Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Dipartimento di Fisica e Astronomia
Corso di Laurea in Fisica

Measurement of the $t\bar{t}$ production cross section at 13 TeV in the all-jets boosted regime with CMS at LHC

Relatore:
Prof. Andrea Castro

Presentata da:
Eric Ballabene

Anno Accademico 2016/2017

Abstract

In this work, a measurement of the top quark-antiquark pair ($t\bar{t}$) production cross section $\sigma_{t\bar{t}}$ is presented, using data from pp collisions collected by the CMS detector in 2016 at CERN, with a centre-of-mass energy $\sqrt{s} = 13$ TeV, and corresponding to an integrated luminosity of $37\,\text{fb}^{-1}$. The analysis has been performed in the all-jets boosted channel, characterized by the presence of two wide jets, stemming from the boost of the particles originating from the hadronic decay into a W boson and a b quark of the top quark. Candidate events have been selected using a multijet trigger, specific kinematic and b-tagging requests, and a multivariate analysis in order to enhance the signal from the large background. Then a likelihood fit has been performed which yields a cross section $\sigma_{t\bar{t}} = 554 \pm 12\,(\text{stat}) \pm 283\,(\text{sys})\,\text{pb}$, where some of the main sources of systematic uncertainty have been considered. This measurement is consistent with what expected from theoretical calculations.

Sommario

In questo lavoro, viene presentata una misura della sezione d'urto $\sigma_{t\bar{t}}$ per la produzione accoppiata di quark e antiquark del top ($t\bar{t}$), usando dati relativi a collisioni pp, raccolti dal rivelatore CMS nel 2016 al CERN, con un'energia nel centro di massa $\sqrt{s} = 13$ TeV, e corrispondente a una luminosità integrata di $37\,\text{fb}^{-1}$. L'analisi è stata effettuata nel canale *all-jets boosted*, caratterizzato dalla presenza di due ampi *jet*, derivanti dal *boost* delle particelle originate dal decadimento adronico in un bosone W e un quark b del quark top. Gli eventi candidati sono stati selezionati usando un trigger *multijet*, specifiche richieste cinematiche e di *b-tagging*, e un'analisi multivariata per favorire il segnale rispetto al fondo. È stato poi effettuato un fit di verosimiglianza il quale restituisce una sezione d'urto $\sigma_{t\bar{t}} = 554 \pm 12\,(\text{stat}) \pm 283\,(\text{sist})\,\text{pb}$, dove sono state considerate alcune tra le principali sorgenti di incertezza sistematica. La misura è consistente con quanto previsto dai calcoli teorici.

I would like to express my sincere gratitude to Professor Andrea Castro, for his continuous support throughout the thesis work, his patience and his kindness.

Contents

Chapter 1

High-Energy Physics at CERN

CERN stands for "Conseil Européen pour la Recherche Nucléaire", or European Council for Nuclear Research. At CERN, physicists and engineers work together studying particles and their interactions, to determine the fundamental structure of the universe and its origin [1]. The main instruments used at CERN are particle accelerators and detectors. Accelerators boost beams of particles to high energies before they are made to collide with each other. Detectors observe and record the results of these collisions. From these results, scientists learn about particle properties, such as mass and charge.
A great amount of energy is required for the particles and can be achieved through an efficient complex of accelerators. The accelerator complex is a succession of machines that accelerate particles to increasingly energies. Each machine boosts the energy of a beam of particles, before injecting the beam into the next machine in the sequence.
Everything starts from a simple bottle of hydrogen gas, where an electric field is used to strip hydrogen atoms of their electrons to yield protons. "Linac 2", the first accelerator in the chain, accelerates the protons to the energy of 50 MeV. The beam is then injected into the "Proton Synchrotron Booster" (PSB), which accelerates the protons to 1.4 GeV, followed by the "Proton Synchrotron" (PS), which pushes the beam to 25 GeV. Protons are then sent to the "Super Proton Synchrotron" (SPS) where they are accelerated to 450 GeV. The protons are finally transferred to the two beam pipes of "The Large Hadron Collider" (LHC). The CERN complex is shown in ***Fig. 1.1***.

1.1 LHC: The Large Hadron Collider

The LHC is without doubt the world's largest particle accelerator. It consists of a 27-kilometre ring of superconducting magnets with a number of accelerating structures to boost the energy of the particles along the way. The tunnel starts near CERN (Meyrin), goes close to the Jura mountains, continues underneath French countryside, comes round near Geneva airport (Switzerland) and then back to CERN. Inside the accelerator, two

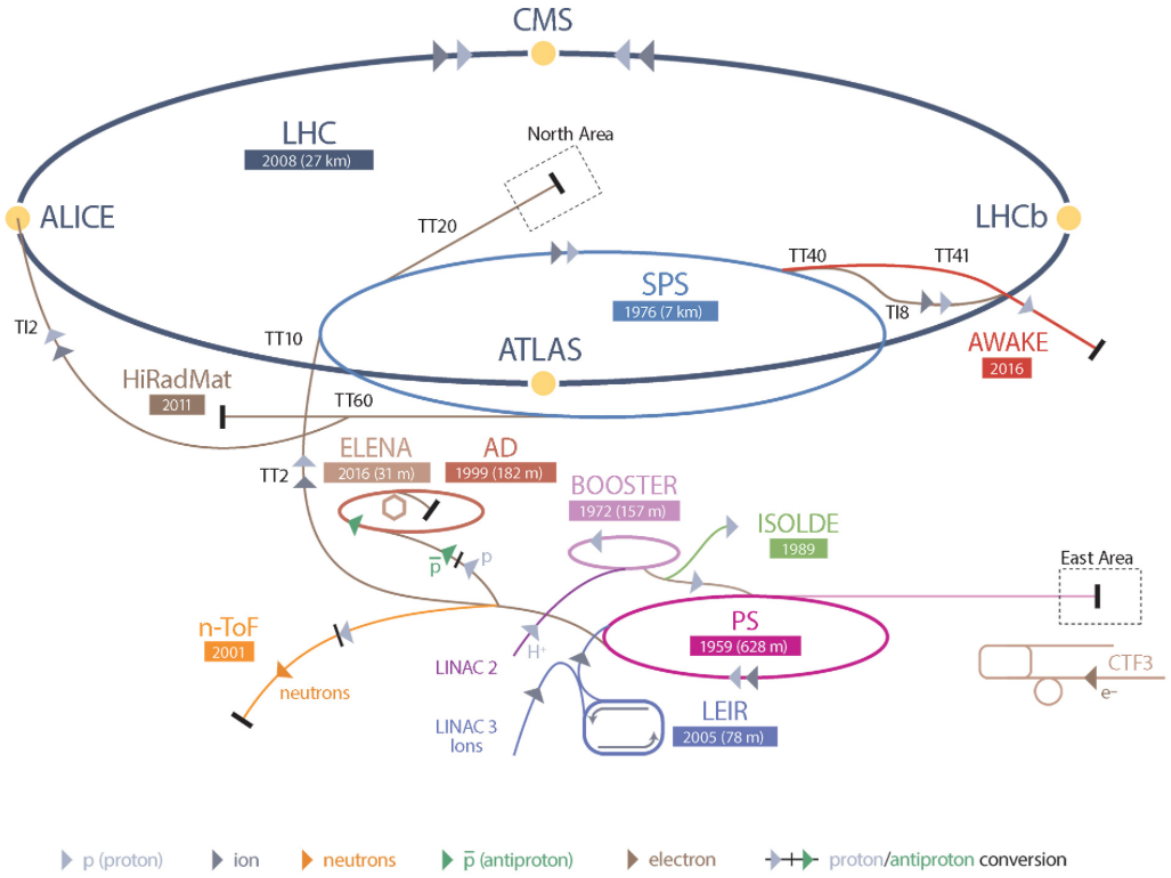

Figure 1.1: *The accelerator complex at CERN.*

high-energy particle beams travel at close to the speed of the light before they are made to collide. For the protons to collide frontally, the beams travel in opposite directions in separate beam pipes (see ***Fig. 1.2***), guided around the accelerator ring by a strong magnetic field maintained by superconducting electromagnets. The electromagnets are placed along the eight 2.45-km-long arcs with 154 in each arc and operate at magnetic fields above 8 T (see ***Table 1.1*** for some basic LHC parameters). They are dipole magnets, allowing opposite and uniform field in both vacuum chambers. The windings for the two beam channels are housed in the same cryostat, a choice due to the space limitations of the tunnel where the LEP was. A schematic cross section of a "cryodipole" is shown in ***Fig. 1.3***. They are built from niobium-titanium (NbTi) coils of special electric cable that operates in a superconducting state, efficiently conducting electricity without resistance or loss of energy. This requires chilling the magnets to −271.3 °C (1.9 K), a temperature colder than outer space. For this reason, much of the accelerator is connected to a distribution system of liquid helium, which cools the magnets, as well as to other supply services. The coils are surrounded by non-magnetic "collars" of austenitic steel, a material that combines the required properties of good thermal contraction and magnetic permeability. The collars hold the coils in place against the strong magnetic

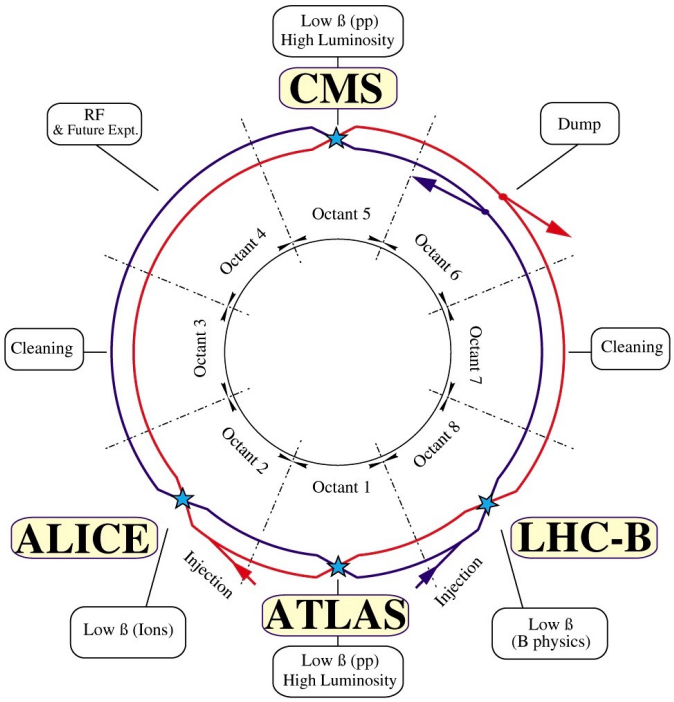

Figure 1.2: *The schematic layout of the LHC.*

forces that arise when the coils are at full field. The collared coil assembly is surrounded by the magnetic circuit contained by a shrinking cylinder, formed by welding two half-shells made out of stainless steel [2]. This provides the necessary rigidity for the whole magnet.

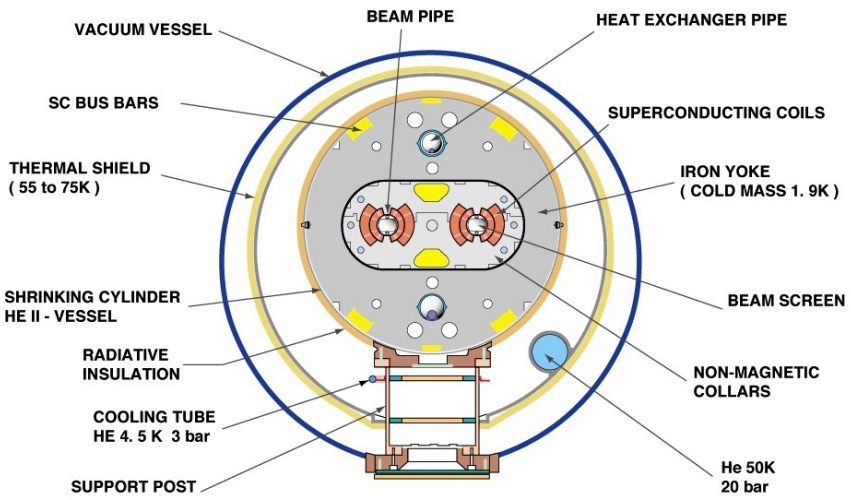

Figure 1.3: *Cross section of LHC cryodipole.*

Relevant LHC parameters (design values)			
Particle used	Protons (or heavy ions)	Number of magnets	$\sim$ 9300
Circumference	26659 m	Number of main dipoles	1232
Injected beam energy	450 GeV	Number of quadrupoles	858
Nominal beam energy	7 TeV	Number of correcting magnets	6208
Number of arcs	8	Number of RF cavities	8 per beam
Number of lattice cells per arc	23	RF frequency	400 MHz
Number of straight sections	8	Current in main dipole	11800 A
Magnetic field at 7 TeV	8.33 T	Power consumption	$\sim$ 180 MW
Operating temperature	1.9 K	Instantaneous luminosity	$10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$

Table 1.1: *Most relevant parameters and physical features of the LHC.*

The dipole superconducting magnets are not the only LHC magnets. There are also 858 quadrupole magnets. The main quadrupole magnets, each one 5-7 meters long, focus the beams in order to maximize the probability of pp collisions in proximity of the experiments. They are placed in correspondence of the eight insertions along the LHC rings (see ***Fig. 1.2***): four in the experiment areas, two for the beam cleaning, one for the radiofrequency cavities and one for the beam dumping. The insertions are made up of a long straight zone and two transition regions that delimit it. Just before the collision, another type of magnet is used to "squeeze" the particles closer together to increase the chances of collisions. The final beams are so tiny that the task of making them collide is equivalent to firing two needles 10 kilometres apart with such precision that they meet halfway. In addition, other magnetic multipoles (sextupoles, octupoles) act to help in beam focusing and counteracting other interactions that each beam suffers along the beam pipes (gravitational interactions over protons, electromagnetic interactions among bunches, etc).

An important indicator of the performance of an accelerator is its instantaneous luminosity L: it is proportional to the number of collisions that occur in a given amount of time according to the following relation:

$$R = L\,\sigma, \tag{1.1}$$

where R is the rate of beam-beam collisions and σ is the cross section for the specific process.
The higher the instantaneous luminosity, the more data the experiments can gather to allow them to observe rare processes. The instantaneous luminosity depends only on beam parameters and can be written for a Gaussian beam profile as

$$L = \frac{N_b^2\, n\, f_r\, \gamma}{4\,\pi\,\epsilon_n\,\beta^*}, \tag{1.2}$$

where:
∘ N_b is the number of particles per bunch;
∘ n is the number of bunches per beam;
∘ f_r is the revolution frequency;
∘ γ is the relativistic factor;
∘ ϵ_n is the normalized transverse emittance;
∘ β^* is the amplitude of the β function at the collision point.

The design LHC value for the instantaneous luminosity is $10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$ (***Table 1.1***) but record values of $2 \times 10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$ have been reached so far. An important project, called "The High-Luminosity Large Hadron Collider" (HL-LHC), aims to crank up the performance of the LHC increasing luminosity by a factor of 10 beyond the LHC design value.
However, not all the four experiments require so high value of instantaneous luminosity. LHCb, designed for the physics of the b quark, and ALICE ("A Large Ion Collider Experiment"), studying the heavy-ions collisions, normally work with lower instantaneous luminosity. On the other hand, ATLAS ("A Toroidal LHC ApparatuS") and CMS ("Compact Muon Solenoid") both need large values of L in order to maximize the physics reach. These last two independently-designed detectors are in fact used to investigate the largest range of physics possible, from studying the Standard Model (SM), to searching for extra dimensions and particles that could make up dark matter. In the next section, we give a complete description of one of these experiments, CMS.

1.2 The CMS Detector

CMS is a particle detector that is designed to detect a wide range of particles and phenomena produced in high-energy pp (or heavy-ions) collisions at the LHC. It is built around a huge superconducting solenoid and different layers of detectors measure the different particles and use this key data to build up a picture of events at the heart of the collision. Finding the energy and momentum of a particle gives clues to its identity and particular patterns of particles are indications of new and exciting physics. In order to perform such tasks, the CMS detector is made up of (see ***Fig. 1.4***):
∘ a silicon central tracking system, to reconstruct trajectories and give accurate momentum measurements for charged particles;
∘ an electromagnetic calorimeter (ECAL), to detect and measure electrons and photons with high resolution;
∘ a hadron calorimeter (HCAL), to detect and measure hadrons and to entirely surround the collision and prevent particles from escaping ("hermetic calorimeter");
∘ a high-performance system to detect and measure muons.

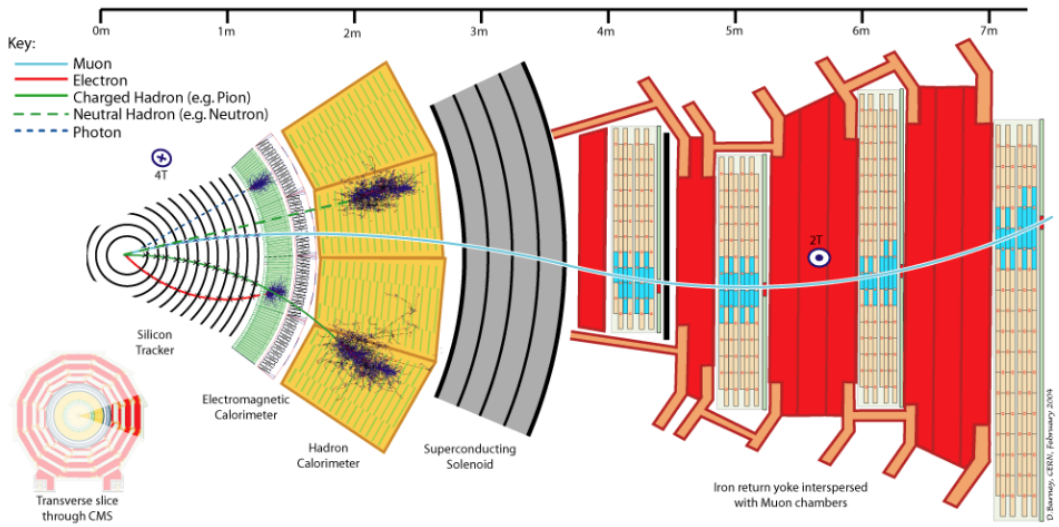

Figure 1.4: *Transverse layout of the CMS detector.*

1.2.1 The tracker

The tracker used at CMS measures efficiently the paths taken by charged particles, by recording their positions at certain key interaction points. It is then able to estimate the momentum of particles by the curvature of the trajectories, since it is known that the more bent a path is, the less momentum that particle has in the magnetic field. Momenta of particles are in fact crucial in helping us to build up a picture of events at the heart of the collision.
The tracker can reconstruct the paths of high-energy muons, electrons and charged hadrons as well as see tracks coming from the decay of very short-lived particles such as hadrons containing b quarks that will be used to study the differences between matter and antimatter. The tracker needs to record particle paths accurately yet be lightweight so as to disturb the particle as little as possible. It does this by taking position measurements so accurate that tracks can be reliably reconstructed using just a few measurement points. Each measurement is accurate to about 10 μm, a fraction of the width of a human hair. It is also the innermost layer of the detector and so it receives the highest flux of particles: the construction materials were therefore carefully chosen to resist radiation. The final design consists of a tracker made entirely of silicon: the pixels, at the very core of the detector and dealing with the highest intensity of particles, and the silicon microstrip detectors that surround them. As particles travel through the tracker the pixels and microstrips produce tiny electric signals that are amplified and detected. The tracker employs sensors covering an area the size of a tennis court, with 75 million separate electronic read-out channels: in the pixel detector there are some 6000 connections per square centimeter.

1.2.2 The electromagnetic calorimeter

The ECAL is specifically designed to measure the energy of particles that interact primarily via the electromagnetic interaction such as electrons and photons. Measuring them with good precision is not easy because of the high magnetic field and the short time between collision, and requires very particular detector materials.
The lead tungstate crystals are made primarily of metal and are heavier than stainless steel, but with a touch of oxygen in this crystalline form they are highly transparent and "scintillate" when electrons and photons pass through it. This means they produce light in proportion to the particle energy. These high-density crystals produce light in fast, short, well-defined photon bursts that allow for a precise, fast and fairly compact detector.
Photodetectors that have been especially designed to work within the high magnetic field, are also glued onto the back of each of the crystals to detect the scintillation light and convert it to an electrical signal that is amplified and sent for analysis.
The ECAL, made up of a barrel section and two "endcaps", forms a layer between the tracker and the HCAL. The cylindrical "barrel" consists of 61,200 crystals formed into 36 "supermodules", each weighing around three tonnes and containing 1700 crystals. The flat ECAL endcaps seal off the barrel at either ends and are made up of almost 15,000 further crystals.

1.2.3 The hadron calorimeter

The HCAL is designed to measure the energy of hadrons, including particles which cannot be recorded in other parts of the CMS detector. To detect these, the HCAL must be "hermetic", that means it must capture, as much as possible, every particle emerging from the collisions.
The HCAL is a sampling calorimeter, which means it finds the particle position, energy and arrival time using alternating layers of absorber and fluorescent scintillator materials that produce a rapid light pulse when the particle passes through. Special optical fibers collect up this light and feed it into readout boxes where photodetectors amplify the signal. When the amount of light in a given region is summed up over many layers of tiles in depth, this total amount of light is a measure of a particle energy.
As the HCAL is massive and thick, fitting it into CMS, which is compact by name, was a challenge, as well as to contain and measure the large cascades of particles produced when a hadron hits the dense absorber material.
Additionally, the HCAL provides an indirect measurement of the presence of uncharged and non-interacting particles such as neutrinos. Measuring these particles is important since they can tell us if new particles such as the Higgs boson or hypothesized supersymmetric particles have been formed.

1.2.4 The superconducting solenoid

The CMS magnet is a solenoid, i.e. a coil of superconducting wire creating a magnetic field when electricity flows through it. At CMS the solenoid has an overall length of 13 m and a diameter of 7 m, able to create a magnetic field of 3.8 T. It is the largest superconducting magnet ever built and allows the tracker and calorimeters to be placed inside the coil, resulting in a detector that is, overall, "compact", compared to detectors of similar weight.

1.2.5 The muon detectors

The muon detectors constitute the final part of the CMS detector. Their task is to reveal muons and measure their momenta through four “muon stations” (MS). Muon stations are placed outside the magnet coil and are interleaved with iron “return yoke plates” (see ***Fig. 1.4***). In fact muons, unlike most other particles such as hadrons, can penetrate several metres of iron without interacting with matter or being stopped by any of CMS calorimeters.
In total there are 1400 muon chambers among which there are the “cathode strip chambers” (CSCs), which track the particles positions and provide a trigger, and the “resistive plate chambers” (RPCs), which constitute a quick trigger that decides to keep the acquired muon data or not.
The great interest in these particles, which even give the name to the CMS detector, is explained by the fact that they are expected to be the product of the decay of a number of potential new particles; for example, one of the cleanest signatures of the Higgs boson is its decay into four muons.

Chapter 2

Standard Model and The Top Quark

The SM explains how the basic building blocks of matter interact, governed by the fundamental forces. These basic building blocks of matter are the irreducibly smallest detectable particles, called elementary particle, while the fundamental forces the SM deals with are the strong, weak and electromagnetic ones.
The SM includes 12 elementary particles, divided into 6 quarks and 6 leptons, and the other 12 anti-particle counterparts. There are also the gauge bosons, force carriers that mediate the strong, weak, and electromagnetic force, which are respectively gluons, W^+, W^-, Z^0 bosons and photons. The SM even includes the Higgs boson, the particle which explains why the other elementary particles, except photons and gluons, are massive.
Among all these particles, the top quark plays a very important role, being the most massive of all observed elementary particles. The large value of its mass, for instance, makes the top quark contribution dominant in loop corrections to many observables, like the W boson mass. Moreover, precise measurements of the W boson and the top quark masses are related to the mass of the Higgs boson, and are used to assess the self-consistency of the SM.

2.1 Standard Model

Most of the experimental facts currently discovered on the ultimate constituents of physical reality and their interactions are described by the SM.
The appropriate theory for the description of these relativistic and quantum phenomena is "Quantum Field Theory" (QFT), which specifies both the particles and the interactions that are the basis of all the physical entities so far known. In QFT, particles appear as an excitation in the particles underlying physical field. Their interactions are accounted for by adding interaction terms between the corresponding physical fields. In contrast to classical field theories, where there is a finite number of degrees of freedom

for a fixed number of particles, a QFT does not have these restrictions and can therefore also describe systems with varying particle numbers. The particles can be divided into two basic groups: fermions and bosons.

2.1.1 Fermions

In general, fermions are particles that has an odd half-integer spin (like 1/2, 3/2, and so forth). Quarks and leptons, as well as most composite particles, like protons and neutrons, are fermions. As a consequence, they obey the Pauli exclusion principle, i.e. fermions cannot co-exist in the same location at the same time with all the same quantum numbers, and thus they are described by Fermi-Dirac statistics.
For what concerns elementary fermions, namely fermions that are not composed of other particles, they all are spin-1/2 particles and are classified according to how they interact in two categories: six quarks (up, down, charm, strange, top and bottom) and six leptons (electron, muon, tau, electron neutrino, muon neutrino, tau neutrino). In ***Table 2.1*** and ***Table 2.2*** are summarized some general characteristics of them.

Quark	Electric Charge	Mass	Spin
u	$2/3$	$2.2^{+0.6}_{-0.4}$ MeV	$1/2$
d	$-1/3$	$4.7^{+0.5}_{-0.4}$ MeV	$1/2$
c	$2/3$	1.27 ± 0.03 GeV	$1/2$
s	$-1/3$	96^{+8}_{-4} MeV	$1/2$
t	$2/3$	$4.18^{+0.04}_{-0.03}$ GeV	$1/2$
b	$-1/3$	$173.21 \pm 0.51 \pm 0.71$ GeV	$1/2$

Table 2.1: *Relevant physical quantities of quarks.*
The values of mass are taken from the **Particle Data Group (PDG)** updated to 2016.

Lepton	Electric Charge	Mass	Spin
ν_e	0	< 2.05 eV (95% CL)	$1/2$
e	-1	$0.5109989461 \pm 0.0000000031$ MeV	$1/2$
ν_μ	0	< 0.19 MeV (90% CL)	$1/2$
μ	-1	$105.6583745 \pm 0.0000024$ MeV	$1/2$
ν_τ	0	< 1.2 MeV (95% CL)	$1/2$
τ	-1	1776.86 ± 0.12 MeV	$1/2$

Table 2.2: *Relevant physical quantities of leptons.*
The values of mass are taken from the **Particle Data Group (PDG)** updated to 2016.

2.1.2 Bosons

Bosons are particles which have integer spin and which therefore are not constrained by the Pauli exclusion principle like the half-integer spin fermions. The energy distribution of bosons is described by Bose-Einstein statistics.
Bosons may be either elementary, like photons, or composite, like mesons. The observed elementary bosons are photons, the force carriers of the electromagnetic field, gluons, the fundamental force carriers underlying the strong force, and W and Z bosons, the force carriers which mediate the weak force. There is also the Higgs boson, which gives W and Z bosons mass via the Higgs mechanism and whose discovery was announced by ATLAS and CMS experiments [3, 4] on 4 July 2012. In ***Table 2.3*** are summarized some general characteristics of them. They are all gauge bosons having a spin of 1 unit, except the Higgs boson having spin 0. Additionally, the graviton is a hypothetical elementary particle not incorporated in the SM. If it exists, a graviton must be a boson, and could conceivably be a gauge boson. However, since the graviton, the hypothetical force mediating particle for the gravitational force has no mass and a spin of 2 units, it would be a tensor boson.

Boson	Electric Charge	Mass	Spin
γ	0	$< 1 \times 10^{-18}$ eV	1
g	0	$0^{[1]}$	1
W^+	1	80.385 ± 0.015 GeV	1
W^-	-1	80.385 ± 0.015 GeV	1
Z^0	0	91.1876 ± 0.0021 GeV	1
H^0	0	125.09 ± 0.24 GeV	0

Table 2.3: *Relevant physical quantities of bosons.*
The values of mass are taken from the **Particle Data Group (PDG)** updated to 2016.
[1] Theoretical value. A mass as large as a few MeV may not be precluded.

2.1.3 Gauge theories

The SM is based on the gauge theories of the electroweak and strong interactions. A gauge theory is a theory which is invariant under certain symmetry transformations. Mathematically, every elementary particle is represented by a field, whose Lagrangian encodes all its possible interactions and is invariant under a continuous group of local transformations. The global symmetry group is $SU(3)_C \otimes SU(2)_L \otimes U(1)_{Y_W}$. It includes the three-dimensional special unitary group $SU(3)_C$, where C stands for colour (red, blue, green) representing the strong force, and the two-dimensional special unitary group

composed of the one-dimensional unitary group, $SU(2)_L \otimes U(1)_{Y_W}$, which unifies the electromagnetic and weak interactions. Here, L indicates that this interaction affects only left-handed fermions and Y_W indicates the "weak hypercharge", a combination of electrical charge and weak isospin. Mathematically, the weak hypercharge is defined by

$$Q = T_3 + \frac{1}{2} Y_W, \tag{2.1}$$

where T_3 is the third component of weak isospin and Q is the particle charge. Weak isospin connects quark and lepton doublets of left-handed particles in all generations: for example, up and down quarks, top and bottom quarks, electrons and electron neutrinos. On the other hand, the definition of the hypercharge of strong interactions Y comes from "Gell-Mann–Nishijima" formula:

$$Q = I_3 + \frac{1}{2} Y, \tag{2.2}$$

where I_3 is the third component of isospin and Q is the particle charge. Isospin connects only up and down quarks, acts on both chiralities (left and right) and is a global symmetry. The hypercarge Y relative to the strong forces is the sum of specific quantum numbers, the strangeness (S), baryon number (B), charm (C), bottomness (B'), and topness (T):

$$Y = S + B + C + B' + T. \tag{2.3}$$

Conservation of hypercharge thus implies the conservation of flavour. Strong interactions conserve hypercharge, but weak interactions do not. The baryon number B is instead a strictly conserved quantum number of a system. It is an addictive quantity, defined as

$$B = \frac{n_q - n_{\bar{q}}}{3}, \tag{2.4}$$

where n_q is the number of quarks, and $\mathrm{n_{\bar{q}}}$ is the number of antiquarks. Particles without any quarks, such as leptons and bosons, have a baryon number of zero. The baryon number is conserved in all the interactions of the SM. The hypothetical models of grand unified theory (GUT) and supersymmetry (SUSY) allow the changing of a baryon into several leptons, thus violating the conservation of both baryon and lepton numbers.

2.2 The Top Quark

According to the SM, the top quark is an elementary particle belonging to the third generation of quarks. It is a fermion, with a spin 1/2 and an electric charge $+2/3\,e$, and it is the most massive of all observed elementary particles (see ***Table 2.1***). It forms a weak isospin doublet together with the bottom quark, where the top quark is the up-type quark with the third component of the weak isospin $I_3 = +1/2$. As well as the

other quarks, it is subject to the electromagnetic interaction (having electric charge), to the strong nuclear interaction (having colour charge), to the weak nuclear interaction (being part of a weak isospin doublet). Due to its huge mass, top quark mean lifetime is so short (about $5 \cdot 10^{-25}$ s) that it decays semi-weakly into a real W boson and a b quark before it can hadronize, without forming mesons or baryons with other quarks. In fact, unlike the other quarks, it has never been observed a bounded state involving this quark. This gives physicists a unique opportunity to study a bare quark, different from all other quarks that they combine to form hadrons, and can only be observed as such. Its existence was postulated in 1973 by Makoto Kobayashi and Toshihide Maskawa to explain the observed CP violations in kaon decay, and was discovered in 1995 by the Collider Detector at Fermilab (CDF) and D0 experiments at Fermilab. Kobayashi and Maskawa won the 2008 Nobel Prize in Physics for the prediction of the top and bottom quark.

2.2.1 Top quark discovery

The first quark model was proposed by Gell-Mann and Zweig in 1964 and included hadrons of the three lightest quarks: up, down, and strange. A fourth quark, the charm quark, was discovered by the observation of the J/Ψ resonance, a meson composed of charm and its antiparticle ($c\bar{c}$). Its discovery was made independently by two research groups, one at the "Stanford Linear Accelerator Center" (SLAC), headed by Burton Richter, and one at the "Brookhaven National Laboratory" (BNL), headed by Samuel Ting. They actually noted that they had discovered the same particle and both groups announced their discoveries on November 11, 1974. This discovery went to history under the name "November Revolution" because brought with it a series of rapid changes in particle physics. As a matter of fact, it completed the second generation of quarks. In both quark generations, there is a quark with a third component of the weak isospin $I_3 = +1/2$ and a charge $Q = +2/3\,e$ ("up-type quark"), and a quark with $I_3 = -1/2$ and a charge $Q = -1/3\,e$ ("down-type quark"), together forming a weak isospin doublet.
A fifth quark, the bottom quark, was discovered by the observation of the ϒ resonance, a meson composed of the bottom quark and its antiparticle ($b\bar{b}$). It was discovered by a team headed by Leon Lederman, at Fermilab in 1977. At that time, the open question was if the bottom quark constitutes a weak isospin singlet or is part of another doublet. To shed light on this question, the quantum numbers of the bottom quark were determined in e^+e^- collision experiments at "Deutsches Elektronen-Synchrotron" (DESY). Through measurements of the angular distribution of b hadrons produced in e^+e^- collision, it was clear the quantum numbers for the bottom quark should be $I_3 = -1/2$ and $Q = -1/3\,e$, strongly suggesting that the bottom quark is the down-type quark of the third generation whose weak isospin partner was yet to be discovered.
The third generation of quarks should be a weak isospin doublet, consisting of the bottom quark and its isospin partner, the top quark. In the early 1980s, the search for top

quarks was conducted in electron-positron colliders, searching for evidence of the process $e^+e^- \rightarrow t\bar{t}$. Those experiments, such as PEP at SLAC with $\sqrt{s} = 30$ GeV and PETRA at DESY with $\sqrt{s} = 45$ GeV, did not have enough energy to produce the $t\bar{t}$ pair and, in absence of a signal, only lower limits on the top quark mass were established ($m_t > 45$ GeV).
A step forward was made with the CERN SPS operating as a proton-antiproton collider (which was called Sp$\bar{p}$S). At the Sp$\bar{p}$S at CERN with $\sqrt{s} = 540$ GeV, top quarks were searched in decays of real W bosons, $W^+ \rightarrow t\,\bar{b}$, but the available energy was insufficient to cope with a mass above 70 GeV. At the Tevatron, the center-of-mass $\sqrt{s}$ available in $p\bar{p}$ collisions was of 1.8 TeV, a value which marked a significant increase compared to the SPS, so the top quark became directly accessible in collider experiment for the first time. The first Tevatron collisions were recorded by the CDF experiment in 1985. In those years, CDF improved the lower limit on the top quark mass to $m_t > 91$ GeV at 95% CL, thus avoiding the possibility for W bosons to decay into top quarks. After several years spent on the long search, on February 24, 1995, CDF and D0, finally announced the observation of the top quark [5 ,6]. CDF reported finding six dilepton events plus 43 single-lepton events. It concluded that the probabilities that background fluctuations could account for these events were only one in a million. D0 observed three dilepton events plus 14 single-lepton events and concluded that the probabilities that these could have been caused by backgrounds were two in a million. The top quark masses reported by the two experiments were (176 ± 13) GeV for CDF and (199 ± 30) GeV for D0.

2.2.2 Stability of the universe

The top quark mass is related to the electroweak vacuum stability described by the SM. According to it, the top quark and Higgs boson masses allow to understand whether the universe is stable, unstable or metastable, leading to possible different endings. The values of the two masses measured so far seem to place the universe near the limit between the stable and the metastable behaviour. The high experimental uncertainty related to the current measurements of the top quark and Higgs boson masses however does not allow us to make an accurate prediction. A very precise assessment of the stability of the electroweak vacuum can only be made at a future high-energy electron-positron collider, where the top quark pole mass could be determined with a few hundred MeV accuracy [7].

2.2.3 Constraints on the Higgs boson mass

Due to its large mass, the top quark contributes significantly to the masses of the W and Z bosons via loop diagrams as in ***Figs. 2.1a and 2.1b***. It has the largest among all quarks coupling to the Higgs boson, contributing to radiative corrections as in ***Figs.***

2.1c and 2.1d. The ratio ρ is defined as

$$\rho = \frac{M_W^2}{M_Z^2}(1 - \sin\theta_W) = 1 + \Delta r, \tag{2.5}$$

where Δr is the radiative correction given by

$$\Delta r = \frac{3G_F}{8\pi^2\sqrt{s}}m_t^2 + \frac{\sqrt{2}G_F}{16\pi^2}M_W^2\left[\frac{11}{3}\log\left(\frac{M_H^2}{M_W^2}\right)\right]. \tag{2.6}$$

We see that this ratio has a quadratic correction from the top quark mass, and a logarithmic correction from the Higgs boson mass. Thus, simultaneously measuring $\sin\theta_W$, the W, Z and top quark masses can give an indirect measurement of the Higgs boson mass. Therefore, the mass of the top quark constrains the mass of the Higgs boson. The considerably large Higgs boson coupling of the top quark and its obvious difference in mass from all the other fermions motivate a close examination of its properties.

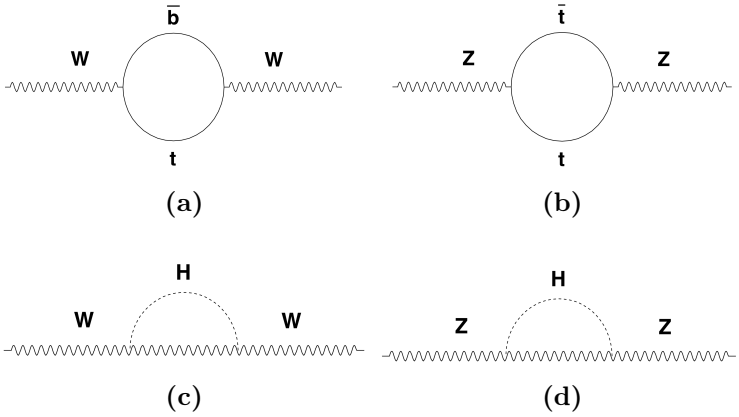

Figure 2.1:
Corrections to W and Z masses due to loop diagrams involving top quarks and the Higgs boson.

2.2.4 Top quark-antiquark ($t\bar{t}$) pair production

There are multiple processes that can lead to the production of top quarks. They can be conceptually divided into two categories: top quark-antiquark ($t\bar{t}$) pair production, through strong interactions, and single top quark production, through weak interactions. In this work, we focus on $t\bar{t}$ pair production, which is the most common in particle accelerators and represents the signal for our study. Consequently, single top quark production constitutes a generic background.
For what concerns $t\bar{t}$ pair production, we can associate it to two main processes. The

annihilation of a quark with its antiquark, producing a gluon which subsequently originates a $t\bar{t}$ pair (see ***Fig. 2.2a***). This process was responsible for the majority of the top quark events produced at Tevatron when the top quark was first discovered in 1995.
The gluon-gluon fusion, instead, prevails in pp collisions at high-energy accelerators like the LHC. The collision is characterized by two gluons, which subsequently decay into a $t\bar{t}$ pair. We show the process of gluon-fusion in the s-channel (***Fig. 2.2b***), in the t-channel (***Fig. 2.2c***) and u-channel (***Fig. 2.2d***).

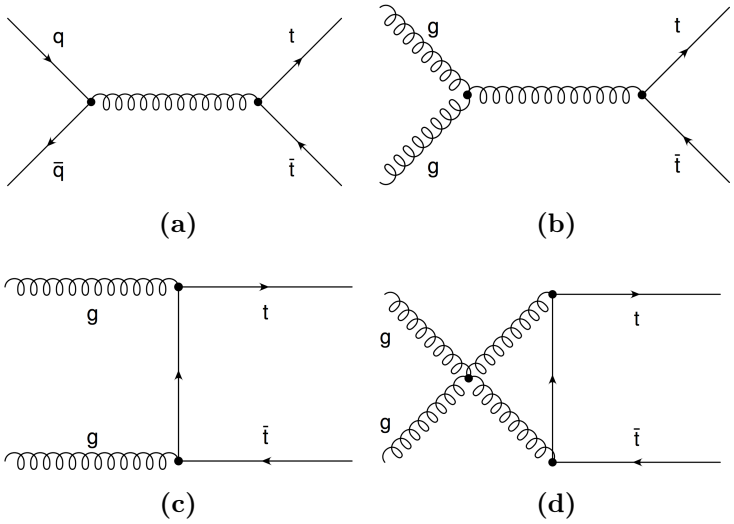

Figure 2.2:
Production Feynman diagrams for $pp \to t\bar{t}$.

Predictions for the total production cross sections are now available at next-to-next-to-leading order (NNLO) with next-to-next-to-leading-log (NNLL) soft gluon resummation. Assuming a top quark mass of 173.2 GeV, close to the Tevatron + LHC average, the resulting theoretical prediction [8] of the top quark pair cross section at NNLO+NNLL accuracy at LHC at $\sqrt{s} = 13\,\text{TeV}$ is $\sigma_{t\bar{t}} = 832 \pm 23$ pb.

2.2.5 Single top quark production

In addition to the $t\bar{t}$ production discussed ***Section 2.2.4***, the top quark can be also produced through the weak interactions during the hard scattering. In this process, only one top quark is produced thus the process is usually referred to as "single top quark production". There are three main channels: the t-channel, the s-channel and tW-channel. The next-to-leading-order (NLO) cross sections are given in ***Table 2.4***. The t-channel single top quark production was the dominant mode at the Tevatron. In this process, a space-like W boson scatters off a b quark, which is either produced via

Production mode at 13 TeV	σ central value (pb)	Scale uncert. (pb)	PDF+α_S uncert. (pb)	Tot. uncert. (pb)
t-channel	136.02	$^{+4.09}_{-2.92}$	$^{+3.52}_{-3.52}$	$^{+5.40}_{-4.57}$
s-channel	6.35	$^{+0.18}_{-0.15}$	$^{+0.14}_{-0.14}$	$^{+0.23}_{-0.20}$
tW-channel	71.7	$^{+1.80}_{-1.80}$	N.A.	$^{+3.40}_{-3.40}$

Table 2.4: *Single top quark channel cross sections.*
The values of single top quark channel cross sections are taken from the **LHCPhysics Web** updated to 2017.

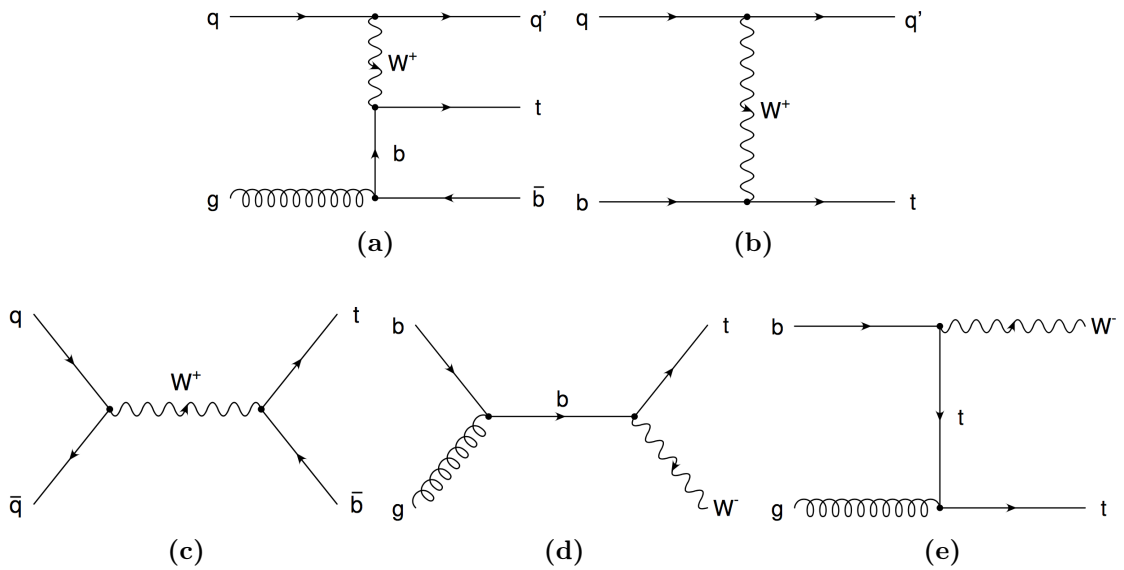

Figure 2.3:
Production Feynman diagrams for single top quark in t-channel (a,b), s-channel (c) and tW-channel (d,e).

gluon splitting g $\rightarrow$ b$\bar{\text{b}}$ or stemming from the proton sea. The leading diagrams for t-channel mode are shown in ***Figs. 2.3a and 2.3b***. There is a $2 \rightarrow 3$ and a $2 \rightarrow 2$ diagram, where the latter is a sub-process of the former. The s-channel is, at the leading order, the process in which a q with a $\bar{\text{q}}$ of the same weak isospin doublet (e.g. u$\bar{\text{d}}$), produce a time-like W boson that subsequently decays into t$\bar{\text{b}}$. The leading diagram for s-channel mode is shown in ***Fig. 2.3c***. The single top quark can be also produced in association with a real W boson in the so-called tW-channel. The main processes for t production is gb $\rightarrow$ tW$^-$ and similarly for the $\bar{t}$ production, g$\bar{\text{b}}$ $\rightarrow$ $\bar{\text{t}}$W$^+$. Other Cabibbo-Kobayashi-Maskawa (CKM) suppressed contributions from gs and gd initial states are negligibly small. The diagrams representing the leading order processes for tW-channel mode are shown in ***Figs. 2.3d and 2.3e***.

2.2.6 Theoretical cross section

The theoretical cross section $\sigma_{\mathrm{pp}\to\mathrm{t\bar{t}}}$ for the process $\mathrm{pp} \to \mathrm{t\bar{t}}$ can be expressed as a cross section for the high-energy ("hard") parton-parton scattering process weighted by parton distribution functions (PDFs) of the partons participating in the scattering processes, integrated over all parton momenta and summed over all parton types. The hard scattering cross section is process-specific and can be computed in perturbative QCD, while the PDFs are universal and can be measured independently of the hard process. The factorization formula for the cross section reads as

$$\sigma_{\mathrm{pp}\to\mathrm{t\bar{t}}} = \sum_{i,j} \int dx_i \; f_i(x_i, \mu^2) \int dx_j \; f_j(x_j, \mu^2) \hat{\sigma}_{ij}(\hat{s}, \hat{\mu^2}, \alpha_s, m_t), \tag{2.7}$$

where indices i, j indicate gluons and quarks (q,$\bar{\mathrm{q}}$,g), f_i,f_j are PDFs, x_i,x_j are the momentum fractions of the partons, $\hat{s} = x_i x_j s$ is the centre-of-mass energy squared of the partons, $\hat{\sigma}_{i,j}$ is the cross section of the partonic process, α_s is the strong coupling constant and μ is the factorization scale, related to the perturbative order of the calculations. Since the cross section is built up starting from the PDFs, a precise measurement can lead to constraints on them, along with constraints on the value of top quark mass m_t.

2.2.7 Top quark decay

Top quark decays before the hadronization can occur and does not form bound states such as top mesons ($\mathrm{t\bar{q}}$) or toponium ($\mathrm{t\bar{t}}$). This is because its mean lifetime τ_t is shorter than the typical time scale of hadronization. The lifetime τ_t can be calculated considering its relation to the total decay width of the top quark Γ_t:

$$\tau_t = \frac{\hbar}{\Gamma_t}, \tag{2.8}$$

where $\hbar$ is the reduced Planck constant, $\hbar \approx 6.582119 \times 10^{-15} \mathrm{eV} \cdot \mathrm{s}$. At LO the total decay width of the top quark Γ_t^{LO} is given by [9]:

$$\Gamma_t^{LO} = \frac{G_F}{8\pi\sqrt{2}} m_t^3 \Big(1 - \frac{m_W^2}{m_t^2}\Big)^2 \Big(1 + 2\frac{m_W^2}{m_t^2}\Big) \approx 1.5\,\mathrm{GeV}, \tag{2.9}$$

yielding $\tau_t \approx 5 \times 10^{-25}\,\mathrm{s}$. On the other hand, the typical time scale of hadronization can be estimated from the inverse of the energy scale Λ_{QCD} at which QCD becomes non-perturbative: $1/\Lambda_{QCD} \approx 1/(200\,\mathrm{MeV}) \approx 3 \times 10^{-24}\,\mathrm{s}$.
The top quark decays semi-weakly into a real W boson and a b quark in $\approx 93\%$ of cases. The decay fraction R_{Wb} can be computed as:

$$R_{Wb} = \frac{BR(t \to Wb)}{BR(t \to Wq)} = \frac{|V_{tb}|^2}{|V_{tb}|^2 + |V_{ts}|^2 + |V_{td}|^2} \approx 0.93, \tag{2.10}$$

where $|V_{tb}|^2$, $|V_{ts}|^2$ and $|V_{td}|^2$ are coefficients of the CKM matrix, which contains information on the strength of flavour-changing weak decays. Obviously, if it is a t, it decays into a W^+ and a b, if it is $\bar{t}$, it decays into W^- and an $\bar{b}$, in agreement with the law of conservation of electric charge.

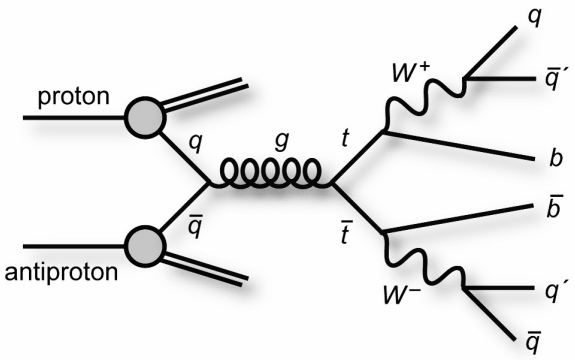

Figure 2.4: *Decay channels of a $t\bar{t}$ pair.*

Depending on how the W boson decays, we can distinguish different channels:

1. "Dilepton channel", in which both the W bosons decay into a lepton-neutrino doublet:

$$t\bar{t} \to W^+b\ W^-\bar{b} \to \ell^+\nu_\ell b\ \ell'^-\bar{\nu}_{\ell'}\bar{b},$$

   with $\ell =$ e or μ. The case in which $\ell = \tau$ is usually treated separately.

2. "Single lepton channel", in which only one W decays into a lepton-neutrino doublet:

$$t\bar{t} \to W^+b\ W^-\bar{b} \to \ell^+\nu_\ell b\ \bar{q}q'\bar{b},$$

   or

$$t\bar{t} \to W^+b\ W^-\bar{b} \to q\bar{q}'b\ \ell^-\bar{\nu}_\ell\bar{b}.$$

3. "All-hadronic channel", in which both W bosons decay hadronically (see ***Fig. 2.4***):

$$t\bar{t} \to W^+b\ W^-\bar{b} \to q\bar{q}'b\ \bar{q}q'\bar{b}.$$

   The products of this decay would be six distinct (i.e. "resolved") jets, but if the top quarks have large transverse momentum p_T, the particles originating from the decay $t \to W^+b$ will receive a large boost and emerge quite collimated, organized in two distinct wide jets:

$$q\bar{q}'b\ \bar{q}q'\bar{b} \to \mathrm{jet}_{boosted,1}\ \mathrm{jet}_{boosted,2}.$$

Channel	Branching ratio
Dilepton	10.3%
Single lepton	43.5%
All-jets	46.2%

Table 2.5: *Branching ratios for the $t\bar{t}$ decay channels.*

Jets are collimated showers of particles produced in high energy collisions. In our case we deal with "boosted jets": we will explain in detail what this entails, after having introduced some necessary kinematic variables. This channel is also called "all-jets" for its characterizing signature of only jets, at least two, without leptons.

The single lepton channel is usually the golden channel, it has high yield and low background, but it is not fully reconstructed kinematically because of the undetected neutrinos. The dilepton channel has high purity but low yield, and it is not kinematically reconstructed because of the undetected neutrinos. The all-jets channel has large background but high yield, and it is fully reconstructed kinematically since there are no undetected neutrinos. The branching ratios (BRs) of these $t\bar{t}$ decay channels are shown in ***Table 2.5***. Our analysis is based on the all-jets channel. In Chapter 3 we will carefully describe a technique to reject the background to favor the $t\bar{t}$ signal.

2.3 Kinematic variables

For particles and reconstructed objects, we define some kinematic variables:

- Transverse momentum p_T

$$p_T = \sqrt{{p_x}^2 + {p_y}^2}, \tag{2.11}$$

  where p_x and p_y are the momentum projections of the particle/object on the axes perpendicular to the direction of the particle beams;

- Rapidity y

$$y = \frac{1}{2} \ln \frac{E + p_z}{E - p_z}, \tag{2.12}$$

  where p_z is the component of the momentum of the particle/object along the axis of the particle beams and E is the energy;

- Pseudorapidity η

$$\eta = -\ln \left[\tan \left(\frac{\theta}{2} \right) \right], \tag{2.13}$$

  where θ is the azimuth angle measured with respect to the axis perpendicular to the horizontal plane passing through LHC. It can be shown that rapidity distributions,

i. e. the number of particles per unit rapidity, dN/dy, are invariant under Lorentz boosts along the z direction. In the limit of momenta much larger than the mass of a particle, the rapidity converges to pseudorapidity:

$$\lim_{|\vec{p}|\gg m} y = \eta. \tag{2.14}$$

The pseudorapidity of a particle is a purely geometrical quantity, it only depends on the polar angle θ, but not on the particle mass.

- Jet invariant mass m_{jet}

$$m_{jet} = \sqrt{\Big(\sum_i P_i\Big)^2}, \tag{2.15}$$

where P_i are the four-vectors of all i particles clustered into the jet, assuming a pion mass for all charged hadrons.

For the remainder of this work, we will follow common practice and work in natural units where the reduced Planck constant $\hbar$ and the speed of light in vacuum c are set to 1:

$$\hbar = c = 1. \tag{2.16}$$

Factors of $\hbar$ and c can always be restored by dimensional analysis.

2.4 Jets

Jets are physical objects defined as sprays of collimated particles produced by the fragmentation and hadronization of quarks and gluons originated by the hard collision. These objects are studied through clustering algorithms which combine information from the tracker and the calorimeters. The jet algorithm should be able to reconstruct the important kinematic quantities of the jet, such as its direction and momentum, from which we can infer those corresponding to the associated parton. There are two typical classes of algorithms used for jet reconstruction:

- Fixed-cone algorithms, which assume that the particles belonging to a jet are produced in conical regions around the parton direction. These algorithms cluster the particles based on $\eta - \phi$ space, producing jets with rigid circular boundaries;
- Sequential clustering algorithms, which are based on the hypothesis that particles belonging to a jet have small differences in p_T. These algorithms group the particles in the momentum space producing jets that have fluctuating area in the $\eta - \phi$ space. This second class is the one which will be used in this analysis.

2.4.1 Boosted jets

At the recent 13 TeV run of LHC, one encounters jets with p_T of up to few TeV. At such a high p_T, the products of decay of top quarks can be so collimated that standard reconstruction techniques begin to be less efficient. As a matter of fact, the width and the mass of a such jets become quite large. Wide jets stemming from particles with a relevant Lorentz boost are called "boosted jets". The angular separation of decay products is

$$\Delta R \approx \frac{2m}{p_T} \tag{2.17}$$

where $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$, m and p_T are the mass and transverse momentum of the particle that decays. For $p_T > 200$ GeV, the ability to solve individually the products of hadronic decays using standard jet algorithm begin to fail, and when $p_T > 300$ GeV, the top quark decay products tend to have a separation $\Delta R < 1$. For rebuilding these objects, it is therefore advisable to use large-scale jet radius, $R = 0.8$, where R is a parameter which enters in the clustering algorithm and is somehow related to the maximum ΔR of particles in the jet. A single jet that contains all the decay products of a massive particle has significantly different properties than a jet with the same p_T originated by a light quark. The feature of having two or three top quark decay particles inside the jet originates a more complex substructure that is totally absent in a typical high-p_T jet made of light gluons and quarks. Boosted particle algorithms remove jet contributions that are irrelevant in order to measure jet characteristics and its substructure variables. The opening angle between the decay products of one of these heavy objects become so small that the highly-boosted object shows up as a single massive, rather than two or three, jet. The study of such boosted jet in the framework of the SM may serve as a test of the model under extreme conditions, but, more electrifying, such a study may, according to some of the physics beyond the SM (BSM) scenarios, provide us with the first observations of BSM signals. Hence, an efficient separation between highly boosted QCD jets and highly-boosted heavy objects can facilitate both tests of the SM and searches for first hints of physics beyond the SM. Such a separation can only be done by looking at the internal structure of these jets, namely, their substructure [10].

2.5 Contamination by underlying event and pileup

Jet reconstruction at hadron colliders is contaminated by contributions from the underlying event (UE) and pileup (PU).

2.5.1 Underlying event

In a pp collision, the interaction is more complicated than the simple "hard scattering", i.e. the process where one parton from the first proton collides with another parton

from the second proton. The probability of such interaction between two specified partons is given by the PDFs. In addition to the remnants of the hadronization of the non-interacting ("spectator") quarks, more than just one parton from each proton may interact. These multiple parton interactions usually result in the radiation of soft particles with low energies. This is because the probability for a second hard interaction within the same proton-proton collision is smaller, which means that in the detector the signatures of the hard interaction are spoiled by soft energy deposits throughout the detector making the jet reconstruction itself more challenging.

2.5.2 Pileup

At LHC the beam does not consist of individual protons but large bunches of protons. As a result, instead of only one pp interaction, two or more are happening simultaneously, and the contribution of these additional interactions ("pileup") increases with the instantaneous luminosity. The soft radiation originating from pileup will also result in energy deposits all over the detector. A single pileup interaction would be less severe for a measurement than one from UE.

2.5.3 Jet grooming

Jet grooming seeks to get rid of softer components in a jet from UE or pileup. It leaves constituents from the hard scatter behind, allowing better mass resolution. Three algorithms are mainly used: "mass-drop/filtering", "trimming" and "pruning".

- Mass-drop/filtering tries to isolate symmetric subjets that exhibit a significantly smaller mass when compared with their parent jet. At first, energy clusters within the jet are identified using mass-drop and symmetry criteria. Secondly, the reclustered jets are filtered. It was optimized for $H \to b\bar{b}$ search using C/A jets. [11].

- Trimming uses the fact that contamination to a jet originating from PU, UE or initial state radiation (ISR) is comparatively soft compared to the products of a hard scattering process. The algorithm selects constituents by comparing their transverse momentum to the one of the reconstructed jet. For this purpose, an already reconstructed jet with radius parameter R_{jet} is split up and reclustered into subjets with radius parameter $R_{sub} < R_{jet}$. It uses the "k_t algorithm" because in opposition to other clustering methods it clusters from softer to harder, which will result in more balanced subjets, where the energy is shared between the subjets. Any i-subjets failing $p_T^i \,/\, p_T < f_{cut}$ are then removed [12].

- Pruning recombine jet constituents while vetoing wide angle (R_{cut}) and softer (z_{cut}) constituents. It does not recreate subjets but prunes at each point in jet reconstruction. It uses the relative transverse momentum as a selection criterion as well

as the trimming algorithm. But while the trimming procedure takes an already reconstructed jet as input and then probes into its substructure by the reconstruction of subjets, the pruning procedures input is a set of protojets and it acts during the reconstruction of a jet [13].

These three algorithms are currently being studied by ATLAS and CMS, and now that the LHC ramps up its luminosity are getting more and more important.

Chapter 3

Data Analysis

This chapter describes the experimental procedure employed to select events for data and Monte Carlo simulated samples, and how to measure the $t\bar{t}$ production cross section $\sigma_{t\bar{t}}$.

3.1 The samples

Our analysis is based on two samples: the data sample, which is collected by the CMS detector, and Monte Carlo (MC) simulated samples, which contain reconstructed $t\bar{t}$ signal and QCD background events.

3.1.1 Data sample

The data used for this analysis were collected during the 2016 LHC run of pp collisions at 13 TeV, corresponding to an integrated luminosity of $37\,\mathrm{fb}^{-1}$. They were stored in ROOT files as trees that contain all the relevant information on the major physics objects reconstructed in the detector (jets, leptons, photons, tracks).

3.1.2 Monte Carlo samples

MC simulated samples resemble experimental data both with respect to the physical processes involved in the hadron-hadron scattering and the interactions of the final state particles with the particle detector. They are used to devise a strategy for reducing the background while keeping as much signal as possible. The CMS collaboration uses MC generation programs to model a number of physics processes relevant to $t\bar{t}$ production and decay. The next-to-leading-order (NLO) POWHEG generator is used to generate $t\bar{t}$ signal events [14, 15]. Generic QCD multijet events are simulated instead with the MADGRAPH generator [16]. The MC samples used for the analysis are reported in ***Table 3.1***.

Sample	σ(pb)	Events
$t\bar{t}$ pair production	832	77229340
QCD multijet production ($300 < H_T < 500$ GeV)	3.67×10^5	54537900
QCD multijet production ($500 < H_T < 700$ GeV)	2.94×10^4	62271340
QCD multijet production ($700 < H_T < 1000$ GeV)	6.52×10^3	15629250
QCD multijet production ($1000 < H_T < 1500$ GeV)	1.06×10^3	15127290
QCD multijet production ($1500 < H_T < 2000$ GeV)	121.5	11826700
QCD multijet production ($H_T > 2000$ GeV)	2.54×10^1	6039005

Table 3.1: *Monte Carlo samples used in the analysis: $t\bar{t}$ and QCD multijet events. QCD simulated events are divided into slices of H_T, which stands for the scalar sum of jet transverse momenta.*

3.2 Events Selection

A great amount of pp collisions is actually produced at LHC, with only few of such events being relevant for the analysis, while the others constitute less interesting background events. The signal (S), in this case the $t\bar{t}$ pairs that decay in the all-jets channel, is overwhelmed by a huge background (B), making it impossible to discriminate S from B without an appropriate selection. The fraction of the $t\bar{t}$ pairs signal and the background events is quantitatively represented by the S/B signal-to-background ratio.
Therefore, a selection of events that maximizes the purity of the data, offering better performance in terms of the S/B, has been implemented. A ROOT macro has been used for this purpose: it first reads ROOT files for data and MC samples, extracts the events, acquires kinematic variables and orders the jets with decreasing p_T. Then, the macro can apply specific selections on the events, such as:

- a trigger selection
- a selection of the number of jets (N_{jets}), specifying requirements on p_T and η;
- a selection of the number of identified charged leptons (e or μ) ($N_{leptons}$);
- a selection of the number of "b-tagged" jets (N_{b-jets}), i.e. jets associated with the hadronization of b quarks, and on the number of "b-tagged" subjets ($N_{b-subjets}$);
- a further selection which considers the specific characteristics of signal events. This will be carried out considering a discriminator obtained from a multivariate analysis (MVA).

ROOT files are structured as ROOT trees, subdivided into branches and then into leaves, which correspond to variables of physical interest. For each event, some leaves are char-

acterized by only one entry, other leaves are characterized by two entries, corresponding to the most energetic jet, also called "leading jet", and the second jet.

3.2.1 Trigger

The trigger is the first requirement for our selection. The one we use is called:
"`HLT_AK8DiPFJet280_200_TrimMass30_BTagCSV`".
This trigger:

- is a "High-Level Trigger" (HLT), which means it is able to reduce to about 1 kHz the collision rate of the proton bunches. It is used after another trigger, called "Level 1" trigger, which previously reduces by a factor of 1000 the original collision rate, which is of 40 MHz at LHC. A HLT trigger is implemented in software running on a farm of commercial computers which includes about 16,000 CPU cores and exploits the same sophisticated software used for offline reconstruction and analysis, optimized in order to comply with the strict time requirements of the online selection [17,18];
- selects two "AK8 jets" (see ***Section 3.3***), i.e. wide jets reconstructed with a larger than usual clustering parameter ($R = 0.8$), whose p_T is at least 280 GeV for the leading one and at least 200 GeV for the second. These jets are reconstructed through the "particle-flow" (PF) algorithm, which aims at identifying and reconstructing individually each particle arising from the LHC proton-proton collision by combining the information from all the subdetectors [19];
- selects a value of the so-called "trimmed mass" greater than 30 GeV. Jet trimming can achieve significant improvements in event reconstruction mitigating sources of contamination in jets initiated by light partons such as the initial state radiation, multiple interactions and pileup events;
- requires at least one b-tagged jet.

In order to verify that data and simulated MC events correctly pass this selection and there is no degradation due to our trigger request, we need to evaluate the trigger efficiency $\epsilon_{trigger}$. It is defined as the ratio between the events that pass both the required trigger and a reference trigger, and those that pass only the reference trigger:

$$\epsilon_{trigger} = \frac{n}{n_0}, \tag{3.1}$$

where:

- n is the number of events that pass the required trigger and the reference one;

- n_0 is the number of events that pass the reference trigger.

The trigger used as a reference trigger is called "`HLT_AK8PFJet200`". Its requests are in fact less strict than those of the required trigger, since it is a HLT trigger which only requires one AK8 jet whose p_T is greater than 200 GeV. Such a trigger, being too loose, includes a suppression ("prescaling") factor which makes it inefficient as signal trigger.

The trigger efficiencies as a function of the p_T, computed respectively for the leading jet and the second jet, are shown in ***Figs. 3.1 and 3.2***. As we can see, starting from values of the p_T greater than 400 GeV, the trigger efficiency as measured in data is compatible with the one measured from MC simulated events. Therefore, there would be no need to introduce any scaling factor between the two efficiencies.

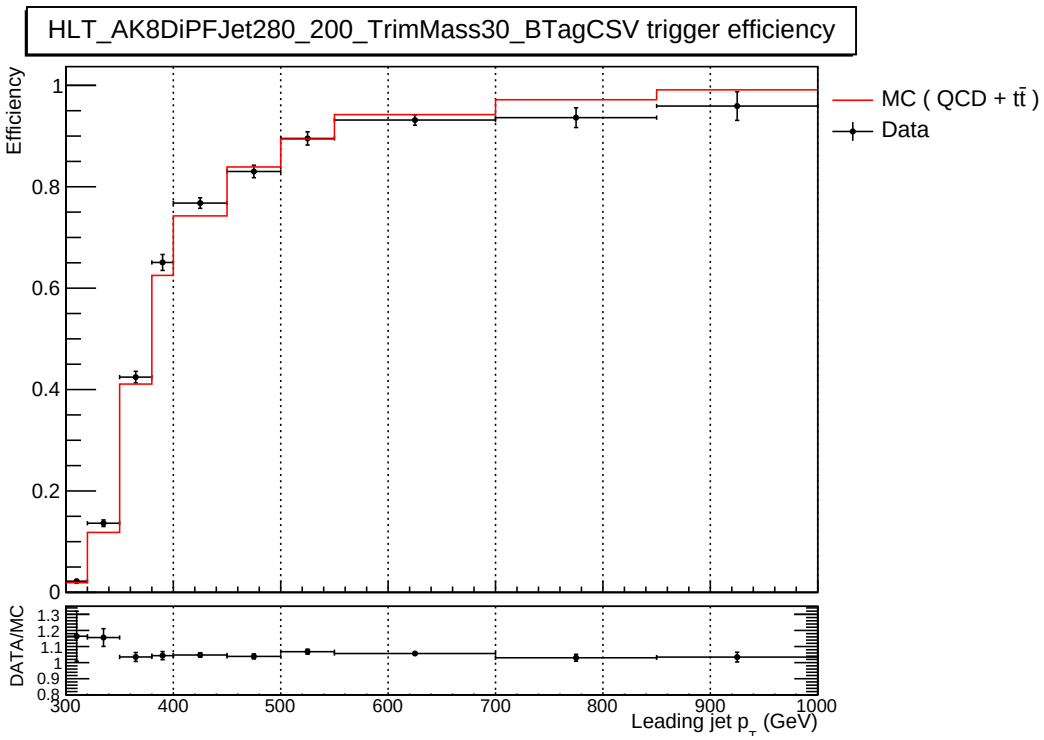

Figure 3.1: *Efficiency of* `HLT_AK8DiPFJet280_200_TrimMass30_BTagCSV` *trigger, measured with data and MC simulated events, as a function of the p_T of the leading jet.*

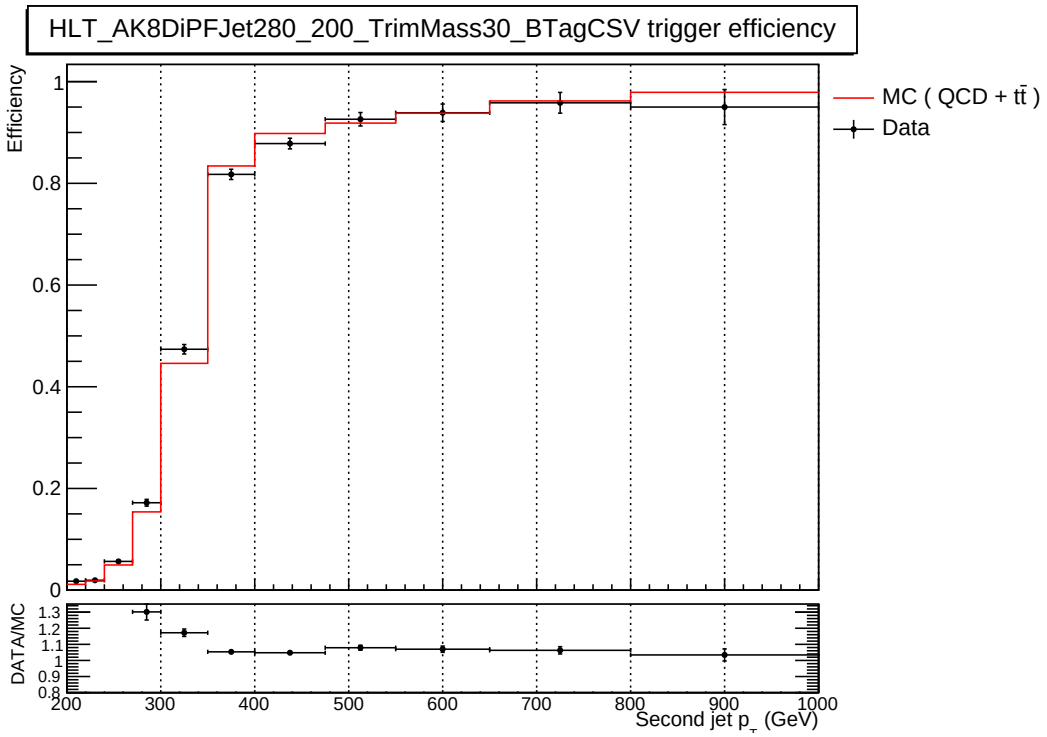

Figure 3.2: *Efficiency of* `HLT_AK8DiPFJet280_200_TrimMass30_BTagCSV` *trigger, measured with data and MC simulated events, as a function of the* p_T *of the second jet.*

3.2.2 Kinematic requests and lepton veto

To properly analyze the all-jets boosted topology, other kinematic requests for the two wide-jets are necessary, in addition to those implemented on the p_T by the trigger. Boosted jets in fact have high p_T, so we require $p_T > 400$ GeV for both jets. Furthermore, boosted jets are highly collimated, which implies that particles cannot be at arbitrary wide angles from the beam axis, but on the contrary as near as possible to it. For this reason, a selection is performed on the pseudorapidity, which is required to be between -2.4 and 2.4 for both jets, corresponding to angles of about $\theta = -10^\circ$ and $\theta = -10^\circ$ from the axis beam.
Furthermore, a lepton veto is required in order to select only the hadronic decays. This is carried out by the request that the number of identified charged leptons $N_{leptons}$ is equal to 0.

3.2.3 b-tagging

The identification of jets that arise from the b quark hadronization, b-jets, is crucial in reducing the overwhelming background, which includes many other non-b jets stemming from gluons, light-flavour quarks (u, d, s) and the c quark fragmentation. The CMS detector, with its precise charged-particle tracking and robust lepton identification systems, is well matched to the task of b-tagging. A variety of algorithms has been developed by CMS to select b-jets based on variables such as the impact parameters of charged-particle tracks, the properties of reconstructed decay vertices, and the presence or absence of a lepton, or combinations thereof. For what concerns the properties of reconstructed decay vertices, it can be used the significance of the flight distance S_{IP}, i.e. the ratio of the flight distance to its estimated uncertainty, as the discriminating variable between b-jets and non b-jets. The efficiencies of those algorithms, generally called "Simple Secondary Vertex" (SSV), are limited by the secondary vertex reconstruction efficiency to about 65% [20]. A more complex approach is used in our analysis. It involves the use of secondary vertices, together with track-based lifetime information. By using these additional variables, the "Combined Secondary Vertex" (CSV) algorithm provides discrimination also in cases when no secondary vertices are found, increasing the maximum efficiency with respect to the SSV algorithms. In many cases, tracks with a $S_{IP} > 2$ can be combined in a "pseudo-vertex", allowing for the computation of a subset of secondary-vertex-based quantities even without an actual vertex fit.
In our analysis, b-tagging is aimed at the subjet structure of the boosted jets, so what we mean here by b-jet is indeed a jet with a b-tagged subjet. We then require the two boosted jets to have specifically a b-subjet associated to a b-quark, so $N_{b-subjets} = 2$. The selection of b-tagged subjets is intended to select only the events in which the top quark decays into a b quark and a W boson (and then the particles produced by the hadronizations are merged in a single wide jet because of the high-p_T of the top quark). As shown in ***Fig. 3.3***, this request rejects the cases where the b-tagged AK8 jets are not associated to a b quark produced along with a W boson: the subjet mass peak at the W mass disappears.

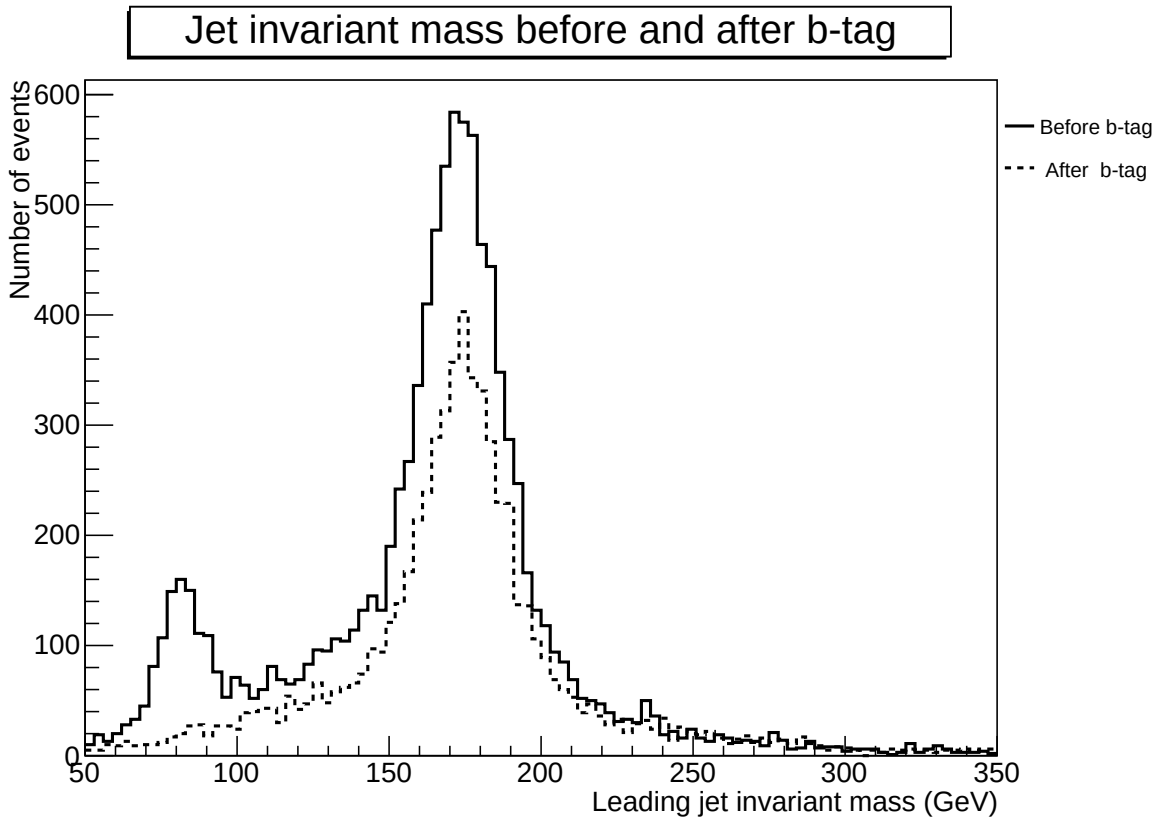

Figure 3.3: *Invariant mass for the leading jet before (solid line) and after (dashed line) the selection of at least one CSV subjet, in simulated $t\bar{t}$ events.*

3.3 Jets reconstruction

In this analysis, jets are reconstructed by a one of the sequential clustering algorithms, namely the "anti-k_t" algorithm, which basically behaves like an idealized cone algorithm and is the mainly used one at CMS. The CMS iterative-cone procedure takes the hardest object (particle, calorimeter tower) in the event, uses it to seed an iterative process of looking for a stable cone, which is then called a jet. It then removes all the particles contained in that jet from the event and repeats the procedure with the hardest available remaining seed, again and again until no seeds remain. The fixed-cone algorithms are similar, but simply define a jet as the cone around the hardest seed, skipping the iterative search for a stable cone [21].

We can introduce distances d_{ij} between entities (particles, pseudojets) i and j and d_{iB} between entity i and the beam B. The clustering proceeds by identifying the smallest of the distances and if it is a d_{ij}, recombines entities i and j, while if it is d_{iB}, calls i a jet and removes it from the list of entities. The distances are recalculated and the procedure repeated until no entities are left.
Distance measures are defined as follows:

$$d_{ij} = \min(p_{T,i}^{2p}, p_{T,j}^{2p}) \frac{\Delta R_{ij}^2}{R^2}, \tag{3.2}$$

$$d_{iB} = p_{T,i}^{2p}, \tag{3.3}$$

where:

- $p_{T,i}$ is transverse momentum of particle i;
- y_i is rapidity of particle i;
- ϕ_i is azimuth of particle i;
- $\Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$;
- R is the radius parameter of the jet;
- p is a parameter which fixes the kind of algorithm:

  $p = 1$ corresponds to k_t algorithm.
  $p = 0$ corresponds to C/A algorithm.
  $p = -1$ corresponds to anti-k_t algorithm.

Negative values of p might at first sight seem pathological but they are not. The behaviour with respect to soft radiation will be similar for all $p < 0$, so here we will concentrate on $p = -1$, and refer to it as the "anti-k_t" jet-clustering algorithm. If there is new soft particle ($p_T \to 0$), meaning $d \to \infty$, it will be clustered last with no effect on hard jets. In anti-k_t algorithm, pairs with a hard particle will cluster first: if no other hard particles are close by, the algorithm will give perfect cones.

In our analysis, jets are reconstructed with the anti-k_t clustering algorithm with distance parameter of $R = 0.8$ (AK8 jets). If a top quark has enough boost its decay products are highly collimated and are reconstructed as a single jet with large distance parameter (hence the usage of AK8 jets). In order to reveal the substructure of such wide jets, we look at the n-subjettiness variable τ_i, defined as

$$\tau_i = \frac{1}{\sum_k p_{T,k} R} \sum_k p_{T,k} \min(\Delta R_{1k}, \Delta R_{2k}, ..., \Delta R_{ik}), \tag{3.4}$$

where the index k enumerates the constituents of the input jet, $p_{T,k}$ is the p_T of the k-th constituent, R is the distance parameter of the original jet, and ΔR_{ik} is the angular distance in the $\phi - \eta$ space between the i-th subjet and the k-th constituent.

The subjettiness variable τ_i measures the compatibility of a jet with the hypothesis that it is composed of i subjets. In the case of exactly i subjets, the value of τ_i tends to zero. In ***Figs. 3.4 - 3.9*** we show our distributions of τ_i for data and MC samples (both $t\bar{t}$ and QCD).

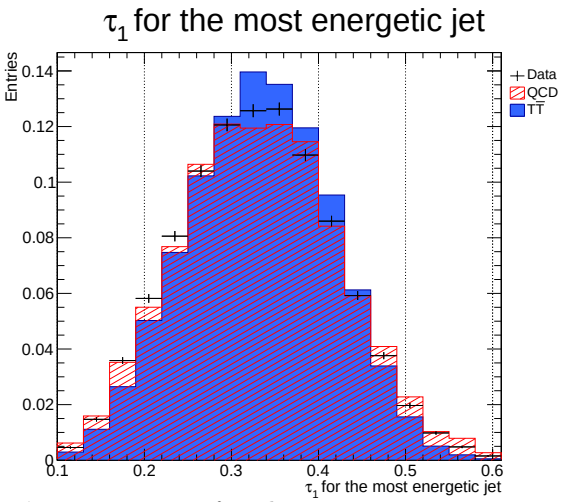

Figure 3.4: *τ_1 for the most energetic jet.*

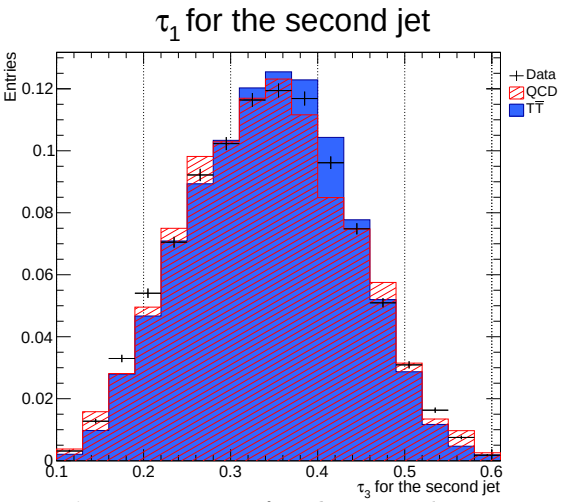

Figure 3.5: *τ_1 for the second jet.*

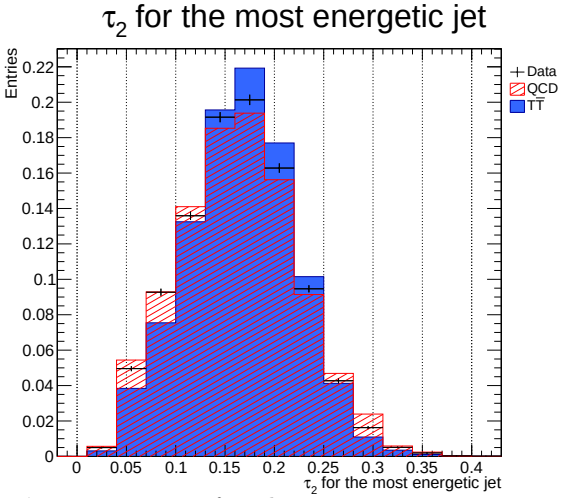

Figure 3.6: *τ_2 for the most energetic jet.*

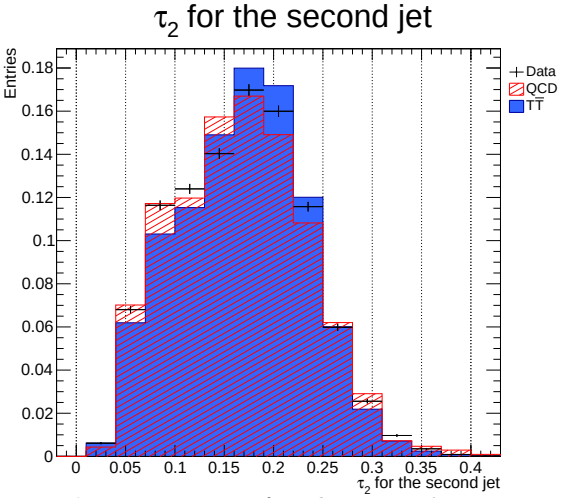

Figure 3.7: *τ_2 for the second jet.*

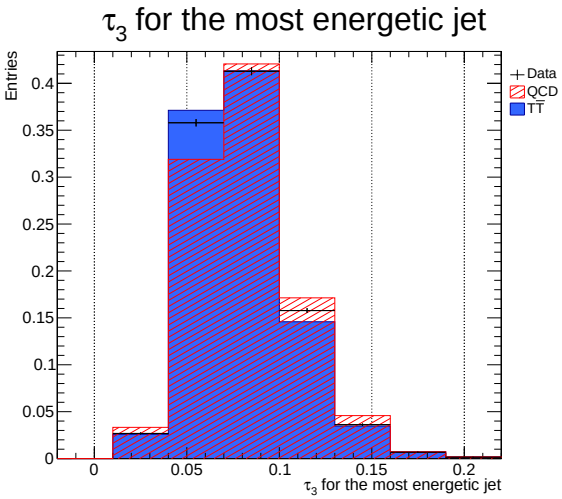

Figure 3.8: *τ_3 for the most energetic jet.*

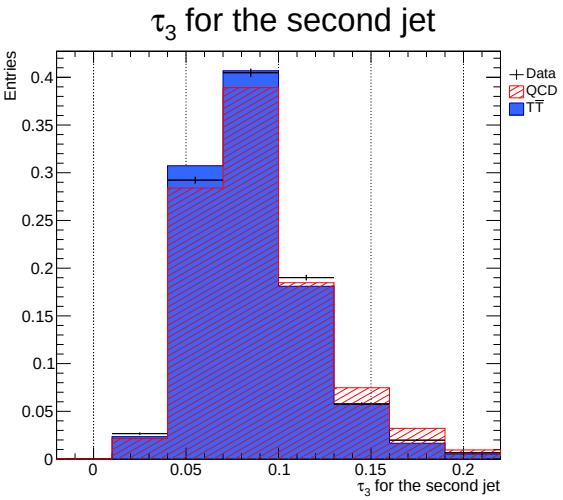

Figure 3.9: *τ_3 for the second jet.*

3.4 Multivariate analysis

In high-energy physics, with the search for ever smaller signals in ever larger data sets, it has become essential to extract the most of the available information from the data. Multivariate classification methods based on machine learning techniques have become a fundamental ingredient to most analyses. Also the multivariate classifiers themselves have significantly evolved in recent years.
Artificial neural networks (ANNs) are computing systems inspired by the biological neural networks that constitute animal brains. The human brain has approximately 100 billion neurons, which communicate through electro-chemical signals. The neurons are connected through junctions called synapses. Each neuron receives thousands of connections with other neurons, constantly receiving incoming signals to reach the cell body. If the resulting sum of the signals surpasses a certain threshold, a response is sent through the axon.
The ANN attempts to recreate the computational mirror of the biological neural network, learning to do tasks by considering examples. An ANN is based on a collection of connected units called artificial neurons. Each connection between neurons can transmit a signal to another neuron. The receiving neuron can process the signals and then signal downstream neurons connected to it. Neurons generally have states represented by real numbers between 0 and 1. Typically, neurons are organized in layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first input to the last output layer, possibly after traversing the layers multiple times, as shown in ***Fig. 3.10***. Each circular node represents an artificial neuron and an arrow represents a connection from the output of one neuron to the input of another.

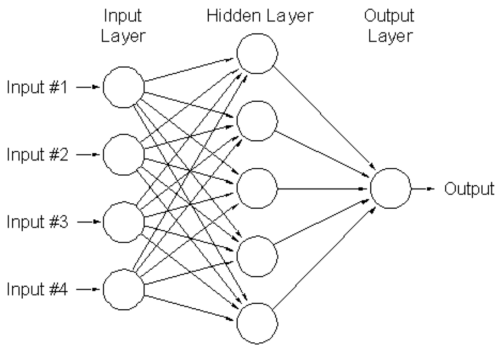

Figure 3.10: *The ANN structure.*

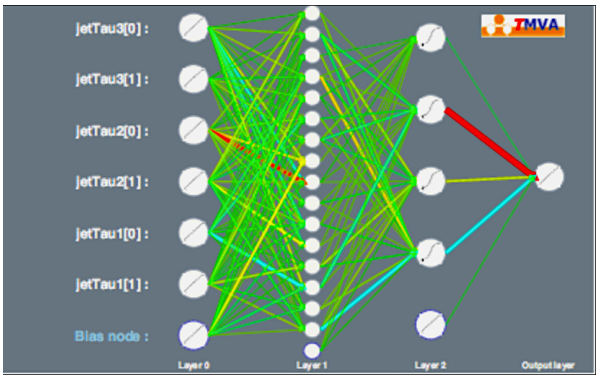

Figure 3.11: *A schematic representation of the neural network used here, with the 6 τ_i of the two jets as input nodes.*

The original goal of the neural network approach was to solve problems in the same way that a human brain would. Over time, attention focused on matching specific mental abilities, leading to deviations from biology such as backpropagation, or passing

information in the reverse direction and adjusting the network to reflect that information. Inputs of our neural network are the τ_i described in the previous section, as shown in ***Fig. 3.11***.

The ANN training consists in finding transfer functions from one node to another in a way to provide separation between signal and background. During the training the network is told which events are to be considered signal (i.e. $t\bar{t}$) and which events are to be considered background (i.e. QCD), then ANN is trained to recognize signal and background events based on a neural network output MVA_{out} which ranges between 0 and 1, with signal events clustered towards $MVA_{out} = 1$ and background events shifted towards $MVA_{out} = 0$. ***Fig. 3.12*** shows the output of the multivariate training.

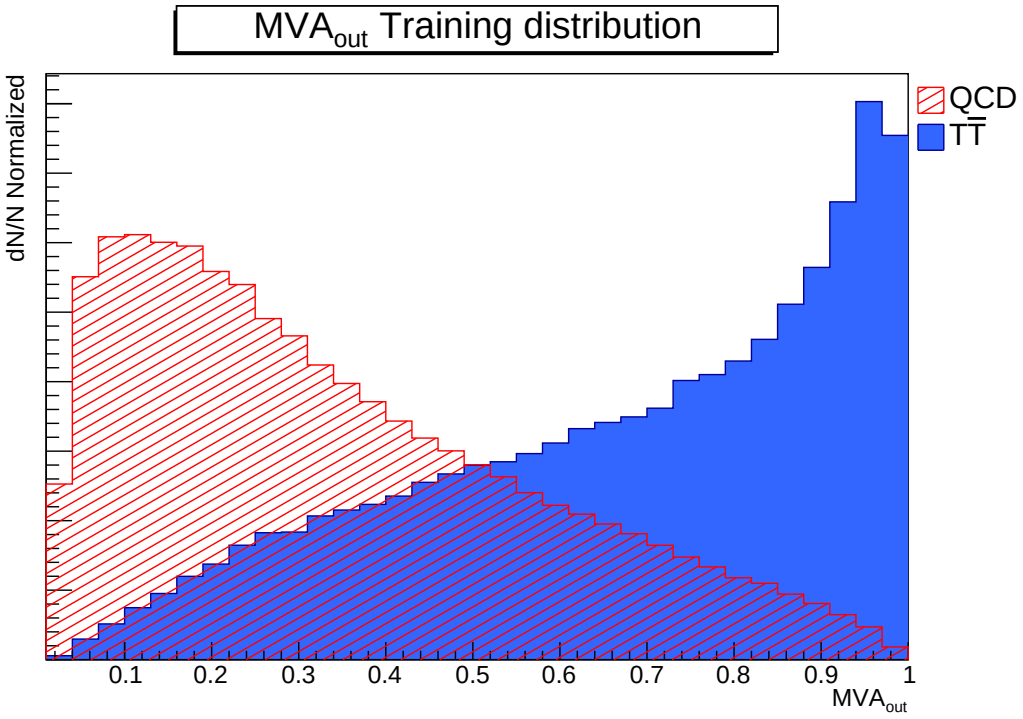

Figure 3.12: *MVA_{out} distributions after the training.*

Both signal and background events are requested to satisfy the previous selection, except for the b-tagging. In ***Figs 3.13 - 3.15*** MVA_{out} distributions are shown for the different b-tagging requests (the absolute normalization of the background distributions is rescaled so that background events summed to the expected $t\bar{t}$ signal events equal the data yield). The $t\bar{t}$ signal clearly emerges from the background requiring two b-tagged jets, as expected from the multivariate training. Therefore, our analysis aims to the MVA_{out} region close to 1, where the $t\bar{t}$ signal significantly emerges from the background, and with the additional request for two b-tagged jets, as occurs in the all-jets boosted topology. It is clear, however, that this procedure of signal maximization, based on

the ANN output and b-tagging requests, leads to systematic uncertainties that are not negligible, and will be discussed later.

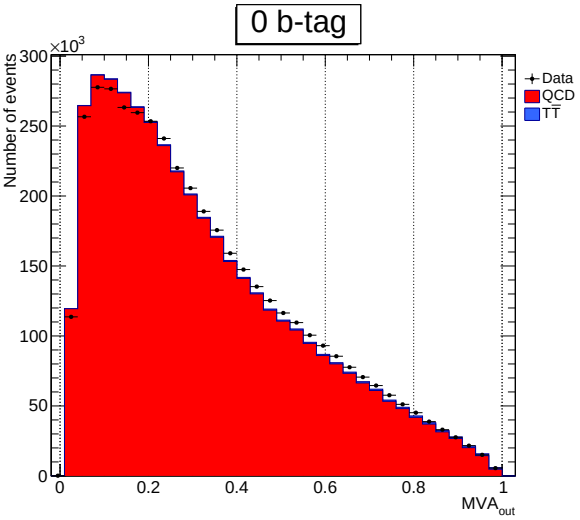

Figure 3.13:
MVA_{out} with the selection of 0 b-jets.

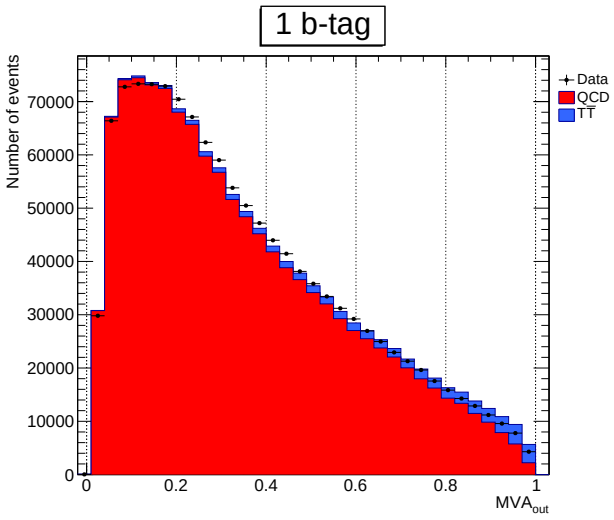

Figure 3.14:
MVA_{out} with the selection of 1 b-jets.

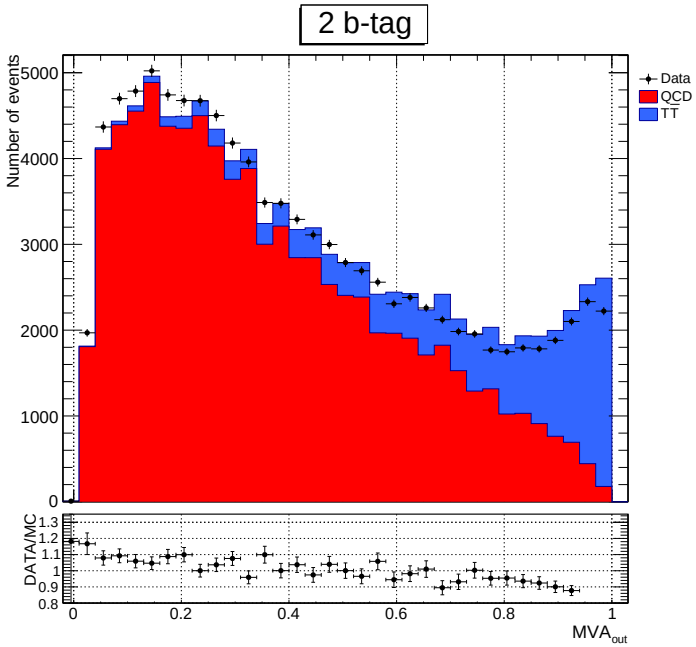

Figure 3.15:
MVA_{out} with the selection of 2 b-jets.

The MVA output layer of the neural network is thus a variable which considers the specific characteristic of the signal events and can be used as the last request for our events. Therefore, we need to evaluate a minimum value of the $\mathrm{MVA_{out}}$ beyond which we accept the events, and this is performed by a maximization approach of the ratio between S and $\sqrt{B}$ ($S/\sqrt{B}$). This optimal minimum value of $\mathrm{MVA_{out}}$ should allow us to have a good signal-to-noise ratio S/B but also a sufficient amount of candidate events that passed the request. The value which maximizes $S/\sqrt{B}$ is in correspondence of the 0.92 output of MVA (see ***Fig. 3.16***). This leads us to the final selection of $\mathrm{MVA_{out}} \geq 0.92$. The signal-to-background S/B distribution as a function of the ANN output is shown in ***Fig. 3.17***. The S/B value is computed with the same normalization of ***Figs 3.13 - 3.15*** and shown for the same corresponding b-tag requests. As we can see, it increases with a proper b-tag request, reducing the generic QCD multijet background. The selection of two b-tagged jets, in addition to $\mathrm{MVA_{out}} \geq 0.92$, leads to S/B greater than 3 (also in agreement with ***Fig. 3.15***).

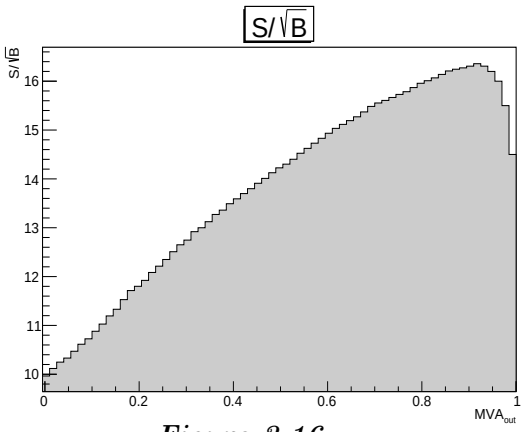

Figure 3.16:
$S/\sqrt{B}$ as a function of the MVA output.

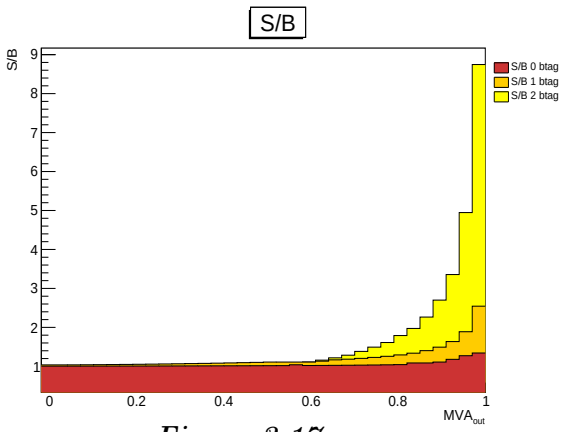

Figure 3.17:
S/B as function of the MVA output, for different b-tag requests.

3.5 Candidate events

Now that we have introduced all the elements on which our events selection is based, we summarize them in ***Table 3.5*** :

Selections
Trigger
$N_{Jets} \geq 2, \quad p_{T1}, p_{T2} \geq 400\,\text{GeV}, \quad \lvert\eta_1\rvert, \lvert\eta_2\rvert < 2.4$
$N_{Leptons} = 0$
$N_{b-subjets} = 2$
$\text{MVA}_{\text{out}} \geq 0.92$

We show in ***Figs. 3.18-3.29*** the distributions of some physical quantities relevant for our analysis, evaluated for the data and the $t\bar{t}$ and QCD background MC samples, corresponding to the selection described in ***Table 3.5*** is applied. The absolute normalization of QCD distributions is rescaled so that QCD events summed to the expected number of $t\bar{t}$ events equal the data yield. All the kinematic distributions observed in data are well reproduced in shape by the sum of $t\bar{t}$ and QCD events.

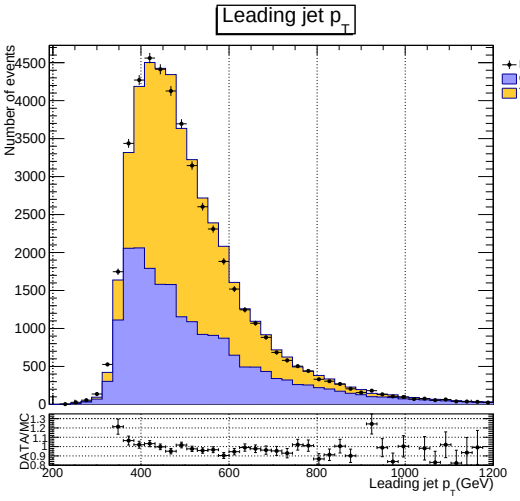

Figure 3.18:
Leading jet p_T.

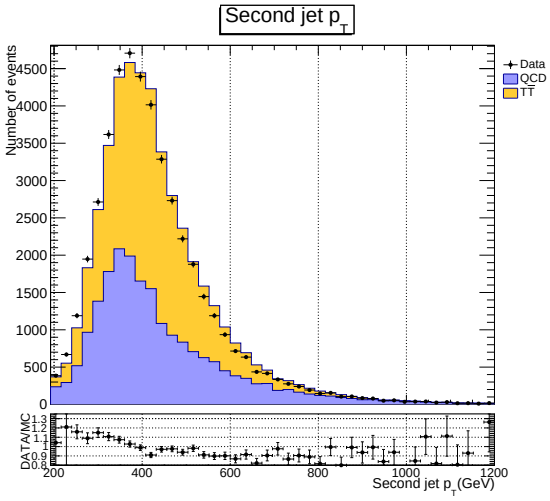

Figure 3.19:
Second jet p_T.

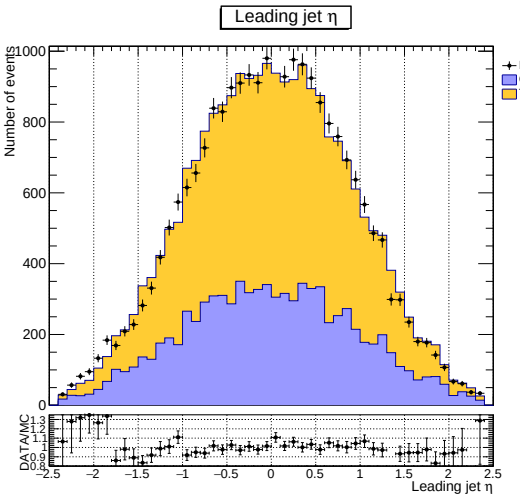

Figure 3.20:
Leading jet η.

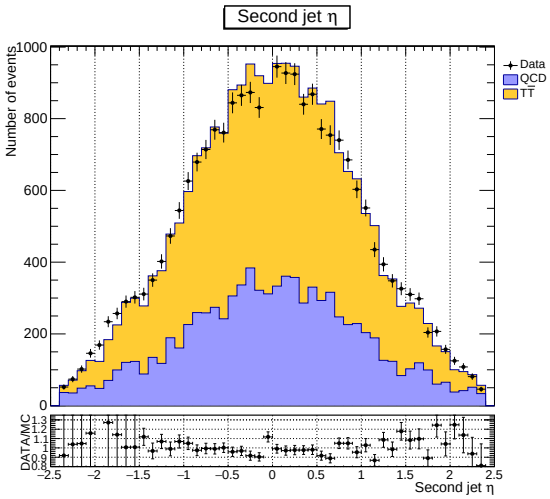

Figure 3.21:
Second jet η.

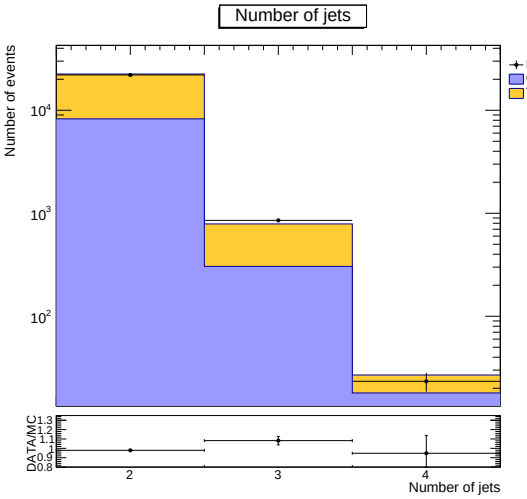

Figure 3.22:
Number of jets.

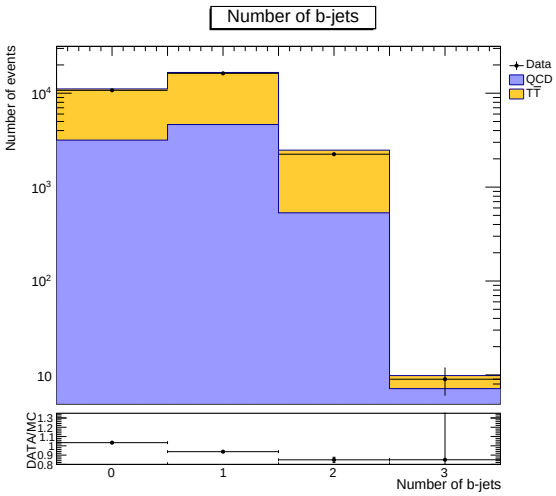

Figure 3.23:
Number of b-jets.

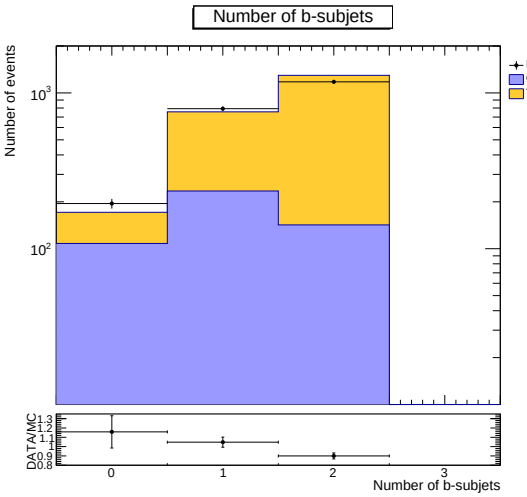

Figure 3.24:
Number of b-subjets.

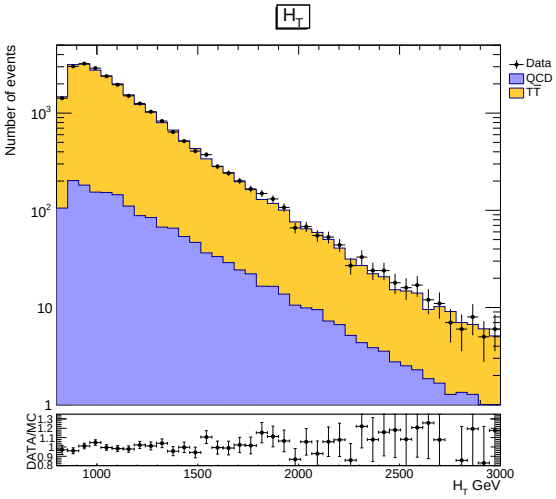

Figure 3.25:
Scalar sum of jet transverse momenta H_T.

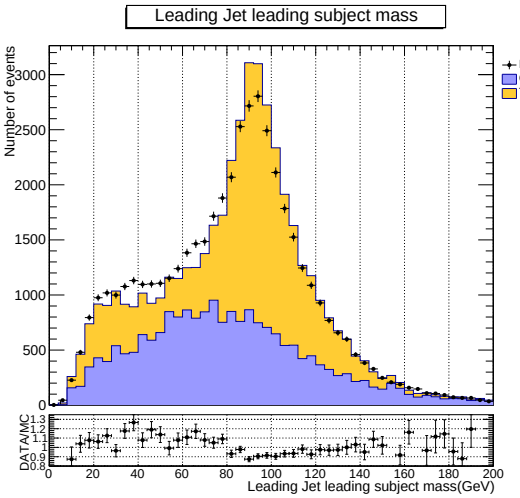

Figure 3.26:
Leading subjet invariant mass for the leading jet.

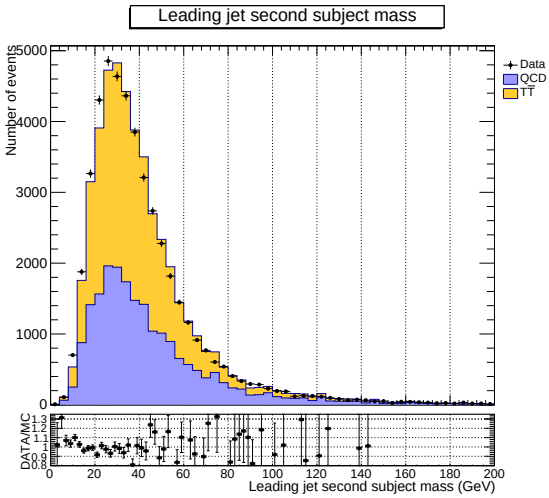

Figure 3.27:
Second subjet invariant mass for the leading jet.

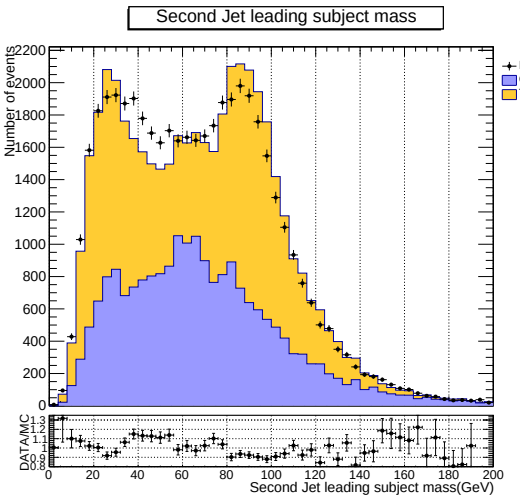

Figure 3.28:
Leading subjet invariant mass for the second jet.

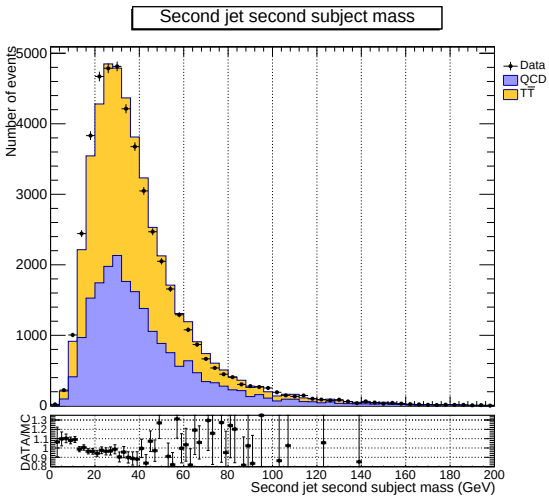

Figure 3.29:
Second subjet invariant mass for the second jet.

3.6 Soft-drop mass

It is essential to introduce a quantity which distinguishes between $t\bar{t}$ and QCD events. This can be chosen to be the jet invariant mass. For the computation of the jet invariant mass, the "soft-drop declustering" algorithm [22] is used, which removes wide-angle soft radiation from a jet in order to mitigate the effects of contamination from initial state radiation (ISR), underlying event (UE), and multiple hadron scattering (pileup). The soft-drop algorithm depends on two parameters, a soft threshold z_{cut} and an angular exponent β. The starting point for soft-drop declustering is a jet with characteristic radius R_0. For definiteness, we will always consider jets defined with the anti-k_t algorithm, but other jet algorithms would work equally well. We then recluster the jet constituents using the C/A algorithm to form a pairwise clustering tree with an angular-ordered structure. The soft-drop declustering is implemented as follows:

1. break the jet j into two subjets by undoing the last stage of C/A clustering. Label the resulting two subjets as j_1 and j_2;

2. if the subjets pass the "soft-drop condition" then deem j to be the final soft-drop jet. Otherwise, redefine j to be equal to subjet with larger p_T and iterate the procedure. The soft drop condition reads:

$$\frac{\min(p_{T1}, p_{T2})}{p_{T1} + p_{T2}} > z_{cut}\Big(\frac{\Delta R_{12}}{R_0}\Big)^{\beta}, \tag{3.5}$$

   where R_0 is the radius of the jet with only two constituents, $p_{T1,2}$ are the transverse momenta of the constituents with respect to the beam, ΔR_{12} is their distance in the rapidity-azimuth plane, z_{cut} is the soft-drop threshold, and β is the angular exponent. By construction this equation fails for wide-angle soft radiation. The degree of jet grooming is controlled by z_{cut} and β, with $\beta \to \infty$ returning back an ungroomed jet.

3. If j is a singleton and can no longer be declustered, then one can either remove j from consideration ("tagging mode") or leave j as the final soft-drop jet ("grooming mode").

In this analysis the default CMS reconstruction is used, taking the parameter values $z_{cut} = 0.1$ and $\beta = 0$.

The soft-drop mass m_{SD} distributions for $t\bar{t}$ signal and for QCD, both with the selection of no b-tag and normalized to the same area, are shown in ***Fig. 3.30***. The $t\bar{t}$ signal distribution is peaked in the region of the mass of top quark, as expected, whereas the QCD distribution is very different and has a more regular shape without a significant peak in the region of top quark mass.

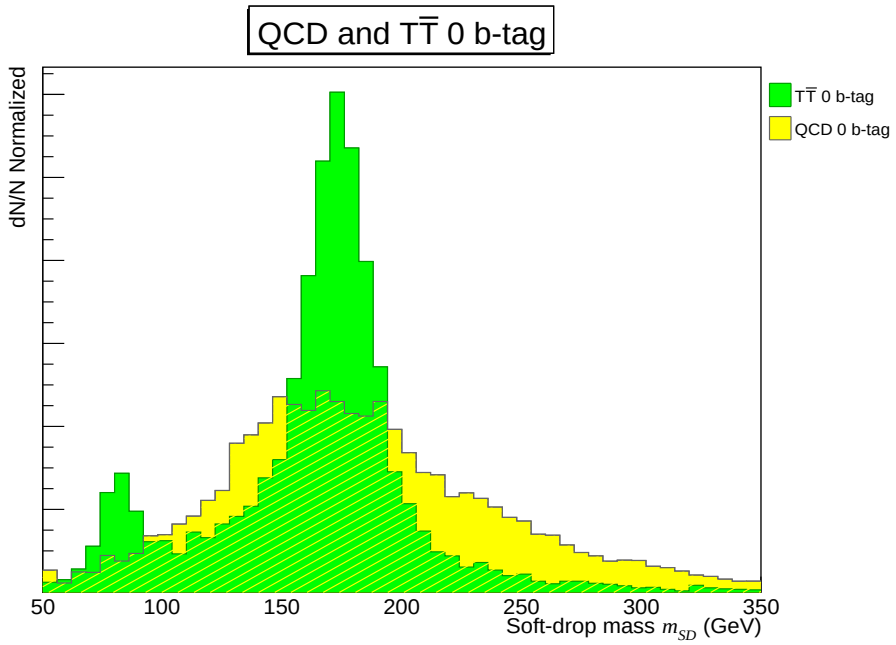

Figure 3.30: *QCD and $t\bar{t}$ signal distributions with the selection of no b-tag as functions of m_{SD}.*

3.7 Background modeling

In order to measure the $t\bar{t}$ production cross section, both the signal and the background distributions are needed. While the signal distribution can be directly obtained from the $t\bar{t}$ sample with the selection of two b-jets, the estimation of the background distribution is a more difficult issue. First of all, background events can be extracted from data or reconstructed from QCD simulations, but it is preferable to use the actual data rather than QCD samples, thus avoiding possible simulation biases. However, in doing so, it is necessary to ensure that the background model we extract from the data is compatible with the expected QCD one. The estimation of QCD background is uncertain because of the uncertainty in the cross section of QCD multijet processes, so we need a proper technique that is able not only to reject as much as possible QCD multijet processes, but also to extract the absolute normalization of the background yield. This can be performed using the soft-drop mass of the two leading boosted jets. The background distribution is derived from data with the selection of no b-tag, in such a way to suppress the presence of signal events as much as possible, and should have a shape similar to the one derived from QCD events with the same no b-tagging request. The comparison between them is shown in ***Fig. 3.31***. The two distributions have a similar shape, however differing by contributions that can be related to the presence of signal events in the data sample. Also considering the contribution of the signal, the sum of QCD and signal events, both

with the selection of no b-jets, is set to be equal to the data yield as shown in ***Figs. 3.32 and 3.33*** for the soft-drop mass m_{SD} of the leading and the second jet.

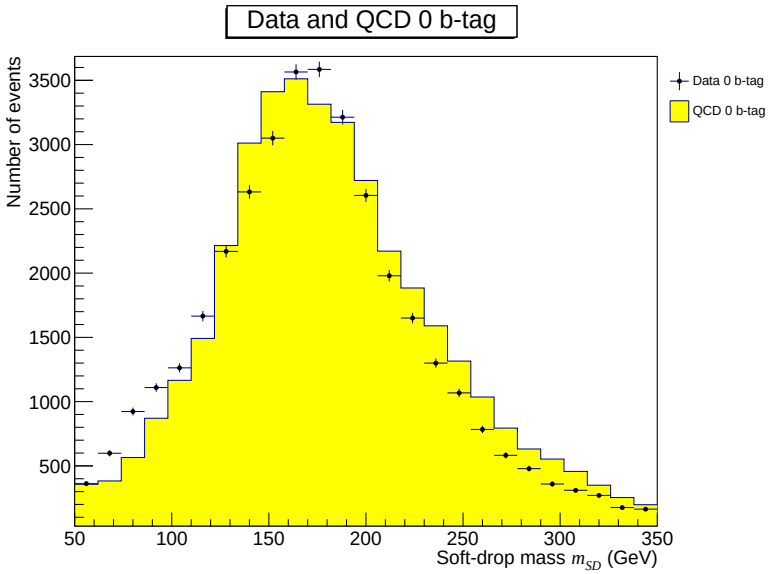

Figure 3.31:
Data and QCD distributions with the selection of no b-jets as functions of m_{SD}.

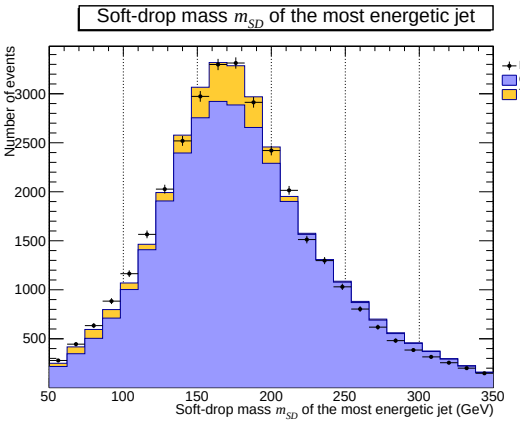

Figure 3.32:
Soft-drop mass m_{SD} of the most energetic jet, with the selection of no b-jets.

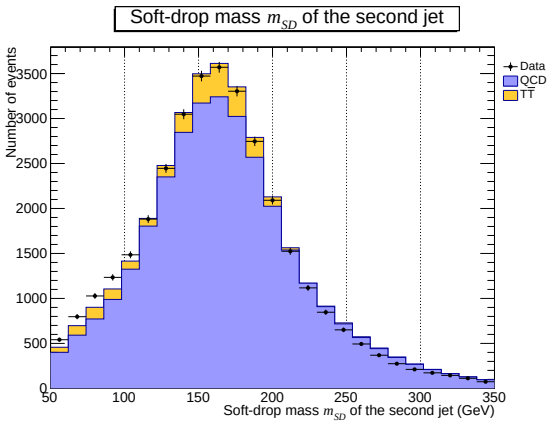

Figure 3.33:
Soft-drop mass m_{SD} of the second energetic jet, with the selection of no b-jets.

As shown in ***Fig. 3.34***, we do not expect any significant bias in the soft-drop mass of QCD events when we require 1 b-jet. Given the reduced size of the sample with the

selection of 2 b-jets, such a conclusion is more difficult to make, but we will treat possible biases as systematic uncertainties. Once we apply the additional selection of 2 b-jets, we obtain for data, signal and QCD the distributions of the m_{SD} of ***Figs. 3.35 and 3.36***.

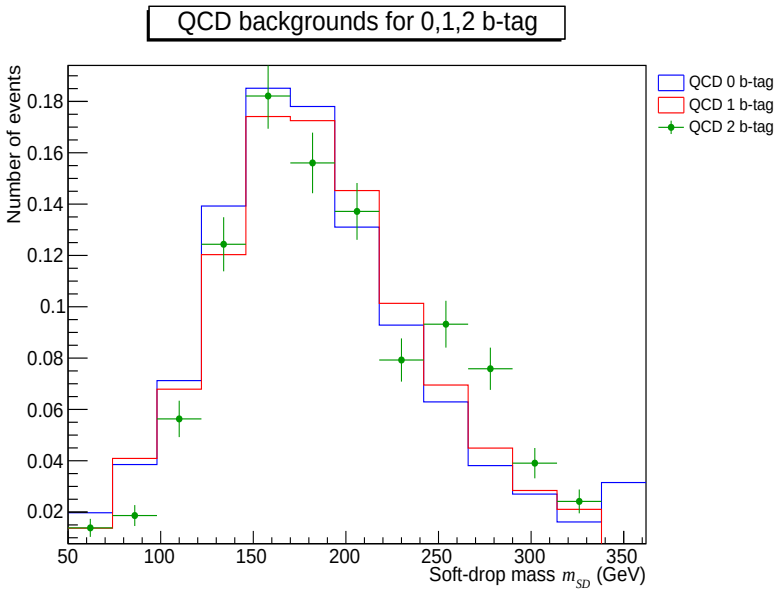

Figure 3.34:
QCD distributions with the selection of 0,1,2 b-jets as functions of m_{SD}.

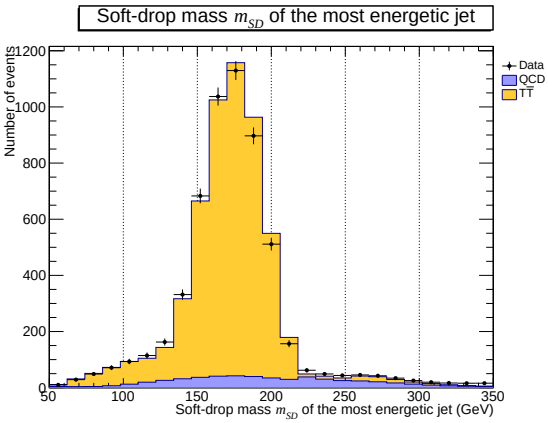

Figure 3.35:
Soft-drop mass m_{SD} of the most energetic jet, with the selection of 2 b-jets.

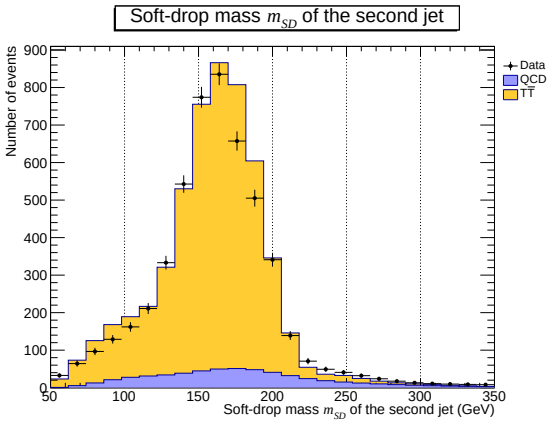

Figure 3.36:
Soft-drop mass m_{SD} of the second jet, with the selection of 2 b-jets.

3.8 The Cross Section

The number $N_{t\bar{t}}^{exp}$ of expected $t\bar{t}$ events at LHC can be determined from the following relation:

$$N_{t\bar{t}}^{exp} = \epsilon \cdot \sigma_{t\bar{t}} \cdot \int L\,dt, \tag{3.6}$$

where:
∘ $\int L\,dt$ represents the integrated luminosity, which, in our case, is $\int L\,dt = 37\,fb^{-1}$;
∘ ϵ is the quantity called efficiency defined as

$$\epsilon = \frac{N_{sel}^{MC}}{N_{gen}^{MC}}, \tag{3.7}$$

where:
∘ N_{sel}^{MC} is the number of MC $t\bar{t}$ events that passed our selection;
∘ N_{gen}^{MC} is the number of $t\bar{t}$ events generated in the MC sample.

On the other hand, knowing the values of ϵ, $\int L\,dt$ and of the observed number of signal events $N_{t\bar{t}}^{obs}$ one can measure the production cross section $\sigma_{t\bar{t}}$ as

$$\sigma_{t\bar{t}} = \frac{N_{t\bar{t}}^{obs}}{\epsilon \int L dt}\,. \tag{3.8}$$

3.9 Likelihood fit

Maximum likelihood methods are powerful tools that can be used to measure parameters from data. Given observations, parameters are estimated by finding the parameter values that maximize the likelihood function. In our case, we perform a likelihood fit to measure $\sigma_{t\bar{t}}$, based on the soft-drop mass distributions of the boosted jets. Since the leading jet and the second jet show a distinctive behaviour for m_{SD}, they will be both used in the fit, that is two m_{SD} entries used for each event.

3.9.1 Template histograms

First of all, a ROOT macro has been used to extract the soft-drop mass m_{SD} for the $t\bar{t}$ signal and QCD background expected after the events selection (see ***Figs. 3.37 and 3.38***).
These two distributions are called "template distributions" because they have been normalized to unit areas, and therefore they represent the two probability distribution functions (pdf) for the $t\bar{t}$ signal and QCD background. Our choice for the background modeling is to use the background as inferred from data with no b-jets. The signal

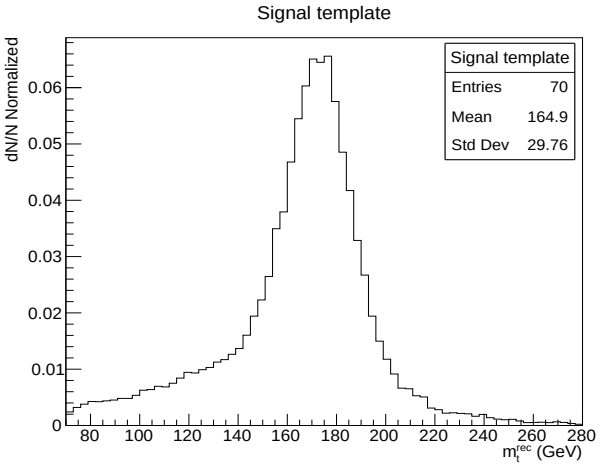

Figure 3.37: *Signal template.*

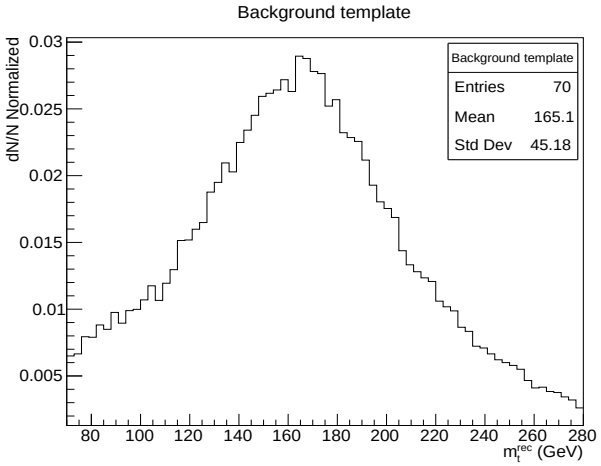

Figure 3.38: *Background template.*

is characterized by a peak in the area around 173 GeV, corresponding to the mass of the quark top. The background is distributed fairly uniformly, growing and decreasing smoothly. The i-th bin content for the pdf of the signal and the background will be indicated in the following with $P_s(i)$ and $P_b(i)$. Their mathematical definitions are:

$$P_s(i) = \frac{n_s^i}{\sum_{i=1}^{N} n_s^i}, \tag{3.9}$$

$$P_b(i) = \frac{n_b^i}{\sum_{i=1}^{N} n_b^i}, \tag{3.10}$$

where N is the number of bins, n_s^i and n_b^i are respectively the numbers of the expected signal entries and of the expected background entries in the i-th bin.

3.9.2 Likelihood fit

In statistics, a likelihood function $\mathcal{L}$ is a function of the parameters of a statistical model. This kind of functions plays a key role in statistical inference, especially methods of estimating a parameter from a set of statistics. Generally, the likelihood of a set of parameter values α, given outcomes x, is equal to the probability of those observed outcomes given those parameter values, that is:

$$\mathcal{L}(\alpha) = P(x|\alpha). \tag{3.11}$$

In our case, the likelihood function has two parameters to estimate the values, $\sigma_{t\bar{t}}$ and n_b, and it is composed of two factors:

$$\mathcal{L}(\sigma_{t\bar{t}}, n_b) = \mathcal{L}_{norm} \cdot \mathcal{L}_{shape}. \tag{3.12}$$

The first one is the normalization factor which has a Poissonian shape:

$$\mathcal{L}_{norm} = e^{-\mu}\frac{\mu^n}{n!}, \tag{3.13}$$

where $n = 9530$ is the number of selected entries, and μ is the sum of the number of observed signal entries n_s and background entries n_b: $\mu = n_s + n_b$.
The second factor is bound to the shape of the binned distributions P_s and P_b:

$$\mathcal{L}_{shape} = \prod_{i=1}^{N} \frac{n_s P_s(i) + n_b P_b(i)}{n_s + n_b}. \tag{3.14}$$

Since we use two entries for each event when filling the m_{SD} distributions, the number of signal entries returned by the fit n_s is twice the number of the observed signal events $N_{t\bar{t}}^{obs}$:

$$n_s = 2 \cdot N_{t\bar{t}}^{obs}. \tag{3.15}$$

The number of signal events, and thus the number of signal entries, can be expressed in terms of the production cross section as

$$N_{t\bar{t}}^{obs} = \sigma_{t\bar{t}} \cdot \epsilon \cdot \int L\mathrm{d}t\,. \tag{3.16}$$

For computational reasons, finding the local minima is easier than finding the maxima, therefore, using MINUIT algorithm, we apply a minimization on $-\log\left[\mathcal{L}(\sigma_{t\bar{t}}, n_b)\right]$, which is allowed by the strictly monotony of the logarithm function. The function to be minimized then reads

$$-\log\left[\mathcal{L}(\sigma_{t\bar{t}}, n_b)\right] = \mu - n\log\mu + \log(n!) - \sum_{i=1}^{N} \log\Big[\frac{n_s P_s(i) + n_b P_b(i)}{n_s + n_b}\Big]. \tag{3.17}$$

Once the likelihood fit is performed, we obtain the m_{SD} distribution shown in ***Fig. 3.39***. The estimated values of $\sigma_{t\bar{t}}$ and n_b are:

$$\sigma_{t\bar{t}} = 554 \pm 12\,(\mathrm{stat})\,\mathrm{pb},$$

$$n_b = 2242 \pm 134.$$

The fit parameters and its results are summarized in the table ***Table 3.2***. The background entries are estimated to be $n_b = 2242 \pm 134$, while the signal contribution corresponds to $n_s = 7288 \pm 152$. This results in a cross section $\sigma_{t\bar{t}}$:

$$\sigma_{t\bar{t}} = \frac{n_s}{2 \cdot \epsilon \cdot \int L\mathrm{d}t} = 554 \pm 12\,(\mathrm{stat})\,\mathrm{pb}. \tag{3.18}$$

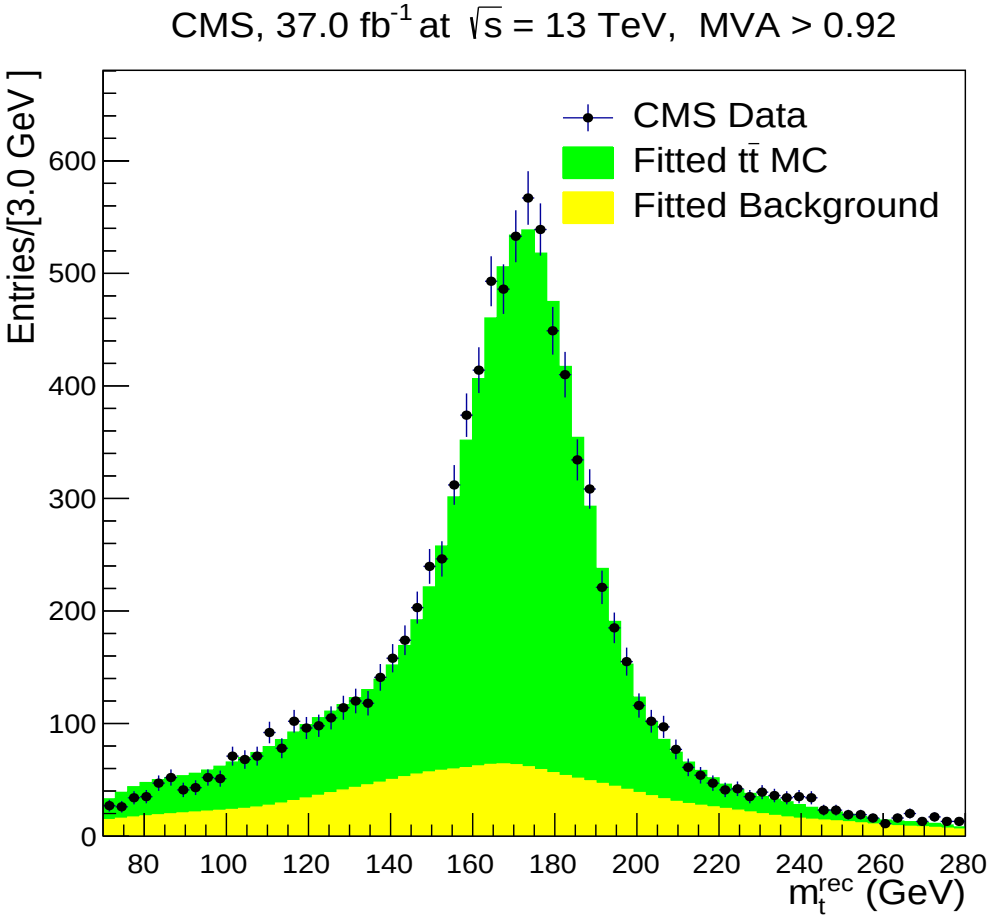

Figure 3.39: *Likelihood fit.*

Parameter	Value
n	9530
n_s	7288 ± 152
n_b	2242 ± 134
$\int L\mathrm{d}t$	$37\,\mathrm{fb}^{-1}$
ϵ	1.78×10^{-4}
$\sigma_{\mathrm{t\bar{t}}}$	$554 \pm 12\,\mathrm{pb}$

Table 3.2: *Likelihood parameters and results of the fit.*

3.9.3 Efficiency correction

The efficiency ϵ of ***Eq. 3.16*** is computed from $t\bar{t}$ simulated events. Such an efficiency refers to the full selection described in ***Table 3.5***. However both the trigger and b-tagging might not be well simulated. For this reason we would need to introduce two MC-to-data scale factors: SF_{trig} and SF_{b-tag}, which improve the estimated MC efficiency as to reflect the value it has on data. From our study in ***Section 3.2.1***, we do not see much difference between the efficiency trigger as evaluated from MC and seen in data, so we assume $SF_{trigger} = 1$.
Studying the b-tag efficiency is, however, a quite complicated issue which has not been carried over in this thesis, so we assume being $SF_{b-tag} = 1$, but we know that values differing by $\approx \pm 5\%$ might be possible. Considering that we require 2 b-jets, the total efficiency we estimate from MC might be off by as much as $\approx 10\%$.

3.10 Systematic uncertainties

The measured cross section $\sigma_{t\bar{t}}$ is affected by several sources of systematic uncertainty. The main ones which we have considered here are associated to:

- "JES", the jet energy scale, dependent on the p_T and η of the jets;
- "JER", the jet energy resolution, dependent on the p_T and η of the jets;

  JES and JER have been corrected to better agree with data but the uncertainties in these corrections have impacts on the measured cross section due to the relatively high p_T requirements on the jets. Jets are shifted according to the JES uncertainty and smeared according to the JER uncertainty. In these two cases, the uncertainties affect both the shape of the soft-drop mass distribution and the efficiency of the selection. To evaluate the corresponding uncertainty, jets in the simulations have a an energy shifted by plus or minus one JES standard deviation, with a resolution shifted by plus or minus one JER standard deviation, and the cross section is measured again. The values obtained for the modified conditions are:

$$\begin{aligned}
\sigma_{t\bar{t}}^{JES-up} &= 545 \pm 11\,\text{pb}, \\
\sigma_{t\bar{t}}^{JES-down} &= 563 \pm 12\,\text{pb}, \\
\sigma_{t\bar{t}}^{JER-up} &= 556 \pm 12\,\text{pb}, \\
\sigma_{t\bar{t}}^{JER-down} &= 555 \pm 12\,\text{pb}.
\end{aligned}$$

  By taking the semi-difference between the extreme values with respect to the average, we estimate a systematic uncertainty of 1.6% for JES and 0.1% for JER:

$$\frac{|\sigma_{t\bar{t}}^{JES-up} - \sigma_{t\bar{t}}^{JES-down}|}{\sigma_{t\bar{t}}^{JES-up} + \sigma_{t\bar{t}}^{JES-down}} = 0.016,$$

$$\frac{|\sigma_{t\bar{t}}^{JER-up} - \sigma_{t\bar{t}}^{JER-down}|}{\sigma_{t\bar{t}}^{JES-up} + \sigma_{t\bar{t}}^{JES-down}} = 0.001;$$

- b-tagging. The uncertainties in the efficiency of b-tagging jets are expected to be quite large, as discussed earlier, and of the order of $\approx 10\%$.

- Background modeling. In ***Section 3.7*** we gave a comparison between the background distribution simulated with MC samples and taken from data with the selection of no b-tag. The two distributions were in good agreement, but however not identical. The uncertainty in the background modeling is estimated from the values of $\sigma_{t\bar{t}}$ obtained by the likelihood fits performed using alternatively data and QCD as backgrounds and different selections of b-jets for the QCD sample. From these values of $\sigma_{t\bar{t}}^{data}$, $\sigma_{t\bar{t}}^{QCD\,0\,b-tag}$, $\sigma_{t\bar{t}}^{QCD\,1\,b-tag}$,$\sigma_{t\bar{t}}^{QCD\,2\,b-tag}$ which are respectively:

$$\begin{aligned}
\sigma_{t\bar{t}}^{data} &= 554 \pm 12\,\text{pb}, \\
\sigma_{t\bar{t}}^{QCD\,0\,b-tag} &= 562 \pm 11\,\text{pb}, \\
\sigma_{t\bar{t}}^{QCD\,1\,b-tag} &= 577 \pm 11\,\text{pb}, \\
\sigma_{t\bar{t}}^{QCD\,2\,b-tag} &= 601 \pm 10\,\text{pb},
\end{aligned}$$

  the uncertainty is estimated considering the $\sigma_{t\bar{t}}^{QCD}$ which differs more from the expected $\sigma_{t\bar{t}}^{data}$ as

$$\frac{|\sigma_{t\bar{t}}^{data} - \sigma_{t\bar{t}}^{QCD\,2\,b-tag}|}{\sigma_{t\bar{t}}^{data}} = 0.085.$$

- Integrated luminosity, amounting to 2.6%, as determined by $x-y$ beam-beam scans utilizing methods from [23];

- Parton shower and hadronization. Here we studied the effect of using different models for the hadronization, by recurring to samples generated with POWHEG+PYTHIA or POWHEG+HERWIG where PYTHIA and HERWIG are used to represent the string and cluster fragmentation models, respectively. The uncertainty is estimated from the values of $t\bar{t}$ cross section obtained by the likelihood fits performed using the two alternative $t\bar{t}$ samples. From these two values, $\sigma_{t\bar{t}}^{PYTHIA}$ and $\sigma_{t\bar{t}}^{HERWIG}$, which are respectively:

$$\sigma_{t\bar{t}}^{PYTHIA} = 554 \pm 12\,\text{pb},$$

$$\sigma_{t\bar{t}}^{HERWIG} = 800 \pm 16\,\text{pb},$$

the uncertainty is estimated as:

$$\frac{|\sigma_{t\bar{t}}^{PYTHIA} - \sigma_{t\bar{t}}^{HERWIG}|}{\sigma_{t\bar{t}}^{PYTHIA}} = 0.44.$$

- ISR and Final state radiation (FSR). To evaluate the corresponding uncertainties, the simulations are changed by increasing/decreasing the amount of the ISR and FSR, and the cross section is measured again. The values obtained are:

$$\begin{aligned}
\sigma_{t\bar{t}}^{ISR-up} &= 563 \pm 12\,\text{pb},\\
\sigma_{t\bar{t}}^{ISR-down} &= 552 \pm 12\,\text{pb},\\
\sigma_{t\bar{t}}^{FSR-up} &= 716 \pm 15\,\text{pb},\\
\sigma_{t\bar{t}}^{FSR-down} &= 462 \pm 10\,\text{pb}.
\end{aligned}$$

  By taking the semi-difference between the extreme values with respect to the average, we estimate a systematic uncertainty of 1% for ISR and 22% for FSR:

$$\frac{|\sigma_{t\bar{t}}^{ISR-up} - \sigma_{t\bar{t}}^{ISR-down}|}{\sigma_{t\bar{t}}^{ISR-up} + \sigma_{t\bar{t}}^{ISR-down}} = 0.01,$$

$$\frac{|\sigma_{t\bar{t}}^{FSR-up} - \sigma_{t\bar{t}}^{FSR-down}|}{\sigma_{t\bar{t}}^{FSR-up} + \sigma_{t\bar{t}}^{FSR-down}} = 0.22.$$

- Missing systematic uncertainties. Other uncertainties which have not been evaluated here include those related to colour reconnection, parton distribution functions, generator modeling, parton shower matching scales.

The main uncertainties are summarized in ***Table 3.3***. The total systematic uncertainty is then estimated as the root sum squared of the all systematic uncertainties. Its value amounts $\approx 51\%$ of the $\sigma_{t\bar{t}}$, which is clearly a large uncertainty. Deeper studies would be needed whether the large contributes have been evaluated properly or if they have been overestimated.

Source of systematic uncertainties	(%)
JES	1.6
JER	0.1
b-tagging	10
Background modeling	8.5
Luminosity	2.6
Parton shower	44
ISR	1
FSR	22
Total systematic unc.	51

Table 3.3: *Systematic uncertainties in the inclusive $t\bar{t}$ production cross section.*

3.10.1 Final result

From the systematic uncertainties summarized in ***Table 3.3***, the final result yields:

$$\sigma_{t\bar{t}} = 554 \pm 12\,(\text{stat}) \pm 283\,(\text{sys})\,\text{pb}. \tag{3.19}$$

Chapter 4

Results and Conclusions

The measurement of the top quark-antiquark pair production cross section provides an important test of perturbative QCD calculations: as a matter of fact several measurements of this cross section have been performed at various values of $\sqrt{s}$, as shown in ***Fig. 4.1***. In addition, $t\bar{t}$ production constitutes an important background source to many new physics scenarios predicted to be accessible at the LHC. In the all-jets final state, the two W bosons decay hadronically and the experimental signature is characterized by the absence of energetic charged leptons and an only-jet topology. The all-jets channel has the advantage of a large branching ratio, 46%, although it suffers from a huge background of multijet QCD events. In this analysis we concentrate on events where the two top quarks, produced with very high p_T, originate wide jets associated to the boost of the particles associated to the top quark decay (the so–called boosted jets). Two such boosted jets are required with $p_T > 400$ GeV, having each a b-tagged subjet. An additional request is based on a neural network which uses subjet characteristics as inputs. The final separation of signal from background events is carried through a likelihood fit based on the distribution of the invariant mass of the two boosted jets. From the number of signal events obtained from the fit we derive the $t\bar{t}$ production cross section. The present measurement, based on data recorded during the 2016 run with pp collisions at $\sqrt{s} = 13$ TeV using the CMS detector, and corresponding to 37 fb^{-1}, yields a cross section value of $\sigma_{t\bar{t}} = 554 \pm 12\,(\text{stat}) \pm 283\,(\text{sys})$ pb, with a preliminary estimate of the systematic uncertainty amounting to 50%. Such a value is about 33% lower than the theoretical one but consistent within uncertainties. We believe the reason might be a systematically higher estimate of the signal efficiency, which could be associated, at least in part, to an overestimated b-tagging efficiency, or the hadronization modeling. Dedicated studies on MC-to-data scale factors would be needed to clarify this point.

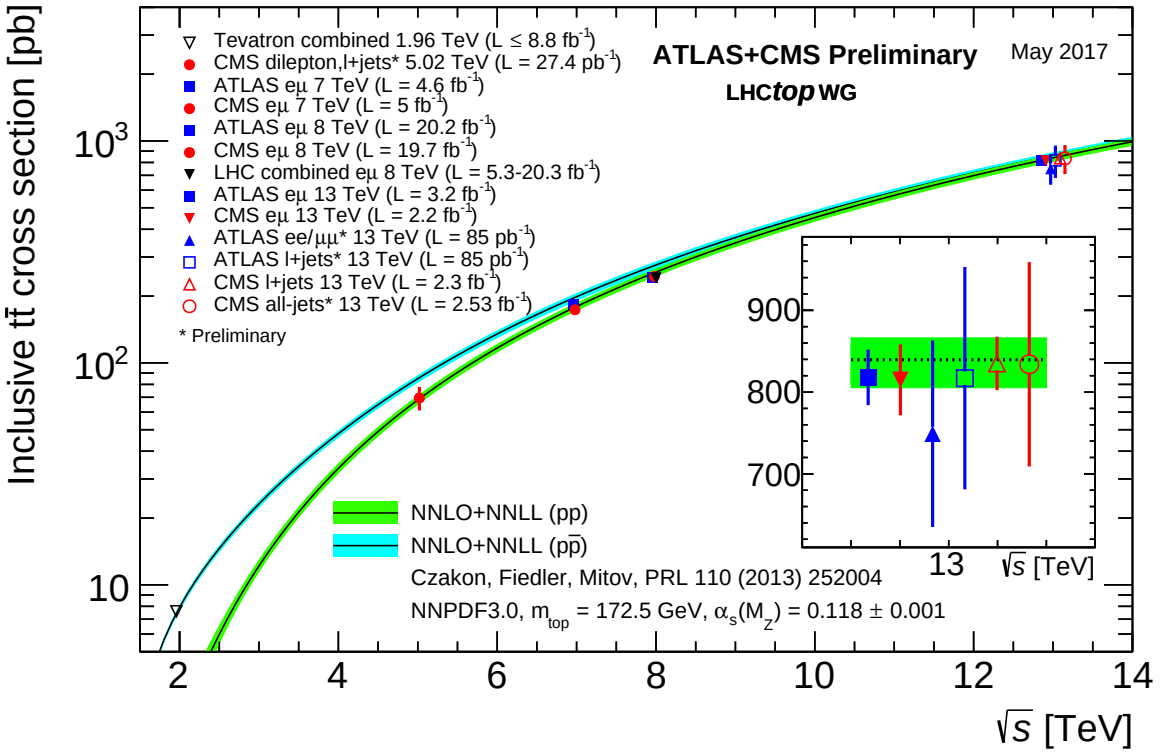

Figure 4.1: *Summary of LHC and Tevatron measurements of the $t\bar{t}$ production cross-section as a function of the $\sqrt{s}$ compared to the NNLO QCD calculation complemented with NNLL resummation. The theory band represents uncertainties due to renormalisation and factorisation scale, parton density functions and the strong coupling. The measurements and the theory calculation are quoted at m_{top}=172.5 GeV.*

List of Figures

List of Tables

Bibliography

[1] Cern website homepage
https://home.cern/about

[2] L. Rossi, "Manufacturing and Testing of Accelerator Superconducting Magnets", CERN Yellow Report CERN-2014-005, arXiv:1501.07164.

[3] 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.

[4] 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, doi:10.1016/j.physletb.2012.08.021, arXiv:1207.7235.

[5] CDF Collaboration, "Observation of Top Quark Production in $p\bar{p}$ Collisions", Phys. Rev. Lett. 74 (1995) 2626, doi:10.1103/PhysRevLett.74.2626, arXiv:hep-ex/9503002.

[6] D0 Collaboration, "Observation of the Top Quark", Phys. Rev.Lett. 74 (1995) 2632, doi:10.1103/PhysRevLett.74.2632, arXiv:hep-ex/9503003.

[7] S. Alekhin, A. Djouadi, S. Moch, "The top quark and Higgs boson masses and the stability of the electroweak vacuum", Phys. Lett. B716 (2012) 214, doi:10.1016/j.physletb.2012.08.024, arXiv:1207.0980.

[8] M.Czakon and A. Mitov, "Top++: A program for the calculation of the top-pair cross-section at hadron colliders", Comput. Phys. Commun. 185 (2014) 2930, doi:10.1016/j.cpc.2014.06.021, arXiv:1112.5675.

[9] U. Husemann, "Top-Quark Physics: Status and Prospects", Prog. Part. Nucl. Phys. 03 002 (2017), doi:10.1016/j.ppnp.2017.03.002, arXiv:1704.01356.

[10] P. Nath, B.D. Nelson, et al., "The Hunt for New Physics at the Large Hadron Collider", Nucl. Phys. Proc. Suppl. 200-202 (2010) 185 doi:10.1016/j.nuclphysbps.2010.03.001, arXiv:1001.2693.

[11] J. M. Butterworth, A. R. Davison et al., "Jet substructure as a new Higgs search channel at the LHC", Phys. Rev. Lett. 100 (2008) 242001, doi:10.1103/PhysRevLett.100.242001, arXiv:0802.2470.

[12] D. Krohn, J. Thaler, L-T Wang, "Jet Trimming", JHEP 1002 (2010) 084, doi:10.1007/JHEP02(2010)084, arXiv:0912.1342.

[13] S. D. Ellis, C. K. Vermilion, J. R. Walsh, "Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches", Phys. Rev. D 81 (2010) 094023, doi:10.1103/PhysRevD.81.094023, arXiv:0912.0033.

[14] 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.

[15] S. Alioli et al., "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX", JHEP 06 (2010) 043, doi:10.1007/JHEP06(2010)043, arXiv:1002.2581.

[16] J. Alwall et al., "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations", JHEP 07 (2014) 079, doi:10.1007/JHEP07(2014)079, arXiv:1405.0301.

[17] The CMS Trigger and Data Acquisition Group, "The CMS High Level Trigger", Eur. Phys. J. C 46, (2006) 605 doi:10.1140/epjc/s2006-02495-8, arXiv:hep-ex/0512077.

[18] A. Perrotta. "Performance of the CMS High Level Trigger", J. Phys. Conf. Ser., 664 (2015) 082044.

[19] CMS Collaboration, F. Beaudette, "The CMS Particle Flow Algorithm", arXiv:1401.8155.

[20] CMS Collaboration, "Identification of b-quark jets with the CMS experiment", JINST 8 (2013) P04013, doi:10.1088/1748-0221/8/04/P04013, arXiv:1211.4462.

[21] M. Cacciari, G. P. Salam and G. Soyez, "The anti-k_t jet clustering algorithm", JHEP 0804 (2008) 063, doi:10.1088/1126-6708/2008/04/063, arXiv:0802.1189.

[22] A. Larkoski, S. Marzani, G. Soyez and T. Jesse, "Soft Drop", JHEP 05 (2014) 146, doi:10.1007/JHEP05(2014)146, arXiv:1402.2657.

[23] CMS Collaboration, "CMS Luminosity Measurement for the 2015 Data Taking Period", CMS Physics Analysis Summary CMS-PAS-LUM-15-001, 2016.