction. Finally, the strong force is accompanied by eight massless gluons. These particles are summarised in Table 1.2 together with their masses. In this table, the quoted unit of mass is GeV rather than GeV/c^2 since in this thesis the convention c=1 is used.

In addition to the bosons carrying the three fundamental forces incorporated within the Standard Model there exists one more spin-zero boson: the Brout-Englert-Higgs (BEH) boson. This boson gives mass to all other particles via the mechanism that will be explained in Section 1.1.3. This boson was only recently discovered in 2012 by both the CMS and ATLAS experiments at the Large Hadron Collider [12, 13].

1.1.2 The Standard Model as a quantum field theory

To lay out the theoretical foundation of the Standard Model, quantum field theory is used. The virtue of this framework is that it combines both quantum mechanics and special relativity. In this theory, the fermions are described by fields and the interactions between the fields come around by demanding the theory to be gauge invariant. First it will be shown how the Lagrangian density for such theory could be obtained. Afterwards the gauge groups constituting the Standard Model will be explained.

Interacting fields through local gauge invariance

In the Standard Model, fermions are represented by fields, more precisely Dirac-spinor fields ψ. The Lagrangian for this theory is thus the Dirac Lagrangian which can be written in terms of the fermion field ψ, the anti-fermion field $\bar{\psi}$, the particle mass m and the Dirac γ-matrices.

$$\mathcal{L}_{Dirac} = i\bar{\psi}\gamma^{\mu}\partial_{\mu}\psi - m\bar{\psi}\psi \tag{1.1}$$

To obtain a gauge invariant theory, invariance of the Dirac Lagrangian under a local phase transformation is enforced. This local phase transformation is given by

$$\psi' = U\psi = e^{i\epsilon^a(x)\cdot\frac{\tau^a}{2}}\psi \tag{1.2}$$

where $\epsilon^a(x)$ (a=1,...,n) are parameters for a n-dimensional Lie-group with generators τ^a. To make the Dirac Lagrangian invariant under this transformation, the normal derivative ∂_μ is to be replaced by a so-called covariant derivative defined as

$$\mathcal{D}_\mu = \partial_\mu - ig\frac{\tau^a}{2}A^a_\mu \tag{1.3}$$

The covariant derivative ensures the invariance of the Lagrangian under this local gauge transformations and introduces new interacting fields A^a_μ . The Dirac Lagrangian now becomes

$$\begin{aligned}\mathcal{L} &= i\bar{\psi}\gamma^\mu\mathcal{D}_\mu\psi - m\bar{\psi}\psi \\ &= i\bar{\psi}\gamma^\mu\partial_\mu\psi - m\bar{\psi}\psi + g\bar{\psi}\gamma^\mu\frac{\tau^a}{2}A^a_\mu\psi.\end{aligned} \tag{1.4}$$

where the last term in the Lagrangian shows the interaction between the fermion fields and the new vector fields. The factor g is called a coupling constant and is proportional to the interaction strength.

The technique of requiring the theory to be invariant under local gauge transformations, or so-called local gauge symmetry, has shown to yield interacting fields in the Lagrangian. When representing the gauge transformation in an Abelian group including commuting generators, the resulting interacting fields can only couple to the fermion fields. Conversely, if the transformation is represented by a non-abelian group, the gauge fields can also couple among themselves.

Gauge groups of the Standard Model

As mentioned at the beginning of this chapter the Standard Model incorporates three of the fundamental forces in nature. As a consequence, three gauge groups are defined as

$$G_{SM} = SU(3)_C \otimes SU(2)_L \otimes U(1)_Y \tag{1.5}$$

where the first group represents the gauge symmetry of the strong force and the other two the gauge symmetry of the unified electroweak force.

To describe the electroweak interaction the Lagrangian has to be gauge invariant under $SU(2)_L \otimes U(1)_Y$. This can be achieved by introducing the following covariant derivative

$$D_\mu = \partial_\mu - ig\frac{\tau^a}{2}W^a_\mu - ig'\frac{Y}{2}B_\mu \tag{1.6}$$

The factors g and g' are the respective coupling constants, τ^a are Pauli matrices and the hyper charge is denoted as Y. The local gauge invariance under the Abelian group $U(1)_Y$ introduces a field B_μ . Moreover, the local gauge invariance under $SU(2)_L$ introduces three gauge fields W^a_μ, a={1,2,3}. These fields do however not immediately

represent the bosons discussed in the first section. To obtain these bosons, linear combinations of the gauge fields have to be taken to generate them. Consequently, the W boson is defined as

$$W_\mu^\pm = \sqrt{\frac{1}{2}}\left(W_\mu^1 \mp iW_\mu^2\right). \tag{1.7}$$

The neutral Z boson is defined as

$$Z_\mu^0 = W_\mu^3 \cos\theta_w - B_\mu \sin\theta_w. \tag{1.8}$$

And finally, the photon can be obtained using the following linear combination

$$A_\mu = W_\mu^3 \sin\theta_w + B_\mu \cos\theta_w \ . \tag{1.9}$$

In the previous equations, the Weinberg angle (θ_w) is defined as

$$\tan\theta_w = \frac{g'}{g}. \tag{1.10}$$

The strong interaction is described in terms of quantum chromodynamics (QCD). Requiring the Lagrangian to be invariant under the transformations of the non-Abelian $SU(3)_C$ group generates an additional eight gauge fields G_μ^a , a=1,...,8. These gauge fields represent the eight gluons introduced in this chapter. To obtain the local gauge symmetry again a covariant derivative is defined as

$$D_\mu = \partial_\mu - ig_s\frac{\lambda^a}{2}G_\mu^a \ , \tag{1.11}$$

where λ^a are the Gell-Mann matrices and g_s is the strong coupling constant. The subscript C in the group definition emphasises the fact that quarks that transform as a triplet under $SU(3)_C$ carry a colour charge and gluons are coloured as wel. The eight gluons can also interact among themselves as the group is non-Abelian.

To accommodate, amongst others, the experimental observation of CP violating processes, the quark eigenstates in the strong interactions are considered to differ slightly from these in the weak interaction. Hence a matrix is defined linking the quark eigenstates from both the strong and weak interactions, also known as the Cabibbo-Kobayashi-Maskawa (CKM) matrix. This matrix is defined as

$$\begin{pmatrix} d^{weak} \\ s^{weak} \\ b^{weak} \end{pmatrix}_L = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix}_L \tag{1.12}$$

where each element V_{ij} is proportional to the probability of quark type i to dacay into a quark from type j through the charged weak interaction.

1.1.3 Spontaneous symmetry breaking: generating particle masses

In Section 1.1.1 the force-carrying particles were introduced. The photons and gluons, governing the electromagnetic and strong forces respectively, are massless. In contrast to the massless bosons, the weak interactions' $W^\pm$ and Z^0 bosons are massive.

Although the electroweak interaction is unified in the Standard Model the symmetry should be broken at low energies.

To generate mass for the $W^{\pm}$ and Z^0 bosons, a mass term can be added to the Lagrangian. As a consequence the invariance under local gauge transformations is broken which breaks the gauge symmetries of the Standard Model. Hence, the implementation of these explicit mass terms is forbidden. Another strategy to add mass to particles is through the mechanism of spontaneous symmetry breaking [14] often called the Brout-Englert-Higgs (BEH) mechanism [15, 16].

The easiest way to break the electroweak gauge symmetry is to add a scalar SU(2) field ϕ

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}, \tag{1.13}$$

to the Lagrangian. This ensures the invariance under local gauge transformations while breaking the gauge symmetry of the vacuum. The fields ϕ^+ and ϕ^0 are complex fields. This provides the following Lagrangian:

$$\begin{aligned} \mathcal{L}_{Brout-Englert-Higgs} &= \left(\mathcal{D}^{\mu}\phi\right)^{\dagger}\mathcal{D}_{\mu}\phi - V(\phi) & (1.14)\\ &= \left(\mathcal{D}^{\mu}\phi\right)^{\dagger}\mathcal{D}_{\mu}\phi - \mu^2\left(\phi^{\dagger}\phi\right) - \lambda\left(\phi^{\dagger}\phi\right)^2, & (1.15) \end{aligned}$$

where λ is a positive number representing the strength of the field self interaction. The mass parameter μ^2 defines two possible scenarios. For $\mu^2 > 0$, the potential $V(\phi)$ reaches a minimum at $\phi = 0$. Additionally, for $\mu^2 < 0$ the minimum is no longer unique and the potential now reaches a minimum when

$$|\phi|^2 = \phi^{\dagger}\phi = \frac{|\mu^2|}{2\lambda} = \frac{v^2}{2}. \tag{1.16}$$

The vacuum can now be written as a quantum fluctuation around the vacuum expectation value v

$$\phi = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}, \tag{1.17}$$

where the only real field is h(x). This field is called the Brout-Englert-Higgs boson field. This field is associated with a scalar boson, called the Brout-Englert-Higgs boson, with a mass $M_H = \sqrt{2\lambda v^2}$. For the vacuum to be stable, λ has to be positive at all scales Q, this introduces a lower bound for M_H. The remaining three complex fields are absorbed by the $W^{\pm}$ and Z^0 bosons to acquire their mass. Inserting the covariant derivative from the electroweak theory into the Lagrangian in Eq. 1.15 leads to the mass terms for the electroweak gauge bosons.

$$m_W = \frac{1}{2}\,v\,g \qquad\qquad m_Z = \frac{1}{2}\,v\,\sqrt{g^2 + g'^2} \tag{1.18}$$

The fermions do not acquire mass in the same way as the bosons. Here a Yukawa coupling term is added to the Lagrangian depicting the coupling of the fermion fields to the Brout-Englert-Higgs field. These gauge invariant interaction terms are defined as

$$\mathcal{L}_{Yukawa} = -g_{Yukawa}\phi\bar{\psi}\psi, \tag{1.19}$$

where g_{Yukawa} is the coupling constant and the fermion mass equals

$$m_{fermion} = g_{Yukawa} v/\sqrt{2}. \tag{1.20}$$

Until the summer of 2012, this mechanism was believed to be responsible for generating particle masses yet no experimental evidence was at hand. Hence, the discovery of a new spin-0 boson by the CMS and ATLAS collaborations at the Large Hadron collider [12, 13] compatible with the Brout-Englert-Higgs boson is one of the major breakthroughs in the understanding of the Standard Model.

1.1.4 Extensions of the Standard Model

The Standard Model is by far one of the most tested theories in physics to date. Numerous precision measurements have been carried out to test its validity and so far everything agrees well. With the discovery of a scalar BEH-like boson in 2012, the foundation of the Standard Model became even stronger as now it is able to describe how particles get their mass. In contrast to its successes, the Standard Model has some theoretical and experimental shortcomings. A few of them will be discussed in the following.

- **Unification of all forces:** Although the electric and magnetic forces are unified in the theory of electromagnetism which in its turn is mixed with the weak force in the electroweak interaction, there exists no unification of the strong force with the latter. There is no strong argument from theory to have such unified theory, nevertheless it is expected that at some very large energy scale the forces are indeed unified and this unification breaks spontaneously at the energy scales where the forces appear to be distinct.

- **No description of gravity:** Apart from the lack of a complete unification of the three fundamental forces in the Standard Model, it can also not describe the gravitational force. This is one of the main issues since this prevents the Standard Model from becoming the unified theory explaining all phenomena in the universe.

- **The hierarchy problem:** One of the open problems in the Standard Model is the hierarchy problem. This problem relates to the large discrepancy between the electroweak energy scale, on the order of 10^2 GeV, and the Planck scale around 10^{19} GeV where gravity starts playing a crucial role. In this large gap between the two energy scales no new phenomena are predicted. Therefore the relatively light mass of the BEH boson induces very large corrections on the parameters of the Standard Model.

- **Most of our universe is unknown:** Only a small fraction of all matter in the universe is known and incorporated in the Standard Model. Recent observations from the Planck satellite [17] have shown that the known matter constitutes 4.9% of the universe while 26.8% is considered to be dark matter and the remainder 68.3% dark energy. Neither of the last two can be described by the Standard Model.

The above mentioned problems do not, however, suggest that the Standard Model is wrong or should be overruled. Given its experimental success and the level of precision achieved when making predictions it is considered to be incomplete. One of the possible extension theories of the Standard Model is SuperSymmetry [18]. This theory predicts the existence of super partners for each currently existing particle that are identical in any way except that the particle and its super partner differ in spin by 1/2. This entails that the super partner of a boson is a fermion and vice versa. The benefit of this extension is that it allows for a unification of the strong and electroweak forces and it incorporates a potential candidate to constitute dark matter: the neutralino. Moreover, due to the existence of a new range of super particles, the radiative corrections to the BEH boson mass can become smaller essentially solving the hierarchy problem.

To test models like SuperSymmetry, particle colliders like the Tevatron collider at Fermilab and the Large Hadron Collider at CERN are constructed to continuously probe higher and higher energy scales. These experiments along with numerous non-collider experiments, such as neutrino factories, hope to unravel the mysteries of the Standard Model.

1.2 The heaviest quark, the top quark

Among all particles in the Standard Model, the top quark is by far the most massive with a current world average mass of 173.20±0.87 GeV [19]. To produce a heavy quark like the top quark, a hadron collider needs to reach a multi-TeV centre-of-mass energy. The first collider that was able to produce them is the Tevatron $p - \bar{p}$ collider located at Fermilab nearby Chicago. At this collider, the top quark was discovered in 1995 by the CDF [20] and DØ [21] collaborations more than 20 years after the discovery of the bottom quark. Hence, the Large Hadron Collider at CERN is the second machine that can produce this heavy quark and the first that will do so in large numbers. This opens a wide area of top quark research at the LHC.

In Section 1.2.1, the production and decay characteristics of the top quark will be discussed. Furthermore, the state-of-the art measurements of the top quark mass and other key properties are outlined in Section 1.2.2. The implications for the Standard Model are discussed in Section 1.2.3 and the potential of using the top quarks as a calibration tool is discussed in Section 1.2.4.

1.2.1 Production and decay

Top quarks can be produced in proton-proton collisions at the LHC in two ways. First, the top quark can be produced through the electroweak interaction as a single top quark or anti-top quark. Secondly, the top quark can be produced in pairs of top and anti-top quarks through the strong interaction. The latter is the more dominant production mechanism and is the main focus of this thesis. The single-top production will not be further discussed.

In the proton-proton collisions delivered by the LHC, top quark pairs can be formed by an annihilation of two quarks, $q\bar{q} \to t\bar{t}$ or by fusion of two gluons, $gg \to t\bar{t}$. At the LHC, gluon fusion is the dominant mechanism for producing top-quark pairs.

One of the important characteristics of the $t\bar{t}$ production mechanism is the cross section. This is the effective area that governs the probability of an absorption or scattering event. In terms of particle physics, this observable is interpreted as the likelihood of a certain interaction between particles. The top quark pair cross section is known theoretically as complete next-to-next-to-leading order calculations are available [22, 23] with next-to-next-to-leading log corrections applied. This means that the cross section is theoretically known up to $\mathcal{O}(\alpha_s^4)$ with a precision of $^{+7.0}_{-7.8}\%$ at $\sqrt{s} = 7$ TeV and $^{+6.8}_{-7.6}\%$ at 8 TeV.

The precise measurement of the $t\bar{t}$ pair production cross section is important for two main reasons. First, this measurement provides a crucial benchmark for the QCD perturbative calculations. Secondly, the $t\bar{t}$ pair production process could be enhanced or suppressed by new physics processes producing top quark pairs. Hence, the cross section could be sensitive to the presence of new phenomena.

The $t\bar{t}$ pair production cross section has been measured both in $p\bar{p}$ collisions at the Tevatron collider and *pp* collisions at the LHC. Table 1.3 shows the latest result from the Tevatron ElectroWeak Working Group combining results from the CDF and DØ collaborations. The most precise results from the LHC are given as well. So far all measurements of this quantity agree well with the theoretical predictions.

Collider	$\sqrt{s}$ (TeV)	$\sigma_{t\bar{t}}^{obs}$ (pb)	$\sigma_{t\bar{t}}^{NNLO+NNLL}$ (pb)
Tevatron	1.96 ($p\bar{p}$)	7.65±0.42	$7.164^{+0.391}_{-0.475}$
Large Hadron Collider	7 (*pp*)	162.0±6.7	$172.0^{+12.1}_{-13.4}$
Large Hadron Collider	8 (*pp*)	227.0 ±15.2	$245.8^{+16.6}_{-18.7}$

Table 1.3: Most precise measurements of the $t\bar{t}$ pair production cross section at the Tevatron [24] and at the LHC [25, 26] compared with the NNLO+NNLL theoretical calculations [22, 23]. The measurements assume a top quark mass of 172.5 GeV

With a life time of the order of 10^{-25}s [30], about 20 times shorter than the typical timescale for the strong interaction, the top quark is the only quark in the Standard Model that can decay through the weak interaction. This makes the top quark very interesting to study as it is the only quark that can be directly accessed as a free quark in experiments.

The top quark is expected to decay into a W boson and a lighter down-type quark. Hence, the top quark can either decay to a W boson with a d quark, an s quark or a b quark. The branching ratio of these three cases are governed by the matrix elements of the unitary CKM matrix introduced in eq. 1.12. Measurements of the CKM matrix element $|V_{tb}| = 1.011^{+0.018}_{0.017}$ (stat+sys) [31] show that $|V_{tb}|$ is consistent with 1 and combined with the unitarity of the CKM matrix this results in the suppression of all decay modes except $t \to Wb$. As a consequence, the decay of a $t\bar{t}$ pair is given as

$$t\bar{t} \to W^+ b W^- \bar{b}.$$

The W boson from the top-quark decay will decay in 2/3 of the cases into a pair of light quarks. Subsequently, in 1/3 of the cases the W boson decays into a charged

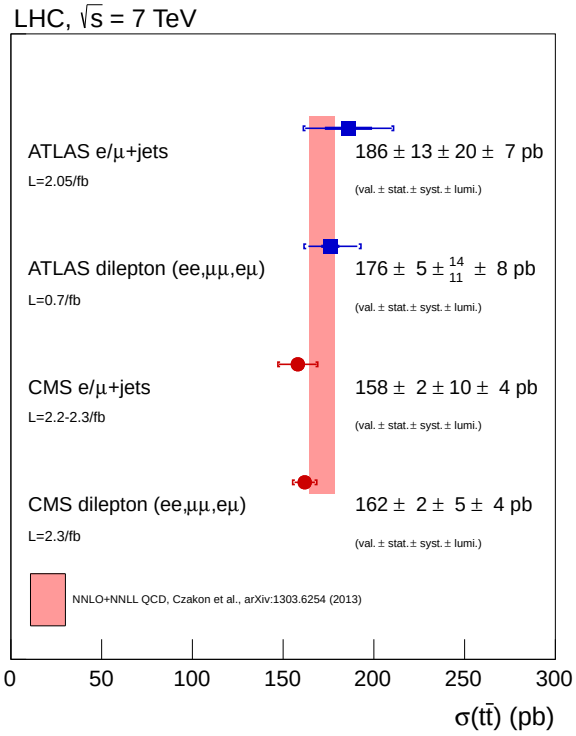


Figure 1.1: Measurement of the $t\bar{t}$ production cross section in the different decay channels at a centre-of-mass energy of 7 TeV [5, 25, 27, 28]. The measurements are compared to the full NNLO+NNLL theoretical calculation [22, 23].

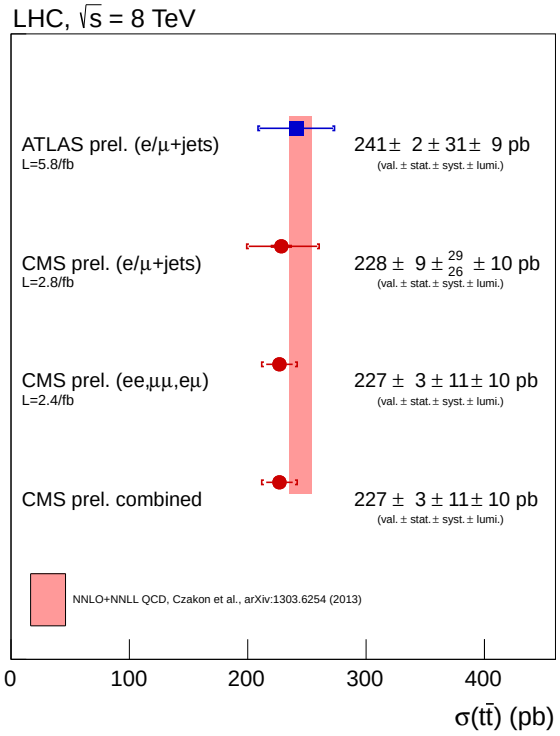


Figure 1.2: Measurement of the $t\bar{t}$ production cross section in the different decay channels at a centre-of-mass energy of 8 TeV [6, 26, 29]. The measurements are compared to the full NNLO+NNLL theoretical calculation [22, 23].

lepton and its associated neutrino. Since the $t\bar{t}$ decay contains two bosons in the final state, it can decay in three different modes. When both W bosons decay into quarks or into a lepton and a neutrino, the decay mode is called fully hadronic and fully leptonic respectively. Conversely, when one of the two W bosons decays into a lepton and a neutrino while the other decays into quarks, the decay is called semi-leptonic. The latter,

$$t\bar{t} \rightarrow W^+bW^-\bar{b} \rightarrow q\bar{q}bl\nu_l\bar{b},$$

will be extensively studied in this thesis and accounts for 14.8% of all top quark pair decays per lepton flavour l. The $t\bar{t}$ production cross section has been measured in each of these decay channels. The most precise results from the CMS and ATLAS experiments at the LHC are provided in Figure 1.1 for $\sqrt{s} = 7$ TeV and Figure 1.2 for $\sqrt{s} = 8$ TeV.

There exist some models where the top quark could decay differently. A search has been conducted for so-called Flavour Changing Neutral Current interaction (FCNC), i.e. a decay into a neutral Z boson and a quark. In the Standard Model such interaction is heavily suppressed so the observation of such interaction would hint new physics. However, the branching ratio of such process is found to be smaller than 0.21% at 95% C.L. [32]. Additionally, baryon number violating top quark decays were investigated [33]. Upper limits of 0.16% and 0.17% at 95% C.L. have been set on the branching ratios of $t \rightarrow bc\mu$ and $t \rightarrow bue$, respectively. Thus no deviation from the Standard Model prediction of the top quark decay characteristics have been found so far.

Also the decay width (Γ) of the top quark is sensitive to new physics effects. As the top quark almost exclusively decays into a W boson and a b quark, its decay width is expected to be dominated by the partial $t \rightarrow Wb$ width, $\Gamma(t \rightarrow Wb)$. According to the Heisenberg uncertainty principle $\tau = \hbar/\Gamma$, a decay width of 1.5 GeV for the top quark is predicted. The decay width is particularly interesting as it is very sensitive to new phenomena like anomalous Wtb couplings. A measurement carried out by DØ yields $\Gamma_t = 2.00^{+0.47}_{-0.43}\ GeV$ [30], consistent with expectations.

Next to the total $t\bar{t}$ cross section measurement, it can also be measured as a function of different kinematic variables of the top quark decay products. Both the CMS and ATLAS experiments have carried out such measurements [34–37]. These measurements allow to further benchmark detailed theoretical predictions to the data collected by the respective experiments. Figure 1.3 is an example of the measurement of the $t\bar{t}$ cross section as a function of the mass of the $t\bar{t}$ system. Another example is provided in Figure 1.4 where the cross section is given as a function of the pseudo rapidity[2] of the leading lepton. These figures show in general a very good agreement between the data and simulation.

1.2.2 Other top quark properties

Next to the production and decay of top quarks, some other very interesting properties can be measured. One of these properties is the top quark mass. The mass has been

[2]The pseudo rapidity of a particle is defined as $\eta = -\ln\left[\tan\left(\frac{\theta}{2}\right)\right]$ where θ is the angle between the momentum vector of the particle and the beam axis.

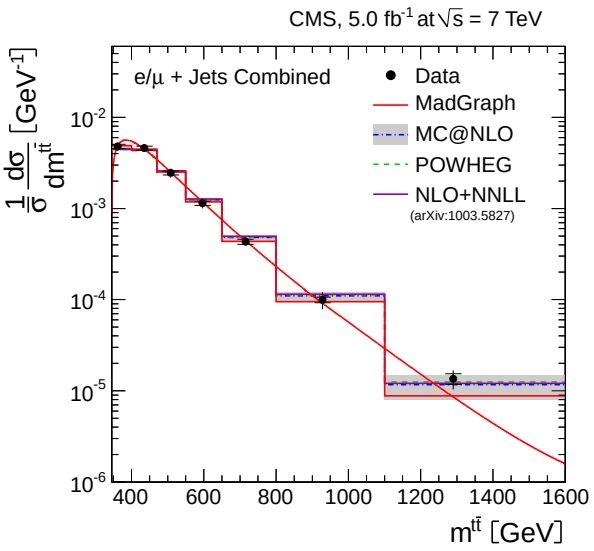


Figure 1.3: Differential $t\bar{t}$ cross section as a function of the mass of the $t\bar{t}$ system

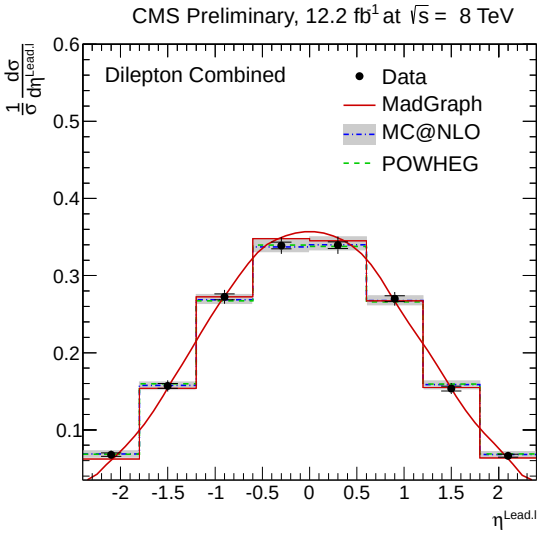


Figure 1.4: Differential $t\bar{t}$ cross section as a function of pseudo-rapidity of the leading lepton

measured at the Tevatron and at the Large Hadron Collider[19, 38–40]. Though the LHC produces larger top quark samples, the Tevatron still holds the most precise combined result of 173.20±0.87 GeV [19]. The most precise measurement for each experiment is shown in Figure 1.5.

The difference in mass between the top and anti-top quark has also been measured to great precision. This difference is sensitive to the CPT symmetry of the Standard Model that claims equal mass for a particle and its anti-particle. A mass difference of $\Delta m_t = -272 \pm 196 \pm 122$ MeV was measured [41] which is compatible with zero.

Due to the fact that the top quark has a shorter lifetime than the typical timescale of the strong interaction, it will decay and hence its spin information is transferred to its decay products. Hence the spin of the top and anti-top quarks can be reconstructed to check for correlation among them. This correlation is predicted by the Standard Model and has been observed at the LHC [42, 43]. The degree of correlation was found to be compatible to predictions. Next to that, the alignment of the top quark spin with its direction of movement, the top quark polarisation, has been measured as well [44, 45] and no deviation from the Standard Model has been observed.

Another measurement has been carried out along these lines, the measurement of the W-boson helicity fractions from top quark pair decays. This measurement has been carried out to carefully study the Wtb vertex and the helicity fractions were measured to be [46, 47]

$$F_0 = 0.626 \pm 0.034(stat) \pm 0.048(syst)$$

$$F_L = 0.359(stat)0.021 \pm 0.028(syst),$$

where F_0 is fraction of longitudinal polarisation and F_L is the fraction of left-handed polarisation. Using the requirement that all helicity fractions sum up to 1, the right-

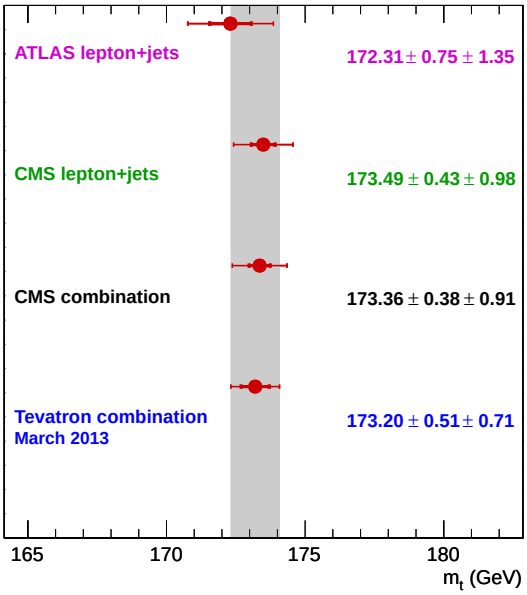


Figure 1.5: Most precise measurement of the top quark mass at the CMS [38] and ATLAS [39] experiments. The combination of the measurements of the DØ and CDF experiments at Tevatron is provided [19] as well as the combination of all mass measurements at CMS [40].

handed fraction F_R can be inferred from the other two and equals 0.015±0.034. These results are in good agreement with Standard Model predictions.

Finally, the top quark charge has been measured. Since the top quark decays into a W boson and a b quark, its charge can either equal -4/3 or +2/3 of the elementary charge. The Standard Model predicts the top quark charge to be +2/3 and this is confirmed by measurement ruling out the -4/3 charge hypothesis at 95% C.L. [48, 49].

1.2.3 Implications for the Standard Model

To check the consistency of the Standard Model to all available measurements, a global fit can be carried out. Since the discovery of a new boson at the LHC consistent with the Brout-Englert-Higgs boson, all unknown parameters of the Standard Model are measured. As the mass of this boson has been measured, the global fit can be carried out including this mass checking the overall consistency [50]. This fit yields a p-value of 0.07 for the compatibility of all available measurements with the Standard Model.

With this global fit, the consistency of the measurements of the W-boson and top-quark masses can be checked compared to all other available measurements. This is shown in Figure 1.6. The electroweak fit is repeated two times. The first fit excludes the measurement of the W-boson mass, the top-quark mass and the BEH boson mass (grey band). Then the fit is rerun including the BEH boson mass (blue band). Comparing the two bands shows clearly that the inclusion of the BEH boson mass improved the predictions significantly. The direct measurement (black marker) shows reasonable agreement with the indirect prediction but it is clear that one can still improve this consistency check with a more precise measurement of both the top quark and W boson

mass.

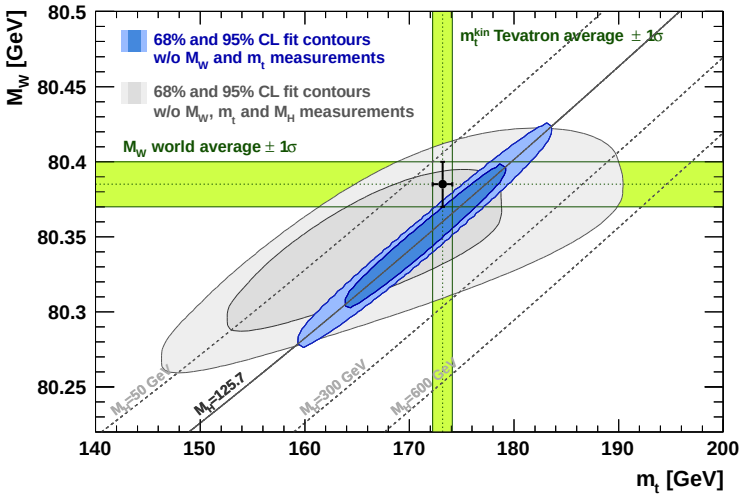


Figure 1.6: Global Standard Model fit to electroweak precision data to check the consistency of the measured W-boson and top-quark masses with all other measurements [50]

1.2.4 The top quark as a calibration tool

The top quark is interesting as it allows to further explore the Standard Model as was shown in the previous section. Next to that, the top quark exhibits some interesting properties that can help in calibrating the reconstructed object in the experiment. Large clean samples of top quark pair events can be constructed. These large statistics samples can be used to infer corrections to detector and reconstruction effects. Since the top quark contains a b quark in its decay, one of the potential areas where top quarks can play a crucial role in the performance measurement of the identification algorithms for b quarks in the experiments, or so-called b-tagging [1, 2]. This is one of the key topics of this thesis as Chapters 5 and 6 provide a detailed discussion on a method to measure the full b-tagging performance using $t\bar{t}$ events.

Chapter 2

The Large Hadron Collider and the CMS experiment

In the previous Chapter, the Standard Model was introduced along with its shortcomings. Theories extending the Standard Model have been developed, but to ultimately prove their validity, they have to be confirmed by experiment. For this reason, high energy particle colliders were built. The Large Hadron Collider (LHC) [51] is currently the largest and most energetic collider in the world giving scientists the tools to pursue a better understanding of the Standard Model and beyond.

In Section 2.1, the Large Hadron Collider will be introduced and some of its main design characteristics will be highlighted. All the experiments located at the LHC are briefly discussed. In Section 2.2 the Compact Muon Solenoid (CMS) experiment [52], playing a crucial role in this thesis, will be discussed in more detail and its main components will be outlined. To finish, a brief overview of the trigger system and computing infrastructure is given.

2.1 The Large Hadron Collider

The Large Hadron collider is currently the most energetic particle collider in the world. It is designed to accelerate proton beams up to 7 TeV and steer them around its 26.7 km ring and force them to collide. The 14 TeV collisions the LHC can generate put it at the energy frontier as they are 7 times more energetic than the proton-anti-proton collisions at 1.96 TeV provided by the Tevatron collider [53].

The LHC collider has been constructed by the CERN laboratory (European Organization for Nuclear Research) near Geneva, Switserland. The 26.7 km tunnel, home to the LHC, was excavated in the 1980's to install the Large Electron Position collider (LEP). After the shutdown of LEP in 2000, construction for the LHC began in the same tunnel and after a commissioning phase in 2009, the collider came in operation in March 2010 with the first ever 7 TeV centre-of-mass energy collisions at colliders.

2.1.1 Design of the collider

The design of the LHC machine is mainly determined by two factors: the physics reach and the constraints of reusing the old LEP tunnel. Since the physics goals requires to probe the TeV energy scale, the collider needs to produce collisions with sufficient centre-of-mass energy. Since the LHC increased the collision energy by a factor of 7 with respect to the Tevatron $p\bar{p}$ collider, it is very suitable to probe this energy scale. Another important accelerator parameter defining the physics potential is the luminosity.

The luminosity, $\mathcal{L}$, is important because it defines the number of events we can observe for any given process. The number of events for a given process, N, is defined as

$$N = \mathcal{L}\sigma, \tag{2.1}$$

where σ is the cross-section for the given process. Consequently, to search for very rare phenomena, phenomena with a very small cross-section, the luminosity needs to be maximised to be able to observe as many of those events as possible. At the LHC, a luminosity of $10^{34}\ cm^{-2}\ s^{-1}$ will be attained at a beam energy of 7 TeV which is roughly 100 times larger than the luminosity of the Tevatron. This provides a large physics potential for the LHC.

The decision to build the LHC in the tunnel excavated for its predecessor was mainly made for budgetary purposes. Hence the LHC had to be designed to suit its physics requirements as well as cope with the dimensions of the LEP tunnel. Compared to the leptons at LEP, the LHC protons carry 70 times more energy. Since electrons loose a fraction ΔE of their energy each orbit through synchrotron radiation, defined in eq. (2.2), the beams have a stringent upper limit on their energy.

$$\Delta E \propto \frac{E^4}{Rm^4} \tag{2.2}$$

In eq. (2.2) the radius of the accelerator ring is denoted as R, the particle mass as m and the beam energy as E. It is apparent that for a factor 70 increase in particle energy, the accelerator radius should increase by 70^4 which is impossible. On the contrary, the increase in energy can be buffered by the choice of beam particle. The best candidate was the proton since its mass is roughly 2000 times larger than the electron mass hence reducing ΔE by a factor of $1/2000^4$ compared to an electron accelerator.

The most practical choice for a hadron collider would be to collide protons with anti-protons, similar to the Tevatron, where the counter-clockwise rotating beams could be steered by the same magnet system. Unfortunately, the design luminosity at the LHC requires the beam to consist of proton bunches containing up to 10^{11} protons each which is beyond the reach of any anti-proton production system. Hence, the machine was built to collide two counter-rotating proton beams.

The difficulty of having two counter-rotating beams with particles of the same charge is that they need two opposite magnetic fields to be guided along the accelerator. Since the LEP tunnel was not wide enough to host two separate accelerators, one per beam, a special twin-bore magnet system was designed to generate the dipole field necessary to steer the beams. The cross section of such superconducting dipole magnet is shown in Figure 2.1 where two coils are integrated in one cryostat. To steer the

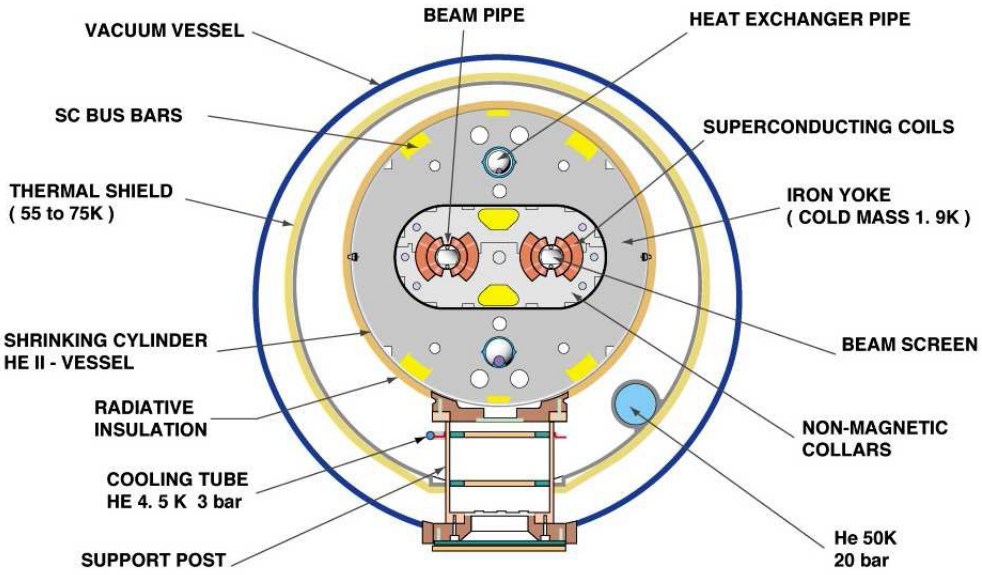


Figure 2.1: The cross section of a superconducting LHC dipole

7 TeV proton beams around the LHC, a magnetic field of 8.33 T is required. Such high magnetic field can only be produced using superconducting coils operating at a temperature near 1.9K. The LHC consists of 1232 of these 15m long dipole magnets weighing over 27 tons each.

In addition to the dipole magnets, a number of higher order fields are used near the interaction points to focus the beam before entering collision. The squeezing of the beams before colliding them helps in attaining the high luminosity required.

Before the proton beams can be injected into the LHC machine, they need to have an initial energy of 450 GeV. This pre-acceleration is performed by multiple particle accelerators in the CERN accelerator complex (Figure 2.2). First, protons are collected by stripping the electrons from hydrogen atoms. Then, the protons enter the first accelerator in the chain which is the LINAC2, a linear accelerator. This machine accelerates the protons to an energy of 50 MeV after which they are injected in the BOOSTER. The latter stacks up four accelerators where the protons are squeezed into bunches and further accelerated to 1.4 GeV. Then the proton bunches enter the Proton Synchrotron (PS) where the beam get s its bunch structure where every bunch is spaced by 25 ns from the previous. The protons also get accelerated to 26 GeV. Finally, the acceleration to 450 GeV is performed by the 6km long Super Proton Synchrotron and the bunches are ultimately injected into the LHC.

When filling the LHC machine, 2808 bunches can be injected spaced by 25 ns. When the bunches are circulating in the machine, the injection is halted and the acceleration of the beam starts. When the beam reaches its design energy of 7 TeV, it is squeezed and adjusted to have the most optimal conditions for colliding the beams in the interaction points. The beam will typically remain in the machine to provide collisions for the next 10-15 hours. When the luminosity of the beams is to degraded, the decision is made to dump them and refill the machine afterwards.

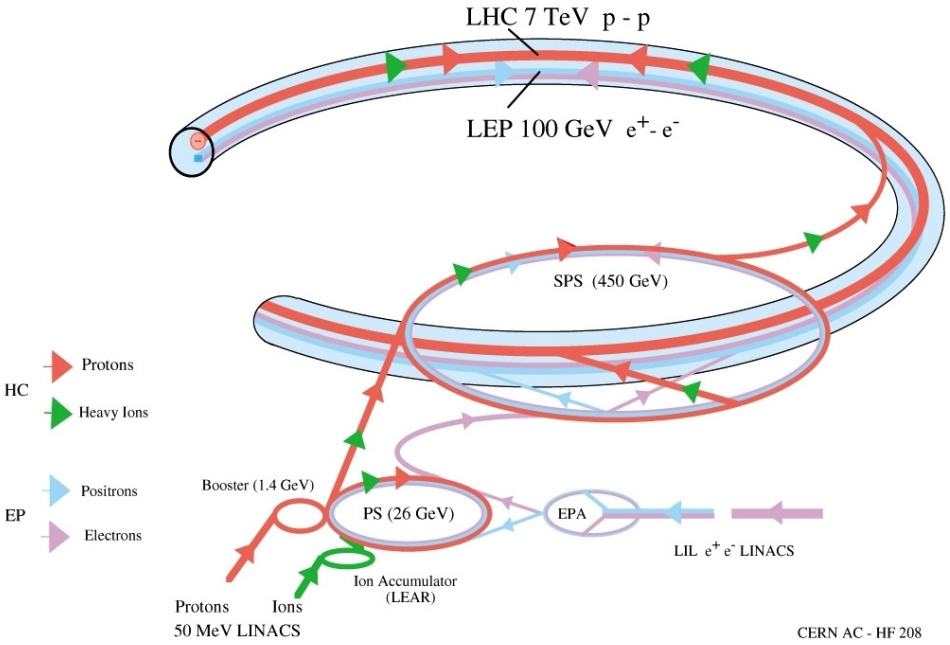


Figure 2.2: The injection chain of the LHC accelerator complex.

2.1.2 The experiments at the LHC

Along the Large Hadron Collider, four interaction regions have been implemented. In these interaction regions, the curved shape of the accelerator is changed to a straight section where the protons of both beams are put on a collision course. Around the four interaction points, large particle detectors have been constructed as shown in Figure 2.3.

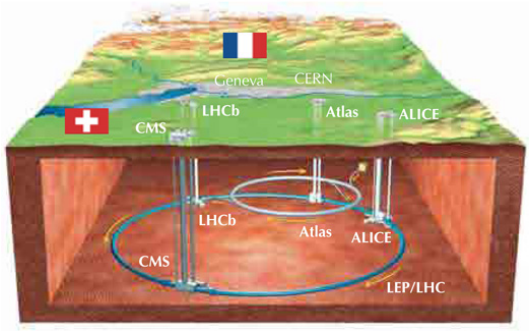


Figure 2.3: Schematic overview of the LHC accelerator and the location of the four main experiments

Two very big general purpose experiments are installed at the Large Hadron Collider, the CMS [52] and the ATLAS [54] experiments. These experiments are often denoted as general purpose experiments since their physics programme is very wide. These detectors are designed to be usable for both the search for new phenomena be-

yond the Standard Model as well as to conduct very precise measurements on already known particles such as the top quark.

Additional to these very huge general purpose detectors, two smaller and very specific experiments were installed. The first is the ALICE experiment [55]. This experiment mainly focuses on the study of the quark-gluon plasma through analysis of heavy-ion collisions that are delivered by the LHC inbetween proton run-periods. The second is the LHCb experiment [56] which is designed to perform precision measurements in the b-quark sector.

Next to the four experiments installed around the interaction points, a detector to measure the total proton-proton cross section and elastic proton scattering is deployed. This detector is named TOTEM [57] and is installed close to the CMS interaction region. Finally, the LHCf experiment [58] measures the particles created in the very forward region of the proton-proton collisions attempting to improve the understanding of ultra-high energetic cosmic rays.

2.1.3 The LHC run periods

The first collisions at 7 TeV centre-of-mass measurement were delivered by the LHC in March 2010 after a long commissioning period starting from September 2008 when a technical failure in the accelerator severely damaged a number of magnets. During the 2010 run, the LHC produced a data sample of 45 pb^{-1} with a peak luminosity of around 2 10^{32} cm^{-2} s^{-1}. Although the peak luminosity is a factor 50 smaller than the design specification, the collisions were the most energetic ever opening the energy frontier at the LHC.

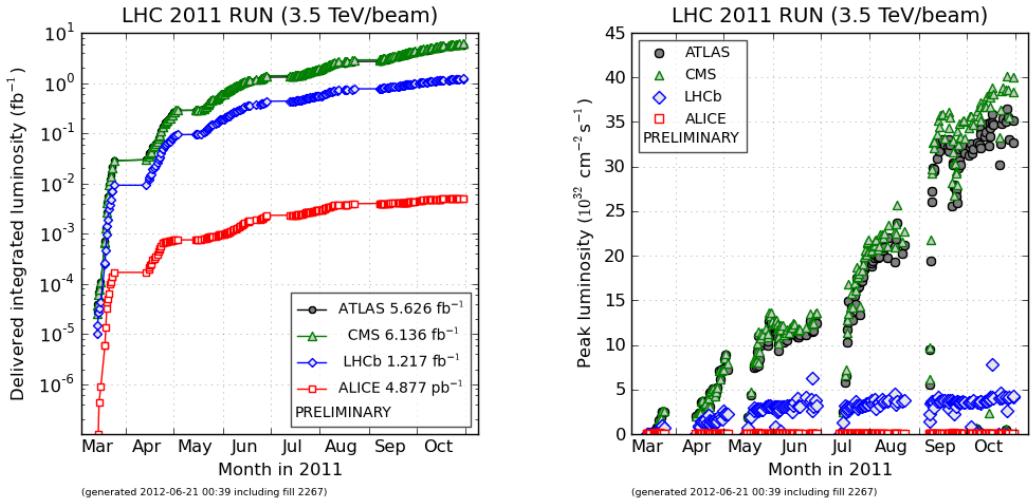


Figure 2.4: Integrated luminosity and peak luminosity as a function of time during the 2011 operation of the LHC at $\sqrt{s} = 7$ TeV [59].

During the first long run in 2011, the machine provided 7 TeV centre-of-mass energy at a peak luminosity of 4 10^{33} cm^{-2} s^{-1} as shown in the right canvas of Figure 2.4.

This provided a total integrated luminosity delivered to the experiments of 6.136 fb^{-1} which is $\mathcal{O}(10^2)$ larger than during the 2010 run. Also the peak luminosity increased by a factor 20. This was up to that point in time the highest reached peak luminosity for a particle collider.

In 2012, the LHC beam energy was increased from 3.5 TeV per beam to 4 TeV resulting in 8 TeV centre-of-mass collisions. Continuing to build on the successful 2011 run, the peak luminosity delivered to the experiments increased again by a factor of almost 2 with respect to the 2011 run to little under 8 10^{33} cm^{-2} s^{-1} which comes very close to the design luminosity. During this run period, the LHC delivered a dataset of 23.269 fb^{-1} proton-proton collisions at 8 TeV as can be seen from Figure 2.5. Currently, the LHC is going through its first 2-year shutdown period gearing up for 13 TeV operation in 2015.

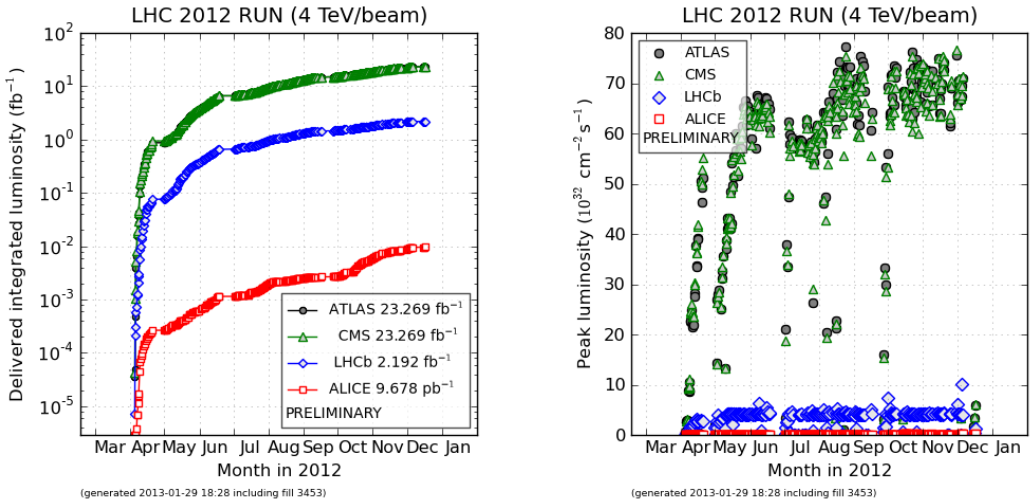


Figure 2.5: Integrated luminosity and peak luminosity as a function of time during the 2012 operation of the LHC at $\sqrt{s} = 8$ TeV [59].

The data sample generated during the 2011 and 2012 runs will be studied in this thesis. Due to the high luminosity and the large number of protons in each bunch, additional interactions can appear in the experiment next to the central hard interaction, called pileup interactions. The average number of proton-proton interactions per bunch crossing averages to 21 [60] as shown in Figure 2.6 where the distribution of the number of interactions per bunch crossing is shown. In this high-pileup environment it is difficult to determine which particle comes from which pp collision.

2.2 The Compact Muon Solenoid experiment

The Compact Muon Solenoid experiment (CMS) is one of the four main experiments reconstructing the proton-proton collisions delivered by the Large Hadron Collider. Additionally, it is one of the two very large general purpose experiments. Together with the ATLAS experiment, its target is further scrutinizing the Standard Model by doing

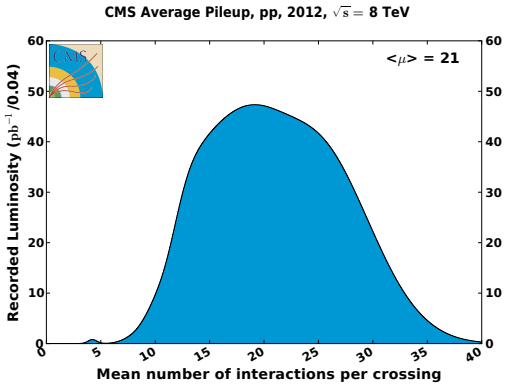


Figure 2.6: Distribution of the average number of pp collisions per crossing of the LHC beams during the 2012 operation at $\sqrt{s} = 8$ TeV [60]. On average, each beam crossing results in 21 pp collisions.

precision measurements but also by searching for new phenomena like SuperSymmetry or the search for the elusive Brout-Englert-Higgs boson. The discovery of a boson very compatible with the latter [12, 13] has been by far one of the largest successes of both CMS and ATLAS but also of the LHC.

This section will explain in more detail how the CMS detector was designed and which technologies allow it to carefully reconstruct the collisions. Furthermore, a brief overview will be given of the trigger system that allows CMS to reduce the vast amount of $\approx 10^9$ collisions per second to a manageable rate. Finally, the computing infrastructure will be briefly introduced that is used by scientists around the world to analyse the data.

2.2.1 Overview of the detector systems

The CMS experiment was designed according to a traditional multi-layered approach typical to collider experiments as shown in Figure 2.7. The CMS experiment has a cylindrical shape of 15 m diameter and is about 22 m long. Although the compact CMS detector is significantly smaller than the very large ATLAS detector. The weight of the CMS experiment totals to an astonishing 15 thousand tons where ATLAS throws in about 5 thousand tons less.

The largest fraction of the weight is located in the iron return yoke which is constructed around the very large superconducting solenoid magnet. The magnet delivers a field of 3.8 T to bend the high energetic particles originating from the collisions to allow their momentum to be measured. Interleaved with the return yoke, muon chambers (cfr. Section 2.2.5) are installed as the most outwards detector.

Inside the bore of the magnet, there is room to host inner detector layers. The detector system closest to the beam axis and the collision point is the inner tracking detector (cfr. Section 2.2.3). This high precision detector allows to reconstruct charged particle trajectories as well as to measure their momentum in the magnetic field. Exiting the tracking detector, particles enter into the calorimeter system, explained in

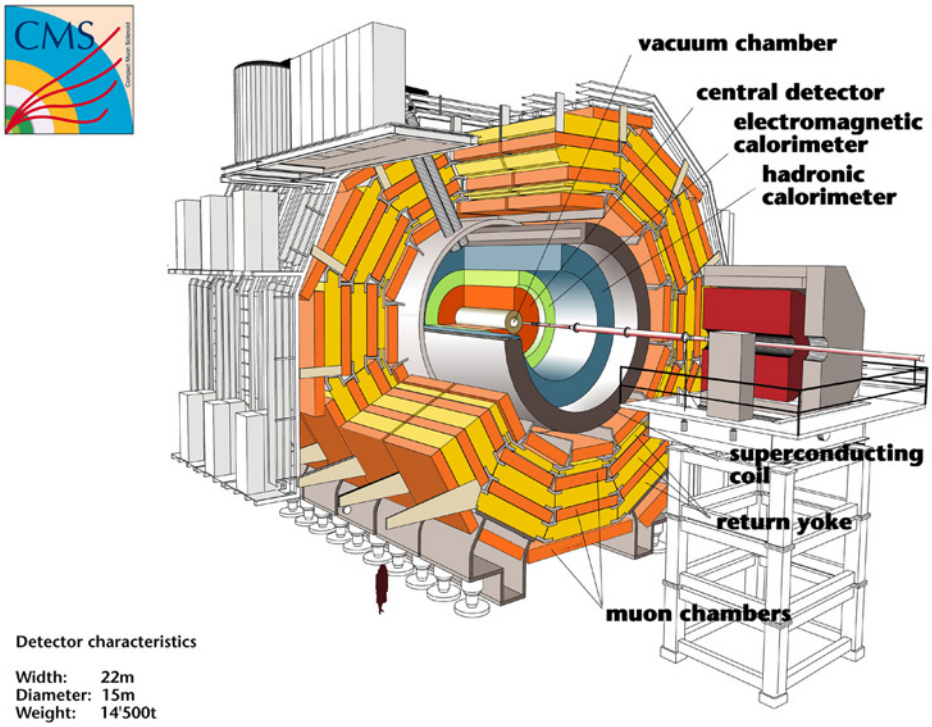


Figure 2.7: Schematic breakdown of the full CMS detector.

Section 2.2.4, where their energy is measured either in the electromagnetic calorimeter for electrons[1] and photons or in the hadronic calorimeter for hadrons.

Finally an endcap is installed on both sides of the central cylinder, to ensure an almost 4π coverage.

2.2.2 The CMS coordinate system

To describe a certain location in the detector, a common CMS coordinate system has been put in place. At the origin of this coordinate system sits the interaction point where the collisions take place. The x-axis points from the origin towards the centre of the LHC while the y-axis is chosen to point upwards to the ground surface. Finally, the z-axis is set perpendicular to the x and y-axis to form a right-handed coordinate system.

To characterise the location of detector material, and even particles, in the detector the azimuthal angle ϕ and the polar angle θ and the radial coordinate r are used. The azimuthal angle is defined as the angle between the object and the x-axis in the (x,y) plane and ranges from 0 to 2π. The polar angle is measured from the z-axis and takes values between 0 and π. The polar angle is most often translated into the pseudorapidity given by

$$\eta = -ln\left(tan\frac{\theta}{2}\right), \tag{2.3}$$

[1] Also for positrons

such that differences in pseudorapidity are invariant under Lorentz boosts along the beam direction.

2.2.3 The tracking detector

Charged particle tracks can be reconstructed using the inner tracking detector of CMS. This cylinder shaped detector is installed closest to the collision point and is 5.8m long and 2.6m in diameter. The tracker is immersed in a magnetic field of 3.8T allowing it to accurately measure charged particle momenta through their bending in the magnetic field. The tracking detector consists of two main detector technologies: the pixel detector and the silicon strip detector.

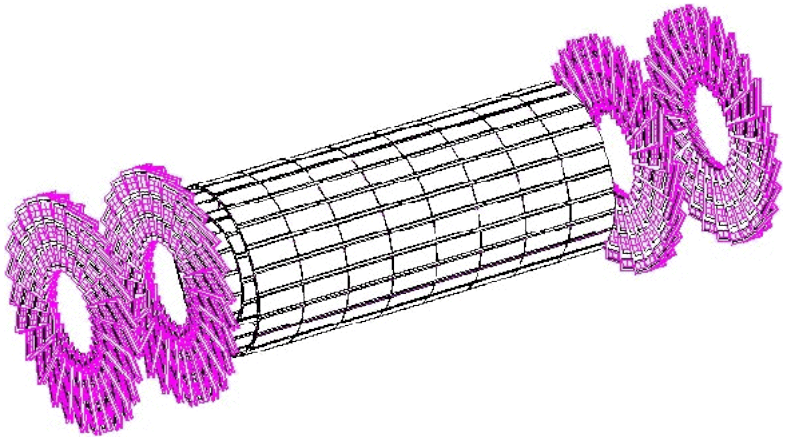


Figure 2.8: Schematic overview of the CMS pixel detector.

The pixel detector is displayed in Figure 2.8 and sits at a radius of less than 10 cm from the beam axis. The pixel detector consists of three layers of 100 x 150 μm^2 pixel cells that are installed respectively at a radius of 4.4 cm, 7.3 cm and 10.2 cm. The three layers are complemented by two disks located on both sides of the interaction point at z=34.5 cm and 46.5 cm. These disks have pixel cells extending from a radius of 6 cm to 15 cm from the beam axis. The pixel detector contains a total of 66 million pixel cells allowing for a single hit resolution of 15-20 μm in this high-activity area.

Outside the pixel detector, the particle density gets low enough to be able to use a silicon strip detector rather than pixel cells. The silicon strip detector, schematically displayed in Figure 2.9 consists of 9.3 million silicon strip sensors and ranges from 20 to 116 cm in radius. In the central barrel of the detector, the strip tracker is divided into two subsystems. The Tracker Inner Barrel (TIB) consists of 4 layers covering up to $|z| < 65$ cm. At each side of the TIB it is complemented with three additional disks, the Tracker Inner Disks (TID). The Tracker Outer Barrel (TOB) is located outside the TIB and consists of six layers of silicon strip detectors in a range of $|z| < 118$ cm. On each side of the barrel, nine disks are installed in the region $124 < |z| < 282$ cm at a radius of 22.5 to 113.5 cm. This system is called the Tracker Endcap (TEC). The silicon strip tracker provides coverage up to $|\eta| < 2.5$ and provides a single hit resolution in the TIB of 23-35 μm on the $r - \phi$ measurement and 230 μm for the z-direction. For the TOB, the respective single hit resolutions are 35-53 μm and 530 μm.

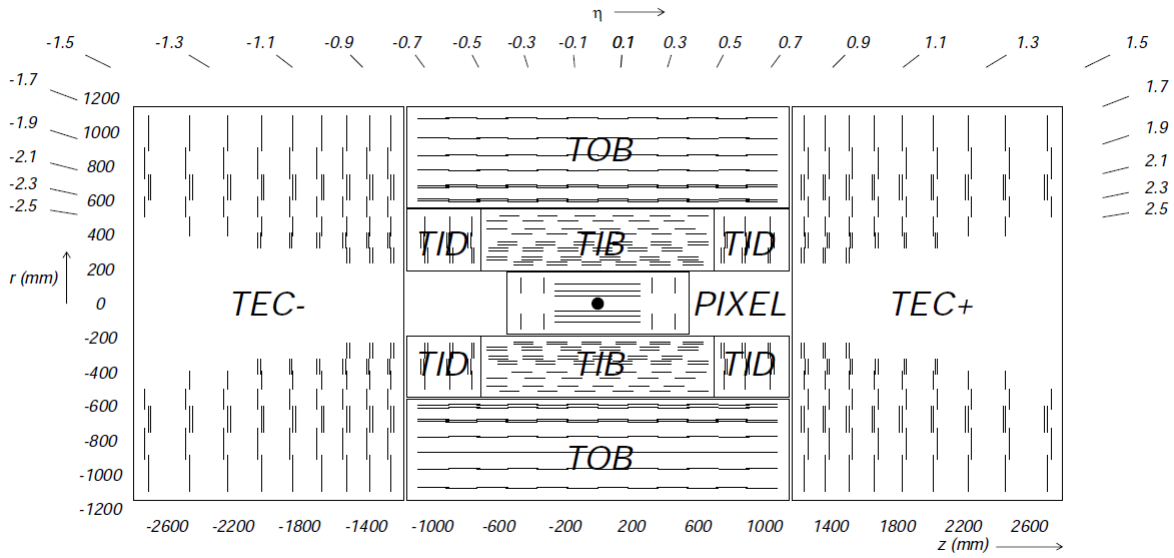


Figure 2.9: Schematic overview of the CMS tracking detector.

Track reconstruction

In both the pixel and silicon strip detectors, the position of a charged particle traversing the layer is recorded along with its uncertainty. These hits can then be used to reconstruct the actual track of a particle followed while traversing the full tracking detector in the magnetic field. Reconstructing tracks from the tracking detector hits can be done in four steps. First the track seeds are generated. These are then used to build tracks. Possible ambiguities are removed and the final track fit is performed [61].

The seeds are generated from the reconstructed hits in the tracker and the seeds need to consist of at least three hits in the tracker or two hits with a beam constraint. The best seeds are provided by the pixel detector as it has the best position resolution. Nevertheless, inclusion of strip tracker seeds allows for improvement of the overall track reconstruction efficiency. These seeds provide initial trajectory candidates that are now passed to the track builder.

The charged particle tracks are reconstructed using a Kahlman filter. The tracks building starts from the initial trajectory candidate provided by the seed generating step and starts to extrapolate them outwards. Compatible hits are identified based on the extrapolation towards each compatible layer using the equations of motion of a charged particle in a magnetic field. The effects of multiple scattering are taken into account in this extrapolation. The compatible hits are added to the trajectory candidates to form a new candidate for which the track fit is repeated. This iterative process is continued until the trajectory candidate is extrapolated to the final layer of the tracking detector. To reduce the collection of trajectory candidates, the trajectories not fulfilling a cut on the normalised χ^2 of their fit are removed as well as tracks with to few associated hits.

In the final list of trajectory candidates, two or more trajectories could share tracker hits as well as seeds. These ambiguities have to be removed before the global track fits

are performed. To do this, tracks sharing too many hits are discarded from the list only retaining the track with the most hits.

Finally, the trajectory candidate list is obtained and all ambiguities are gone. To ensure an optimal determination of the track parameters, the track is refit completely using a least-squares minimisation taking into account all the tracker hits assigned to the track candidate. Initially, the track parameters are propagated outwards from the beam line to the calorimeter surface. Subsequently, the track fit is carried out in the reverse direction starting from the outermost tracker hit. To retrieve the trajectory information at the impact point, the track can be extrapolated from the first layer containing a tracker hit to the interaction region.

Vertex reconstruction

When all the tracks are reconstructed they can be used to reconstruct the vertex. This reconstruction uses a technique called Deterministic Annealing (DA) [62] to cluster the tracks into a vertex candidate. The method assigns each track to a vertex candidate and then minimises a global χ^2 where the candidates are weighed to account for their compatibility to the tracks. A temperature parameter T is introduced to control the assignment of tracks to vertices. Starting at infinite temperature, all weights are equal and only one vertex candidate is available. Then the temperature gets decreased incrementally where in each step the prototype is split in two. If all tracks are not compatible to one vertex candidate, the process continues potentially further splitting the vertex candidates. The process is continued until a minimum temperature is reached. Vertices with less than two tracks assigned will be removed from the list to remove fake vertices from track outliers. The track list is then fitted again with the Adaptive Vertex Fitter [63] to get the vertex position in three dimensions. This fitter weights all tracks based on their distance to the primary vertex to further reduce the effect of charged particles originating from the decay long-lived particles.

2.2.4 The calorimeters

The CMS calorimeter is built up in two different subsystems. The first, the electromagnetic calorimeter, is a nearly hermetic and homogeneous calorimeter designed to measure the energy for electromagnetically interacting particles. Just outside the electromagnetic calorimeter sits the hadronic calorimeter to measure energy deposits from neutral and charged hadrons. Since these particles can traverse more material, the hadronic calorimeter is designed as a sampling calorimeter interchanging layers of sensitive material with brass to provide the necessary stopping power.

Electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) of CMS, shown in Figure 2.10, is a homogeneous detector constructed from lead tungstate ($PbWO_4$) crystals. Due to the relatively short radiation length (X_0) of these crystals, the ECAL can stop high energetic electrons while still remaining relatively compact. This is important for fitting this detector into the solenoid bore. Furthermore, 80% of the light is emitted within a 25

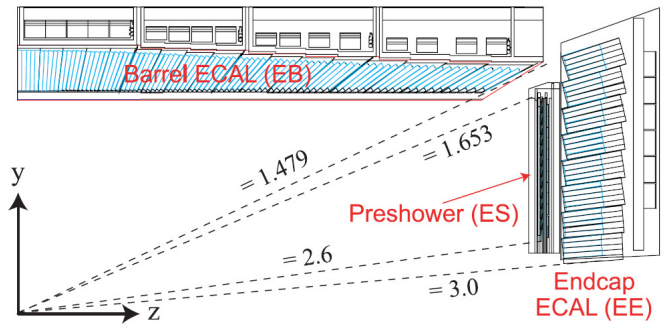


Figure 2.10: Detailed view of a quarter of the ECAL calorimeter in the CMS detector.

ns time window making it ideal for LHC operation. These properties together with its radiation hardness make the crystals a good choice for the ECAL.

The calorimeter is segmented into three parts, the ECAL Barrel (EB), the Endcaps (EE) and the Preshower (ES) . The ECAL Barrel detector has an inner radius of 129 cm and contains about 61200 crystals. These crystals cover an $|\eta|$ range up to 1.479 and are mounted such that their surface is facing the interaction region though they are tilted by 3° to reduce the effect of particles crossing the boundary in between adjacent crystals. The crystals cover a surface of 22 x 22 mm^2 with a length of 230 mm corresponding to 25.8 radiation lengths.

Additionally, both endcaps contain 14648 crystals with a surface size of 28.6 x 28.6 mm^2 and a length of 220 mm. This system extends the ECAL coverage to a pseudorapidity range of $1.479 < |\eta| < 3.0$.

The last part of the ECAL is the Preshower detector located in front of the endcaps in the region $1.653 < |\eta| < 2.6$. This detector consists of lead radiators to initiate electromagnetic cascades interleaved with silicon sensors of 63 x 63 mm^2. This detector allows to identify neutral pions in the endcaps and helps improving the identification and position resolution of electrons in the endcaps using its high granularity.

The ECAL provides an energy resolution (σ_E/E) of 1-2% for $|\eta| < 1.1$ increasing to 4% when going up in pseudorapidity [64].

Hadronic calorimeter

The hadronic calorimeter (HCAL) is designed to measure the energy of hadrons and it plays a crucial role in the measurement of the missing transverse energy. The latter is the energy that is attributed to very weakly interacting particles such as neutrinos[2]. Unlike the ECAL, the HCAL is a sampling calorimeter consisting of around 7000 plastic scintillator surfaces combined with layers of brass absorbers.

The HCAL Barrel (HB) has to fit inbetween the ECAL and the magnet. Since that does not leave much room for the HCAL, as much absorbing power needs to be positioned inside the magnet region reducing the space for sensitive detector material. The HB covers a pseudorapidity range up to $|\eta| < 1.3$ with calorimeter towers of 0.087 x 0.087 in (η, ϕ) coordinates. The segmentation of the HCAL is shown in Figure 2.11.

[2]The concept of missing transverse energy will be further discusses in the next chapter.

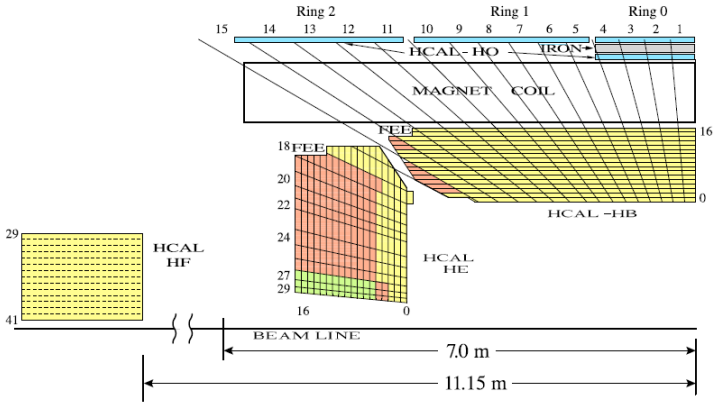


Figure 2.11: Detailed view of a quarter of the HCAL calorimeter in the CMS detector.

Given the limited space for the HCAL inside the magnet coil, hadronic shower leakage can occur. As a consequence, an extra layer is added just outside the magnet called the HCAL Outer detector (HO). This system covers -1.26 $< \eta <$ 1.26 and uses the same segmentation of HB, albeit with iron absorbers instead of brass.

The HCAL Endcaps (HE) on both sides of the barrel have a range $1.3 < |\eta| < 3.0$ partially overlapping with the barrel detector. The tower size in HE is the same as for HB up to $|\eta| < 1.74$. Beyond this region, the tower sizes are incrementally enlarged to a maximal size $\Delta\eta \times \Delta\phi$ size of 0.350×0.174.

Finally, this system is complemented with a forward hadron calorimeter (HF) inserted at 11.2 m from the interaction region along the z-axis. This calorimeter allows to measure hadronic activity in the very forward region upgrading the HCAL coverage to $|\eta| < 5.0$.

The resolution of the HCAL system was determined in a test-beam with single pions crossing the prototype detector [61]. The single-pion resolution for ECAL+HB was found to be 20-30% below 50 GeV improving to less than 10% for 300 GeV particles.

2.2.5 The muon system

The muon system is the outer most subdetector of CMS. The muon system ensures a very precise reconstruction of the muons complementing the inner tracker and the calorimeter. The layout of a slice of the system is provided in Figure 2.12 where it can be seen that the system comprises of three different detector technologies. The choice of the detector technologies is mainly determined by the different operating environments in the detector.

In the barrel region up to $|\eta| < 1.2$, the muon detection layers are interleaved with the iron return yoke responsible for closing the magnetic field lines of the solenoid. Hence, the residual magnetic field at the position of the muon chambers is small. Combined with a low muon flux this far out from the beam axis, Drift Tubes (DT) can be used. These DTs have a very good single-hit spatial resolution of around 100 μm and an angular resolution of 1 mrad.

Conversely, the magnetic field and muon flux in the forward region are larger.

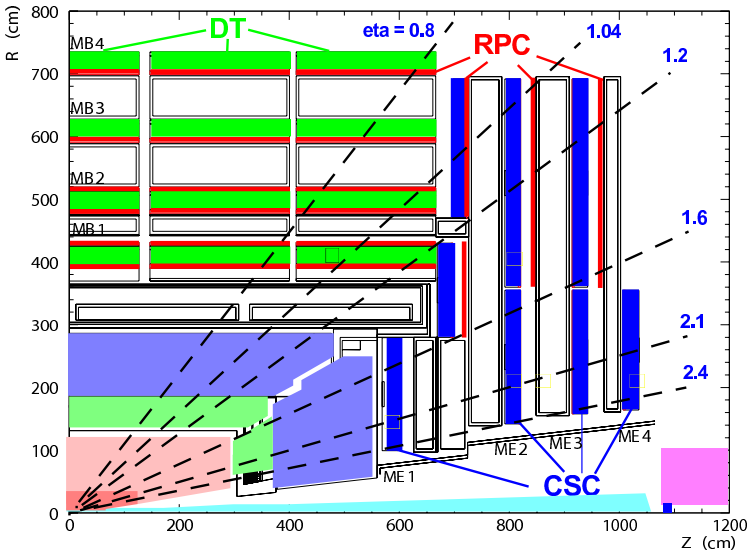


Figure 2.12: Longitudinal view of a slice of the CMS muon system showing the three different detection technologies used and their coverage.

Hence, Cathode Strip Chambers (CSC) are used in the endcaps covering a pseudorapidity ($|\eta|$) range of 0.9 to 2.4. These chambers also exhibit a good single-hit position resolution of typically 200 μm with an angular resolution of the order of 10 mrad.

Both CSCs and DTs are accompanied by Resistive Plate Chambers (RPC). These detectors are very fast and provide a very good time resolution of the order of 1 ns albeit with a more crude position resolution. The function of the RPCs is mainly to trigger on muons as well as to determine the correct bunch crossing in which the muon was produced. The first two barrel layers of the DT system are sandwiched between two layers of RPCs while the other DT layers have one accompanying RPC layer. The RPCs accompany the CSCs in the endcaps as well but only up to $|\eta| < 1.6$.

2.2.6 The trigger system

During nominal operation at design luminosity the LHC will deliver up to 10^9 inelastic collisions per second. With an average single event size of the order of 1 MB, it is clear that no storage system available to date can cope with such high rate of incoming data let alone store it. Therefore a trigger system needs to be implemented taking a decision on each event to be stored on tape. The decision is made in two steps. The first selection is made by the very fast Level-1 trigger. Events passing this step are fed to the High Level Trigger that makes the final call.

Level-1 trigger (L1)

As the LHC delivers collisions every 25 ns, a trigger decision has to be made at this rate. Since this time frame is very short to make a full reconstruction of the detector, a specialised electronics trigger was installed in the Underground Service Cavern (USC) adjacent to the experimental hall. Since 25 ns is very short even for specialised elec-

tronics, a buffer memory is installed next to the detector to host 128 collisions. This leaves the L1 trigger just over 3 μs to make a decision for a particular event. Albeit that about 2 μs are already used for the transportation of the information to and from the trigger system.

To cope with this short timescale, the L1 trigger uses very fast algorithms to make a decision based on crude information from the muon stations and the calorimeters. The L1 trigger has a maximal output of 100 kHz.

High Level Trigger (HLT)

Once the events passed the L1 trigger, they are shipped to the surface buildings where the High Level Trigger farm (HLT) is located. Since the event rate dropped by a factor 10^4 thanks to the L1 trigger, it is now possible to work with commercial CPUs as the decision time window is now of the order of seconds.

To decide if an event should be kept and stored on tape, the HLT has access to the fully reconstructed event and uses complex algorithms, similar to the ones used in offline analysis, to make the decision.

2.2.7 The LHC computing grid

To make all the collected data available to scientists worldwide, the LHC has a special distributed computing system: the Worldwide Large Hadron Collider GRID (WLCG) [65]. This system foresees both massive storage capacity throughout the world but also the necessary computing power to process all the data and simulation. The WLCG groups together GRID systems from smaller geographical areas like the EGEE (Europe) and the OSG (United states). The GRID consists of a hierarchical tiered structure.

Tier-0

The Tier-0 centre is located at the CERN Meyrin site and is the place where all the LHC experiments sent their data to. The data from CMS arrives here from the HLT and the prompt reconstruction is carried out. The Tier-0 centre permanently stores this data and is continuously used for very high priority workflows such as detector calibration.

Tier-1

The Tier-1 centres are scattered across Europe, the USA and Asia. In total CMS is attached to about 6 of these centres. The Tier-1 centres receive the data from the Tier-0 centre at CERN such that for each dataset more than one copy exists on the GRID at all times. The Tier-1s then distribute the data further downstream to the Tier-2 sites connected to them. Sometimes analysis jobs are run on these sites whenever they require very high priority. These GRID sites are also responsible for staging the simulated samples as well as running re-reconstruction of the data whenever needed.

Tier-2

The Tier-2 GRID sites are those that are used for the data analysis. These clusters are usually smaller in size yet contain some significant storage capabilities. They receive data and simulation samples from the Tier-1 centres to provide access to the physics community. These sites are also used to generate simulation samples.

The benefit of this structure is that the data and simulation is at all times stored on different sites across the world. This provides scientists easy access to the data they want to analyse since the system allows to submit analysis jobs to a central Workload Management System (WMS). The WMS matches the job to a given GRID site based on the required input dataset and the available resources at sites hosting the required sample. This makes analysis for the user particularly simple as he or she does not have to know where the data or simulation is hosted, nor has to transfer the data to local machines.

One of the Tier-2 centres is hosted in Brussels at the VUB/ULB computing centre. This Tier-2 site provides 1800 job slots and about 1.2 petabyte of storage capabilities. This cluster allows both local collaborators to carry out their analysis as well as providing access to the wider community to the samples hosted here.

2.2.8 CMS Data taking during the 2011 and 2012 LHC runs

Data taking efficiency

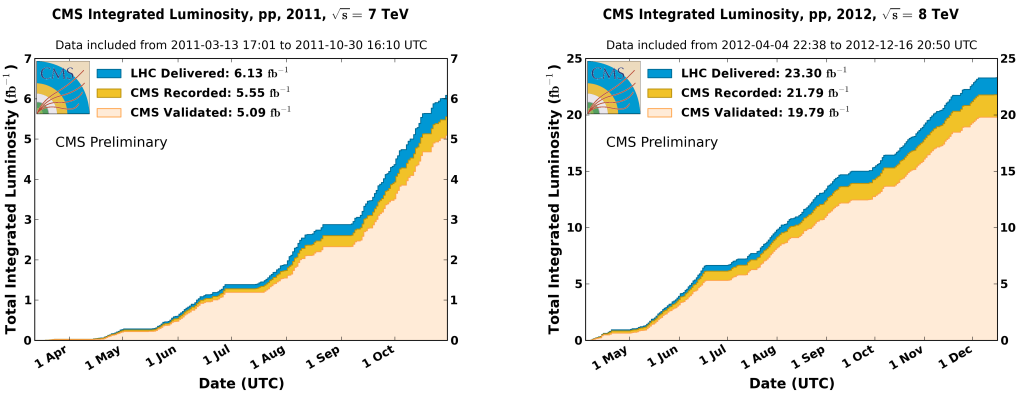


Figure 2.13: Integrated luminosity as a function of time during the 2011 and 2012 proton-proton running of the LHC at $\sqrt{s} = 7$ and 8 TeV [66]. The luminosity delivered by the LHC is compared to the recorded luminosity by CMS. The validated luminosity depicts the data that is certified for physics analysis.

When the LHC machine is filled with two counter rotating proton beams and they are set to collide, the LHC experiments start recording the collisions. It is clear from the previous that an experiment like the CMS detector is very complex. Hence, recording data with the CMS detector does not happen by flipping the 'on/off'-switch. Conversely, operating the CMS experiment requires a 5-person shift crew to continuously

steer and monitor the apparatus around the clock. In the control room, a person is in charge of the data acquisition (DAQ) system accompanied by another person that monitors the trigger. Furthermore, the integrity of the data is checked by a Data Quality Monitoring shifter and finally the operation is overseen by the shift leader. Last but not least, the Detector Control System shifter is assigned to oversee the hardware from cooling to high-voltage power and needs to respond quickly in the event of failure of any system.

Even with this continuous monitoring of the experiment, it sometimes happens that while the LHC is providing collisions, CMS is unable to record them. This could happen because the run is stopped by a malfunction of one of the subdetectors. The data recording is then resumed when the issue is fixed. This means that overall CMS is recording a little less integrated luminosity than the LHC is providing as is shown in Figure 2.13 where the recording efficiency of CMS is 90.5% in 2011 and 93.5% during 2012 operation.

Data certification

After the data is recorded by the CMS experiment, it immediately gets partially reconstructed for the Data Quality Monitoring (DQM) system. Each run recorded by the CMS experiment gets certified both online, i.e. during data-taking, and offline, i.e. after data-taking.

For physics analyses, there are certain requirements on the data: the tracking detector should be operational as well as the calorimeters and the muon system. For this purpose, the data has to be certified before physicists can analyse it. The certification is done by the Data Quality Monitoring shifters that look for anomalies in the detector response during data taking. These anomalies can be any range of things from a sub-detector being switched off or in the wrong operating mode to certain modules producing too much noise or malfunctioning.

A typical distribution from the tracker certification is shown in Figure 2.14. The plot shows the occupancy map of the pixel detector which is used to look for any holes in the detector. These holes are areas in the detector that do not function properly and hence are not producing tracker hits. The occupancy map for each run is then cross-checked to previous runs to assess if new holes have appeared.

Finally, when the certification is over, a list of all certified runs is propagated to all analysts. This list can then be used in analysis to select only the data that is found to be adequate to produce physics results where the detector is found to be operating well. This kind of certification file is used in the analyses presented in this thesis.

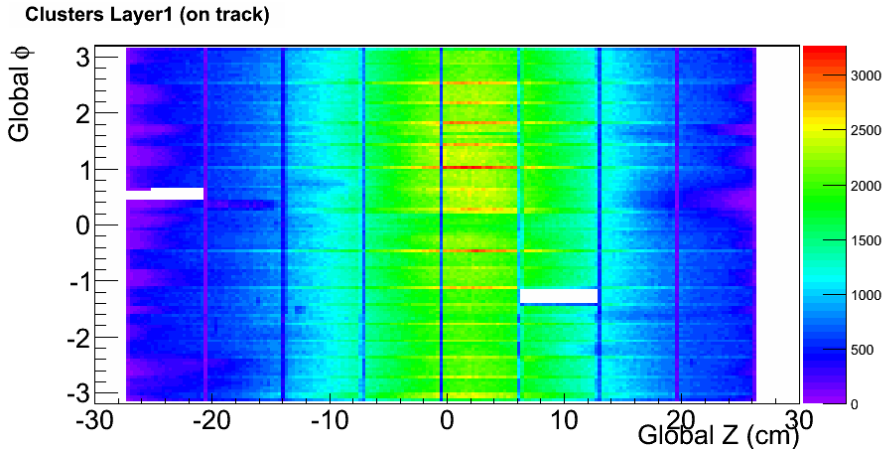


Figure 2.14: A typical distribution from the tracker detector Data Quality Monitoring system. The pixel detector occupancy is shown as a function of the (z,ϕ) position in the detector. This map allows to spot white areas, called holes, where the pixel modules are not working properly.

Chapter 3

Simulation and reconstruction of proton-proton collisions

To exploit the physics potential of the CMS experiment, two key components are developed for the analysis of the data: event generation using Monte Carlo techniques and the dedicated reconstruction of physical objects in the detector. The object reconstruction is of key importance since reconstructing the particles produced in the detector is the only way to unravel the underlying interaction between the colliding protons.

Moreover, to benchmark the reconstruction and to validate models or analysis techniques to be used on real data, the experiment needs to be able to simulate events just like they were real collisions. Events are first generated according to the Standard Model, or any other model, and then pass a detector simulation step. This is the topic of the first part of this chapter. Afterwards we will discuss how physics objects are reconstructed.

3.1 Event generation chain

The complex procedure to generate pp→X events can be subdivided into a number of sequential steps [67, 68]. Figure 4.1 shows this process from the initial interaction between the protons down to the decays of long-lived particles in the experiment. This factorized approach allows for tuning of each individual step to better describe the data collected by various experiments.

Parton Distributions

Each proton consists of three valence quarks and many sea quarks and gluons, called partons. The momentum distribution of the proton among its partons is given by the so-called Parton Distribution Functions which are described in Section 3.1.1.

Hard Interaction

The interaction between two incoming protons is often soft and elastic leading to events which are not interesting in the framework of the research presented in this thesis. More

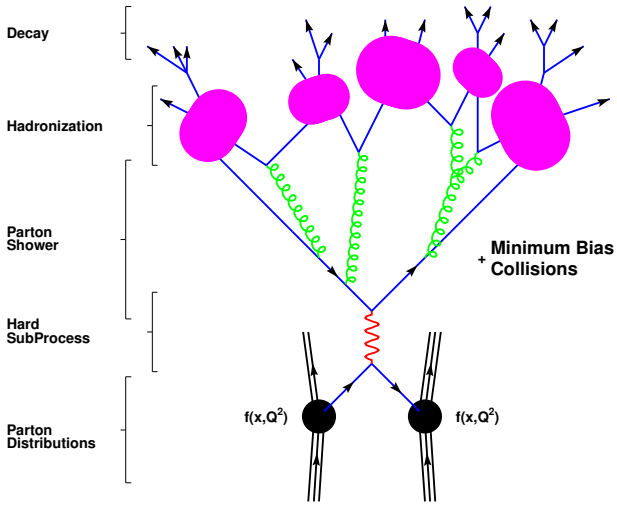


Figure 3.1: Overview of the event generation procedure

interesting are the hard interactions where two partons from the incoming protons engage in a fundamental interaction, often referred to as a collision. This hard interaction is explained in Section 3.1.1.

Parton Shower

The quarks and gluons in both the initial and final state are able to branch into other quarks and gluons. The former is called Initial State Radiation (ISR) and the latter Final State Radiation (FSR). Both types of radiation can be modelled by a Parton Shower approach (PS). This is outlined in Section 3.1.2. Furthermore, the radiation can also be incorporated in the computation of the hard process itself. This, however, leads to double counting of radiation in the Parton Shower step and requires matching between the Parton Shower and the Matrix Element. This will be discussed as well in Section 3.1.2.

Hadronization

When the partons from the Parton Shower move further away from each other, the evolution of the partons cannot be described anymore with perturbative QCD. In this regime, phenomenological hadronization models have to be employed to cluster the partons into colour neutral hadrons. This process is described in Section 3.1.4.

Underlying event

As only one parton from each proton is resolved to produce the hard interaction, the coloured remnants of the protons continue in almost the same direction as the initial proton. These remnants will again be subjected to radiation and hadronization effects. This will also be briefly discussed in Section 3.1.6.

Finally, most hadrons produced in the previous step are not stable and will decay. When these steps are completed, the simulated event can be passed along to a simulation program that describes how this event would be recorded by the CMS experiment. This is further described in Section 3.2.

3.1.1 The hard interaction

Given the Lagrangian of the Standard Model, Feynman rules can be derived that combined with the phase space allow for the calculation of cross sections. Since two protons will interact by resolving a quark or a gluon, the hard interaction is a convolution of different diagrams. In QCD, a pp→X interaction can be factorized [69] in terms of partonic cross sections ($\hat{\sigma}_{ij\to X}$)

$$\sigma_{pp\to X} = \sum_{i,j} \int\int\int dx_1 dx_2 d\hat{t} f_i^{(A)}(x_1, Q^2) f_j^{(B)}(x_2, Q^2) \frac{d\hat{\sigma}_{ij\to X}}{d\hat{t}}, \tag{3.1}$$

where i and j represent the incoming partons resolved from protons A and B respectively and $f_i(x_i, Q^2)$ are called the *Parton Distribution Functions*. The partonic cross sections $\hat{\sigma}_{ij\to X}$ can be calculated in the framework of QCD as an expansion in terms of the strong coupling constant α_s. The lowest order in α_s is called Leading Order (LO) followed by Next-to-Leading-Order (NLO) and so on. Inclusion of each subsequent order in the calculation increases the theoretical precision of the result.

The Parton Density Functions (PDF) $f_i(x_i, Q^2)$ provide the probability for a given parton i to be present inside the proton with a momentum fraction x_i, w.r.t. the proton momentum, when the proton is probed at an energy scale Q^2. As these functions cannot be determined from first principles, they have to be obtained by performing global fits to data. Two collaborations, among others, carry out these PDF fits namely the CTEQ [70] and MSTW [71] groups.

The PDFs are obtained from measurements on deep-inelastic scattering events using the lepton-proton collisions provided by the HERA [72] collider. Also hadron collision data from Tevatron [53] is added to the global fit to further constrain the gluon distribution function. For the LHC regime these distributions are extrapolated to the higher Q^2 scale at either low or very high momentum fraction x. This is only an approximation and hence uncertainties are evaluated.

In the case of top quark pair production, the Q^2 scale can be set to $(346.4\ \text{GeV})^2$, corresponding to the invariant mass of the $t\bar{t}$ system. In this thesis the main PDF set in use if the CTEQ6L1 PDF set [74]. The parton distribution functions for this particular Q^2 scale within CTEQ6L1 are provided in Figure 3.2.

To determine the uncertainty on the parton distributions, the Hessian technique is used [75, 76]. A matrix with a dimension equal to the number of free parameters needs to be diagonalised. In the case of the CTEQ6L1 PDF set this translates in 20 orthonormal eigenvectors. This leads to 20 variations for the pdf parameters in the "+" and 20 variations in the "-" direction.

Multiple programs exist to generate pp→X interactions using eq. 3.1 to give a theoretical description of the collision data. Among them are general-purpose programs

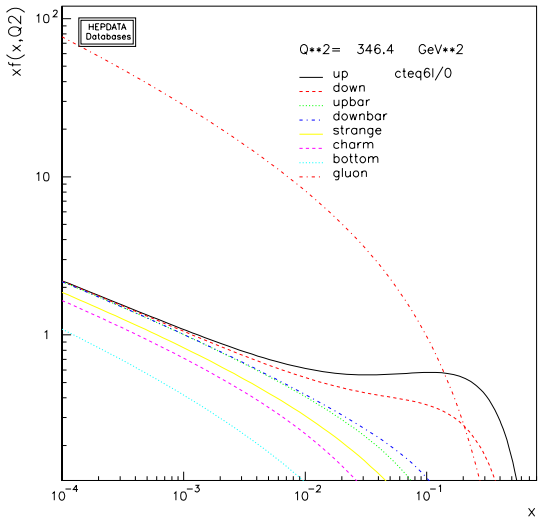


Figure 3.2: Parton Distribution Functions for different partons as a function of the longitudinal momentum fraction x at a Q^2 scale of (346.4 GeV)^2 [73].

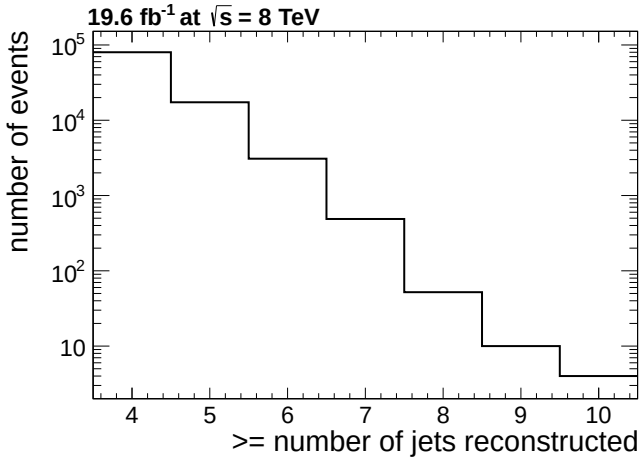


Figure 3.3: Distribution of the minimal number of reconstructed jets per event in 8 TeV collisions at the LHC after selecting the top-pair like events decaying semi-muonic. The jets are the experimental signature of partons as will be explained further in this chapter. The selection of top quark pair events is outlined in the next chapter.

like PYTHIA [77] and HERWIG [78] that are complete event generation chains. Unfortunately, these programs are bound to leading order calculations.

In Figure 3.3, the number of properly reconstructed jets[1] is shown for 8 TeV LHC data. At leading order, the $t\bar{t}$ process would yield four partons in the final state of

[1] A jet is the experimental signature for a quark or gluon through the process of hadronisation. This will be explained later in this chapter along with jet clustering.

the semi-leptonic decay channel. In the distribution it can be seen that these events typically contain more than 4 jets, while a small subset can contain even as many as 10 jets. These additional partons can be generated through higher order effects (in the α_s expansion) like Initial State and Final State Radiation (ISR/FSR) showing the need to go beyond leading order calculations.

To introduce higher order effects in the generation of events, two main techniques can be used. In PYTHIA and HERWIG, additional partons can be generated in a Parton Showering (PS) step (Section 3.1.2) which follows the hard interaction step to approximate the radiative effects. The second method consists in introducing higher order effects directly in the matrix element. This is done by programs like MadGraph/MadEvent [79], PowHeg [80] and MC@NLO [81].

MadGraph/MadEvent

MadGraph is a matrix element generator that can generate the leading order Feynman diagrams for any given process based on the Standard Model, or any other user-defined model. Events are then generated by MadEvent based on the diagrams produced by MadGraph. In addition to the leading order process, MadGraph adds real correction diagrams by generating processes such as $t\bar{t}$+1,2,3 parton. Here extra partons are incorporated in the LO matrix element. Because the generation of all diagrams is CPU intensive, the number of QCD particles in the final state is the limiting factor. Currently, this generator can generate $t\bar{t}$ with up to 3 additional partons while for less complex processes, like W-boson production, 5 additional partons can be generated. Although this is not an exact higher order calculation, since it neglects virtual loop corrections, it gives already a better description compared to the pure leading order generators.

PowHeg and MC@NLO

In the generator landscape other programs like PowHeg and MC@NLO can be found. Both programs include a complete next-to-leading (NLO) order calculation including the virtual loop diagrams that are neglected in for example the MadGraph approach.

The NLO calculation allows the PowHeg and MC@NLO to accurately predict inclusive observables to NLO accuracy. Since NLO calculations only calculate diagrams up to 1 additional parton, these generators have to be interfaced with a Parton Shower program to generate additional radiation. Where MadGraph produces real corrections to the LO calculation for any amount of additional partons, the NLO generators fall back to the Parton Shower approximation beyond 1 additional parton.

3.1.2 Parton showering

The goal of parton showering programs is to describe the process where an incoming or outgoing parton radiates quarks and gluons. To form a parton shower, successive branchings are performed on the partons until the energy scale drops below a lower energy scale where $\alpha_s \approx 1$ and the non-perturbative regime begins. Three different

types of branching exist: a quark radiating a gluon $q \to qg$, a gluon emitting another gluon $g \to gg$ and finally a gluon that splits into a $q\bar{q}$ pair.

When the radiated parton is well separated in phase space and is hard enough, the radiation can be described well with the matrix element calculation. However, the $2 \to n$ process diverges when the radiated parton is at small angle with the mother parton (collinear divergence) or when the energy of one of the partons vanishes (soft or infrared divergence). It is exactly in the soft and collinear limit where the Parton Shower approach is deployed.

The evolution of the parton shower from the hard interaction energy scale Q^2 down to a lower scale can be described by the DGLAP [67, 68, 82–85] equations.

When a parton a branches into partons b and c, parton b will carry an energy fraction $z = \frac{E_b}{E_a}$ leaving an energy fraction *1-z* for parton c. The branching probability for the branching $a \to bc$ can now be written as

$$d\mathcal{P}_{a\to bc} = \frac{\alpha_S}{2\pi}\mathcal{P}_{a\to bc}(z)dtdz = \frac{\alpha_S}{2\pi}\frac{dQ^2}{Q^2}\mathcal{P}_{a\to bc}(z)dz, \tag{3.2}$$

where $\mathcal{P}_{a\to bc}(z)$ are called the splitting functions. The three splitting functions for the existing branchings are given as

$$\begin{aligned}
\mathcal{P}_{q\to qg}(z) &= C_F\frac{1+z^2}{1-z}\ , \\
\mathcal{P}_{g\to gg}(z) &= N_C\frac{(1-z(1-z))^2}{z(1-z)}\ , \\
\mathcal{P}_{g\to q\bar{q}}(z) &= \frac{n_f}{2}(z^2+(1-z)^2),
\end{aligned} \tag{3.3}$$

where $C_F = 4/3$ is the colour factor, n_F is the number of quark flavours and $N_C = 3$ is the number of colours.

From eq. (3.2) and (3.3) it is clear that when $z \to 0$ (soft divergence) or when $Q^2 \to 0$ (collinear divergence) the branching probability diverges. To remove this divergence, a cutoff scale Q^2_{min} of the order of 1 GeV^2 has been implemented. This solves the divergent behaviour but nevertheless the branching probability can still be larger than 1. This is unphysical and a *Sudakov factor* [86] is added to eq. (3.2) to cancel this effect. The $a \to bc$ branching probability is now given by

$$d\mathcal{P}_{a\to bc} = \frac{\alpha_S}{2\pi}\frac{dQ^2}{Q^2}\mathcal{P}_{a\to bc}(z)dz\ \underbrace{\exp\left(-\sum_{b,c}\int_{Q^2}^{Q^2_{max}}\frac{dQ'^2}{Q'^2}\int\frac{\alpha_S}{2\pi}\mathcal{P}_{a\to bc}(z')dz'\right)}_{Sudakov\,factor}, \tag{3.4}$$

where the exponent is the Sudakov factor and the sum runs over all possible branchings. The Sudakov factor represents the probability of evolving from a scale Q^2 down to a lower scale without emitting any parton.

The parton shower approach can now be regarded as a cascade of successive emissions of quarks and gluons. In the case of Final State Radiation (FSR), the cascade starts at the hard interaction scale Q^2 and is evolved through random parton branchings

until a lower scale Q^2_{min} is reached. As this scale is on the boundary of the perturbative region, the further development is carried out by non perturbative hadronization models.

The Initial State Radiation (ISR) is more complex since the incoming protons have a structure. Therefore the ISR evolution is done backwards starting from the hard interaction scale Q^2 and reconstructing what could have happened before. When a parton b engages in the hard interaction, the backward evolution is performed by looking at the probability for an $a \to bc$ branching at a lower scale under the condition that parton b is present at a scale Q^2.

Up to now, the Parton Shower evolution was outlined in terms of the virtuality, Q^2. Nevertheless, other variables exist to evolve the Parton Shower. In PYTHIA, the evolution is performed in terms of the transverse momentum $p_T^2 \approx z(1-z)m^2$ while in HERWIG the energy-weighted emission angle $E^2\theta^2 \approx m^2/(z(1-z))$ is chosen.

In the generated event samples used for the physics studies presented in this thesis the parton showers have been developed with the PYTHIA program. The cutoff scale Λ_{QCD} (controlling the amount of ISR/FSR), the strong coupling constant α_s and the hard interaction scale are given as input to the program. The central values of these parameters are determined such that the best agreement with experimental data from various particle collider experiments is attained. To estimate the effect of the choice of these parameters, samples are generated with up and down variations of these parameters to yield systematic uncertainties.

3.1.3 Matching Parton Showers and Matrix Elements

In the previous two sections, the concepts of higher order matrix elements and Parton Showers were introduced as complementing approaches to handle the description of radiation from the initial and final state partons. On the one hand, this radiation effect was (partially) treated by computing higher order corrections to the hard interaction generating events with full NLO accuracy. This allowed for the description of a $2 \to 3$ process including all virtual corrections as well. Since this is not straightforward to extend to higher orders, MadGraph approximated this by focussing on adding only real corrections to the leading order diagrams to generate $t\bar{t}$ with up to 3 additional partons.

Unlike the matrix element technique, the parton showering approach is computationally cheap and has no constraint on parton multiplicities. Whereas the radiation in the matrix element diverges when the emitted parton becomes soft or collinear, the Parton Shower operates in this particular region. Hence, both approaches should not be treated as alternatives, rather they are complementary and should be merged to give the optimal description of the data.

The latter is nicely illustrated in Figure 3.4 where the Differential Jet Rate is compared between the pure Matrix Element and Parton Shower approaches. The Differential Jet Rate (DJR) is the value of the jet resolution parameter, present in the jet reconstruction algorithm, for which an n jet event turns into an n-1 jet event. The desired curve shows clearly that a combination of both techniques is desirable.

Thanks to the Les Houches Accord [88], the merging of matrix element and Parton Shower is facilitated. Unlike the general purpose programs, such as PYTHIA and

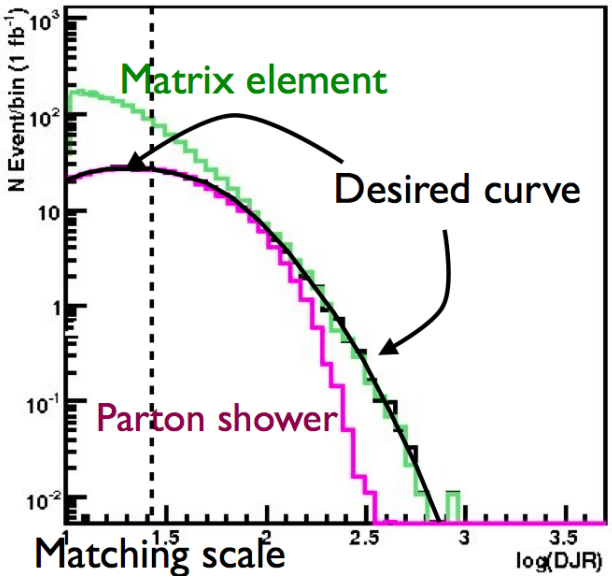


Figure 3.4: Differential Jet Rate compared between a pure Matrix Element Approach and a Parton Shower approach [87]. The desired curve proves to be a combination of both techniques.

HERWIG, most programs are built to perform a specific task. Matrix element generators such as MadGraph are built only to generate the hard interaction step but not beyond. The Les Houches Accord has introduced a generic format to encode parton level information to be passed on to any general purpose generator. This format is implemented by most available programs such that a matrix element generator like MadGraph can pass on its events to PYTHIA for showering and hadronization.

When merging Parton Showers with higher order matrix elements, double counting of radiation might occur since an n+1 jet event could be created in two different ways. When an n+1 parton hard interaction is calculated and passed on to the Parton Shower program, every parton can produce radiation in a cascade, observed as a jet, and thus leads to an n+$\geq$1 jet event. Conversely, an n parton event could also generate n+1 jets when a sufficiently hard additional parton is generated in the Parton Shower.

To solve the double counting of radiation, the MLM Matrix Element-Parton Shower (ME-PS) matching technique can be used. Both PowHeg and MC@NLO use their separate approaches to match NLO to the Parton shower. These three methods will be outlined in the following.

The MLM matching approach [67, 89, 90]

To solve the double counting in MadGraph/MadEvent, the MLM ME-PS matching scheme is used. The basic idea of the MLM scheme is to veto the parton showers that lead to the same multi-parton final states as described by the matrix element itself. The MLM scheme can be divided in the following steps:

- Define the exclusive n-parton sample as the collection of events where exactly n

partons from the matrix element computation pass a number of acceptance cuts. The following acceptance cuts are applied

$$p_T > p_T^{min},\ |\eta| < \eta^{max},\ \Delta R > \Delta R_{min},$$

where p_T is the transverse momentum of the parton, η the pseudo-rapidity and ΔR the minimal angle between the partons in (η, ϕ) space.

- Run the events through a Parton Shower programme.
- Cluster the partons produced in the Parton Shower with a generic clustering algorithm using a jet cone size R. The resulting clusters are labeled "jets" and are put in the final list of jets if they fulfil $E_T > E_T^{min}$.
- Match a matrix element parton to a jet
  - Start from the hardest parton and look for the closest jet in ΔR. The closest jet needs to be within $\Delta R < \Delta R^{min}$ to be matched to the parton. Once the jet is successfully matched to the parton, remove the jet from the list to avoid double matching.
  - Continue this procedure until all partons have been matched.

The n jet exclusive samples can then be defined as the samples of events where all n partons uniquely match a jet and no unmatched jets remain. If N is the maximal number of partons that can be generated in the ME, this constrains n to n$\leq$ N. When n=N, the N jet inclusive sample is defined by the events where all partons are matched to a jet albeit some softer clusters can be present as well.

The PowHeg approach [67, 91]

Matching the Parton Shower to a next-to-leading order computation is a bit more complex than the ME+PS matching in the case of MadGraph. The same problem arrises as before since the Parton Shower here as well generates parton multiplicities that are part of the NLO calculation. The basic idea in PowHeg to mitigate this situation is to swap the hardest emission in the Parton Shower evolution with the NLO emission generated by PowHeg. To do this the full NLO computation including real and virtual corrections is performed. Afterwards the Parton Shower evolution is performed only starting from the p_T scale of the emission of the extra parton in the matrix element. This allows PowHeg to have the hardest parton emission at NLO accuracy while avoiding double counting of radiation.

The MC@NLO approach [67, 91]

MC@NLO takes a much different approach than PowHeg in solving the double counting of radiation between the Matrix Element and the Parton Shower. In this scheme, the first step is to perform the full NLO matrix element computation including n+1 parton real corrections and virtual corrections. Then it is calculated analytically how a first branching in the shower evolution of a n parton event would populate the n+1 parton

phase space. The resulting analytical shower expression can then act as a correction term to the n+1 parton matrix element calculation to remove the pollution from n parton events. What remains after this method are two event populations: n parton and n+1 parton events. Both populations are passed on to a Parton Shower programme and are then added.

Unfortunately, this approach requires correction terms based on the Parton Shower step and as such MC@NLO cannot be interfaced with just any Parton Shower programme. Another downside to this method is that there is no guarantee that the ME is always above the PS in the phase space even though both converge in the collinear and soft region. This is buffered by giving a small set of events a negative weight. The MC@NLO approach is formally equivalent to the PowHeg approach up to NLO but not beyond.

3.1.4 Hadronization

The Parton Shower approach, explained in Section 3.1.2, describes the evolution of the proton-proton interaction down to a cut-off scale Λ_{QCD}. The cut-off scale was introduced to cope with the soft and collinear divergencies in the cross section. Below this scale, the perturbative approach breaks down and the coloured partons created in the Parton Shower and the hard interaction start to group into colour neutral hadrons. This process is called *hadronization* and can only be described using phenomenological models.

The first important step in the hadronization process is the fragmentation step. In PYTHIA, this process is implemented following the Lund string model [92] where the assumption of linear confinement in QCD plays a central role. When two colour-charged partons q and $\bar{q}$ separate from each other, the potential energy stored in the field between them increases. This field could then be interpreted as a string connecting the two coloured objects. When the potential energy increases further the string can break at some point resulting in the creation of a new $q\bar{q}$ string. When the invariant mass of either of these two strings is still large enough, further string breaking might occur. This process continues until only on-shell hadrons remain.

In the last step of the hadronization process, all partons are grouped together to form colour-neutral mesons (composed of 2 quarks) and baryons (composed of three quarks). These particles are not all stable and they will hence decay into daughter nparticles that are in turn observed in the experiment.

3.1.5 Decay

The decay step is again taken care of by the general-purpose PYTHIA program. In this section we will review the hadronization process of a b quark since it plays a central role in this thesis. In an experiment it is very tedious to identify the flavour of a quark. However, b-flavour quarks are particularly interesting because their hadronisation provides some distinct experimental signatures to identify them.

In the hadronization chain of a b quark, b-flavoured hadrons are formed [10] and the most prominent B-hadrons are displayed in Table 3.1. The most interesting property of these B-hadrons is that the admixture has an average lifetime of (1.568 ± 0.009)

10^{-12}s [10]. This lifetime corresponds to $c\tau$=0.47 mm. Hence, a secondary decay vertex is produced that is not compatible with the hard interaction vertex. The presence of a displaced vertex in jets of an experimental collision can thus be used to identify b quarks.

Hadron	Branching fraction (%)	Lifetime (10^{-12}s)
$B^{\pm}$	40.1 ± 0.8	1.641 ± 0.008
B^0	40.1 ± 0.8	1.519 ± 0.007
B_s^0	10.5 ± 0.6	1.497 ± 0.015
$b-baryons$	9.3 ± 1.6	1.382 ± 0.029

Table 3.1: Most prominent b-flavoured hadrons, their branching fraction and their lifetime [10]

The decay of the B-hadron is handled by PYTHIA using the spectator model. When the B-hadron decays the non-bottom quark acts as a spectator that does not take part in the decay. However, the spectator quark serves to identify the flavour composition of the decay particles. The bottom quark will most often decay into a virtual W boson and a charm quark $b \to W^{*-}c$ since this decay mode dominates the suppressed b→u mode ($|V_{cb}| >> |V_{ub}|$). The D-hadrons created of the charm quark have a lifetime of the order of 1 10^{-12} s, smaller than for B-hadrons.

In addition to the presence of displaced vertices, b quarks can also be identified by the presence of non-isolated leptons in the jet. These leptons arise from the decay of the virtual W boson originating from the b quark decay. This W boson can decay either in a lepton and its neutrino or a $q\bar{q}$ pair. Consequently, the decay $b \to l\nu c$ through a virtual W boson has a branching ratio of about 10% per lepton flavour. The non-isolated leptons produced in the b quark decay can be distinguished from other leptons produced in the Parton Shower by their energy and momentum relative to the direction of the parton. The latter leptons originate from inflight decay of π's, K's, photon conversion and mis-identified leptons.

In Section 3.4, the bottom flavour identification algorithms will be explained and how they use these properties to distinguish b quarks from light and charm quarks.

3.1.6 Underlying event

In proton-proton collisions, the incoming protons emit a quark or a gluon to engage in the hard interaction. Since the proton is composed of several valence quarks and sea quarks plus gluons its remnants continue down the same path as the incoming proton beams. However, since a quark or a gluon is missing from the proton, the remnant is no longer colour neutral and will be subject to hadronization. This is called the beam remnant.

In addition, the proton remnants can introduce parton-interactions additional to the hard interaction. This effect is called multi-parton interactions and is a typical side-effect of the compositeness of the proton.

The beam remnant combined with the multi-parton interactions are labeled the *underlying event*. This phenomenon can unfortunately not be described from first

principles and thus phenomenological models have to be used to properly simulate this effect and tune the event generators. The underlying event tuning in PYTHIA has been cross-checked to the CMS data taken at 0.9 and 7 TeV [93, 94] and a good overall agreement to the data has been observed.

3.1.7 Overview of the generated event samples

The signal and background event samples for this thesis are centrally produced by the CMS experiment and are mainly generated with MadGraph interfaced to PYTHIA. The list of samples for both 7 TeV and 8 TeV are displayed in respectively Tables 3.2 and 3.3.

The $t\bar{t}$ signal is generated using MadGraph interfaced with PYTHIA. Additional samples with MC@NLO and PowHeg have been generated for comparison. Systematically varied samples have been generated as well with respect to the nominal MadGraph sample to variate the factorisation scale, ME-PS matching threshold and the top quark mass that are used to generate the events. These samples will be used to provide systematic uncertainties later on in this thesis. The $t\bar{t}$ cross section used to normalise the samples is calculated at NNLO accuracy [22, 23].

One of the main backgrounds to this analysis is the production of a W boson with additional jets. To generate large enough statistics this sample is generated in exclusive bins of W+Xjets where X=1,2,3,4. The cross section of the inclusive W+jets process was calculated at NNLO [95] unlike the exclusive bins which are calculated only up to LO. To extrapolate their cross section to NNLO, a K-factor is derived by taking the ratio of the LO and NNLO inclusive cross section and applying this factor in each jet bin.

Less dominant is the Z boson production in association with jets. This process is generated with MadGraph and the cross section is known with NNLO accuracy [95]. At 7 TeV an inclusive Z+jets sample was generated. To increase statistics for the 8 TeV analyses, the Z+jets process was generated in different additional jet multiplicity bins just as the W+jets process. Again a K-factor was determined to extrapolate the LO cross section in each bin to NNLO.

Finally, another background comes from electroweak single-top production in the t-channel as well as tW production. These samples are generated using the PowHeg NLO generator interfaced to PYTHIA for Parton Showering and hadronization. The single-top process is normalised using its approximate NNLO cross section [96] at 8 TeV and NLO cross section [97] at 7 TeV.

3.1.8 Comparison of different $t\bar{t}$ generators

In the previous part of this chapter, the event generation chain has been discussed at length. In the following, the event generators can be put to work simulating $t\bar{t}$ events allowing comparison between the leading order MadGraph events showered with PYTHIA and the NLO generators PowHeg+PYTHIA and MC@NLO+HERWIG. While MadGraph generates additional partons in the matrix element, PowHeg and MC@NLO benefit from a full NLO calculation.

Process	Generator	Parton Shower	σ (pb)	$\mathcal{L}$ (fb^{-1})
$t\bar{t}$+jets				
nominal	MadGraph	PYTHIA	172.0	20.8
Factorisation scale, ISR/FSR $\uparrow$	MadGraph	PYTHIA	172.0	18.9
Factorisation scale, ISR/FSR $\downarrow$	MadGraph	PYTHIA	172.0	21.4
ME-PS matching threshold $\uparrow$	MadGraph	PYTHIA	172.0	22.9
ME-PS matching threshold $\downarrow$	MadGraph	PYTHIA	172.0	8.8
m_t=163.5 GeV	MadGraph	PYTHIA	172.0	9.5
m_t=181.5 GeV	MadGraph	PYTHIA	172.0	9.3
nominal	PowHeg	PYTHIA	172.0	85.5
nominal	MC@NLO	HERWIG	172.0	125.5
W + 2 jets excl.	MadGraph	PYTHIA	1435.0	17.5
W + 3 jets excl.	MadGraph	PYTHIA	343.0	17.8
W + 4 jets excl.	MadGraph	PYTHIA	194.6	66.3
Z + jets	MadGraph	PYTHIA	3048	11.8
$m_{ll} > 50$ GeV				
Single-top (t)				
t-channel	PowHeg	PYTHIA	42.6	91.3
tW-channel	PowHeg	PYTHIA	10.6	76.6
Single-top ($\bar{t}$)				
t-channel	PowHeg	PYTHIA	22.0	88.1
tW-channel	PowHeg	PYTHIA	10.6	76.2

Table 3.2: Overview of the signal and background samples for $\sqrt{s}$=7 TeV. The generator, cross section and integrated luminosity is provided as well.

Process	Generator	Parton Shower	σ (pb)	$\mathcal{L}$ (fb^{-1})
$t\bar{t}$+jets				
nominal	MadGraph	PYTHIA	245.8	27.6
Factorisation scale, ISR/FSR ↑	MadGraph	PYTHIA	245.8	20.3
Factorisation scale, ISR/FSR ↓	MadGraph	PYTHIA	245.8	21.9
ME-PS matching threshold ↑	MadGraph	PYTHIA	245.8	21.9
ME-PS matching threshold ↓	MadGraph	PYTHIA	245.8	22.3
m_t=163.5 GeV	MadGraph	PYTHIA	245.8	20.9
m_t=181.5 GeV	MadGraph	PYTHIA	245.8	21.8
nominal	PowHeg	PYTHIA	245.8	88.2
nominal	MC@NLO	HERWIG	245.8	132.7
W + 1 jets excl.	MadGraph	PYTHIA	6662.8	53.5
W + 2 jets excl.	MadGraph	PYTHIA	2159.2	15.8
W + 3 jets excl.	MadGraph	PYTHIA	640.4	24.2
W + 4 jets excl.	MadGraph	PYTHIA	264.0	50.7
Z + 1 jets excl. $m_{ll} > 50$ GeV	MadGraph	PYTHIA	666.3	36.0
Z + 2 jets excl. $m_{ll} > 50$ GeV	MadGraph	PYTHIA	215.0	10.7
Z + 3 jets excl. $m_{ll} > 50$ GeV	MadGraph	PYTHIA	60.7	175.0
Z + 4 jets excl. $m_{ll} > 50$ GeV	MadGraph	PYTHIA	27.4	228.2
Single-top (t)				
t-channel	PowHeg	PYTHIA	56.4	66.0
tW-channel	PowHeg	PYTHIA	11.1	44.5
Single-top ($\bar{t}$)				
t-channel	PowHeg	PYTHIA	30.7	62.1
tW-channel	PowHeg	PYTHIA	11.1	44.5

Table 3.3: Overview of the signal and background samples for $\sqrt{s}$=8 TeV. The generator, cross section and integrated luminosity is provided as well.

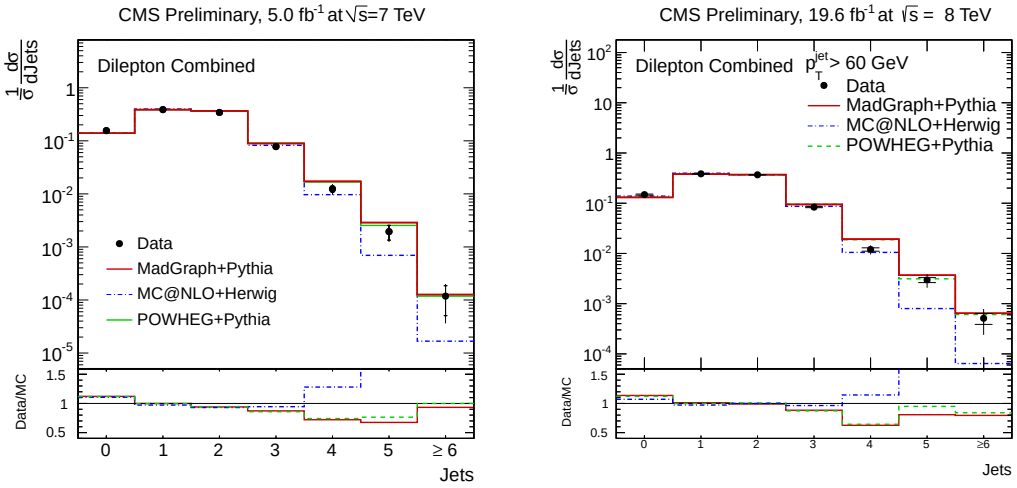


Figure 3.5: Differential $t\bar{t}$ cross section in the di-lepton channel as a function of the reconstructed jet multiplicity compared to different generators at 7 and 8 TeV [98, 99].

One of the variables that is expected to provide sensitivity to the difference between LO and NLO event generation is the reconstructed jet multiplicity. This multiplicity has been measured in data using di-lepton $t\bar{t}$ events and can be compared to the predictions of the three different generators. Figure 3.5 shows the jet multiplicity distribution as measured in data at both 7 and 8 TeV. It turns out that MadGraph and PowHeg model the data very well while MC@NLO predicts a smaller average multiplicity. Since MC@NLO does not yield a satisfactory data description in this case, it is not further considered in this thesis.

The transverse momenta and pseudo-rapidities of the generated top quark in μ+jets $t\bar{t}$ events can be found in Figure 3.6. While the transverse momentum distribution shows good agreement between MadGraph and PowHeg, top quarks are on average more central in pseudo-rapidity in the case of MadGraph. This effect is most pronounced in the tails where the difference grows above 10% and is reproduced in both the lepton and the b quark.

Finally, the mass of the b jet and lepton system in $t \to Wb$ events is shown in Figure 3.7. The distribution is wider for PowHeg $t\bar{t}$ events with a longer tail due to the fact that these events have been generated with a finite top quark width rather than the zero width used in the sample generated with MadGraph.

3.2 Detector simulation

Once the event generation chain outlined in Section 3.1 is completed, the events are passed on from PYTHIA to a detector simulation program to simulate how the detector would respond to the final state particles. In CMS, the GEANT4 [100] program is used to perform the detector simulation. This program tracks the particles through the detector material using a detailed geometrical description of the detector generating hits in several sensitive layers. Then the response of the subdetector electronics to

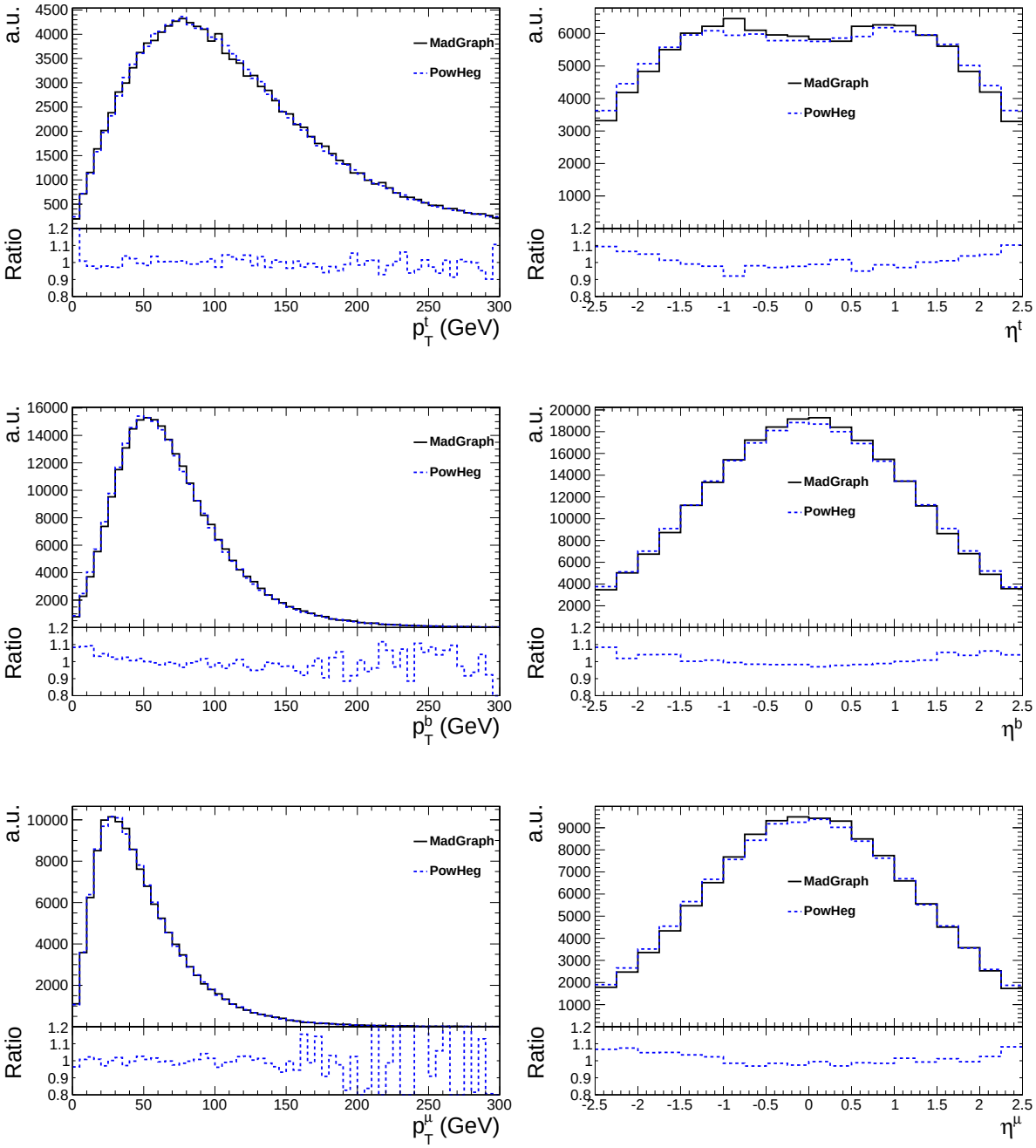


Figure 3.6: The kinematics of top quarks, bottom quarks, muons. The distributions are compared between two μ+jets $t\bar{t}$ samples generated with MadGraph and POWHEG respectively.

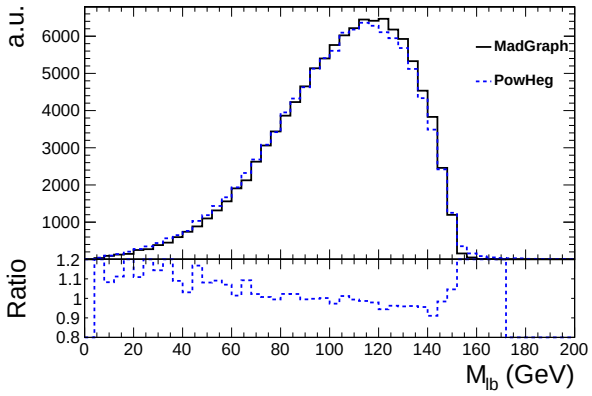


Figure 3.7: The mass of the b quark and muon system compared between two μ+jets $t\bar{t}$ samples generated with MadGraph and POWHEG respectively.

these hits is simulated.

After the detector simulation step, the simulated data have the exact same format of real collision data that are recorded with the CMS experiment. This allows to use the same software framework to reconstruct both data and simulation and allows for direct comparison of simulated events to data. In the next section, it will be shown how the events are reconstructed to allow for physics analyses.

3.3 Object reconstruction

In this section, the object reconstruction in the CMS experiment will be described. This is necessary to translate the digital signals recorded by the detector to physical quantities that can be used in physics analyses.

In CMS, an algorithm is used called ParticleFlow. This algorithm uses the superior performance of the inner tracking detector combined with all the other subdetectors of CMS to give a global description of the event and reconstruct the final state particles of the hard interactions. This algorithm is outlined in Section 3.3.2.

Among the reconstructed particles are the ones important for reconstructing the $t\bar{t}$ decays. The muons and electrons are described in Sections 3.3.3 and 3.3.4 respectively while the jets clustered from hadrons are discussed in Section 3.3.7. To end, the Missing Transverse Energy ($\vec{\not{E}}_T$) will be introduced in Section 3.3.8 as a handle on the weakly interacting neutrinos. In the next section, the treatment of pileup interactions will be discussed.

3.3.1 Pileup interactions

To produce proton-proton collisions, many protons are injected in the counterclockwise rotating beams to increase the probability of collisions in each interaction point. Crossing large bunches of protons indeed increases the probability of collisions and

therefore the instantaneous luminosity. This effect is referred to as in-time pileup as already introduced in Section 2.1.3. In addition to the in-time pileup, the limited time window between two subsequent bunch crossings results in an additional pileup effect namely out-of-time pile up. This mainly affects the calorimeters as their signals have to be integrated over a time window spanning multiple bunch crossings.

It was shown in Section 2.1.3 that on average 20 interactions appear in the experiment per bunch-crossing during the 8 TeV operation. The additional interactions produce extra signals in the detector which have to be simulated in order to make a meaningful comparison between data and simulation for the experimental observable quantities of interest.

In the simulation, additional pileup interactions have been added by injecting minimum bias proton-proton interactions. These additional interactions are not added to the generated events, they are added at the detector level. The latter allows to use the same generated $t\bar{t}$ events with different number of pileup interactions without having to regenerate them. This is particularly useful when the pileup conditions change throughout the data-taking.

The number of pileup interactions per event is a quantity that cannot be directly measured. Therefore, one needs to correlate this effect to experimental observables. One of the candidates is the number of reconstructed primary vertices that proves to be highly correlated to the number of pileup interactions. Moreover, this variable allows to crosscheck the pileup distribution in simulation with respect to data. In Figure 3.8, the number of reconstructed primary vertices for events after a typical $t\bar{t}$ event selection in the muon channel[2] is compared between data and simulation. From this figure it is clear that there is an offset in the average number of pileup interactions between data and simulation.

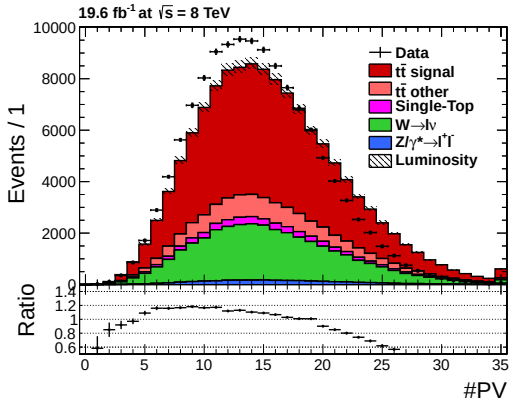


Figure 3.8: Distribution of the number of reconstructed primary vertices in semi-muonic $t\bar{t}$ events from the 2012 8 TeV data

To correct the simulation for this offset, each simulated event is weighted according to the number of pileup interactions. To perform this weighting, the number of pileup interactions from data has to be known. This distribution can be reconstructed from

[2]This event selection is described in Chapter 4

the luminosity measurement using

$$< N_{pileup} >= \sigma_{pp}\ \mathcal{L}, \tag{3.5}$$

where $\mathcal{L}$ is the instantaneous luminosity for a given run in the data-taking period. The total proton-proton cross section (σ_{pp}) has been measured by CMS, ATLAS and TOTEM [101–103]. A central value for this cross section of 68mb and 69.4mb is used for the pileup determination in data at $\sqrt{s} = 7$ and 8 TeV respectively. The resulting pileup distribution estimated for the 2012 data is shown in the left side of Figure 3.9. The uncertainty on this estimate is shown as well and is determined by shifting the total proton-proton cross section by ±5% covering all systematic effects.

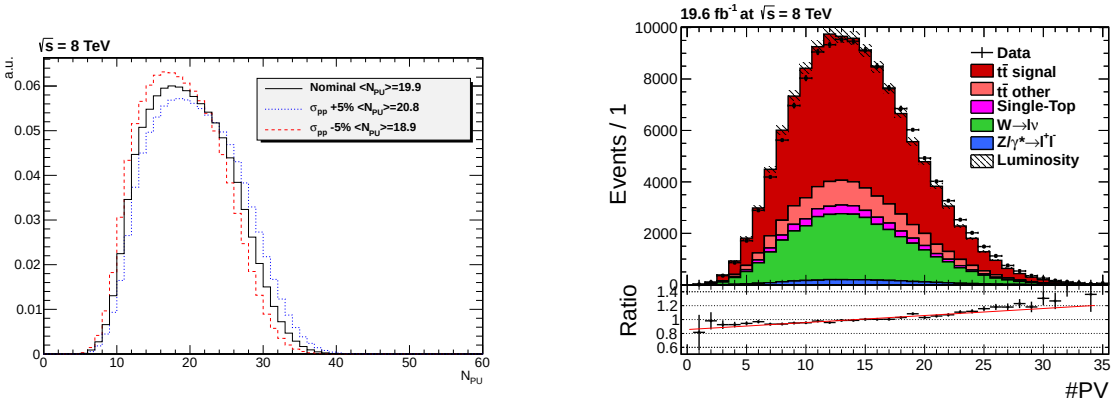


Figure 3.9: The distribution of the number of pileup interactions estimated from the luminosity for the data is shown on the left side. This distribution is used to reweigh the simulation. The right canvas shows the distribution of the number of reconstructed primary vertices in semi-muonic $t\bar{t}$ events after reweighting. The pileup reweighting shows clear improvement in the agreement between data and simulation, the remaining slope is treated by imposing a systematic uncertainty on pileup reweighting.

Now that the number of pileup interactions can be estimated from data, the simulation can be reweighted to match the data. At 8 TeV, the simulation contains information about the true number of pileup interactions that were injected in each event. Consequently, each event can be weighted according to N_{PU} where the weight is immediately determined by comparing the N_{PU} distributions in data and simulation. However, at 7 TeV the N_{PU} number was not stored in the simulated events. Hence this number was approximated by averaging the number of interactions over the central and ± 1 bunch crossings. The effect of the pileup reweighting to the number of reconstructed primary vertices can be directly observed by comparing Figure 3.8 with the right canvas of Figure 3.9. The latter plot is created using a pileup weight for each simulated event and shows much improved agreement with data. The residual slope is taken into account by imposing a systematic uncertainty on the final measurement due to pileup.

3.3.2 The ParticleFlow Algorithm

The main characteristic of the ParticleFlow reconstruction [104–106] is that it reconstructs all stable particles in the event by combining the information obtained from each of the CMS subdetectors. The combination of all subdetectors to obtain the final list of particles ultimately leads to an improved momentum determination as well.

In each event, *elements* are defined such as tracks, ParticleFlow calorimeter clusters, and muon track segments from the muon detectors. The ParticleFlow algorithm then consists of two main steps. First, pairs of elements are grouped into *blocks* by the linking algorithm. Then, the linked blocks are fed to a *particle reconstruction and identification* step that generates the final list of particles in the event.

Muon track segment reconstruction

The reconstruction of track segments in the muon detector is driven by the same Kalman filter [107] as used for reconstructing tracks in the inner tracking detector. The muon track is fitted inside out using the segments reconstructed from the Drift-Tubes (DT) augmented with the reconstructed hits from the Resistive Plate Chambers (RPC). In the endcaps, the hits within the Cathode Strip Chambers (CSC) are used alongside the RPCs.

Once the track is fit starting from the innermost layer all the way to the final layer of the detector, the last hit is used as starting point to perform a fit outside-in. This second fit is performed to improve the determination of the track parameters at the boundary between the muon station and the outer hadron calorimeter such that the matching of the muon track in the central tracker with the muon track segment is improved.

Calorimeter Clustering

The clustering of energy deposits in the calorimeter serves multiple purposes,

- to measure the energy and position of neutral hadrons and photons,
- to separate neutral hadrons from charged hadrons,
- to complement energy measurement for charged hadrons which have poorly reconstructed tracks (due to low-quality or very high p_T)
- and to identify electrons and the Bremsstrahlung photons they emit while crossing the tracker material.

As a result, a specific clustering procedure was developed for Particle Flow to ensure detection of low-energy particles and to be able to separate close energy deposits. The clustering is performed separately on ECAL, HCAL and the PreShower (PS) detector separately, but not on HF where every cell gives rise to a cluster.

First the calorimeter cells with an energy above a certain threshold are taken as *cluster seeds*. The threshold is taken as two standard deviations of the electronic noise

corresponding to 80 up to 300 MeV in the ECAL barrel and end-caps respectively and about 800 MeV in HCAL.

From the cluster seeds, *topological clusters* are formed and calorimeter cells are added to the cluster when they have at least one side in common with a crystal already in it. Moreover, the calorimeter cell must have a signal which is at least two standard deviations larger than the expected electronic noise in order to be added to the cluster.

When all the topological clusters are formed, each cluster seed is assigned to a *ParticleFlow Cluster*. This means that every topological cluster contains as many ParticleFlow Clusters as it contains seeds. The energy in the topological cluster is shared among its ParticleFlow Clusters proportional to the (η,ϕ) distance between ParticleFlow Cluster i and cell j. Then the position of the ParticleFlow Clusters is recomputed as the centre-of-gravity of the five or nine central cells. This process is iterated until the position of the ParticleFlow Cluster is stable. For the ECAL, an extra correction to the ParticleFlow Cluster position is added because the crystals are tilted.

In general, an event contains tracks, ParticleFlow Cluster and possibly track-segments from the muon stations. Consequently, these "building blocks" need to be combined by a *link algorithm* to be able to reconstruct particles and ultimately provide an event description.

Link algorithm

The link algorithm produces *blocks* by linking elements. The quality of each link is then defined by the distance between the linked elements. Thanks to the high granularity of CMS, the blocks typically contain only a few elements. The advantage of the method is that, with very complex events, the number of blocks increases. In general the number of elements in the block will remain the same.

Track - Cluster link

To make the link between a central track and the calorimeter clusters, the track is propagated from the outermost hit in the tracker to the calorimeters. The propagation is carried out to a depth corresponding to the maximum of a shower profile in the ECAL and the HCAL. In the end-caps, the track is also propagated to the two layers of the PS. The track and the calorimeter cluster are linked if the propagated track is within the cluster boundaries. The cluster envelope can be enlarged by one cell in each direction due to the uncertainty on the position of the shower maximum, multiple scattering and energy leaks. The link distance is then defined by the distance in (η,ϕ)-space between the extrapolated track and the cluster.

Cluster - Cluster link

The link between clusters of HCAL and ECAL or PS and ECAL is present if the cluster from the most granular detector is within the envelope of the less granular one. This means that the ECAL cluster must fit into the HCAL cluster or the PS cluster has to fit into the ECAL cluster. Just as with the track-cluster link, the envelope is allowed to

grow in size by one calorimeter cell. The link distance is again defined as the distance in (η,ϕ)-space between the elements.

Track - Muon Track-Segment link

Subsequently, the linking is performed between tracks in the central tracker and track segments from the muon system. When the global fit between the inner track and the muon track results in an acceptable χ^2, the tracks are linked and the result is considered a *global muon*. When a muon track can be fit to multiple tracker tracks, this leads to multiple global muons. Only the global muon with the lowest χ^2 is retained. In this case, the link distance is defined by the χ^2 value rather than the distance in (η,ϕ)-space.

Particle reconstruction and identification

When all the blocks in the event are built, they are propagated to the reconstruction and identification step. This step will build the final list of particles and thus provides a global description of the event. This last step starts by reconstructing muons (Section 3.3.3) and electrons (Section 3.3.4). When their constituents are removed from the blocks, hadrons and photons are reconstructed (Section 3.3.6). When the ParticleFlow particle candidate collections are built, identification criteria are applied to remove fake particles.

3.3.3 Muon reconstruction and identification

When a ParticleFlow block contains a link between a muon detector track segment and a track in the inner tracker, this gives rise to a global muon candidate. For global muons their properties are reconstructed using a combination of the inner silicon tacker and the outer muon systems. In turn if the momentum of the global muon matches its tracker-based momentum determination within three standard deviations, this global muon is considered a *ParticleFlow muon* and its track is removed from the block. A transverse momentum resolution of 1-6% [108] is obtained for the muons used in this thesis.

The resulting ParticleFlow muon collection still contains a large fraction of misidentified charged hadrons and other particles punching through the calorimetric system. To purify the collection, additional identification criteria are enforced on the muon candidates. The ParticleFlow algorithm uses three different sets of muon identification criteria *isolated*, *PF-tight* and *PF-loose*.

For example in semi-muon $t\bar{t}$ decays, the muon from the W-decay is expected to be well separated from all other final state objects in the event. To identify this type of muons, a criterion can be developed describing the particle activity surrounding the muon. To this extent, the relative isolation variable I_{rel} is introduced and defined as

$$I_{rel} = \frac{\sum p_T^{Tracks} + \sum E_T^{ECAL} + \sum E_T^{HCAL}}{p_T^{\mu}}, \tag{3.6}$$

where the sums run over all track transverse momenta and calorimeter energy deposits within a cone of radius 0.3 around the muon track in the (η,ϕ)-space. The distance in

this space is defined as

$$\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}. \tag{3.7}$$

The isolation cone radius is tuned in such that the muon isolation efficiency, i.e. the efficiency of identifying an isolated muon with the muon isolation criterion, is maximal keeping the fake rate low. In the 8 TeV analysis, the cone has been extended to a radius of 0.4 to increase isolation performance.

The isolation definition in eq. (3.6) is completely driven by tracks and calorimetry deposits. However, the ParticleFlow algorithm reconstructs all charged and neutral hadrons as well as photons. Hence the relative isolation can be described in terms of particles rather than energy deposits and tracks and is redefined as

$$I_{rel} = \frac{\sum p_T^{charged\ hadrons} + \sum p_T^{neutral\ hadrons} + \sum p_T^{photons}}{p_T^{\mu}}. \tag{3.8}$$

The sums now run over reconstructed particle momenta instead of the transverse energy of calorimeter deposits. Moreover, the sum over track momenta has been replaced by the sum over the transverse momenta of the reconstructed charged hadrons.

The particle based relative isolation provides an improved isolation definition compared to the traditional relative isolation criterion. Nevertheless, this definition has been found very sensitive to pileup effects. This pileup dependence was found especially in the neutral hadrons and photons since the charged hadrons allow matching to the primary vertex and are thus less affected. During 8 TeV operation, where the pileup was more pronounced compared to the 7 TeV running period, this has lead to a redefinition of the particle based isolation corrected for the pileup effect by adding a subtraction term for the transverse momenta of the pileup particles. If the production vertex of a charged particle is displaced from the primary vertex in the z-direction, then the particle is considered as pileup.

The corrected particle based relative muon isolation is then defined as

$$I_{rel} = \frac{\sum p_T^{charged\ hadrons} + max[0, \sum p_T^{neutral\ hadrons} + \sum p_T^{photons} - 0.5 \sum p_T^{PU}]}{p_T^{\mu}}. \tag{3.9}$$

The particle based relative muon isolation is shown in Figure 3.10 for mu+jets $t\bar{t}$ events. The distribution is split for muons that match the generated muon, ΔR <0.2, from the W-boson decay and the muons not matched with the generated muon. It can be seen that the unmatched muons tend to be less isolated either because they are produced in a jet or they are mis-identified as a muon. The matched muon distribution shows a peak around 0 where the isolation is maximal as is expected for the signal muons. The distribution is cut off around 0.2 because of the isolation criterion embedded in the ParticleFlow muon identification. Comparing the relative isolation between 7 TeV and 8 TeV for matched muons it is clear that the pileup subtracted isolation definition at 8 TeV improved the muon isolation. The peak at low I_{rel} is more pronounced and the distribution falls more rapidly.

To retain efficiency for non-isolated muons produced inside jets, the *PF-tight* and *PF-loose* identification is used which are optimised for reconstructing muons inside jets. After all isolated muons have been handled the PF-tight muons are identified

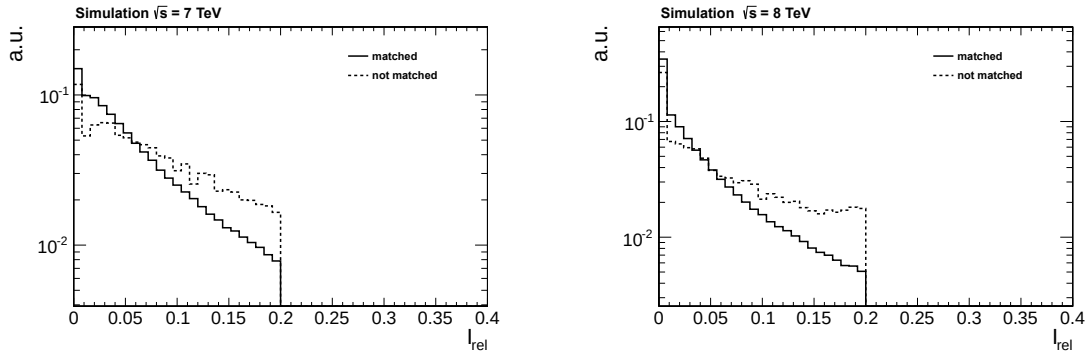


Figure 3.10: The particle based relative muon isolation in muon+jets $t\bar{t}$ events is shown at 7 and 8 TeV where the reconstructed muon is matched to the generated muon from the W-boson decay (matched) and where they do not match (not matched). The matching is performed by requiring $\Delta R(\mu^{rec}, \mu^{gen})$ <0.2.

by requiring a minimum number of hits assigned to the muon track and requiring geometric compatibility between the muon segment and the calorimeter deposits. To further recover efficiency, the PF-loose identification relaxes these cuts further albeit with an additional requirement that the momentum of the track exceeds the calorimeter deposit to rule out the charged hadron hypothesis.

In addition to the identification criteria embedded in the ParticleFlow algorithm, muon identification requirements outlined in Table 3.4 are applied in the analysis to further purify the muon reconstruction. First, a selection is made based on the muon track parameters to ensure a proper reconstruction of the muon track. All muons are required to have a good global fit probability of its tracker track and muon track segment assuring a good quality muon track. Hence, the muon track is required to have a measurement in at least 1 pixel layer and more than 5 or 8 tracker layers depending on the data taking period. At 8 TeV the requirement on the number of pixel layers is dropped and replaced by the criterion that at least one valid pixel hit needs to be reconstructed. Finally, the muon track segment needs to have at least one hit in the muon detectors with at least 2 muon stations matching the global muon track.

Additionally, the muon identification serves the purpose to ensure that the well reconstructed muon does not originate from a mis-identified particle like a charged hadron. Therefore, the muon production vertex is constrained both in the (x,y)-plane to 0.02 cm and in the z-direction to 1 cm and 0.5 cm at 7 and 8 TeV respectively. To ensure the muon does not originate from a B or D hadron decays inside a jet it has to be separated from any reconstructed jet (cfr. Section 3.3.7) in the (η, ϕ) plane by imposing $\Delta R > 0.3$.

3.3.4 Electron reconstruction and identification

The reconstruction of electrons is in general much more complicated than the reconstruction of muons. Since charged leptons traversing the tracking detector have to pass a significant material budget in order to reach the calorimetry they suffer from energy

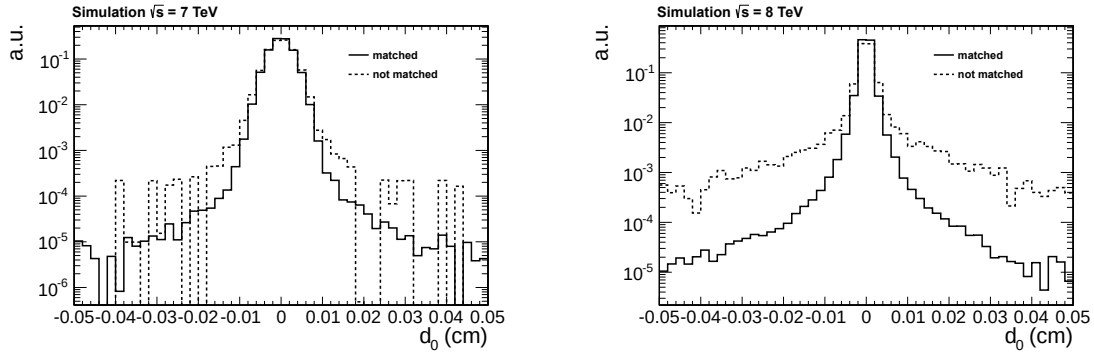


Figure 3.11: The muon transverse impact parameter with respect to the primary vertex (d_0) in muon+jets $t\bar{t}$ events is shown at 7 and 8 TeV where the reconstructed muon is matched to the generated muon from the W-boson decay (matched) and where they do not match (not matched). The matching is performed by requiring $\Delta R(\mu^{rec}, \mu^{gen})$ <0.2.

Criterion	$\sqrt{s} = 7$ TeV	$\sqrt{s} = 8$ TeV
GlobalMuon	required	required
χ^2 of global fit	<10	<10
# tracker layers with measurement	>8	>5
# pixel layers with measurement	≥1	-
# valid hits in the pixel detector	-	> 0
# matched muon stations	>1	>1
# hits in the muon detector	>0	>0
Transverse IP of the muon w.r.t. primary vertex (cm)	< 0.02	<0.02
Longitudinal IP of the muon w.r.t. primary vertex (cm)	<1	<0.5
Relative muon isolation	< 0.125	0.12
$\Delta R(\mu, jet)$	>0.3	>0.3

Table 3.4: Overview of the muon identification criteria used in the 7 TeV and 8 TeV datasets.

losses traversing this material. However, for muons this energy loss is mainly defined by Coulomb scattering which can be modelled in the Kalman Filter, the electrons mainly lose energy due to Bremsstrahlung.

When electrons travel through the tracking detector in the strong magnetic field they will suffer energy loss from radiating Bremmstrahlung photons. These photons will reach the ECAL but with a spread in the ϕ direction with respect to the point where the electron track crosses the calorimeter. Roughly 35% of all electrons lose up to 70% of their energy when traversing the CMS tracker. Moreover, a smaller fraction of 10% loses even up to 95% of its energy.

To reconstruct the electron, the track fit has to account for the Bremsstrahlung effect. This can be achieved with a Gaussian-Sum Filter (GSF) [109]. The Gaussian-Sum Filter is an extension of the Kalman Filter that is being used in the track reconstruction where the difference lies in the modelling of energy losses. In the Kalman Filter, the energy loss of a charged particle traversing the tracker material is considered to be Gaussian. While this approach performs well for multiple scattering, the Bremsstrahlung effect is better described by a composition of Gaussians which is used by the GSF.

The fit starts from a collection of seeds that is generated in two different ways: cluster-driven and tracker-driven seeding. In the cluster-driven seeding, the energy weighted average position of the calorimeter cluster is propagated through the magnetic field to the most inner layer of the pixel detector to identify the first hit. If this layer does not yield any hit, the next layer is looked at to overcome pixel seeding inefficiencies. This seeding method provides the best possible seeding efficiency for high p_T electrons while keeping the fake-rate as low as possible.

On the other hand, the tracker-driven seeding provides good seeding efficiency for low p_T electrons. Here, the collection of tracks produced by the iterative tracking algorithm are used to define a set of seeds. The reconstructed tracks not pre-identified as electrons are filtered out. When an electron suffers significant energy loss, the track properties can be used to perform the pre-identification. More precisely, when the KF is used to fit the electron track it will either lose track of the electron resulting in a short track with few hits or follow the electron all the way through the tracker albeit with a high χ^2 value for the fit. These two properties can then be used to define a subset of candidate tracks. These tracks are then refit with the GSF. The pre-identification is finally performed using a Boosted Decision Tree (BDT) which creates one single discriminator from multiple variables such as the ratio between the GSF and KF track χ^2 values.

Once the collection of trajectory seeds is obtained from both the tracker- and calorimeter-based seeding approaches, the electron trajectory is built using the Gaussian-Sum Filter. The electron momentum is then measured using both the GSF track and the corresponding calorimeter cluster.

In the CMS reconstruction software, two different electron reconstruction methods can be distinguished in the way they reconstruct the calorimeter cluster. The first is the ParticleFlow electron reconstruction [106, 110] where the cluster is reconstructed using the PF Clusters with bremmstrahlung recovery. This method has been used at 7 TeV to reconstruct the electrons in $t\bar{t}$ events. The second method is the ECAL-

driven electron reconstruction [111, 112] where an ECAL SuperCluster is used for the calorimeter-based seeding. This reconstruction is being used at 8 TeV because an inefficiency in the reconstruction of ParticleFlow electrons in the endcap region was observed motivating the switch. Both methods will be outlined in this section followed by the electron identification criteria.

ECAL-driven electron reconstruction

The ECAL-driven electron reconstruction uses an ECAL SuperCluster (SC) to drive the calorimeter-based track seeding.

In the ECAL, a typical electron will deposit most of its energy in a limited number of crystals. It has been studied in a test-beam setup that for electrons with 120 GeV of energy, 97% of their energy is deposited in a 5x5 matrix of crystals.

The ECAL SuperCluster is built by clustering, rows of 3 to 5 connecting ECAL crystals in the η direction. Since the Bremsstrahlung photons are emitted in the transverse plane, they can be collected by connecting isolated clusters in the ϕ direction to the SuperCluster.

ParticleFlow electron reconstruction

In the ParticleFlow algorithm, the seeding for the GSF track fitting starts from a ParticleFlow Cluster rather then an ECAL SuperCluster. Next to this, the ParticleFlow algorithm incorporates a different technique for collecting the possible Bremsstrahlung photons.

When the electron emits a Bremsstrahlung photon, this photon in turn can materialise to an electron-positron pair within the tracker volume. Hence, these electrons would be picked up by the tracker seeding and even by the cluster-driven seeding if the conversion happens within the pixel detector volume resulting in additional tracks. To remove these tracks prior to the electron reconstruction, the ParticleFlow algorithm cleans the collection of reconstructed tracks. Photon conversion in the tracker material often gives rise to displaced tracks. These tracks can then be identified and removed by looking at the distance of their inner most hit to the beam line.

Subsequently, electrons are reconstructed from the remainder of the pre-identified and cleaned GSF tracks. This reconstruction is performed by using the ParticleFlow linking algorithm to match the GSF track to a calorimeter cluster. When such link is found, the ParticleFlow algorithm attempts to recover possible Bremsstrahlung photons.

The Bremsstrahlung recovery procedure is depicted in Figure 3.12. At each intersection between the tracker layers and the electron track, the tangent to the track is propagated to the ECAL surface. If the tangent lies within the envelope of a calorimeter cluster, a track-cluster link is defined in the usual way.

Finally, the ParticleFlow reconstruction embeds a first layer of electron identification to separate true electrons from charged hadrons. This identification uses a Boosted Decision Tree (BDT) multivariate discriminator based on multiple tracking variables as well as variables describing the match between the track and the calorimeter.

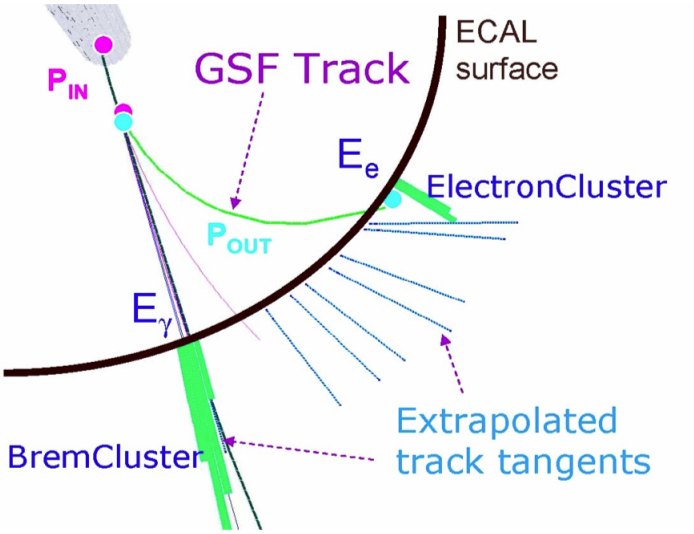


Figure 3.12: Visualisation of the Bremsstrahlung recovery method applied in the ParticleFlow electron reconstruction. At each intersection of the track with a tracker layer, a tangent to the track is extrapolated to the ECAL and a corresponding cluster is searched for [110].

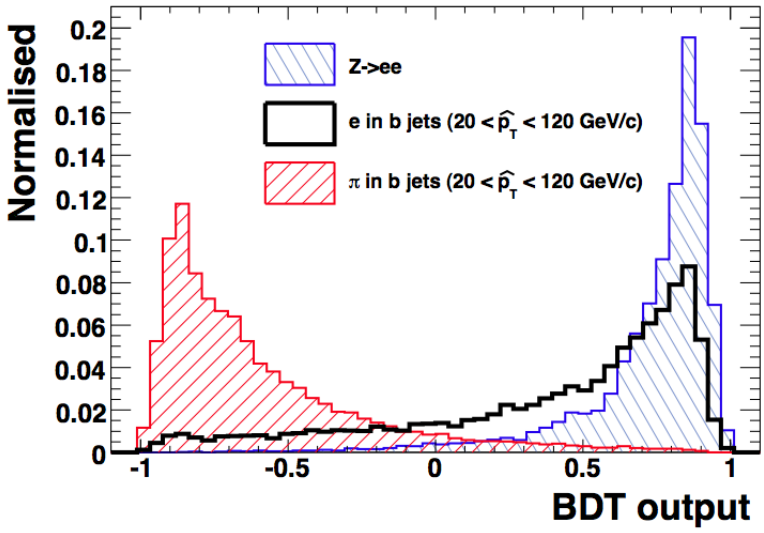


Figure 3.13: The output of the Boosted Decision Tree (BDT) used for the separation of true electrons and charged hadrons [110].

The BDT output is shown in Figure 3.13 and shows a good separation for isolated electrons originating from Z boson decays and non-isolated pions. It can be noted as well that also for soft non-isolated electrons in b jets, a good separation is obtained allowing for efficient reconstruction of electrons in jets. The electron candidates that have a discriminator value exceeding -0.1 are finally labeled as *ParticleFlow electrons*.

The resulting electrons, which are used for the analyses presented in this are, have an average energy resolution of less than 5% [64].

3.3.5 Additional electron identification

To further purify the electron reconstruction and because the ECAL-driven reconstruction has no embedded electron and charged-hadron separation, additional identification criteria are enforced. These are geared towards the optimal reconstruction of the isolated electrons that appear in the semi-electronic $t\bar{t}$ decay. All the identification criteria for both the 7 and 8 TeV data are outlined in Table 3.5.

To select isolated electrons, the particle based isolation can be used for electrons using the definition in eq. (3.8). The isolation is determined in a cone of radius 0.3 around the electron track. For the collisions at 8 TeV, where the pileup influence is increased, the particle based isolation definition can be extended just as in the muon case with a pileup subtraction term. In this case the so-called effective area correction term ρA_{eff} [113] is added to the equation since it provides better isolation performance. Thus, the relative electron isolation is defined as

$$I_{rel} = \frac{\sum p_T^{charged\ hadrons} + max[0, \sum p_T^{neutral\ hadrons} + \sum p_T^{photons} - \rho A_{eff}]}{p_T^{\mu}}, \qquad (3.10)$$

where ρ is the energy density in the event and A_{eff} is called the effective area (cfr. Section 3.3.7).

The relative isolation for ParticleFlow electrons at 7 TeV is shown in the left plot in Figure 3.14 while the right plot contains ECAL-driven electrons. In both cases the isolation is compared between matched and unmatched electrons demonstrating that the unmatched electrons are in general less isolated. For the ECAL-driven electrons, no cut-off around I_{rel}=0.2 is observed since there is no embedded identification in the reconstruction while ParticleFlow only retains loosely isolated pre-identified electrons.

In this transition region between the barrel and the endcap, there is a gap in the sensitive material as can be seen in Section 2.2.4, leading to a degraded electron resolution. As a consequence, electrons having a Super-Cluster located in the transition region between the ECAL Barrel (EB) and the ECAL Endcap (EE) or more precisely in the region $1.4442 < |\eta_{SC}| < 1.5660$ are rejected.

To ensure that the electron is a final state particle in the hard interaction, the electron candidate has to have an impact parameter within 0.02 cm from the primary vertex and needs to be separated from any jet in the (η, ϕ) plane by imposing ΔR to exceed 0.3. This ensures that the electron does not originate from a jet that is wrongly reconstructed as an electron.

Finally, two additional criteria have to be met: a dedicated electron identification and a conversion rejection criterion. First, identification tools have been developed to

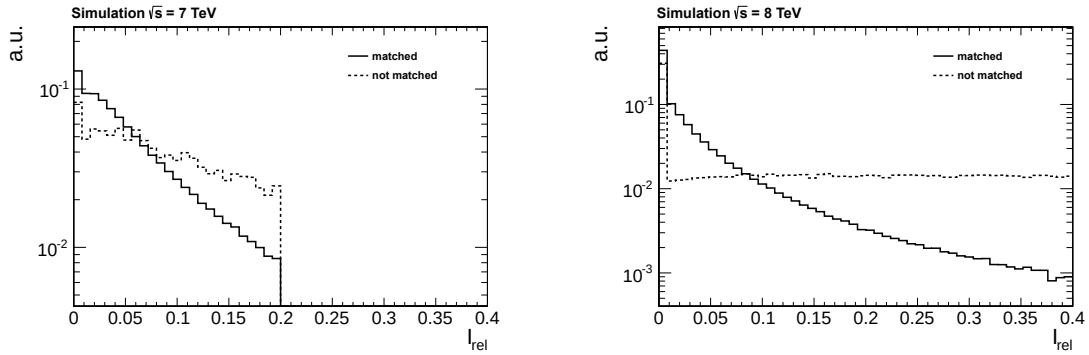


Figure 3.14: The particle based relative electron isolation in electron+jets $t\bar{t}$ events is shown at 7 and 8 TeV where the reconstructed electron is matched to the generated electron from the W-boson decay (matched) and where they do not match (not matched). The matching is perfumed by requiring $\Delta R(e^{rec}, e^{gen})$ <0.2.

provide a single decision based on different variables. In the 7 TeV analysis, a cut-based electron identification [114] is used that provides a single decision based on a set of variables using pre-defined cuts. Different sets of cuts were provide "working points" where in this thesis, the working point was used corresponding to a reconstruction efficiency of the order of 80%. Among the variables used are H/E (hadronic over electromagnetic energy fraction), the Bremsstrahlung energy fraction, the ratio of SuperCluster energy and momentum, the absolute difference between the electrons ECAL energy and the track momentum at the vertex and the separation in (η,ϕ)-space between the SuperCluster and the electron track at the vertex.

Conversely, in the 8 TeV analysis, the cut-based electron identification was swapped with a multivariate analysis (MVA) based identification [115] . The MVA combines different sensitive variables into one final discriminator separating fake electrons from good ones. Among the variables used are the χ^2 of the fit, geometrical matchings between the ECAL SuperCluster and the electron track, ECAL shower shape variables, such as the supercluster width in η and ϕ, and energy matching variables, such as the SuperCluster ECAL preshower energy over the raw energy.

The MVA electron identification output is shown in Figure 3.15 for matched and unmatched electrons. The matched electrons show a peaked distribution around 1 meaning that they are more electron-like while the unmatched electrons have an MVA value uniformly distributed between 0 and 1. The threshold is put at MVA > 0.9 which for electron+jets $t\bar{t}$ events yields a reduction of 68% unmatched electrons while keeping the matched electron efficiency at 98.6%.

Finally, a conversion rejection criterion has to be applied since electrons emit Bremsstrahlung photons when traversing the magnetic field. These Bremsstrahlung photons in turn can materialise in the tracker to form a new electron pair. These electrons, however, should not be treated as normal electrons since they originate from Bremsstrahlung and thus should be accounted for in the energy of the signal electron. This disambiguation is performed by reconstructing conversion candidate tracks. Then a simple set of geometrical cuts is applied. The first variable is the number of lost hits,

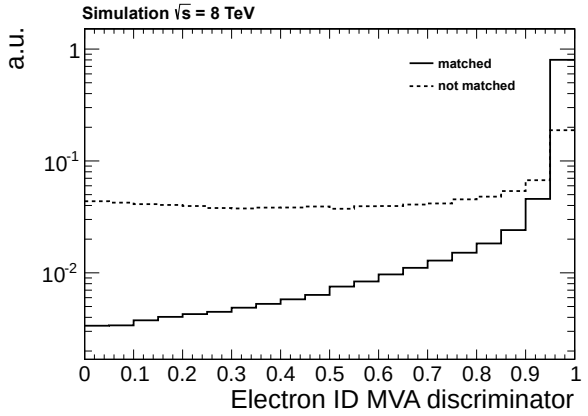


Figure 3.15: The MVA electron identification output distribution for electrons that are matched to the generated electron from the W-boson decay (matched) and where they do not match (not matched). The matching is perfumed by requiring $\Delta R(e^{rec}, e^{gen})$ <0.2.

representing the difference between the number of expected tracker hits for a signal electron and the number of hits assigned to the track, which should be 0. Additionally, the variables $\Delta cot\theta$ and $|\Delta dist|$ provide the angle between the conversion tracks and the absolute distance of closest approach of the tracks. These three variables have been used to reject conversion electrons in the 7 TeV analysis.

In contrast with the cut-based conversion rejection presented for the 7 TeV analysis, a new technique is used at 8 TeV. A full vertex fit is performed on pairs of charged particle tracks with the aim of reconstructing conversion vertices. A conversion vertex is retained if it has a fit probability higher than 10^{-6}. Additionally, its transverse decay length needs to be greater than 2 cm and the corresponding tracks should have no associated hits before the position of the conversion vertex. If such a vertex is found, the electron is rejected. Finally, analogous to the 7 TeV analysis, electrons are rejected when they have 1 or more hits missing.

3.3.6 Hadron and photon reconstruction

After reconstructing the leptons, the remaining blocks consist mainly of hadrons (charged or neutral) and photons. To identify neutral hadrons, the comparison is made between the momentum of the track (or the sum of the momenta of the tracks) linked to the PFCluster and the cluster energy.

It is possible for a certain track to get linked to multiple ECAL or HCAL clusters. In this case only the closest link is preserved for HCAL. For ECAL, if the extra links come from photons, the links should be dropped to allow photon detection. If the links are caused by fluctuations in the hadronic shower, the links should be kept to avoid double-counting. In general, the links to ECAL clusters are ordered according to the link distance.

The links are removed once the total calibrated calorimetric energy (HCAL+ECAL)

Criterion	$\sqrt{s} = 7$ TeV	$\sqrt{s} = 8$ TeV
Exclusion of the EB-EE transition region	$!(1.4442 < \lvert\eta_{SC}\rvert < 1.5660)$	$!(1.4442 < \lvert\eta_{SC}\rvert < 1.5660)$
Transverse IP of the electron w.r.t. primary vertex (cm)	< 0.02	<0.02
Relative electron isolation	< 0.1	<0.1
$\Delta R(e, jet)$	>0.3	>0.3
Dedicated electron ID	cut-based ID [114]	MVA > 0.9 [115]
Conversion rejection:		
# of lost tracker hits	$==0$	$==0$
$\Delta cot\theta$	> 0.02	-
$\lvert\Delta dist\rvert$	> 0.02 cm	-
conversion vertex fit	-	applied

Table 3.5: Overview of the electron identification criteria used in the 7 TeV and 8 TeV datasets.

is larger than the momentum of the track. If the total tracker momentum is larger than the calibrated calorimeter energy by more than three standard deviations, a relaxed muon and fake track search is performed. This is followed by an ordering of the tracks according to the uncertainty on their momentum.

Finally, the tracks are removed one by one until the total momentum is equal to the calorimeter energy or there are no tracks left with a p_T-uncertainty above 1 GeV. The remaining tracks in the block give rise to *ParticleFlow Charged Hadrons*. In the opposite case, when the calorimeter energy exceeds the total momentum of the tracks with a relative difference bigger than the calorimeter resolution, the excess can be labelled as *ParticleFlow Photons* or *ParticleFlow Neutral Hadrons*.

In general if the excess is larger than the total ECAL energy, a photon is identified with this energy, leaving the rest of the excess to a neutral hadron. This favouring of photons in ECAL is justified by the fact that for example in a jet roughly 25% of the energy is coming from photons while only 3% is coming from neutral hadrons.

The remaining ECAL and HCAL clusters in the block, which are not linked to any track, are identified as respectively *ParticleFlow Photons* and *ParticleFlow Neutral Hadrons*.

3.3.7 Jet reconstruction

In Section 3.1.4 the concept of hadronization of quarks was introduced. Since a quark carries a colour charge, QCD confinement dictates it cannot appear as a free particle in collisions and as such it can only be observed as a cascade of hadron production and decay. This cascade is then clustered into a jet by a jet clustering algorithm and forms the experimental representation of a quark. In this Section, the anti-k_T jet clustering technique will be explained to be followed by an explanation on how the jet energy scale and resolution is calibrated to give the best possible description of the true quark properties.

The anti-k_T jet clustering

In the semi-leptonic decay mode of a $t\bar{t}$ pair, at least four quarks are produced in the final state at leading order with additional quarks when considering higher order calculations. This means that jet clustering algorithms are among the most important tools to fully reconstruct the $t\bar{t}$ event topology.

Multiple jet clustering algorithms exist and they can be subdivided into two main classes: the cone algorithms like SisCone [116] and the successive recombination algorithms like the k_T [117], anti-k_T [118] and Cambidge/Aachen [119] clustering algorithms. In the CMS experiment the anti-k_T algorithm is used.

In a successive recombination algorithm, objects i and j are either recombined or clustered into separate jets. To do this the distance d_{ij} between the two objects is introduced as well as d_{iB} which is the distance between the object i and the beam. Two objects i and j are recombined when $d_{ij} < d_{iB}$, in the case when $d_{iB} < d_{ij}$ the object i is considered as a jet and removed from the list. The distance parameters d_{ij} and d_{iB} can be defined in a general way as

$$\begin{aligned} d_{ij} &= min\{k_{T_i}^{2p}, k_{T_j}^{2p}\}\frac{\Delta_{ij}^2}{R^2}, \\ d_{iB} &= k_{T_i}^{2p}, \end{aligned} \tag{3.11}$$

where k_{T_i} and k_{T_j} are the transverse momenta of objects i and j, Δ_{ij} the distance between them in the (y,ϕ)-space, defined as $\Delta_{ij} = \sqrt{\Delta\phi_{ij}^2 + \Delta y_{ij}^2}$ with y the rapidity, and R a dimensionless parameter which can be regarded as the radius of the jet cone. The parameter p is added to the equations to regulate the importance of the energy versus the Δ_{ij} scales.

The before mentioned three successive recombination algorithms are now special cases of the general eqns. (3.11); when the parameter p equals +1, the k_T algorithm [117] is reproduced, when p=0, the equations turn into the Cambridge/Aachen algorithm [119] and finally when p=-1, eqns. (3.11) yield the anti-k_T algorithm.

What is special about the anti-k_T algorithm is that it produces conical jets. This can be seen from eqns. (3.11) when setting p to -1 and considering the special case where a hard object i has no neighbouring hard object within a distance 2R, this object would generate a non-overlapping jet and for any neighbouring soft object j, $k_{T_i} > k_{T_j}$. The condition to recombine two objects i and j $d_{ij} < d_{iB}$ then translates into

$$\Delta_{ij}^2 < R^2, \tag{3.12}$$

while for the k_T algorithm, it would look like

$$\Delta_{ij}^2 < \frac{k_{T_i}^2}{k_{T_j} d^2} R^2. \tag{3.13}$$

Thus it is clear that for the anti-k_T algorithm the parameter R actually defines the radius of the jet cone while for the k_T algorithm, since $\frac{k_{T_i}^2}{k_{T_j}^2}$ can be greater than 1, objects can be clustered outside this radius leading to less conical jets. This behaviour

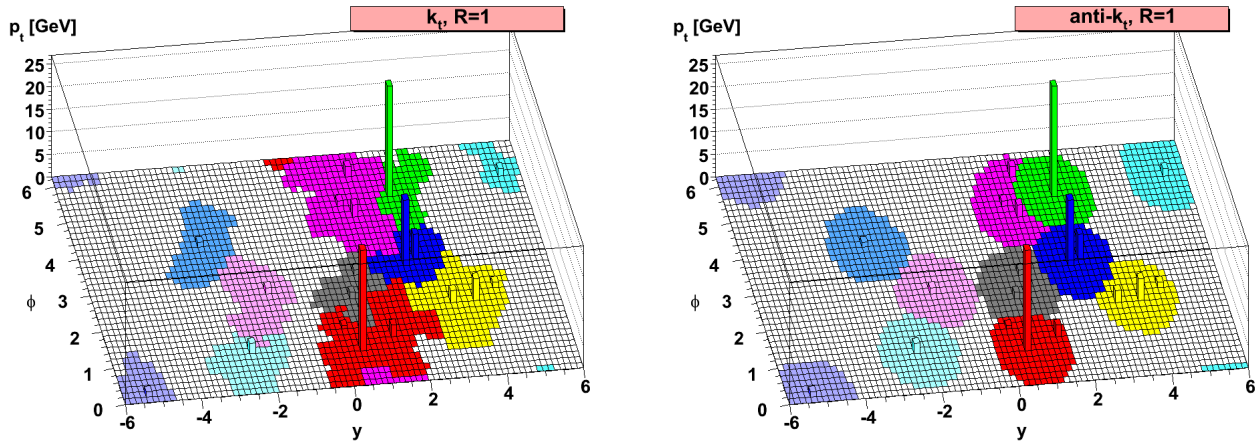


Figure 3.16: A graphical representation of the k_T (left) and anti-k_T jet clustering algorithms applied on a $t\bar{t}$ event generated with HERWIG where a large number of soft particles was added. The anti-k_T algorithm yields conical jets while this is not the case for the k_T algorithm [118].

is illustrated in Figure 3.16 where both algorithms are applied on a $t\bar{t}$ event at the particle level generated with HERWIG while adding a large number of soft particles. This illustration shows the conical jet shape within the anti-k_T algorithm.

In the case that there are two hard objects i and j such that $R < \Delta_{ij} < 2R$, two overlapping jets will be clustered centred around the respective objects. The hardest jet of the two will be conical while the other might have a distorted shape since it loses the objects it shared with the harder jet. This is clearly shown in Figure 3.16 where the hard green jet is perfectly conical eating away an elliptical shape of the dark pink one. When both jets are equally hard, both cones will be clipped and the object they share are separated by a straight line boundary between them. In the last case where $k_{T_i} \approx k_{T_j}$, the shared objects are divided according to a boundary b defined as $\Delta_{ib}/k_{T_i} = \Delta_{jb}/k_{T_j}$. This division of objects is depicted by the blue and yellow jets in Figure 3.16.

Finally, if the two hard objects i and j are closer to each other than R, only one jet will be clustered that is centred around the hardest object. If the difference in k_T is small, the shape becomes more complex as two cones of R'<R are centred around i and j with a final cone of radius R being the union of the two. The red, yellow and dark blue jets in Figure 3.16 are a good example where multiple hard objects lie inside the cone boundary R which is centred around the hardest one.

The most important property of the anti-k_T algorithm, next to its circular jet cones, is that by construction the jet shape can not be significantly altered by soft objects. Only hard objects located close by can alter the jet shape. In the CMS reconstruction the anti-k_T algorithm is implemented with R=0.5.

The objects used as input to the jet clustering can be either partons as provided by PYTHIA or HERWIG, particles or tracks and calorimeter clusters. Typically, jets are clustered at the calorimeter level and can contain information from the inner tracking

detector to improve the jet energy resolution. In the case of the ParticleFlow event reconstruction in CMS a full event description in terms of particles at the production vertex is provided which can be used as input for jet clustering. The usage of ParticleFlow based jets has improved the jet reconstruction significantly in terms of angular and energy resolutions compared to the calorimeter based jets [105, 120, 121].

There are multiple effects that can bias and distort the jet energy determination, or response. First, the energy response of the calorimeters need to be accounted for in the Jet Energy Scale. Moreover, soft particles can be bent out of the jet cone by the strong magnetic field, particles that are not successfully reconstructed or particles produced from pileup interactions, the underlying event and detector noise can end up in the clustered jet. For this reason it is crucial to calibrate the jets to allow for the best possible energy determination.

Jet Energy Scale (JES) calibration

The cornerstone of the Jet Energy Scale calibration technique used in CMS [120, 122, 123] is that the true unbiased jet momentum can be related to the raw, uncorrected, momentum of the reconstructed jet

$$p_i^{true} = \mathcal{C}\ p_i^{raw}, \tag{3.14}$$

where i denotes each component of the jet four-vector and $\mathcal{C}$ is called the calibration factor and its measurement will be described in this section.

A factorized approach can be used to correct the jets. First an *offset correction* is performed to remove any contamination from pileup. Subsequently, a simulation driven correction is derived to flatten the Jet Energy Scale as a function of p_T^{jet} and η^{jet}. Finally, in real collision events, the residual difference with the simulation is absorbed in a residual relative correction as a function of η, and an absolute correction as a function of p_T.

To calculate the jet energy response in simulation, ParticleFlow jets can be compared to GenJets in simulated events. These GenJets are clustered using the same anti-k_t clustering algorithm but use the particles produced by PYTHIA as input whereas the reconstruction level ParticleFlow jets are clustered on ParticleFlow candidates. The reconstructed jet can be matched to the corresponding generator jet by requiring them to be closer than 0.25 in (η, ϕ)-space. Then the jet energy response ($\mathcal{R}$) can be defined as the absolute p_T ratio between the two jets, $\mathcal{R} = p_T^{rec}/p_T^{gen}$.

Offset correction [124, 125]

The first correction is applied to remove the effect of pileup on the jet response. Figure 3.17 provides the jet p_T response as a function of the GenJet p_T for different pileup scenarios. A degradation of the response with increasing number of Primary Vertices demonstrates the need for a pileup correction.

The pileup correction consists of removing additional energy deposits from the raw jet p_T. Therefore an absolute offset correction, $\mathcal{C}_{Offset}$, is introduced.

$$p_T^{Corr} = p_T^{Uncorr} - \mathcal{C}_{Offset}. \tag{3.15}$$

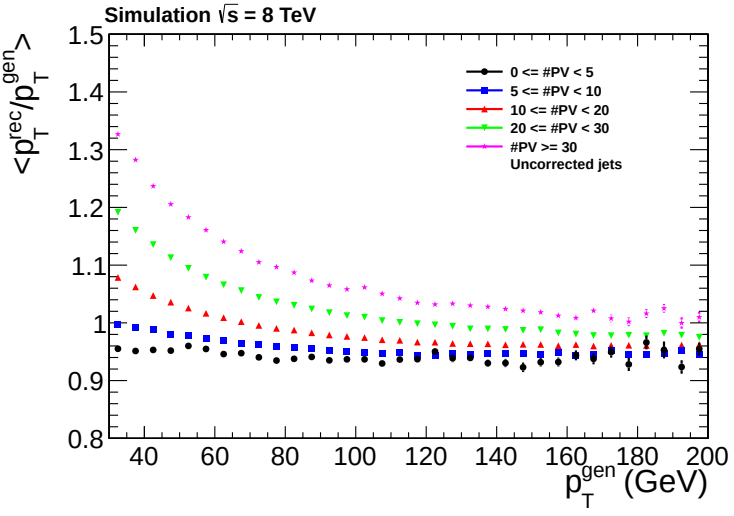


Figure 3.17: Jet energy response for uncorrected jets as a function of the transverse momentum of the GenJet for different reconstructed primary vertex multiplicities.

The offset correction factor is a function of the effective jet area, A_j, and the average p_T density, ρ, that represents the soft jet activity in the event. The effective jet area is measured by injecting a large number of soft four-vectors into the event without changing the properties of the true jets prior to the jet clustering. The *effective jet area* is then defined as the maximal spread of these soft particles in (y,ϕ) within the jet.

The p_T-density per unit area, ρ, can be determined in the same event by re-clustering the jets with the k_T algorithm with distance parameter R=0.6. This algorithm has the benefit of clustering a large number of soft particles covering the full (y,ϕ)-space. The average p_T-density can then be defined as the median of $p_{T,j}/A_j$ where j loops over all jets in the event and A_j is their effective area.

In the 7 TeV analysis, this correction factor is given as a function of A_j and ρ

$$\mathcal{C}^{7\,TeV}_{Offset}(\rho, A_j, \eta) = A_j \left(p_0(\eta) + p_1(\eta) \cdot N_{PV}(\rho) + p_2(\eta) \cdot (N_{PV}(\rho))^2\right), \tag{3.16}$$

where N_{PV} describes the dependence of the primary vertex multiplicity on ρ. The parameters $p_{i=0,1,2}(\eta)$ are determined in bins of η by looking at the offset in a cone of R=0.5 in events selected by a random trigger.

In the 8 TeV analysis, the offset correction factor was refined to better cope with the high pileup environment. As opposed to $\mathcal{C}^{7\,TeV}_{Offset}$, the correction factor is now also a function of the jet transverse momentum and pseudo-rapidity. The correction factor is obtained by using a jet-by-jet matching between identical QCD events with and without pileup production and is defined as

$$\mathcal{C}^{8\,TeV}_{Offset}(\rho, A_j, \eta, p_T) = A_j \left(p_0(\eta) + p_1(\eta) \cdot \rho \cdot \left(1 + p_2(\eta) \log\left(p^{PFjet}_{T,raw}\right)\right)\right). \tag{3.17}$$

The parameters $p_{i=0,1,2}(\eta)$ are obtained by fitting the offset between the matched jets from the pileup and no-pileup samples. As opposed to the offset correction at 7 TeV,

no explicit dependence on the number of reconstructed Primary Vertices is introduced in the 8 TeV offset correction. However, this dependence is still present through ρ.

Since the 8 TeV correction factor is purely simulation based, a residual correction factor is applied to jets in data. This residual correction is determined from the remaining p_T offset in events selected by random triggers after the jets were corrected with the simulation based correction.

Multiple systematic uncertainties enter these correction and are added into quadrature. In the 7 TeV analysis, the main sources of uncertainty originate from a residual bias on the measurement of the p_T offset and the residual differences in offset between data and simulation. Furthermore, the method is sensitive to out-of-time pileup for which a systematic is added.

The 8 TeV correction factor has two main components of systematic uncertainty. First a residual bias on the method was observed in simulation and thus quoted as a systematic uncertainty. Secondly, since the method is purely simulation driven and residual correction has to be applied in data, 20% of the observed p_T offset is added as an uncertainty.

Figure 3.18 now shows the jet response as a function of p_T in different Primary Vertex multiplicity bins after the offset correction was applied. The dependence on the number of primary vertices has essentially vanished proving that the correction works correctly.

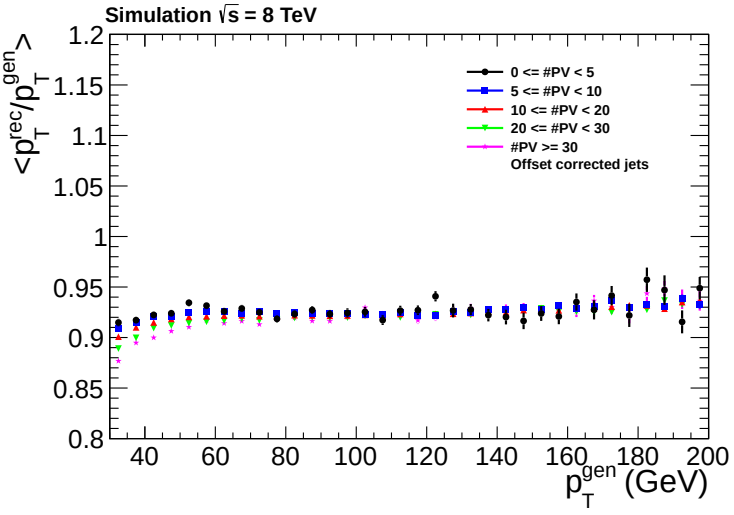


Figure 3.18: Jet energy response for offset-corrected jets as a function of the transverse momentum of the GenJet for different reconstructed primary vertex multiplicities.

Simulation-driven calibration

After applying the offset correction, the response is independent of the primary vertex multiplicity but the response is far from unity. To correct for this effect, a simulation based correction is applied to make the response equal to 1 in the full p_T range. This

correction is derived also as a function of η, which makes the response dependence on η flat as well.

The correction factor is derived by measuring the jet energy response as a function of p_T and η in QCD multijet events and is defined as

$$C_{MC}(p_T, \eta) = \frac{1}{\langle \mathcal{R}(p_T, \eta) \rangle}. \tag{3.18}$$

This correction is then applied on top of the offset correction to provide a jet energy response close to 1 as shown in Figure 3.19.

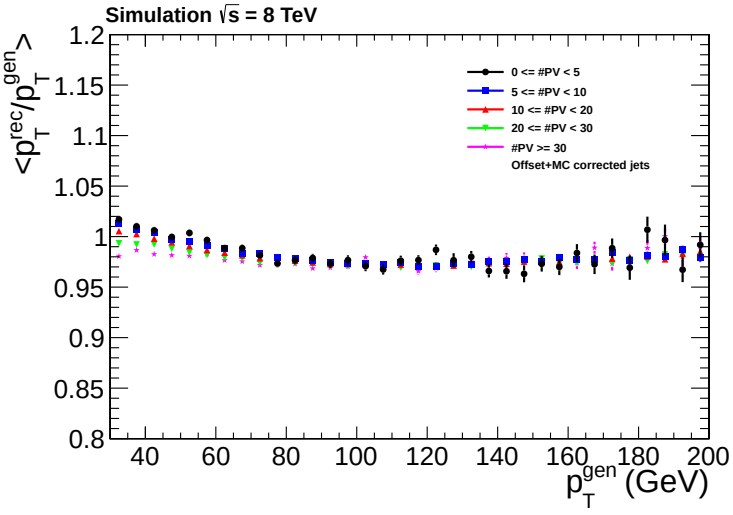


Figure 3.19: Jet energy response for offset+MC-corrected jets as a function of the transverse momentum of the GenJet for different reconstructed primary vertex multiplicities.

The main systematic uncertainty on this correction is the flavour composition of the sample. While gluon and heavy quarks produce much more soft particles compared to light quarks, their energy response is considerably worse. Hence, the total jet energy response will be influenced by the composition of the sample. The corrections were determined on simulated QCD multijet events dominated by gluons. The effect of the flavour dependent response is then determined by generating events with two different hadronization models, namely PYTHIA versus HERWIG, and quoting the difference as an uncertainty.

This correction is also the final correction for simulated jets. In the following, two residual corrections for data will be discussed correcting for any data to simulation differences.

Residual relative correction [126]

Because the previous correction was determined solely on simulation, a residual relative η-dependent correction is determined to correct for a residual η dependence in the jet

response in data. In di-jet events, the jet in the barrel ($|\eta| < 1.3$) is used as a reference and a relative correction in bins of η^{jet} can be determined by requiring p_T-balance between the two jets.

This relative correction is subject to different sources of uncertainties. First the statistical uncertainty is propagated as a systematic. Furthermore, the method depends on the resolution of the two jets. Hence the measured Jet Energy Resolution (JER) uncertainty is taken into account.

In the 7 TeV analysis, a systematic uncertainty was added for the Final State Radiation effects. This effect was less pronounced in the 8 TeV analysis due to improvements in the method.

For the 8 TeV correction, a residual p_T dependence was observed leading to additional systematic uncertainties.

Residual absolute correction [126]

The residual absolute energy response can be measured in data by using the p_T-balance of a back-to-back jet-photon or jet-Z boson pair in Z/γ+jet events. The jet is restricted to the barrel detector ($|\eta| < 1.3$) but because of the previous residual correction, the numbers can safely be extrapolated to the full η-range. The p_T balance between the jet and the precisely measured photon or Z boson allows to measure the absolute Jet Energy Response.

The uncertainty on the method of the absolute energy response correction is twofold. First, the correction is measured for a fixed jet p_T assuming that the correction does not depend on the transverse momentum. To account for any residual p_T-dependence, an additional uncertainty is added. The first part of this uncertainty is determined by varying the calorimeter single particle response in the simulation within its uncertainties. Additionally, the difference between two different fragmentation and hadronization models with different intrinsic p_T-dependence (PYTHIA vs HERWIG) is added as well.

The second source of systematic uncertainty comes directly from the lepton and photon energy scales as the absolute energy response is determined by the p_T balancing of the jet with respect to the photon/Z boson.

Combining the corrections and uncertainties outlined above, an average total Jet Energy Scale uncertainty of 1.8% in the barrel ($|\eta| < 1.2$) and 2.3% in the endcaps ($1.2 < |\eta| < 2.3$) is obtained for a 50 GeV jet. For a jet of 100 GeV, the average total uncertainty decreases to 1.1% and 1.4% respectively.

Jet Energy Resolution (JER)

After fully correcting the jets with the corrections introduced in the previous part, the jet energy resolution can be studied and compared with data. The resolution can be measured within di-jet events using a p_T-balancing [120].

Two main sources of systematic uncertainties act on this measurement. The p_T-balance between the two jets is a convolution of an intrinsic balance with an imbalance caused by particle-level effects such as fragmentation effects that lead to radiation

not confined within the jet cone. A ±25% variation on the imbalance is added as a systematic uncertainty.

The second source of systematic uncertainty originates from an overall bias observed between the measurement on simulation and the true jet resolution. The measurement is corrected for the bias and to be conservative, 50% of this bias has been quoted as the uncertainty.

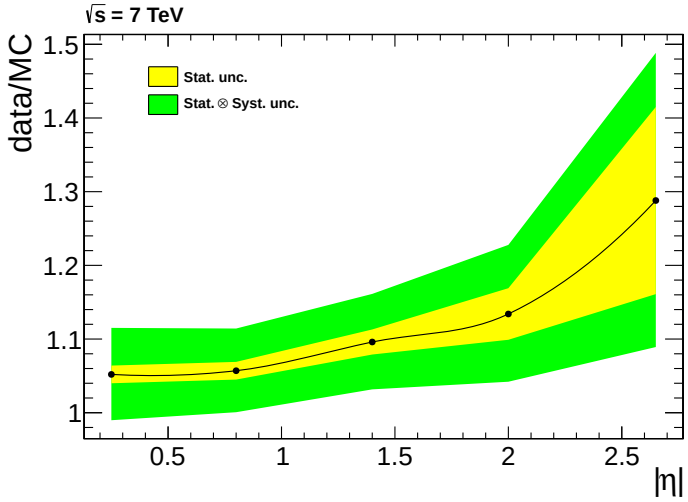


Figure 3.20: The bias-corrected jet energy resolution difference between data and simulation as a function of the jet pseudo-rapidity [127].

Compared to simulation, a considerably worse jet energy resolution was found in data. Figure 3.20 shows the data to simulation scale factors for the jet energy resolution as a function of η ranging from 5 to 29% in the forward region. The simulated events both at 7 and 8 TeV are corrected with this scale factors to better match the data and the systematic uncertainty on them will be propagated into the analysis.

Within simulated $t\bar{t}$ events, the Jet Energy Resolution can also be determined by looking at the width of the p_T^{rec}/p_T^{gen} distribution. The JER is found to be between 10 and 20% for jets with a transverse momentum ranging from 40 to 200 GeV.

Jet identification

Just as for the lepton objects, jets need to pass a set of identification criteria to ensure that they are not mis-reconstructed. The identification criteria are summarised in Table 3.6 and require that each jet consists of at least 2 particles. Furthermore, a jet is not retained if its energy is purely attributed to either neutral hadrons or photons. When the jet lies within $|\eta|$ <2.4, i.e.: within the tracker acceptance, properties of charged particles can be used as well. The latter jets need to contain at least 1 charged particle and some charged hadrons. If a jet passes all these criteria it is retained for the physics analysis.

Criterion	$\sqrt{s} = 7$ TeV	$\sqrt{s} = 8$ TeV
Number of constituents	>1	>1
Neutral EM Energy Fraction	<99%	<99%
Neutral Hadron Energy Fraction	<99%	<99%
$\|\eta\|$ <2.4 Charged EM Energy Fraction	<99%	<99%
$\|\eta\|$ <2.4 Charged Hadron Energy Fraction	>0%	>0%
$\|\eta\|$ <2.4 Charged particle multiplicity	>0%	>0%

Table 3.6: Overview of the jet identification criteria used in the 7 TeV and 8 TeV datasets

3.3.8 Missing Transverse Energy reconstruction ($\vec{\not{E}}_T$)

In the semi-leptonic $t\bar{t}$ decay, one of the W bosons decays into a lepton and its associated neutrino. The latter is a very weakly interacting particle and will escape detection. As a consequence, the neutrino energy can not be directly measured. Nevertheless, the energy can be inferred by imposing momentum conservation in the transverse plane. The missing fraction in the transverse plane is called the Missing Transverse Energy (MET,$\vec{\not{E}}_T$) [128] and is used to infer the presence of a neutrino.

The $\vec{\not{E}}_T$ is defined as minus the vectorial sum of the transverse momenta of all objects.

$$\vec{\not{E}}_{\mathrm{T}} = -\sum_i \vec{p}_{\mathrm{T,\,i}}. \tag{3.19}$$

This sum can be subdivided into two categories: the clustered objects (jets) and the unclustered objects (leptons)

$$\vec{\not{E}}_{\mathrm{T}} = -\sum_{\mathrm{jet}} \vec{p}^{\,\mathrm{uncorr.}}_{\mathrm{T,\,jet}} - \sum_{i \notin \mathrm{jets}} \vec{p}_{\mathrm{T,\,i}}, \tag{3.20}$$

where uncorrected jets are used to calculate the $\vec{\not{E}}_T$. Since jets are used to calculate the $\vec{\not{E}}_T$, the corrections to the jet energy scale play a role in its determination. Therefore, the jet energy scale corrections are propagated to the $\vec{\not{E}}_T$. Since jets aren't properly corrected below a transverse momentum of 10 GeV, the jets are subdivided even further in jets that can be corrected ($p_T > 10\ GeV$) and jets that remain uncorrected ($p_T < 10\ GeV$).

$$\vec{\not{E}}_{\mathrm{T}} = -\sum_{\substack{\mathrm{jet} \\ \vec{p}^{\,\mathrm{corr}}_{\mathrm{T,\,jet}} > 10\mathrm{GeV}}} \vec{p}^{\,\mathrm{uncorr.}}_{\mathrm{T,\,jet}} - \sum_{\substack{\mathrm{jet} \\ \vec{p}^{\,\mathrm{corr}}_{\mathrm{T,\,jet}} < 10\mathrm{GeV}}} \vec{p}^{\,\mathrm{uncorr.}}_{\mathrm{T,\,jet}} - \sum_{i \notin \mathrm{jets}} \vec{p}_{\mathrm{T,\,i}}. \tag{3.21}$$

The simulation based and residual jet energy scale corrections, explained in the jet reconstruction section, should now be propagated to the $\vec{\not{E}}_T$. This is referred to as

type-I corrected $\vec{\not{E}}_T$. The correction factor is obtained by removing the offset correction for pileup from the fully corrected jet:

$$\vec{C}_{\mathrm{T}}^{\mathrm{type1}} = - \sum_{\substack{\mathrm{jet} \\ \vec{p}_{\mathrm{T,\,jet}}^{\mathrm{corr}} > 10\mathrm{GeV}}} \left(\vec{p}_{\mathrm{T,\,jet}}^{\mathrm{corr}} - \vec{p}_{\mathrm{T,\,jet}}^{\mathrm{offset}} \right), \tag{3.22}$$

such that

$$\vec{\not{E}}_T{}^{corr} = \vec{\not{E}}_T + \vec{C}_{\mathrm{T}}^{\mathrm{type1}}. \tag{3.23}$$

In simulated $\mathrm{t\bar{t}}$ events, the $\vec{\not{E}}_T$ resolution can be studied using the width of the $\not{E}_T$ and the generated neutrino p_T. For $\not{E}_T$ below 40 GeV, a resolution of 13% is obtained at 7 TeV increasing to roughly 40% for 200 GeV $\not{E}_T$. At 8 TeV the resolution ranges between 16 and 45% which is slightly larger than at 7 TeV due to the increased pileup.

As a consequence of propagating the jet energy scale corrections to the $\vec{\not{E}}_T$, the jet energy scale uncertainty on results like the $\mathrm{t\bar{t}}$ cross section in Chapter 7 will contain a component originating from the jets as well as a component originating from the $\vec{\not{E}}_T$. The same goes for the jet energy resolution as the smearing of the jets, explained in the previous section, is also propagated to the $\vec{\not{E}}_T$.

Finally, one additional uncertainty arrises on the unclustered energy. The unclustered energy is defined as all energy that is not clustered and cannot be attributed to leptons. Additionally, jets with a transverse momentum below 10 GeV are considered to be part of this category. A 10% flat uncertainty is taken into account on the unclustered energy.

3.4 Bottom quark identification

The identification of jets originating from b quarks is important for many analyses and is extensively used in this thesis. In $\mathrm{t\bar{t}}$ events, at least two b quarks are produced in the final state. Evidently, being able to efficiently identify b jets is crucial in selecting signal events among the multitude of background processes.

In this section different b-tagging approaches [2] are outlined using properties of displaced tracks in Section 3.4.1, secondary vertices in Section 3.4.2 and the presence of soft leptons from B hadron decays in Section 3.4.3. Finally, all these ingredients are grouped in the combined and super-combined algorithms explained in Section 3.4.4.

3.4.1 Track impact parameter significance based algorithms

The properties of reconstructed tracks associated to the jet can be used to discriminate b jets from other types of jets. The tracks are reconstructed using the standard track reconstruction in the inner tracking system albeit with some additional selection criteria. To reject poorly reconstructed tracks, every track needs to have a momentum exceeding 1 GeV with a normalised χ^2 below 5. In addition, eight tracking hits are required to be associated to the track as well as two hits in the pixel detector since it holds the highest discriminating power. To remove contributions from long lived

particles, an additional loose selection is performed on the absolute value of the transverse and longitudinal distance between the point of closest approach and the primary vertex. Finally, the track needs to be within $\Delta R < 0.5$ of the jet direction at the vertex.

Since B hadrons are produced in the fragmentation of b quarks, displaced tracks with respect to the primary vertex are expected to be found within the b quark jets. The displacement of a track can be quantified as is shown in Figure 3.21. The track impact parameter (IP) is defined as the distance between the jet vertex and the tangent to the track extrapolated from the point of closest approach between the track and the jet axis. Because of p_T and η dependence of the impact parameter resolution, the impact

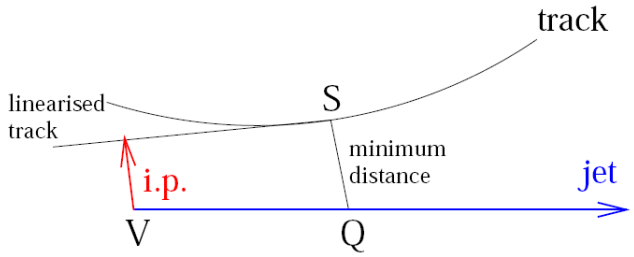


Figure 3.21: Definition of the track impact parameter (IP) compared to the jet axis.

parameter is divided by its uncertainty to yield the impact parameter significance. The *sign* variable is defined by the scalar product of the vector along the jet axis and the vector from the vertex to the point of closest approach. For particles produced along the jet axis, *sign* will be positive while for random non-prompt tracks *sign* can be either positive or negative.

Track counting b tagger

The track counting b tagging algorithm exploits the impact parameter significance as a discriminating variable between b jets and light jets. First, all the tracks associated to the jet are sorted in decreasing significance. The IP significance of the second highest IP significance track is used as a discriminating value. This discriminator is labeled the *Track Counting High Efficiency* (TCHE) discriminator. To increase purity at the cost of efficiency, the IP significance of the third track can be used as well giving rise to the *Track Counting High Purity* (TCHP) discriminator.

Both discriminators are shown in Figure 3.22, comparing b jets to light and gluon jets. For b jets this distribution has a long positive tail whereas the light and gluon contributions are centred around smaller values. This defines the discriminating power of these algorithms.

Jet Probability b tagger

A second class of track IP significance based b tagging algorithms is formed by the *jet probability* discriminators. Unlike the simple track counting taggers, it uses all tracks associated to the jet to calculate a probability that a set of tracks originates from the

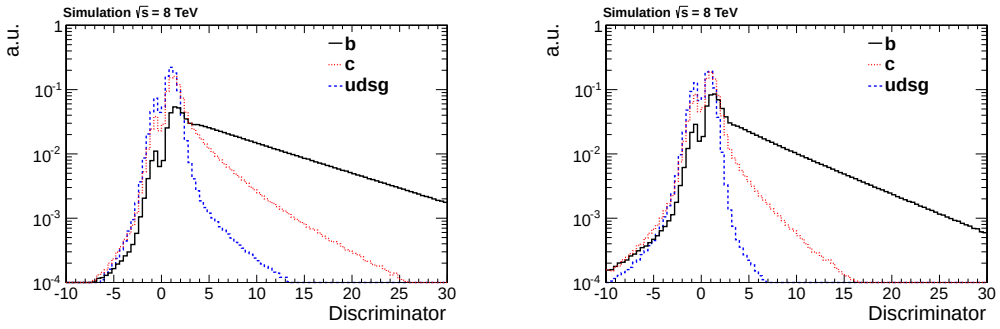


Figure 3.22: The b-tagging discriminator distribution for the TCHE algorithm (left) and the TCHP algorithm (right).

primary vertex. This technique starts by calculating the individual track probabilities P_i

$$P_i(S) = sign(S) \int_{|S|}^{\infty} R(x)dx. \tag{3.24}$$

which are a function of the signed IP significance (S) and the normalised IP significance distribution, R(x), that can be estimated from data using negative signed tracks assuming that the distribution has the same shape as for positive signed tracks.

Then, the likelihood for a set of tracks in the jet to originate from the primary vertex, P_{jet}, can be defined as

$$P_{jet} = \Pi \sum_{j=0}^{N-1} \frac{(-ln\Pi)^j}{j!}, \tag{3.25}$$

where

$$\Pi = \prod_{i=1}^{N} max(P_i, 0.005). \tag{3.26}$$

To reduce the effects of poorly reconstructed single tracks on the final discriminator value, a cut-off value of 0.5% was introduced for P_i.

The Jet Probability (JP) discriminator can finally be defined as $-ln\ P_{jet}$. Additionally, a more stringent Jet B Probability (JBP) algorithm is introduced alongside that gives more weight to the highest impact parameter significance tracks. The discriminator distributions are shown in Figure 3.23, showing very good separating power for b jets which is even more pronounced in the JBP discriminator compared to the JP.

3.4.2 Secondary vertex based algorithms

Another useful property originating from the lifetime of the B hadron produced in b quark fragmentation is the presence of a secondary vertex inside a b jet. The properties of this secondary vertex can be exploited to identify b jets. In a first iteration, all vertices are reconstructed using the Adaptive Vertex fitter [63]. Then, each reconstructed

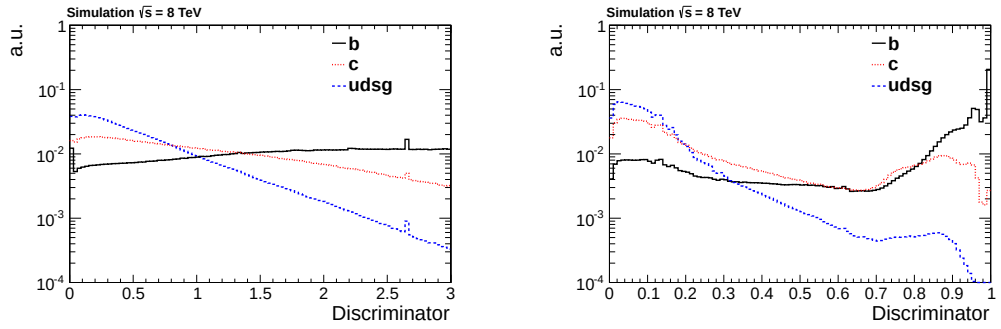


Figure 3.23: The b-tagging discriminator distribution for the JP algorithm (left) and the JBP algorithm (right).

track inside the event is attributed a weight based on the distance to the primary vertex. All tracks with a weight above 0.5 are removed from the event as they are expected to be compatible to the primary vertex and the vertex fit is restarted. The iterative procedure continues until no new vertices are found.

To use this secondary vertex information in the identification of b jets, a *Simple Secondary Vertex Algorithm* (SSV) has been developed using the flight distance significance (D/σ_D) as the sensitive variable which is defined as the distance between the reconstructed secondary vertex and the primary vertex weighted by its uncertainty. Just as for the track based tagger, two variants of the SSV algorithm exist. The first requires the secondary vertex to have two assigned tracks yielding a higher efficiency and is called the *Simple Secondary Vertex High Efficiency (SSVHE)* discriminator. To increase purity, at the cost of efficiency, the *Simple Secondary Vertex High Purity* (SSVHP) discriminator requires the association of three tracks. The actual discriminator value is calculated as $log(1 + D/\sigma_D)$ and is shown in Figure 3.24 for both variants.

To boost the overall b purity of the algorithm, additional requirements are set on the secondary vertex. First, the secondary vertex can only share 65% of its tracks with the primary vertex. In addition, to ensure high purity in the secondary vertex reconstruction, the significance of the radial distance between the primary and secondary vertex has to exceed three times its uncertainty. Next, the flight direction for each vertex has to be within the jet, $\Delta R < 0.5$. Finally, to remove contamination of long-lived mesons, well separated (>2.5 cm) secondary vertices are rejected if their mass exceeds 6.5 GeV.

3.4.3 Soft lepton based algorithms

Additional to the distinct characteristics of displaced tracks and secondary vertices, jets originating from b jet hadronization have another interesting property that can be averted to identify them. As was discussed in Section 3.1.5, the B hadron has a branching ratio of 10% to decay into leptons. These soft non-isolated leptons can be efficiently reconstructed by the ParticleFlow event reconstruction and hence provide a distinct signature specific to tag b jets [129]. Four variables are combined in a Neural Network to build the *Soft Muon Tagger* (SM) and *Soft Electron Tagger* (SE):

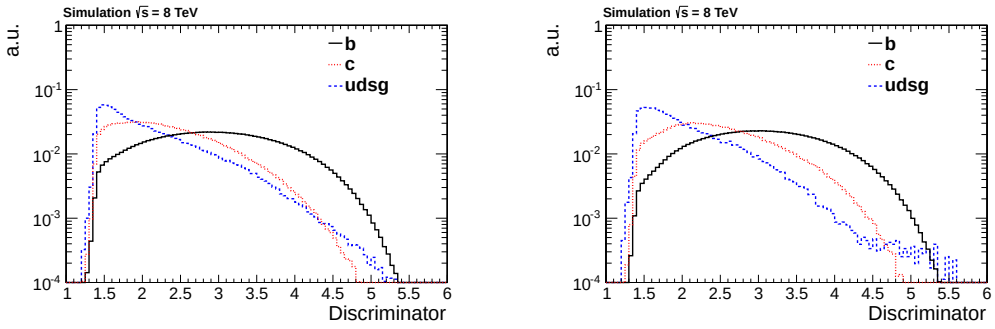


Figure 3.24: The b-tagging discriminator distribution for the SSVHE algorithm (left) and the SSVHP algorithm (right).

- Three-dimensional impact parameter significance of the lepton compared to the jet axis.
- Momentum ratio: the ratio of the lepton momentum and the jet momentum.
- The separation of the jet and the lepton in (η, ϕ)-space.
- p_T^{rel}: transverse momentum of the lepton relative to the jet axis.

The Neural Network is trained on simulated $t\bar{t}$ events for the b quark component and on multijet events for light quarks. The output of the Neutral Network is shown in Figure 3.25 for both muons and electrons. Since this is a Neural Network output, it is distributed between 0 and 1 where a value close to 0 favours the non-b jet hypothesis and a value close to 1 most probably identifies a b jet.

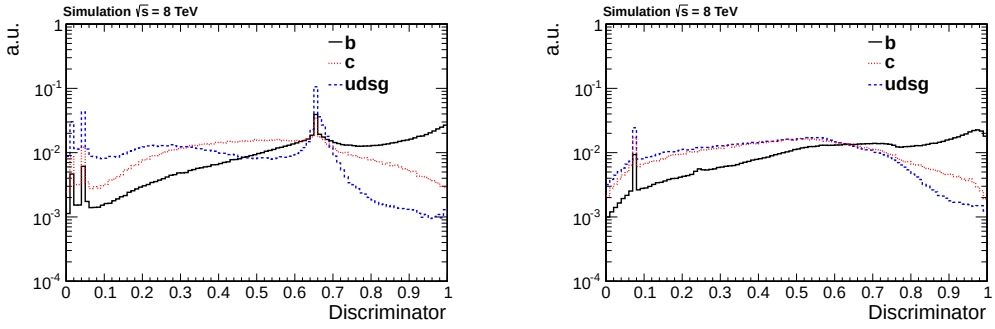


Figure 3.25: The b-tagging discriminator distribution for the SM algorithm (left) and the SE algorithm (right).

3.4.4 Combined secondary vertex algorithm

In the *Combined Secondary Vertex* (CSV) b-tagging algorithm, multiple track based and secondary vertex related variables are combined to form a single discriminating

variable. Since the vertex reconstruction is only 65% efficient, putting a limit on the efficiency of the Simple Secondary Vertex algorithms, the CSV has an added value of providing a discriminator value even in the absence of a reconstructed secondary vertex.

The usage of secondary vertex based variables requires the reconstruction of a good secondary vertex. Nevertheless, when no good fit exists for a vertex where at least two tracks are assigned with IP significance exceeding 2 it is classified as a pseudo-vertex for which a subset of vertex properties can still be calculated.

- The vertex category: reconstructed vertex, pseudo-vertex or no vertex.
- Flight distance significance in the transverse plane: the same variable as used for the Simple Secondary Vertex algorithm but calculated in the transverse plane.
- The mass of the secondary vertex: the invariant mass of all charged particles associated to the vertex. This variable can discriminate as particles from the B hadron decay are more massive than from D hadrons.
- Number of tracks associated to the vertex: the average number of tracks can be significantly higher in the presence of B hadron decays.
- η_{rel}: pseudo-rapidity of each track associated to the secondary vertex relative to the jet axis.
- The fraction of the total sum of track energies carried by the tracks associated to the secondary vertex.
- The transverse impact parameter significance of the track that, by associating it to the secondary vertex, brings the mass of the vertex above the charm threshold of 1.5 GeV.
- The number of tracks in a jet.
- The three-dimensional impact parameter significance for all tracks in the jet.

When not even a pseudo-vertex is reconstructed, only the last two variables can be used. The likelihood ratio variable can then be defined as

$$LR = \frac{\mathcal{L}_b}{\mathcal{L}_b + \mathcal{L}_{c/udsg}} \tag{3.27}$$

$$\mathcal{L}_{b,c/udsg} = \prod_i p_{b,c/udsg}(x_i) \tag{3.28}$$

where $p(x)$ represents the probability density function and the product runs over all variables. The Likelihood Ratio is then trained and evaluated on charm jets and light jets (udsg) separately. The resulting CSV discriminator is then calculated by summing the charm and light discriminators using a weight of 25% and 75% respectively to match the non-b flavour composition in $t\bar{t}$ events. The discriminator is shown in the left panel of Figure 3.26 and is distributed between 0 and 1 where a value of 1 is more

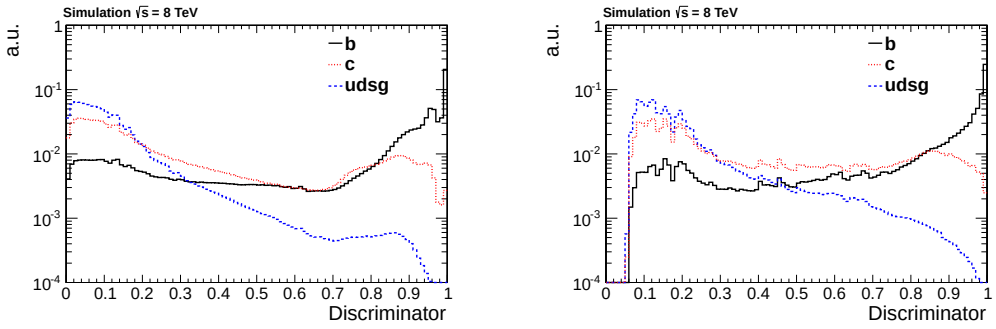


Figure 3.26: The b-tagging discriminator distribution for the CSV algorithm (left) and the CSVR algorithm (right).

likely obtained by a b-jet while the low discriminator region is populated by light and charm jets.

The training framework for the combined secondary vertex has recently been updated including an updated training. This training has been propagated into a second *Retrained Combined Secondary Vertex* (CSVR) discriminator as is shown on the right side in Figure 3.26 where for discriminator values moving towards 1 a good separation of b-jets is obtained.

This new training procedure can also be used for other new taggers; new taggers have been added combining information from tracks, vertices and leptons called super-taggers. Among the newly introduced super-taggers, a combined CSV+JP, CSV+SL and CSV+JP+SL is created.

3.4.5 b-tagging performance

The performance of b tagging algorithms is usually characterised by the efficiency of tagging an actual b jet and the mis-tagging efficiency or fake rate. In simulated events, both performance measures can be easily extracted since the flavour of jets can be determined by matching a jet to its nearest parton. A performant b tagger will keep the tagging efficiency high while the mis-tagging efficiency should remain as low as possible.

Figure 3.27 shows the b tagging efficiency versus the mis-tagging rate for various algorithms. Some interesting observations can be made from this plot. First, for SSVHE, the b tagging efficiency does not exceed 65% showing the effect of the vertex reconstruction efficiency. Moreover, for both soft lepton taggers, the maximal efficiency lies in the range of 10% corresponding to the branching ratio of B hadrons decaying into leptons.

The CSV taggers clearly outperform the simpler algorithms since they are tuned to perform the best possible performance over the full efficiency range. The CSV tagger provides a b-tagging efficiency of roughly 80%, 70% and 52% for its Loose, Medium and Tight Working Points, corresponding to a total mis-tagging efficiency of 15%, 3% and 1%. Finally, the new training for the recently added CSVR tagger has clearly improved the performance.

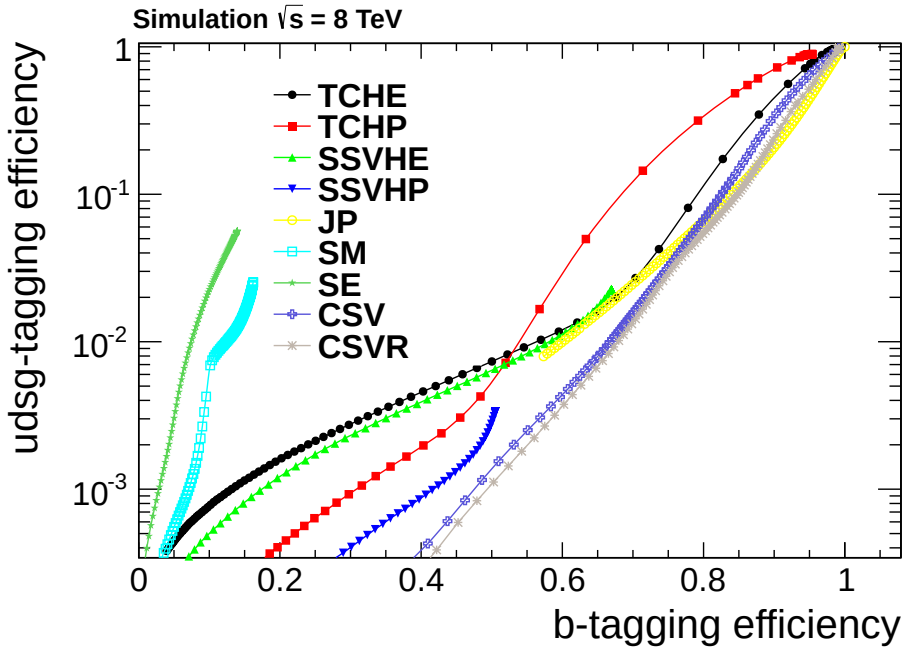


Figure 3.27: The b-tagging efficiency as a function of the light mis-tagging rate for various b tagging algorithms.

The b tagging performance can also be measured from data. Different methods [2] have been developed in QCD multijet and $t\bar{t}$ events to measure the b tagging efficiency as well as the mis-tagging efficiency. The methods are briefly described in the following.

The first category of measurements uses the distinct topology of muons inside jets as a key property to identify jets in a QCD multijet sample where the non-isolated muon is required to be well within the jet cone by imposing $\Delta R(\mu, jet) < 0.4$.

- Efficiency measurement with p_T^{rel}: in this method, dijet events are selected where one of the two jets contains a muon inside the jet. The p_T^{rel} distribution is directly used to estimate the number of b, c and light quarks in the sample by fitting the data distribution to templates from simulation split in flavour. This fit is then repeated after applying a b tag, on the jet containing the muon, defining a subset of the previous sample, yielding an estimate of the number of b jets after tagging. This allows to directly calculate the efficiency of any given working point.

- Efficiency measurement with the three-dimensional impact parameter: this method is essentially the same as the previous one albeit that the p_T^{rel} distribution is interchanged with the three-dimensional impact parameter of the muon track.

- System8 method: the System8 method creates two different sets of jet samples: before and after applying b tagging. Then, using different discriminating variables, a set of 8 equations can be constructed from the total number of events in both samples to provide the b tagging efficiency.

- Reference lifetime method: The Reference lifetime method is applied in dijet events where one jet contains a muon. The latter criterion already entails an above average b jet purity but this is even further increased by requiring an additional jet to be reconstructed and tagged by the JPM working point. This creates a pure b jet sample where the efficiency can readily be measured for any algorithm except JP.

Another set of measurements is performed in $t\bar{t}$ events. Since $\mathcal{R}(t \to Wb)$ is found to be very close to 1, top quark pairs provide a unique vast sample of b quarks to study the b tagging efficiency.

- Profile Likelihood Ratio: In this method, a Profile Likelihood Ratio fit is performed on the two-dimensional distribution of the number of reconstructed jets and the number of b tagged jets in di-lepton $t\bar{t}$ events. This likelihood function is then minimised among others to the b-tagging efficiency scale factor which is defined as $\frac{\epsilon^{data}}{\epsilon^{MC}}$ where the denominator is the efficiency determined on true b jets in the simulation.

- Flavour Tag Consistency Method: The FTCM method is performed in l+jets $t\bar{t}$ events requiring consistency between the measured number of b tags and the expected number.

- Flavour Tag Matching Method: The FTMM method is analogous to the FTCM method and is carried out in di-lepton $t\bar{t}$ events.

In addition, the mis-tagging efficiency can be derived using negative tag rates [2] from so-called inverted b tagging algorithms. These algorithms are duplications from the regular b tagging algorithms with the exception that they use for example negative impact parameter tracks and negative decay length vertices. The mis-tagging efficiency can then be calculated as

$$\epsilon_{data}^{mistag} = \epsilon_{data}^{-} R_{light}, \tag{3.29}$$

where

$$R_{light} = \frac{\epsilon_{MC}^{mistag}}{\epsilon_{MC}^{-}} \tag{3.30}$$

and ϵ_{data}^{-} is the negative tag rate obtained from data.

In Chapter 5, a different method is presented to estimate the b tagging efficiency from l+jets $t\bar{t}$ events. This method uses a variable called the jet-lepton mass to define a b jet enriched and b jet depleted sample allowing to reconstruct the discriminator distribution for pure b jets. Next to measuring the efficiency for any tagger at any working point, this method can be used to reconstruct the discriminator for non b jets in Chapter 6.

Chapter 4

Event selection

In the previous Chapter, the object reconstruction was discussed including a series of identification criteria to be fulfilled by each object. These criteria serve the purpose of separating for example a real reconstructed electron from charged hadrons misidentified as electrons. Nevertheless, additional criteria are needed at the object and event level to select only these events that are likely to contain an interesting collision, namely $pp \to \mathrm{t\bar{t}}$.

In Section 4.1, the event selection criteria are outlined for both μ+jets and e+jets events. Subsequently, the results of the event selection are discussed in Section 4.2. Using selected events, it was observed that the top quark transverse momentum spectrum was to soft compared to data. This leads to a mismatch for all its decay products that can be mitigated by applying a reweighting to the simulated events. This reweighting will be introduced in Section 4.3.

Finally, the analyses in the next chapter rely on an estimate which of the four leading reconstructed jets originate from the b jet in the $t \to W(\to l\nu_l)b$ decay.

4.1 Selection of l+jets $\mathrm{t\bar{t}}$ events

In Figure 4.1, an event display is shown for a typical $\mathrm{t\bar{t}} \to \mu\nu_\mu bqqb$ event. Although the W boson decays into a muon and a neutrino, this drawing represents the more general l+jets topology. To select this type of event, an isolated muon or electron is looked for, accompanied by at least four jets and some $\not\!\!E_T$. The full event selection will be described below and will be applied in exactly the same way to data and simulation. Nevertheless, some additional event filters have been implemented in data to mitigate events recorded from noise and to clean high tails of the $\not\!\!E_T$ distribution. For the purpose of this analyses, these filters have little effect.

4.1.1 Online trigger

Before the event selection is applied, the event can be pre-selected by requiring a certain trigger to be fired. The triggers used in this thesis in general require the presence of a muon with or without the requirement of jets.

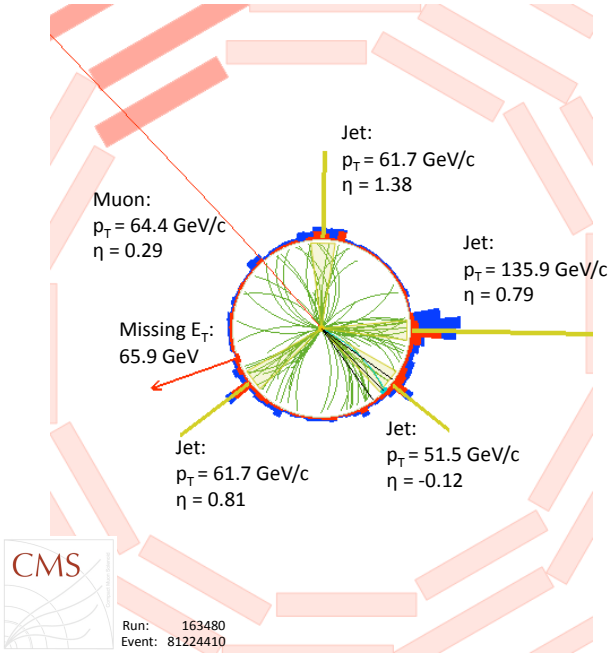


Figure 4.1: Event display for a $t\bar{t} \to \mu\nu_\mu bqqb$ candidate recorded at $\sqrt{s} = 7$ TeV.

In the muon channel, the *HLT_IsoMu24_eta2p1* trigger has been chosen which is available throughout the 7 and 8 TeV data taking period. This trigger requires an isolated muon with a p_T exceeding 24 GeV within $|\eta| < 2.1$ which provides a good trade-off between the trigger rate and the lepton momentum threshold.

In the electron channel, the triggers are more complicated compared to the muon case. During the 7 TeV data-taking, the isolated single electron triggers could not be kept at a sensibly low p_T-threshold due to high trigger rates. For this reason it was decided to use *HLT_Ele25_CaloIdVT_TrkIdT_CentralTriJet30* with the requirement of one electron with p_T >25 GeV and three jets with p_T >30 GeV in the central region. Later in the data taking, an isolation criterion was also added by using *HLT_Ele25_CaloIdVT_CaloIsoT_TrkIdT_TrkIsoT_TriCentralJet30*. Finally, the trigger was changed again in the last part of the data taking period to incorporate ParticleFlow jets since they are used in analyses.

Thanks to the improvement of the electron identification at the trigger level, the electron trigger scheme was greatly simplified in the 8 TeV running period. A single isolated electron trigger *HLT_Ele27_WP80* is used throughout the whole dataset requiring an electron with p_T >25 GeV and $|\eta|$ <2.5.

4.1.2 Primary vertex selection

Once an event passes the trigger selection, additional criteria are enforced to ensure the presence of a good primary vertex. The reconstructed vertices in the event are sorted according to a decreasing sum of the p_T^2 of the tracks and only the vertex with

the highest sum is considered. This vertex needs to be flagged as being not-fake to ensure it is properly fitted. Furthermore, the vertex fit needs to have at least four degrees of freedom and a maximum distance to the interaction point in the $|z|$ and radial direction is set to 24 and 2cm respectively. If the leading vertex in the event does not satisfy these criteria, the event is dropped.

4.1.3 Lepton selection criteria

In the previous chapter, numerous criteria on muons and electrons were introduced to introduce well reconstructed leptons from for example mis-identified charged hadrons. Leptons that survive this cleaning are subject to additional kinematic cuts. Although the transverse momentum threshold on the lepton is preferred as low as possible, the thresholds applied in the various triggers provide a lower bound. Consequently, the lepton criteria are taken as loose as possible such that the working point sits on the plateau of the trigger efficiency turn-on curve.

The lepton selection criteria are outlined for the μ+jets and e+jets decay modes in Table 4.1. The selection on the signal lepton from the W boson decay is accompanied by a veto on the presence of a loosely identified muon or electron. If any lepton of the latter is present, the event is removed to avoid overlap between the μ+jets and e+jets samples as well as with the di-lepton $t\bar{t}$ sample.

	$\sqrt{s}=7$ TeV						$\sqrt{s}=8$ TeV					
	μ+jets			e+jets			μ+jets			e+jets		
	p_T	$\|\eta\|$	I_{rel}	p_T	$\|\eta\|$	I_{rel}	p_T	$\|\eta\|$	I_{rel}	p_T	$\|\eta\|$	I_{rel}
Signal μ	>26	<2.1	<0.125	-	-	-	>25	<2.1	<0.120	-	-	-
Signal e	-	-	-	>30	<2.5	<0.100	-	-	-	>32	<2.5	<0.100
Veto μ	>10	<2.5	<0.200	>10	<2.5	<0.200	>10	<2.5	<0.200	>10	<2.5	<0.200
Veto e	>15	<2.5	<0.200	>20	<2.5	<0.200	>20	<2.5	<0.150	>20	<2.5	<0.150

Table 4.1: Lepton kinematic selection criteria used at 7 and 8 TeV.

4.1.4 Jet selection criteria

A generic selection of jets is performed in both $e-$ and μ-channels, and both for the 7 and 8 TeV analysis. Jets that pass the identification discussed in the previous chapter are required to fulfil $p_T > 40$ GeV and be within $|\eta| < 2.5$. Subsequently, events are retained if they contain at least four of these jets.

The jet requirement is quite stringent and limits the $t\bar{t}$ signal selection efficiency as it cuts hard on the softer third and fourth jet p_T spectra, when ordered in decreasing p_T. For this reason, a dynamic cut set is usually adopted with gradually decreasing thresholds going from the first to the fourth jet. While this allows to improve on the selection efficiency, it impairs the method to estimate the b-tagging efficiency presented in Chapter 5. The latter relies on kinematic reweighting between different jet samples

created from the four leading jets and therefore to an increased sensitivity to the jet p_T threshold asymmetry. For this reason a symmetric cut set is used requiring at least four jets with a p_T > 40 GeV.

4.1.5 $\not\!\!E_T$ selection criteria

Among the different backgrounds that enter in the selected event sample, QCD multi-jet events are particularly tough to handle. Due to the large multi-jet cross section, extreme amounts of events should be simulated to have a sufficiently large sample useful for the analysis. This is beyond current technical reach so an extra criterion on the size of $\vec{\not\!\!E}_T$ (Table 4.2) is added to reduce this background.

	$\sqrt{s} = 7$ TeV		$\sqrt{s} = 8$ TeV	
	μ+jets	e+jets	μ+jets	e+jets
$\not\!\!E_T$ $>$	-	30 GeV	30 GeV	40 GeV

Table 4.2: $\not\!\!E_T$ criterion used at 7 and 8 TeV.

The $\not\!\!E_T$ distribution before any threshold was applied is shown in Figure 4.2 for events at 8 TeV after the μ+jets event selection and after the e+jets event selection. The low $\not\!\!E_T$ region is expected to be populated by QCD events since they do not contain any neutrinos. As a consequence, the $\not\!\!E_T$ cut is chosen such that it removes the bulk of the QCD events while retaining acceptable $t\bar{t}$ signal efficiency.

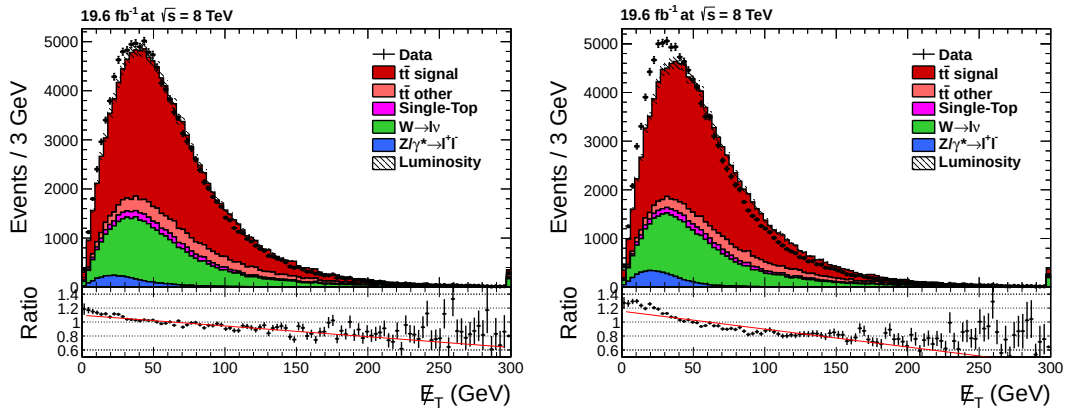


Figure 4.2: Distribution of the $\not\!\!E_T$ in the μ+jets channel (left) and in the electron channel (right) at $\sqrt{s} = 8$ TeV.

The same reasoning has been used at 7 TeV for which the $\not\!\!E_T$ distributions are shown in Figure 4.3. However, due to the lower QCD contamination compared to 8 TeV there was no clear indication of the need for a $\not\!\!E_T$ threshold in the muon channel. In the electron channel, a relaxed threshold is applied compared to the 8 TeV selection.

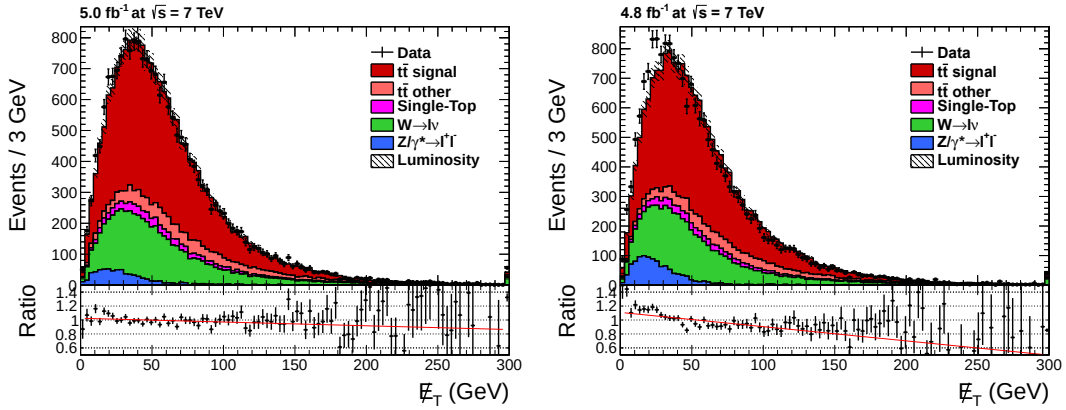


Figure 4.3: Distribution of the $\not\!\!E_T$ in the μ+jets channel (left) and in the electron channel (right) at $\sqrt{s} = 7$ TeV.

4.2 Event selection results

The event selection performance can be characterised both by the efficiency of selecting $t\bar{t}$ signal events but also the rejection of background events. In the $t\bar{t} \to l\nu_l b\bar{b}q\bar{q}$ oriented selection presented in the previous section, events are selected when they contain essentially one isolated lepton, at least four jets and can produce some moderate amount of $\not\!\!E_T$. Processes other than $t\bar{t}$ events can produce the same topology by for example the presence of additional fake jets, radiation or mis-identified leptons. These background processes hence contribute to the selected event sample as can be seen from the event selection yields for different processes in Tables 4.3 and 4.4 for 7 and 8 TeV respectively.

One of the main backgrounds to the $t\bar{t}$ signal process is the production of W and Z bosons[1] with additional jets from ISR/FSR. While the W boson can decay into a lepton and a neutrino, the Z boson decays into an opposite sign lepton pair. Combined with additional jet activity, both processes potentially look like a signal $t\bar{t}$ event in the experiment. The W boson background is the most prominent of the two due to its larger cross section and the additional loose lepton veto. The W/Z event samples are generated with different additional jet multiplicities and from Tables 4.3 and 4.4 it is clear that mainly the W/Z+4jets events play a role in the analysis.

Another background process arrises from the electroweak production of single top quarks. Two main production modes are taken into account[2], namely the t-channel production with exchange of a W boson and the associated production of a W boson and a top quark (tW-channel production). As can be seen from the selection yields, the tW-channel single-top production is the most dominant component of the single-top background.

Finally, another background component originates from the $t\bar{t}$ process itself. Di-lepton and fully hadronic $t\bar{t}$ decays can mimic the μ+jets or e+jets topology by ad-

[1]The simulated event sample containing $Z \to l^+l^-$ also incorporates virtual photons decaying into an opposite sign lepton pair.

[2]The s-channel production is omitted because of its small cross section.

ditional jet activity and the reconstruction of fake leptons. The selected events from these channels are labeled $t\bar{t}$ *other* throughout this thesis.

The event selection performance can also be checked by looking at the invariant mass distribution of the three-jet combination with the highest vectorial sum of their transverse momenta in the event, called the m3 mass. By construction, this mass definition strongly relates to the mass of the hadronically decaying top quark and hence it is expected that this distribution would be dominated by signal $t\bar{t}$ events. Figure 4.4 provides a data to simulation comparison for the m3 mass. A good agreement between data and simulation is observed and the breakdown of the simulation into signal and background processes indicates that the sample is $t\bar{t}$ enriched.

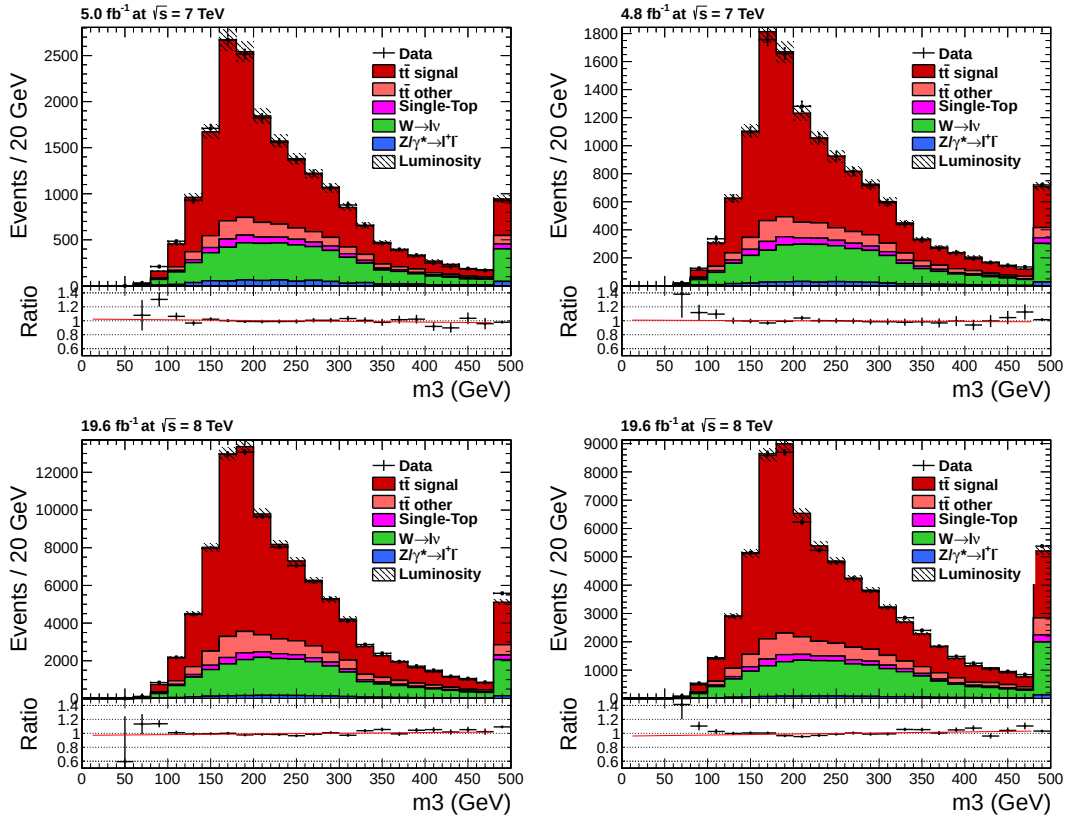


Figure 4.4: Invariant mass of the three-jet system with the highest vectorial sum of transverse momenta among the selected jets after the full event selection. The mass distribution is shown for μ+jets events (left) and e+jets events (right) at $\sqrt{s} = 7$ and 8 TeV.

4.3 Reweighting for soft p_T^{top} spectrum in simulation

It has been observed that the standard $t\bar{t}$ event sample generated with the MadGraph event generator exhibits a harder p_T-spectrum for the jets and the lepton than in data. Since the selection cuts are directly on the jet transverse momentum, the simulated MadGraph $t\bar{t}$ events have to be reweighted to cancel this discrepancy.

Table 4.3: An overview of the event selection yields for an integrated luminosity of $5.0\,\mathrm{fb}^{-1}$ at 7TeV.

	μ+jets		e+jets	
Process	Sel. eff. (%)	Yield	Sel. eff. (%)	Yield
$t\bar{t}$ (signal)	1.6	13982.6	1.1	9303.8
$t\bar{t}$ (other)	0.2	2140.7	0.2	1600.1
$SingleTop\ tW-Channel\ (t)$	0.7	393.6	0.5	271.7
$SingleTop\ tW-Channel\ (\bar{t})$	0.8	401.0	0.5	281.9
$SingleTop\ t-Channel\ (t)$	0.1	117.1	0.04	75.5
$SingleTop\ t-Channel\ (\bar{t})$	0.1	64.2	0.04	40.0
$W \to l\nu_l$ +2jets	0.0004	28.8	0.0003	19.7
$W \to l\nu_l$ +3jets	0.003	45.3	0.001	23.9
$W \to l\nu_l$ +4jets	0.7	6331.8	0.4	4204.6
$Z/\gamma* \to l^+l^-$ incl.	0.01	917.0	0.003	493.9

Table 4.4: An overview of the event selection yields for an integrated luminosity of $20.0\,\mathrm{fb}^{-1}$ at 8TeV.

	μ+jets		e+jets	
Process	Sel. eff. (%)	Yield	Sel. eff. (%)	Yield
$t\bar{t}$ (signal)	1.6	77547.2	1.1	52066.8
$t\bar{t}$ (other)	0.3	12864.4	0.2	8778.5
$SingleTop\ tW-Channel\ (t)$	0.7	1627.4	0.5	1164.2
$SingleTop\ tW-Channel\ (\bar{t})$	0.8	1682.5	0.5	1137.0
$SingleTop\ t-Channel\ (t)$	0.1	653.8	0.04	423.9
$SingleTop\ t-Channel\ (\bar{t})$	0.1	342.7	0.04	233.7
$W \to l\nu_l$ +1jet	5e-05	65.6	5e-05	61.0
$W \to l\nu_l$ +2jets	0.001	278.4	0.0004	164.6
$W \to l\nu_l$ +3jets	0.004	482.6	0.002	293.7
$W \to l\nu_l$ +4jets	0.6	32239.5	0.4	21113.5
$Z/\gamma* \to l^+l^-$ +1jet	0.0001	8.8	0.0001	7.6
$Z/\gamma* \to l^+l^-$ +2jets	0.001	36.5	0.0004	17.6
$Z/\gamma* \to l^+l^-$ +3jets	0.03	313.9	0.02	188.3
$Z/\gamma* \to l^+l^-$ +4jets	0.4	2320.8	0.2	1224.7

A reweighting function based on the top quark transverse momentum was calculated by comparing the p_T^{top} differential distribution between data and simulation [34, 35]. This analysis was performed both at 7 and 8 TeV resulting in a different reweighting function to be applied at 7 and 8 TeV.

$$w^{7TeV}(p_T^{top}) = exp(0.199 - 0.00166 \times p_T^{top})$$

$$w^{8TeV}(p_T^{top}) = exp(0.156 - 0.00137 \times p_T^{top})$$

The effect of the reweighting is shown in Figure 4.5 depicting the p_T spectra of the four leading jets before and after reweighting is applied. The shift in the p_T spectrum of the jets is more pronounced.

The reweighting does not only improve the agreement between data and simulation in the jet p_T spectra, it also affects the $t\bar{t}$ event selection efficiency. A relative decrease of the selection efficiency of 6% is observed after reweighting the events.

The reweighting reduces the shift between the p_T spectra in data and simulation but does not completely cancel it. The residual effect is believed to be covered by the systematics that are quoted on each physics result. Finally, the lepton transverse momentum is shown in Figure 4.6 where a good agreement is found between data and simulation after applying the weighting procedure.

4.4 Topology reconstruction

In the previous sections, kinematic event selection criteria have been introduced to select top quark pairs from the multitude of background and it has been shown that these cuts define a kinematic phase space where the $t\bar{t}$ process is dominant. For each selected event, one lepton and four selected jets are present among which the lepton can be trivially matched to the $t\bar{t} \to l\nu_l b\bar{b}q\bar{q}$ topology. Unfortunately, this is not so trivial for the jets as there is no a priori knowledge about the flavour of their initiating quark.

One of the key variables that will be used throughout the analyses in this thesis is the invariant mass of the lepton and the leptonic-side b jet system. In the context of the semi-leptonic $t\bar{t}$ decay, the leptonic side b jet is defined as the jet that is initiated by the b quark that originates from the top quark decay where the W boson decays into a charged lepton and its associated neutrino. As a consequence, it is crucial to be able to identify this jet among the selected jets in the event.

In Section 4.4.1, a jet sorting technique will be introduced to allow mapping the selected jets in the event to the $t\bar{t}$ topology. Thereafter, the performance of this technique will be discussed.

4.4.1 Associating jets to partons with a χ^2-based jet matching

The event selection requires at least four jets with a transverse momentum above 40 GeV. To initiate the jet combination algorithm, an assumption is made that the four selected jets with the highest transverse momentum correspond to the quarks from the $t\bar{t}$ decay. Without the use of b-tagging, this leaves 24 possible combinations to combine

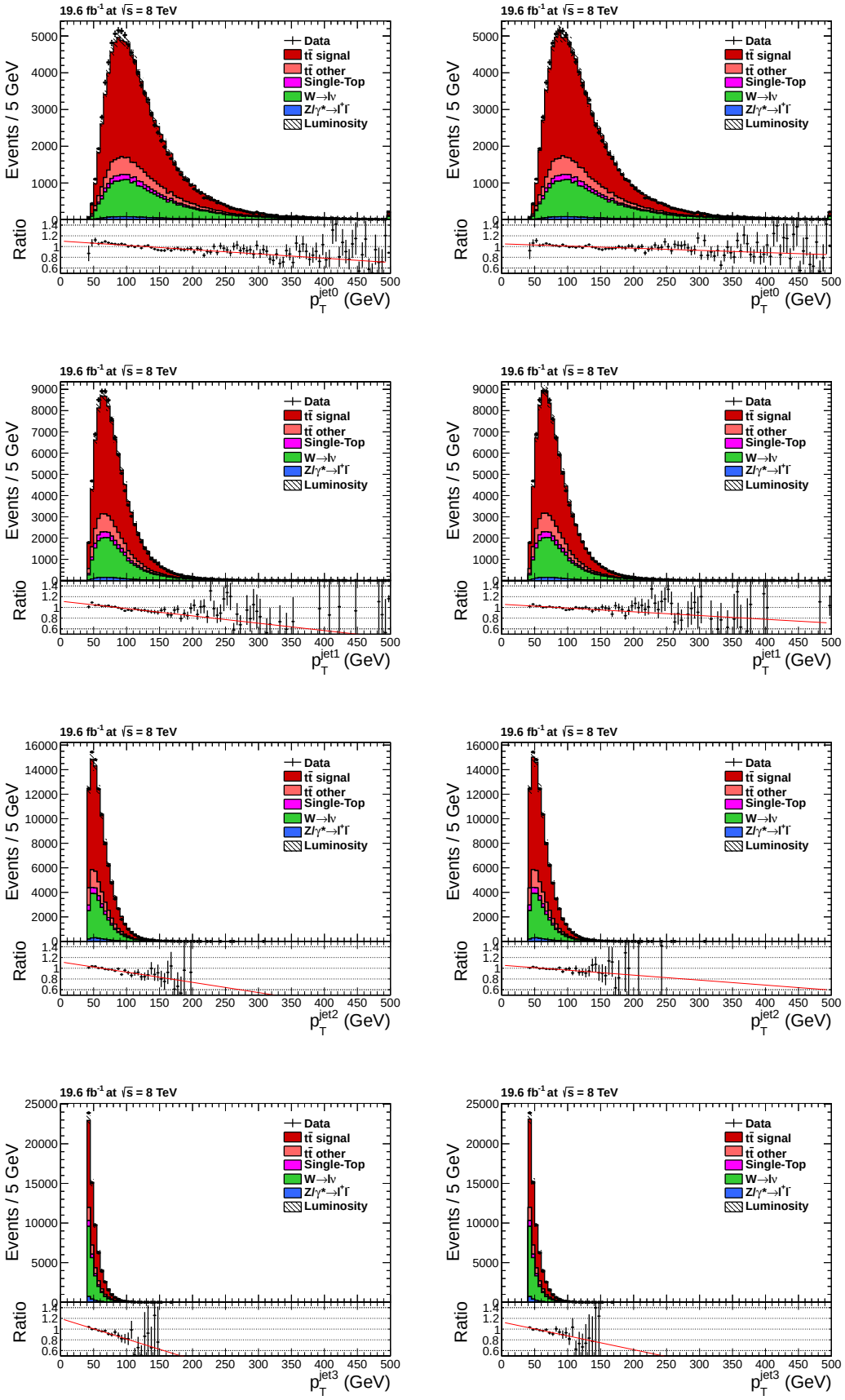


Figure 4.5: Distribution of the transverse momentum of four selected leading jets (jet 0→3). The left distribution compares data to unweighted simulation while the p_T^{top} reweighting was applied to create the right plots.

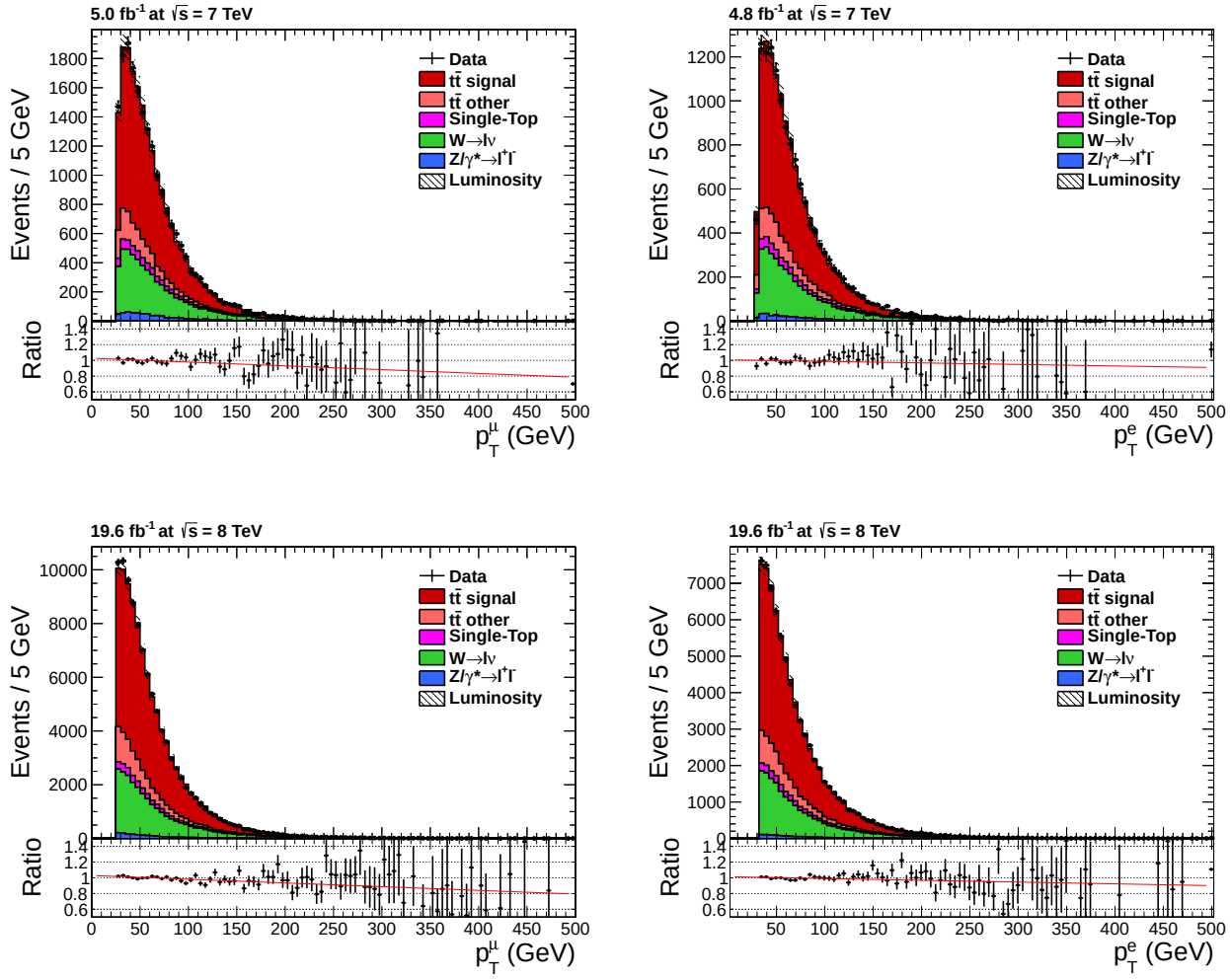


Figure 4.6: Distribution of the muon p_T (left) and the electron p_T right in μ+jets and e+jets events respectively. The p_T^{top} reweighting was applied and a good agreement between data and simulation is obtained.

the four leading jets to reconstruct the $t\bar{t}$ topology. On the other hand, the light quark jets from the W boson decay can be interchanged among each other since this does not prevent the top quark mass reconstruction. This reduces the number of combinations to 12 instead of 24.

To perform the matching between the selected jets and the $t\bar{t}$ final-state quarks, a chi-square function is defined based on the W boson and top quark masses reconstructed with different jet combinations:

$$\chi^2 = \left(\frac{m_{bqq} - m_t}{\sigma_t}\right)^2 + \left(\frac{m_{qq} - m_W}{\sigma_W}\right)^2. \tag{4.1}$$

The top quark and W boson masses (m_t, m_W) and resolutions $(\sigma(m_t), \sigma(m_W))$ are determined from simulated $t\bar{t}$ events where the four leading jets have been matched to

the generated quarks. The jet-parton matching is performed by requiring the generated parton to be within $\Delta R < 0.5$ from the jet axis. Using the matched jet-parton combinations, the distributions of the reconstructed W boson mass and top quark mass are obtained. These distributions are finally fitted with a Gaussian function to obtain the expected mass values and variances as shown in Figure 4.7.

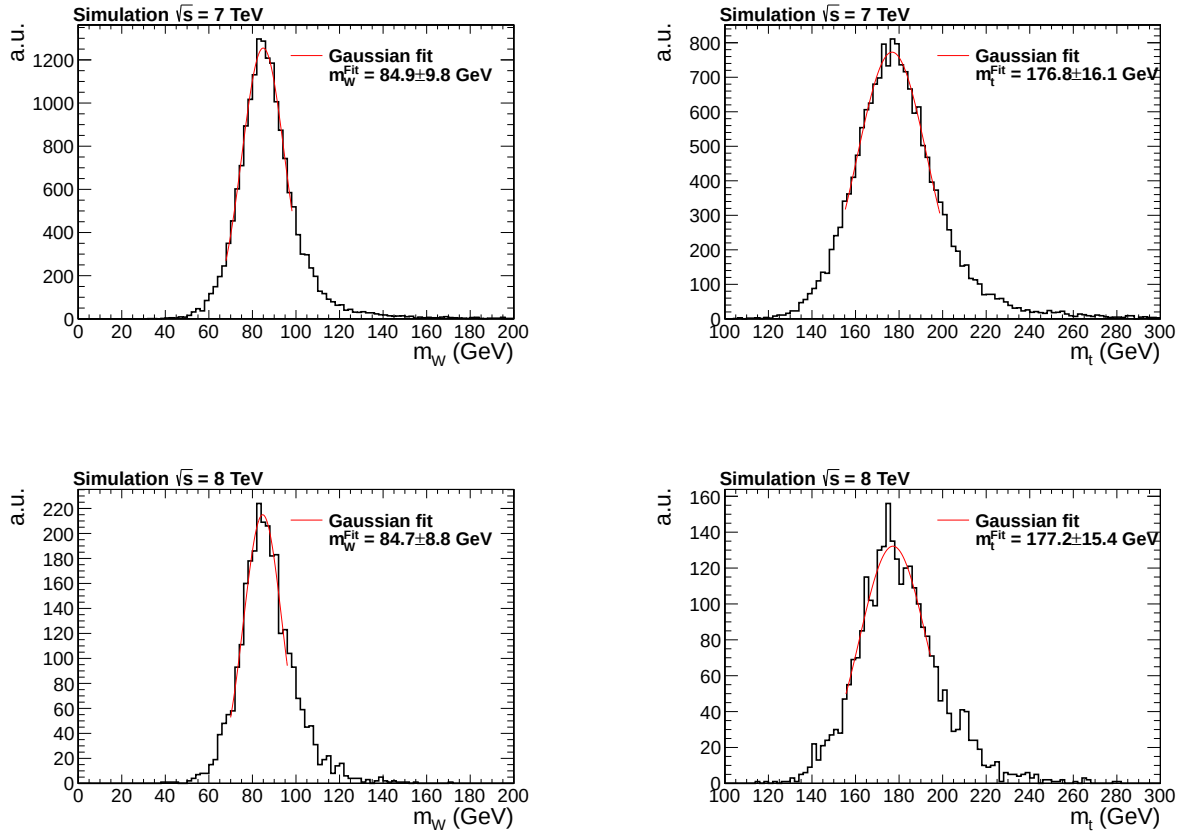


Figure 4.7: Distribution of the W boson and top quark masses reconstructed from simulated $t\bar{t}$ signal events where the four leading jets match the generated final-state partons. The distribution is fit with a Gaussian function to obtain the expected mass values and variances to use in the jet combination algorithm.

The χ^2 value for each of the 12 combinations is calculated, and the combination with the smallest χ^2 is selected to represent the event topology. The distribution of the smallest χ^2 values is shown in Figure 4.8 and compared with data. The good agreement between data and simulation indicates that the jet combination technique works in the same way on data as it does on simulated events.

4.4.2 Jet matching performance

The efficiency of the χ^2 jet combination mechanism can be estimated by the fraction of selected events in a simulated sample of true $t\bar{t} \to \mu\nu_\mu b\bar{b}q\bar{q}$ events where the jet

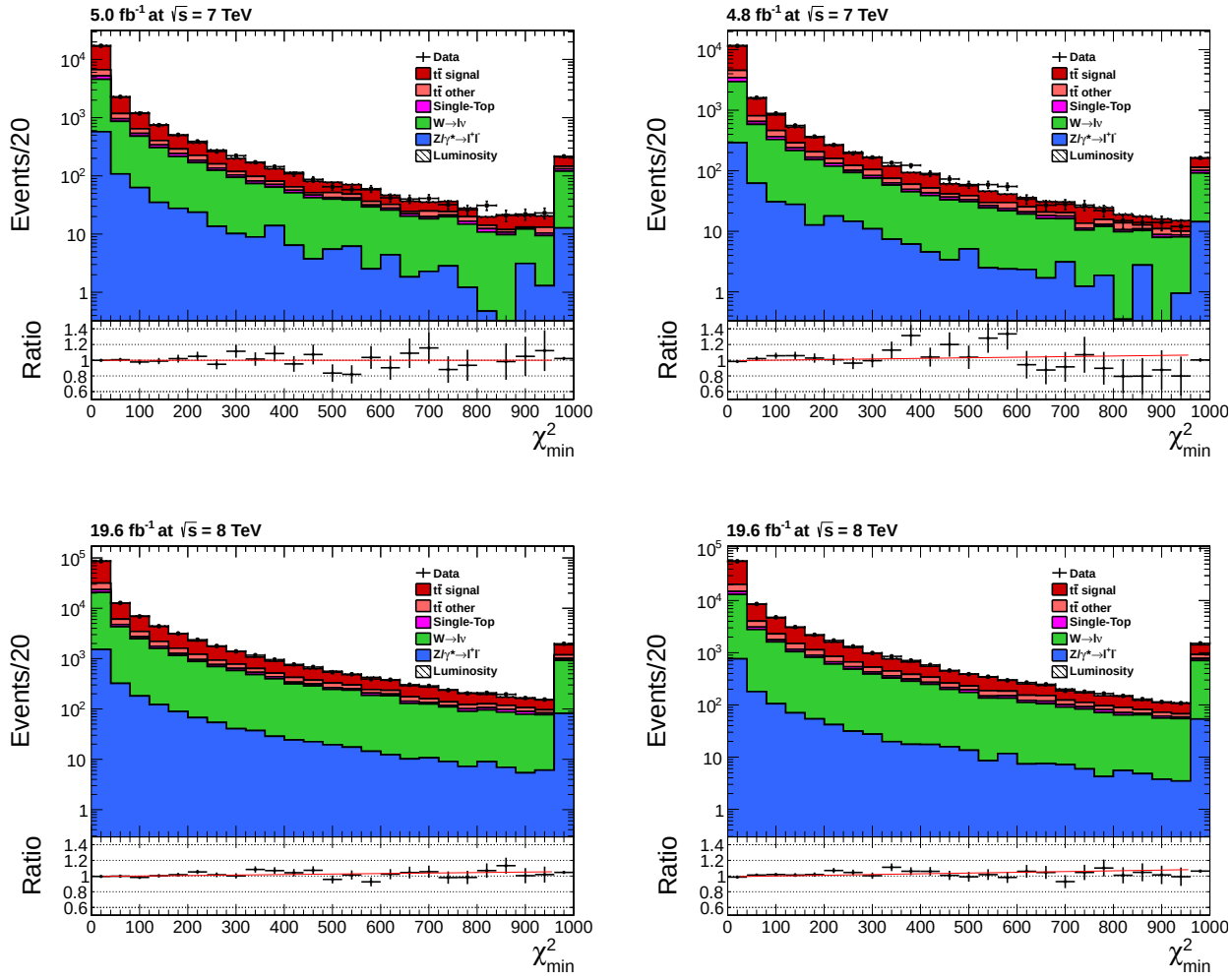


Figure 4.8: Distribution of the smallest χ^2 value among all different jet combinations in the selected events. The distribution for data and simulation are provided for μ+jets events (left) and for e+jets events (right).

combination chosen by the χ^2 method is exactly the same as the combination obtained with the generator based jet-parton matching. The efficiency is found to be 11%.

The low jet combination efficiency might raise questions on the usefulness of this technique. However, the jet-parton matching efficiency was not taken into account. In the previous chapter, the effect of ISR/FSR and Underlying Event were introduced. All these effects can create additional (hard) jets in the event that can end up among the leading jets. If this happens, a good jet combination cannot be found among the four leading jets alone. It turns out that this happens in 80% of the $t\bar{t}$ events resulting in a maximal jet-parton matching efficiency of only 20%. In the events where a good combination exists, the χ^2 jet combination technique reaches an efficiency of 56%.

So far, the χ^2 jet combination efficiency has been expressed in terms of fully reconstructed events where this technique yields the same combination as the true jet-parton matching. Nevertheless, this may be to harsh in the framework of this thesis as the main focus is on the leptonic side b jet. As a consequence, the jet-combination efficiency can also be defined as the probability for the leptonic side b jet candidate to be an actual b jet. This probability is found to be 46.3% in simulated semi-μ $t\bar{t}$ decays after the event selection.

Chapter 5

Measurement of the b tagging efficiency

The identification of b jets [2], also called b-tagging, is an important aspect of many analyses in the CMS Collaboration. Since a wide variety of physics processes contain one or more b jets in the final state, b-tagging is widely used to refine the signal over background ratio while keeping a good signal efficiency. Accordingly, studying the performance of the various b jet identification algorithms is a key element of the physics program.

The performance of b-tagging algorithms, namely the efficiency and light jet mistag rate, is usually measured using multi-jet events [2, 130]. Conversely, top quark pair events can be used to measure these quantities since a top quark decays about 100% of the time into a W-boson and a b quark. The large amount of top quark pair events delivered by the LHC in proton-proton collisions at 8 TeV provides a unique sample of b quarks to perform this measurement [1, 2].

In the following, a measurement of the inclusive b-tagging efficiency is proposed using semi-leptonic decaying top quark pairs, namely $t\bar{t} \to bWbW \to bqqb\mu\nu_\mu$, accounting for 15% of the total branching ratio. A signal region rich in jets originating from the hadronization of a b quark, further denoted as b jets, is constructed where the non-b jet contamination has to be subtracted. This contamination is estimated from a control region using a fully data-driven procedure.

In Chapter 4, the procedure to extract semi-muonic top quark pair decays from the multitude of background processes has been described. Using the four highest p_T jets in the selected events, the $t\bar{t}$ topology can be reconstructed by the χ^2 jet sorting technique.

The jet matched to the hypothetical b quark from the leptonic decaying top quark, further denoted as the leptonic b jet candidate, can be used to define a b jet sample, also denoted as the signal sample. This sample can be further divided into two subsamples, one enriched in jets truly coming from b quarks and the other depleted. In the b-enriched sample, the b-tag discriminator distribution for b jets can be reconstructed after removing the non b jet contamination.

The method developed in this section will be illustrated using the Track Counting High Efficiency b-tagging algorithm in the μ+jets channel at a centre of mass energy of 8 TeV. Results for the electron channel as well as all other available algorithms are

provided in the last section. There also the results obtained at a centre of mass energy of 7 TeV are provided.

5.1 Constructing the b jet candidate sample

A b jet candidate sample can be constructed by matching the selected jets with the $t\bar{t}$ event topology. For every event that passes all the event selection criteria, a χ^2 value is assigned to each possible combination of the four leading jets to the four leading order partons in the final state. To obtain these χ^2 values, the mass of the two- and three-jet systems are compared to the W mass and top-quark mass obtained from simulation. The combination with the lowest χ^2 value is taken as the correct jet combination since it provides the best match to the expected event kinematics. Additionally, the minimum χ^2 value in the event needs to be lower than 90 (160) in μ+jets (e+jets) events to reduce the amount of background and wrong jet combinations from $t\bar{t}$ signal events. The selected jet according to this procedure has an enhanced probability to originate indeed from a true b quark. This jet is considered to be the b jet candidate.

The b jet candidate sample constructed using the leptonic b jet candidates contains a non b jet contamination. The average b-purity of this jet sample is 31%. For this reason this sample is further subdivided in two subsamples where one is rich in b jets, called the b-enriched sample, and the other is depleted in b jets, called the b-depleted sample. The b-depleted sample will serve its purpose as a handle to remove the non b jet contamination from the signal sample.

To be able to make the subdivision, an observable discriminating between b jets and other jet flavours has to be introduced. One such observable is the invariant mass of the leptonic b candidate and the lepton, or jet-lepton mass, from now on denoted as M_{lj}.

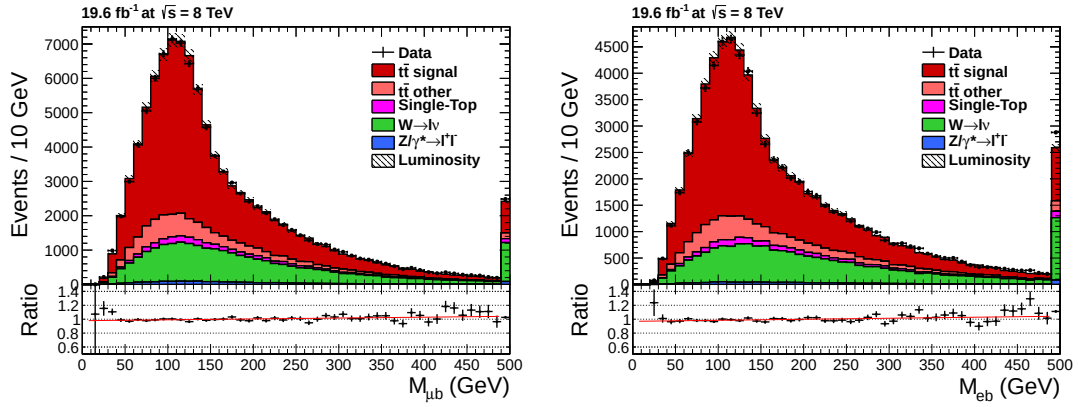


Figure 5.1: Distribution of the jet-lepton mass for signal and all relevant background processes in the μ+jets channel (left) and in the e+jets channel (right). The simulated distribution is compared to data. The distributions of all simulated processes are normalized to the integerated luminosity of the data.

A data to simulation comparison for the M_{lj} variable is shown in Figure 5.1 for

an integrated luminosity of around $20 fb^{-1}$. The distribution shows that this variable is overall well described by simulation. In contrast there is a small discrepancy for M_{lj} values below 50 GeV corresponding to the abundant multi-jet background where not enough simulation is at hand due to its large cross section. This background does not contain a large fraction of b jets and as such will not degrade the result. The main contribution in this plot comes from the semi-muonic $t\bar{t}$ events (signal). The jet-lepton mass distribution is much less peaked for background processes like W$\rightarrow l\nu$ and $Z/\gamma* \rightarrow l^+l^-$ in comparison to the $t\bar{t}$ signal. This difference originates from the absence of top quark decays in those processes. Processes like $t\bar{t}$ decays, other than semi-muonic (signal) decays, and SingleTop also give an important contribution because of the presence of b quarks from top quark decay. These processes hence also contribute b jets to the b jet candidate sample.

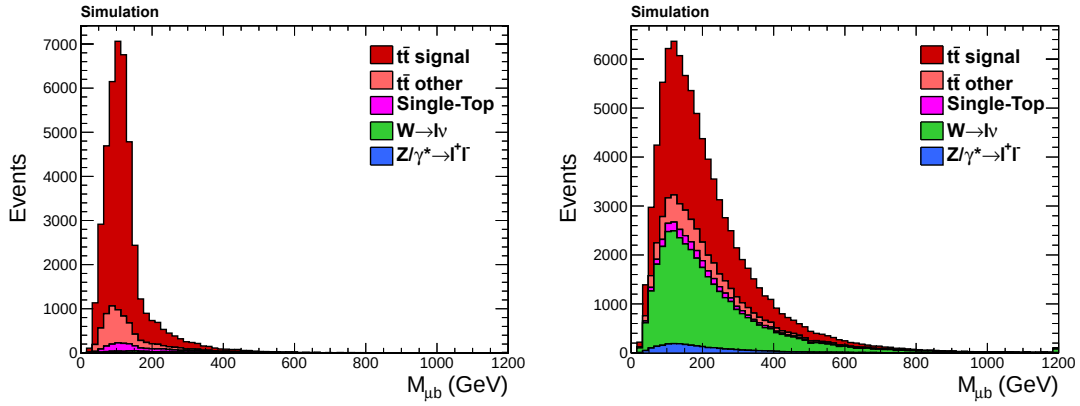


Figure 5.2: The jet-lepton mass distribution for b quarks only (left) and non-b quarks only (right)

The b jet flavor discriminating power of the M_{lj} variable is immediately visible in Figure 5.2. It displays the difference in the distribution of M_{lj} between b jets and non b jets exhibiting a much more peaked distribution for the former and a broader distribution for the latter. The b jet flavor discriminating power of the M_{lj} observable is also demonstrated in Figure 5.4 showing the b jet purity in each bin of M_{lj}. The b jet purity, P_i, is defined in equation 5.1.

$$P_i = \frac{M^i_{lj}(b)}{M^i_{lj}(b) + M^i_{lj}(non-b)} \tag{5.1}$$

In this equation $M^i_{lj}(b)$ refers to the number of b jets in the i'th bin of the jet-lepton mass distribution while $M^i_{lj}(non-b)$ refers to the number of non b jets in that bin.

In Figure 5.4, a clear drop in the b jet purity around 160 GeV is observed. This allows for the b jet candidate sample to be subdivided in a b-enriched and b-depleted subsample.

Before the jet-lepton mass can be used to define the b-enriched and b-depleted samples, it has to be checked that the shapes of the b-tagging discriminator distributions for b jets as well as non b jets are invariant with respect to the jet-lepton

mass variable. If not, the b-tagging discriminator shapes will be different in the b-enriched and b-depleted regions possibly distorting the shape of the reconstructed b jet b-tagging distribution in the b-enriched sample. This would ultimately cause a bias on the measured b-tagging efficiency.

However, a small correlation between the b-tagging discriminator and the jet-lepton mass has been observed as shown in Figure 5.3 for b jets and non-b jets seperately. For b jets, a small correlation of 0.3% is observed while it grows to 9% for light jets. This correlation for light jets, jets that originate from wrong jet combinations, can be understood by the W boson and top quark mass constraints applied to select the leptonic b jet candidate.

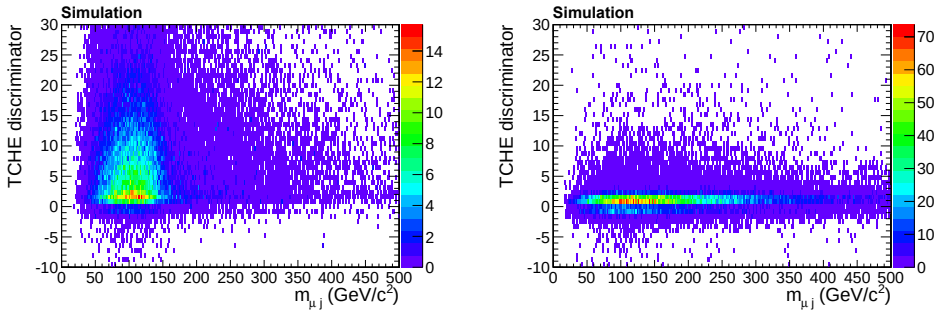


Figure 5.3: The TCHE discriminator value as a function of the jet-lepton mass for b jets only (left) and non b jets only (right). A very small correlation of 0.3% for b jets and 9.0% for non b jets is observed.

In the 8 TeV analysis, the b enriched region is defined as 70 $GeV < M_{lj} < 160$ GeV for both decay channels and the b depleted region as 160 $GeV < M_{lj} < 300$ GeV in the muon channel while it is extended by 30 GeV in the electron channel. The lower limit of the b depleted and the upper limit of the b enriched sample are chosen around the cut-off value while the upper limit of the b depleted region and the lower limit of the b enriched region are taken away from the tails of the distribution. The choice for these particular definitions mainly depends on the stability of the method which will be the topic of Section 5.4.

The b-enriched and b-depleted samples constructed in this way have a b jet purity of 50.0% and 15.8% respectively.

5.2 Reconstructing the b-tag discriminator distribution

In the following, a method will be developed to measure the b-tagging efficiency within the b-enriched subsample by reconstructing the b-tag discriminator distribution for true b jets, denoted as $\hat{\Delta}_b^{enr}$. Once this distribution is reconstructed, the b tagging efficiency, $\hat{\epsilon}_b^{enr}$, can be calculated for any given cut value on the discriminator, Δ_b, also

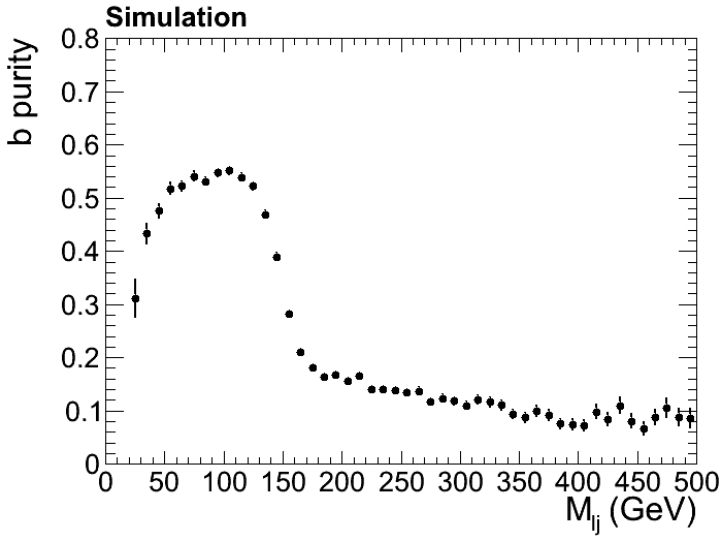


Figure 5.4: The b jet purity as a function of the jet-lepton mass in the b candidate sample using all signal and background simulation samples.

called Working Points (WP). For each b-tagging algorithm three working points are provided called the loose, medium and tight working points. These working points correspond to a discriminator cut value that provides a 10%, 1% and 0.1% light-jet mis tagging efficiency respectively. Although these working points are measured explicitly, the method proposed here measures the efficiency for any given working point.

It has to be checked that the b-tagging efficiency obtained from the b-enriched sample describes the b-tagging efficiency in the full b jet candidate sample. A priori, no correlation has been found between the b-tagging discriminator distribution and the jet-lepton mass for b-jets. Nevertheless, as the construction of the b-enriched and depleted samples depends on the jet-lepton mass it also depends on the p_T of the jets. The b-tagging algorithms do exhibit some p_T dependence as well since the track momentum resolution degrades for high p_T tracks when the jet is more collimated resulting in closely spaced tracks. Moreover, for low p_T jets, the B-hadron will be less boosted resulting in a smaller decay length.

Figure 5.5 shows, that the true b-tagging efficiency[1] extracted from the b-enriched subsample sample agrees with the efficiency obtained from the full b jet candidate sample. Hence we can extract the b-tagging efficiency from the b-enriched jet sample which can in turn be used for genuine b quarks in $\mathrm{t\bar{t}}$ events.

To obtain the $\hat{\Delta}_b^{enr}$ distribution, the non b jet contamination in the b enriched sample has to be removed. To do this, both the shape and scale of the b-tagging discriminator distribution for non b jets has to be known. The shape of the non b jet b-tagging discriminator distribution can be obtained from the b depleted sample, Δ_b^{depl}. Because the non b jet purity is not equal among the b-depleted and the b-enriched subsample (Table 5.1), the distribution from the b-depleted sample has to be

[1]This is the b-tagging efficiency extracted from jets that match a generated b quark in $\Delta R < 0.3$.

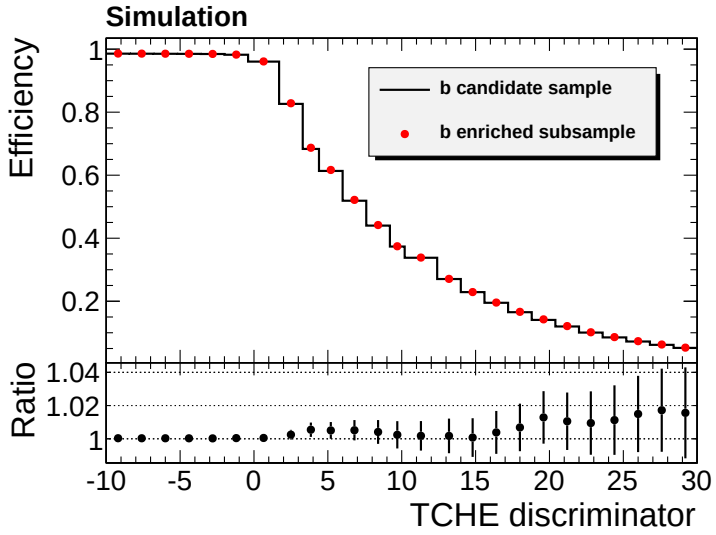


Figure 5.5: Closure test comparing the b tagging efficiency in the whole b jet candidate sample (black line) to the b tagging efficiency in the b enriched subsample (red markers). This figure shows a very good agreement between the two samples.

corrected for the non b jet scale difference between the two samples. Finally, the b jet b-tagging discriminator distribution, $\hat{\Delta}_b^{enr}$, can be reconstructed using eq. (5.3). The factor F in eq. (5.3) represents the ratio between the number of non b jets in the b enriched and b depleted subsamples which will correct the scale of the non b jet b-tagging discriminator distribution in the b-depleted sample to the one in the b-enriched sample.

$$F = \frac{N_{non-b}^{enr}}{N_{non-b}^{depl}} \tag{5.2}$$

$$\hat{\Delta}_b^{enr} = \Delta_b^{enr} - F \times \Delta_b^{depl} \tag{5.3}$$

Table 5.1: b and non-b purity in the b enriched and depleted subsamples

	b enriched subsample	b depleted subsample
b jet purity	50.0 %	15.8 %
non-b jet purity	50.0 %	84.2 %

The scale factor F expected from simulation using the generator information, denoted as F^{exp}, is equal to 1.169±0.008 where statistical uncertainty is calculated for a simulated sample corresponding to an integrated luminosity of 30 fb^{-1}.

The b jets b-tagging discriminator distribution reconstructed using eq. (5.3) and the scale factor F^{exp} is shown in Figure 5.6. The measured distribution is compared to

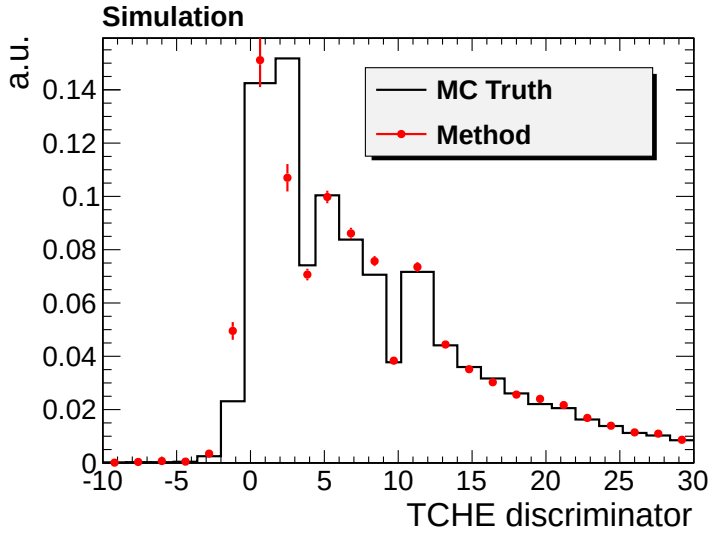


Figure 5.6: The Track Counting High Efficiency b tag discriminator distribution for true b jets in the b jet candidate sample (black line) compared to the measured distribution (red markers). The statistical uncertainty is calculated for a simulated sample corresponding to an integrated luminosity of 30 fb^{-1}.

the distribution for true b jets in the b jet candidate sample. Using the reconstructed b-tagging discriminator distribution, the efficiency for different working points of the b tagging algorithm can be measured. This efficiency as a function of the discriminator cut is shown in Figure 5.7 along with the relative bias, b_{ϵ_b}, between the measured efficiency and the true efficiency. The relative bias on simulation is defined in eq. (5.4).

$$b_{\epsilon_b} = \frac{\hat{\epsilon}_b - \epsilon_b^{true}}{\epsilon_b^{true}} \tag{5.4}$$

The relative bias as a function of the b-tagging discriminator in Figure 5.7 shows a significant increase in the discriminator interval [0,3]. The same effect is already visible from the b-tagging efficiency plot in the same figure. This effect originates from the small correlation that was observed in Section 5.1 between the b-tagging discriminator for light jets and the jet-lepton mass.

To remove this correlation, each jet is weighed according to its p_T. The p_T distribution in both the b-enriched and b-depleted sample is shown in Figure 5.8. A data-driven reweighing function is constructed using the bin-by-bin ratio between the p_T spectra in the b-depleted and b-enriched region. This ratio, shown in Figure 5.9, is fitted with an exponential function. The fitted function only depends on the transverse momenta of the jets and hence it is not bound to any particular b tagging algorithm. The weights obtained by this procedure are then used to reweigh the jets in the b-depleted sample to mimic the kinematics from the b-enriched region and to remove the observed correlation. The p_T distribution in the b-depleted sample after reweighing agrees much

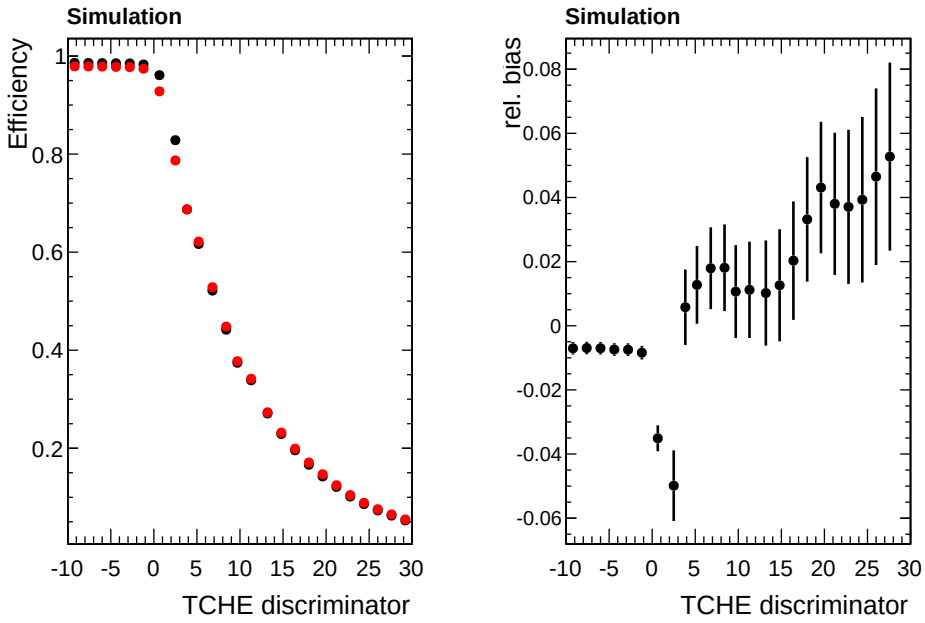


Figure 5.7: The Track Counting High Efficiency b tag efficiency distribution compared between true b jets in the b jet candidate sample (black line) and the measured efficiency (left) and the relative bias between the two efficiencies (right). The statistical uncertainty is calculated for a simulated sample corresponding to an integrated luminosity of 30 fb^{-1}.

better to the distribution in the b-enriched sample as shown in Figure 5.8.

The master formula to reconstruct the b-tagging discriminator distribution, eq. (5.3), can now be transformed to eq. (5.5) taking into account the kinematics reweighing.

$$\hat{\Delta}_b^{enr,rew} = \Delta_b^{enr} - F \times \Delta_b^{depl,rew} \tag{5.5}$$

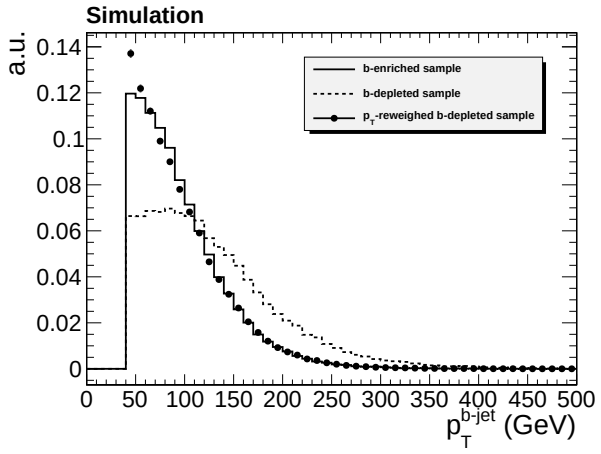


Figure 5.8: p_T distribution for the b jet candidates in the b-enriched and b-depleted subsamples. The p_T distribution in the b-depleted sample is shown before and after applying the reweighing.

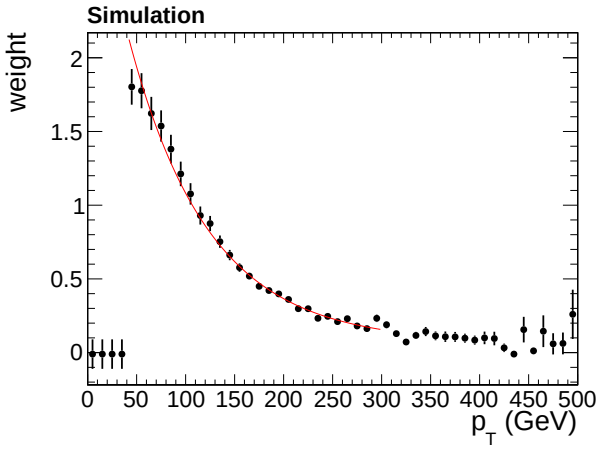


Figure 5.9: Reweighing function used for reweighing the jets according to their p_T. Since the reweighing is based on the jet p_T the same function can be used for the measurement on any b tagging algorithm.

The measured b-tagging efficiency along with the relative bias on the results are shown in Figure 5.10 using the kinematics reweighing between the b-enriched and b-depleted samples. The reweighing of the jets in the b-depleted sample solves the correlation and removes the dip in the range of [0,3] for the Track Counting High Efficiency tagger. The b tagging efficiency results for the loose, medium and tight working points for the Track Counting High Efficiency algorithm are provided in Table 5.2.

Table 5.2: Results for the Track Counting High Efficiency tagger Loose, Medium and Tight working points for an integrated luminosity of $30.0 fb^{-1}$ simulated events. The efficiency provided by the method is compared to the true b tagging efficiency.

Working Point	$\epsilon_b^{true}(\%)$	$\hat{\epsilon}_b(\%)$
Loose	82.6 ± 0.2	84.0 ± 0.9
Medium	68.3 ± 0.2	70.0 ± 0.8
Tight	33.8 ± 0.2	34.8 ± 0.5

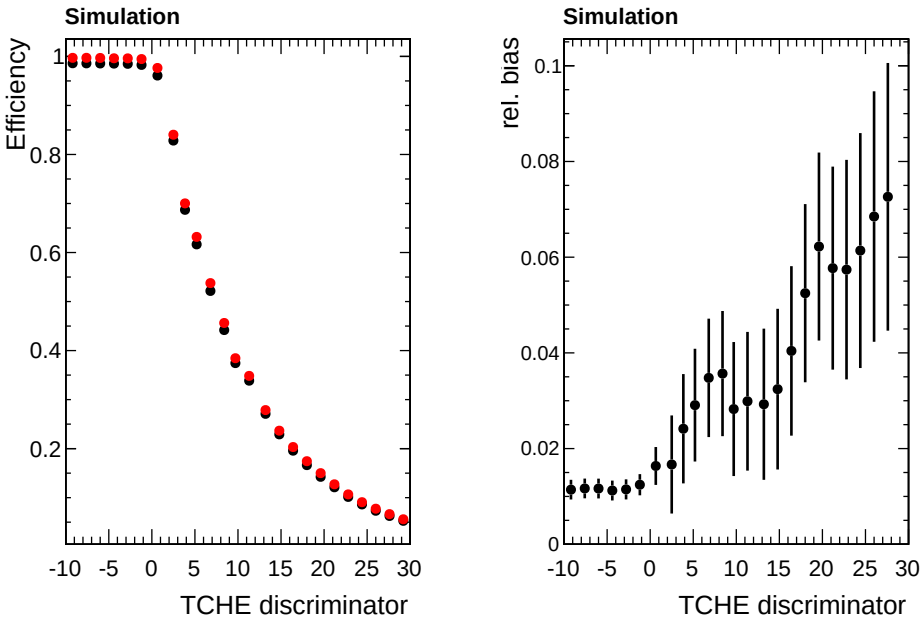


Figure 5.10: The Track Counting High Efficiency b tag efficiency distribution compared between true b jets in the b candidate sample (black markers) and the measured efficiency (red markers) after reweighing the jets for their transverse momentum (left) and the relative bias between the two efficiencies (right).

5.3 Data-driven estimation of the scale factor F

The method to measure the b-tagging efficiency developed in the previous section is not completely data-driven as it takes the scale factor F from simulation. Here, a method will be discussed to obtain this scale factor from data and as such remove the last simulation dependency. The method to measure the b-tagging efficiency consists of creating a b candidate sample and subdividing it in a b enriched and a b depleted region according to the jet-lepton mass. Subsequently, the b-tagging discriminator distribution in the b-depleted subsample is subtracted from the b-enriched subsample to remove the non b jet contamination in the b-enriched subsample. In this subtraction, the b depleted distribution is scaled by a factor F, to account for the difference in non b jet scale between the two subsamples. This scale factor is determined using simulated events which makes it possibly susceptible to differences between data and simulation, including reconstruction effects and signal modelling. Most systematic effects act on the shape of the jet-lepton mass distribution and hence will affect the scale factor F. Employing a data-driven measurement will reduce the sensitivity to the signal modelling.

In Section 5.3.1 it will be shown how a control sample consisting mostly of non b jets is constructed to measure the scale factor F. However, a kinematic mismatch was observed between the b-enriched and b-depleted subsamples of the b candidate sample. Hence, this kinematic mismatch will also appear between the b jet candidate sample (i.e. the signal sample) and the non-b jet candidate (i.e. the control sample). Since the scale factor strongly depends on the jet-lepton mass shape, the kinematic mismatch between signal sample and control sample has to be cancelled. Section 5.3.2 will elaborate on how to reweigh for this kinematic mismatch.

5.3.1 Constructing a non b jet control sample

To measure the scale factor F in a data-driven way, a control sample dominated by non b jets is constructed. In this control sample two subsamples can be created using the same M_{lj} boundaries used for the b-enriched and b-depleted subsamples in the signal sample. Since the control sample is pure in non b jets, it provides a data-driven handle on the scale factor. The natural candidates to populate this control sample are the two light jet originating from the W-boson in the top quark decay. These two light jet candidates can be obtained from the χ^2 jet combination previously discussed when constructing the b jet candidate sample.

Using simulation, the control sample is found to be 90.8% pure in non b jets after an anti-b tag on the two light candidate jets. For this the Track Counting High Efficiency b-tagging algorithm is used at a medium working point. The jet-lepton mass in the control sample can be defined as the invariant mass of the selected lepton with either one of the two light jet candidates. In the jet-lepton mass region $70\ GeV < M_{lj} < 160\ GeV$ a non b jet purity of 89.4% is obtained while the sample is and 93.2% pure in the region $160\ GeV < M_{lj} < 300\ GeV$. A good agreement between data and simulation for the control sample jet-lepton mass distribution is observed and shown in Figure 5.11 . The same purity numbers have also been observed in the e+jets channel.

The use of anti-b-tagging to construct the control sample might seem counter-

intuitive but since the threshold is reasonably loose and the b quark component in the control sample very small (<10%), the anti-b-tagging criterion does not affect the jet-lepton mass shape. If the shape remains invariant, no effect on the scale factor F is expected.

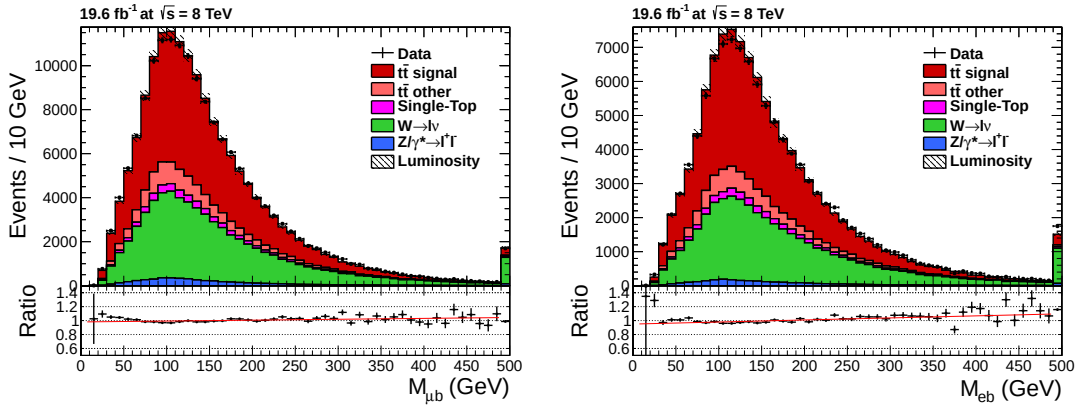


Figure 5.11: Distribution of the jet-lepton mass in the control sample for all relevant processes in the μ+jets channel (left) and in the e+jets channel (right). The simulated distribution is compared to data. The distributions of all simulated processes are normalized to the integerated luminosity of the data.

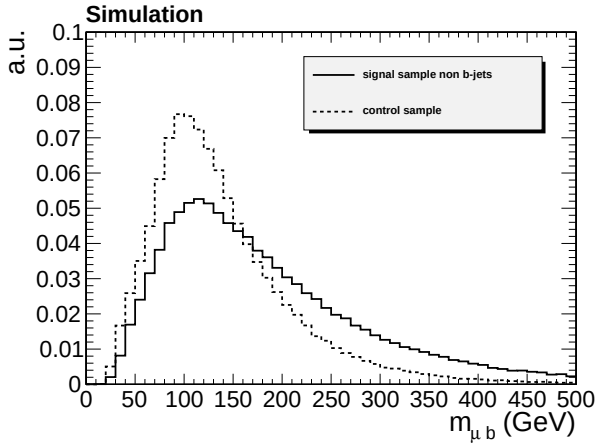


Figure 5.12: Comparison of the jet-lepton mass shape between the signal sample non b jets and the jets in the control sample. The shape comparison clearly shows the expected kinematics mismatch between the signal sample non b jets and control sample jets.

5.3.2 Reweighing the control sample kinematics

A very important prerequisite to measure the scale factor F from the control sample is that the shape of the jet-lepton mass distribution in this sample equals the shape

for non b jets in the signal sample. Since the scale factor F is defined as the ratio of the number of non b jets in the b-enriched subsample to the b-depleted subsample, a different jet-lepton mass shape would translate to a bias on the estimator of F.

A kinematic discrepancy has been observed between the signal and control samples as can be seen from Figure 5.12. Based on simulation a comparison is shown between the jet-lepton mass distribution obtained from the control sample and the same observable using only the non b jet fraction of the signal sample. A significant difference between both distributions is observed. This shape difference can be traced back to the kinematic properties of the jets. Figure 5.13 shows the p_T and η distributions for the jets in the control sample compared to the non b jets in the signal sample. Both distributions show a clear kinematic mismatch between the two samples.

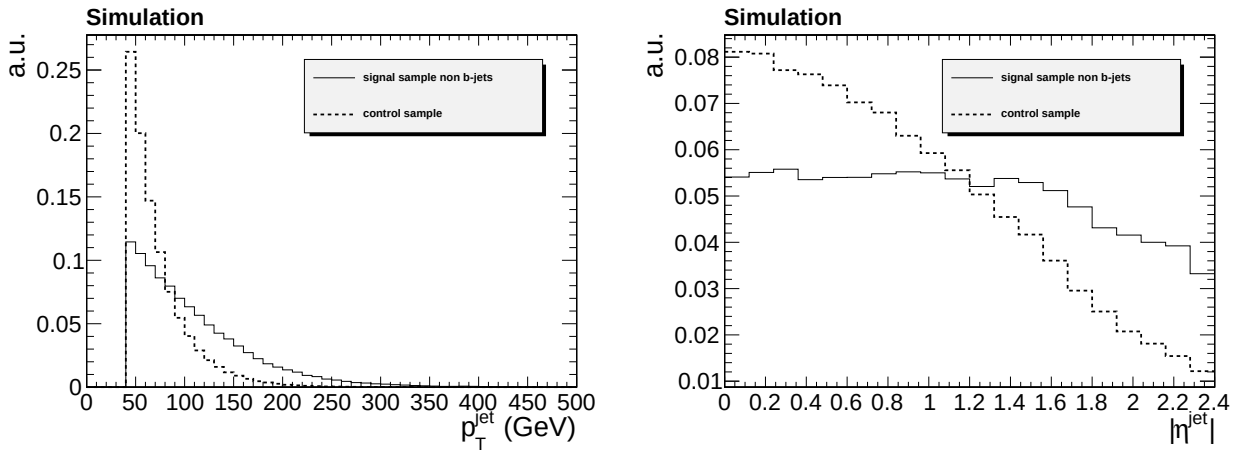


Figure 5.13: Comparison of the jet transverse momentum (p_T) and pseudo-rapidity (η) distributions between the signal sample non b jets and the jets in the control sample. The shape comparison clearly shows the expected kinematics mismatch between the signal sample non b jets and control sample jets.

The difference in jet kinematics observed in Figures 5.12 and 5.13 can be corrected by applying a data-driven reweighing technique. A two-dimensional (p_T,η) distribution is constructed from all jets in both the signal sample and the control sample. The division of these two histograms provides a weight to be applied in each (p_T,η) bin of the control sample. This weight distribution is shown in Figure 5.14.

After applying the (p_T,η) weights to the jets in the control sample, their kinematics match the non b jet fraction of the signal sample allowing to measure the scale factor F from the control sample. Figure 5.15 shows the p_T and η distributions for the reweighted control sample jets compared to the non b jets of the signal sample. A much better agreement is observed which also translates in a good agreement between the jet-lepton mass distributions as shown in Figure 5.16. On a simulated sample of 30 fb^{-1}, the scale factor F derived from the reweighted control sample, $\hat{F}^{CS}$, equals 1.149 $\pm$ 0.006 with a relative difference to F^{exp} of (1.7 ± 0.8) %.

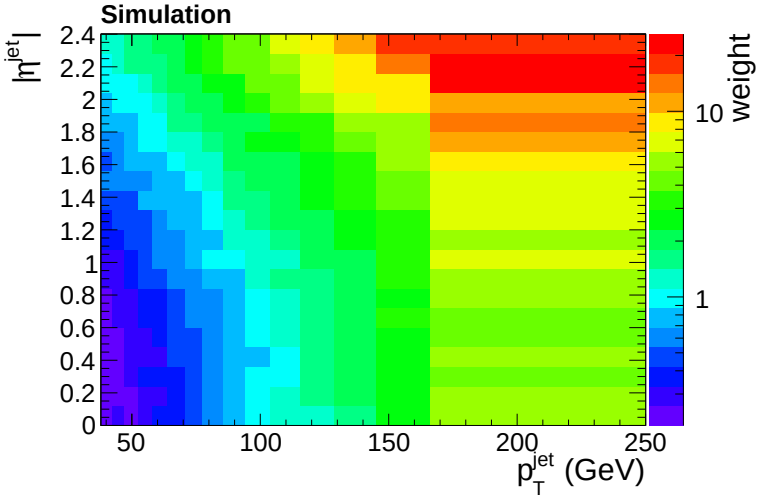


Figure 5.14: (p_T,η) dependent weights to reweigh the control sample jets to match the signal sample kinematics.

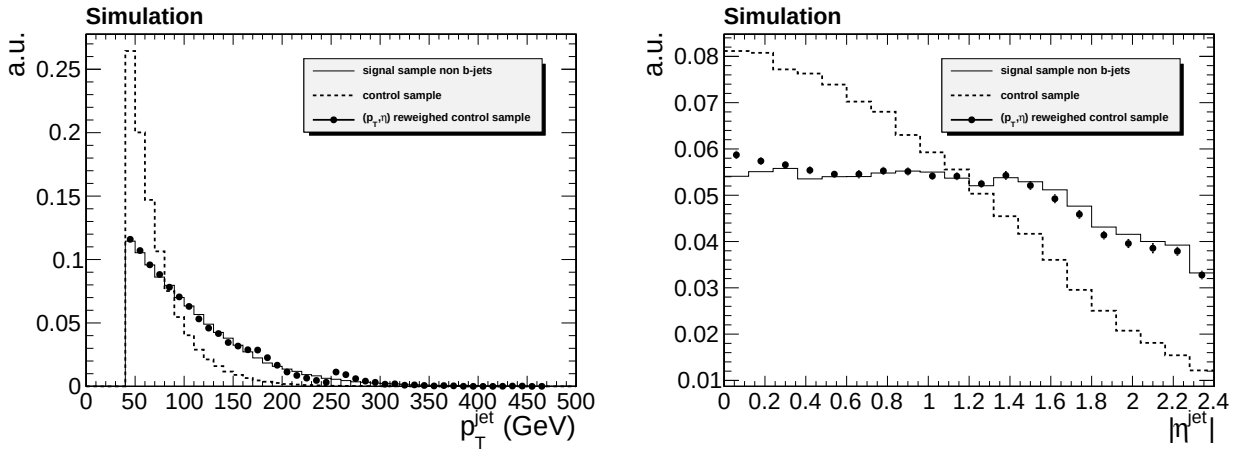


Figure 5.15: Comparison of the jet transverse momentum (p_T) and pseudo-rapidity (η) distributions between the signal sample non b jets and the control sample. After (p_T,η) reweighing, a good agreement is found between the signal sample non b jets and the control sample jets.

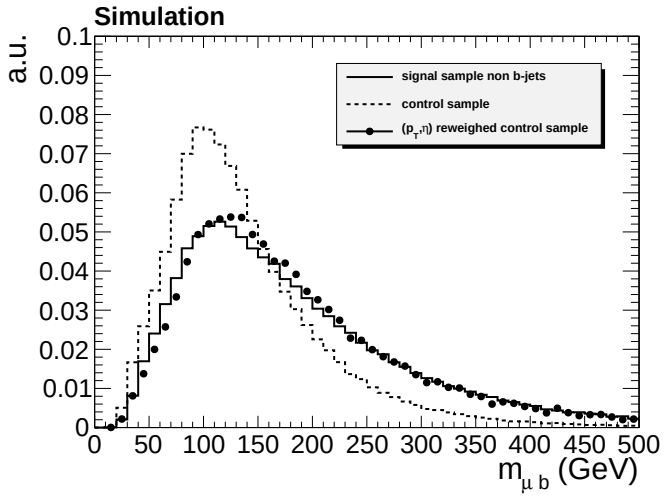


Figure 5.16: Comparison of the jet-lepton mass shape between the signal sample non b jets and the (p_T,η) reweighted control sample. A good agreement between the shapes can be observed when the (p_T,η) reweighing is applied on the control sample.

5.4 Studying the bias of the method

The method described to measure the b-tagging efficiency using the scale factor $\hat{F}^{CS}$ exhibits an overall bias affecting the method. Nevertheless, making appropriate choices for the definitions of the b-enriched and b-depleted samples as well as the cut on the minimal χ^2 value, also referred to as the method settings, minimises the bias.

The overall bias can be subdivided into three sources. The first two sources are more general while the third is specific for the Combined Secondary Vertex algorithm. First there appears to be a residual correlation between the b-tagging discriminator distribution for b jets and the jet-lepton mass. Secondly there is a residual bias due to the choice of the cut on the minimal χ^2. Finally there is a correlation between the b-tagging discriminator distribution for non b jets and the jet-lepton mass which appears mostly for the Combined Secondary Vertex algorithm. This final bias can be resolved by reweighing the shape in a data-driven manner. These effects will be explained in the following as well as the means to minimise them.

5.4.1 Correlation of Δ_b with M_{lj}

The invariance of the b-tagging discriminator distribution with respect to the jet-lepton mass observable is one of the corner stones in the proposed measurement of the b-tagging efficiency. If a shape difference in the b-tagging discriminator between the b-enriched and b-depleted samples would occur, the reconstructed discriminator distribution would be biased. Clearly this would then bias the measured b-tagging efficiency.

It has been observed that there is a correlation between the b-tagging discriminator value for b jets and the jet-lepton mass. This effect is illustrated by Figure 5.17 where

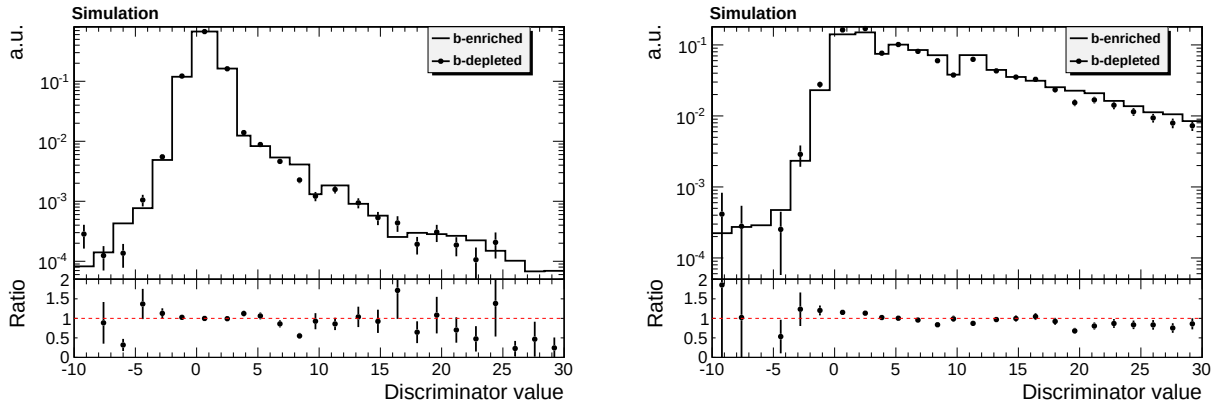


Figure 5.17: Comparison of the Track Counting High Efficiency discriminator shape between the b-enriched and b-depleted samples for non b jets (left) and for b jets (right)

the discriminator is compared between the b-enriched and b-depleted samples, after the p_T reweighing is applied on the latter. The comparison is made for non-b jets and b jets separately and a ratio between the two shapes is added to clearly show possible shape differences. As opposed to the non b jets ratio, showing no tendency, the b jets ratio plot shows a clear slope. This indicates differences in the discriminator shape between the b-enriched and b-depleted samples. Hence, a bias is introduced.

While studying this effect, it appeared to be sensitive to the choice of the lower and upper boundary of the b-depleted sample. Hence changing the boundaries of the b-depleted region provides a handle on the magnitude of the bias. To have a robust estimation of the boundaries to be used, pseudo-experiments have been used to test a large number of possible boundaries.

Each pseudo-experiment is a randomly selected subset of the full simulation sample reflecting a certain integrated luminosity. In this study and with the available simulation 20 pseudo-experiments of $2fb^{-1}$ each were drawn from the full simulated sample. Each of the 20 pseudo-experiments scans over some predefined possibilities to define the lower boundary and the upper boundary of the b-depleted sample. The lower bound of 70 GeV on the b-enriched sample is used and the upper boundary is chosen to match the lower boundary of the b-depleted sample. For each unique definition of the b-depleted sample, the relative bias is averaged over the biases obtained from the 20 pseudo-experiments. This averaging is needed since the experiments are low in statistics and hence statistical fluctuations occur. Figure 5.18 shows that the definition of the b-depleted sample affects the bias on the measured efficiency. It also proves that an optimal choice for the definition of the b-depleted sample exists coinciding with the green band on the plot where the relative bias is around zero. Moreover, defining the b-enriched sample as 70 $GeV < M_{lj} <$ 160 GeV and the b depleted sample as 160 $GeV < M_{lj} <$ 300 GeV puts the muon channel measurement in the stable zone where the relative bias is minimal. For the electron channel this stability is achieved with the same b-enriched region but a b-depleted region extended to 330 GeV. This procedure was applied for all b-tagging algorithms and working points yielding com-

patible results.

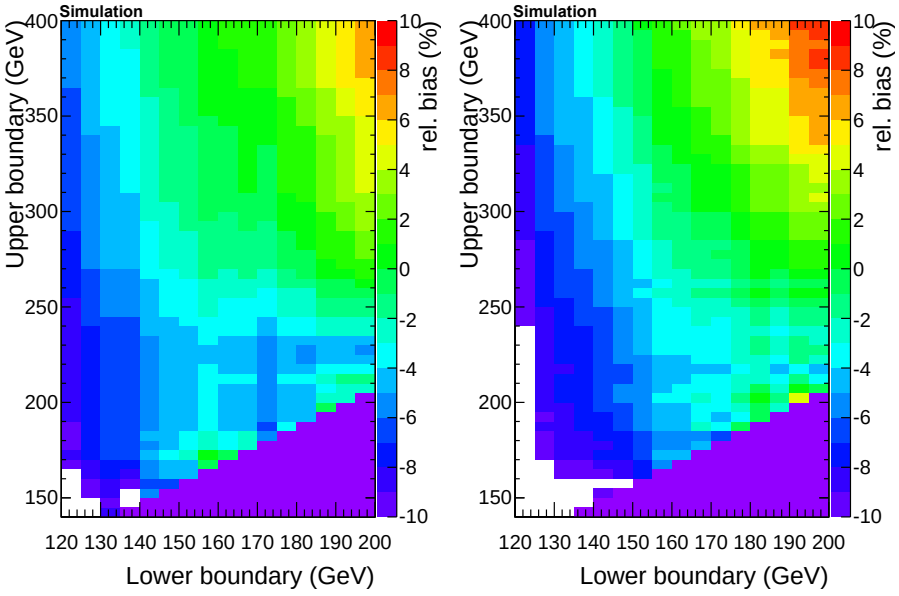


Figure 5.18: Scan of the definition of the b-depleted region for the TCHE Medium WP for μ+jets events (left) and e+jets events (right). The relative bias between the true and measured b-tagging efficiency is given on the z-axis. It is an average over 20 pseudo-experiments of each $2fb^{-1}$.

5.4.2 Effect of the χ^2_{min} cut

After minimizing the effect of the correlation between the b-tagging discriminator value for b jets and the jet-lepton mass, a residual bias on the measured b-tagging efficiency remains. This bias is caused by the choice of the cut on the minimal χ^2 value. As shown in Figure 5.19, the relative bias on the b-tagging efficiency varies from -2% to +6%. In the muon channel, cutting increasingly harder on the minimum χ^2 value decreases the bias on the b-tagging efficiency unto a cut value of about 90 where the bias starts to grow again in the opposite direction. Hence a cut value of 90 is chosen as it brings the bias to zero within the statistical precision. In the electron channel the same trend is observed around a cut value of 160, the value that will be used in the analysis.

To understand this evolution of the relative bias on the b-tagging efficiency with the cut on the minimal χ^2, one can look at Figure 5.20 which displays the data to simulation comparison of the minimal χ^2 value in each event. This distribution has a long tail consisting of background processes as well as $t\bar{t}$ signal events. The signal events in the tail are those where at least one of the four leading jets does not originate from the $t\bar{t}$ decay and hence these events fail the top quark mass and W boson mass hypothesis posed in the χ^2 formula. Loosening the minimal χ^2 cut increases the bias since it opens up the signal and control samples to a multitude of background events.

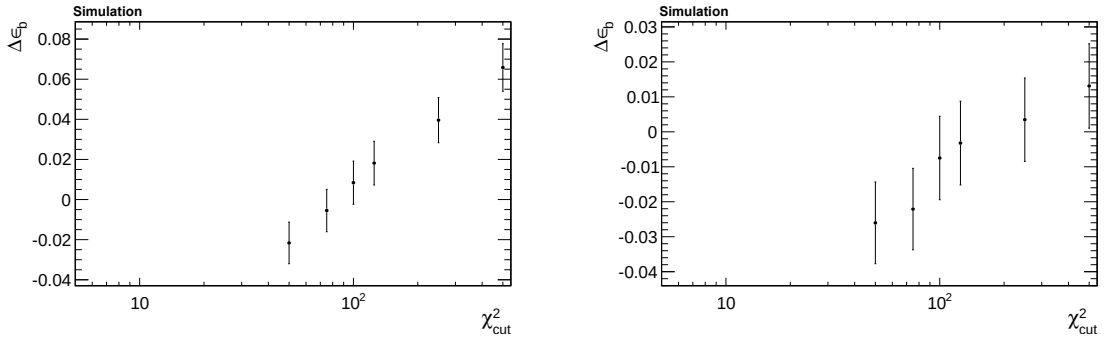


Figure 5.19: Relative bias between the true and measured b-tagging efficiency ($\Delta\epsilon_b$) as a function of the cut value on the χ^2_{min} value in each event for μ+jets events (left) and e+jets events (right). The statistical uncertainty on the relative bias is given for an integrated luminosity of $30 fb^{-1}$.

Conversely, cutting increasingly harder on the minimal χ^2 will decrease the background influence but will also put increasingly larger constraints on the two-jet and three-jet masses in these events. From a certain χ^2 cut on, cutting harder than this value would start to cut away the signal. The latter can be visualized by the number of events that are removed from the 2σ window around the top quark mass, compared to the original number of events in this window. Figure 5.21 shows that cutting much tighter than 90 would remove events from the signal peak and re-introduce the bias on the measured efficiency.

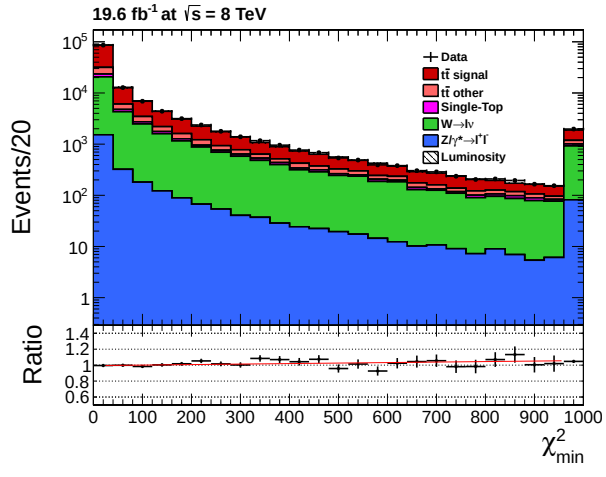


Figure 5.20: Distribution of the χ^2 value for the best jet-combination in each event (χ^2_{min}).

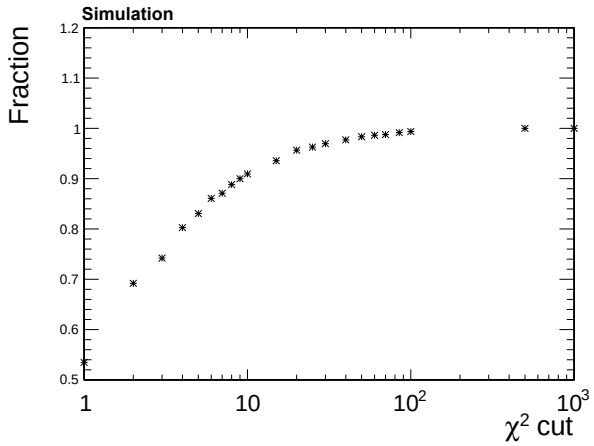


Figure 5.21: The fraction of events remaining in the top quark mass peak distribution with the χ^2 cut.

5.4.3 Correlation of light jets Δ_b with M_{lj} for CSV taggers

The overall method bias due to the correlation between Δ_b and M_{lj} has already been introduced. It was shown that the bias can be reduced to zero by optimizing the definition of the b-depleted region and choosing the appropriate cut on the minimum χ^2 value. However, for the Combined Secondary Vertex (CSV) algorithm a remaining bias is observed. This bias is as high as 5% for the loose working point decreasing towards tighter cut values on the discriminator. This residual bias, specific to the CSV tagger, is introduced by a residual correlation of the non b jet discriminator distribution with the jet-lepton mass. This correlation is shown in Figure 5.22 where the non b jets CSV shapes in the b-enriched and b-depleted samples are compared and the ratio is shown.

To remove this correlation, the discriminator shape in the b-depleted sample is reweighted. Since the effect stems from the non b jet fraction of the sample, the reweighing function can be derived from the control sample. In the control sample, the difference in discriminator shape between the b-enriched and b-depleted sample is taken after the control sample is reweighted to match the signal sample kinematics. The weights obtained by this method are shown in Figure 5.23. Nonetheless it has first to be shown that the TCHE anti-tag applied on the control sample, as explained in Section 5.3.1, does not interfere with this procedure. More precisely, the correlation between the CSV and TCHE algorithms is desired to be small.

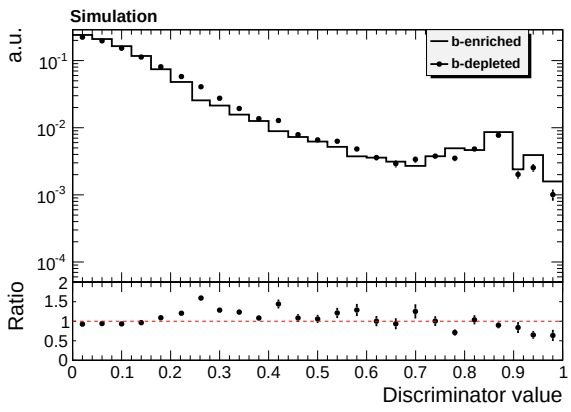


Figure 5.22: Comparison of the Combined Secondary Vertex discriminator shape between the b-enriched and b-depleted samples for non b jets.

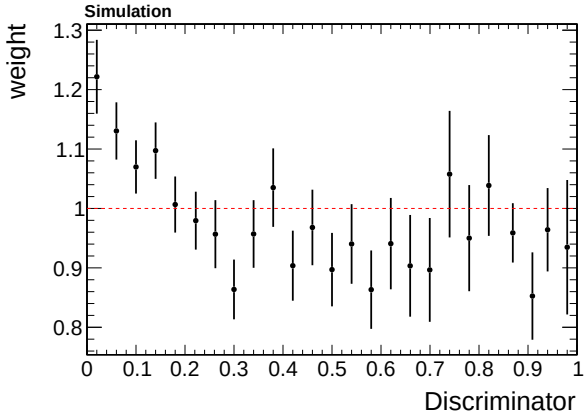


Figure 5.23: Weights to be applied to the Combined Secondary Vertex discriminator shape in the b-depleted samples. The weights are derived from the control-sample.

The left canvas in Figure 5.24 shows the correlation between the CSV and TCHE discriminators for non-b jets. A correlation of 36% is observed which indicates that a TCHE anti-b-tag will have an effect on the CSV shape. In the right canvas, the shape of the CSV distribution for non-b jets in the signal sample is compared before and after application of a TCHE anti-b-tag, revealing a disagreement in the shape.

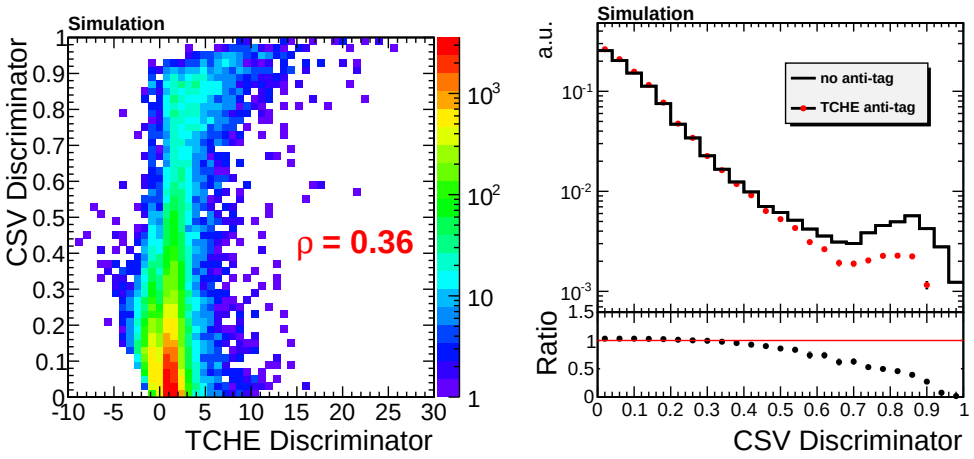


Figure 5.24: This figure shows the correlation between the TCHE and the CSV discriminators (left) and the effect of the anti-b-tagging criterion on the shape of the CSV discriminator for non-b jets in the signal sample.

It is clear that the CSV shape itself is affected by the anti-b-tagging, but one is interested in the shape difference between the b-enriched and b-depleted subsamples rather than the shape itself. To determine the effect of the control sample anti-b-tag on the applied weights, the weights are derived before and after the jets are anti-b-tagged. The ratio of these weights as a function of the discriminator value is shown in Figure 5.25. The weight ratio is fitted with a straight line that has an intercept of 1 and no slope. This indicates that the weights themselves are not affected by the anti-b-tag in the control sample and can thus be used without inflicting any bias.

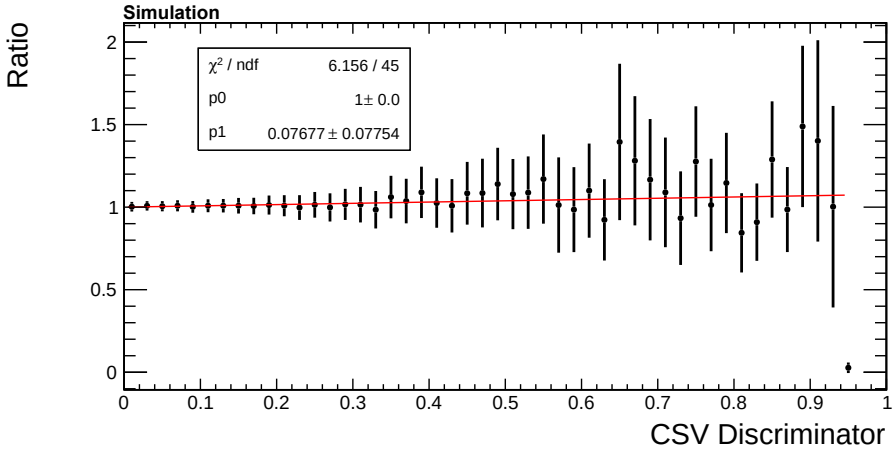


Figure 5.25: Ratio of the weights to be applied to remove the M_{lj}-dependence of the CSV discriminator shape for light jets determined before and after applying an anti-b-tag criterion. The ratio is fitted with a straight line.

5.5 Data-driven estimation of the inclusive b-tagging efficiency

A data-driven method has been developed to determine the b-tagging efficiency for any working point of any b-tagging algorithm. The first section outlined a method to reconstruct the b-tagging discriminator distribution by subdividing a b jet candidate sample into a b-enriched sample and a b-depleted sample. The b-depleted sample was then used to subtract the non b jet contribution in the b-enriched sample. To account for the scale difference of non b jets in both samples a scale factor F was applied. A data-driven estimator for this scale factor F was introduced in the next section utilizing a control sample of predominantly non b jets.

The data-driven measurement of F is included in the method to obtain results for the b-tagging efficiency using a fully data-driven estimator, $\hat{\epsilon}_b$. Figure 5.26 shows the b-tagging efficiency measured for the Track Counting High Efficiency (TCHE) algorithm as a function of the cut on the discriminator value compared to the expectation from the simulated samples along with the bias on the estimator. The numerical results for the three working points of this algorithm are provided in Table 5.3. For example in the muon channel, the bias estimated for the loose, medium and tight working point is respectively $(0.1 \pm 0.9)\%$, $(0.4 \pm 1.1)\%$ and $(0.9 \pm 1.3)\%$. The relative bias is hence statistically compatible with zero when projecting for a dataset corresponding to an integrated luminosity of $30 fb^{-1}$.

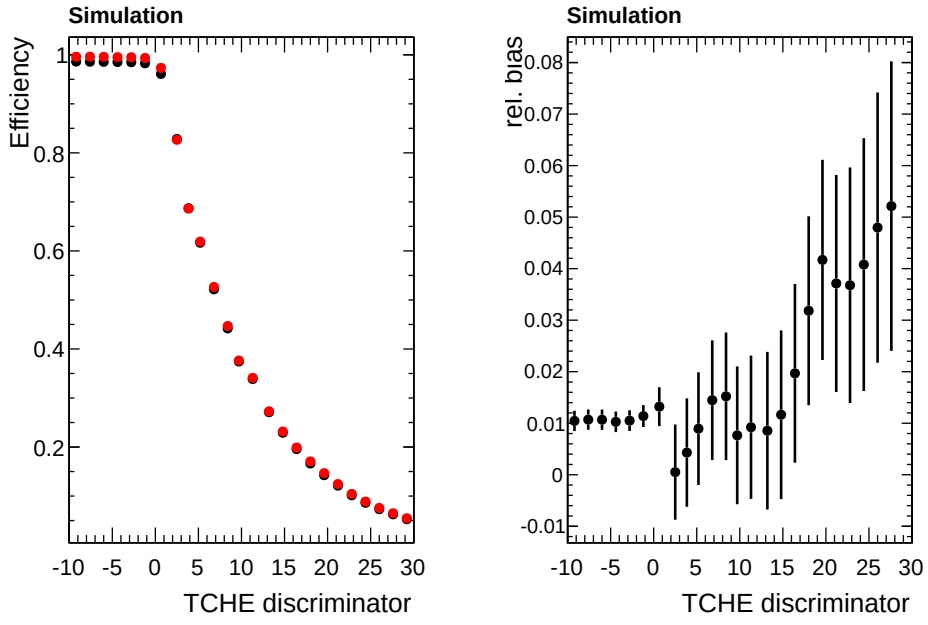


Figure 5.26: The Track Counting High Efficiency b tag efficiency (left) compared between true b jets in the b candidate sample (black markers) and the measured efficiency (red markers) and the relative bias between the two efficiencies (right)

Table 5.3: Results for the Track Counting High Efficiency tagger Loose, Medium and Tight working points for an integrated luminosity of $30.0 fb^{-1}$ simulated events. The efficiency provided by the method is compared to the true b tagging efficiency.

WP	$\epsilon_b^{True}(\%)$	$\hat{\epsilon}_b$ (%)	rel. bias (%)
L	82.6±0.2	82.7±0.7	0.1±0.9
M	68.3±0.2	68.6±0.7	0.4±1.1
T	33.8±0.2	34.1±0.4	0.9±1.3

5.6 Statistical properties of $\hat{\epsilon}_b$

So far no apparent bias on the b-tagging efficiency estimator was observed whilst performing the measurement on all the available simulation statistics. To study the statistical properties of this estimator in more detail, the simulation can be resampled.

A set of N samples is generated by randomly selecting events from the total simulation sample corresponding to a certain integrated luminosity, called pseudo-experiments. To study the statistical properties of the estimator, 750 samples have been generated corresponding to an integrated luminosity of $19.7 fb^{-1}$. The luminosity value is chosen such that each pseudo-experiment has enough statistics for the method to yield stable results while limiting the degree of correlation among the experiments as much as possible. For each sample, an estimate of the b tagging efficiency is obtained as shown in Figure 5.27. The pull of $\hat{\epsilon}_b$ for each experiment i can be defined as

$$Pull_i = \frac{\hat{\epsilon}_b^i - < \epsilon_b >}{\delta\epsilon_b^i}, \tag{5.6}$$

where $\hat{\epsilon}_b^i$ and $\delta\epsilon_b^i$ are the estimated b-tagging efficiency and statistical uncertainty in sample i. If the statistical uncertainty and the residual $(\hat{\epsilon}_b^i - < \epsilon_b >)$ are properly estimated, the pull distribution should have unit width. In Figure 5.27 , the pull distribution is fitted with a gauss function.

The gaussian fit to the pull distribution yields a mean of -0.014±0.04 and a width of 1.14±0.04. It needs to be taken into account that there is a certain degree of correlation between the different experiments due to the limited simulated statistics at hand. Hence it can be concluded that the statistical uncertainty is reliable.

5.7 Systematic uncertainties

In this section, the systematic uncertainty is broken down into its different contributions for the b tagging efficiency and the data to simulation scale factors. A detailed explanation of the different contributions is provided in the following. For the b tagging efficiency the variation of the methods bias, as defined in eq. (5.4), between the systematic variation and the nominal simulation sample is taken as systematic uncertainty. For the data to simulation scale factor, the absolute difference between the

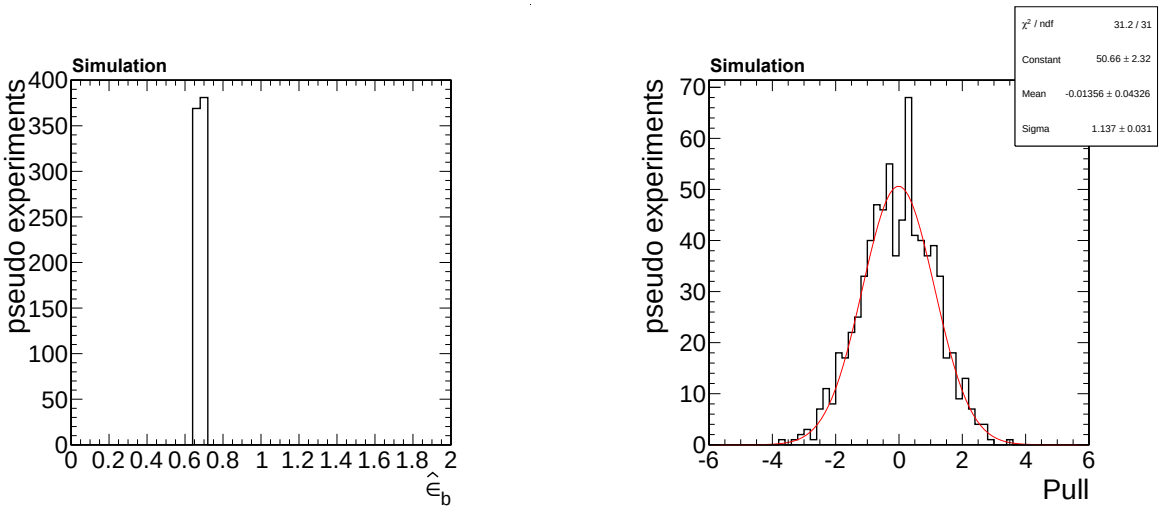


Figure 5.27: Distribution of the b tagging efficiency estimator (left) and the pull distribution (right) for the TCHE Medium WP using 750 pseudo-experiments.

nominal scale factor and the scale factor using the varied sample is used. For each contribution we consider an up- and downwards shift and the biggest effect among the two is quoted as systematic uncertainty. Finally all contributions are added in quadrature to obtain the total systematic uncertainty on the measurement. The actual systematic uncertainties for each algorithm are provided in the results section 5.8.

Jet Energy Scale

To evaluate the effect of jet energy scale variations on the result, the energy scale of the jets is varied within the p_T and η dependent 1-σ uncertainty band. This variation accounts for the uncertainty on the pileup corrections as well as the uncertainty on the flavor dependence of the jet energy scale which were measured in data as described in [120]. In addition to this, an extra uncertainty due to absolute scale mismatches between Z+jets and γ+jets is added in quadrature along with an uncertainty for jets with $|\eta| > 1.3$ due to unaccounted conditions in the relative scale.

Jet Energy Resolution

To account for differences in the jet energy resolution between data and simulation, a Jet Energy Resolution systematic is taken into account. It was measured that the jet energy resolution in data is up to 10% worse than in simulation depending on the jet η [120]. Thus a jet η-dependent smearing is applied on the jets in simulated events corresponding to a $\pm 1\sigma$ variation of their energy resolution.

Missing Transverse Energy (MET) Unclustered Energy

The Missing Transverse Energy (MET) is a composite object since it cannot be measured directly. Instead it is inferred for each event by looking at the sum of the

transverse momenta of all other objects like leptons and jets. Since jet transverse momenta play a role in the determination of MET, the Jet Energy Scale and Jet Energy Resolution systematics also need to be propagated to MET. This effect on the MET is included in the previous two systematics, yet one additional source of uncertainty remains.
Additional to the reconstructed objects within the MET, there is an unclustered energy component. This component does not correspond to any reconstructed object in the event and the uncertainty on its determination equals 10%.
To determine a systematic uncertainty due to this unclustered energy, the MET gets stripped of all reconstructed jets and leptons in the event. Jets not passing the tight selection criteria outlined in Chapter 4 are also considered in this procedure. Subsequently, the remaining energy is varied by $\pm 10\%$ and finally all the objects that were stripped are added again. This procedure yields a $\pm 1\sigma$ MET value to be used in the analysis to determine the $\pm 1\sigma$ effect on the b-tagging efficiency due to unclustered energy in MET.

Pile up

The simulated events in this analysis are produced with contributions from pile up, additional proton-proton interactions in the event. The pileup effect in simulation is then corrected to the expected pileup effect in data by reweighing the number of pile up interactions to data. To build the event weights, the expected number of interactions from data is estimated by using the connection between pile up and the instantaneous luminosity. Hence, the main sources of uncertainty on the pile up in simulation arise from the modelling of the pile up as well as the uncertainties on the total inelastic proton-proton cross section and luminosity measurement on the applied event weights. The systematic uncertainty due to pile up is then calculated by varying the mean number of interactions by $\pm 5\%$ covering the uncertainty sources mentioned earlier.

Btag Method Settings

In the analysis to estimate the b-tagging efficiency as well as the mis-tagging efficiency, the choice of the b-depleted region was taken by minimisation of the relative bias on the respective efficiencies. As this method is data-driven, no theory uncertainties are considered on the b-tagging part of the analysis. Nevertheless, simulated events were used in the determination of the optimal b-depleted region and as such a systematic for this choice is added.
To attribute a method uncertainty due to the tuning of the b-depleted region, the region is shifted by ± 30 GeV. This shift covers most of the stable region in the phase space of possible definitions of the b-depleted region and thus gives a good estimate of the uncertainty due to the particular chosen definition.

Top Quark Mass

As the top quark mass from simulation is used to construct the χ^2 for each jet combination a systematic uncertainty on possible variation of this mass is evaluated. To

obtain this systematic uncertainty, simulated top quark pair decays are used where the mass is shifted ±9 GeV. Since the top quark mass is known with a precision of ≈1 GeV in the top quark mass is required, the effect is divided by a factor 9.

5.8 Results at $\sqrt{s} = 8$ TeV

This section provides a data to simulation comparison plot, the measured b-tagging discriminator distribution and b-tagging efficiencies for the Track Counting High Efficiency and Combined Secondary Vertex b-tagging algorithms at 8 TeV. Also the data to simulation scale factors and the systematics tables are provided.

5.8.1 Track Counting High Efficiency (TCHE)

In this section the results for the Track Counting High Efficiency (TCHE) tagger are provided. In Figure 5.28, the data to simulation comparison is shown for the TCHE discriminator in μ+jets and e+jets events respectively. In both channels, a fairly good description of the data by the simulation is observed in the discriminator region [0,30] whereas below TCHE discriminator value 0, the data overshoots the distribution from simulation.

Interesting to note as well is that while W+jets and Z+jets strongly peak at low TCHE discriminator values, $t\bar{t}$ and to a smaller extent single-top dominate the tail. This shows the discriminating power of b-tagging.

Since the TCHE discriminator is defined as the signed IP significance of the second highest IP significance track, this region is dominated by negative IP tracks as discussed in Section 3.4.1. These tracks are most likely badly reconstructed and hard to simulate.

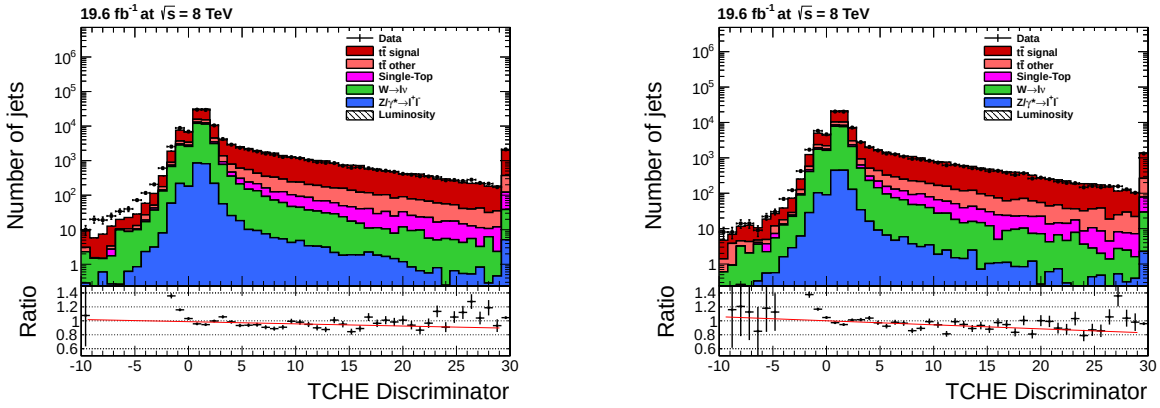


Figure 5.28: Distribution of the TCHE discriminator for μ+jets events (left) and e+jets events (right). The distributions of all simulated processes are normalized to the integrated luminosity of the data.

In this method, the b-tagging efficiency is measured by first reconstructing the discriminator distribution for true b jets. This reconstructed distribution is shown for both μ+jets and e+jets events in Figure 5.29 along with the resulting b-tagging

efficiency and data to simulation scale factor ($SF_b = \hat{\epsilon}_b/\epsilon_b^{True}$) as a function of the Working Point (WP), or threshold. The "loose", "medium" and "tight" Working Points are indicated by the arrows in the efficiency plot. For these specific points, the residual relative bias on the efficiency is provided in Table 5.4 which is statistically compatible with zero. The efficiencies measured from data are provided in Table 5.5 along with their statistical and systematic uncertainties. In the latter Table, the data to simulation scale factors are provided as well. These are the most important result as these can be used by other analyses to correct their simulation based selection efficiency for the difference in b-tagging efficiency between data and simulation.

Finally, the breakdown of the systematic uncertainty in all it's different sources is provided in Table 5.6. The dominant systematic is found to be the variation of the b-depleted region caused by the usage of simulated events to define it.

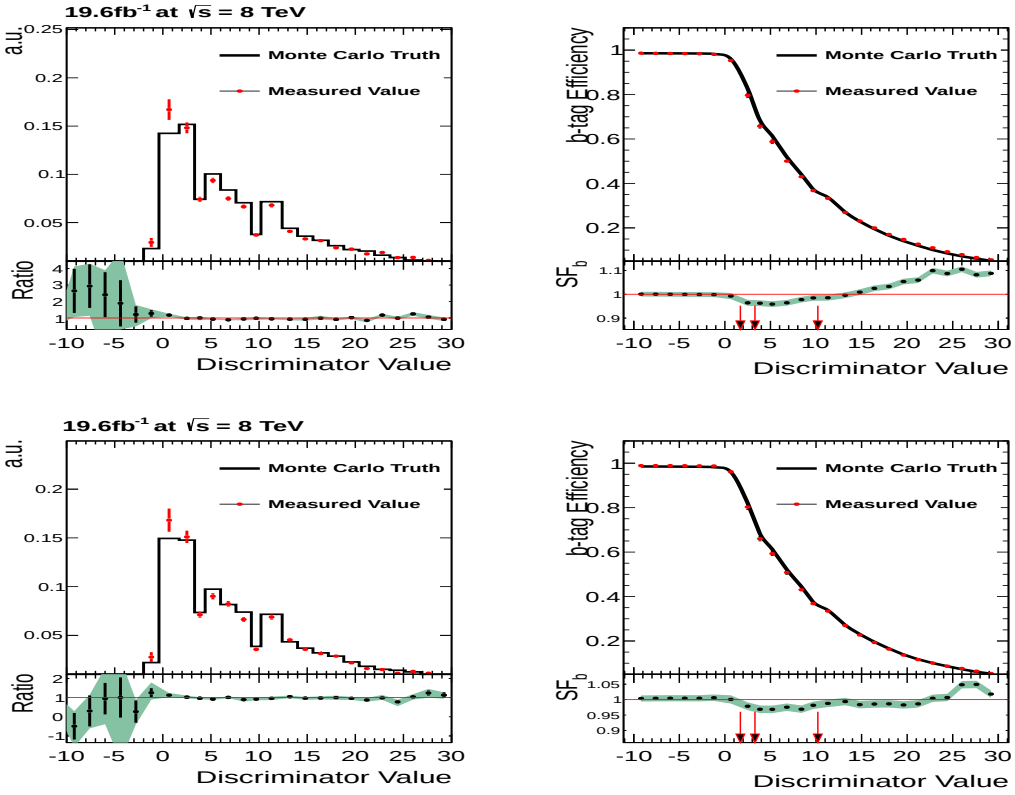


Figure 5.29: Measured b-jets TCHE Discriminator distribution in the b-enriched sample compared to simulation (left) and the measured b-tagging efficiency, and data to simulation scale factors (right) for μ+jets events (top) and e+jets events (bottom). The arrows depict the loose, medium and tight working points respectively. The green band shows the combined statistical and systematic uncertainty.

Table 5.4: The efficiency provided by the method compared to the true b tagging efficiency for an integrated luminosity of $30.0 fb^{-1}$ simulated events.

Channel	WP	$\epsilon_b^{True}(\%)$	$\hat{\epsilon}_b$ (%)	rel. bias (%)
μ+jets	L	82.6±0.2	82.7±0.7	0.1±0.9
	M	68.3±0.2	68.6±0.7	0.4±1.1
	T	33.8±0.2	34.1±0.4	0.9±1.3
e+jets	L	82±0.2	82.7±0.8	1±1
	M	68.1±0.2	68.4±0.8	0.4±1.2
	T	33.8±0.2	34.5±0.5	2.1±1.6

Table 5.5: Measured efficiencies and data over simulation scale factors for the TCHE discriminator.

Channel	WP	$\epsilon_b^{True}(\%)$	$\hat{\epsilon}_b$ (%)	$\hat{\epsilon}_b/\epsilon_b^{True}$
μ+jets	L	82.6± 0.2	79.7± 0.9 (stat.) ± 1.0 (syst.)	0.965± 0.011 (stat.) ± 0.010 (syst.)
	M	68.3± 0.2	65.8± 0.9 (stat.) ± 1.2 (syst.)	0.963± 0.013 (stat.) ± 0.012 (syst.)
	T	33.8± 0.2	33.3± 0.5 (stat.) ± 1.2 (syst.)	0.985± 0.016 (stat.) ± 0.012 (syst.)
e+jets	L	82.0± 0.2	80.3± 1 (stat.) ± 1.3 (syst.)	0.979± 0.012 (stat.) ± 0.013 (syst.)
	M	68.1± 0.2	66.0± 1 (stat.) ± 1.4 (syst.)	0.969± 0.015 (stat.) ± 0.014 (syst.)
	T	33.8± 0.2	33.4± 0.6 (stat.) ± 1.7 (syst.)	0.988± 0.019 (stat.) ± 0.017 (syst.)

Table 5.6: Systematic uncertainty sources on the b-tagging efficiency for the TCHE working points.

	μ+jets (%)						e+jets (%)					
WP →	Loose		Medium		Tight		Loose		Medium		Tight	
Systematic ↓	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF
Jet Energy Scale	0.4	0.4	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6	0.6
Jet Energy Resolution	0.2	0.2	0.3	0.3	0.3	0.3	0.6	0.6	0.6	0.6	0.9	0.9
PileUp	0.1	0.1	0.2	0.2	0.3	0.3	0.3	0.3	0.0	0.0	0.3	0.3
MET Unclustered Energy	0.1	0.1	0.3	0.3	0.3	0.3	0.4	0.4	0.3	0.3	0.3	0.3
Top quark mass	0.1	0.1	0.2	0.2	0.2	0.2	0.3	0.3	0.2	0.2	0.2	0.2
Right region definition	0.9	0.8	0.9	0.9	0.9	0.9	0.9	0.9	1.0	1.0	1.2	1.2
Total	1	1	1.2	1.2	1.2	1.2	1.3	1.3	1.4	1.4	1.7	1.7

5.8.2 Combined Secondary Vertex (CSV)

In this section the results for the Combined Secondary Vertex (CSV) tagger are provided. The data to simulation comparison is shown in Figure 5.30 for μ+jets and e+jets events respectively. In both channels, a fair description of the data is observed but with data consistently above simulation in the region [0.2,0.8]. It is explained in Section 3.4.4 that the CSV algorithm is quite complex and takes many input parameters. Any discrepancy between data and simulation for one of its input variables can cause discrepancies in the final discriminator. This indicates the importance of measuring the CSV performance on data.

Just as with the TCHE, the discriminating power between $t\bar{t}$ signal and background processes becomes apparent when looking at the corresponding discriminator distributions. While the background processes are mostly distributed around 0, the $t\bar{t}$ signal peaks around 1 where the jet is most likely to be a b jet.

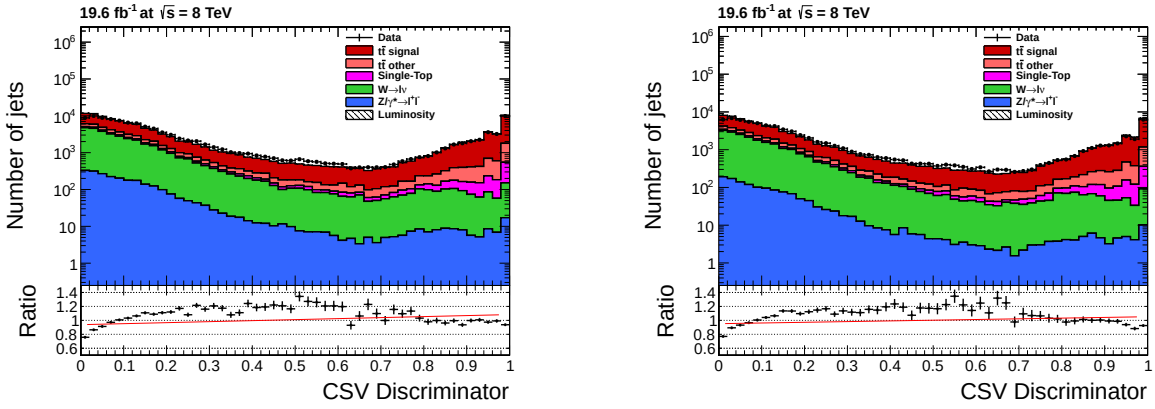


Figure 5.30: Distribution of the CSV discriminator for μ+jets events (left) and e+jets events (right). The distributions of all simulated processes are normalized to the integrated luminosity of the data.

The reconstructed CSV discriminator distribution for true b jets is shown for both μ+jets and e+jets events in Figure 5.31 along with the resulting b-tagging efficiency and data to simulation scale factor as a function of the Working Point (WP), or threshold. The "loose", "medium" and "tight" Working Points are indicated by the arrows in the efficiency plot. For these specific points, the residual relative bias on the efficiency is provided in Table 5.7 which is statistically compatible with zero. The efficiencies measured from data are provided in Table 5.8 along with their statistical and systematic uncertainties. In this Table, the data to simulation scale factors are provided as well.

Finally, the breakdown of the systematic uncertainty in all it's different sources is provided in Table 5.9. The dominant systematics are found to be the variation of the b-depleted region caused by the usage of simulated events to define it and the Jet Energy Scale (JES) uncertainty.

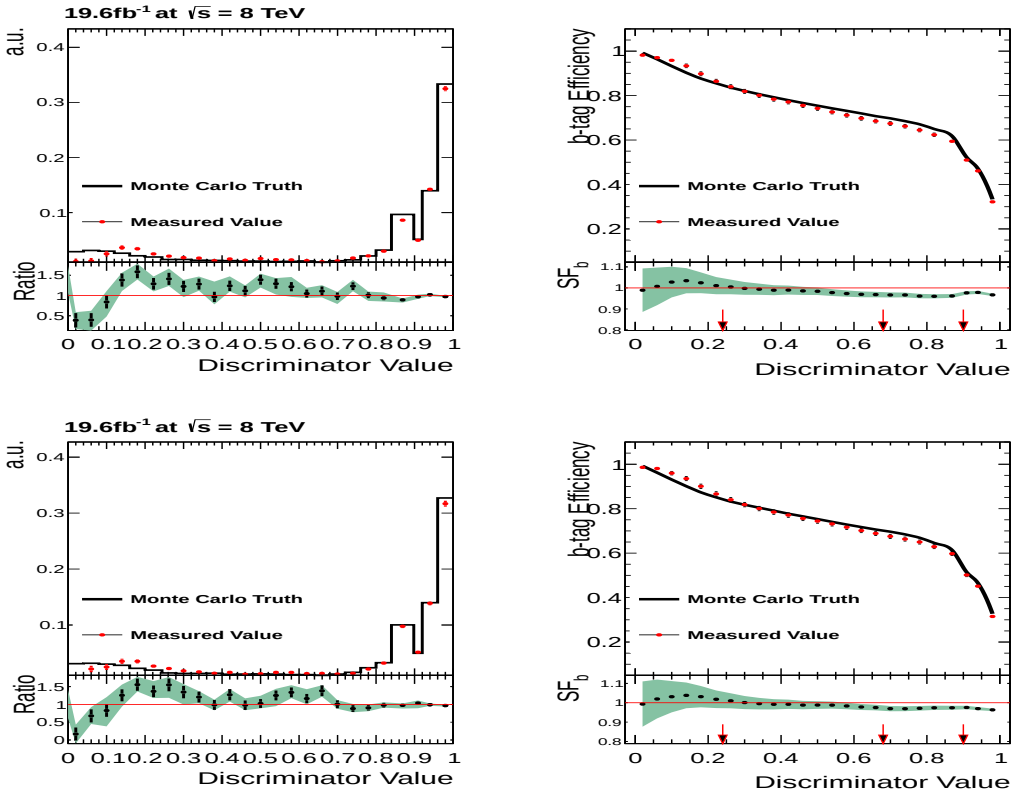


Figure 5.31: Measured b-jets CSV Discriminator distribution in the b-enriched sample compared to simulation (left) and the measured b-tagging efficiency, and data to simulation scale factors (right) for μ+jets events (top) and e+jets events (bottom). The arrows depict the loose, medium and tight working points respectively. The green band shows the combined statistical and systematic uncertainty.

Table 5.7: The efficiency provided by the method compared to the true b tagging efficiency for an integrated luminosity of $30.0 fb^{-1}$ simulated events.

Channel	WP	$\epsilon_b^{True}(\%)$	$\hat{\epsilon}_b$ (%)	rel. bias (%)
μ+jets	L	83.6±0.2	83.3±0.7	-0.4±0.9
	M	69.8±0.2	69.7±0.7	-0.1±1
	T	52.2±0.2	52.8±0.6	1.2±1.2
e+jets	L	83.2±0.2	83.6±0.8	0.6±1
	M	69.6±0.2	69.1±0.7	-0.6±1
	T	51.4±0.2	51.8±0.6	1±1.2

Table 5.8: Measured efficiencies and data over simulation scale factors for the CSV discriminator.

Channel	WP	$\epsilon_b^{True}(\%)$	$\hat{\epsilon}_b$ (%)	$\hat{\epsilon}_b/\epsilon_b^{True}$
μ+jets	L	83.6± 0.2	84.0± 0.9 (stat.) ± 0.8 (syst.)	1.005± 0.011 (stat.) ± 0.008 (syst.)
	M	69.8± 0.2	67.4± 0.8 (stat.) ± 1.0 (syst.)	0.966± 0.012 (stat.) ± 0.010 (syst.)
	T	52.2± 0.2	51.0± 0.7 (stat.) ± 1.2 (syst.)	0.977± 0.014 (stat.) ± 0.012 (syst.)
e+jets	L	83.2± 0.2	84.0± 1 (stat.) ± 1.1 (syst.)	1.01± 0.012 (stat.) ± 0.011 (syst.)
	M	69.6± 0.2	67.5± 0.9 (stat.) ± 1.3 (syst.)	0.97± 0.013 (stat.) ± 0.013 (syst.)
	T	51.4± 0.2	50.2± 0.8 (stat.) ± 1.6 (syst.)	0.977± 0.016 (stat.) ± 0.06 (syst.)

Table 5.9: Systematic uncertainty sources on the b-tagging efficiency for the CSV working points.

	μ+jets (%)						e+jets (%)					
WP $\rightarrow$	Loose		Medium		Tight		Loose		Medium		Tight	
Systematic $\downarrow$	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF
Jet Energy Scale	0.4	0.4	0.6	0.6	0.8	0.8	0.6	0.6	0.6	0.6	0.8	0.8
Jet Energy Resolution	0.1	0.1	0.3	0.3	0.4	0.4	0.5	0.5	0.6	0.6	0.8	0.8
PileUp	0.2	0.2	0.3	0.3	0.2	0.2	0.2	0.2	0.2	0.2	0.4	0.4
MET Unclustered Energy	0.1	0.1	0.0	0.0	0.4	0.4	0.1	0.1	0.3	0.3	0.4	0.4
Top quark mass	0.2	0.2	0.2	0.2	0.1	0.1	0.2	0.2	0.1	0.1	0.1	0.1
Right region definition	0.6	0.6	0.7	0.7	0.8	0.8	0.7	0.7	1.0	1.0	1.0	1.0
Total	0.8	0.8	1	1	1.2	1.2	1.1	1.1	1.3	1.3	1.6	1.6

5.9 Results at $\sqrt{s} = 7$ TeV

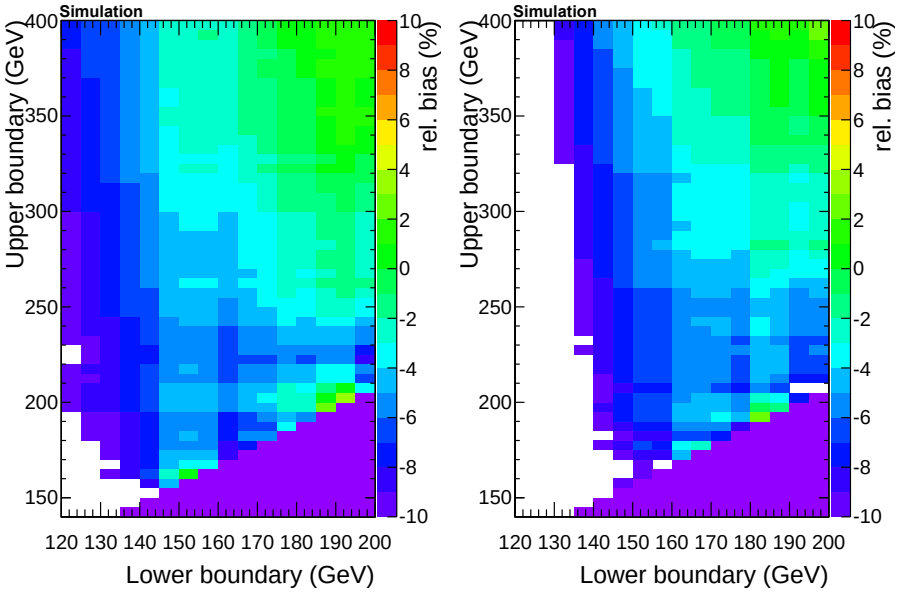


Figure 5.32: Scan of the definition of the b-depleted region for the TCHE Medium WP for μ+jets events (left) and e+jets events (right). The relative bias between the true and measured b-tagging efficiency is given on the z-axis and is averaged over 20 pseudo-experiments of $5fb^{-1}$.

In the 7 TeV analysis, the definition of the b-enriched and b-depleted samples is determined by throwing pseudo-experiments using the same method as for the 8 TeV analysis. Each pseudo-experiment is a randomly selected subset of the full simulation sample reflecting a certain integrated luminosity. In this study 20 pseudo-experiments of $5fb^{-1}$ were drawn. The luminosity increase with respect to the $2fb^{-1}$ experiments at 8 TeV is chosen because of the lower $t\bar{t}$ cross section. Each of the 20 pseudo-experiments scans over different possibilities to define the lower boundary and the upper boundary of the b-depleted sample. For all experiments, the lower bound of 70 GeV on the b-enriched sample is used and the upper boundary is chosen to match the lower boundary of the b-depleted sample. For each unique definition of the b-depleted sample, the relative bias is averaged over the 20 pseudo-experiments. This averaging is needed since the experiments are low in statistics and hence statistical fluctuations occur. Figure 5.32 shows that the definition of the b-depleted sample affects the bias on the measured efficiency. It also proves that an optimal choice for the definition of the b-depleted sample exists coinciding with the green band on the plot. Moreover, defining the b-enriched sample again as 70 $GeV < M_{lj} < 170$ GeV and the b depleted sample as 170 $GeV < M_{lj} < 300$ GeV yields a minimal bias in both the muon and the electron channel.

After minimizing the effect of the correlation between the b-tagging discriminator

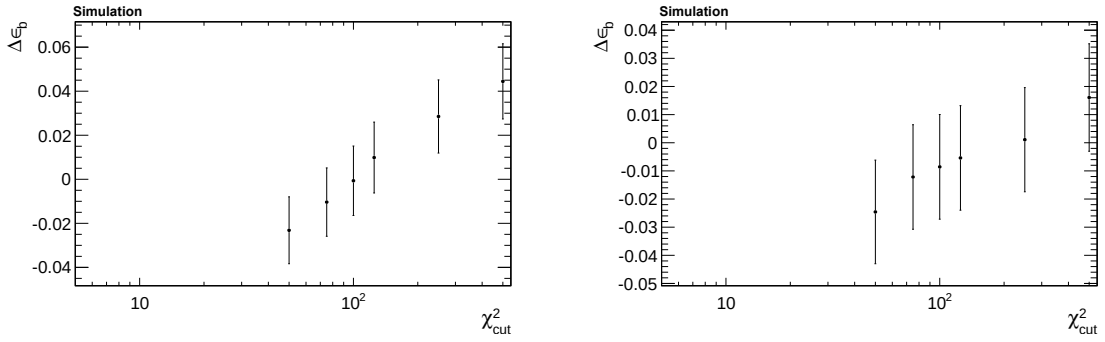


Figure 5.33: Relative bias between the true and measured b-tagging efficiency ($\Delta\epsilon_b$) as a function of the cut value on the χ^2_{min} value in each event for μ+jets events (left) and e+jets events (right). The statistical uncertainty on the relative bias is given for an integrated luminosity of $20 fb^{-1}$.

value for b jets and the jet-lepton mass, a residual bias on the measured b-tagging efficiency remains. In the muon channel, the bias on the b-tagging efficiency is minimal around a cut value of about 100 where the bias starts to grow again in the opposite direction. Hence a cut value of 100 is chosen as it brings the bias to zero within the statistical precision. In the electron channel the same trend is observed around a cut value of 100, the value that will be used in the analysis.

5.9.1 Track Counting High Efficiency (TCHE)

In this section the results for the Track Counting High Efficiency (TCHE) tagger are provided. In Figure 5.34, the data to simulation comparison is shown for the TCHE discriminator in μ+jets and e+jets events respectively. The same level of agreement is found compared to the 8 TeV analysis.

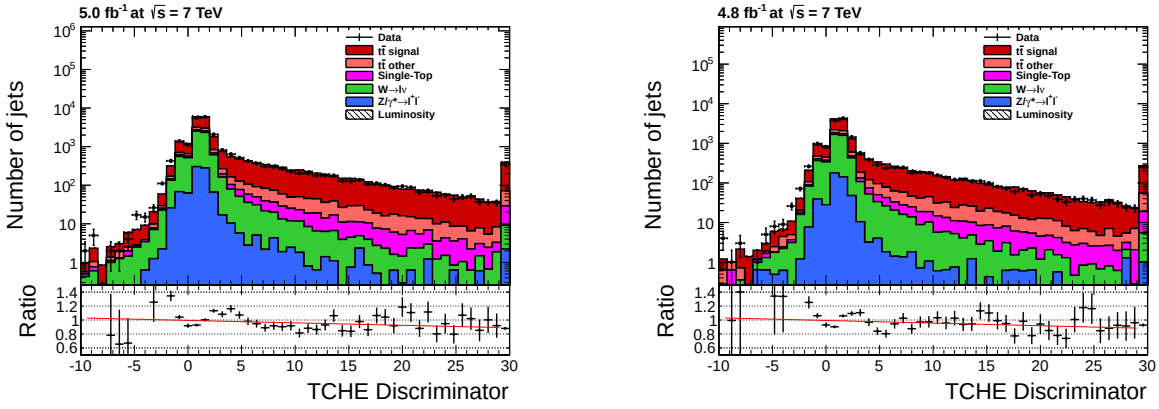


Figure 5.34: Distribution of the TCHE discriminator for μ+jets events (left) and e+jets events (right). The distributions of all simulated processes are normalized to the integrated luminosity of the data.

The reconstructed b-tagging discriminator distribution is shown for both μ+jets

and e+jets events in Figure 5.34 along with the resulting b-tagging efficiency and data to simulation scale factor as a function of the Working Point (WP), or threshold. The "loose", "medium" and "tight" Working Points are indicated by the arrows in the efficiency plot. For these specific points, the residual relative bias on the efficiency is provided in Table 5.10 which is statistically compatible with zero. The efficiencies measured from data are provided in Table 5.11 along with their statistical and systematic uncertainties. In this Table, the data to simulation scale factors are provided as well.

Finally, the breakdown of the systematic uncertainty in all it's different sources is provided in Table 5.12. The dominant systematics are found to be the variation of the b-depleted region caused by the usage of simulated events to define it and the Jet Energy Scale (JES) and Jet Energy Resolution (JER) uncertainty.

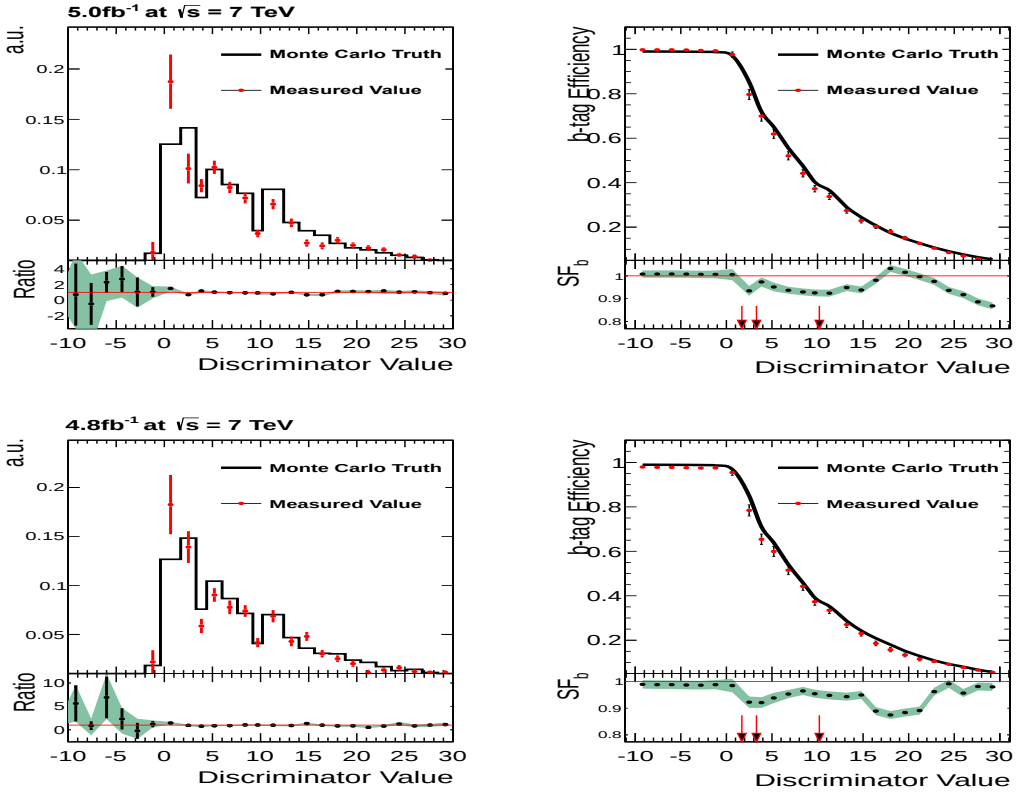


Figure 5.35: Measured b-jets TCHE Discriminator distribution in the b-enriched sample compared to simulation (left) and the measured b-tagging efficiency, and data to simulation scale factors (right) for μ+jets events (top) and e+jets events (bottom). The arrows depict the loose, medium and tight working points respectively. The green band shows the combined statistical and systematic uncertainty.

Table 5.10: The efficiency provided by the method compared to the true b tagging efficiency for an integrated luminosity of $30.0 fb^{-1}$ simulated events.

Channel	WP	$\epsilon_b^{True}(\%)$	$\hat{\epsilon}_b$ (%)	rel. bias (%)
μ+jets	L	85.3±0.2	85.1±1.2	-0.2±-1.4
	M	71.9±0.3	71.9±1.1	0±-1.6
	T	36.6±0.3	36.5±0.7	-0.3±-2.1
e+jets	L	85±0.3	84.5±1.4	-0.6±-1.7
	M	71±0.3	70.4±1.3	-0.9±-1.9
	T	35.3±0.3	35.7±0.8	1.1±-2.4

Table 5.11: Measured efficiencies and data over simulation scale factors for the TCHE discriminator.

Channel	WP	$\epsilon_b^{True}(\%)$	$\hat{\epsilon}_b$ (%)	$\hat{\epsilon}_b/\epsilon_b^{True}$
μ+jets	L	85.3± 0.2	79.7± 2.3 (stat.) ± 1.6 (syst.)	0.934± 0.027 (stat.) ± 0.016 (syst.)
	M	71.9± 0.3	70± 2.3 (stat.) ± 1.9 (syst.)	0.974± 0.032 (stat.) ± 0.019 (syst.)
	T	36.6± 0.3	33.8± 1.4 (stat.) ± 1.5 (syst.)	0.923± 0.039 (stat.) ± 0.015 (syst.)
e+jets	L	85± 0.3	78.4± 2.5 (stat.) ± 1.2 (syst.)	0.922± 0.03 (stat.) ± 0.012 (syst.)
	M	71± 0.3	65.4± 2.4 (stat.) ± 1.8 (syst.)	0.921± 0.034 (stat.) ± 0.018 (syst.)
	T	35.3± 0.3	33.5± 1.5 (stat.) ± 1.7 (syst.)	0.949± 0.043 (stat.) ± 0.017 (syst.)

Table 5.12: Systematic uncertainty sources on the b-tagging efficiency for the TCHE working points.

	μ+jets (%)						e+jets (%)					
WP →	Loose		Medium		Tight		Loose		Medium		Tight	
Systematic ↓	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF
Jet Energy Scale	0.7	0.7	0.8	0.8	0.3	0.3	0.6	0.6	0.9	0.8	0.9	0.9
Jet Energy Resolution	0.9	0.9	1.0	1.0	0.5	0.5	0.5	0.5	0.6	0.6	0.3	0.3
PileUp	0.2	0.2	0.1	0.1	0.5	0.5	0.2	0.2	0.2	0.2	0.0	0.0
MET Unclustered Energy	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.4	0.4	0.3	0.3
Top quark mass	0.5	0.5	0.6	0.6	0.5	0.5	0.2	0.2	0.0	0.0	0.3	0.3
Right region definition	0.9	0.9	1.3	1.3	1.1	1.1	0.9	0.9	1.4	1.4	1.4	1.4
Total	1.6	1.6	1.9	1.9	1.5	1.5	1.2	1.2	1.8	1.8	1.7	1.7

5.9.2 Combined Secondary Vertex (CSV)

In this section the results for the Combined Secondary Vertex (CSV) tagger are provided. The data to simulation comparison is shown in Figure 5.36 for μ+jets and e+jets events respectively. The same level of agreement between data and simulation is observed as for the 8 TeV analysis.

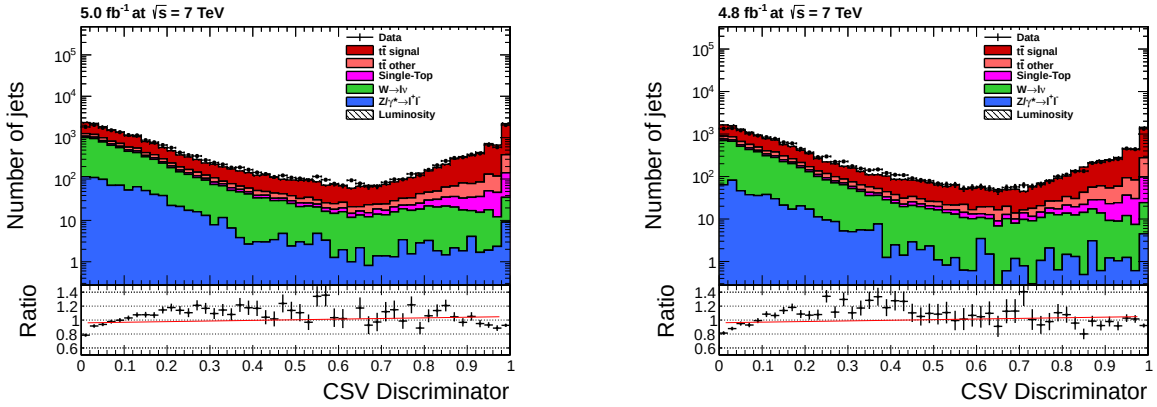


Figure 5.36: Distribution of the CSV discriminator for μ+jets events (left) and e+jets events (right). The distributions of all simulated processes are normalized to the integrated luminosity of the data.

The reconstructed CSV discriminator distribution for true b jets is shown for both μ+jets and e+jets events in Figure 5.37 along with the resulting b-tagging efficiency and data to simulation scale factor as a function of the Working Point (WP), or threshold. The "loose", "medium" and "tight" Working Points are indicated by the arrows in the efficiency plot. For these specific points, the residual relative bias on the efficiency is provided in Table 5.13 which is statistically compatible with zero. The efficiencies measured from data are provided in Table 5.14 along with their statistical and systematic uncertainties. In this Table, the data to simulation scale factors are provided as well.

Finally, the breakdown of the systematic uncertainty in all it's different sources is provided in Table 5.15. The dominant systematics are found to be the variation of the b-depleted region caused by the usage of simulated events to define it and the Jet Energy Scale (JES) and Jet Energy Resolution (JER) uncertainty.

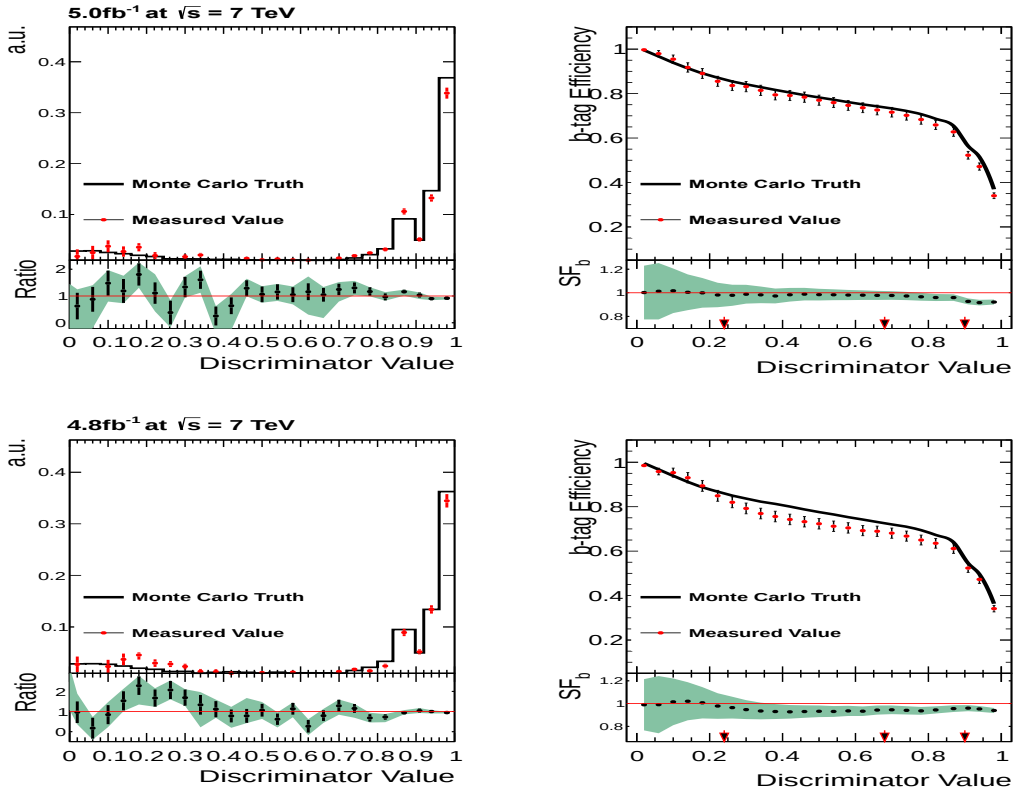


Figure 5.37: Measured b-jets CSV Discriminator distribution in the b-enriched sample compared to simulation (left) and the measured b-tagging efficiency, and data to simulation scale factors (right) for μ+jets events (top) and e+jets events (bottom). The arrows depict the loose, medium and tight working points respectively. The green band shows the combined statistical and systematic uncertainty.

Table 5.13: The efficiency provided by the method compared to the true b tagging efficiency for an integrated luminosity of $30.0 fb^{-1}$ simulated events.

Channel	WP	$\epsilon_b^{True}(\%)$	$\hat{\epsilon}_b$ (%)	rel. bias (%)
μ+jets	L	85.4±0.2	84.2±1.1	-1.4±-1.3
	M	73.3±0.3	73.4±1	0.1±-1.4
	T	56.4±0.3	56.6±0.9	0.4±-1.7
e+jets	L	85±0.3	85.4±1.3	0.5±-1.6
	M	72±0.3	72.5±1.2	0.7±-1.7
	T	54.5±0.4	55.4±1	1.6±-1.9

Table 5.14: Measured efficiencies and data over simulation scale factors for the CSV discriminator.

Channel	WP	$\epsilon_b^{True}(\%)$	$\hat{\epsilon}_b$ (%)	$\hat{\epsilon}_b/\epsilon_b^{True}$
μ+jets	L	85.4± 0.2	83.6± 2.3 (stat.) ± 1.8 (syst.)	0.979± 0.027 (stat.) ± 0.018 (syst.)
	M	73.3± 0.3	71.6± 2.2 (stat.) ± 1.9 (syst.)	0.977± 0.03 (stat.) ± 0.019 (syst.)
	T	56.4± 0.3	52.3± 1.7 (stat.) ± 1.6 (syst.)	0.927± 0.031 (stat.) ± 0.016 (syst.)
e+jets	L	85± 0.3	82± 2.4 (stat.) ± 1.5 (syst.)	0.965± 0.028 (stat.) ± 0.015 (syst.)
	M	72± 0.3	68.1± 2.3 (stat.) ± 1.5 (syst.)	0.946± 0.032 (stat.) ± 0.015 (syst.)
	T	54.5± 0.4	52.4± 1.9 (stat.) ± 1.5 (syst.)	0.961± 0.036 (stat.) ± 0.015 (syst.)

Table 5.15: Systematic uncertainty sources on the b-tagging efficiency for the CSV working points.

	μ+jets (%)						e+jets (%)					
WP →	Loose		Medium		Tight		Loose		Medium		Tight	
Systematic ↓	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF	$\delta\epsilon_b$	δSF
Jet Energy Scale	0.7	0.7	0.8	0.8	0.5	0.5	0.8	0.8	0.7	0.7	0.4	0.4
Jet Energy Resolution	1.2	1.2	1.0	1.0	0.9	0.9	0.5	0.5	0.3	0.3	0.4	0.4
PileUp	0.1	0.1	0.1	0.1	0.2	0.2	0.4	0.4	0.3	0.3	0.2	0.2
MET Unclustered Energy	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.4	0.3	0.3	0.4	0.4
Top quark mass	0.8	0.8	0.6	0.6	0.5	0.5	0.2	0.2	0.2	0.2	0.2	0.2
Right region definition	0.8	0.8	1.2	1.2	1.1	1.1	1.1	1.1	1.3	1.2	1.3	1.3
Total	1.8	1.8	1.9	1.9	1.6	1.6	1.5	1.5	1.5	1.5	1.5	1.5

5.10 Summary

In this chapter, a completely data-driven method was developed to measure the efficiency of tagging a jet originating from a b quark. This efficiency is vital for all analyses where b-tagging is applied especially when they are sensitive to the signal efficiency. Figure 5.38 and Figure 5.39 show the measured b-tagging efficiency compared to the simulation truth value for all available algorithms and working points as well as the obtained data to simulation scale factors to be applied in analysis. Next to the results for the Track Counting High Efficiency (TCHE) and Combined Secondary Vertex (CSV) algorithms, results for all other available algorithms are included. At 8 TeV a number of new algorithms was commissioned including a Combined Secondary Vertex Retrained (CSVR), a Soft Muon (SM) and Soft Electron (SE) tagger and a series of so-called super-combined taggers (CSV+JP, CSV+SL and CSV+JP+SL).

The scale factors are provided both for μ+jets and e+jets $t\bar{t}$ events. Since b-tagging relies on properties of single jets and not on event topologies, it is expected that the scale factors match between the two channels. To check this consistency, the ratio of the scale factors in μ+jets and e+jets events is displayed in Figure 5.40 for every available working point. The uncertainty on this ratio is calculated from the individual uncertainties assuming that they are uncorrelated. Overall the results are found to be consistent. The residual differences originate from the methodology itself, namely the difference in residual method bias among both channels.

In chapter 7, this measured b-tagging efficiency will be used to measure the $t\bar{t}$ production cross section with the requirement of one b-tagged jet in the event selection.

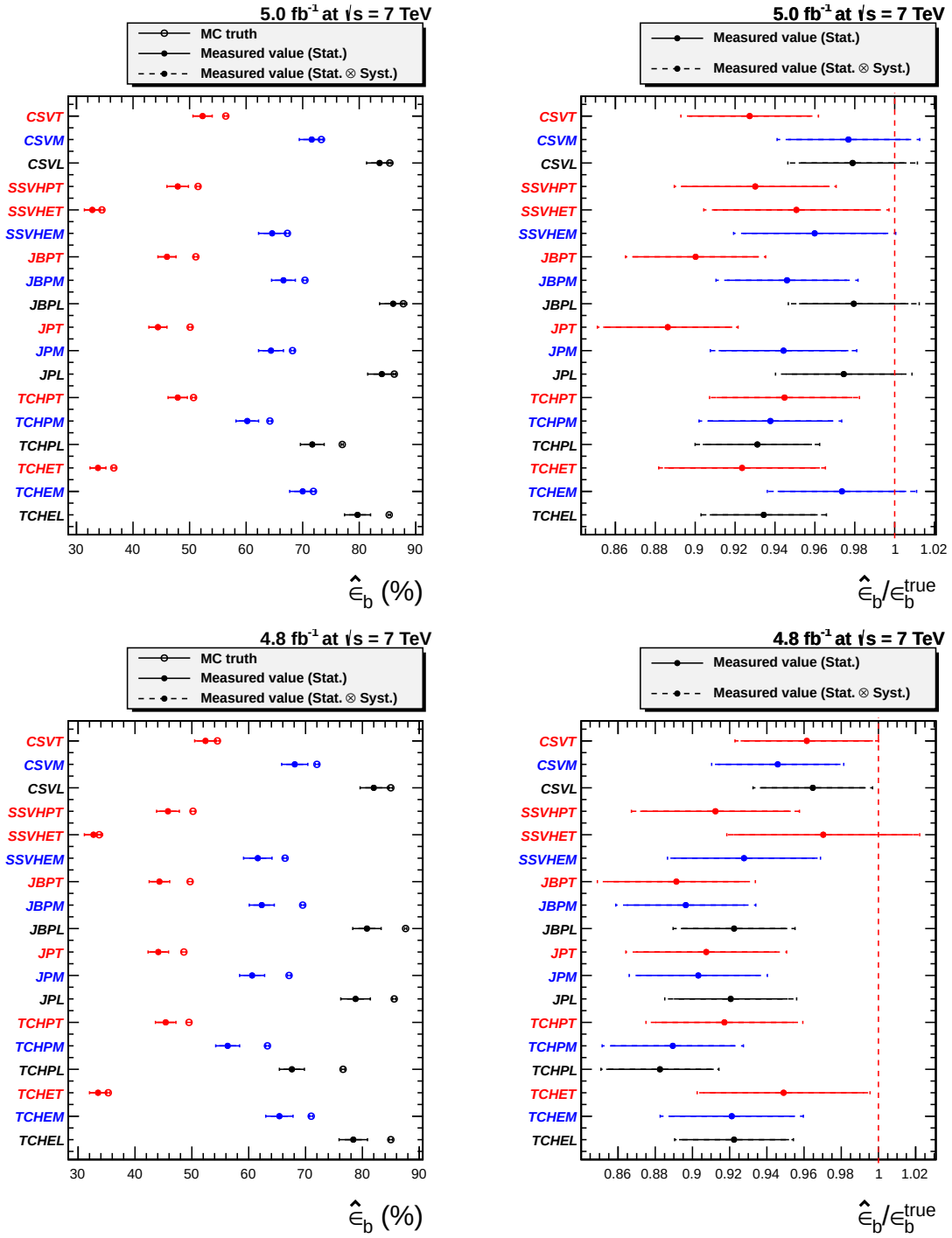


Figure 5.38: Visual representation of all working points for the studied b-tagging algorithms at 7 TeV. The measured b-tagging efficiencies are shown (left) as well as the data to simulation scale factors (right) in the μ+jets channel (top) and in the e+jets channel (bottom).

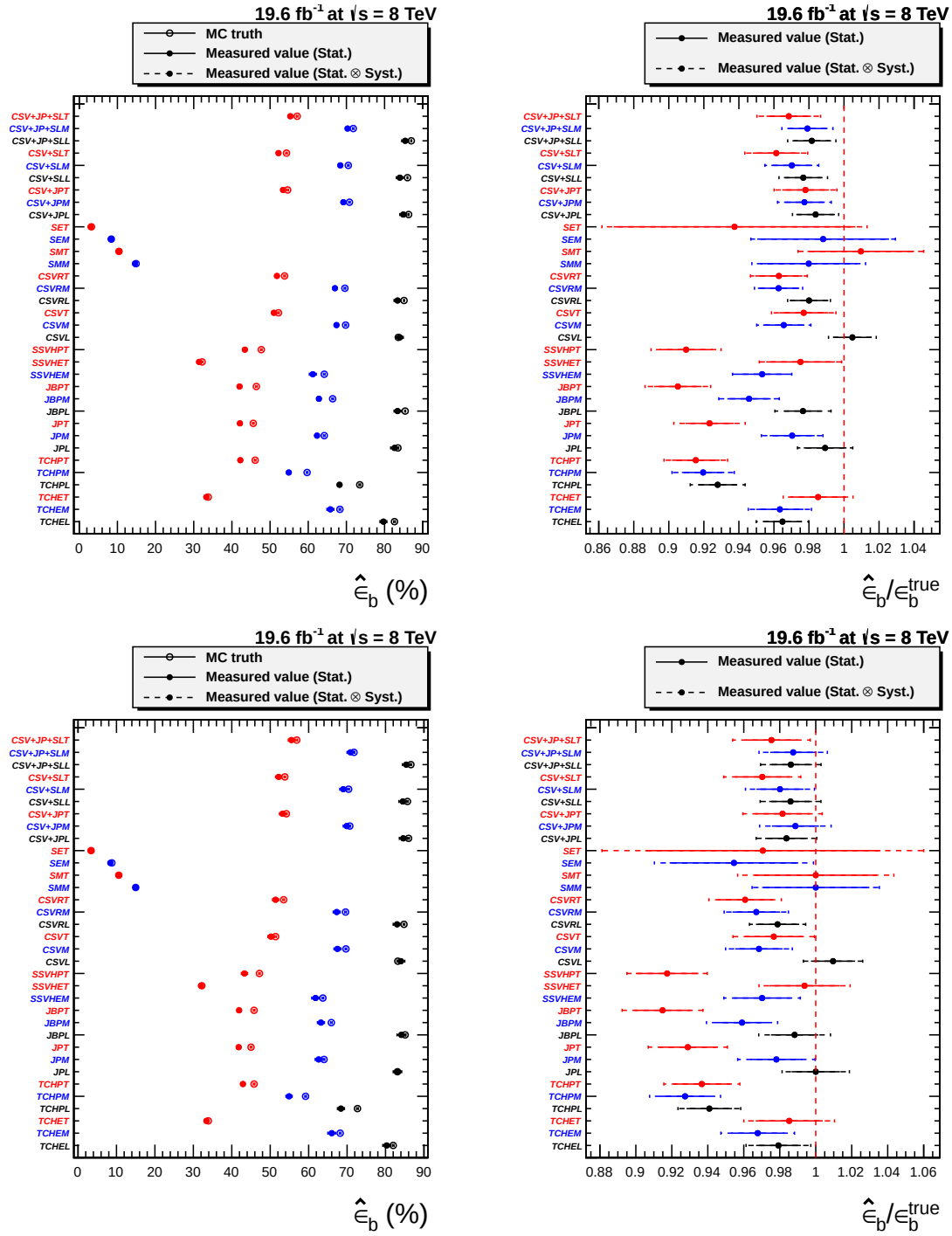


Figure 5.39: Visual representation of all working points for the studied b-tagging algorithms at 8 TeV. The measured b-tagging efficiencies are shown (left) as well as the data to simulation scale factors (right) in the μ+jets channel (top) and in the e+jets channel (bottom).

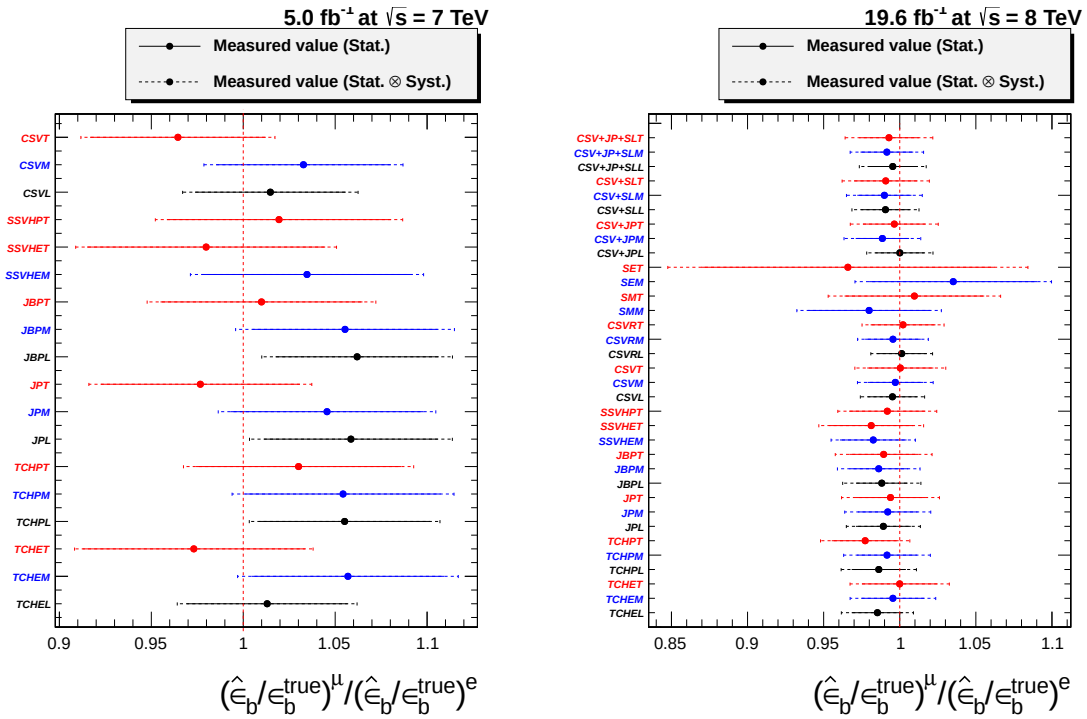


Figure 5.40: The ratio of the measured b-tagging efficiency scale factor measured in μ+jets and e+jets events is shown for all working points at 7 TeV (left) and 8 TeV (right)

Chapter 6

Measurement of the mis-tagging efficiency

The performance of a b-tagging algorithm is usually described in terms of the efficiency of tagging a jet originating from a b-quark (i.e. b-tagging efficiency) and the efficiency of tagging a jet not originating from a b-quark (i.e. mis-tagging efficiency). Usually the latter efficiency is provided for udsg- and c-jets separately.

In Chapter 5 a method was developed to reconstruct the b-tagging discriminator distribution for b jets in a data driven way. This method can be extended to measure the total mis-tagging efficiency. This extension is explained in Section 6.1 using the Track Counting High Efficiency algorithm. Section 6.3 proves that the measured b-tagging efficiency is uncorrelated to the measured mis-tagging efficiency. Since the mis-tagging rate is measured inclusively, meaning for udscg jets together, and depends strongly on the flavour composition in the event sample, Section 6.4 discusses the flavour composition in the signal sample. The statistical properties of this estimator are studied in Section 6.5 and the systematic uncertainties are discussed in Section 6.6. Finally the results are provided in Section 6.7 for all available algorithms at both 7 and 8 TeV collision energy.

6.1 Reconstructing the non-b jet b-tag discriminator distribution

The measurement of the total mis-tagging efficiency also relies on the reconstruction of the b-tagging discriminator distribution. This time it is the discriminator distribution for non-b jets that will be reconstructed. Using eq. (6.1) the discriminator distribution for non-b jets ($\hat{\Delta}_{\not b}$) is reconstructed from the distribution in the complete signal sample (Δ_{all}) by subtracting the b-jet contribution (Δ_b) using the same signal sample as was used in Chapter 5. The subtraction method is also depicted in Figure 6.1. The discriminator distribution in the signal sample is shown along with the distribution for b jets, obtained from simulation. Subtracting these two distributions provides the dotted distribution representing the non-b jets in the sample. From the dotted distribution the total mis-tagging efficiency can be determined at any working point for the

b-tagging algorithms.

$$\hat{\Delta}_{\not b} = \Delta_{all} - \Delta_b \tag{6.1}$$

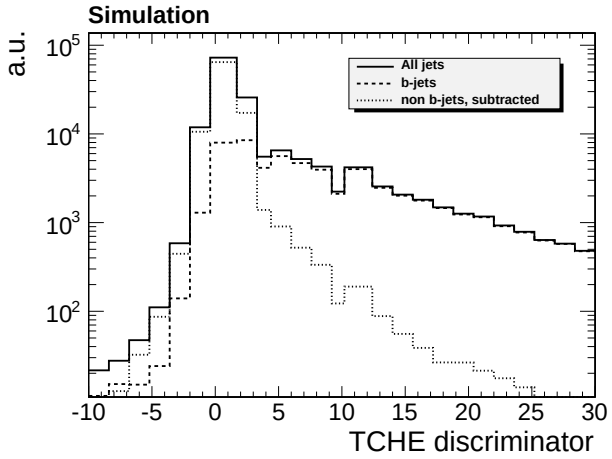


Figure 6.1: The TCHE discriminator distribution for all jets (solid line) and for b jets (dashed line) in the signal sample. Subtracting these two distributions yields the discriminator distribution for non-b jets (dotted line).

The only unknown in the subtraction formula, eq. (6.1), is the discriminator distribution for b jets in the signal sample. In the previous chapter the b-tagging discriminator distribution for b jets was reconstructed. However, this reconstruction was performed in the b-enriched sample rather than the entire signal sample. Evidently, it has to be shown that the discriminator distribution for b jets reconstructed within the b-enriched sample matches the discriminator distribution for b jets in the full signal sample.

By construction of the samples, the latter assumption does not completely hold. As Figure 6.2 clearly shows, the shape of the b-tagging discriminator distribution reconstructed in the b-enriched sample matches very well the true shape in the full signal sample. While indeed the b-tagging discriminator shape is roughly independent of the jet-muon mass, the number of b jets as well as other jets is obviously not. Hence, the normalisation of the reconstructed b-tagging discriminator distribution is off and has to be corrected. As no data-driven estimate is at hand, a simulation based constant scale factor of 1.6 is applied.

Finally, the subtraction formula eq. (6.1) can be rewritten taking into account the reconstructed b jets b-tagging discriminator distribution ($\hat{\Delta}_b^{b-enriched}$) and the corresponding scale factor (G) to correct for the different number of b jets between the b-enriched sample and the complete signal sample. Based on simulation, the resulting reconstructed b-tagging discriminator distribution for non-b jets is shown in Figure 6.3 and compared to the expected distribution. A good agreement is observed.

$$\hat{\Delta}_{\not b} = \Delta_{all} - G \times \hat{\Delta}_b^{b-enriched} \tag{6.2}$$

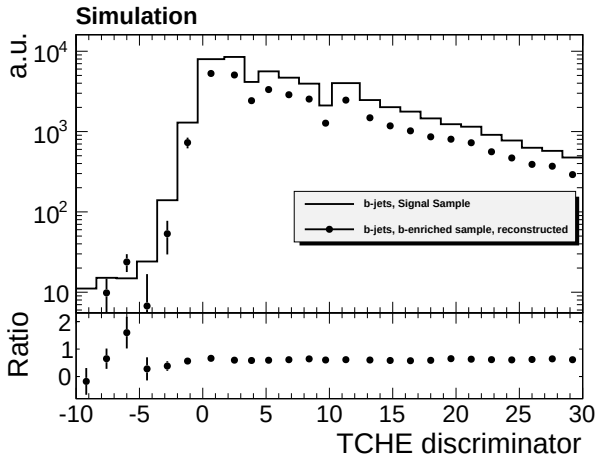


Figure 6.2: The TCHE discriminator distribution for true b jets in the signal sample compared to the distribution for b jets reconstructed with the method outlined in the previous chapter. The attached ratio plot shows that the shapes agree well but a constant offset in the normalisation is observed.

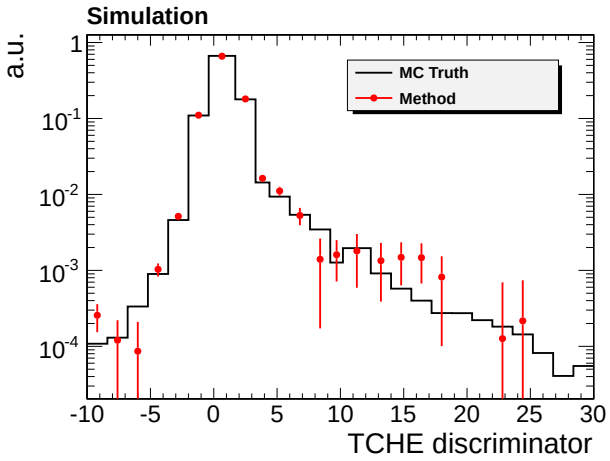


Figure 6.3: The TCHE discriminator distribution for non-b jets in the signal sample. The truth distribution is compared to the reconstructed distribution. The uncertainties are determined on a simulated sample corresponding to an integrated luminosity of $30 fb^{-1}$.

6.2 Estimation of the total mis-tagging efficiency

The reconstructed b-tagging discriminator distribution for non-b jets, Figure 6.3, can be used to measure the total mis-tagging efficiency for any of the working points of the Track Counting High Efficiency b-tagging algorithm. The measured efficiency is

compared to the truth efficiency in Figure 6.4. The same figure also shows the relative bias between the two efficiency values as a function of the working point.

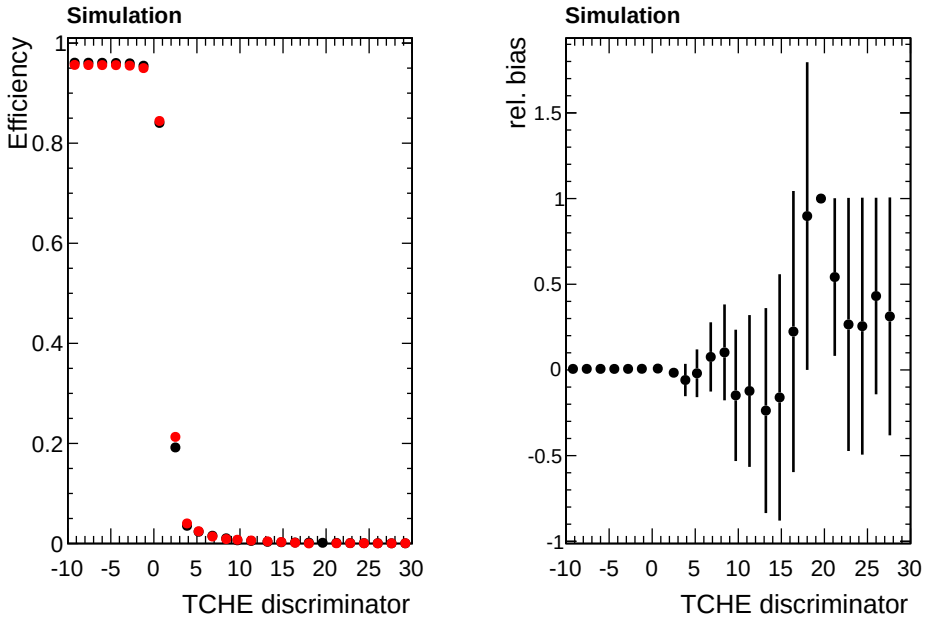


Figure 6.4: The TCHE b-tagging efficiency compared between true non-b jets in the b candidate sample (black markers) and the measured efficiency (red markers) using the reconstructed distribution (left) and the relative bias between the two efficiencies (right). The uncertainties are determined on a sample corresponding to an integrated luminosity of $30 fb^{-1}$.

The relative bias on the estimator can be defined as:

$$\frac{\hat{\epsilon}_{\not b} - \epsilon_{\not b}^{true}}{\epsilon_{\not b}^{true}}. \tag{6.3}$$

A bias has been observed and is of order (1.4±2.0)%, (5±10.9)% and (20.0±40.0)% for the loose, medium and tight working points respectively. These values are displayed in Table 6.1. The relative bias is statistically compatible with zero over most of the discriminator range but seems to increase with tighter working points. The reason for the increase of the bias is immediately visible from Figure 6.3. This distribution is reconstructed from the signal sample by removing the b-jet contribution depleting the high discriminator values. For that reason the efficiency estimator is more susceptible for statistical fluctuations at tighter working points compared to looser ones.

Table 6.1: Results for the Track Counting High Efficiency tagger Loose, Medium and Tight working points for an integrated luminosity of $30 fb^{-1}$ simulated events. The efficiency provided by the method is compared to the true total mis-tagging efficiency.

WP	$\epsilon_{\not b}^{True}(\%)$	$\hat{\epsilon}_{\not b}$ (%)	rel. bias (%)
L	21±0.1	21.3±0.4	1.4±2
M	3.8±0.1	4±0.4	5.3±10.9
T	0.5±0	0.6±0.2	20±40

6.3 Correlation between $\hat{\epsilon}_{\not b}$ and $\hat{\epsilon}_b$

To estimate the mis-tagging efficiency from data, the b-jet component is being removed from the discriminator distribution in the signal sample. Since the shape for the b jets component is taken from the measurement of the b-tagging efficiency in the previous chapter, these two measurements are potentially correlated. If they are, the mis-tagging efficiency cannot be interpreted as a self-contained efficiency and cannot be used in further physics measurements. To determine the degree of correlation between the two estimators, $\hat{\epsilon}_{\not b}$ and $\hat{\epsilon}_b$, 750 pseudo-experiments are constructed from the total simulation sample each reflecting a luminosity of $19.7 fb^{-1}$. In each pseudo-experiment both the b-tagging and mis-tagging efficiencies are estimated and placed in a two dimensional distribution. This distribution is shown in Figure 6.5 and it can be concluded that no correlation exists between the estimators $\hat{\epsilon}_{\not b}$ and $\hat{\epsilon}_b$.

6.4 Parton flavour composition of the signal sample

The mis-tagging efficiency strongly depends on the jet flavour. Consequently the mis-tagging efficiency is usually separately determined for charm and other light quarks using multijet events. Moreover, the measurement in this chapter cannot discriminate between the various non b-jet classes so only an inclusive estimation of the mis-tagging efficiency is possible. Therefore, the measurement can only be interpreted alongside the exact flavour composition of the sample it was determined from. Also the measurement can only be applied to event samples of the same composition.

The sample used for this measurement is created using a $t\bar{t}$ oriented event selection introduced in Chapter 4. From this sample the signal sample was created using a χ^2-matching technique assuring a good b-jet purity. As shown in Table 6.2, the b-jet fraction is indeed the dominant part of the sample making up 36%. These jets, however, play no role in the mis-tagging efficiency. The mis-tagging efficiency is determined from a jet sample consisting of 10.8% u, 12.8% d, 4.5% s and 6.3% c-jets. Also a significant fraction of the sample, 29.2%, are jets originating from gluons. Finally, 0.4% of the sample consists of unmatched jets, jets where no match to any parton from the top quark pair decay is found.

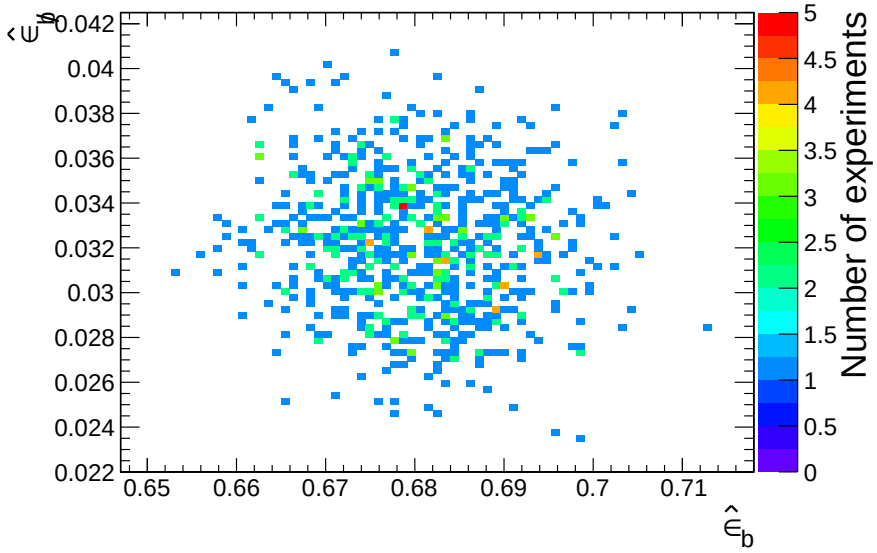


Figure 6.5: This distribution shows the correlation between $\hat{\epsilon}_{\not b}$ and $\hat{\epsilon}_b$ for 750 randomly selected simulation samples of $19.7 fb^{-1}$. The efficiencies are determined for the Track Counting High Efficiency Medium Working Point. The correlation is well below 10%.

Table 6.2: Parton flavour composition in the signal sample. The algorithmic definition is used to assign a parton flavour to each jet. The unmatched jets are those where the algorithm has not found any matching generated quark to determine its flavour.

Parton	Number of jets	Fraction
d	21648.8	10.8%
u	25489.9	12.8%
s	9049.04	4.5%
c	12525.9	6.3%
b	71767.4	36.0%
gluon	58191.1	29.2%
Unmatched	875.318	0.4%

6.5 Statistical properties of $\hat{\epsilon}_{\not b}$

In Figure 6.6 the pull distribution is shown for the mis-tagging efficiency estimator at the Medium Working Point of the TCHE. The pull distribution is determined using 750 pseudo-experiments corresponding to an integrated luminosity of $19.7 fb^{-1}$ each.

The pull distribution is fitted with a guass function. The fit returns a mean of 0.021 ± 0.024 and a width of 0.644 ± 0.017. The width of the pull deviates significantly from unity indicating that the statistical uncertainty on the mis-tagging efficiency might be overestimated. However, it needs to be taken into account that there is a certain

degree of correlation between the different experiments due to the limited simulated statistics at hand.

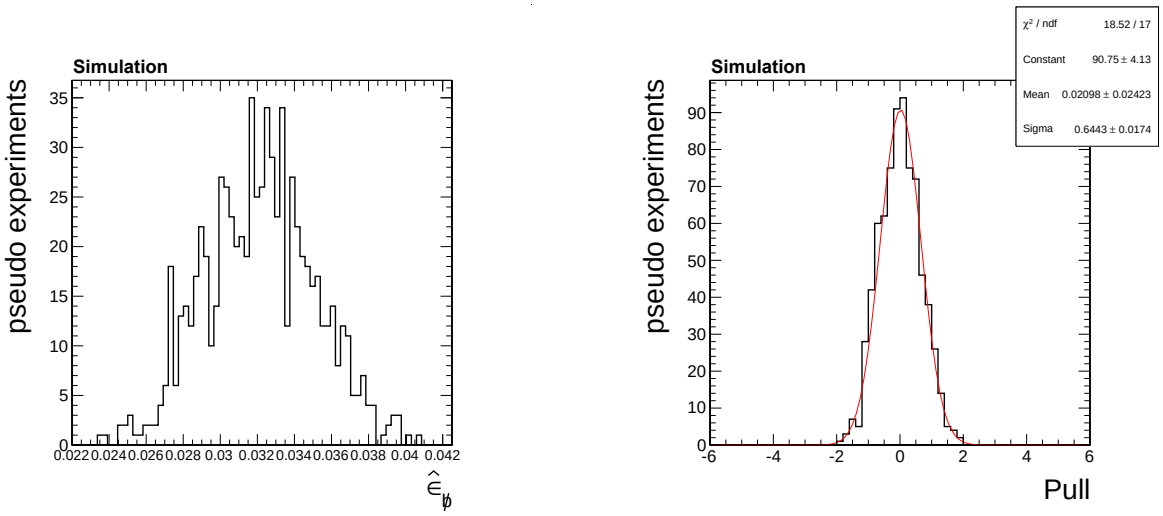


Figure 6.6: Distribution of the mis-tagging efficiency estimator (left) and the pull distribution (right) for the TCHE Medium WP

6.6 Systematic uncertainties

In this section, the systematic uncertainty is broken down into its different contributions for the mis-tagging efficiency and the data to simulation scale factors. A detailed explanation for the different contributions is provided in the following. For the mis-tagging efficiency the variation of the method bias, as defined in eq. (6.3), between the systematic variation and the nominal simulation sample is taken since the bias itself is not independent of the systematic variations. For the data to simulation scale factor, the absolute difference between the nominal scale factor and the scale factor using the varied sample is used. For each contribution we consider an up- and downwards shift and the biggest effect among the two is quoted as systematic uncertainty. The same systematic uncertainties are evaluated as for the b-tagging efficiency measurement in Chapter 5. All contributions are added in quadrature to obtain the total systematic uncertainty on the measurement. The actual systematic uncertainties for each algorithm are provided in the results section 6.7.
As opposed to the measurement of the b-tagging efficiency which is fully data-driven, this method uses simulation input for the normalisation of the measured b-tagging discriminator distribution for b jets. This means that theory uncertainties have to be calculated on the mis-tagging efficiency estimator and also on the data to simulation scale factors. As will be visible in the systematics tables in the results section (Section 6.7), the theory uncertainties are quite large.
The source of the large theory uncertainties lies in the fact that these simulated samples contain significantly fewer statistics compared to the nominal simulation. Since the estimation of the discriminator shape for b jets is sensitive to sample statistics, so

will the estimation of the mis-tagging efficiency. As a consequence, large systematic uncertainties due to these theoretical variations are quoted. These are outlined below.

Background composition

Variations in the normalisation of the simulated backgrounds used in this analysis can have an impact on the final result. To estimate this effect, a systematic uncertainty is determined corresponding to the variation of the main V+jets background sources. The normalisation of these background processes, mainly the production of a W/Z-boson with additional jet activity, is scaled by $\pm 30\%$ to mimic the $\pm 1\sigma$ uncertainty on their respective cross sections. The effects for W+jets and Z+jets are added in quadrature.

Factorisation scale

Variations in the Q^2-scale are studied by comparing two distinct samples where the Q^2-scale is varied by respectively a factor 0.5 and 2.

ME-PS Matching Threshold

To estimate the effect of varying the threshold used for the matching between the matrix-element level and parton showers in the event simulation, dedicated simulation samples were used where this threshold was varied by a factor of 0.5 and 2.0 with respect to its nominal value.

Top Quark Mass

As the top quark mass from simulation is used to construct the χ^2 for each jet combination a systematic uncertainty on possible variation of this mass is evaluated. To obtain this systematic uncertainty, simulated top quark pair decays are used where the mass is shifted $\pm 9\text{GeV}/c^2$. Since the top quark mass is known with a precision of $\approx 1\text{GeV}c^2$ in the top quark mass is required, the effect is divided by a factor 9.

Parton Distribution Function (PDF) uncertainties

The simulation samples used in this analysis are generated using the CTEQ6.6 PDF set [75]. To evaluate the uncertainty on the measurement due to the uncertainties on the 22 parameters of this PDF set, LHAPDF is used to obtain a set of 44 ErrorPDFs where each parameter is varied up and down. For each ErrorPDF an event weight is obtained in addition to the overall event weight. With this additional weight the analysis is performed and the result is compared to the result using the nominal PDF. The systematic uncertainty due to PDF uncertainties is derived up and down using the following "master equations".

$$\Delta X^{+}_{max} = \sqrt{\sum_{i=1}^{N} [max(X_i^{+} - X_0, X_i^{-} - X_0, 0)]^2}, \tag{6.4}$$

$$\Delta X^{-}_{max} = \sqrt{\sum_{i=1}^{N} [max(X_0 - X_i^{+}, X_0 - X_i^{-}, 0)]^2}, \tag{6.5}$$

where X_0 is the result using the nominal PDF and $X_i^{\pm}$ the result with the $\pm 1\sigma$ variation of the i^{th} parameter in the PDF set.

6.7 Results for the different b-tagging algorithms

This section provides the measured b-tagging discriminator distribution for non-b jets and mis-tagging efficiencies for the Track Counting High Efficiency and Combined Secondary Vertex b-tagging algorithms at 8 TeV. Also the data to simulation scale factors and the systematics tables are provided.

6.7.1 Track Counting High Efficiency (8 TeV)

In this section the mis-tagging efficiency results for the Track Counting High Efficiency (TCHE) algorithm are provided. The reconstructed b-tagging discriminator distribution for non-b jets is shown for both μ+jets and e+jets events in Figure 6.7 along with the resulting mis-tagging efficiency and data to simulation scale factor ($SF_b = \hat{\epsilon}_{\not b} / \epsilon_{\not b}^{True}$) as a function of the Working Point (WP), or threshold. The "loose", "medium" and "tight" Working Points are indicated by the arrows in the efficiency plot. The discriminator distribution is displayed only up to a value of 15 since beyond that point the efficiency drops to 0%.

For these specific points, the residual relative bias on the efficiency is provided in Table 6.3. The relative bias enlarges for tighter working points because the bins for higher discriminator values start to become depleted. Combined with the limited size of the simulated statistics this can cause statistical fluctuations in the subtraction formula used to reconstruct the non-b jets discriminator distribution.

The efficiencies measured from data are provided in Table 6.4 along with their statistical and systematic uncertainties. In the same Table, the data to simulation scale factors are provided as well.

The breakdown of the systematic uncertainty in all its different sources is provided in Table 6.5. The total systematic uncertainty is dominated by the large theory uncertainties, namely factorisation scale and the ME-PS matching threshold. The size of these uncertainties is dominated by the limited statistics in the event samples available with these theory parameter variations.

6.7.2 Combined Secondary Vertex (8 TeV)

In the following, the results for the Combined Secondary Vertex (CSV) algorithm will be provided. Figure 6.8 shows the reconstructed b-tagging discriminator distribution for non-b jets in both $t\bar{t}$ decay channels along with the resulting mis-tagging efficiency and data to simulation scale factors. The reconstructed b jet b-tagging distribution peaks at zero and falls rapidly. This shows that the distribution is indeed dominated by non-b jets.

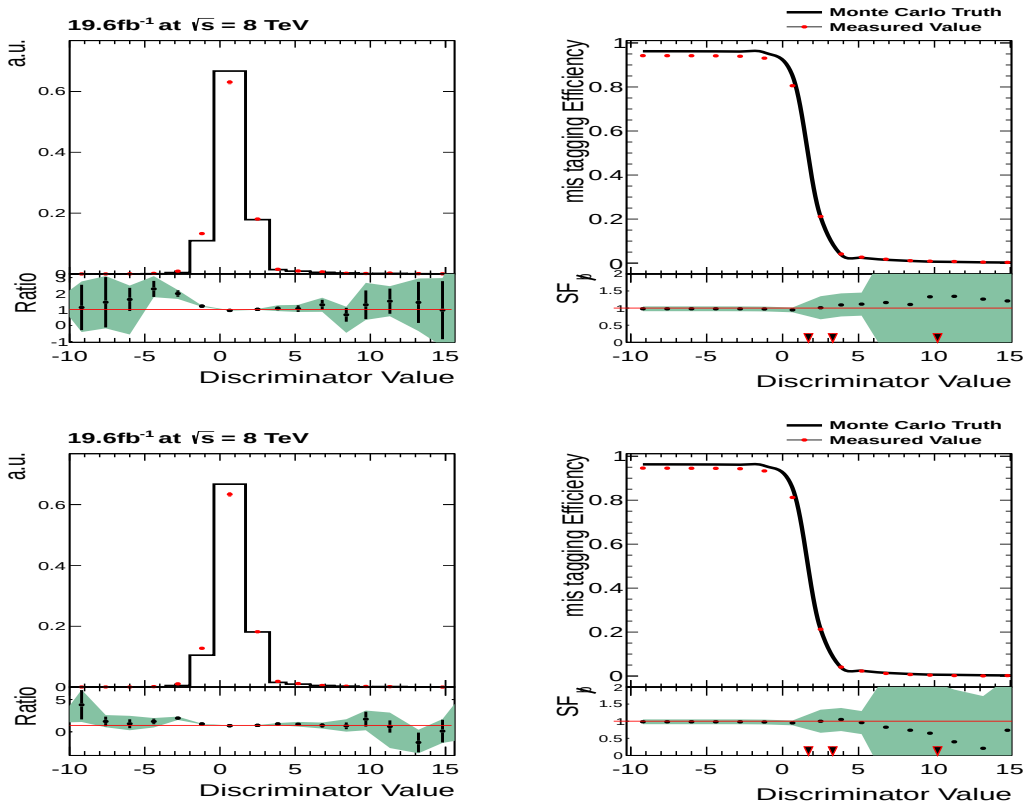


Figure 6.7: Measured non-b jets TCHE Discriminator distribution in the b-jet candidate sample compared to simulation (left) and the measured mis-tagging efficiency, and data to simulation scale factors (right) for μ+jets events (top) and e+jets events (bottom). The arrows depict the loose, medium and tight working points respectively. The green band shows the combined statistical and systematic uncertainty.

Table 6.3: The efficiency provided by the method compared to the true mis-tagging efficiency for an integrated luminosity of $30 fb^{-1}$ simulated events.

Channel	WP	$\epsilon_{\not b}^{True}(\%)$	$\hat{\epsilon}_{\not b}$ (%)	rel. bias (%)
μ+jets	L	21±0.1	21.3±0.4	1.4±2
	M	3.8±0.1	4±0.4	5.3±10.9
	T	0.5±0	0.6±0.2	20±40
e+jets	L	21.3±0.2	21.3±0.4	0±2.1
	M	3.9±0.1	4.1±0.4	5.1±10.6
	T	0.5±0	0.3±0.3	-40±60

The numerical values for the mis-tagging efficiencies as well as the data to simulation scale factors for the three Working Points are provided in Table 6.7. The relative bias on these results is provided in Table 6.6. As was the case for the TCHE result, the

Table 6.4: Measured efficiencies and data over simulation scale factors for the TCHE discriminator.

Channel	WP	$\epsilon_{\not b}^{True}(\%)$	$\hat{\epsilon}_{\not b}$ (%)	$\hat{\epsilon}_{\not b}/\epsilon_{\not b}^{True}$
μ+jets	L	21± 0.1	21.2± 0.5 (stat.) ± 7 (syst.)	1.01± 0.024 (stat.) ± 0.07 (syst.)
	M	3.8± 0.1	4.1± 0.4 (stat.) ± 33.7 (syst.)	1.079± 0.109 (stat.) ± 0.337 (syst.)
	T	0.5± 0	0.7± 0.3 (stat.) ± 152.6 (syst.)	1.4± 0.6 (stat.) ± 1.526 (syst.)
e+jets	L	21.3± 0.2	21.4± 0.5 (stat.) ± 3.6 (syst.)	1.005± 0.025 (stat.) ± 0.036 (syst.)
	M	3.9± 0.1	4.2± 0.5 (stat.) ± 26.9 (syst.)	1.077± 0.131 (stat.) ± 0.269 (syst.)
	T	0.5± 0	0.3± 0.3 (stat.) ± 105.1 (syst.)	0.6± 0.6 (stat.) ± 1.051 (syst.)

Table 6.5: Systematic uncertainty sources on the inclusive mis tagging efficiency for the TCHE working points.

	μ+jets (%)						e+jets (%)					
WP →	Loose		Medium		Tight		Loose		Medium		Tight	
Systematic ↓	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF
Background Composition	1.1	1.1	3.9	3.9	0.0	0.0	0.9	0.9	5.9	5.9	20.0	20.0
Factorisation Scale	6.2	6.2	28.9	28.9	120.0	120.0	1.8	1.8	20.5	20.5	90.0	90.0
ME-PS Matching Threshold	1.4	1.4	10.7	10.7	80.0	80.0	1.9	1.9	5.1	5.1	40.0	40.0
PDF Uncertainties	0.8	0.8	2.6	2.6	0.0	0.0	1.6	1.6	8.5	8.5	0.0	0.0
Jet Energy Scale	0.5	0.5	0.0	0.0	0.0	0.0	0.5	0.5	2.6	2.6	0.0	0.0
Jet Energy Resolution	1.0	1.0	5.3	5.3	20.0	20.0	0.9	0.9	7.7	7.7	20.0	20.0
PileUp	0.5	0.5	2.6	2.6	0.0	0.0	0.5	0.5	2.6	2.6	0.0	0.0
MET Unclustered Energy	1.0	1.0	2.6	2.6	20.0	20.0	0.5	0.5	2.6	2.6	0.0	0.0
Top quark mass	0.6	0.6	2.7	2.7	8.9	8.9	0.9	0.9	5.6	5.6	11.9	11.9
Right region definition	1.9	1.9	10.5	10.5	40.0	40.0	0.5	0.5	7.7	7.7	20.0	20.0
Total	7	7	33.7	33.7	152.6	152.6	3.6	3.6	26.9	26.9	105.1	105.1

relative bias enlarges with tighter discriminator thresholds as the method becomes more sensitive to fluctuations in this depleted region.

Finally, the breakdown of the systematic uncertainty in all its different sources is provided in Table 6.8. The total systematic uncertainty is dominated by the large theory uncertainties, namely factorisation scale and the ME-PS matching threshold mainly because of small statistics in the respective samples.

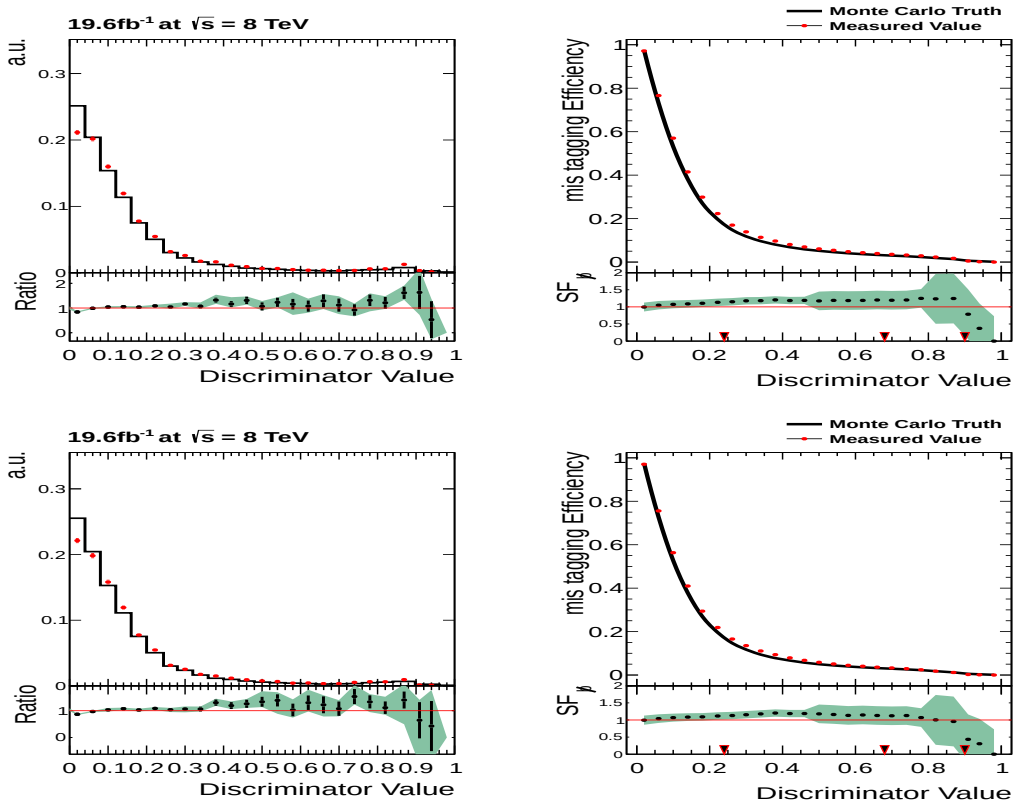


Figure 6.8: Measured non-b jets CSV Discriminator distribution in the b-jet candidate sample compared to simulation (left) and the measured mis-tagging efficiency, and data to simulation scale factors (right) for μ+jets events (top) and e+jets events (bottom). The arrows depict the loose, medium and tight working points respectively. The green band shows the combined statistical and systematic uncertainty.

Table 6.6: The efficiency provided by the method compared to the true mis-tagging efficiency for an integrated luminosity of $30 fb^{-1}$ simulated events.

Channel	WP	$\epsilon_{\not b}^{True}$(%)	$\hat{\epsilon}_{\not b}$ (%)	rel. bias (%)
μ+jets	L	14.8±0.1	15.1±0.3	2±2.1
	M	3±0.1	3.2±0.3	6.7±10.6
	T	0.6±0	0.5±0.2	-16.7±33.3
e+jets	L	14.7±0.1	14.6±0.4	-0.7±2.8
	M	2.9±0.1	3.3±0.3	13.8±11.1
	T	0.5±0	0.5±0.2	0±40

Table 6.7: Measured efficiencies and data over simulation scale factors for the CSV discriminator.

Channel	WP	$\epsilon_{\not b}^{True}(\%)$	$\hat{\epsilon}_{\not b}$ (%)	$\hat{\epsilon}_{\not b}/\epsilon_{\not b}^{True}$
μ+jets	L	14.8± 0.1	17± 0.4 (stat.) ± 10.6 (syst.)	1.149± 0.028 (stat.) ± 0.106 (syst.)
	M	3± 0.1	3.6± 0.3 (stat.) ± 27.2 (syst.)	1.2± 0.108 (stat.) ± 0.272 (syst.)
	T	0.6± 0	0.5± 0.2 (stat.) ± 72.8 (syst.)	0.833± 0.333 (stat.) ± 0.728 (syst.)
e+jets	L	14.7± 0.1	16.7± 0.5 (stat.) ± 4.7 (syst.)	1.136± 0.035 (stat.) ± 0.047 (syst.)
	M	2.9± 0.1	3.3± 0.4 (stat.) ± 23.7 (syst.)	1.138± 0.143 (stat.) ± 0.237 (syst.)
	T	0.5± 0	0.3± 0.3 (stat.) ± 68.7 (syst.)	0.6± 0.6 (stat.) ± 0.687 (syst.)

Table 6.8: Systematic uncertainty sources on the inclusive mis tagging efficiency for the CSV working points.

	μ+jets (%)						e+jets (%)					
WP →	Loose		Medium		Tight		Loose		Medium		Tight	
Systematic ↓	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF
Background Composition	0.6	0.6	0.2	0.2	0.0	0.0	0.7	0.7	3.8	3.8	36.7	36.7
Factorisation Scale	8.8	8.8	23.9	23.9	66.7	66.7	2.7	2.7	20.9	20.9	16.7	16.7
ME-PS Matching Threshold	5.5	5.5	10.0	10.0	16.7	16.7	2.7	2.7	7.1	7.1	36.7	36.7
PDF Uncertainties	0.0	0.0	0.0	0.0	0.0	0.0	0.7	0.7	3.5	3.4	16.7	16.7
Jet Energy Scale	1.4	1.4	3.3	3.3	0.0	0.0	1.4	1.4	3.5	3.4	16.7	16.7
Jet Energy Resolution	0.7	0.7	3.3	3.3	16.7	16.7	1.4	1.4	3.5	3.4	16.7	16.7
PileUp	0.7	0.7	0.0	0.0	0.0	0.0	0.7	0.7	0.0	0.0	16.7	16.7
MET Unclustered Energy	0.7	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	16.7	16.7
Top quark mass	0.2	0.2	1.2	1.2	4.1	4.1	1.3	1.3	3.3	3.3	9.3	9.3
Right region definition	1.4	1.4	6.7	6.7	16.7	16.7	0.7	0.7	3.5	3.4	16.7	16.7
Total	10.6	10.6	27.2	27.2	72.8	72.8	4.7	4.7	23.7	23.7	68.7	68.7

6.7.3 Track Counting High Efficiency (7 TeV)

In this section the mis tagging efficiency results for the Track Counting High Efficiency (TCHE) algorithm are provided. The reconstructed b-tagging discriminator distribution for non-b jets is shown for both μ+jets and e+jets events in Figure 6.9 along with the resulting mis tagging efficiency and data to simulation scale factors.

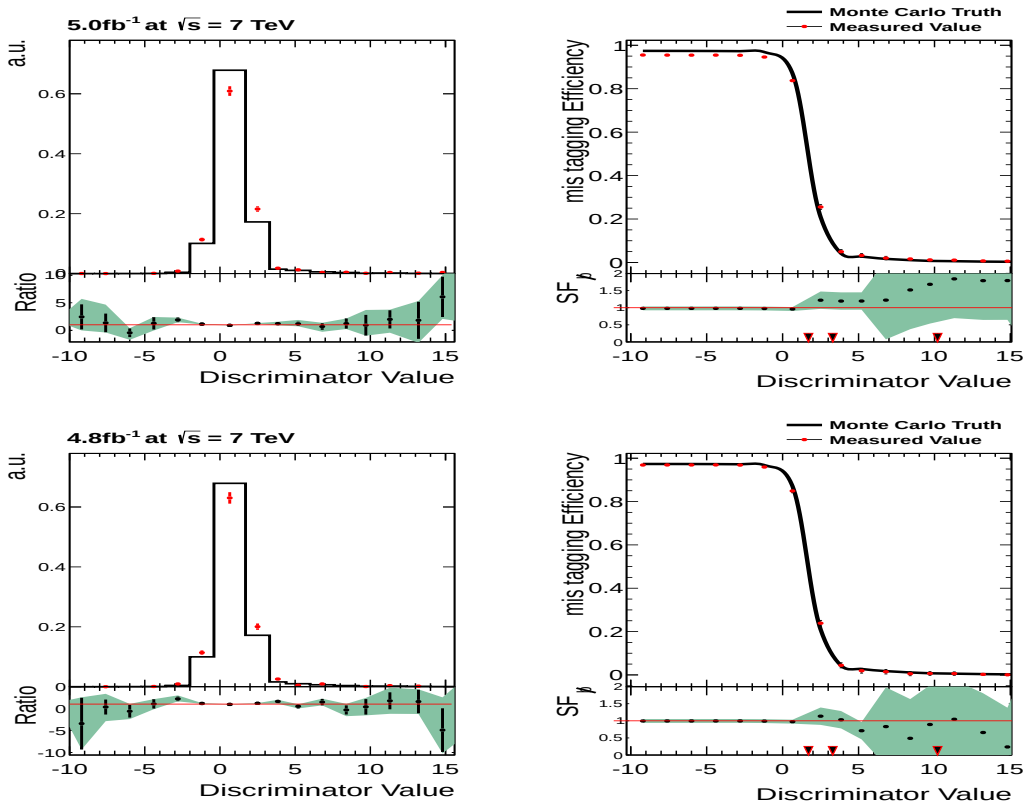


Figure 6.9: Measured non-b jets TCHE Discriminator distribution in the b-jet candidate sample compared to simulation (left) and the measured mis tagging efficiency, and data to simulation scale factors (right) for μ+jets events (top) and e+jets events (bottom). The arrows depict the loose, medium and tight working points respectively. The green band shows the combined statistical and systematic uncertainty.

The residual relative bias on the efficiency for the three Working Points is provided in Table 6.9. Similarly to the 8 TeV result, the relative bias enlarges for tighter working points due to limited statistics.

The efficiencies measured from data are provided in Table 6.10 along with their statistical and systematic uncertainties. In the same Table, the data to simulation scale factors are provided as well.

The breakdown of the systematic uncertainty in all it's different sources is provided in Table 6.11. Again, the total systematic uncertainty is dominated by the large theory uncertainties.

Table 6.9: The efficiency provided by the method compared to the true mis tagging efficiency for an integrated luminosity of $20 fb^{-1}$ simulated events.

Channel	WP	$\epsilon_{b\not{b}}^{True}(\%)$	$\hat{\epsilon}_{b\not{b}}\,(\%)$	rel. bias (%)
μ+jets	L	20.9±0.2	21±0.5	0.5±2.6
	M	4.1±0.1	4.1±0.5	0±12.4
	T	0.6±0	0.5±0.3	-16.7±50
e+jets	L	21±0.2	21.2±0.6	1±3
	M	4.3±0.1	4.6±0.6	7±14.2
	T	0.6±0	0.4±0.4	-33.3±66.7

Table 6.10: Measured efficiencies and data over simulation scale factors for the TCHE discriminator.

Channel	WP	$\epsilon_{b\not{b}}^{True}(\%)$	$\hat{\epsilon}_{b\not{b}}\,(\%)$	$\hat{\epsilon}_{b\not{b}}/\epsilon_{b\not{b}}^{True}$
μ+jets	L	20.9± 0.2	25.5± 1.1 (stat.) ± 5.6 (syst.)	1.22± 0.054 (stat.) ± 0.056 (syst.)
	M	4.1± 0.1	5± 1 (stat.) ± 25.4 (syst.)	1.22± 0.246 (stat.) ± 0.254 (syst.)
	T	0.6± 0	1.1± 0.6 (stat.) ± 114.7 (syst.)	1.833± 1 (stat.) ± 1.147 (syst.)
e+jets	L	21± 0.2	23.8± 1.3 (stat.) ± 6.5 (syst.)	1.133± 0.063 (stat.) ± 0.065 (syst.)
	M	4.3± 0.1	4.4± 1.2 (stat.) ± 41 (syst.)	1.023± 0.28 (stat.) ± 0.41 (syst.)
	T	0.6± 0	0.6± 0.8 (stat.) ± 122.5 (syst.)	1± 1.333 (stat.) ± 1.225 (syst.)

6.7.4 Combined Secondary Vertex (7 TeV)

Finally, this section summarised the results for the Combined Secondary Vertex (CSV) algorithm. Figure 6.10 shows the reconstructed b-tagging discriminator distribution for non-b jets in both $t\bar{t}$ decay channels along with the resulting mis tagging efficiency and data to simulation scale factors.

Table 6.13 contains the measured mis tagging efficiencies and scale factors for the three Working Points are provided. The relative bias on these results is provided in Table 6.12. As was the case for the other algorithms both at 7 and 8 TeV, the relative bias enlarges with tighter discriminator thresholds as the method becomes more sensitive to fluctuations in this depleted region.

Finally, the breakdown of the systematic uncertainty in all it's different sources is provided in Table 6.14. The total systematic uncertainty is dominated by the large theory uncertainties, namely factorisation scale and the ME-PS matching threshold mainly because of small statistics in the respective samples.

Table 6.11: Systematic uncertainty sources on the inclusive mis tagging efficiency for the TCHE working points.

	μ+jets (%)						e+jets (%)					
WP $\rightarrow$	Loose		Medium		Tight		Loose		Medium		Tight	
Systematic $\downarrow$	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF
Background Composition	0.7	0.7	3.4	3.4	23.6	23.6	1.3	1.3	3.4	3.4	16.7	16.7
Factorisation Scale	1.9	1.9	11.9	11.9	83.3	83.3	5.1	5.1	24.1	24.1	66.7	66.7
ME-PS Matching Threshold	4.2	4.2	14.6	14.6	43.3	43.3	1.0	1.0	28.4	28.4	83.3	83.3
PDF Uncertainties	1.5	1.5	7.3	7.3	33.3	33.3	1.4	1.3	4.6	4.7	0.0	0.0
Jet Energy Scale	1.4	1.4	4.9	4.9	16.7	16.7	2.4	2.4	9.7	9.7	33.3	33.3
Jet Energy Resolution	1.9	1.9	11.9	11.9	33.3	33.3	1.4	1.4	9.3	9.3	33.3	33.3
PileUp	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	4.7	4.7	16.7	16.7
MET Unclustered Energy	0.0	0.0	0.0	0.0	0.0	0.0	1.0	1.0	2.3	2.3	16.7	16.7
Top quark mass	0.5	0.5	2.8	2.8	13.0	13.0	0.6	0.6	1.5	1.5	17.0	17.0
Right region definition	1.0	1.0	7.3	7.3	33.3	33.3	1.4	1.4	7.0	7.0	16.7	16.7
Total	5.6	5.6	25.4	25.4	114.7	114.7	6.5	6.5	41	41	122.5	122.5

Table 6.12: The efficiency provided by the method compared to the true mis tagging efficiency for an integrated luminosity of $20 fb^{-1}$ simulated events.

Channel	WP	$\epsilon_{\not b}^{True}$(%)	$\hat{\epsilon}_{\not b}$ (%)	rel. bias (%)
μ+jets	L	15.2±0.2	15.6±0.5	2.6±3.6
	M	3.2±0.1	2.9±0.4	-9.4±12.8
	T	0.7±0	0.2±0.3	-71.4±42.9
e+jets	L	15.2±0.2	14.8±0.6	-2.6±4.2
	M	3.2±0.1	2.8±0.5	-12.5±15.9
	T	0.7±0	0.1±0.3	-85.7±42.9

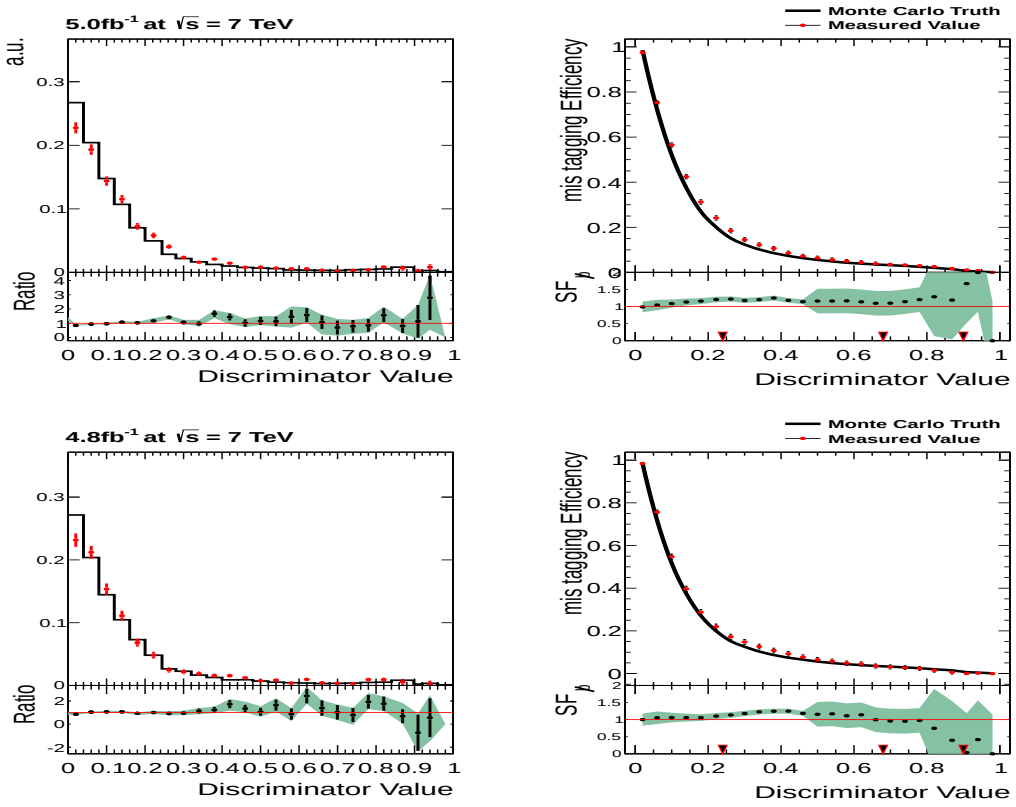


Figure 6.10: Measured non-b jets CSV Discriminator distribution in the b-jet candidate sample compared to simulation (left) and the measured b-tagging efficiency, and data to simulation scale factors (right) for μ+jets events (top) and e+jets events (bottom). The arrows depict the loose, medium and tight working points respectively. The green band shows the combined statistical and systematic uncertainty.

Table 6.13: Measured efficiencies and data over simulation scale factors for the CSV discriminator.

Channel	WP	$\epsilon_{b}^{True}(\%)$	$\hat{\epsilon}_{b}$ (%)	$\hat{\epsilon}_{b}/\epsilon_{b}^{True}$
μ+jets	L	15.2± 0.2	18.5± 0.9 (stat.) ± 6.7 (syst.)	1.217± 0.061 (stat.) ± 0.067 (syst.)
	M	3.2± 0.1	3.5± 0.7 (stat.) ± 36 (syst.)	1.094± 0.221 (stat.) ± 0.36 (syst.)
	T	0.7± 0	1.1± 0.5 (stat.) ± 114.9 (syst.)	1.571± 0.714 (stat.) ± 1.149 (syst.)
e+jets	L	15.2± 0.2	17.2± 1.2 (stat.) ± 4.6 (syst.)	1.132± 0.08 (stat.) ± 0.046 (syst.)
	M	3.2± 0.1	3.1± 1 (stat.) ± 30.5 (syst.)	0.969± 0.314 (stat.) ± 0.305 (syst.)
	T	0.7± 0	0± 0.7 (stat.) ± 121.7 (syst.)	0± 1 (stat.) ± 1.217 (syst.)

Table 6.14: Systematic uncertainty sources on the inclusive mis tagging efficiency for the CSV working points.

	μ+jets (%)						e+jets (%)					
WP $\rightarrow$	Loose		Medium		Tight		Loose		Medium		Tight	
Systematic $\downarrow$	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF	$\delta\epsilon_{\not b}$	δSF
Background Composition	1.9	1.9	2.8	2.7	25.8	25.8	0.9	0.9	4.4	4.4	14.3	14.3
Factorisation Scale	2.6	2.6	21.5	21.5	88.1	88.1	3.3	3.3	24.6	24.6	57.1	57.1
ME-PS Matching Threshold	5.1	5.1	27.6	27.6	54.8	54.8	0.7	0.7	12.5	12.5	102.4	102.4
PDF Uncertainties	1.5	1.5	5.4	5.4	28.6	28.6	0.9	0.9	6.0	6.0	14.3	14.3
Jet Energy Scale	1.4	1.4	3.1	3.1	14.3	14.3	1.3	1.3	6.3	6.3	14.3	14.3
Jet Energy Resolution	2.0	2.0	3.1	3.1	14.3	14.3	1.3	1.3	3.1	3.1	0.0	0.0
PileUp	0.7	0.7	0.0	0.0	14.3	14.3	0.7	0.7	3.1	3.1	14.3	14.3
MET Unclustered Energy	0.0	0.0	0.0	0.0	0.0	0.0	1.3	1.3	3.1	3.1	0.0	0.0
Top quark mass	0.2	0.2	2.8	2.8	11.6	11.6	0.6	0.6	2.4	2.4	6.3	6.3
Right region definition	0.7	0.7	3.1	3.1	14.3	14.3	1.3	1.3	6.3	6.3	14.3	14.3
Total	6.7	6.7	36	36	114.9	114.9	4.6	4.6	30.5	30.5	121.7	121.7

6.8 Summary

In this chapter, a mostly data-driven method was developed to measure the efficiency of tagging a jet originating from a u,d,s,g, or c-quark. The mis-tagging efficiency was measured by reconstructing the b-tagging discriminator distribution for non-b jets. The reconstructed b jet b-tagging distribution, measured in the previous chapter, was subtracted from the distribution in the full b-jet candidate sample after the b jet distribution was corrected for its normalisation offset using simulated events. The resulting discriminator distribution for non-b jets provides access to the mis-tagging efficiency.

The measured efficiencies and scale factors are provided in Figure6.11 and Figure 6.12 both for μ+jets and e+jets $t\bar{t}$ events at 7 and 8 TeV respectively. Next to the results for the Track Counting High Efficiency (TCHE) and Combined Secondary Vertex (CSV) algorithms that were presented in the previous sections, the results for all other available algorithms are included as well.

Just as for the b-tagging efficiency, It is expected that the mis-tagging efficiency scale factors match between the two channels. To check this consistency, the ratio of the scale factors in μ+jets and e+jets events is displayed in Figure 6.13 for every available working point. The uncertainty on this ratio is calculated from the individual uncertainties assuming that they are uncorrelated. The measured scale factors in both channels are found to be consistent.

However, at 7TeV, a 0% efficiency is measured for some tight Working Points in the e+jets channel because the reconstructed discriminator distribution is depleted for high discriminator values. This effect is not found in the μ+jets channel since looser kinematic cuts were applied on the muon compared to the electron and no $\not\!\!E_T$ threshold was applied resulting in a larger event yield. These working points are removed from the comparison.

In chapter 7, this measured efficiency will be used to measure the $t\bar{t}$ production cross section with the requirement of one b-tagged jet in the event selection.

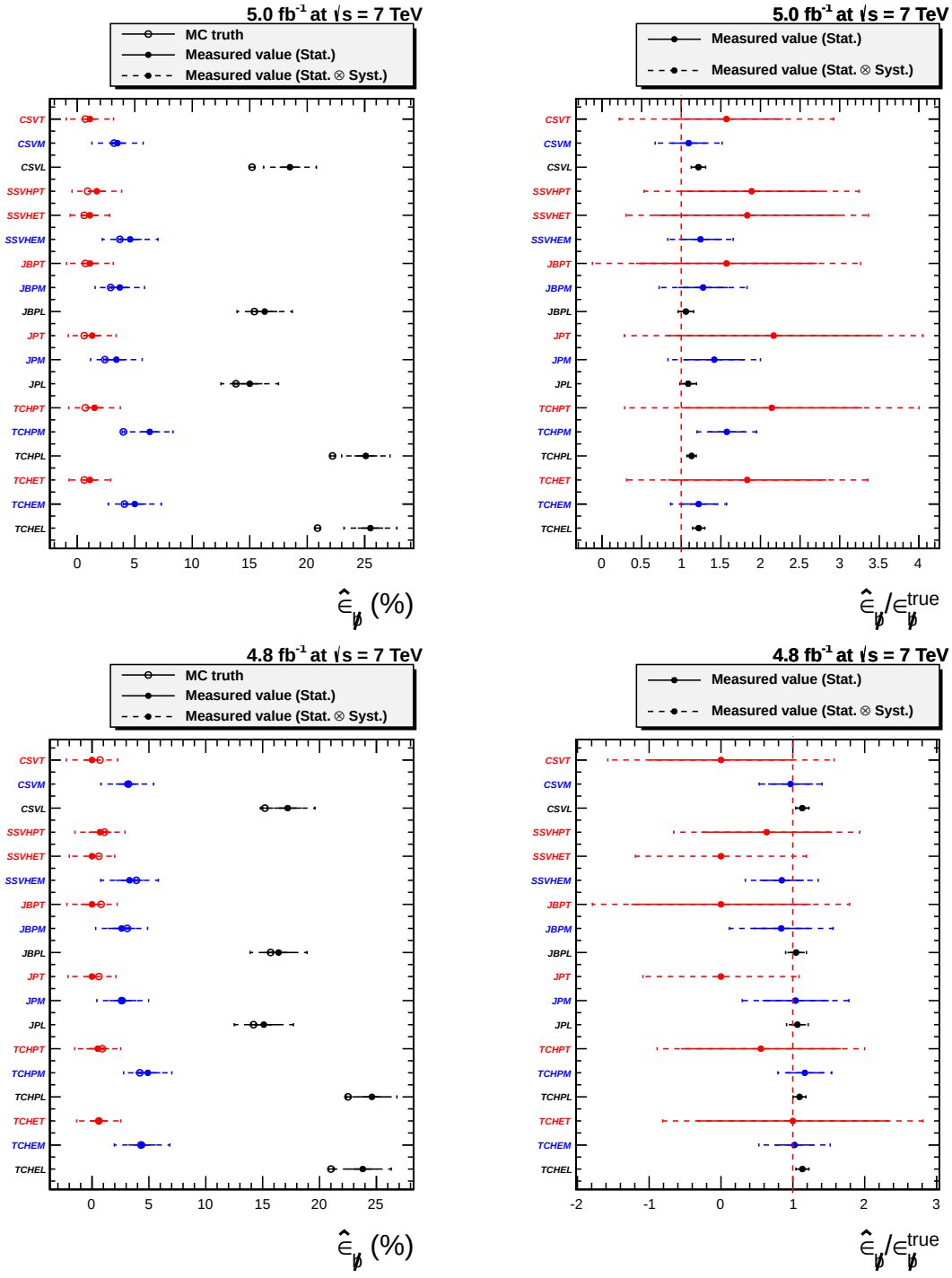
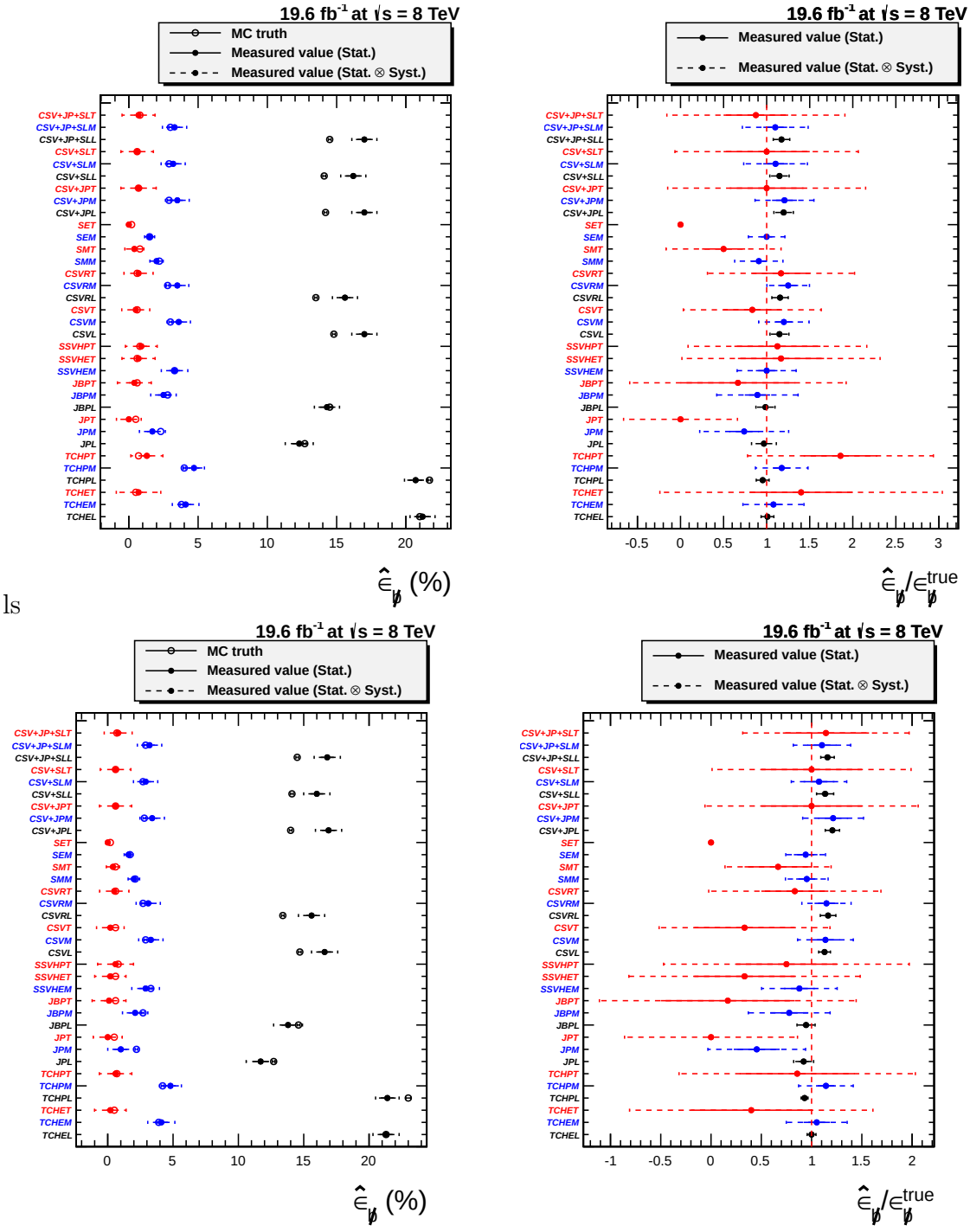


Figure 6.11: Visual representation of all working points for the studied b-tagging algorithms at 7 TeV. The measured mis-tagging efficiencies are shown (left) as well as the data to simulation scale factors (right) in the μ+jets channel (top) and in the e+jets channel (bottom).



ls

Figure 6.12: Visual representation of all working points for the studied b-tagging algorithms at 8 TeV. The measured mis-tagging efficiencies are shown (left) as well as the data to simulation scale factors (right) in the μ+jets channel (top) and in the e+jets channel (bottom).

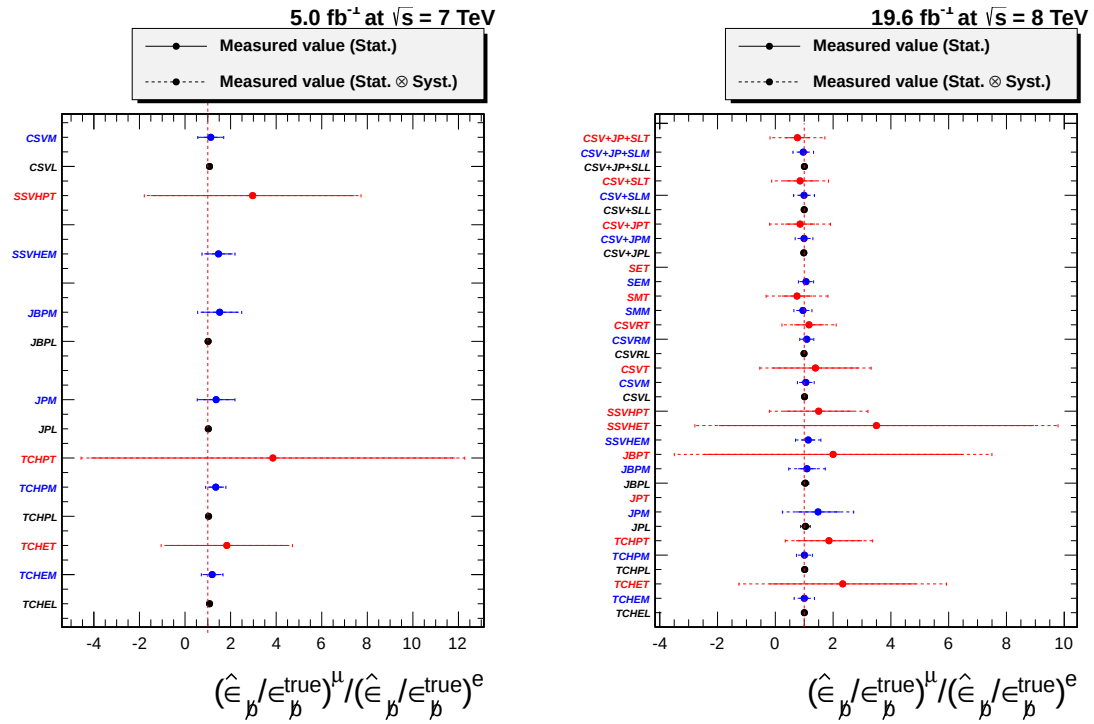


Figure 6.13: The ratio of the measured mis-tagging efficiency scale factor measured in μ+jets and e+jets events is shown for all working points at 7 TeV (left) and 8 TeV (right)

Chapter 7

Measurement of the $t\bar{t}$ production cross section

The measurement of the inclusive top quark pair cross section at the LHC is important for diverse reasons. The measurement in itself is a test of the Standard Model as well as a benchmark for the theoretical calculations. The result is also relevant for those searches beyond the Standard Model for which the top quark pair processes are the dominant background. In the decay of top quarks bottom quarks are present as the CKM element V_{tb} is constrained to be close to unity. Therefore the top quark pair signal can be isolated from the background in the proton collisions at the LHC, by use of b-quark identification algorithms or so-called b-tagging. Although the application of b-tagging will purify the selected sample, it will also introduce important systematic uncertainties due to our understanding of the performance of these b-tagging algorithms. With a combined measurement of the top quark pair cross section with the b-tagging performance we will be able to reduce the total uncertainty on the top quark pair cross section.

The b-tagging efficiency and light jet mis-tagging rate will be estimated with the method outlined in Chapter 5 and 6 respectively. The method will be applied on the semi-leptonic decaying top quark pairs, namely $t\bar{t} \rightarrow bWbW \rightarrow bqqb\ell(\ell = e, \mu)\ \nu_\ell$ reflecting about 8/27 of the total branching ratio of top quark pairs. Using the lepton flavour, two decay channels are defined in semi-leptonic decaying top quark pairs. The first μ+jets channel containing events with a reconstructed muon and at least four jets resembling the semi-muon top quark pair decay. The second is the e+jets channel which is similar to the first but with an electron in stead of a muon.

The top quark pair production cross section ($\sigma_{t\bar{t}}$) can be determined experimentally from the luminosity of the data ($\mathcal{L}$), the number of observed $t\bar{t}$ events ($N_{t\bar{t}}^{obs}$) and the signal selection efficiency (ϵ) as depicted in eq. (7.1).

$$\sigma_{t\bar{t}} = \frac{N_{t\bar{t}}^{obs}}{\mathcal{L}} \times \frac{1}{\epsilon} \tag{7.1}$$

The signal selection efficiency is proportional to how well $t\bar{t}$ signal events are selected and is the topic of Section 7.2. In Section 7.1, the procedure to extract $N_{t\bar{t}}^{obs}$ from data is discussed and finally the resulting cross section from both the μ+jets and e+jets channel is provided in Section 7.3. Additionally, in Section 7.6, both channels are combined

into a l+jets channel measurement. The systematic uncertainties are described in 7.5 and finally the cross section results are compared using different b-tagging algorithms and working points in Section 7.9.

7.1 Estimating the number of top quark pairs in data

To measure the top quark pair production cross section, the number of top quark pair events, further denoted as $t\bar{t}$ events, in the data has to be known. This figure can be determined by usage of a binned maximum likelihood fit, often called a template fit. This type of fit uses simulation driven template distributions of a variable that discriminates between signal and background. For each distinguishable process, a template is derived from simulation and then fit to data. The result is the relative fraction of each process in the given data sample yielding directly the number of observed $t\bar{t}$ events.

The template fit requires a variable to discriminate between $t\bar{t}$ events and background events. In Chapter 5, the jet-lepton mass distribution was introduced as a discriminator between b jets and other jets in the light of estimating the b tagging efficiency. Looking at its distribution in Figure 7.1 it is apparent that this variable also holds discriminating power between signal and background. The $t\bar{t}$ events give a very peaked distribution, because of the event kinematics, while the background events have a larger tail and generally a broader distribution. It is also visible that in the electron channel the background level is also higher.

Although the jet-lepton mass distribution in Figure 7.1 hints good discriminating power between the $t\bar{t}$ events and background processes, it does not seem to provide the same level of discrimination between the various backgrounds. This means that the template fit will have to be performed using a combined background template next to the signal. As the backgrounds were taken directly from simulation using the theoretical cross sections, the result might be sensitive to their modelling and the uncertainty on the theoretical cross section calculations. Two main options exist to circumvent this effect.

First, the background distributions could be determined from data. Using data-driven techniques, one can determine the jet-lepton mass shape for the dominant W+jets background as well as for the other backgrounds. The downside of this approach is that every background would need a dedicated method to estimate the shape. Given the shapes one would also need to measure the normalisation of the templates to be fitted to data because the jet-lepton mass itself does not properly discriminate between different background components.

Another approach is to reduce the background in a way that the final measured cross section is not very sensitive to the uncertainties on the background levels. One way of reducing the background is to exploit the fact that in $t\bar{t}$ events, unlike in most backgrounds, b-quarks are produced in the final state. These quarks can be identified in the reconstructed event by application of b-tagging. This purifies the selected event sample significantly as can be seen from Figure 7.2 where the distributions from Figure 7.1 are shown but with the requirement that the leptonic b-jet candidate, that con-

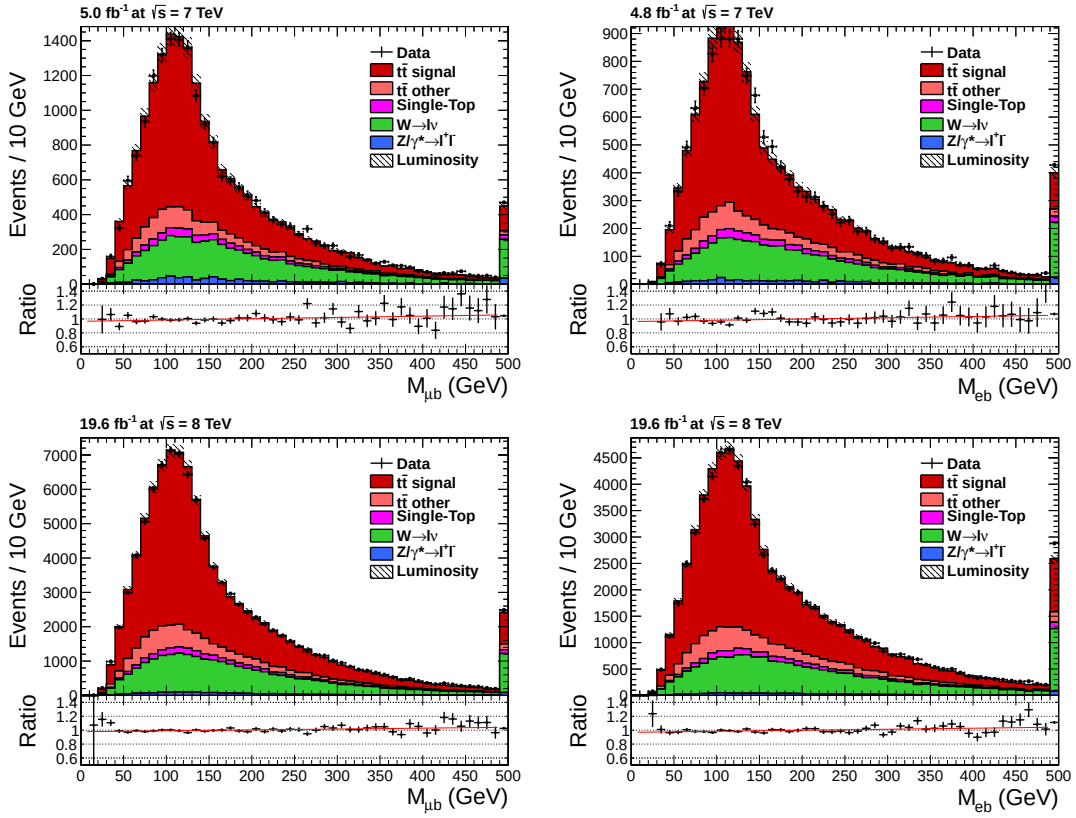


Figure 7.1: Distribution of the jet-lepton mass for signal all relevant background processes in the μ+jets channel (left) and in the e+jets channel (right). The simulated distribution is compared to data. The distributions of all simulated processes are normalized to the integrated luminosity of the data.

stitutes the jet-lepton mass together with the lepton, passes the Combined Secondary Vertex (CSV) b-tagging algorithm at its medium working point. After application of b-tagging the background component is reduced to a low level reducing the normalisation uncertainty to well below 1% on the final measurement, as will be shown in Section 7.5. Hence, this approach is favoured in the analysis.

The template fit was performed on the jet-lepton mass after b-tagging the leptonic b jet candidate and the result is shown in Figure 7.3. The simulated distributions are normalised to the template fit results and the overall agreement between data and simulation is good. From the template fit it is now possible to determine the number of observed $t\bar{t}$ events by taking the new normalisation for the $t\bar{t}$ component. The results are listed in Table 7.1. Since we are interested in the cross section of $t\bar{t}$ production, all $t\bar{t}$ decay modes other than the semi-leptonic will be considered signal within this analysis. As a consequence, the background component consists only of non-$t\bar{t}$ processes like W/Z+jets and Single-Top quark production.

Using the number of observed $t\bar{t}$ events obtained from Table 7.1, the cross section can be calculated once the signal efficiency is known. This will be determined in the next section.

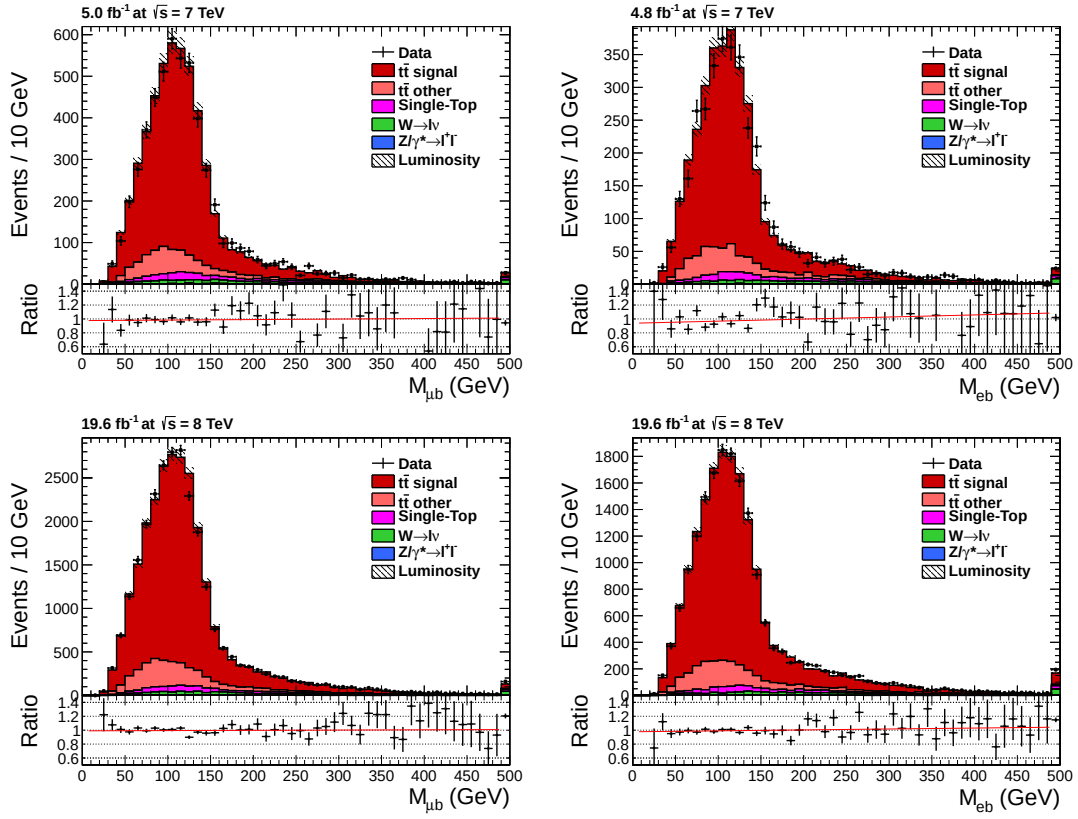


Figure 7.2: Distribution of the jet-lepton mass for signal all relevant background processes in the μ+jets channel (left) and in the e+jets channel (right) after requiring the leptonic b-jet candidate to pass the CSV algorithm medium working point. The simulated distribution is compared to data. The distributions of all simulated processes are normalized to the integrated luminosity of the data.

Table 7.1: Number of observed $t\bar{t}$ signal and background events obtained from the jet-lepton mass template fit. The uncertainty is statistical only.

	$\sqrt{s} = 7$TeV			$\sqrt{s} = 8$TeV		
Channel	N_{data}	$N_{t\bar{t}}^{obs}$	N_{bkg}^{obs}	N_{data}	$N_{t\bar{t}}^{obs}$	N_{bkg}^{obs}
μ+jets	5406	4693.6±91.3	712.4±92.1	26741	24462.9±231.2	2277.9±232.1
e+jets	3403	2985.9±69.4	417.1±69.9	18098	16489.9±171.7	1608±172.6

7.2 Determining the total event selection efficiency

In the previous section, the technique to estimate the number of $t\bar{t}$ events in data was outlined. Since kinematic cuts have been applied to separate the signal events from the abundant background processes, it was argued that the total event selection efficiency for the signal (ϵ_{tot}) has to be known to calculate the total inclusive $t\bar{t}$ cross section.

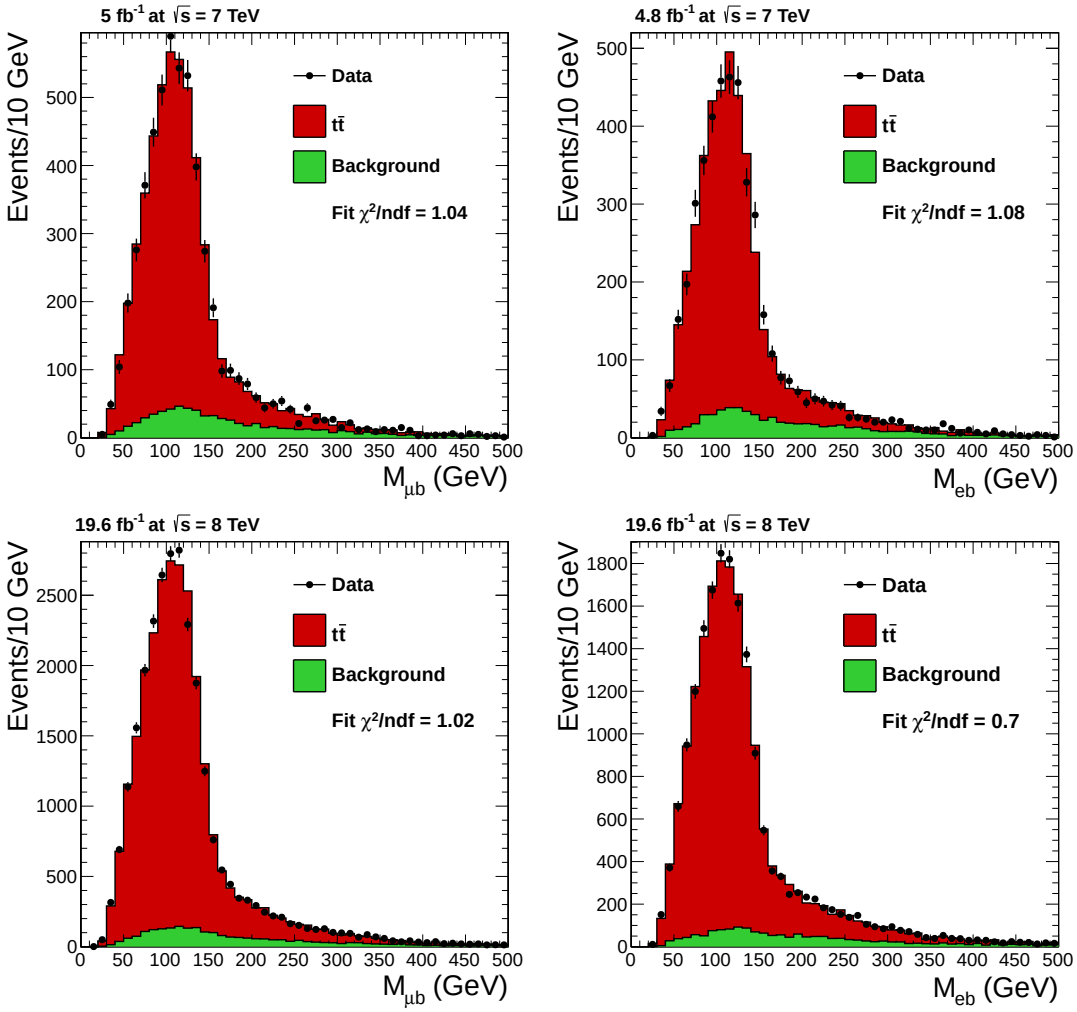


Figure 7.3: Distribution of the jet-lepton mass for signal all relevant background processes in the μ+jets channel (left) and in the e+jets channel (right). The signal and background components are normalised to the template fit result. The templates were obtained from events where the leptonic b-jet candidate to passes the CSV algorithm medium working point

The total event selection efficiency can be factorized in different components as shown in eq. (7.2).

$$\epsilon_{tot} = A \times \epsilon \tag{7.2}$$

The first component is the theoretical acceptance (A). This figure allows the extrapolation of the measured cross section from the measurable phase space to the full theoretical phase space. The acceptance depends purely on the theoretical model.

The second piece is the detector efficiency (ϵ) which describes how well the detector can reconstruct $t\bar{t}$ events. This efficiency can be further broken down into different components, as shown in eq (7.3), that are measured on data or determined from simulated events. Where available, a data-driven correction factor applied to remove

discrepancies between data and simulation.

$$\epsilon = \epsilon_{sel} \times \epsilon_{M_{lb}} \times \epsilon_{\chi^2} \times \epsilon_{btag} \tag{7.3}$$

7.2.1 Theoretical acceptance (A)

To be able to separate the $t\bar{t}$ signal events from the background, kinematic cuts have to be made. These kinematic cuts, however, remove a significant part of the signal events. This means that the method proposed in the previous section is in fact only measuring the number of signal events in a slice of the full phase space, called the visible phase space. When looking at simulated $t\bar{t}$ events, only a certain amount will pass the final event selection and the ratio is what is called the acceptance. This acceptance figure is then used to extrapolate the measurement from the visible phase space to the complete theoretical phase space.

The acceptance is estimated on simulated events by mimicking the final reconstruction level event selection at generator level. In the generated events exactly one muon or electron is required depending on which decay channel is studied. Secondly, the missing transverse energy (MET) requirement is translated into the requirement on the presence of 1 neutrino . Finally, at least 4 generator jets have to be present in the event. An overlapping acceptance definition is chosen among both decay channels to allow the definition of a combined l+jets visible cross section (cfr. Section 7.6). The kinematic cuts are set according to the reconstruction level cuts in the electron channel since these are tighter. The same acceptance definition is also used between 7 and 8 TeV. These cuts are summarised in Table 7.2 and allow to define the acceptance using any generator or model.

Table 7.2: Overview of the generator level cuts used to determine the acceptance

	$\sqrt{s}$ = 7 TeV				$\sqrt{s}$ = 8 TeV			
	μ+jets		e+jets		μ+jets		e+jets	
	#	Kin. cuts	#	Kin. cuts	#	Kin. cuts	#	Kin. cuts
Lepton	==1	$p_T > 32GeV$ $\lvert\eta\rvert < 2.1$	==1	$p_T > 32GeV$ $\lvert\eta\rvert < 2.1$	==1	$p_T > 32GeV$ $\lvert\eta\rvert < 2.1$	==1	$p_T > 32GeV$ $\lvert\eta\rvert < 2.1$
Neutrino	== 1	$p_T > 40GeV$	==1	$p_T > 40GeV$	==1	$p_T > 40GeV$	==1	$p_T > 40GeV$
GenJets	≥4	$p_T > 40GeV$ $\lvert\eta\rvert < 2.5$ $\Delta R(j,l) > 0.4$	≥4	$p_T > 40GeV$ $\lvert\eta\rvert < 2.5$ $\Delta R(j,l) > 0.4$	≥4	$p_T > 40GeV$ $\lvert\eta\rvert < 2.5$ $\Delta R(j,l) > 0.4$	≥4	$p_T > 40GeV$ $\lvert\eta\rvert < 2.5$ $\Delta R(j,l) > 0.4$

The acceptance values for both decay channels at $\sqrt{s}$ = 7 and 8 TeV are shown in Table 7.3 and are calculated using $t\bar{t}$ events generated with the MadGraph event generator where the branching ratios of the W boson decay is set to 1/3 for $W \to l^+l^-$ and 2/3 for $W \to q\bar{q}$.

As these acceptance values strongly depend on the theoretical model, and as such on the event generator in use, the cross section will be measured both in the visible phase space alone and secondly in the full phase space using the determined acceptance correction. The former cross section is valuable for theorists as it allows for extrapolation to the full phase space of any generator to be compared to any specific theoretical model.

Table 7.3: Acceptance corrections for both decay channels obtained with the MadGraph event generator

Channel	A ($\sqrt{s}$ = 7 TeV)	A ($\sqrt{s}$ = 8 TeV)
μ+jets	0.0158	0.0166
e+jets	0.0158	0.0166

7.2.2 Event selection efficiency at reconstruction level (ϵ_{sel})

At reconstruction level the events are required to contain a high quality isolated lepton, four jets and a certain amount of missing transverse energy as described in Chapter 4. Additionally, the events are only selected if they pass a trigger requiring at least one isolated high momentum lepton. To be able to measure the cross section in the visible phase space, i.e. the phase space defined by these kinematic cuts, the efficiency of these cuts needs to be estimated.

The overall selection efficiency, provided in Table 7.4, is derived from simulation by dividing the number of selected events after reconstruction level cuts by the number of events passing the generator level cuts outlined in the previous section. This requires a good level of agreement between data and simulation in order to be certain that the efficiency estimated from simulation can be applied to data.

Table 7.4: Event selection efficiency in both decay channels.

Channel	ϵ_{sel} ($\sqrt{s}$ = 7 TeV)	ϵ_{sel} ($\sqrt{s}$ = 8 TeV)
μ+jets	1.208	0.737
e+jets	1.161	0.821

The μ+jets selection efficiency exceeds 100% both at 7 and 8 TeV. This might seem controversial but this is caused by the definition of the acceptance cuts which are tighter than the reconstruction level cuts in this case. Hence, a μ+jets event can be easily selected at reconstruction level while it is not passing the cuts at generator level. This is not the case for the electrons where the kinematic cuts applied at generator level mimic the reconstruction level cuts. The differences between both collision energies can be explained by the change in triggers which require different kinematic cuts in the event selection.

To correct for small discrepancies between data and simulation for the efficiency subcomponents, data-driven correction factors are applied. More precisely, corrections are applied for the lepton trigger, identification and isolation cuts. All the other objects have been shown in Chapter 4 and show good description in simulation compared to data.

The lepton trigger efficiency can be measured on data along with the lepton identification and isolation efficiencies. These efficiencies are typically measured using a so-called *tag and probe* [108] technique using $Z \rightarrow l^+l^-$ decays. Events are required to have two leptons providing a mass in a narrow window around the Z-mass. Subsequently, tight identification requirements are applied on one of the two leptons in these dilepton pairs. Finally, the second lepton in the pair is used as a probe. The cut under study is applied and the number of probes passing the cut is divided by the number of probes yielding the cut efficiency. Moreover, applying this method both on data and on simulation provides a data-to-simulation scale factor to be applied on the overall selection efficiency derived from simulation. Figures 7.5 and 7.4 provide the combined lepton trigger, identification and isolation scale factors at $\sqrt{s} = 7$ and 8 TeV respectively.

At 8 TeV, the muon and electron scale factors are measured in bins of p_T and η. In the transition region from the barrel to the endcaps ($|\eta| \approx 1$) the muon scale factor shows a drop over the full p_T range which is depicted by the blue band. The same drop towards the endcaps is seen for the electron scale factor which shows best data to simulation agreement in the barrel region.

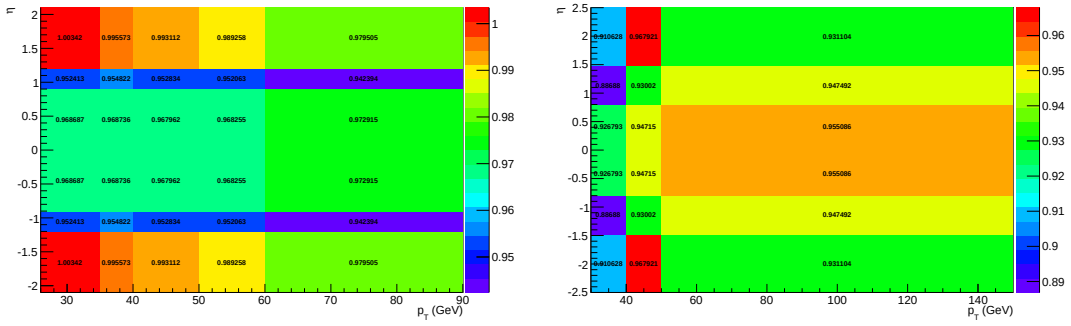


Figure 7.4: Combined scale factor for lepton ID/Isolation and lepton trigger efficiency at $\sqrt{s} = 8$ TeV in the muon channel (left) and in the electron channel (right)[131, 132].

The muon scale factor at 7 TeV was found to be invariant with respect to the muon p_T. For this reason the scale factor was only provided as a function of η and is close to one over the full range. In the barrel-endcap transition region ($|\eta| \approx 1$), the scale factor drops a little.

Since all single electron triggers in the 7 TeV data were pre-scaled at some point, a trigger was used requiring both an isolated lepton and some hadronic activity in the central region in the form of three jets. The efficiency for this trigger can be split up in the leptonic trigger leg and the hadronic leg. The leptonic leg was already included in Figure 7.5 showing a scale factor close to 1 almost everywhere except for the white

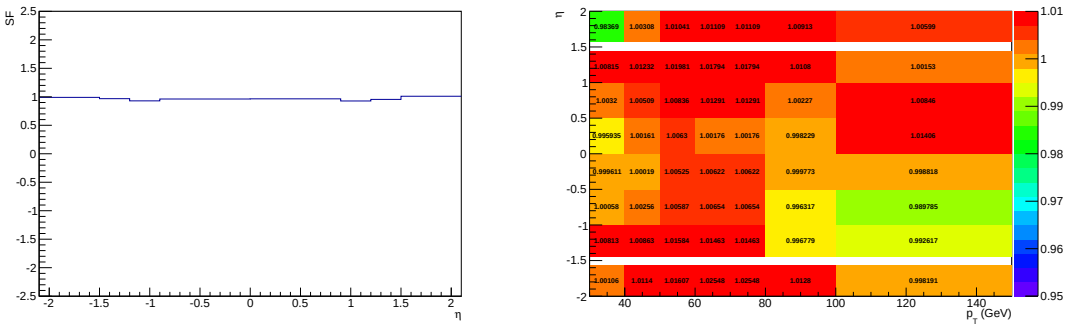


Figure 7.5: Combined scale factor for lepton ID/Isolation and lepton trigger efficiency at $\sqrt{s} = 7$ TeV in the muon channel (left) and in the electron channel (right). The hadronic leg of the electron trigger is not included in the scale factor [133, 134].

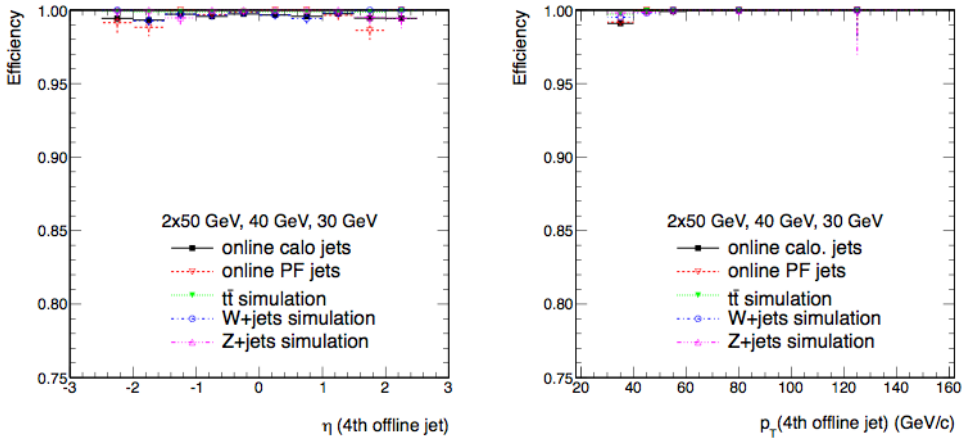


Figure 7.6: Efficiency for the hadronic leg of the electron trigger $\sqrt{s} = 7$ TeV as a function of the fourth leading jet p_T and η. The efficiency is compared between data and various simulations and is very close to 1 [134].

bands in this plot that are due to the exclusion of electrons with their SuperCluster located in the EB-EE transition region. The hadronic leg is provided in Figure 7.6. From the figure it is clear that both data and simulation are fully efficient over the analysed p_T and η range of the fourth leading jet in the event. Hence the scale factor is equal to 1 and is no longer considered in the analysis.

7.2.3 Jet-lepton mass overflow bin removal ($\epsilon_{M_{lb}}$)

In the binned maximum likelihood fit to the jet-lepton mass, the overflow bin in the templates is not accounted for. This entails that events generating a jet-lepton mass exceeding 500 GeV are not considered in the fit. This deficit in events leads inevitably to a small bias on the cross section. To remove this effect, the fit range could be

extended so that the overflow bin is empty. This, however, would require to fit up to very large masses where the uncertainties on the main background processes become very large and statistical fluctuations are at play.

Another alternative is to remove events where the jet-lepton mass exceeds the value of 500 GeV. In this way the overflow bin is empty while keeping the fit range as before. This allows the template fit to yield an unbiased number of observed $t\bar{t}$ events although the effect of this cut has to be modelled when calculating the cross section. The efficiency of this cut on signal events has to be estimated and since no data-driven method exists, it is determined on simulation.

Figure 7.1 shows the jet-lepton mass distributions in both muon and electron channels showing a good agreement between data and the simulated events. This shows that the efficiency of the jet-lepton mass cut, provided in Table 7.5, can be properly estimated from simulation. Small differences in $\epsilon_{M_{lb}}$ can be observed between the μ+jets and e+jets channels which arise from differences in the reconstruction level kinematic cuts on the respective leptons.

Table 7.5: Cut efficiency $\epsilon_{M_{lb}}$ for both decay channels using simulated events.

Channel	$\epsilon_{M_{lb}}$ ($\sqrt{s} = 7$ TeV)	$\epsilon_{M_{lb}}$ ($\sqrt{s} = 8$ TeV)
μ+jets	0.983	0.973
e+jets	0.977	0.968

7.2.4 Jet combination threshold (ϵ_{χ^2})

To construct the jet-lepton mass, namely the mass of the leptonic b jet candidate and muon system, a jet combination was used based on the mass and width of the W-boson and top quark. This jet combination method yields a χ^2 value for each possible permutation of the four leading jets in the event characterising how well the combination matches the final state partons in the $t\bar{t}$ decay. The combination with the lowest χ^2 value is taken as the best combination, albeit that a cut is made to remove events where even the best combination is far from matching the $t\bar{t}$ topology. This cut was determined using the techniques explained in Chapter 5 and its efficiency has to be determined. The χ^2 value for the best jet combination in each event is shown in Figure 7.7 for both muon and electron channels at 7 and 8 TeV. The Figure shows overall good agreement between data and simulation and the flat slope in the ratio plots show that the efficiency of the χ^2 cut can be determined using simulated events.

The simulation based efficiencies to be applied on data are provided in Table 7.6. The efficiency differs among both decay channels as well as between the different collision energies. The origin of this difference lies in the different χ^2_{min} threshold that is applied in all four cases which is determined to minimise the bias on ϵ_b as explained in Chapter 5.

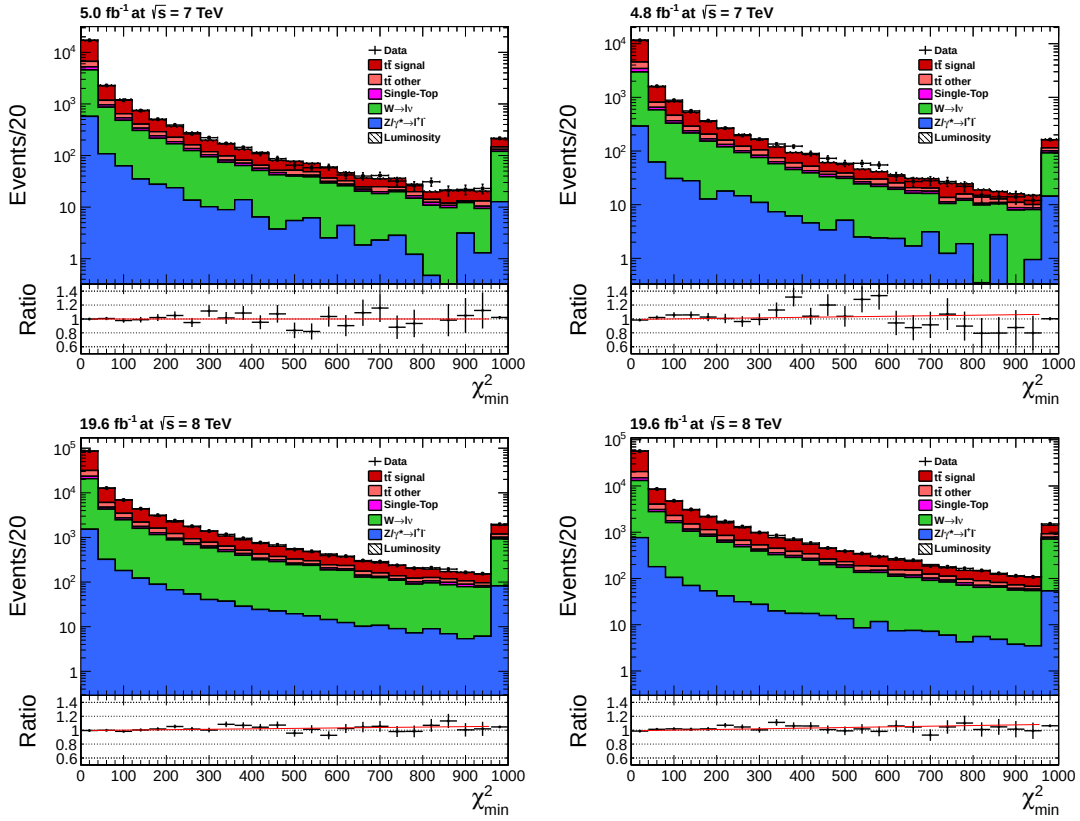


Figure 7.7: Distribution of the χ^2 value for the best jet-combination in each event (χ^2_{min}).

Table 7.6: Cut efficiency ϵ_{χ^2} for both decay channels using simulated events.

Channel	ϵ_{χ^2} ($\sqrt{s}$ = 7 TeV)	ϵ_{χ^2} ($\sqrt{s}$ = 8 TeV)
μ+jets	0.883	0.827
e+jets	0.879	0.896

7.2.5 Efficiency of the b-tagging cut (ϵ_{btag})

The $t\bar{t}$ process under study is rich in b jets. Hence, to reduce the background a b-tagging cut is applied on the leptonic b jet candidate in the jet combination with the minimal χ^2. To estimate the efficiency it is not sufficient to use the measured b-tagging efficiency as measured in Chapter 5.

The b-tagging efficiency previously measured represents the efficiency of tagging a jet as b jet when it actually originates from a b quark. Since the χ^2 jet sorting algorithm is not fully efficient, the leptonic b jet candidate will not always originate from a b jet. As a consequence, the mis tagging efficiency, has to be accounted for.

To properly combine both the b-tagging efficiency and the mis tagging efficiency, the populations of both b jets and non-b jets in the leptonic b jet candidate sample

have to be known. The fraction of b jets, denoted as g, is provided in Table 7.7. This fraction is obtained from simulation where the reconstructed jets are matched to the generator level quarks to determine their flavour. Finally, using both efficiencies and the factor g the overall b-tagging cut efficiency on the leptonic b candidate can be determined using eq. (7.4) where ϵ_b and ϵ_q are obtained from the measurements described in Chapters 5 and 6 respectively.

$$\epsilon_{btag} = g \times \epsilon_b + (1 - g) \times \epsilon_q \tag{7.4}$$

Table 7.7: Fraction of b jets in the leptonic b-jet candidate sample for both decay channels obtained with the MadGraph generator.

Channel	g ($\sqrt{s} = 7$ TeV)	g ($\sqrt{s} = 8$ TeV)
μ+jets	0.495	0.490
e+jets	0.477	0.470

7.3 Results in the electron and muon channels

In the previous sections, the major components to measure the $t\bar{t}$ cross section (eq. (7.1)), namely the number of $t\bar{t}$ pairs and the signal efficiency, were outlined. The first number can be measured through a binned maximum likelihood fit to the jet-lepton mass.

Secondly, the efficiency was broken down into its components and they were determined on simulation with data-driven correction factors or directly measured from data. Using eq. (7.3), the detector efficiency can be plugged into eq (7.1) to yield the visible $t\bar{t}$ cross section.

This visible cross section is proportional to the production rate of top quark pairs in the kinematic phase space selected by the analysis. As a consequence, this measurement is not sensitive to the underlying theory model in the Monte Carlo event generator. This allows the measurement to be compared to theoretical calculations in any model given that the acceptance corresponding to that given model is estimated using the generator level cuts in Table 7.2.

Below, the measured visible cross section at a centre-of-mass energy of 7 TeV is provided.

μ+jets channel:

$$\sigma_{t\bar{t}}^{vis} = 2.524 \pm 0.088(stat.)^{+0.177}_{-0.209}(syst.) \pm 0.056(lumi.)pb$$

e+jets channel:

$$\sigma_{t\bar{t}}^{vis} = 2.652 \pm 0.105(stat.)^{+0.138}_{-0.210}(syst.) \pm 0.058(lumi.)pb$$

At a centre-of-mass energy of 8 TeV the visible cross section is measured to be

$$\sigma_{t\bar{t}}^{vis} = 3.894 \pm 0.058(stat.)^{+0.132}_{-0.214}(syst.) \pm 0.101(lumi.)pb$$

in the μ+jets channel and

$$\sigma_{t\bar{t}}^{vis} = 3.990 \pm 0.066(stat.)^{+0.152}_{-0.255}(syst.) \pm 0.104(lumi.)pb$$

int the e+jets channel.

Because an identical acceptance is used in both channels the measured visible cross section is consistent among the two decay channels both for 7 and 8 TeV. Using the acceptances displayed in Table 7.3 and eq. (7.2), the visible cross section can be extrapolated to the total cross section.

The acceptance in both channels is determined using simulated $t\bar{t}$ events with the *MadGraph* generator. At a centre-of-mass energy of 7 TeV, this yields a cross section of

$$\sigma_{t\bar{t}} = 159.7 \pm 5.6(stat.)^{+11.2}_{-13.3}(syst.) \pm 3.5(lumi.)pb$$

in the μ+jets channel and

$$\sigma_{t\bar{t}} = 167.9 \pm 6.6(stat.)^{+8.7}_{-13.4}(syst.) \pm 3.7(lumi.)pb$$

in the e+jets channel. Finally, at 8 TeV the cross section is found to be

$$\sigma_{t\bar{t}} = 234.6 \pm 3.5(stat.)^{+8.0}_{-12.9}(syst.) \pm 6.1(lumi.)pb$$

in the μ+jets channel and

$$\sigma_{t\bar{t}} = 240.4 \pm 4(stat.)^{+9.4}_{-15.1}(syst.) \pm 6.3(lumi.)pb$$

in the e+jets channel. The total cross section is found to be consistent within uncertainties between both decay channels as expected from theory. The consistency is found at both centre-of-mass energies.

The systematic uncertainties quoted are determined using simulated events and the different sources are explained in the next section. Subsequently, to reduce the overall uncertainties, both $t\bar{t}$ decay channels will be combined into a lepton+jets result using the BLUE method explained in Section 7.6.

7.4 Statistical properties of $\sigma_{t\bar{t}}$

In Figure 7.8 the pull distribution is shown for the $t\bar{t}$ cross section estimator. The pull distribution is determined using 750 pseudo-experiments corresponding to an integrated luminosity of $19.7 fb^{-1}$ each.

The pull distribution is fitted with a guass function. The fit returns a pull width of 1.044±0.029. Considering the high degree of correlation between the different experiments due to the limited simulated statistics at hand the pull can be considered close to unity showing that the statistical uncertainty on $\hat{\sigma}_{t\bar{t}}$ is properly estimated.

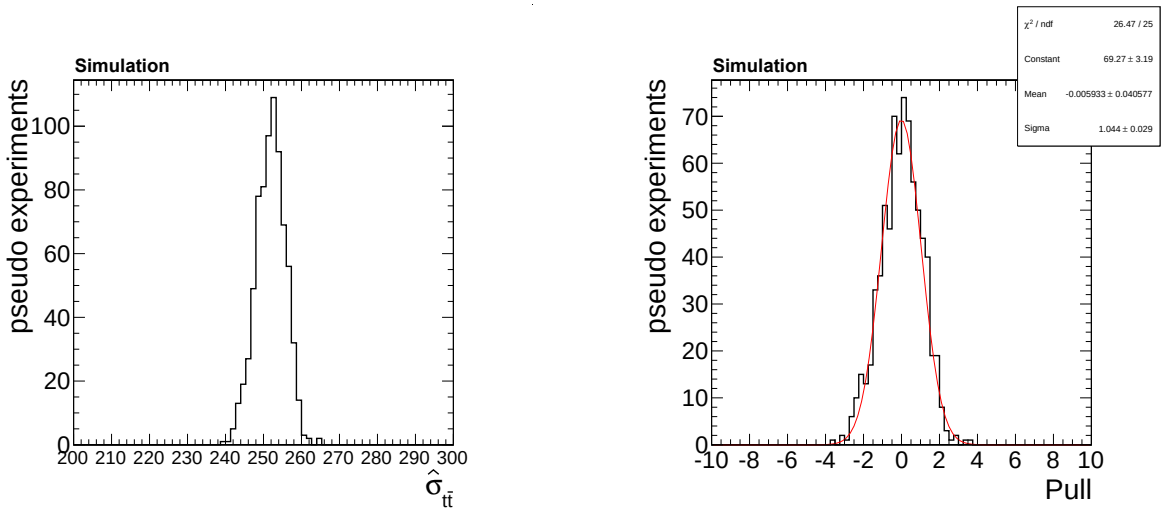


Figure 7.8: Distribution of the b tagging efficiency estimator (left) and the pull distribution (right) for the TCHE Medium WP

7.5 Systematic uncertainties

Tables 7.8 and 7.9 provide a numerical overview of the total systematic uncertainty and its components on the visible and total cross section measurement respectively. These systematic uncertainties are devised to cover possible differences between data and simulation concerning the objects used for event selection as well as theoretical uncertainties concerning the $t\bar{t}$ events and various background processes. In what follows, the systematic sources will be explained one by one.
For each uncertainty source, an up and downwards shift is applied to the simulated samples. For both the upward and downward shift, the analysis is repeated yielding the $t\bar{t}$ cross section for both variations through eq. (7.5).

$$\sigma_{t\bar{t}}^{\pm 1\sigma} = \frac{N_{t\bar{t}}^{\pm 1\sigma}}{\mathcal{L}} \times \frac{1}{\epsilon^{nominal}} \tag{7.5}$$

The systematic uncertainty is then obtained by looking at the relative difference between the nominal and the varied measurement.

$$\Delta\sigma_{t\bar{t}}^{\pm 1\sigma}(\%) = \frac{\sigma_{t\bar{t}}^{\pm 1\sigma} - \sigma_{t\bar{t}}^{nominal}}{\sigma_{t\bar{t}}^{nominal}}$$

The following systematic sources are estimated using the prescription provided in Chapter 5: Jet Energy Scale, Jet Energy Resolution, $\not{E}_T$ unclustered energy and Pile up. The theory modelling uncertainties such as PDF uncertainties, Top Quark mass, Factorisation Scale and ME-PS Matching threshold are evaluated as described in Chapter 6. Moreover, the Jet Energy Scale uncertainty quoted here is already constrained by the W boson mass calibration technique explained in the next Section.

Since a b-tagging criterion is applied in the cross section measurement, the b-tagging and mis tagging efficiency need to be accounted for as explained in Section 7.2.5. In this analysis, both efficiencies are measured in-situ which has the benefit of omitting

the systematics that have to otherwise be quoted on the b-tagging and mis tagging scale factors. This approach allows to vary the b-tagging and mis tagging efficiencies simultaneously with the cross section when evaluating the effect of various systematic uncertainties. Possible negative correlations between the three estimators could then potentially lead to partial cancellation of systematics.

However, since the in-situ calibration of the b-tagging efficiency is performed, an additional systematic, *Btag Method Settings*, needs to be added to account for the simulation based tuning of the b-enriched and b-depleted regions as described in Chapter 5.

Two additional systematic uncertainty sources are evaluated in addition to the previously mentioned: the Lepton identification and trigger efficiency systematics and the uncertainty on the luminosity. Both systematics affect the normalisation of the M_{lj} distribution and thus directly affect the estimated cross section.

Lepton identification and trigger efficiency

To account for differences in lepton identification and isolation efficiency as well as trigger efficiency between data and simulation, a scale factor was applied to the overall detector selection efficiency in Section 7.2. To obtain the scale factor, the bare efficiency was measured on data and compared to the simulation based efficiency. A lepton p_T and η based systematic uncertainty is added to cover all uncertainty sources on these measurements.

Luminosity

In the equation to calculate the cross section of the number of observed events and the event selection efficiency, eq. (7.1), the integrated luminosity of the data sample has to be entered. The luminosity value is measured on data and for the 2011 dataset, the uncertainty is found to be $\pm 2.2\%$ [135]. For the 2012 dataset, the pixel based luminosity measurement has been performed, just as for the 7 TeV dataset, resulting in a luminosity uncertainty of $\pm 2.6\%$ [136].

7.5.1 Constraining the Jet Energy Scale uncertainty

In the $t\bar{t}$ cross section measurement, the Jet Energy Scale uncertainty is by far the dominant source of systematic uncertainty. This can be understood by the fact that the Jet Energy Scale not only affects the templates used in the template fit but also strongly affects the part of the event selection efficiency concerning the jet selection. This is not, however, the conclusion that would be drawn by looking at Tables 7.8 and 7.9 where the Jet Energy Scale uncertainty is not dominant. This is because, an additional calibration for Jet Energy Scale is implemented.

Since a W boson decaying into two quarks is present in each $t\bar{t} \to W(\to l\nu_l)bW(\to q\bar{q})b$ event, its mass can be exploited to constrain the Jet Energy Scale uncertainty. As the mass of this W boson is very sensitive to the Jet Energy Scale corrections, a simulation to data correction can be determined by comparing the reconstructed W boson mass in data and in simulation. The correction factor α is defined as the ratio of

Table 7.8: Overview of the systematic uncertainties on the visible cross section measurement.

	$\sqrt{s}$ = 7 TeV		$\sqrt{s}$ = 8 TeV	
	μ+jets	e+jets	μ+jets	e+jets
Systematic	$\delta\sigma_{t\bar{t}}$ (%)	$\delta\sigma_{t\bar{t}}$ (%)	$\delta\sigma_{t\bar{t}}$ (%)	$\delta\sigma_{t\bar{t}}$ (%)
Jet Energy Scale	+4.9 –4.7	+4.4 –6.1	+2.4 –2.3	+2.0 –2.2
Jet Energy Resolution	+2.3 –0.6	+0.2 –1.9	+0.7 –1.0	+0.2 –1.6
MET Unclustered Energy	+0.0 –0.0	+0.2 –0.4	+0.2 –0.0	+0.7 –0.1
Pileup	+0.4 –0.3	+0.6 –0.6	+0.3 –0.6	+0.3 –0.3
Lepton ID/Trigger SF	+1.5 –1.5	+1.7 –1.7	+0.6 –0.3	+0.5 –0.6
Btag Method Settings	+0.0 –0.9	+1.0 –0.3	+0.1 –0.5	+0.6 –0.8
Background composition	+0.4 –0.5	+0.4 –0.4	+0.3 –0.1	+0.4 –0.3
Factorisation Scale	+3.4 –4.1	+0.0 –0.7	+0.5 –3.0	+1.6 –3.6
ME-PS Matching threshold	+0.0 –4.0	+0.6 –2.9	+0.9 –1.7	+1.9 –2.8
Top Quark Mass	+0.1 –2.1	+0.4 –2.0	+0.1 –2.5	+0.4 –2.6
PDF Uncertainties	+2.1 –2.6	+1.8 –2.1	+1.9 –2.3	+1.7 –2.1
Total	+7.0 -8.3	+5.2 -7.9	+3.4 -5.5	+3.8 -6.4
Luminosity	+2.2 -2.2	+2.2 -2.2	+2.6 -2.6	+2.6 -2.6

Table 7.9: Overview of the systematic uncertainties on the total cross section measurement.

	$\sqrt{s}$ = 7 TeV		$\sqrt{s}$ = 8 TeV	
	μ+jets	e+jets	μ+jets	e+jets
Systematic	$\delta\sigma_{t\bar{t}}$ (%)	$\delta\sigma_{t\bar{t}}$ (%)	$\delta\sigma_{t\bar{t}}$ (%)	$\delta\sigma_{t\bar{t}}$ (%)
Jet Energy Scale	+4.9 –4.7	+4.3 –6.2	+2.4 –2.3	+2.0 –2.2
Jet Energy Resolution	+2.3 –0.6	+0.2 –2.0	+0.7 –1.0	+0.2 –1.6
MET Unclustered Energy	+0.0 –0.0	+0.2 –0.4	+0.2 –0.0	+0.6 –0.1
Pileup	+0.4 –0.3	+0.6 –0.7	+0.4 –0.6	+0.4 –0.3
Lepton ID/Trigger SF	+1.5 –1.4	+1.7 –1.8	+0.6 –0.2	+0.5 –0.6
Btag Method Settings	+0.0 –0.9	+1.0 –0.3	+0.1 –0.5	+0.6 –0.8
Background composition	+0.4 –0.5	+0.5 –0.3	+0.2 –0.2	+0.4 –0.3
Factorisation Scale	+3.5 –4.0	+0.0 –0.7	+0.5 –3.0	+1.7 –3.6
ME-PS Matching threshold	+0.0 –4.0	+0.6 –2.8	+0.9 –1.7	+1.9 –2.8
Top Quark Mass	+0.1 –2.1	+0.4 –2.0	+0.1 –2.5	+0.4 –2.6
PDF Uncertainties	+2.2 –2.5	+1.8 –2.1	+1.9 –2.3	+1.7 –2.1
Total	+7.0 -8.3	+5.2 -8.0	+3.4 -5.5	+3.9 -6.3
Luminosity	+2.2 -2.2	+2.2 -2.2	+2.6 -2.6	+2.6 -2.6

the reconstructed mass in data to the reconstructed mass in simulation. This factor can then be applied to the four-vectors of the jets in simulation which will align both masses and hence constrain the Jet Energy Scale uncertainty as this calibration is determined simultaneously with the cross section measurement. In the 8 TeV analysis, the Jet Energy Scale uncertainty decreased by more than 60% because of this calibration.

In the following Sections, the two key components of this calibration will be outlined. First, the W boson mass distribution is reconstructed from data and compared to simulation. Subsequently, in order to derive the calibration factor α, the distribution has to be fitted to extract an estimate of the W boson mass.

Reconstructing the W boson mass

To reconstruct the W boson mass, the hadronically decaying W boson in the selected $t\bar{t}$ events is used. The mass of all two-jet systems (M_{jj}) in the event are determined and collected in a distribution. As this is very sensitive to background and random jet combinations, the selection of events is altered slightly compared to the nominal event selection for the main analysis.

The requirement of $\not{E}_T$ is no longer applied since its to correlated with the JES and is replaced by requiring the presence of at least two b-tagged jets in the event passing the medium working point of the Combined Secondary Vertex algorithm. The latter criterion does not only serve the purpose of removing the multitude of background events, yet it also helps in isolating the light jet candidates most likely to originate from the hadronic W-boson decay. Hence, only two-jet systems where both jets are not b-tagged are used to construct the M_{jj} distribution and each event can contribute multiple entries in this distribution as all selected jets are used.

The M_{jj} distribution is shown in Figure 7.10 where it is compared between data and simulation at both 7 and 8 TeV and in the μ+jets and e+jets channels separately. From the data over simulation ratio it can be observed that both in the muon and electron channels a shift is present between both distributions yielding a different average mass value. This shift is apparent both at $\sqrt{s} = 7$ and 8 TeV proving that this distribution indeed provides the necessary sensitivity to the Jet Energy Scale.

Thanks to the slightly altered event selection used to build this distribution, the background level is almost negligible. The distribution shows a populated tail from the peak extending unto high masses. This tail is populated by random jet combinations since all untagged two-jet permutations in the event are considered. This tail is more predominant at 8 TeV compared to 7 TeV, possibly attributed to the higher average number of pile up interactions during the 8 TeV run.

The peak value of the M_{jj} distribution can be regarded as an estimate for the W boson mass. In data, a mass of roughly 85 GeV is obtained which is shifted compared to the world average W boson mass of 80.385±0.015 GeV [10]. This mass shift is caused by the JES corrections that are applied to the jets as described in Chapter 3.

In the left panel of Figure 7.10, the absolute Jet Energy Response is shown as a function for p_T^{gen} for jets that are only corrected with the Offset correction. The response is provided separately for quark jets and for gluon jets. Since the response is not equal to unity, an additional simulation based correction has to be applied. It can be seen that compared to gluon jets, quark jets have an energy response that is closer to

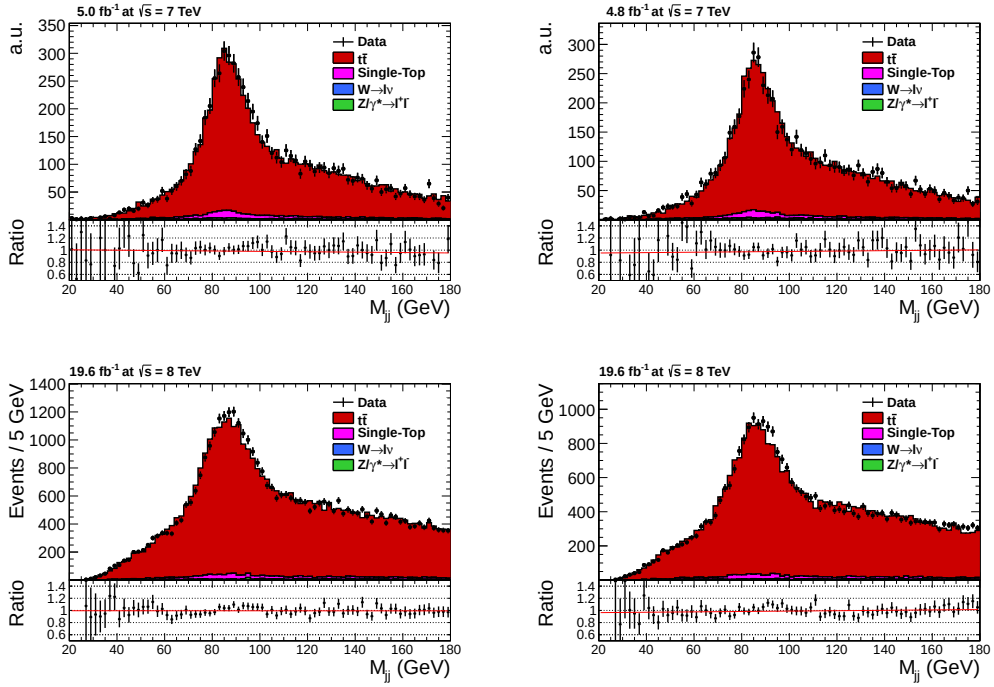


Figure 7.9: Distribution of the reconstructed W-boson mass for $t\bar{t}$ signal events and all relevant background processes in the μ+jets channel (left) and in the e+jets channel (right). The distributions are shown at 7 TeV (top) and 8 TeV (bottom). A small shift is observed in both channels between data and simulation at both collision energies. This allows to derive a global JES correction.

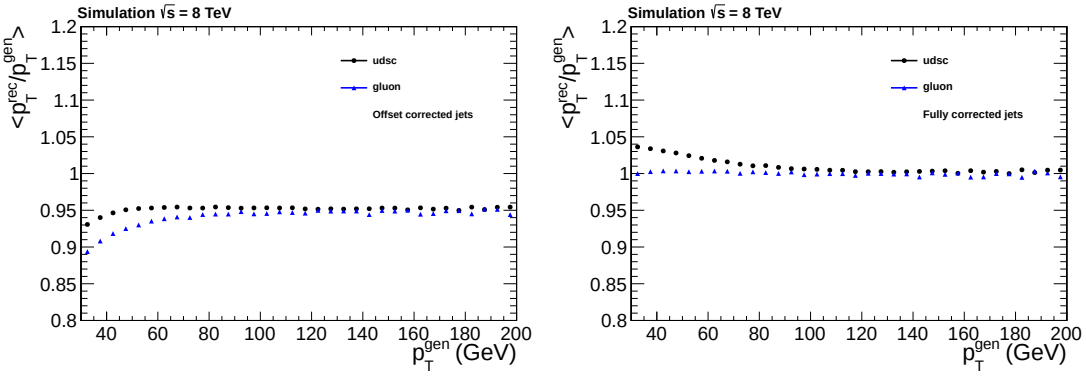


Figure 7.10: The Jet Energy Response as a function of p_T^{gen} is shown for jets in $t\bar{t}$ events that have been corrected for the Offset pileup correction (left) and jets that have also been corrected using the simulation based correction (right). The response is shown separately for quark jets and gluon jets.

unity. Since the simulation based correction is determined from a gluon jet rich QCD

sample, the correction is dominated by gluon jets. This leads to an overcorrection of quark jets as can be seen from the right panel in Figure 7.10 where the jets are shown after all JES corrections are applied. This overcorrection of quark jets inevitably leads to a shift in the reconstructed W boson mass compared to the world average value.

Determining the calibration factor α

To derive a global calibration from the distributions in Figure 7.10, the dijet mass distribution (M_{jj}) will be fitted both for data and for simulation using a gauss function convoluted with a function for the combinatorial background. The ratio of both expectation values of the gaussian function will be used as a constant calibration factor to all jets in all events before the event selection criteria are applied. Consequently, the correction factor α is defined as:

$$\alpha = \frac{m_W^{fit,mc}}{m_W^{fit,data}}. \tag{7.6}$$

To fit the distribution of M_{jj} both in data and simulation, a simple gaussian function will not yield accurate results due to the large combinatorial background. For this reason a more advanced function is used by convoluting a gauss function (signal) with a crystal ball function (background). This fit function is defined in as:

$$f(x;\mu_1,\sigma_1,\alpha,n,\mu_2,\sigma_2) = \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2} \otimes N \cdot \begin{cases} \exp(-\frac{(x-\mu_2)^2}{2\sigma_2^2}), & \text{for } \frac{x-\mu_2}{\sigma_2} > -\alpha \\ A \cdot (B - \frac{x-\mu_2}{\sigma_2})^{-n}, & \text{for } \frac{x-\mu_2}{\sigma_2} \leqslant -\alpha \end{cases},$$

where

$$A = \left(\frac{n}{|\alpha|}\right)^n \cdot \exp\left(-\frac{|\alpha|^2}{2}\right),$$

$$B = \frac{n}{|\alpha|} - |\alpha|,$$

$$N = \frac{1}{\sigma(C+D)},$$

$$C = \frac{n}{|\alpha|} \cdot \frac{1}{n-1} \cdot \exp\left(-\frac{|\alpha|^2}{2}\right),$$

$$D = \sqrt{\frac{\pi}{2}}\left(1 + \operatorname{erf}\left(\frac{|\alpha|}{\sqrt{2}}\right)\right).$$

The parameters μ_1, σ_1, α, n, μ_2, σ_2 are allowed to float in the fit while the crystal ball normalisation factor N has to be fixed. The latter factor is set to a value providing the best fit probability for the overall fit.

The fit to the reconstructed di-jet mass distributions is shown in Figure 7.11 and 7.12 for μ+jets events respectively at 7 and 8 TeV. The same distributions are shown in Figures 7.13 and 7.14 for e+jets events. Each of these plots shows the fit performed on data and simulation in the top row. This allows to determine the α calibration factor to be applied on the jets in the simulation. For each systematic variation of

the simulated sample, this procedure is repeated to determine the respective α factors. In each of these Figures, the fit is also shown on the reconstructed di-jet mass in the simulated sample where the Jet Energy Scale is varied within $\pm 1\sigma$.

Calibration results

In Table 7.10, the α factors are summarised for the nominal simulated event sample at 7 and 8 TeV in both decay channels. The calibration factors are found to be consistent between the muon and electron channels which is expected as the Jet Energy Scale a priori does not depend on the $t\bar{t}$ decay channel studied. Because of the in-situ character of this method, the calibration factors are determined and applied to each sample, including samples with systematic variations.

As an example, the calibration factors derived for the $\pm 1\sigma$ Jet Energy Scale variations are provided as well. By comparing the α factors for the nominal sample and the samples with a $\pm 1\sigma$ JES variation the potential of the method to constrain the Jet Energy Scale uncertainty is apparent. Thus, the four-vectors of the jets in the two JES varied samples and the nominal simulation sample are corrected with their respective α factors. This results in partial cancellation of the final Jet Energy Scale uncertainty. The residual JES uncertainty quoted in Tables 7.8 and 7.9 originates from the finite statistical precision of the α factors.

Furthermore, it can be noted that the sensitivity at 8 TeV is larger than at 7 TeV. This is due to the more limited statistics in the 7 TeV data compared to 8 TeV resulting in a M_{jj} mass shift that is less precisely determined and thus yields less variation with the Jet Energy Scale shifts.

Table 7.10: The α calibration factors obtained for the nominal simulated event sample as well as the same sample with a $\pm 1\sigma$ Jet Energy Scale variation applied

		Jet Energy Scale variation		
$\sqrt{s}$	Channel	-1σ	nominal	$+1\sigma$
7 TeV	μ+jets	1.010 ± 0.004	1.005 ± 0.005	1.000 ± 0.004
	e+jets	1.005 ± 0.005	1.002 ± 0.007	0.997 ± 0.005
8 TeV	μ+jets	1.036 ± 0.004	1.015 ± 0.005	0.997 ± 0.004
	e+jets	1.031 ± 0.004	1.012 ± 0.004	0.992 ± 0.004

Figure 7.15 shows the di-jet mass distribution after applying the α correction to the jets in the simulated sample. The small mass shift between data and simulation is now resolved as compared to the case before calibration in Figure 7.10.

7.6 Performing the l+jets combination

Up to now, the cross section has been measured in both the μ+jets and e+jets decay channels separately. The results have proven to be consistent among both channels and to increase the overall precision, a combined result in the l+jets channel is attempted.

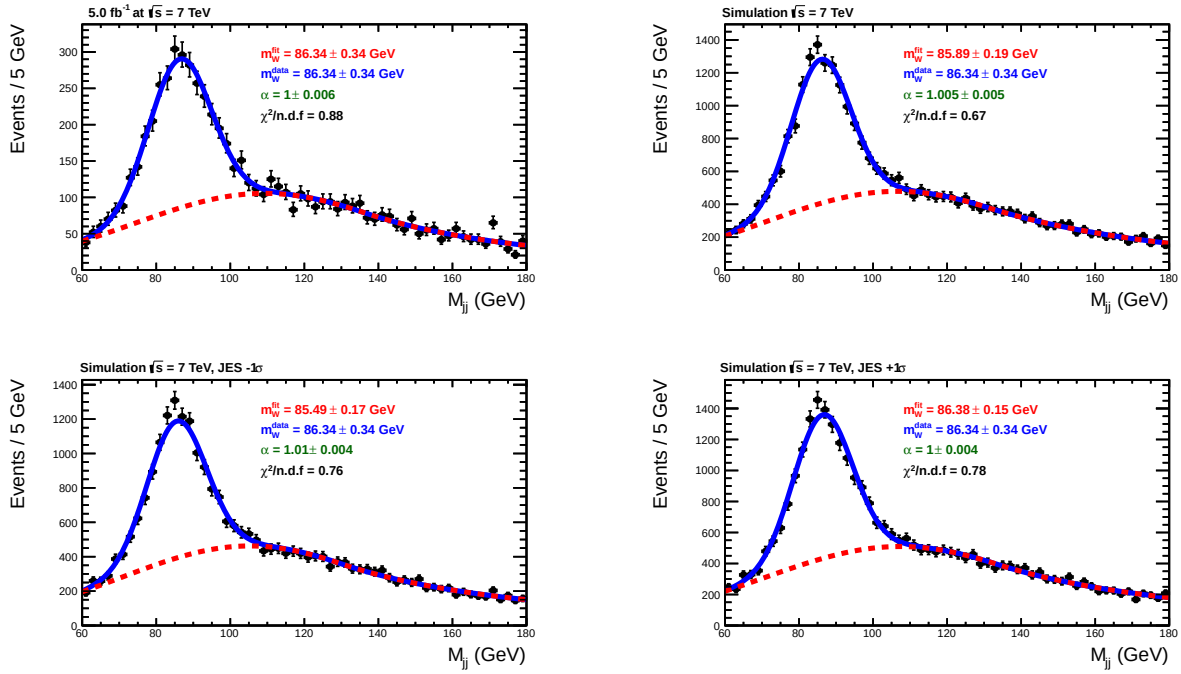


Figure 7.11: The reconstructed di-jet mass distribution in μ+jets events at $\sqrt{s}$ = 7 TeV overlaid with the fit function in data (top left) and the distribution in the nominal simulation sample (top right). The W boson mass estimate extracted from both distributions can be used to determine the α correction factor to be applied to the jets in the simulated sample. The bottom row shows the same procedure applied to the $\pm 1\sigma$ variation of the Jet Energy Scale in the same simulated event sample.

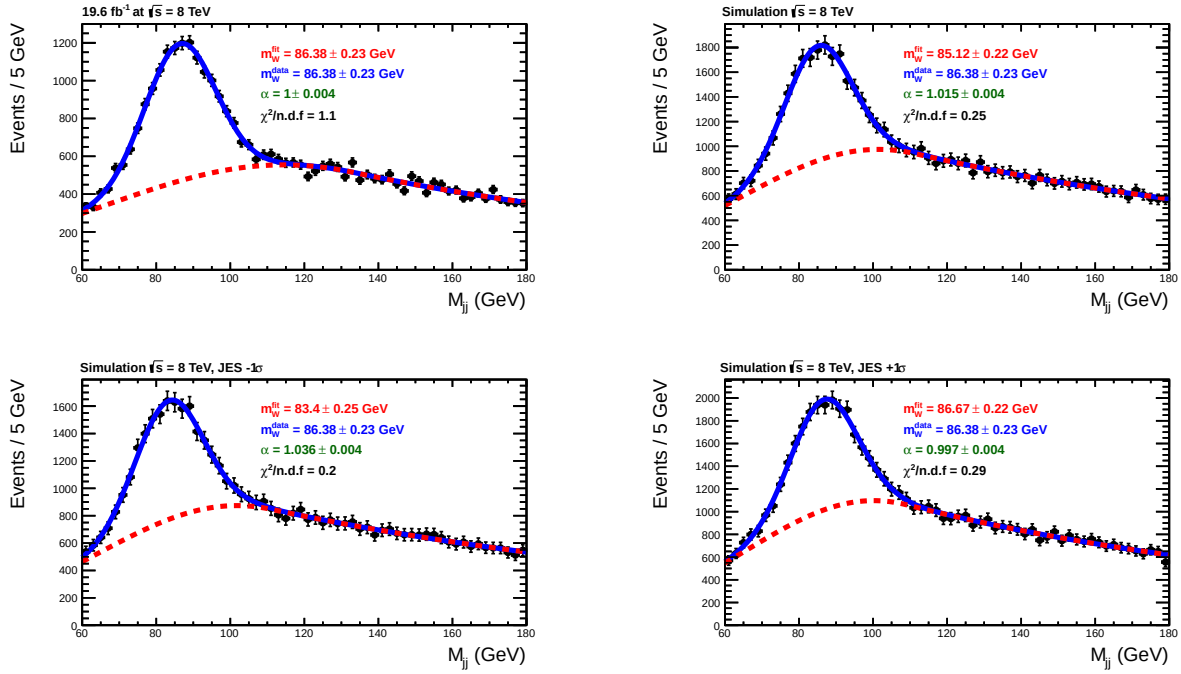


Figure 7.12: The reconstructed di-jet mass distribution in μ+jets events at $\sqrt{s}$ = 8 TeV overlaid with the fit function in data (top left) and the distribution in the nominal simulation sample (top right). The W boson mass estimate extracted from both distributions can be used to determine the α correction factor to be applied to the jets in the simulated sample. The bottom row shows the same procedure applied to the $\pm 1\sigma$ variation of the Jet Energy Scale in the same simulated event sample.

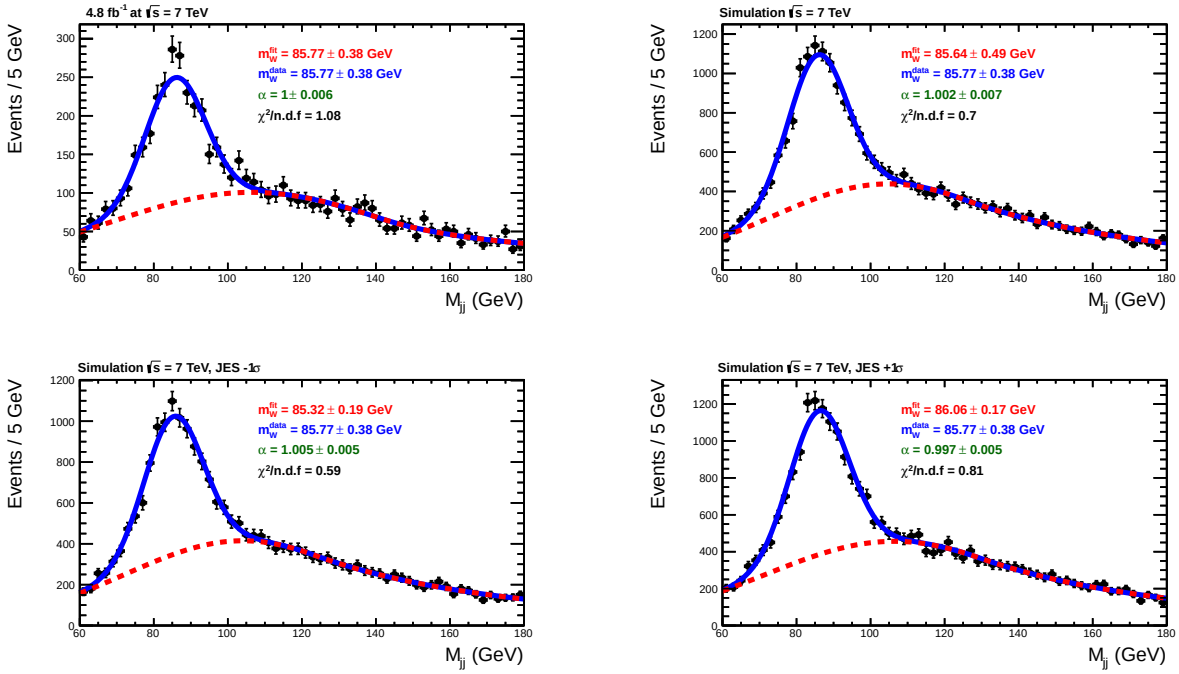


Figure 7.13: The reconstructed di-jet mass distribution in e+jets events at $\sqrt{s}$ = 7 TeV overlaid with the fit function in data (top left) and the distribution in the nominal simulation sample (top right). The W boson mass estimate extracted from both distributions can be used to determine the α correction factor to be applied to the jets in the simulated sample. The bottom row shows the same procedure applied to the $\pm 1\sigma$ variation of the Jet Energy Scale in the same simulated event sample.

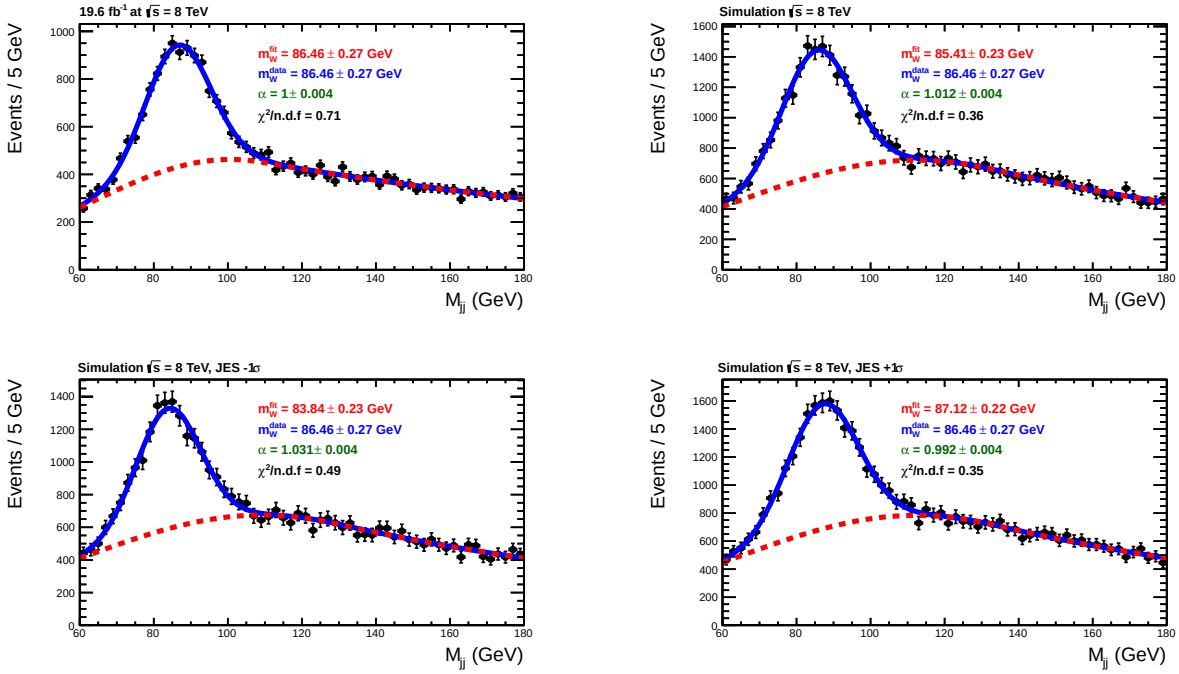


Figure 7.14: The reconstructed di-jet mass distribution in e+jets events at $\sqrt{s}$ = 8 TeV overlaid with the fit function in data (top left) and the distribution in the nominal simulation sample (top right). The W boson mass estimate extracted from both distributions can be used to determine the α correction factor to be applied to the jets in the simulated sample. The bottom row shows the same procedure applied to the $\pm 1\sigma$ variation of the Jet Energy Scale in the same simulated event sample.

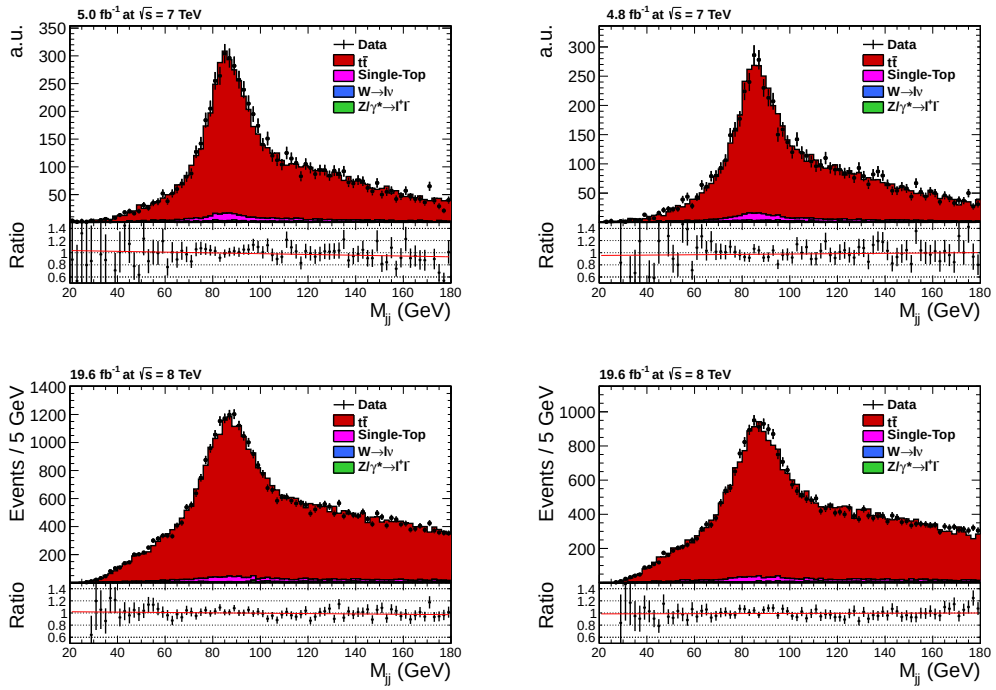


Figure 7.15: Distribution of the reconstructed W-boson mass for $t\bar{t}$ signal events and all relevant background processes in the μ+jets channel (left) and in the e+jets channel (right) after applying the measured global calibration. The small shift that was observed in both channels before the calibration has now disappeared and the agreement between data and simulation is improved

Since the method relies heavily on the jet-lepton mass, the events from both decay channels can not be joined into one sample to run the analysis. There are two main reasons for this.

First, the electron and muon kinematic cuts are not exactly the same. Since different identification and transverse momentum criteria are used at trigger level, the cuts offline are chosen accordingly. This results in a slightly different measured phase space and as such minor differences at the level of the mass shape.

Secondly, due to the difference in kinematic cuts, the background levels are not exactly the same. Also backgrounds with fake electrons are more predominant to backgrounds with fake muons. Hence, the background template should be split up by channel.

One option to measure the cross section in the combined l+jets channel is to perform a simultaneous maximum likelihood fit to both channels allowing the backgrounds to float differently for both channels but keeping the signal fraction equal. This method has the benefit of simplicity but does not properly take into account differences in systematic uncertainty between the two channels as it is a statistical fit procedure.

To properly take into account the various systematics in each channel and the

correlations of these uncertainties among channels, the Best Linear Unbiased Estimator (BLUE) [137, 138] is used. The benefit of this method is that it allows for combination of correlated results of one or more observables. The BLUE estimator is a weighted linear combination of the input measurements. Furthermore, it defines the estimates of each measurement such that the total uncertainty on the combination remains minimal while accounting for the statistical as well as systematic uncertainties and correlations.

7.6.1 The BLUE method

Although the BLUE method allows to combine n measurements for N observables, in this a simplified case is treated. Consider two measurements of the $t\bar{t}$ cross section, each in one decay channel (e.g. n=2, N=1), $\sigma_1 \pm \delta\sigma_1$ and $\sigma_2 \pm \delta\sigma_2$. Then an estimator for the cross section, $\hat{\sigma}$ can be written as:

$$\hat{\sigma} = \sum_{i=1}^{n} w_i \sigma_i = w_1\sigma_1 + w_2\sigma_2. \tag{7.7}$$

This estimator is a linear combination of the individual measurements using weight factors w_i such that $\sum_{i=1}^{n} w_i = 1$. The variance on $\hat{\sigma}$ then equals:

$$Var[\hat{\sigma}] = \sum_{i=1}^{n}\sum_{j=1}^{n} w_j M_{ij} w_j. \tag{7.8}$$

The matrix M in this equation is the covariance matrix of the measurements defined as

$$M = \begin{pmatrix} \delta\sigma_1^2 & \rho\delta\sigma_1\delta\sigma_2 \\ \rho\delta\sigma_1\delta\sigma_2 & \delta\sigma_2^2 \end{pmatrix},$$

where $\delta\sigma_i$ represents the uncertainty on measurement i and ρ is the correlation factor between the measurements. By construction of the matrix M, the BLUE method requires symmetric uncertainties for all σ_i. Hence the asymmetric uncertainties of the $t\bar{t}$ cross section measurements in the muon and electron channels will be made symmetric by averaging between the up and downwards systematic.

The BLUE method consists of deriving the weights such that the variance is minimal. The weights w_i can be derived by using Lagrange Multipliers:

$$w = M^{-1}U / U_T M^{-1} U, \tag{7.9}$$

where U is a 1x2 matrix where every element equals 1. This yields a solution for w_2=1-w_1:

$$w_2 = \frac{1-\rho z}{1+z^2-2\rho z} \quad \left(z = \frac{\sigma_2}{\sigma_1}\right). \tag{7.10}$$

This method is in fact equivalent to minimising a weighted sum of squares S,

$$S = \sum_{i=1}^{n}\sum_{j=1}^{n} (\hat{\sigma} - \sigma_i)(\hat{\sigma} - \sigma_j) M_{ij}^{-1}, \tag{7.11}$$

which measures the consistency of the individual σ_i to $\hat{\sigma}$. This equation is particularly useful as it is χ^2 distributed with (n-N) degrees of freedom which allows to define a χ^2 value for the BLUE combination characterising the consistency of the fit.

Another interesting property of BLUE is that it allows to easily break down the covariance matrix M into different components. This is for example useful for measurements where different systematic uncertainties have been identified which is the case for this analysis. Then the covariance matrix M can be regarded as a sum:

$$M_{ij} = \sum_{i=1}^{U} M_{ij}^{[U]}, \tag{7.12}$$

where U denotes the number of individual uncertainties. Using eq. 7.8, the contribution from uncertainty q can then be immediately determined using

$$Var^{[q]}[\hat{\sigma}] = \sum_{i=1}^{n}\sum_{j=1}^{n} w_j M_{ij}^{[q]} w_j. \tag{7.13}$$

7.6.2 Results for the $\sigma_{\mathrm{t\bar{t}}}$ combination

Using the BLUE method, the l+jets visible $t\bar{t}$ cross section can be obtained by combining the individual measurements in the muon and electron channels. The results from Section 7.3 are taken as input together with their respective uncertainties displayed in Table 7.8. All systematics are considered to be 100% correlated between the measurement in both decay channels except for the lepton trigger, identification and isolation systematics where the correlation is set to be 0%.

At 7 TeV, the combination yields the following visible cross section

$$\sigma_{t\bar{t}}^{vis} = 2.590 \pm 0.069(stat.) \pm 0.168(syst.) \pm 0.057(lumi.)pb$$

At 8 TeV the visible cross section is as follows:

$$\sigma_{t\bar{t}}^{vis} = 3.888 \pm 0.062(stat.) \pm 0.163(syst.) \pm 0.101(lumi.)pb$$

Table 7.11: Properties of the combined total cross section fit with BLUE. The χ^2 value of the fit is provided as well as the contribution of each channel.

	$\sqrt{s} = 7$TeV	$\sqrt{s} = 8$TeV
$\chi^2/n.d.f$	0.556	0.743
$w_{\mu+jets}$	0.485	1.069
w_{e+jets}	0.515	-0.069

The weights for the individual measurements and the χ^2 value of the fit are provided in Table 7.11. At 8 TeV, BLUE attributes a negative weight to the electron channel. This can be understood from eq. 7.10 where $w_2(= w_{e+jets})$ becomes negative when the correlation between the measurements (ρ) becomes larger than the ratio of their total

uncertainties $\sigma^{\mu+jets}/\sigma^{e+jets}$. Thus, given a total uncertainty of 6.2% and 6.8% in the muon and electron channels respectively, a correlation higher than 91% would yield a negative weight for the electron channel measurement. The correlation of the two measurements can be calculated as:

$$\rho = \frac{\sum_{i=1}^{P} \rho_i \sigma_i^{\mu+jets} \sigma_i^{e+jets}}{\sigma^{\mu+jets}\sigma^{e+jets}}, \tag{7.14}$$

where the sum runs over the P individual uncertainty sources and ρ_i are the pre-defined correlations factors for uncertainty source i. The correlation between the muon and electron channel measurements is found to be as high as 91.2% since most uncertainty sources are considered correlated. Hence, the electron channel measurement gets a negative weight. This, however, does not mean that the latter result is ignored as it still helps to reduce the variance on the combined result.

At 7TeV, the situation is different since here the ratio $\sigma^{\mu+jets}/\sigma^{e+jets}$ is 1.06 due to a slightly better total systematic uncertainty in the electron channel yielding a total uncertainty on the cross section of 8.0% compared to 8.4% for the muon channel measurement. The correlation between the two measurements is found to be 0.66 using the same definition as was used at 8 TeV. Hence, a positive weight is expected for the electron channel result and due to the lower correlation compared to 8 TeV, both measurements get a more comparable weight.

Using the acceptance defined in Section 7.2.1, the combined l+jets $t\bar{t}$ cross section becomes:

$$\sigma_{t\bar{t}} = 163.9 \pm 4.4(stat.) \pm 10.7(syst.) \pm 3.6(lumi.)pb$$

at 7 TeV and

$$\sigma_{t\bar{t}} = 234.2 \pm 3.8(stat.) \pm 9.6(syst.) \pm 6.1(lumi.)pb$$

at 8 TeV.

Finally, the systematics for the combined fit result are provided in Table 7.12 for the 7 TeV combination and in Table 7.13 for the 8 TeV combination.

7.7 Cross section ratio between different LHC beam energies

The measurement of the ratio of the $t\bar{t}$ cross section between two different LHC beam energies is interesting for two main reasons. First the ratio is very sensitive to precision SM predictions. Because some of the systematic uncertainties overlap at different energies, both at experimental and theoretical level, the total uncertainty on the ratio can be brought below the uncertainty on the individual cross section measurements. Hence the ratio provides a unique handle to benchmark theory predictions to data.

Furthermore, the cross section ratio is also interesting because of the potential to enhance the BSM physics sensitivity to the absolute cross sections. Since this analysis is conducted both at 7 TeV and 8 TeV, it provides a unique opportunity to measure the

Table 7.12: Overview of the systematic uncertainties on the cross section measurement $\sqrt{s} = 7$ TeV. The combination is performed using BLUE. All correlation factors are set to 1 except for the Lepton ID and Isolation SF's.

	μ+jets	e+jets	Combined fit
Systematic	$\delta\sigma_{t\bar{t}}$ (%)	$\delta\sigma_{t\bar{t}}$ (%)	$\delta\sigma_{t\bar{t}}$ (%)
Jet Energy Scale	+4.9 –4.7	+4.3 –6.2	± 5.0
Jet Energy Resolution	+2.3 –0.6	+0.2 –2.0	± 1.2
MET Unclustered Energy	+0.0 –0.0	+0.2 –0.4	± 0.2
Pileup	+0.4 –0.3	+0.6 –0.7	± 0.5
Lepton ID/Trigger SF	+1.5 –1.4	+1.7 –1.8	± 0.8
Btag Method Settings	+0.0 –0.9	+1.0 –0.3	± 0.6
Background composition	+0.4 –0.5	+0.5 –0.3	± 0.4
Factorisation Scale	+3.5 –4.0	+0.0 –0.7	± 2.0
ME-PS Matching threshold	+0.0 –4.0	+0.6 –2.8	± 1.8
Top Quark Mass	+0.1 –2.1	+0.4 –2.0	± 1.1
PDF Uncertainties	+2.2 –2.5	+1.8 –2.1	± 2.1
Total	+7.0 -8.3	+5.2 -8.0	± 6.5
Luminosity	+2.2 -2.2	+2.2 -2.2	± 2.2

Table 7.13: Overview of the systematic uncertainties on the cross section measurement at $\sqrt{s} = 8$ TeV. The combination is performed using BLUE. All correlation factors are set to 1 except for the Lepton ID and Isolation SF's.

	μ+jets	e+jets	Combined fit
Systematic	$\delta\sigma_{t\bar{t}}$ (%)	$\delta\sigma_{t\bar{t}}$ (%)	$\delta\sigma_{t\bar{t}}$ (%)
Jet Energy Scale	+2.4 –2.3	+2.0 –2.2	± 2.4
Jet Energy Resolution	+0.7 –1.0	+0.2 –1.6	± 0.8
MET Unclustered Energy	+0.2 –0.0	+0.6 –0.1	± 0.1
Pileup	+0.4 –0.6	+0.4 –0.3	± 0.5
Lepton ID/Trigger SF	+0.6 –0.2	+0.5 –0.6	± 0.5
Btag Method Settings	+0.1 –0.5	+0.6 –0.8	± 0.3
Background composition	+0.2 –0.2	+0.4 –0.3	± 0.2
Factorisation Scale	+0.5 –3.0	+1.7 –3.6	± 1.7
ME-PS Matching threshold	+0.9 –1.7	+1.9 –2.8	± 1.2
Top Quark Mass	+0.1 –2.5	+0.4 –2.6	± 1.3
PDF Uncertainties	+1.9 –2.3	+1.7 –2.1	± 2.1
Total	+3.4 -5.5	+3.9 -6.3	± 4.1
Luminosity	+2.6 -2.6	+2.6 -2.6	± 2.6

ratio as well since the method used at both energies is identical. As a consequence, the method can be run simultaneously on both centre-of-mass energies varying systematics up and down by 1σ at the same time. This entails that the systematics are assumed fully correlated between the two energies with the exception of the luminosity uncertainty which is considered uncorrelated.

The ratio is calculated by dividing the cross section in each $t\bar{t}$ decay channel separately. For each systematic variation, the ratio is re-calculated and the absolute shift to the nominal ratio value is taken as the systematic uncertainty. Using this method at each systematic variation yields the total systematic uncertainty. The ratio measurement yields

$$R_{\sigma_{t\bar{t}}}^{8/7TeV} = 1.469 \pm 0.056(stat.)^{+0.065}_{-0.056}(syst.)^{+0.051}_{-0.051}(lumi.)$$

in the μ+jets channel and

$$R_{\sigma_{t\bar{t}}}^{8/7TeV} = 1.432 \pm 0.061(stat.)^{+0.084}_{-0.042}(syst.)^{+0.05}_{-0.05}(lumi.)$$

in the e+jets channel.
Just as for the individual cross section measurements, the combination of both decay channels is performed using the BLUE technique where the systematic uncertainties are considered fully correlated between the two decay channels except for the specific lepton systematics. The combined l+jets channel cross section ratio between 8 and 7 TeV is found to be

$$R_{\sigma_{t\bar{t}}}^{8/7TeV} = 1.453 \pm 0.041(stat.) \pm 0.057(syst.) \pm 0.051(lumi.)$$

The systematics on the combined ratio fit are outlined in Table 7.15. The fit probability and individual weights of both channels in the combination are provided in Table 7.14. If the systematics among the two beam energies are correlated, they are believed to reduce in the direct ratio of the two measured cross sections. Conversely, if they are anti-correlated they can enhance the systematic uncertainty. Most systematic uncertainty sources partially cancel bringing the total uncertainty down to 3.9%, luminosity not included. Hence, the cross section ratio provides a better precision as the individual cross section measurements if all uncertainties are correlated.

Table 7.14: Properties of the combined total cross section fit with BLUE. The χ^2 value of the fit is provided as well as the contribution of each channel.

$\chi^2/n.d.f$	0.168
$w_{\mu+jets}$	0.564
w_{e+jets}	0.436

7.8 Summary of the results

In this chapter, a method is outlined to measure the top quark pair production cross section with application of b-tagging in the event selection. The number of $t\bar{t}$ pairs is

Table 7.15: Overview of the systematic uncertainties on the cross section ratio measurement. The combination is performed using BLUE. All correlation factors are set to 1 except for the Lepton ID and Isolation SF's.

	μ+jets	e+jets	Combined fit
Systematic	$\delta R_{\sigma_{t\bar{t}}}^{8/7TeV}$ (%)	$\delta R_{\sigma_{t\bar{t}}}^{8/7TeV}$ (%)	$\delta R_{\sigma_{t\bar{t}}}^{8/7TeV}$ (%)
Jet Energy Scale	+2.5 –2.3	+4.3 –2.2	± 2.8
Jet Energy Resolution	+1.6 –0.4	+0.0 –0.4	± 0.6
MET Unclustered Energy	+0.2 –0.0	+0.4 –0.3	± 0.2
Pileup	+0.0 –0.2	+0.4 –0.3	± 0.2
Lepton ID/Trigger SF	+1.2 –0.9	+1.2 –1.2	± 0.8
Btag Method Settings	+0.0 –0.5	+0.0 –0.5	± 0.2
Background composition	+0.3 –0.2	+0.0 –0.1	± 0.2
Factorisation Scale	+2.9 –1.1	+3.6 –0.9	± 2.1
ME-PS Matching threshold	+0.9 –2.5	+1.3 –0.0	± 1.2
Top Quark Mass	+0.0 –0.5	+0.0 –0.8	± 0.3
PDF Uncertainties	+0.3 –0.3	+0.3 –0.3	± 0.3
Total	+4.4 -3.8	+5.9 -2.9	± 3.9
Luminosity	+3.5 -3.5	+3.5 -3.5	± 3.5

determined from data using a binned maximum likelihood fit on the jet-lepton mass distribution. The selection efficiency needed to propagate the number of top quark pairs to a cross section was factorized into its components and for each component it was described how it is determined. Finally, the b-tagging component of the selection efficiency is determined from data using the techniques outlined in Chapters 5 and 6.

In addition, the cross section ratio between the LHC beam energies of 7 and 8 TeV was measured. This measurement has the benefit of having systematics cancel between the two individual measurements.

To summarize, the results for these three quantities are provided while comparing them to theoretical expectation. Finally, Figure 7.16 compares the cross section dependence on the centre-of-mass energy between the measurements and theoretical predictions.

7.8.1 Results at 7 TeV

The top quark pair production cross section at 7 TeV has been calculated up to NNLO+NNLL accuracy [22, 23]. Assuming a top quark mass of 173.3 GeV, the theoretical expectation for the cross section equals:

$$\sigma_{t\bar{t}}^{theory} = 172.0^{+4.4}_{-5.8}(scale) \pm 4.7(PDF) \pm 2.7(\alpha_s)^{+5.3}_{-5.1}(m_t)pb.$$

This cross section is measured in this thesis using μ+jets and e+jets $t\bar{t}$ events resulting respectively in

$$\sigma_{t\bar{t}} = 159.7 \pm 5.6(stat.)^{+11.2}_{-13.3}(syst.) \pm 3.5(lumi.)pb$$

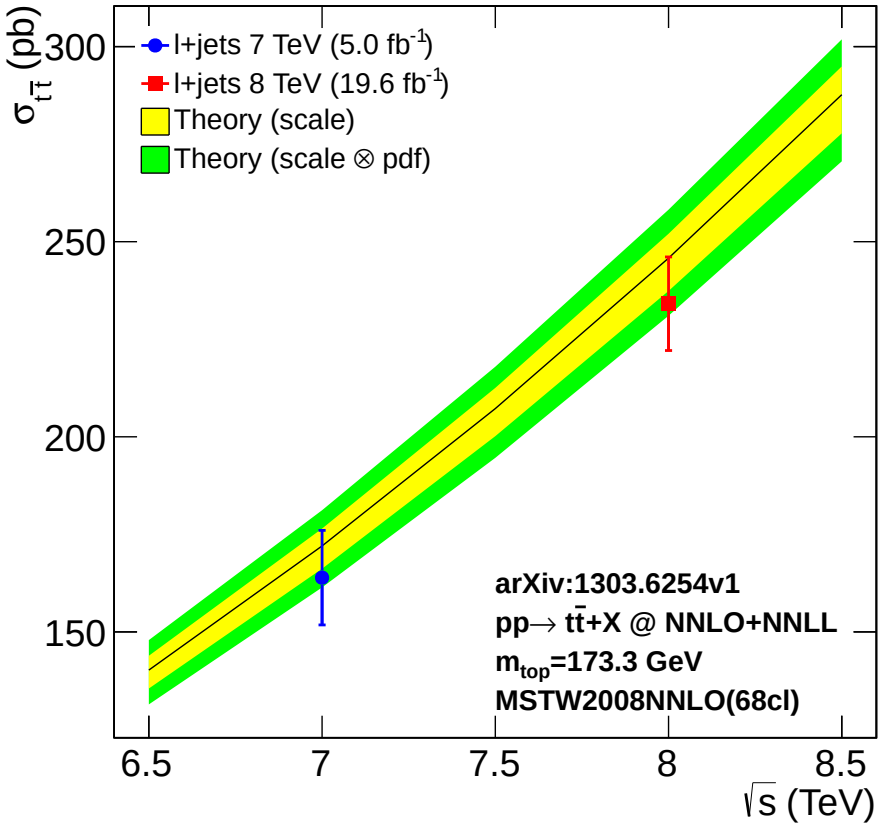


Figure 7.16: The $t\bar{t}$ pair production cross section in the combined l+jets channel as a function of the centre-of-mass energy. The Measured results are compared to the theoretical calculations at NNLO+NNLL[22, 23]. The measurements agree well with the theory expectation within experimental and theoretical uncertainties.

and

$$\sigma_{t\bar{t}} = 167.9 \pm 6.6(stat.)^{+8.7}_{-13.4}(syst.) \pm 3.7(lumi.)pb$$

Finally, a combination is performed between both channels using the BLUE method that takes into account all systematic uncertainty sources and their correlation. This yields the measured cross section:

$$\sigma_{t\bar{t}} = 163.9 \pm 4.4(stat.) \pm 10.7(syst.) \pm 3.6(lumi.)pb$$

The measurements in the individual channels as well as in the combined channel all agree well with the theoretical expectation given that the measurements were carried out assuming a top quark mass of 172.5 GeV.

7.8.2 Results at 8 TeV

Since the $t\bar{t}$ cross section depends on the collision energy, the measurement is repeated at a collision energy of 8 TeV. Again the measurement is performed in the muon channel providing a $t\bar{t}$ cross section of:

$$\sigma_{t\bar{t}} = 234.6 \pm 3.5(stat.)^{+8.0}_{-12.9}(syst.) \pm 6.1(lumi.)pb$$

Furthermore, in the electron channel, the measured cross section equals:

$$\sigma_{t\bar{t}} = 240.4 \pm 4(stat.)^{+9.4}_{-15.1}(syst.) \pm 6.3(lumi.)pb$$

To conclude, the cross section is combined among both channels to yield the final result of

$$\sigma_{t\bar{t}} = 234.2 \pm 3.8(stat.) \pm 9.6(syst.) \pm 6.1(lumi.)pb$$

which can be compared to the NNLO+NNLL calculation [22, 23] assuming a top mass of 173.3 GeV

$$\sigma^{theory}_{t\bar{t}} = 245.8^{+6.2}_{-8.4}(scale) \pm 6.2(PDF) \pm 4.0(\alpha_s)^{+7.4}_{-7.1}(m_t)pb$$

It can be observed that, although the measurements are performed under a different assumption for the top quark mass, they show good agreement with the theoretical expectation.

7.8.3 Results for $R^{8/7TeV}_{\sigma_{t\bar{t}}}$

The $t\bar{t}$ cross section is measured using the same technique at 7 and 8 TeV. Hence, the ratio between both cross section measurements can be measured allowing for (partial) cancellation of various sources of systematic uncertainties. Not only the experimental systematics cancel, also the theoretical precision improved for the cross section ratio compared to the individual cross sections. The expected cross section ratio has been calculated as well at NNLO+NNLL [23]:

$$R^{8/7TeV,theory}_{\sigma_{t\bar{t}}} = 1.429^{+0.001}_{-0.001}(scale) \pm 0.004(PDF) \pm 0.001(\alpha_s) \pm 0.001(m_t)$$

The cross section is measured assuming the systematic uncertainties to be fully correlated except for the luminosity. In the μ+jets channel the cross section ratio equals:

$$R_{\sigma_{t\bar{t}}}^{8/7TeV} = 1.469 \pm 0.056(stat.)^{+0.065}_{-0.056}(syst.)^{+0.051}_{-0.051}(lumi.)$$

and in the e+jets channel:

$$R_{\sigma_{t\bar{t}}}^{8/7TeV} = 1.432 \pm 0.061(stat.)^{+0.084}_{-0.042}(syst.)^{+0.05}_{-0.05}(lumi.)$$

Both measurements can then be combined with the BLUE method just as was done for the cross section measurements themselves. The combined fit yields

$$R_{\sigma_{t\bar{t}}}^{8/7TeV} = 1.453 \pm 0.041(stat.) \pm 0.057(syst.) \pm 0.051(lumi.)$$

which is in good agreement with theory prediction.

7.9 Results for other b-tagging algorithms

In the cross section measurement, the Combined Secondary Vertex algorithm at medium working point (CSVM) was chosen to reduce the background. It was shown in the previous chapters that the method to estimate the b-tagging performance is valid for all available algorithms at loose, medium and tight working points. Subsequently, for each combination of algorithm and working point, the $t\bar{t}$ cross section is measured. Moreover, for these working points equal at 7 and 8 TeV collisions, the cross section ratio is measured as well.

The $t\bar{t}$ cross section is not expected to depend on the implementation of b-tagging nor should it depend on the working point. Thus, the measurement of the cross section and the cross section ratio as a function of the b-tagging working point forms a unique cross check of the methods stability.

Figure 7.17 shows the cross section measured in the combined l+jets channel and it is clear that the measured cross section is consistent among the different b-tagging working points and algorithms. This figure also shows the agreement between the measurement and theory. Moreover, the stability of the $t\bar{t}$ cross section with respect to the different b-tagging working points illustrates that the b-tagging uncertainty on $\sigma_{t\bar{t}}$ is small compared to the other systematics.

Secondly, Figure 7.18 shows the same agreement between measurement and theory but now for the cross section ratio. Here, the ratio is stable as well compared to the chosen b-tagging working point and algorithm.

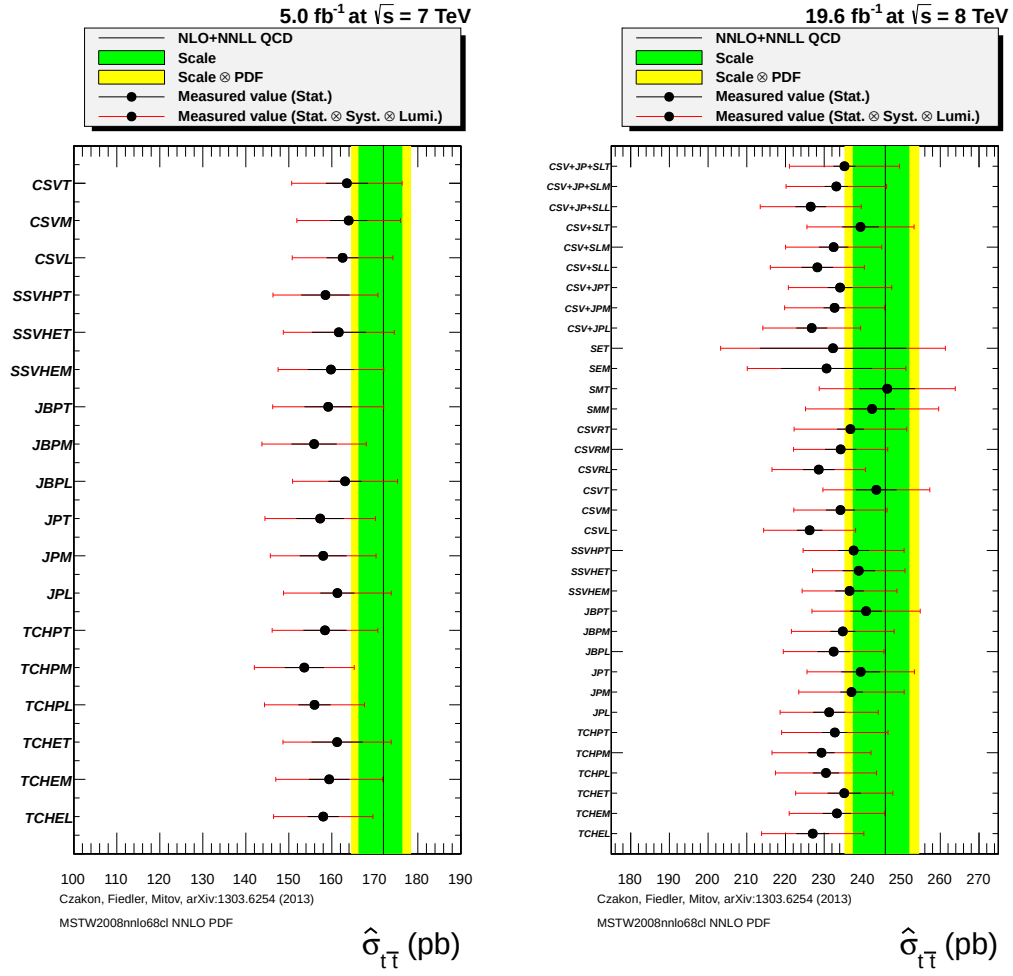


Figure 7.17: The total $t\bar{t}$ pair production cross section in the combined l+jets channel at a centre-of-mass energy of 7 TeV (left) and 8 TeV (right). The measured results are compared to NNLO+NNLL calculations [22, 23]. At both energies the results agree very well within the experimental uncertainties and the uncertainties on the theoretical calculations. The results are obtained for all available b-tagging algorithms and all available Working Points. All obtained cross section values are consistent among each other.

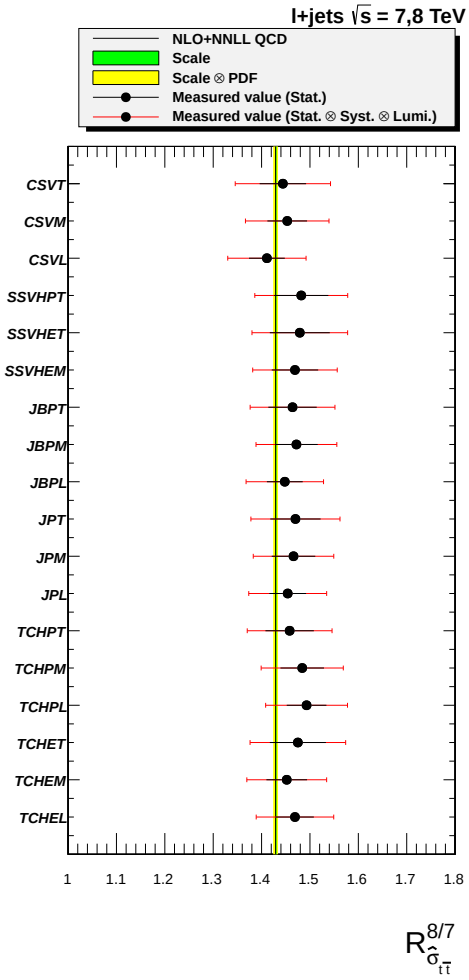


Figure 7.18: The $t\bar{t}$ pair production cross section ratio between 7 and 8 TeV in the combined l+jets channel. The measured results are compared to NNLO+NNLL calculations [22, 23]. The cross section ratio measurement for all b-tagging working points an algorithms equal at 7 and 8 TeV is provided. All obtained cross ratio section values are consistent among each other and with theoretical expectation.

Chapter 8

Conclusions and Perspectives

At the time of writing this thesis, the LHC had just finished its first run lasting from 2009 to early 2013. It was certainly a successful run showing the immense potential of the world's most energetic particle collider to provide stable proton-proton collisions at an enormous rate. The CMS experiment recorded about $5\,\mathrm{fb}^{-1}$ of collisions at a centre-of-mass energy of 7 TeV and no less than $20\,\mathrm{fb}^{-1}$ at 8 TeV providing unique large event samples to unravel the dynamics of the Standard Model. Undoubtedly, the LHC RunI will be mostly remembered for the discovery of a new heavy spin-0 particle compatible with the long-sought Standard Model scalar predicted by the Brout-Englert-Higgs mechanism. This remarkable discovery will certainly put its stamp on the future of the LHC and particle physics. On the other hand, this period should also be remembered as a new era of high precision measurements at the energy frontier, more precisely in the top quark sector.

Before the observation of the top quark at CMS in 2010, the top quark research was solely centred around the Tevatron $\mathrm{p\bar{p}}$ collider experiments CDF and DØ. With the large data samples produced at 7 and 8 TeV, the LHC became a true top quark factory allowing the top quark research field to rapidly develop. In the years between 2010 and 2013, many analyses targeted its basic properties like mass, charge and its production cross section with ever increasing precision. Given the scala of analysis techniques currently available, the increased luminosity and energy foreseen for the Run2 starting in 2015 this will certainly bring even more exciting results in the top quark sector.

8.1 Measurement of the b-tagging performance

Thanks to the huge sample of top quark events collected both at 7 and 8 TeV, the properties of the top quark become known with ever increasing precision. A better description of the top quark production and properties in turn allows physicists to use top quarks as calibration tools. One of the domains where top quarks nowadays are used in calibration is b-tagging.

The application of b-tagging has become a mainstream tool in CMS analyses to separate signal events with b quarks in the final state from abundant backgrounds. In the special case of top quark production, a b-tagging criterion can seriously decrease

the influence of the QCD multijet and W+jets backgrounds. Hence, the measurement of the b-tagging performance for different algorithms has become a crucial part in the physics programme. Top quark production is almost perfect for this purpose. The top quark decays almost exclusively into a W boson and a b quark creating a vast sample of b quarks to study.

Chapter 5 introduces a method to measure the efficiency to identify a true b quark jet for any given working point from any given algorithm. The method consists in reconstructing the b-tagging discriminator distribution for true b jets from data in a completely data-driven way. In the jet sample constructed by the leptonic side b jet candidate, a b-enriched and b-depleted sample was defined, based on the jet-lepton mass, allowing to model the non-b jet discriminator shape from the latter to remove the non-b contamination from the former.

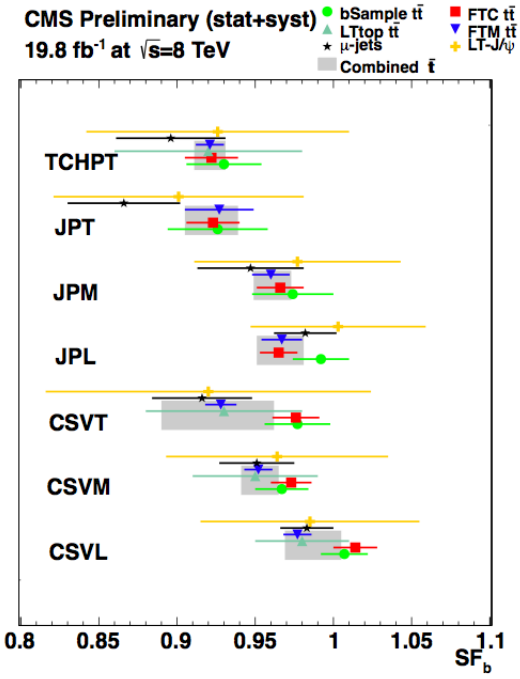


Figure 8.1: Comparison of the b-tagging efficiency data to simulation scale-factor (SF_b) for various techniques applied on $t\bar{t}$ and QCD events [3]. The method labeled as "bSample" is the method presented in this thesis and yields compatible results with respect to the other analysis techniques.

This technique differs from other $t\bar{t}$-driven measurements in the CMS collaboration by the fact that it is completely data-driven. Figure 8.1 shows the 8 TeV data to simulation scale factors for the b-tagging efficiency for commonly used working points. The method constructed in this thesis, labeled "bSample", is compared to other techniques and is found to produce compatible results albeit with a slightly larger systematic uncertainty than some other $t\bar{t}$ methods. In μ+jets events, the efficiency is found to be

$$\epsilon_b = 67.4 \pm 0.8(stat.) \pm 1.0(syst.)\%, \tag{8.1}$$

for the medium Working Point of the Combined Secondary Vertex algorithm. A com-

patible measurement in e+jets events

$$\epsilon_b = 67.5 \pm 0.9(stat.) \pm 1.3(syst.)\% \tag{8.2}$$

has been performed as well.

The larger total uncertainty originates from two aspects. First, due to the definition of subsamples and the subtraction method, the method depends on large statistics and hence produces a sizeable statistical uncertainty. Moreover, the method suffered from an intrinsic method bias that was found to depend on the definition of the b-enriched and b-depleted subsamples. An optimal definition was found by using simulated events at the cost of a leading systematic uncertainty on this definition. The higher centre-of-mass energy RunII of the LHC will generate an even larger $t\bar{t}$ sample which will greatly benefit this method and allow for further constraining of the bias.

To cure these biases, simulation based corrections could be devised. This approach was now avoided at all cost to be independent of signal modelling and maybe large theory uncertainties. With an even larger $t\bar{t}$ event sample at a centre-of-mass energy anywhere around 13 or 14 TeV, it is thinkable that theory uncertainties will become more and more constrained.

Given the large expected data sample during the next run, a natural extension of this method is to measure the b-tagging efficiency as a function of different observables. Such extension has not yet been successfully implemented in the analysis before because the measurement is very sensitive to different kinematic reweighting procedures deployed in the method. This sensitivity to the event kinematics makes it particularly difficult to perform the measurement binned in p_T^{jet} or η^{jet} resulting in very large biases.

Next to the b-tagging efficiency, the b-tagging performance is characterised by the mis-tagging rate, or the efficiency of tagging a non-b jet as a b jet. In Chapter 6, an extension of the previous method has been introduced to measure the inclusive mis-tagging rate. The reconstructed b-tagging discriminator shape for true b jets can be used to reconstruct the distribution for its complement, the non-b jets. This in turn allows to measure the efficiency for non-b jets to be tagged for any algorithm at any working point.

Although the method to reconstruct the discriminator distribution for b jets is data-driven, the normalisation is intrinsically biased by the subtraction procedure. To then reconstruct the distribution for non-b jets, the normalisation for b-jets has to be taken from simulation. This however, introduces large systematic uncertainties that dilute the precision of the measurement.

The results for this method cannot be directly compared to other mis-tagging efficiency results as they quote udsg and charm mis-tagging separately. In this technique it is not possible to discriminate between udsg and charm such that only the inclusive mis-tagging efficiency is measured. Nevertheless, the latter serves its purpose in the cross section analyses to calculate the total efficiency of the b-tagging criterion (cfr. eq 7.4). The measurement has been applied in the muon channel, yielding for the Combined Secondary Vertex Medium Working point

$$\epsilon_{b\not{l}} = 3.6 \pm 0.3(stat.) \pm 27.2(syst.)\%, \tag{8.3}$$

which is confirmed by a measurement in the electron channel

$$\epsilon_{b\not{l}} = 3.3 \pm 0.4(stat.) \pm 23.7(syst.)\%. \tag{8.4}$$

The measurement of the mis-tagging efficiency is currently dominated by the systematic uncertainties due to the $t\bar{t}$ signal modelling. These uncertainties enter the otherwise data-driven approach since the normalisation of the b-tagging discriminator distribution for b jets has to be determined from simulation. Due to the small statistics in the systematically varied samples, large systematic uncertainties arise in this measurement. These uncertainties will probably decrease in the future with larger event samples and data-driven techniques to constrain the modelling uncertainties.

8.2 Measurement of the inclusive $t\bar{t}$ cross section

While the previous two analyses focus on the top quarks ability as a calibration tool, Chapter 7 introduces a measurement on one of the key properties to model the $t\bar{t}$ process, namely its production cross section.

The $t\bar{t}$ cross section is determined by using the jet-lepton mass variable as a discriminator between $t\bar{t}$ and background events. The sample is first purified by requiring the leptonic side b jet candidate to pass the medium b-tag working point of the Combined Secondary Vertex algorithm.

The number of observed $t\bar{t}$ events in data is then derived by performing a binned maximum likelihood fit on the jet-lepton mass distribution using a $t\bar{t}$ and inclusive background template from simulation. The resulting number of $t\bar{t}$ events then needs to be corrected for the overall selection efficiency partially derived from data and from simulation. Additionally, a correction for the b-tagging efficiency and mis-tagging efficiency is also required and the values determined within Chapters 5 and 6 can be used. The virtue of the in-situ calibration of the b-tagging and mis-tagging efficiencies in this measurement is that a systematic uncertainty for both is avoided on the final measurement.

Moreover, the Jet Energy Scale uncertainty is reduced by introducing a calibration technique based on the W boson mass.

The $t\bar{t}$ cross section is measured in μ+jets and e+jets events separately, both at 7 and 8 TeV. The results from the separate decay channels can be combined with the BLUE method to yield

$$\sigma_{t\bar{t}} = 163.9 \pm 4.4(stat.) \pm 10.7(syst.) \pm 3.6(lumi.)pb$$

in the combined l+jets channel at a centre-of-mass energy of 7 TeV with a total precision of 7.4% and

$$\sigma_{t\bar{t}} = 234.2 \pm 3.8(stat.) \pm 9.6(syst.) \pm 6.1(lumi.)pb$$

at 8 TeV with a total precision of 5.1%.

The cross section results obtained in this thesis are compared to the published results from CMS and ATLAS in Figure 8.2. Both at 7 and 8 TeV, the measurement from this thesis, highlighted in green, agrees perfectly with all other measurements and a high precision is reached.

The future prospects for the measurement at 7 and 8 TeV lie mainly in the reduction of systematic uncertainties as the method is firmly founded by now. In the 8 TeV analysis, the leading systematic is the uncertainty on the luminosity. This uncertainty

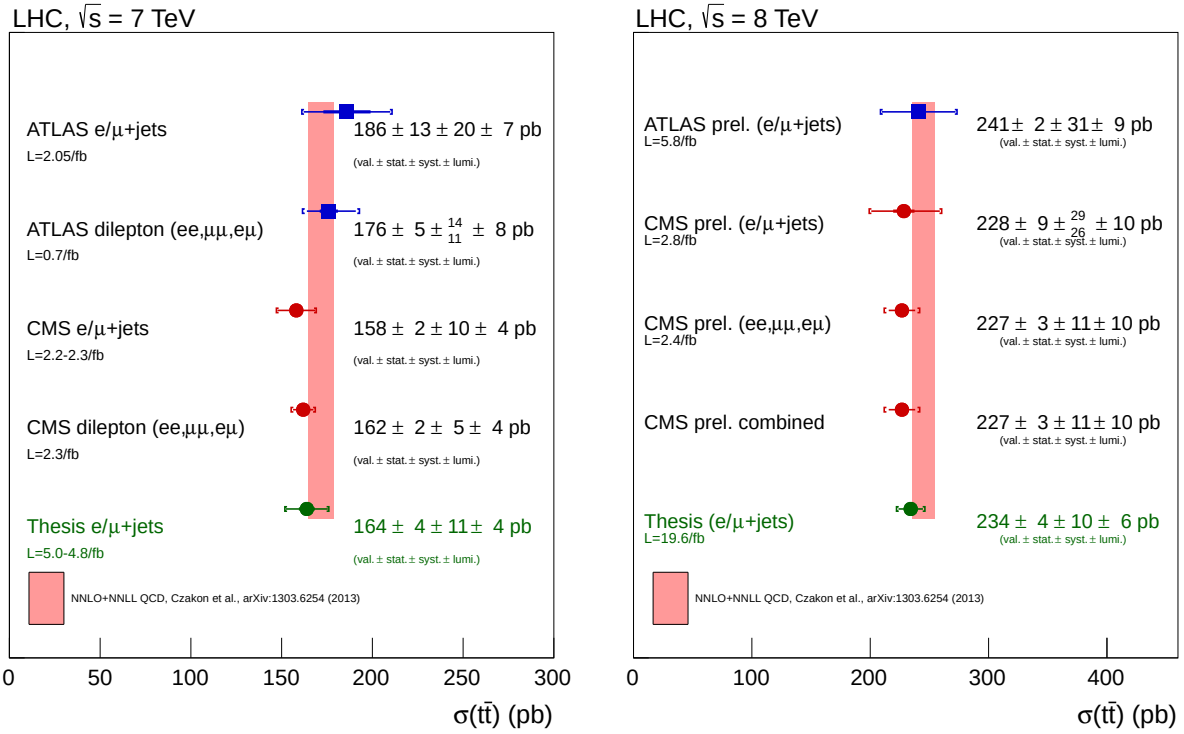


Figure 8.2: Measurement of the inclusive $t\bar{t}$ production cross section in the different decay channels at a centre-of-mass energy of 7 TeV [5, 25, 27, 28] (left) and 8 TeV [6, 26, 29] (right). The results obtained in this thesis are highlighted in green and show a very good agreement compared to the published measurements as well as to the NNLO+NNLL theoretical calculation [22, 23].

will go down as the pixel-detector based luminosity calibration will become available as is the case at 7 TeV.

To reach the level of precision exhibited in these results, a Jet Energy Scale calibration using the W boson mass is introduced in the cross section measurement to reduce the dominant Jet Energy Scale systematic. This calibration leads to a significant reduction in JES of roughly 40% at 7 TeV and over 60% at 8 TeV reducing the overall systematic uncertainty but a residual JES uncertainty remains. This uncertainty can in the future be further reduced by moving the measurement to a phase-space where the JES uncertainty is constant and small to begin with.

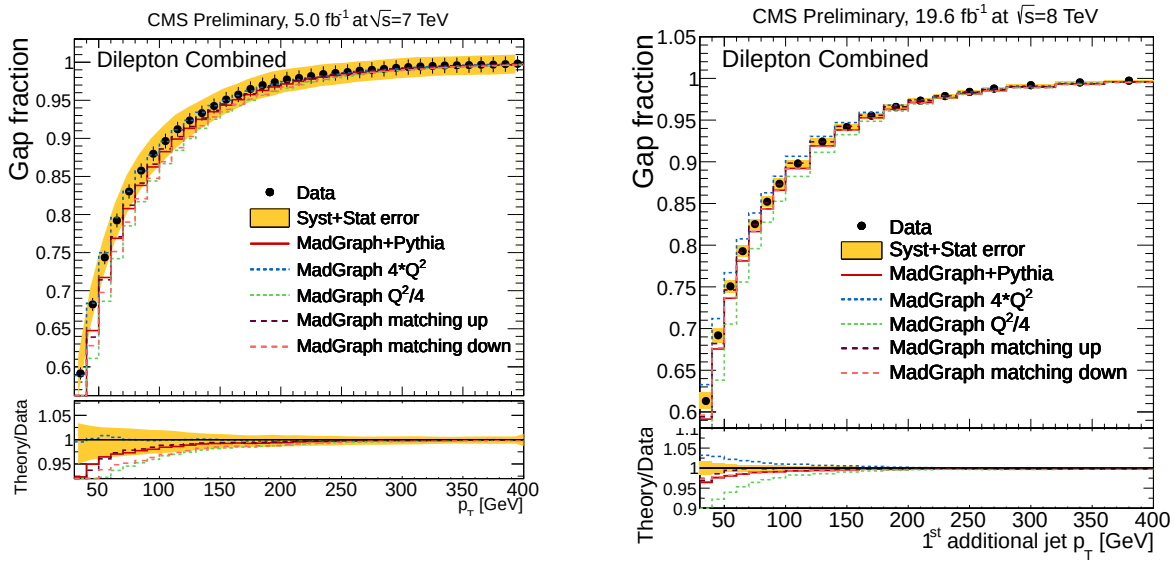


Figure 8.3: The measured Gap Fraction as a function of the additional jet p_T in dilepton $t\bar{t}$ events [98, 99]. The Gap Fraction is compared between data and different simulated samples where the factorisation (Q^2) and ME-PS matching (matching) scales have been altered.

The other important systematics mainly originate from theory modelling. The factorisation scale in particular can in the future be constrained from data by looking at the *Gap Fraction*, the fraction of events that is removed by requiring the presence of an additional jet. Figure 8.3 shows the gap fraction as a function of the p_T-threshold on the first additional jet. The data points are compared to different simulation sets where the factorisation scale and ME-PS matching scale have been altered. A strong sensitivity to the former can be observed showing a possibility to constrain this systematic uncertainty using data.

8.3 Measurement of the 8 to 7 TeV $t\bar{t}$ cross section ratio

An additional measurement presented in Chapter 7 is the ratio between the 8 TeV and 7 TeV $t\bar{t}$ cross section. This ratio is interesting since it is assumed that some

uncertainties will cancel if they are correlated between 7 and 8 TeV which allows even more precise benchmarks on theoretical predictions. The cross section ratio is measured in the combined l+jets channel and equals

$$R_{\sigma_{t\bar{t}}}^{8/7TeV} = 1.453 \pm 0.041(stat.) \pm 0.057(syst.) \pm 0.051(lumi.)$$

with a relative precision of 6.0% dominated by the systematic uncertainty. The systematic uncertainties between the 7 and 8 TeV result are all considered fully correlated which portrays an optimistic scenario. Therefore the total systematic uncertainty indeed improves with respect to the 7 and 8 TeV results respectively but to improve this measurement in the future, a thorough study of systematics correlations should be carried out.

The measured value can finally be compared to the prediction from theory:

$$R_{\sigma_{t\bar{t}}}^{8/7TeV,theory} = 1.429^{+0.001}_{-0.001}(scale) \pm +0.004(PDF) \pm 0.001(\alpha_s) \pm 0.001(m_t).$$

The measured result is in very good agreement with theoretical expectation.

Bibliography

[1] CMS Collaboration, "Measurement of b-tagging efficiency using ttbar events", CMS Physics Analysis Summary CMS-PAS-BTV-11-003, (2011).

[2] CMS Collaboration, "Identification of b-quark jets with the CMS experiment", *JINST* **8** (2013) 04013, doi:10.1088/1748-0221/8/04/P04013, arXiv:1211.4462.

[3] CMS Collaboration, "Performance of b tagging at $\sqrt{s}$=8 TeV in multijet, ttbar and boosted topology events", CMS Physics Analysis Summary (to be released) CMS-PAS-BTV-13-001, (2013).

[4] CMS Collaboration, "Measurement of ttbar Pair Production Cross Section at sqrt(s)=7 TeV using b-quark Jet Identification Techniques in Lepton + Jet Events", CMS Physics Analysis Summary CMS-PAS-TOP-11-003, (2011).

[5] CMS Collaboration, "Measurement of $\mathrm{t\bar{t}}$ Production Cross Section at $\sqrt{s}$=7 TeV using b-quark Jet Identification Techniques in Lepton + Jet Events", *Phys. Lett. B* **720** (2013) 83, doi:10.1016/j.physletb.2013.02.021, arXiv:1212.6682.

[6] CMS Collaboration, "Top pair cross section in e/mu+jets at 8 TeV", CMS Physics Analysis Summary CMS-PAS-TOP-12-006, (2012).

[7] F. Mandl and G. Shaw, "Quantum Field Theory". 1984. 358 p.

[8] F. Halzen and A. D. Martin, "Quarks and Leptons: an Introductory Course in Modern Particle Physics". 1984. 396p.

[9] S. Weinberg, "The Quantum Theory of Fields. Vol. 1: Foundations". 1995. 609 p.

[10] J. Beringer et al., "The Review of Particle Physics", *Phys. Rev. D* **86** (2012) 010001, doi:10.1103/PhysRevD.86.010001.

[11] CMS Collaboration, "Measurements of the properties of the new boson with a mass near 125 GeV", CMS Physics Analysis Summary CMS-PAS-HIG-13-005, (2013).

[12] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", *Phys. Lett. B* **716** (2012) 30–61, doi:10.1016/j.physletb.2012.08.021, arXiv:1207.7235.

[13] ATLAS Collaboration, "Observation of a new particle in the search for the standard model Higgs boson with the ATLAS detector at the LHC", *Phys. Lett. B* **716** (2012) 1–29, doi:10.1016/j.physletb.2012.08.020, arXiv:1207.7214.

[14] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, "GLOBAL CONSERVATION LAWS AND MASSLESS PARTICLES", *Phys. Rev. Lett.* **13** (1964) 585–587, doi:10.1103/PhysRevLett.13.585.

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

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

[17] Planck Collaboration, "Planck 2013 results. XVI. Cosmological parameters", arXiv:1303.5076.

[18] S. P. Martin, "A Supersymmetry Primer", arXiv:9709356v6.

[19] Tevatron Electroweak Working Group Collaboration, "Combination of CDF and DØresults on the mass of the top quark using up to 8.7 fb^{-1} at the Tevatron", arXiv:1305.3929.

[20] CDF Collaboration, "Observation of Top Quark Production in Pbar-P Collisions", *Phys.Rev.Lett.* **74** (1995) 2626–2631, doi:10.1103/PhysRevLett.74.2626, arXiv:9503002.

[21] DØ Collaboration, " Observation of the Top Quark ", *Phys.Rev.Lett.* **74** (1995) 2632–2637, doi:10.1103/PhysRevLett.74.2632, arXiv:9503003.

[22] M. Czakon et al., "The total top quark pair production cross-section at hadron colliders through $\text{O}(\alpha_s^4)$", arXiv:1303.6254.

[23] M. Czakon et al., "Constraints on the gluon PDF from top quark pair production at hadron colliders", arXiv:1303.7215.

[24] Tevatron Electroweak Working Group, "Combination of the $\text{t}\bar{\text{t}}$ production cross section measurements from the Tevatron Collider", *DØ Note 6363 and CDF Note 10926* (2012).

[25] CMS Collaboration, "Measurement of the $\text{t}\bar{\text{t}}$ production cross section in the dilepton channel in pp collisions at $\sqrt{s} = 7$ TeV", *JHEP* **11** (2012) 67, doi:10.1007/JHEP11(2012)067, arXiv:1208.2671.

[26] CMS Collaboration, "Top pair cross section in dileptons", CMS Physics Analysis Summary CMS-PAS-TOP-12-007, (2012).

[27] ATLAS Collaboration, "Measurement of the top quark pair cross section with ATLAS in pp collisions at $\sqrt{s} = 7$ TeV using final states with an electron or a muon and a hadronically decaying tau lepton", *Phys. Lett. B* **717** (2012) 89–108, doi:10.1016/j.physletb.2012.09.032, arXiv:1205.2067.

[28] ATLAS Collaboration, "Measurement of the cross section for top-quark pair production in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector using final states with two high-pt leptons", *JHEP* **1205** (2012) 59, doi:10.1007/JHEP05(2012)059, arXiv:1202.4892.

[29] ATLAS Collaboration, "Measurement of the top quark pair production cross section in the single-lepton channel with ATLAS in proton-proton collisions at 8 TeV using kinematic fits with b-tagging", ATLAS Note ATLAS-CONF-2012-149, (2012).

[30] D0 Collaboration, "A Direct Measurement of the Total Decay Width of the Top Quark", arXiv:1308.4050.

[31] CMS Collaboration, "Measurement of the ratio B(t→Wb)/B(t→Wq)", CMS Physics Analysis Summary CMS-PAS-TOP-12-035, (2012).

[32] CMS Collaboration, "Search for flavor changing neutral currents in top quark decays in pp collisions at 7 TeV", *Phys. Lett. B* **718** (2013) 1252, doi:10.1016/j.physletb.2012.12.045, arXiv:1208.0957.

[33] CMS Collaboration, "Search for baryon number violating top quark decays in pp collisions at 8 TeV", CMS Physics Analysis Summary CMS-PAS-B2G-12-023, (2012).

[34] CMS Collaboration, "Measurement of differential top-quark pair production cross sections in pp collisions at $\sqrt{s} = 7$ TeV", *EPJC* **73** (2013) 2339, doi:10.1140/epjc/s10052-013-2339-4, arXiv:1211.2220.

[35] CMS Collaboration, "Measurement of differential top-quark pair production cross sections in the lepton+jets channel in pp collisions at 8 TeV", CMS Physics Analysis Summary CMS-PAS-TOP-12-027, (2012).

[36] CMS Collaboration, "Measurement of the differential ttbar cross section in the dilepton channel at 8 TeV", CMS Physics Analysis Summary CMS-PAS-TOP-12-028, (2012).

[37] ATLAS Collaboration, "Measurements of top quark pair relative differential cross-sections with ATLAS in pp collisions at $\sqrt{s} = 7$ TeV", *Eur. Phys. J. C* **73** (2013) 2261, doi:10.1140/epjc/s10052-012-2261-1, arXiv:1207.5644.

[38] CMS Collaboration, "Measurement of the top quark mass in the muon+jets channel", *JHEP* **12** (2012) 105, doi:10.1007/JHEP12(2012)105, arXiv:1209.2319.

[39] ATLAS Collaboration, "Measurement of the Top Quark Mass from $\sqrt{s}$=7 TeV ATLAS Data using a 3-dimensional Template Fit", ATLAS Note ATLAS-CONF-2013-046, (2013).

[40] CMS Collaboration, "Top mass combination", CMS Physics Analysis Summary CMS-PAS-TOP-11-018, (2012).

[41] CMS Collaboration, "Measurement of the top - antitop mass difference in pp collisions at $\sqrt{s}$ = 8 TeV", CMS Physics Analysis Summary CMS-PAS-TOP-12-031, (2012).

[42] CMS Collaboration, "Measurement of Spin Correlations in ttbar production", CMS Physics Analysis Summary CMS-PAS-TOP-12-004, (2012).

[43] CMS Collaboration, "Observation of spin correlation in ttbar events from pp collisions at sqrt(s) = 7 TeV using the ATLAS detector", *Phys. Rev. Lett.* **108** (2012) 212001, doi:10.1103/PhysRevLett.108.212001, arXiv:1203.4081.

[44] CMS Collaboration, "Measurement of the top polarization in the dilepton final state", CMS Physics Analysis Summary CMS-PAS-TOP-12-016, (2012).

[45] ATLAS Collaboration, "Measurement of top quark polarisation in $t\bar{t}$ events with the ATLAS detector in proton-proton collisions at $\sqrt{s}$=7TeV", ATLAS Note ATLAS-CONF-2012-133, (2012).

[46] CMS Collaboration, "LHC Combination note: W helicities", CMS Physics Analysis Summary CMS-PAS-TOP-12-025, (2013).

[47] ATLAS Collaboration, "Combination of the ATLAS and CMS measurements of the W-boson polarization in top-quark decays", ATLAS Note ATLAS-CONF-2013-033, (2013).

[48] CMS Collaboration, "Constraints on the Top-Quark Charge from Top-Pair Events", CMS Physics Analysis Summary CMS-PAS-TOP-11-031, (2012).

[49] ATLAS Collaboration, "Measurement of the top quark charge in pp collisions at $\sqrt{s}$ = 7TeV in the ATLAS experiment", ATLAS Note ATLAS-CONF-2011-141, (2011).

[50] The Gfitter Group Collaboration, "The Electroweak Fit of the Standard Model after the Discovery of a New Boson at the LHC", *Eur. Phys. J. C* **72** (2012) 2205, doi:10.1140/epjc/s10052-012-2205-9, arXiv:1209.2716.

[51] L. Evans and P. Bryant, "LHC Machine", *Journal of Instrumentation* **3** (2008) doi:10.1088/1748-0221/3/08/S08001.

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

[53] Fermilab, "Design report Tevatron 1 project", *FERMILAB-DESIGN* **01** (1983).

[54] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider", *Journal of Instrumentation* **3** (2008) doi:10.1088/1748-0221/3/08/S08003.

[55] ALICE Collaboration, "The ALICE experiment at the CERN LHC", *Journal of Instrumentation* **3** (2008) doi:10.1088/1748-0221/3/08/S08002.

[56] LHCb Collaboration, "The LHCb Detector at the LHC", *Journal of Instrumentation* **3** (2008) doi:doi:10.1088/1748-0221/3/08/S08005.

[57] LHCf Collaboration, "The TOTEM Experiment at the CERN Large Hadron Collider", *Journal of Instrumentation* **3** (2008) doi:doi:10.1088/1748-0221/3/08/S08007.

[58] LHCf Collaboration, "The LHCf detector at the CERN Large Hadron Collider", *Journal of Instrumentation* **3** (2008) doi:doi:10.1088/1748-0221/3/08/S08006.

[59] LHC, "LHC Programme Coordination web pages", 2012.

[60] CMS Collaboration, "CMS Luminosity - Public Results", 2012.

[61] M. Della Negra et al., "CMS physics Technical Design Report". Technical Design Report CMS. CERN, Geneva, 2006.

[62] CMS Collaboration, "Offline Primary Vertex Reconstruction with Deterministic Annealing Clustering", CMS Internal Note CMS-NOTE-11-014, (2011).

[63] CMS Collaboration, "Adaptive Vertex Reconstruction", CMS Note CMS-NOTE-08-033, (2008).

[64] CMS Collaboration, "Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at $\sqrt{s} = 7$ TeV", *Submitted to JINST* (2013) arXiv:1306.2016.

[65] C. Eck et al., "LHC computing Grid : Technical Design Report",.

[66] CMS Collaboration, "Public CMS Data Quality Information", 2013.

[67] Sjöstrand T., "Monte Carlo Tools", arXiv:0911.5286.

[68] Seymour, M.H. and Marx M., "Monte Carlo Event Generators", arXiv:1304.6677.

[69] Collins J. C. and others, "Factorization of Hard Processes in QCD", *Adv.Ser.Direct.High Energy Phys.* **5** (1988) 1–91, arXiv:0409313.

[70] CTEQ, "The Coordinated Theoretical-Experimental Project on QCD".

[71] MSTW, "Martin-Stirling-Thorne-Watt Parton Distribution Functions".

[72] DESY, "Super electron microscope HERA".

[73] CTEQ, "Online PDF plotting and calculation".

[74] Pumplin J. and others, "New Generation of Parton Distributions with Uncertainties from Global QCD Analysis", *JHEP* (2002) doi:10.1088/1126-6708/2002/07/012, arXiv:0201195.

[75] P. N. et al., "Implications of CTEQ global analysis for collider observables", *Physical Review D* **78** (2008), no. 1, doi:10.1103/PhysRevD.78.013004.

[76] J. Pumplin et al., "Uncertainties of predictions from parton distribution functions. 2. The Hessian method", *Phys. Rev.* **D65** (2001) 014013, doi:10.1103/PhysRevD.65.014013, arXiv:hep-ph/0101032.

[77] Sjöstrand T. and others, "PYTHIA 6.4 Physics and Manual", *JHEP* **6** (2006) 26, doi:10.1088/1126-6708/2006/05/026, arXiv:0603175.

[78] Corcella G. and others, "HERWIG 6.5: an event generator for Hadron Emission Reactions With Interfering Gluons (including supersymmetric processes)", *JHEP* (2001) doi:10.1088/1126-6708/2001/01/010, arXiv:0011363.

[79] J. Alwall et al., "MadGraph 5 : Going Beyond", *JHEP* **06** (2011) 128, doi:10.1007/JHEP06(2011)128, arXiv:1106.0522.

[80] S. Frixione, P. Nason, and C. Oleari, "Matching NLO QCD computations with parton shower simulations: the POWHEG method", *JHEP* **11** (2007) 070, doi:10.1088/1126-6708/2007/11/070, arXiv:0709.2092.

[81] Frixione F. and Webber B.R., "The MC@NLO 3.4 Event Generator", arXiv:0812.0770.

[82] Y. L. Dokshitzer *Sov. Phys. JETP* **46** (1977) 641–653.

[83] V. N. Gribov and L. N. Lipatov *Sov. J. Nucl. Phys.* **15** (1972) 438–450.

[84] L. N. Lipatov *Sov. J. Nucl. Phys.* **20** (1975) 94–102.

[85] G. Altarelli and G. Parisi *Nucl. Phys.* **B126** (1977) 298.

[86] V. V. Sudakov, "Vertex parts at very high-energies in quantum electrodynamics", *Sov. Phys. JETP* **3** (1956) 65–71.

[87] Alwall, J., "Parton shower and MLM matching", *NTU MadGraph school* (2012).

[88] Alwall, J. and others, "A standard format for Les Houches Event Files", *Comput.Phys.Commun* **176** (2007) 300–304, doi:10.1016/j.cpc.2006.11.010, arXiv:0609017.

[89] Hoeche S. and others, "Matching Parton Showers and Matrix Elements", arXiv:0602031.

[90] Mangano L. and others, "Matching matrix elements and shower evolution for top-quark production in hadronic collisions", *JHEP* **07** (2007) 013, doi:10.1088/1126-6708/2007/01/013, arXiv:0611129.

[91] Nason P. and Webber B., "Next-to-Leading-Order Event Generators", arXiv:1202.1251.

[92] Andersson, B. and others, "The Lund Fragmentation Process for a Multi-gluon String According to the Area Law", *Eur.Phys.J* **21** (2001) 631–647, doi:10.1007/s100520100757, arXiv:0106185.

[93] CMS Collaboration, "Measurement of the Underlying Event Activity in Proton-Proton Collisions at 0.9 TeV", *Eur.Phys.J.C* **70** (2010) 555–572, doi:10.1140/epjc/s10052-010-1453-9, arXiv:1006.2083.

[94] CMS Collaboration, "Measurement of the Underlying Event Activity at the LHC with $\sqrt{s}$= 7 TeV and Comparison with $\sqrt{s}$ = 0.9 TeV", *JHEP* **9** (2011) 109, doi:10.1007/JHEP09(2011)109, arXiv:1107.0330.

[95] Melnikov K. and Petriello F., "Electroweak gauge boson production at hadron colliders through $\mathcal{O}(\alpha_s^2)$", *Phys.Rev.D* **74** (2006) doi:10.1103/PhysRevD.74.114017, arXiv:0609070.

[96] Kidonakis N., "NNLL threshold resummation for top-pair and single-top production", arXiv:1210.7813.

[97] Campbell J. and others, "Single top production and decay at next-to-leading order", *Phys.Rev.D* **70** (2014) doi:10.1103/PhysRevD.70.094012, arXiv:0408158.

[98] CMS Collaboration, "Measurement of Jet Multiplicity Distributions in Top Quark Events With Two Leptons in the Final State at a centre-of-mass energy of 7 TeV", CMS Physics Analysis Summary CMS-PAS-TOP-12-004, (2012).

[99] CMS Collaboration, "Measurement of the Jet Multiplicity in dileptonic Top Quark Pair Events at 8 TeV", CMS Physics Analysis Summary CMS-PAS-TOP-12-004, (2012).

[100] Allison J. and others, "Geant4 developments and applications", *IEEE Trans.Nucl.Sci.* **53** (2006) 270, doi:10.1109/TNS.2006.869826.

[101] CMS Collaboration, "Measurement of the inclusive production cross sections for forward jets and for dijet events with one forward and one central jet in pp collisions at $\sqrt{s}$ = 7 TeV", *JHEP* **6** (2012) 36, doi:10.1007/JHEP06(2012)036, arXiv:1202.0704.

[102] ATLAS Collaboration, "Measurement of the Inelastic Proton-Proton Cross-Section at $\sqrt{s}$=7 TeV with the ATLAS Detector", *Nat. Commun.* **2** (2011) 463, doi:10.1038/ncomms1472, arXiv:1104.0326.

[103] TOTEM Collaboration, "First measurement of the total proton-proton cross section at the LHC energy of $\sqrt{s}$ =7 TeV", arXiv:1110.1395.

[104] CMS Collaboration, "Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET", CMS Physics Analysis Summary CMS-PAS-PFT-09-001, (2009).

[105] CMS Collaboration, "Commissioning of the Particle-Flow reconstruction in Minimum-Bias and Jet Events from pp Collisions at 7 TeV", CMS Physics Analysis Summary CMS-PAS-PFT-10-002, (2010).

[106] CMS Collaboration, "Particle-flow commissioning with muons and electrons from J/Psi and W events at 7 TeV", CMS Physics Analysis Summary CMS-PAS-PFT-10-003, (2010).

[107] Wolfgang A. and others, "Track Reconstruction in the CMS tracker",.

[108] CMS Collaboration, "Performance of CMS muon reconstruction in pp collision events at sqrt(s) = 7 TeV", *JINST 7* **7** (2012) 10002, doi:10.1088/1748-0221/7/10/P10002, arXiv:1206.4071.

[109] Adam, W. and others, "Reconstruction of electrons with the Gaussian-sum filter in the CMS tracker at the LHC", *Journal of Physics G* **31** (2005), no. 9, doi:10.1088/0954-3899/31/9/N01.

[110] CMS Collaboration, "Electron Reconstruction within the Particle Flow Algorithm", CMS Analysis Note CMS-NOTE-AN-10-034, (2010).

[111] Baffioni, S. and others, "Electron reconstruction in CMS", *Eur. Phys. J. C* **49** (2006), no. 4, 1099–1116, doi:10.1140/epjc/s10052-006-0175-5.

[112] CMS Collaboration, "Electron reconstruction and identification at sqrt(s) = 7 TeV", CMS Physics Analysis Summary CMS-PAS-EGM-10-004, (2010).

[113] CMS Collaboration, "Effective Areas for the particle-based isolation for electrons".

[114] CMS Collaboration, "Cut based electron identification".

[115] CMS Collaboration, "Multivariate electron identification".

[116] Salam, G.P. and Soyez, G., "A practical Seedless Infrared-Safe Cone jet algorithm", *JHEP* **5** (2007) 86, doi:10.1088/1126-6708/2007/05/086, arXiv:0704.0292.

[117] Ellis S. D. and others, "Successive Combination Jet Algorithm For Hadron Collisions", *Phys.Rev.D* **48** (1993) 3160–3166, doi:10.1103/PhysRevD.48.3160, arXiv:9305266.

[118] Cacciari M. and others, "The anti-k_T jet clustering algorithm", *JHEP* **8** (2008), no. 4, 063, arXiv:0802.1189.

[119] Dokshitzer, Yu. L. and others, "Better Jet Clustering Algorithms", *JHEP 9708:001,1997* (1997) doi:10.1088/1126-6708/1997/08/001, arXiv:9707323.

[120] CMS Collaboration, "Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS", *JINST* **6** (2011) 11002, doi:10.1088/1748-0221/6/11/P11002, arXiv:1107.4277.

[121] Maes M., "Usage of the ParticleFlow method at the CMS detector", *Master Thesis* (2009).

[122] CMS Collaboration, "Jet Energy Scale performance in 2011", CMS Detector Performance Summaries CMS-DP-2012-006, (2012).

[123] CMS Collaboration, "Status of the 8 TeV Jet Energy Corrections and Uncertainties based on 11 fb^{-1} of data in CMS", CMS Detector Performance Summaries CMS-DP-2013-011, (2013).

[124] A. Perloff (on behalf of the CMS Collaboration), "Pileup measurement and mitigation techniques in CMS", *J. Phys.: Conf. Ser.* **404** (2012) 012045, doi:10.1088/1742-6596/404/1/012045.

[125] M. Cacciari and G. P. Salam, "Pileup subtraction using jet areas", *Phys. Lett. B* **659** (2008) 119, doi:10.1016/j.physletb.2007.09.077, arXiv:0707.1378.

[126] H. Kirschenmann (on behalf of the CMS Collaboration), "Determination of the Jet Energy Scale in CMS", *J. Phys.: Conf. Ser.* **404** (2012) 012013, doi:10.1088/1742-6596/404/1/012013.

[127] Baffioni, S. and others, "Jet Energy Resolution Measurement",.

[128] CMS Collaboration, "Missing transverse energy performance of the CMS detector", *JINST* **6** (2011) 09001, doi:10.1088/1748-0221/6/09/P09001, arXiv:1106.5048.

[129] CMS Collaboration, "Tagging b jets with electrons and muons at CMS", CMS Note CMS-NOTE-2006-043, (2006).

[130] CMS Collaboration, "b-Jet Identification in the CMS Experiment", CMS Physics Analysis Summary CMS-PAS-BTV-11-004, (2012).

[131] CMS Collaboration, "Single Muon efficiencies in 2012 Data", CMS Detector Performance Summary CMS-DP-2013-009, (2013).

[132] CMS Collaboration, "Electron Efficiency Measurement for Top Quark Physics at $\sqrt{s}$ =8 TeV", CMS Analysis Note CMS-NOTE-AN-12-429, (2012).

[133] CMS Collaboration, "Measurement of the inclusive top pair production cross section in the semi-leptonic muon and electron channels with the complete dataset from the 2011 data taking period", CMS Analysis Note CMS-NOTE-AN-11-443, (2012).

[134] B. Klein, "Measurement of the top quark pair production cross section in the electron and jets channel at $\sqrt{s} = 7$ TeV with the CMS detector at the LHC". PhD thesis, Ghent University, 2012.

[135] CMS Collaboration, "Absolute Calibration of the Luminosity Measurement at CMS: Winter 2012 Update", CMS Physics Analysis Summary CMS-PAS-SMP-12-008, (2012).

[136] CMS Collaboration, "CMS Luminosity Based on Pixel Cluster Counting - Summer 2013 Update", CMS Physics Analysis Summary CMS-PAS-LUM-13-001, (2013).

[137] P. C. Louis Lyons, Duncan Gibaut, "How to combine correlated estimates of a single physical quantity", *Nuclear Instruments and Methods* **270** (July, 1988) 110–117, doi:10.1016/0168-9002(88)90018-6.

[138] A. Valassi, "Combining correlated measurements of several different physical quantities", *Nuclear Instruments and Methods* **500** (March, 2003) 391–405, doi:10.1016/S0168-9002(03)00329-2.

Summary

When the Large Hadron Collider came into operation in March 2010, it became not only the worlds largest but also most energetic particle collider. Throughout 2011, the LHC collided protons at a centre-of-mass energy of 7 TeV. Using the 5 fb^{-1} of integrated luminosity delivered by the LHC and recorded by the CMS experiment, physicists could for the first time explore the TeV scale with large statistics. For 2012, it was decided to push the accelerator to accelerate the beams beyond the 3.5 TeV of 2011 to produce collisions at 8 TeV.

To increase the potential of discovering new phenomena in collisions, the currently known phenomena have to be further explored and fortified. This is not only useful to develop and deploy new analysis methodologies, it also provides further benchmarks to the Standard Model. The top quark in particular plays a key role in this respect. As the LHC is the first collider able to produce top quarks in large numbers, a unique opportunity arrises to measure this quarks properties with high precision. Moreover, detailed knowledge of top quarks is crucial to many beyond the Standard Model theories as they often contain particles decaying into top quarks. In this thesis an important property of the top quark production process is measured, namely the production cross section of pairs of top quarks.

The top quark predominantly decays into a W boson and a b quark. Depending on the decay mode of the W boson, top quark pair decays can contain up to two leptons, neutrinos and four quarks. In this thesis, the semi-leptonic decay $\text{t}\bar{\text{t}} \rightarrow WbWb \rightarrow l\nu_l(l = e, \mu)bqqb$ is targeted where two b quarks are produced along with two light quarks, a lepton and a neutrino. Hence $\text{t}\bar{\text{t}}$ events are selected in data by requiring the event to contain exactly one lepton, either a muon or an electron, and four jets. The presence of a neutrino is inferred by requiring a minimal Missing Transverse Energy ($\not{E}_T$) which is inferred via the energy conservation in the transverse plane. To further purify the selected event sample, the presence of b quarks in the $\text{t}\bar{\text{t}}$ decay is exploited using b quark identification, or b-tagging, algorithms.

The $\text{t}\bar{\text{t}}$ production cross section is measured by estimating the number of $\text{t}\bar{\text{t}}$ events in data using a binned maximum likelihood fit. As input to the fit, the distribution of the invariant mass of the b jet and the charged lepton in the $t \rightarrow Wb \rightarrow l\nu_l b$ decay, or the jet-lepton mass, is used after a b-tag is applied on the b jet in this top quark decay. The number of observed $\text{t}\bar{\text{t}}$ events in data is then turned into a cross section using the selection efficiency for $\text{t}\bar{\text{t}}$ events. However, as the efficiency is determined from simulation, the b-tagging criterion implies large uncertainties on the final result.

To reduce the b-tagging uncertainty on the measured cross section, a data-driven technique is developed in this thesis to measure the b-tagging efficiency. Using the jet-

lepton mass, a sample rich in b jets (50%) is constructed as well as a sample depleted in b jets (16%). The latter sample is used to model the non-b jets contribution in the b jet rich sample allowing to ultimately measure the b-tagging efficiency. Additionally, the mis-tagging efficiency, is measured by using the previously reconstructed b jets component in the b jet rich sample to reconstruct the non-b jet component in the full event sample.

The measured b-tagging performance is then used to correct the simulation based $\mathrm{t\bar{t}}$ event selection efficiency to cancel the b-tagging systematic uncertainty on the cross section. Nevertheless, a large uncertainty remains due to the energy scale of the jets that are required to be present in the $\mathrm{t\bar{t}}$ events. To reduce this effect, a Jet Energy Scale calibration is implemented exploiting the mass of the W boson produced in the $t \to bW \to b\mathrm{q\bar{q}}$ decays. Reconstructing the W boson mass in data and in simulation allows to derive a constant calibration factor for the jets four-momenta allowing to significantly reduce the Jet Energy Scale uncertainty.

The top quark pair production cross section is measured at a collision energy of 7 and 8 TeV. The measured cross section at 7 TeV equals

$$\sigma_{t\bar{t}} = 163.9 \pm 4.4(stat.) \pm 10.7(syst.) \pm 3.6(lumi.)pb.$$

Furthermore, the 8TeV cross section is found to be

$$\sigma_{t\bar{t}} = 234.2 \pm 3.8(stat.) \pm 9.6(syst.) \pm 6.1(lumi.)pb.$$

Both measurements are found to be in very good agreement with the NNLO+NNLL predictions. Moreover, the measurements are also in good agreement with other $\mathrm{t\bar{t}}$ cross section measurements both from CMS and ATLAS and have a competitive total uncertainty.

Finally, the ratio of both cross sections is determined in order to reduce the total systematic uncertainty by cancellation of parts of the systematic uncertainties between both collision energies. The cross section ratio is measured to be

$$R_{\sigma_{t\bar{t}}}^{8/7TeV} = 1.453 \pm 0.041(stat.) \pm 0.057(syst.) \pm 0.051(lumi.).$$

Samenvatting

Meting van de werkzame doorsnede van top quark paar productie met het CMS experiment bij de LHC

Wanneer de Large Hadron Collider (LHC) in maart 2010 in gebruik werd genomen, werd het niet alléén 's werelds grootste maar ook de meest energetische deeltjesversneller. In 2011 werden proton bundels in de LHC gebotst met een botsingsenergie van 7 TeV. Dankzij de dataset van 5 fb^{-1} aan proton-proton botsingen geleverd door de LHC en geregistreerd door het CMS experiment, konden wetenschappers voor de eerste keer de TeV energie schaal onderzoeken met grote precisie. Om het bereik van de zoektocht naar nieuwe fysica te vergroten, werd vervolgens beslist om in 2012 de protonbundels te versnellen tot 4 TeV in plaats van 3.5 TeV waardoor er botsingen met een energie van 8 TeV konden plaatsvinden.

Om de kans op nieuwe ontdekkingen te vergroten moeten de huidig gekende processen zeer goed begrepen zijn. Dit is niet enkel nuttig om nieuwe analysetechnieken te ontwikkelen of te verbeteren, het is ook belangrijk als een test van het Standaard Model op zich. Aangezien de LHC de eerste deeltjesversneller is die top quarks in grote hoeveelheden kan produceren, biedt dit een unieke kans om de eigenschappen van deze quark met grote precisie op te meten. Daarbij komt nog dat de precieze kennis van de top quark belangrijk is in de zoektocht naar nieuwe fysica omdat top quarks daar vaak voorkomen als een vervalproduct. In deze thesis wordt één van de belangrijke eigenschappen van top quarks opgemeten, namelijk de werkzame doorsnede voor top quark paar productie in proton botsingen.

De top quark vervalt bijna uitsluitend in een W boson en een b quark. Afhankelijk van het verval van het W boson kan het verval van een top quark paar tot twee leptonen bevatten alsook hun bijhorende neutrino's en tot zelfs zes quarks. In deze thesis wordt het zogenaamde semi-leptonische verval $\text{t}\bar{\text{t}} \to WbWb \to l\nu_l(l = e, \mu)bqqb$ bestudeerd waarbij twee b quarks, twee lichte quarks, een lepton en een neutrino voorkomen. Dit proces kan dan geïsoleerd worden in de data door exact één gereconstrueerd muon of elektron te eisen samen met minstens vier gereconstrueerde jets. De aanwezigheid van een neutrino kan vervolgens afgeleid worden via het energiebehoud in het transverse vlak. Om het geselecteerde $\text{t}\bar{\text{t}}$ sample nog zuiverder te maken, kan gebruik gemaakt worden van b quark identificatie technieken of $b - tagging$.

De werkzame doorsnede voor de productie van een top quark paar kan gemeten worden in data door gebruik te maken van een Maximum Likelihood fit. Deze fit wordt

toegepast op de distributie van de invariante massa van het geladen lepton en de b jet in het $t \to Wb \to l\nu_l b$ verval, ook wel de jet-lepton massa genoemd. Vervolgens wordt dan geëist dat de b jet in dit top quark verval voldoet aan een b-tagging criterium. Het resultaat van de fit is dan het aantal geobserveerde $t\bar{t}$ events in data waaruit, gebruik makend van de selectie efficiëntie voor $t\bar{t}$ events, de werkzame doorsnede kan bepaald worden. Gezien deze efficiëntie bepaald wordt in gesimuleerde botsingen, is er hierop mogelijk een grote systematische onzekerheid komende van het b-tagging criterium.

Om deze onzekerheid in te perken kan de efficiëntie van het b-tagging criterium zelf gemeten worden door middel van een methode die volledig los staat van gesimuleerde botsingen. In deze thesis wordt dergelijke techniek ontwikkeld. Gebruik makend van de jet-lepton massa kan een jet sample rijk aan b jets (50%) alsook een jet sample arm aan b jets (16%) gedefiniëerd worden. Het laatstgenoemde sample wordt daaropvolgend gebruikt om de contributie van niet-b jets in het eerstgenoemde sample in te schatten en uiteindelijk te verwijderen. De b-tagging discriminator distributie kan dan opgemaakt worden in het b jet-rijk sample waar de niet-b jet contaminatie verwijderd werd. Dit laat toe voor elk b-tagging criterium de efficiëntie te bepalen. Vervolgens kan ook de efficiëntie om een niet-b jet foutief te identificeren als een b jet gemeten worden. Dit wordt gedaan door de gereconstrueerde b-tagging discriminator distributie te gebruiken om de niet-b jet distributie te reconstrueren in het volledige sample.

De gemeten efficiënties voor het b-tagging criterium kunnen vervolgens gebruikt worden om de $t\bar{t}$ selectie efficiëntie verkregen uit gesimuleerde botsingen te corrigeren zodat de onzekerheid omtrent b-tagging op de uiteindelijke werkzame doorsnede opgeheven wordt. Dit is echter niet de enige grote onzekerheid op het finale resultaat. Een andere bron van onzekerheid is de energie kalibratie van de jets die gebruikt worden om de $t\bar{t}$ topologie te reconstrueren. Deze onzekerheid kan echter gereduceerd worden door gebruik te maken van de massa van het W boson dat geproduceerd wordt in het $t \to bW \to b\mathrm{q}\bar{\mathrm{q}}$ verval. Door de massa van dit boson zowel in echte als in gesimuleerde botsingen te meten kan een bijkomende globale kalibratie van de jet viervectoren bekomen worden. Deze laat toe de onzekerheid van de jet energie kalibratie significant te verlagen.

De werkzame doorsnede voor top quark paar productie is gemeten bij zowel 7 TeV als bij 8 TeV proton botsingen. Bij 7 TeV werd een resultaat bekomen van

$$\sigma_{t\bar{t}} = 163.9 \pm 4.4(stat.) \pm 10.7(syst.) \pm 3.6(lumi.)pb.$$

Bij 8 TeV botsingen werd een werkzame doorsnede van

$$\sigma_{t\bar{t}} = 234.2 \pm 3.8(stat.) \pm 9.6(syst.) \pm 6.1(lumi.)pb.$$

gemeten. Beide metingen zijn in goede overeenstemming met de theoretische verwachting bij NNLO+NNLL. Daarbij komt dat deze resultaten ook zeer consistent zijn met andere metingen die uitgevoerd zijn in zowel het CMS als het ATLAS experiment.

Als laatste wordt de verhouding tussen de werkzame doorsneden bij 7 en 8 TeV gemeten. Het voordeel van deze meting is dat systematische onzekerheden die gecorreleerd zijn tussen de beide metingen deels of volledig zullen opheffen in de verhouding. Dit leidt tot een preciezer resultaat dat met de theoretische voorspelling kan vergeleken

worden. Deze verhouding is gelijk aan

$$R_{\sigma_{t\bar{t}}}^{8/7TeV} = 1.453 \pm 0.041(stat.) \pm 0.057(syst.) \pm 0.051(lumi.).$$

Acknowledgements

The nice thing about doing research and writing a PhD thesis is the fact that you are not alone in this adventure. There are so many people who contributed to this thesis in so many different ways over the past years.

I would like to start by thanking the members of my thesis jury. I really enjoyed the private defence and the feedback I received from you all. Thank you for reviewing my manuscript, your feedback was valuable to improve the manuscript as well as strengthen the scientific results.

Next. I would like to thank CERN, the place where my passion for particle physics began which ultimately got me to pursue a PhD in Physics. In particular I am very grateful for the opportunity I was given to join the CERN Summer Student programme in 2008. This was an amazing experience both on scientific as personal level. During my stay at CERN I came into contact with experimental physics research. Working and living at CERN really opened my eyes and aroused my passion for experimental particle physics.

After completing my Masters thesis in 2009, I was given the opportunity by the IIHE and the Vrije Universiteit Brussel to start a PhD under the supervision of Jorgen D'Hondt. Jorgen, I really appreciated your strong physics insights and your endless creative ideas. We had numerous interesting discussions on how to proceed with the research and how to continuously improve the results. The nice thing is that we always agreed on the path to follow. The result of this fruitful collaboration is clearly visible if you compare the results two years ago with those provided in this manuscript. It wouldn't have been possible without your support!

During my 4 years at the IIHE , I came across many people who contributed both on a personal and on a scientific level. I would like to thank all of them and in particular all the people in the top quark research group. There are a few people which I want to thank in particular. I had a lot of fun sharing the office with Stijn, Alexis, Gerrit, Petra and the others. I want to thank you all for the nice working atmosphere. Stijn, after almost 9 years of studying together I really came to value you as a colleague and a friend. I really enjoyed working with you and our many discussions about physics topics and beyond. Gerrit, you are a walking encyclopedia and your umbrella gave a nice holiday touch to the office. Petra, thank you for the weird and random conversations we had in the office and for the moral support. Alexis, the best of luck to you for your own defence. You will do great! I also want to thank Rebeca. I really enjoyed the coffee meetings at CERN and your interesting views on things.

Furthermore I like to thank the people I have worked with closely in the CMS collaboration especially the people from the BTV and TOP groups. I am also grateful

for the nice collaboration I had with the people from Universiteit Gent.

Whenever I was stuck and needed a break, I could count on family and friends to get my mind of work. I really enjoyed going out for drinks, food and parties with you and this helped me to relax and clear mind. In particular, I want to thank my wife, Freya, who has been there for me the whole time, in the good and the bad. Thank you for keeping me sane these last years. You have always pushed me to get up and continue whenever I was in a stub. Thank you for being patient given the long working hours and the limited spare time I had.

I also want to express my gratitude to my parents. You supported me throughout my studies and allowed me to pursue a degree in Physics even if nobody believed I was capable to succeed. Your support and confidence was key to my success.