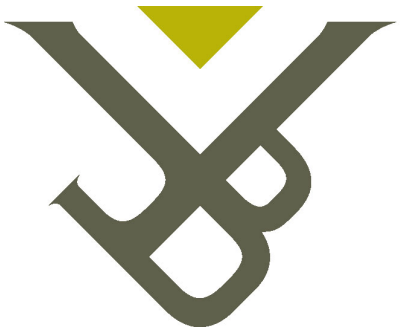


Vrije Universiteit Brussel

Faculteit Wetenschappen
Departement Natuurkunde

Development of new charm-tagging methods for the search for Flavour Changing top-quark dark matter interactions at the LHC

Graduation thesis submitted in partial fulfilment of the requirements for the degree of Master of Science in Physics and Astronomy

Seth Moortgat

Promotor: Prof. Dr. Jorgen D'Hondt
Co-promotor: Dr. Gerrit Van Onsem

Academic Year: 2014-2015
June 5th 2015

Summary

This thesis presents the results of a threefold research project: the construction of a phenomenological model for flavour changing (FC) interactions between top quarks and dark matter, the development of a new charm-jet tagging algorithm for the CMS experiment and finally an investigation of the potential use of that charm-tagging algorithm in a search for interactions, as predicted by this phenomenological model, in proton-proton collisions at the LHC. First an effective field theoretical description of this recently developed phenomenological model is presented in order to parametrise these interactions. Limits from dark matter relic abundances, direct and indirect dark matter searches are investigated and the calculated cross sections for possible processes at the LHC are discussed. In order to possibly improve the search for these flavour changing interactions, a new charm-tagging method is developed for the CMS experiment and its performance is optimised. Information from secondary vertices and displaced tracks in the hadronisation of charm quarks is combined with multivariate analysis techniques to identify jets from charm quarks and to distinguish them from light-flavour and gluon jets or from bottom jets. Finally an analysis is presented to identify proton-proton collision events at the LHC involving these FC top-quark dark matter interactions. This search is performed with a simple cut-based method and by applying template fitting. The new charm-tagging algorithm is shown to improve slightly the sensitivity of these searches with final-state charm jets.

Samenvatting

Deze thesis presenteert de resultaten van een drievoudig onderzoeksproject: de constructie van een fenomenologisch model dat interacties beschrijft tussen top quarks en donkere materie waarbij de flavour van de quark verandert (FC), de ontwikkeling van een nieuw charm-jet tagging algoritme voor het CMS experiment en uiteindelijk een onderzoek naar de toepasbaarheid van dit charm-tagging algoritme in een zoektocht naar interacties, die voorspeld worden door dit fenomenologisch model, in proton-proton botsingen bij de LHC. Eerst wordt een effectieve veldentheorie van dit recent ontwikkelde model voorgesteld om deze interacties te parametriseren. Limieten van de huidige dichtheid aan donkere materie en directe en indirecte zoektochten naar donkere materie worden onderzocht en de berekende cross secties voor mogelijke processen bij de LHC worden besproken. Om de zoektocht naar dit soort interacties mogelijk te verbeteren, wordt een nieuwe charm-tagging methode voor het CMS experiment ontwikkeld en zijn performantie wordt geöptimaliseerd. Informatie van secundaire vertices en tracks met een grote impact parameter ten opzichte van de primaire botsingsvertex, die zich vormen tijdens de hadronisatie van charm quarks, wordt gecombineerd in multivariate analysetechnieken om jets komende van charm quarks te indentificeren en ze te onderscheiden van light-flavour of bottom jets. Uiteindelijk wordt een analyse voorgesteld om protonbotsingen bij de LHC met FC interacties tussen top quarks en donkere materie te identificeren. Deze zoektocht wordt uitgevoerd door selectiesnedes op gevoelige variabelen toe te passen en door gebruik te maken van template-fitting. Het nieuwe charm-tagging algoritme toont een kleine verbetering van de gevoeligheid voor deze analyse met charm jets in de eindtoestand.

Contents

Chapter 1

Introduction

The first chapter discusses the basic concepts relevant to the research presented in this thesis. First the Standard Model of particle physics will be outlined with a historic overview and its current status. After this the concept of dark matter is introduced and as an indisputable motivation for its existence, some cosmological observations are discussed. Then the experimental setup is explained with a discussion of the CMS detector at the LHC. In the final section the main objectives and further content of this thesis are presented.

1.1 The Standard Model of particle physics

The Standard Model (SM) of particle physics is a theory describing the fundamental interactions between elementary particles. It is a mathematical framework in which the elementary particles are described by quantum fields and the fundamental interactions between these particles occur though the exchange of mediators. In such a field theory, the dynamics of a system can be summarized in the form of a Lagrangian, from which equations of motion can be deduced. The description of these interactions is based on the principle of gauge invariance, which is why the mediators are often called gauge bosons. The Standard Model aims at providing a unified description of all elementary particles and the fundamental interactions between them (electromagnetic, weak and strong interactions)[1] within one and the same mathematical framework. This theory has been successfully tested over the last decades with a large variety of experiments, which makes it one of the most precisely tested theories in the history of physics.

It was in the middle of the 20th century that the foundations of the Standard Model were laid when an effort was made to unify the electromagnetic and weak interactions. The success of this unified description triggered the development of a consistent mathematical framework that led to the Standard Model as known today. Some of the most prominent contributors to this work are Glashow [2], Weinberg [3] and Salam [4] who received the 1979 Nobel Price in Physics for their work.
Around the early 70's the first experimental evidence for the unified nature of the electroweak

[1] The Standard Model fails to include a description of gravitational interactions without spoiling renormalizability of the theory [1].

interactions arose with the discovery of the neutral current electroweak interaction with the Gargamelle bubble chamber at CERN (1973) [5]. The direct observation of the electroweak mediators themselves, now known as the W and Z bosons, had to wait until 1983 when they were discovered at the Super Proton Synchrotron [6]. Over the years the accelerators became more powerful and the Standard Model could be tested more precisely. The success of all these tests led to the acceptance of the Standard Model as the theory that describes the electroweak and strong interactions.

As illustrated in Figure 1.1, the SM particles can be subdivided into fermions (particles with half-integer spin) and bosons (particles with integer spin), named according to the type of statistical behaviour they obey in a quantum mechanical description [7]. The fundamental particles like the charged leptons, neutrinos and quarks are fermions. According to the current knowledge, they each exist in three generations differing only in mass; the three charged lepton generations are called electron, muon and tau, each with a corresponding neutrino. The up-type quarks are called up, charm and top quark and the down-type quarks are called down, strange and bottom quark.
On the other hand, the mediators of the fundamental forces are bosons. The massless photon together with the massive neutral Z boson and two oppositely charged W bosons make up the 4 gauge bosons of the electroweak theory, described by a SU(2) × U(1) gauge theory. Eight massless gluons mediate the strong force, which is described by a SU(3) gauge theory with corresponding colour charges. Together this forms the SU(3) × SU(2) × U(1) Standard Model gauge theory. However, this theory cannot be made invariant under gauge transformations (dictated by the SM gauge group structure) if masses for the bosons or fermions are introduced by hand. Brout, Englert [8] and Higgs [9] proposed a mechanism referred to as spontaneous symmetry breaking, which introduces a scalar field that couples to the gauge bosons and therefore generates mass terms for the bosons in a gauge invariant way. Also the fermions can acquire mass through Yukawa couplings to the scalar field. The scalar boson that corresponds to that scalar field completes the Standard Model as a consistent renormalizable framework.
Currently the experimental tests of the Standard Model are driven by the strongest particle accelerator present on earth: the Large Hadron Collider (LHC) at the CERN laboratory in Geneva, Switzerland. In 2011 and 2012 enough data of proton-proton collisions at a centre of mass energy of 7 and 8 TeV were collected to discover the Brout-Englert-Higgs (BEH) boson [10, 11], which was the last missing piece of the SM puzzle. More precise measurements of the properties of the scalar sector are obviously needed and the same holds for the top-quark sector, but the Standard Model has passed all its tests successfully and is therefore well established. At the same time, the success of the Standard Model to describe high-energy physics in currently accessible energy ranges (around the TeV energy scale) is in a way troublesome, since it fails to explain some fundamental questions about the universe. One of these unresolved issues is the existence of dark matter, but many more problems are known such as the failure of the Standard Model to describe gravity, the fine-tuning of the scalar boson mass (commonly referred to as the hierarchy problem), the presence of neutrino masses, ... The need for beyond the Standard Model (BSM) physics is inevitable and so the

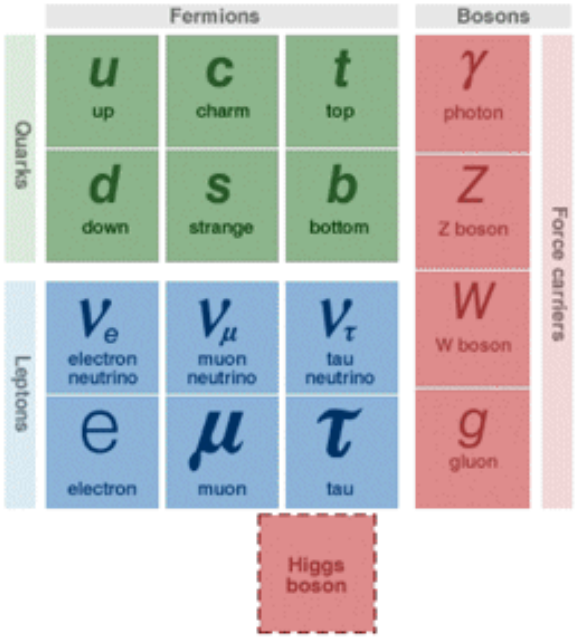


Figure 1.1: Schematic overview of the SM fermions and bosons.

search for BSM physics at particle accelerators is at the center of attention for the next few years.

1.2 Dark Matter

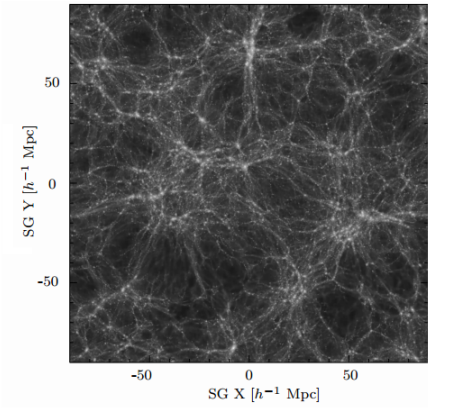
According to the standard model of cosmology, supported by a variety of experiments such as the observations from the Hubble Space Telescope and the WMAP satellite [12], the content of the universe consists of three different types of energy/matter; only about 4 percent is made up of ordinary matter, as described in the Standard Model of particle physics, around 72 percent is the so-called dark energy[2] and the remaining 24 percent is dark matter.

The existence of dark matter (DM) is supported by some very convincing cosmological observations [13]:

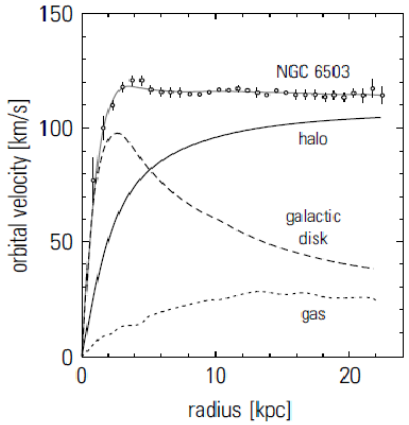
- First of all DM is needed for the structure formation in the universe, to explain the formation of galaxies and clusters as we see them today. Its gravitational attraction is needed to form large scale structures on a time scale we observe today. The influence of DM on structure formation is measured by dedicated simulations, with visualisations like in Figure 1.2(a) [14].
- When the orbital velocity of stars is considered as a function of the distance from the center of the galaxy [15], the observations cannot be explained by the presence of visible matter alone. Some sort of missing, invisible (dark) matter is required to explain why objects far away from the rotation axis seem to move too fast to be explained by only the ordinary, visible matter. This is illustrated in Figure 1.2(b).

[2]The presence of dark energy tends to accelerate the expansion of the universe, but its origin is perhaps even more of a mystery than the origin of dark matter.

- Gravitational Lensing is the effect where the light of distant objects is bent around heavy objects (such as galaxies) in the line of sight under the influence of gravity. This creates the illusion of seeing the distant object in an optical ring around the lens. The presence of DM can amplify this effect [16]. A picture showing so-called arcs because of gravitational lensing is shown in Figure 1.2(c).
- The Bullet Cluster consists of two colliding clusters of galaxies. Results from gravitational lensing, as described above, allow to determine the distribution of mass in these colliding clusters, as shown in Figure 1.2(d) [17]. This picture shows that the luminous matter (red) is located around the center due to electromagnetic interactions in the colliding material. However the matter distribution shows a clear separation of two centres of mass, indicating the presence of dark matter that is not affected by electromagnetic friction forces.

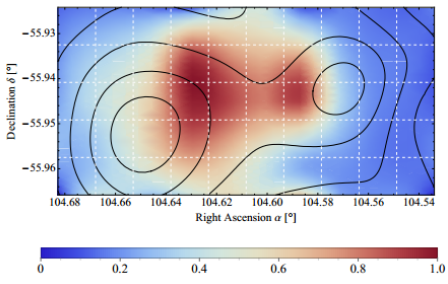


(a) The simulation of structure formation requires the presence of dark matter to obtain the particular filament structure of galaxy clusters which is consistent with observations [14].



(b) Rotational curves of the spiral galaxy NGC 6503. The contributions of the galactic disk, the gas, and the assumed dark matter halo are separately shown [18].

(c) Strong gravitational lensing around galaxy cluster CL0024+17 [16].



(d) Matter distribution in the colliding Bullet Cluster [17].

Figure 1.2: Different Experimental observations that support the existence of dark matter.

The so-called *WIMP miracle* states that for thermally produced dark matter in the early universe, with a mass around the GeV scale, the required annihilation cross section to explain the current measured DM relic abundance should be of the order of the weak interaction scale. Therefore dark matter is expected to be only very weakly interacting, not coupling to photons, but having large gravitational effects and therefore being massive. This is why a typical dark matter particle candidate is often called a weakly interactive massive particle (WIMP). None of the currently known SM particles have the right properties to serve as a dark matter candidate. Not even the neutrino, whose abundance could not explain the present day observations of the dark matter abundance and the observable effects of dark matter as explained above. Therefore there should be particles beyond the Standard Model of particle physics to describe this.

Many ideas have been brought forward to describe the nature of DM [18]. The contribution from neutrinos and MACHOS[3] has proven to be too small for the observed DM abundance. Moreover neutrinos would make up so-called hot (or thermal) dark matter, whereas cold dark matter is needed for example for structure formation in the early universe.
Superssymmetry (SUSY) relates fermions to bosonic patners and vice versa, extending the particle content of the Standard Model. A possible candidate for WIMPs is the lightest supersymmetric particle (LSP): the neutralino, which is the lightest mass eigenstates of the neutral colorless gauginos[4]. However, the existence of a stable LSP requires a theory in which all interactions obey the conservation of a symmetry called R-parity, which is defined as $P_R = (-1)^{3B+L+2s}$ where B is the baryon number, L is the lepton number and s is the spin. This symmetry ensures that SUSY particles are always produced in pairs in collisions of SM particles, but it also ensures that the LSP is stable, since there exists no lighter SUSY particle for it to decay to.
Another possibility is the axion, which arises as a solution to the strong CP problem [19] but happens to have the correct properties for a dark matter candidate.

In addition to cosmological observations, wich are mostly based on the gravitational effect of dark matter, it is necessary to search for the nature of the dark matter particles themselves. This search for dark matter is already progressing and is based on different detection methods, illustrated in Figure 1.3. Indirect searches look for dark matter annihilation to SM particles, for example as is being done with the IceCube neutrino detector [20]. Direct detection experiments look for interactions of dark matter with the nuclei of the detector material, for example in the DAMA/LIBRA [21] experiment, the XENON experiment [22] and the LUX experiment [23]. Collider searches look for dark matter creation in collisions of SM particles, which is done in the CMS [24] and ATLAS [25] experiments at the LHC and is also the aim of this research project. So far, none of these searches have been able to identify the nature of DM.

[3] A massive astrophysical compact halo object or MACHO is an object of normal baryonic matter, but with very little radiation making it hard to detect and a possible dark matter candidate. Examples are brown dwarves, black holes, neutron stars or even planets.

[4] Neutral colorless gauginos, namely the photino, zino and neutral higgsino, are the supersymmetric partners of the neutral colorless gauge bosons.

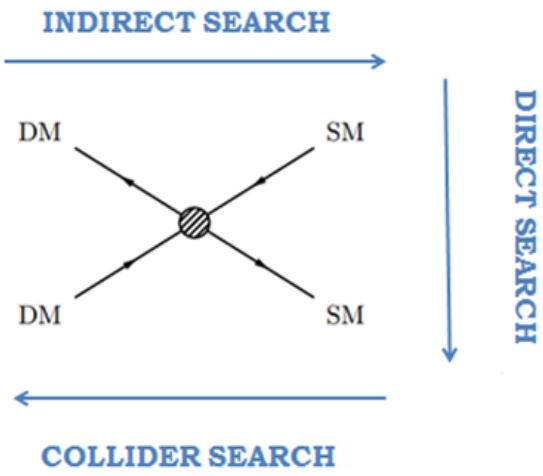


Figure 1.3: Illustration of different detection methods for dark matter interacting with SM particles.

1.3 The Large Hadron Collider and the Compact Muon Solenoid detector

1.3.1 The Large Hadron Collider at CERN

Studying physics at very high energies allows one to investigate and discover the fundamental building blocks of nature and their properties. The production of such high energies is achieved in particle accelerators by accelerating particles to velocities close to the speed of light and making them collide. The choice of the colliding particles defines the nature of the accelerator and is often based on the research goals for which it was built. One can collide leptons, hadrons or even heavy nuclei (ions) and one may choose to collide particles or anti-particles. Besides that one may choose to build a linear or a circular accelerator. Every setup has its advantages and downsides.

Particles accelerated in linear accelerators do not experience energy losses through synchrotron radiation, but the particles can only be accelerated once. A circular accelerator allows the particles to be accelerated over many cycles, but is limited by the amount of synchrotron radiation energy losses, while it is challenging to keep the particles on a bent track with very strong magnets.

Lepton colliders have clean event signatures and are therefore well fit to do precision measurements of properties of elementary particles and their interactions. However, the large amount of synchrotron radiation[5] limits the maximum accessible energy of the collider. Hadrons lose far less energy due to synchrotron radiation, but the fact that they are constituent particles, made up of quarks and gluons, causes the collision event signal to be very busy. This increases the amount of background activity in the particle detector and therefore the amount of uncertainty on any measurement. The higher accessible energies but relative

[5]The energy loss of a charged particle with mass m due to synchrotron radiation falls off as m^{-4}, which is why leptons, being much lighter than hadrons, lose far more energy.

lack of precision often gives hadron colliders the name of *discovery machines*.

The LHC at CERN was mainly designed to discover the Brout-Englert-Higgs (BEH) boson. Based on that goal the decision was made to build a circular proton-proton collider. The LHC [26] lies approximately 100 meters underground crossing the Swiss-France border near Geneva, Switzerland and has a circumference of almost 27 kilometers. It is currently the most powerful particle accelerator in the world and is designed to collide protons[6] at a center of mass energy $\sqrt{s}$ of up to 14 TeV and with a luminosity[7] of 10^{34} cm^{-2}s^{-1}. In 2011 and 2012 the LHC ran at centre of mass energies of 7 and 8 TeV respectively, yielding enough data to discover the BEH-boson and fill the last gap in the Standard Model. In 2013 and 2014 the first long shutdown (LS1) took place for upgrades and repairs to the accelerator and its surrounding particle detectors. Recently the second run of the LHC started and the center of mass energy will increase to around 13 TeV.

Before the proton beams enter the 27km long tunnel of the LHC, they pass through a series of pre-accelerators. This is illustrated in Figure 1.4. The protons pass through the LINAC2 ($\rightarrow$ 50 MeV), through the BOOSTER ($\rightarrow$ 1.4 GeV), the Proton Synchrotron (PS) ($\rightarrow$ 25 GeV), the Super Proton Synchrotron (SPS) ($\rightarrow$ 450 GeV) and finally get inserted in the LHC.

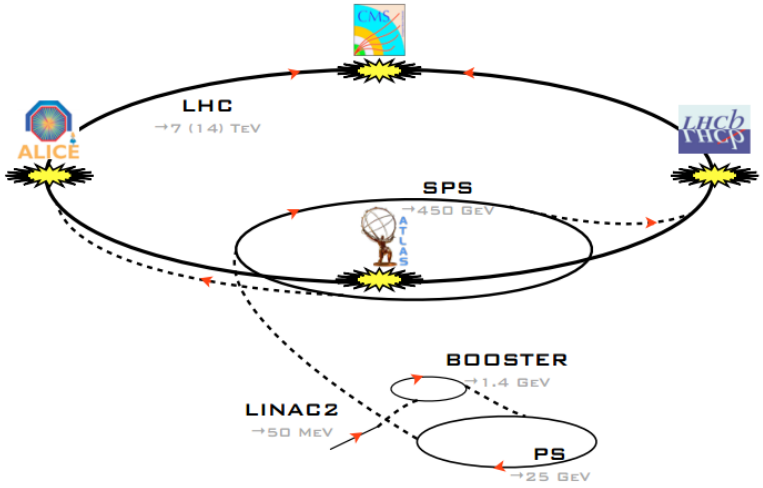


Figure 1.4: Sketch of the Large Hadron Collider, its preaccelerators and the four main experiments (CMS, ATLAS, LHCb and ALICE) at the interaction points (not to scale) [27].

The proton beams move through two separate beam pipes in opposite directions. These bunches are accelerated with a system of radio cavities and held on the circular trajectory by superconducting magnets with fields up to 8.3 T. Each proton bunch contains around 10^{11} particles and the minimal bunch spacing is approximately 25 ns. This way a design

[6] The LHC can also collide heavy lead ions with center of mass energies of up to 2.8 TeV per nucleon and at a peak luminosity of 10^{27} cm^{-2}s^{-1}.

[7] The luminosity ($\mathcal{L}$) of the accelerator is defined by the number of detected events per unit time (N) and the cross section (σ) of the process of interest through the following formula: $N = \mathcal{L} \times \sigma$. The cross section is a measure for the probability of a certain process to occur.

luminosity of up to around 10^{34} $cm^{-2}s^{-1}$ can be achieved, resulting in an average of around 20 to 40 inelastic collisions per bunch crossing.
A total of four interaction points are located at different places along the 27 km long beam pipes, where the proton bunches can be slightly bent of track and are made to collide. Around these four interaction points large detectors are placed to detect the products of the collisions and study known physics or search for new physics. Two of these detectors, CMS and ATLAS, are general purpose machines. A third one, ALICE, focuses on collisions of lead ions. The fourth detector, LHCb, is specialised in physics of B-mesons to search for CP violation and an explanation for the matter anti-matter asymmetry in the universe. Their position at the LHC is displayed on Figure 1.4, but the relative spacing between the interaction points is not to scale (neither is the relative size of the different (pre)accelerators).

The LHC is pushing the boundaries of the current experimental techniques and is the perfect environment to search for signs of new physics. The CMS detector is designed to do so, and after the recent discovery of the BEH-boson, it is expected to provide a lot of new interesting information when the collision data at 13 TeV will be analysed.

1.3.2 The Compact Muon Solenoid detector

Detector design The Compact Muon Solenoid (CMS) detector [28] is one of the four detectors surrounding the interaction points of the LHC. It is located in France in an underground cavern near Cessy. The detector has a cylindrical shape and is more than 21 meters long and 15 meters in diameter. The cylindrical shape is typical for modern particle detectors to accommodate a layered structure of subdetectors, each aiming at detecting different types of particles. It weighs about 14 thousand tonnes.

The CMS detector is a general purpose detector, meaning it is designed to study a variety of physics topics. Nevertheless its design is inspired by the search for the BEH-boson, which was the main motivation to build the LHC in the first place. The name of the detector already reveals some of its specialised features. It is compact, using all available space to fit most of the subdetector parts inside its large solenoid magnet. That very strong solenoid magnet allows for a very precise measurement of the transverse momentum (p_T) of the charged particles. This is a consequence of the electromagnetic force on the charged particles that bends their tracks, which can be measured by the inner tracker. Jets are measured by the electromagnetic (ECAL) and hadronic (HCAL) calorimeters, also located inside the solenoid magnet. Just outside of the solenoid, an extensive system of muon chambers is installed to improve the reconstruction of muons in addition to the inner tracker. An overview of the structure of the CMS detector is given in Figure 1.5 and a detailed schematic view of the subdetectors and their pseudorapidity[8] (η) coverage is given in Figure 1.6. A more extensive discussion of the different subdetectors is given below.

[8]The pseudorapidity is a quantity deduced from the polar angle θ: $\eta = -ln\left[tan\left(\frac{\theta}{2}\right)\right]$. At high energies it is equivalent to the rapidity (y) of a particle, defined as: $y \equiv \frac{1}{2}ln\{\frac{E+p_z}{E-p_z}\}$ where E is energy of the particle and p_z is the z component of its momentum. Differences in rapidity are Lorentz-invariant which makes it an attractive quantity. A more detailed description of the CMS coordinate system can be found in reference [28].

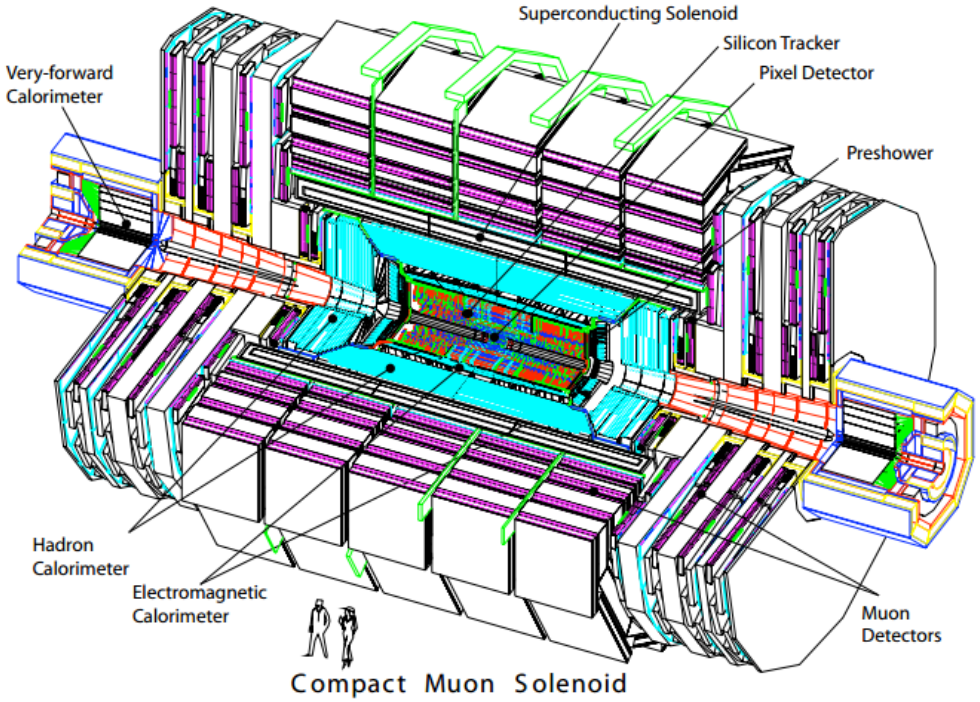


Figure 1.5: Overview of the CMS detector and its different subdetectors [29].

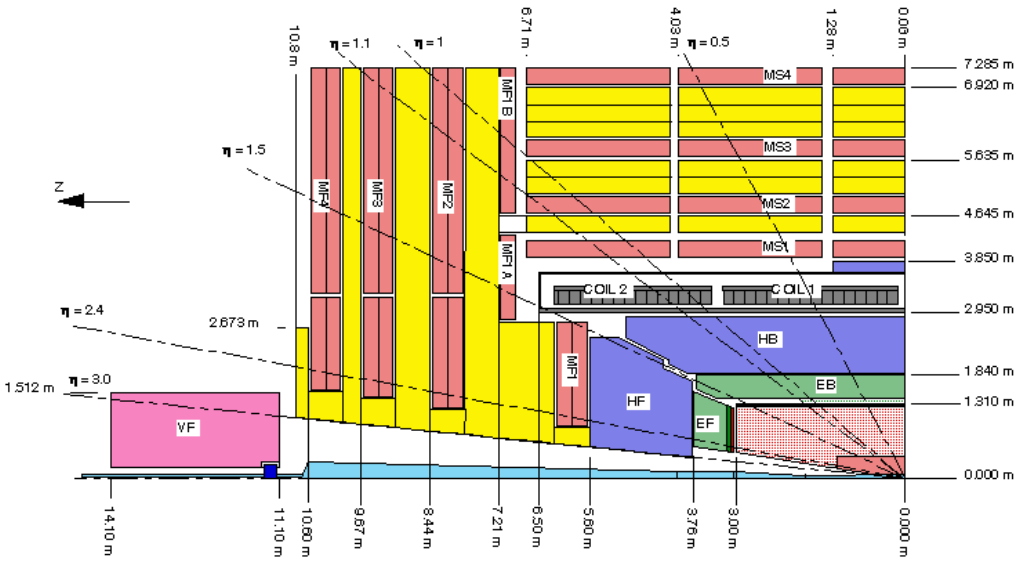


Figure 1.6: Schematic side view of the different CMS subdetectors and corresponding η coverage. From inside-out: The inner tracker (red), the ECAL (green), the HCAL (blue), the solenoid magnet (gray) and the muon chambres (pink) interleaved by return yokes (yellow) [30].

Superconducting solenoid magnet One of the central pieces of the CMS detector is its solenoid magnet. This superconducting magnet is designed to reach a field strength of about 3.8T. The choice to equip the detector with a very strong and state-of-the-art magnet is based on the design goal to achieve an excellent p_T resolution for charged particles. Due to its compactness, the tracker and calorimeter systems fit inside the solenoid. Only the muon

chambers (and an extra outer layer of the HCAL) are located outside and are interleaved with iron return yokes to guide the magnetic field properly through these chambers. In Figure 1.6, the magnet is presented in gray and the return yokes are drawn in yellow.

The magnet itself is superconducting and exhibits a 4-layer winding of a NbTi conductor. This material has the right properties to conduct very large currents and allows the magnet to produce a very strong magnetic field. To keep a magnetic field within such a complex structure homogeneous is challenging and requires state-of-the-art technologies.

Inner tracker The CMS tracker system [28] is located just around the beam pipe of the LHC and is designed to reconstruct tracks of charged particles from the proton collisions. The solenoid magnet provides a homogeneous field along the entire tracker volume and bends the tracks of the charged particles that move through the tracker. A precise reconstruction of these tracks allow for a very precise measurement of the p_T of these particles. The precision of the tracker is very important, since is has to distinguish tracks from around 100 particles created in the average of 20 to 40 inelastic collisions per bunch crossing, with one bunch crossing every 25 ns.

In view of the research presented in this thesis, the precision of the tracking system is of vital importance. The process of identifying charm-quark jets (charm tagging), which will be extensively explained in Chapter 3, is based on the reconstruction of secondary vertices, created after the decay of short-lived D-mesons. These mesons travel a short distance in the tracker (a few millimetres) and then decay to other particles, creating a displaced secondary vertex from displaced tracks. The CMS tracker is precise enough to identify individual interaction vertices and secondary vertices and is therefore well suited to perform charm tagging (and bottom tagging).

The tracker itself is composed of an inner pixel detector, containing around 66 million pixels of 100×150 μm each and an outer silicon strip tracker that contains around 9 million strips with changing dimensions according to the position and radial distance from the centre. Its detection principle is based on the use of semiconductors in ionisation detectors. It has a pseudorapidity acceptance of $|\eta| < 2.5$ and reaches outward to a radius of approximately 1.1 meter. Figure 1.6 shows the inner tracking system in red, with the pixel detector in full red and the silicon strips in dotted red.

Electromagnetic (ECAL) and hadronic (HCAL) calorimeters The calorimeter systems measure the energy of the particle showers created by passing particles. These can either be electromagnetic particles, like electrons or photons, which create showers due to successive chains of electromagnetic radiation (photons), pair production and annihilation. These showers are detected by the inner layers of the calorimeter system: the electromagnetic calorimeter (ECAL) [28]. On the other hand the quarks and the gluons are known to hadronise and create a shower of new baryons and mesons. This is a consequence of the color confinement of hadrons which states that due to the nature of the strong interaction, coloured particles can not be measured individually but are always confined in pairs

(mesons) or triplets (baryons) which are colour neutral. The resulting hadronic showers, also called jets, typically penetrate further into the detector, leaving some energy deposits in the electromagnetic calorimeter, but most of the energy is deposited in the outer layers of the calorimeter system: the hadronic calorimeter (HCAL) [28].

The reconstruction of these hadronic jets is of great importance to the research presented in this thesis. Different jet reconstruction algorithms exist, which try to cluster the energy deposits in the calorimeter cells and combine this information with charged tracks from the inner tracker system to reconstruct each jet as precise as possible. This is at its own a very challenging procedure and will briefly be discussed in Section 3.1.2. The properties of the reconstructed jet give a lot of information on the originating quark or gluon (starting the fragmentation chain). Heavy-flavour tagging is based on this principle and tries to find out whether a jet originated from a meson with a bottom quark inside (B-mesons), with a charm quark inside (D-mesons), with light-flavour quarks (down, up or strange) inside or whether it originated from a gluon[9]. The better the jet reconstruction, the better the performance of heavy flavour tagging. This research focuses on the identification of jets originating from charm quarks (c tagging).

The ECAL is made from 61200 lead tungstate ($PbWO_4$) crystals in the barrel and another 7324 crystals in each of the two endcaps. This material creates scintillation light, which is detected by photodiodes at the end of each crystal. Moreover, it is fast, radiation resistant and has a very fine granularity which all allows for a very good resolution on the electromagnetic energy measurements.
The HCAL is located in the next layer, just behind the ECAL from the point of view of the interaction point. The available space is restricted between the outside of the ECAL and the inside of the solenoid magnet, and therefore only a limited amount of absorbing materials can be used. In order to catch any parts of the hadronic showers that get through the solenoid (punch through), a small layer outside of the solenoid completes the HCAL. It is made from layers of showering material (with different compositions), interleaved by layers of scintillating material to measure the shower energy. The HCAL has a broader granularity and in general a worse energy resolution than the ECAL.
Figure 1.6 shows the ECAL and the HCAL barrel and sidecap parts respectively in green and blue.

Muon systems The CMS detector has a very advanced muon system[28], since many new physics searches greatly benefit from good muon reconstruction. Muons are far less affected by radiative energy losses in comparison to electrons and will therefore penetrate much deeper into the detector. For this reason a system of muon chambers is located in the outermost layer of the detector, as illustrated in Figure 1.6 in pink. Muons will, just like any charged particle, also leave tracks in the inner tracker. Combined information from the inner tracker and the outer muon system can be used to improve the p_T resolution of

[9]The top quark will decay very quickly into a lighter quark (almost always a bottom quark) and will not form a stable meson or baryon, which is why people never refer to a jet as originating from a top quark.

muons. A muon that is reconstructed with combined information from the tracker and the muon system is called a global reconstructed muon.

Muons can be created directly in hard scatter processes, in which case they tend to be isolated. This means that they do not have much energy surrounding them. The identification of electrons and muons will be important in the search for flavour changing interactions between top quarks and dark matter, which will be addressed in Chapter 4. On the other hand, leptons can be created inside jets with a lot of energy around them (making them non-isolated). Information on the energy and direction of these electrons and muons can differ between different types of jets and can therefore be used in heavy-flavour tagging methods like charm tagging. These features show the importance of muon (and electron) reconstruction in the CMS detector for the purpose of this research.

The CMS muon system uses three types of detector elements: Cathode Strip Chambers (CSC), Drift Tubes (DT) and Resistive Plate Chambers (RPC).
The CSCs are made up of anode wires (usually in groups of 10) and cathode strips, which are oriented in a perpendicular direction to the wires (this allows for position measurement in two perpendicular directions). When a particle passes through the gas in this detector, the gas gets ionised and a charge will be induced on the cathode strips, whereas the electrons move towards the anode wires, creating an avalanche of ionisations on their way.
DTs work in a similar way, using an anode wire in between cathode strips. However in DTs the gas inside has a high drift velocity for the ionised electrons. Using the time difference between the passage of the particle and the arrival of the (avalanche of) electrons at the wire gives additional information on the position where the particle passed through. This results in a good spatial resolution of DTs.
Finally, RPCs where made in order to have a good time resolution as well. RPCs are made from plates of high resistivity, with a very large voltage between them, and separated by a narrow area of gas. Particles flying through will ionise the gas, which causes an immediate spark between the two plates. This leads to a very good time resolution. Together with the good spacial resolution of DTs and CSCs, the muon system of CMS allows for a very precise reconstruction of the muon tracks.

Trigger system and data acquisition CMS is a complex detector with different subdetectors all used together to reconstruct the proton collisions at the LHC. Each 25 ns multiple collisions takes place, leading to an overwhelming amount of data that need to be processed in a very short period of time. Furthermore, most of the collisions are low-energetic (soft) and not of great interest for discovering new physics. Since it would not be possible to store all data from the proton collisions, very fast decisions need to be made on whether an event seems interesting enough to store on disk or not. This is the job of the trigger system of CMS [28]. Triggering is therefore very important at the LHC and each analysis may want to trigger on different signatures in the detector. When an event passes the trigger, the electronic signals need to be digitised into data structures and brought together to form a complete event reconstruction. This is done by the data acquisition system [28].

CMS uses a two-level triggering system [31] in order to bring down the event rate from several MHz to about 100 Hz. After an event occurred, the data are sent to a pipeline, which delays the data stream for about 3 μs. This is enough time to send the data to the Level 1 Trigger (L1) processors. This trigger is purely hardware (the data have not been put in software data structures) and makes local decisions (not combining the different detector elements). It consists of a muon trigger that uses information from the CSC, DT and RPC, and a calorimeter trigger that uses the information of the ECAL and the HCAL (but typically with a more rough granularity than present in the calorimeters). This information is then combined to form the Global L1 trigger, which may accept or reject the event. After about 3 μs this decision arrives back at the end of the pipeline to keep or throw away the event.
After this, the data that passed the L1 trigger is read out by the Front-End Drivers (FEDs) and it is sent through a dedicated switching system. Here the data are digitised and all of the data from the same event are brought together. These data (which are now in a software data structure format) have to pass the High Level Trigger (HLT) to again reduce the amount of data. The HLT tries to apply thresholds on physics objects as one would apply in an offline analysis, but in a limited amount of time and with only a limited output bandwidth. After this, the data are written out offline and distributed over different computing systems around the world.
Currently a new triggering system is under development and will be used during the second run of the LHC.

1.4 Searching for interactions between DM and SM particles with the CMS detector

Dark matter is one of the major mysteries is physics nowadays and although its existence is indisputable due to many cosmological observations, the origin of dark matter remains unclear. The standard model of particle physics fails to deliver a feasible dark matter particle candidate, and so the need for beyond the Standard Model physics arises.

A priori there is no indication on how dark matter would interact with the SM particles. The many compelling cosmological observations of dark matter are mostly based on its macroscopic gravitational interaction with ordinary matter. However, if we want to find out the nature of dark matter, it should have some other interactions with the SM particles as well, much stronger and therefore more relevant than gravity at the subatomic scale. These interactions might manifest themselves in particle collisions at the LHC and could be detected with the CMS detector. The presence of dark matter manifests itself as missing transverse energy (MET) in the detector, since it will not interact with the detector material. Missing transverse energy is the magnitude of the vector sum of the energy in the transverse plane of all the detected particles in the detector. Since the two initially colliding particles have no momentum in the transverse plane, conservation of momentum dictates that the total vector sum of the p_T of all resulting particles from the collision should add up to 0. Any significant deviation from zero means that some particles escaped the detector, leading to missing energy in the transverse plane.

One possibility explored in this thesis is that DM interacts with the Standard Model through flavour changing interactions with top quarks. This change in flavour would then result in the production of charm quarks, creating jets inside the CMS detector called c jets. In order to identify these processes and distinguish them from SM background processes it could prove useful to develop a method to identify these jets originating from charm quarks. Such an algorithm is called a charm tagger or c tagger and is new in the CMS experiment. A charm tagger will find many applications in other analyses as well and is therefore of general importance to the CMS collaboration.

The research presented in this thesis consists of three main objectives, all aiming for the search of (flavor changing) top-quark dark matter interactions with the CMS detector at the LHC:

1. A first objective is to *explore a phenomenological model that describes flavour changing top-quark dark matter interactions* and that can be tested at the LHC. This is done by investigating a recently developed effective field theory describing the effective four-point interaction between a top quark, charm quark and two dark matter particles ($tc\chi\chi$) and check its consistency with the measured dark matter relic abundance. Finally the viability of detecting this model at the LHC is investigated by calculating branching ratios and/or cross sections for possible processes in 13 TeV proton-proton collisions.
2. A second objective is to *develop a novel c tagger for the CMS collaboration*, using multivariate analysis (MVA) methods. The current methods for b tagging are adapted to be used for c tagging. When the setup for an operational c tagger is developed, its performance in discriminating c jets from either light-flavour jets or b jets is measured in different situations using simulated data. Eventually the measured performance is optimised by selecting a set of sensitive variables for the MVA, tuning the MVA parameters and improving the secondary vertex reconstruction.
3. The final objective is to look at the *effect of c tagging on the search for flavour changing top-quark dark matter interactions in simulated data.* Therefore the MadAnalysis [32] software tools are used to determine the signal significance of the phenomenological model built in objective 1. Different event selection criteria are applied to simulated signal and background data sets to optimise this significance. Finally the effect of additional c tagging selection criteria (based on the new implementation of the developed c tagger in objective 2) on the significance are studied to look for possible improvements.

Each of the above objectives will be separately discussed in the following chapters, and it is instructive to keep in mind the interplay between all objectives. They all contribute to the search for flavour changing interactions between top quarks and dark matter at the LHC. But before looking at data, one needs to build a (phenomenological) model to be tested and develop the necessary methods (like c tagging) to optimise the analysis. This research project focuses on all of these preparations in order to check the potential of detecting such interactions at the LHC and to look for possible improvements from new charm-tagging methods.

Chapter 2

Phenomenological model for flavour changing top-quark dark matter interactions

The gravitational properties of dark matter are well known and supported by cosmological observations on large-scale structures. The particle nature of dark matter on the other hand remains a complete mystery. In order to understand the true nature of dark matter, the hope is that some type of interaction, other than gravitational, to the known SM particles exists. The so-called *WIMP miracle*, as explained in Section 1.2, provides a seemingly coincidental explanation for the weak interaction scale of thermally produced dark matter and puts this requirement in a bright perspective. Therefore a weakly interacting massive particle seems a good dark matter candidate.

Besides the nature of the interaction, it is not clear either with what SM particles the dark matter particle would interact. The following study is built on the hypothesis that dark matter interacts with the Standard Model through top quarks. Although some efforts have already been made to create models to describe top-quark dark matter couplings, the novelty of this project lies within the underlying assumption that this interaction may occur through flavour changing couplings. This opens up some new interesting signatures such as rare flavour changing top decays into charm quarks and dark matter particles, which has not been studied before.

First an effective field theory (EFT) will be introduced to parametrise the flavour changing interactions between top quarks and dark matter. Possible constraints from previous and ongoing searches will also briefly be mentioned. Then the consistency of different couplings with the dark matter relic abundance will be investigated and finally the relevant processes at the LHC are considered with their corresponding cross sections.

2.1 Effective field theory

In the simplest way possible, the SM particle content needs to be extended with only two particles to include interactions between quarks and DM. The first one is the dark matter particle itself, denoted by χ and the second particle is some unstable particle, denoted by V, that mediates the interaction between the DM and the quarks. This is illustrated in Figure 2.1 on the left. Within an EFT framework this mediator is considered heavy enough (typically at least at the TeV scale) such that it can be integrated out, even at the high energies accessible at the LHC[1]. In practice this means that the energy flowing through the mediator or equivalently the invariant mass of the $\chi\bar{\chi}$ system in Figure 2.1 on the left, should not exceed the mass of that mediator. This way the interaction can be considered a four-point interaction, as illustrated in Figure 2.1 on the right, and the SM Lagrangian only needs to be extended with a kinetic term for the DM particle and an effective operator to describe the interaction between the dark matter particle and the quarks. This operator is taken to be of dimension six[2], which is the lowest order possible to extend the SM Lagrangian with a coupling between four fermions. Higher order operators are suppressed by higher orders of the cut-off scale Λ of the EFT and will not be considered.

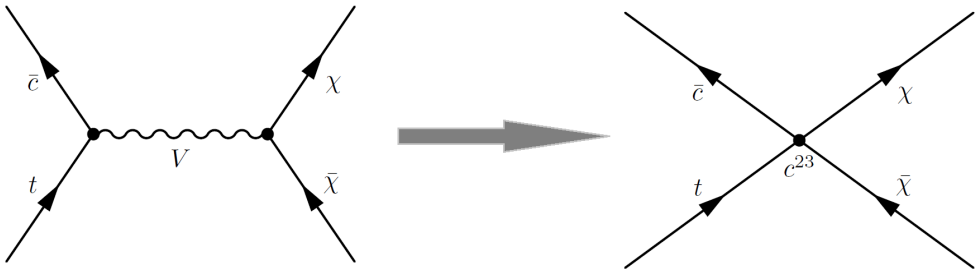


Figure 2.1: Illustration of the interaction between a top quark, a charm quark and two dark matter fermions (χ), mediated by a vector boson V on the left, and the corresponding EFT four-point interaction on the right.

In its most general form (at lowest order), the extension of the SM Lagrangian is expressed in Equation (2.1).

$$\mathcal{L} = \mathcal{L}_{SM} + \mathcal{L}_{DM}^{kin} + \mathcal{L}_{DM}^{int} \tag{2.1}$$

$$\mathcal{L}_{DM}^{kin} = \bar{\chi}(i\gamma^{\mu}\partial_{\mu} - m_{\chi})\chi \tag{2.2}$$

$$\mathcal{L}_{DM}^{int} = \frac{c^{ij}}{\Lambda^2}(\bar{q}^i\Gamma q^j)(\bar{\chi}\Gamma\chi) \tag{2.3}$$

In this notation, χ denotes the DM particle field. In this project it is considered a Dirac

[1]This can be compared to the Fermi interaction, accurately describing muon decay to an electron (or positron) and two neutrinos for energies well below the mass of the W boson.

[2]The dimension of an operator is determined by the dimensions of the fields used to build that operator. Fermionic fields are of dimension (in natural units) $[M]^{3/2}$ and the Lagrangian itself is of dimension $[M]^4$. Therefore an operator built from four fermions is of dimension $[M]^6$ and dimensional analysis dictates that such an operator should be divided by the square of the cut-off energy scale (which has dimension $[M]^1$) of the EFT to be used in the Lagrangian.

fermion with mass m_χ, as expressed by the form of its kinetic term in Equation (2.2). In Equation (2.3), c^{ij} is the coupling strength between the dark matter particles and quarks of type i (q^i) and j (q^j), where i and j are flavour indices that can vary from 1 to 3 and correspond to the up-type quarks (1,2,3) = (u,c,t). Λ is the cut-off scale of this effective interaction, corresponding to the mass scale of the mediator and Γ describes the type of mediator. $\Gamma = \{1(\gamma^5), \gamma^\mu(\gamma^5), \sigma^{\mu\nu}(\gamma^5)\}$ correspond respectively to (pseudo)scalar, (axial-)vector and (axial-)tensor interactions[3].

Equation (2.3) is general and describes different types of interactions (scalar, vector, ...) between any two types of quarks (but with the same electric charge if the DM is considered neutral) and between any kind of dark matter particles. However to reduce the complexity of the model, this research project focuses on the following assumptions:

- The interaction is considered vector-like, meaning $\Gamma = \gamma^\mu$. This choice was based on considerations of relic abundances (see Section 2.2).
- The focus is put on interactions in the up-type sector (up, charm and top quarks), since the down quark sector is much more constrained by previous flavour-constraining experiments involving B-mesons and Kaons. This includes limits on $Br(B^+ \to K^+\nu\bar{\nu}) < 1.4\times10^{-5}$ [33], $Br(B \to K^*\nu\bar{\nu}) < 8.0\times10^{-5}$ [34] and $Br(K^+ \to \pi^+\nu\bar{\nu}) = (1.73^{+1.15}_{-1.05})\times10^{-10}$ [35]. In these kind of processes, bottom quarks would decay to strange quarks and strange quarks to down quarks, with only missing transverse energy (MET) from the neutrinos. As the flavour changing interactions between quarks and dark matter would yield very similar signatures (DM also manifests itself as MET), these processes put strong constrains on the down-type sector.
- Only right-handed quarks are included, since left-handed quarks would require a treatment of the down quark sector as well if one does not want to break SU(2) gauge invariance.
- The flavours taking part in the interactions will be restricted to top quarks and charm quarks ($c^{ij} = c^{23}$, $q^i = c_R$, $q^j = t_R$), although diagonal (non-flavour changing) couplings may be considered when looking at DM relic abundances. The choice for top quarks is justified by the fact that the LHC is a top quark production machine. The presence of charm quarks is promising for the application of a charm-tagging algorithm.
- As already mentioned before, the DM particles are taken to be (Dirac) fermions. If they would be Majorana particles, the vector interactions would vanish [36].
- The presence of two dark matter particles in the effective coupling is needed in order to have a stable dark matter candidate, which can not decay into for example the quarks which take part in the effective coupling.
- The cut-off scale of the EFT (Λ) is typically taken around 1 TeV.

With all of the above considerations and assumptions, $\mathcal{L}^{int}_{DM}$ can now be written as in Equation (2.4), describing the effective interaction between dark matter and top quarks, with the flavour changing coupling to charm quarks (see also Figure 2.1 on the right).

[3] $\sigma^{\mu\nu} \equiv \frac{i}{2}(\gamma^\mu\gamma^\nu - \gamma^\nu\gamma^\mu)$

$$\mathcal{L}_{DM}^{int} = \frac{c^{23}}{\Lambda^2}(\bar{c_R}\gamma^\mu t_R)(\bar{\chi}\gamma_\mu\chi) \tag{2.4}$$

This phenomenological model is implemented in FeynRules [37] and simulated with MadGraph 5 [38].

2.2 Dark matter relic abundance

In the early universe, dark matter is assumed to be in thermal equilibrium. As dark matter disappeared through annihilation into other particles, an equal amount was produced by the inverse process. As time evolved, the universe expanded and started cooling down. As a consequence, the density of dark matter dropped and annihilations started to become less likely. On top of that, lighter particles did not have enough energy to produce the heavier dark matter particles any more. This phenomena is known as freeze-out, meaning that the dark matter stops interacting with its environment and conditions for thermal equilibrium are violated. At the moment of freeze-out, the dark matter abundance remains constant and it is that what we observe today as the dark matter relic abundance. This value was measured by the WMAP experiment [12] and more recently by Planck [39], resulting in the most sensitive measurement of the cold dark matter (CDM) relic density[4] of:

$$\Omega_{CDM}h^2 = 0.1199 \pm 0.0027 \tag{2.5}$$

This measurement is very important, not only to cosmology but also to any type of dark matter search. The dark matter relic abundance is determined by the annihilation cross section[5] σ and by the mass m_χ of the dark matter, as illustrated in Figure 2.2. These parameters are also of crucial importance in direct, indirect and collider searches for dark matter.

For the model outlined in the previous section, the consistency with the dark matter relic abundance $\Omega_{CDM}h^2 = 0.1199 \pm 0.0027$ needs to be checked. The parameter space of the couplings c^{ij} and the dark matter mass m_χ is scanned and the corresponding relic abundance is calculated. This is done separately for $\mathrm{tt}\chi\chi$ couplings (only $c^{33} \neq 0$ and all other couplings put to 0), $\mathrm{tc}\chi\chi$ couplings (only $c^{23} \neq 0$) and $\mathrm{cc}\chi\chi$ couplings (only $c^{22} \neq 0$). As a phenomenologically viable parameter space, the one where the calculated relic abundance is equal to or smaller than approximately 0.11 is taken. A relic abundance higher than the measured

[4] Ω_{CDM} is the closure parameter of cold dark matter that defines the energy density of CDM relative to the critical energy density of the universe. Measurements show that the total closure parameter is equal to 1 (flat universe), which makes the interpretation of Ω_{CDM} as the fraction of energy from CDM in the universe possible. The parameter h is related to the Hubble constant (H_0) by: $H_0 = 100 \times h \times kms^{-1}Mpc^{-1}$ and is currently measured to be around 0.67 [39].

[5] The annihilation cross section is often expressed as a thermal average reaction rate $\langle\sigma v\rangle$ where σ is the annihilation cross section and v is the relative velocity between the dark matter particles. For a more detailed discussion see [40].

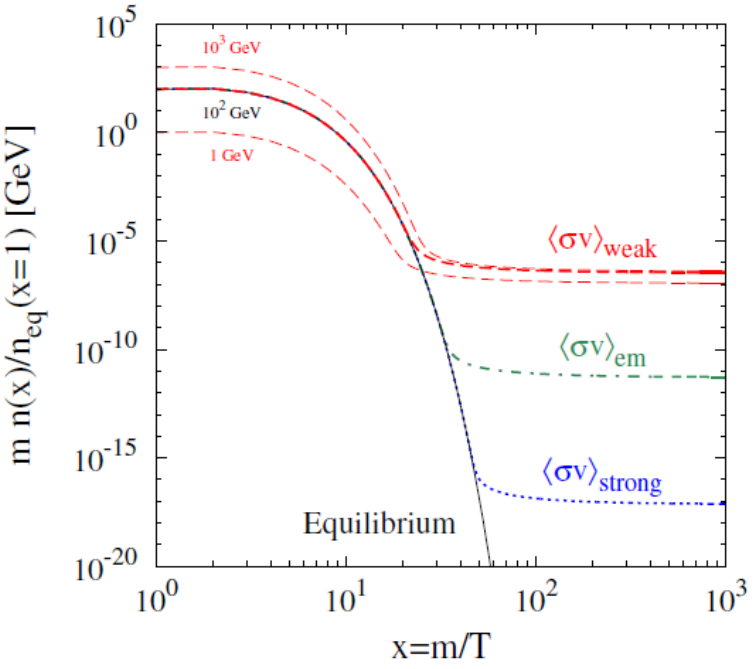


Figure 2.2: Evolution of the dark matter abundance (normalised to its initial density) as a function of x = m_χ/T, where T is the temperature of the universe [41]. As time evolves, the universe expands and the temperature drops. The equilibrium density (solid line) drops until the moment of freeze-out at which point the relic density remains constant. The different colors show the dependency of the relic abundance on different interaction cross sections (red, green and blue representing respectively an annihilation cross section of the order of the weak, electromagnetic and strong scale). The three red lines show the dependency of the relic abundance on the mass of the dark matter (ranging from 1 GeV over 10 GeV to 100 GeV).

value could be explained by a non-thermal dark matter history, but will not be considered here. A value lower than the measured value can still be explained by another source of dark matter that interacts with the Standard Model in a different way and this possibility will not be discarded. These assumptions will lead to a lower bound on the couplings since a small coupling would lead to a low annihilation rate and therefore to a relic abundance that is too high according to the measurements.

The calculation of dark matter relic abundances was performed with MadDM [42] and the resulting contour plots are shown in Figures 2.3, 2.4 and 2.5 for $c^{33} \neq 0$, $c^{23} \neq 0$ and $c^{22} \neq 0$ respectively. A curve in these plots connects the parameter values that results in the same relic density. In Figure 2.4 it can be seen that for the model described in Equation (2.4) and for relatively light dark matter masses (below 100 GeV), the couplings c^{23} needs to be sufficiently large (above 2). For masses below 80 GeV, the coupling would have to be unacceptably large (jeopardising the perturbativity of the theory) and by looking at Figure 2.5 it can be concluded that in order to keep the couplings below 10, also the cc$\chi\chi$ coupling (c^{22}) needs to be turned on to be consistent with the observed dark matter relic abundance. The annihilation of two dark matter particles to two charm quarks then offers a way to reduce

the relic abundance. Although the cc$\chi\chi$ coupling is needed for low dark matter masses, it will be extremely hard to detect at the LHC because of the large QCD multijet background. From now on the focus is only on the original flavour changing model of tc$\chi\chi$ interactions as described in Equation (2.4) and the flavour conserving cc$\chi\chi$ couplings will no longer be considered.

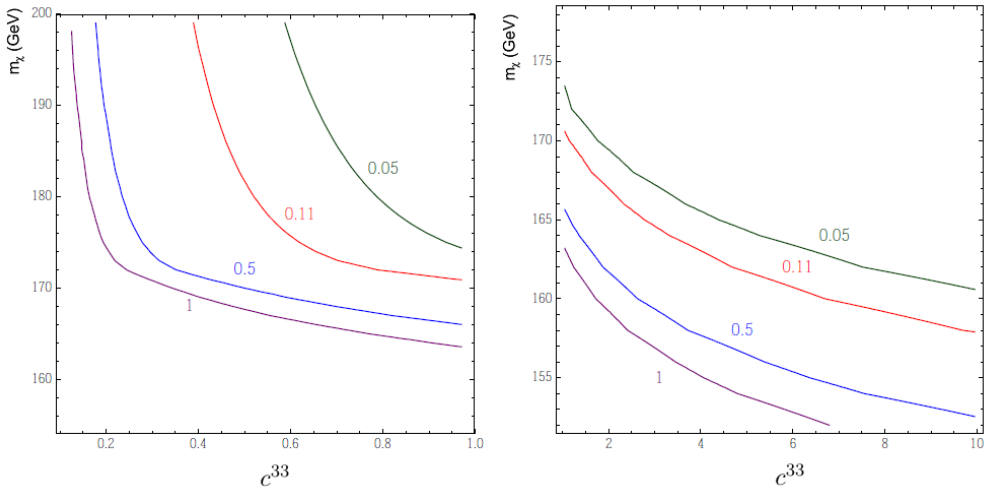


Figure 2.3: Contour plots of relic densities (Ωh^2) of fermionic dark matter coupling to the Standard Model via eq. (2.3) where $\Gamma = \gamma^\mu$ and $\Lambda = 1$ TeV and $c^{ij} = c^{33}$. The red line at relic density 0.11 indicates the contour for the measured relic density by Planck [39]. On the left c^{33} ranges from 0 to 1 and on the right from 1 to 10. All other c^{ij} couplings are set to 0.

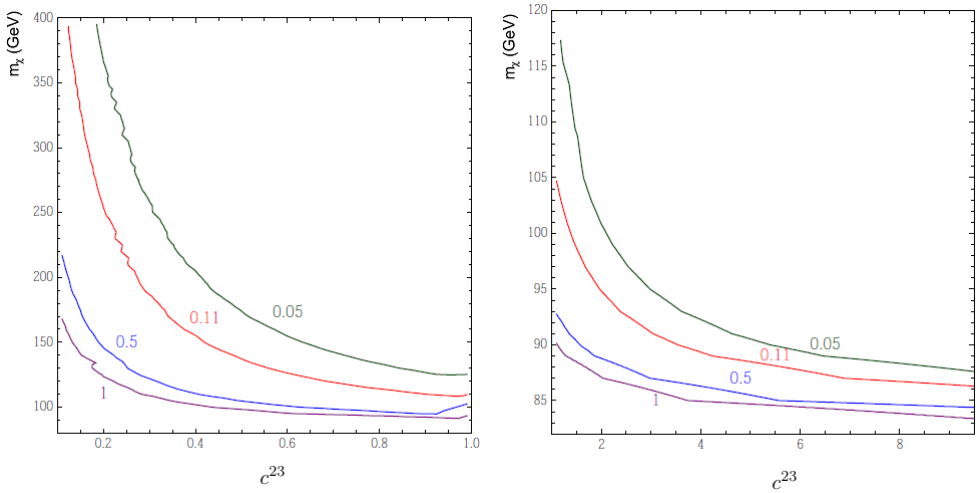


Figure 2.4: Same as in Figure 2.3, but for only $c^{23} \neq 0$ and all other couplings set to 0.

2.3 Constraints from direct and indirect dark matter searches

It is clear that the measured dark matter relic abundance already can constrain the different couplings between quarks and dark matter. However, once the cc$\chi\chi$ couplings are turned

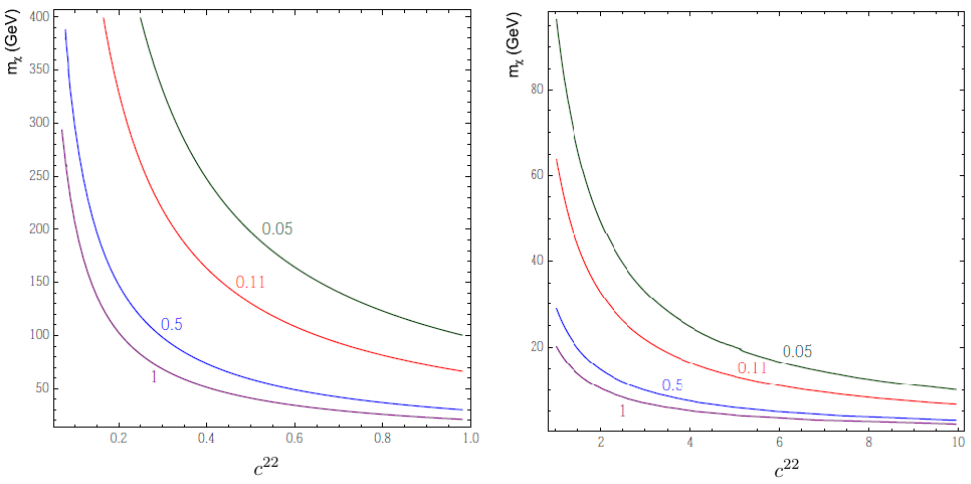


Figure 2.5: Same as in Figure 2.3, but for only $c^{22} \neq 0$ and all other couplings set to 0.

on, the tc$\chi\chi$ model is not restricted any more since the relic abundances are low enough by annihilation of dark matter to charm quarks. Other constraints also have to be taken into account. Although the flavour changing tc$\chi\chi$ couplings have not been studied in collider experiments in the past, the couplings strength could be constrained by other experiments such as direct and indirect dark matter searches.

In direct searches, the interactions between dark matter and nuclei of the detector material are probed. Direct detection is based on a measurement of the recoil energy of the nuclei and the interactions mostly conserve the flavour of the quark. Therefore these experiments only very weakly constrain the flavour changing tc$\chi\chi$ coupling. On the other hand, the cc$\chi\chi$ couplings are much more sensitive in direct detection experiments, since they only include flavour conserving interactions with light-flavour quarks. Hence, when the c^{22} coupling is turned on to satisfy the relic abundance requirement there can be non-trivial bounds resulting from direct detection experiments. The c^{22} couplings could also be induced by renormalisation-group running effects on the c^{23} couplings, leading to further constraints on the allowed parameter space. This requires a detailed investigation planned to be done in the near future, but which is outside the scope of this thesis.

Indirect detection of dark matter is based on the annihilation of dark matter to SM particles that can be observed. These are much more sensitive to the quark dark matter couplings in general. However, difficulty arises because obviously not the quarks themselves but rather the products produced in resulting showers are detected. An interesting example is the recent observation from the Fermi-LAT experiment [43] of the excess in γ-rays from the center of the galaxy between 1-10 GeV. A possible explanation could be the annihilation of dark matter, resulting into two photons. It is shown in [44] that the annihilation of two dark matter particles into a top and a charm quark (producing photons in the hadronisation process) has the potential to explain this excess within a certain range of thermal annihilation cross sections and dark matter masses. It is yet to be investigated how this translates to

the parameters of the model described in Equation (2.4) and it is not yet clear whether or not this is the true cause of the excess in γ-rays. Nevertheless it is a good example on how indirect detection experiments can constrain the model parameters of tc$\chi\chi$ couplings. It also provides a further motivation to study these flavour changing top-quark dark matter interactions at the LHC.

2.4 Processes in 13 TeV proton-proton collisions

One of the objectives of this research is to investigate the possibility of a collider search for tc$\chi\chi$ couplings at the LHC. In order to do so, one must first investigate the different possible processes and determine their cross sections. If these are of the order of 1 pb or above, one can expect to find these signatures in the next run of the LHC, in which collisions at 13 TeV are about to take place and a data set corresponding to an integrated luminosity around 100 fb^{-1} of new data will be gathered by the end of 2017. When sizeable cross sections are found, one still has to investigate the possible significance of such a signal over the SM background processes. This is done in Chapter 4 of this thesis.

The resulting production modes at the LHC can be divided into three subprocesses, also illustrated[6] in Figure 2.6:

1. *Top-pair production with flavour changing top-quark decay*
   $pp \to t\bar{t} \to (c\chi\chi)(bl\nu)$
   This process is similar to the SM $t\bar{t}$ production, but one of the top quarks has a flavour changing decay into dark matter and a charm quark while the other decays to a W boson and a bottom quark[7]. The branching ratios for this rare top-quark decay as a function of the DM mass are shown in Figure 2.7 for different coupling strengths c[23]. There is however a kinematic bound on this decay, which requires the mass of the dark matter particle to be smaller than half of the top quark mass ($m_\chi < \frac{m_t}{2} \simeq 86\ GeV$ [46]). This process is particularly interesting as it has never been studied before at the moment of writing (at either 8 or 13 TeV) and it opens up a new diagonal direction in Figure 1.3. When focusing on the leptonic decay of the W boson from the SM top quark decay, the detector signature of this process is 1 isolated electron or muon + MET + 2 jets (1 b jet + 1 c jet).

2. *Mono-top production with a final-state c jet*
   $pp \to t(\to bl\nu)c\chi\chi$
   The term mono-top typically refers to a single top quark production in association with new physics phenomena, in this case dark matter production which manifests itself as missing transverse energy. This process therefore contains one top quark that decays according to the Standard Model ($t \to bW$) and is accompanied by dark matter production and a charm quark. This leads to a detector signature identical to process

[6] It should be noted that the Feynman diagrams in Figure 2.6 serve only as an illustration of the processes and do not cover all possible diagrams. Many more diagrams should be included.

[7] In the Standard Model a top quark decays to a bottom quark and a W boson almost 100 percent of time [45]. This is due to the almost diagonal structure of the CKM matrix.

number 1 described above: 1 isolated electron or muon + MET + 2 jets (1 b jet + 1 c jet). However there is no kinematic bound on the mass of the dark matter for this process. Mono-top production in association with large missing transverse energy has already been studied at a centre of mass energy of $\sqrt{s} = 8$ TeV in a more general framework by CMS [47] and ATLAS [48].

3. *Mono-top production without a final-state c jet*
   $pp \to t(\to bl\nu)\chi\chi$
   This process is somehow similar to process 2 (mono-top production) but there is no final-state c jet. Therefore it has a different detector signature: 1 isolated electron or muon + MET + 1 jet (b jet). There is again no kinematic bound and as shown in the right feynman diagram in Figure 2.6, this process can be achieved with 1 QCD vertex less than the other two processes. Therefore it is expected that this process will have the largest contribution to the cross section of this phenomenological model.

Process 1 and 2 have the same detector signature and it will be extremely hard to distinguish between these two processes in an analysis. Process 3 can be distinguished from processes 1 and 2 by the jet multiplicity, but extra radiation can cause extra jets that spoil this clear separation. Therefore all processes should be considered inclusively when searching for tc$\chi\chi$ couplings. However during the analysis in Chapter 4, two signal regions are considered motivated by the different sensitivity of possible charm tagging methods. A first signal region consists of events with exactly one jet and a second signal region contains events with exactly 2 jets.

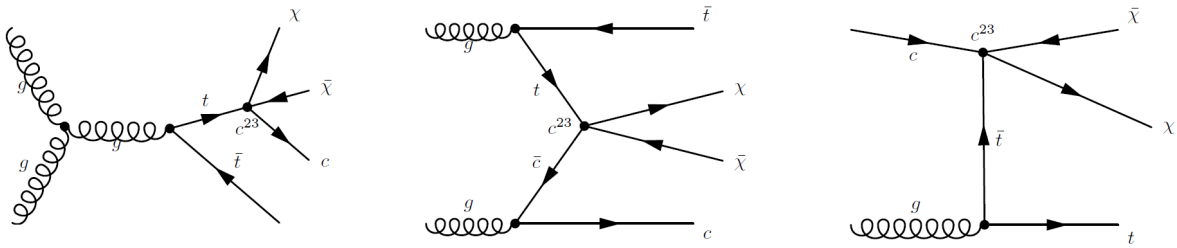


Figure 2.6: Illustration of the three categories of processes involving the tc$\chi\chi$ couplings, that can appear in proton-proton collisions at the LHC. Left: Top-pair production with flavour changing top-quark decay (process 1). Middle: Mono-top production with a final-state c jet (process 2). Right: Mono-top production without a final-state c jet (process 3).

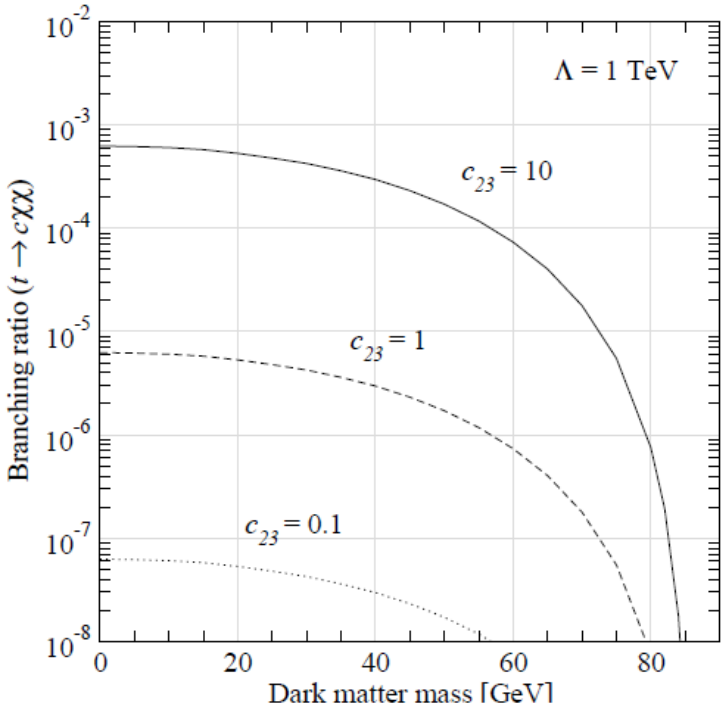


Figure 2.7: Branching ratio of the flavour changing top-quark decay into dark matter and a charm quark ($t \to c\chi\chi$) as a function of the dark matter mass and for different coupling strengths c^{23} at $\Lambda = 1$ TeV.

The cross sections at $\sqrt{s} = 13$ TeV for the 3 processes discussed above were calculated with MadGraph as a function of the dark matter mass for $\Lambda = 1$ TeV and $c^{23} = 10$ and are shown in Figure 2.8. Lower values for c^{23} result in a large drop in cross section since the cross section depends on $(c^{23})^2$. The sum of process 1 and process 2 is also plotted in red as they have the exact same detector signature. As expected, mono-top production without a final-state c jet (process 3) has the dominant contribution to the cross section. The cross section of the top-pair production with flavour changing top-quark decay drops to 0 at $m_\chi \gtrsim 86$ GeV as predicted by the kinematic bound on this rare top-quark decay. With these values of the model parameters, the cross section reaches between 0.1 and 1 pb, which is expected to be accessible in the next run of the LHC at $\sqrt{s} = 13$ TeV.

Based on the above considerations and more specifically on the kinematic bound on process 1, two benchmark point were selected to investigate a potential search for flavour changing interactions between top quarks and dark matter at the LHC at $\sqrt{s} = 13$ TeV:

1. $\Lambda = 1$ TeV, $c^{23} = 10$ and $m_\chi = 30$ GeV
2. $\Lambda = 1$ TeV, $c^{23} = 10$ and $m_\chi = 90$ GeV

The first benchmark point is at low dark matter mass, with the highest cross section and including the flavour changing top-quark decay. The second benchmark point explores a dark matter mass right above the kinematic bound, resulting in a lower cross section and lacking process 1. These two benchmarks will be explored in Chapter 4. As discussed earlier, process 1 and process 2 both contain a final-state c jet. Therefore these events with 2 jets

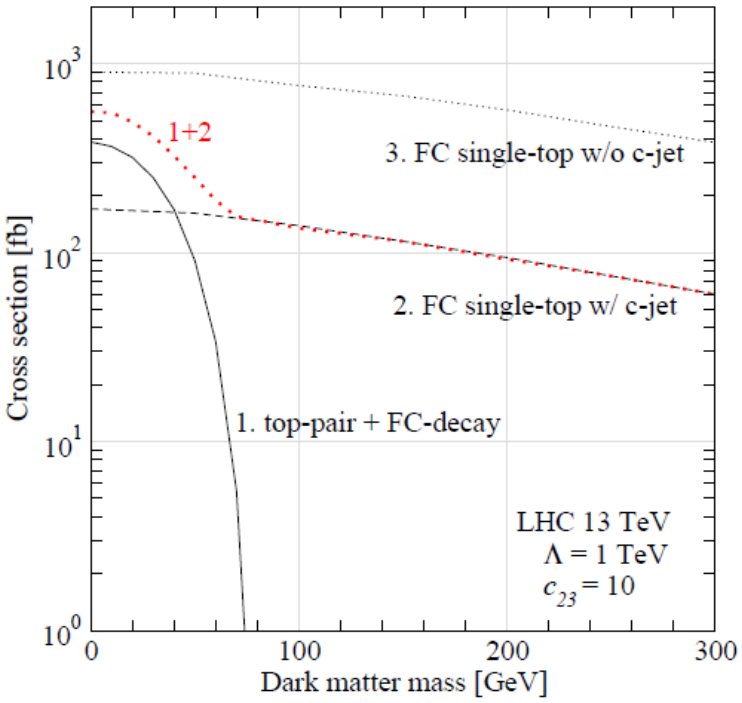


Figure 2.8: Cross sections for the three different processes in Figure 2.6 as a function of the dark matter mass for 13 TeV collisions at the LHC and for $\Lambda = 1$ TeV and $c^{23} = 10$.

could benefit from an algorithm to identify these jets originating from charm quarks. In Chapter 3 such a charm-tagging algorithm will be developed for the CMS experiment, which is new within CMS. In Chapter 4 the potential of that charm tagger to improve the search for tc$\chi\chi$ couplings at the LHC will be investigated.

Chapter 3

Development of a charm-tagging algorithm

The LHC is a proton-proton (pp) collider and will therefore produce a lot of QCD interactions in its collisions. As the protons collide, the partons inside the protons will interact and will lead in many cases to the formation of showers of hadrons[1]. These showers are detected in the calorimeter systems of the CMS detector and all of the shower constituents are clustered in jets by jet reconstruction algorithms. The flavour of the quark that lies at the origin of the jet will affect the properties of the jets. Heavy-flavour tagging algorithms aim at identifying the flavour of a jet, or equivalently the flavour of the originating quark. Within the CMS collaboration an algorithm to identify b jets ('b tagger') is already developed and widely used in physics analyses, but an algorithm to identify c jets ('c tagger') is not yet present. In this chapter the development of a c tagger for the CMS experiment is presented.

First the physics object reconstruction in CMS and the idea of heavy flavour tagging will be explained in more detail. After this the TMVA framework, in which a new charm-tagging algorithm is developed, is introduced in detail. Next, the performance of the charm-tagging algorithm is discussed in different situations (for different centre of mass energies, using different jet-properties and discriminating charm jets from either light-flavour or b jets). Finally the optimisation of the charm tagger is discussed and the performance is compared to a different c-tagging setup that is being investigated within the CMS experiment.

3.1 Physics object reconstruction with the CMS detector

3.1.1 Particle flow reconstruction

The particle flow (PF) algorithm [49] aims at reconstructing all stable particles inside the CMS detector. These include electrons, photons, muons, charged and neutral hadrons (being mostly pions, kaons, and neutrons). The PF algorithm combines and links all detector parts (inner tracker, calorimeter systems and muon systems), thereby improving the resolution on

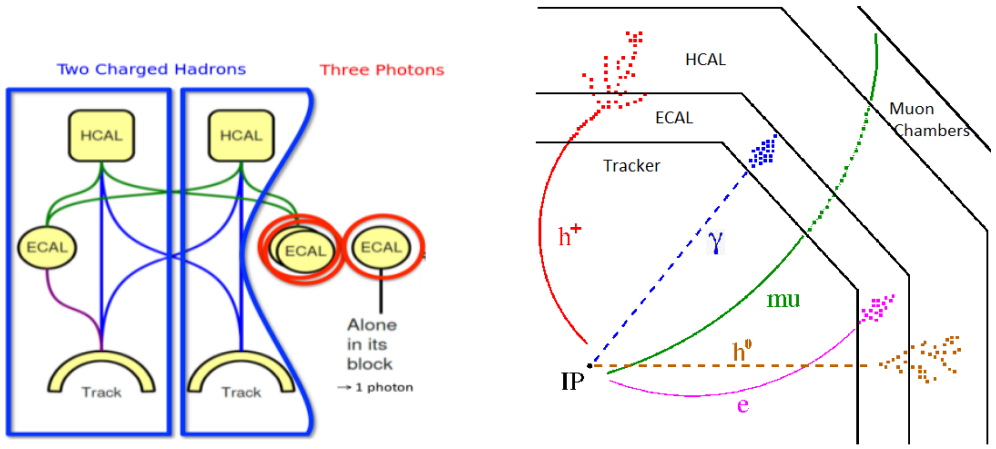
[1] The showering or fragmentation is a consequence of color confinement which states that at low enough energies, quarks are always found in bound states (baryons or mesons) and never as a single separate quark.

the energy, momentum and direction of all these particles. This has been studied and proven to give better resolutions in comparison to only using tracks, calorimeter clusters or hits in the muon system separately.

The particle flow reconstruction algorithm is performed in three stages. First the different basic elements are reconstructed. This consists of an iterative tracking algorithm to find all the tracks in the tracker, a clustering algorithm using the calorimeter information (both electromagnetic and hadronic) and searching for tracks in the muon system. Then associations (called blocks) are made between the different objects from different detector parts as illustrated in Figure 3.1(a). Tracks are extrapolated from the tracker to the calorimeter (ECAL or HCAL) and associated to a cluster if the extrapolated position lies within the cluster boundaries. Clusters in the electromagnetic calorimeter can be associated to the ones in the hadronic calorimeter if they also overlap. Finally, charged tracks can be extrapolated to the muon systems and combined into a global muon if the tracker tracks and the muon system tracks are compatible. In the final step the actual particle flow algorithm tries to further refine and interpret these associations in order to identify the final particles by their distinct detector signatures. This is schematically illustrated in Figure 3.1(b). First electrons [50] and muons [51] are identified by dedicated algorithms exploiting the fact that electrons leave tracks and ECAL energy deposits (with the possibility of radiating bremsstrahlung photons), whereas muons are constructed from combined fits between the tracker and the muon systems. Once these are identified they are removed from the blocks and the PF algorithm is left to identify photons, neutral hadrons and charged hadrons. The energy in the calorimeter clusters is compared to the energy of the associated tracks to that cluster and several situation may arise. If the calorimeter energy matches the energy of the tracks, only charged hadrons are identified. If the calorimeter energy is higher than that of the tracks, the charged hadrons are identified by the tracks and their energy removed. The remaining energy is used to construct either photons (especially when there is a lot of energy in the ECAL) and eventually neutral hadrons. In some cases the energy of the tracks is higher than the calorimeter energy, in which case tracks are removed by looking further for extra muons (with loosened reconstruction criteria) or eliminating tracks with high uncertainties. Finally one ends up with a set of final-state particles, which can be used for further analysis and to construct combined objects like jets (see next section) or to measure the missing transverse energy (MET) by taking the negative vector sum of the p_T of all PF objects.

3.1.2 Jet reconstruction

As mentioned during the discussion of the CMS ECAL and HCAL in Section 1.3.2, quarks produced in pp collisions are not directly observed by the CMS detector. Colour confinement dictates that these quarks will start a chain of fragmentation, producing many hadrons. The resulting set of hadrons are then detected by the calorimeter systems of CMS. The fact that one quark produces a stream of hadrons leads to the necessity to recombine all these hadrons into one object that can give information on the originating quark. These combined objects are called jets and jet reconstruction algorithms try to correctly combine and cluster these



(a) Blocks are constructed by associating different detector elements containing matching tracks and clusters [52].

(b) Each type of particle leaves a distinct signature while travelling through the detector.

Figure 3.1: Schematic illustrations of the particle flow algorithm.

hadrons.

Various jet reconstruction algorithms exist and all aim at clustering a set of particles (like particle flow objects) or energy deposits in the calorimeters into jets. In the environment of high energy pp collisions at the LHC, these algorithms should fulfil three basic requirements:

- As collisions appear every 25 ns, these algorithms should be fast (especially for the purpose of triggering).
- The algorithm should be insensitive to extra soft radiation (infrared safety).
- The algorithm should be insensitive to the collinear splitting of a hard hadron (collinear safety).

Two types of jet reconstruction algorithms can be distinguished. *Cone Algorithms* use a predefined cone around a seed particle and add all the particles inside that cone to form a jet. Often this procedure is repeated several times updating the seed using the result of the previous iteration, which is know as an iterative cone algorithm. Examples are the Iterative- Midpoint- and SIS-Cone algorithms [53]. A second category covers the *Sequential Recombination Algorithms* which also use a seed, but add particles based on a measure of distance until an isolated jet is constructed. These algorithms do not use a predefined cone, although ultimately they would also result in a cone shaped structure around the seed. The anti-k_T algorithm [54] is an example of such a Sequential Recombination Algorithm and is used most commonly in the CMS experiment (and also in this thesis). It has shown to be fast and collinear and infrared safe.

The anti-k_T algorithms starts from a seed particle and during each recombination step it measures both the distance to the beamline (d_{iB}) and to its closest particle (d_{ij}) as expressed in Equations (3.1) and (3.2) respectively.

$$d_{iB} = k_{Ti}^{2p} \tag{3.1}$$

$$d_{ij} = min(k_{Ti}^{2p}, k_{Tj}^{2p}) \frac{\Delta R_{ij}^2}{R^2} \tag{3.2}$$

$$\Delta R_{ij} = \sqrt{\Delta y_{ij}^2 + \Delta \phi_{ij}^2} \tag{3.3}$$

In these equations, k_{Ti} and k_{Tj} are the transverse momenta of particle i (usually the seed) and particle j respectively. ΔR_{ij} represents the distance in the plane defined by the rapidity y and the azimuthal angle ϕ. The parameter R is referred to as the radius or the distance parameter of the jet reconstruction algorithm and the parameter p regulates the treatment of the transverse momenta in terms of a distance. For the anti-k_T algorithm, p is set to -1. During each recombination step the algorithm determines the minimum amongst both distances (d_{ij} and d_{iB}). If d_{ij} is the minimum, the particles are clustered and the seed is updated to the combined particle. If d_{iB} is the minimum, the combined object is well isolated from other particles in the event and is clustered as a jet. The particles used to create the jet are removed from the list and the algorithm starts again with a new seed to potentially form a new jet. It is interesting to note that the anti-k_T algorithm produces cone-like shapes (see for example Figure 3.2), which allows for an easier interpretation of the distance parameter R as an actual cone radius.
When the jet is created its energy can be determined, which is in practice a challenging procedure which requires the necessary corrections to the jet energy due to pile-up, detector response, reconstruction efficiencies etc... The anti-k_T algorithm is discussed in more detail in reference [54], where also its performance is compared to other jet reconstruction algorithms.

The simulated event samples used for the investigation of the charm-tagging performance (see Section 3.4.3) all use the anti-k_T algorithm. The samples at 8 TeV centre of mass energy use a distance parameter of $R = 0.5$, whereas the samples at 13 TeV use $R = 0.4$. This is motivated by the fact that for higher energies the decay products in jets will be boosted more towards the direction of the jet axis, creating a narrower cone than would be the case at lower energies. A smaller cone radius is advantageous, as it is more robust against pile-up events. Only jets with a p_T larger than 30 GeV are considered in the development of the charm tagger.
When studying the effect of c tagging on the search for flavour changing top-quark dark matter interactions, other simulated samples are used created by Delphes detector simulations. The object reconstruction in these simulations is different and will now briefly be discussed.

3.1.3 Object reconstruction in Delphes

The Delphes software [55] serves as a fast simulation software for generic collider experiments, in this case for the CMS detector. Whereas simulating the CMS detector in all its complexity is a time-consuming operation, a series of simplifications allow the user to run much faster but slightly less accurate simulations of the detector response to collision events. This results in simplified object reconstruction techniques, different from the ones mentioned above. A

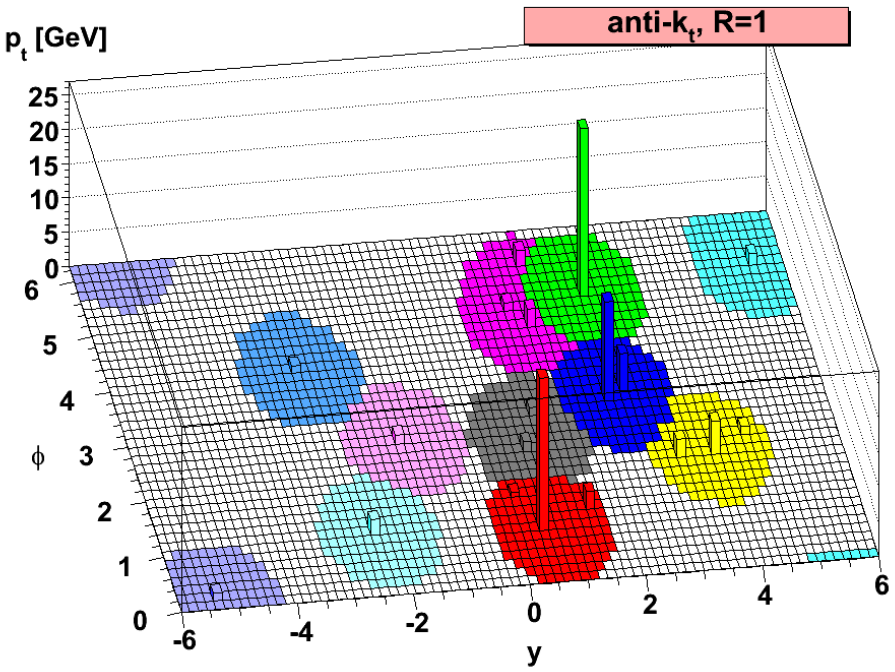


Figure 3.2: Visualisation of the output of the anti-k_T jet reconstruction algorithm on a simulated event [54].

detailed discussion can be found in reference [55], but a short summary is given below.

The particle flow reconstruction is also included in Delphes, but in a simplified version using so-called PF tracks and PF towers. The former simply corresponds to reconstructed tracks within the Delphes simulation with corresponding energy deposits in the calorimeter systems, the latter represents calorimeter energy deposits not related to tracks, i.e. neutral particles. Electrons and muons are reconstructed separately by applying (combined) reconstruction efficiency curves (in terms of p_T and η) to the generator level electrons and muons. Their reconstructed p_T is smeared out by means of a resolution curve.
From the PF objects, PF jets are clustered using the anti-k_T algorithm with $R = 0.5$. A combined correction to the jet energy, that accounts for all the possible effects, can be applied by the user as well.

3.2 Secondary vertex reconstruction: inclusive vertex finder (IVF)

Decays of for example B- or D-mesons (containing b or c quarks respectively) in the CMS detector will result in displaced tracks originating from displaced or secondary vertices (SV) with respect to the primary vertex (PV). Such displaced tracks have a large impact parameter (IP), meaning the distance of closest approach from the PV to the reconstructed track is on average larger than for tracks that originated from the PV itself. To reconstruct these SV an algorithm has to identify these displaced tracks, cluster them and calculate the position

of the originating secondary vertex as illustrated in Figure 3.3. The Inclusive Vertex Finder (IVF) algorithm is designed to do so, independent of any type of jet reconstruction. This algorithm consists of four subsequent steps as outlined below [56].

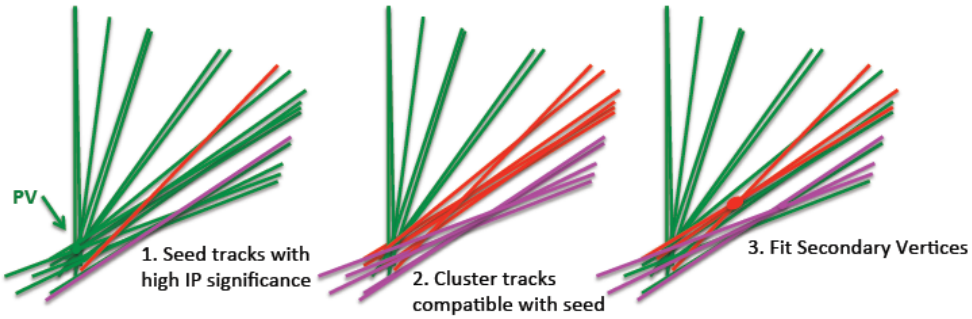


Figure 3.3: Visualisation of the general idea of the IVF algorithm: first seed tracks with large IP are identified (left), then tracks that are compatible to originate from the same SV as the seed are clustered together (middle) and finally these clusters are fitted to reconstruct a SV. In practice some intermediate steps are needed to clean up the resulting set of SV as outlined below [56].

Step 1 First of all seed tracks are identified based on a large impact parameter (significance[2]) with respect to the PV (but not in the direction along the beampipe to avoid tracks from pile-up). The parameters related to this selection are called *seedMin3DValue* and *seedMin3DSignificance*. For each of these seeds a clustering algorithm iterates over all the tracks in the event and clusters them with the seed based on the distance of closest approach between the seed and the track. This clustering criterion is based on the *distanceRatio*[3] parameter, but also includes other parameters like impact parameters and mutual angles amongst the tracks. This results in a collection of clustered tracks. Each of these clusters is passed to a vertex fitting algorithm (AVR [57] or AVF [58]) which tries to calculate the most probable (secondary) vertex consistent with all tracks in that cluster. This finally results in a collection of secondary vertices that might share a lot of common tracks since nothing prevents the algorithm to cluster tracks with multiple seeds, or even to cluster different seed tracks together. After the set of SV is present, a cut is applied on the distance (significance) between PV and SV which is a measure of the decay length of the meson (i.e. the distance the meson travels before it decays) and is consequently named *vertexMinDLenSig* (or *vertexMinDLen2DSig* for the projection on the x-y plane).

Step 2 The multiple usage of tracks in different SV is reduced by iterating over all SV and dropping vertices that have too much in common. This decision is made based on the fraction of the tracks that are shared amongst two SV (*maxFraction*) and the distance significance between those two SV (*minSignificance*). This results in a more unique list of SV, which share at most a fraction of jets defined by *maxFraction* and are separated

[2]The significance of a parameter refers to the ratio of its value to its uncertainty.

[3]The cut on the distance between track and seed (d_{ts}) is taken relative to the distance between track and PV (d_{PV}) and is defined as: $d_{ts} \times distanceRatio < d_{PV}$.

by a distance larger than defined by *minSignificance*.

Step 3 The third step iterates over the new set of secondary vertices and calculates the flight direction and decay length of each vertex. All tracks used in the SV reconstruction are now reconsidered and checked whether they are more compatible to that SV than to the PV. This decision is based on the impact parameter of the track and the angular distance $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ between the track and the flight direction (i.e. the direction of the vector pointing from the PV to the SV). If at least two tracks are assigned to the SV rather than the PV, they form a new collection of displaced tracks and once again a SV is fitted using these tracks but this time using prior information from the previously fitted SV. A new set of refitted SV is now available, which could again share common tracks.

Step 4 The collection of SV from step 3 has again an overlap in tracks assigned to different vertices. Step 4 is a repetition of step 2 on the set of refitted SV, but with much tighter cuts on the fraction of shared tracks and the distance between the vertices. This results in a final set of reconstructed SV. One last cut is applied on the flight distance (distance between PV and SV) of these final SV, called *distSig2dMin*.

The IVF algorithm thus creates a set of secondary vertices from displaced tracks without using directly information from reconstructed jets. The default settings (used during this study) of the IVF configuration can be found in Table 3.1. Only during the discussion of the optimisation of the charm tagger, some of the IVF parameters will be varied and the effect on the vertex category composition and charm-tagging performance will be studied. It should however be noted that the IVF algorithm is in practice more complex than the outline given above, and many more parameters show up during the entire process. Table 3.1 only mentions the ones explained in the outline above. Also additional cuts on the tracks used in the algorithm are applied and can be found in [59].

IVF Parameter	**Default Value**
seedMin3DValue	0.005 cm
seedMin3DSignificance	1.2
distanceRatio	20
vertexMinDLenSig	0.25
vertexMinDLen2DSig	1.25
maxFraction	0.7 (0.2 in step 4)
minSignificance	2 (10 in step 4)
distSig2dMin	1.5

Table 3.1: Overview of the default values used for the parameters in the IVF algorithm for the charm tagger.

The reconstruction efficiency of a displaced vertex from a B- or D-meson depends on the values of the parameters used in the algorithm. In heavy flavour taggers the values of these parameters are chosen depending on the type of discrimination one wants to achieve, as will be explained later on. The default values chosen for c tagging (see Table 3.1) are based on the defaults from b tagging, but keeping in mind the shorter lifetime of D-mesons. In the

development of the charm tagger discussed in this thesis the IVF algorithm configuration results in three vertex categories based on the reconstruction of a secondary vertex:

- **RecoVertex:** A jet with at least two tracks in which one or more secondary vertices were reconstructed using the IVF algorithm (see Table 3.1) ends up in the *RecoVertex* category.
- **PseudoVertex:** In jets with at least two tracks in which no IVF secondary vertex can be reconstructed, an attempt is made to reconstruct a pseudovertex. This reconstruction uses only tracks from the reconstructed jet and is therefore dependent on the jet reconstruction, unlike the IVF algorithm. A selection of tracks is made, which are all required to have a 2D IP significance of at least 2. If at least two such highly displaced tracks are found, an additional cut requires the invariant mass of any two such tracks to be at least 0.05 GeV away from the K^0 mass[4]. If this succeeds the jet is assigned to the *PseudoVertex* category. Note that a pseudovertex does not contain a reconstructed secondary vertex as no vertex fitting algorithm is applied to the selected tracks. It is merely a collection of highly displaced tracks that is interpreted as possibly originating from a point different from the PV.
- **NoVertex:** Finally any jet that does not end up in the *RecoVertex* or *PseudoVertex* categories, is assigned to the *NoVertex* category. Consequently no SV was reconstructed for these jets.

These vertex categories can be extremely helpful for heavy-flavour tagging algorithms as will now be explained in more detail.

3.3 Heavy flavour tagging

3.3.1 Combined secondary vertex (CSV) algorithms

Jets inside the CMS detector are divided in three categories: b jets originating from bottom quarks, c jets originating from charm quarks and light-flavour[5] jets originating from either up, down or strange quarks or from gluons. B-mesons (containing a bottom quark) have a relatively long lifetime and will travel a few hundred μm inside the CMS tracker system before they decay. Besides that, the mass of B-mesons is considerably larger than that of D-mesons (containing a charm quark), who also have a shorter lifetime. Therefore D-mesons will also travel some distance (although shorter than B-mesons) before they decay. In these decays, the more stable and lighter mesons like pions and kaons are produced, which will either decay themselves (into leptons or photons) or live long enough to be detected by the calorimeter systems of CMS. For light-flavour jets, most tracks are produced at the

[4]The cut on the K^0 mass window is defined as: $|m_{i,j} - m_{K^0}| > K0MassWindow$, where i and j loop over all the selected tracks ($i \neq j$), m_{K^0} is the neutral kaon mass and K0MassWindow is the applied cut. This cut is also applied on jets from the RecoVertex or NoVertex categories, but the value of K0MassWindow is lower (0.03 GeV) for these categories.

[5]Light-flavour stands for either down, up, strange quark or gluon and will sometimes be referred to as *DUSG*.

interaction point itself. Before discussing the implication of these properties for heavy-flavour tagging algorithms, a brief revision of some properties of the interactions themselves will be useful.

At the point of interaction between two colliding protons, an interaction vertex or primary vertex is created from which all the particles created in the collision will originate, as shown in Figure 3.4. Charged particles will interact with the tracker system of CMS and will create tracks. The impact parameter (IP) of such a track denotes the distance of closest approach from the reconstructed track to the primary vertex. Particles originating from the PV are expected to have a small IP. However, when a particle first travels some distance before it decays, the tracks of the decay products may be displaced with respect to the PV. This will result in a large IP. It is said that the relatively long-lived particle created a secondary vertex at the point of its decay, and its decay products produce displaced tracks in the detector.

Heavy-flavour tagging algorithms exploit the different properties mentioned above to identify the flavour of a jet. The longer the lifetime of the unstable meson (B- or D-meson), the larger the distance between the PV and a possibly reconstructed SV. Also the IP of the corresponding displaced tracks will be larger on average for mesons with a longer lifetime. Light-flavour jets will however not create a secondary vertex and consequently also no displaced tracks. These properties can be translated in many variables that can be measured with the CMS detector. These variables can be put in three separate categories:

1. **Secondary vertex related variables:** in case a secondary vertex was reconstructed (see Section 3.2), the properties of that SV can be used to distinguish between c , b- or light-flavour jets. Examples are the distance between the PV and the SV (which is the decay length of the meson), the reconstructed mass of the SV, the number of displaced tracks that make up the SV, etc...
2. **Displaced-track related variables:** a secondary vertex is potentially reconstructed from displaced tracks and the properties of these tracks are also used by heavy-flavour tagging algorithms. Examples are the IP of these tracks, p_T of the tracks w.r.t. the p_T of the jet, distance of the track to the jet-axis, etc...
3. **Soft lepton related variables:** throughout the decay chain, mesons can have leptonic decays and produce electrons or muons. Information from these low-energetic (soft) leptons could be different for different mesons and can also be used as a discriminating variable. This can be information on the p_T, IP, direction, etc... of the electron or muon.

A more detailed description of the variables that are used in the charm tagger is given in Appendix A. The Combined Secondary Vertex (CSV) taggers in CMS combine all the kinematic properties from the secondary vertices, displaced tracks and potentially from soft leptons in multivariate analysis (MVA) techniques to obtain the most efficient discrimination between b jets, c jets and light-flavour jets. These MVA techniques typically map all the above mentioned kinematic properties of a jet onto a single discriminator value. The stronger the discrimination in the separate kinematic variables, the stronger the distinction in the

final discriminator distributions from the MVA and the better the performance of the heavy flavour tagger. In the following section the goals, properties and difficulties of a charm tagger will be mentioned in more detail.

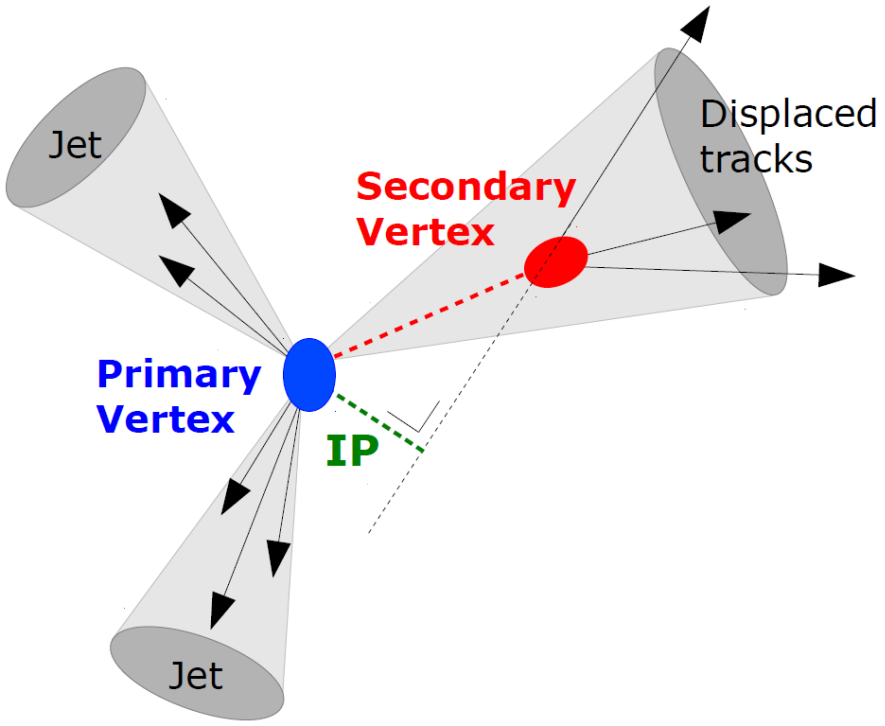


Figure 3.4: Illustration of an interaction with three jets of which one jet contains a displaced secondary vertex. The displaced vertex is created after the decay of a B- or D-meson inside the jet, and heavy-flavour tagging algorithms use the information form the displaced tracks and the reconstruction of the secondary vertex to identify the flavour of the jet.

3.3.2 Charm tagging

There are many examples of interesting physics processes, either SM processes or new physics models, with final-state c jets. Examples are the decay of the recently discovered Higgs boson ($m_H \approx 125$ GeV [10]) into a charm quark pair, SUSY models where scalar quarks decay to charm quarks, flavour changing neutral current (FCNC) interactions in the top-quark sector or the model of flavour changing interaction between top quarks and DM introduced in Chapter 2. However, without the presence of a dedicated algorithm to identify jets from charm quarks, these processes could be overwhelmed by large SM backgrounds from processes with lots of light-flavour or b jets such as QCD multi-jet production, vector boson production with jets or top pair production. The existence of a b tagger which is already used by CMS in many analyses and has shown to be of great value, can already reduce the SM backgrounds significantly. However, beyond the tagging of b jets, these searches become insensitive to jet flavours. Therefore the next logical step is to develop a dedicated charm tagger to distinguish c jets from b jets and light-flavour jets.

A charm tagger exploits the properties of D-mesons[6], which have a lifetime that is approximately half of the lifetime of B-mesons[7]. This results in smaller impact parameters of the displaced tracks and eventually in a decreased performance of the secondary vertex reconstruction. Also, most of the kinematic properties of c jets are distributed on average somewhere in between the same distributions for light-flavour jets and for b jets. This will make it even harder to distinguish them simultaneously from light-flavour jets and from b jets. Finally, to make things even more complicated, B-mesons often have decay chains via D-mesons, which can fake c jets. All of this information illustrates the fact that a charm tagger will have an overall worse tagging performance than similar b-tagging methods. Nevertheless the extra information from charm tagging may improve the sensitivity of a search for processes with final-state c jets.

3.4 Discussion of the charm-tagging setup

3.4.1 Multivariate analysis (MVA) and the TMVA framework

A heavy-flavour tagging algorithm exploits the differences in jet properties between light-flavour, charm and bottom quark jets. Applying cuts to discriminating variables allows to separate the signal (c jets in case of a charm tagger) from the background (light-flavour or b jets), but often applying successive cuts to multiple variables is not the most optimal way to separate signal from background. When dealing with a large set of multiple variables, Multivariate Analysis (MVA) techniques provide smarter solutions to improve the discrimination. These techniques keep in mind possible correlations and discriminating power of the variables and combine all of the available information to maximise the discrimination between signal and background. Typically these MVA techniques provide a map from the input variables (jet properties) onto a single discriminator value of which the distribution for signal and background has the highest possible separation. A single cut on the output discriminator distribution will yield in general a better discrimination between signal and background than successive cuts on all the variable distributions separately. This provides the user with a very easy technique (one simple cut) to optimise the performance of an analysis. There are different kinds of MVA techniques, all of which have advantages and downsides. The chosen MVA method for this research project and a motivation for this choice are given in Section 3.4.2.

The investigation of the performance of an MVA based analysis happens in two steps. First a training of the MVA is performed, during which the MVA determines the best way to discriminate signal from background and creates a map from the input variables to the output discriminator value. In a second step (called the validation step) the trained MVA is applied to a test sample to determine the discriminator distributions and the performance of the MVA. For heavy flavour taggers the performance is typically presented in terms of

[6] Other hadrons may also contain charm quarks, but are produced less frequently. An example is the Λ_c baryon.

[7] For B-mesons, the decay length expressed as $c\tau$ (where τ is the meson lifetime) lies roughly between 450 and 500 μm. For charged (neutral) D-mesons this is only about 312 (123) μm [45].

a ROC[8] curve, which expresses the efficiency of the signal selection with respect to the efficiency of background selection. For a charm tagger the signal efficiency is the efficiency of identifying a jet from a charm quark correctly as a c jet (ϵ^C), whereas the background efficiency expresses the efficiency of mistakenly tagging either a light-flavour (ϵ^{light}) or a b jet (ϵ^B) as a c jet. In this context the term efficiency denotes the number of signal of background jets that passes the cut on the MVA discriminator output divided by the total number of signal or background jets. Sometimes the rejection (rejection = 1/efficiency) of signal or background is used rather than the efficiency.

TMVA [60] is a software package inside the ROOT data analysis software for Multivariate Analysis techniques. It provides a variety of MVA techniques to be applied within ROOT and can be used as a standalone application. This research investigates a setup of the TMVA framework, which is completely standalone and not dependent on the CMS software framework (CMSSW). This provides more freedom, but also introduces the necessity of developing an interface that allows the CMS software to communicate with this standalone c-tagging setup. In parallel to this, also a CMSSW-integrated c-tagger setup is being investigated within the CMS collaboration and will later on be compared to the standalone TMVA setup discussed in this thesis. In the following, the standalone TMVA based setup will simply be referred to as the *TMVA c tagger* whereas the CMSSW integrated setup will be referred to as the *CMSSW c tagger*. These two setups differ by the choice of the MVA method and by the fact that the TMVA c tagger performs the training and validation inclusively for all secondary vertex categories together, whereas the CMSSW c tagger trains and validates separately for all vertex categories and combines the MVA output in the end.

3.4.2 MVA technique: boosted decision trees

As mentioned above, TMVA offers a variety of MVA methods to be used within ROOT. Some of the most well-known ones are Likelihood Ratios (LR), Artificial Neural Networks (ANN) and Boosted Decision Trees (BDT), but many others are available and can be found in the TMVA Users Guide [60]. Each of the above mentioned methods has its advantages, but in general they can be categorised by their complexity and expected performance.

Likelihood Ratio methods are the most straight-forward because they simply use likelihood estimators (combinations of variable probability density functions) to construct the ratio of the combined signal likelihood to the sum of the combined signal and background likelihoods. This provides an analytical description of the method which makes it more easy to understand what is going on inside the MVA algorithm. This high comprehensibility makes it easier to tune the MVA parameters to optimise its performance. However, a LR is generally not expected to give the best performance (it can for example not very effectively include the correlations between variables). The other extreme is the Artificial Neural Network which is based on a complicated algorithm, but it has shown to give in general the most optimal performance. An ANN combines different layers each consisting of a variety of nodes or neurons. Each input variable makes up an input node to the ANN, after which a series

[8] ROC stands for receiver operating characteristic.

of hidden layers processes the information and finally feeds this information to the output nodes. Typically there are two output nodes; one for signal and one for background, but in principle there could be more output nodes[9]. One might think that the good performance of an ANN makes it the obvious choice and the user does not care about the complexity of the algorithm on which it is based. However that complexity makes it hard to tune the MVA parameters to optimise the performance and it might be more difficult to interpret the results or find out the cause of unexpected behaviour of the ANN. This argument serves as a motivation to use a Boosted Decision Tree in the development of the TMVA c tagger, together with the fact that it serves as a complementary alternative to the CMSSW c tagger that uses an ANN.

A BDT is still rather complex, but the principles on which it is based are much more intuitive than an ANN making the BDT parameters more understandable and easier to tune to optimise the performance. Its performance, although in general slightly worse than an optimised ANN, has proven to be still very good. It should however be stressed that the statements about the general performance can vary in different analyses and it could be that certain MVA methods are better than others for certain applications. A BDT was used in the TMVA based c tagger and to understand how a BDT works, three concepts have to be explained in more detail: Decision Trees, Boosting and Bagging.

Decision trees A decision tree uses successive decision nodes to categorise events as either being signal or background. Starting from a root node, a cut is applied to the full event sample set on one of the input variables that most efficiently distinguishes signal from background. This cut creates two new nodes, each consisting of a subset of the original sample of which one is more signal-like and the other is more background-like. This cutting procedure is repeated for each of the sub-nodes and stops only when one of the nodes reaches a minimum number of events or a certain signal purity. These final *leaves* are then identified with either signal or background depending on whether they contain more signal or background events.

Boosting Boosting is actually a weighting procedure in which the signal events that ended up in a background leaf or vice versa are given a larger weight than the ones that ended up in the correct leaf. This way these events are considered more important as they cause possible mistakes in the identification of signal or background. With this new weighed sample, a new decision tree is trained and this procedure of boosting is repeated a number of times to finally end up with a so-called *forest* of trees.

Bagging While boosting a decision tree, information from the previous tree (the obtained weights) is used in the next training tree. Bagging is a variation on or addition to boosting, in which each new tree uses a stochastic re-sampling of the original full sample. This means

[9] As an example applicable to c tagging, one might expect three output nodes; one for c jets (signal), one for b jets and one for light-flavour jets (both background). This would allow to have a combined c tagger that is able to most optimally discriminate c jets from both light-flavour and b jets at the same time.

each tree uses a sub-sample containing only a fraction of all the events. Bagging is used during the analysis presented in this thesis.

After the BDT is trained, events can be successively subjected to the different decision trees and are assigned an estimator (often called discriminator) value based on the number of times they end up as signal or background. This is how a trained BDT is used in analysis or during the validation step of the investigation of the performance.

In Table 3.2 some relevant options for the BDT that are explicitly set during the training of the c tagger are explained and their default value is given. Note that default does not refer to the default value set by TMVA, but rather refers to the value which is used in the discussion of the performance in this thesis. Only when optimizing the performance of the charm tagger, different BDT settings will be investigated. All options that are not explicitly written down in Table 3.2 are either kept at the TMVA defaults (which can be found in the TMVA Users Guide) or are options that are not turned on in the presented research in this thesis.

BDT Option	**Default Value**	**Description**
NTrees	1000	The number of trees used in the boosting algorithm to build up the forest of decision trees.
nCuts	80	The number of points in the input variable range to find the most optimal cut in the splitting of a node.
MinNodeSize	1.5%	The minimum fraction of events (with respect to the full sample set) required in each node. Once a node contains less than this fraction the node splitting stops and it becomes a final leaf.
BoostType	Grad [10]	The type of boosting used for the trees in the forest.
Shrinkage	0.1	Learning rate for the gradient boosting (Grad) algorithm.
UseBaggedGrad	True	Use bagging within the Gradient boosting algorithm. Each tree in the forest will use only a subsample of all the events.
GradBaggingFraction	0.5	The (stochastically chosen) fraction of events used in each tree in the forest when using bagging.
MaxDepth	2	The maximum depth of each tree in the forest. This can be seen as the maximal amount of subsequent node splittings before constructing a final leaf of the decision tree.

Table 3.2: Explanation of the BDT options that are explicitly set during the c tagger training and their default values used throughout the discussion of the c tagger performance [60].

[10] Gradient boosting is a type of boosting in which the mistakes or residuals (signal seen as

3.4.3 Simulated event samples

During the development of the charm tagger different event samples have been used:

- QCD multi-jet training samples and $t\bar{t}$ validation samples at 8 TeV to determine the c-tagger performance for 8 TeV pp collisions. These samples are produced in the 53X version of CMSSW.
- QCD multi-jet training samples and $t\bar{t}$ validation samples at 13 TeV to determine the c-tagger performance for 13 TeV pp collisions. These samples also contain soft lepton information and are used throughout the entire study, except when studying the performance at 8 TeV or for the optimisation of the IVF settings (see below). These samples are produced in the 70X version of CMSSW.
- QCD multi-jet training samples and $t\bar{t}$ validation samples at 13 TeV with soft lepton information, produced in the 73X version of CMSSW. These samples are slightly different from the 13 TeV samples in the 70X CMSSW version and were used during the optimisation of the IVF settings.

One important remark is that the QCD multi-jet samples are skimmed. This means only a maximum amount of jets (20.000) are kept in each bin of p_T and η. This binning is predefined (and arbitrary) and is kept very broad. This skimming process is introduced to speed up the training of the MVA. However the skimming will distort the distributions of the different vertex categories (RecoVertex, PseudoVertex and NoVertex), a problem that will be dealt with in the next section by applying weights to the events.

3.4.4 Biases and additional weights

When training the c tagger, some additional weights are applied to the jets. These weights are introduced to achieve two goals. First of all the c tagger should be physically correct, meaning that the training and the validation samples (and all their variable distributions) should represent real physics processes such that the performance of the c tagger is also representative for real physics analysis applications. Secondly, although somehow related to the first requirement, the c tagger should be as generic as possible. It should perform well on a variety of physics analysis and should not depend too much on the kinematics of the training sample. The obtained performance curves are in principle only applicable to the validation sample that was used ($t\bar{t}$ in this thesis), but can easily be reproduced for any desired sample. The training however is performed on one sample (QCD multi-jet in this thesis) and should not be changed once a c tagger is put in operation for CMS. The two requirements mentioned above lead to two types of additional weights which will now further be explained.

$t\bar{t}$ bias weights The training of the c tagger is performed on a QCD sample because of the large variety in jet p_T which makes it a generic training sample. The validation is

background or vice versa) in the previous tree are minimised in the next tree in the forest such that after each iteration signal and background are more clearly separated. For more information see for example [61].

performed on a $t\bar{t}$ sample because it is more representative for present-day physics analyses. The discrepancy between these two samples could however affect the performance of the c tagger in a negative way. Especially since the vertex category distributions for both samples could differ a lot. The fact that the QCD samples are skimmed will make the QCD vertex distributions even more physically incorrect. Therefore a first type of weight is introduced to somehow bias the vertex distributions in the QCD training samples towards those of the $t\bar{t}$ validation samples. These weights are therefore referred to as $t\bar{t}$ biases and are calculated separately for each flavour (c, b or light) and in bins of p_T and η[11]. The calculation of these weights happens in two steps. First the QCD vertex distributions are normalised to one for each category by simply adding a weight (w^{norm}_{QCD}) equal to the inverse of the bin content for that vertex category ($N^{QCD}_{f,p_T,\eta,cat}$) as expressed in Equation (3.4) where f stands for a specific jet flavour and cat for a specific vertex category.

$$w^{norm}_{QCD}(f, p_T, \eta, cat) = \frac{1}{N^{QCD}_{f,p_T,\eta,cat}} \tag{3.4}$$

After this normalisation, the vertex distributions still have to be biased towards those of the $t\bar{t}$ sample. This introduces a second weight ($w^{bias}_{t\bar{t}}$) for each category and flavour which is the number of jets for that vertex category in the $t\bar{t}$ sample normalised to the total number of jets for all vertex categories inclusively (but still separately for each flavour and each p_T and η bin), as expressed in Equation (3.5). The combination of these two weights forms the global $t\bar{t}$ bias and this procedure is schematically shown in Figure 3.5.

$$w^{bias}_{t\bar{t}}(f, p_T, \eta, cat) = \frac{N^{t\bar{t}}_{f,p_T,\eta,cat}}{N^{t\bar{t}}_{f,p_T,\eta,incl}} \tag{3.5}$$

p_T - η flattening weights Secondly, the training should be insensitive to the kinematic properties of the QCD training samples, such as the jet p_T and jet η. Therefore a second weight ($w^{p_T,\eta}_{QCD}$) is introduced that makes the QCD distributions flat for all flavours separately but inclusively for all the vertex categories (meaning the distributions are made flat for the histograms of the NoVertex, PseudoVertex and RecoVertex summed together). To compute these weights, a binning has to be chosen for the histograms of the jet p_T and the jet η to calculate the weight for each jet as the inverse of the bin content of its p_T and η bin ($N^{QCD}_{p_T,\eta,f,incl}$). This is expressed in Equation (3.6). The flattening of the p_T and η spectra is illustrated in Figure 3.6. In practice a 2-dimensional histogram is created that contains the weight for each possible bin in p_T and η.

[11] The binning in p_T is defined by the following bins (in GeV): $\{[15, 40]; [40, 60]; [60, 90]; [90, 150]; [150, 400]; [400, 600]; [600, +\infty]\}$. The binning in η is defined by the following set of bins: $\{[0, 1.2]; [1.2, 2.1]; [2.1, +\infty]\}$ and combined with every p_T bin, except for the last two high-p_T bins which use a binning in η defined by $\{[0, 1.2]; [1.2, +\infty]\}$.

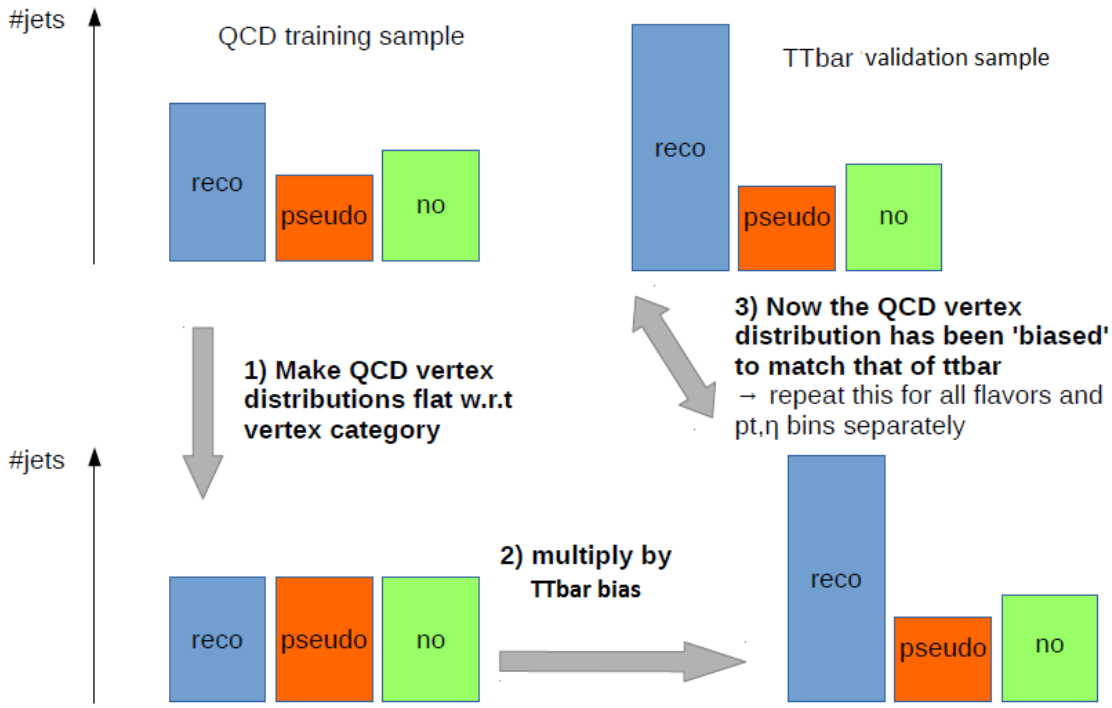


Figure 3.5: Schematic illustration of the application of the biases from the QCD vertex distributions (for a specific flavour f and a specific (p_T,η) bin) towards the $t\bar{t}$ vertex distributions. The displayed SV category distributions do not represent the real vertex distributions but merely serve as an illustration of the biasing procedure.

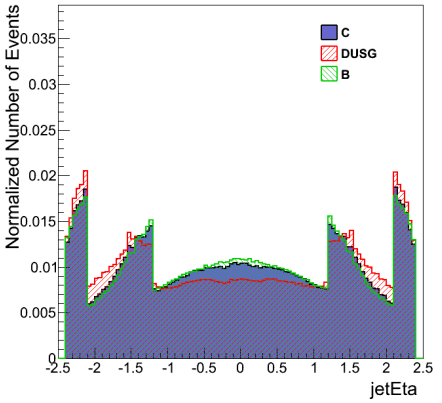
$$w_{QCD}^{p_T,\eta}(f) = 1/N_{p_T,\eta,f,incl}^{QCD} \tag{3.6}$$

It could be argued that in order to make the training independent of the jet p_T and jet η, one could simply not use these variables in the training and these weights become irrelevant. However other training variables can and will be correlated to p_T and η of the jet and for this reason the weights are necessary to eliminate this dependency as much as possible.

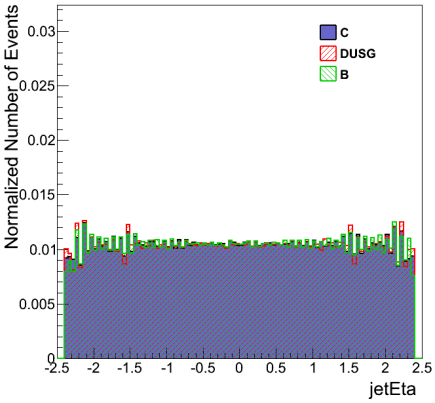
To summarise, the weighting procedure is performed in the following steps:

1. The QCD vertex distributions are normalised to one in bins of p_T and η for each flavour (w_{QCD}^{norm})
2. These normalised vertex distributions are biased towards those of the $t\bar{t}$ sample ($w_{t\bar{t}}^{bias}$)
3. The distributions of jet p_T and jet η in the QCD training sample are made flat by applying p_T-η weights ($w_{QCD}^{p_T,\eta}$)

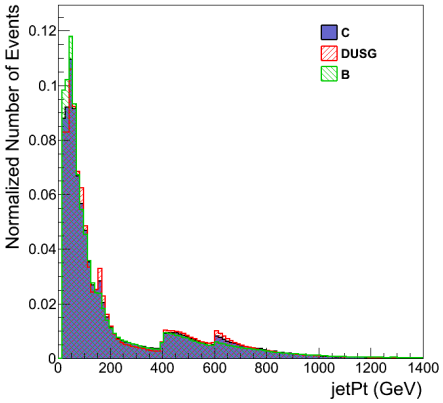
Finally all jets are weighed by the product of the three weights mentioned above and the weighed variable distributions are used in the c-tagger training.



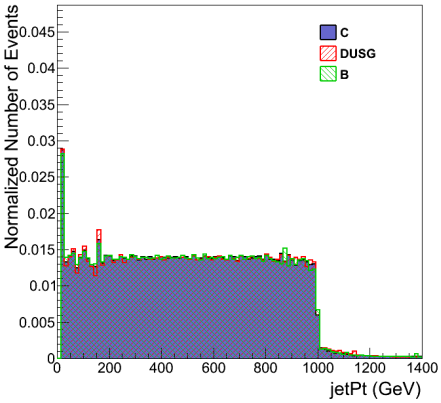
(a) jet η without weights



(b) jet η with weights



(c) jet p_T without weights



(d) jet p_T with weights

Figure 3.6: Distributions of jet η and p_T for b, c, and light-flavour jets before and after applying the weights to flatten the distributions of these kinematic variables. The sharp edges in the left plots (before applying weights) are due to the skimming of the QCD samples. The small spikes in the right plots (after weights are applied) are due to differences in the choice of binning between the $t\bar{t}$ biases and the p_T-η weights. It has been tested and shown that these spikes do not influence the performance of the charm tagger.

3.5 Performance of the C vs DUSG tagger

A charm tagger tries to distinguish c jets from either b or light-flavour jets. The fact that the variable distributions for c jets are typically shaped in between those of b and light-flavour jets makes it impractical to develop one c tagger to distinguish c jets from non-c jets (light-flavour or b jets) at the same time. This is in contrast to a b tagger, which can easily distinguish b jets from non-b jets with one and the same tagger. There is the possibility to use a neural network MVA with three output nodes (c, b and light-flavour jets), treating two backgrounds at the same time. The use of a BDT in this projects rises the necessity to look at C vs DUSG and C vs B taggers[12] separately. The discrimination between c jets and b jets

[12]C vs DUSG and C vs B taggers denote respectively a charm tagger that discriminates c jets from light-flavour jets and from b jets.

is somehow already possible with a b tagger (although within the c-tagger setup it will be slightly different). Therefore the choice has been made to first investigate the performance of a C vs DUSG tagger and only after that combine this with a C vs B tagger.

In the following sections, the performance of a C vs DUSG tagger will be discussed using the standalone TMVA setup explained in Section 3.4. Training is performed on skimmed and weighed QCD samples with a validation on $t\bar{t}$ samples. These samples are created with default IVF cuts for the secondary vertex reconstruction as shown in Table 3.1 and the default BDT settings as shown in Table 3.2 are used. The c tagger was built from a similar b-tagger setup and the same set of variables was used. These variables (with definition) are listed in Tables A.1 and A.2 in appendix A together with their weighed and normalised distributions for c, b and light-flavour jets. Only when the optimisation of the c tagger will be discussed, the variable set, BDT and IVF settings will be varied and the effect on the performance will be studied. For now however these settings are held constant for the discussion of the C vs DUSG and C vs B taggers.

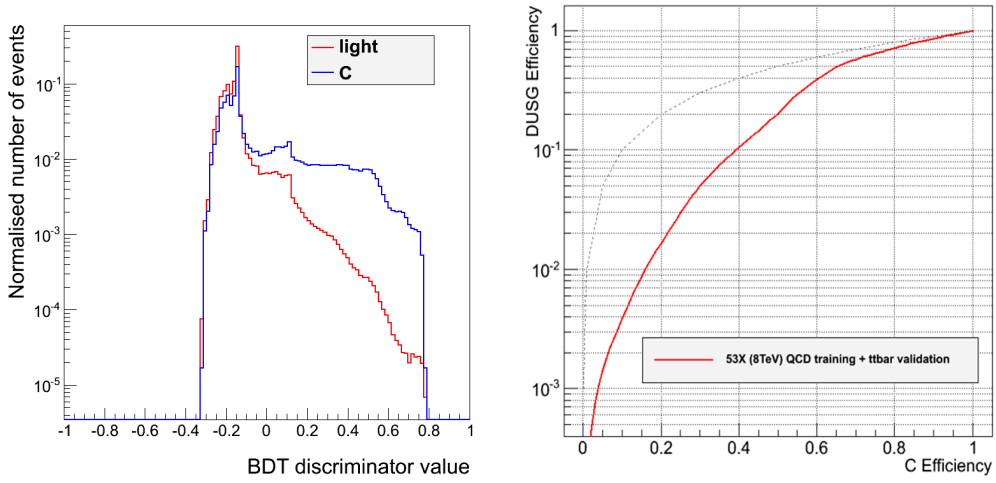
3.5.1 Performance at 8 TeV

During 2011 and 2012, the LHC has had a run at a centre of mass energy of 7 and 8 TeV respectively. Most of the analyses on the collected data at 8 TeV are finished as the LHC recently started towards running at energies up to 13 TeV. Nevertheless a c tagger was not yet present during the 8 TeV data taking or during Long Shutdown 1 (LS1). Therefore the performance of the charm tagger was first estimated on 8 TeV simulated samples. This provides a first idea of the c-tagger performance and allows for a first comparison to a similar b-tagger setup.
In all of the following ROC curves, a gray dashed line indicates the diagonal (in a log scale on the y-axis) which indicates the line on which there is no discrimination between c jets and light-flavour jets at all. The more the curve is located to the right bottom corner, the better the performance of the c tagger (larger c efficiency for a smaller light-flavour mistagging probability). If however the curve would be located above the diagonal (toward the left upper corner), this would indicate the cut on the BDT discriminator value is taken in the opposite direction and there would in fact be some discrimination present[13].
In practice a working point has to be defined which picks out one point on the ROC curve that corresponds to one cut on the BDT discriminator distributions. To choose a working point one has to figure out whether it is more important to have a high purity (large background reduction) or a high efficiency (with larger background contamination). When discussing the p_T dependence of the combined C vs DUSG and C vs B taggers later on, three such working points (Loose, Medium and Tight) will be introduced. Right now, for the sake of comparison, the light-flavour (DUSG) efficiency will be quoted for a c efficiency of 0.2 and 0.6. This way both the low and high charm-efficiency regions are covered.

The performance at 8 TeV is shown in the ROC curve in Figure 3.7(b), together with the

[13]The ROC curve at the other side of the diagonal is simply obtained by changing each axis to its complement to 1 (1-value on axis).

BDT discriminator distributions in Figure 3.7(a). For a charm efficiency of 20% or 60%, light-flavour efficiencies of around 2% and 40% are reached. Prior to any optimisations, this provides a first indication of the typical performance of a charm tagger. This is as expected worse than a similar b-tagger performance (B vs DUSG) that reaches 2% light-flavour efficiency at around 70% b efficiency [62]. With an eye on the new data that will be accumulated in 13 TeV collisions, it is interesting to also investigate the performance of the charm tagger on 13 TeV simulated samples, which is discussed in the following section.



(a) Discriminator distributions

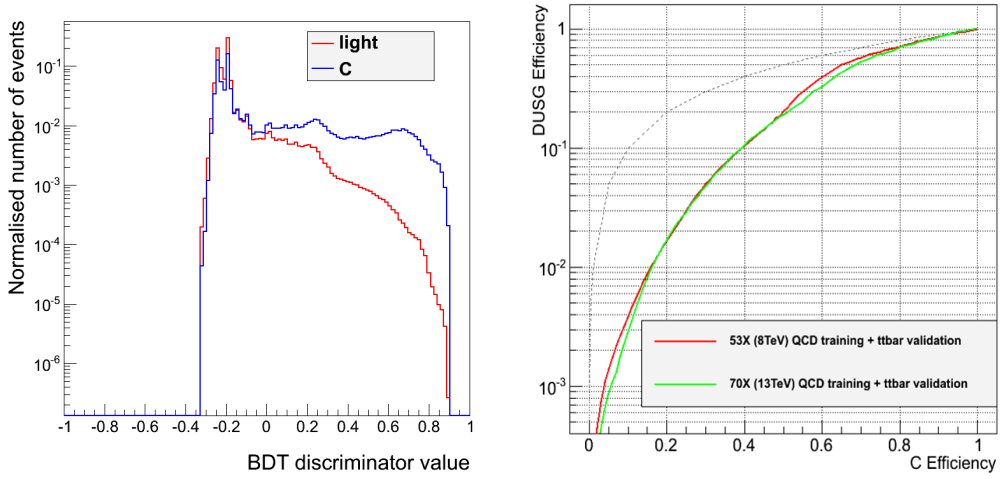
(b) ROC curve

Figure 3.7: (a) BDT discriminator distributions and (b) performance of the C vs DUSG charm tagger for training (QCD) and validation ($t\bar{t}$) on 8 TeV pp collision samples. The full variable set from Tables A.1 and A.2 (not including soft lepton variables) is used with default BDT and IVF settings.

3.5.2 Performance at 13 TeV

The high energy collisions at 13 TeV create a high pile-up environment for particle reconstruction, making it harder to deal with pile-up during the jet reconstruction. One way to be less sensitive to pile-up is to use a smaller distance parameter in the anti-k_T jet reconstruction algorithm ($R = 0.4$ instead of 0.5). This is justified by the fact that objects from these high-energy collisions will be boosted more towards the direction of the jet axis, narrowing the cone of the jet. Besides that also tracking and especially triggering require new advanced techniques. These arguments should convince anyone that the performance of the charm tagger needs to be investigated at 13 TeV as well to see if there are large differences with respect to 8 TeV.

The performance of the charm tagger at 13 TeV is shown in Figure 3.8(b) in green and compared to the performance at 8 TeV in red. The BDT discriminator distributions are shown in Figure 3.8(a). It can be seen that the overall performance is very similar. Again the DUSG mistag efficiency at 20% c effiency is around 2%. At 60 % c efficiency a minor improvement with respect to the 8 TeV case is seen, reaching a DUSG mistag efficiency of around 35% instead of 40% at 8 TeV.



(a) Discriminator distributions

(b) ROC curves

Figure 3.8: (a) BDT discriminator distributions and (b) performance of the C vs DUSG charm tagger for training (QCD) and validation ($t\bar{t}$) on 13 TeV pp collision samples, compared to the performance at 8 TeV. The full variable set from Tables A.1 and A.2 (not including soft lepton variables) is used with default BDT and IVF settings.

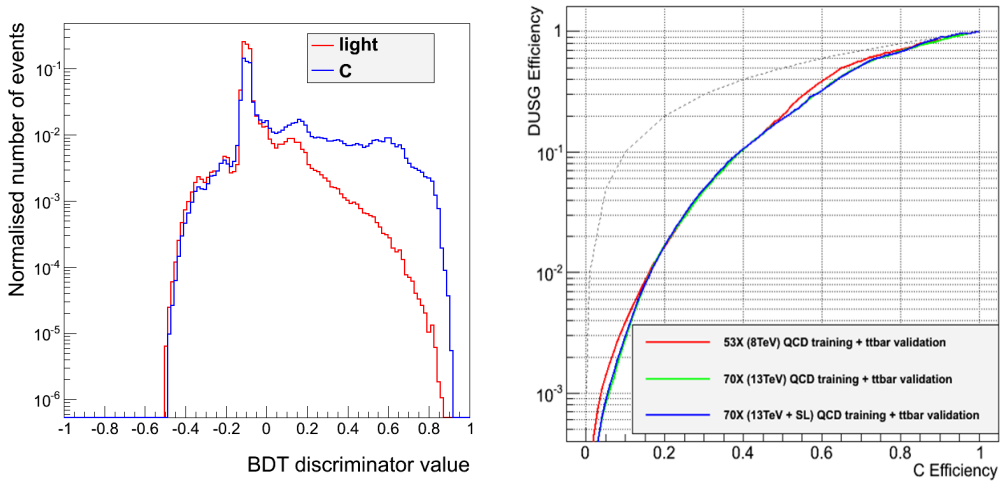
3.5.3 Using soft lepton (SL) information

Throughout the jet showering process hadrons will decay and radiate, producing a lot of new particles that can be detected by the CMS detector. Mesons can have hadronic decays toward other lighter mesons, but also leptonic decays will occur during the fragmentation process. These leptonic decays typically have rather small branching ratios, of the order of 10^{-2} to 10^{-4} (see for example [63] and [64]). This will result in soft charged leptons within jets that can add information to the jet properties and therefore to the discrimination between c and light-flavour jets. In theory these decays will result in electrons, muons and taus with associated neutrinos, but in practice only the electrons and muons are used in the c tagger because tau reconstruction is much harder due to its short lifetime and subsequent decays.

The use of soft electron and soft muon information in the charm tagger introduces three new soft-lepton (SL) categories on top of the three secondary vertex categories to end up with a combination of 9 vertex-SL categories for each jet flavour. These three new categories are called *NoSoftLepton*, *SoftElectron* and *SoftMuon* based on the soft lepton content of the jet. Just like for the vertex categories, the standalone TMVA c-tagger setup treats these SL categories inclusively in the training, giving default values to SL related variables in the NoSoftLepton categories. Due to a lack of statistics, the PseudoVertex-SoftElectron and PseudoVertex-SoftMuon categories are omitted and merged within the PseudoVertex-NoSoftLepton category in the samples that are used here. The SL related variables contain information on the (transverse) momentum, pseudorapidity, impact parameter and ΔR(lepton,jet) of the soft-lepton tracks inside the jet and can be found in Appendix A. These variables are added to the existing variable set and the performance is

again compared to previous cases.

The performance of the charm tagger at 13 TeV including SL information is shown in Figure 3.9(b) in blue and compared to the performance at 13 TeV and 8 TeV (without SL information) in green and red respectively. The BDT discriminator distributions are shown in Figure 3.9(a). No gain in performance is seen from the use of SL-related variables. When the c tagger does not gain from new variables this can have two possible explanations: either the new variables have a very poor discrimination between c and light-flavour jets or the new variables are highly correlated to already existing variables adding no new information to the MVA. To investigate the lack of improvement from SL information the performance of the c tagger was calculated using a subsample including only the SoftElectron and SoftMuon categories, making sure that all the jets contain SL information. For this subsample at 13 TeV the performance was compared with and without the use of SL variables. The result is shown in Figure 3.10. It can be seen that for jets with soft leptons, the performance of the c tagger increases at high c efficiencies when the SL variables are used in the training and validation. This motivates the conclusion that the lack of increase in the global performance is due to the fact that a large amount of the jets do not contain SL information, suppressing the potential gain in performance.



(a) Discriminator distributions

(b) ROC curves

Figure 3.9: (a) BDT discriminator distributions and (b) performance of the C vs DUSG charm tagger for training (QCD) and validation ($t\bar{t}$) on 13 TeV pp collision samples including soft lepton variables, compared to the performance at 8 and 13 TeV without soft lepton variables. The full variable set from Tables A.1 and A.2 is used with default BDT and IVF settings.

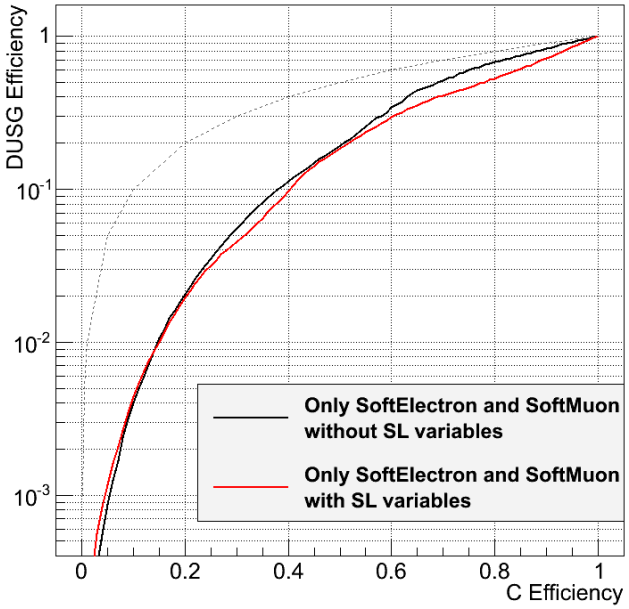
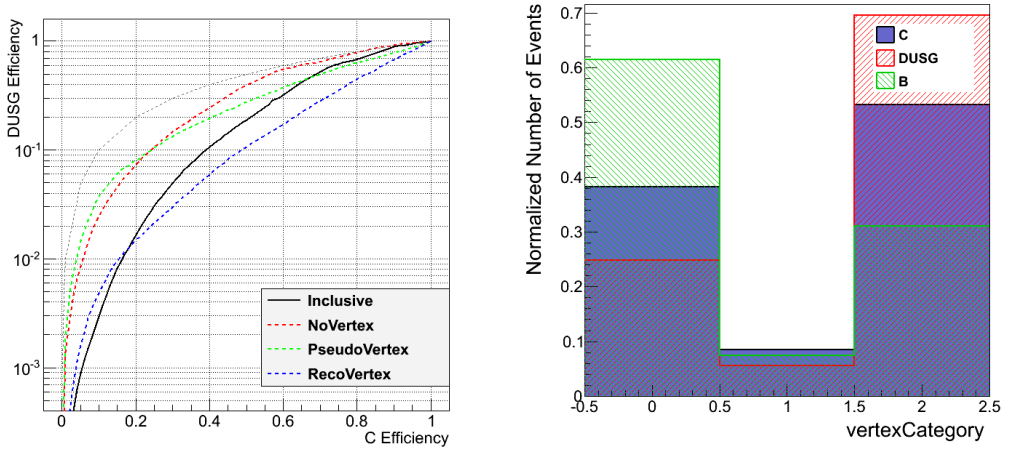


Figure 3.10: Performance of the C vs DUSG tagger (13 TeV with SL information) with a training and validation on only jets from the the SoftElectron and SoftMuon categories, making sure all jets contain soft-lepton information. The red curve shows the performance when the SL variables are included in the training, whereas the black curves shows the performance when SL information is left out. In the SoftElectron and SoftMuon categories, an improvement in the performance is seen when soft-lepton information is used, which was not the case when also the NoSoftLepton category was included (see Figure 3.9(b)).

3.5.4 Sensitivity of different vertex categories

The standalone TMVA c-tagger setup trains and validates inclusively for all secondary vertex categories. It is interesting to investigate which of these categories (NoVertex, PeudoVertex or RecoVertex) has the highest sensitivity to C vs DUSG discrimination and how each category affects the global (inclusive) performance. Figure 3.11(a) clearly shows that the RecoVertex category has by far the highest sensitivity to C vs DUSG discrimination, but the global performance is suppressed by the worse performance of the NoVertex and PseudoVertex categories. Whether or not the better performance of the RecoVertex Category can be beneficial in an analysis depends on the relative fraction of jets of a certain flavour in each vertex category. Figure 3.11(b) shows the normalised distributions of the secondary vertex categories for each flavour in the $t\bar{t}$ validation sample.



(a) ROC curves

(b) Secondary vertex category distributions

Figure 3.11: (a) Performance of the C vs DUSG tagger (13 TeV + SL) for the different secondary vertex categories separately (and the inclusive performance in black). The RecoVertex category is by far the most sensitive one, giving the best c-tagger performance. (b) Weighed and normalised distributions of the secondary vertex categories (corresponding to the distributions in the simulated $t\bar{t}$ validation samples) for different jet flavours (RecoVertex = 0, PseudoVertex = 1 and NoVertex = 2).

3.6 Combined C vs DUSG and C vs B tagger

The biggest novelty of a charm tagger is the possibility to discriminate c jets from light-flavour jets, whereas C vs B discrimination is somehow already possible with a b tagger. Nevertheless the charm tagger can only be used in analyses once its effect on b jets (b jet mistag probability) is known as well. Furthermore, the charm tagger is developed within a different setup and may on its own provide additional C vs B discrimination compared to the b tagger.

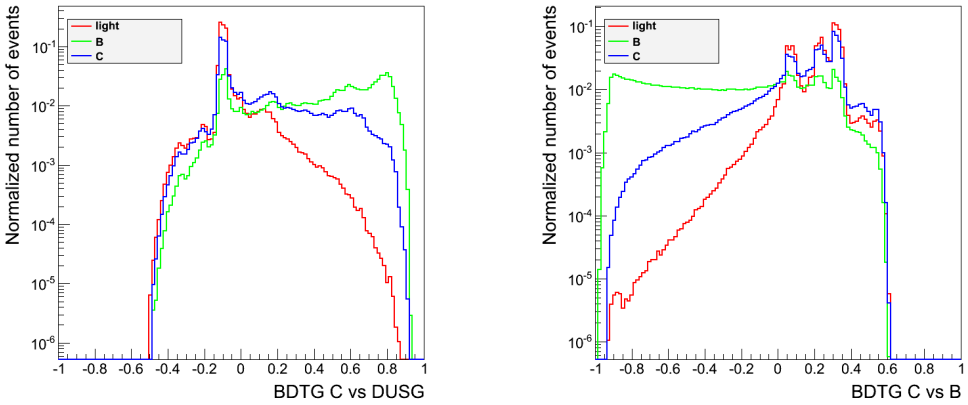
It is very inefficient to develop a single BDT-based c tagger that looks at both b and light-flavour jets as a background since the variable distributions for c jets are typically located in between those of b and light-flavour jets. When focussing on a light-flavour jet background, b jets will behave as extreme cases of c jets. Therefore two charm taggers need to be trained, one for C vs DUSG discrimination and one for C vs B discrimination. This will give rise to two dimensional discriminator distributions (one dimension for each tagger) which consequently requires a two dimensional cut in that phase space of discriminators in order to try to isolate c jets.

The performance of this two dimensional framework will be discussed in the following section. The C vs DUSG discrimination uses the results from the c tagger at 13 TeV including soft lepton information. The C vs B discrimination uses the exact same setup (same full variable set, BDT settings and IVF cuts) but focuses on a background of b jets rather than light-flavour jets. It should be realised that for the C vs B discrimination, other IVF settings (for

example the ones used for b tagging) should actually be used to obtain the best performance. This will however be disregarded for now since no optimisation for either one of the taggers has been applied yet. Optimisations are discussed later on for the C vs DUSG tagger and should be performed for the C vs B tagger in the future.

3.6.1 Performance

Before immediately discussing the performance of the combined charm tagger, it is advisable to first describe the process of calculating such performances in two dimensions. The development of two separate taggers results in two separate discriminator distributions. Figure 3.12 shows the BDT discriminator distributions, also called BDTG distributions, for the different jet flavours for the C vs DUSG tagger on the left and for the C vs B tagger on the right. In practice a two dimensional cut will have to be applied. Therefore these two distributions are combined in a two dimensional histogram with the C vs DUSG discriminator on the x-axis and the C vs B discriminator on the y-axis. This is shown in Figure 3.13 for the three jet flavours separately and in order to achieve an overlay of all the flavours also a scatter plot is included in the bottom right corner. This overlay serves as an illustration of where each of the jet flavours are located in this two dimensional phase space of discriminators. Due to the fact that the BDT output pushes the signal distribution towards 1 and the background towards -1, c jets will be located towards the upper right corner of this plot whereas b jets and light-flavour jets are located on average more towards the bottom right corner and the top left corner respectively. This separation is not very clear due to the rather poor discrimination of the charm tagger in general (see Figure 3.12), but it shows that in order to cut out c jets from the background one has to cut out a rectangular shape towards the upper right corner of this phase space.



(a) C vs DUSG (b) C vs B

Figure 3.12: BDT discriminator (BDTG) distributions for c, b and light-flavour jets for (a) the C vs DUSG tagger and (b) the C vs B tagger.

In order to produce performance plots for this two-dimensional setup (the equivalent of a ROC curve in one dimension) one has to scan over the BDTG values for both the C vs DUSG

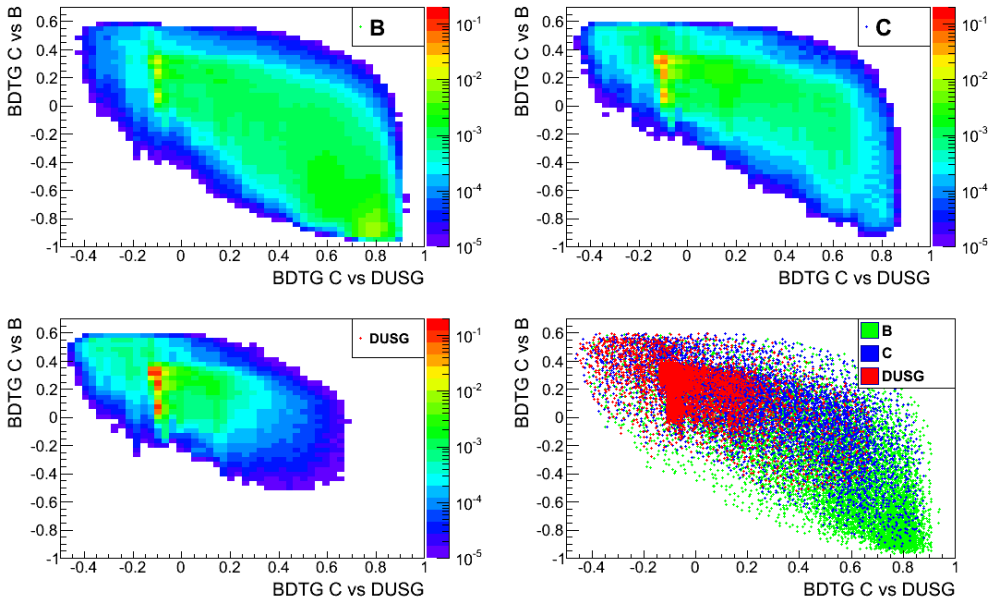
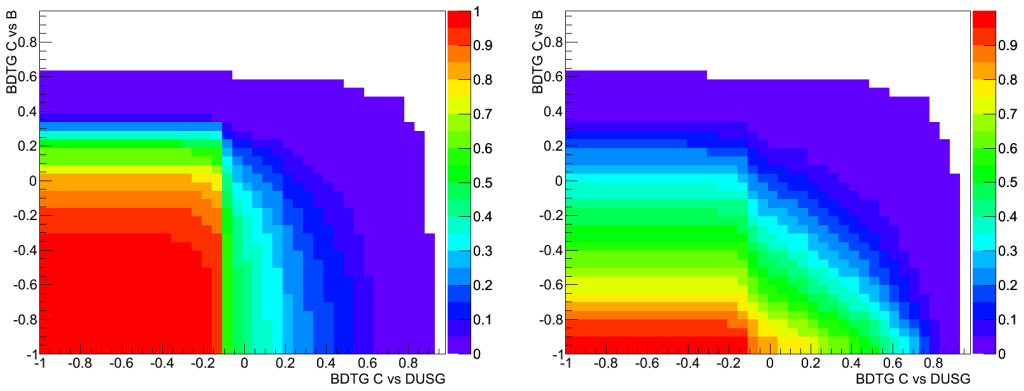


Figure 3.13: Two-dimensional normalised distributions of the BDTG values for the C vs DUSG tagger on the x-axis and the C vs B tagger on the y-axis. The bottom right plot shows an overlay of scatter plots of the other three plots, which illustrates the distribution of c, b and light-flavour jets along this 2D plane.

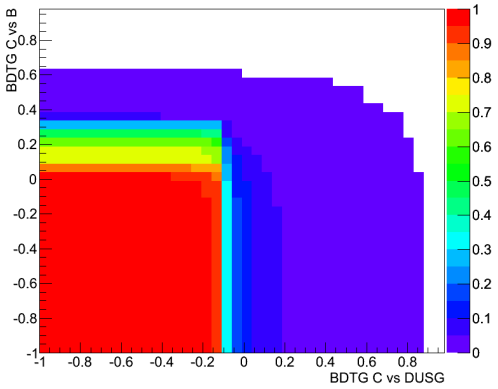
and C vs B taggers and each time cut out the upper right rectangle defined by these cuts. For each of these combined two cuts, one can count the number of jets of a certain flavour that pass that cut relative to the total number of jets of that flavour and obtain an efficiency of that cut. This is shown in Figure 3.14 for the three jet flavours, where the efficiency is shown as a function of the two-dimensional cut in the discriminator phase space.

The plots in Figure 3.14 contain all the necessary information to produce performance curves for the combined tagger. There are two degrees of freedom in constructing the performance: one either chooses a cut value for each of the two BDTG distributions which fixes the efficiency for each flavour, or one picks a predefined efficiency for two of the three jet flavours, which fixes the cuts on the discriminators and therefore the efficiency of the third flavour. The performance can not be expressed by a single ROC curve any more, but rather a set of ROC curves is needed. In Figure 3.15 the rejection (which is simply the inverse of the efficiency) for b jets is shown as a function of the rejection of light-flavour jets for different constant values of the charm efficiency. The higher the charm efficiency, the lower the overall rejection of b- and light-flavour jets. However for a constant predefined charm efficiency, the freedom exists to either choose a large light-flavour rejection with a poor b rejection or vice versa (or somewhere in between). This freedom is very convenient for applying the charm tagger on different types of analyses. Based on the backgroud composition of the analysis one can choose for a desired charm efficiency to either reject b jets or light-flavour jets.



(a) c jets

(b) b jets



(c) light-flavour jets

Figure 3.14: Tagging efficiency as a function of the BDTG values in two dimensions for (a) c jets, (b) b jets and (c) light-flavour jets.

3.6.2 p_T-Dependence of the tagging efficiency

In the previous discussions of the performance, all jets with a p_T larger than 30 GeV are used. It should however not come as a surprise that the performance itself is also dependent on the p_T of the jet. This dependence will now be discussed for the samples at 13 TeV with SL information. One could in principle produce ROC curves for different ranges of jet p_T to see which ranges are most sensitive for charm tagging. In order to achieve a functional relation between the tagging efficiency and the jet p_T three working points (WP) will be defined and for each of these working points that functional relation will be calculated. These three working points are defined by three sets of cuts on the discriminator distributions. They are called the Loose WP (keeping a lot of c jets but also including a lot of b- and light-flavour jets), the Medium WP (reducing the bottom and light-flavour efficiency at the cost of a lower charm efficiency) and the Tight WP (keeping only a very pure signal selection at the cost of a low charm efficiency). The working points are defined in Table 3.3.

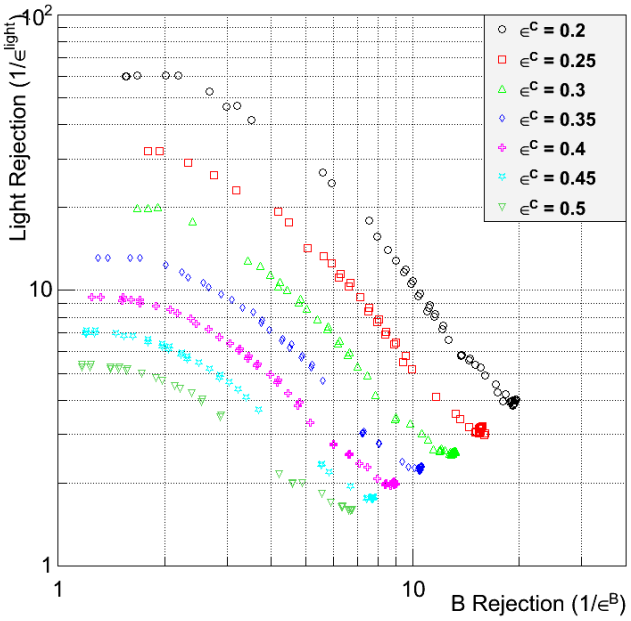


Figure 3.15: Relation between b rejection ($1/\epsilon^B$) and light-flavour rejection ($1/\epsilon^{light}$) for different values of a constant c efficiency (ϵ^C).

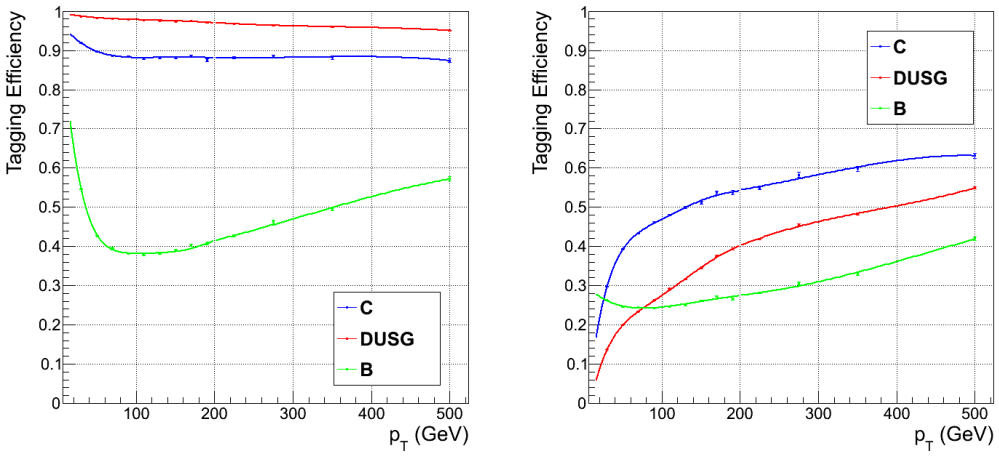
WP	ϵ^C	ϵ^B	ϵ^{light}	**BDTG cut C vs DUSG**	**BDTG cut C vs B**
Loose	0.9	0.4	0.98	-0.337	-0.356
Medium	0.4	0.25	0.2	-0.073	-0.302
Tight	0.2	0.37	0.02	0.294	-0.682

Table 3.3: Definitions of the three working points (WP) with the corresponding cuts on the BDTG values and the global efficiencies for each flavour.

For each of these WP the p_T dependence of the c, b and light-flavour efficiency is shown in Figure 3.16. The fits through the obtained points are polynomials of different orders in order to match the points as close as possible. In order to achieve this, the fits are performed separately in the region where p_T < 200 GeV and p_T > 200 GeV. These fitted functions will be used later on when investigating the potential of charm tagging in the search for flavour changing top-quark dark matter interactions. Due to the freedom of the two dimensional setup explained above, it can be seen that the Loose WP focuses mostly on C vs B discrimination with a very poor light-flavour rejection, whereas the Tight working point focuses more on C vs DUSG discrimination with a poor b rejection. The Medium WP is chosen somewhere in between, with both a moderate b and light-flavour rejection.

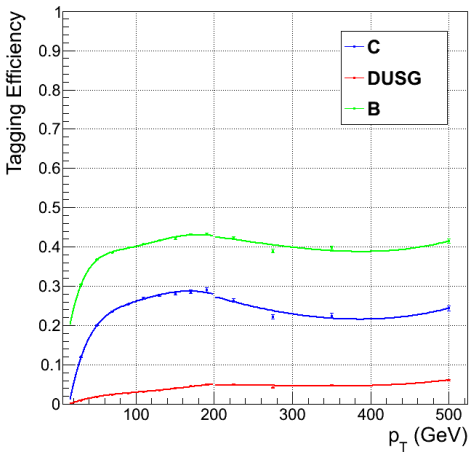
3.6.3 Comparison to the ATLAS charm-tagging algorithm

Within the ATLAS experiment a charm-tagging algorithm (*JetFitterCharm*) [65] has already been developed. It is advisable to compare the obtained performance of the TMVA c tagger for CMS to the performance obtained by the JetFitterCharm algorithm for ATLAS. The



(a) Loose WP

(b) Medium WP



(c) Tight WP

Figure 3.16: Tagging efficiency as a function of p_T for (a) the Loose WP, (b) the Medium WP and (c) the Tight WP.

performance of a combined C vs DUSG and C vs B c tagger at a center of mass energy of 13 TeV for CMS was shown in Figure 3.15 and an equivalent plot, although at a center of mass energy of 8 TeV, for ATLAS can be found in Figure 3.17. In general the ATLAS c tagger shows a better performance, reaching a light-flavour and b rejection of up to 200 and 23 receptively for a c efficiency of 20%. For CMS these numbers are respectively 60 and 11. However, one should be careful in comparing these numbers as numerous differences are present between the CMS and the ATLAS c tagger setups:

1. ATLAS determined the performance at a center of mass energy of 8 TeV, whereas the performance presented in this thesis for CMS is calculated on 13 TeV samples. Figure 3.8(b) shows that only small differences in performance are seen between 8 TeV and 13 TeV simulated samples, showing that this is probably not the largest contribution to the differences between CMS and ATLAS.
2. ATLAS reconstructs jets using the anti-k_T reconstruction algorithm explained in Sec-

tion 3.1.2 with a radius parameter of $R = 0.4$, whereas CMS uses the same algorithm but with a radius parameter of $R = 0.5$ at 8 TeV and $R = 0.4$ at 13 TeV. Moreover, ATLAS uses jets with a $p_T > 20$ GeV in the determination of the c-tagging performance, whereas CMS uses jets with $p_T > 30$ GeV.

3. ATLAS uses a Neural Network with three output nodes (one for each jet flavour) to determine the combined performance. For the TMVA c tagger of CMS a BDT was used and a combination of two trained BDTs is needed to obtain a combined C vs B and C vs DUSG performance. Moreover, only a simple combination of these two separate c taggers has been deployed so far.
4. The performance obtained by ATLAS is calculated by training and validating both on simulated $t\bar{t}$ samples. In the TMVA setup the training is performed on simulated QCD samples and the validation is performed on simulated $t\bar{t}$ samples. It can be expected that using different samples for training and validation might lower the obtained performance because of the differences in variable distributions. However, the discrepancy between the QCD and $t\bar{t}$ samples is partially solved by applying the weights discussed in Chapter 3.4.4.
5. The jet properties used in the MVA are similar between CMS and ATLAS, but often they are used in a different way. For example, ATLAS combines all the track-IP-related variables in a likelihood ratio and uses that LR as an input to the MVA, whereas CMS simply uses each IP-related variable separately as an input variable. Moreover, ATLAS chooses to use the minimum, maximum and mean rapidity of tracks along the jet axis, whereas CMS uses the (pseudo)rapidity (with respect to the jet axis) of the three highest-p_T tracks.
6. The obtained performance for CMS is not yet optimised. A first set of optimisations will be discussed in Chapter 3.7 for the C vs DUSG discrimination.

A combination of all these differences can explain the differences in performance between ATLAS and CMS. This discussion is ended by noting that the performance on simulated data is not necessarily the same as the performance on real collision data. Calibrations of the obtained performance between simulation and data are necessary for each jet flavour. ATLAS already presented some of these calibrations in the form of scale factors [65] which shows that they slightly overestimate the c efficiency and underestimate the light-flavour efficiency. This could possibly contribute to the better performance in simulations for ATLAS, but a fair comparison requires also a calibration to collision data of the efficiencies of the TMVA c tagger for CMS.

3.7 Optimising the performance

The performance of the charm tagger has been discussed using the full available variable set and with default BDT and IVF settings. Using the complete set of over 70 variables seems unnecessary and default IVF and BDT settings are not necessarily the best choices. This section explores possible improvements to the charm tagger to obtain the most optimal

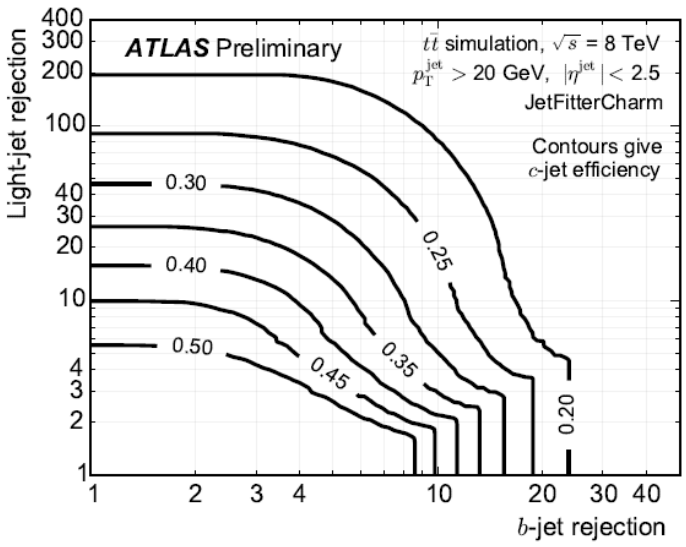


Figure 3.17: Two-Dimensional performance of the ATLAS JetFitterCharm algorithm on 8 TeV simulated $t\bar{t}$ samples [65]. The curves show the relation between b rejection and light-flavour rejection for different values of a constant c efficiency and can be compared with Figure 3.15 of the current performance obtained for CMS at 13 TeV.

performance. This includes a detailed investigation of the following three aspects of charm tagging:

1. Looking for a smaller set of sensitive variables that achieve a similar or perhaps even better performance than the complete set of available variables.
2. Optimizing the BDT settings from Table 3.2 to increase the discriminating power of the BDT.
3. Looking for the most optimal secondary vertex reconstruction by varying the parameters from the IVF reconstruction algorithm (see Table 3.1).

The following sections explore each of these aspects to optimise the performance of the C vs DUSG tagger. The optimisation of the C vs B tagger is not considered yet, but is planned to be performed in the future.

3.7.1 Sensitive variables

At first there seems to be no harm in using all of the available information from all variables in the MVA. However such a large variable set with all possible correlations could be hard to handle for the MVA and it drastically increases the computation time and memory usage of the BDT training. Therefore a smaller set of variables will be selected based on two important criteria:

1. The variable distributions should show a good discrimination between c and light-flavour jets.

2. The selected variables should have low correlations[14] amongst each other to avoid using the same information multiple times.

TMVA provides for each MVA training two lists, ranking the variables that are used by their importance. One such a list is produced before the training and ranks the variables according to their discrimination independent of any MVA method (method unspecific ranking table). Another table is produced after the BDT training and ranks the variables according to their importance during the training (method specific ranking table). For a BDT this calculation is based on the number of times a variable is used in the splitting of a node. One could take the top variables from the method specific list and trust the calculation of TMVA to provide a set of the most sensitive variables. However after optimizing the BDT settings this method specific ranking table could be different again. In order to have more control over the variable selection and be independent on the specific calculation of the method specific ranking table, the choice has been made to select variables based on the two criteria mentioned above.

First the method unspecific ranking table is used to select the top 30 variables based on large discriminating power. The performance slightly increases for high charm efficiencies when using this reduces variable set, as can be seen from Figure 3.19 in red (top 30 variables) with respect to the blue curve (all variables). TMVA also produces (linear) correlation matrices for all the variables used in the training. Figure 3.18 shows this matrix for the charm jet correlations for the selected set of 30 variables. From this matrix, the most correlated variables are identified and for each of such two highly correlated variables, the one that is ranked lowest in the method unspecific ranking table is eliminated until a top 20 of variables remains. Again the performance on this reduced variable set is almost the same as before as shown in green in Figure 3.19. The final chosen set contains the following variables: trackSip2dSig_0(_1), trackSip3dSig_0(_1), trackEtaRel_1, vertexMass_0, vertexEnergyRatio_0, trackSip2d(3d)Sig(Val)AboveCharm_0, flightDistance2dSig(Val)_0, vertexBoostOverSqrtJetPt_0, leptonPtRel_0, leptonSip3d_0, leptonEtaRel_0, leptonRatio_0, vertexNTracks_0 and jetNSecondaryVertices. The method unspecific and method specific ranking tables (for these 20 variables) produced by TMVA can be found respectively in Tables B.1 and B.2 in Appendix B.

[14] It should be kept in mind that if correlations between the variables exist, but are different for signal and background this could still add to the performance of the charm tagger. This effect is not yet included in the determination of the most sensitive variables, but should be investigated in the future.

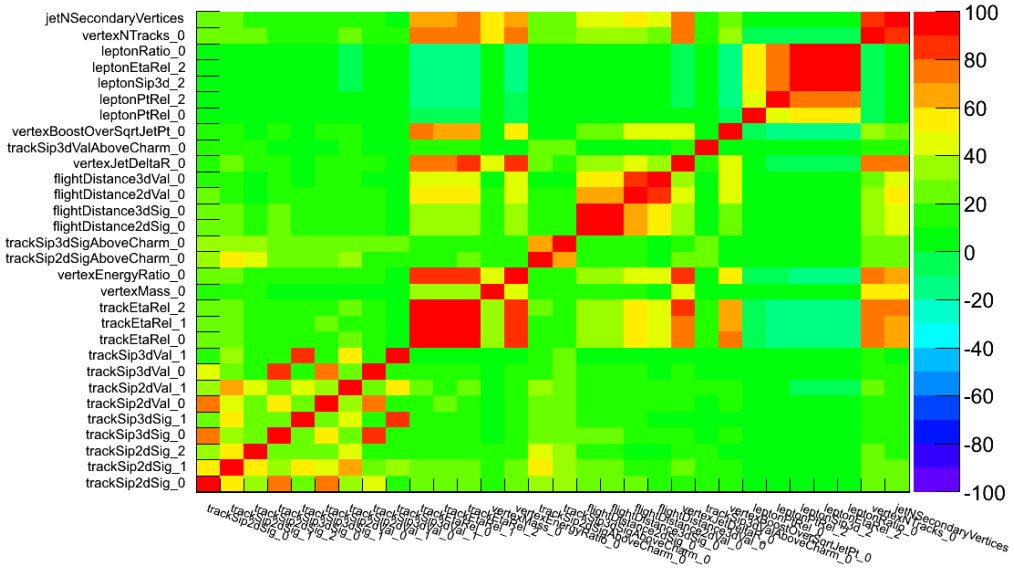


Figure 3.18: Correlation matrix of the charm jet distributions for the BDT training (CMSSW 70X, 13 TeV samples with SL information) on the reduced set of 30 variables with the highest discriminating power.

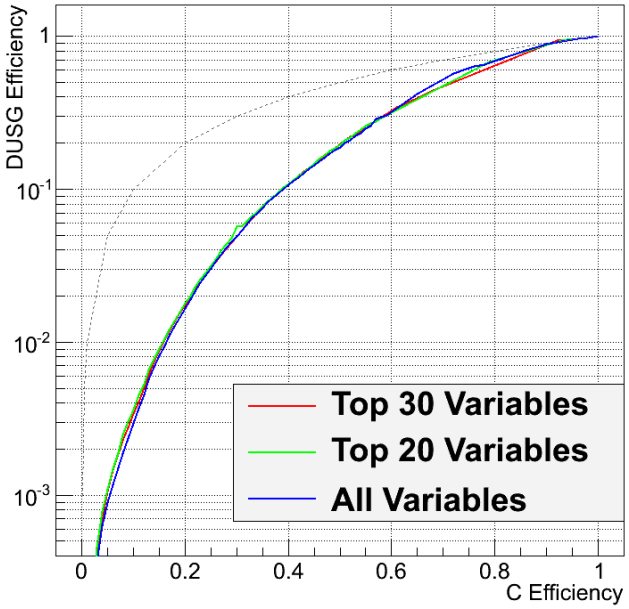


Figure 3.19: Performance of the charm tagger on the full variable set (CMSSW 70X, 13 TeV with SL information) compared to the performance for the reduces variable sets containing the top-30 and top-20 variables.

3.7.2 BDT settings

The default BDT settings as shown in Table 3.2 were adopted from a similar b-tagger setup. These settings are not necessarily the most optimal ones for the c tagger. All of the parameters from Table 3.2 (except BoostType and UseBaggedGrad) were varied over a

range of values, ranging from settings that are probably too minimal towards values that are probably exaggerated. The goal is to find a set of values for the BDT settings that optimise the charm tagging performance, but without increasing the computation time of the BDT training too much. Figure 3.20 shows the individual effects of each parameter, while keeping other parameters at their default value and using the full variable set.

- The effect of varying the GradBaggingFraction is almost non-existent but reducing this value also reduces the computation time (a smaller subsample of jets is used for each tree in the forest). For this reason the value was changed from 0.5 to 0.3.
- The MaxDepth seems to be the most sensitive variable and the default value of 2 seems too minimal. Its value was changed to 8, showing a visible improvement in the performance.
- The MinNodeSize seems again insensitive to changes, except if it is chosen too large. The larger the minimum size of a node has to be, the sooner the tree will be broken down and the maximum depth will not be reached any more. The value has been raised from 1.5% to 5% since this reduces the computation time without losing any performance.
- The number of cuts (nCuts) seems to be once again rather insensitive unless it is chosen too small. Using a small number of cuts will not effectively scan the entire variable range but rather take large jumps, resulting in a loss of performance. Once again the value was lowered from 80 to 50 to reduce computation time without losing performance.
- The number of trees (NTrees) used in the forest seems to perform optimally for any value above around 500. Nevertheless the default value of 1000 was raised to 2000 since small improvements keep showing up every time the value is increased.
- Finally the Shrinkage shows little to no change in performance when being varied and was changed from 0.1 to 0.5 to reduce computation time[15].

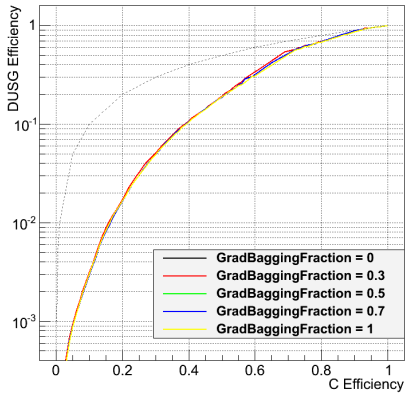
An overview of the optimised values is given in Table 3.4.

BDT Option	**Default Value**	**Optimal Value**
GradBaggingFraction	0.5	0.3
MaxDepth	2	8
MinNodeSize	1.5%	5%
nCuts	80	50
NTrees	1000	2000
Shrinkage	0.1	0.5

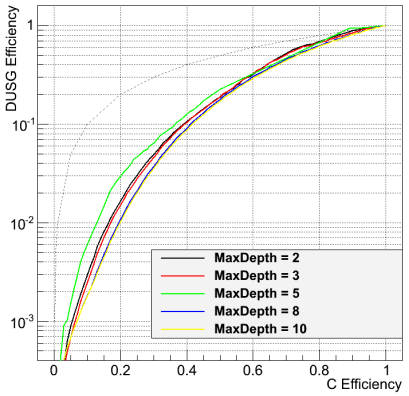
Table 3.4: Overview of the optimised values of the BDT settings compared to the default values.

The overall observation is that the individual variations yield only very small changes in performance. Only the MaxDepth shows a visible improvement. Figure 3.21 compares the performance with all optimised BDT settings together in green to the default settings in

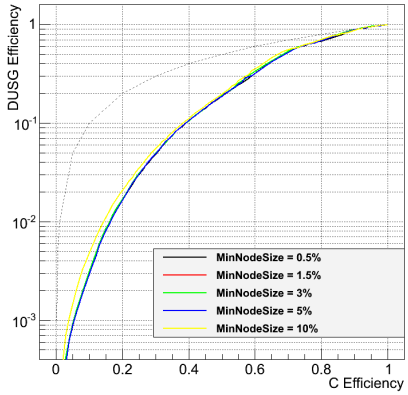
[15] A smaller shrinkage means a slower learning rate for the Gradient Boosting algorithm and therefore requires more iterations and thus a longer computation time.



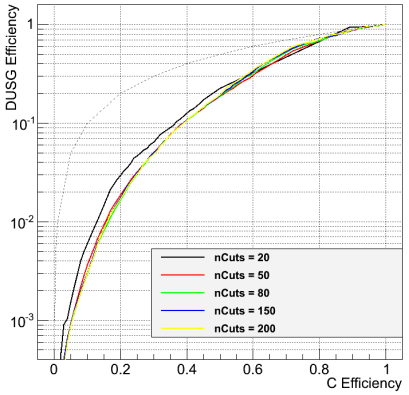
(a) GradBaggingFraction



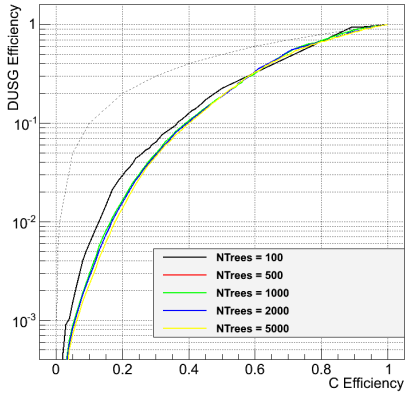
(b) MaxDepth



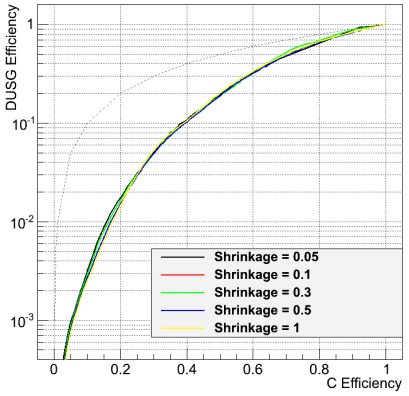
(c) MinNodeSize



(d) Ncuts



(e) NTrees



(f) Shrinkage

Figure 3.20: Variations of the different BDT parameters from Table 3.2 and their effect on the performance of the charm tagger (13 TeV with SL information) using the full variable set.

red. The combined effect of optimizing the BDT parameters shows a clear improvement in performance with respect to the default BDT settings over the entire range of charm efficiencies. Increasing the NTrees and MaxDepth variable values causes a large increase in

computation time for training and validation. This effect is highly reduced by the chosen values of the other parameters as explained above. The combined optimised settings increase the total CPU time for training and validation from around three hours to around six hours.

The question arises whether this improvement is still visible when looking only at the reduced variable set of 20 variables constructed in the previous section. When the optimal settings are applied to the reduced variable set it appears that the improvement in performance vanishes again, as shown in Figure 3.21 in blue. Some efforts have been made to increase the performance again using the reduced variable set, but without success so far. This could be explained by the fact that the optimised BDT settings caused even the less sensitive variables to contribute slightly to the performance, whereas this small amount of information for each non-sensitive variable gets lost when using only the 20 most sensitive variables. This leaves us to conclude that in order to benefit from the most optimal BDT settings, it could be necessary to use a larger variable set.

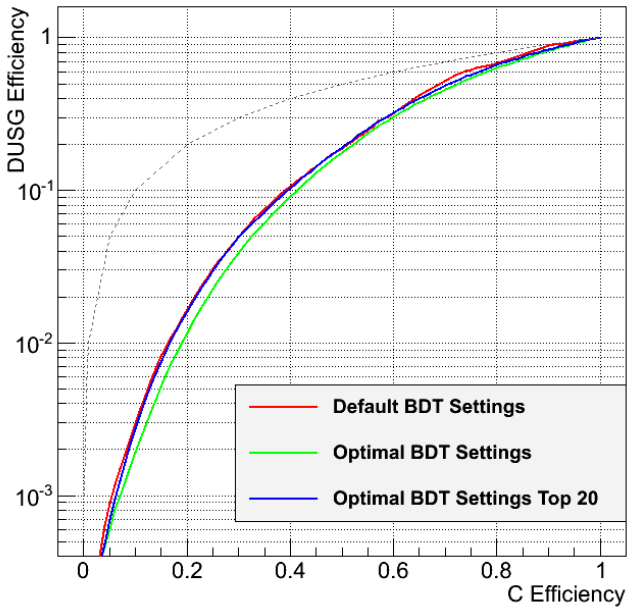


Figure 3.21: Performance of the charm tagger (CMSSW 70X, 13 TeV with SL information) compared for the default BDT options, optimised BDT options and optimised BDT options with the reduced variable set (top 20 variables).

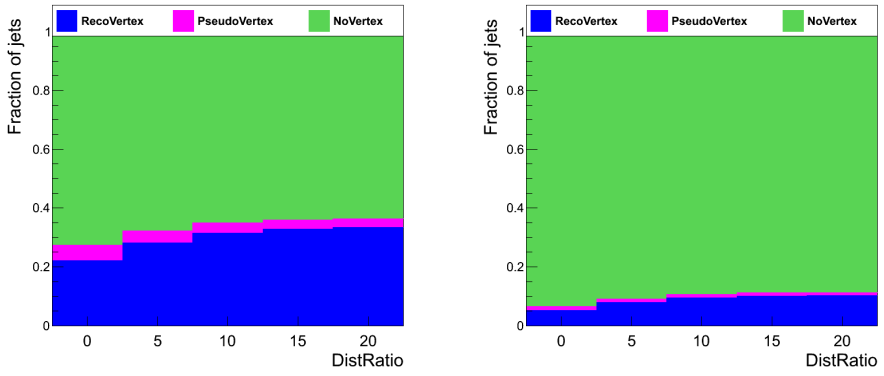
3.7.3 IVF optimisation

The inclusive vertex finder (IVF) algorithm is complex and contains many different parameters that can be tuned during the secondary vertex reconstruction as outlined in Section 3.2. The values of these parameters will affect the performance of SV reconstruction and may therefore affect the distribution of secondary vertex categories (Reco-, Pseudo- and NoVertex). As these categories are very important to heavy flavour taggers, changing the IVF parameters could have an effect on the tagging performance. On one hand the SV re-

construction has to be efficient, since the RecoVertex category shows the best performance. On the other hand the heavy-flavour tagging algorithms are based on the presence of a SV in b jets and c jets and the lack of any SV for light-flavour jets. This means one has to find a balance between having a large population of jets in the RecoVertex category, but also maintaining a clear distinction between the vertex category distributions for light-flavour jets and c or b jets. This section investigates the effect on the vertex category distributions and the charm-tagger performance from varying some of the IVF parameters.

Previous studies on b tagging [56] have shown that the only parameters (see Table 3.1 for the default values) that visibly affect the reconstruction of secondary vertices are the distanceRatio, distSig2dMin, vertexMinDLen(2D)Sig and seedMin3DValue(Significance). This served as a motivation to only look at these parameters in a first investigation, however it is not yet certain if for c tagging other parameters could become more sensitive as well. It was also seen that for c tagging the seed-track-related parameters did not affect the vertex category compositions and they will also not be discussed any further. In the following paragraphs, the effect of varying the distanceRatio, distSig2dMin and vertexMinDLen(2D)Sig on the vertex category distributions and the c-tagger performance will be discussed. The effect on the c-tagger performance is however only checked at validation level. This means the IVF settings were only varied in the $t\bar{t}$ validation samples, but the training is always performed on a QCD sample with the default IVF settings.

distanceRatio The distanceRatio (also noted DistRatio) was varied between 0 and 20 in steps of 5, and the effect on the distribution of the vertex categories is shown in Figure 3.22 for c jets and light-flavour jets. The variation of this parameter shows clearly that a value of 20 has the largest population of jets in the RecoVertex category. The effect on the performance at validation level shown in Figure 3.23 seems however negligible, but a small gain in performance can be seen at higher charm efficiencies for a value of 20 for the distanceRatio. This shows that a value of 20, which was also the default value, is already the most optimal choice.



(a) c jets (b) light-flavour jets

Figure 3.22: Change in vertex category composition from varying the value of the DistRatio parameter in the IVF algorithm for (a) c jets and (b) light-flavour jets. The displayed distributions are from the CMSSW 73X, 13 TeV with SL samples used for the IVF variations.

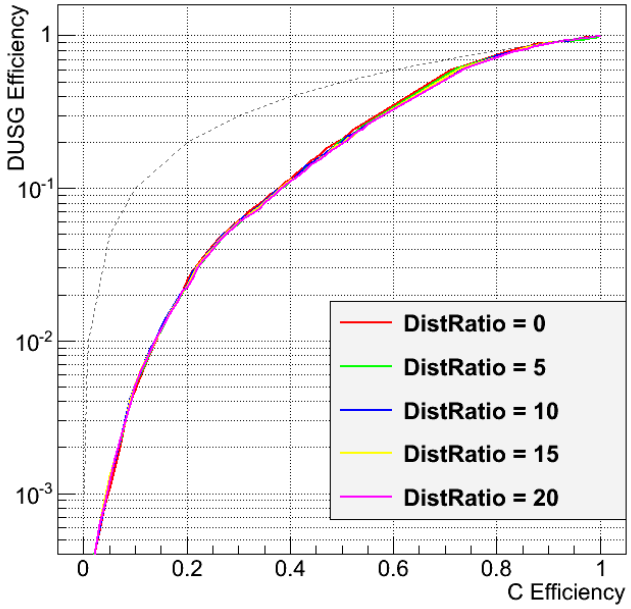
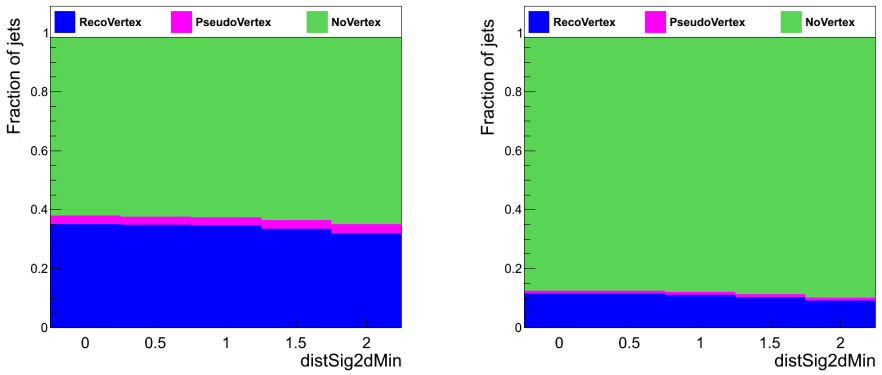


Figure 3.23: Performance of the charm tagger (CMSSW 73X, 13 TeV with SL information) compared for the different values of the DistRatio parameter in the IVF algorithm.

distSig2dMin The distSig2dMin parameter represents a cut on the flight distance of the final set of SV after the full IVF algorithm has been completed. Figures 3.24 and 3.25 show that there is only a small effect on the vertex category distributions and no effect on the c-tagger performance when varying this parameter between 0 and 2 in steps of 0.5. For this reason the choice has been made to put this parameter to 0 (meaning there is no final cut on the flight distance of the SV). Above that it might also be correlated to the variation of the vertexMinDLen(2D)Sig parameters, which also represent a cut on the flight distance

but in the first step of IVF and which will be discussed below. Putting the distSig2dMin to 0 is further motivated by the fact that during the MVA training, the flight distance is itself a parameter used in the MVA and in principle no prior cuts are therefore needed.



(a) c jets (b) light-flavour jets

Figure 3.24: Change in vertex category composition from varying the value of the distSig2dMin parameter in the IVF algorithm for (a) c jets and (b) light-flavour jets. The displayed distributions are from the CMSSW 73X, 13 TeV with SL samples used for the IVF variations.

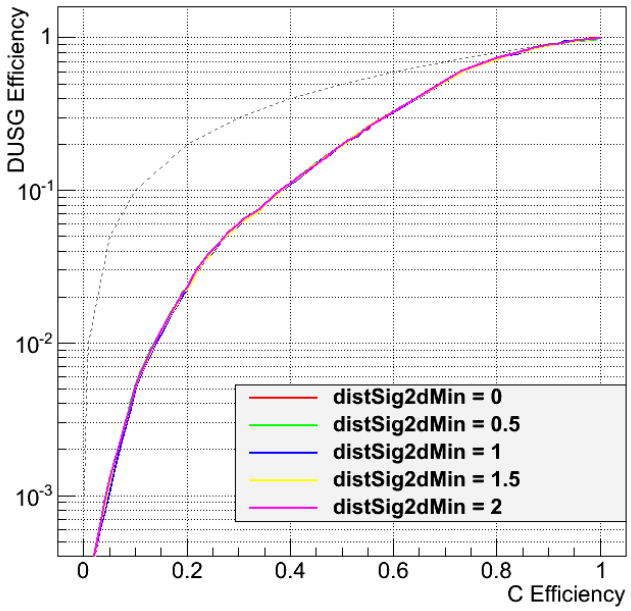
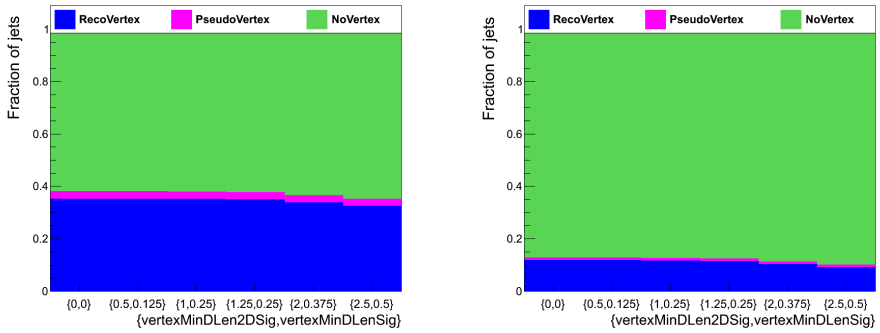


Figure 3.25: Performance of the charm tagger (CMSSW 73X, 13 TeV with SL information) compared for the different values of the distSig2dMin parameter in the IVF algorithm.

vertexMinDLen(2D)Sig Finally the effect of varying the vertexMinDLen2DSig and vertexMinDLenSig simultaneously between 0 and 2.5 and 0 and 0.5 respectively has been studied. For this variation the distSig2dMin parameter has already been put to 0 to avoid any

effect of the correlation amongst these parameters. Once again the effect on the vertex category distributions seems small as shown in Figure 3.26 and almost no effect is seen on the c-tagger performance in Figure 3.27 on validation level.



(a) c jets (b) light-flavour jets

Figure 3.26: Change in vertex category composition from varying the value of the vertexMinDLen(2D)Sig parameters in the IVF algorithm for (a) c jets and (b) light-flavour jets. The displayed distributions are from the CMSSW 73X, 13 TeV with SL samples used for the IVF variations.

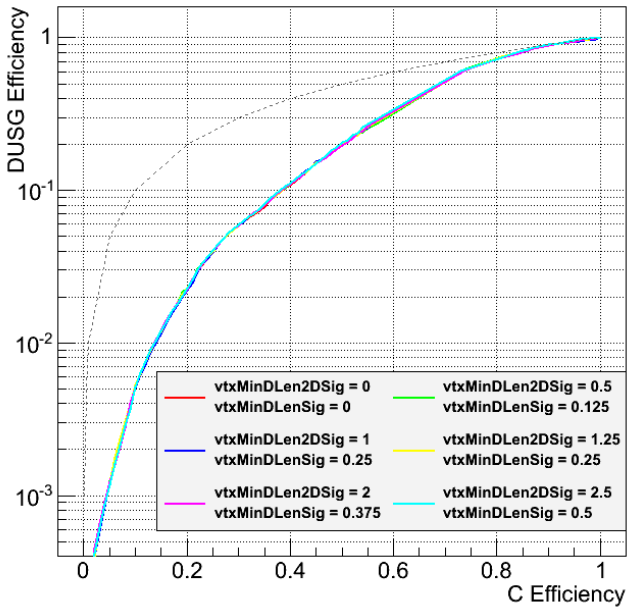


Figure 3.27: Performance of the charm tagger (CMSSW 73X, 13 TeV with SL information) compared for the different values of the vertexMinDLen(2D)Sig parameters in the IVF algorithm.

The overall observation is that varying the IVF parameters in accessible ranges can visibly affect the distribution of the vertex categories, but the effect on the c-tagger performance at validation level is negligible. The next step would be to check whether larger effects are seen for the performance if also the training is performed on QCD samples with varied IVF

settings (or perhaps to recalculate the $t\bar{t}$ biases for each validation sample without changing the training sample) and to possibly check the effect of other parameters in the IVF algorithm on c tagging that showed only small sensitivity for b tagging.

3.8 Comparison between TMVA and CMSSW charm-tagging setups

To end the discussion of the development of the TMVA standalone charm tagger, the obtained performance for C vs DUSG discrimination is compared to the CMSSW integrated charm-tagging setup mentioned earlier on. Recall the fact that the CMSSW c tagger uses an ANN instead of a BDT. The comparison is made on the 13 TeV samples (CMSSW 70X) with soft lepton information. The default as well as the optimised[16] TMVA charm tagger are compared to the CMSSW charm tagger in Figure 3.28. After optimisation, the TMVA c tagger performs very similar to the CMSSW charm tagger. It even performs slightly better at low charm efficiencies, although slightly worse at high charm efficiencies. The large differences in these two setups make this comparison a very convincing sanity check for the performance of charm tagging.

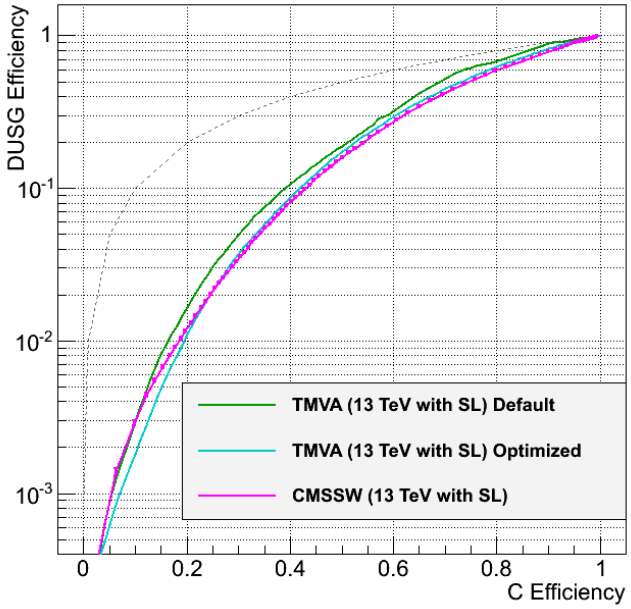


Figure 3.28: Comparison between the performance of the TMVA c tagger (default and optimised) and the CMSSW c tagger.

[16]This optimisation includes only the optimised BDT settings on the full variable set and with default IVF settings.

Chapter 4

Effect of charm tagging on the search for flavour changing top-quark dark matter interactions

The phenomenological model for FC top-quark dark matter interactions explained in Chapter 2 resulted in accessible cross sections to be probed in the second run of the LHC at 13 TeV. However, such a signal can only be investigated if it can be significantly distinguished from SM background processes. Therefore it is important to investigate the significance of this signal in 100 fb^{-1} of data expected from the 13 TeV pp collisions at the LHC in the next few years. This can be achieved by investigating discriminating variables in simulated data and by applying cuts to select the signal or by performing template fits to their distributions. The analysis presented in this thesis is a first approximative study from which the obtained insights can serve as a basis for a full dedicated study.

In the remaining sections of this chapter the analysis setup will be explained, the background processes are discussed and the selection criteria on a set of discriminating variables are presented. Finally the signal significance is calculated either by a simple cut and count experiment or by template fitting. The effect of applying the charm tagger developed in Chapter 3 is also quoted to investigate the potential gain in sensitivity from such a charm-tagging algorithm.

4.1 Analysis setup

The phenomenological model from Chapter 2, which is defined by the coupling strength c^{23} and the dark matter mass m_χ, is implemented in FeynRules [37] and simulated with MadGraph [38]. The signal therefore consists of a combination of the three processes discussed in Section 2.4 and illustrated in Figure 2.6. Hadronisation of the partons is simulated with Pythia [66] and the detector response has been simulated with a fast simulation of the CMS detector using Delphes [55]. The object reconstruction in Delphes has been discussed in Section 3.1.3.
The event samples processed by Delphes are used in MadAnalysis [32] software tools where

cuts are applied to the selected variables. These cuts are always chosen such that they optimise the signal significance. The statistical signal significance (Sign) is defined as: $Sign = \frac{S}{\sqrt{S+B}}$, where S denotes the number of signal events and B the number of background events. This definition only includes statistical (Poisson) uncertainties on the number of events. When systematic uncertainties are taken into account, an extra factor in the denominator is added as will be explained in Section 4.4.1.

The analysis is applied to simulated samples for 13 TeV proton-proton collisions and for an integrated luminosity of 100 fb^{-1}. Two benchmark points of the signal sample will be compared: one for a dark matter particle mass of 30 GeV to which the FC top-quark decay ($t \to c\chi\chi$) is kinematically allowed and one for a dark matter particle mass of 90 GeV to which this decay is not kinematically allowed. In Section 4.4.2 a scan of part of the $c^{23} - m_\chi$ parameter space will also briefly be discussed. Besides that, two signal regions will be considered based on the jet multiplicity: signal region 1 in which exactly one jet is present (which is b tagged) and signal region 2 in which exactly two jets are present, one of which is b tagged and another one that is a candidate to be c tagged.

The charm tagger developed in Chapter 3 will be implemented to be used within the MadAnalysis software and will be applied to events from signal region 2. The implementation of a charm tagger relies on the MVA training files which are needed to map the jet properties onto a single discriminator value on which a cut can be applied. In order to apply the charm tagger to real LHC collision data, a framework will have to be developed to read the training files from the standalone TMVA c tagger and interpret them from within the CMSSW software. This framework is still in development. For this first study, the use of Delphes simulated samples within the MadAnalysis software does not allow the use of such MVA training files and consequently the discriminator values for each jet are not available in this analysis. This problem is partially solved by implementing a parametrisation of the charm-tagging efficiencies as a function of jet p_T. This means that for each jet a random number is generated and based on the tagging efficiency for the flavour and the p_T of the considered jet it will be tagged or not. Accordingly no information of the jet properties is used, but the number of tagged jets from a certain flavour will be similar to that in a real analysis. This process is considered accurate enough to get a first idea of the effect of the charm tagger in this analysis.
The parametrisation of the charm tagger as a function of p_T is taken from the fits to the data points in Figure 3.16. These fits are polynomials of different orders (from order 3 to order 7) and separate fitting is performed for jets with a p_T between 30 and 200 GeV and for jets with a p_T larger than 200 GeV. The loose working point (see Table 3.3) was used, for which the reason will be further explained in Section 4.3 after discussing the backgrounds in this analysis. For the b tagger the same parametrisation procedure was used.
One more difficulty arises in combining the c tagger with the b tagger during the analysis. As will be outlined in Section 4.3, first a requirement on the number of b-tagged jets is made and only later a requirement on the number of c-tagged jets is made. A b-tagged jet is not considered any more for c tagging later on, which is why the c tagger can only be applied

to signal region 2 in which two jets are present, one of which has to be b tagged. The fact that b-tagged jets are not considered any more for c tagging may alter the parametrisation of the charm-tagging efficiencies as explained above. If the true discriminator values of each jet would be used this would not be the case, but by removing jets that are b tagged the random selection of jets based on the parametrisation of the performance of the charm tagger can be misleading. No real solution exists within this analysis setup, but this effect can be minimised by applying a tight b-tag selection. This way a very clean selection of b jets will be made and the mistag efficiency for c and light-flavour jets is very low, causing almost no loss of c and light-flavour jets during the b jet identification. Even though the tight b-tag working point may not be the most optimal one in terms of signal significance, it will be used here to have a more trustworthy implementation of the c-tagger efficiency in this framework.

During the analysis the signal significance will be measured in different situations. First a simple cut and count based significance without including any systematic uncertainties will be calculated and compared for the different benchmark points (different DM masses), the different signal regions (1 or 2 jets) and with or without a c-tagging requirement. After this the effect of including systematic uncertainties will be discussed, which will lower the significance drastically. To recover some performance, the cuts will then be optimised including systematics and the significance will again be calculated. Finally a method using template fitting will be presented, which will illustrate how to lower the effect of systematic uncertainties by exploiting the shape of variable distributions in combination with the number of observed events. First however the relevant background processes will be discussed and the signal selection criteria on discriminating variables will be presented.

4.2 Discussion of the background

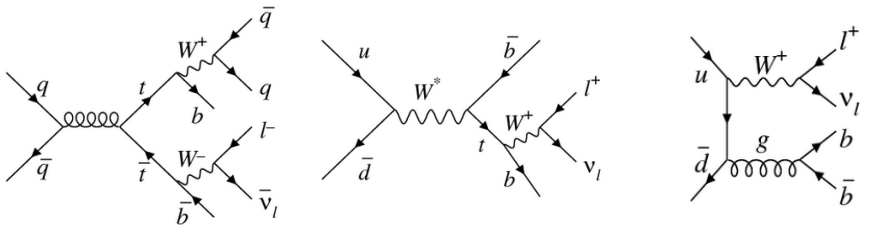
In Section 2.4 the different processes resulting from the phenomenological model were discussed. This topology results in finals states with missing transverse energy, a b jet from the SM decay of a top quark, possibly a c jet and a lepton since this analysis focuses on the leptonic decay of the W boson from the SM top decay. It should always be kept in mind that extra radiation can add more jets to these processes. Keeping this topology in mind, three main backgrounds have to be considered during the analysis:

- $\boldsymbol{t\bar{t}}$: at the LHC a lot of top quark pair production takes place and although these events contain two SM top-quark decays, the resulting b jets will not always be reconstructed or tagged as a b jet. The main contribution will come from semileptonic decays in which one of the W bosons decays leptonically and the other one hadronically, resulting in one final-state lepton. A representative Feynman diagram is shown in Figure 4.1(a). The total cross section of semileptonic and dileptonic $t\bar{t}$ events in proton-proton collisions at 13 TeV (calculated with MadGraph at leading order) is 366.1 pb.
- **Single top:** Single top production [67] is often associated with extra jets (possibly b jets) and possibly with a W boson. If such a W boson is present in the final state and it decays leptonically this process will be an important contribution to the background.

A representative Feynman diagram is shown in Figure 4.1(b). The total cross section of single top events in proton-proton collisions at 13 TeV (calculated with MadGraph at leading order) is 277.8 pb.

- **W + jets:** The production of a W boson (with a leptonic decay) in association with jets should of course also be considered, especially if one of these jets is a b jet (or gets b tagged). A representative Feynman diagram is shown in Figure 4.1(c). The total cross section of W + jets events with a jet multiplicity between 1 and 4 jets in proton-proton collisions at 13 TeV (calculated with MadGraph at leading order) is 10751.3 pb.

The presence of only one final-state lepton for the signal removes all processes with Z bosons from the list of relevant backgrounds. Furthermore the presence of large missing transverse energy due to the dark matter can be exploited to distinguish the signal from the background, as will be explained in the next section.



(a) $t\bar{t}$ (b) Single Top (c) W+jets

Figure 4.1: Representative diagrams [68] for the different background processes. These diagrams just serve as an illustration and in practice many more diagrams contribute to these SM background processes.

4.3 Signal selection criteria

In order to distinguish the signal from the SM background processes discussed above, discriminating variables have to be identified on which efficient cuts can be made. The main features of the signal are the presence of large missing energy due to the dark matter particles that escape the detector and the presence of a b jet and in some cases a c jet in the final state. Besides that this search focuses on the leptonic decay of the W boson to an electron or a muon and their corresponding neutrinos, resulting in exactly one isolated electron or muon and some more missing energy due to the neutrino. These specific features are the motivation behind the cuts that will now be discussed. The specific choice of the cut values is based on an optimisation of the statistical signal significance ($\frac{S}{\sqrt{S+B}}$).

cut 1 **1 isolated electron or muon ($p_T^\ell > 30$ GeV):** The isolation of a particle is calculated by taking the sum of the p_T of all the particles that fall within a cone with an angular radius ΔR around that particle, while typically subtracting the p_T of particles in a much smaller inner cone to avoid including radiation from that particle or the particle itself. If an electron or muon is produced in a jet, it will have a lot of energy surrounding it, whereas promptly produced electrons or muons have only a small amount of energy

surrounding them and are therefore well isolated. This search focuses on the leptonic decay of the W boson into an electron or a muon and the corresponding neutrino. This will result in exactly one isolated electron or muon on which a minimal p_T cut is applied of 30 GeV.

cut 2 **1 ≤ jet multiplicity ≤ 2 (p_T^{jet} > 30 GeV):** The signal processes as illustrated in Figure 2.6 contain either one or two jets in the final state. Radiation can add extra jets, but these are not taken into account in the determination of the value of this cut. The normalised distribution of the number of jets for signal and background is shown in Figure 4.2(a). It can be seen that this cut is especially efficient in reducing the $t\bar{t}$ background in which the average jet multiplicity is much higher, especially in the case where one of the W bosons decays hadronically (see for example Figure 4.1(a)). This cut also includes the two signal regions introduced before. The p_T of the jets is required to be larger than 30 GeV (which is also the minimal p_T used in the development of the charm tagger).

cut 3 **1 b-tagged jet:** The signal always contains exactly one b jet. A tight b tag is applied to identify jets from b-quarks. The normalised distributions of the number of b jets for signal and background after requiring 1 isolated electron or muon are shown in Figure 4.2(b). These distributions are also drawn just before the cut on the number of b jets as shown in Figure 4.3(b). Due to the efficiency of the tight b tag more than half of the b jets are not tagged. Nevertheless requiring one b-tagged jet highly reduces the W+jets background since processes with a W boson in association with b jets (see for example Figure 4.1(c)) are rare in comparison to processes with W bosons in association with light-flavour jets. The $t\bar{t}$ background is the only one in which sometimes events with 2 b-tagged jets appear. These events are also eliminated with this cut.

cut 4 **M_T(lep,MET) > 150 GeV:** The transverse mass of the lepton and MET system is defined in Equation (4.1).

$$M_T(lep, MET) = \sqrt{2 \times p_T^{lep} \times MET \times \{1 - cos(\Delta\phi(lep, MET))\}} \tag{4.1}$$

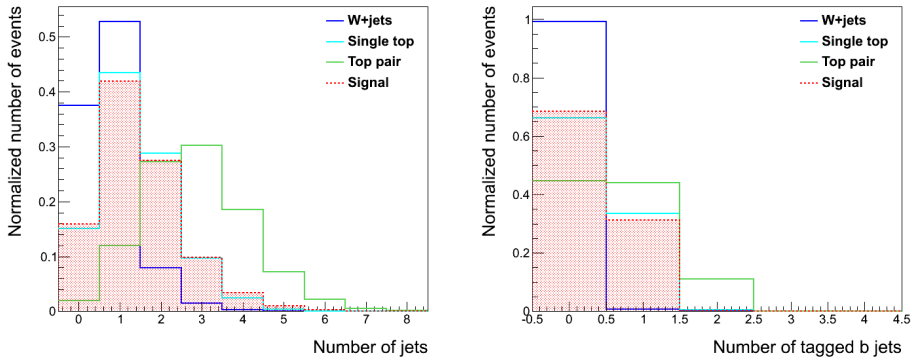
It is obviously somehow correlated to the cut on the MET itself, but it also uses the information on the difference in direction between the MET and the lepton. For the signal the main contribution to the MET comes from the dark matter particles which are expected to be approximately back-to-back with respect to the lepton from the W decay. This cosine in the definition of the M_T will be closer to -1 and together with the large MET this maximises the value of the transverse mass. For the background the MET comes from the neutrinos which are expected to travel in a direction close to the lepton due to the boost of the W boson system itself. The cosine will be closer to 1 and will minimise the value of the M_T. A drop in the M_T(lep,MET) around the mass of the W boson (~ 80 GeV) is also expected for the SM background processes. This is a consequence of the fact that if the MET originates from the neutrino from the W boson, the M_T(lep,MET) would be roughly equal to the mass of the W boson if that

W boson would decay in the transverse plane. This is shown in Figure 4.2(c) where the normalised distributions for signal and background are drawn after requiring 1 isolated electron or muon. These distributions are also drawn just before the cut on the M_T as shown in Figure 4.3(c).

cut 5 **MET > 200 GeV:** The presence of dark matter in the signal introduces a lot of missing energy in the detector. The missing transverse energy will therefore be much larger on average for the signal in comparison to the SM background processes in which the MET originates from neutrinos only. This can be seen from the normalised distributions in Figure 4.2(d) drawn after requiring 1 isolated electron or muon or in Figure 4.3(d) drawn just before the cut on the MET. A cut on the MET at 200 GeV will eliminate the main bulk of background events and still keeps the large tail of signal events.

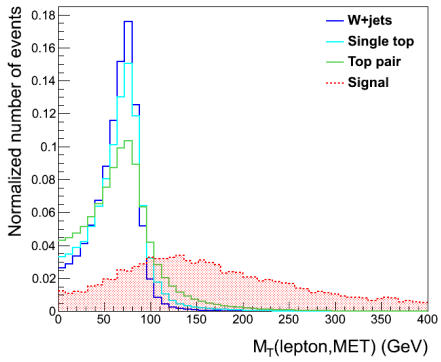
cut 6 **$|\Delta\phi$(lep,b jet)| < 1.6:** The difference in azimuthal angle ϕ between the lepton and the b jet is also a discriminating factor to reduce the SM background. For the signal the b jet always originates from the SM top-quark decay together with the production of a W boson that decays leptonically. The lepton and the b jet originate from the same mother particle and due to the boost of that system they are expected to travel in a direction close to each other. This will result in small values of $|\Delta\phi|$. For all of the background processes similar things can happen but it is also much more likely that the b jet and the lepton originate from different mother particles. Especially for the W+jets background in which the W boson and the b jet do not originate from a top quark decay. This causes the value of $|\Delta\phi|$ to be more centred around large values. This effect can clearly be seen in the normalised distributions (requiring 1 isolated electron or muon) shown in Figure 4.2(e). The distribution is shown again just before the cut on $|\Delta\phi|$ in Figure 4.3(e). Statistical fluctuation start to become visible in this distribution, especially for the W + jets sample, due to the fact that this sample is rather small. This is however not a problem since the main background contribution comes from $t\bar{t}$. A selection of $|\Delta\phi| < 1.6$ optimises the signal significance.

cut 7 **1 c-tagged jet:** Finally the effect of the charm tagger can be investigated. This cut is only included when comparing the significance with and without c tagging. In signal region 2 the signal has a c jet on top of the b jet as shown in Figure 2.6. Background processes are much less likely to have c jets in the final state. More specifically the W+jets background is dominated by light-flavour jets and the single top and $t\bar{t}$ backgrounds have both b- and light-flavour jets. It will become clear in the next section that the main background is $t\bar{t}$. For this reason the choice has been made to focus on C vs B discrimination with a high charm efficiency to keep a lot of signal events. This corresponds to the loose WP defined in Table 3.3. The requirement of having 1 loosely c-tagged jet is applied only to events that have two jets (signal region 2). The distribution of tagged c jets after requiring 1 isolated electron or muon is shown in Figure 4.2(f) and is drawn again in Figure 4.3(f) right before the cut on the number of c jets.

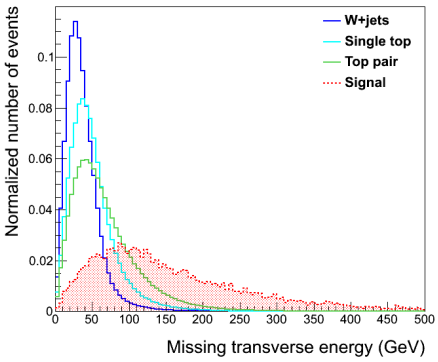


(a) Number of jets

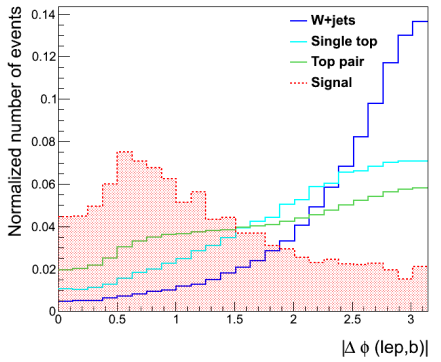
(b) Number of b-tagged jets



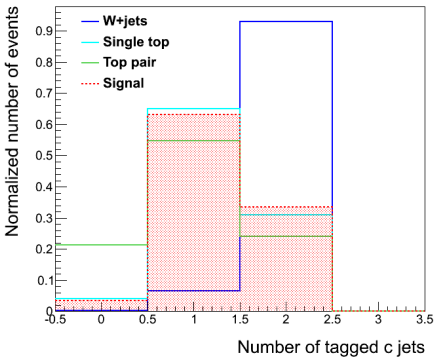
(c) Transverse mass (lepton,MET)



(d) Missing transverse energy



(e) $\Delta\phi$(lepton,b jet)

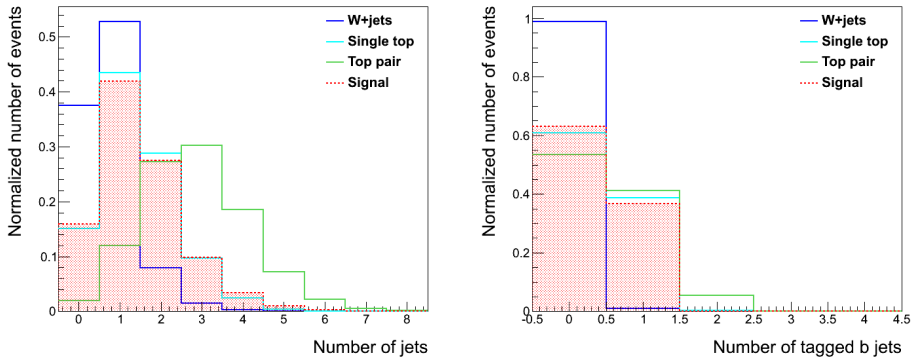


(f) Number of c-tagged jets

Figure 4.2: Normalised distributions of the variables used in the analysis for signal (Λ = 1 TeV, c^{23} = 10 and m_χ = 30 GeV) and background (see section 4.2). These distributions are drawn for events with exactly 1 isolated lepton ($p_T > 30$ GeV) but without applying any other cut. However, the distribution of $\Delta\phi$(lep,b) has the extra requirement of having exactly one b-tagged jet and the distribution of the number of c-tagged jets is drawn only for events with exactly 2 jets.

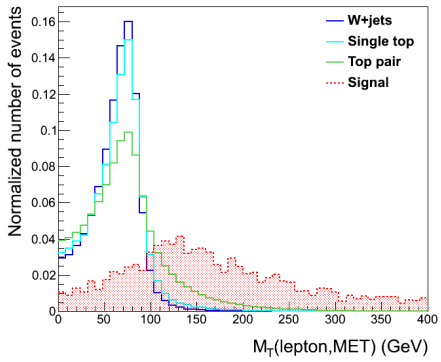
4.4 Analysis results

The results of the analysis will now be discussed in different situations. First the significance will be calculated from a simple cut and count based method without taking into account any systematic uncertainties. To put the acquired results in perspective of a real analysis

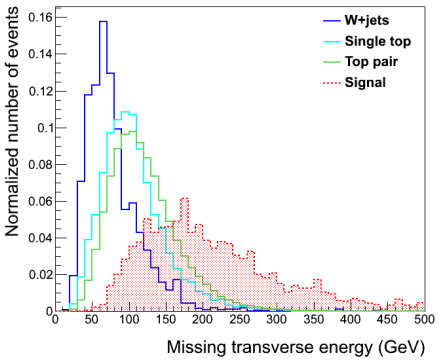


(a) Number of jets

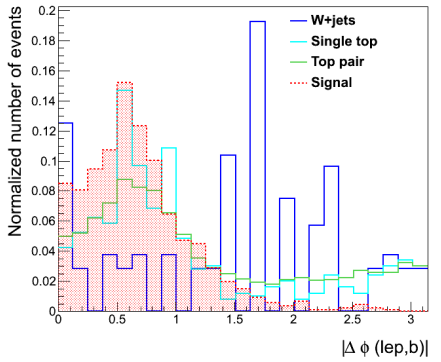
(b) Number of b-tagged jets



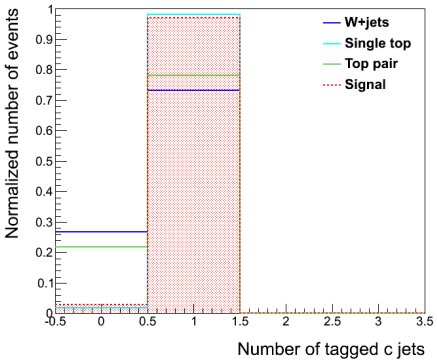
(c) Transverse mass (lepton,MET)



(d) Missing transverse energy



(e) $\Delta\phi$(lepton,b jet)



(f) Number of c-tagged jets

Figure 4.3: Normalised distributions of the variables used in the analysis for signal (Λ = 1 TeV, c^{23} = 10 and m_χ = 30 GeV) and background (see section 4.2). Each of these variable distributions is drawn right before the cut on that variable, following the order of the cuts outlined in Section 4.3.

the significance will then be quoted when some systematics are taken into account. It will be shown that this dramatically reduces the significance. To recover some performance the applied cuts will be changed and optimised for a significance with systematic uncertainties. Finally a method is presented based on template fitting to show that reasonable signal significance can be achieved while reducing the effect of systematic uncertainties by using the shape of the variable distributions in combination with the number of observed events.

4.4.1 Cut and count based method

No systematics The cuts discussed in Section 4.3 are chosen such that they optimise the statistical signal significance $\frac{S}{\sqrt{S+B}}$. The cuts are applied to the signal and background samples and this signal significance is calculated from the final number of selected signal (S) and background (B) events. The results are shown in Table 4.3 in the row labelled *"w/o syst"*. The significance is compared between the m_χ = 30 GeV and m_χ = 90 GeV benchmark points. For each of those the results are quoted before and after applying the requirement of having 1 c-tagged jet. The significance is always shown for signal region 1 and signal region 2 combined and separately.

The significance for the m_χ = 30 GeV benchmark point, for the combined signal regions before applying the c-tagging requirement reaches a value of 9.5. It should always be kept in mind that this does not include any systematic uncertainties. Almost the exact same result is acquired for the m_χ = 90 GeV benchmark point. Although the cross section of the signal is lower at a dark matter mass of 90 GeV (see Figure 2.8), the MET is expected to be higher which makes the cut on the MET more efficient for that benchmark point[1].
After the requirement of having 1 c-tagged jet, a small gain is seen in the combined significance from 9.50 to 9.93 for the 30 GeV dark matter mass (9.53 to 10.01 for 90 GeV dark matter mass). Only signal region 2 is responsible for this small gain since this is the only situation in which c tagging is applied. This gain is however too small to really contribute to the sensitivity of this search. From this we conclude that in order to make the c tagger useful for this analysis, something more advanced than simply counting c-tagged jets is to be developed.
The combination of both signal regions is more sensitive than each of them separately and signal region 1 is more sensitive than signal region 2 because of the larger cross section of signal events with one final-state jet (process 3 in Section 2.4). This effect is even more pronounced for the m_χ = 90 GeV benchmark point because the FC top-quark decay (expected in signal region 2) is not kinematically allowed.

To get an idea of the actual number of events Table 4.1 shows the number of events for 100 fb^{-1} of 13 TeV pp collision data after each cut. In this table it can be seen that the dominant remaining background is $t\bar{t}$.

With systematics Including only statistical uncertainties in the calculation of the signal significance is not realistic since often the systematic uncertainties are of similar or even greater importance than the statistical uncertainties. To illustrate this the significance will be calculated after adding a systematic uncertainty. It is not the goal of this research to determine with great precision every source of systematic uncertainties and its importance. Therefore a simple argumentation will be made to combine all possible systematics in one additional uncertainty.
Table 4.1 shows that $t\bar{t}$ is the most dominant remaining background after applying all the

[1] The efficiency of the MET cut at 200 GeV is ~ 47 % for the m_χ = 30 GeV benchmark point and ~ 51 % for the m_χ = 90 GeV benchmark point.

Cut	**W+jets**	**Single top**	**Top pair**	**Signal**
initial	1.4e+09	2.8e+07	3.9e+07	20031.8
1 isolated e or μ ($p_T > 30$ GeV)	3.9e+08	3.3e+06	1.3e+07	12285.8
$\geq$ 1 jets ($p_T > 30$ GeV)	2.6e+08	2.8e+06	1.3e+07	10328.5
$\leq$ 2 jets ($p_T > 30$ GeV)	2.5e+08	2.4e+06	5.0e+06	8528.42
1 b-tagged jet	2.7e+06	928461	2.1e+06	3134.76
M_T(lep,MET) > 150 GeV	10665.1	11741.3	125652	1722.2
MET > 200 GeV	134.2	454.7	7853.8	813.4
$\lvert\Delta\phi$(lep,b jet)$\rvert$ < 1.6	61.1	344.9	5574.7	779.7
1 jet ($p_T > 30$ GeV)	12.7	139.1	2169.8	403.4
2 jets ($p_T > 30$ GeV)	48.4	205.8	3404.9	376.3
1 c-tagged jet	48.2	341.2	4832.9	768.4
1 jet ($p_T > 30$ GeV)	12.7	139.1	2169.8	403.4
2 jets ($p_T > 30$ GeV)	35.5	202.1	2663.0	365.0

Table 4.1: Cutflow table quoting the number of events for 100 fb^{-1} of 13 TeV pp collision data for background and for the signal sample with Λ = 1 TeV, c^{23} = 10 and m_χ = 30 GeV. The cuts are optimised for a signal significance without systematics.

cuts. Therefore an important contribution to the systematic uncertainty will originate from the large uncertainty on the $t\bar{t}$ production cross section. Based on the results from reference [69], the systematic uncertainty on this cross section is of the order of 7%. Other systematic uncertainties on the $t\bar{t}$ rate may for example come from the corrections to the jet energy scale[2], which induce an effect of the order of 4% for jets with a p_T above 30 GeV [70]. Combining these two contributions results in a systematic uncertainty of $\sqrt{0.07^2 + 0.04^2} \simeq$ 8%, which will be rounded up to a combined uncertainty on the total background of 10% to take into account possible contributions from other systematic uncertainties. This is sufficient for the purpose of illustrating the effect of systematics on the signal significance.

The same cuts are applied (see Section 4.3), but this time the signal significance is calculated as $\frac{S}{\sqrt{S+B+(0.1\times B)^2}}$. This includes the statistical (Poisson) uncertainties on the number of signal (S) and background (B) events ($\sigma_S^2 = \sqrt{S}^2$ and $\sigma_B^2 = \sqrt{B}^2$) and the systematic uncertainty equal to 10% of the background events ($\sigma_{syst}^2 = (0.1 \times B)^2$). The results are shown in Table 4.3 in the row labelled *"w syst"*.
This clearly shows a large drop in significance because of the large amount of background events with respect to signal events. Since the background is almost ten times larger than the signal (see Table 4.1) the terms S and B in the denominator of the significance can actually be neglected and the significance is dominated by $\frac{S}{0.1\times B}$, which is consequently very low. It can now clearly be seen that the effect of c tagging is very small and will not be sufficient to improve the sensitivity of this search with the proposed cuts. When searching for a small signal over a large background with a large systematic uncertainty the cuts proposed in Section 4.3 are clearly not strong enough. This is why the cuts will be optimised for

[2]These results are for 8 TeV pp collisions, as results for 13 TeV collisions are not available yet.

$\frac{S}{\sqrt{S+B+(0.1\times B)^2}}$, resulting in much tighter cuts on the MET, M_T and $|\Delta\phi|$.

Optimised cuts with systematics When searching for a small signal over a large background with a large systematic uncertainty it is much more efficient to search in a signal region where almost no background is present. This means one has to cut much tighter in order to cut out a lot more background. The cuts on MET, M_T and $|\Delta\phi|$ were now chosen to optimise the signal significance with systematics ($\frac{S}{\sqrt{S+B+(0.1\times B)^2}}$) resulting in the following cut values:

- MET > 350 GeV
- M_T(lep,MET) > 300 GeV
- $|\Delta\phi$(lep,b jet)| < 1

The cutflow is shown in Table 4.2 and the significance is shown in Table 4.3 in the row labelled *"w opt syst"*. The same general features as before show up, however the new cuts increased the sensitivity of the search and a significance up to 7.83 is reached including systematic uncertainties for the 30 GeV dark matter mass. This shows the potential of searching for this kind of signal during run 2 of the LHC.

Cut	**W+jets**	**Single top**	**Top pair**	**Signal**
initial	1.4e+09	2.8e+07	3.9e+07	20031.8
1 isolated e or μ ($p_T > 30$ GeV)	3.9e+08	3.3e+06	1.3e+07	12285.8
≥ 1 jets ($p_T > 30$ GeV)	2.6e+08	2.8e+06	1.3e+07	10328.5
≤ 2 jets ($p_T > 30$ GeV)	2.5e+08	2.4e+06	5.0e+06	8528.42
1 b-tagged jet	2.7e+06	928461	2.1e+06	3134.76
M_T(lep,MET) > 300 GeV	431.9	852.9	7312.2	568.1
MET > 350 GeV	0 ± 4.2	10.1	220.9	171.3
$\|\Delta\phi$(lep,b jet)| < 1	0 ± 4.2	7.3	122.2	158.2
1 jet ($p_T > 30$ GeV)	0 ± 4.2	2.7	40.0	74.9
2 jets ($p_T > 30$ GeV)	0 ± 4.2	4.6	82.2	83.3
1 c-tagged jet	0 ± 4.2	7.3	107.2	157.2
1 jet ($p_T > 30$ GeV)	0 ± 4.2	2.7	40.0	74.9
2 jets ($p_T > 30$ GeV)	0 ± 4.2	4.6	67.2	82.4

Table 4.2: Cutflow table quoting the number of events for 100 fb^{-1} of 13 TeV pp collision data for background and for the signal sample with Λ = 1 TeV, c^{23} = 10 and m_χ = 30 GeV. The cuts are optimised for a signal significance with systematics. The uncertainty on the zero values is the equivalent of one simulated event scaled to 100 fb^{-1}.

4.4.2 Scan of the $c^{23} - m_\chi$ parameter space

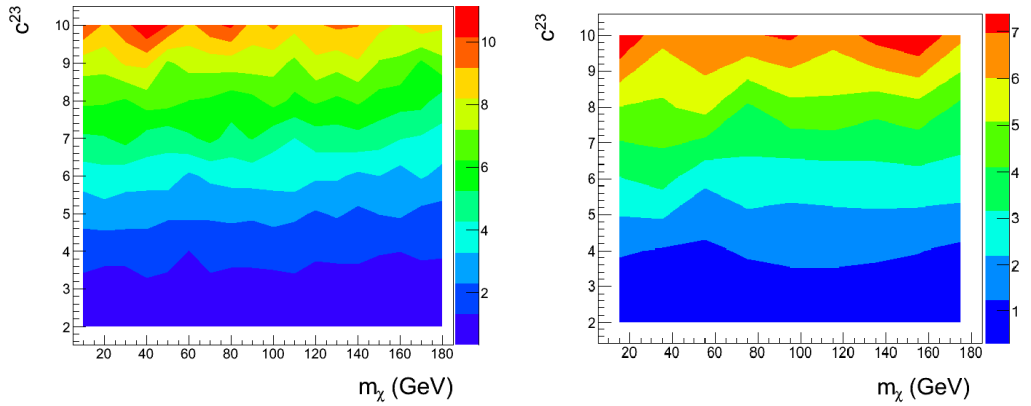
The two benchmark points discussed above illustrate that a larger dark matter mass results in a lower cross section, but the drop in signal significance is only moderate as the larger MET results in more efficient cuts on the kinematic variables. However, this comparison does not give information on the effect of the coupling strength c^{23} and only covers two

	m_χ = 30 GeV						m_χ = 90 GeV					
	without c tag			with c tag			without c tag			with c tag		
	comb	1 jet	2 jets	comb	1 jet	2 jets	comb	1 jet	2 jets	comb	1 jet	2 jets
w/o syst	9.50	7.73	5.92	9.93	7.73	6.39	9.53	8.15	5.62	10.01	8.15	6.10
w syst	1.29	1.70	1.01	1.46	1.70	1.23	1.30	1.80	0.96	1.47	1.80	1.18
w opt syst	7.41	6.42	5.32	7.83	6.42	5.75	6.72	5.97	4.75	7.15	5.97	5.20

Table 4.3: Summary of the acquired signal significance in the different situations discussed in section 4.4.1.

discrete points in the $c^{23} - m_\chi$ parameter phase space. It is therefore convenient to scan a region of that phase space and calculate the signal significance for each point in that scan.

A scan was performed ranging c^{23} between 2 and 10 in steps of 1 and varying m_χ between 10 GeV and 180 GeV in steps of 10 GeV. For each of these points the significance has been calculated with and without systematic uncertainties for 100 fb^{-1} of simulated data. The resulting contour plot is shown in Figure 4.4(a) without systematics and in Figure 4.4(b) with the cuts optimised with systematics, where the contours connect points in phase space that result in the same signal significance. The effects of c tagging have been illustrated in section 4.4.1 and will be similar for other points in the parameter space.
The signal significance does not depend too much on the dark matter mass m_χ. Larger dark matter masses result in a slightly lower cross section for the signal, but as discussed before the drop in significance is only moderate due to the higher efficiency of the MET cut. As the cross section depends on the square of the coupling strength, the significance drops rapidly with decreasing values of the coupling strength.
Figure 4.4(b) illustrates the region of phase space in which a discovery can be made ($\geq 5\sigma$) or a sign of evidence can be found ($\geq 3\sigma$) for this phenomenological model in the expected 100 fb^{-1} of data to be collected after the second run of the LHC.



(a) Without systematics (b) With systematics

Figure 4.4: Signal significance (a) without systematics and (b) with optimised cuts for systematics for different points in the $c^{23} - m_\chi$ parameter space. The contours connect points in the parameter space that result in the same signal significance. The significance does not depend too much on the dark matter mass m_χ, but drops rapidly with decreasing coupling strength c^{23}.

4.4.3 Template fitting

A cut-and-count based analysis as presented in Section 4.4.1 only takes into account a number of events that passes a series of cuts. It has been shown that the calculated signal significance is very sensitive to systematic uncertainties. This problem was partially solved by optimizing the cut values to a significance that includes systematics, but this solutions assumes there is no way to control the systematic uncertainties. The strength of template fitting methods lies in the fact that it can be less sensitive to systematic uncertainties.

Template fitting methods exploit the shape of a variable distribution for signal and background in combination with the number of events. As an input a template distribution of that variable for signal and for the different backgrounds is given to the fitting algorithm. The sum of all these distributions is made to determine the shape of the variable distribution for all selected signal and background events in an analysis. To test the performance of the method, this total distribution is used to generate pseudoexperiments. In each pseudoexperiment a set of pseudodata is generated by applying Poisson-distributed variations on the bin contents of the total distribution. For each of those pseudodata a binned maximum likelihood fit is performed with the signal and the total background templates such that the fitted templates show the best possible agreement to the pseudodata. This will result in a value of the fitted signal s_{fit} expressing the number of signal events that were obtained in the fit and an uncertainty on that value $\sigma_{s_{fit}}$ that results from the fitting procedure. The significance of the signal can be expressed as $s_{fit}/\sigma_{s_{fit}}$. Repeating this procedure by generating many pseudoexperiments will result in a distribution of the signal significance and the mean value of this distribution is taken as the final significance of the signal using the template fitting procedure.
The effect of systematic uncertainties on the amount of background is simulated by adding to the pseudodata a certain amount of events that corresponds to the uncertainty on that background. This results in an overall higher total distribution and the template fit is performed again for each pseudoexperiment. This will result in a second value for the fitted signal $s_{fit,syst}$. The definition of the significance is changed to $s_{fit}/\sqrt{\sigma^2_{s_{fit}} + \sum_i (s_{fit} - s^i_{fit,syst})^2}$, where i runs over the different background contributions. The extra uncertainty in the denominator will always be positive and will lower the significance, but it is expected to be small as the fitting procedure can still be able to fit the new accumulated distribution by just shifting the background distribution up, without altering the amount of signal events used in the fit. For this effect to be most optimal it is recommended to use a variable for which signal and background have a very different shape. This shows that template fitting methods are able to control systematic uncertainties on the background by exploiting the shape of the variable distribution, especially in regions where no signal is present.

In this research, a template fit on the distribution of $|\Delta\phi(lep, b)|$ is presented. The templates for signal and background are shown in Figure 4.5 and are drawn after requiring exactly 1 isolated electron or muon ($p_T > 30$ GeV), one or two jets ($p_T > 30$ GeV) of which 1 is tightly b tagged and a strong cut on the MET requiring MET > 350 GeV. These templates are assumed to be exact (no error on the templates is taken into account while creating the

pseudoexperiments). From the sum of the signal and background templates the pseudodata are generated, of which one example is also shown in Figure 4.5. A systematic uncertainty of 10% is added to the template distribution for $t\bar{t}$, since this is the main contribution to the background.

According to the procedure outlined in this section, the significance of this template fit was calculated with and without systematic uncertainties. A combination of 10.000 pseudoexperiments results in a mean value for the significance of the fit of 5.14 without systematics and 4.35 with systematics. With the exact same set of cuts this loss in sensitivity due to systematic uncertainties is much smaller than for a simple cut and count based analysis as discussed in Section 4.4.1. This is due to the fact that the region of high $|\Delta\phi(lep, b)|$ contains no signal and therefore serves to scale the level of the background in the template fit. Applying a systematic shift to the background over the entire variable range will result in an upwards shift of the fit to the background, but with almost the same fitted value for the signal strength. It should be noted that the acquired results are not yet optimized and will therefore not give the most optimal signal significance. A determination of the best fitting variable and the most optimal cuts to perform that fit is work to be performed in the future. Nevertheless this example serves as a first illustration of the use of template fitting methods and the low sensitivity to systematic uncertainties on the level of the background, while still obtaining a reasonable signal significance.

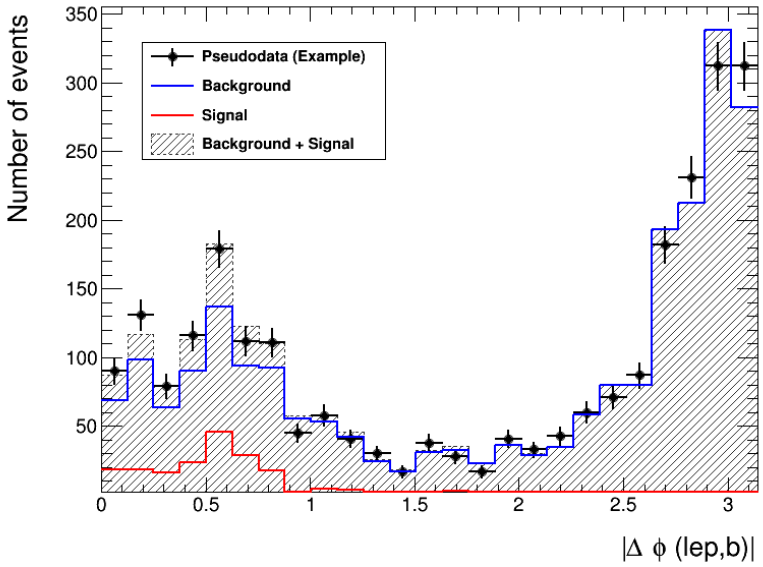


Figure 4.5: Signal (Λ = 1 TeV, c^{23} = 10 and m_χ = 30 GeV) template in red and total background template is blue for the distribution of $|\Delta\phi(lep, b)|$ together with the summed signal+background template (gray) and an example of the output of one of the pseudoexperiments (black) that are drawn from the summed signal+background template. The region of high $|\Delta\phi(lep, b)|$ contains no signal, which makes this a good variable for template fitting as this region serves to scale the level of the background.

Conclusion and outlook

Although cosmological observations provide indisputable evidence of the existence of dark matter, direct, indirect or collider searches have not yet been able to discover the true nature of dark matter and its interactions (other than gravitational) with the SM particles. This thesis presents a threefold research project with the common outlook of trying to reveal the nature of dark matter. First a recently developed phenomenological model is investigated to describe flavour changing interactions between top quarks and dark matter. Secondly a new charm-tagging algorithm is developed for the CMS experiment. Finally a first approximative study for a search for these flavour changing top-quark dark matter couplings at the LHC is presented and the potential use of the charm-tagging algorithm in this search is investigated.

The investigated phenomenological model uses an effective field theoretical approach to study the possibility of having flavour changing interaction between top quarks and dark matter. This can be achieved by extending the SM Lagrangian with an effective operator describing vector-mediated interactions between two Dirac dark matter fermions and a (right-handed) top quark and charm quark. Limits from the measured dark matter relic abundance have been studied and cross sections for the possible processes at the LHC have been calculated and were found to be around the order of 1 pb, high enough to be detected in the 100 fb^{-1} of 13 TeV collisions that are about to take place during the second run of the LHC.
It is yet to be investigated how direct and indirect searches will constrain the model parameters and how this model might provide a solution to the γ-ray excess from the center of the galaxy, measured by the Fermi-LAT experiment.

Because of the presence of final-state c jets in the processes predicted by the phenomenological model, a search for these flavour changing couplings could benefit from an algorithm to identify c jets. A new charm-tagging algorithm for the CMS experiment was developed in the TMVA framework and its performance was calculated on simulated samples. Information from displaced tracks, secondary vertices and soft leptons inside jets have been combined in multivariate analysis methods to discriminate charm jets from light-flavour and bottom jets. An optimisation of the performance in discriminating c jets from light-flavour jets has been performed, yielding a charm efficiency of 20% for a light-flavour efficiency or around 1% for the tight working point. This performance was found to be similar (although slightly worse) to the performance of the ATLAS charm-tagging algorithm, keeping in mind the many differences between the CMS and ATLAS setups and the fact that the charm-tagger for the CMS experiment is not yet completely optimised.
In order to further improve the performance of the TMVA-based charm-tagger, further

optimisations are required on both the discrimination between charm and light-flavour jets and between charm and bottom jets. The former requires a more detailed investigation of the secondary vertex reconstruction and for the latter a complete optimisation (sensitive variables, BDT settings and secondary vertex reconstruction) still has to be performed, since this will most likely be different from the charm to light-flavour discrimination. In order to use the charm-tagging algorithm in physics analyses on real proton-proton collision data, an interface has to be developed from the standalone TMVA-based framework to the CMS software framework. The performance that was found from simulations has to be calibrated to real collision data.

Finally a first approximative study has been presented to investigate the potential of a search for flavour changing top-quark dark matter couplings in 100 fb^{-1} of 13 TeV proton-proton collisions at the LHC and to look at the effect of charm tagging on the sensitivity of this search. A set of cuts is applied based on the presence of large MET and specific angular distributions between final-state particles in the signal events. A cut and count based method including the effect of systematic uncertainties shows that a reasonable signal significance (above five) can be reached in a large part of the parameter space (defined by the coupling strength c^{23} and the mass of the dark matter particle m_χ), showing that a potential discovery can be made or exclusion limits can be derived for the model parameters. Template fitting methods have shown to reduce the effect of systematic uncertainties, while still obtaining a reasonable signal significance. In all of these search strategies the effect of the currently available charm tagger is only very small. From this we conclude that in order to make the charm tagger useful for this analysis, something more advanced than simply counting c-tagged jets (like for example using the shape of the discriminator distributions) is to be developed.
It might be useful to search for new parameters that exploit the differences in kinematics between the signal and the background and improve the sensitivity of this search. Instead of applying subsequent cuts on different parameters, these could be combined in multivariate analysis (MVA) techniques to optimise the performance of the cuts that are applied or to apply template fitting methods to the discriminator distributions that result from the MVA. The optimisation of the charm tagger will hopefully give a visibly better performance for this search. The insights obtained from this first study show that this analysis can be effective and should therefore serve as a basis for a more dedicated full study.

In conclusion, a new phenomenological model involving flavour changing top-quark dark matter interactions has been investigated. A new charm-tagging algorithm for the CMS experiment has been developed and its application in the analysis of this new model has been studied in 13 TeV proton-proton collisions at the LHC.

Acknowledgements

This master thesis has been an amazing opportunity for me to work on my own research project and gain expertise in many different branches of experimental high-energy physics. This can only be achieved within a well established scientific environment like the IIHE and with the help of many other researchers sharing the same passion for this topic. Therefore I would like to thank all of the members of the IIHE to welcome me and help me whenever needed.

Special gratitude goes towards my promotor, Jorgen D'Hondt, whose motivation, inspiration, knowledge and support allowed me to work on state-of-the-art topics as an undergraduate student. The trust I was given to work on this research project is extremely motivating and the novel ideas he provided me with allowed me to push my boundaries in knowledge and practical skills. I would also like to express my special gratitude towards my copromotor, Gerrit Van Onsem, from whom I learned a great deal and who has provided me with help and answers whenever needed and always on very short notice. Not only did he provide me with the necessary samples, pieces of code or references such that I could work fluently, but he was always prepared to give extensive explanations or to have interesting discussions about any physics-related topic that helped me to make this thesis.

I am thankful to have witnessed and to have followed the work of Alberto Mariotti, Pantelis Tziveloglou and Kentarou Mawatari on the development of the phenomenological model for FC top-quark dark matter interactions. They involved me in this research from the start and gave me clear explanations on any topic related to phenomenological model building. I also want to thank Clemens Lange for the development of the standalone TMVA setup in which I developed the charm tagger and Valère Lambert for his contributions to the development of the charm tagger and for the pleasant collaboration.

Finally it should not be forgotten that working on a research project for such a period of time can only be done efficiently with the support and help of family and friends. I consider myself lucky to have such a supportive and motivational group of people behind me.

Appendices

Appendix A

Variable definitions and distributions

Variable	Definition
jetPt	the transverse momentum of the jet
jetEta	the pseudorapidity of the jet
track-Sip3d(2d)Sig(Val)	the signed (transverse) impact parameter significance (value) of each selected track
track-Sip3d(2d)Sig(Val) AboveCharm	the signed (transverse) impact parameter significance (value) of the track that raises the mass obtained from the summed four-momenta of the current track with the previous tracks (sorted in Sip2dSig) above the mass of the charm quark ($\simeq$ 1.5 GeV)
trackPtRel	the track transverse momentum, relative to the jet axis
trackPPar	the track momentum parallel to the jet direction, i.e. the scalar product of the jet direction and the track momentum, which basically projects the track momentum on the jet direction
trackEtaRel	η_{rel} = 0.5ln$\left(\frac{E+trackPPar}{E-trackPPar}\right)$ with $E = \sqrt{\lvert\vec{p}_{track}\rvert^2 + m_\pi^2}$ and trackPPar defined just above. This is in fact the track pseudorapidity, relative to the jet axis
trackDeltaR	ΔR between the jet direction and the track momentum
trackPtRatio	the track transverse momentum, relative to the jet axis, normalised to the magnitude of its momentum
trackPParRatio	the track momentum parallel to the jet direction, normalised to the magnitude of its momentum
trackJetDist	the distance between the track and the jet axis
trackDecayLenVal	the decay length of the track calculated as the distance between the primary vertex and the point of closest approach of the track with respect to the jet axis
trackSumJetEtRatio	the ratio of the transverse energy of the summed four-momenta of all selected tracks and the transverse energy of the jet

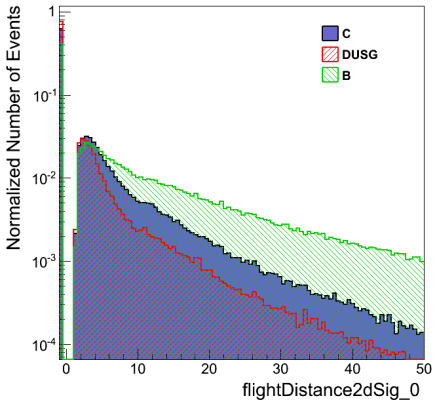
Table A.1: Variables used in the development of the c tagger with their definition [59]. Some of these variables have multiple values for one jet and show up as vectors in the samples. These vectors are removed and each value is assigned a new variable with a suffix _0, _1 or _2 for the first second or third element of the vector respectively, as can be seen in the variable distributions below.

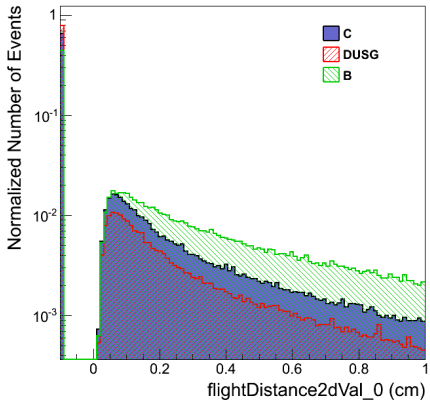
Variable	Definition
trackSumJetDeltaR	ΔR between the summed four-momenta of all selected tracks and the jet direction
vertexMass	mass of the track sum at the secondary vertex
vertexNTracks	the number of tracks associated with the secondary vertex
vertexEnergyRatio	the ratio of the energy of the summed four-momenta of all secondary vertex tracks and the energy of the summed four-momenta of all tracks associated with the jet
vertexJetDeltaR	ΔR between the summed four-momenta of all secondary-vertex tracks and the jet direction
flightDistance3d(2d)Sig(Val)	the significance (value) of the (transverse) distance between the primary and the secondary vertex.
jetNSecondaryVertices	the number of reconstructed secondary vertices (of the type RecoVertex)
jetNTracks	the number of tracks associated to the jet
massVertexEnergyFraction	vertex mass times the fraction of the vertex energy with respect to the jet energy
vertexBoostOverSqrtJetPt	variable related to the boost of the vertex system in flight direction
leptonSip3d(2d)	the signed (transverse) impact parameter significance of each selected soft lepton
leptonPtRel	transverse momentum of the soft lepton with respect to the jet axis
leptonDeltaR	ΔR between the jet direction and the soft lepton momentum
leptonRatio(Rel)	momentum of the soft lepton (parallel to jet axis) over jet energy
leptonEtaRel	pseudorapidity of the soft lepton along jet axis

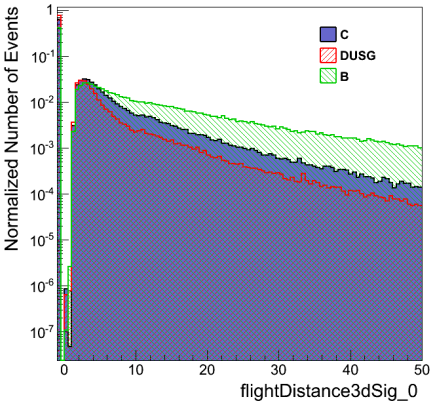
Table A.2: Continuation of Table A.1.

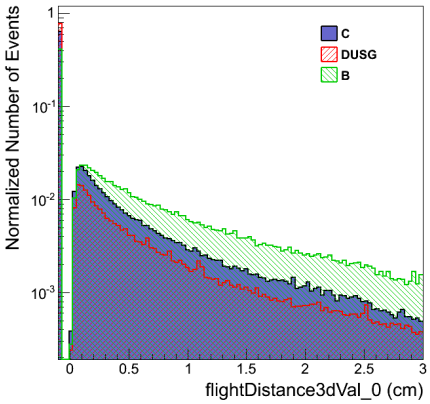
Variable distributions

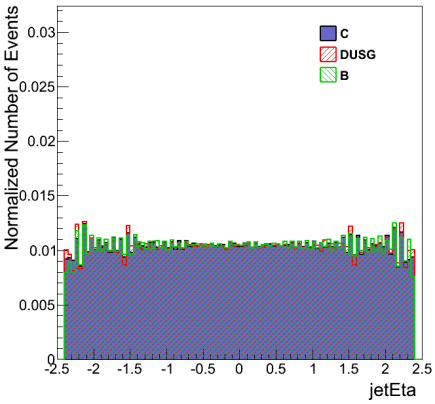
The following variable distributions, drawn from the 13 TeV samples (CMSSW 70X) with soft-lepton information, are weighed and normalised for each flavour.

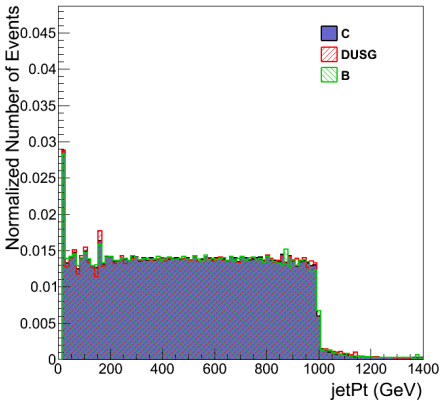


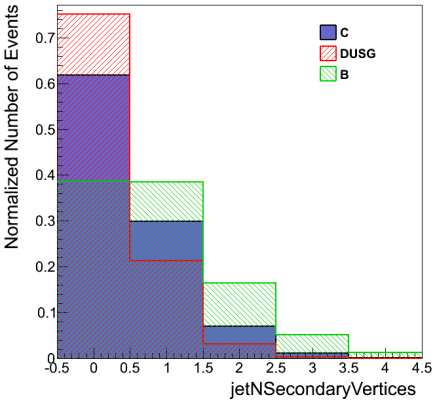


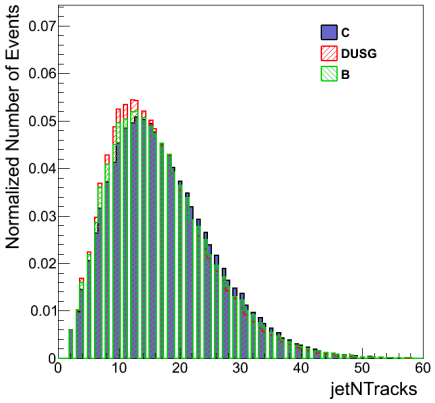

C
DUSG
B
Normalized Number of Events
flightDistance3dSig_0

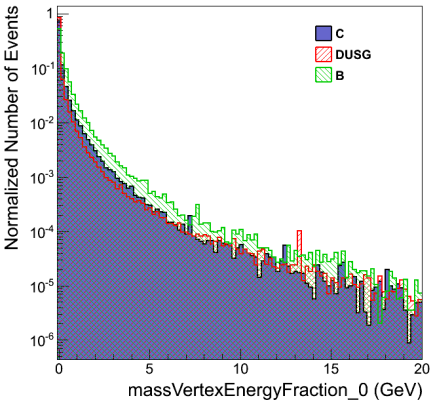

C
DUSG
B
Normalized Number of Events
flightDistance3dVal_0 (cm)

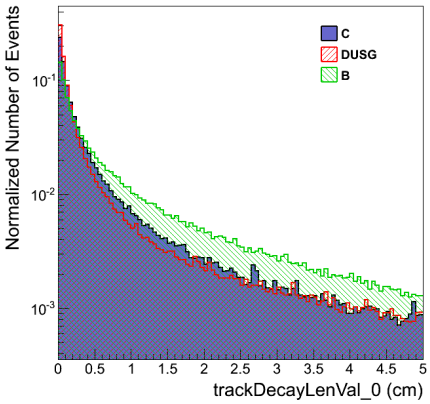

C
DUSG
B
Normalized Number of Events
jetEta

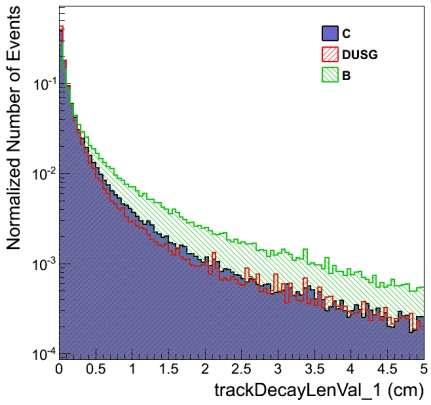

C
DUSG
B
Normalized Number of Events
jetPt (GeV)

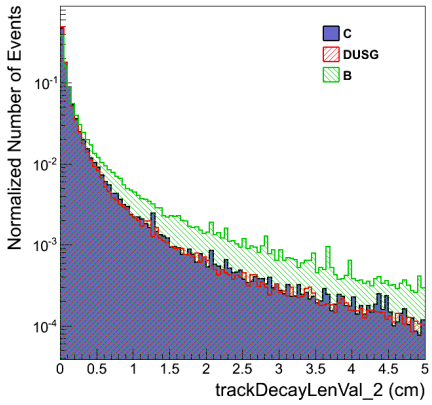

C
DUSG
B
Normalized Number of Events
jetNSecondaryVertices

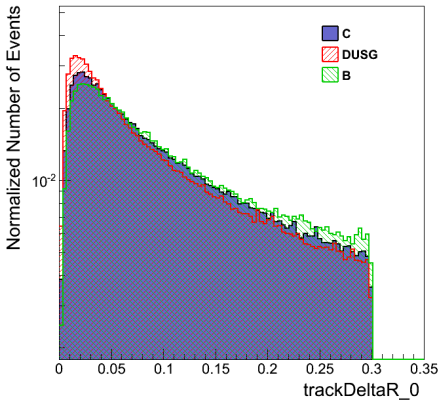

C
DUSG
B
Normalized Number of Events
jetNTracks

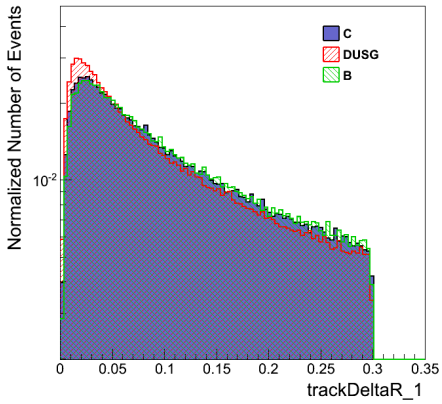

C
DUSG
B
Normalized Number of Events
massVertexEnergyFraction_0 (GeV)

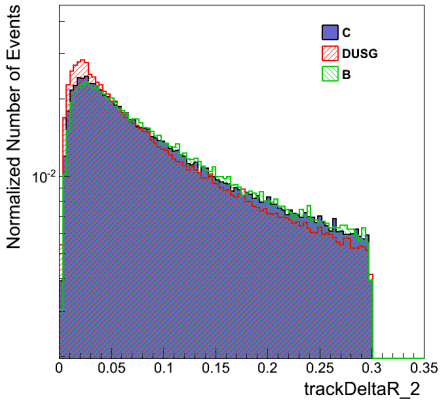

C
DUSG
B
Normalized Number of Events
trackDecayLenVal_0 (cm)

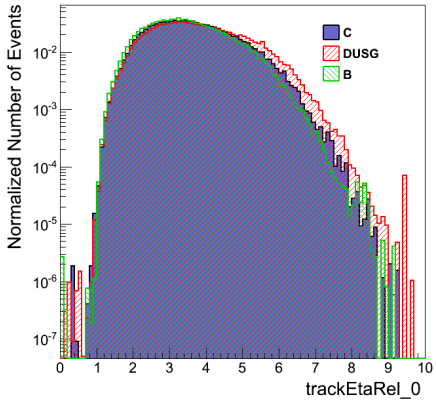

C
DUSG
B
Normalized Number of Events
trackDecayLenVal_1 (cm)

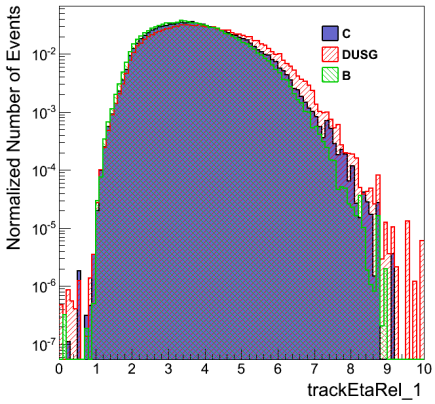

C
DUSG
B
Normalized Number of Events
trackDecayLenVal_2 (cm)

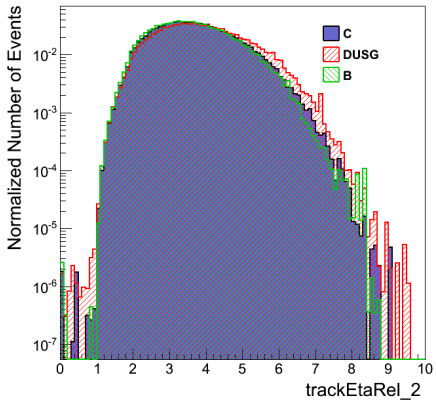

C
DUSG
B
Normalized Number of Events
trackDeltaR_0

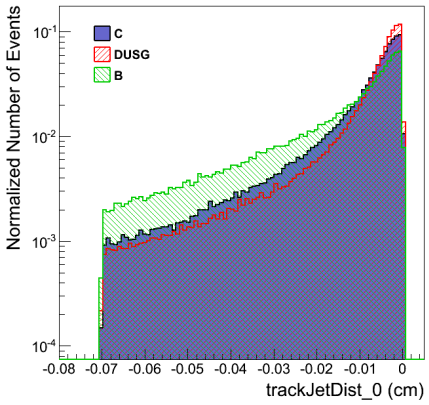

C
DUSG
B
Normalized Number of Events
trackDeltaR_1

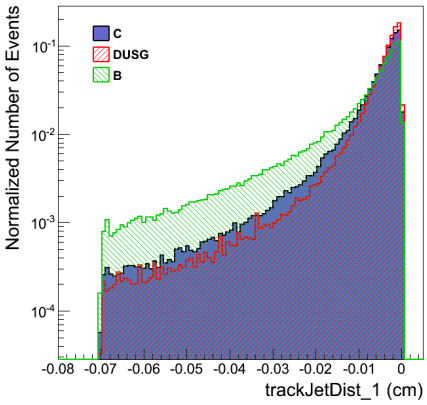

C
DUSG
B
Normalized Number of Events
trackDeltaR_2

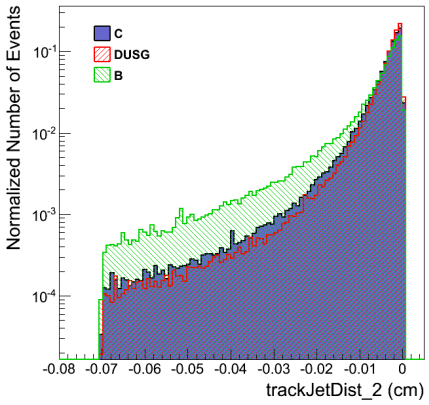

C
DUSG
B
Normalized Number of Events
trackEtaRel_0

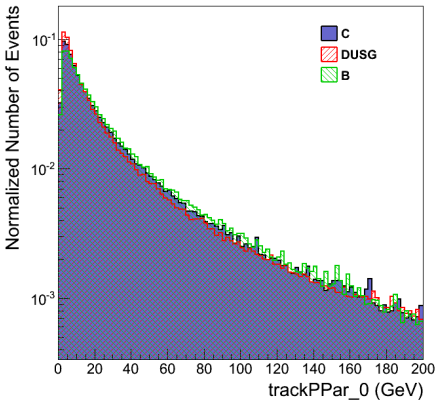

C
DUSG
B
Normalized Number of Events
trackEtaRel_1

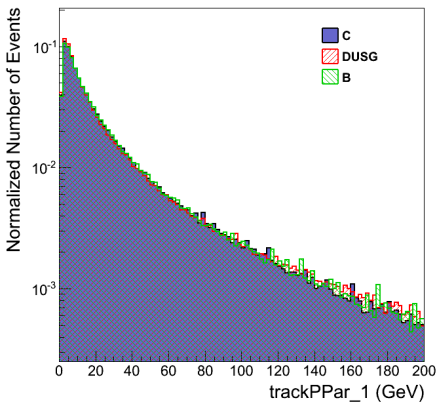

C
DUSG
B
Normalized Number of Events
trackEtaRel_2

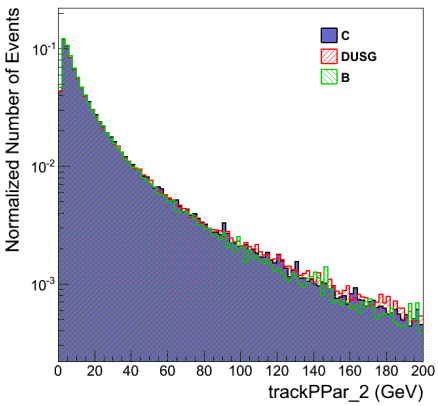

C
DUSG
B
Normalized Number of Events
trackJetDist_0 (cm)

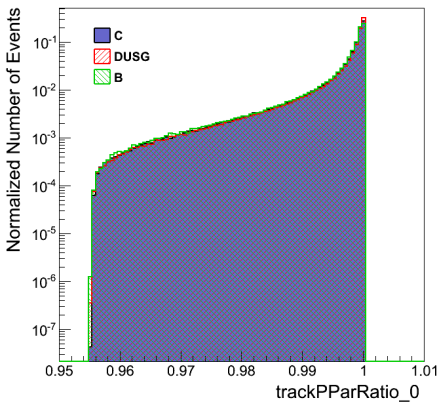

C
DUSG
B
Normalized Number of Events
trackJetDist_1 (cm)

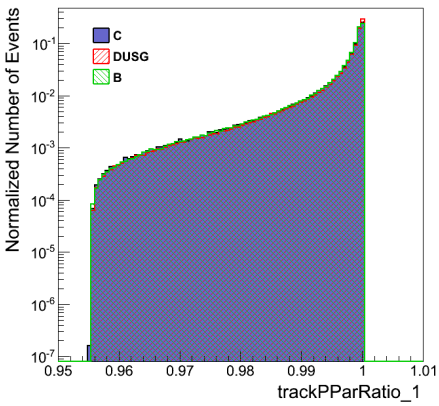

C
DUSG
B
Normalized Number of Events
trackJetDist_2 (cm)

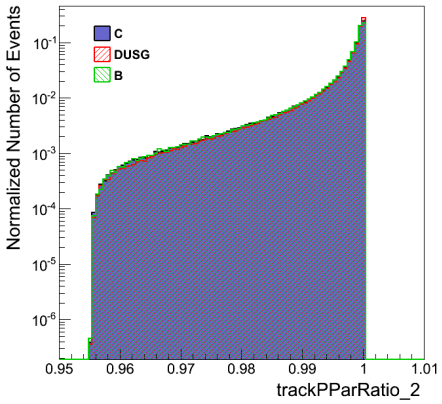

C
DUSG
B
Normalized Number of Events
trackPPar_0 (GeV)

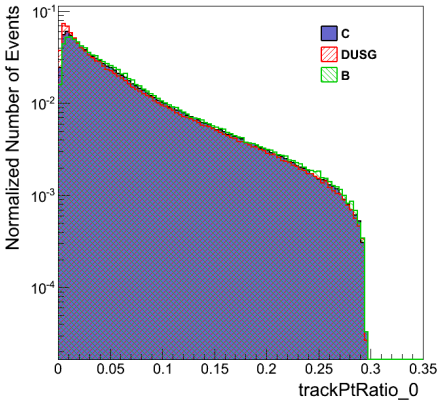

C
DUSG
B
Normalized Number of Events
trackPPar_1 (GeV)

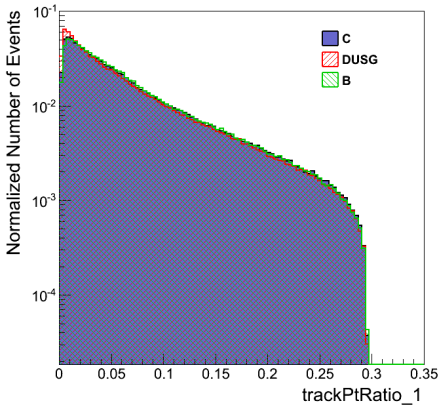

C
DUSG
B
Normalized Number of Events
trackPPar_2 (GeV)

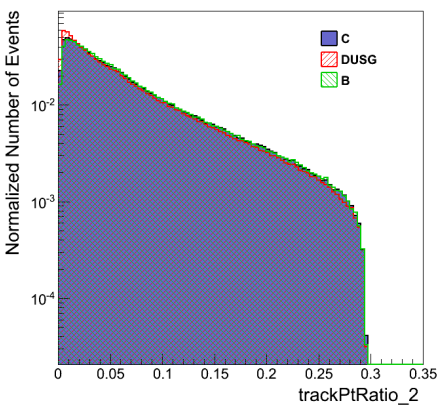

C
DUSG
B
Normalized Number of Events
trackPParRatio_0

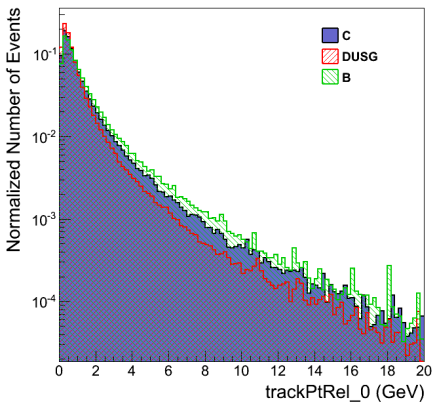

C
DUSG
B
Normalized Number of Events
trackPParRatio_1

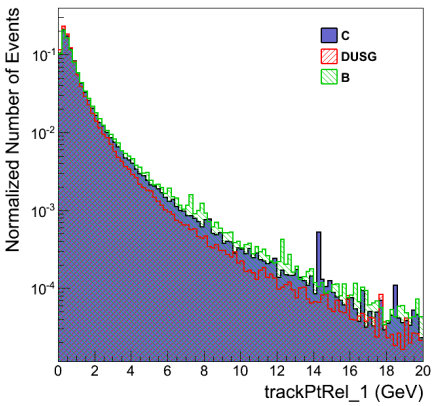

C
DUSG
B
Normalized Number of Events
trackPParRatio_2

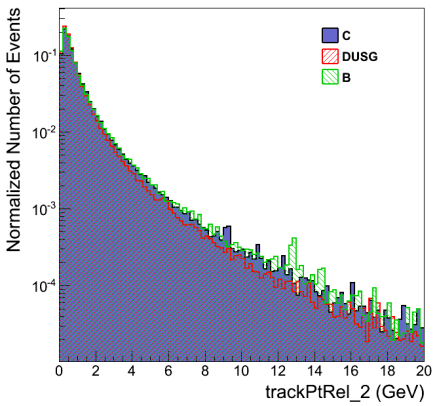

C
DUSG
B
Normalized Number of Events
trackPtRatio_0

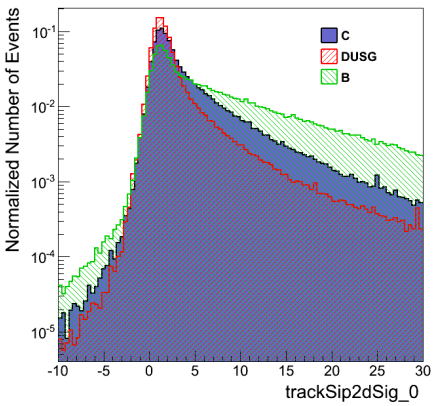

C
DUSG
B
Normalized Number of Events
trackPtRatio_1

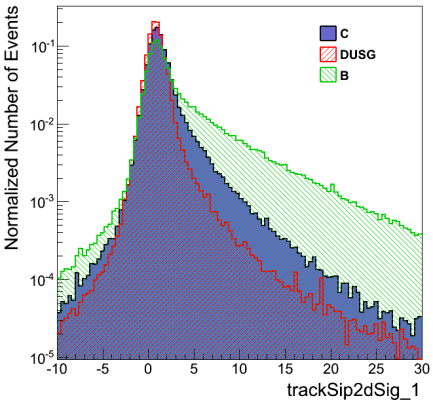

C
DUSG
B
Normalized Number of Events
trackPtRatio_2

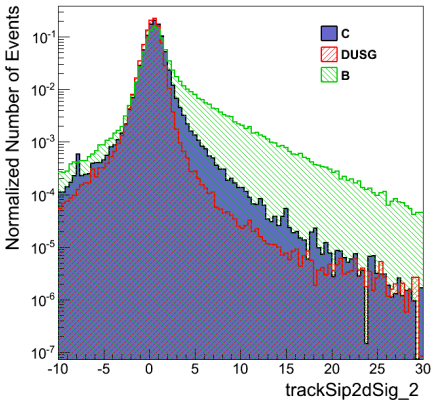

C
DUSG
B
Normalized Number of Events
trackPtRel_0 (GeV)

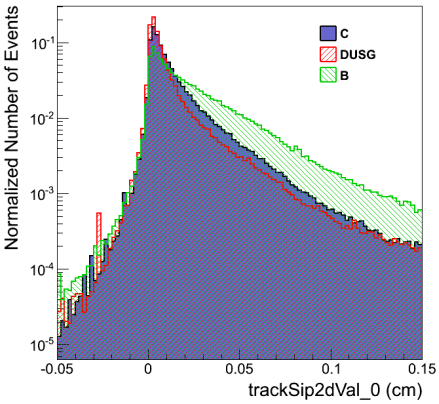

C
DUSG
B
Normalized Number of Events
trackPtRel_1 (GeV)

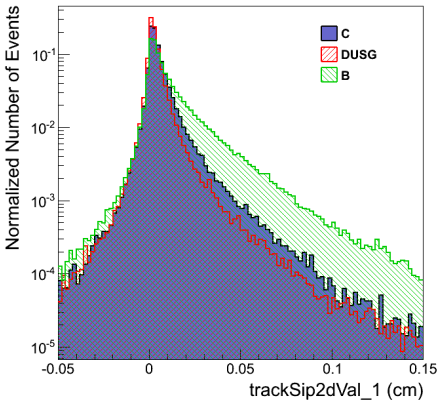

C
DUSG
B
Normalized Number of Events
trackPtRel_2 (GeV)

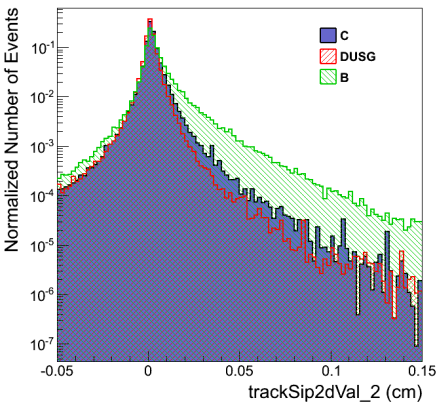

C
DUSG
B
Normalized Number of Events
trackSip2dSig_0

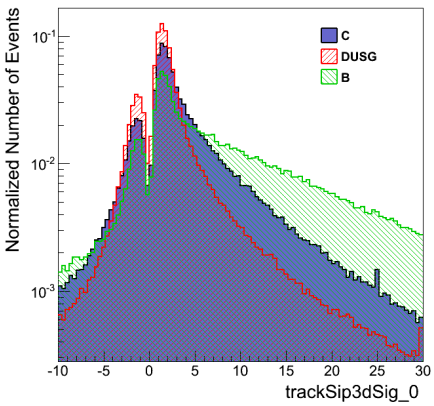

C
DUSG
B
Normalized Number of Events
trackSip2dSig_1

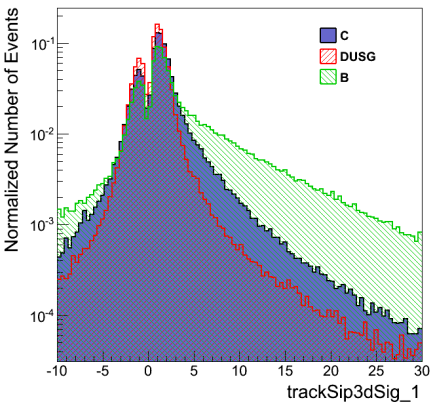

C
DUSG
B
Normalized Number of Events
trackSip2dSig_2

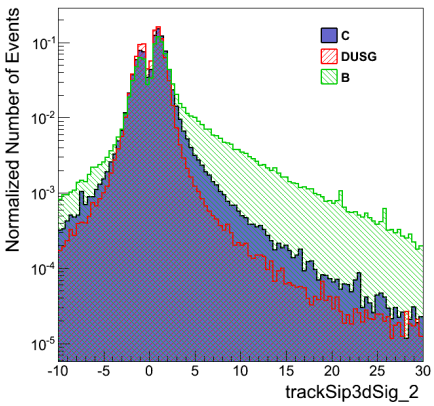

C
DUSG
B
Normalized Number of Events
trackSip2dVal_0 (cm)

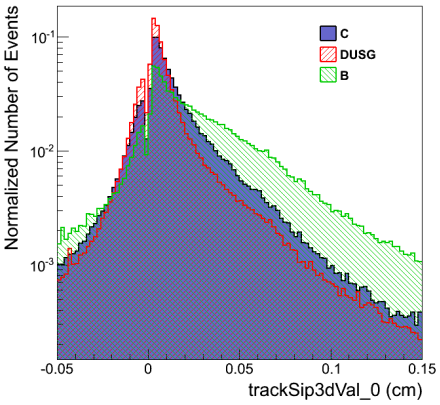

C
DUSG
B
Normalized Number of Events
trackSip2dVal_1 (cm)

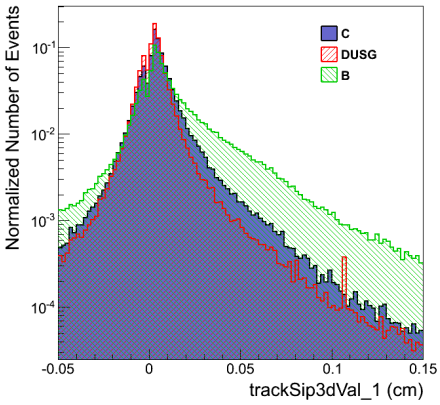

C
DUSG
B
Normalized Number of Events
trackSip2dVal_2 (cm)

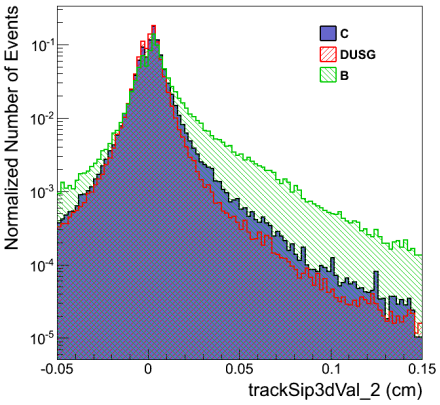

C
DUSG
B
Normalized Number of Events
trackSip3dSig_0

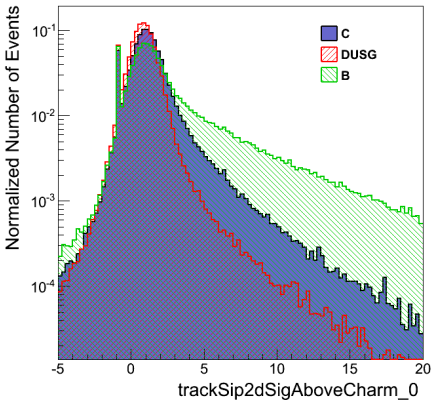

C
DUSG
B
Normalized Number of Events
trackSip3dSig_1

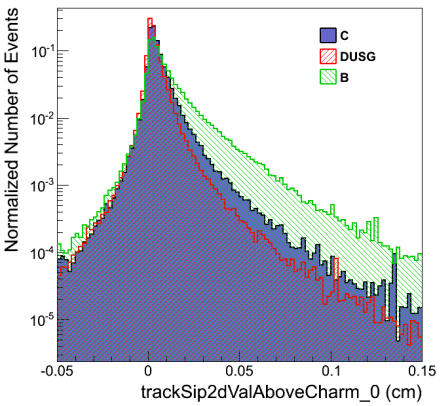

C
DUSG
B
Normalized Number of Events
trackSip3dSig_2

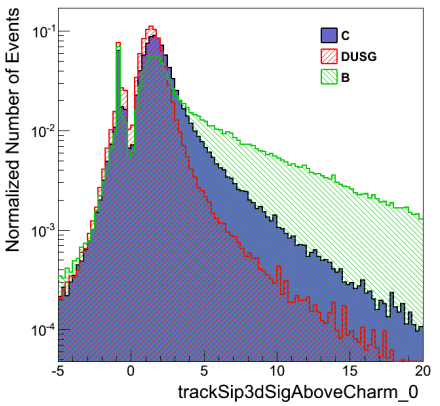

Normalized Number of Events
C
DUSG
B
trackSip3dVal_0 (cm)

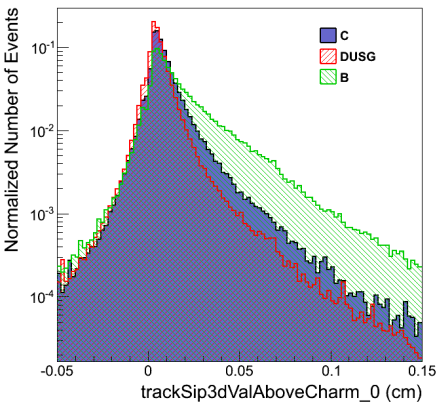

Normalized Number of Events
C
DUSG
B
trackSip3dVal_1 (cm)

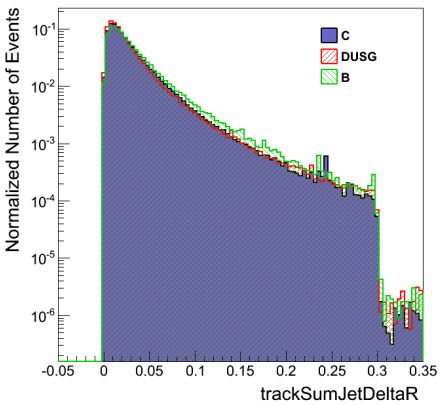

Normalized Number of Events
C
DUSG
B
trackSip3dVal_2 (cm)

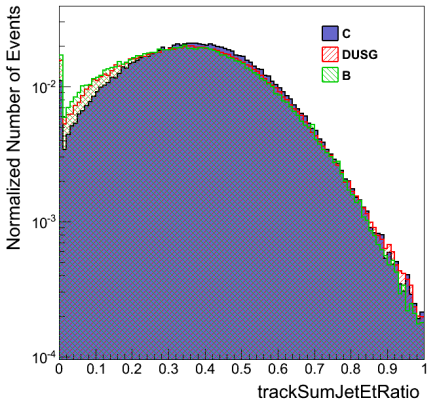

Normalized Number of Events
C
DUSG
B
trackSip2dSigAboveCharm_0

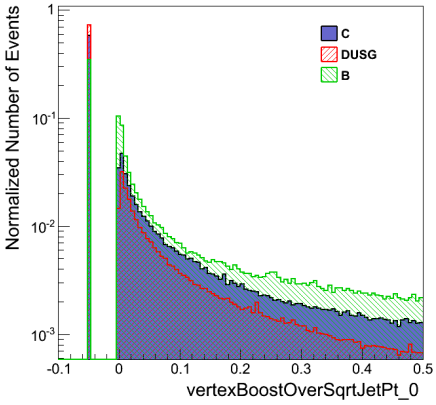

Normalized Number of Events
C
DUSG
B
trackSip2dValAboveCharm_0 (cm)

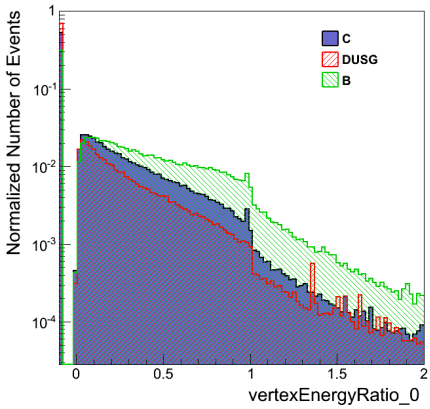

Normalized Number of Events
C
DUSG
B
trackSip3dSigAboveCharm_0

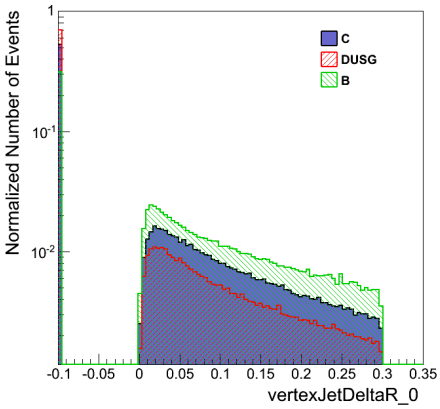

Normalized Number of Events
C
DUSG
B
trackSip3dValAboveCharm_0 (cm)

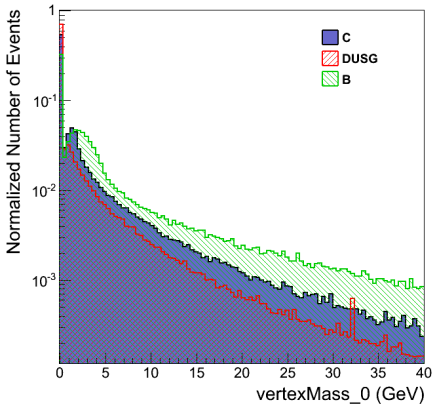

Normalized Number of Events
C
DUSG
B
trackSumJetDeltaR

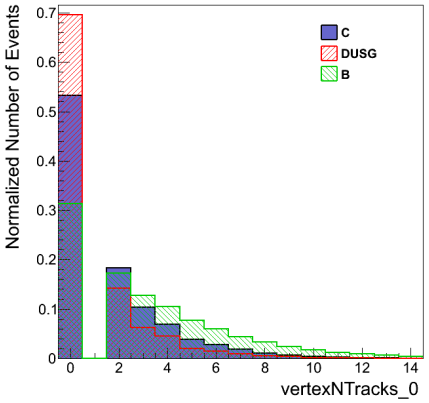

Normalized Number of Events
C
DUSG
B
trackSumJetEtRatio

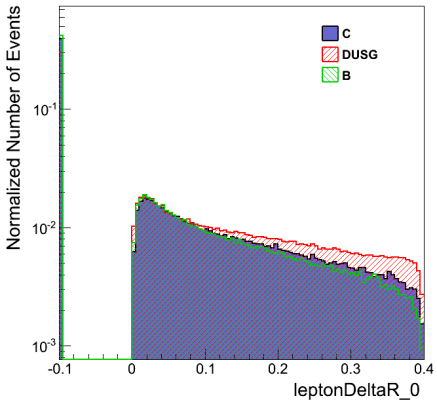

Normalized Number of Events
C
DUSG
B
vertexBoostOverSqrtJetPt_0

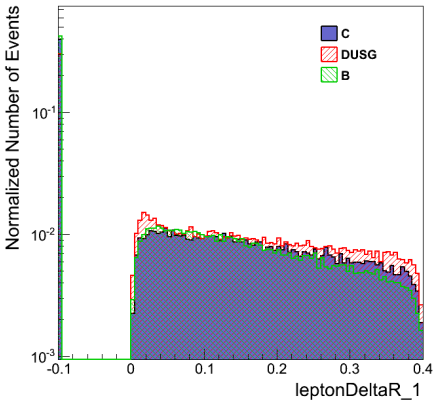

Normalized Number of Events
C
DUSG
B
vertexEnergyRatio_0

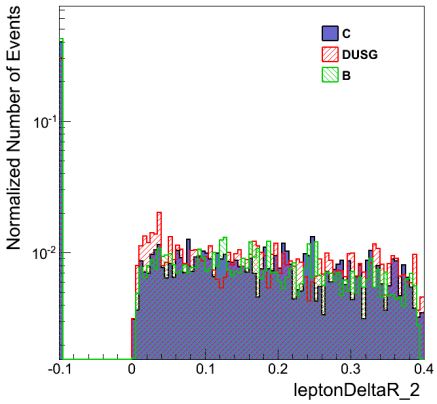

Normalized Number of Events
C
DUSG
B
vertexJetDeltaR_0

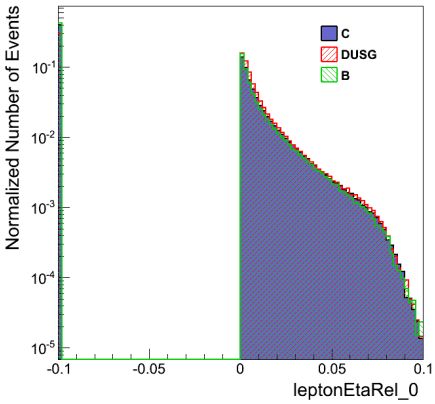

Normalized Number of Events
C
DUSG
B
vertexMass_0 (GeV)

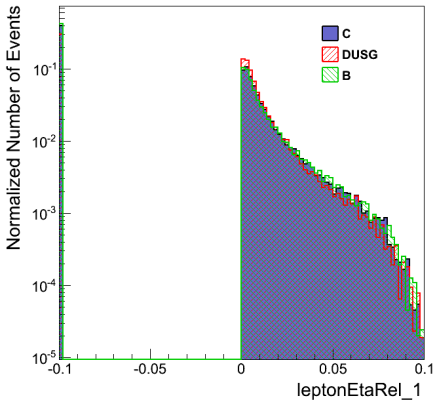

Normalized Number of Events
C
DUSG
B
vertexNTracks_0

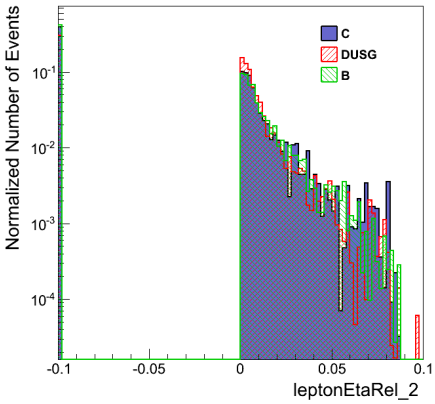

Normalized Number of Events
C
DUSG
B
leptonDeltaR_0

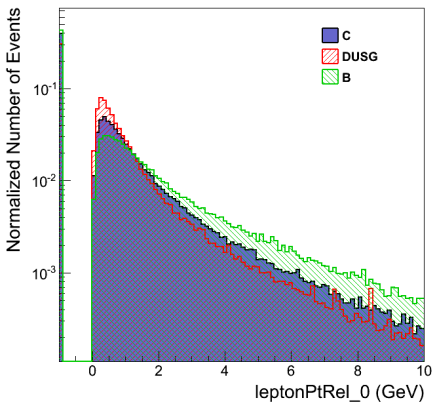

Normalized Number of Events
C
DUSG
B
leptonDeltaR_1

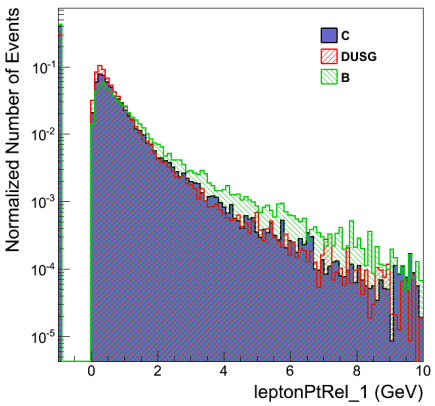

C
DUSG
B
Normalized Number of Events
leptonDeltaR_2

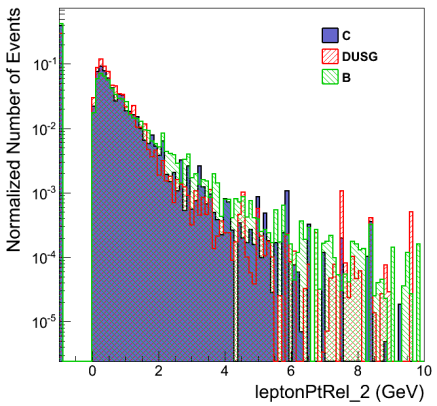

C
DUSG
B
Normalized Number of Events
leptonEtaRel_0

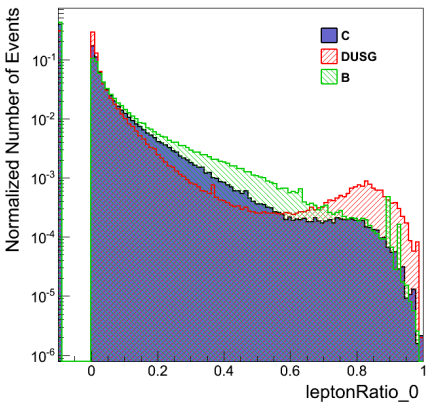

C
DUSG
B
Normalized Number of Events
leptonEtaRel_1

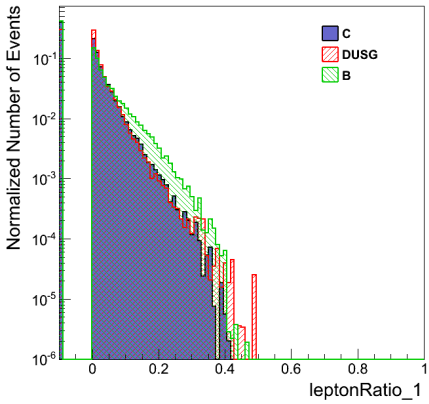

C
DUSG
B
Normalized Number of Events
leptonEtaRel_2

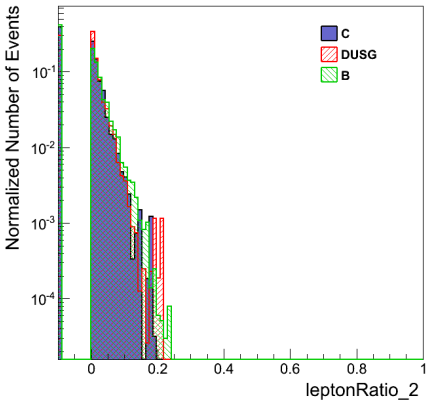

C
DUSG
B
Normalized Number of Events
leptonPtRel_0 (GeV)

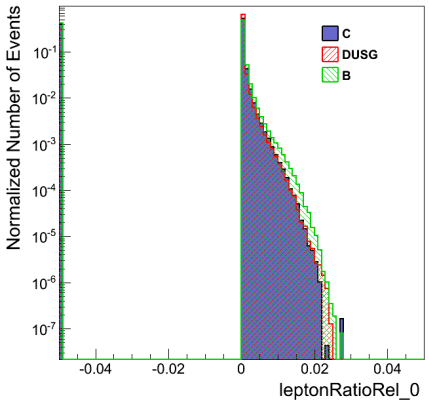

C
DUSG
B
Normalized Number of Events
leptonPtRel_1 (GeV)

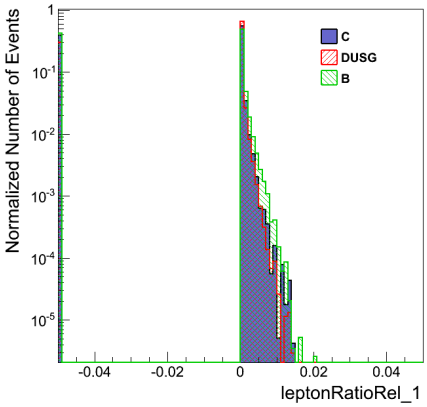

C
DUSG
B
Normalized Number of Events
leptonPtRel_2 (GeV)

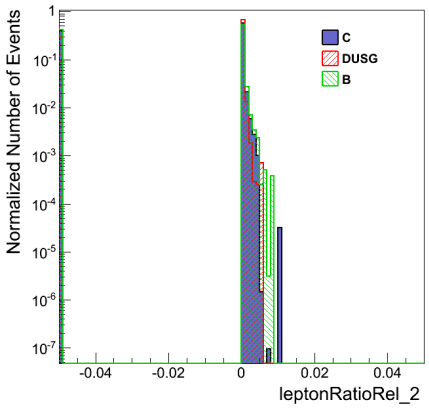

C
DUSG
B
Normalized Number of Events
leptonRatio_0

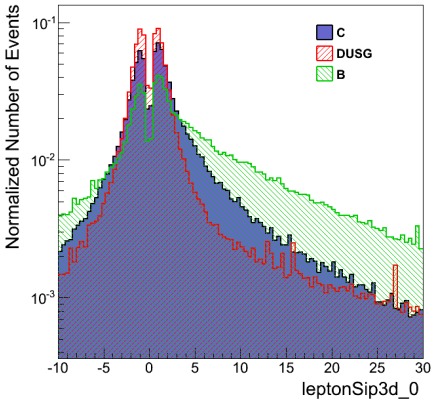

C
DUSG
B
Normalized Number of Events
leptonRatio_1

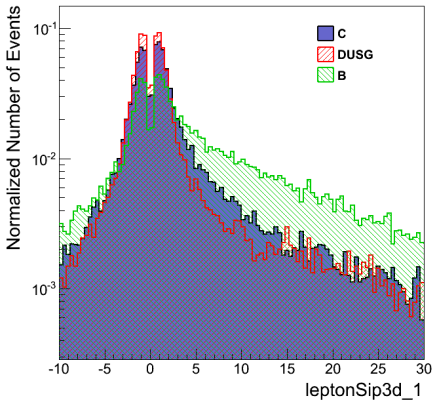

C
DUSG
B
Normalized Number of Events
leptonRatio_2

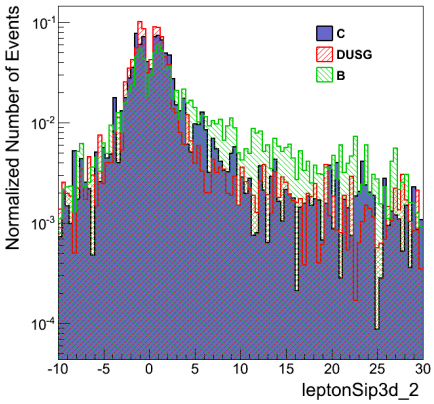

C
DUSG
B
Normalized Number of Events
leptonRatioRel_0


C
DUSG
B
Normalized Number of Events
leptonRatioRel_1


C
DUSG
B
Normalized Number of Events
leptonRatioRel_2

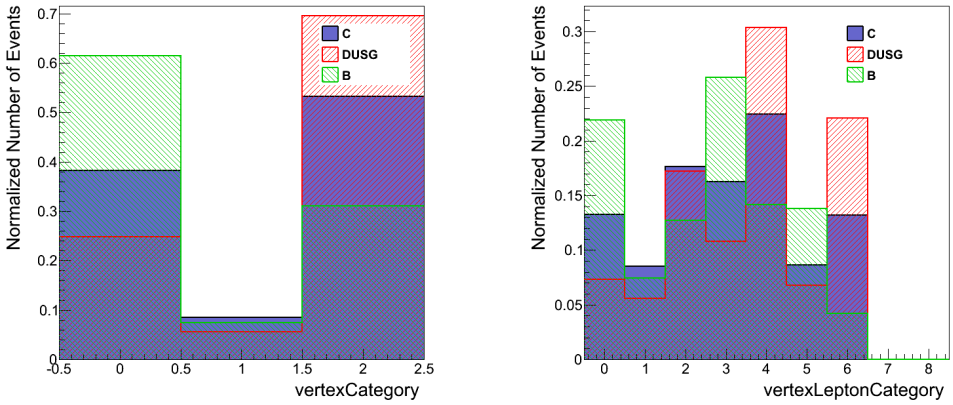

C
DUSG
B
Normalized Number of Events
leptonSip3d_0


C
DUSG
B
Normalized Number of Events
leptonSip3d_1


C
DUSG
B
Normalized Number of Events
leptonSip3d_2

Vertex and SL category distributions

The secondary vertex and soft-lepton categories are assigned a number denoted by the vertexCategory (3 vertex categories) or vertexLeptonCategory (7 vertex-SL categories) parameters. These numbers are explained in Table A.3 and Table A.4 for the vertexCategory variable and vertexLeptonCategory variables respectively. The distributions for these variables are shown below.



SV category	RecoVertex	PseudoVertex	NoVertex
Value	0	1	2

Table A.3: vertexCategory number codes.

	RecoVertex	**PseudoVertex**	**NoVertex**
NoSoftLepton	0	1	2
SoftMuon	3	n.a.	4
SoftElectron	5	n.a.	6

Table A.4: vertexLeptonCategory number codes.

Appendix B

Variable ranking tables (top 20 variables)

Rank	Variable	Importance
1	trackSip2dSig_0	4.626e-02
2	trackSip3dSigAboveCharm_0	4.593e-02
3	trackSip3dSig_0	4.586e-02
4	trackSip2dSigAboveCharm_0	3.826e-02
5	trackSip2dSig_1	3.368e-02
6	trackEtaRel_1	3.157e-02
7	vertexNTracks_0	3.101e-02
8	vertexBoostOverSqrtJetPt_0	3.045e-02
9	vertexEnergyRatio_0	2.956e-02
10	trackSip3dSig_1	2.881e-02
11	flightDistance2dSig_0	2.761e-02
12	leptonRatio_0	2.637e-02
13	jetNSecondaryVertices	2.328e-02
14	leptonPtRel_0	2.242e-02
15	trackSip2dValAboveCharm_0	2.211e-02
16	flightDistance2dVal_0	2.134e-02
17	trackSip3dValAboveCharm_0	1.764e-02
18	leptonSip3d_0	1.627e-02
19	leptonEtaRel_0	1.591e-02
20	vertexMass_0	1.540e-02

Table B.1: Method unspecific ranking table of the 20 most sensitive variables discussed in section 3.7.1.

Rank	Variable	Importance
1	leptonRatio_0	1.150e-01
2	trackSip3dSigAboveCharm_0	1.014e-01
3	leptonSip3d_0	9.910e-02
4	trackSip2dSig_0	9.133e-02
5	trackSip2dSigAboveCharm_0	8.133e-02
6	trackSip3dSig_0	6.786e-02
7	trackEtaRel_1	6.692e-02
8	vertexNTracks_0	5.613e-02
9	trackSip3dValAboveCharm_0	5.151e-02
10	flightDistance2dSig_0	5.098e-02
11	leptonEtaRel_0	4.413e-02
12	vertexEnergyRatio_0	3.694e-02
13	trackSip3dSig_1	2.936e-02
14	vertexBoostOverSqrtJetPt_0	2.792e-02
15	trackSip2dSig_1	2.762e-02
16	flightDistance2dVal_0	1.779e-02
17	leptonPtRel_0	1.649e-02
18	trackSip2dValAboveCharm_0	1.207e-02
19	jetNSecondaryVertices	6.145e-03
20	vertexMass_0	0.000e+00

Table B.2: Method specific ranking table of the 20 most sensitive variables discussed in section 3.7.1.

Appendix C

Standalone TMVA setup: workflow

This section briefly discusses the workflow that is used to produce the event samples, apply the weights, perform the training and validation and produce the performance curves. All of the scripts mentioned can be found in the following GitHub repository:
https://github.com/vlambert/TMVA_CTagging

The workflow is illustrated in Figure C.1. First a variable extraction script makes ROOT trees from the original AOD[1] samples that contain all the jet properties. These ROOT trees still contain vectors of variables (for examples for track-related variables the values of all tracks in that jet are put into a vector), which is not supported by the TMVA framework. To create so-called "flat" trees (each variable represents exactly one value and not a vector of values), a script (createNewTrees(SL).py) extracts the information from these vectors and puts them in separate variables keeping usually only the first three vector elements and giving them a suffix _0, _1 and _2 for the first, second and third vector element respectively. Then the weights discussed in the previous section are applied. The createEtaPtWeightHists.py script makes 2-dimensional histograms for p_T and η which contain the p_T-η weights. The Normalization_Weights.C and biasTTbar.C scripts create text files containing the respective weights. The addWeightBranch.py script extracts the weights from the p_T-η histograms and the text files and adds them as a branch to the flat trees. At this point the flat trees are ready for the c-tagger training and validation, which are defined in the tmva_training.py script. Here the MVA settings, variables and samples are defined. The output-files of the TMVA training and validation are then used in the makePlots.py script to produce the final performance curves.

[1] AOD stands for Analysis Object Data and is a common format of CMS event samples for physics analyses.

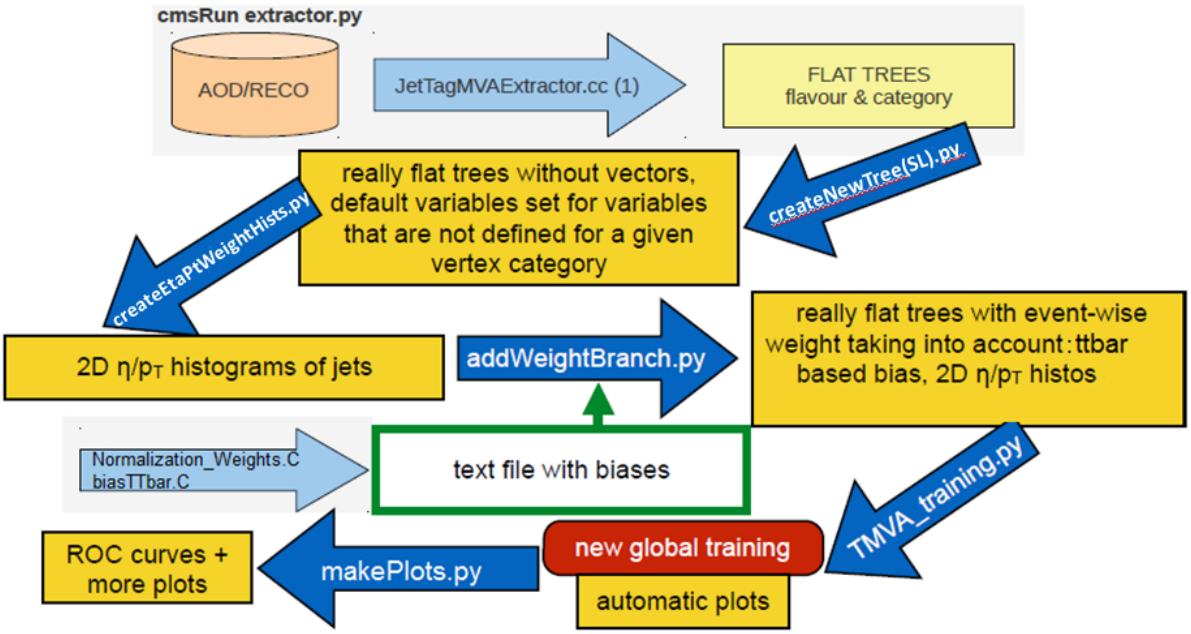


Figure C.1: Schematic overview of the workflow of the standalone TMVA c-tagger setup.

Bibliography

[1] S. Carrol, *Spacetime and Geometry.* Addison Wesley, 2004. Chapter 9.

[2] S. Glashow and M. Gell-Mann, "Gauge Theories Of Vector Particles," *Annals Phys. 15 (1961) 437-460.*

[3] S. Weinberg, "Charge symmetry of weak interactions," *Phys.Rev. 112 (1958) 1375-137.*

[4] A. Salam and J. Ward, "Weak and electromagnetic interactions," *Nuovo Cim. 11 (1959) 568-577.*

[5] Gargamelle Neutrino Collaboration, "Observation of Neutrino Like Interactions without Muon or Electron in the Gargamelle Neutrino Experiment," *Nucl.Phys. B73 (1974) 1-22.*

[6] UA1 Collaboration, "Experimental Observation of Lepton Pairs of Invariant Mass Around 95 GeV/c^2 at the CERN SPS Collider," *Phys.Lett. B126 (1983) 398-410.*

[7] D. Perkins, *Particle Astrophysics.* Oxford University Press, 2009. Chapter 1.3.

[8] F. Englert and R. Brout, "Broken Symmetry and the Mass of Gauge Vector Mesons," *Phys.Rev.Lett. 13 (1964) 321-323.*

[9] P. Higgs, "Broken Symmetries and the Masses of Gauge Bosons," *Phys.Rev.Lett. 13 (1964) 508-509.*

[10] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," *Phys.Lett. B716 (2012) 30-61.*

[11] 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. B716 (2012) 1-29.*

[12] G. Hinshaw *et al.*, "Nine-year Wilkinsons Microwave Anisotropy Probe (WMAP) Observations: Cosmological Parameter Results," *Astrophys.J.Suppl. 208 (2013) 19.*

[13] D. Hooper, "TASI 2008 Lectures on Dark Matter," *Lectures given at a conference (2-27 Jun 2008. Boulder, Colorado); CNUM: C08-06-02.2, arXiv:0901.4090v1 [hep-ph].*

[14] S. Hess, F. Kitaura, and S. Gottloeber, "Simulating Structure Formation of the Local Universe," *Mon.Not.Roy.Astron.Soc. 435 (2013) 2065.*

[15] A. Borriello and P. Salucci, "The Dark Matter Distribution in Disk Galaxies," *Mon.Not.Roy.Astron.Soc. 323 (2001) 285.*

[16] R. Massey and T. Kitching, "The dark matter of gravitational lensing," *Rep. Prog. Phys. 73 (2010) 086901.*

[17] P. Graham *et al.*, "Towards a Bullet-proof test for indirect signals of dark matter," *arXiv:1502.03824 [hep-ph].*

[18] C. Grupen, *Astroparticle Physics.* Springer, 2005. Chapter 13.

[19] J. Kim and G. Carosi, "Axions and the Strong CP Problem," *Rev.Mod.Phys.82:557-602,2010.*

[20] IceCube Collaboration, "Search for dark matter annihilations in the Sun with the 79-string IceCube detector," *Phys.Rev.Lett. 110 (2013) 13, 131302.*

[21] R. Bernabei *et al.*, "Final model independent result of DAMA/LIBRA–phase1," *Eur.Phys.J. C73 (2013) 12, 2648.*

[22] XENON100 Collaboration, "Dark Matter Results from 225 Live Days of XENON100 Data," *Phys.Rev.Lett. 109 (2012) 181301.*

[23] LUX Collaboration, "First results from the LUX dark matter experiment at the Sanford Underground Research Facility," *Phys.Rev.Lett. 112 (2014) 9, 091303.*

[24] CMS Collaboration, "Search for the Production of Dark Matter in Association with Top Quark Pairs in the Di-lepton Final State in pp collisions at $\sqrt{s} = 8$ TeV," *CDS: CMS-PAS-B2G-13-004.*

[25] ATLAS Collaboration, "Search for dark matter in events with heavy quarks and missing transverse momentum in pp collisions with the ATLAS detector," *Submitted to Eur. Phys. J. C, arXiv:1410.4031v1 [hep-ph].*

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

[27] J. Erdmann, "Measurement of the inclusive $t\bar{t}\gamma$ cross section at $\sqrt{s} = 7$ TeV with the ATLAS detector," *CDS:CERN-THESIS-2012-076, arXiv:1206.5696 [hep-ex].*

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

[29] CMS Collaboration, "The CMS Technical Design Report Volume 1: Detector Performance and Software," *CDS: CERN-LHCC-2006-001.*

[30] Y. Onel, "Present status of CMS HF quartz fiber calorimetry," *Conf.Proc. C020325 (2002) 504-514.*

[31] S. Lowette, "Experimental Techniques in Particle Physics: Triggering and DAQ," *Lectures at Vrije Universiteit Brussel (2014).*

[32] E. Conte, B. Fuks, and G. Serret, "MadAnalysis 5, A User-Friendly Framework for Collider Phenomenology," *Comput.Phys.Commun. 184 (2013) 222-256.*

[33] K. Chen *et al.*, "Search for $B \to h^{(*)}\nu\bar{\nu}$ Decays at Belle," *Phys.Rev.Lett. 99 (2007) 221802.*

[34] BaBar Collaboration, "Search for $B \to K^*\nu\bar{\nu}$ decays," *Phys.Rev. D78 (2008) 072007.*

[35] A. Artamonov *et al.*, "New measurement of the $K^+ \to \pi^+\nu\bar{\nu}$ branching ratio," *Phys.Rev.Lett. 101 (2008) 191802.*

[36] U. Sarkar, *Particle and Astroparticle Physics.* Taylor and Francis, 2005. p.143.

[37] A. Alloul *et al.*, "FeynRules 2.0 - A complete toolbox for tree-level phenomenology," *Comput.Phys.Commun. 185 (2014) 2250-2300.*

[38] J. Alwall *et al.*, "MadGraph 5 : Going Beyond," *JHEP 1106 (2011) 128.*

[39] Planck Collaboration, "Planck 2013 results. XVI. Cosmological parameters," *Astron.Astrophys. 571 (2014) A16.*

[40] P. Gondolo *et al.*, "DarkSUSY: Computing supersymmetric dark matter properties numerically," *Chapter 4, JCAP 0407 (2004) 008.*

[41] G. Steigman, B. Dasgupta, and J. Beacom, "Precise Relic WIMP Abundance and its Impact on Searches for Dark Matter Annihilation," *Phys.Rev. D86 (2012) 023506.*

[42] M. Backovic, K. Kong, and M. McCaskey, "MadDM v.1.0: Computation of Dark Matter Relic Abundance Using MadGraph5," *Physics of the Dark Universe 5-6 (2014) 18-28.*

[43] Fermi-LAT Collaboration, "Indirect Searches for Dark Matter with the Fermi Large Area Telescope 1," *Phys.Procedia 61 (2015) 6-12.*

[44] A. Rajaraman, J. Smolinsky, and P. Tanedo, "On-Shell Mediators and Top–Charm Dark Matter Models for the Fermi-LAT Galactic Center Excess," *arXiv:1503.05919.*

[45] Particle Data Group Collaboration, "Review of Particle Physics," *Chin.Phys. C38 (2014) 090001.*

[46] CMS Collaboration, "Measurement of the Top Quark Mass With 2012 CMS Data," *2nd Conference on Large Hadron Collider Physics Conference (LHCP 2014) New York City, New York, USA); CNUM: C14-06-02.2, arXiv:1409.0288 [hep-ex].*

[47] CMS Collaboration, "Search for monotop signatures in proton-proton collisions at $\sqrt{s}$ = 8 TeV," *Phys.Rev.Lett. 114 (2015) 10, 101801.*

[48] ATLAS Collaboration, "Search for invisible particles produced in association with single-top-quarks in proton-proton collisions at $\sqrt{s}$ = 8 TeV with the ATLAS detector," *Eur.Phys.J. C75 (2015) 79.*

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

[50] S. Baffioni *et al.*, "Electron Reconstruction in CMS," *Eur.Phys.J. C49 (2007) 1099-1116.*

[51] CMS Collaboration, "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s}$ = 7 TeV," *JINST 7 (2012) P10002.*

[52] C. Bernet, "Particle Flow in CMS: The Data Challenge," *CERN Detector Seminar, 23th september 2011.*

[53] CMS Collaboration, "Performance of Jet Algorithms in CMS," *CMS PAS JME-07-003 (2008).*

[54] M. Cacciari, G. Salam, and G. Soyez, "The Anti-k_T jet clustering algorithm," *JHEP 0804 (2008) 063.*

[55] J. de Favereau *et al.*, "DELPHES 3 A modular framework for fast simulation of a generic collider experiment," *JHEP 1402 (2014) 057.*

[56] L. Wehrli, "Measurement of $b\bar{b}$ Angular Correlations based on Secondary Vertex Reconstruction in Proton-Proton Collisions at $\sqrt{s} = 7$ TeV," *CDS:CERN-THESIS-2012-010*.

[57] W. Waltenberger, "Adaptive Vertex Reconstruction," *Technical Report: CMS NOTE-2008/033*.

[58] T. Speer and K. Prokofiev, "Vertex Fitting in the CMS Tracker," *Technical Report CMS NOTE 2006/032*.

[59] P. V. Mulders *et al.*, "Implementation and training of the Combined Secondary Vertex MVA b-tagging algorithm in CMSSW," *CMS Analysis Note: CMS AN AN-12-441*.

[60] A. Hoecker *et al.*, "TMVA - Toolkit for Multivariate Data Analysis," *PoS ACAT (2007) 040*.

[61] ATLAS Collaboration, "The Evolution of Boosting Algorithms - From Machine Learning to Statistical Modelling," *Methods Inf Med 2014; 53(6): 419-427*.

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

[63] CLEO Collaboration, "Measurement of BR($D_s^+ \to \ell + \nu$) and the Decay Constant $f_{D_s^+}$ From 600 pb^{-1} of e^+e^- Annihilation Data Near 4170MeV," *Phys.Rev. D78 (2008) 052003*.

[64] CLEO Collaboration, "Precision Measurement of B($D^+ \to \mu^+\nu$) and the Pseudoscalar Decay Constant f_{D^+}," *Phys.Rev. D78 (2008) 052003*.

[65] ATLAS Collaboration, "Performance and Calibration of the JetFitterCharm Algorithm for c-Jet Identification," *ATLAS Note: ATL-PHYS-PUB-2015-001*.

[66] T. Sjostrand, S. Mrenna, and P. Skands, "PYTHIA 6.4 Physics and Manual," *JHEP 0605 (2006) 026*.

[67] E. Boos and L. Dudko, "The Single Top Quark Physics," *Int.J.Mod.Phys. A27 (2012) 1230026*.

[68] CDF Collaboration, "Search for Standard Model Higgs Boson Production in Association with a W Boson Using a Matrix Element Technique at CDF in $p\bar{p}$ Collisions at $\sqrt{s} = 1.96$ TeV," *Phys.Rev. D85 (2012) 072001*.

[69] CMS Collaboration, "Measurement of the $t\bar{t}$ production cross section in the dilepton channel in pp collisions at $\sqrt{s} = 8$ TeV ," *JHEP 1402 (2014) 024*.

[70] CMS Collaboration, "8 TeV Jet Energy Corrections and Uncertainties based on 19.8 fb^{-1} of data in CMS," *CMS Performance Note: CMS DP 2013/033*.