Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea Magistrale in Fisica

Study of $t\bar{t}H$ production at $\sqrt{s} = 13$ TeV in the all-jets channel with CMS

Relatore:
Prof. Andrea Castro

Correlatore:
Dott. Giuseppe Codispoti

Presentata da:
Maria Giovanna Foti

Anno Accademico 2015/2016

Acknowledgments

Contents

Sommario

La misura della produzione associata del bosone di Higgs con una coppia di top quark e antiquark ($t\bar{t}H$) è uno degli obiettivi del Run 2 ad LHC. Lo scopo di questa tesi è quello di compiere una prima ricerca del segnale $t\bar{t}H$, usando tutto il campione di collisioni pp ad una energia nel centro di massa pari a $\sqrt{s} = 13$ TeV raccolto dal rivelatore CMS nel 2016, corripondente ad una luminosità integrata $L = 36$ fb^{-1}. In particolare, l'analisi è condotta e ottimizzata sui campione di dati raccolto nel 2015, corrispondente ad una luminosità integrata $L = 2.63$ fb^{-1}, ma viene fornita alla fine anche una proiezione dei risultati sul campione completo. Gli eventi candidati $t\bar{t}H$ sono selezionati in modo da favorire lo studio del canale completamente adronico del sistema $t\bar{t}$ ed il decadimento del bosone di Higgs in una coppia di quark e antiquark bottom ($H \longrightarrow b\bar{b}$). Nonostante questo sia il canale con il più alto branching ratio e l'unico potenzialmente completamente ricostruibile, lo studio di questo modo di decadimento è particolarmente complicato, a causa della presenza di contributi di fondo che dominano completamente sul segnale atteso. La selezione degli eventi è migliorata attraverso l'impiego di tecniche multivariate ed in particolare di un Boosted Decision Tree. Una prima stima della sensibilità del BDT è presentata, attraverso il parametro di *signal strength* μ. In particolare, è presentato un limite superiore al 95% del livello di confidenza per la sezione d'urto di produzione $t\bar{t}H$ rispetto alle previsioni attese dal Modello Standard, che risulta valere $\mu = 8.2^{+3.4}_{-2.4}$. La proiezione del risultato sull'intero campione di dati è $\mu = 2.1^{+2.1}_{-1.6}$.

Abstract

Measuring the associated production of the Higgs boson and a top quark-antiquark pair is one of the major goals for the Run 2 of the LHC. The aim of this thesis is to report on a first search for $t\bar{t}H$, using pp collision recorded with the CMS detector in the whole 2016, at a centre-of-mass energy of 13 TeV, and corresponding to an integrated luminosity $L = 36$ fb^{-1}. In particular, the analysis is performed and optimised on 2015 data, corresponding to an integrated luminosity $L = 2.63$ fb^{-1}, and then a projection of the results on the full data set is given. Candidate $t\bar{t}H$ events are selected with criteria which enhance the all-jets decay channels of the $t\bar{t}$ system and the decay of the Higgs boson into a bottom quark-antiquark pair ($H \longrightarrow b\bar{b}$). Despite being the channel with the highest branching ratio and potentially fully reconstructable, such decay mode is particularly challenging, due to the completely dominating background contributions. In order to refine the selection, signal and background events are separated using a multivariate approach, using a Boosted Decision Tree. A first estimation of the BDT sensitivity is performed. Such estimate is presented as an upper limit at 95% confidence level on the $t\bar{t}H$ production cross section, with respect to the SM expectations, through the signal strength (μ), and results in $\mu = 8.2^{+3.4}_{-2.4}$. The projection of the result on the full data set is $\mu = 2.1^{+2.1}_{-1.6}$.

Introduction

A new particle with a mass of about 125 GeV was discovered in 2012 at the Large Hadron Collider by the ATLAS and CMS Collaborations. The consistency of this new particle with the Standard Model Higgs boson has been confirmed by an extensive experimental program designed to test all of its properties (its production and decay rates, spin, parity and couplings), but a deviation from the Standard Model expectations could provide hints of new physics.

While the coupling of the new particle to down-type fermions has been verified, the Yukawa coupling to up-type fermions has not yet been measured directly. The top quark is the best candidate for this study, being the heaviest fundamental particle and given the dependence of the Yukawa coupling on the fermion mass.

Therefore, since the Higgs boson mass is smaller than the top quark one, the only direct way to measure the Yukawa coupling to top quarks is by measuring the associated production of the Higgs boson with a top-quark pair ($t\bar{t}H$). The cross section of this process is two orders of magnitude smaller than the gluon-fusion Higgs boson production, making the measurement highly challenging.

The aim of this thesis is to study this process in the all-jets final state, maximizing the expected rate, and selecting the Higgs boson decays into bottom-quark pairs. The project is very ambitious since the backgrounds from QCD multijet production and irreducible $t\bar{t} + b\bar{b}$ events dominate the small signal.

In *Chapter 1* an overview on the Large Hadron Collider and the CMS detector is presented. In *Chapter 2* the necessary theoretical context is given, and the history and the properties of the Higgs boson, the top quark and the $t\bar{t}H$ channels are summarised. *Chapter 3* describes the steps of the analysis performed, explores the event selection and the used MVA techniques in depth. The experimental results, as well as a brief prospect on the future, are presented in *Chapter 4*.

Chapter 1

High-energy physics at the LHC

This chapter describes the Large Hadron Collider (LHC), the CMS experiment and the main focus of CERN research. The LHC is a machine which accelerates two beams of particles in opposite directions to more than 99.9% the speed of light. What we learn from the LHC will take us to a deeper understanding of the Universe and the results could open up new fields of scientific endeavour.

1.1 The Large Hadron Collider

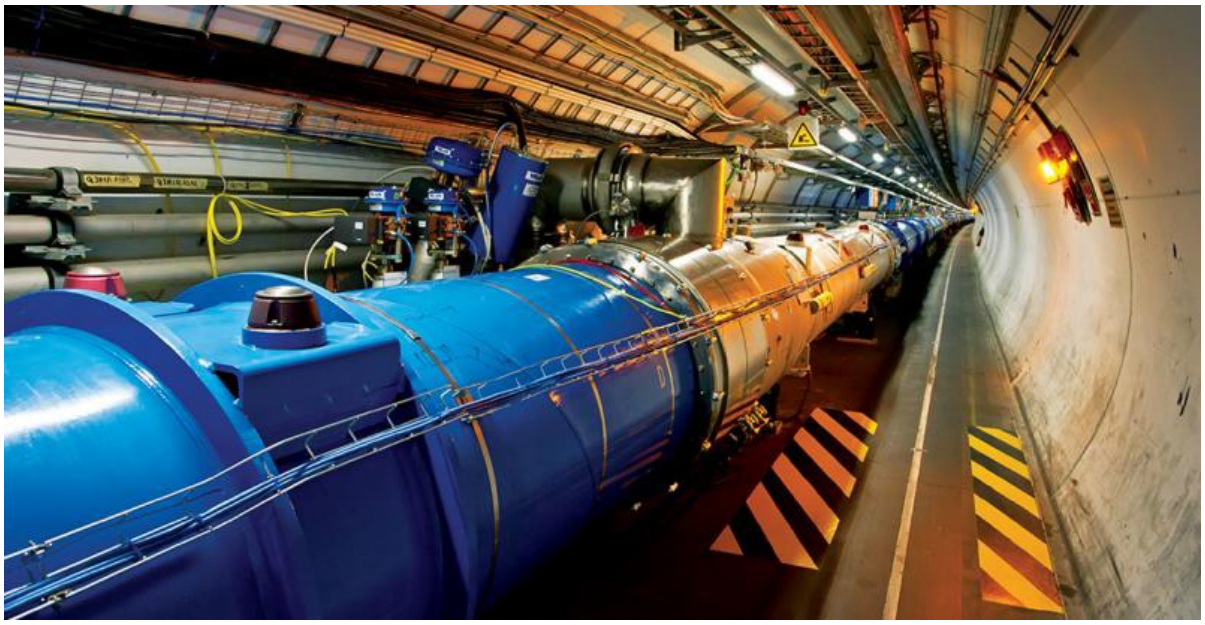


Figure 1.1: The LHC tunnel.

The LHC [1, 2, 3] is the largest and most powerful particle accelerator in the world. LHC began operations on the 10th of September 2008 and is the latest addition to the CERN accelerator complex. The LHC consists of a 27-kilometre ring of superconducting magnets (see Fig. 1.1) with a number of accelerating structures to boost the energy of the particles along the machine.

Inside the accelerator, two high-energy particle beams travel extremely close to the speed of light before they are made to collide. The beams travel in opposite directions in separate beam pipes - two tubes kept at ultrahigh vacuum - and are guided around the accelerator ring by a strong magnetic field produced and maintained by superconducting electromagnets. The electromagnets consist of coils of special electric cable that operates in a superconducting state, efficiently conducting electricity without resistance or loss of energy. This requires cooling the magnets to ≈ 1.9 K - a temperature colder than outer space (2.7 K). For this reason, much of the accelerator is connected to a distribution system of liquid helium that cools both the magnets and the other supply services.

1.1.1 History of the LHC

Back in the early 1980s, while the Large Electron-Positron (LEP) collider was being designed and built, groups at CERN were already looking at the long-term future. Scientists intended to use the existing LEP tunnel to install a higher-energy machine – the LHC. When, on the 21st of October 1993, the US government voted to cancel the Superconducting Super Collider project, due to concerns linked to rising costs, the LHC became the only candidate for a new high-energy hadron collider.

After many years of work on the technical aspects and physics requirements of such a machine, these dreams started to take form on the 16th of December 1994 when CERN governing body, the CERN Council, voted to approve the construction of the LHC. The green light for the project was given under the condition that the new accelerator would not enlarge CERN budget. Therefore, building the accelerator within a constant budget implied that the LHC was to be set up in two stages, with the possibility that any non-Member State contributions would be used to speed up and improve the project. Actually, following contributions from non-Member State such as Japan, the USA, India, Canada and Russia, the Council allowed the project to proceed in a single phase.

In October 1995 the LHC study group published the "*LHC Conceptual Design Report*" that described the architecture and the operation of the accelerator, while between 1996 and 1998, the four experiments ALICE, ATLAS, CMS and LHCb received official approval and construction work started on the four sites.

As construction workers were preparing the work site for the CMS detector cavern in July 1998, they unearthed 4th century Gallo-Roman ruins that turned out to be from an ancient villa with surroundings fields, causing a six-month delay in the work.

On the 2nd of November 2000 the LEP collider was shut down for the last time and with the tunnel available for work, teams began excavating the caverns to house the four

big detectors for the LHC.

In June 2003 the ATLAS detector cavern was complete and on the 1st of February 2005 the completion of the CMS cavern was celebrated, after six and a half years of work. Finally, more than three years later, at 10:28 AM on the 10th of September 2008 a beam of protons successfully circulated around LHC for the first time.

Figure 1.2: The LHC "island" of the CERN Control Centre on the 10th of September 2008. The CCC monitors the performance of the whole accelerator complex operating at CERN.

After an incident occurred on the 19th of September 2008 during powering tests of the main dipole circuit, resulting in mechanical damage and helium release in the tunnel, 37 of the 53 involved magnets were replaced and on the 20th of November 2009 particle beams once again circulated in the LHC. It then started the exciting Run 1, whose most important result was the observation of a particle consistent with the SM Higgs boson publicly announced on the 4th of July 2012.

Run 1 terminated in 2013, with a centre-of mass energy $\sqrt{s} = 8$ TeV and Run 2 began in April 2015 at record energy $\sqrt{s} = 13$ TeV, to further *push back the frontiers of science.*

1.1.2 The LHC machine

The LHC is not a perfect circle: it is made of eight arcs and eight "insertions", as shown in Fig. 1.3. While the arcs contain the dipole bending magnets (154 dipoles per arch), an insertion consists of a straight section, plus two transition regions at its ends. The specific layout of the insertions depend its actual use, that might be physics (when the

experiments take place), injection, beam dumping, or beam cleaning. A sector is defined as the part of the accelerator between two consecutive insertion points.

The eight sectors are the working units of the LHC: the magnet installation is conducted sector by sector, as well as the hardware production. Furthermore, all the dipole magnets of a sector are connected in series and lie in the same continuous cryostat. For this reason, the powering of each sector is essentially independent.

There are three most important elements in a particle accelerator:

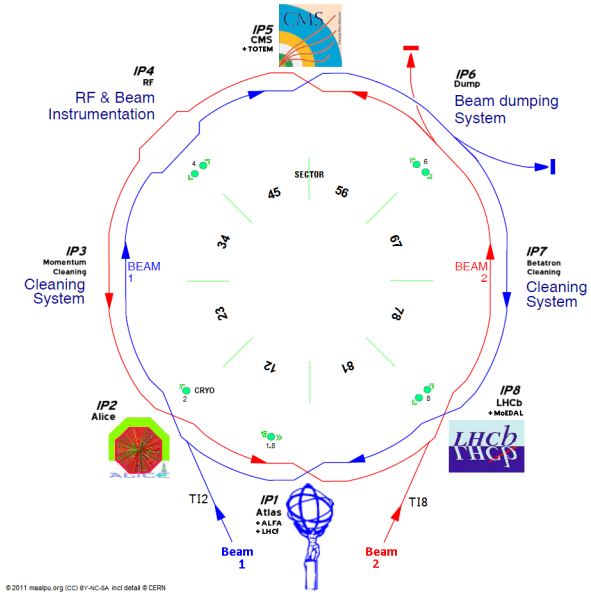


Figure 1.3: The structure of the LHC.

1. **Vacuum** - The LHC has the particular feature of having three separate vacuum systems: the insulation vacuum for cryomagnets, the insulation one for the helium distribution line and finally the beam vacuum. The highest vacuum in the LHC is the beam vacuum, its pressure being 10^{-13} atm (called *ultrahigh vacuum*), since collisions between the hadron beams and gas molecules need to be avoided.

2. **Magnets** - The large variety of magnets in the LHC consists of dipoles, quadrupoles, sextupoles, octupoles, decapoles and gives a total of about 9600 magnets. Each type of magnet contributes to optimizing one aspect of a particle trajectory: dipoles have the function to maintain the beams in their circular orbit, whilst insertion quadrupoles are used to focus the beam down to the smallest possible size at the collision points, maximizing the chance of two protons colliding.

   The dipoles of the LHC represented the toughest technological challenge for the design of the LHC. In a proton accelerator like the LHC, the maximum energy that

Figure 1.4: The first prototype of the 1232 bending-magnets for the LHC was produced by the Italian Institute of Nuclear Physics (INFN). On the 14th of April 1994, magnet training pushed it to a field of 8.73 T (higher than the 8.4 T design field) and LHC Director Lyn Evans received a hand-written note as he sat in meeting that read: "*Message de J.P Goubier et R.Perin à L Evans on a attaint 8.73 tesla 100 quench.*". They meant "sans quench" - a pun on the French word "cent" or "one hundred", which is pronounced the same as the word for "without".

can be achieved is directly proportional to the dipole field, given a specific ray for the accelerator. At the LHC the dipole magnets are superconducting electromagnets which are able to provide the exptional field of 8.3 T over their length. No practical solution could have been designed using ordinary "warm" magnets instead of superconducting ones. The LHC dipoles are made of niobium-titanium (NbTi) cables, that become superconducting below a temperature of 10 K, meaning that they conduct electricity without resistance (please rember that the LHC operates at 1.9 K). The cooling process takes a few weeks to be completed and happens in three phases: firstly the accelerator is brought to 4.5 K, then the magnets are filled with liquid helium and after that, the final cool down to 1.9 K occurs and helium becomes *superfluid*.

The key process that ensures the correct behaviour of the dipoles is the so-called "*training*". Such procedures helps the magnets to maintain the superconducting state, necessary to achieve such high fields. Any abnormal termination of the superconducting state, that switches the magnet back to its resistive state, is called a "quench".

3. **Cavities** - The main role of the LHC cavities is to keep the proton bunches tightly squeezed to ensure high instantaneous luminosity at the collision points and thus maximise the number of collisions. They also accelerate the beams to the required energy by delivering radiofrequency (RF) power. Superconducting cavities with

small energy losses and large stored energy constitute the best solution for the LHC.

1.1.3 The accelerator complex

LHC is not the only accelerator used at CERN, but a succession of machines (see Fig. 1.5) serves the purpose of accelerating the protons beams to increasingly higher energies. Each machine boosts the energy of a beam of particles, before injecting the beam into the following machine in the sequence, up to the record energy of 6.5 TeV per beam.

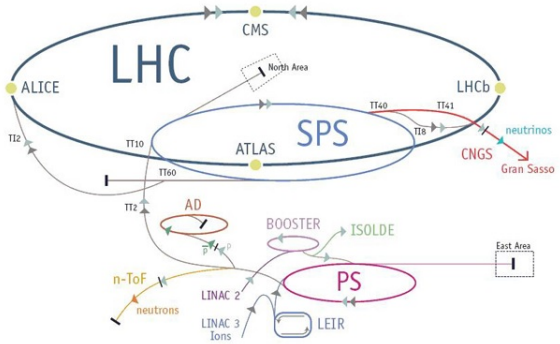


Figure 1.5: There is more to CERN than the LHC: a series of accelerators work together to push particles to nearly the speed of light.

The proton source is a simple tank of hydrogen gas that is ionised through an electric field, separating the atomic elctrons from the protons. The first accelerator the protons go through is Linac 2, that accelerates the protons to the energy of 50 MeV and then injects the beam into the Proton Synchrotron Booster (PSB), which accelerates the protons to 1.4 GeV. The booster is then followed by the Proton Synchrotron (PS), which pushes the beam to 25 GeV and 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 LHC. The beam in one pipe circulates clockwise while the beam in the other pipe circulates anticlockwise. It takes about 4 minutes to fill each ring of the LHC and 20 minutes for the protons to reach their maximum energy of 6.5 TeV. Under normal operating conditions, beams circulate for many hours inside the LHC beam pipes. The two beams are brought into collision inside the four detectors - ALICE, ATLAS, CMS and LHCb - where the total (centre-of-mass) energy at the collision point is of 13 TeV.

The accelerator complex also includes the Antiproton Decelerator and the Online Isotope Mass Separator (ISOLDE) facility and feeds the CERN Neutrinos to Gran Sasso (CNGS) project and the Compact Linear Collider (CLIC) test area, as well as the neutron time-of-flight facility (nTOF).

Protons are not the only particles accelerated in the LHC, since proton-lead and lead-lead collisions are also studied. Lead (Pb) ions for the LHC start from a source of vaporised lead and enter Linac 3 before being accelerated in the Low Energy Ion Ring (LEIR). They then follow the same route to maximum energy as the protons.

1.2 The CMS experiment

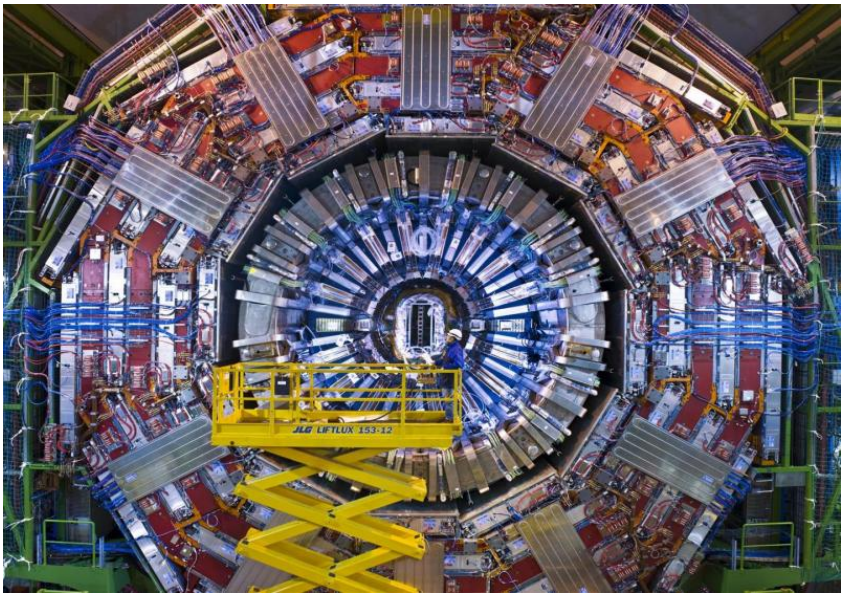


Figure 1.6: The CMS detector has a broad physics programme, and is designed to study particles and phenomena produced in high-energy collisions in the LHC.

The Compact Muon Solenoid (CMS) [4, 5] is a general-purpose detector at the LHC. It is designed to investigate a wide range of physics, ranging from studying and testing the Standard Model with unprecedented precision, to looking for extra dimensions and particles that could constitute dark matter. Despite having the same scientific goals as the ATLAS experiment, it uses different technical solutions and in particular a different design of the magnet-system.

The CMS detector (see Figs. 1.6) is built around a massive solenoidal magnet, that takes the form of a cylindrical coil of superconducting cable generating a field of 3.8 T, which is confined within a steel structure that forms the bulk of the detector 14,000-tonne weight. The complete detector is 21 metres long, 15 metres wide and 15 metres high.

The CMS experiment is one of the largest international scientific collaborations in history (see Fig. 1.7), counting more than 5000 particle physicists, engineers, techni-

cians, students and support staff from more tha 180 institutes in more than 40 countries (November 2016).

Figure 1.7: A small fraction of the CMS Collaboration.

1.2.1 The CMS detector

The detector resembles a giant filter, whose layers are designed to either stop, track or measure different particles emerging from proton-proton and heavy ion collisions. Finding the energy and momentum of a particle gives clues to its identity and particular patterns of particles or "signatures" are indications of new and exciting physics.

The detector consists of several layers of material that exploit the different properties of particles in order to measure the energy and momentum of each one (see Fig. 1.8). For these reasons, CMS needs:

- a high-quality central tracking system to give accurate momentum measurements;
- a high-resolution method to detect and measure electrons and photons (electromagnetic calorimeter);
- a "hermetic" hadron calorimeter, designed to entirely surround the collision and prevent hadronic particles from escaping;
- a high-performance system to detect and measure muons.

Given these priorities, the first essential item is a powerful magnet. Following the law of the Lorentz force, the higher a charged particle momentum, the less its trajectory is bent in the magnetic field, therefore the particle momentum can be easily measured by knowing the details of its curved path. Thus, a strong magnet is needed to accurately measure even the very high momentum particles, such as muons.

Particles emerging from the collision point first encounter a tracker, made of silicon pixels and strips detectors, that measures their positions as they move through the detector providing a measurement of their momentum. Outside the tracker, two calorimeters measure the energy of particles. In measuring the momentum, the tracker should interact with the particles as little as possible, whereas the calorimeters are specifically designed to stop the particles in their paths.

As the name indicates, CMS is particularly designed to measure muons. The muon tracks are measured by several layers of muon detectors, while the neutrinos remain undetected by CMS and their presence can be only indirectly inferred from the momentum imbalance in the event.

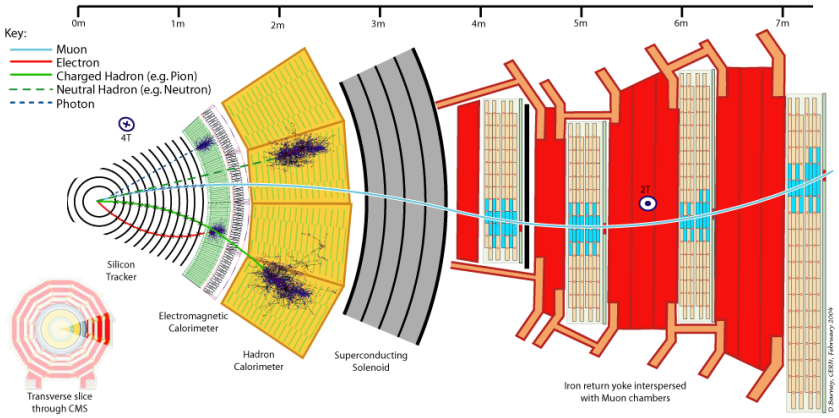


Figure 1.8: Different particles correspond to different trajectories troughout the layers of the CMS detector.

The CMS magnet

The CMS magnet is a solenoid, that is a coil of superconducting wire creating a magnetic field when electricity flows through it. In CMS the solenoid has an overall length of 13 m and a diameter of 7 m and a magnetic field of 3.8 T. It is the largest superconducting magnet ever constructed 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.

The tracker

The momentum of particles is crucial in building a picture of events at the heart of the collision. One method to evaluate the momentum of a particle is to track its path through a magnetic field and the CMS tracker records the paths taken by charged particles by finding their positions at a number of key points. 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 b quarks. Moreover, the tracker needs to record particle paths accurately yet be lightweight so that the particle does not lose a significant amount of energy in it. This job is done by taking position measurements so accurate that tracks can be reliably reconstructed using just a few measurement points with a resolution of $\approx 10\,\mu$m. It is also the innermost layer of the detector and therefore receives the highest flux of particles, and thus the construction materials were carefully chosen to resist radiation.

The final design consists of a tracker made entirely of silicon, with pixels at the very core of the detector and silicon microstrip detectors that surround them. As particles travel through the tracker, pixels and microstrips produce electric signals that are amplified and detected.

The **pixel detector**, about the size of a shoebox, contains 65 million pixels, allowing the tracking of particles emerging from the collision with extreme accuracy. It is also the closest detector to the beam pipe, with three cylindrical layers at 4, 7 and 11 cm and two endcaps at either end, and therefore it is crucial in reconstructing the tracks of very short-lived particles. However, as we have seen, being so close to the collision points means that the number of particles passing through is huge: the flux at 8 cm from the beam line amounts to $10^7\,\text{particles/cm}^2\,\text{s}$. Despite this flux, the pixel detector is able to disentangle and reconstruct all the tracks they leave behind. When a charged particle passes through, it ionizes the silicon atoms, creating electron-hole pairs. These charges are collected on the pixel surface as a small electric signal which is then amplified.

After the pixels, and on their way out of the tracker, particles pass through ten layers of silicon **microstrip detectors**, reaching out to a radius of 1.3 m and having both a barrel and an endcap part, with a total of 10 million detector strips read by 80,000 microelectronic chips. Each module consists of three elements: a set of sensors, its mechanical support structure and readout electronics.

For typical events under average 2011 pileup conditions, the average track-reconstruction efficiency for promptly-produced charged particles with transverse momenta of $p_T > 0.9$ GeV is 94% for pseudorapidities of $|\eta| < 0.9$ and 85% for $0.9 < |\eta| < 2.5$. The inefficiency is caused mainly by hadrons that undergo nuclear interactions in the tracker material. For isolated muons, the corresponding efficiencies are essentially 100%. For isolated muons of $p_T = 100$ GeV emitted at $|\eta| < 1.4$, the resolutions are approximately 2.8% in p_T, and respectively, 10 μm and 30 μm in the transverse and longitudinal impact parameters. The position resolution achieved for reconstructed primary vertices that correspond to interesting pp collisions is 10–12 μm in each of the three spatial dimensions.

The ECAL

After the tracker, the Electromagnetic Calorimeter (ECAL) is placed, in order to measure the energy of photons, electrons and positrons. It occupies the cilindrical space between radii 1.30 m and 1.70 m and it is a hermetic and homogeneous calorimeter. Measuring the energy of such particles with the necessary resolution in the extreme conditions of the LHC – a high magnetic field, high levels of radiation and only 25 ns between collisions – requires uncommon detector materials. In particular, crystals of lead tungstate ($PbWO_4$) are made primarily of metal and are heavier than stainless steel, but with a small fraction of oxygen they become highly transparent and scintillate when electrons and photons pass through them, producing light in proportion to the particle energy. Photodetectors are 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 processed. The ECAL, consisting of a barrel section and two endcaps, forms a layer outside the tracker. The cylindrical barrel consists of 61,200 crystals formed into 36 supermodules, each containing 1700 crystals, while the flat ECAL endcaps are made up of almost 15,000 further crystals.

The HCAL

Following a particle emerging from the ECAL, the next layer is constituted by the Hadron Calorimeter (HCAL), that measures the energy of hadrons and provides indirect measurement of the presence of non-interacting, uncharged particles such as neutrinos. It is a sampling calorimeter, made by alternating layers of absorber (stainless steel and brass) and fluorescent plastic scintillator materials that produce a rapid light pulse when the particle passes through. Special optic fibres collect such light and feed it into the readout electronics where photodetectors amplify the signal. The HCAL extends between radii 1.77 m and 2.95 m. Measuring hadrons is important as they can tell us if interesting particles such as the Higgs boson or even supersymmetric particles have been formed. As these particles decay, they may produce new particles that do not leave record of their presence in any part of the CMS detector, but a momentum imbalance. Actually, in order to spot these particles the HCAL must be "hermetic", so as to capture every particle emerging from the collisions. In this way if any imbalance in the momentum and energy (measured in the "transverse" direction relative to the beam line) is observed, "invisible" particles have been produced.

Quantitatively speaking, as far as electrons are concerned, the energy scale is calibrated with an uncertainty smaller than 0.3%. In particular, the energy resolution for electrons produced in Z boson decays ranges from 1.7 to 4.5%, depending on the electron pseudorapidity and energy loss through bremsstrahlung in the detector material.

The jet energy resolution, instead, is measured in data and simulated events and is studied as a function of pileup, jet size and jet flavour. Typical jet energy resolutions at central rapidities are 15–20% at 30 GeV, about 10% at 100 GeV and 5% at 1 TeV.

The studies exploit events with dijet topology, as well as photon+jet, Z+jet and multijet events. The final uncertainties on the jet energy scale are below 3% across the phase space considered by most analyses ($p_T > 30$ GeV and $|\eta| < 5.0$). In the barrel region ($|\eta| < 1.3$) an uncertainty below 1% for $p_T > 30$ GeV is reached.

The muon detectors

The final layers of the detector house the muon system. As the name "Compact Muon Solenoid" implies, muon detection is crucial for the CMS experiment, since muons take part in several important experimental signatures of interesting physics such the Higgs boson sector. Because muons can penetrate several metres of iron without interacting, the chambers that detect muons are placed at the very edge of the detector where they are the only particles likely to produce a signal. A particle is measured by fitting a curve to hits among the several muon stations. There are 1400 muon chambers in total: 250 *drift tubes* (DTs) and 540 *cathode strip chambers* (CSCs) track the particles positions and provide a trigger, while 610 *resistive plate chambers* (RPCs) form a redundant trigger system, quickly deciding whether to keep the acquired muon data or not. Because of the many layers of detector and different specialities of each type, the system is naturally robust and able to filter out background noise. DTs and RPCs are arranged in concentric cylinders around the beam line ("the barrel region") whilst CSCs and RPCs, make up the endcaps disks that cover the ends of the barrel. The spatial resolution per chamber is 80–120 μm in the DTs, 40–150 μm in the CSCs, and 0.8–1.2 cm in the RPCs. The time resolution achievable is 3 ns or better per chamber for all 3 systems. The efficiency for reconstructing hits and track segments originating from muons traversing the muon chambers is in the range 95–98%. The CSC and DT systems provide muon track segments for the CMS trigger with over 96% efficiency, and identify the correct triggering bunch-crossing in over 99.5% of such events.

The trigger

Within the LHC, bunches of particles collide up to 40 million times per second, so a "trigger" system that saves only potentially interesting events is essential. This reduces the number of events recorded from one billion to around 100 per second.
When CMS is performing at its peak, $\approx 10^9$ proton-proton interactions/s take place inside the detector. It is not possible that data from all these events could be read out and even if they could, most of the events would be low-energy glancing collisions for instance, rather than energetic, head-on interactions. We therefore need a "trigger", selecting the potentially interesting events, such as those which will produce the Higgs boson, and reduce the rate to just a few 10^2 events/s, which can be read out and stored on computer disks for subsequent analysis. However, with groups of protons colliding at a frequency of 40 MHz, there are only 25 ns before the next lot arrive. New particles

are being generated before those from the last event have even left the detector! The solution is to store the data in pipelines that can retain and process information from many interactions at the same time. In order not to confuse particles from two different events, the detectors must have great time resolution and the signals from the millions of electronic channels must be synchronised so that they can all be identified as being from the same event.

1.3 The LHC potential

Since the 1970s, the fundamental structure of matter has been described using the elegant equations of the SM. The model describes how everything we observe in the Universe is made from a few fundamental particles, governed by four forces. Particle accelerators and detectors at CERN are used to test the predictions and limits of the SM that has explained many experimental results and precisely predicted a range of phenomena over the years, such that it is considered a well-tested physics theory.

However, our current understanding of the Universe through the SM is incomplete. The SM has been tested by various experiments and it has proven sucessful in anticipating the existence of previously undiscovered particles. Yet the model describes only 4% of the matter of the known Universe and leaves many unsolved questions. Will we see a unification of forces at the high energies of the LHC? Why is gravity so weak? Why is there more matter than antimatter in the Universe? Is there more exotic physics waiting to be discovered at higher energies? Will we discover evidence for Supersymmetry at the LHC?

On the 4th of July 2012, the ATLAS and CMS experiments at CERN announced the discovery of a Higgs boson, a particle with a mass of about 125 GeV [15, 16]. The Higgs boson is the simplest manifestation [10, 11, 12, 13] of the Brout-Englert-Higgs mechanism that gives mass to particles. It is the final particle in the SM to be experimentally verified. Increasing the energy of the LHC will increase the chance of creating Higgs bosons in collisions, which means more opportunity to measure the Higgs boson precisely and to probe its extremely rare decays. High-energy collisions could detect small, subtle differences between what the boson looks like in experiments, and what the Standard Model predicts.

In early 2013, after three years of running, the LHC shut down for planned maintenance. Hundreds of engineers and technicians spent two years repairing and strengthening the accelerator in preparation for running at the energy of 13 TeV – almost double its previous energy.

Yet, it remains difficult to construct a theory of gravity similar to those for the other forces in the SM. Supersymmetry – hypothesising the existence of more massive partners of the standard particles we know – could facilitate the unification of fundamental forces. If Supersymmetry is right, then the lightest supersymmetric particles should be found

with the new energies at the LHC. Another possible theory of gravity could be based on the existence of extra dimensions and we could find evidence of particles that can exist only if extra dimensions are real. Theories that require extra dimensions predict that there would be heavier versions of standard particles in other dimensions. Again, such heavy particles could be revealed at the high energies the LHC will reach in Run 2.

As we have previously seen, cosmological and astrophysical observations have shown that all of the visible matter accounts for only 4% of the Universe. The search is open for particles or phenomena responsible for dark matter (23%) and dark energy (73%). A very popular idea is that dark matter is made of neutral – but still undiscovered – supersymmetric particles. The first hint of the existence of dark matter came in 1933, when astronomical observations and calculations of gravitational effects revealed that there must be more matter present in the Universe than we could account for by sight. Researchers now believe that the gravitational effect of dark matter makes galaxies spin faster than expected and that its gravitational field deviates the light of objects behind it. Measurements of these effects show the existence of dark matter and can be used to estimate its density even without the possibility of observing it directly. Dark energy, on the other hand, is a form of energy that appears to be associated with the vacuum in space, and is homogenously distributed throughout the Universe and in time. In other words, its effect is not diluted as the Universe expands. The even distribution means that dark energy does not have any local gravitational effects, but rather a global effect on the Universe as a whole. This leads to a repulsive force, which tends to accelerate the expansion of the Universe. The rate of expansion and its acceleration can be measured by experiments using the Hubble law. These measurements, together with other scientific data, have confirmed the existence of dark energy and have been used to estimate its quantity.

The LHC will also help us investigate the mystery of antimatter. Matter and antimatter must have been produced in the same amount at the time of the Big Bang, but from what we have observed so far, our Universe is made only of matter. Why? The LHC could help to provide an answer. It was once thought that antimatter was a perfect "reflection" of matter – that if you replaced matter with antimatter and looked at the result as if in a mirror, you would not be able to tell the difference. We now know that the reflection is imperfect and this could have led to the matter-antimatter imbalance in our Universe. The strongest limits on the amount of antimatter in the Universe come from the analysis of the "diffuse cosmic gamma rays" and the inhomogeneities of the cosmic microwave background (CMB). Assuming that after the Big Bang, the Universe separated somehow into different domains where either matter or antimatter was dominant, it is evident that at the boundaries there should be annihilations, producing cosmic (gamma) rays. Taking into account annihilation cross sections, distance and cosmic redshifts, this leads to a prediction of the amount of diffuse gamma radiation that should arrive on Earth. The free parameter in the model is the size of the domains. The comparison with the observed gamma-ray flux leads to an exclusion of any domain size below 3.7

giga light years, which is not so far away from the entire Universe. Another limit comes from analyzing the inhomogeneities in the CMB: antimatter domains (at any size) would cause heating of domain boundaries and show up in the CMB as density fluctuations. The observed value of 10^{-5} sets strong boundaries to the amount of antimatter in the early Universe.

In addition to the studies of proton-proton collisions, heavy-ion collisions at the LHC will provide a window onto the state of matter that would have existed in the early Universe, called "quark-gluon plasma". When heavy ions collide at high energies they form an instantaneous "fireball" of hot, dense matter that can be studied by the experiments and the higher energy collisions at the LHC will allow new and more detailed characterization of this quark-gluon plasma.

Chapter 2

The top quark and the Higgs boson

In this chapter the top quark and the Higgs boson will be described. In particular their properties, discovery and phenomenology will be discussed.

2.1 The top quark

The top quark, also known as the t quark, is one of the six existing quarks in the SM. According to the SM, it belongs to the third family, along with its paired quark (in the weak ispospin doublet): the bottom or b quark. Like all quarks, the top quark, as well as its antiparticle, is a fermion with spin $\frac{1}{2}\hbar$ and experiences all four fundamental interactions: gravitation, electromagnetism, weak interactions and strong interactions. It has an electric charge of $+\frac{2}{3}e$ and is the most massive of all observed elementary particles. Its large mass is the main reason why it was only discovered in 1995 at the Fermilab collider with $\sqrt{s} = 1.8$ TeV. Since its discovery, the measurement of the top quark mass has become more and more accurate [6, 7].

2.1.1 Discovery of the top quark

In 1964 Murray Gell-Mann and George Zweig proposed the quark hypothesis to account for the explosion of subatomic particles discovered in accelerator and cosmic ray experiments during the 1950s and early 1960s. Over a hundred new particles, most of which strongly interacting and very short-lived, had been observed. These particles, called hadrons, are not elementary: they possess a definite size and internal structure and most of them are unstable. The quark hypothesis suggested that different combinations of three quarks – the up (u), down (d) and strange (s) quarks and their antiparticles – could account for all of the hadrons then known.

Most physicists were initially reluctant to believe that quarks were anything more than convenient abstractions helping particle classification. The fractional electric charges

seemed bizarre and experiments repeatedly failed to expose any individual free quarks. Two major developments established the reality of quarks during the 1970s. Fixed-target experiments using high-energy leptons beams and protons and neutrons targets showed that these hadrons contain point-like internal constituents whose charges and spins are just what the quark model had predicted. And in 1974, experiments at Brookhaven National Laboratory in New York and Stanford Linear Accelerator Center (SLAC) in California discovered a striking new hadron at the then very large mass of 3.1 GeV, over three times that of the proton. This hadron was found to be a bound state of a new kind of quark, called charm or c, with its antiquark. With two quarks of each possible charge, a symmetry could be established between the quarks and the leptons. Two pairs of each were then known: (u, d) and (c, s) for quarks and (e, ν_e) and (μ, ν_μ) for leptons, satisfying theoretical constraints. But this symmetry was quickly broken by unexpected discoveries. In 1976 experiments at SLAC turned up a third charged lepton, the tau lepton or τ. A year later at the Fermi National Accelerator Laboratory in Illinois a new hadron was discovered, called the upsilon (Υ), at the huge mass of about 10 GeV and it was soon found to be the bound state of another new quark: the *bottom* or b quark and its antiparticle. Experiments at DESY in Germany and Cornell in New York measured its fundamental properties.

With these discoveries and through the development of the SM, physicists then understood that matter comes in two parallel but distinct classes: quarks and leptons. They occur in "generations" of two related pairs with differing electric charge, but the third-generation quark doublet seemed to be missing its charge $+2/3$ member, whose existence was inferred from the existing pattern. In advance of its sighting, physicists named it the top (t) quark. Thus began a search that lasted almost twenty years.

Using the ratios of the observed quark masses, some physicists naively suggested that the t quark might be about three times as heavy as the b and thus expected that the top quark would appear as a heavy new hadron containing a $t\bar{t}$ pair, at a mass around 30 GeV. The electron-positron colliders then under construction (PEP at SLAC and PETRA at DESY) raced to capture the prize, but they found no hint of the top quark.

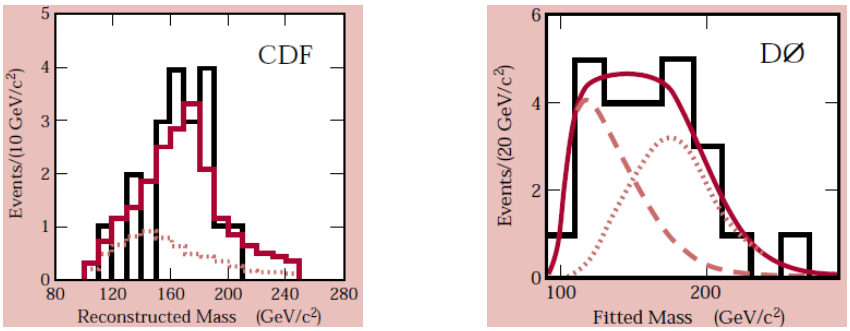
In the early 1980s a new class of accelerators came into operation at CERN in Switzerland, in which counter-rotating beams of protons and antiprotons collided with a centre-of-mass energy of about 600 GeV. The protons and antiprotons brought their constituent quarks and antiquarks into collision with typical energies of 50 to 100 GeV, so the top quark search could be extended considerably. Besides the important discovery of the W and Z bosons that act as carriers of the unified electroweak force, the CERN experiments demonstrated another aspect of quarks. Though quarks had continued to elude direct detection, they can be violently scattered in high-energy collisions, producing *jets*, that are collimated sprays of particles.

In 1992 the D0 detector joined CDF as a long Fermilab Tevatron run began. Further searches would have to rely on the production of separate top quarks and antiquarks from annihilation of incoming quarks and antiquarks in the proton and antiproton, with

subsequent decays into observable particles. The two experiments, while searching for the same basic decay sequence, had rather complementary approaches. First, D0 published a new lower limit of 131 GeV on the possible top quark mass, from the absence of events with the characteristic dilepton or single lepton signatures.

After years spent on the long search, on the 24th of February 1995, the observation of the top quark has been announced by the two experiments at the Tevatron. In its paper [8], the CDF Collaboration reported finding six dilepton events plus 43 single-lepton events; they concluded that observed data were inconsistent with the background prediction by 4.8σ. The D0 Collaboration, in its paper [9], observed three dilepton events plus 14 single-lepton events and concluded that the probability for an upward fluctuation of the background to produce the observed signal was equivalent to 4.6 standard deviations. The top quark masses reported by the two experiments were 176 ± 8 (stat.) $\pm$ 10 (syst.) GeV for CDF, and 199^{+19}_{-21} (stat.) $\pm$ 22 (syst.) GeV for D0 (see Figs. 2.1(a) and 2.1(b)).

The top quark appears to be a point-like particle; it has no internal structure that we can discern. As one of the six fundamental constituents of matter, it has properties very similar to the up and charm quarks, with the exception of its remarkable massiveness and its very short lifetime. The top quark is about 200 times more massive than the proton, about 40 times heavier than the second heaviest quark (the b quark) and roughly as heavy as the entire gold nucleus.



(a) Reconstructed mass (CDF).

(b) Fitted mass (D0).

Figure 2.1: Since the top quark has a unique mass, the data (indicated by the black histograms) should show a "peak" in the reconstructed distribution. The non-top background (the red dashed curve for D0 and the red dotted curve for CDF) has very different shapes. For both experiments, the solid red curve shows what a simulated top quark mass distribution would look like when added to the background. These curves should be compared to the actual data.

2.1.2 Top quark production and decay

Top quark production

At the LHC, top quarks are mostly produced in pairs via strong interaction with a production cross section $\sigma_{t\bar{t}} \approx 830$ pb for $\sqrt{s} = 13$ TeV; however, there is a significant number of top quarks which are produced singly, via the weak interaction ($\sigma_t \approx 300$ pb for $\sqrt{s} = 13$ TeV).

At leading order (LO), there are only a few processes which describe the production of $t\bar{t}$ production: the dominant production mechanism is from gluon-gluon fusion ($\approx$ 90%), while $q\bar{q}$ annihilation accounts for about 10%. The Feynman diagrams for the processes are shown in Fig. 2.2.

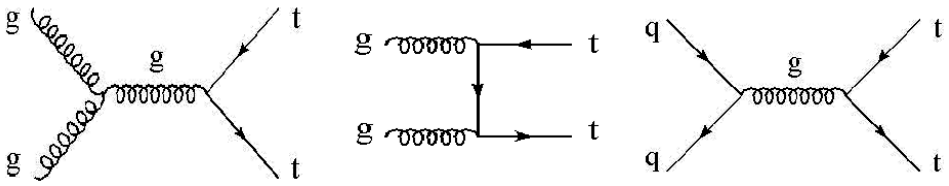


Figure 2.2: Feynman diagrams for $t\bar{t}$ production at LHC.

Single top quark production proceeds through three separate sub-processes at the LHC: t-channel (the dominant process involves the exchange of a space-like W boson. This process is also called W-gluon fusion, because the b quark ultimately arises from a gluon splitting to a $b\bar{b}$ pair), s-channel (involving the production of a time-like W boson, then decaying to a top and a bottom quark) and tW-processes, involving the production of a real W boson.

Top quark decay

The top quark decays through the weak interaction almost exclusively to a W boson and a bottom quark, hence $t\bar{t}$ final states are classified by the decay products of the W boson, that can decay to either leptons or quarks. Therefore, three final states are distinguished as following:

- The dilepton channel, in which both W bosons decay to a lepton-neutrino doublet (branching ratio BR $\approx$ 5%):

$$t\bar{t} \longrightarrow W^+b\ W^-\bar{b} \longrightarrow \ell^+\nu_\ell b\ \ell'^-\bar{\nu}_{\ell'}\bar{b} \qquad \text{with } \ell = e \text{ or } \mu$$

- The single-lepton channel, in which only one W decays to a lepton-neutrino doublet (BR $\approx$ 30%):

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

or:

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

- The all-jets (or fully hadronic) channel, in which both W bosons decay hadronically, as represented in Fig. 2.3 (BR $\approx 46\%$):

$$t\bar{t} \longrightarrow W^+b\ W^-\bar{b} \longrightarrow q\bar{q}'b\ q\bar{q}'\bar{b} \longrightarrow j_1j_2j_3\ j_4j_5j_6,$$

  producing, nominally, 6 jets in the final state.

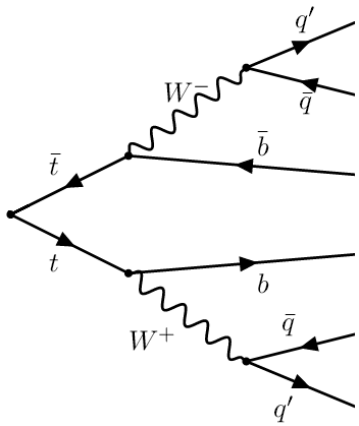


Figure 2.3: Feynman diagrams for $t\bar{t}$ all-jets decay.

2.2 The Higgs boson

In 1964 [10, 11, 12, 13], it was proposed that spontaneous symmetry breaking in gauge theories could be achieved through the introduction of a scalar field. Applying this mechanism to the electroweak theory through a complex scalar doublet field leads to the generation of the W and Z masses, and to the prediction of the existence of the SM Higgs boson (H). The scalar field also gives mass to the fundamental fermions through the Yukawa interaction.

For these reasons, finding experimental evidence of such particle was of paramount importance and physicist had been searching for the Higgs boson since the late 1980s; however, no direct observervation was made before 2012.

2.2.1 The Higgs boson before the LHC

Despite the fact that the Higgs boson had not yet been discovered, there were many experimental results that could give hints on its existence and on its mass [14].

During the 1990s the search for the Higgs boson went on through the collaborations working at the CERN LEP. LEP collided electrons and positrons at energies ranging from $\sqrt{s} = 87$ GeV to $\sqrt{s} = 209$ GeV in the same 27 km underground tunnel now used for the LHC. In 2000, after taking and analysing data for 10 years, the LEP experiments were able to set a lower limit on the mass of the Higgs boson m_H of:

$$m_{H-LEP} > 114.4 \text{ GeV} \qquad \text{at } 95\% \text{ confidence level C.L.}$$

On the other side of the ocean, the Tevatron Collider at Fermilab near Chicago, colliding protons and antiprotons, started its search in 1987. The Tevatron searched for the Higgs boson created directly as a result of a proton - antiproton collision, with the Higgs boson decaying subsequently to a pair of W bosons. Using around 80% of the total data collected, the two experiments at the Tevatron, CDF and D0, ruled out the existence of a Higgs boson with a mass between:

$$158 \text{ GeV} < m_{H-excluded,\ Tevatron} < 177 \text{ GeV} \qquad \text{at } 95\% \text{ C.L.}$$

Although LEP did not find any direct evidence for the Higgs boson, measurements from the four LEP experiments, along with measurements from the SLC and Tevatron colliders in the USA, had found circumstantial, or indirect evidence for the Higgs boson.

The global fit (in Fig. 2.4), combining all the experimental indirect measurements, suggested that the mass of the Higgs boson was:

$$m_{H-combination-indirect} < 155 \text{ GeV} \qquad \text{at } 95\% \text{ C.L.}$$

with:

$$m_{H-expected} = 121^{+17}_{-6} \text{ GeV}$$

"While this is not a proof that the Standard Model Higgs boson actually exists, it does serve as a guideline in what mass range to look for it" - LEP Electroweak Working Group.

2.2.2 The discovery of the Higgs boson at the LHC

In order to allow the discovery (or the exclusion) of a Higgs boson whose mass lied between 115 GeV and 1 TeV, it was necessary to build a new machine that could reach the suitable energy, and on the 10th of September 2008 the LHC at CERN started operations.

Almost two years later, on the 4th of July 2012, the ATLAS and CMS experiments at CERN announced they had each observed a new particle in the mass region around 125 GeV [15, 16]. This particle was consistent with the Higgs boson predicted by the Standard Model. Furthermore, on the 8th of October 2013, the Nobel Prize in Physics was awarded to Francois Englert and Peter Higgs *"for the theoretical discovery of a mechanism that*

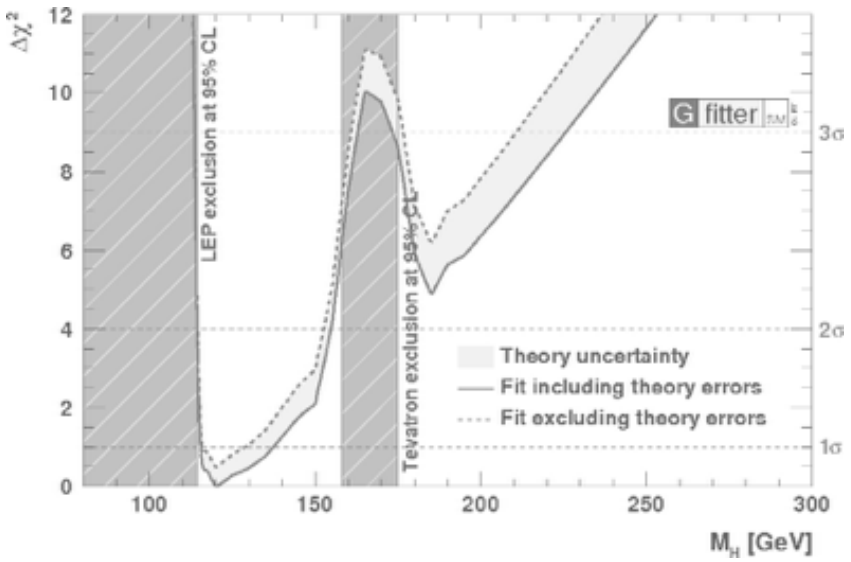


Figure 2.4: Global fit of the Higgs boson mass, including all the indirect measurements of the Higgs boson mass.

contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN Large Hadron Collider."
We will now go through the steps that led to the discovery.

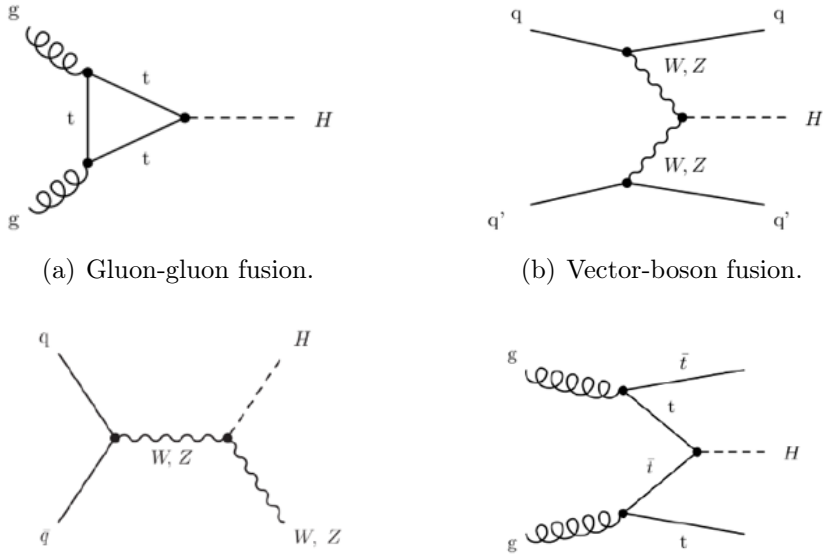
Higgs production at the LHC

From a phenomenological point of view, four main mechanisms are predicted for the Higgs boson production in pp collisions (see Fig. 2.5): the gluon-gluon fusion mechanism, the vector-boson fusion (VBF), the associated WH and ZH production (VH or "*higgs-strahlung*"), and finally the production in association with top quarks ($t\bar{t}H$). The cross sections for the individual production mechanisms depend on $\sqrt{s}$.

The gluon-gluon fusion and the vector-boson fusion are the dominant processes at the LHC, and in particular the mechanism that has the largest cross section (see Fig. 2.6) is the gluon-gluon fusion, through a loop of top quarks.

Higgs boson decay modes

The Higgs boson couplings to particles are proportional to the masses of the particles themselves, meaning that the Higgs boson preferably decays to the heaviest particles that are kinematically permitted. From Fig. 2.7, it can been easily seen that the set of sensitive decay modes of the SM Higgs boson depends strongly on its mass. For example, for a "light" Higgs boson ($m_{\mathrm{H}} < 130$ GeV) the dominant decay channel would



(a) Gluon-gluon fusion.

(b) Vector-boson fusion.

(c) Associated WH and ZH production.

(d) Production in association with top quarks ($t\bar{t}H$)

Figure 2.5: Higgs boson production processes at the LHC.

be $H \longrightarrow b\bar{b}$, while for $m_H \geq 130$ GeV the most relevant decays are $H \longrightarrow WW^*$ and $H \longrightarrow ZZ^*$ with a real and virtual W/Z boson.

However, for a given value of m_H, the search sensitivity depends on the production cross section, the decay branching ratio into the chosen final state, the signal selection efficiency, the mass resolution, and the level of background from identical or similar final-state topologies. For this reason the fully hadronic final states were not exploited for the discovery, due to the extreme difficulty in extracting the signal events from the quantum chromodynamics (QCD) background.

Therefore, at the time of the discovery, all the experimental efforts were put on the three most sensitive decay modes in the low-mass region:

$$H \longrightarrow \gamma\gamma, \qquad H \longrightarrow ZZ^* \longrightarrow 4\ell, \qquad H \longrightarrow WW^* \longrightarrow 2\ell\, 2\nu.$$

$H \longrightarrow \gamma\gamma$

The experimental signature of a $H \longrightarrow \gamma\gamma$ process is characterised by a narrow peak in the diphoton invariant mass distribution, together with a hard diphoton p_T spectrum and isolated photons (see Fig. 2.8(a)). Such a narrow resonance was found on a large irreducible background from QCD production of two photons (Fig. 2.8(b)) and a reducible

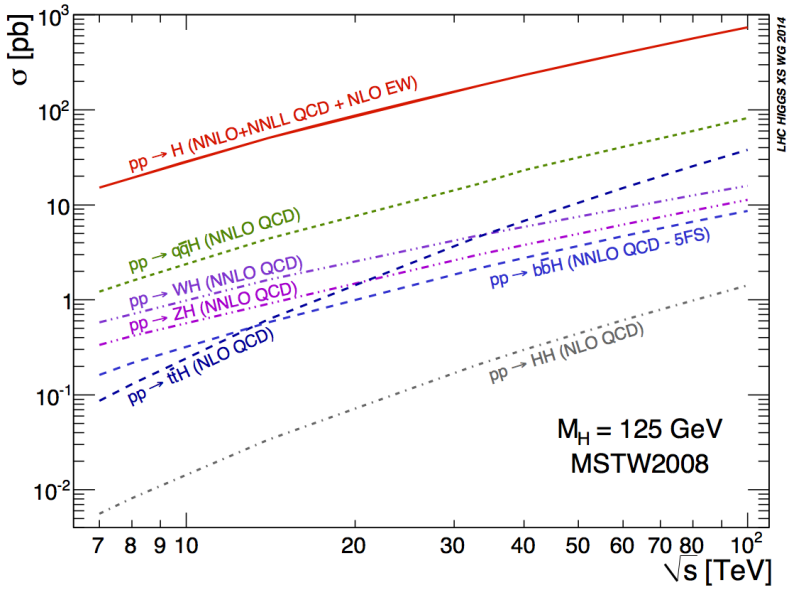


Figure 2.6: Higgs boson production cross sections as a function of $\sqrt{s}$.

background where one or more of the reconstructed photon candidates originated from misidentification of jet fragments (Fig. 2.8(c)).

As far as the analysis was concerned, first the event selection was carried out, selecting events that matched the experimental signature, thus involving two photons with high p_T; then the energy and direction of each photon was measured; finally the invariant mass of the pair was calculated. Eventually, the plots of invariant mass were obtained.

$H \longrightarrow ZZ^* \longrightarrow 4\ell$

The search for the SM Higgs boson through the decay $H \longrightarrow ZZ^* \longrightarrow 4\ell$, where $\ell = e$ or μ, provided great ($\sim 1\%$) sensitivity over a wide mass range (110–600 GeV), largely due to the excellent momentum resolution of both the ATLAS and CMS detectors. For this reason, this channel was known as the "*golden channel*". Despite the fact that this channel was extremely clean, it had a disadvantage, consisting of a very low signal in the range masses near 125 GeV. Figures 2.9(a) and 2.9(b) show an event display of two events of this type, whose experimental signature was characterised by two pairs of isolated leptons, each of which contained two leptons with the same flavour and opposite charge.

The background sources included an irreducible four-lepton contribution from direct ZZ^* production via $q\bar{q}$ and gluon-gluon processes. Reducible contributions arose from $Z + b\bar{b}$ and $t\bar{t}$ production where the final states contained two isolated leptons and two b quark jets producing secondary leptons. Additional background arose from Z + jets and WZ + jets events where jets were misidentified as leptons.

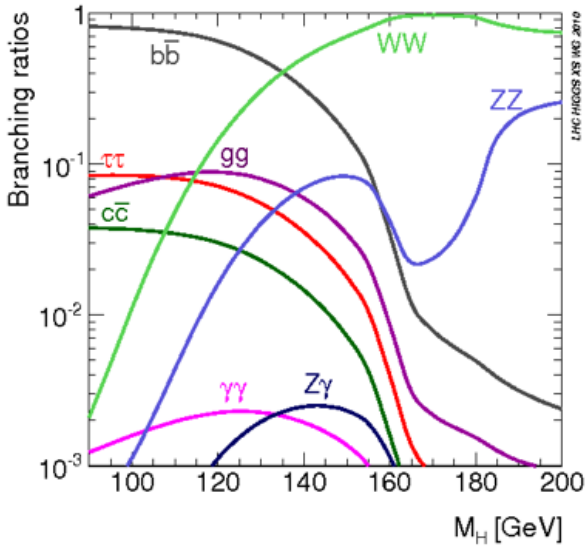


Figure 2.7: Branching ratios of the Higgs boson decay modes.

As far as the event selection was concerned, both muons and electrons were required to be isolated and identified as originating from the same primary vertex. This was ensured by requirements on the impact parameter. Then, the Z candidates were reconstructed and the four leptons were combined in pairs to reconstruct a pair of Z_1Z_2 candidates, where Z_1 was identified as the closest (in mass) to a real boson Z. One or both the Z candidates could be off mass-shell.

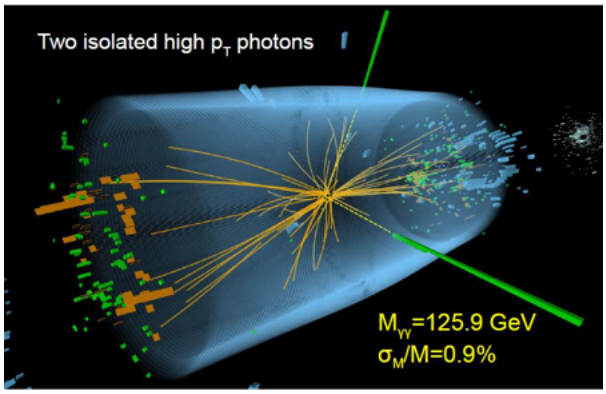
Finally, the two pairs were combined to reconstruct the Higgs boson candidate and its invariant mass, that was eventually plotted.

$H \longrightarrow WW^* \longrightarrow 2\ell 2\nu$

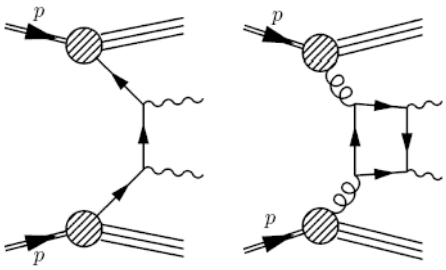
The channel in which the Higgs boson decays to two leptons and two neutrinos is characterised by high sensitivity, but low mass resolution, due to the presence of the ν in the event. This decay mode was analysed by selecting events in which both W bosons decayed leptonically, resulting in a signature (see Fig. 2.10) with two isolated, oppositely charged leptons (electrons or muons) and large p_T^{miss}, due to the undetected neutrinos. p_T^{miss} is the magnitude of the negative vector sum of the transverse momenta of the reconstructed objects, including muons, electrons, photons, jets, and clusters of calorimeter cells not associated with these objects.

The dominant backgrounds were controllable and consisted of non-resonant WW, $t\bar{t}$ and Wt production, all of which had real W pairs in the final state.

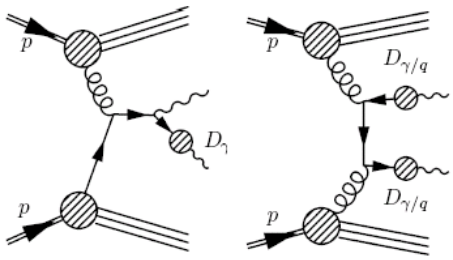
As far as the event selection was concerned, the experimental signature of the events was exploited, with requirements on isolated leptons with opposite charge and large



(a) Event display for a H $\longrightarrow \gamma\gamma$ by the CMS experiment.

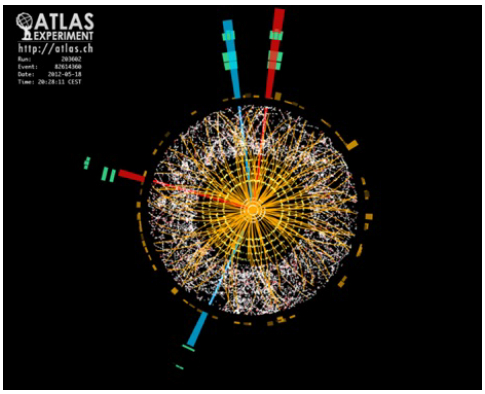


(b) $\gamma\gamma$ production: irreducible background.

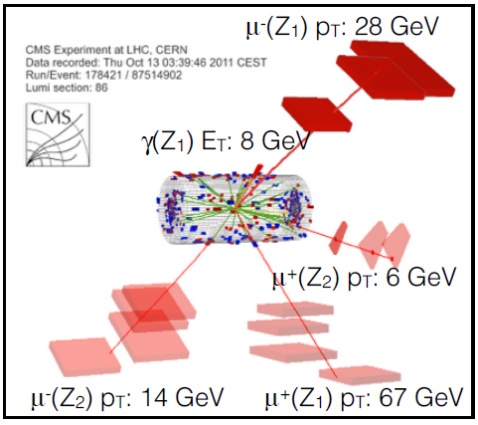


(c) γj and jj production: reducible background.

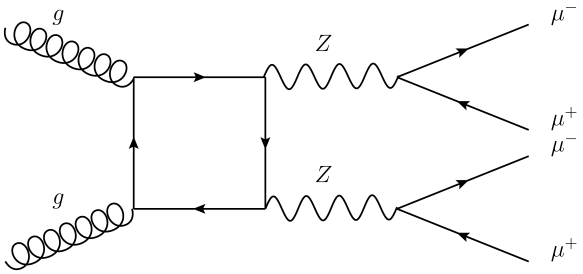
Figure 2.8: Event display and Feynman diagrams for pp $\longrightarrow$ H $\longrightarrow \gamma\gamma$ processes.



(a) Event display for a H $\longrightarrow$ 4e event by the ATLAS experiment.



(b) Event display for a H $\longrightarrow$ 4μ event by the CMS experiment.



(c) ZZ* production: irreducible background.

Figure 2.9: Event displays and Feynman diagram for pp $\longrightarrow$ H $\longrightarrow$ ZZ* $\longrightarrow$ 4ℓ processes.

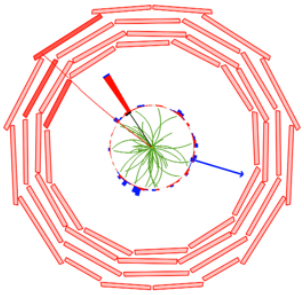
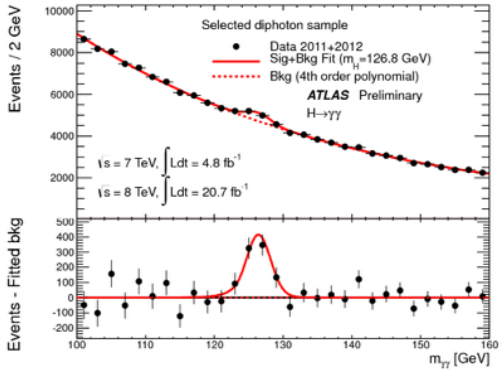


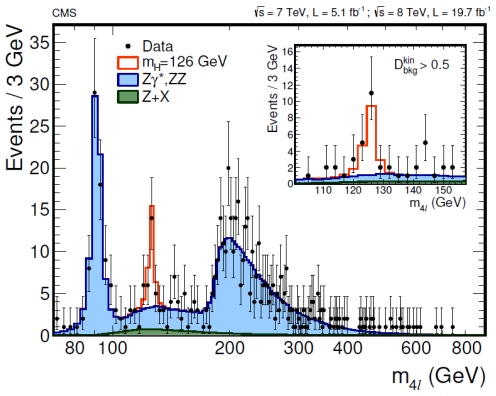
Figure 2.10: Event display for a pp $\longrightarrow$ H $\longrightarrow$ WW* $\longrightarrow$ e$\mu 2\nu$ event by the CMS experiment.

missing transverse momentum. Furthermore, an additional kinematic feature is used: owing to spin correlations in the WW* system arising from the spin-0 nature of the SM Higgs boson and the V–A structure of the W boson decay vertex, the charged leptons tended to emerge from the primary vertex pointing in the same direction. Going on with the analysis, after using algorithms for the computation of the missing momentum, the invariant mass was plotted.

Results



(a) Results for the H $\longrightarrow \gamma\gamma$ channel for the ATLAS collaboration.



(b) Results for the H $\longrightarrow$ 4ℓ channel for the CMS collaboration.

Figure 2.11: Experimental results of the Higgs discovery: the invariant mass plots clearly show a peak.

On the 4th of July 2012, the ATLAS and CMS collaborations announced that a new particle at about 125 GeV was observed to decay to gauge bosons with the expected relative branching ratio for the SM Higgs boson. The new particle had consistent mass

between CMS and ATLAS and spin-parity measurements disfavoured alternative hypotheses to the one associated to the SM Higgs boson. Finally, the signal strength and couplings were consistent with the SM expectations.

In particular, the ATLAS collaboration stated that the experimental results "*provide conclusive evidence for the discovery of a new particle*", with mass of:

$$m_{\mathrm{H_{ATLAS}}} = 126.0 \pm 0.4(stat.) \pm 0.4(syst.)\ \mathrm{GeV}.$$

Moreover, "*the decays to pairs of vector bosons whose net electric charge is zero identify the new particle as a neutral boson. The observation in the diphoton channel disfavours the spin-1 hypothesis. Although these results are compatible with the hypothesis that the new particle is the Standard Model Higgs boson, more data are needed to assess its nature in detail*".

As far as the CMS collaboration is concerned, instead, "*the excess is most significant in the two decay modes with the best mass resolution, $\gamma\gamma$ and ZZ, and a fit to these signals gives a mass of*

$$m_{\mathrm{H_{CMS}}} = 125.3 \pm 0.4(stat.) \pm 0.5(syst.)\ \mathrm{GeV}.$$

The decay to two photons indicates that the new particle is a boson with spin different from one. The results presented here are consistent, within uncertainties, with expectations for a standard model Higgs boson. The collection of further data will enable a more rigorous test of this conclusion and an investigation of whether the properties of the new particle imply physics beyond the standard model.

2.3 Production of the Higgs boson in association with a pair of top quarks ($\mathrm{t\bar{t}H}$)

The observation of a Higgs boson with a mass of approximately 125 GeV in 2012 marked the starting point of a broad experimental program to determine the properties of the newly discovered particle. As we have seen so far, the results of all measurements performed at the LHC are consistent with the expectations for a SM Higgs boson [17].

In the SM, the coupling of the Higgs boson to fermions is of Yukawa type, with a coupling strength proportional to the fermion mass. The first evidence that the 125 GeV Higgs boson couples to down-type fermions with SM-like strength has been provided through direct measurements of decays into bottom quarks and τ leptons [18]. However, evidence of a direct coupling to up-type fermions, in particular to top quarks, is still lacking. An indirect constraint on the top quark Yukawa coupling can be inferred from measuring either the production or the decay of Higgs bosons through top quark loops. Current measurements of the Higgs boson cross section via gluon fusion and of its branching ratio to photons are consistent with the SM expectation for the top quark

Yukawa coupling. Since these effective couplings occur at the loop level, they can be affected by beyond-standard model (BSM) particles. Therefore, in order to disentangle the top quark Yukawa coupling from a possible BSM contribution, a *direct* measurement of the former is required. This can be achieved by measuring observables that probe the top quark Yukawa interaction with the Higgs boson already at the tree level. The production cross section of the Higgs boson in association with a top quark pair ($t\bar{t}H$) provides an example of such an observable. The case in which the Higgs boson decays into a quark-antiquark pairs ($b\bar{b}$), shown in Fig. 2.12, is a particularly attractive final state, because it features the largest branching ratio of:

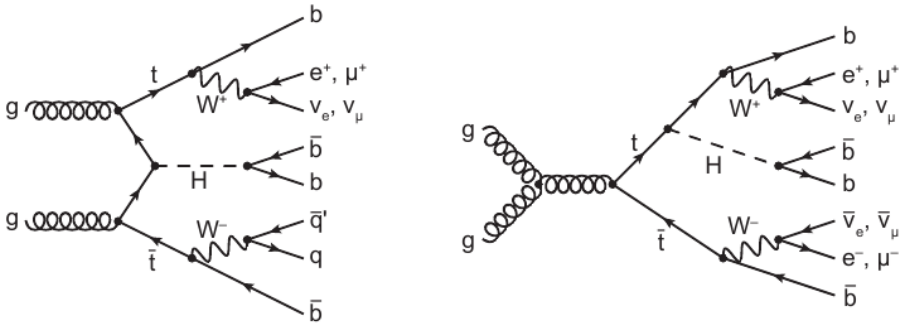


Figure 2.12: Feynman diagrams for pp $\longrightarrow$ $t\bar{t}H(b\bar{b})$ processes.

$$BR_{H\longrightarrow b\bar{b}} = 0.58 \pm 0.02.$$

Furthermore, several BSM physics models predict a significantly enhanced $t\bar{t}H$ production rate while *not* modifying the branching ratios of Higgs boson decays by a measurable amount. For example, a number of BSM physics models predict vector-like partners of the top quark (T) that decay into tH, bW and tZ final states. The production and decay of $T\bar{T}$ pairs would lead to final states indistinguishable from those of $t\bar{t}H$ production. Therefore, the measurement of the $t\bar{t}H$ production cross section has the potential to distinguish the SM Higgs mechanism from alterative mechanisms to generate fermion masses.

Various dedicated searches for $t\bar{t}H$ production have been conducted during Run 1 of the LHC. In particular, both the CMS and ATLAS collaborations have studied Higgs boson decays to hadrons, photons, and leptons using multivariate analysis (MVA) techniques, showing a mild excess of the observed $t\bar{t}H$ signal strength relative to the SM expectation [19].

The observation of $t\bar{t}H$ production is one of the major goals in Higgs boson physics for Run 2. The increased centre-of-mass energy of 13 TeV results in a $t\bar{t}H$ production cross section 3.9 times larger than at $\sqrt{s} = 8$ TeV [20], while the cross section for the most important background, $t\bar{t}$ production, is only increased by a factor of 3.3, resulting

in a more favourable signal-to-background ratio. In addition, a larger fraction of events contains top quarks or Higgs bosons with transverse momenta above 200 GeV, originating jets closer in the $\eta - \phi$ space, and making "boosted" jet reconstruction techniques increasingly attractive for $\mathrm{t\bar{t}H}$ studies.

2.3.1 The $\mathrm{t\bar{t}H}$ all-jets channel

The analysis described here focuses on the final states involving the Higgs boson decay into a $\mathrm{b\bar{b}}$ pair ($\mathrm{H} \longrightarrow \mathrm{b\bar{b}}$) and the all-jets decay channel of the $\mathrm{t\bar{t}}$ pair ($\mathrm{BR}_{\mathrm{t\bar{t}} \longrightarrow \mathrm{bj_1j_2\,bj_3j_4}} = 0.46$). Despite being the one with the highest BR,

$$BR_{\mathrm{t\bar{t}H} \longrightarrow \mathrm{bj_1j_2\,bj_3j_4\,b\bar{b}}} = 0.58 \times 0.46 = 0.27,$$

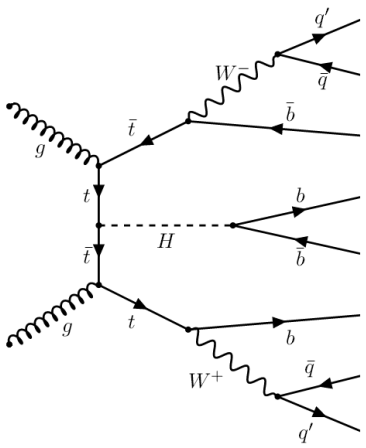


Figure 2.13: Feynman diagram for $\mathrm{t\bar{t}H} \longrightarrow \mathrm{bj_1j_2\,bj_3j_4\,b\bar{b}}$ processes.

and potentially fully reconstructable, such decay mode (cfr. Fig. 2.13) is extremely complex, due to the dominating QCD multijet background.

All-jets signal events are characterised by: three jets coming from the top quark decay, three jets coming from the antitop quark decay, and two more jets coming form the decay of the Higgs boson to two b quarks. Therefore, the experimental signature of the signal events we look for consists of:

- at least 8 jets, with at least 4 b-jets;
- no leptons (and no missing transverse momentum).

Chapter 3

Event selection and characterization

3.1 Samples

In this section a brief description of the used samples is given. Data and Monte Carlo (MC) samples are discussed separately. Both the data and simulation samples were stored in ROOT files as trees containing all the relevant information on the major physics objects reconstructed in the detector, such as jets, leptons, photons and tracks.

Data samples

The data used in this work represent the full data set collected during 2015 and 2016 from pp collisions at 13 TeV, corresponding to an integrated luminosity of 36 fb^{-1}. In particular, the analysis is performed and optimised on 2015 data, corresponding to an integrated luminosity $L = 2.63$ fb^{-1}, and then a projection of the results on the full data set is given.

Monte Carlo samples

The analysis has been conducted using several MC samples, simulating both signal and background processes.

As far as signal is concerned, $t\bar{t}H$ events have been simulated using MADSPIN [21, 22] up to the next-to-leading order (NLO). Events contained in such sample consist of $t\bar{t}H$ events with a simulated Higgs boson mass of 125 GeV, the $t\bar{t}$ pair decaying to all channels (dilepton, single lepton and all-jets) and the Higgs boson decaying to $b\bar{b}$.

The dominant background processes are the QCD and $t\bar{t}$ multijets production, respectively simulated with MADGRAPH [23] and POWHEG [24, 25]. Subdominant background sources then include WW (simulated with POWHEG), ZZ, $t\bar{t}W$ + jets, $t\bar{t}Z$, DY + jets (Drell-Yan processes), and W + jets (all simulated with MADGRAPH).

In all cases, the parton shower simulation is conducted using PYTHIA [26, 27].

Sample	σ (pb)
$t\bar{t}$	832
WW	51.7
ZZ	22.3
$t\bar{t}W$ + jets	0.406
$t\bar{t}Z$ + jets	0.530
DY + jets	1460
W + jets	35.4
$t\bar{t}H(b\bar{b})$	$0.509 \cdot 0.58_{(BR)}$

Table I: Cross sections of the signal and background processes for $\sqrt{s} = 13$ TeV.

Table I summarises all the MC samples used in the analysis and the corresponding cross sections of the processes involved.

3.2 Jet reconstruction

Jets can be defined as collimated sprays of particles arising from the fragmentation and the hadronisation of a parton (quark or gluon) after a collision [28]. Jet reconstruction algorithms combine the calorimetry and tracking information to define jets, that provide a link between the observed colourless stable particles and the underlying physics at partonic level. A basic illustration of two protons colliding, the subsequent particle shower and a reconstructed jet is shown in Fig. 3.1. This link provides information on the kinematic quantities of the originating partons, that can be used to shed light on QCD and infer the presence and thus the properties of all the particles that are too short-lived to be detected. An accurate jet algorithm is also able to calculate the correct amount of momentum imbalance in the detector, which gives an estimate of the direction and energy of the neutrinos and other invisible particles.

There are two main classes of jet algorithms in use: the first being the *cone algorithms* and the second being the *sequential clustering algorithms*.

Cone algorithms assume that particles in jets will be produced in conical regions and thus they cluster based on $\eta - \phi$ space, resulting in jets with rigid circular boundaries.

Sequential clustering algorithms, instead, are based on the hypothesis that particles within jets will have small differences in p_T, and thus group together particles in the momentum space, resulting in jets that have fluctuating areas in the $\eta - \phi$ space. Sequential clustering algorithms have always been preferred by theorists, but were not been extensively used in the past, given their slow computational performance. Sequen-

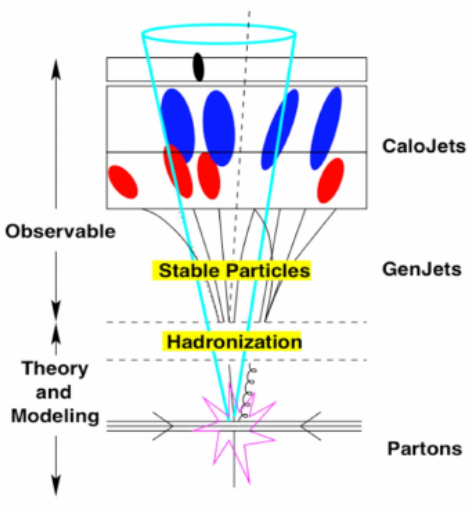


Figure 3.1: A simple example of an event showing the point of collision, the fragmentation and hadronization of the quarks and gluons, and the resulting jet found through the detection of the stable particles.

tial clustering algorithms are also infrared safe, meaning that are well defined at any order of perturbation theory.

All sequential clustering algorithms use similar methods, with small variations dependent on the actual algorithm. Firstly, a variable representing the distance between two particles is computed:

$$d_{ij} = min\left(p_{T,i}^{a}, p_{T,j}^{a}\right) \cdot \frac{R_{ij}^{2}}{R}\,, \qquad\qquad R_{ij}^{2} = \left(\eta_i - \eta_j\right)^2 + \left(\phi_i - \phi_j\right)^2\,,$$

where a is an exponent corresponding to the particular clustering algorithm, R_{ij}^2 is the $\eta - \phi$ space distance between the two particles and R is the radius parameter which determines the final size of the jet and usually ranges between 0.4 and 0.7. Secondly, an additional distance is defined:

$$d_{iB} = p_{T,i}^{a}\,,$$

representing the distance (in the momentum space) between the beam axis and the detected particle. At this point, sequential clustering algorithms work by first finding the minimum of the entire set $\{d_{ij}, d_{iB}\}$. If d_{ij} is the minimum, then particles i and j are combined into one particle (ij) using summation of four-vectors, after which i and j are removed from the list of particles. If d_{iB} is the minimum, i is labelled a final jet and removed from the list of particles. This process is repeated until either all particles are part of a jet with the distance between the jet axes $R_{ij} > R$ (inclusive clustering), or until a desired amount of jets have been found (exclusive clustering).

The so-called k_T algorithm uses $a = 2$, resulting in:

$$d_{ij} = min\left(p_{T,i}^2, p_{T,j}^2\right) \cdot \frac{R_{ij}^2}{R}\,, \qquad\qquad d_{iB} = p_{T,i}^2\,,$$

and determining the dominance of low-p_T contributions. This means that the k_T algorithm prefers to cluster soft particles first, resulting in an area that fluctuates considerably and an algorithm that is susceptible to underlying events and the pileup. Due to its method of clustering, k_T does a good job at resolving subjets.

However, the particular algorithm used in the case of this analysis is the anti-k_T one [29], corresponding to $a = -2$, and resulting in the following equations:

$$d_{ij} = min\left(\frac{1}{p_{T,i}^2}, \frac{1}{p_{T,j}^2}\right) \cdot \frac{R_{ij}^2}{R}\,, \qquad\qquad d_{iB} = \frac{1}{p_{T,i}^2}\,.$$

As it can be deduced from the equation above, this reconstruction is dominated by high-p_T contributions and the algorithm prefers to cluster hard particles first. Thus, the area only fluctuates slightly and the algorithm is only marginally influenced by underlying events and the pileup. The anti-k_T clustering preference results in an algorithm that is the best at resolving jets, but is the worst for studying jet substructure due to its poor de-clustering performance.

Particles that are used as input to the jet reconstruction are identified in CMS using the *particle-flow* algorithm (PF) [30].

The PF event reconstruction aims at reconstructing and identifying all stable particles in the event, such as electrons, muons, photons, charged hadrons and neutral hadrons, *with a thorough combination of all CMS sub-detectors* towards an optimal determination of their direction, energy and type. This list of individual particles is then used (as if it came from a MC event generator) to build jets (from which the quark and gluon energies and directions are inferred), to determine the missing transverse momentum, reconstruct and identify τ leptons from their decay products, quantify charged lepton isolation with respect to other particles, tag b-jets, etc.

The CMS detector appears to be almost ideally suited for this purpose. With its large silicon tracker (with excellent resolution) immersed in the uniform axial magnetic field of 3.8 T, charged-particle tracks can be reconstructed with large efficiency and adequately small misidentification rate, down to a momentum transverse to the beam of 150 MeV, for $|\eta| \leq 2.6$.

The PF algorithm proceeds through several steps, that will be briefly outlined. Firstly, the identification of fundamental *elements* is conducted, exactly as reconstructed in the sub-detectors, obtaining charged particles, calorimeter clusters and muon tracks. These elements are then connected to each others by making use of link algorithms identifying *blocks* of elements which are topologically compatible. For example, a charged-particle track is linked to a calorimeter cluster if the extrapolated position from the track

to the calorimeter is within the cluster boundaries. From the blocks, PF candidates are fully reconstructed and identified in the following order:

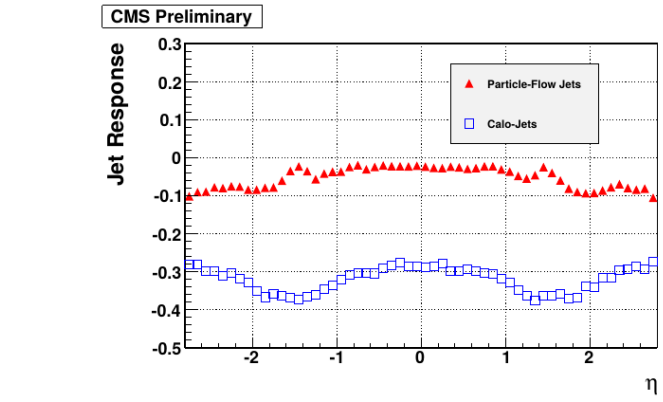
- *muons*: a global muon, reconstructed from the combination of a track in the tracker and a track in the muon system, gives rise to a PF muon. After the identification, the corresponding track is removed from the block;
- *electrons*: the link between a charged-particle track and one or more ECAL clusters identifies PF electrons. The corresponding track and ECAL clusters are removed from further processing;
- *charged hadrons*: the remaining tracks give rise to PF charged hadrons and the momentum of the particle is taken directly from the track momentum. Tracks can be linked to ECAL and HCAL clusters if they are not identified as electrons, and the momentum is redefined taking into account information from calorimeters;
- *photons and neutral hadrons*: ECAL clusters not compatible with charged tracks give rise to PF photons, while unaccounted HCAL deposits are interpreted as PF neutral hadrons.
- *missing transverse energy*: PF p_T^{miss} is reconstructed at the end of the event reconstruction and is computed by forming the transverse momentum-vector sum over all reconstructed PF candidates in the event and then taking the opposite of this vectorial sum. At this point, the missing transverse energy is the modulus of this vector.

Once the list of PF candidates is defined, PF jets can be reconstructed using the sequential clustering jet algorithm described above. Typically the largest fraction of jet energy (about 65%), is carried by charged particles, 25% by photons and 10% by neutral hadrons. Therefore, 90% of the jet energy can be reconstructed with good precision, thanks to the high resolution of the tracker and of the electromagnetic calorimeter, while just 10% of it is affected by the poor HCAL resolution. The combination of tracks and calorimeter clusters is a key point of the PF algorithm because it allows to get a very high efficiency for PF jets, comparing to the sole use of information from calorimeters.

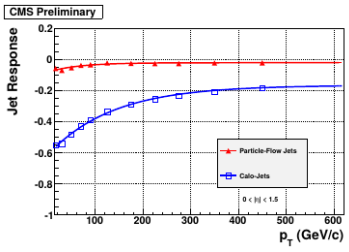
Fig. 3.2 compares traditional calorimeter-based jet reconstruction (Calo-Jets) with PF reconstruction, that shows a better performance.

3.3 The b-tagging

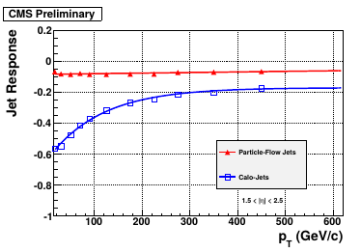
Jets arising from the hadronization of bottom quarks (b-jets) characterise many physics processes, such as the decay of top quarks, the Higgs boson, and several particles predicted by supersymmetric models. For this reason, the ability to identify b-jets accu-



(a)



(b)



(c)

Figure 3.2: Jet response for PF and calorimeter-based jet reconstruction as a function of η integrated over all $p_T < 750$ GeV (3.2(a)) and as a function of p_T, in the barrel (3.2(b)), and in the endcaps (3.2(c)). The response curves are fit with exponential functions of p_T.

rately is crucial in reducing the dominant background to these channels, from processes involving jets from gluons (g) and light quarks (u, d, s), and from c-quark fragmentation.

The hadronic jets into which the b quarks fragment can be identified by using the properties of the bottom hadrons. Such hadrons have relatively large masses, long lifetimes and daughter particles with hard momentum spectra. Their semileptonic decays can be exploited as well. The CMS detector is well suited for the task of b-jet identification (b-jet tagging), thanks to its precise charged-particle tracking and robust lepton identification systems [31].

3.3.1 B-tagging algorithms

A variety of reconstructed objects (such as tracks, vertices and identified leptons) can be used to build observables that discriminate between b and light-parton jets. A single observable is used by several simple and robust algorithms, while others combine a few of these objects, in order to achieve a higher discrimination power. In either cases, these CMS algorithms provide a single discriminator value for each jet. The minimum thresholds on these discriminator values define loose ("L"), medium ("M") and tight ("T") operating points with a misidentification probability for light-parton jets of close to 10%, 1% and 0.1%, respectively, at an average jet p_T of about 80 GeV.

The CSV algorithm

The presence of a secondary vertex, and the kinematic variables associated with it, can be used to discriminate between b and non-b jets. The most significant variables used for the discrimination are the flight distance and direction, using the vector between primary and secondary vertices. The remaining variables are related to some properties of the system of associated secondary tracks, such as the multiplicity, the mass, or the energy.

In order to enhance the "b-purity", secondary vertex candidates should meet the following requirements:

- secondary vertices must share less than 65% of their associated tracks with the primary vertex and the significance of the radial distance between the two vertices has to be $> 3\sigma$, with σ being the uncertainty on the distance;
- secondary vertex candidates with a radial distance of more than 2.5 cm with respect to the primary vertex, with masses compatible with the mass of K^0, or exceeding 6.5 GeV are rejected, thus reducing the contamination by vertices corresponding to both interactions of particles with the detector material, and decays of long-lived mesons;

- the flight direction of each candidate has to lie within a cone of $\Delta R = 0.5$ around the jet direction.

The Combined Secondary Vertex (CSV) algorithm – used in this analysis – involves the use of secondary vertices, together with track-based lifetime information. By using these additional variables, the CSV algorithm provides discrimination also in cases when no secondary vertices are found, increasing the maximum efficiency with respect to the so-called "Simple Secondary Vertex" algorithms – these using only the flight distance as discriminating variable.

In many cases, tracks with an impact parameter significance S_{IP} – that is the ratio of the IP to its estimated uncertainty – that is > 2 can be combined into a "pseudo vertex", allowing the computation of a subset of secondary-vertex-based quantities even without an actual vertex fit. Finally, when even this is not possible, a "no vertex" category reverts to track-based variables and the discrimination is conducted in a way similar to that of the track-based algorithms.

Therefore, the CSV algorithm uses the following set of variables with high discriminating power and low correlations (of course, in the "no vertex" category only the last two variables are available):

- the vertex category: *real*, *pseudo*, or *no vertex*;
- the flight distance significance in the transverse plane;
- the vertex mass, i.e. the invariant mass of the particles associated to the vertex;
- the number of tracks at the vertex;
- the ratio of the energy carried by tracks at the vertex with respect to all tracks in the jet;
- the pseudorapidities of the tracks at the vertex with respect to the jet axis;
- the 2D IP significance of the first track that raises the invariant mass above the charm threshold of 1.5 GeV (tracks are ordered by decreasing IP significance and the mass of the system is recomputed after adding each track);
- the number of tracks in the jet;
- the 3D IP significances for each track in the jet.

Then, two likelihood ratios are built from these variables and are used to discriminate between b- and c-jets and between b- and light-parton jets. Finally, they are combined with prior weights of 0.25 and 0.75, respectively.

The CSV algorithm has evolved into the CSVv2 (Combined Secondary Vertex *version 2*) algorithm in Run 2. Just like the CSV, the CSVv2 is based on secondary vertex and track-based lifetime information [32]. Despite this, the new version of the CSV algorithm combines the variables using a neural network instead of a likelihood ratio to produce a discriminator *csv*, and the secondary vertex information is obtained with the Inclusive Vertex Finder algorithm. The operating point values for the loose, medium and tight tagging criteria are set to 0.460, 0.800, 0.935, respectively.
The b-tagging efficiency measured in MC events is corrected using scale factors in order to reproduce the efficiency measure in data. For the medium working point – the one used in the case of this analysis – the scale factor amounts to 0.97 ± 0.02.

3.4 Preselection

The samples used in the analysis contain events preselected according to the topology of the signal events:

- jet multiplicity $n_{jets} \geq 6$, with $|p_T| > 30$ GeV and $|\eta| < 2.4$;
- total jet transverse energy $H_T = \sum |p_T| > 400$ GeV;
- number of b-tagged jets $n_{b-jets} \geq 2$.

3.5 Trigger

When the LHC works at nominal parameters, proton bunches cross at the outstanding rate of 40 MHz. The CMS experiment needs to reduce this rate by a factor 1000, using a Level 1 hardware trigger and subsequently by another factor 1000 using a software-implemented High Level Trigger (HLT).

Triggers can be defined as *path* blocks, each of these being a sequence of modules and operands. The result of each operand is a boolean quantity, and thus the final outcome of a trigger path is either "reject" or "accept"; in this last case we say that the *trigger is fired*.

3.5.1 Trigger paths

Several HLT trigger paths have been studied for this analysis, and they can be classified into three categories: **signal**, **control** and **reference** triggers.
Two **signal** trigger paths have been studied:

- `HLT_PFHT450_SixJet40_BTagCSV0p72` - Trg-0: requiring the presence of six leading PFJets with $p_T > 40$ GeV and one b-jet ($csv > 0.72$). It also requires a $H_T > 450$ GeV in the event.

- `HLT_PFHT400_SixJet30_BTagCSV0p55_2BTagCSV0p72` - Trg-2: requiring the presence of six leading PFJets with $p_T > 30$ GeV and two b-jets: one of which with $csv > 0.72$ and the other one – looser – with $csv > 0.55$. It also requires a $H_T > 400$ GeV in the event.

Therefore, our first signal trigger has tighter kinematic and looser b-tag requirements than the second one. Both these paths are unprescaled, i.e. all the events passing the trigger are kept.

Then, we have two **control** trigger paths, used to select a control region. They correspond to the signal triggers as to the kinematic criteria, but no requirements on the b-jets are implemented. Differently from the signal trigger paths, these control triggers are prescaled (through different factors):

- `HLT_PFHT450_SixJet40` - Trg-1: requiring the presence of six leading PFJets with $p_T > 40$ GeV and a $H_T > 450$ GeV in the event.
- `HLT_PFHT400_SixJet30` - Trg-3: requiring the presence of six leading PFJets with $p_T > 30$ GeV and a $H_T > 400$ GeV in the event.

Lastly, there is the **reference** trigger, that has been heavily and differently prescaled during the data taking, i.e. only a predefined fraction of events passing the trigger are kept. Such trigger is used to evaluate the signal trigger efficiencies:

- `HLT_PFHT350` - Trg-4: selecting events with a $H_T > 350$ GeV in the event.

3.5.2 Trigger efficiencies

The efficiency of the triggers is measured for each signal trigger, within the phase space defined by the preselection criteria and matching the topology of the signal events. These preselection criteria were optimized to reproduce the trigger requirements, and to be sufficiently in or close to the trigger turn-on plateaus without reducing the signal acceptance. The trigger efficiency is initially estimated by using MC simulations, then the addition of recorded data requires the selection of a reference trigger that is 100% efficient for our preselected events. We will show that the aforementioned reference trigger is unbiased.

Two-dimensional efficiency and fraction of retained signal

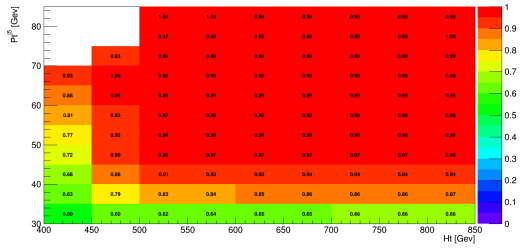
As a first step, the two-dimensional projection of the trigger efficiency versus $(H_T; p_T^{j5})$ is considered (where "$j5$" is the least energetic among the six leading jets, named from "j0" to "j5" in a "C++-like" way). By analysing this two-dimensional map, further offline kinematic cuts for the following multivariate analysis can be inferred. At this stage of the analysis, the trigger efficiencies are "absolute", in the sense that the efficiencies are defined as the ratio of the events triggered by the signal trigger as well as passing the

preselection criteria, over the events only passing the preselection criteria. The signal trigger is considered to be the logic OR between Trg-0 and Trg-2. In order to properly estimate the kinematic cuts for the analysis, the sensitivity $S/\sqrt{B}$ is then studied (see Fig. 3.3(c)), where S is the expected signal yield while B is the background yield derived from QCD multijet simulations.

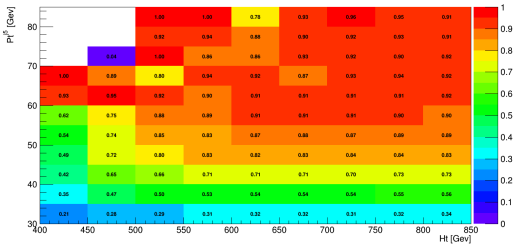
Based on the two-dimensional efficiencies and on the study of the sensitivity (Figs. 3.3(a), 3.3(b) and 3.3(c), the kinematic cuts for the BDT analysis, are set to be:

$$H_T > 450 \text{ GeV} \quad p_T^{j5} > 40 \text{ GeV}\,,$$

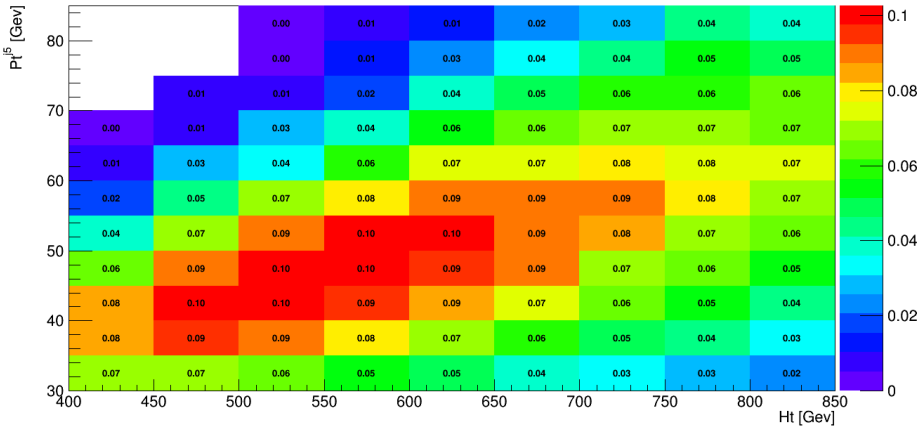
so that the trigger efficiency (of Trg-0 OR Trg-2) is greater than 0.65 on QCD background events and greater than 0.86 on the signal $t\bar{t}H$, without rejecting areas of the phase space in which there is the highest sensitivity.



(a) $t\bar{t}H$ sample: efficiency of Trg-0 OR Trg-2.



(b) QCD sample: efficiency of Trg-0 OR Trg-2.



(c) Signal sensitivity.

Figure 3.3: Two-dimensional trigger efficiency and sensitivity plots versus $(H_T; p_T^{j5})$.

One-dimensional projections and choice of the reference trigger

After setting the new kinematic cuts for the signal trigger, the one-dimensional projections of the efficiencies are drawn. The evaluation of the reference trigger to be used on data for the all-jets $t\bar{t}H$ analysis is performed thanks to these projections. The chosen candidate reference trigger is:

`HLT_PFHT350`.

In the one-dimensional projections in Fig. 3.4, we can see that **the used reference trigger is unbiased** above the fixed kinematic cuts, especially on the QCD sample (that constitutes the greatest part of the data). Furthermore, it is shown that the signal trigger (Trg-0 OR Trg-2) depends on H_T, p_T^{j5}, csv_{max} (that is the maximum value of the b-tag discriminant in a single event), and also on n_{jets}, since our signal triggers actually have a requirement on the number of jets. Instead, no dependency is shown either on η^{j5}, or on the number of vertices in the event (n_{vtx}). The latter means that the performance of the considered trigger is not affected significantly by the pileup.
Please note that in order to show some of the turn-on effects of the trigger on the jet and event kinematic quantities, all preselection criteria except for the one shown in the graph are applied to the events ($N-1$ preselection criteria applied). Additionally, the trigger efficiencies are shown in function of few other relevant variables (e.g. the number of vertices). Since these variables do not appear in the preselection criteria, these plots have N preselection criteria applied.

Turn-on maps and scale factors

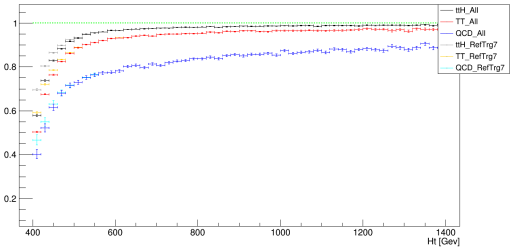
Since the two control triggers have different prescaling on data, events are divided into two non-overlapping categories:

1. events hitting Trg-0 inclusively,

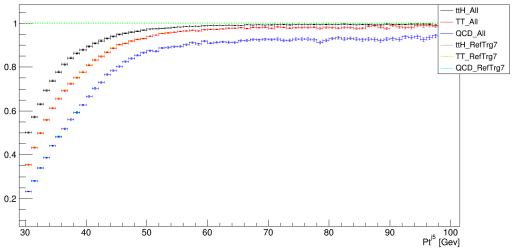
2. events hitting Trg-2 exclusively.

The discrepancy between data and MC simulation is addressed by introducing a correction factor that scales the efficiency curve obtained from simulation to the level of the data curve. Furthermore, data and MC have a better agreement when the two triggers are considered separately, therefore smaller correction factors are needed.

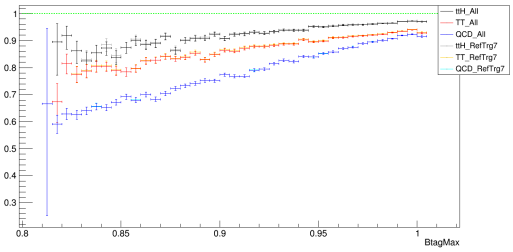
The difference between data and uncorrected background (that is QCD + $t\bar{t}$) is visible in both Fig. 3.5 and Fig. 3.6 (red curves vs black points). In the same figure, the effect of the scale factor application is also visible, where the corrected background (blue) matches the data (black) within less than 10%. The scale factor is derived in bins of the two variables that show the largest difference (data/simulation), also being logical choices given the original trigger logic.



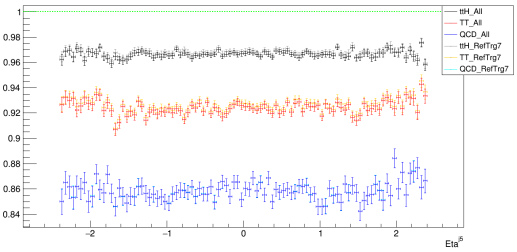
(a) Trigger efficiency vs H_T



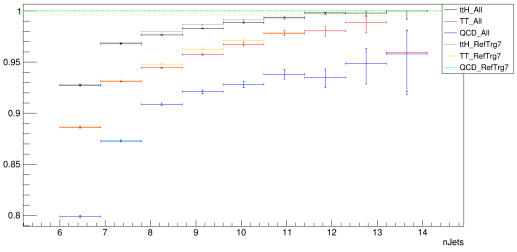
(b) Trigger efficiency vs p_T^{j5}



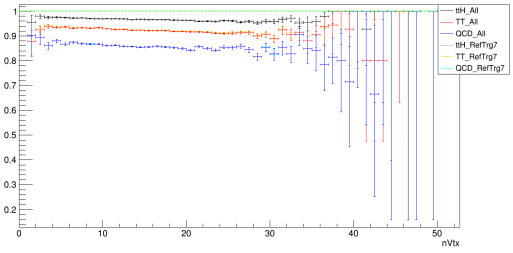
(c) Trigger efficiency vs csv_{max}



(d) Trigger efficiency vs η^{j5}

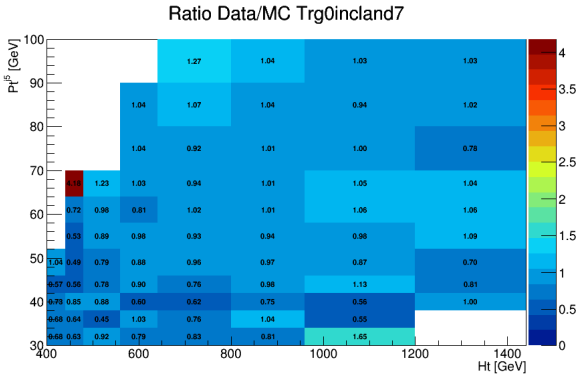


(e) Trigger efficiency vs n_{jets}

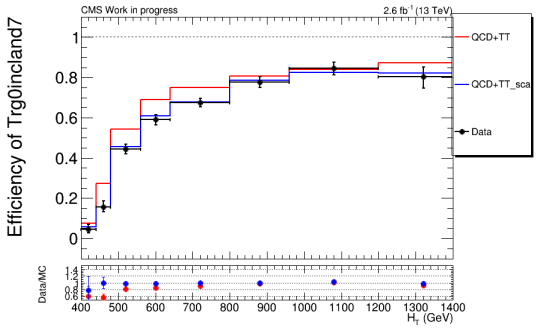


(f) Trigger efficiency vs n_{vtx}

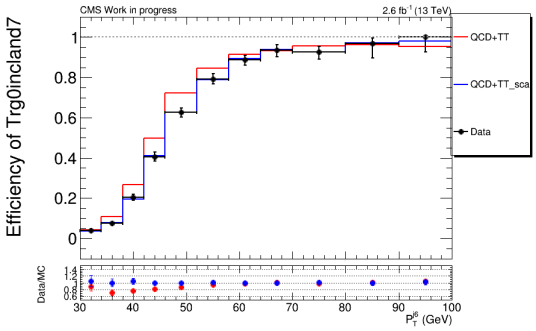
Figure 3.4: One-dimensional projections of the efficiency of Trg-0 OR Trg-2. The dark points refer to the "absolute" efficiencies, whilst the lighter colours of every series correspond to the efficiencies calculated with respect to the reference trigger.



(a) Scale factor as a function of $(H_T; p_T^{j5})$.

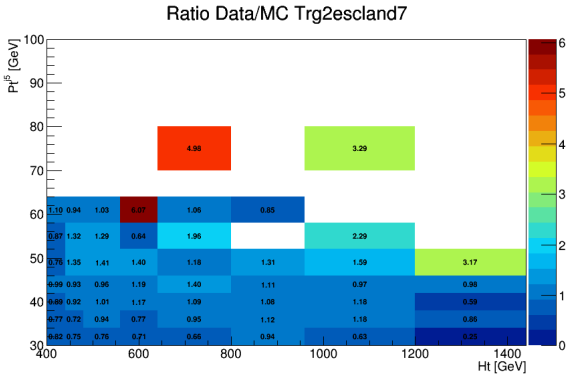


(b) Efficiency vs H_T.

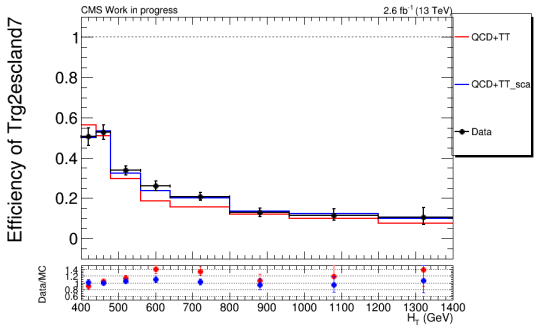


(c) Efficiency vs p_T^{j5}.

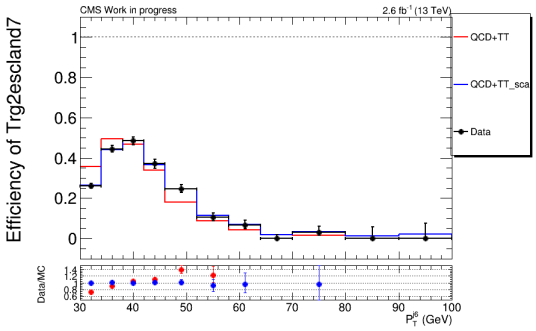
Figure 3.5: Two-dimensional scale factor and one-dimensional trigger efficiency of **Trg-0 inclusive**. The red curves represent MC efficiency plots before applying the two-dimensional scale factor, while the blue curves show the results after the scale factor is applied. Data are in black.



(a) Scale factor as a function of $(H_T; p_T^{j5})$.



(b) Efficiency vs H_T.



(c) Efficiency vs p_T^{j5}.

Figure 3.6: Two-dimensional scale factor and one-dimensional trigger efficiency of **Trg-2 exclusive**. The red curves represent MC efficiency plots before applying the two-dimensional scale factor, while the blue curves show the results after the scale factor is applied. Data are in black.

Sample	Event yields
Data	297570
QCD	265459
$t\bar{t}$	44416
WW	22
ZZ	55
$t\bar{t}W$ + jets	85
$t\bar{t}Z$ + jets	144
DY + jets	216
W + jets	164
$t\bar{t}H(b\bar{b})$	126

Table II: Selection yields for $L = 2.63$ fb^{-1}.

3.6 Event selection

Before exploring the details of the BDT we have built, it is necessary to go through the selection criteria data and MC events should meet before being fed to the BDT analysis.

Based on both the topology of the signal events and the study of the trigger efficiency, the selection requirements for the multivariate analysis are:

- Trigger Pass: selection of the event firing the logic OR of the two signal triggers (Trg-0 OR Trg-2);
- Jet selection: $n_{jets} \geq 6$ with $|p_T| > 30$ GeV and $|\eta| < 2.4$;
- $H_T = \sum |p_T| > 450$ GeV and $|p_T| > 40$ GeV (of the six leading jets), based on the trigger efficiency studies;
- B-tagging: $n_{b-jets} \geq 2$, with $csv \geq 0.800$ (corresponding to the medium CSVv2 working point);
- Successful kinematic fit for the $t\bar{t}$ hypothesis;
- Veto on (MC) leptons – to ensure orthogonality to dileptonic and semi-leptonic analyses;

Table II gives the estimated yields after the event selection, leading to $S/B \approx 1/2500$.

3.7 The BDT Analysis

The event selection reaches a very small value for S/B so we need to introduce an additional quantity to improve the discrimination between the small signal and the huge background. This critical task requires a discriminator that only a multivariate analysis can provide. We thus recur to the Boosted Decision Tree (BDT) technique.

3.7.1 Boosted Decision Trees

A decision tree [33] is a tree-structured classifier similar to the one represented in Fig. 3.7. Repeated left/right (yes/no) decisions are taken on one single variable at a time, until a certain stop criterion is fulfilled. In this way, the phase space is split into many regions that are eventually classified as signal or background, depending on the majority of training events that end up in the final leaf node.

Boosting a decision tree consists of enhancing its classification performance and increasing its stability, with respect to statistical fluctuations in the training sample. It is usually conducted by sequentially applying an MVA algorithm to reweighted (boosted) versions of the training data and then taking a weighted majority vote of the sequence of MVA algorithms thus produced.

Having said that, it is easy to understand that the boosting of a decision tree extends this concept from one tree to several trees which form a forest. The trees belonging to such forest are derived from the same training ensemble by reweighting events, and are finally combined into a single classifier (*bdt*) that is given by a weighted average of the individual decision trees.

We have seen that decision trees are well known classifiers that allow a straightforward interpretation, as they can be visualized by a simple two-dimensional tree structure and this feature makes them very similar to rectangular cuts. However, whereas a cut-based analysis is able to select only one hypercube as region of phase space, the decision tree splits the phase space into a large number of hypercubes, each of which is identified as either "signal-like" or "background-like".

After configuring a decision tree, its training has to be performed. In particular, the training of a decision tree is the process that defines the splitting criteria for each node. The training starts with the root node, where an initial splitting criterion for the full training sample is determined. The split results into two subsets of training events that each go through the same algorithm to determine the next splitting iteration. This procedure is repeated until the whole tree is built. At each node, the split is determined by finding the variable and the corresponding cut value that provides the best separation between signal and background. The node splitting stops once it has reached the minimum number of events which is specified in the BDT configuration. The leaf nodes are classified as signal or background according to the class the majority of events belongs to.

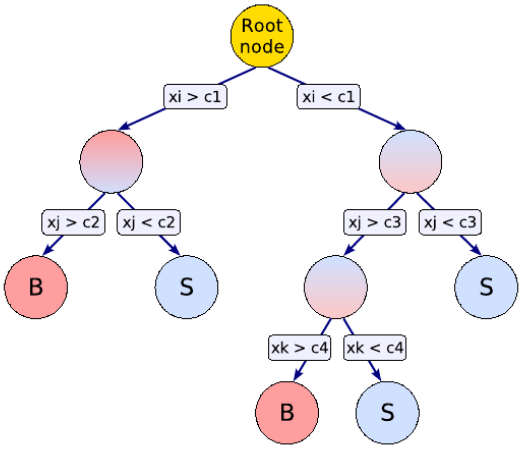


Figure 3.7: The schematic view of a decision tree. Starting from the root node, a sequence of binary splits is applied to data, using the discriminating variables "xi". Each split uses the variable that gives the best signal-background separation, when being cut on. The same variable may be used in several nodes, while others might not be used at all. The leaf nodes at the bottom end of the tree are labeled "S" for signal and "B" for background, depending on the majority of events that end up in such nodes.

Only limited experience has been gained so far with boosted decision trees in high energy physics, and decision trees are sometimes referred to as the best "out of the box" classifiers. This is because little tuning is required in order to obtain reasonably good results. This is due to the simplicity of the method, since each training step (node splitting) involves only a one-dimensional cut optimisation. What is more important, decision trees are also insensitive to the inclusion of poorly discriminating input variables, differently from neural networks. Indeed, what happens is that the decision tree training algorithm basically ignores non-discriminating variables, and for each node splitting only the best discriminating variable is used.

After the training, the BDT is ready to be fed with data events, and the classification between "signal" and "background" is performed.

3.7.2 Preliminary study on the BDT variables

The events meeting the BDT preselection requirements are then classified as signal-like or background-like by the BDT, based on a set of variables which describe the kinematical properties of the individual jets of the whole event. A preliminary study has been carried out, in order to individuate the most and the least significant ones, from an initial given set.

Figure 3.8 shows the areas of the ROC curves obtained after training twenty-two

different BDTs. Every BDT differs from the initial BDT (the one called "All" – in black) due to the fact that the training has been performed removing only one variable from the inital set. We remind that a ROC curve is a plot that illustrates the performance of a binary classifier system, plotting the background rejection ability of the classifier versus the signal efficiency (that is the fraction of signal that is retained by the classifier). Therefore, the computation of the area under the ROC curve gives an indication of the performance of the classifier: the bigger the area, the more powerful the discriminator is.

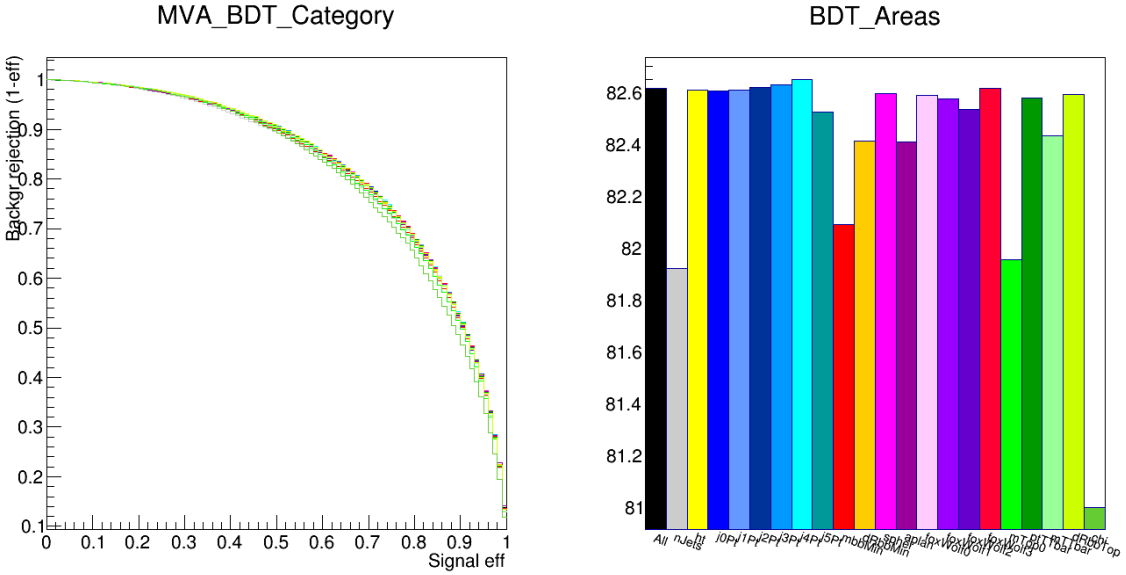


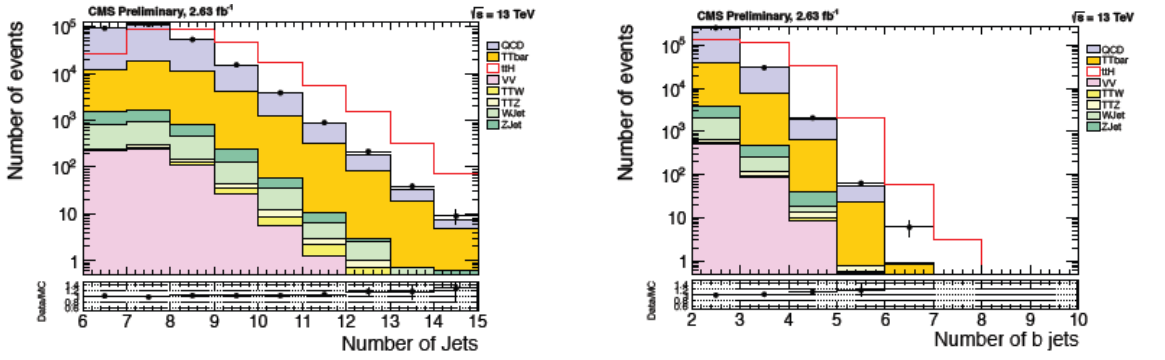
Figure 3.8: ROC curves and ROC curves areas obtained when training several BDTs, each of which lacks one variable. The removed variable is the one used to label the actual BDT, and therefore the corresponding histogram column in the plot on the right. The entity of such areas has led to determine that the most powerful variables are the χ^2 of the kinematic fit for the $t\bar{t}$ hypothesis (see Section 3.7.3), the top quark mass and the number of jets n_{jets}. Conversely, other variables – such as the p_T of the leading jets – do not play a major role in the discrimination and therefore are removed from the final machinery.

This means that, qualitatively, the variable that has the highest discriminating power is the one removing which the ROC curve area is the smallest. Following this criterion, this preliminary study has allowed to remove from the machinery all the variables that did not play a major role in the discrimination.

3.7.3 Discriminating variables

After studying the behaviour of the initial set of variables thanks to the preliminary study, and also carrying out additional investigations, the final BDT is built. It is based on more than twenty different variables, that can be split into several categories:

1. Jet multiplicity variables (Fig. 3.9):
   - the number of jets in each event: n_{jets} (Fig. 3.9(a));
   - the number of b-tagged jets in each event: n_{b-jets} (Fig. 3.9(b));

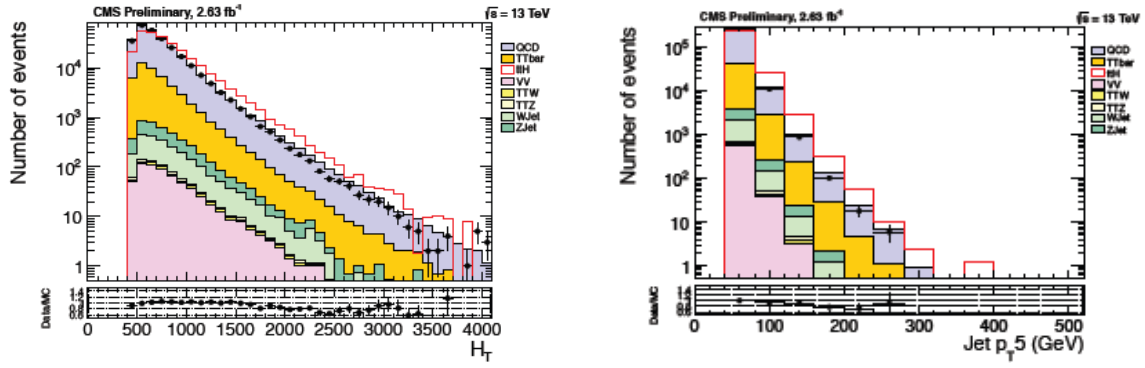


(a) (b)

Figure 3.9: Distributions of the number of jets and the number of b-tagged jets, after the preselection, for data and simulated samples. All the background contributions are stacked one upon the other, normalised to $L = 2.63$ fb^{-1}, while the signal $t\bar{t}H$ contribution is normalised to the total number of events, so that the difference in shape can be appreciated.

2. Jet hardness variables (Fig. 3.10):
   - H_T (Fig. 3.10(a));
   - jet P_T of the four subleading ones (p_T^{j2}, p_T^{j3}, p_T^{j4}, p_T^{j5} – Fig. 3.10(b)), since the two leading jets have no discriminating power;

3. Quark-gluon jet discriminant [34]: a likelihood discriminant that has been developed in order to separate jets originating from gluons or light-quarks, ranging between 0 (meaning that the jet is gluon-like) and 1 (meaning that the jet is quark-like).
   - qgl_{min}: minimum value of the likelihood discriminant in the event;



(a) (b)

Figure 3.10: Distributions of H_T and P_T of the sixth leading jet, after the preselection, for data and simulated samples. All the background contributions are stacked one upon the other, normalised to $L = 2.63$ fb^{-1}, while the signal $t\bar{t}H$ contribution is normalised to the total number of events, so that the difference in shape can be appreciated.

- qgl_{median}: median value of the likelihood discriminant in the event (Fig. 3.11);
- qgl_{avg}: average value of the likelihood discriminant in the event;

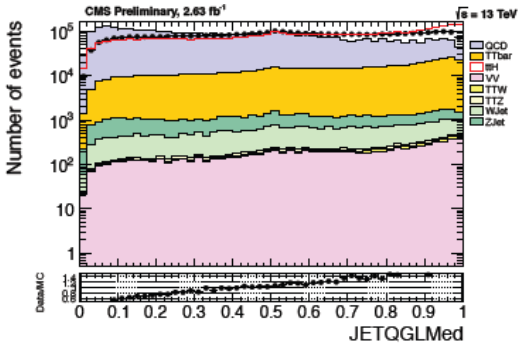


Figure 3.11: Distributions of the median of the quark-gluon discriminant, after the preselection, for data and simulated samples. All the background contributions are stacked one upon the other, normalised to $L = 2.63$ fb^{-1}, while the signal $t\bar{t}H$ contribution is normalised to the total number of events, so that the difference in shape can be appreciated.

4. Centre-of-mass variables:

- $\cos\theta_1^*$ (Fig. 3.12), where θ_1^* is the angle between the leading jet ($j0$) and the z axis in the centre-of-mass framework of the multijet system;

- $\cos\theta_2^*$, where θ_2^* is the angle between the first sub-leading jet ($j1$) and the z axis in the centre-of-mass framework of the multijet system;
- E_T^1: module of p_T^{j0} in the centre-of-mass framework of the multijet system;
- E_T^2: module of p_T^{j1} in the centre-of-mass framework of the multijet system;

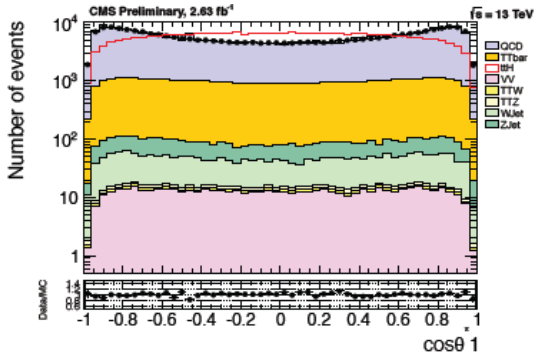
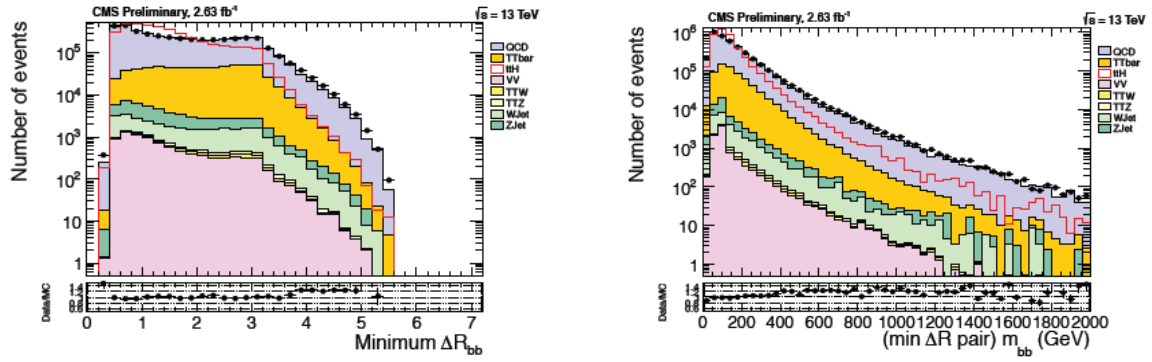


Figure 3.12: Distributions of one of the centre-of-mass variables ($\cos\theta_1^*$), after the preselection, for data and simulated samples. All the background contributions are stacked one upon the other, normalised to $L = 2.63\,\text{fb}^{-1}$, while the signal $\text{t}\bar{\text{t}}\text{H}$ contribution is normalised to the total number of events, so that the difference in shape can be appreciated.

5. B-tagged jets variables (Fig. 3.13):
   - $\Delta R_{bb,min}$: the angular distance between the closest pair of b-tagged jets (Fig. 3.13(a));
   - $m_{bb,min}$: the invariant mass of the closest pair of b-tagged jets (Fig. 3.13(b));
6. Event shape variables (Fig. 3.14):
   - centrality: $C = \frac{\sum_i p_T^i}{\sum E_{vis}}$ (Fig. 3.14(a));
   - sphericity: $S^{\alpha\beta} = \frac{\sum_i p_i^\alpha p_i^\beta}{\sum_i |p_i|^2}$ (Fig. 3.14(b));
   - Fox-Wolfram variables: $H_l = \sum_{i,j} \frac{|p_i||p_j|}{E_{vis}^2} P_l(\cos\theta_{ij})$ (H_0 in Fig. 3.14(c));

   where p_i and p_j are the four-momenta of jets i and j, E_{vis} is the visible energy of the jet in the events, and P_l is the Legendre polynomial of order l.
7. Variables correlated to the kinematic fit for the $\text{t}\bar{\text{t}}$ hypothesis (Fig. 3.15). The kinematic fit is performed under two constraints, these being: a fixed value m_W for the dijet masses associated to the W's coming from both the t and the $\bar{\text{t}}$ quarks,



(a) (b)

Figure 3.13: Distributions of the minimum angular distance between the two of the ($\geq$ 2) preselected b-tagged jets and their invariant mass, after the preselection, for data and simulated samples. All the background contributions are stacked one upon the other, normalised to $L = 2.63$ fb^{-1}, while the signal t$\bar{\text{t}}$H contribution is normalised to the total number of events, so that the difference in shape can be appreciated.

and $m_t = m_{\bar{t}}$ for the two jet triplets associated to the top quark and antiquark in the t$\bar{\text{t}}$ pair:

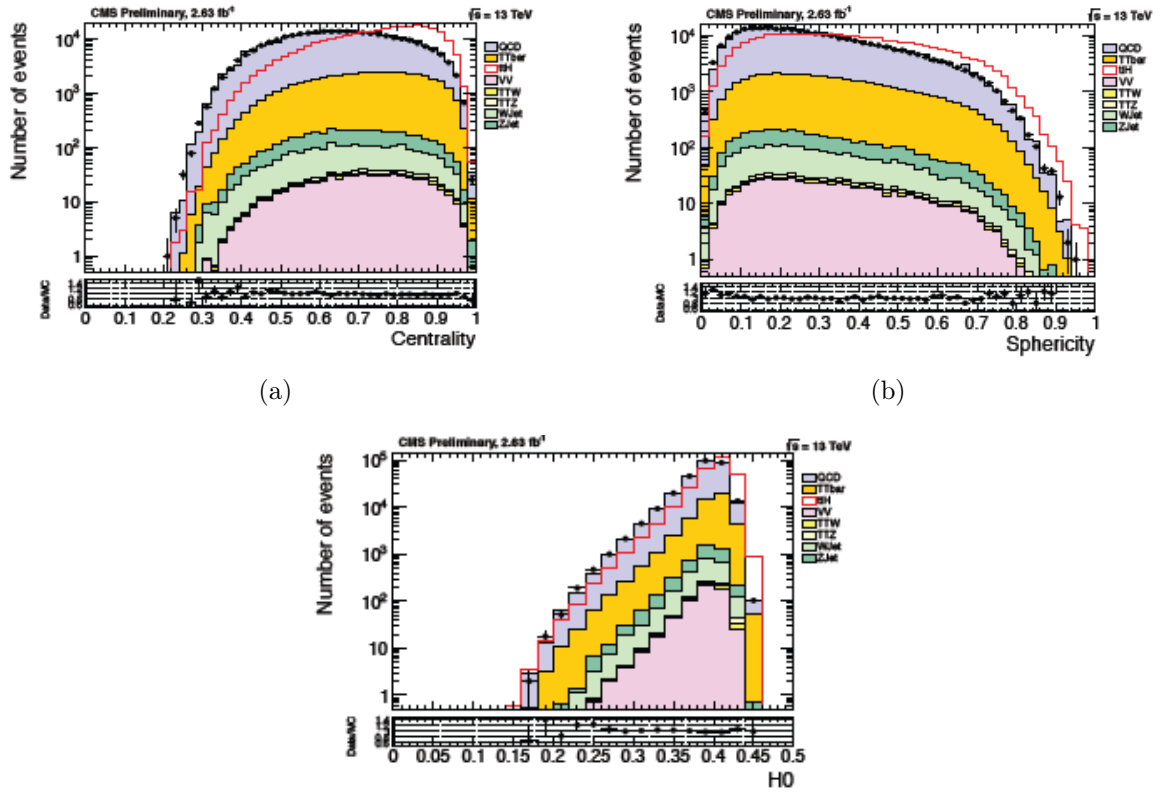
- m_t (Fig. 3.15(a));
- $m_{t\bar{t}}$;
- $p_T^{t\bar{t}}$;
- $\Delta R_{bb}^{t\bar{t}}$ (Fig. 3.15(b));
- χ^2.

3.7.4 BDT performance

We will now discuss the implementation of the BDT procedure that has been built.

Categories of the BDT

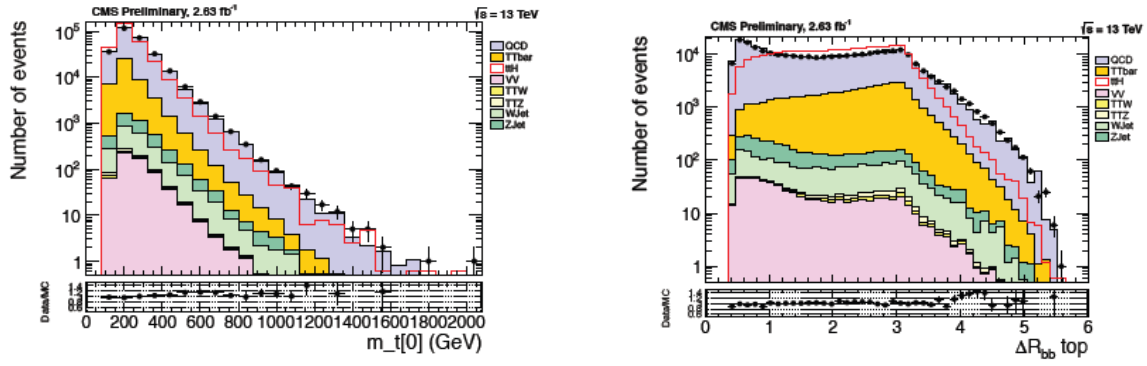
The BDT we have built performs differently on two categories of events: events having only two b-tagged jets (Category 1) and events having more than two b-tagged jets (Category 2). This choice allows to reach a greater sensitivity with respect to the case with only one category involved and also to a better combination of the BDT analysis results with a parallel analysis based on the use of a Matrix Element discriminator.



(a) (b)

(c)

Figure 3.14: Distributions of some of the event shape variables (centrality, sphericity and the first Fox-Wolfram variable – H_0), after the preselection, for data and simulated samples. All the background contributions are stacked one upon the other, normalised to $L = 2.63$ fb^{-1}, while the signal $t\bar{t}H$ contribution is normalised to the total number of events, so that the difference in shape can be appreciated.



(a) (b)

Figure 3.15: Distributions of some of the variables used in the kinematic fit for the $t\bar{t}$ hypothesis (mass of the top quarks and antiquarks, and the angular distance between the $b\bar{b}$ pair deriving from the decay of the $t\bar{t}$ pair), after the preselection, for data and simulated samples. All the background contributions are stacked one upon the other, normalised to $L = 2.63$ fb^{-1}, while the signal $t\bar{t}H$ contribution is normalised to the total number of events, so that the difference in shape can be appreciated.

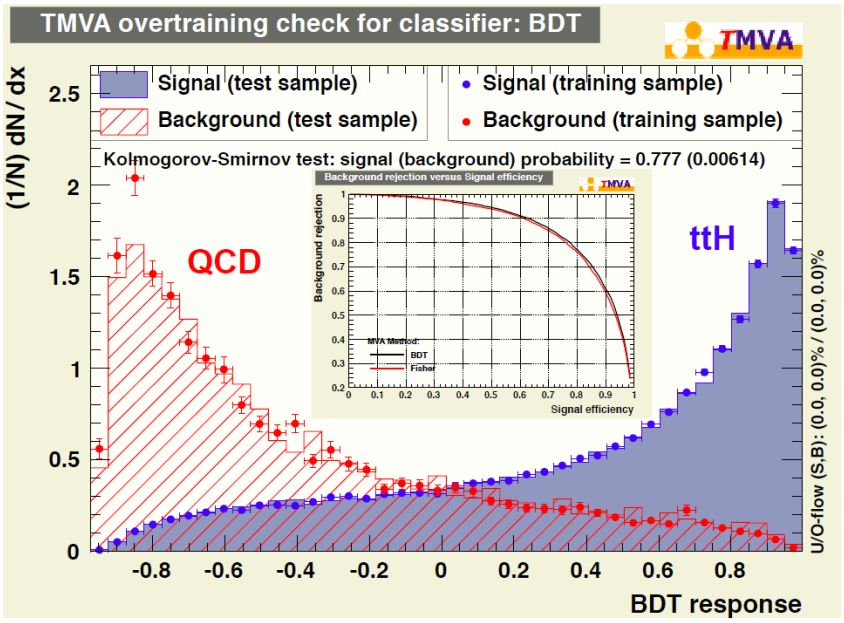
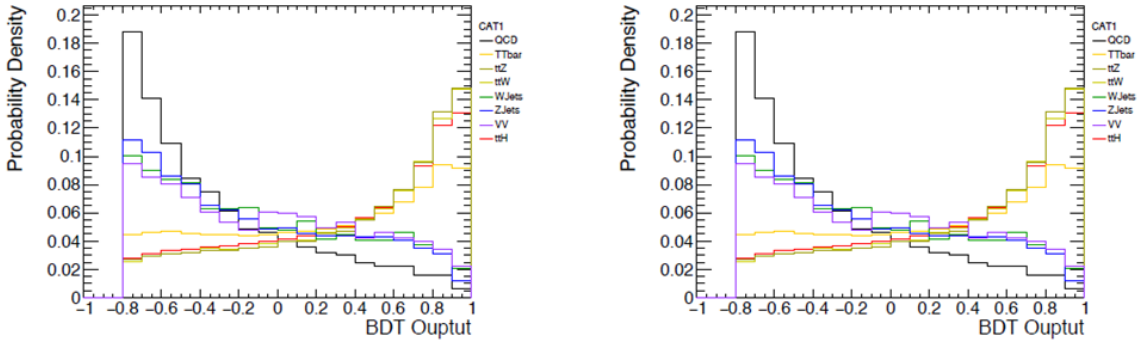


Figure 3.16: Results of the training of the BDT and output of the overtraining check: the shapes of the BDT discriminant show a very good agreement between the training samples and the test samples for $bdt > -0.8$, meaning that the BDT has not been overtrained.

Training of the BDT

Figure 3.16 shows the performance of the training of the BDT on both the signal and background contributions. The background used here is modeled with QCD multijet simulated events, which represent the large majority of the total background. Signal and background are well separated by the BDT, as it can be easily seen also from the ROC curve in the centre. Furthermore, the figure shows that the BDT has a consistent behaviour when considering the differences between the training samples and the test samples, this meaning that it was not overtrained. Differences in behaviour between the training and the test sample can be appreciated below $bdt < -0.8$, therefore we will remove such region in the following steps of our analysis.

BDT output shapes on MC samples



(a) Cat 1: events with exactly 2 b-tagged jets.

(b) Cat 2: events with > 2 b-tagged jets.

Figure 3.17: Output of the BDT on the different MC samples, for $bdt > -0.8$.

Figure 3.17 shows the performance of the BDT on the different MC samples. In particular we can see that W + jets, Z + jets and diboson events have "QCD-like" shapes, meaning that their contributions can be largely reduced by the application of a cut on the BDT discriminant. This behaviour can be easily understood if we consider that such processes do not have the characteristics of a $t\bar{t}$ final state.

On the other hand, $t\bar{t}$, $t\bar{t}W$ and $t\bar{t}Z$ background processes are "$t\bar{t}H$-like", and $t\bar{t}W$ and $t\bar{t}Z$ are practically indistinguishable from the signal $t\bar{t}H$.

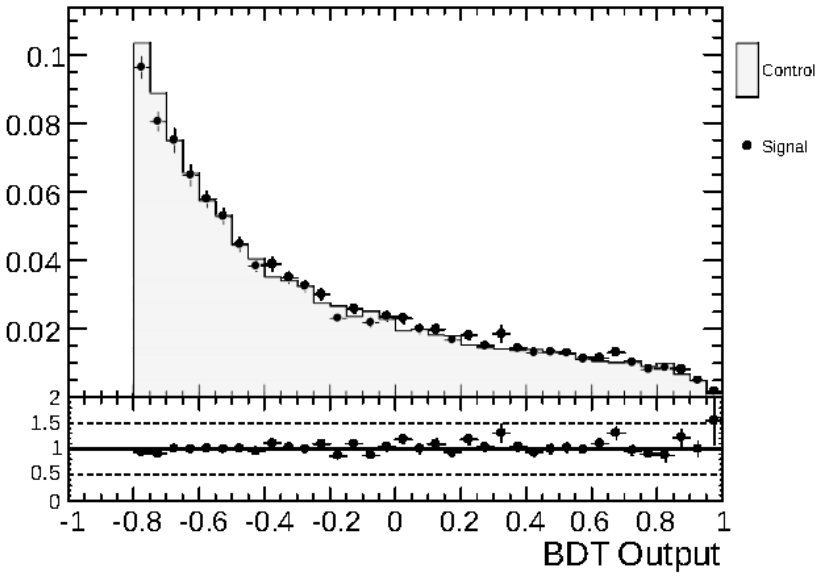
3.8 QCD background estimation

Given the large cross section in pp collisions, the QCD multijet production is the dominant background for the $t\bar{t}H$ events in the all-jets final state and actually, as we have seen

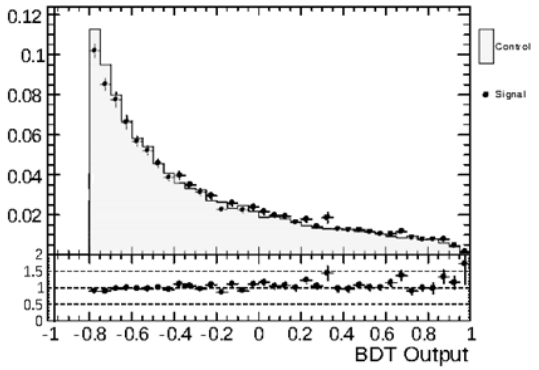
in Table II, the most significant subdominant background ($t\bar{t}$) contributes for only 15% to the total background. Since QCD multijet production cannot be safely reproduced by MC simulations, we define a control sample, QCD-enriched, that can be used for a data-driven estimation of the shape of the QCD contribution. As we have seen so far, our *signal region* is the region of the phase space where we have events firing the signal triggers and having at least two "medium" b-tagged jets. At this point, then, we define the *control region* reverting this conditions and therefore asking for no medium b-tagged jets, but at least two "loose" jets and the firing of the control triggers (meeting the same kinematic requirements as the signal triggers, but having no b-tag selection). In this way we obtain an almost pure QCD sample, that exhibits a behaviour similar to that of QCD multijet events in the signal region.

We have seen what the output of the BDT is, with respect to the different MC samples. What is particularly interesting, is its behaviour for the QCD contribution. In Fig. 3.18 a closure test is performed. Such test compares the BDT response on the simulated QCD background contribution in the signal region (black points) and in the control region (filled area).

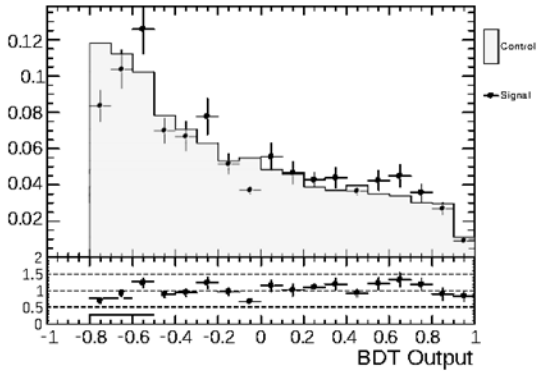
Therefore, after considering such distributions, the cut on the BDT output ($bdt > -0.8$) is confirmed, this being also the threshold above which a good accordance is shown in the QCD closure test.



(a) All categories.



(b) Category 1: events with exactly 2 b-tagged jets.



(c) Category 2: events having $>$ 2 b-tagged jets.

Figure 3.18: Closure of the QCD background prediction for the BDT output for the QCD simulated sample, in the control and signal regions.

Chapter 4

Experimental results

4.1 Full selection and BDT results

At this point, after adding the cut $bdt > -0.8$ on the output of the BDT, the selection of the analysis is complete. Fig. 4.1 shows final BDT distributions of data and of the signal and background contributions.

In Table I the corresponding yields are shown, for both the categories of the BDT. At this point the signal-over-background ratio corresponds to: $S/B \approx 1/2000$, definitely still too small to highlight any presence of a signal contribution. For this reason we will now proceed with the estimation of the sensitivity of the BDT as a way to extract a small signal from a large background.

4.2 Sensitivity estimation

In this section the experimental procedure to obtain the limits on the BDT sensitivity is described.

4.2.1 Statistical method

Estimating the sensitivity of our machinery means deriving 95% confidence level upper limits on the $t\bar{t}H$ production cross section, with respect to the SM expectation. This is usually done introducing a signal strength modifier μ defined as:

$$\mu = \frac{\sigma^{95\%}_{t\bar{t}H}}{\sigma^{SM}_{t\bar{t}H(b\bar{b})}}. \tag{4.1}$$

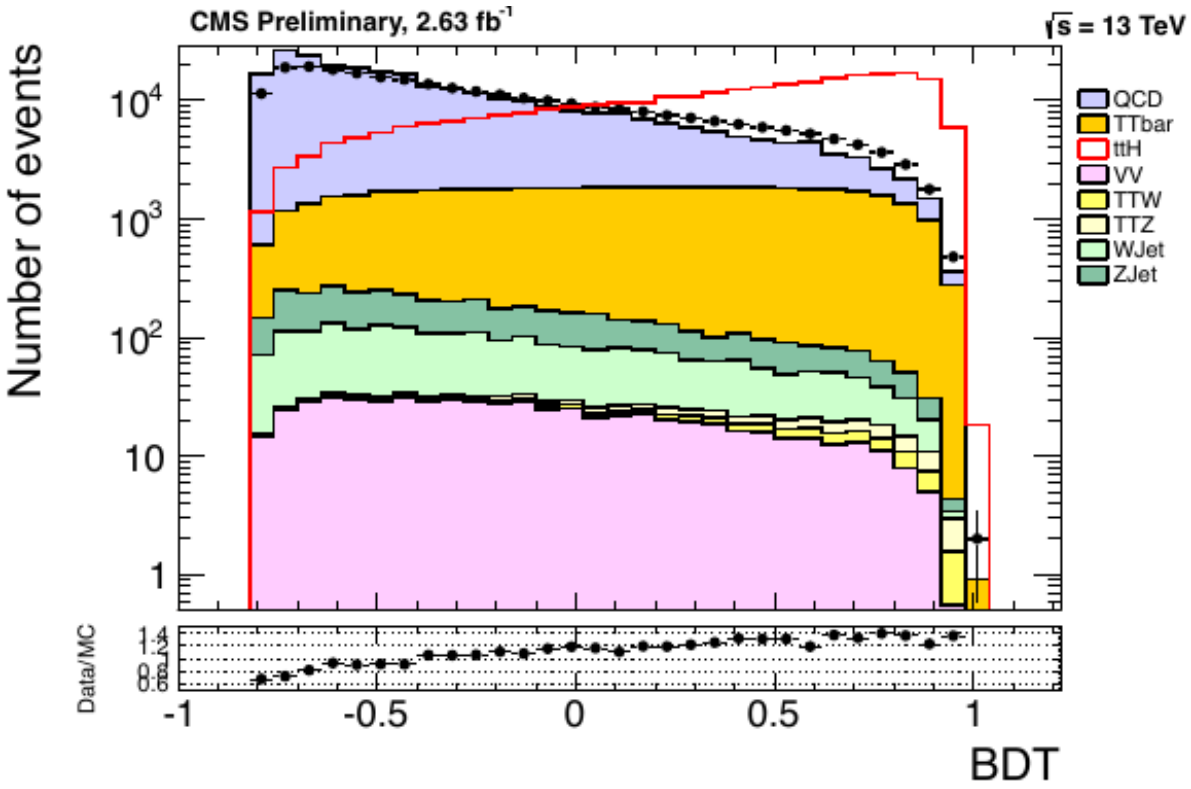


Figure 4.1: Results of the classification operated by the BDT. All the background contributions are stacked one upon the other and normalised to $L = 2.63$ fb^{-1}, while the signal $t\bar{t}H$ contribution is normalised to the total number of events, so that the difference in shape can be appreciated.

Sample	Category 1 (2 b-jets)	Category 2 (>2 b-jets)	All Categories
Data	233240	29500	262669
QCD	199237	22390	22472
$t\bar{t}$	35910	7535	43439
VV	64	9	73
$t\bar{t}W$ + jets	72	17	89
$t\bar{t}Z$ + jets	99	43	142
DY + jets	176	24	200
W + jets	135	14	149
$t\bar{t}H(b\bar{b})$	59	66	125

Table I: Yields after the full selection for $L = 2.63$ fb^{-1}.

The statistical method used here is the modified frequentist approach, also known as CL_s [35].

For the CL_s method, the likelihood function $\mathcal{L}\left(\text{data}|\mu,\theta\right)$ is defined as:

$$\mathcal{L}\left(\text{data}|\mu,\theta\right) = Poisson\left(\text{data}|\mu\cdot s\left(\theta\right)+b\left(\theta\right)\right)\cdot p\left(\tilde{\theta}|\theta\right) = \\ = \prod_i \frac{\left(\mu s_i + b_i\right)^{n_i}}{n_i!} e^{-\left(\mu s_i + b_i\right)} \cdot p\left(\tilde{\theta}|\theta\right)$$

where μ is the signal strength modifier defined in Equation (4.1) and θ represents a full set of nuisance parameters that are used to incorporate systematic uncertainties (see Paragraph 4.2.2). At this point, the (joint) probability density function of the nuisance parameters $p\left(\tilde{\theta}|\theta\right)$, where $\tilde{\theta}$ is the default value, reflects the degree of belief in what the true value of θ could be.

Then, in order to compare the compatibility of the data with the "background-only" and "signal+background" hypotheses, where the signal is allowed to be scaled by a factor μ, the test statistic $\tilde{q}_\mu$ is constructed, based on the profile likelihood ratio:

$$\tilde{q}_\mu = -2ln\frac{\mathcal{L}\left(\text{data}|\mu,\hat{\theta}_\mu\right)}{\mathcal{L}\left(\text{data}|\hat{\mu},\hat{\theta}_\mu\right)}, \qquad 0\leq\hat{\mu}\leq\mu\,,$$

where $\hat{\theta}_\mu$ refers to the conditional maximum likelihood estimator of θ, given the signal strength parameter μ and data. The pair of parameter estimators $\hat{\mu}$ and $\hat{\theta}$ correspond to the global maximum of the likelihood.

The computation of the limits, then, is performed by using the RooStats-based statistical analysis tools recommended within the CMS Collaboration. The corresponding software is the "Higgs Combination" package [36]. Finally, the analysis is conducted through the "asymptoptic" approach [37], that makes an analytic approximation of the full CL_s technique and therefore avoids performing pseudo-experiments.

The distinct background and signal contributions that have been taken into account for the limit calculation are those reported in Table I, whose rates and BDT distributions are allowed to fluctuate according to a set of nuisance parameters, that we will discuss in the following paragraph.

4.2.2 Systematic uncertainties

Before mentioning the procedure that has been followed, it is important to underline that we are performing an estimation on the expected sensitivity of the tool we have developed. Therefore, we do not aim to give an extensive and detailed description of the

Source	Type	Entity	Notes
Luminosity	`lnN`	± 2.6%	Recommended value.
Jet Energy Scale	`lnN`	± 10%	Conservative estimate.
b-Tag uncertainty	`lnN`	± 4%	Due to the requirement on the two b-tagged jets.
Cross section	`lnU`	± 20%	On the QCD contribution.
	`lnN`	± 10%	On the MC processes.

Table II: Summary for the systematic uncertainties considered on the inputs to the limit calculation. Please note that `lnN` stands for "log-normal", which is the recommended choice for multiplicative corrections, while `lnU` stands for the "log-uniform" distribution, that is normally useful to set a large a priori uncertainty on a given background – QCD multijet production in our case.

systematic uncertainties. Instead, what is discussed in this paragraph is the individuation of some significant sources of systematic uncertainties and their rough estimate.

As we have already mentioned, such systematic uncertainties are used to constrain the input nuisance parameters for the Combine Tool. Table II summarises what have been currently considered to be the most significant contributions to the systematic uncertainties and the estimate of their amount, as they have been used to compute the limits.

4.2.3 Sensitivity estimate

We have seen in Paragraph 4.2.1, that we can define a figure of merit by which we can quantitatively assess the performance of the machinery we have developed so far, and therefore the state of the analysis, without appealing to the data. This figure of merit is the expected limit on the signal strength μ, that – again – is the 95% confidence level upper limits on the $t\bar{t}H$ production cross section, with respect to the theoretical value. In other words, it is the magnitude of μ that we expect to be able to exclude, at the 95% of the confidence level, under the assumption that the observed data look exactly like the modelled backgrounds. With this in mind, this is the limit we have obtained:

$$\mu = 8.2^{+3.4}_{-2.4}.$$

What this value means is that our machinery is expected to potentially exclude a $t\bar{t}H$ production cross section roughly eight times stronger than the SM expectations, at the 95% of the confidence level. In other words, up to this value of μ, the signal presence can be "absorbed" by fluctuations of the background distribution and therefore remain undetected.

4.3 Projection of the results on the full data set

After studying the methodology on a small sample, we will now attempt to extend the results on the full 2016 data set, which has been only recently made available, and corresponding to an integrated luminosity of $L = 36$ fb^{-1}. It is important to keep in mind that such projection on the full data set is obtained by following the exact steps we have seen above, unless otherwise stated. Indeed, extending the analysis to the full data set is not straightforward, but it requires some tuning, due to the different experimental conditions.

4.3.1 QCD background prediction

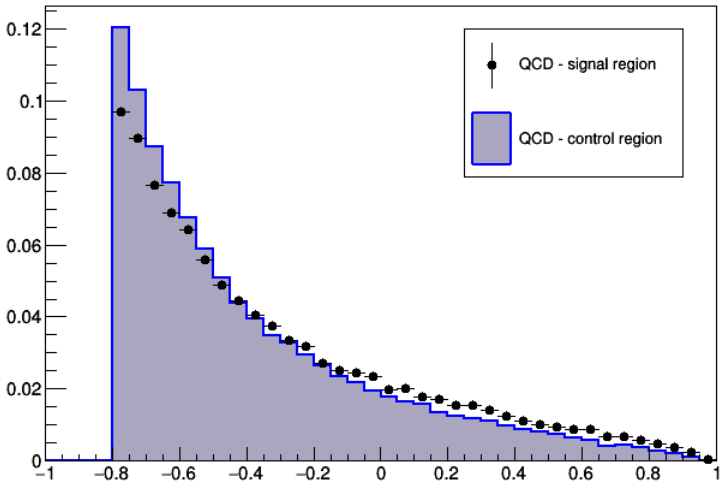
One of the most difficult steps for the extension is to obtain a good a data-driven estimation of the QCD background. The reason for this is that during 2016 LHC has delivered to the experiments an unprecedented instantaneous luminosity, even higher than what expected by design. For this reason, it has been necessary to "prescale" many trigger paths, in order to limit the rate of data taking to operating values.

Unfortunately, the control triggers we have used to define the control region were heavily prescaled during the data taking. What is more delicate, the prescaling was not the same for the two trigger paths and also, it varied throughout the weeks of data taking. For this reason it is not possible to use the control region we have seen in Section 3.8. Before dealing with the details of the procedure, we remind that the QCD multijet production is the dominant background for $t\bar{t}H$ events, and that such contribution cannot be safely reproduced by MC simulations. Therefore, we need to define a control sample, QCD-enriched, that can be used for a data-driven estimation of the shape of the QCD contribution. The control region in Section 3.8 was defined by asking no medium b-tagged jets, but at least two loose b-jets and the firing of the logic OR of the two control triggers. Given their prescale, we cannot meet the requirement on the control triggers anymore. The new control region is then defined by asking the firing of the sole *reference trigger*, along with the requests on the loose b-tags. Figure 4.2 shows the closure test, comparing the BDT response on the simulated QCD background contribution in the signal region (black points) and in the control region (filled area). As we can see, the two shapes are in good agreement in the central bins, but not in the rest of the distribution.

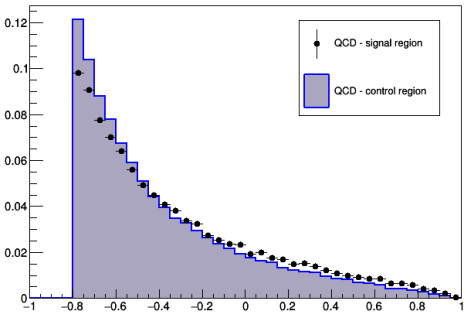
4.3.2 BDT results

At this point, we are ready to repeat the analysis, based on the new QCD background estimation: we implement the full selection and finally feed the BDT with our signal, background and data events.

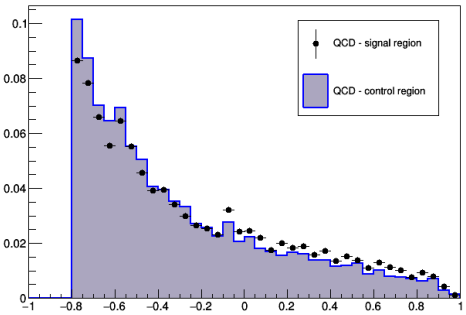
Figure 4.3 shows the final BDT distributions of data and of the signal and background contributions. We can see that there are not negligible discrepancies between the ex-



(a) All categories.



(b) Category 1: events with exactly 2 b-tagged jets.



(c) Category 2: events having $>$ 2 b-tagged jets.

Figure 4.2: Closure of the QCD background prediction for the BDT output for the QCD simulated sample, in the control and signal regions, for the 2016 data set.

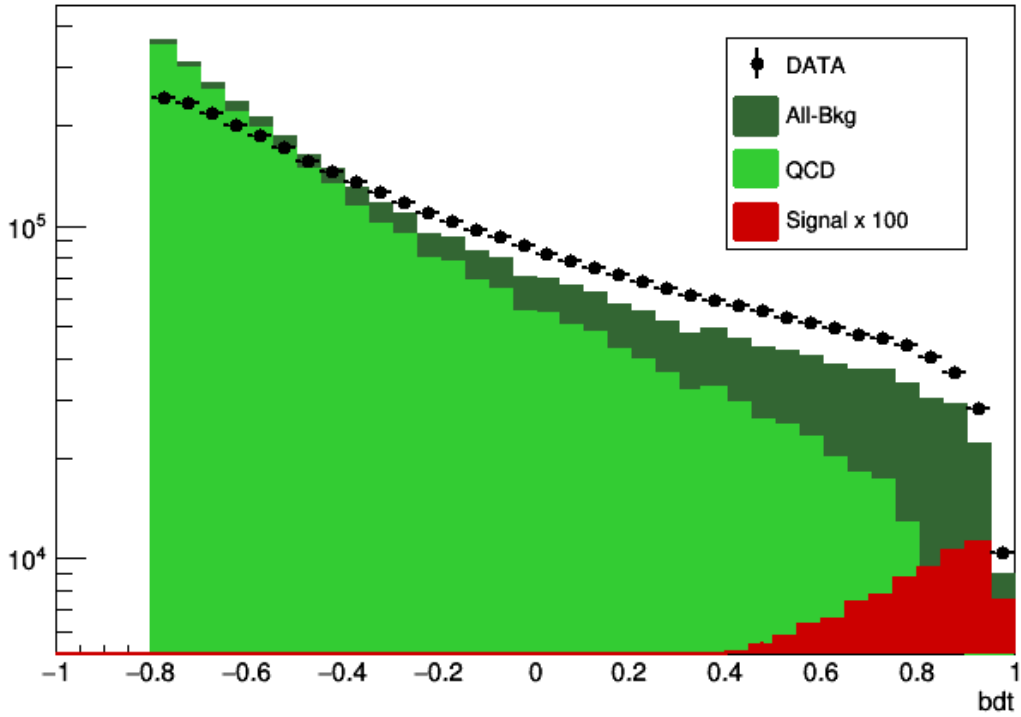


Figure 4.3: Result of the classification operated by the BDT, for the 2016 data set. All the background contribution are normalised to the integrated luminosity $L = 36$ fb^{-1}, as well as the signal, multiplied by a factor 100.

Sample	Category 1 (2 b-jets)	Category 2 (>2 b-jets)	All Categories
Data	3161400	348260	3503660
QCD	2667863	270811	2938674
$t\bar{t}$	466704	72895	539599
VV	810	112	922
$t\bar{t}W$ + jets	885	166	1051
$t\bar{t}Z$ + jets	1335	489	1824
DY + jets	22705	2963	25668
W + jets	170	15	185
$t\bar{t}H(b\bar{b})$	922	810	1732

Table III: Yields after the full selection for $L = 36$ fb^{-1}.

pected distribution and the one actually obtained, probably due to the poor modelisation of the QCD multijet background. In Table III the corresponding yields are shown, for both the categories of the BDT. As in the case of the partial data set, after performing the full selection we obtain a $S/B \approx 1/2000$, still too small to highlight any presence of a signal contribution. We will now proceed with the estimation of the sensitivity of the BDT.

4.3.3 Sensitivity estimate

The projection on the full 2016 data set of the expected limit on the signal strength results to be:

$$\mu = 2.1^{+2.1}_{-1.6}.$$

Therefore, potentially, recurring to the full data set would increase the sensitivity of our machinery by a factor of four. It is important to notice, though, that not only do the evaluation of systematic uncertainties is at an early stage (as in the case of the computation of μ in the partial data set), but this result is also affected by a not well-performing QCD background modelisation. Such weakness leads to a remarkable instability of the expected limits and their dependence on the initial conditions. Further studies are therefore needed.

4.4 Future perspectives

Now, we would like to present some prospects on possible future improvements for this analysis.

After tuning the QCD data-driven estimation, the event selection may be still refined. Secondly, the variables to be fed to the BDT may be further optimised. What is more relevant, an additional BDT which discriminates against the $t\bar{t}$ background may be added: indeed, it may be noticed in Fig. 3.17, that the shape of the $t\bar{t}$ sample presents slight differences from the $t\bar{t}H$ distribution, therefore a further BDT trained vs the $t\bar{t}$ sample may be effective in the reduction of the still overwhelming background.

Finally, boosted topologies may be included in the study, especially boosted top quarks. We remind that a boosted top quark jet is produced in events in which the decay products of the top quark have a high Lorentz boost and are thus reconstructed in the detector as a single, wide jet. These jets have become more and more important with the increased LHC centre-of-mass energy.

Conclusions

Measuring the associated production of the Higgs boson and a top quark-antiquark pair is one of the major goals for the Run 2 of the LHC. This thesis reports on a first search for $t\bar{t}H$, using pp collision data recorded with the CMS detector in the whole 2015 and 2016, at a centre-of-mass energy of $\sqrt{s} = 13$ TeV, and corresponding to an integrated luminosity $L = 36$ fb^{-1}. In particular, the analysis is performed and optimised on 2015 data, having only $L = 2.63$ fb^{-1}, and then a projection of the results on the full data set is given.

The event selection is adapted to $t\bar{t}H$, with $H \longrightarrow b\bar{b}$ and the all-jets decay of the $t\bar{t}$ pair. Such channel is particularly challenging, given its topology and the experimental conditions, and leads to a final state completely dominated by background processes.

The event selection is based on the study of MC simulated events passing the logic OR of two multijet trigger paths requiring the presence of at least six jets with $p_T > 40$ GeV and 30 GeV, respectively, $H_T = \sum p_T > 450$ GeV and 400 GeV, and *one* and *two* b-jets. Then, based on an extensive trigger study, we required $H_T > 450$ GeV, and at least six jets with $p_T > 40$ GeV. Furthermore, we asked for at least two "medium" b-tagged jets, a successful kinematic fit under the $t\bar{t}$ hypothesis, and a lepton veto to ensure the orthogonality with the semi-leptonic and dileptonic analyses, leading to $S/B \approx 1/2500$.

Then, a multivariate method (Boosted Decision Tree) is employed to separate the $t\bar{t}H$ signal from the simulated QCD background. The BDT selection is based on more than twenty different variables, exploiting several properties of the events, such as jet multiplicity, jet hardness, quark-gluon discriminants, event shape, the kinematic fit for the $t\bar{t}$ hypothesis, centre-of-mass variables and variables referred to the b-tagged jets. Such classification is carried out in two categories (for events with exactly 2 b-tagged jets and with more than 2 b-tagged jets), to allow the combination of the results with a parallel group working on a different multivariate technique.

After a comparison between data and simulations, the QCD modelling resulted not to be very accurate, so that a new background modelling method was needed. Since the presence of b-tagged jets is a strong characteristic of the decay of the signal events, we chose to define a control sample, QCD-enriched, by asking the usual selection, combined with a veto on the "medium" b-tagged jets, those used for the signal selection, and a request on two "loose" b-jets. The closure test and the output of the training of the BDT

provided a further cut on the BDT output variable ($bdt > -0.8$), that completes the selection and leads to $S/B \approx 1/2000$.

Given the still dominating background contribution, and therefore the practical impossibility to highlight any presence of signal, the sensitivity of the BDT machinery is evaluated and the final estimate is presented in terms of the expected signal strength modifier μ for $t\bar{t}H$ production, defined as the ratio of the 95% confidence level upper limit on the $t\bar{t}H$ production cross section, with respect to the SM expectation. The sensitivity of the machinery built so far, then, is expected to correspond to $\mu = 8.2^{+3.4}_{-2.4}$, using a preliminary estimation of the sources of systematic uncertainty.

After evaluating the expected sensitivity on the partial data set corresponding to $L = 2.63$ fb^{-1}, we evaluated the projection to the full data set, under a slightly different data-driven QCD modelling. The projection on the expected signal strength results to be $\mu = 2.1^{+2.1}_{-1.6}$.

My contribution to this analysis started in Summer 2015 when I was a CERN Summer Student at CERN. My major contributions consisted, at first, on the entire extensive and detailed trigger study, and the preliminary study of the variables to be fed to the BDT. Then, thanks to the knowledge I acquired about the trigger paths, I have collaborated to the definition of the QCD control region and to the study of its properties. Finally, I have broaden the study of the BDT sensitivity, by starting to add the first sources of systematic uncertainties. Finally, I have evaluated the projection of the result to the full data set.

The experimental procedures explained ihere have been reported by myself at several CMS Higgs Physics Meeting and at a national CMS Italia Meeting on the topic of $t\bar{t}H$. Lastly, I took part in the IFAE 2016 (Incontri di Fisica delle Alte Energie) workshop with a poster displaying the work performed, and I was awarded the "Best Poster Award".

Bibliography

[1] *LHC the guide*, http://cds.cern.ch/record/1092437/files/CERN-Brochure-2008-001-Eng.pdf.

[2] http://home.web.cern.ch/topics/large-hadron-collider.

[3] L. Evans, P. Bryant, *LHC Machine*, JINST **3** (2008) S08001.

[4] http://cms.web.cern.ch/.

[5] CMS Collaboration, *The CMS experiment at the CERN LHC*, JINST **3** (2008) S08004.

[6] A. Barbaro Galtieri, F. Margaroli, I.Volobouev, *Precision measurements of the top quark mass from the Tevatron in the pre-LHC era*, Rept. Prog. Phys. **75** (2012) 056201.

[7] B. Carithers, P. Grannis, *Discovery of the top quark*, SLAC Beam Line **25** (1995) 4.

[8] CDF Collaboration, *Observation of top quark production in* $\mathrm{p\overline{p}}$ *collisions with the Collider Detector at Fermilab*, Phys. Rev. Lett. **74**, (1995) 2626.

[9] D0 Collaboration, *Observation of the top quark*, Phys. Rev. Lett. **74**, (1995), 2632.

[10] F. Englert, R. Brout, *Broken Symmetry and the Mass of Gauge Vector Mesons*, Phys. Rev. Lett. **13** (1964) 321.

[11] P. W. Higgs, *Broken symmetries, massless particles and gauge fields*, Phys. Rev. Lett. **12** (1964) 132.

[12] P. W. Higgs, *Broken Symmetries and the Masses of Gauge Bosons*, Phys. Rev. Lett. **13** (1964) 508.

[13] G.S. Guralnik, C.R. Hagen, T.W.B. Kibble, *Global Conservation Laws and Massless Particles*, Phys. Rev. Lett. **13** (1964) 585.

[14] ALEPH, CDF, D0, DELPHI, L3, OPAL, SLD Collaborations, the LEP Electroweak Working Group, the Tevatron Electroweak Working Group, and the SLD Electroweak and Heavy Flavour Groups, *CERN PH-EP-2010-095* (2010).

[15] 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.

[16] 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.

[17] ATLAS and CMS Collaborations, *Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at* $\sqrt{s}$ *=7 and 8 TeV*, JHEP **08** (2016) 045.

[18] CMS Collaboration, *Evidence for the direct decay of the 125 GeV Higgs boson to fermions*, Nature Phys. **10** (2014) 557.

[19] CMS Collaboration, *Search for the associated production of the Higgs boson with a top-quark pair*, JHEP **09** (2014) 087. [Erratum: JHEP **10** (2014) 106].

[20] LHC Higgs Cross Section Working Group Collaboration, `arXiv:1101.0593`.

[21] P. Artoisenet, R. Frederix, O. Mattelaer, R. Rietkerk, *Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations*, JHEP **03** (2013) 015.

[22] S. Frixione, E. Laenen, P. Motylinski, B. R. Webber, *Angular correlations of lepton pairs from vector boson and top quark decays in Monte Carlo simulations*, JHEP **04** (2007) 081.

[23] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer and T. Stelzer, *MadGraph 5: Going Beyond*, JHEP **06**, (2011) 128.

[24] S. Frixione, P. Nason and C. Oleari, *Matching NLO QCD computations with parton shower simulations: the POWHEG method*, JHEP **11** (2007) 070.

[25] S. Alioli et al., *A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX*, JHEP **06** (2010) 043.

[26] T. Sjöstrand, S. Mrenna and P. Skands, *PYTHIA 6.4 physics and manual*, JHEP **05** (2006) 026.

[27] T. Sjöstrand, S. Mrenna and P. Skands, *A Brief Introduction to PYTHIA 8.1*, Comput. Phys. Commun. **178** (2008) 852.

[28] R. Atkin, *Review of jet reconstruction algorithms*, Journal of Physics: Conference Series **645** (2015) 012008.

[29] M. Cacciari, G. P. Salam, G. Soyez, *The anti-k_t jet clustering algorithm*, JHEP **04** (2008) 063.

[30] CMS Collaboration, *Particle–Flow Event Reconstruction in CMS and Performance for Jets, Taus, and E_T^{miss}*, CMS Physics Analysis Summary CMS-PAS-PFT-09-001 (2009).

[31] CMS Collaboration, *Identification of b-quark jets with the CMS experiment*, JINST **8** (2013) P04013.

[32] CMS Collaboration, *Performance of b-tagging algorithms in proton collisions at 13 TeV using the 2016 data*, CMS DP -2016/042.

[33] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, H. Voss, *TMVA 4 - Toolkit for Multivariate Data Analysis with ROOT - Users Guide.*

[34] CMS Collaboration, *Performance of quark/gluon discrimination using pp collision data at $\sqrt{s} = 8$ TeV*, CMS-PAS-JME-13-002 (2013).

[35] ATLAS and CMS Collaborations, LHC Higgs Combination Group, *Procedure for the LHC Higgs boson search combination in Summer 2011*, Technical Report ATL-PHYS-PUB 2011-11, CMS NOTE 2011/005 (2011).

[36] `https://twiki.cern.ch/twiki/bin/view/CMS/SWGuideHiggsAnalysisCombinedLimit`

[37] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, *Asymptotic formulae for likelihood-based tests of new physics*, Eur. Phys. J. C **71** (2011) 1554.

List of Tables

List of Figures