Search for a heavy Higgs boson in the channel with four b quarks with the CMS experiment at the LHC

Vorgelegt von
Jonas Rübenach

Bachelor-Arbeit im Studiengang Physik
Universität Hamburg
2017

Gutachter:
Prof. Dr. Elisabetta Gallo
Dr. Roberval Walsh

Abstract

A search for a heavy neutral Higgs boson as postulated by the Minimal Supersymmetric Standard Model is performed. The data used in the search was taken by the CMS experiment at the LHC in 2016 at a center-of-mass energy of 13 TeV and amounts to an integrated luminosity of $36.3\,\mathrm{fb}^{-1}$. The search focuses on the bottom quark pair decay channel of the heavy Higgs boson, associated with an additional pair of bottom quarks from its production. In order to gain a higher sensitivity on the Higgs boson, the measured events are split into two categories, one for at least measured 4 b jets in the event and one for at least 3 b jets.

Zusammenfassung

Es wird eine Suche nach einem schweren, neutralen Higgs-Boson, wie es im minimalem supersymmetrischem Standardmodel vorkommt, durchgeführt. Die hierfür verwendeten Messdaten stammen vom CMS-Experiment am LHC aus dem Jahr 2016, wurden bei einer Schwerpunktsenergie von 13 TeV aufgenommen und entsprechen einer integrierten Luminosität von $36.3\,\mathrm{fb}^{-1}$. Die Suche richtet sich auf den Zerfallskanal in ein Bottomquarkpaar von schweren Higgs-Bosonen, die zusammen mit einem weiteren Bottomquarkpaar entstanden sind. Um höhere Sensitivitäten zu erreichen, werden die gemessenen Ereignisse in zwei Kategorien aufgeteilt, eine für Ereignisse mit mindestens 4 b Jets und eine für Ereignisse mit mindestens 3 b Jets.

Contents

Chapter 1

Introduction

One of the main goals of physics is to create a single theory from which every observable in physics can be derived and explained. Quantum field theory has been very successful in this subject, bringing forth various theories which later built up the standard model (SM) of particle physics. With the standard model the most fundamental concepts such as matter or 3 of the fundamental forces in our universe are explained in one theory. After the prediction of various particles, which have later been observed in the second half of the 20th century and have become part of the standard model, the discovery of the Higgs boson in 2012 at the ATLAS and CMS experiments located at the LHC posed the last building block for the model. The Higgs boson together with its field, the Higgs field, represents a central part of the SM as it explains the mass of the other particles. While the model already explains a wide range of physical concepts, there are still many problems left to be solved until reaching a single theory of everything. These problems include the one force not explained by the standard model, namely gravity, the existence of dark matter or the hierarchy problem. In order to advance on solving these problems, various theories have been made including supersymmetric models (SUSY), expanding the standard model and adding new particles such as additional Higgs bosons.

The LHC, which is part of the CERN accelerator complex, is the world's largest and highest energy particle accelerator. Using its energy scales the discovery of one of the heaviest particle in the standard model, the Higgs boson, became possible. Currently it runs at even higher energies than it was during the discovery, which enables the hunt for even heavier particles such as the additional Higgs bosons postulated by supersymmetry. Observing such a supersymmetric particle would advance particle physics by a huge step as it would hint at what lies beyond the standard model.

In this thesis an analysis searching for neutral Higgs bosons is performed. This is done by studying the decay into two bottom quarks of supersymmetric Higgs bosons associated with at least one additional bottom quarks from the production. From the data, which was taken by the CMS experiment at the LHC in 2016, events are selected fulfilling various criteria for this particular channel as described in chapter 4. Furthermore the events are split up into categories, one only requiring 3 measured b-jet candidates, and one with at least 4 b jets, which are then

analyzed individually. If the 4 b-tagged jet category on its own turns out to be sensitive, a future combination of the two categories can result in higher sensitivities than an analysis without categorization.

Similar analyses on the Run 1 data at c.m. energies of 7 TeV and 8 TeV data is published by the CMS Collaboration in [1, 2].

Chapter 2

The theory of particle physics

The following sections discuss the Standard Model and supersymmetric variants. For further discussion and sources on these topics see [3].

Masses are given in GeV by using natural units. These units are obtained by defining the speed of light c, the reduced Planck constant $\hbar$ and the vacuum permittivity ε_0 to be equal to one. Thus 1 GeV of mass approximately corresponds to 1.78×10^{-27} kg.

2.1 Standard Model

Currently the most fundamental phenomena in physics are observed in particle physics. Here the Standard Model is the most central theory behind particle physics and has been very efficient at predicting observations made in experiments. The model describes 17 elementary particles that have already been observed. They differ in properties such as mass, charge or spin. Some of them are responsible for the electro-magnetic, the strong or the weak force, three of the fundamental forces in physics, while others make up matter such as atoms. To classify these differences, they are categorized into 5 bosons and 12 fermions. Elementary fermion particles make up matter, always have a spin of $\frac{1}{2}$ and are further divided into 6 quarks and 6 leptons. Quarks and leptons are in turn divided into 3 generations each containing 2 particles. On top of that it is important to note that for every quark and lepton, there is an anti-particle "partner" with the same properties and an opposite charge. These symmetries in categorization are one reason why the Standard Model is so favorable.

Besides the elementary fermions, there are also 5 bosons. Four of these are gauge bosons, which correspond to three fundamental forces and the remaining one is the Higgs boson. The Higgs boson has a mass equal to 125 GeV, its charge and spin are equal to zero and it is the result of introducing the Higgs field which is used to explain why most of the other elementary particles aren't massless. The Higgs field has a non-zero vacuum expectation value v. This is different from all the other fields in the Standard Model.

2.2 Supersymmetry and heavy Higgs boson

Although the Standard Model has been very successful in explaining particle physics, there are still numerous open questions. These include the unification of all forces or the existence of dark matter. Another example are the corrections from other particles to the mass of the Higgs boson, which become very large compared to the Higgs mass when assuming the Standard Model holds true up to the scale of the grand unification theory. This question is known as the hierarchy problem and is being answered by supersymmetry theories by introducing new particles, of which none have been observed yet. Similar to the symmetries within the Standard Model, every elementary particle from the Standard Model gets a new supersymmetry "partner", whose spin differs by $\frac{1}{2}$. The corrections to the Higgs mass are being kept small because now for every correction there is another from a supersymmetry particle with opposite sign. As the corrections do not cancel out completely and as supersymmetry particles have not been observed yet, there must be differences between the masses of Standard Model and supersymmetry particles. This is one reason why the latter are postulated to be much heavier.

One possible model for supersymmetry is the Minimal Supersymmetric Standard Model. It introduces 5 Higgs bosons in total, of which 2 are charged, denoted as H^+ and H^-. The lightest of the three neutral ones, h^0, is assumed to be the one already known from the Standard Model. This assumption leads to the other two neutral ones, H^0 and A^0, being heavy Higgs bosons with masses beyond 125 GeV [4]. In addition all of these 5 Higgs bosons have supersymmetric partners called Higgsinos. Similar to the Standard Model Higgs boson before its observation, the masses of the new ones are unknown and thus, because there was no other Higgs boson observed below the mass of the SM Higgs boson, must be heavier than the SM Higgs boson. In addition there are two new vacuum expectation values which are again non-zero. These are free parameters and are denoted as v_u and v_d. It is further defined that

$$\tan\beta := \frac{v_u}{v_d}.$$

	Leptons		Quarks	
Generation	Particle	M in GeV	Particle	M in GeV
First	electron	5×10^{-4}	down	3×10^{-3}
	neutrino	$< 10^{-9}$	up	5×10^{-3}
Second	muon	0.106	strange	0.1
	neutrino	$< 10^{-9}$	charm	1.3
Third	tau	1.78	bottom	4.5
	neutrino	$< 10^{-9}$	top	174

Table 2.1: Elementary Fermions from the SM with their respective masses. From [3].

Gauge bosons		Higgs bosons	
Particle	M in GeV	Particle	M in GeV
photon	0	h^0	125
gluon	0	H^0	unknown
Z	91.2	A^0	unknown
$W^{\pm}$	80.4	$H^{\pm}$	unknown

Table 2.2: The 4 Higgs bosons introduced by the MSSM together with all the elementary bosoms from the SM and their respective masses [3]. The masses of the MSSM Higgs bosons are unknown.

Chapter 3

The experiment

3.1 Outline of the task

Using a high energy particle collider like LHC and a high precision detector like CMS, it is possible to record the creation of high-mass particles by analyzing its decay. In this thesis the decay of the neutral MSSM Higgs boson into a bottom quark pair is analyzed. The bottom-quark decay mode is the overall dominant channel for the neutral MSSM Higgs boson [5] and thus is a promising channel for this task. In particular the analysis is restricted to events where the other 2 bottom quarks from the production are observed. The corresponding Feynman-Diagram can be found in figure 3.1. Currently the analysis is being performed by selecting events with at least 3 b-tagged jets (for details on b-tagged jets see section 3.3). In this analysis two categories are made: one for 3 b-tagged jets and one for 4 b-tagged jets. Events from both categories will then be evaluated separately. This is done in order to compare their relative sensitivities and infer whether a 4 b-tagged jet category is worth further analyses.

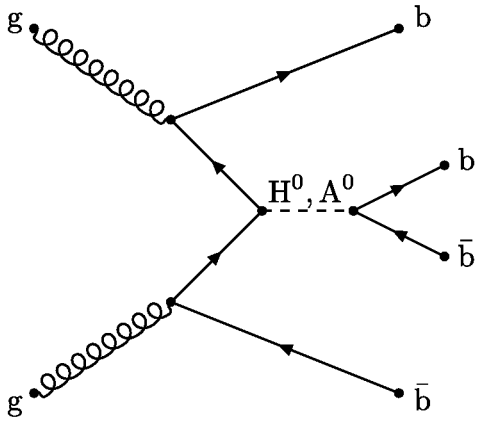


Figure 3.1: Feynman-Diagram showing the production and decay of a heavy Higgs boson.

In case of b-tagging, several algorithms are available [6, 7]. This thesis uses DeepCSV if not specified otherwise and also includes a comparison between the results from the algorithms CSVv2 and DeepCSV (see chapter 6).

3.2 LHC and CMS

The Large Hadron Collider is the largest and most powerful circular particle accelerator of the world [8]. Located at the facilities of the European Organization for Nuclear Research (CERN) in Switzerland on the border to France it is built at a depth ranging from 175 m to 50 m and has a circumference of 27 km [9]. It was built in the tunnel of a previously used accelerator of the CERN, LEP, and started its operation in September 2008. The Compact Muon Solenoid (CMS) is one out of several experiments at the LHC.

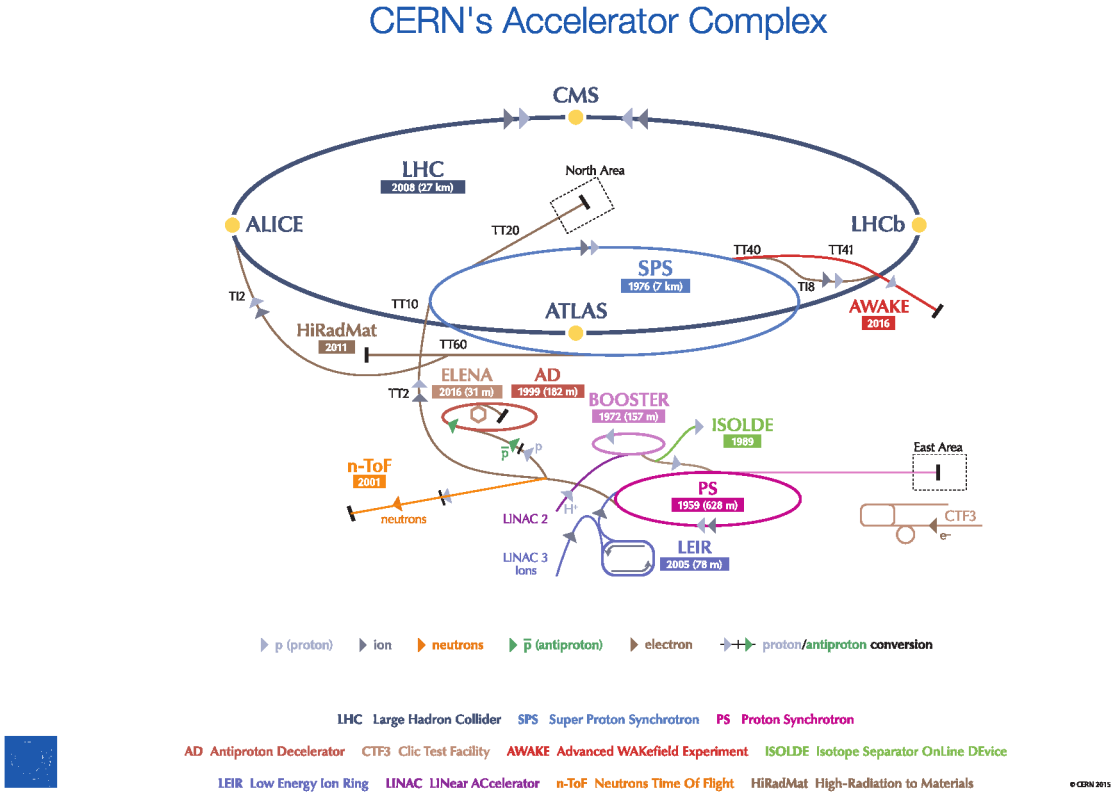


Figure 3.2: CERN's accelerator complex. The LHC is clearly visible as the largest accelerator ring. From [10].

3.2.1 Structure of the LHC

The LHC is part of an accelerator complex (as seen in Figure 3.2) and a particle will go through different stages before it reaches the LHC. In each stage, the particle will be further accelerated until its energy finally reaches an appropriate energy for the LHC. For example to accelerate a proton, it is first brought to an energy of 50 MeV by the Linear accelerator 2 (LINAC2), upon which it is accelerated to 1.4 GeV by the PS Booster. After that it reaches an energy of 450 GeV by going through the Proton Synchrotron (PS) and the Super Proton Synchrotron (SPS) whereupon it finally arrives in the LHC. The design energy of the LHC is 7 TeV per beam and it is currently running at 6.5 TeV. Two beams colliding at the latter energy result

in a collision-energy or center-of-mass energy of 13 TeV. During normal operation of proton collisions each beam is made up of up to 2808 bunches of approximately 10^{11} protons each. At the interaction points one bunch has a width of about 20 µm and the space between two bunches can be as low as 7.5 m. This results in an frequency of 40 million bunches per second that pass an interaction point or an average time of 25 ns between each bunch. With this one can define the instantaneous luminosity

$$\mathscr{L} := f\frac{n^2}{4\pi\sigma_x\sigma_y}$$

where f is the collision frequency, n the number of particles per bunch and σ_x and σ_y the root-mean-square of the horizontal and vertical beam size. The instantaneous luminosity of the LHC is around $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$[3]. The number of events of a particular process is given by

$$N = \sigma L = \sigma \int \mathscr{L}\mathrm{d}t$$

where σ is the cross-section, which depends on the process, and L the integrated luminosity. A unit commonly used for cross-section and luminosity is barn (symbol: b) which equals $10^{-28}\,\mathrm{m}^2$.

3.2.2 Structure of the CMS

There are several experiments at the LHC of which the largest are ALICE, ATLAS, CMS and LHCb. All of these experiments are built with different types of measurements in mind. ATLAS and CMS are both general-purpose detectors, with the goal among others to verify the existence of the SM Higgs boson. The name CMS points to the key features of the experiment: Compact, because of its relatively small size of 21 m in length, 15 m in height and width (compared to ATLAS which is 46 m long, 26 m high and wide), Muon because of the importance of measuring muons and Solenoid because of its large solenoid magnet, a superconducting cable in form of a cylindrical coil generating a near constant magnetic field of around 3.8 T. The CMS detector has a total weight of around 12.5 kt. The components of CMS are built around the beam like layers to a cylindrical onion. The central components of a detector are its tracking chamber, the electromagnetic plus the hadron calorimeter, a magnet and the muon detector [11] which can be seen in Figure 3.3.

- The tracking chamber is the part that is the closest to the beam. If a charged particle traverses the chamber, the tracker is able to measure a series of points in space where the particle moved through the chamber. These points are then reconstructed to tracks and can be identified to individual particles. The tracker of CMS totals around 200 m^2 and is made from two different types of silicon trackers. The innermost part are three layers of pixel trackers, that allow high resolution vertexing in all directions. Around these are 10 layers of double- and single-layered strip trackers. Single layers only allow coordinate measurements in two dimensions. The tracker has a radius of about 130 cm.

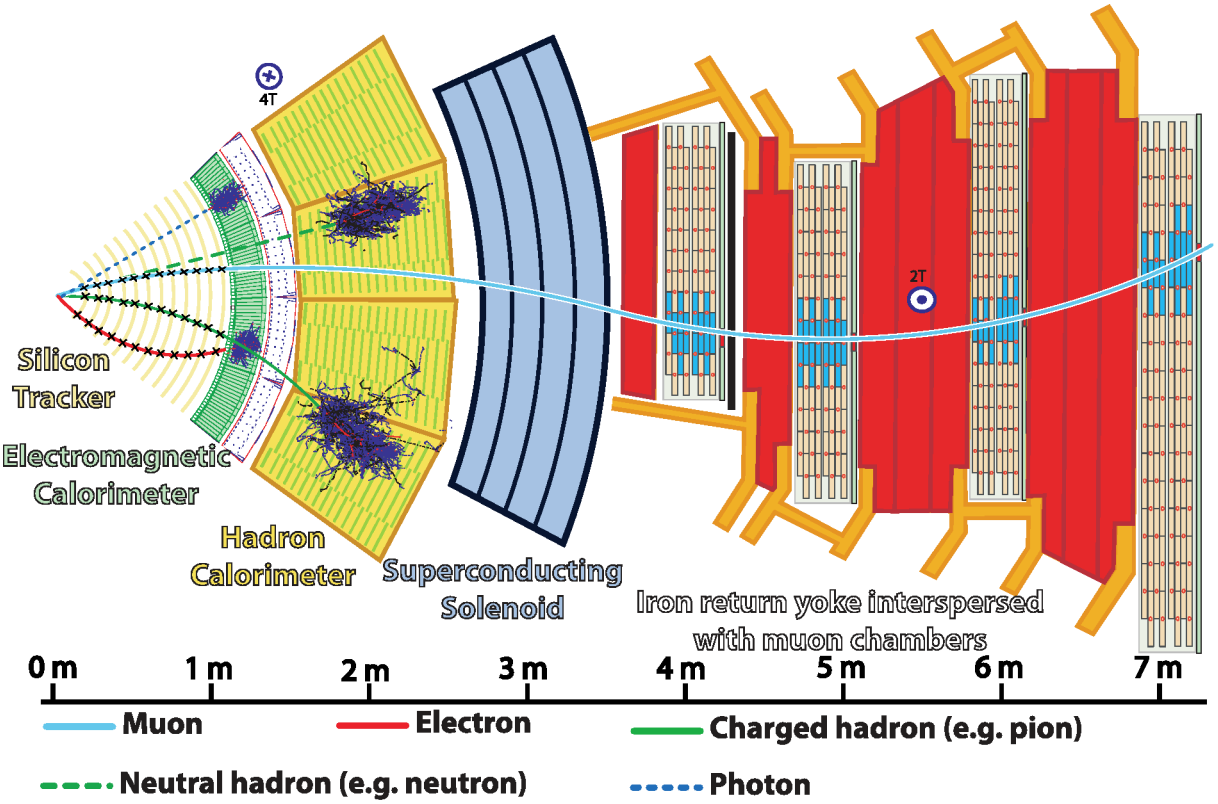


Figure 3.3: Transverse slice through the CMS with various particle detections. After leaving tracks in the tracker, the particles are stopped in the calorimeters. Only the muons and neutrinos traverse further. The trajectories of the charged particles are bent by the magnetic field. From [12].

- The next layer is the electromagnetic calorimeter followed by the hadron calorimeter. In a calorimeter a particle creates a shower, meaning various radiation and decays of a multitude of other particles, or ionizes and thereby stores its energy into the calorimeter. This energy is subsequently measured. A particle that has done this won't traverse the detector any further. In the electromagnetic calorimeter the energy of photons, electrons and positrons are being measured this way. Likewise in the hadron calorimeter the energy of hadrons are measured. Particles that don't interact and do pass the calorimeters are muons and neutrinos.

- Surrounding the CMS hadron calorimeter is the large solenoid magnet. Rather than detecting any particles by itself, its magnetic field is crucial for the momentum measurement of the charged particles. The track of these particles in the magnetic field are bent according to the Lorentz force, which depends on their velocity. By reconstructing the tracks with the tracker, a measurement of the momentum becomes possible. The magnet of CMS creates a magnetic field of around 3.8 T between beam and magnet and around 2 T in the layers around the magnet.

- The outermost and largest layer is formed by the muon detector. Normally the only particles that are able to reach this far are muons (and neutrinos) because all the other free particles have been stopped by the calorimeters. The muon detectors consists of several types of chambers used to track particles that traverse them. The CMS detector has a total of 1400 muon chambers. Their resolution is much lower than the one of the tracker but together with the tracker a high precision track reconstruction and momentum measurement is possible.

Using this setup one can differentiate several types of particles, which are reconstructed as objects.

In a collider experiment cylindrical coordinates are used. The z-axis is defined to be parallel to the beam direction and the coordinate origin is the point of interaction. From this one defines the transverse momentum p_T as the momentum within the transverse plane, meaning the plane orthogonal to the beam

$$p_T := \sqrt{p_x^2 + p_y^2}.$$

A helpful variable, which is invariant under Lorentz boosts along the beam axis, is the pseudorapidity η defined as

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

where θ is the polar angle of the momentum [3]. Moreover with an azimuthal angle (in spherical coordinates) and pseudorapidity difference of two objects $\Delta\phi$, $\Delta\eta$ and one defines

$$\Delta R := \sqrt{(\Delta\phi)^2 + (\Delta\eta)^2}$$

which can be used as a measure of distance between two objects and also is Lorentz invariant

for boosts along the beam axis.

3.3 Jets and b-tagging

In a detector such as CMS, quarks produced from a collision are not observed on their own. When a pair of quarks is produced in a collision and move apart, because of color confinement each quark starts a process called hadronisation [3]. During hadronisation new quarks are created in order to create a new hadron, which is a color neutral system. This process is repeated and results in a multitude of hadrons being measured by the hadron calorimeter of the detector. Normally the hadrons originating from one quark hit the detector in a narrow cone in the direction of the original quark, which then is called a jet. A jet can be reconstructed using various algorithms utilizing information from the detector such as tracks and energy measured by the calorimeter. This way one can identify the production of a quark. In case of this analysis, the anti-k_T algorithm [13] with a distance parameter of $R = 0.4$ is used.

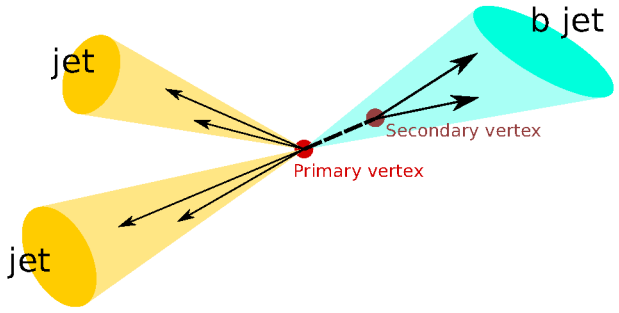


Figure 3.4: Diagram of jet creation from a collision. The tracks of the jet originating from a bottom quark (b jet) come from a different vertex than the other tracks, which is due to the delayed decay of the hadron containing the bottom quark.

In this analysis it is important to know whether a jet originates from a bottom quark or not. To identify such a jet, b-tagging algorithms are used. B-tagging algorithms make use of various properties of bottom quark jets. One is the fact that hadrons from bottom quarks, because of their relatively long life time, travel a longer distance before they continue to decay when compared to other hadrons. This means that the jet won't originate from the point in space of the collision, from which the other objects will originate and which is called primary vertex, but rather from another point slightly displaced from it, called secondary vertex. This can be seen in figure 3.4. As the distance between primary and secondary vertex typically is only around 5 mm, the identification requires track reconstruction of a high resolution tracker such as the one of the CMS. Another feature of bottom-quark hadrons is their relatively high mass, which causes the decay products to have a higher energy and so spread out further before reaching the detector components. Currently a frequently used b-tagging algorithm by CMS is the second version of Combined Secondary Vertex (CSVv2) [6]. There is also a newer algorithm called DeepCSV which uses the same observables as CSVv2, while taking more tracks into account and using a deep neural network [7].

Both algorithms, jet clustering and b tagging, assign a value to the object corresponding to a likelihood that the object is a jet or that it originates from a bottom quark, respectively. In an analysis one normally requires it to be greater (or in some cases less) than a recommended value called working point. For an algorithm commonly there are three such working points: loose, medium and tight, whereas loose is the lowest and tight is the highest. In case of b-tagging these working points correspond, respectively, to 10%, 1% and 0.1% of fake rate, meaning jets which are false-positively tagged as b jets. For the jet identification algorithm a loose working point corresponds to keeping at least 99% of real jets while a tight keeps 98%.

3.4 Triggers and data used

During operation CMS produces approximately $1\,\mathrm{GB\,s^{-1}}$ of data [9]. The amount of data that is actually measured is even larger because triggers are used to select only events of interest for storage and analysis, while other events are discarded. For this purpose there are two types of triggers: Level 1 triggers (L1) and high level triggers (HLT). Level 1 triggers are used right after an event was collected and immediately decide whether will be stored or not based on criteria that can be evaluated in few milliseconds. High level triggers process the events that were accepted by the level 1 triggers and need to reconstruct objects in order to be able to selected the events based on the objects properties. To still achieve a high processing rate of the events the processes utilize parallelization and object reconstruction might not be as accurate as the one used for analysis.

For the purpose of this analysis data taken in 2016 by CMS is used. The data amounts to a total integrated luminosity of $36.26\,\mathrm{fb^{-1}}$ after applying the high level triggers and was measured at a center-of-mass energy of 13 TeV. Data from CMS gets certified based on the quality of the measured data e.g. if all instruments were operating properly. Furthermore to model the signal Monte Carlo simulation (MC) for MSSM Higgs boson events is used. MC events are based on a Geant4 [14] simulation of the CMS detector on events generated by the software framework PYTHIA8 [15]. Geant4 simulates effects imposed by the structure of the CMS detector like electronic noise or geometric limitations.

Chapter 4

Event selection

The events of interest, the signal events, in this case events in which a heavy neutral Higgs boson was created, are selected based on cuts on different observables. This is done for both, the 4 b-tag category and the 3 b-tag category. The cuts have been optimized in order to extract the signal on top of the overwhelming background created from multi-jet events including b jets. While these discard a number of events that actually involve a heavy Higgs boson and are of interest, the cuts were optimized to increase the significance as further shown in section 5.2. The detailed numbers of how many events pass the cuts and their order can be seen in Table 4.1 and 4.2.

Cut	Number of events	Relative efficiency	Absolute efficiency
Trigger	48,628,825	1	1
At least 4 jets	27,871,894	0.57	0.57
Jet-kinematics	12,182,499	0.44	0.25
ΔR	4,775,284	0.39	0.10
$\Delta\eta_{12}$	3,697,716	0.77	0.76
Signal region			
b-tagged (bbbb)	24,528	0.0066	0.00050
Trigger matched	17,865	0.73	0.00037
Control region			
b-tagged (bbbnb)	71,241	0.019	0.0015
Trigger matched	51,694	0.73	0.0011

Table 4.1: Number of events and efficiencies after each cut applied for 4 b-tagged jets. Relative efficiency refers to the ratio between the number of events after the current cut and after the previous cut, whereas absolute efficiency refers to the ratio between current and total numbers. The strongest cut is the four b-tag requirement.

By defining the cuts for the selection a set of all events that match the selection, called signal region, is formed. While the selection has been optimized, the signal region still contains events that are not from the MSSM Higgs boson and thus are not signal. These events make up the background and are modeled by doing another selection that is depleted from signal. This latter selection forms the control region.

Cut	Number of events	Relative efficiency	Absolute efficiency
Trigger	48,628,825	1	1
At least 3 jets	39,703,685	0.82	0.82
Jet-kinematics	21,913,249	0.55	0.45
ΔR	15,167,379	0.69	0.31
$\Delta\eta_{12}$	12,379,423	0.82	0.25
Signal region			
b-tagged (bbb)	368,884	0.030	0.0076
Trigger matched	283,677	0.77	0.0058
Four b exclusion	214,118	0.75	0.0044
Control region			
b-tagged (bbnb)	2,476,127	0.20	0.051
Trigger matched	2,396,515	0.97	0.049

Table 4.2: Number of events and efficiencies after each cut applied for 3 b-tagged jets. Relative efficiency refers to the ratio between the number of events after the current cut and after the previous cut, whereas absolute efficiency refers to the ratio between current and total numbers.

The cut flow is described in detail in the following. After a certified event has passed the trigger it is required to have at least 4 (or 3 depending on the category) loose working-point jets. These jets are ordered by their p_T value, with the highest first. Following this kinematics cuts are applied requiring the first two jets to fulfill $p_T \geq 100\,\text{GeV}$ and $|\eta| \leq 2.2$, the third to fulfill $p_T \geq 40\,\text{GeV}$ and $|\eta| \leq 2.2$ and in the 4 b-tag category for the fourth to fulfill $p_T \geq 30\,\text{GeV}$ and $|\eta| \leq 2.4$. The pseudorapidity requirement helps increasing the efficiency of b-tagging algorithm for the events that pass it. The weaker requirement on the fourth jet ensures to increase the overall efficiency of the selection as the fourth jet has a softer p_T spectrum. A cut requiring $\left|\Delta R_{ij}\right| \geq 1$ between each of the first 4 (or 3) jets ensures that there is enough distance between the jets for the jet reconstruction to have worked properly. The first two jets are required to fulfill $|\Delta\eta_{12}| = |\eta_1 - \eta_2| \leq 1.55$, which suppresses the multi-jet background. The last requirement on the jets is a b-tag value of at least medium working point (for DeepCSV 0.6324) for the first three and loose working point for the fourth. This results in the "bbbb" (or "bbb") signal region. The loose working point on the fourth one again increases efficiency.

For the control region, that is later used for the background model, the minimum b-tag value requirement for the fourth (or third) jet is replaced by a maximum b-tag value requirement, called a non-b-tag. In this case a loose working point is used. This results in a selection depleted from signal events called "bbbnb" (or "bbnb"). Finally it is ensured that the objects, which were used by the trigger, are matched to the objects that were reconstructed for the purpose of the analysis.

Both categories must be mutually exclusive to be able to combine them. This means after selecting events with at least three 3 b-tagged jets, all events that would also match the 4 b-tagged jets selection must be excluded. However exclusion is not needed for the control region as the non-b cut already excludes all events from the 4 b-tagged jet signal and control region.

Chapter 5

Background and signal model

In the following sections it is explained how the models for signal and background are derived. For this purpose the distribution for the di-jet mass distribution is used, which is the distribution of the two leading jet's invariant mass of all events. If there indeed exists an MSSM Higgs boson, there should be an excess of events around its mass in the mass distribution of the measured data signal region. This excess shall be described by the signal model. In case of no excess, the mass distribution's shape is ideally equal to that of the control region.

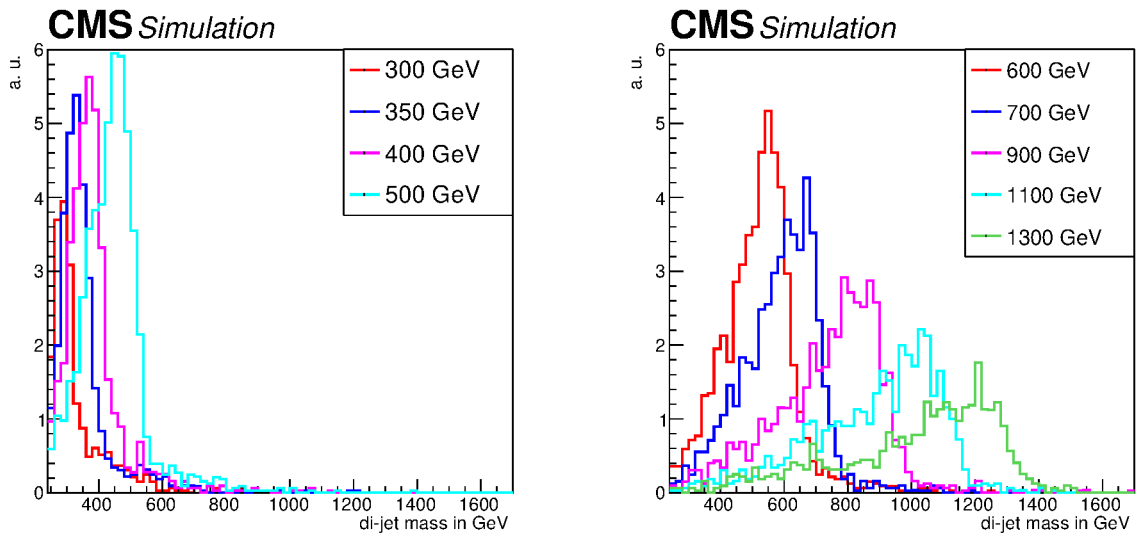
Ensuring that the results won't be biased by the observer the complete analysis is done blinded. This means that the measured data signal region will only serve as source for efficiencies, which are used to scale the background distributions, and no distributions will be used from it.

5.1 Monte Carlo and signal efficiency

To create a model for the signal, a Monte Carlo simulation produced by PYTHIA8 is used. The events are generated for possible Higgs boson masses (mass points) between 300 GeV to 1300 GeV. To account for the properties of the CMS detector various adjustments and trigger simulations are made. Furthermore b-tagging scale factors are applied because b-tagging efficiency from simulation and measured data differ. This is done in a simplified way by weighting the di-jet mass histogram (the invariant mass of the first two leading jets) with weights

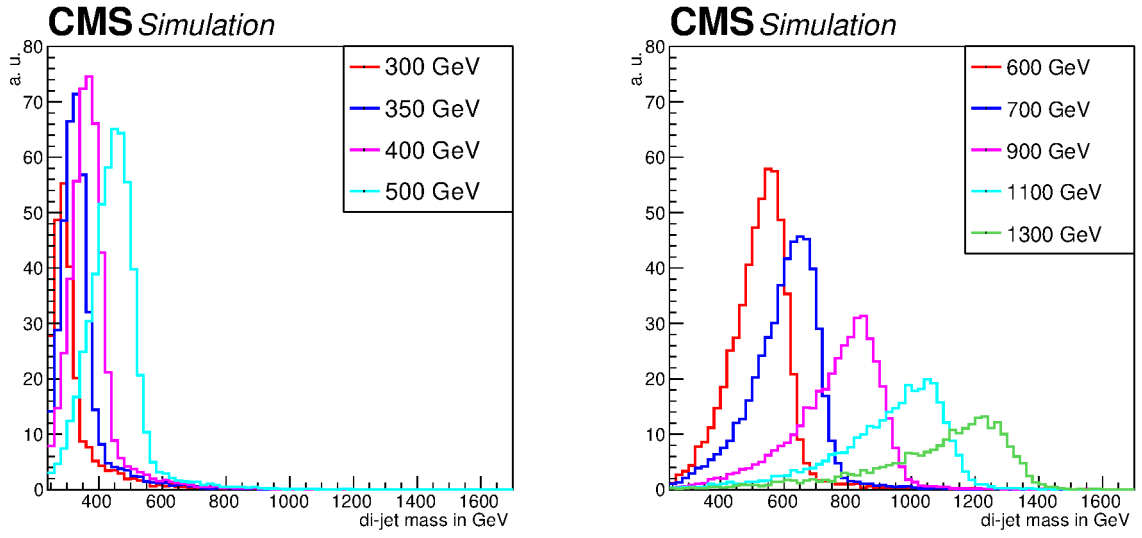
$$w = \prod_{i=1}^{n} s_i$$

where s_i is the scale factor for the i-th jet, which depends on the working point used, and n is the number of b jets being required (4 or 3 depending on the category). This term is simply the result from the Poisson binomial distribution with the scale factors being the probabilities and n successful Bernoulli trials. The di-jet mass distribution has also been scaled by a constant factor of 0.9 which is a rough estimate for online b-tagging inefficiencies in 2016. The resulting di-jet mass histograms can be seen in figure 5.1 and 5.2. Resulting signal efficiencies ε, which are



(a) (b)

Figure 5.1: Invariant mass distribution of the first two leading jets from Monte Carlo simulation after 4 b-tagged jets selection for different Higgs boson masses.



(a) (b)

Figure 5.2: Invariant mass distribution of the first two leading jets from Monte Carlo simulation after 3 b-tagged jets selection for different Higgs boson masses. Note that compared to figure 5.1 this has less fluctuation because of the larger amount of events in this category.

defined by the ratio of selected events to the total number of events, can be seen in figure 5.3. Both categories show similar shapes in their efficiencies with the maximum around 600 GeV. The initial increase is due to the trigger turn-on curves and the kinematic cuts. The decrease is due to the lower b-tagging efficiency for b jets with higher p_T. The 4 b-tagged jets category is roughly on average lower by a factor of 0.3. For higher masses the factor increases. This means that by requiring a fourth jet the suppression of signal is smaller for high masses than it is for lower masses.

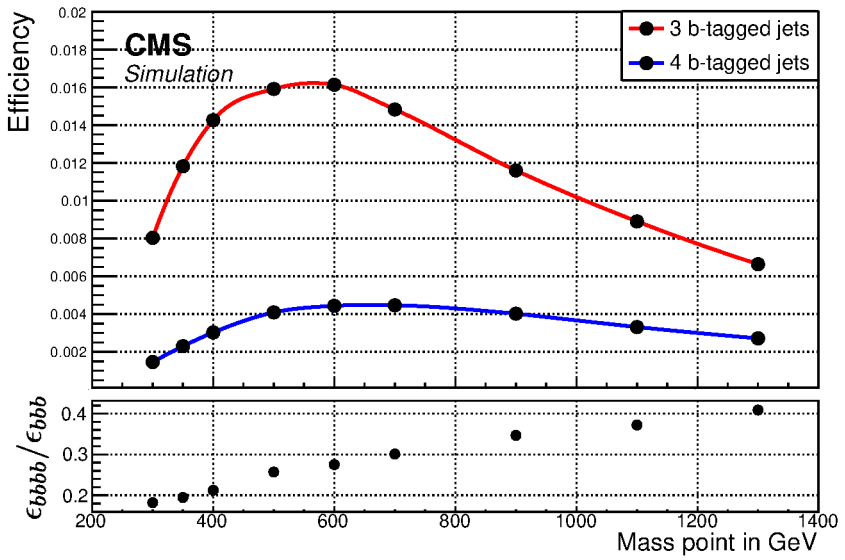


Figure 5.3: The upper plot shows the signal efficiencies for different Higgs boson masses for the 3 and 4 b-tagged jets category. Both have their maximum around 600 GeV while the 4 b-tagged jets category's efficiencies are overall much lower. Below are the ratios of the efficiencies from the two categories. It can be seen that for higher masses the 4 b-tagged jets category's efficiencies doesn't fall as fast as the one for the 3.

5.2 Signal over background and significance estimation

To get an estimate for the significance S/B and $S/\sqrt{B}$ are computed. Here S and B are obtained by integrating the number of events in the di-jet mass distributions of the Monte Carlo as used for the signal model in case of S and of the control region for B, where both integrations are done over a specific mass window. The mass distributions of the signal Monte Carlo are scaled to match the data luminosity while the mass distribution of the control region is scaled to the number of events of the signal region of the measured data. The mass window is set to be $[0.8s, 1.2s]$ where s is the position of the maximum in the mass distribution of the Monte Carlo. The results can be seen in table 5.1, table 5.2 and figure 5.4. One can observe a steep fall from 300 GeV to 600 GeV for both categories, which can be explained by the peak in the mass distributions being very narrow for lower masses as seen in figure 5.1 and figure 5.2 for the signal. This makes S very large for low masses. Both significance estimations show a similar

M in GeV	S (bbbb)	B (bbbnb)	S/B	$S/\sqrt{B}$
300	411.2	10236	0.040	4.1
350	296.0	9981	0.030	3.0
400	200.2	9743	0.020	2.0
500	73.74	5472	0.013	1.0
600	26.75	2737	0.0098	0.51
700	12.14	1977	0.0061	0.27
900	2.654	702.2	0.0038	0.10
1100	0.5820	185.6	0.0031	0.043
1300	0.1737	98.84	0.0018	0.017

Table 5.1: Numbers for computed signal, background and significance estimations for the 4 b-tagged jets category.

M in GeV	S (bbb)	B (bbnb)	S/B	$S/\sqrt{B}$
300	2614	132849	0.020	7.2
350	1762	125968	0.014	5.0
400	1137	124474	0.0091	3.2
500	338.6	60320	0.0056	1.4
600	118.3	28440	0.0042	0.70
700	45.08	14280	0.0032	0.38
900	8.338	4455	0.0019	0.12
1100	1.934	1722	0.0011	0.047
1300	0.5362	919.6	0.00058	0.018

Table 5.2: Numbers for computed signal, background and significance estimations for the 3 b-tagged jets category.

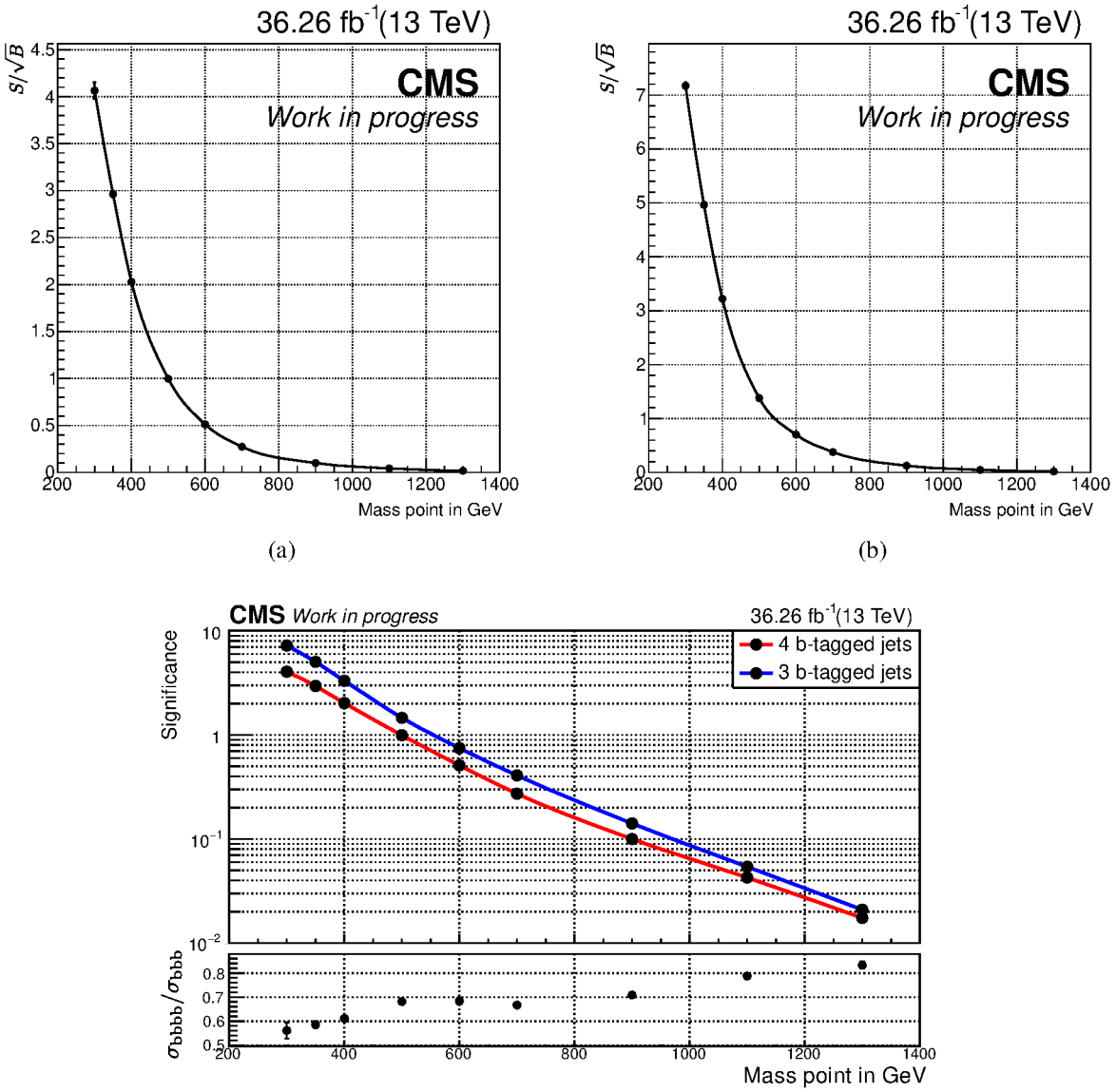


Figure 5.4: Significance estimation $S/\sqrt{B}$ for 4 b-tagged jets (a) and 3 (b) in linear scale and both in logarithmic scale (c) with their ratio.

shape as do the ratios S/B. In accordance with the efficiencies seen in figure 5.3 the significances have a similar shape across categories. The significance is lower by a factor of around 0.7 in the 4 b-tagged categories while this factor increases for higher masses. One explanation for this increase are the efficiencies as seen in figure 5.3, where the efficiency for the 4 b-tagged category doesn't fall as fast as the one for the 3 b-tagged category. This results in a relatively faster increase of S for the 4 b-tagged jet category than for the other category.

5.3 Novosibirsk function and background fit

The background model is taken from measured data. This is because the vast majority of the background arises from QCD events and the Monte Carlo simulated QCD events possibly aren't as accurate as this data driven method. Another benefit are better statistics because of the higher amount of events in the data compared to Monte Carlo. In fact the 3 b-tagged jet category has so many events in the control region that its distributions had to be pre-scaled by randomly selecting only 10% of the triggered events to correspond to roughly the same amount of events in the signal region. With the number of events lower, the uncertainties rise, which makes it easier to fit as there is less structure visible. In addition the number of events in the control region is now roughly on the same magnitude of order as the signal region.

A comparison between the di-jet mass distributions of the control regions of both categories can be seen in figure 5.5. Both distributions have little difference in shape, which means there is little to no bias because of the different selection between the two categories. The number of events in the 4 b-tagged jet category's control region is generally lower by a factor of 0.2 to 0.3. For masses higher than 1200 GeV the uncertainties go up due to the relatively low number of events in the 4 b-tagged jet category.

In order to produce a smooth di-jet mass distribution free from statistical fluctuation the resulting distribution from the control region is fitted to the turn-on Novosibirsk function $F(M_{12})$. This function is given by

$$F(M_{12}) = f(M_{12}) \cdot g(M_{12})$$

with the turn-on

$$f = \frac{1}{2}\left(\mathrm{Erf}\left(p_{\mathrm{slope}} \cdot (M_{12} - p_{\mathrm{turnon}})\right) + 1\right)$$

where Erf is the Gaussian error function

$$\mathrm{Erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\mathrm{d}t$$

and the Novosibirsk function

$$g(M_{12}) = \exp\left(-\frac{1}{2\sigma_0^2}\ln^2\left(1 - (M_{12} - p_{\mathrm{peak}})\frac{p_{\mathrm{tail}}}{p_{\mathrm{width}}}\right) - \frac{\sigma_0^2}{2}\right)$$

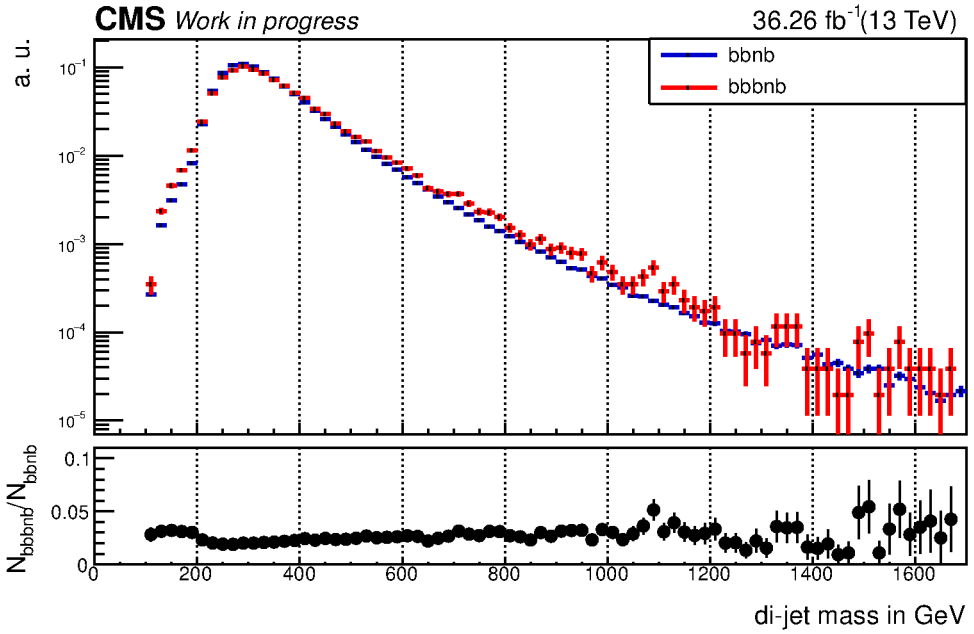


Figure 5.5: Comparison of background events between both categories. Both distributions have been scaled to the same value for display in the upper part. The lower part shows the ratio between the two (without scaling from the upper part). Note that the bbnb has been pre-scaled as described in the text. For masses below 300 GeV there are less events because of the trigger selection.

where

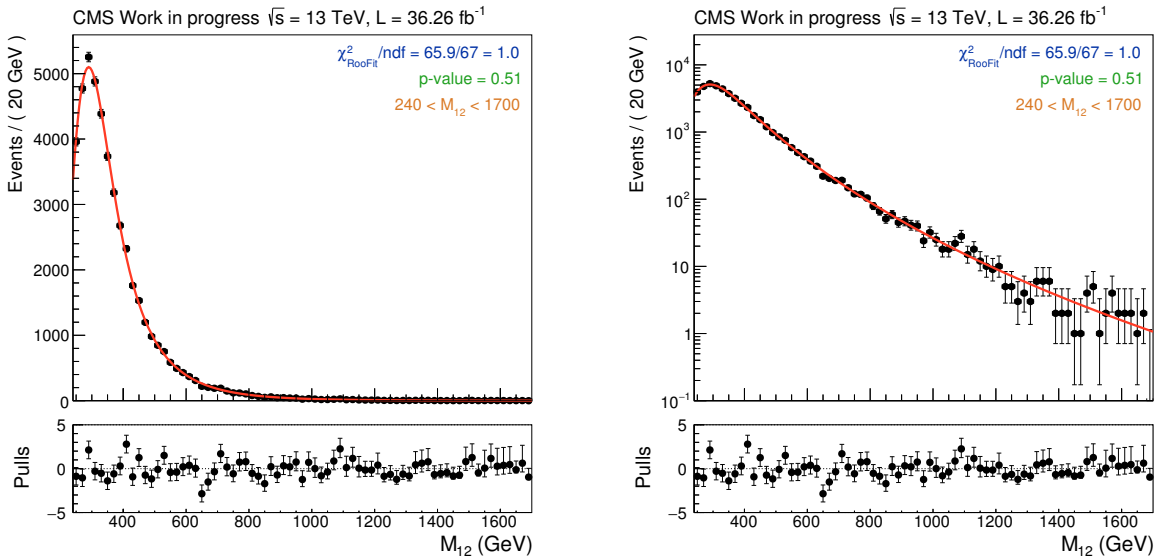
$$\sigma_0 = \frac{1}{\sqrt{\ln 4}} \sinh^{-1}(p_{\text{tail}}\sqrt{\ln 4})$$

where p_{slope}, p_{turnon}, p_{peak}, p_{tail} and p_{width} are free parameters determined by the fit. The turn-on function is able to reproduce the peak originating from the triggers for low masses, while the Novosibirsk function works well to model the dependence up to high masses in the distribution. This function does not possess any physical meaning in this context and is only used as an analytical tool to be able to subtract the background. For this reason bias tests are necessary but due to time constraints are not done in this thesis. The fit is being done over a wide mass range from 240 GeV to 1700 GeV and χ^2 as well as p-value are used to check the validity of the fit. Both χ^2/ndf and p-value should be optimally close to 1.

The results can be seen in figure 5.6 for the 4 b-tagged jet category and in figure 5.7 for the 3 b-tagged jet category. For the 4 b-tagged jet category the χ^2/ndf is approximately equal to 1 and has a p-value of 0.51, yet the 3 b-tagged jet category is a bit worse with a χ^2/ndf of 1.1 and a p-value of only 0.19. One can also see some structure in the pulls for masses between 400 GeV and 800 GeV, which isn't significant. This difference between the two categories might be caused by the higher number of events in the 3 b-tagged jet category and it is possible that with more events in the 4 b-tagged jet category some structure in the pulls would arise as well. Better fits might be achieved by using an extended Novosibirsk function or splitting the fit into sub-ranges and fitting low masses and high masses separately. The extended Novosibirsk function would use an additional term that is quadratic in the argument of the exponential function,

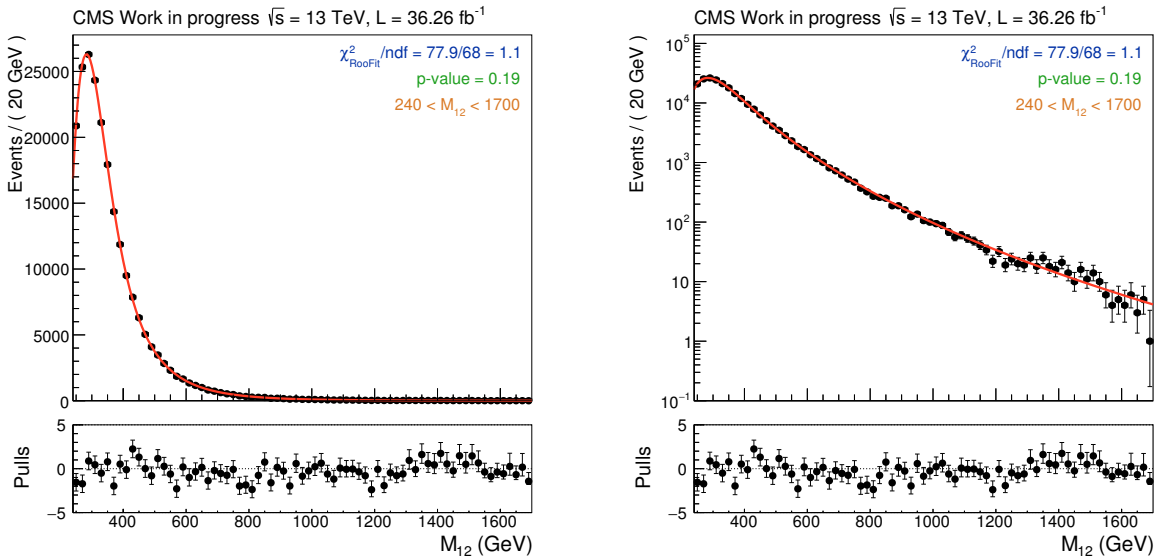
which makes this option involve an additional free parameter. The latter involves splitting the fit which can result in a discontinuity later on.

After the fits are done, histograms with the same binning as the ones created from Monte Carlo are created using the fits. This results in the templates later used for obtaining the limits as described in 7.



(a) (b)

Figure 5.6: Background fit for the 4 b-tagged jet category in linear (a) and logarithmic scale (b). Black dots depict measured data while the red line is the turn-on Novosibirsk function fit. The pulls represent the difference between data and fit divided by the uncertainty.



(a) (b)

Figure 5.7: Background fit for the 3 b-tagged jet category in linear (a) and logarithmic scale (b). Black dots depict measured data while the red line is the turn-on Novosibirsk function fit. The pulls represent the difference between data and fit divided by the uncertainty.

Chapter 6

Comparison between the two b-tagging algorithms CSVv2 and DeepCSV

Besides the b-tagging algorithm DeepCSV that this thesis uses mainly, another often used, older algorithm is CSV, which uses the same set of observables [7]. To evaluate their respective performance for differently flavored jets, events produced by Monte Carlo simulation, that include information about the actual jet flavor, are used. In particular these are top-anti-top events which yield a large amount of QCD decays similar to the background in this analysis. The results can be seen in figure 6.1.

For the medium working point and bottom quark jets DeepCSV clearly shows a higher efficiency than CSVv2, with the difference being even greater for higher p_t. Around 180 GeV DeepCSV is able to reach an efficiency close to 0.7. Overall both algorithms show a decreasing efficiency towards higher p_t. The charm quark jets are falsely identified as b jets with an efficiency of little over 0.1. While both algorithms seem to be similar for this flavor, below 300 GeV DeepCSV yields better results than CSVv2, because of a lower efficiency with a difference of around 0.03. For jets of other flavors, the algorithms show little difference for this working point. This is different for the loose working point, where DeepCSV shows a higher efficiency for those (falsely identified as b jet) light flavor and charm quark jets, which means a higher fake rate. In the interval from 400 GeV to 600 GeV the difference is noticeably around 0.05 for light flavor and around 0.1 for charm quark jets. However for the correctly tagged bottom quark jets DeepCSV still shows better results. At the tight working point light and charm quark jets are suppressed and efficiencies show again little difference between these algorithms. For the bottom quark jets DeepCSV again outperforms CSVv2.

A check if any bias arises from the newer DeepCSV algorithm compared to the older CSVv2 is shown in figure 6.2. The ratio between the mass distributions of the background events selected with these two algorithms is approximately equal to 1 without any strong structure differing from 1. This makes it plausible to assume that there is in fact no bias from the use of DeepCSV compared to CSVv2.

Comparisons between the resulting efficiencies of both categories from Monte Carlo events

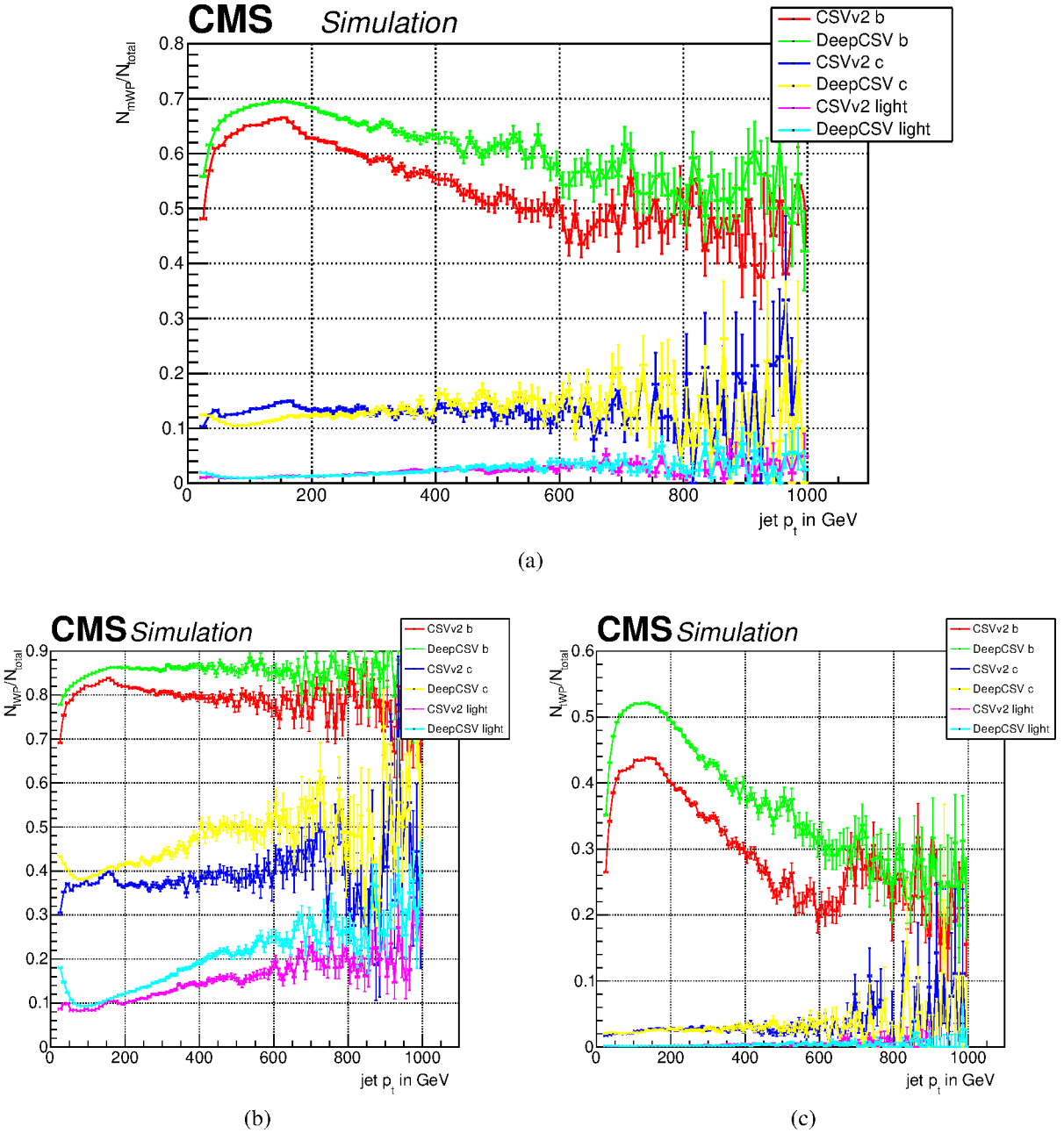


Figure 6.1: Ratio between the the number of b-tagged events and the total number of events of CSVv2 and DeepCSV at medium (a), loose (b) and tight (c) working points with respect to the transverse momentum p_t of the jet for variously flavored jets.

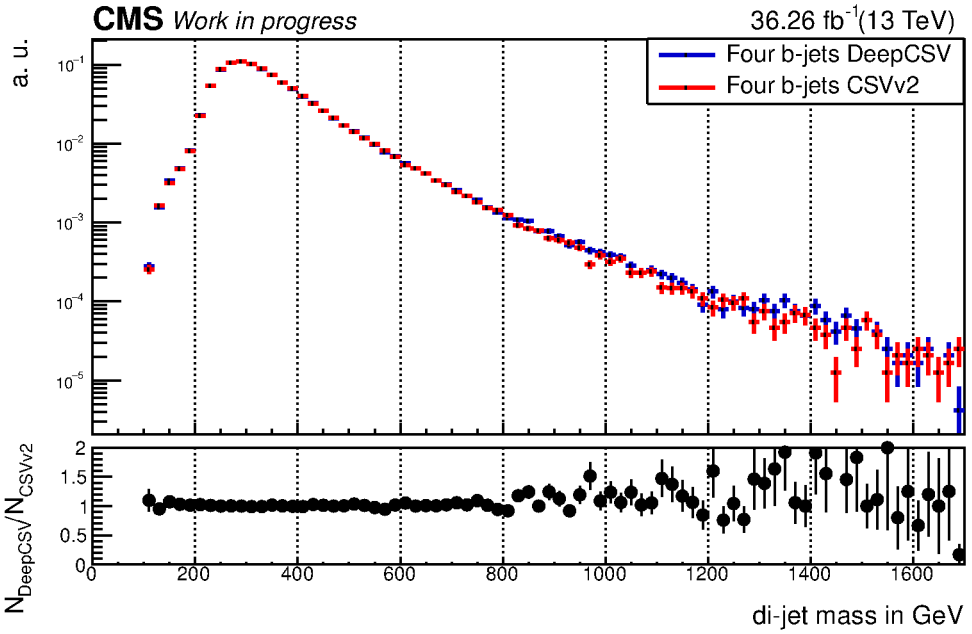
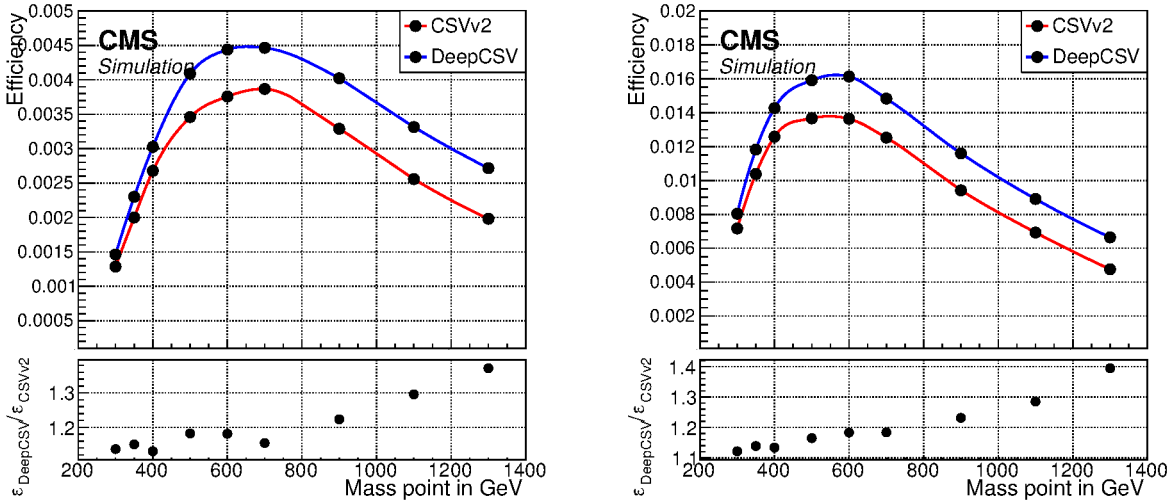


Figure 6.2: Comparison of background events of the 3 b-tagged jets category between CSVv2 and DeepCSV. Both distributions have been scaled to the same value for display in the upper part. The lower part shows the ratio between the two (without scaling of the upper part).

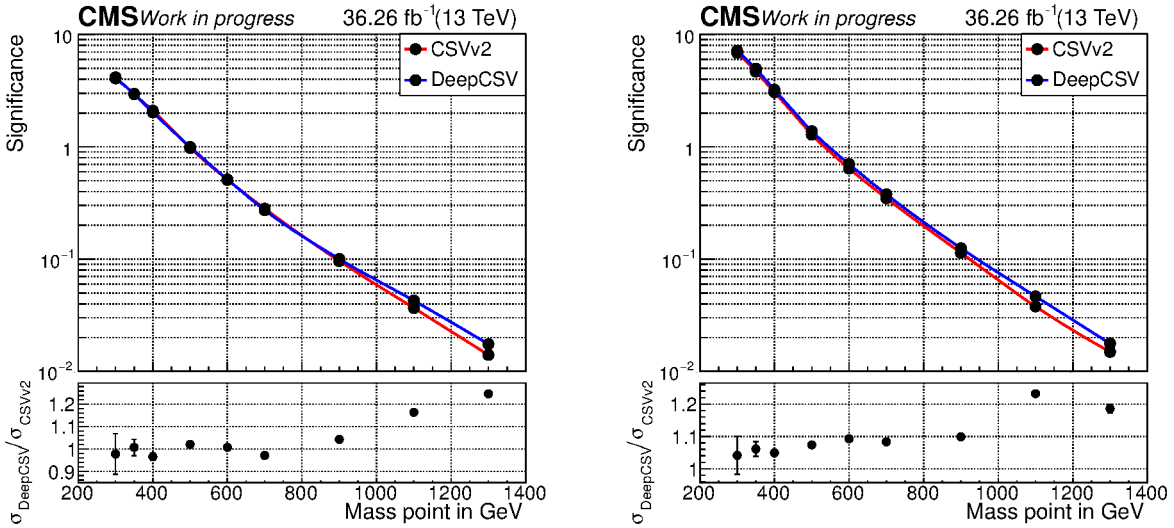
can be seen in figure 6.3. For both categories one can see that with DeepCSV the efficiencies increase by a factor between 1.1 and 1.4 compared to CSVv2 with a mean around 1.2. Especially for higher masses (above 800 GeV) DeepCSV seems to give better results. This agrees with the observation from before that for higher p_t the improvement gained by DeepCSV becomes larger. There is no noticeable difference in shape between the two categories.

Finally the significance estimation $S/\sqrt{B}$ is compared in figure 6.4. While there is still an increase in significance for DeepCSV the factor by which it increased is only around 1.1, which is lower than the efficiency increase from before. This is to be expected as the events in the background control region also have increased due to the higher efficiency. Assuming the same factor of 1.2 for the control region one immediately gains the factor $1.2/\sqrt{1.2} \approx 1.1$ for the significance estimation. The efficiency for the control region can rise easily because of the increased false identification of DeepCSV for the loose working point, which is the working point for the control region.



(a) (b)

Figure 6.3: Efficiencies of the signal region of the two algorithms CSVv2 and DeepCSV for the 4 b-tagged jets category (a) and the 3 b-tagged jet category (b). The lower part shows the efficiency of DeepCSV divided by the one of CSVv2.



(a) (b)

Figure 6.4: Significance estimations $S/\sqrt{B}$ of CSVv2 and DeepCSV for the 4 b-tagged jet category (a) and the 3 b-tagged category (b). The lower part shows the significance estimation of DeepCSV divided by the one of CSVv2.

Chapter 7

Results on the expected limits

7.1 Systematic uncertainties

Various systematic uncertainties arise from the methods used in this analysis, which have been considered before computing the limits:

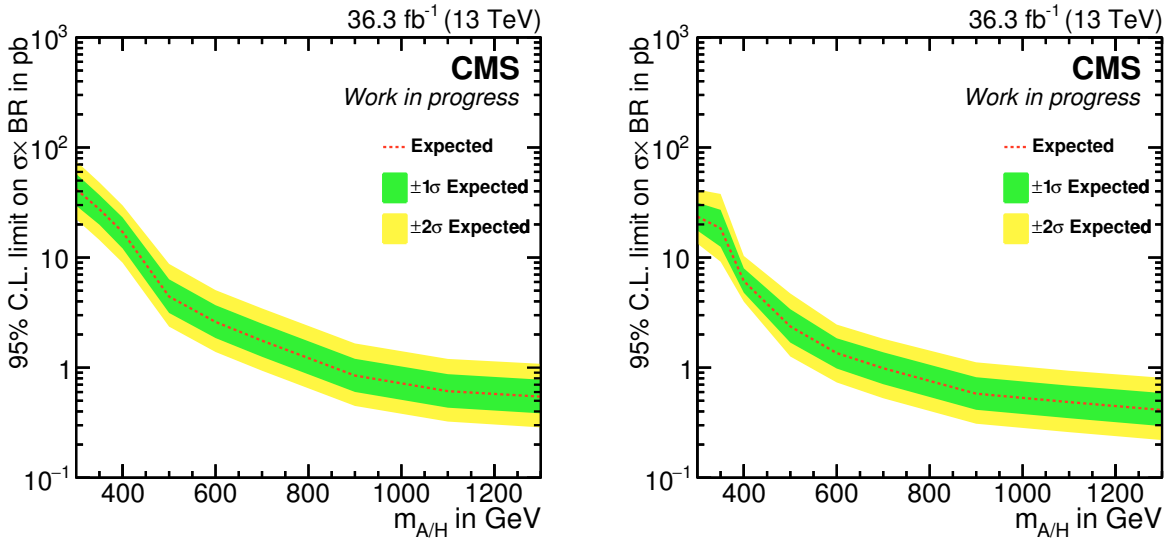
- Five uncertainties arise from the free parameters of the background fit. The value for these uncertainties are obtained as a result of the fit and are applied by using shape uncertainties obtained by varying the each affected parameter.
- The signal is affected by an uncertainty on the online b-tagging which is assumed to be log-normal distributed with a relative uncertainty of 5%.
- Likewise the signal luminosity is also assumed to be log-normal distributed and to have a relative uncertainty of 3%.
- The b-tagging scale factors applied as described in section 5.1 have uncertainties. The values of these depend on the p_T of the jets and so are applied as shape uncertainty.
- The jet energy scale, which is a scaling for the measurements of calorimeters, comes with an uncertainty and is also applied as shape uncertainty.
- Further uncertainties include those from sources such as bias as mentioned in section 5.3, jet energy resolution, pile-up reweighting and efficiencies. These (aside from bias uncertainties) are very likely to have an impact of less than 3% on the results and also have been omitted in this thesis.

7.2 Extraction of the limits

From this the upper limits for the background-only hypothesis are obtained by fitting the background (events from the control region) plus the signal (events from Monte Carlo) against the background. This maximum likelihood fit results in upper limits for the cross section times

branching ratio corresponding to a number of events in the signal region that can be observed caused by the uncertainties, even if there actually are no signal events. The extracted upper limits of 95% C.L. are shown for the two categories in figure 7.1 and listed in detail in table 7.1. After unblinding, which is outside the scope of this thesis, the upper limits can be used to determine if there are any excesses of events when fitting it against the data of the signal region.

In the upper limit plot for both categories one can observe an increase in the limit for lower



(a) (b)

Figure 7.1: Upper limits for the cross-section times the branching ratio for the 4 b-tagged jets category (a) and the 3 b-tagged jets category (b) with respect to the Higgs boson mass $m_{A/H}$.

	Limit for bbbb in pb					Limit for bbb in pb				
$m_{A/H}$ in GeV	-2σ	-1σ	Exp.	$+1\sigma$	$+2\sigma$	-2σ	-1σ	Exp.	$+1\sigma$	$+2\sigma$
300	22	30	42	58	75	13	18	24	32	41
350	14	20	28	37	48	9	13	18	27	38
400	9.0	12	17	23	30	4.0	4.8	6.1	8.0	10
500	2.4	3.1	4.4	6.3	8.8	1.3	1.7	2.4	3.4	4.7
600	1.4	1.9	2.6	3.7	5.1	0.73	0.98	1.4	1.8	2.5
700	0.94	1.3	1.8	2.5	3.5	0.53	0.71	0.99	1.4	1.8
900	0.45	0.60	0.85	1.2	1.7	0.31	0.41	0.58	0.82	1.1
1100	0.32	0.43	0.61	0.87	1.2	0.26	0.35	0.49	0.69	0.94
1300	0.29	0.38	0.54	0.78	1.08	0.22	0.29	0.41	0.59	0.81

Table 7.1: Upper limit values for the cross-section times the branching ratio for the 4 b-tagged category (bbbb) and the 3 b-tagged category (bbb).

masses. This is caused by the fitting penalty and the sharp decrease of the efficiency as seen in figure 5.3, which is mainly due to the transverse momentum selection applied. One also notices that the upper limits of the 4 b-tagged jet category are higher compared to the 3 b-tagged jet category. This can be observed in more detail in figure 7.2, where it is shown that the expected

limits of the 4 b-tagged jets category divided by the one of the 3 b-tagged jets category is mostly between 1.3 and 2.

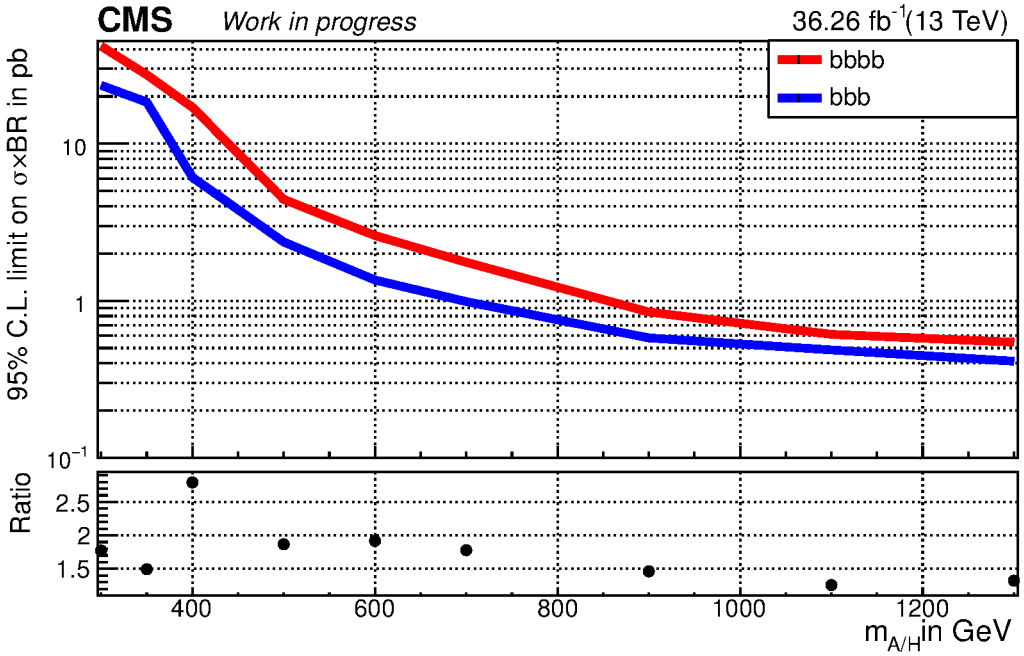
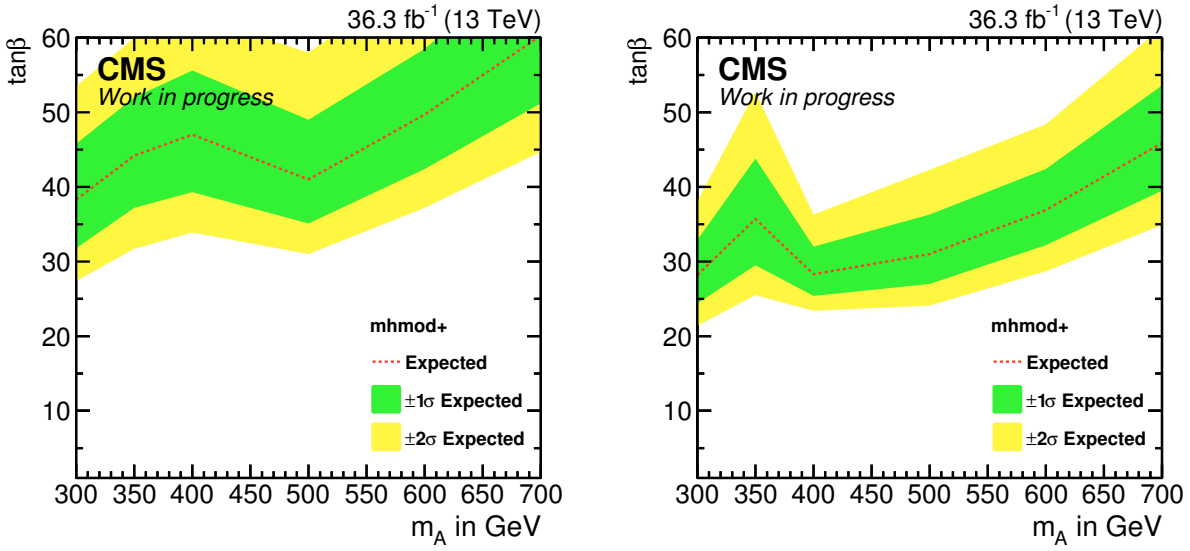


Figure 7.2: Comparison between the expected upper limits for the 4 b-tagged jets category and the 3 b-tagged jets category. The lower plot shows the ratio between the former and the latter.

Using theoretically obtained values for $\tan\beta$, the ratio of vacuum expectation values, one can compute upper limits for this ratio with respect to the Higgs boson mass. This is done in figure 7.3 using the $m_h^{\text{mod+}}$ benchmark scenario [16, 17] with a Higgsino mass parameter $\mu = 200\,\text{GeV}$. This benchmark scenario defines certain free parameters of the MSSM needed for the computation of $\tan\beta$. Furthermore these calculated values for $\tan\beta$ also have uncertainties which have to be considered when computing the limits. Compared to those discussed in the previous paragraph, these limits are model dependent as they explicitly need values obtained through a specific Supersymmetry model.



(a) (b)

Figure 7.3: Model-dependent upper limits for $\tan\beta$ using the $m_h^{\text{mod+}}$ benchmark scenario for the 4 b-tagged jets category (a) and the 3 b-tagged jets category (b).

Chapter 8

Summary and discussion

A search for an MSSM Higgs boson decaying into two bottom quarks has been performed. This was done by splitting the selected events into categories, one for events with 3 b-tagged jets and another one for at least 4 b-tagged jets. The resulting efficiencies are around 0.44% for the 3 b-tagged jets category signal region and around 0.037% in case of the other category. The selections were applied to events simulated by Monte Carlo for MSSM Higgs bosons of different masses and the efficiencies are observed to have a similar shape across categories with a global maximum around 600 GeV while the efficiency in case of the 4 b-tagged jets category is about one third of the 3 b-tagged jets category. The significance estimations for both categories have been observed to steeply decrease for higher masses. In the context of this analysis, the two b-tagging algorithms CSVv2 and DeepCSV have been compared and the latter has been deemed to be favorable with delivering an increase in the significance estimation of about 10% compared to CSVv2. The expected upper limits for the cross-section times branching ratio for MSSM Higgs boson decay channel assuming the background-only hypothesis have been obtained and range from 40 pb to 0.5 pb. The limits in case of the 4 b-tagged jets category is higher in comparison to the 3 b-tagged jets category by a factor of 1.3 to 2.

One has to consider that the 4 b-tagged jets contains roughly only 8% of the number of events of the 3 b-tagged jets category. But even though the number of the latter category is so much larger, the expected upper limits of the former category are only a factor of two lower. The limits are especially close for mass points higher than 800 GeV. While closer limits for lower mass points would also be desirable, this shows that an analysis that categorizes events into 3 and 4 b-tagged jets and then combines the two categories could yield improved results over an analysis without categorization.

Kapitel 9

Erklärung

Hiermit bestätige ich, dass die vorliegende Arbeit von mir selbständig verfasst wurde und ich keine anderen als die angegebenen Hilfsmittel – insbesondere keine im Quellenverzeichnis nicht benannten Internet-Quellen – benutzt habe und die Arbeit von mir vorher nicht einem anderen Prüfungsverfahren eingereicht wurde. Die eingereichte schriftliche Fassung entpricht der auf dem elektronischen Speichermedium. Ich bin damit einverstanden, dass die Bachelorarbeit veröffentlicht wird.

Hamburg, 26. Oktober 2017, Jonas Rübenach

Bibliography

[1] CMS Collaboration. Search for a Higgs boson decaying into a b-quark pair and produced in association with b quarks in proton-proton collisions at 7 TeV. 2013. doi: 10.1016/j.physletb.2013.04.017.

[2] CMS Collaboration. Search for neutral MSSM Higgs bosons decaying into a pair of bottom quarks. 2015. doi: 10.1007/JHEP11(2015)071.

[3] M. Thomson. *Modern Particle Physics*. Cambridge University Press, 2013.

[4] P. Bechtle, S. Heinemeyer, O. Stål, T. Stefaniak, G. Weiglein, and L. Zeune. MSSM Interpretations of the LHC Discovery: Light or Heavy Higgs? 2012. doi: 10.1140/epjc/s10052-013-2354-5.

[5] The LHC Higgs Cross Section Working Group et al. Handbook of LHC Higgs Cross Sections: 3. Higgs Properties. 2013. doi: 10.5170/CERN-2013-004.

[6] CMS Collaboration. Identification of b-quark jets with the CMS experiment. 2012. doi: 10.1088/1748-0221/8/04/P04013.

[7] CMS Collaboration. Heavy flavor identification at CMS with deep neural networks, 03 2017. URL `https://twiki.cern.ch/twiki/bin/view/CMSPublic/BTV13TeVDPDeepCSV`.

[8] CERN. The Large Hadron Collider, 2014. URL `https://cds.cern.ch/record/1998498`.

[9] CERN. LHC, the guide, 2017. URL `https://cds.cern.ch/record/2255762/files/CERN-Brochure-2017-002-Eng.pdf`.

[10] CERN. CERN's Accelerator Complex, 2015. URL `https://espace.cern.ch/acc-tec-sector/Pictures/CERN%27s%20accelerator%20complex2015.pdf`.

[11] CERN. What is CMS?, 2011. URL `http://cms.web.cern.ch/news/tracker-detector`.

[12] D. Barney. CMS slice, 10 2011. URL `https://cms-docdb.cern.ch/cgi-bin/PublicDocDB//ShowDocument?docid=5581`.

[13] M. Cacciari, G. P. Salam, and G. Soyez. The anti-k_t jet clustering algorithm. 2008. doi: 10.1088/1126-6708/2008/04/063.

[14] S. Agostinelli, J. Allison, K. Amako, and J. Apostolakis. Geant4 - a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250 – 303, 2003. ISSN 0168-9002. doi: https://doi.org/10.1016/S0168-9002(03)01368-8. URL `http://www.sciencedirect.com/science/article/pii/S0168900203013688`.

[15] T. Sjöstrand, S. Ask, J. R. Christiansen, R. Corke, N. Desai, P. Ilten, S. Mrenna, S. Prestel, C. O. Rasmussen, and P. Z. Skands. An Introduction to PYTHIA 8.2. 2014. doi: 10.1016/j.cpc.2015.01.024.

[16] M. Carena, S. Heinemeyer, O. Stål, C. E. M. Wagner, and G. Weiglein. MSSM Higgs Boson Searches at the LHC: Benchmark Scenarios after the Discovery of a Higgs-like Particle. 2013. doi: 10.1140/epjc/s10052-013-2552-1.

[17] LHC Higgs Cross Section Working Group. MSSM neutral Higgs cross sections, 2017. URL `https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWGMSSMNeutral`.