Facultad de Ciencias

Search for dark matter production in association with top quark pairs in the dilepton final state at $\sqrt{s} = 13$ TeV

(Búsqueda de materia oscura producida en asociación con un par de quarks top en el estado final dileptónico a $\sqrt{s} = 13$ TeV)

Trabajo de fin de Máster
para acceder al

MÁSTER EN FÍSICA, INSTRUMENTACIÓN Y MEDIO AMBIENTE

Autor : Cédric PRIEËLS

Director : Jónatan PIEDRA GÓMEZ
Co-directora : Alicia CALDERÓN TAZÓN

Junio 2017

Abstract

Dark matter production in proton-proton collisions at a center of mass energy $\sqrt{s} = 13$ TeV is searched for with the CMS experiment at the LHC, using a data sample corresponding to an integrated luminosity of 2.4 fb^{-1}, taken during 2016. This search is performed considering dark matter production in association with a pair of top quarks. It requires the presence of two leptons, jets and a large amount of missing transverse energy.

We haven't managed to exclude any dark matter mediator production model by applying our analysis to the blinded dataset currently available to us, but we do expect to be able to exclude the low dark matter mediator masses once we start studying the complete 2016 dataset.

Key words : Particle Physics, CERN, CMS, dark matter, multi-variate analysis

Resumen

Se busca la producción de materia oscura en colisiones protón-protón a una energía en el centro de masa $\sqrt{s} = 13$ TeV con el experimento CMS del LHC, usando una muestra de datos que corresponde a una luminosidad integrada de 2.4 fb^{-1}, tomada en 2016. Esta búsqueda está hecha considerando producción de materia oscura en asociación con un par de quarks top y requiere la presencia de dos leptones, jets y una gran cantidad de energía transversa faltante.

No hemos podido excluir ningún modelo de producción de materia oscura aplicando nuestro análisis a los datos actualmente disponibles, pero esperamos poder excluir los mediadores de masa débiles cuando consideremos todos los datos de 2016.

Palabras claves : Física de Partículas, CERN, CMS, materia oscura, análisis multivariable

Agradecimientos

Primero, quiero dar las gracias a todo el grupo de Física de Altas Energías del Instituto de Física de Cantabria en general por haberme ofrecido un contrato de trabajo tan rápido y sin conocerme, que me ha permitido desarrollar mis conocimientos mientras estudiaba el Máster. Estos dos últimos años han sido una experiencia que nunca olvidaré y que me servirá seguramente mucho en el futuro.

Quiero agradecer a Jónatan en particular por su ayuda y sus consejos desde mi llegada al IFCA, por siempre tener la puerta abierta y por haber respondido con precisión y paciencia a todas las preguntas que tenía (y que no eran pocas). Gracias a Alicia también por toda su ayuda sobre algunos temas en particular de este trabajo, y a Rocío por su buen humor todos los días. Gracias también a Pablo por los consejos acertados dados antes de nuestras charlas importantes en MET+X, a Pupi por estar siempre accesible por Skype, y a Celso por su ayuda con el papeleo y la burocracia en general (que es incluso a veces peor que en Bélgica).

Muchas gracias también a todos los estudiantes con quien comparto el despacho. Me ha gustado vuestra compañía y me alegro de poder seguir trabajando con vosotros en los próximos años. Gracias en particular a Juan por haberme explicado con paciencia todas las diferentes partes de su análisis, y a Pedro por haber compartido conmigo todas las clases de Máster y todas estas horas comparando nuestros resultados para las diferentes prácticas.

Gracias también a todos mis amigos que me han acompañado a lo largo de mis estudios. Gracias a Nicolas en particular que me ha apoyado al principio de mis estudios, y a Alberto por haber compartido conmigo mi primer año aquí en el IFCA. Por supuesto, muchas gracias a Inés, porque todo este trabajo no hubiera sido posible sin su apoyo diario.

Por supuesto, gracias también a mis padres, a Antoine y Lucas por estar siempre presentes y ser un soporte imprescindible.

Note on my personal work

The almost two years spent working at the IFCA have been a great opportunity for me to learn how to use different programming languages, such as C++ and python, and how to work in a large collaboration with other people. From the physics point of view, I also learnt a lot about particle physics in general and about the methods used to make precision measurements (I first measured the cross section production of the well-known WZ process), to estimate some backgrounds using data-driven methods (I spent quite some time studying how to estimate the non-prompt background which appears in most of the analyses performed) and to now search for new exotic particles, such as the production of dark matter particles.

I have personally produced all the distributions and numbers appearing throughout this work, with the exception of the Figure 5.2. I have produced all of these results by either completely writing the scripts from scratch, or by modifying scripts which had previously been written by people with whom we are collaborating in several physics analyses such as the WW cross-section measurement or the Mono-Higgs, Higgs to WW and Stop searches.

Contents

Chapter 1

Introduction

Particle physics is the field which studies the matter surrounding us, along with the fundamental interactions between the particles. The origin of particle physics is to be found in the fifth century before JC, when Empedocles made the first attempt to divide matter into different categories (which were, according to him, Water, Earth, Fire and Air). A few years later Democritus added his own contribution to this newly born field, by saying that matter is made of indivisible entities that he decided to call atoms. This was a brilliant hypothesis, and over the next centuries, many different philosophers and scientists have joined their efforts and studied the composition of our Universe. In the 19th century, mainly thanks to the emergence of empirical reasoning, it was discovered that atoms are actually not fundamental particles, since it is possible to divide them into smaller entities, later identified as electrons, orbiting around a nucleus composed by of nucleons (protons and neutrons). Subatomic particle physics was starting to grow thanks to the development of quantum mechanics and quantum field theory [1], and thanks to many experiments carried out during the 20th century. For example, detailed analysis of the data coming from the Stanford Linear Accelerator Center in the 1960s showed the presence of three scattering centers within the nucleons, which then can not be considered as elementary particles as previously thought [2]. Other particles were as well experimentally discovered during this period, such as the J/Ψ particle which lead to the discovery of the charm quark in 1974. All of the discoveries made during the 20th century raised many questions, and the necessity of developing a complete model explaining all of these observations became obvious.

Nowadays, the model most accepted and used to describe the particles and the fundamental interactions between them is the so-called Standard Model, as seen in Figure 1.1. This model is simple in concept, but has been extremely good describing the phenomena observed so far, and made a lot of predictions that have now been proven to be true, such as the postulate of the Higgs mechanism [3] followed by the discovery of the Higgs boson in 2012 [4, 5]. We currently know that there are 12 elementary particles in nature, divided into two categories (leptons and quarks) along with their 12 corresponding antiparticles. On the other side, the three main fundamental forces in particle physics can be described with four gauge bosons[1] (the gluon, the photon, and the $W^{\pm}$ and Z^0 bosons). The Higgs boson is the last piece of this model, needed to explain the origin of the mass of the particles.

However, as accurate as it is, this model is known to have several shortcomings which require further investigation. Over the next few years, eventual exotic particles which do not fit in the current model and which could be the sign of new physics were then extensively searched for. For example, in 1970, the first serious dark matter hypothesis was introduced because of gravitational anomalies observed by several astrophysicists, as a way to explain the apparent missing mass in the Universe. Indeed, the visible mass seemed to be way too low to explain several astrophysical processes, such as the rotation curves of the galaxies, as explained in Section 2.1. As far as we currently know from

[1] A gauge boson is a particle carrying any of the fundamental forces. Elementary particles of the Standard Model can interact with each other by exchanging this kind of boson, as described by the so-called gauge theory.

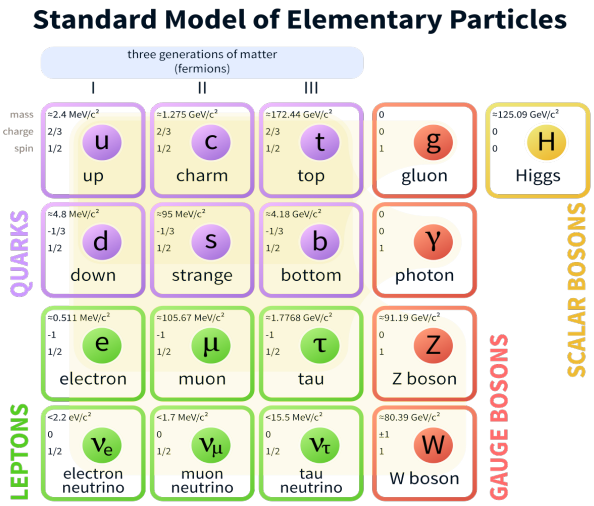


Figure 1.1: Summary of the Standard Model of elementary particles [6].

cosmological measurements, ordinary baryonic matter only constitutes around 5% of the Universe, while dark matter represents around 27% of the mass of the Universe (the rest is being considered as dark energy) [7].

This document begins with a short theoretical introduction about the dark matter production process in general, and then in particular in the channel we are interested in (production of dark matter in association with a pair of top quarks in the dilepton final state), followed by some general explanations and descriptions of the detector used to record the data. The different datasets, samples and backgrounds of this analysis will then be presented, checked and studied in detail and the results obtained by performing a multi-variate analysis and using a neural network will be shown. Finally, the upper limits for the production cross section of different dark matter mediator masses will be plotted, to check if we observe or not any significant deviation between the observed and expected limits for the Standard Model, and to see if we expect to get some sensitivity with the current luminosity to the different dark matter production models.

Chapter 2

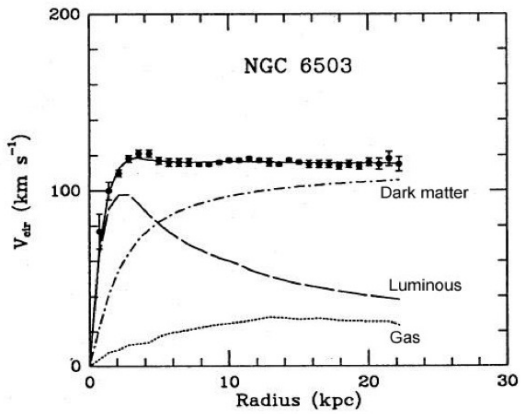
Theoretical introduction

2.1 At the origins of dark matter

The origins of the dark matter hypothesis can be traced back to the 17th and 18th centuries, shortly after Newton's works on gravitation [8]. The dark matter considered back then was however quite different than the one considered nowadays, since it was thought to be ordinary matter which simply does not emit any kind of electromagnetic radiation (and is therefore invisible) but which has a strong impact in the gravitational point of view. Dark matter was for example considered during the 20th century to be found in massive astronomical objects able to absorb the light of other objects situated right behind them, such as black holes.

At the beginning of the 20th century, the first experimental evidences for the existence of dark matter were showed. In 1933, Fritz Zwicky determined the mass of the Coma Cluster using the virial theorem, which states that if we consider thousands of stars interacting gravitationally with each other, the average kinetic energy of any given star can be directly related to the average gravitational potential energy of this star. This means that the typical velocity of a star within the cluster can be related directly to the mass of the whole cluster. Thanks to this theorem, Zwicky was able to determine that the mass of the Coma Cluster should be around 400 to 500 times larger than the mass previously estimated by Edwin Hubble, who simply considered the number of visible galaxies in this cluster. Zwicky concluded that there should exist some kind of invisible matter that we can not detect but responsible for most of the mass of this cluster [9].

Zwicky's results were controversial since they were based on statistical calculations relying on different hypotheses not always justified, such as the fact that the clusters must be gravitationally bound and they were actually proven to be overestimated later on [10]. However, these results were soon followed by different observations leading to similar conclusions, the most famous one being the study of the observed and expected rotation curves of the galaxies in the 1970s [11]. It was expected that the mass of the galaxies would be concentrated where more stars are visible, around the galactic center, and this would imply that outside of this high density region the velocities of the stars should decrease as the inverse of the square root of the distance. However, we now know that in reality this velocity has a flat behavior when this distance increases, as seen in Figure 2.1. One of the easiest way to explain this kind of behavior is to introduce the concept of dark matter, which must not only be non-luminous as previously thought, but must also not absorb usual light either since we would have been able to detect this kind of matter already. Therefore, dark matter can not just be dark ordinary baryonic matter, it should be something different.



K.G. Begeman, A.H. Broels, R.H. Sanders. 1991. Mon.Not.RAS 249, 523.

Figure 2.1: Expected and observed rotation curves of the galaxy NGC 6503 [11]. The black dots correspond to the data, and the line labeled as *Luminous* corresponds to the expected rotation curve without dark matter.

Nowadays, many more experiments have contributed to this field and strengthen the dark matter hypothesis even though this is still not the only valid hypothesis accepted and under investigation (for example, another possibility is the MOdified Newtonian Dynamics hypothesis [12], modifying slightly newtonian's laws of gravity at large scales in order to explain some of the observations made by astrophysicists over the years).

The question remains to know the exact nature of dark matter particles. Astrophysicists have been able to determine the dark matter distribution within the galaxies from the latest sky surveys [13] and gravitational lensing studies [14], and we also know that not only dark matter does not interact at all with light, but it also does not interact with ordinary matter or itself. Some studies also showed that dark matter should be made of slow and cold particles (therefore, the neutrinos, which could have been considered as good candidates since they only interact weakly with ordinary matter can be rejected because they have such a low mass that they travel almost at the speed of light). Since we do not know any baryonic particle fulfilling all these conditions, we can postulate the existence of new kinds of elementary particles to fill this role: the so-called MACHOs, the MAssive Compact Halo Object, or the WIMPs, the Weakly Interactive Massive Particles [15]. This latest group includes particles which are stable and are assumed to have been produced in pairs (particles, antiparticles) shortly after the Big Bang. This is typically the kind of particles we are hoping to find at the LHC.

2.2 Dark matter production at the LHC

Not only astrophysicists have their word to say when it comes to searching for dark matter, since particle physicists all around the globe hope that the LHC will help us find a way to determine if the dark matter hypothesis can be excluded or not [16]. Most physicists assume nowadays that dark matter was produced shortly after the Big Bang, along with ordinary matter, in a hot and dense Universe. If this assumption is correct, then the LHC is the perfect place to study this kind of particles, since its objective is to go back in time and study the Universe as it was just a fraction of a second after the Big Bang. For this reason, the LHC is a perfect tool to study the properties of dark matter, at least if we assume that it is actually made of particles and that it can actually be produced by LHC's proton-proton collisions, with a measurable cross section.

The main idea is to search for missing transverse energy, corresponding to the imbalance of vector momentum in the plane perpendicular to the beam direction, in the data coming from the accelerator, since the eventual dark matter particles produced are not expected to interact at all with the detector. This seems like a proper way to indirectly detect dark matter particles, but things may get complicated in practice because other Standard Model processes are also responsible for the apparition of missing transverse energy. For example, the detector resolution, production of neutrinos (which almost do not interact with ordinary matter) or particles escaping through some cracks of the detector are interpreted as missing energy as well and it is actually almost impossible to distinguish neutrinos from eventual dark matter particles. One of the ways we have to determine the kind of particle we are dealing with (besides the different kinematics or the angular distributions studies) is to estimate the fraction of proton-proton collisions that will produce a certain amount of missing transverse momentum, directly from the Standard Model and from theoretical Monte Carlo simulations. The main idea of the analysis consists in comparing these simulations for all the different expected backgrounds and the data obtained by the detector for different variables, in order to see if we observe an excess of data with respect to the simulations, which could be the sign of some new physics.

Different channels are expected to be good candidates for the eventual discovery of dark matter particles [17]. Usually, searches at the LHC are focused on considering production of a pair of WIMPs in association with initial or final state radiation (photons, gluons, $W^{\pm}$ or Z^0 bosons for example) in order for the detector to be able to trigger the measurement, since it is not able to know whether or not an event consisting of only two WIMPs particles happened. The basic idea for a typical dark matter search at the LHC consists in looking for a high transverse momentum (p_T) particle production, in association with missing transverse momentum (MET, $\not{E}_T$ or E_T^{miss}).

2.3 The $t\bar{t}$ + DM dilepton channel

Many different mechanisms of dark matter production are expected by different theories. This work will be focused on dark matter production in association with a top and an anti-top quarks, as showed in Figure 2.2. This process will be referred to as the $t\bar{t}$ + DM process throughout this work.

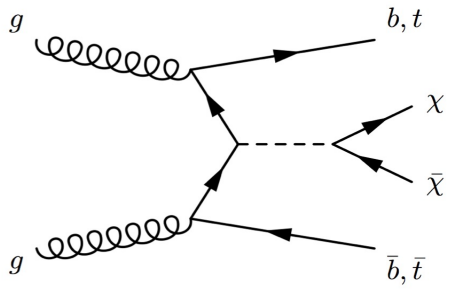


Figure 2.2: The channel studied in detail throughout this work: a pair of top quarks $t\bar{t}$ is sometimes expected to interact, producing a spin-0 mediator (dashed line in the diagram) and a pair of dark matter particle χ and antiparticle $\bar{\chi}$ [18].

A natural question which might arise is to know why we consider this channel with a pair of top quarks and missing transverse energy in the final state. The first reason is that the top quark interacts strongly, meaning that we can expect to produce a lot of $t\bar{t}$ at the LHC and even if the dark matter production cross section is really low, we still would expect to see at least a few interesting events. A second reason is that the top can be spotted quite easily within the huge amount of data produced at the LHC, since its signature can be well isolated with respect to the other backgrounds and processes. It is quite a natural process to first look where it is easier to find something. Another more fundamental reason is coming directly from the theory of the Standard Model. Some of the hypotheses about the nature of

dark matter assume that it is a spin-0 particle, whose mediator should couple to ordinary particles in a similar way than the Higgs boson does. Thanks to last years data from the LHC, we actually know that the Higgs has a stronger coupling constant with the more massive particles, and the top quark has been found to be the most massive fermion of the Standard Model (way more massive than any other quark, actually) [19].

As previously stated, we currently do not have any way to directly detect any eventual dark matter particle since it is not supposed to interact at all with ordinary matter or itself. We then have to rely on the top quarks detection and on Monte Carlo simulations to be able to make any measurement in this channel. The top is actually not stable, has a lifetime of around 10^{-24} s, which is way too short for us to be able to spot it directly [20] and decays almost instantly into a $W^{\pm}$ boson and a bottom quark through weak interaction. This is not quite over yet, since both these particles produced are not stable either, in the sense that the bottom quark immediately goes through a hadronization process and produces a jet that we can detect, while the W boson can decay into two different channels: the leptonic one ($W^{\pm} \to \nu l$) and the hadronic one ($W^{\pm} \to q\bar{q}$). Since we have two W produced, it is possible to study three different channels: the dileptonic, the semileptonic and the hadronic channel. The channel studied throughout this work is the dileptonic one, because even though this is the channel having the smallest cross section for dark matter production, this is also the channel having the smallest number of significant background processes.

In summary, the channel studied results from the interaction between the protons in the colliding beams, which can sometimes produce a pair of top quarks and a pair of dark matter particle and antiparticle, through the apparition of a spin-0 mediator. We are then searching for dark matter production in the dilepton final state.

2.4 Previous results

A similar analysis has already been published by the CMS collaboration in August 2016 with data taken in 2015 at a center of mass energy of 13 TeV and corresponding to an integrated luminosity of 2.2 fb^{-1} [21]. As we can see in Figure 2.3, they did not observe any significant deviation between the observed and expected limits (at least for a dark matter candidate having a 1 GeV mass, and for both scalar and pseudoscalar spin-0 mediators) with this limited luminosity. In this work, we expect to improve these results by adding more sophisticated techniques to the analysis (such as an improved top reconstruction and a neural network) and by exploiting more data.

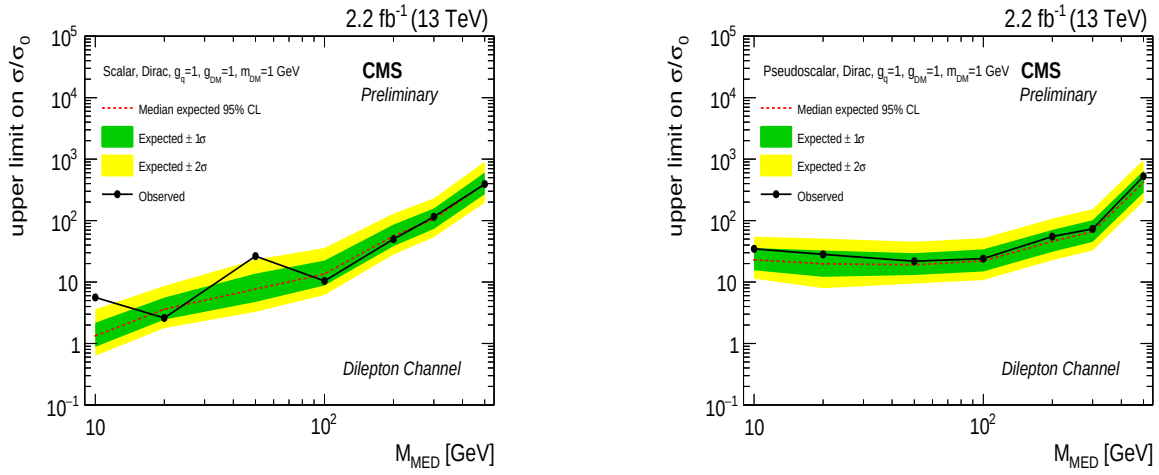


Figure 2.3: Expected and observed limits on the production cross section for an integrated luminosity of 2.2fb^{-1}, for both the scalar (left) and pseudoscalar (right) spin-0 mediators [21].

Chapter 3

The experimental device

3.1 The LHC collider

The Large Hadron Collider is a circular underground proton-proton collider, situated at CERN, the European Organization for Nuclear Research, next to the city of Geneva. With its 27 kilometers of circumference, it is currently the most powerful accelerator in the world. A schematic representation of the accelerator can be found in Figure 3.1. This accelerator has been built by a collaboration of 22 countries in order to study and reproduce the Universe at its origin and the conditions right after the Big Bang, to make precision measurements to check the validity of the Standard Model and to search for exotic new physics [22].

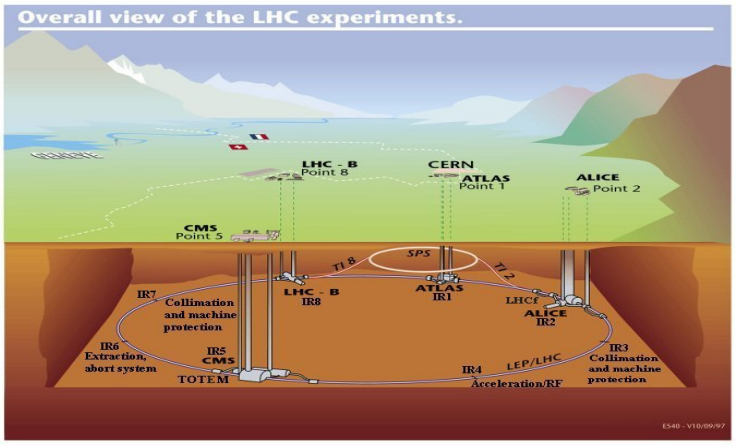


Figure 3.1: Schematic representation of the LHC and its different detectors [23].

Two beams of protons can circulate in opposite directions at velocities close to the speed of light and collide in four different points of the accelerator, corresponding to the four detectors that have been designed to study the collisions: the Compact Muon Solenoid (CMS), A Toroidal LHC ApparatuS (ATLAS), A Large Ion Collider Experiment (ALICE) and LHCb. The LHC is currently able to accelerate beams made out of 2808 bunches of around 10^{11} protons up to a center of mass energy of 13 TeV, with an instantaneous luminosity of around 10^{34} cm^{-2} s^{-1} (these bunches are separated by 25 ns, leading to an impressive collision rate of 40 MHz). It is trivial to understand why having a higher energy is interesting from the physics point of view, since with a higher center of mass energy we would be able to create particles with higher masses. The luminosity is a crucial parameter as well, since a higher luminosity results in a higher number of collisions and an increase in the sensitivity. The integrated luminosity collected in the year 2016 corresponds to 35.9 fb^{-1}, which corresponds to around 30% of the total luminosity expected during the first phase of operation of the accelerator, until the years 2018-2019 [24].

3.2 The CMS detector

The data studied in this work has been taken by the CMS detector of the LHC, the Compact Muon Solenoid. This detector, along with ATLAS, has been designed to be a polyvalent detector, able to make measurements in most of the major different fields of particle physics (from precision measurements of Standard Model properties, to the Higgs hunting and to the search of exotic processes) [25]. A schematic representation of the detector and its different layers can be found in Figure 3.2.

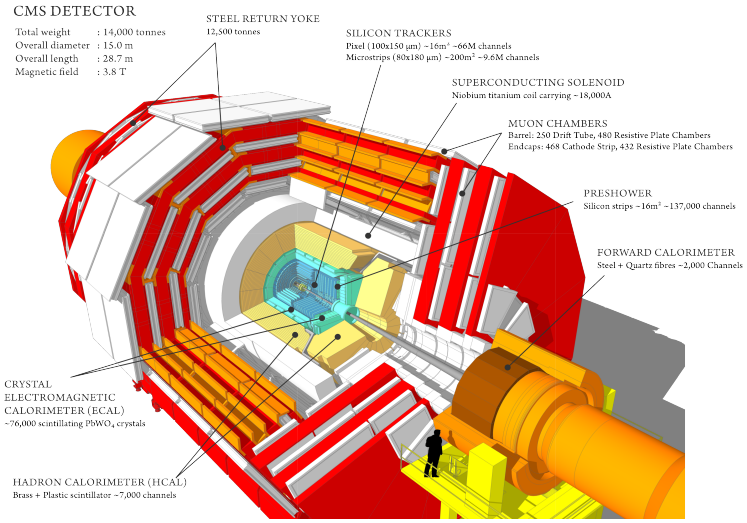


Figure 3.2: Schematic representation of the CMS detector [26].

What is the exact meaning of the name CMS? First, this is a *Compact* detector, which is really heavy (around 14000 tons) but all this weight is compacted into a relatively small volume, since the detector "only" has 15 meters of diameter and 28.7 meters of length. The *Muon* part of the name comes from the fact that this is a detector specially dedicated to the detection and measurement of different properties of the muons, throughout a large range of energies. Finally, the *Solenoid* part of the CMS name comes from the fact that its central piece is a huge solenoid able to produce a magnetic field of 3.8 T (equivalent to around 100 000 times the Earth's magnetic field) parallel to the beam direction, to curve and study different properties of the charged particles produced.

The detector is made out of different layers, each able to detect and measure different properties of the particles. The tracker is the center part of this detector: composed by pixels and microstrips of silicon, small electrical currents are produced by charged particles passing through this layer, giving a way to detect the exact position of the interaction and a way to reconstruct the exact track followed by the charged particles produced. It is really light, so that it does not affect the particles before we get a chance to measure their properties, and is extremely resistant to the radiation. The next layer is the electromagnetic calorimeter, made out of transparent lead-tungstate crystals to detect light produced by the electromagnetic showers produced when an electron or a photon comes crashing into this layer. Next comes the hadron calorimeter, able to measure the energy of most of the hadrons. Finally, the last layer, right after the solenoid, is made out of the muon detectors. This is the most external layer since muons interacts very little with ordinary matter except by ionization and scattering, and they are able to go through all the previous layers without getting affected. Of course, to avoid the presence of any instrumental missing energy coming from a particle escaping the detector without being detected, all the different layers of the detector must be completely hermetic.

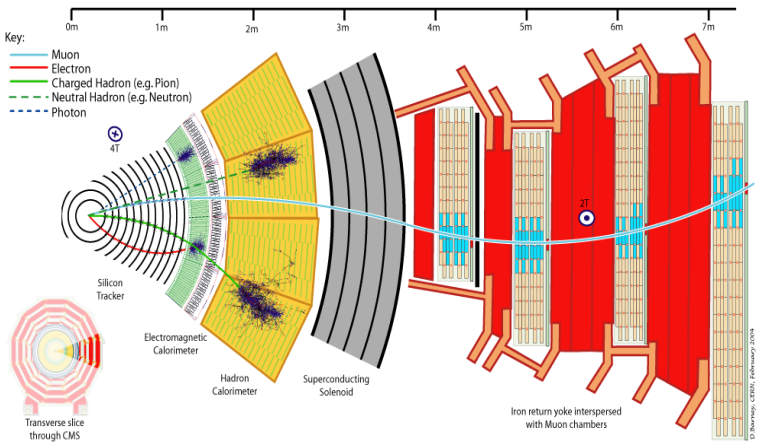


Figure 3.3: Layers of the CMS detector, and expected track for different kind of particles [27].

The reconstruction and identification of the particles produced in a single event is a complicated task. This can be done however thanks to a powerful algorithm called the *Particle Flow* [28]. With this algorithm and because of the acceptance of the detector, jets and electrons can be reconstructed up to a pseudorapidity[1] $|\eta| < 3$ while muons can be reconstructed up to $|\eta| < 2.4$.

Since the LHC has an extremely high collision rate, a system of triggers is being used, in order for us to be able to keep potentially interesting events and reject the most common ones almost immediately, without having to store all the events, since it would be computationally impossible to keep all the events for the analysis. This system, known as the trigger system, is divided in two different categories: the L1 is the first trigger level, and is able to take extremely fast decisions based on a few basic variables only and using a dedicated hardware, to take the decision to keep or reject the event while the HLT (High Level Trigger) is more complex and slow, since it gathers information from different parts of the detector before reaching to a decision. At the end, a collision rate of around 1 kHz (out of the 40 MHz produced by the LHC) is being selected with this system to be stored and analyzed.

The most important magnitudes at CMS are those corresponding to the transverse plane. There is a simple reason for this: since the LHC is a proton-proton collider, the initial center of mass energy along the beam pipes (defined as the z-axis) is not known because the energy is distributed between the different quarks in unknown proportions. However, we do know that the initial transverse energy or initial transverse momentum is equal to exactly 0, and this is why we use this kind of variables.

[1]The pseudorapidity is a variable defined as $\eta = -ln\left(tan(\frac{\theta}{2})\right)$. This quantity is interesting because it is invariant under Lorentz transformations corresponding to boost along the z-axis.

Chapter 4

Objects, datasets, triggers and samples

In this chapter, the objects, datasets and triggers used for the analysis will be detailed, along with the expected signal and the different backgrounds coming in the way. All the major backgrounds of the analysis will then be detailed, while the different methods developed to estimate their respective proportion will be explained.

4.1 Objects and datasets

On one hand, the definitions of the different objects will be detailed in Section 5.1 but basically, we apply some cuts on various variables (on the p_T and on the isolation of the leptons between each other and between the jets produced, for example) to select only interesting leptons and achieve a particular signal efficiency. Different working points are then made available for us, to select the kind of lepton we are interested in for the analysis we want to perform, going from the so-called loose working point, for analyses with a clean final state to the tight working point, for analyses having a lot of backgrounds. A loose lepton is defined with only some basic requirements in order to keep a high efficiency (of the order of 90% for electrons having a p_T larger than 20 GeV) while the medium and the tight leptons are defined by applying cuts much more restrictive, resulting in better but less efficient objects (the efficiency is estimated to be of the order of respectively 80 and 70% for the same kind of electrons). It is important to note that the cuts which go inside and define these working points depend on several factors, including the running conditions, meaning that they have to be updated and tuned again every time these conditions change.

As described previously, the datasets used have been recorded during the Run II at the LHC, at a center of mass energy of $\sqrt{s} = 13$ TeV and correspond to an integrated luminosity of 2.4 fb^{-1} for the signal region, and 35.9 fb^{-1} for the different control regions studied (the uncertainty on the value of these luminosities has been estimated by the luminosity monitoring group in [29] to be equal to 2.5%). The events were actually selected randomly from the 35.9 fb^{-1} complete dataset to follow the blinding policy of the MET+X group at CERN[1]. The acquisition of the data has been divided into small entities called runs and grouped into 8 different periods, going from 2016A to 2016H (but the run period 2016A has been taken with the magnet of CMS turned off and does not have any interest for this analysis, since we loose a lot of information about the different leptons produced), as described in Table 4.1. Each period is then made out of different files corresponding to the five different interesting processes available for any dilepton analysis (SingleMuon, SingleElectron, DoubleMuon, DoubleEG and MuonEG), usually referred to as datasets. All of these datasets have been updated at the beginning of 2017 and correspond to the ReReco of the data (which consists mainly in updated calibrations and minor corrections of the data) [31].

[1]The blinding policy is a methodology that forces the analyzer to define all his procedures and signal region definitions using only a small subset of the data. This kind of blinding policy is applied to most of the searches for new physics, and it is quite easy to understand why such a policy is applied. When discovering or searching for a new process, we need the observation to be statistically significant by collecting more data, and we want to make sure to avoid any conscious or even subconscious bias when one tries to optimize hiss analysis based on what has already been seen [30].

Era	From run	To run	Luminosity (fb^{-1})
Run2016B	272007	275376	5.748
Run2016C	275657	276283	2.573
Run2016D	276315	276811	4.248
Run2016E	276831	277420	4.009
Run2016F	277772	278808	3.102
Run2016G	278820	280385	7.540
Run2016H	280919	284044	8.606
Total	272007	284044	35.827

Table 4.1: Different datasets taken so far during the Run II data period and luminosities associated, calculated with Brilcalc [31, 32].

4.2 Triggers

Since the LHC is producing an enormous amount of collisions and since it is currently not computationally possible to record all the events, a trigger system has been put in place, as described in Section 3.2. The single and dilepton triggers used in this analysis are the following:

- **SingleMuon** : HLT_IsoTkMu22_v*
  and HLT_IsoMu22_v*
- **SingleElectron** : HLT_Ele27_eta2p1_WPLoose_Gsf_v*
  and HLT_Ele45_WPLoose_Gsf_v*
- **DoubleMuon** : HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_v*
  and HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL_DZ_v*
- **DoubleEG** : HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ_v
- **MuonEG** : HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL_v*
  and HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL_v*

It is important to note that the single and dilepton triggers are combined together in most of the analyses in order to increase their efficiency to select signal events, and that the different numbers appearing in the names of the triggers correspond to the minimum p_T of the lepton required to pass the corresponding trigger.

4.3 Monte Carlo samples

As previously explained, the expected signal and most of the major background processes are produced using Monte Carlo simulations, and have been simulated from different theoretical models [33]. All the MC samples and possible backgrounds considered for this analysis along with their respective cross sections and the total number of events generated are represented in Table 4.2. These samples have been produced using different MC generators and have different levels of accuracy. For example, the $t\bar{t}$ process, our most important background, has been generated using POWHEG V2 [34] and the single top background that we will mention in Section 4.5 has been generated using POWHEG V1. Most of the backgrounds have been generated at the next to leading order (NLO) accuracy level and a few samples have even been generated at the next to next to leading order (NNLO). Events are then hadronized and showered using PYTHIA8 [35], and then interfaced to GEANT4 [36], which contains a detailed model of CMS and the interactions of the particles with its subdetectors.

Process	Accuracy	Cross section [pb]	Number of generated events
Z $\to$ ll (low m_{ll})	NLO	18610	30899063
Z $\to$ ll (high m_{ll})	NLO	6025.2	28751199
W$\gamma \to$ l ν γ	NNLO	586	4316841
Z$\gamma \to$ ll γ	NLO	131.3	14861
$t\bar{t} \to$ ll $\nu\nu$	NLO	87.31	4995600
Single top	NNLO	35.6	1000000
Single antitop	NNLO	35.6	999400
WZ $\to$ lll ν	NLO	4.42965	2000000
ZZ $\to$ llll	NLO	1.212	6669188
ZZ $\to$ ll $\nu\nu$	NLO	0.564	8785050
ZZ $\to$ ll $q\bar{q}$	NLO	3.22	15301695
WZZ	NLO	0.05565	249800
$t\bar{t}$W $\to$ l ν	NLO	0.2043	252673
$t\bar{t}$W $\to q\bar{q}$	NLO	0.4062	833298
$t\bar{t}$Z $\to$ ll $\nu\nu$	NLO	0.2529	398600
$t\bar{t}$Z $\to q\bar{q}$	NLO	0.5297	749400
WW $\to$ ll $\nu\nu$	NNLO	12.178	1979988
ggWW $\to$ ll $\nu\nu$		0.5905	500000
ggH $\to$ WW $\to$ ll $\nu\nu$		0.9913	100000
H $\to$ WW		0.5686	994337
$t\bar{t}$H $\to b\bar{b}$	NNLO	0.212	1501096
W $\to$ l ν	NNLO	61526.7	24156124

Table 4.2: Monte Carlo samples used for this analysis along with their respective cross sections and number of events generated.

4.4 Dark matter signal

Since we do not know the mass and the type of the mediator of the interaction we are interested in, or the mass of the eventual dark matter particles produced, many different signal samples have been produced for different mass points, for us to compare all these possible signals to the data. These samples can be separated into two different categories (when the spin-0 mediator is considered to have either scalar or pseudoscalar couplings) and are represented in Tables 4.3 and 4.4, where m_χ represents the mass of the dark matter particle and m_S represents the mass of the mediator.

m_χ [GeV]	m_S [GeV]						
1	10	20	50	100	200	300	500
10	10	15	50	100			
50	10		50	95	200	300	

Table 4.3: Different signal samples available for the scalar mediator.

m_χ [GeV]	m_P [GeV]							
1	10	20	50	100	200	300	500	1000
10	10	15	50	100				
50	10		50	95	200	300		

Table 4.4: Different signal samples available for the pseudoscalar mediator.

To produce all these samples, a serie of assumptions has been made, following the recommendations of the LHC Dark Matter Forum [37]. First of all, the couplings between the mediator and the usual Standard Model particles have been considered equal to one. Then, the dark matter particle is assumed to be a Dirac fermion (meaning that it is different than its corresponding antiparticle, and that it has a semi-integer spin) and finally, for the scalar model, no mixing with the Higgs boson is assumed. The different samples produced for the different mass points considered have of course different cross sections, as seen on Table 4.5, where only the samples having amass of 1 GeV are represented, since these are the only samples used for our analysis so far. All these signal samples have been produced by Monte Carlo simulations at the leading order level, using Madgraph.

m_χ [GeV]	m_S [GeV]	Cross section [pb]	Number of generated events
Scalar mediator			
1	10	19.590	240379
1	20	10.480	251225
1	50	2.941	253308
1	100	0.672	257761
1	200	0.093	254295
1	500	0.005	250524
Pseudoscalar mediator			
1	10	0.441	252054
1	20	0.399	249253
1	50	0.303	255516
1	100	0.191	249971
1	200	0.084	240536
1	500	0.005	241952

Table 4.5: Signal samples studied in this work at several grid points, along with their corresponding production cross sections from Madgraph (at leading order) for both scalar and pseudoscalar mediator models, and number of generated events.

4.5 Major Standard Model backgrounds

Many backgrounds appear in this analysis and it is crucial to study them in detail if we want to be able to extract any information from the data, in order to be able to detect any eventual excess between this data and the different backgrounds simulated. We expect however to be able to reduce strongly the backgrounds by applying cuts in different variables, but some of them, characterized as irreducible backgrounds, present similar kinematic distributions as our signal does, making them therefore impossible to remove completely. The major backgrounds of this analysis are detailed in the following bullets, while the actual methods or variables available to reduce them will be studied in Section 5.3.

- $t\bar{t}$ **production**. This background, which can be produced through different channels as represented in Figure 4.1, is the one presenting the most similar signature as the one expected for the $t\bar{t}$ + DM process we are looking for, since both are characterized by two top quarks in the final state and therefore, the biggest challenge of this analysis consists in finding a way to separate the $t\bar{t}$ production from our signal. This background is characterized as irreducible since its main

kinematic distributions are really similar to the ones expected for the signal, especially for the low mediator mass points, as we will see.

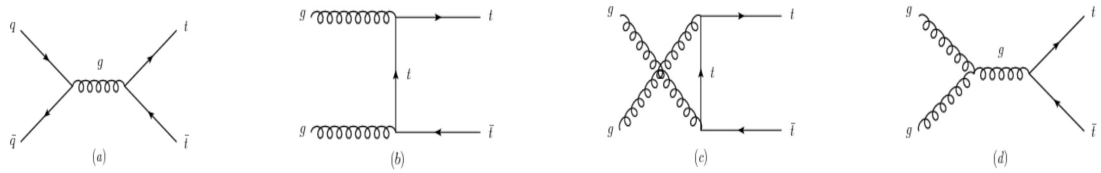


Figure 4.1: Different leading order Feynman diagrams for the production of $t\bar{t}$ [39].

- **tW$^{\pm}$**. This process is the second most important background in our analysis at the final selection level, because its cross section is much higher than the one expected for the signal (cf. Tables 4.2 and 4.5). It is an important background because it features a single top in the final state, and because one out of every three W bosons produced decay leptonically into a lepton and a neutrino, which is responsible for the presence of some missing transverse energy. The main way to reduce this background is to ask for the events to present at least two jets, coming from the production of two quarks: since our signal has two quarks in the final state, it should then pass this requirement, while this background only features one jet and should be a bit reduced with this cut.
- **ttW$^{\pm}$ and ttZ^{0}**. Having a much lower cross section, we expect this kind of background to have a lower impact on the final results than the previous ones, even though it has a pair of top quarks in the final state, just as our signal does. It can be divided into four parts, corresponding to the main different decay modes of the W and the Z bosons, as shown in Figure 4.2. However, two of these four decay modes can be strongly reduced by introducing the m_{T2}^{ll} variable[2] in our analysis. As we can see on the right plot in this figure, when applying the $m_{T2}^{ll} > 80$ GeV cut of our analysis, only two contributions are important: the ttW when the W decays to one lepton and one neutrino (49%), and the ttZ background when the Z decays to two neutrinos (34 %). The other two decay modes can at first order be neglected in this analysis, since they seem to be reducible backgrounds.

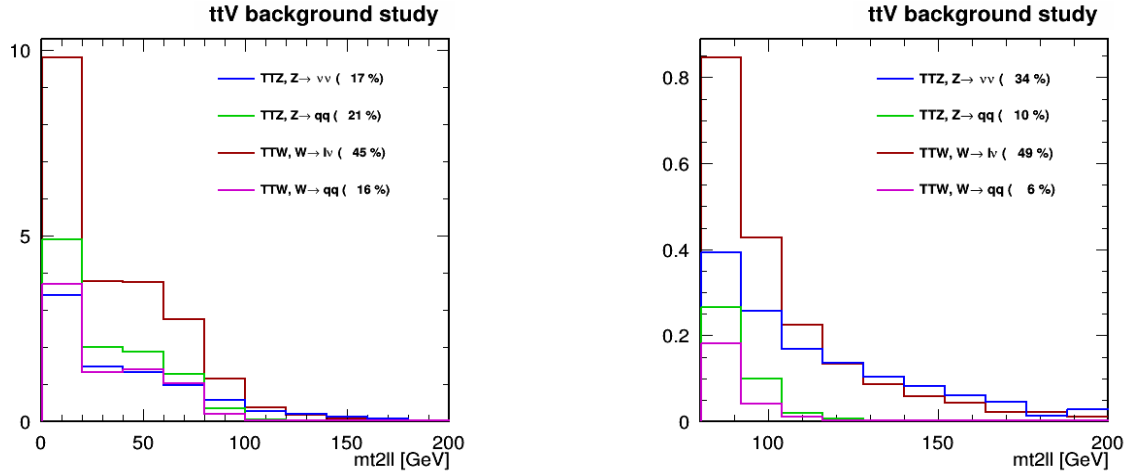


Figure 4.2: Proportion of each decay mode of the ttV background, with respect to the m_{T2}^{ll} variable (in GeV), without applying any m_{T2}^{ll} cut (left) and by applying the $m_{T2}^{ll} > 80$ GeV cut of our analysis (right).

[2]While the definition and the use of this variable will actually be detailed in the section 5.3, it is already important to know that this variable is really important, because it presents a strong discriminating power between the signal and the backgrounds in our analysis.

- **Z^0 + jets**. The Drell-Yan process is the production of a Z boson in association with jets, and it is an important background of this analysis mainly because of its huge cross section. This background can be strongly reduced however by adding a Z-veto to our analysis (we reject all the events featuring either two electrons or two muons that have a reconstructed mass m_{ll} closer than 15 GeV to the mass of the Z, equal to $(91,1876 \pm 0,0021)$ GeV [40]). This cut, along with the introduction of the variable m_{T2}^{ll}, is quite powerful and removes a lot this background but, as previously stated, its huge cross section makes it impossible to get rid of this process completely. This background will be studied in more detail in Appendix A.1, since it is one of the two backgrounds which has been estimated using a data-driven method.

- **Backgrounds with non-prompt leptons**. The calculations needed to study this background will be studied in Appendix A.2, since it is different in the sense that it can not actually be properly calculated using the usual Monte Carlo methods and has to be estimated using data-driven methods. Basically, it appears when a jet is misidentified as a lepton, and is a relatively important background of this analysis. The best way to reduce it strongly is to apply a p_T requirement on the leptons of the analysis, because most of the non prompt leptons have a low p_T.

- **Other backgrounds**. Other backgrounds appear in this analysis, but with the cuts we decided to apply, they almost do not affect the final results. We can cite for example the $W^{\pm}W^{\mp}$ process, which can produce one or several leptons, missing transverse energy and some jets, or the VZ^0 where the V can be either a W or a Z boson. We also looked at different processes having a Higgs in the final state but none of them seem to be able to survive to our selection.

Chapter 5

Event reconstruction and selection

The techniques used to reconstruct the physics objects from the data collected by CMS will be detailed in this chapter, along with the method we developed to reconstruct and estimate the p_T of the mediator of the dark matter production. Then, the variables which have a strong discriminating power for our analysis will be presented and our complete selection will be detailed. Finally, some control plots will be shown to check the validity of the simulation of different backgrounds, and a few words will be said about the systematics of the analysis.

5.1 Event reconstruction

The objects we observe in the final state are reconstructed using different methods and algorithms contained in the CMSSW software package [21]. The so-called Particule Flow algorithm introduced in Section 3.2 allows us to reconstruct all the particles of the event, by combining the information coming from different parts of the detector (such as the tracks left by the charged particles, the energy deposited in the calorimeters and the impacts in the muon chambers). The primary vertex of the interaction is also reconstructed at this point, by calculating the sum of the square value of the p_T of all the different vertices, the primary vertex being defined as the vertex with the largest value obtained.

Leptons

Only the leptons having a p_T higher than 10 GeV and η values smaller than 2.5 (2.4) for electrons (muons) are used for this analysis. Following the recommendations of a CMS object group, and as explained in Section 4.1, both the electrons and muons are then selected by applying a tight identification requirement, based on the information coming from the tracker and from either the electromagnetic calorimeter (for electrons) or the muons chambers (for muons). Finally, a set of loose criteria is also defined to reject the events having more than two leptons, and the leptons are required to be isolated from hadronic activity to be used in the analysis, to remove some of the contribution of the fake leptons, usually produced inside jets. This isolation is defined as the sum of the p_T of the Particle Flow candidates within a cone size of $\Delta R = \sqrt{(\Delta\Phi)^2 + (\Delta\eta)^2} = 0.4$.

Jets

Jets are reconstructed from the Particle Flow candidates (excepting the charged candidates not coming directly from the primary vertex) with the anti-kT algorithm [41] within a cone ΔR equal to 0.4 and, to select the jet candidates, jet-area based corrections are applied in order to take into account the pile-up[1]. The four momentum of the jets is also corrected by applying energy scale calibrations calculated from the data [42]. This analysis only considers jets having a p_T larger

[1]Since the LHC is colliding bunches of protons, we observe many different events every time the beams cross. These multiple interactions happening every time are referred to as the *pile-up*, and complicate the reconstruction of the event.

than 30 GeV, a η value smaller than 4 and passing a loose set of identification and isolation criteria (jets are required to be separated by more than a distance $\Delta R = 0.3$ to any well-identified lepton).

The CSVv2 algorithm [43] is used as b-tagger, to select jets coming from the hadronization of a bottom quark. The medium working point of the tagger is used for the analysis, and leads to reconstruction efficiencies of the order of 69% to tag b-jets and to a mistagging rate of the order of 35 % for the c-jets and 1% for the light flavour jets [21].

Missing transverse energy

The missing transverse energy corresponds to the magnitude of the negative vectorial sum of the transverse momenta of all the particle candidates of the event. Type-I corrections (propagation of the jet energy corrections [44]) have been applied to remove events with large but artificial MET, according to the JET-MET POG recommendations [45].

5.2 Top reconstruction

The top reconstruction is a method used to try to assign a p_T value to the tops appearing in any $t\bar{t}$ like event. Getting this kind of information is really important for us because we want to be able to calculate the difference between the p_T obtained for the top system and the potential dark matter, since this should give us a way to separate the usual $t\bar{t}$ process from the $t\bar{t}$ + DM process we are looking for. Indeed, in a pure $t\bar{t}$ event, we expect both the top quarks to leave the primary vertex almost back-to-back, while this should not be the case in a $t\bar{t}$ + DM since we expect in this case that the difference in p_T of the tops should be equal to the p_T of the mediator produced at the same time.

If we leave the mediator aside for the moment and focus on the top reconstruction method applied to the usual $t\bar{t}$ process, we can realize quickly that reconstructing the p_T of the tops is not an easy task, because we do not have any way to detect the neutrinos which appear once each one of the top decays (we actually have 6 unknowns, corresponding to the three components of the momenta of the two neutrinos produced). Fortunately, by studying the kinematics of the process, we can get to 6 different equations:

$$\begin{cases} M(b_1 + W_1) = M_t \\ M(b_2 + W_2) = M_t \end{cases}$$

...because the tops produced immediately decay to a bottom and a W.

$$\begin{cases} M(\nu_1 + l_1) = M_W \\ M(\nu_2 + l_2) = M_W \end{cases}$$

...because we consider the leptonic decay of the W in this analysis.

$$\begin{cases} \nu_{1_x} + \nu_{2_x} = (E_T^{miss})_x \\ \nu_{1_y} + \nu_{2_y} = (E_T^{miss})_y \end{cases}$$

...because we assume that the E_T^{miss} is coming from the neutrinos only.

It is then possible to resolve this problem although it presents several technical complications. For example, the mass of the W and the mass of the top that appear in the previous equations can not be fixed, since they are distributed following a Breit-Wigner distribution and are not constant. Moreover, the E_T^{miss} is a variable that we can measure, but this measurement is usually affected by large uncertainties, and other sources of missing transverse energy could appear sometimes, making our calculation completely wrong since we assume that this E_T^{miss} comes only from the two neutrinos produced. Finally, we do not actually know how to match each lepton with its corresponding b-jet, and we can also observe more than two jets in the same event, making it difficult to determine the right combination of jets to take for the calculation. All of these issues imply that the top reconstruction is not expected to work for every event and, sometimes, we just can not find any solution to the problem or even find many different solutions, each one corresponding to one combination of jets.

However, we do have some ways to reduce the impact of the previously mentioned issues, so that we can still manage to get some information with the top reconstruction. We actually just need to realize that the top reconstruction is not going to give us the exact momentum of the top, but will most likely return a probable value, which can be enough if we interpret the top reconstruction as a statistical problem. In practice, what we do is to consider all the (b-)jets of each event, and we then try to reconstruct the kinematics of the system by probing different combinations of leptons and jets. Then, we simply take the solution which gives the smallest mass for the top-antitop system. Applying this method to general $t\bar{t}$ events allows us to reconstruct around 99% of the events with a 20% resolution, but this reconstruction efficiency drops to around 50 % (and these reconstructed events usually have a poor resolution) when we consider heavy $t\bar{t}$ + DM events. The reason for this efficiency drop and loss in resolution is simple: in the second case, the E_T^{miss} is not only coming from the neutrinos, but also from the dark matter particles produced, making the previous equations wrong. The solutions we keep seeing are fortuitous, and appear when the E_T^{miss} casually matches a combination of jets (along with their smearing) and W boson. However, in the dark matter case in particular, we expect the p_T of the tops to be higher than the one obtained for the usual $t\bar{t}$ since they will be accommodating the extra E_T^{miss}. If we manage to compensate the loss in efficiency, we should then obtain an excellent variable to get some discrimination between these two processes.

What we have seen so far can be resumed in Figure 5.1. In this figure, we can see a purple ellipse corresponding to the phase space with solutions, and several points corresponding to the expected solutions for different kind of problems have been represented. As we can see in this figure and from the previous equations, the distance between the dark matter events without solution and the general phase space with solutions correspond directly to the p_T of the mediator which, as we have seen at the beginning of this section, is a variable we are really interested in since we expect it to give a good discrimination between the usual $t\bar{t}$ and the $t\bar{t}$ + DM processes.

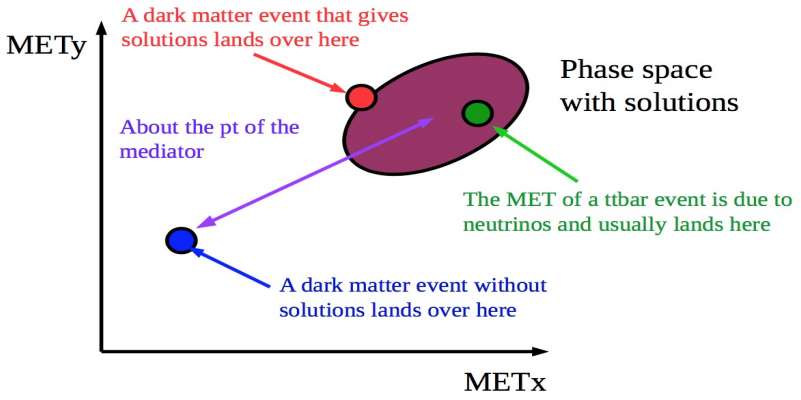


Figure 5.1: Schematic view of the top reconstruction and the different expected solutions for different cases in the E_T^{miss} phase space.

Then, the question remains to know if we can correct the efficiency drop or not, since throwing 50% of the eventual DM events is of course not good. It is actually possible to correct this problem and to get an estimation of the p_T of the mediator of the interaction at the same time. The solution of the previous equations gives a 4-degree polynomial and events with no solution have a polynomial which never crosses the x-axis. If we define the cost as the minimal distance between this polynomial and the x-axis, we can perform an iterative descent (as observed in Figure 5.2) of the observed E_T^{miss} (blue dot in Figure 5.1) to reach the phase space corresponding to the physical solutions (purple ellipse in Figure 5.1), by using directly the definition of the derivative of the cost:

$$-\nabla \vec{cost} = \begin{pmatrix} \frac{\partial cost}{\partial MET_x} \\ \frac{\partial cost}{\partial MET_y} \end{pmatrix} \tag{5.1}$$

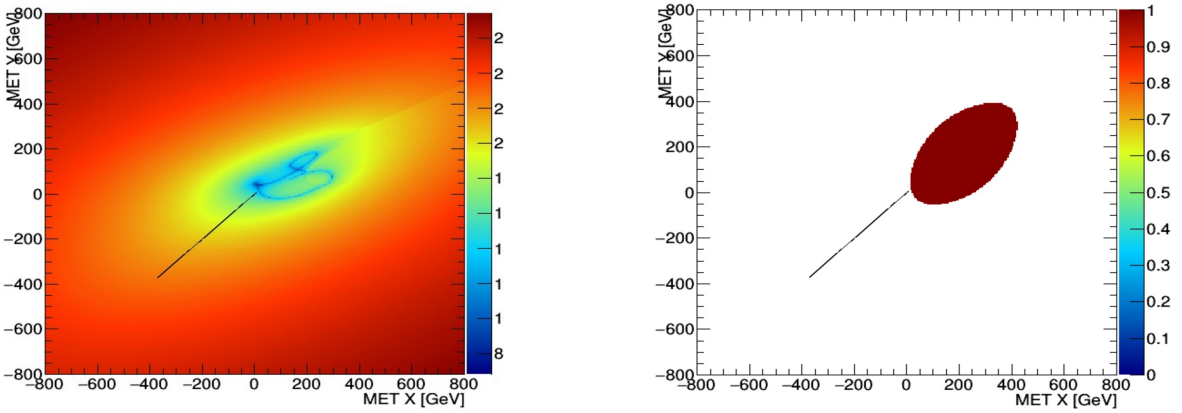


Figure 5.2: Iterative descent performed by minimizing the cost, in order to estimate the mediator p_T as the distance between the solution found and the phase space with solutions, represented by the dark ellipse.

This method then allows us to correct for the efficiency issue of the top reconstruction by making most of the dark matter events converging points (for the heavy scalar mediator, we reach an efficiency of the order of 90%, to be compared with the initial 50%), and gives us moreover a way to estimate the p_T of the mediator of the interaction (we assign to this new variable a negative value if the usual top reconstructions fails, and a value equal to 0 if our more developed reconstruction fails). This variable will be a key of our analysis, as we will now see in Section 5.3.

5.3 Discriminating variables studies

Our complete analysis lays mostly on the discriminating power of four different variables that will be used as input for our neural networks, as we will see in Chapter 6, and which have been obtained by applying the first two levels of selection, as we will we in Table 5.1.

The first of these variables is of course the missing transverse energy since, as previously explained, we expect to see E_T^{miss} appearing with the $t\bar{t}$ + DM process. This variable is able to reduce some backgrounds, such as the DY, but as we can see in Figure 5.3, there is however not much we can do to reduce the $t\bar{t}$ or the ttV for example with this variable only.

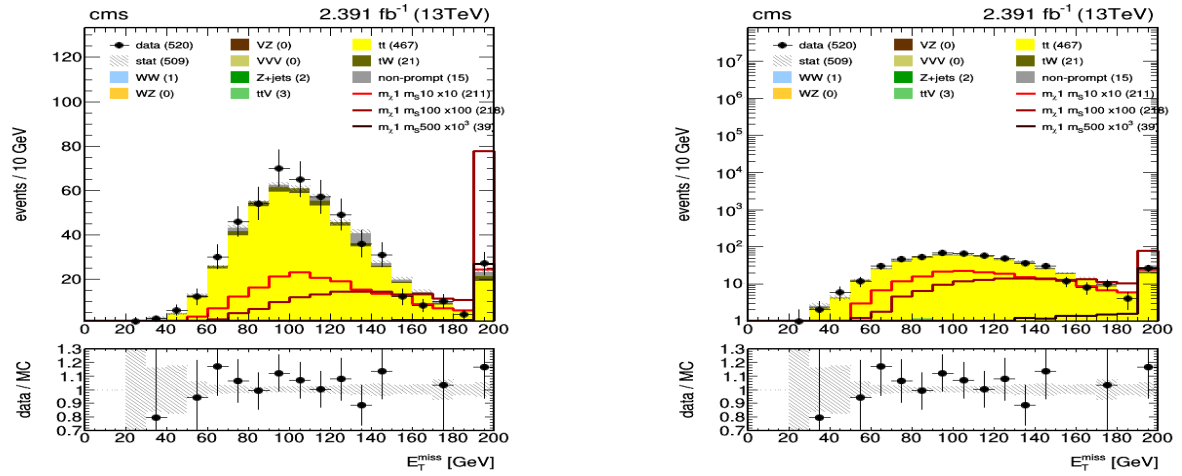


Figure 5.3: Missing transverse energy distribution with (right) and without (left) logarithmic scale for the different processes of the analysis and three different scalar signal mass points, with the first level of event selection of Table 5.1 applied. The error bars represented are statistical only.

Another interesting variable is the angle $\Delta\Phi$ between the two leptons coming from the tops and the E_T^{miss}. We expect that the tops will be much closer to each other when the mass of the mediator produced raises (as seen in Figure 5.4), and this variable should then theoretically give us a way to get some separation between the $t\bar{t}$ and the $t\bar{t}$ + DM. As we can see in Figure 5.5, only a limited discrimination between some of the backgrounds and the signal is introduced in practice with this variable.

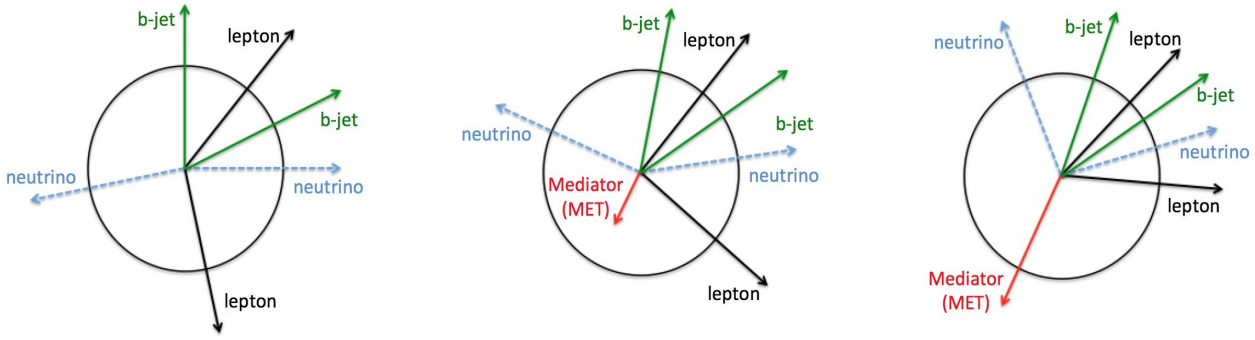


Figure 5.4: Schematic representation in the Φ plane of the distribution of the particles within the detector for the $t\bar{t}$ process (left) and the for the $t\bar{t}$ + DM (center and right).

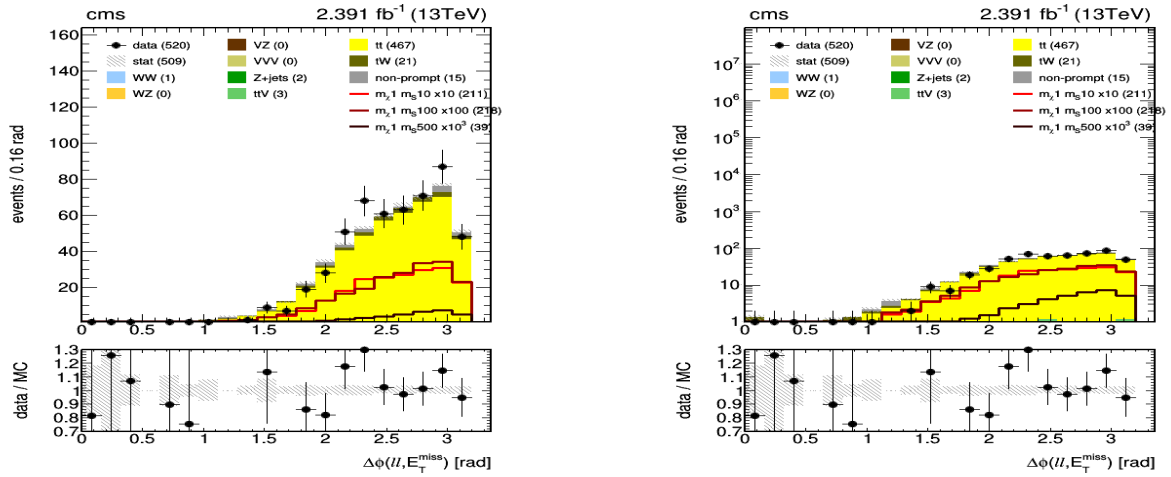


Figure 5.5: Distribution of the $\Delta\Phi$ angle between the two leptons and the E_T^{miss} with (right) and without (left) logarithmic scale for the different processes of the analysis and three different scalar signal mass points, with the first level of event selection of Table 5.1 applied. The error bars represented are statistical only.

The next variable, called m_{T2}^{ll}, is an interesting variable as well since it presents a large discriminating power, as we can see in Figure 5.6. The definition of this variable is not trivial but basically, it represents a measure of the imbalanced transverse momentum in an event and is mostly used for events for which two heavy particles are produced, each of which decaying to at least one undetected particle [46, 47]. To calculate this variable, the following steps are needed.

- We first need to calculate the transverse mass M_T for the two pairs of particles produced, according to the Equation 5.2.

$$\left(M_T^{(i)}\right)^2 = \left(m_{vis}^{(i)}\right)^2 + m_\chi^2 + 2 \cdot \left(E_T^{vis(i)} E_T^{\chi(i)} - \overrightarrow{p}_T^{\,vis(i)} \cdot \overrightarrow{p}_T^{\,\chi(i)}\right) \tag{5.2}$$

- However, there is a problem with the previous equation since the individual quantities $\overrightarrow{p}_T^{\chi(i)}$ of the neutrinos are not experimentally accessible (in this case, we can only determine the value of their sum, $\sum_i \overrightarrow{p}_T^{\chi(i)} = \overrightarrow{E}_T^{\text{miss}}$). We can then generalize in the equation 5.3 the definition of M_T to a variable called M_{T2}, or equivalently m_{T2}^{ll} in this work.

$$m_{T2}\left(m_\chi\right) = \min_{\sum_i \overrightarrow{p}_T^{\chi(i)}} \left[\max\left(M_T^{(1)}, M_T^{(2)}\right)\right] \tag{5.3}$$

- The previous equation has a free parameter m_χ and the minimization is done over the different combinations of missing particles fulfilling the total $\overrightarrow{p}_T^{miss}$ constraint. For this analysis however, we make some assumptions to freeze this parameter.

As we can see in Figure 5.6, adding a cut selecting only events with a large amount of m_{T2}^{ll} seems like a good idea, since this removes a huge part of the $t\bar{t}$ background, the most problematic one, along with a good proportion of all the other backgrounds of the analysis (of course, we also throw away some signal but at the end of the day, we realized that this cut is useful anyway).

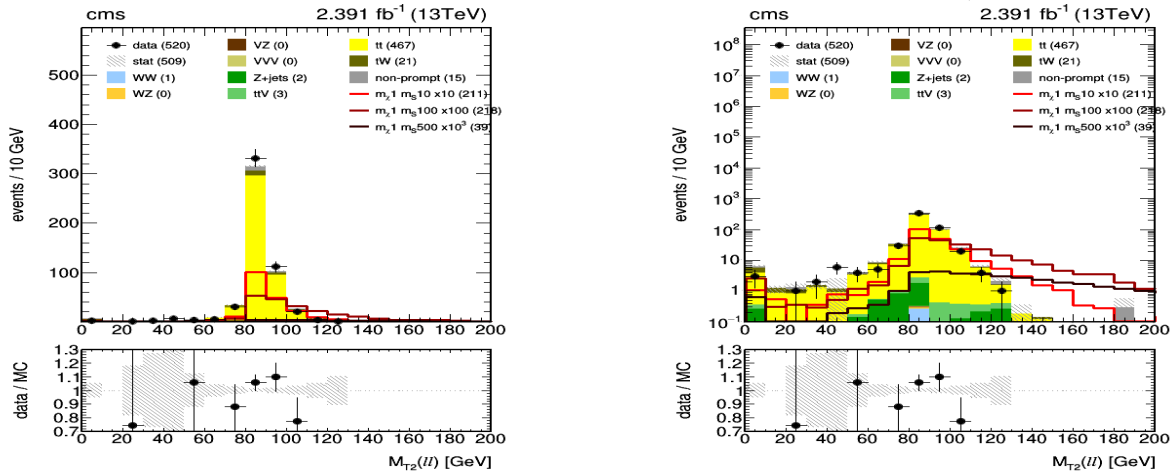


Figure 5.6: Distribution of the m_{T2}^{ll} variable with (right) and without (left) logarithmic scale for the different processes of the analysis and three different scalar signal mass points, with the first level of event selection of Table 5.1 applied. The error bars are statistical only.

The last variable having a reasonable discriminating power is the so-called dark p_T, the p_T of the mediator of the interaction obtained with the top reconstruction method developed in Section 5.2.

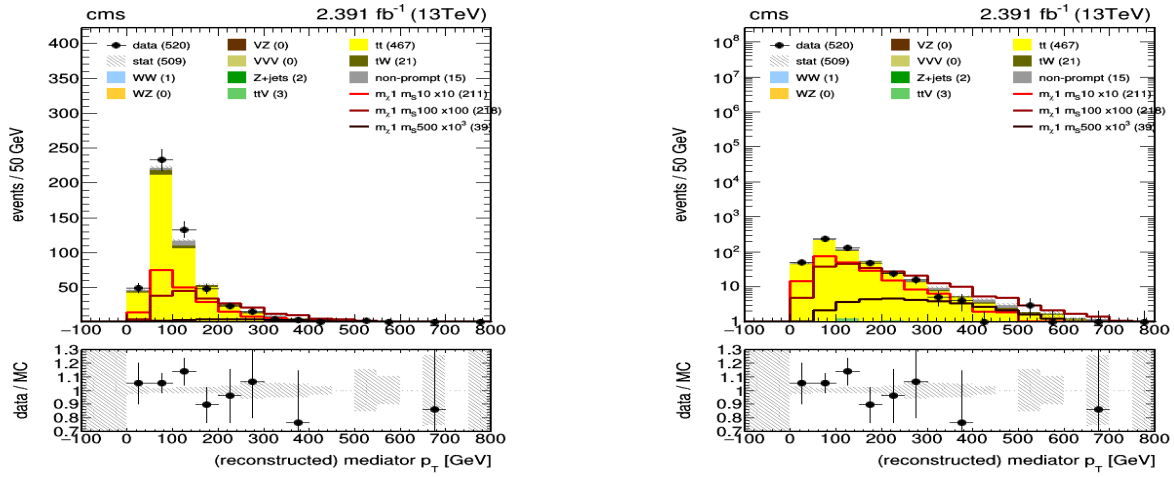


Figure 5.7: Distribution of the dark p_T variable with (right) and without (left) logarithmic scale for the different processes of the analysis and three different scalar signal mass points, with the first level of event selection of Table 5.1 applied. The error bars are statistical only.

5.4 Event selection

We will now discuss the selection and cuts we apply to our analysis, to remove as much background as we possibly can while leaving the signal as high as possible. The main objective of the cuts we apply is to improve the value of the significance, usually defined as the ratio between the number of yields of signal and the square root of the number of yields of all the backgrounds. A higher value of significance is of course what we are aiming for since it would mean that we managed to have more signal with respect to the backgrounds.

The cuts we apply in our analysis are described in Table 5.1. Basically, we have applied two different levels of cuts, corresponding to a general prelesection and then to a selection more specific to our analysis. Here are some details about these two levels of cuts:

- First of all comes the preselection, an initial set of cuts that select only events passing the single or double lepton triggers listed in Section 4.2 and having at least two tight leptons with a p_T higher than 25 GeV for the leading lepton (the one having the highest p_T) and higher than 20 GeV for the trailing lepton (the lepton having the second highest p_T). This is required to avoid the low p_T regions where the triggers lose in efficiency and to remove some of the fake background contribution, since we expect the fake rate to be higher for low p_T leptons. Moreover, we require these two leptons to have opposite charges. Finally, we also reject events having a significative third lepton ($p_T > 10$ GeV) to remove most of the backgrounds featuring three or more leptons in the final state, such as the WZ.
- Then come all the cuts more specific to our analysis in particular. For example, we require the invariant mass of the lepton pair m_{ll} to be higher than 20 GeV in order to reject possible low mass resonances which are not taken into account by the simulated samples. We also apply a Z veto to the ee and $\mu\mu$ channels in order to remove most of the Drell-Yan contribution, and we select only events having at least two jets ($p_T > 30$ GeV) and at least one b-jet (medium working point of the CSSV v2 b-tagger). Then, we also look for events having a high E_T^{miss} (> 80 GeV) and a high value of m_{T2}^{ll} since, as explained in Section 5.3, this variable has a strong discriminating power in our analysis. Finally, we ask the reconstructed mediator p_T to be higher than 0, meaning that we select only events for which our top reconstruction did not fail.

Number	Cut level	Cut	Comment
0	Preselection	$p_T^{lep_1} > 25$ GeV $p_T^{lep_2} > 20$ GeV $p_T^{lep_3} < 10$ GeV $q_l^{lep_1} \cdot q_l^{lep_2} < 0$	On the leading lepton On the trailing lepton Third lepton veto Opposite charge requirement
1	First level	$m_{ll} > 20$ GeV $\lvert m_{ll} - m_Z \rvert > 15$ GeV $n_{jet} \geq 2$ $n_{b_{jet}} \geq 1$	Only applied to ee and $\mu\mu$ channels At least one medium csv-v2 b-jet
2	Second level	$E_T^{\text{miss}} > 80$ GeV $m_{T2}^{ll} > 80$ GeV Dark $p_T > 0$ GeV	The top reconstruction has to work

Table 5.1: Description of the complete selection applied to our analysis.

5.5 Control regions

5.5.1 $t\bar{t}$ scale factor and control region

Since the $t\bar{t}$ is our most important background, we want to make sure that we understand it right and that the simulations are correct. One way to check this consists in calculating a scale factor, defined as 5.4, in bins of the variable m_{T2}^{ll}, by which we can scale the MC simulations.

$$SF = \frac{\left(n_{\text{data}} - n_{\text{other backgrounds}}\right)}{n_{t\bar{t}}} \tag{5.4}$$

The scale factor dependance with m_{T2}^{ll} obtained this way can be represented in Figure 5.8 in which the red line corresponds to a fitted constant to the points in the region where m_{T2}^{ll} is comprised between 40 and 80 GeV (this plot has been obtained by applying the cuts corresponding to the first level of selection, along with the $40 < m_{T2}^{ll} < 80$ GeV and dark $p_T \geq 0$ cuts). We obtain a value of (0.974 ± 0.003) for this scale factor, where the uncertainty shown correspond to the statistical error only.

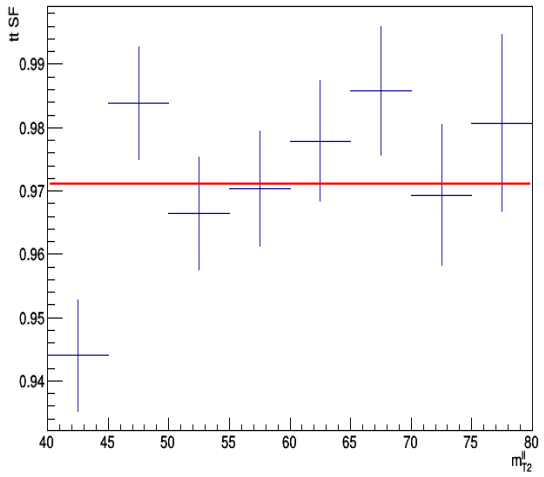


Figure 5.8: Calculated scale factor for the $t\bar{t}$ process in bins of m_{T2}^{ll}. The error bars represented are statistical only.

Let's now start studying the $t\bar{t}$ control region, in order to check if the major background of the analysis seems to be correctly simulated. This control region is defined with the same cuts as the ones applied to our analysis (as seen in Table 5.1 in Section 5.4) except that we remove the $E_T^{\text{miss}} > 80$ GeV requirement, and that we switch the $m_{T2}^{ll} > 80$ GeV to $m_{T2}^{ll} < 80$ GeV, to get a region as pure in $t\bar{t}$ as possible. The distributions for three different variables (E_T^{miss}, $\Delta\phi_{ll,E_T^{\text{miss}}}$ y m_{T2}^{ll}) are represented in Figure 5.9, where the $t\bar{t}$ and DY scale factors calculated in Section 5.5.1 and appendix A.1 have been applied. As we can see, in the region defined this way, we obtain 85,7% of $t\bar{t}$ and we can observe a nice agreement between the data and the MC.

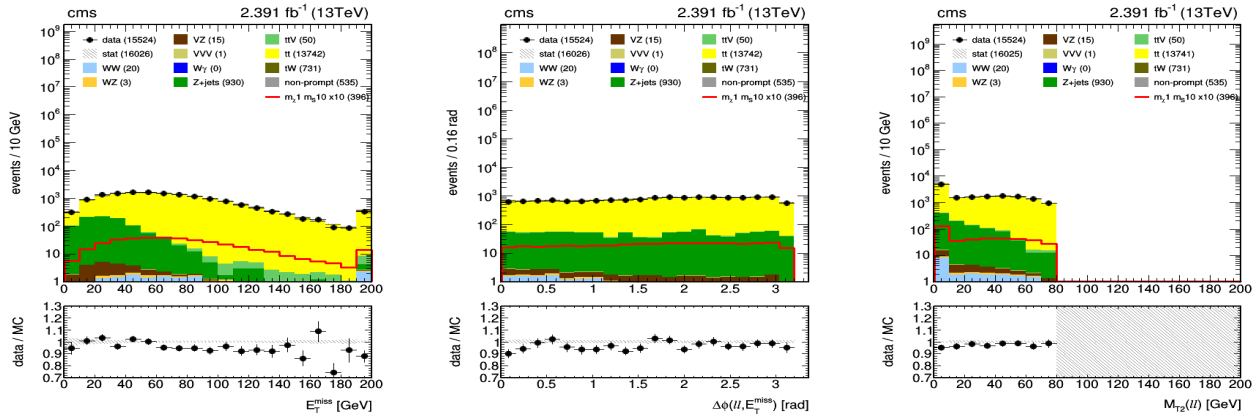


Figure 5.9: Some distributions in the control region defined to check for the validity of the $t\bar{t}$ process.

5.5.2 ttZ control region

We can perform a similar analysis to check for the validity of the ttZ process. To get a control region as pure as possible in ttZ, we only look at events having three leptons in the final state with one pair having a mass close to the one expected for the Z in order to remove most of the backgrounds, such as $t\bar{t}$. We also require the E_T^{miss} to be higher than 50 GeV and we only select events having $m_{T2}^{ll} > 80$ GeV and at least one b-jet. The distributions for three variables of the analysis are represented in Figure 5.10, where we see that we obtain 50% of ttZ and a scale factor of (1.85±0.29).

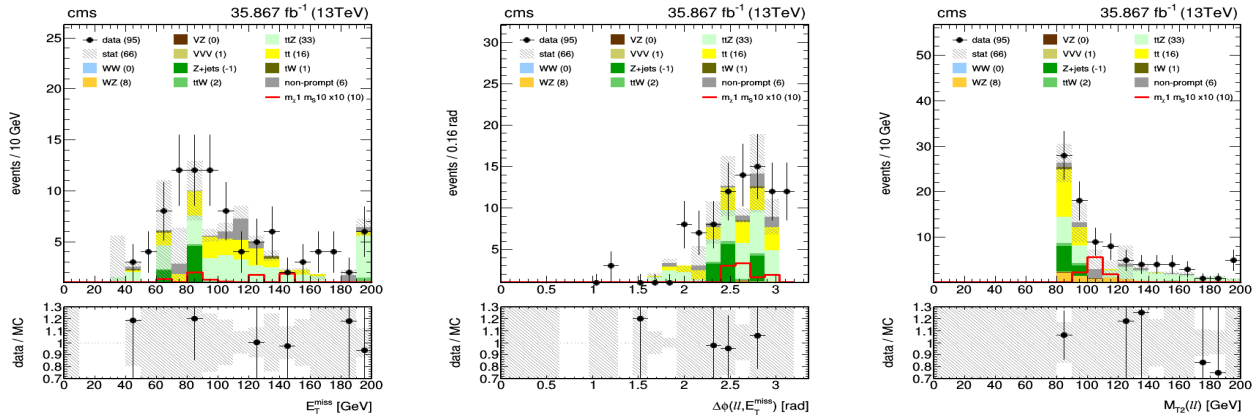


Figure 5.10: Some distributions in the control region, to check for the validity of the ttZ.

5.6 Uncertainties of the analysis

All the measurements we can make in high energy physics and in physics in general present some uncertainties, and the determination of their value is a critical point of every analysis, especially in this case since we try to detect a really low signal over a large background. A detailed study about the origin and the impact on the final results of all the systematic uncertainties being well beyond the scope of this work, we will only consider and make some general comments about the statistical and some systematic uncertainties.

Statistical uncertainties are easy to deal with, and appear in any counting experiment [48] since each measurement, made of a finite set of observations, results in different observations. The statistical uncertainty is then just a measure giving us an idea about the range of this variation. What we do is that we consider that our counting experiment can be approximated with a Poisson distribution (we actually have a binomial distribution but this approximation is justified whenever the number of measurements N gets really high, as this is usually the case in particle physics since the number of bunch crossings and the number of particles per bunch are really high, or when the

probability of observing an event gets really small), for which we know that the error on the number of measurements is directly given by its square root.

The systematic uncertainties are just as important in this analysis, but they are more difficult to estimate and are really different in nature [49] since they arise directly from the theory or from the detector itself. A typical example of systematic uncertainty might be related to the calibration of the detector, since a bad calibration will have the same impact on all the measurements we want to perform. This kind of uncertainty can be separated into two different categories usually referred to as the theoretical and the experimental systematic uncertainties.

- **Theoretical systematic uncertainties**. This category is related directly to the theoretical models we use in the analysis and is mostly related to the production of the Monte Carlo simulations, since we use them to predict the background. Indeed, this kind of simulation lays on several different models which are not able to describe nature in a perfect way. For example, as accurate as the Feynman diagrams are, it is impossible for us to make them represent the complete picture since they get really complicated once we start considering next to leading order (NLO), or even next to next to leading order (NNLO) perturbations. The number of possible diagrams increases exponentially with each level of precision, and so do the computational needs to take them all into account. Usually, the Monte Carlo simulations are performed at NLO order, which might introduce a small systematic uncertainty. There also exists another uncertainty of this kind arising from the fact that we do not know the exact shape of the Parton Density Functions (PDF), a crucial element when it comes to the production of theoretical simulations.
- **Experimental systematic uncertainties**. Many different sources of experimental systematics have been studied in this work, to get results as precise as possible. We can separate this kind of uncertainties itself into two different categories (the same distinction can also be made for the theoretical systematics).
    - **Scale uncertainties**. To calculate their impact, we observe the change in yields obtained when moving the nominal values up and down by a value given by this uncertainty. The 2.5% error on the determination of the integrated luminosity [29] is a typical example which is treated this way. The lepton reconstruction and identification efficiencies, the triggers acceptance, the b-tagging efficiency or the fake rate uncertainties all belong to this category as well [50].
    - **Shape uncertainties**. Sometimes, the systematic uncertainty can not just be resumed to a single number or percentage, but is determined by the complete shape of a variable. The jet energy scale is a typical example corresponding to this category [50].

As previously explained, understanding and taking into account all the different sources of systematic uncertainties is crucial in this analysis, especially when producing the final limits that will be shown in Section 7. Indeed, we need to get a precise idea about the error we are committing on the different number of yields we measure, in order to be able to detect eventual significative deviation between the expected and measured limits for the different mediator masses considered.

Chapter 6

Neural network

Performing a general cut and count analysis is usually not optimal when looking for dark matter production, since the significance of this signal is expected to be really low, mostly because of the different backgrounds such as the $t\bar{t}$. Because of this low significance, we need to come up with solutions to extract as much information as we can from the data. This is the main reason why we decided to perform a Multi-Variate Analysis (MVA) by creating several Artifical Neural Networks (ANN), using general machine learning techniques. First of all, we will study a bit of theory concerning this kind of analysis, to understand better the reasons why we decided to use it and the precautions we took using it. Then, we will talk about and study different parameters of the neural networks we have built in particular, and we will finally show all the results we obtained.

6.1 Multi-Variate Analysis

As the name of this method [51] suggests, the main idea of this kind of analysis consists in observing several different variables at the same time, to combine the information coming from all of them in order to find a way to classify a single event as either background or signal. We want with this method to be able to reduce a large number of inputs to a single output and the MVA is performed in this case using several neural networks, as we will see in Section 6.2. In summary, the objective of the MVA is to use and feed this network with the input information coming from the four variables having the most discriminating power in our analysis (as described in Section 5.3), to get in return a single new variable that we can use in our analysis to discriminate the signal and the backgrounds.

6.2 Artificial Neural Network

A general Artificial Neural Network has been represented in Figure 6.1, where we can see that it is able to combine different input variables into a single one, through a hidden layer made out of neurons. Each neuron of this layer is able to combine the different inputs it receives in a certain way, through the definition of several weights ω_i, which are numbers expressing the importance of the respective inputs to the output of the neuron [52].

In this work, we will consider networks composed by two different kinds of neurons having either a sigmoid or a tangent hyperbolic activation function. The output of a sigmoid neuron is a real number between 0 and 1, and can be defined with the sigmoid activation function 6.1, where x_j are the inputs of the neuron, ω_j the weights and b the bias, defined as a measure of how easy it is to get the neuron to give an output different than 0 [53].

$$\sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-\sum_j \omega_j x_j - b}} \tag{6.1}$$

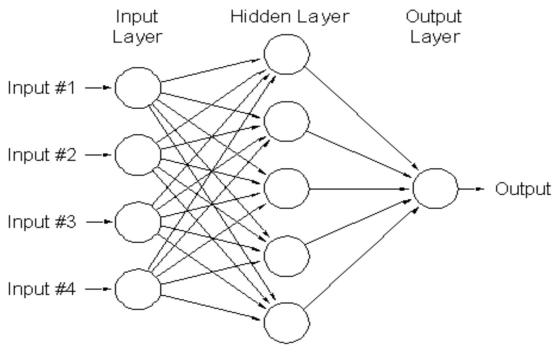


Figure 6.1: General representation of the ANN defined for the MVA analysis, able to combine the discriminating power of several different input variables into a single output [51].

On the other side, the tanh activation function 6.2 has a slightly different expression, giving back a number between -1 and 1 (but for this analysis, we will actually convert back this output to get a number between 0 and 1 as well). As in the previous equation, z is defined as $\sum_j \omega_j x_j - b$.

$$tanh(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{6.2}$$

An important point to note is that a neural network in itself is useless, since we first need to train it with a fraction of the events we have at our disposition for it to be able to calculate the output and to therefore create a new variable useful to distinguish between signal and background. We have to be careful however when choosing the number of training events, because a number too small could result in a inadequate fit because of the loss of fitting capacity of the network. We could then be tempted to use as many events as possible to train our network to improve it, so that we can get a more reliable output, but we actually have to be careful not to use too many events for training, for two main reasons. First of all, simply because the events we use to train the network are actually lost, since we can not use them later on for the analysis and then because we have to be really careful to avoid the overtraining issue of the network.

This overtraining appears when the network is creating a perfect model fitting the training events instead of trying to find a way to generalize the trend observed with this dataset. A simple example of this overtraining issue can be made by considering a network allowed to create a model which has a number of parameters greater than the number of observations. In this case, the network will give us back a model fitting perfectly the training data since it is always possible to find a function with n degrees of freedom passing exactly through n points, but such a model will typically fail drastically when making predictions about new data because of the loss of generalization. It will then be crucial to check for any sign of overtraining with the networks that will be used in this analysis.

6.3 Characterization of the neural networks

6.3.1 Input variables

The four variables used as input for our neural networks are the variables (E_T^{miss}, $\Delta\phi_{ll,E_T^{\text{miss}}}$, m_{T2}^{ll} and the so-called dark p_T, the reconstructed p_T of the mediator) presented in Section 5.3, because they are the variables offering the highest discriminating power between background and signal. The different distributions of the input variables used for the MVA have been represented in Figure 6.2 thanks to the TMVA (Toolkit for Multivariate Analysis) tool [54], for both the 10 GeV and 500 GeV scalar mediators. It is important to note at this point that we train a different neural network for each mass point.

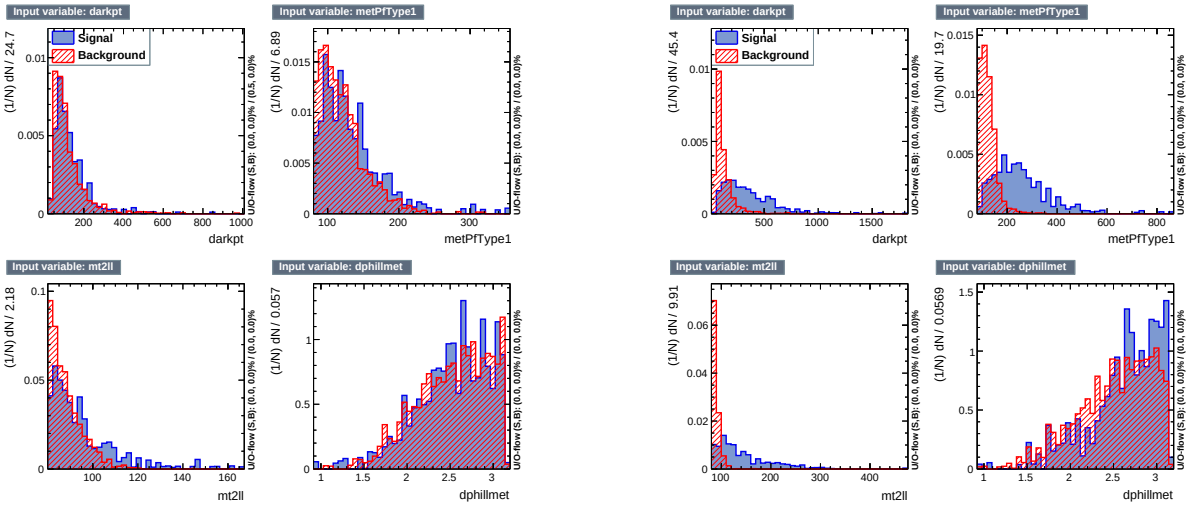


Figure 6.2: Distribution of the signal and background for the four input variables used for the MVA. The background in red corresponds to the $t\bar{t}$ process, while the signal in blue corresponds to either the 10 GeV scalar mediator, the mass point offering the worst discrimination (left) or the 500 GeV scalar mediator, offering the best discrimination (right).

We can also plot in Figure 6.3 the correlation matrices for the background and signal, for the four different variables considered. This is useful mainly because we want to use the smallest possible number of input variables to feed our network in order to simplify the analysis, and this plot allows us to check for eventual correlations and therefore detect eventual superfluous input variables.

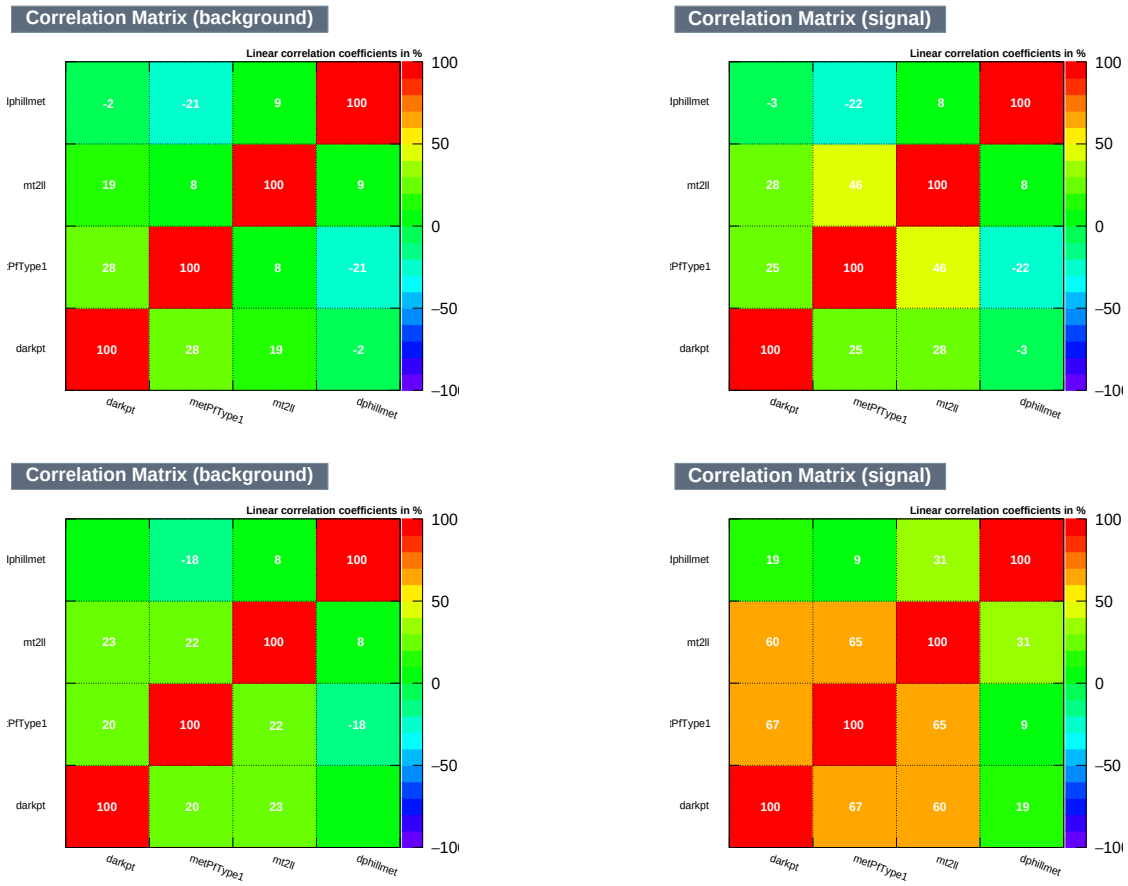


Figure 6.3: Study of the correlation between the input variables, for both the 10 GeV (top) and 500 GeV (bottom) scalar mediators, considering the background (left) and the signal (right).

By looking at the previous plots, we can conclude that we do not observe any strong correlation (or anti-correlation) for the background, if we consider the input variables of the neural networks built for the 10 and 500 GeV mediator masses. The 10 GeV signal distributions do not present any sign of strong correlation either, but we can see that three of the four input variables used to build the 500 GeV network actually present some correlation between each other.

6.3.2 Architecture

The architecture of one of the neural networks used in this analysis is represented in Figure 6.4. As we can see, the neural networks we built are made out of two hidden layers, containing respectively six and three neurons. They are able to read the input from the four previous variables and give us in return a single variable which combines the information from all the input variables. We consider in this work both sigmoid and tanh-like neurons, to compare the results we obtain in both cases.

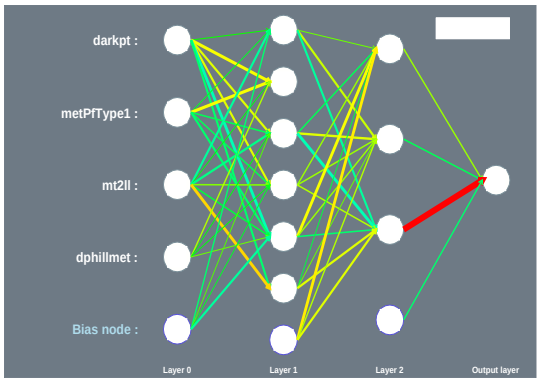


Figure 6.4: Architecture of the sigmoid neural network built for the 10 GeV signal sample of this analysis. This network has two layers of hidden neurons and consider four input variables.

6.3.3 Training

In order to get an efficient neural network, we need to train it properly first. The network is for the moment trained only against the $t\bar{t}$, the most important background of our analysis, in order to avoid overtraining issues and to simplify the analysis. We use 200 events to train all the networks to avoid any eventual overtraining, and as previously explained, we train a different network for each dark matter mediator mass, and for each mediator type (scalar and pseudoscalar).

The relative importance of the different input variables for the training can be calculated automatically [54] as the sum of the weights-squared of the connections between the input variables and the first layer of the network, as described in Equation 6.3, where I_i corresponds to the importance and where $\bar{x}_i$ is the mean value of the variable i considered.

$$I_i = \bar{x}_i^2 \sum_{j=1}^{n_h} \left(\omega_{ij}^{(1)}\right)^2 \tag{6.3}$$

The results obtained for the importance of our variables are represented in Table 6.1, for the different networks. As we can see in this table, the importance of the variable changes quite a lot depending on the signal sample used to train the ANN: when we consider the 10 GeV network, we see that the most useful variable is our dark p_T but when we consider the 500 GeV network, the most useful variable is m_{T2}^{ll} for both the sigmoid and tanh-like cases. Moreover, we can also see that the $\Delta\phi_{ll,E_T^{\text{miss}}}$ variable is clearly the less important variable in all the cases and finally, we can conclude that both the sigmoid and tanh-like networks are giving similar results, even though the ranking and the importance of some variables change from time to time.

Variable	10 GeV network		500 GeV network	
	Sigmoid-like	Tanh-like	Sigmoid-like	Tanh-like
dark p_T	11.06	10.93	4.62	6.64
m_{T2}^{ll}	7.18	3.70	17.75	14.83
E_T^{miss}	3.84	4.48	6.77	5.69
$\Delta\phi_{ll,E_T^{\text{miss}}}$	0.58	0.82	0.40	0.69

Table 6.1: Importance of the different variables used for the training of the neural networks for both the sigmoid (left) and tanh-like networks (right).

6.4 Results

Before checking the actual output we get, we need to check some control plots to verify that everything went well and to compare the different networks created. First of all, we can plot the convergence test in Figure 6.5, to check the convergence in all the cases, for both the training and test samples. As we can see in this figure, in the 500 GeV case, both the training and test samples converge to the almost same value, while they seem to be converging to different values in the 10 GeV case, which is the typical sign that we are committing some overtraining in this case.

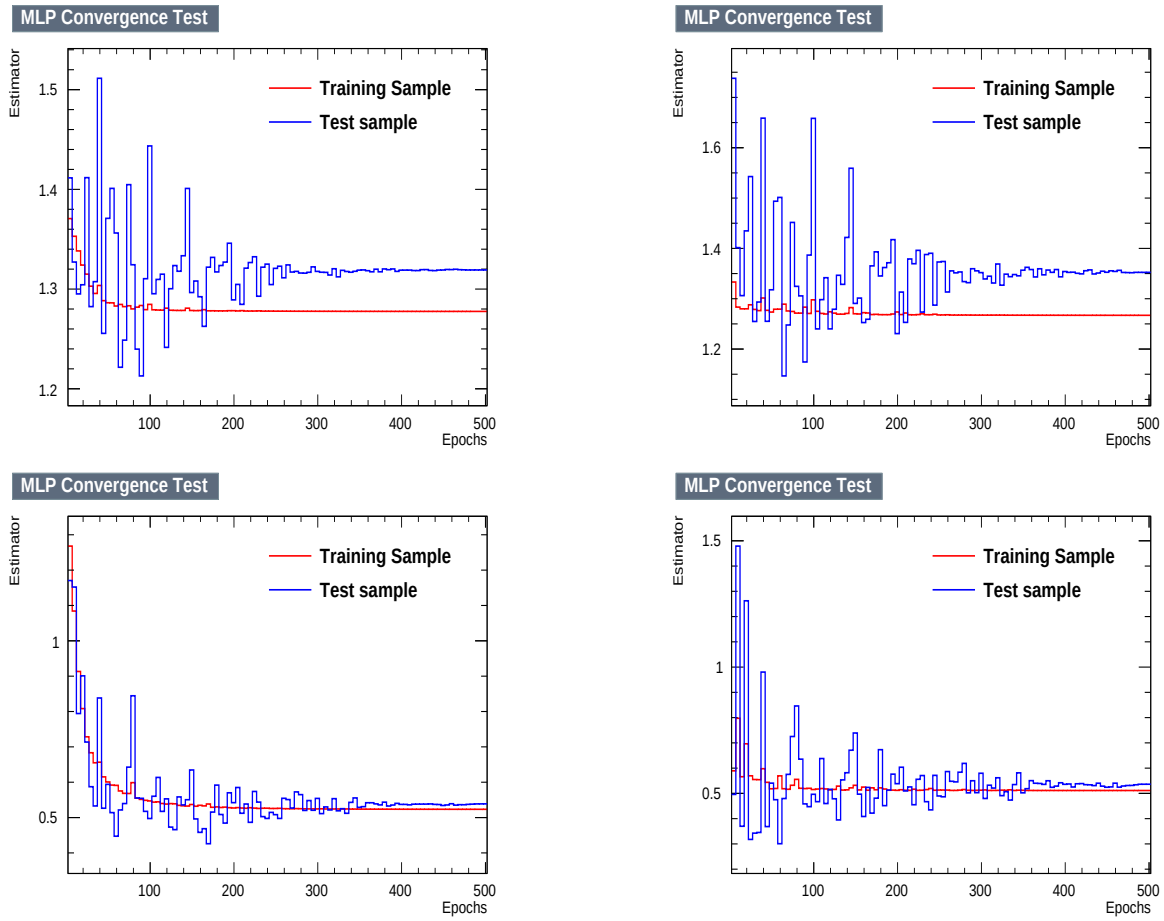


Figure 6.5: Study of the convergence for the training and test samples, for both the 10 GeV (top) and 500 GeV (bottom) networks, and for both the sigmoid (left) and tanh cases (right).

At this point, it is time to check in Figure 6.6 for eventual signs of overtraining in the response obtained by any of the networks. As we can see in all the plots, we do not observe any strong evidence of overtraining since the distributions obtained for the test and train samples are quite similar for both the signal and the background. The Kolmogorov-Smirnov test seems to indicate however that the sigmoid response is a bit better and presents less overtraining, for both the 10 and 500 GeV networks. We can also clearly see that the 10 GeV signal sample presents responses much closer to the background distributions, while the 500 GeV signal gets a much better separation with the background.

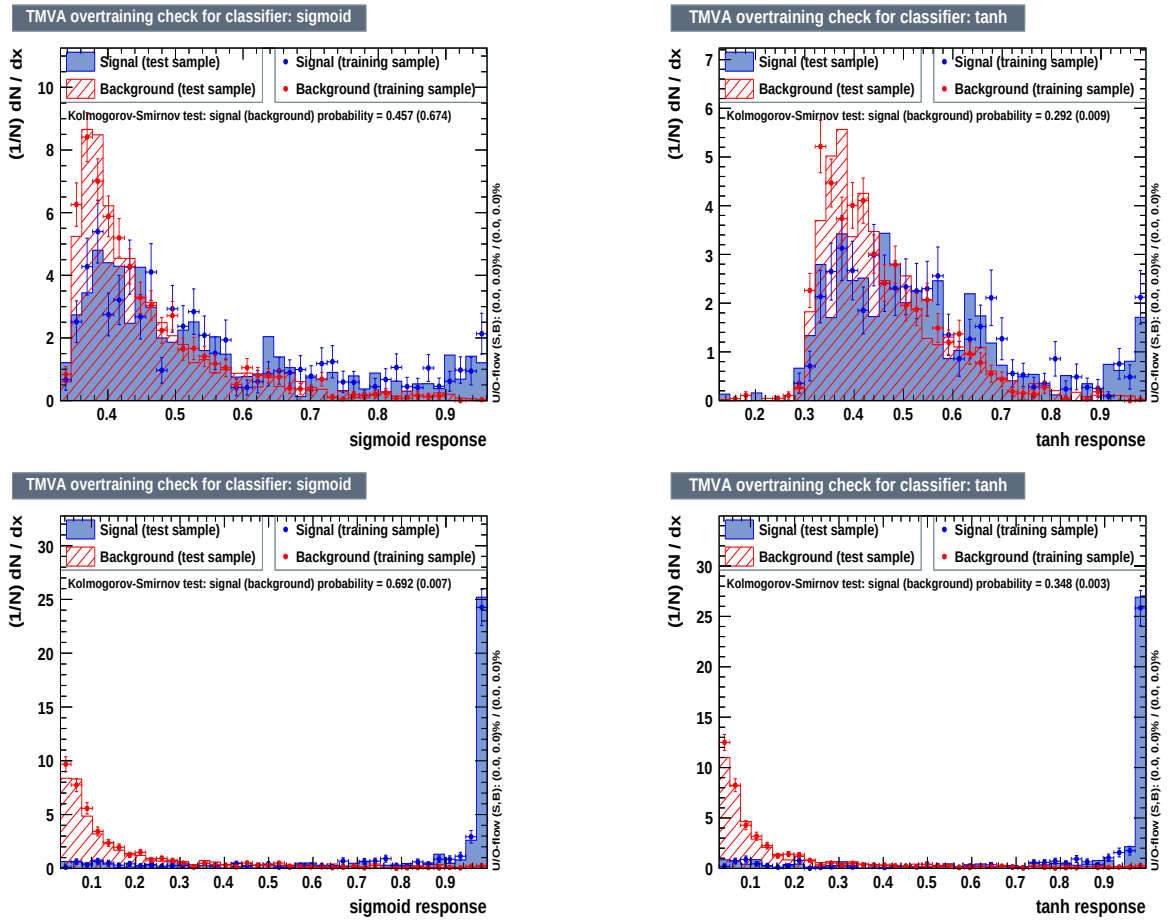


Figure 6.6: Study of the response of the network, to search for eventual overtraining, for both the 10 GeV (top) and 500 GeV (bottom) networks, and for both the sigmoid (left) and tanh cases (right).

We can also study the background rejection versus the signal efficiency in all the cases, as represented in Figure 6.7 for the 10 GeV and 500 GeV dark matter scalar mediators. This is an important plot to consider, because the output given by any of the networks is used for the moment in a simple cut and count analysis, by finding the optimal significance point using this significance curve. As we can see in this figure, we actually won't be able to find an optimal cut for the analysis, and it will always be a compromise. We can either try to raise the signal efficiency, but this will let pass trough more background, or try to reject as much background as possible while losing in signal efficiency at the same time. However, as expected, we can reach a much better signal efficiency in the 500 GeV case than the one possible to obtain for the 10 GeV mediator, by keeping a similar level of background rejection.

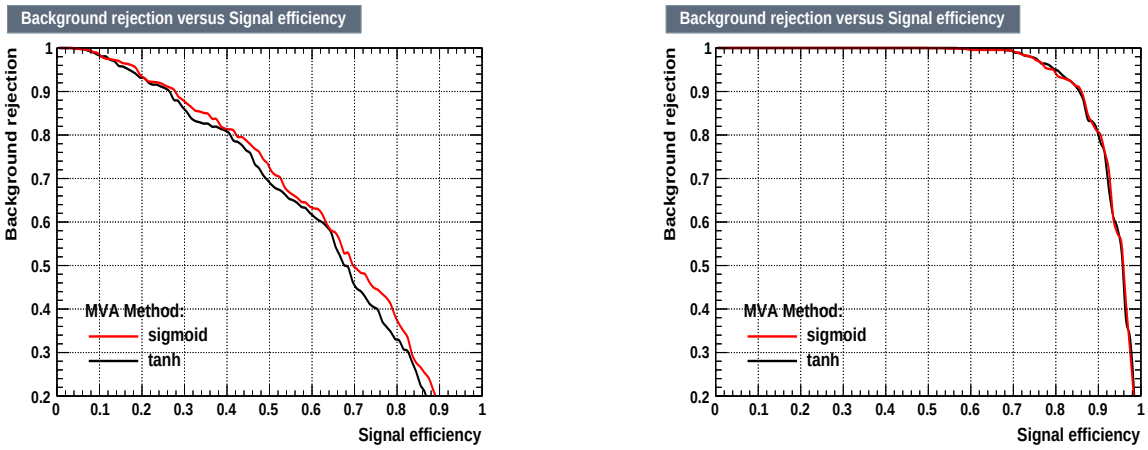


Figure 6.7: Comparison of the background rejection with respect to the signal efficiency of our MVA, for both the 10 (left) and 500 GeV (right) neural networks.

Finally, we can plot the significance curves and distributions obtained for the output of the sigmoid-like networks we have built in Figures 6.8 and 6.9. With all the cuts of the analysis applied, we do not observe a strong discrimination with the new variable coming from the network for the 10 GeV mass point, but we do get a much more significative discrimination for the 500 GeV scalar mediator (we observe a clear maximum in the red curve around 0.98 in this latest case).

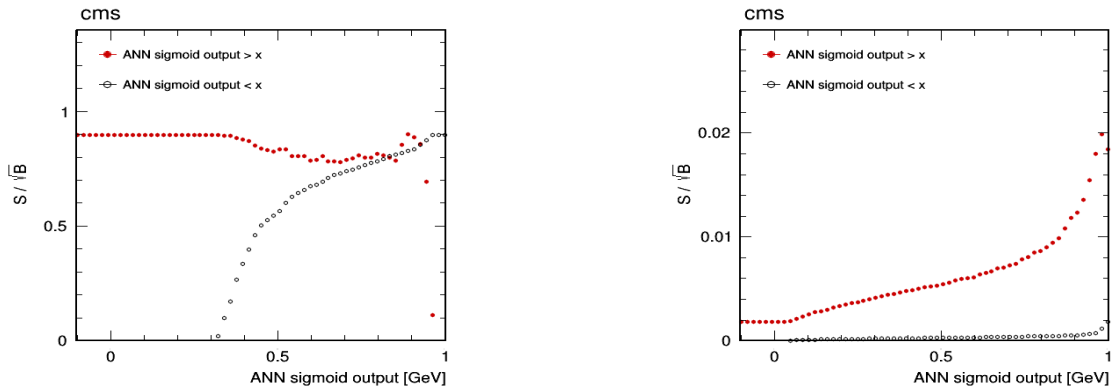


Figure 6.8: Significance curves of the new variable obtained, for both the 10 GeV (left) and 500 GeV (right) mediators. Errors are statistical only and all the cuts of the analysis have been applied.

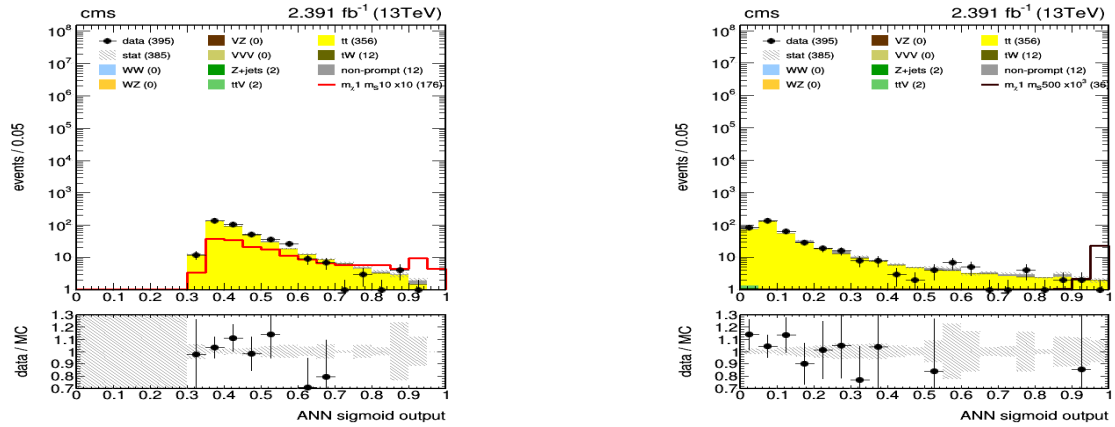


Figure 6.9: Distributions of the new variable obtained, for both the 10 GeV (left) and 500 GeV (right) mediators. Errors are statistical only and all the cuts of the analysis have been applied.

Chapter 7

Results

First of all, in Table 7.1 are represented the expected yields for the main backgrounds, different signal masses and for the data at the final selection level, corresponding to an integrated luminosity of 2.4 fb^{-1} and for the three different channels available separately and then combined.

Process	Channel ee	Channel $e\mu$	Channel $\mu\mu$	Channel ll
WW	0.02 ± 0.01	0.06 ± 0.03	0.20 ± 0.05	0.28 ± 0.06
WZ	0.02 ± 0.01	0.01 ± 0.01	0.01 ± 0.01	0.04 ± 0.01
VZ	0.00 ± 0.00	0.01 ± 0.00	0.00 ± 0.00	0.01 ± 0.00
VVV	0.00 ± 0.00	0.01 ± 0.00	0.02 ± 0.01	0.03 ± 0.01
Z+jets	0.59 ± 0.35	1.14 ± 0.53	0.00 ± 0.00	1.73 ± 0.63
$t\bar{t}$V	0.32 ± 0.02	0.66 ± 0.04	1.09 ± 0.04	2.07 ± 0.06
$t\bar{t}$	46.95 ± 0.33	119.20 ± 0.56	191.34 ± 0.69	357.49 ± 0.95
tW	1.64 ± 0.14	4.01 ± 0.22	6.78 ± 0.28	12.43 ± 0.39
Non prompt	0.00 ± 1.64	0.00 ± 1.32	11.74 ± 3.07	11.70 ± 3.72
Scalar mediator				
m_χ 1, m_ϕ 10 (x10)	25.07 ± 1.57	58.42 ± 2.56	93.76 ± 3.12	177.25 ± 4.33
m_χ 1, m_ϕ 100 (x100)	25.45 ± 0.95	67.24 ± 1.66	106.83 ± 2.02	199.53 ± 2.78
m_χ 1, m_ϕ 500 (x10^3)	4.95 ± 0.11	12.35 ± 0.19	18.66 ± 0.23	35.96 ± 0.32
Pseudoscalar mediator				
m_χ 1, m_ϕ 10 (x10)	1.96 ± 0.06	5.08 ± 0.10	7.92 ± 0.12	14.95 ± 0.17
m_χ 1, m_ϕ 100 (x100)	12.34 ± 0.35	31.61 ± 0.60	47.42 ± 0.72	91.38 ± 1.00
m_χ 1, m_ϕ 500 (x10^3)	5.00 ± 0.12	13.65 ± 0.21	20.13 ± 0.25	38.78 ± 0.34
Total background	49.53 ± 1.71	125.10 ± 1.55	211.17 ± 3.16	385.77 ± 3.91
Data	51.00	133.00	211.00	395.00

Table 7.1: Expected yields for the main backgrounds and some signals of the analysis at the final selection level, and measured yields in data. The V boson can either stand for a W or a Z boson, and the errors are statistical only.

We can then plot in Figure 7.2 all the main variables of the analysis at the final selection level, for three different scalar mediator masses (10, 100 and 500 GeV). The same is done for the different pseudoscalar mediators in Figure 7.3, where all the signal distributions have been rescaled by a given factor to make them more visible.

Using the data available we can finally also establish an upper-limit on the dark matter production cross section value for different mediator masses. Plotting the upper limit on the cross section at the 95% confidence level allows us to eventually reject some of the dark matter mass points considered. Indeed, if we get a value smaller than one for the limit, we would expect to have enough sensitivity and therefore to be able to detect the model considered.

The limits obtained for different mediator masses and for scalar mediators are shown in Table 7.2. These limits correspond to the best limits obtained by changing the value of the cut applied on the output variable using a tanh-like neural network, as previously described in Chapter 6. The best output of the MVA is for now considered as the best significance point, but we will move to a complete shape analysis as soon as possible.

m_S [GeV]	Best AAN cut	Expected central limit	1σ interval	2σ interval	Observed limit
10	0.75	2.99	[1.95, 4.87]	[1.38, 7.81]	2.22
20	0.80	3.12	[2.04, 5.04]	[1.43, 8.14]	2.73
50	0.90	4.14	[2.66, 6.81]	[1.87, 11.19]	2.90
100	0.90	7.03	[4.43, 11.80]	[3.05, 19.81]	5.94
200	0.90	25.37	[16.26, 41.36]	[11.40, 67.93]	19.69
300	0.80	54.25	[35.20, 88.42]	[24.69, 144.03]	50.34
500	0.98	175.50	[106.94, 308.41]	[71.64, 531.91]	116.65

Table 7.2: Limits obtained at an integrated luminosity of 2.4 fb^{-1}, for different masses corresponding to the scalar mediators.

The same can of course be repeated, considering this time different masses of pseudoscalar mediators. These results can be found in Table 7.3.

m_S [GeV]	Best ANN cut	Expected central limit	1σ interval	2σ interval	Observed limit
10	0.80	9.47	[6.17, 15.28]	[4.38, 24.46]	7.66
20	0.85	9.94	[6.50, 16.04]	[4.60, 25.90]	8.18
50	0.85	11.47	[7.50, 18.51]	[5.31, 29.63]	10.18
100	0.90	11.03	[7.05, 18.33]	[4.93, 30.09]	8.91
200	0.95	17.31	[10.77, 29.32]	[7.41, 49.18]	9.81
300	0.90	33.12	[21.10, 55.04]	[14.69, 91.09]	28.98
500	0.98	167.94	[103.19, 289.77]	[69.86, 495.14]	105.27

Table 7.3: Limits obtained at an integrated luminosity of 2.4 fb^{-1}, for different masses corresponding to the pseudoscalar mediators.

We can now plot these expected and observed limits in Figure 7.1 along with the 1 and 2σ intervals (in green and yellow, respectively), for the different mass points considered and for both the scalar and the pseudoscalar mediators.

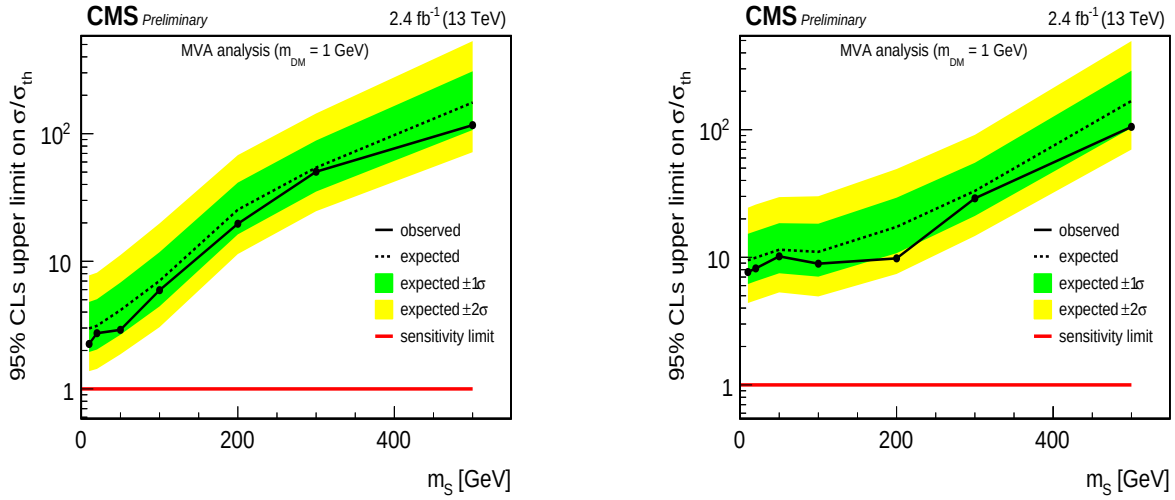


Figure 7.1: Limits obtained at 2.4 fb^{-1} for our analysis, for both the scalar (right) and pseudoscalar mediators (left).

As we can see in the previous tables and plots, we can not exclude any dark matter model by applying our analysis to the blinded dataset corresponding to an integrated luminosity of 2.4 fb^{-1}. Since we know that the expected limit is being reduced as the square root of the luminosity, we expect to be able to exclude the 10 and 20 GeV scalar mediators when unblinding our analysis. All the expected limits for the full 2016 dataset calculated this way can be found in Table 7.4.

m_S [GeV]	Expected limit
10	0.77
20	0.81
50	1.07
100	1.82
200	6.55
300	14.00
500	45.31

m_S [GeV]	Expected limit
10	2.44
20	2.57
50	2.96
100	2.84
200	4.47
300	8.55
500	43.36

Table 7.4: Expected limits for 35.9 fb^{-1} for the scalar (left) and pseudoscalar (right) mediators.

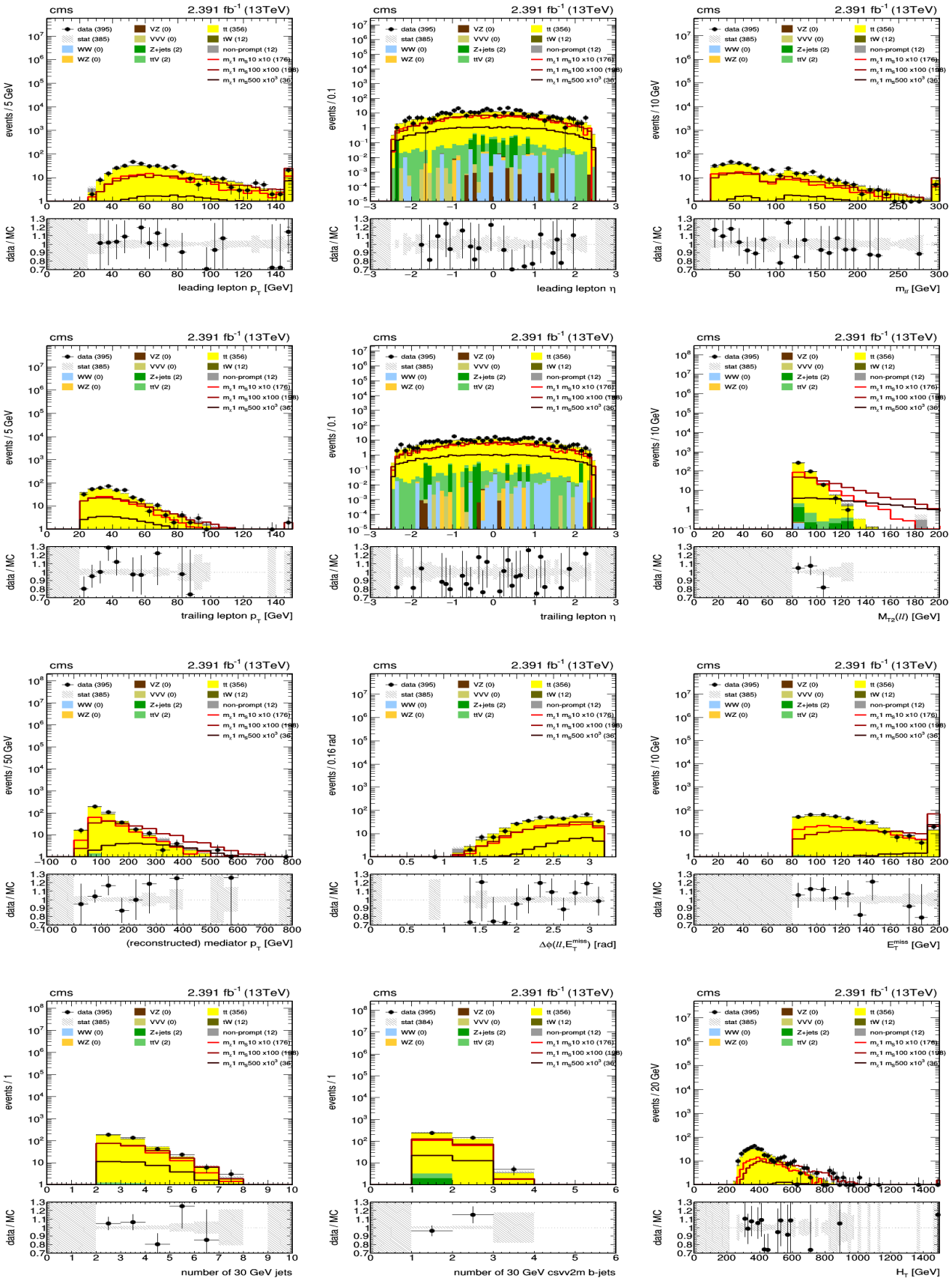


Figure 7.2: Final distributions for different variables at the final selection level of the analysis, for the data, the backgrounds and three different scalar mediator masses. All the signals have been rescaled by a factor given in the legend of the plots, and the errors are statistical only.

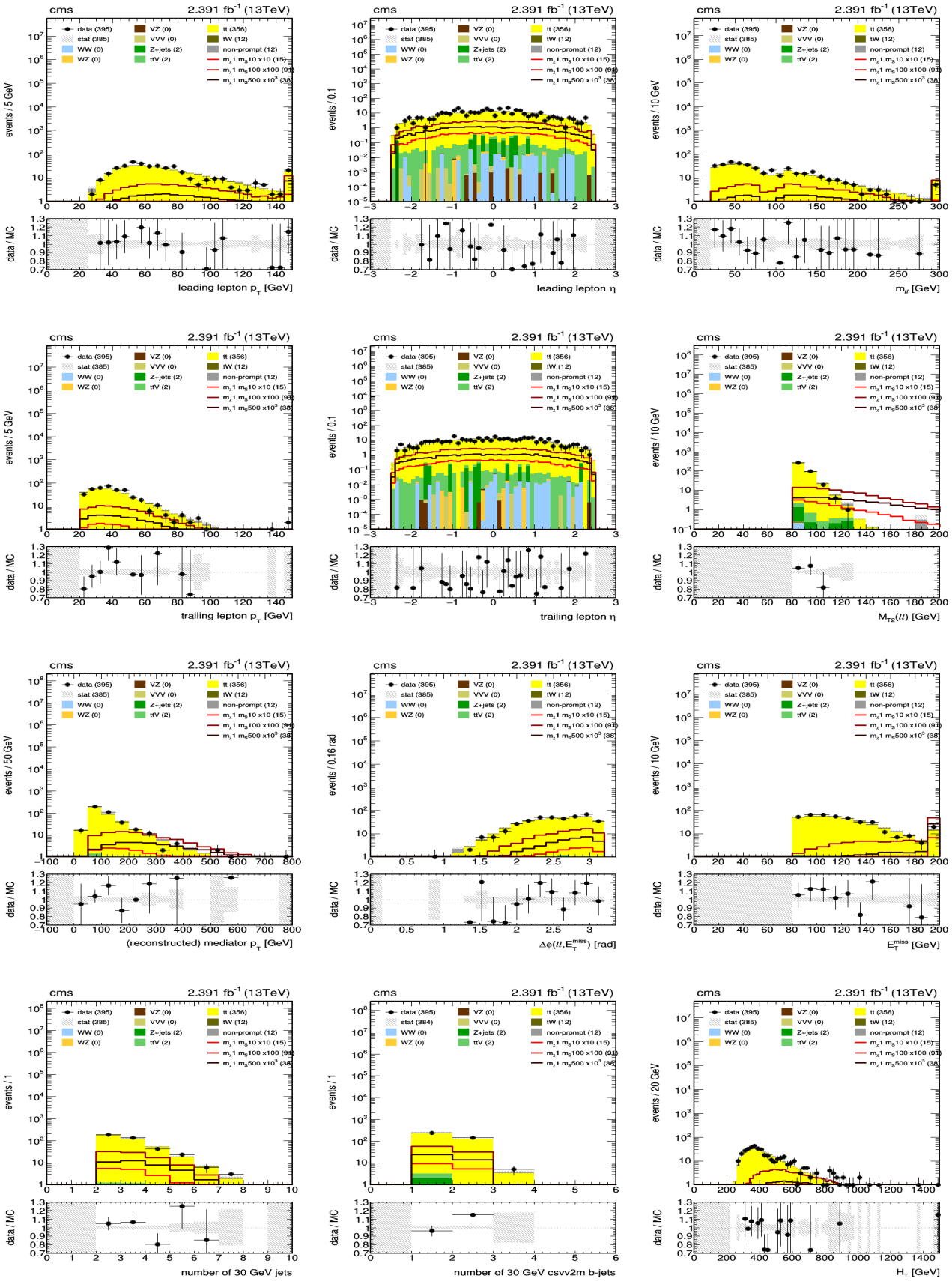


Figure 7.3: Final distributions for different variables at the final selection level of the analysis, for the data, the backgrounds and three different pseudoscalar mediator masses. All the signals have been rescaled by a factor given in the legend of the plots, and the errors are statistical only.

Chapter 8

Conclusions

A search for dark matter production is association with a pair of top quarks in the dilepton final state has been developed throughout this work. The analysis has been made using a 2.4 fb^{-1} dataset for the signal region and the complete 35.9 fb^{-1} dataset for the control regions, both taken by the CMS detector at CERN, at a center of mass energy of $\sqrt{s} = 13$ TeV.

After a general and a theoretical introduction about the process studied and the different channels of production of dark matter at the LHC, we studied the different objects, datasets and triggers used for the analysis. We also studied in detail the different backgrounds of the analysis and the ways we have to estimate them and check their validity. We then saw some generalities about the top reconstruction method we developed in order to be able to estimate the p_T of the mediator of the interaction studied, and the procedure deployed to correct for the low efficiency observed using a generic top reconstruction method. We also described the four variables able to introduce some discrimination between the major background $t\bar{t}$ and the signal, we detailed the cuts we apply to the events to get a sample as pure as possible in $t\bar{t}$ + DM signal and different control regions have been plotted to check for the validity of the different Monte Carlo simulations. We also explained some theory concerning the Multi-Variate Analysis we decided to apply to this analysis by developing different kinds of Artificial Neural Networks, in order to create a new variable allowing us to separate the backgrounds from the signal. Finally, we showed some distributions and yields obtained for the different processes at the final selection level, and we managed to obtain an upper limit on the cross section, considering different masses for both the scalar and pseudoscalar mediators.

As explained in Chapter 7, the results with the unblinded dataset do not allow to exclude any dark matter mediator production model so far. There are however several ways to improve this analysis in the near future. First of all, the increase in luminosity given by the unblinding of the complete 2016 dataset and the new collisions happening daily at the LHC are expected to lower the limits obtained by a factor proportional to the square root of the change in integrated luminosity, as we saw in Table 7.4. Moreover, a better understanding of the different backgrounds (especially the $t\bar{t}$, of course) and the different systematics is also a crucial point of the analysis, since studying them in detail and reducing them would also allow us to improve our results. Finally, moving to a shape study instead of a simple cut and count analysis on the MVA output is expected to improve a lot the final results and will be done as soon as possible.

Appendices

Appendix A

Data-driven methods

The Monte Carlo simulations might not describe properly all the effects that happen in a collider, since they are extremely complex to produce and depend on a lot of different theoretical parameters. Another way to study the different backgrounds is to calculate the yields we expect directly from the data we measured. This is a difficult task as well and has been done so far to characterize two different processes: the Drell-Yan and the non-prompt leptons. Each method of calculation will now be explained in detail.

A.1 Rin-out method for the Drell-Yan process

This method has been developed in order to find a way to calculate directly from the data a factor by which we can scale the Drell-Yan (DY) produced thanks to the usual MC simulations [55]. This is an important process because it has a huge cross section (as seen in Table 4.2) and because it is expected to contribute to the three different channels of the analysis (e^+e^-, $\mu^+\mu^-$ and $e^\pm\mu^\mp$ via tau decays), so we want to make sure that we understand it well enough. The main idea of this method consists in considering the number of DY events in MC inside (N^{in}_{DY}) and outside (N^{out}_{DY}) of the Z mass window (defined as the region where the reconstructed mass of the two leptons takes values between 76 and 106 GeV) with all the cuts of the final selection level of the Section 5.4 applied, except for the Z veto.

We can define the ratio $R_{out/in}$ as being the ratio between the number of DY events predicted outside and inside of the Z mass window by the simulations. This ratio can then be used to calculate the number of DY events expected outside of the Z window in data, from the number of events observed within this window, according to the following equation:

$$N^{out}_{DY,est} = N^{in}_{DY,data} \cdot \left(\frac{N^{out}_{DY,MC}}{N^{in}_{DY,MC}} \right) = N^{in}_{DY,data} \cdot R_{out/in} \tag{A.1}$$

It is important to note that the previous equation assumes that the Z window is dominated by DY events, which is not an evident assumption to make, since different backgrounds should be present in this window as well (peaking backgrounds such as ZZ and WZ and non peaking backgrounds that we actually do not consider in this calculation since they give a continuous distribution in the dilepton invariant mass). The number of yields within the peak of the Z can then be described with the equation A.2, where k_{ll} is a factor applied to opposite flavor final states to normalize for the relative efficiencies for electrons and muons.

$$N^{out}_{DY,est} = (N^{in}_{ll,data} - k_{ll} \cdot N^{in}_{e\mu,data} - N^{in}_{ZV,MC}) \cdot R_{out/in} \tag{A.2}$$

It is then possible to show that the scale factor for the channel ee takes the form A.3 (a similar equation can of course be derived in the $\mu\mu$ case).

$$SF_{ee} = \frac{n^{in}_{ee} - n^{in}_{WZ} - n^{in}_{ZZ} - k_{ee} \cdot n^{in}_{e\mu}}{n^{in}_{DY}} \text{ where } k_{ee} = \frac{1}{2} \cdot \sqrt{\frac{n^{obs}_{ee}}{n^{obs}_{\mu\mu}}} \tag{A.3}$$

The scale factors obtained for each channel and for different bins of m_{T2}^{ll} are shown in Figure A.1.

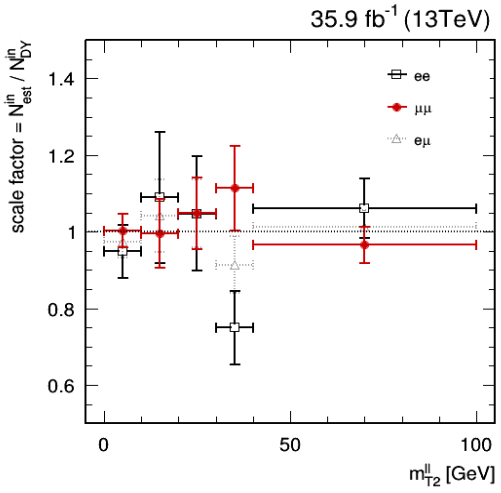


Figure A.1: Scale factor obtained from the Rin-out data-driven method for the Drell-Yan process, represented with respect to bins of m_{T2}^{ll}, for the three channels of our analysis (all the possible combinations of flavors of two leptons). The error bars represented are statistical only.

A.2 Fake and prompt rates

We can start this section by giving some useful definitions. First, a *prompt lepton* is defined as a real lepton, in the sense that the lepton is originating from the primary interaction vertex of a collision while on the other hand, a *non-prompt* or *fake lepton* is defined as a lepton falsely detected (usually, this kind of lepton comes from a mis-identified jet or from a heavy meson decay). The *fake rate* corresponds to the probability to see a fake lepton passing the final selection level of the analysis, giving rise to a new kind of background, especially at low lepton p_T. This new background source therefore needs to be understood quite well, the problem being that this misidentification rate can not be properly estimated by Monte Carlo simulations mainly because of its complexity. This is why this background is estimated using data-driven methods. In this work, the fake rate is estimated within a loose single lepton triggered sample and is calculated separately for electrons and muons. It is important to note at this point that even though this background is small in this analysis, the method presented here is a general method which can be used in most of the analyses, since it does not depend on the number of final-state particles or on the event selection applied.

The fundamental idea of the method is quite simple. We first want to define a control region as pure as possible, in which we can estimate the yields of the QCD backgrounds, featuring a lot of jets in the final state. Then, we just want to calculate directly from the data an extrapolation factor allowing us to go back to our signal region where we perform the analysis. The measurement of the fake rate is made by defining the so-called *fakeable objects*, which are lepton-like objects passing only the loose requirements, and the *fully selected objects*, which correspond to objects able to pass the complete selection of our analysis. The fake rate is then easily calculated as the ratio between the number of fully reconstructed leptons and the number of fakeable objects.

We have now seen how it is possible to calculate the fake rate, but we are not quite done yet since an important step is still missing. As previously stated, for this method to be useful, we have to define a QCD-enriched control region, but we also expect this control region to have some contamination, coming mainly from the leptonic decay of the eventual W or Z bosons that survived our selection. This kind of contamination is usually referred to as the electroweak contamination and is considered

in this work to be coming from the W+jets and Z+jets processes only. It is necessary to take into account this contamination because it biases the number of fully reconstructed leptons which appear in the fake rate definition, leading to a wrong fake rate calculation, especially at high p_T, where the contribution of real leptons is higher. This contamination has been removed in this analysis. The fake rates obtained using this method, with and without electroweak correction, are represented in Figure A.2, for both electrons and muons.

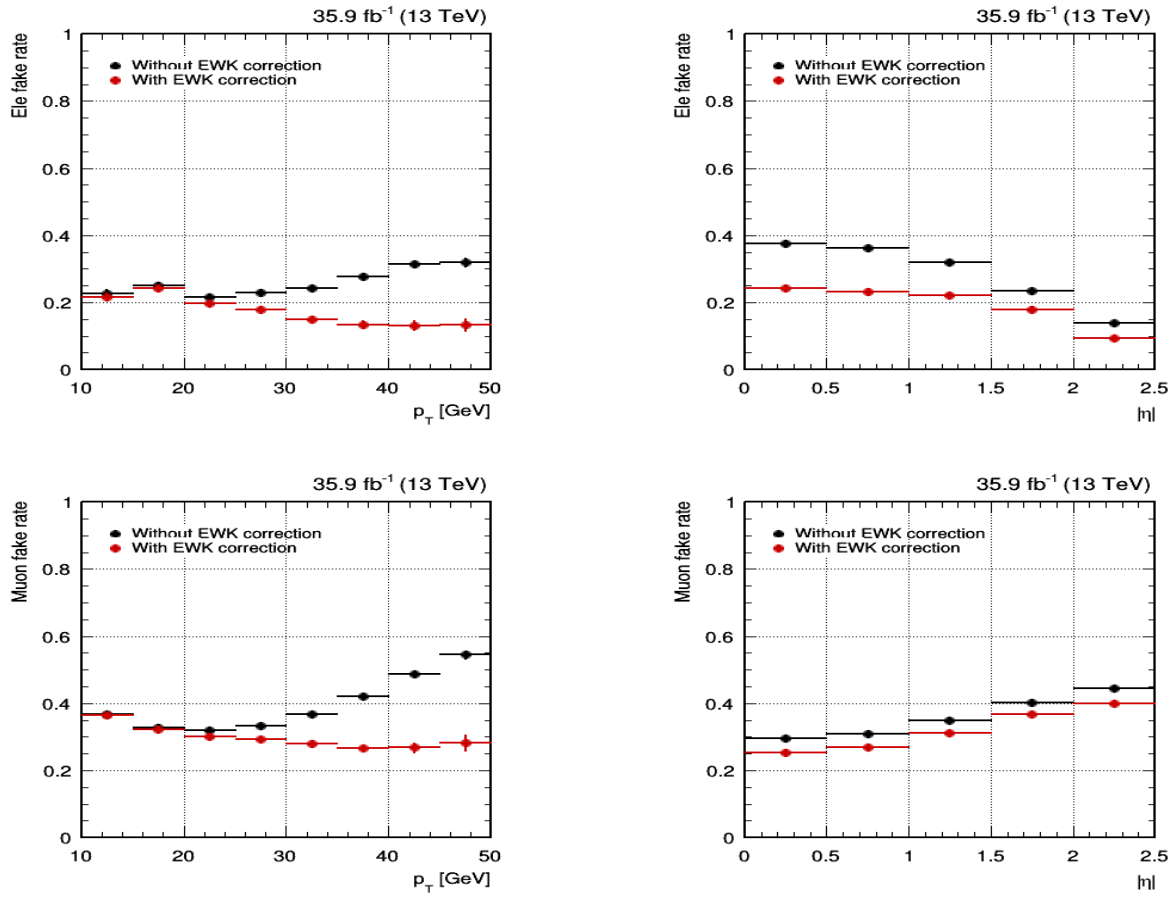


Figure A.2: Fake rates obtained for both electrons (top) and muons (bottom), with respect to the p_T (left) and η (right) with 35.9 fb^{-1} of integrated luminosity.

The fake rate is not the only important rate to calculate, since we also need to estimate the *prompt rate* to take into account the real lepton contamination in the control region we defined. This rate is also measured in data, using a general tag and probe method within a Z enriched control region. This method consists simply in reconstructing $Z \rightarrow ll$ events in this region, and to select all the events for which the first lepton can be characterized as tight. Then, we search for the second lepton coming from the Z decay within all the leptons detected, by calculating the reconstructed mass of all the possible combinations and selecting the one which is closer to the expected mass of the Z. If we find a pair of leptons giving a reconstructed mass close to the expected one, and if we know that one of the lepton is tight, we can conclude that the second lepton should be tight as well, and we can look if this is effectively the case or not. This allows us to calculate the electron and muon prompt rates, which are represented in Figure A.3.

The systematic associated to the estimation of this background is estimated to be a flat 30%, meaning that we do get some improvement with respect to the usual MC simulations which have a 50% uncertainty associated.

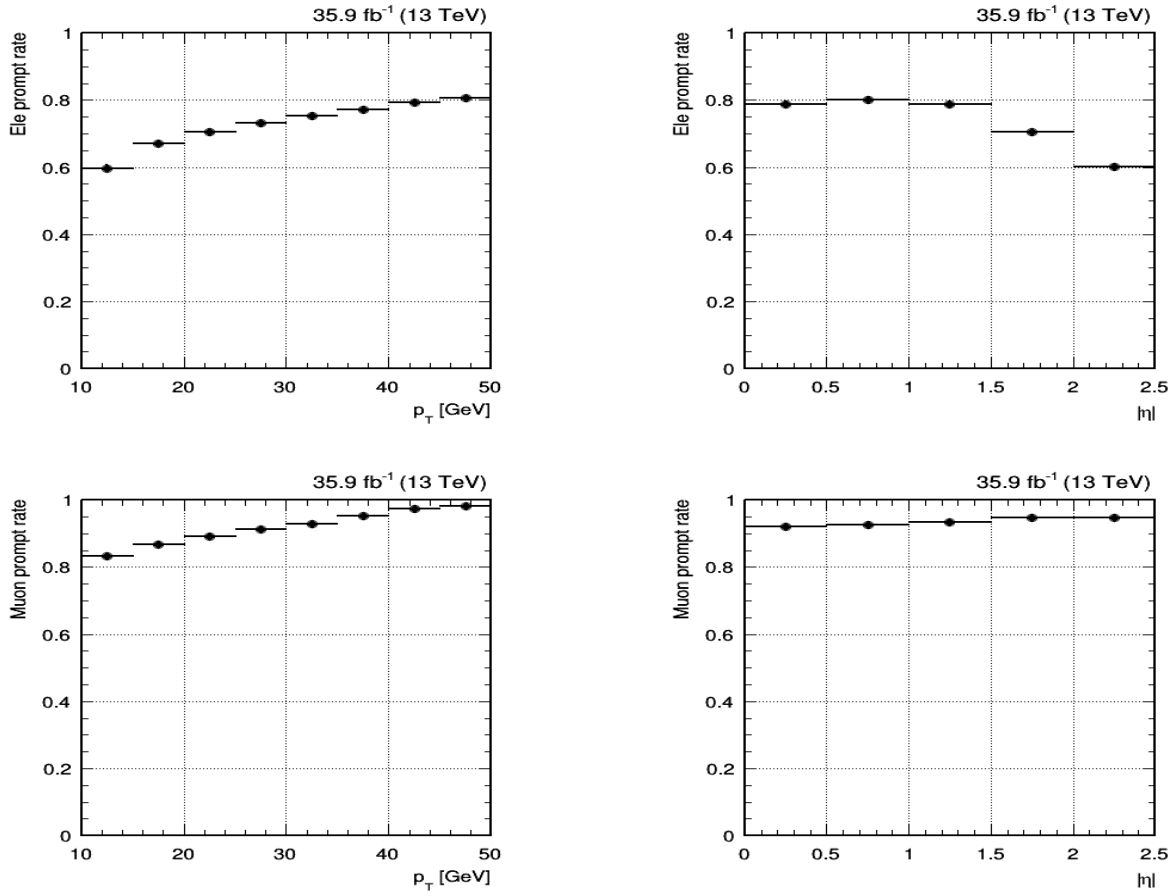


Figure A.3: Prompt rates obtained for both electrons (top) and muons (bottom), with respect to the p_T (left) and η (right) with 35.9 fb^{-1} of integrated luminosity.

Bibliography

[1] M. Born, P. Jordan & W. Heisenberg, *Zur Quantenmechanik. II.*, Physik, 1926
http://link.springer.com/article/10.1007%2FBF01379806

[2] M. Riordan, *The Discovery of Quarks*, SLAC-PUB-5724, April 1992
http://www-spires.slac.stanford.edu/cgi-wrap/getdoc/slac-pub-5724.pdf

[3] F. Englert & R. Brout, Phys. Rev. Lett. **13**, 321 (1964);
P.W. Higgs, Phys. Rev. Lett. **13**, 508 (1964) and Phys. Rev. **145**, 1156 (1966);
G.S. Guralnik, C.R. Hagen & T.W. Kibble, Phy. Rev. Lett. **13**, 585 (1964).

[4] CMS Collaboration, *Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*, Phys. Lett. B, September 2012 [arXiv:1207.7235]
http://www.sciencedirect.com/science/article/pii/S0370269312008581

[5] 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, September 2012 [arXiv:1207.7214]
http://www.sciencedirect.com/science/article/pii/S037026931200857X

[6] *Standard Model*, Wikipedia, as seen in May 2017
https://en.wikipedia.org/wiki/Standard_Model

[7] *Dark Matter*, CERN, as seen in May 2017
https://home.cern/about/physics/dark-matter

[8] I. Newton, *Newton's Demonstration that planets move in ellipses*, Philosophical Transactions of the Royal Society, No. 128, p. 698-705, September 1676
http://www.newtonproject.ox.ac.uk/view/texts/diplomatic/NATP00092

[9] F. Zwicky, *Die Rotverschiebung von extragalaktischen Nebeln*, Helvetica Physica Acta, Vol. 6, p. 110-127, 1933
http://articles.adsabs.harvard.edu/cgi-bin/nph-iarticle_query?1933AcHPh...6.
.110Z&data_type=PDF_HIGH&whole_paper=YES&type=PRINTER&filetype=
.pdf

[10] S. Van Den Bergh, *The early history of dark matter*, Dominion Astrophysical Observatory, June 1999
https://arxiv.org/pdf/astro-ph/9904251.pdf

[11] K.G. Begeman, A.H. Broeils & R.H. Sanders, *Extended rotation curves of spiral galaxies - Dark haloes and modified dynamics*, Monthly Notices of the Royal Astronomical Society (ISSN 0035-8711), April 1991
http://adsabs.harvard.edu/full/1991MNRAS.249..523B

[12] R. Scarpa, *Modified Newtonian Dynamics, an Introductory Review*, European Southern Observatory, September 2005
https://cds.cern.ch/record/923578/files/0601478.pdf

[13] *Galaxy Clusters Reveal New Dark Matter Insights*, NASA Jet Propulsion Laboratory, 2016
https://www.jpl.nasa.gov/news/news.php?feature=4829

[14] T. Treu & E. Koopmans, *Probing dark matter distribution with gravitational lensing and stellar dynamics*, California Institute of Technology, May 2002 [arXiv:0205335]
https://arxiv.org/abs/astro-ph/0205335

[15] K. Griest, *WIMPs and MACHOs*, Encyclopedia of astronomy and astrophysics, 2002
http://www.astro.caltech.edu/~george/ay20/eaa-wimps-machos.pdf

[16] M. Peskin, *Dark Matter and Particle Physics*, SLAC-PUB-12493, July 2007 [arXiv:0707.1536]
https://arxiv.org/pdf/0707.1536.pdf

[17] S. Yan Hoh, J. R. Komaragiri & W. A. T. Wan Abdullah, *Dark Matter Searches at the Large Hadron Collider*, SLAC-PUB-12493, 2016 [arXiv:1512.07376]
https://arxiv.org/pdf/1512.07376.pdf

[18] M. Buckley *Paper Explainer: Two is not always better than one: Single Top Quarks and Dark Matter*, PhysicsMatt
http://www.physicsmatt.com/blog/2017/1/21/paper-explainer-two-is-not-always-better-than-one

[19] D Φ collaboration, *A precision measurement of the mass of the top quark*, Nature 429, p. 638-642, June 2004
http://www.nature.com/nature/journal/v429/n6992/full/nature02589.html

[20] T. M. Liss & P. L. Tipton, *The Discovery of the Top Quark*, Ecole Polytechnique Federale de Lausanne, September 1997
http://lphe.epfl.ch/~mtran/seminaires/Cours_Master_Bordeaux/Articles/SciAmTop.pdf

[21] CMS Collaboration, *Search for dark matter in association with a top quark pair at* $\sqrt{s} = 13$ *TeV in the dilepton channel*, CMS-PAS-EXO-16-028, 2016.

[22] *Sidebar: 5 Goals for the LHC*, Scientific American, as seen in May 2017
https://www.scientificamerican.com/article/5-goals-for-the-lhc/

[23] *Particle Accelerators and Detectors*, Universe Review
https://universe-review.ca/R15-20-accelerators03.htm

[24] *LHC roadmap*
https://lhc-commissioning.web.cern.ch/lhc-commissioning/schedule/LHC%20schedule%20beyond%20LS1%20MTP%202015_Freddy_June2015.pdf, as seen in April 2017.

[25] *CMS*, CERN, as seen in May 2017.
https://home.cern/fr/about/experiments/cms

[26] *CMS detector design*, CERN
http://cms.web.cern.ch/news/cms-detector-design

[27] *Compact Muon Solenoid: Wikis*, The Full Wiki
http://www.thefullwiki.org/Compact_Muon_Solenoid, as seen in April 2017.

[28] CMS Collaboration , *Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus and MET*, Technical Report CMS-PAS-PFT-09-001, 2009
https://cds.cern.ch/record/1194487?ln=fr

[29] CMS Collaboration, *CMS Luminosity Measurements for the 2016 Data Taking Period*, CMS-PAS-LUM-17-001, 2017.

[30] A. Rao , *Blinding and unbliding analyses*, CERN news, June 2012
http://cms.web.cern.ch/news/blinding-and-unblinding-analyses, as seen in May 2017.

[31] *PdmV2016Analysis*, CERN twiki
https://twiki.cern.ch/twiki/bin/view/CMS/PdmV2016Analysis, as seen in May 2017.

[32] *BRIL Work Suite*, CMS service lumi
https://cms-service-lumi.web.cern.ch/cms-service-lumi/brilwsdoc.html, as seen in May 2017.

[33] P. Nason & P.Z. Skands, *Monte Carlo events generators*, Particle Data Group, September 2013
http://pdg.lbl.gov/2014/reviews/rpp2014-rev-mc-event-gen.pdf

[34] C. Oleari, *The POWHEG BOX*, Nuclear Physics B Proceedings Supplements 205, August 2010 [arXiv:1007.3893]
https://arxiv.org/pdf/1007.3893.pdf

[35] T. Sjöstrand, P. Skands & S. Prestel , *PYTHIA 8 Worksheet*, Lund University, January 2014
http://home.thep.lu.se/~torbjorn/pythia81html/Welcome.html

[36] *Geant4*, CERN, May 2017http://geant4.cern.ch

[37] D. Abercrombie et al. , *Dark Matter Benchmark Models for Early LHC Run-2 Searches: Report of the ATLAS/CMS Dark Matter Forum*, July 2015 [arXiv:1507.00966]
https://arxiv.org/pdf/1507.00966.pdf

[38] CERN Twiki , *Summary table of samples produced for the 1 Billion campaign, with 25ns bunch-crossing*
https://twiki.cern.ch/twiki/bin/viewauth/CMS/SummaryTable1G25ns#TTbar, as seen in April 2017.

[39] B. Yang & N. Liu, *One-loop QCD correction to top pair production in the littlest Higgs model with T-parity at the LHC*, EPL, 2015 [arXiv:1507.07104]
https://arxiv.org/abs/1507.07104

[40] J. Beringer et al., *Z mass*, Particle Data Group, PR D86, 2012
http://pdg.lbl.gov/2012/listings/rpp2012-list-z-boson.pdf

[41] M. Cacciari, G. P. Salam & G. Soyez, *The anti-kt jet clustering algorithm*, JHEP, 2008 [arXiv:0802.1189]
https://arxiv.org/abs/0802.1189

[42] CMS Collaboration, *Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS*, JINST 6, 2011 [arXiv:1107.4277]
http://cds.cern.ch/record/2138504

[43] CMS Collaboration, *Identification of b quark jets at the CMS Experiment in the LHC Run 2*, Technical Report CMS-PAS-BTV-15-001, CERN, 2016
https://arxiv.org/abs/1107.4277

[44] *MET Analysis*, CERN Twiki, June 2015
https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookMetAnalysis#Type_I_Correction, as seen in May 2017.

[45] JET-MET POG, *MET Filter Recommendations for Run II*, CERN Twiki
https://twiki.cern.ch/twiki/bin/viewauth/CMS/MissingETOptionalFiltersRun2, as seen in May 2017.

[46] C. Jones, *Using the M_{T2} and Φ^* variables to investigate Monte Carlo methods of $t\bar{t}$ production*, September 2015
http://www.desy.de/2011summerstudents/2015/reports/CaitlinJones.pdf, as seen in May 2017

[47] CMS Collaboration, *Searches for supersymmetry using the MT2 variable in hadronic events produced in pp collisions at 8 TeV*, May 2015 [arXiv:1502.04358]
https://arxiv.org/abs/1502.04358

[48] T. R. Junk, *Statistical Methods for Experimental Particle Physics*, July 2009
https://www-cdf.fnal.gov/~trj/tsi09/trjtsi_Day1.pdf

[49] P. K. Sinervo, *Definition and Treatment of Systematic Uncertainties in High Energy Physics and Astrophysics*, Toronto University, September 2003
http://hep.physics.utoronto.ca/~pekka/papers/systematicsreview.pdf

[50] CMS Collaboration, *Search for dark matter in association with a top quark pair at $\sqrt{s}$ = 13 TeV*, CMS PAS-EXO-16-005, August 2016
https://cds.cern.ch/record/2204933

[51] M. Jachowski, *Multivariate Analysis, TMVA, and Artificial Neural Networks*, Michigan REU Final Presentations, August 2006
https://indico.cern.ch/event/5007/contributions/1177811/attachments/962648/1366777/reu_presentation3.pdf

[52] M. A. Nielsen, *Using neural nets to recognize handwritten digits*, Determination Press, Chapter 1, 2015
http://neuralnetworksanddeeplearning.com/chap1.html

[53] *Neural Networks*, Stanford University, April 2013
http://ufldl.stanford.edu/wiki/index.php/Neural_Networks

[54] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, H. Voss & M. A. Nielsen *Toolkit for Multivariate Data Analysis with ROOT: Users Guide*, Stanford University, September 2013
http://tmva.sourceforge.net/docu/TMVAUsersGuide.pdf

[55] J. Brochero, *Top Anti-Top Production Cross Section Measurement at $\sqrt{s}$ = 8 TeV in the Dilepton Channel with the CMS Detector*, CERN-THESIS-2014-209, July 2014
http://digital.csic.es/bitstream/10261/141369/1/Tesisdetectorbrochero.pdf

Search for dark matter production in association with top quark pairs in the dilepton final state at $\sqrt{s} = 13$ TeV

Cédric Prieels

Director - Jónatan Piedra Gómez
Co-director - Alicia Calderón Tazón

Instituto de Física de Cantabria
Universidad de Cantabria

- June 26th, 2017 -

Outline

- Theoretical introduction
- Experimental device
- Objects, datasets, triggers and samples
- Event reconstruction and selection
- Multi-variate analysis and neural network
- Systematic uncertainties
- Limits of the analysis
- Conclusions

Introduction

In this work is presented a search for dark matter production in proton-proton collisions, using the data collected by the CMS detector at CERN during 2016. The data analyzed has been taken at a center of mass energy $\sqrt{s} = 13$ TeV.

Motivations for the analysis

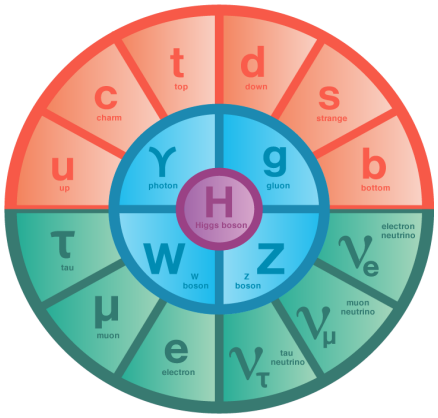
- There are several **astrophysical evidences** for the existence of dark matter, but **no direct nor indirect detection** so far.
- We hope **to be able to produce WIMPs** (Weakly Interactive Massive Particles) in high energy collisions of ordinary particles within the LHC.

Main objective

- Consider different dark matter production models to **search for dark matter**, and to eventually exclude some of these models or at least put upper limits on their cross section production.

Standard Model

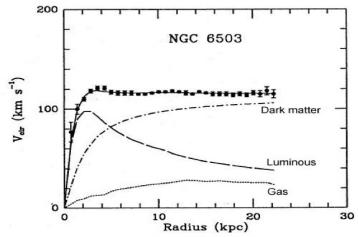
- The most accepted and used model to describe the elementary particles and the fundamental interactions between them is the **Standard Model**.
- This model is quite elegant and is working really well, and has successfully made different predictions, such as the existence of the Higgs boson, finally discovered in 2012.



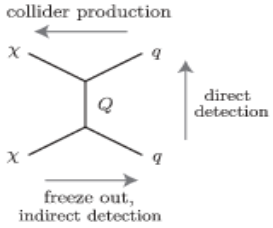
→ However, this model is known to have **several shortcomings** which require further investigation. For example, eventual exotic particles which do not fit within this model (such as dark matter) are extensively searched for nowadays.

Dark matter

The dark matter hypothesis was introduced as a way to explain the strange behavior of the **rotation curves** of the galaxies at high radius, and the apparent **missing mass** in the Universe.



K.G. Begeman, A.H. Broels, R.H. Sanders. 1991. Mon.Not.RAS 249, 523.



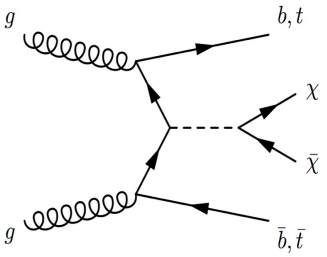
Ordinary baryonic matter constitutes only 5% of the total mass of the Universe, while the dark matter accounts for 27%.

We also assume that dark matter is **made out of cold particles** and interact only weakly with ordinary matter or itself, making it **extremely difficult to detect**.

$\rightarrow$ Since we do not know any baryonic candidate fulfilling these conditions, we can postulate the **existence of new elementary particles**, the WIMPs. This is typically the kind of particles we **hope to find at the LHC**.

Our channel of interest

This work will be focused on **dark matter production in association with a top and an anti-top quarks**, in the dilepton final state. The dark matter χ is produced through the interaction of two gluons and through the apparition of a **mediator** Φ for which we will consider **different masses and couplings** (scalar and pseudoscalar).

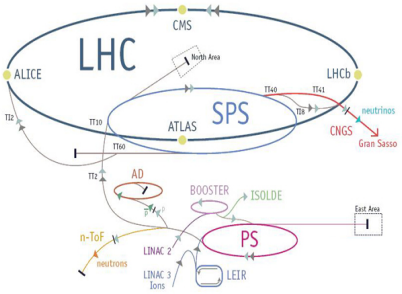


We are looking at this channel mainly because...

- The **signature left by the top quarks** can be isolated well with respect to the other backgrounds' leftovers.
- If we assume that the dark matter mediator couples in the same way than the Higgs does, it should have **stronger couplings with heavier particles** (and the top is the most massive fermion known).

The dilepton final state is interesting because, even though this is the channel with the **smallest branching ratio** (fraction of times for which a particle decays by an individual decay mode with respect to the total number of decays), this is also the channel having the **least number of backgrounds**.

The Large Hardon Collider



The data has been taken by the Large Hardron Collider, a **circular underground proton-proton collider**, situated at CERN. With its 27 kilometers of circumference, it is currently the **most powerful accelerator** in the world.

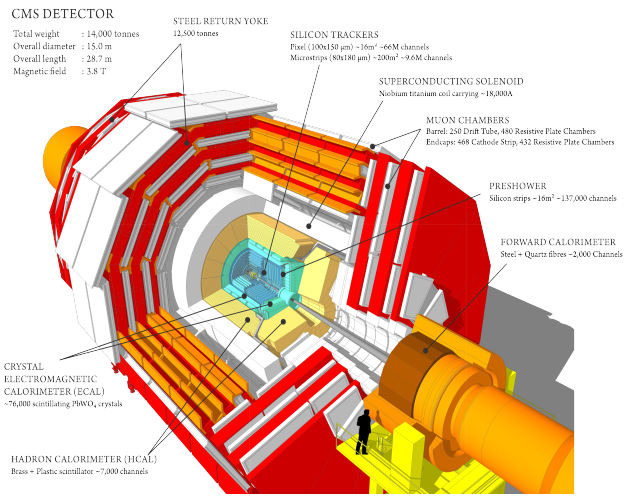
It is the result of a collaboration of 22 countries, and has been built in order to study and reproduce the Universe at its origin.

The LHC in a nutshell

- Beams of protons circulate in opposite directions at velocities close to the speed of light and collide in 4 different points, where the detectors are.
- Accelerates 2808 bunches of 10^{11} protons up to a center of mass energy of 13 TeV so far (but has been designed to reach 14 TeV).
- Can reach an instantaneous luminosity of 10^{34} $cm^{-2}s^{-1}$ and a collision rate of 40 MHz.

The Compact Muon Solenoid

CMS is one of the two **polyvalent detectors** of the LHC, designed to make measurements in most of the major different fields of particle physics, from precision measurements to the search of new exotic processes.



CMS in a nutshell

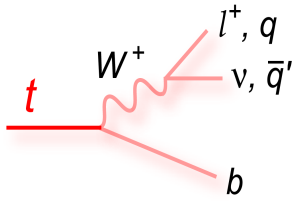
- Is a **compact** detector (its 14.000 tons are compacted in a relatively small volume).
- Has a **powerful tracker and muon detection system** allowing us to measure some of the properties of the leptons throughout a large range of energies.
- Has a **huge solenoid** as central piece able to produce a 3.8T magnetic field, to curve the particles and study their properties.
- Is made out of **different layers** (such as the tracker, the calorimeters and the muon chambers), each having its own purpose.

What are we looking for?

For an event to be considered interesting, we basically require the presence of exactly **two energetic leptons**, at least **two b-jets** and some **missing transverse energy** (corresponding to the unbalanced transverse momentum of the collision).

Two leptons and two b-jets

We are looking for dark matter produced with a top and an anti-top, which are not stable and decay immediately to a $W^{\pm}$ and a bottom quark. We are then searching for the results of the **hadronisation of the bottom** quarks, and for **two leptons coming from the $W^{\pm}$**.



Missing transverse energy

Moreover, since dark matter particles do not interact with ordinary matter, we do not expect this kind of particles to **interact at all with the detector**. We are then also looking for a large amount of **missing transverse energy**.

Datasets and triggers

The datasets...

- Have been recorded during 7 different periods and have been selected to match the current **blinding policy** (we look at the already published 2.4 fb^{-1} for the signal region, and at the full 35.9 fb^{-1} for the control regions).
- This blinding policy consists in optimizing the analysis considering a **limited dataset** only, to avoid any **conscious or unconscious bias** when one tries to optimize his analysis based on what has already been seen.

Era	Luminosity (fb^{-1})
Run2016B	5.748
Run2016C	2.573
Run2016D	4.248
Run2016E	4.009
Run2016F	3.102
Run2016G	7.540
Run2016H	8.606
Total	35.827

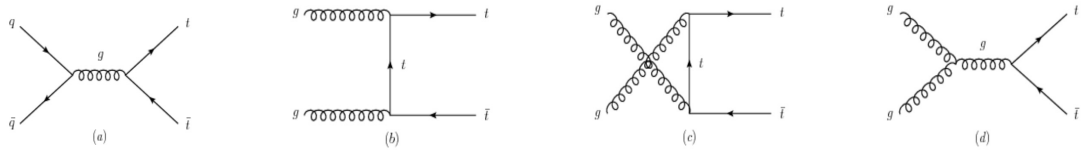
The triggers...

- Are used to **store only interesting events** out of the 40 MHz of collisions produced.
- Have been chosen and combined to only select events with at least 1 or 2 leptons while trying to maximize the efficiency of selection.

Background processes

Different backgrounds have similar final state, distributions and/or signatures than the ones expected for our signal. Here are the most important ones.

- $t\bar{t} \rightarrow$ is our main background and has kinematics close to the expected ones for our signal. The biggest challenge of the analysis consists in finding ways to reduce this background while leaving the signal as it is.



- $tW^{\pm} \rightarrow$ having a lower cross section and different kinematics, this background has a lower impact on the final results, even though it features a two b-jets and missing transverse energy in the final state as well.

- ttV (ttW/ttZ) $\rightarrow$ having two top quarks in the final state, some of this process is also expected to survive the selection of the analysis and to have an impact on the final results.

- Other backgrounds $\rightarrow$ such as the Drell-Yan ($Z/\gamma*$) or the non-prompt background also have different kinematics and only a limited impact on the final results.

Monte Carlo samples

- The different backgrounds have either been simulated with **Monte Carlo simulations** and checked in **control regions** (regions enriched in a single background) or measured using **data-driven methods** (estimated directly from the data).

- The dark matter samples have also been generated using Monte Carlo simulations. Different samples have been produced, for different **dark matter mediator masses** (10, 20, 50, 100, 200, 300 and 500 GeV), considering both the **scalar and pseudoscalar mediator couplings**.

MC samples used

Background	σ [pb]
$t\bar{t} \rightarrow$ ll $\nu\nu$	87.31
Single top	35.6
Single antitop	35.6
$t\bar{t}$W $\rightarrow$ l ν	0.2043
$t\bar{t}$W $\rightarrow q\bar{q}$	0.4062
$t\bar{t}$Z $\rightarrow$ ll $\nu\nu$	0.2529
$t\bar{t}$Z $\rightarrow q\bar{q}$	0.5297
Z $\rightarrow$ ll (low m_{ll})	18610
Z $\rightarrow$ ll (high m_{ll})	6025.2
W$\gamma \rightarrow$ l ν γ	586
Z$\gamma \rightarrow$ ll γ	131.3

Scalar mediators

m_S [GeV]	σ [pb]
10	19.950
20	10.480
50	2.941
100	0.672
200	0.093
300	0.029
500	0.005

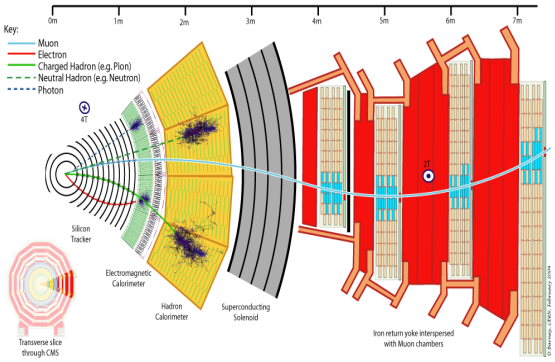
Pseudoscalar mediators

m_P [GeV]	σ [pb]
10	0.441
20	0.399
50	0.303
100	0.191
200	0.084
300	0.040
500	0.005

Event reconstruction

Reconstruction of the objects

A particular algorithm, the Particle Flow, is usually used in order to **gather the information** coming from the different parts of the detector and to **reconstruct the tracks** of the different particles.



The geometry of the detector is quite simple. Usually, the z-axis is defined as the axis followed by the beams, meaning that the O_{xy} plane (also called Φ plane) corresponds to the transverse plane of the detector.

The pseudorapidity η is a Lorentz invariant quantity defined as $-ln\left(tan(\frac{\theta}{2})\right)$ where θ is the angle between the z- and x-axes.

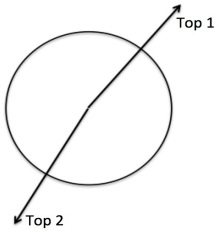
In this analysis, only **isolated tight leptons** (complying with a lot of different requirements, resulting in better leptons) having a p_T larger than 10 GeV and $|\eta|$ smaller than 2.5 (2.4) for electrons (muons) are considered. Jets are required have a p_T larger than 30 GeV and $|\eta|$ smaller than 4.

Top reconstruction

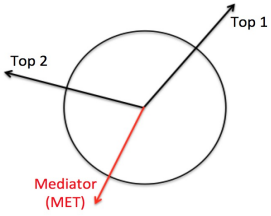
General idea

The top reconstruction is a general method developed to calculate the p_T of the tops in any $t\bar{t}$-like event. We extended this method so that it is also **able to measure the p_T of the mediator** produced in the dark matter case.

Having a way to determine the p_T of the mediator is interesting, because this variable is expected to **give us a good discrimination** between the $t\bar{t}$ and our signal.



In the usual $t\bar{t}$, both the tops are expected to leave the primary vertex back-to-back.



In the $t\bar{t}$ + DM case, the mediator leaves the primary vertex with a p_T equal to the difference in p_T of the tops.

An event passing the usual top reconstruction gets a value of dark p_T exactly equal to 0, while an event passing moreover our improved reconstruction (and therefore more likely to be a signal event) gets a value higher than 0.

Event preselection

We want to be able to reduce as much as we possible can the background, while keeping as much signal as possible by applying several cuts on different variables. This is first done by cutting in several general variables.

Preselection of the analysis

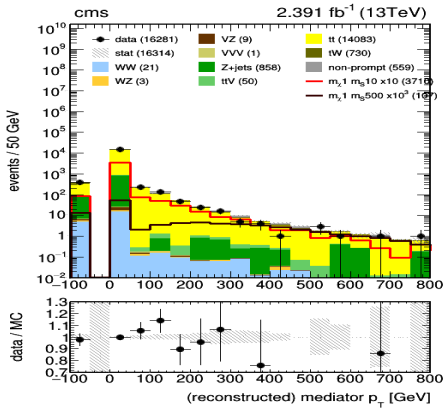
Number	Cut level	Cut	Comment
0	Preselection	$p_T^{lep_1} > 25$ GeV $p_T^{lep_2} > 20$ GeV $p_T^{lep_3} < 10$ GeV $q_l^{lep_1} \cdot q_l^{lep_2} < 0$	On the leading lepton On the trailing lepton Third lepton veto Opposite charge requirement
1	First level	$m_{ll} > 20$ GeV $\lvert m_{ll} - m_Z \rvert > 15$ GeV $n_{jet} \geq 2$ $n_{b_{jet}} \geq 1$	Only applied to ee and $\mu\mu$ channels

All these cuts allow us to reduce strongly different backgrounds (such as the Drell-Yan). However, we still need to find other variables able to remove some of the $t\bar{t}$, the most problematic background.

Discriminating variables

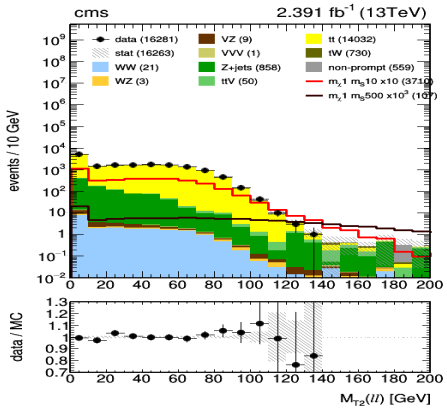
In the following plots have been represented the distributions of the number of event for four different variables. These plots have been obtained by applying the cuts corresponding to the preselection of the analysis, and both the 10 (red line) and 500 GeV (brown line) scalar signals have been represented (and **rescaled** by a factor 10 and 1000, respectively).

Dark p_T



Reconstructed mediator p_T, obtained by applying our top reconstruction method

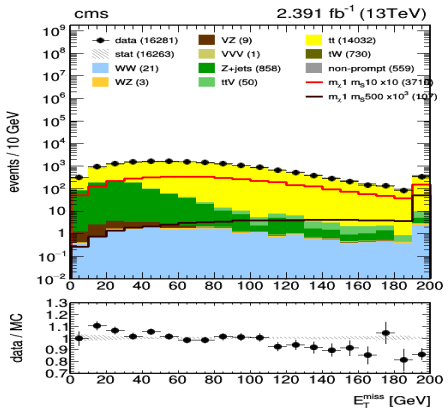
m_{T2}^{ll}



Transverse mass of the pair of leptons produced (cf. backup)

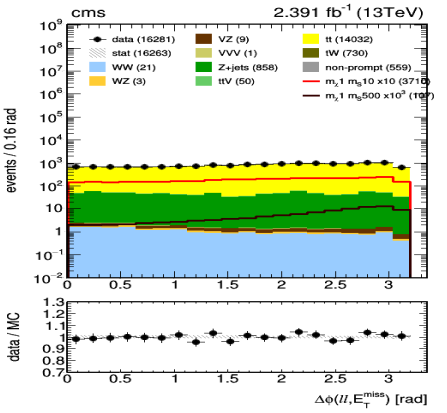
Discriminating variables II

E_T^{miss}



Missing transverse energy, equivalent to the imbalance of vector momentum in the O_{xy} plane

$\Delta\Phi_{ll,E_T^{miss}}$



Angle between the two leptons coming from the top quarks and the E_T^{miss} in the O_{xy} plane

Our complete analysis actually lays on the discriminating power of four different variables that will be used as input for our different neural networks, as we will see later.

Final event selection

We can then use the previous discriminating variables to introduce a third level of selection, to gain even more in purity.

Complete selection of the analysis

Number	Cut level	Cut	Comment
0	Preselection	$p_T^{lep_1} > 25$ GeV $p_T^{lep_2} > 20$ GeV $p_T^{lep_3} < 10$ GeV $q_l^{lep_1} \cdot q_l^{lep_2} < 0$	On the leading lepton On the trailing lepton Third lepton veto Opposite charge requirement
1	First level	$m_{ll} > 20$ GeV $\lvert m_{ll} - m_Z \rvert > 15$ GeV $n_{jet} \geq 2$ $n_{b_{jet}} \geq 1$	Only applied to ee and $\mu\mu$ channels
2	Second level	$E_T^{miss} > 80$ GeV $m_{T2}^{ll} > 80$ GeV Dark $p_T > 0$ GeV	Our top reconstruction works

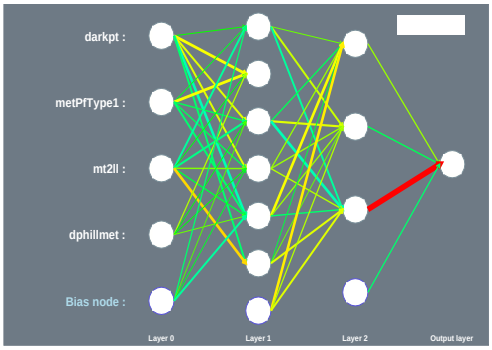
Multi-Variate Analysis

Performing a general **cut and count analysis** (applying several cuts on different variables to check for the yields obtained) is not optimal when looking for dark matter, since the **significance** of the signal is expected to be really low.

We then need to come up with solutions to **extract as much information as we can** from the data. This is why we decided to perform a **Multi-Variate Analysis**, by using general machine learning techniques and neural networks.

The main idea of the MVA consists in using a **neural network** to combine the information coming from the four previous variables into a single output.

This should give a way to classify any single event as either background or signal, after training correctly the networks created.



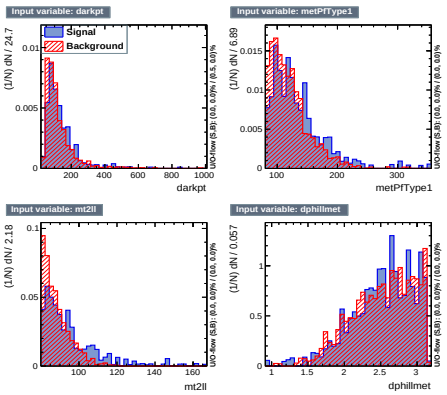
Artificial Neural Network

Our MVA is performed by defining several different **neural networks**.

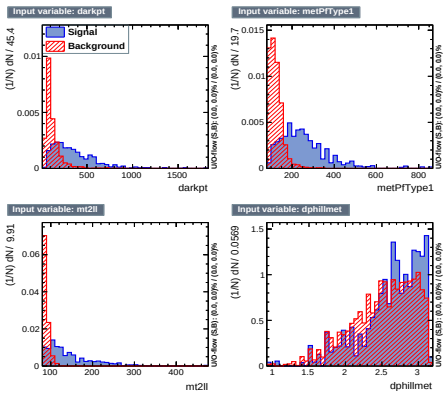
- Each network created contains **two hidden layers** made out of respectively 6 and 3 neurons. We considered two different kind of neurons, having different **activation functions**.
- We train a **different network for each mediator mass** point considered, and for both the scalar and pseudoscalar mediators. The neural networks are trained using 200 events and are trained for the moment only against the $t\bar{t}$.

Input distributions

10 GeV scalar mediator



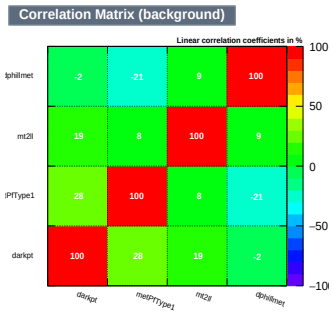
500 GeV scalar mediator

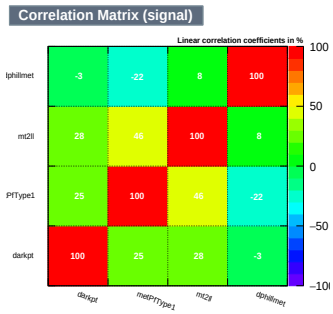


Artificial Neural Network II

It is also important to check the presence of eventual **correlations** between the input variables, to make sure not to complicate the networks without valid reason.

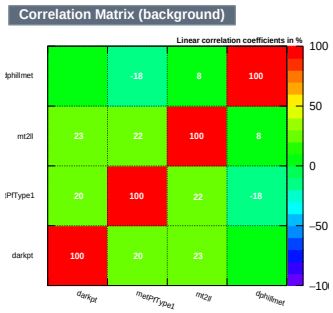
$m_\Phi = 10$ GeV

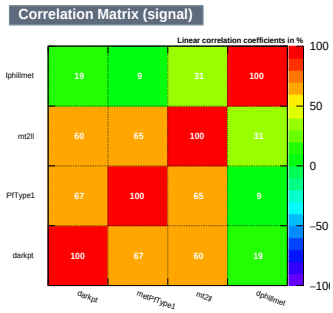




The strongest correlation is obtained for the signal, between the m_{T2}^{ll} and the E_T^{miss} variables (46 %).

$m_\Phi = 500$ GeV

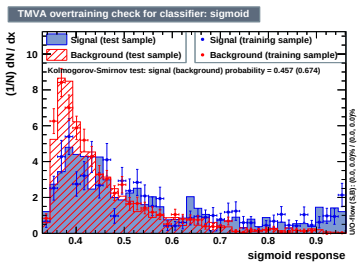
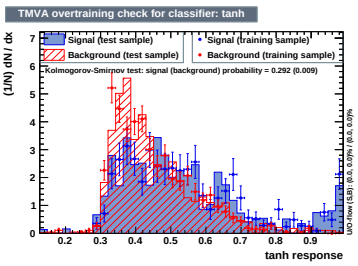
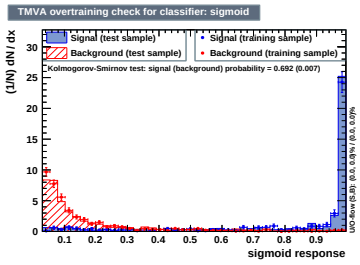
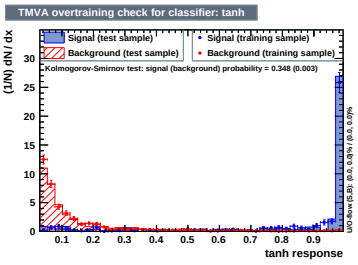




The strongest correlations are obtained for the signal, between the m_{T2}^{ll}, the E_T^{miss} and the dark p_T variables.

Artificial Neural Network III

Finally we can study the **actual output** of the different networks considered so far which is, at least for the moment, used in a simple count and count analysis.

	Activation function 1	Activation function 2
$m_\Phi = 10$ GeV		
$m_\Phi = 500$ GeV		

We clearly observe a **better discrimination** of the output in the 500 GeV case, and the Kolmogorov-Smirnov test indicates that the first activation function output presents less overtraining.

Systematic uncertainties

Several sources of **systematic uncertainties** have been considered, and can be classified into two different categories. The estimation of these systematics is a crucial point of the analysis since we need to get a precise idea about the error we are committing if we want to be able to make some conclusions about the eventual existence of dark matter.

- The **theoretical uncertainties**, such as the Parton Density Functions normalization ($< 5\%$) and shape ($< 3\%$), and the high-order corrections of the MC simulations uncertainties.
- The **experimental uncertainties** are usually coming from the detector itself. In this category are grouped several different uncertainties, such as:
  - Determination of the luminosity
  - Lepton RECO and ID normalization ($< 2\%$) and shape ($< 1\%$)
  - B-tagging efficiency normalization ($< 7\%$) and shape ($< 7\%$)
  - Jet energy scale normalization ($< 4\%$) and shape ($< 10\%$)
  - E_T^{miss} modeling and resolution
  - Non-prompt (30%), $t\bar{t}$ (15%) and Drell-Yan normalizations (7%)

The other backgrounds estimated directly from Monte Carlo get an usual 50% systematic uncertainty associated. All of the previously mentioned systematics have of course been taken into account when plotting the limits of the analysis (and for most of them, we did considered both their normalization and shape).

Yields table

We can now represent the yields obtained for the different backgrounds with the associated systematics, for the different background processes and different dark matter mediator mass points (and therefore, different neural networks).

10 GeV scalar (ANN output > 0.75)

Process	Yields	Statistical error	Systematic error
$t\bar{t}$	11.08	0.18	1.72
Non-prompt	1.30	0.92	0.39
ttV	0.75	0.03	0.19
Single top	0.48	0.07	0.16
Drell-Yan	0.46	0.35	0.34
WW	0.02	0.02	0.02
Total background	14.09	1.00	1.81
Signal	3.21	0.18	2.28
Total with signal	17.30	1.02	2.91
Data	10	-	-

500 GeV scalar (ANN output > 0.98)

Process	Yields	Statistical error	Systematic error
$t\bar{t}$	1.82	0.14	0.28
Non-prompt	0.00	0.00	0.00
ttV	0.57	0.03	0.15
Single top	0.12	0.06	0.04
Drell-Yan	0.00	0.00	0.00
WW	0.01	0.01	0.01
Total background	2.52	0.16	0.32
Signal	0.03	0.01	0.01
Total with signal	2.55	0.16	0.32
Data	2	-	-

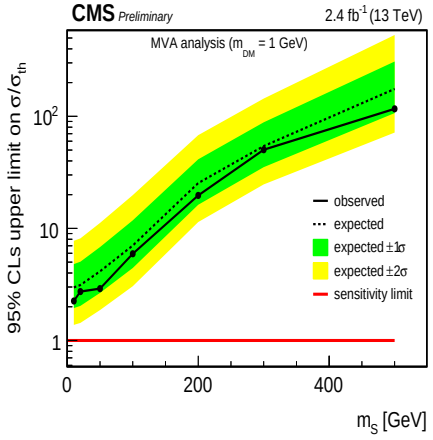
$\rightarrow$ The yields tables for the pseudoscalar mediator can be found in the backup.

Scalar limits

We can now represent the **limits** of the analysis, for both the scalar and pseudoscalar mediators.

The limit plots consist in plotting the value of the upper limit on the cross-section for which we expect to get some sensitivity, divided by the theoretical cross-section. This way, obtaining a value smaller than 1 for a given mass point then means that with the luminosity considered, we would **expect to get some sensitivity** to the signal.

Scalar mediator

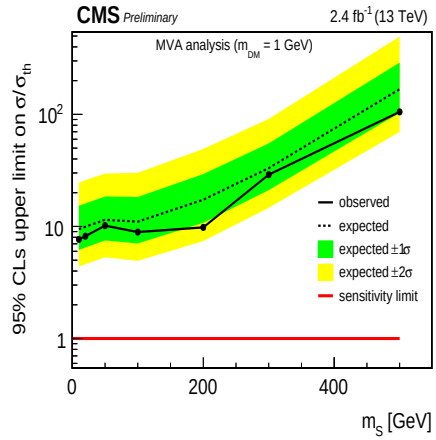


m_S [GeV]	Best AAN cut	Expected limit	Observed limit
10	0.75	2.99	2.22
20	0.80	3.12	2.73
50	0.90	4.14	2.90
100	0.90	7.03	5.94
200	0.90	25.37	19.69
300	0.80	54.25	50.34
500	0.98	175.50	116.65

The green and yellow lines on the plot correspond to the 1 and 2σ confidence intervals for the value of the expected limits.

Pseudoscalar limits

Pseudoscalar mediator



m_P [GeV]	Best ANN cut	Expected limit	Observed limit
10	0.80	9.47	7.66
20	0.85	9.94	8.18
50	0.85	11.47	10.18
100	0.90	11.03	8.91
200	0.95	17.31	9.81
300	0.80	33.12	28.98
500	0.98	167.94	105.27

With our analysis and considering for the moment the 2.4 fb^{-1} blinded dataset, we **cannot exclude** any dark matter production model.

However, we expect to be **able to exclude the 10 and 20 GeV scalar mediators** once considering the complete 2016 dataset, since we know that our sensitivity scales with the change in luminosity.

Conclusions

A search for dark matter production in association with top quark pairs in the dilepton final state has been performed. The analysis has been made using a 2.4 fb^{-1} blinded dataset for the signal region, and the complete 35.9 fb^{-1} dataset for the control regions, both taken by the CMS detector at CERN, at a center of mass energy of $\sqrt{s} = 13$ TeV.

The results obtained **do not allow us to exclude any dark matter production model** so far, with the limited luminosity considered, but we do **expect to be able to exclude the 10 and 20 GeV scalar mediators** when unblinding the analysis.

We did think about several ways we have to improve the analysis. First of all, a **better understanding** of the different **backgrounds** (mainly the $t\bar{t}$) and the different **systematics** is a crucial point since studying them in more detail and reducing them would allow us to improve our results. Moving to a complete shape study of the output of the MVA instead of performing a simple cut and count analysis is also expected to improve the final results, and will be done as soon as possible.

Thank you
for your attention!

Any questions?

Top reconstruction addendum

We can get 6 different equations from the kinematics expected for our signal.

$$\begin{cases} M(b_1 + W_1) = M_t \\ M(b_2 + W_2) = M_t \end{cases}$$

...because both the tops produced decay to a bottom and a W boson.

$$\begin{cases} M(\nu_1 + l_1) = M_W \\ M(\nu_2 + l_2) = M_W \end{cases}$$

...because we only consider the leptonic decay of the W in this analysis.

$$\begin{cases} \nu_{1_x} + \nu_{2_x} = (E_T^{miss})_x \\ \nu_{1_y} + \nu_{2_y} = (E_T^{miss})_y \end{cases}$$

...at least if we assume that the E_T^{miss} is coming from the neutrinos only.

$\rightarrow$ We have 6 equations and 6 unknowns (the three components of the momenta of the two neutrinos), meaning that this is a problem which can be solved. However, we have to find a way to correct the efficiency drop appearing when considering the signal, since the last hypothesis is not correct in this case (the E_T^{miss} is not coming only from the neutrinos).

This variable is expected to give some discrimination between the usual $t\bar{t}$ and our signal because in the dark matter case, we expect the p_T of the tops to be higher, since they will be accommodating for the extra missing transverse energy which appears (at least, if we manage to correct the efficiency drop issue since in this latest case, the previous equations are actually not exactly true).

E_T^{miss} variable definition

The E_T^{miss}, the missing transverse energy, corresponds to the imbalance of vector momentum in the plane perpendicular to the beam direction. It is defined in the following way:

$$E_T^{\text{miss}} = -\sum_i \vec{p}_T(i)$$

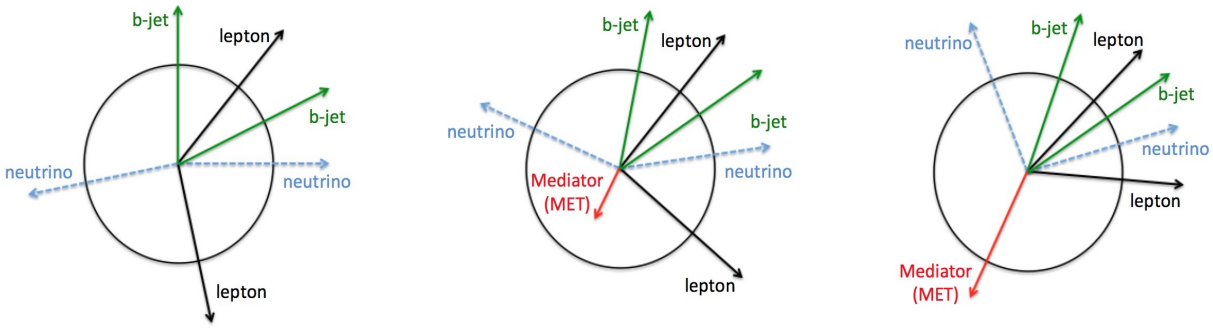
We have to consider the missing transverse energy and not directly the missing energy simply because the LHC is a proton-proton collider, and because each proton has three quarks (and the totat energy is distributed between these quarks in unknown proportions). This means that, for any given collision, we cannot actually know the exact initial longitudinal energy of collision. However, we do know that the initial transverse energy of the collision is equal to 0, so we can only calculate differences in transverse energy.

We expect this variable to give some discrimination between the $t\bar{t}$ and our signal because we know that the eventual dark matter particles produced should escape the detector while being undetected, since they almost do not interact with ordinary matter. The only way to detect the production of this kind of particle is therefore to look at the missing transverse energy.

$\Delta\Phi_{ll,E_T^{miss}}$ variable definition

This variable corresponds to the angle in the Φ plane, between the two leptons coming from the tops and the missing transverse energy.

$t\bar{t}$ background	$t\bar{t}$ + DM (low m_Φ)	$t\bar{t}$ + DM (high m_Φ)



We expect that the tops will be much closer to each other when the mass of the mediator produces raises, so this variable is expected to give some separation between the $t\bar{t}$ and the signal we are looking for, at least for high mediator masses.

m_{T2}^{ll} variable definition

The variable m_{T2}^{ll} has been introduced to measure the mass of a pair of particle produced in the particular case where both these particles decay into a final state including an undetected particle (neutrinos in this case).

The problem is that in this case we can only know the total amount of E_T^{miss} of the event considered and there is no way to know the exact contribution to this value given by each neutrino separately. We then need to calculate the transverse mass for the two pairs of particles, for different repartitions of E_T^{miss}. The repartition giving rise to the smallest possible mass is kept as the value of m_{T2}^{ll}.

$$\begin{cases} \left(M_T^{(i)}\right)^2 = \left(m_{vis}^{(i)}\right)^2 + m_\chi^2 + 2 \cdot \left(E_T^{vis(i)} E_T^{\chi(i)} - \vec{p}_T^{\,vis(i)} \cdot \vec{p}_T^{\,\chi(i)}\right) \\ m_{T2}\left(m_\chi\right) = \min_{\sum_i \vec{p}_T^{\,\chi(i)}} \left[\max\left(M_T^{(1)}, M_T^{(2)}\right)\right] \end{cases}$$

We expect this variable to give some discrimination between $t\bar{t}$ and our signal because we know that the transverse mass obtained this way for the W has to take values less or equal than the actual mass of the W. However, the presence of additional E_T^{miss} coming for the dark matter can break this condition.

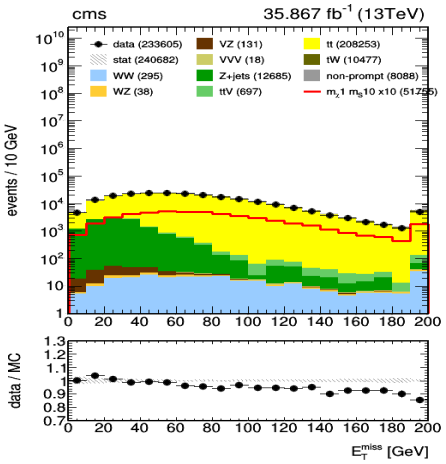
$t\bar{t}$ control region

A **control region** is usually defined by several cuts in order to check for the validity of a process simulated by Monte Carlo. The control region then has to be defined in such a way that this region is **enriched in the process we are interested in**.

This is why we created a control region for the $t\bar{t}$, our most important background. This region has been defined by removing the E_T^{miss} and by reversing the cut in m_{T2}^{ll}. As always, the 10 GeV signal is represented with the red line and has been rescaled by a factor 10.

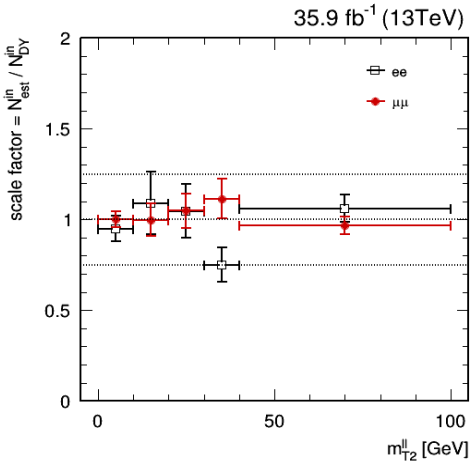
The main objective of this control region is to calculate a **scale factor**, a factor by which we need to scale the considered process to get a perfect agreement between data and Monte Carlo.

In this case, we obtain a scale factor for the $t\bar{t}$ of 0.97 ± 0.01, the error being statistical only.



Rin-out method

To estimate the Drell-Yan (DY), we calculated its importance using a data-driven method. This method consists basically in estimating the number of DY outside of the peak of the Z directly from the number of data events inside of the peak.

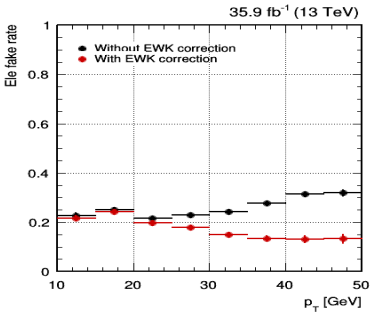


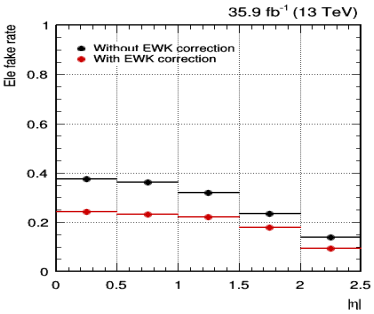
The scale factor obtained (1.07 ± 0.07) with this method has been applied to the usual MC simulations for the Drell-Yan process.

Fake rate

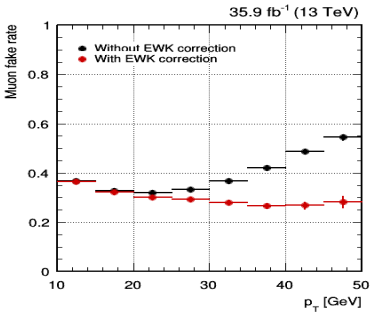
The fake rate corresponds to the probability to see a fake lepton (usually misidentified from a jet) passing the cuts corresponding to the final selection level of the analysis. This rate is estimated for both electrons and muons, within a QCD-enriched control region, by removing the contamination coming from the Z+jets and W+jets processes.

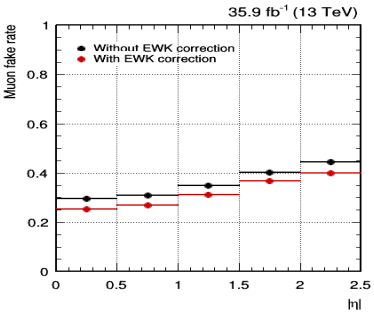
Electron FR





Muon FR

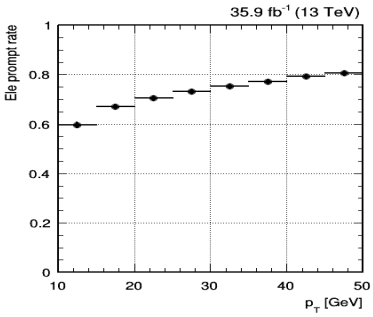


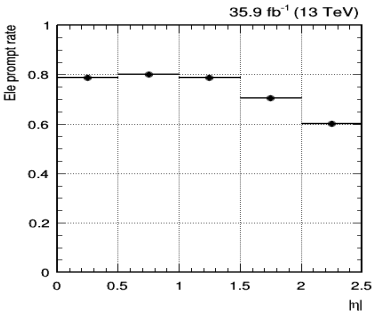


Prompt rate

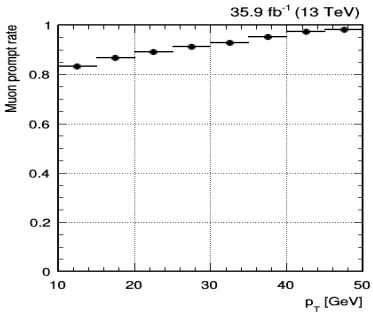
The prompt rate is calculated thanks to a general tag and probe method and is useful to take into account the real lepton contamination in the control region we defined to calculate the fake rate.

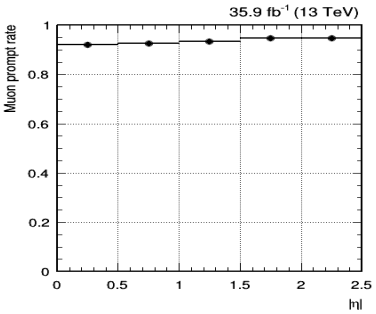
Electron PR





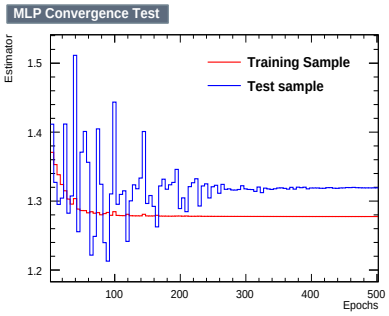
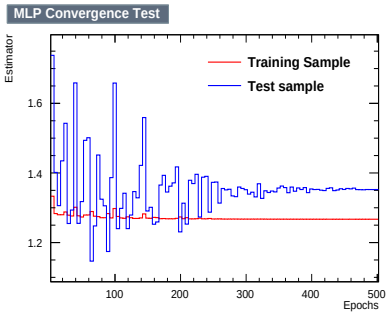
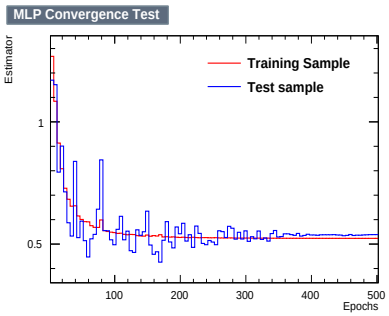
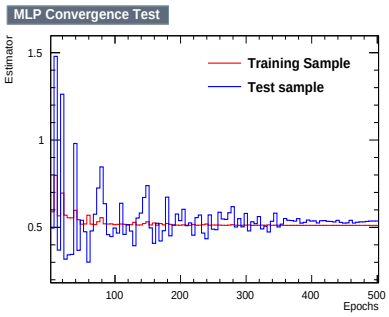
Muon PR





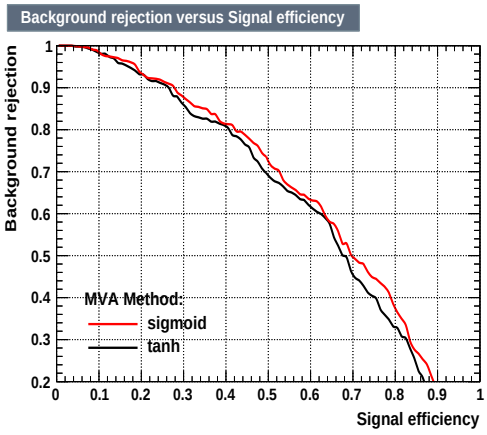
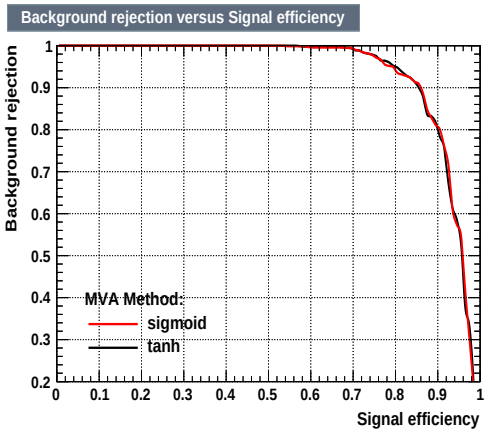
Overtraining issue of the ANN

We can check for eventual signs of overtraining for the different neural networks built. This can be done thanks to the convergence plots.

	Activation function 1	Activation function 2
$m_\Phi = 10$ GeV		
$m_\Phi = 500$ GeV		

ROC curves of the ANN

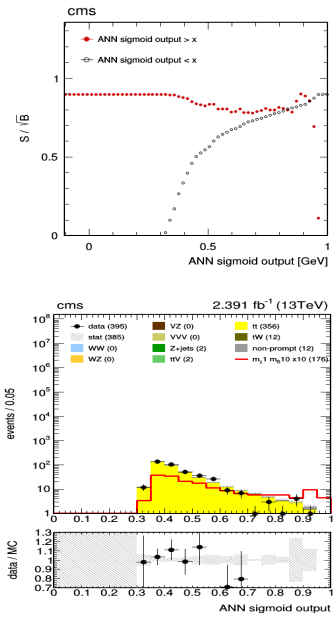
The ROC curves represent the signal efficiency possible to achieve with the neural network considered, given any background rejection we are interested in. Here are compared the ROC curves obtained for both the sigmoid and tanh activation functions, and for both the 10 and 500 GeV networks. As we can see, we can achieve a much better background rejection for any given signal efficiency in the 500 GeV case.

10 GeV network	500 GeV network
	

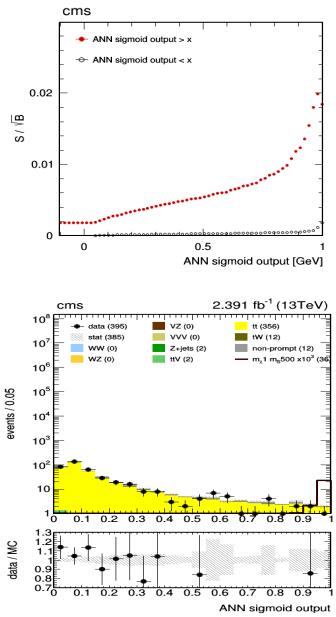
Output of the ANN

We can also study the output of the different neural networks, and its significance curve which are helpful to find the optimal cut to apply to the analysis, to get the best significance possible.

10 GeV network



500 GeV network



Yields table (pseudoscalar mediator)

We can now represent the yields obtained for the different backgrounds with the associated systematics, for the different background processes and different dark matter mediator mass points (and therefore, different neural networks).

10 GeV pseudo (ANN output $>$ 0.80)

Process	Yields	Statistical error	Systematic error
$t\bar{t}$	8.28	0.15	0.42
Non-prompt	0.30	0.65	0.10
ttV	0.74	0.03	0.18
Single top	0.35	0.06	0.01
Drell-Yan	0.39	0.28	0.12
WW	0.02	0.02	0.01
Total background	10.08	0.72	0.48
Signal	0.89	0.02	0.45
Total with signal	10.97	0.73	0.66
Data	8	-	-

500 GeV pseudo (ANN output $>$ 0.98)

Process	Yields	Statistical error	Systematic error
$t\bar{t}$	0.81	0.05	0.04
Non-prompt	0.10	0.31	0.03
ttV	0.46	0.02	0.12
Single top	0.04	0.02	0.01
Drell-Yan	0.00	0.00	0.00
WW	0.01	0.01	0.01
Total background	1.42	0.32	0.13
Signal	0.03	0.01	0.01
Total with signal	1.45	0.32	0.13
Data	1	-	-

Complete 2016 dataset expected limits

From the 2.4 fb^{-1} limits we calculated, it is actually easy to estimate the expected limits when considering the complete 35.9 fb^{-1} dataset, since we know that the limits are expected to decrease as the square root of the increase in luminosity. An increase of a factor 15 in the luminosity should then result in limits 4 times smaller.

Scalar mediator

m_S [GeV]	Expected limit
10	0.77
20	0.81
50	1.07
100	1.82
200	6.55
300	14.00
500	45.31

Pseudoscalar mediator

m_P [GeV]	Expected limit
10	2.44
20	2.57
50	2.96
100	2.84
200	4.47
300	8.55
500	43.36

By considering the complete 2016 dataset, we therefore expect to be able to exclude with our analysis the 10 and 20 GeV scalar mediators.