Search for the Higgs Boson in WH and tHq Production Modes with the CMS Experiment

Zur Erlangung des akademischen Grades eines
DOKTORS DER NATURWISSENSCHAFTEN
von der Fakultät für Physik des
Karlsruher Institut für Technologie (KIT)

genehmigte

DISSERTATION

von

Dipl.-Phys. Christian Böser
aus Schwetzingen

Mündliche Prüfung: 24. April 2015

Referent: Prof. Dr. Th. Müller
Institut für Experimentelle Kernphysik

Korreferent: Prof. Dr. U. Husemann
Institut für Experimentelle Kernphysik

Introduction

> *The scientist does not study nature because it is useful; he studies it because he delights in it, and he delights in it because it is beautiful.*
>
> – Henri Poincaré

Indeed, nature is beautiful! And as scientists it is our goal to reveal its beauty by understanding it. Particle physicists are able to take the deepest look into the architecture of nature. For this purpose huge machines are built reproducing the conditions only a blink of an eye after the Big Bang — the beginning of our universe. At the Large Hadron Collider (LHC) near Geneva particles with a tremendous amount of energy are brought to collision. To record the debris of such collisions large detectors like the CMS experiment are placed around the interaction points. The data recorded by these detectors helps to find out what our world is made of. The start of the Large Hadron Collider in 2010, and its terrific performance since then, took our understanding of the universe to the next level.

The most accurate theory characterizing the subatomic universe to date is the standard model of particle physics (SM). Developed in the 1960s, it unifies the explanation of three fundamental forces: The strong, the weak and the electromagnetic force. The SM successfully describes elementary particles — fermions, which are the building blocks of matter, and gauge bosons, that are the mediators of the forces — and their interactions. As of today all particles predicted by the standard model have been discovered. But for a long time there was one missing piece of the puzzle left: the Higgs boson.

This beauty of nature was revealed in 2012, when the ATLAS and CMS collaborations announced the observation of a new boson with a mass of 125 GeV that is comparable with the mass of Xenon atoms. The discovery depicts the greatest success of the LHC so far. In the meantime, many measurements were performed and so far no deviations in the properties of the Higgs boson with respect to the SM predictions are found. This SM-like Higgs boson is proof for the Higgs mechanism[1] that was predicted already in 1968 by Robert Brout, François Engler, Peter Higgs and others. The Higgs mechanism is the simplest theory to explain the massiveness of elementary particles via the concept of electroweak symmetry breaking. A scalar Higgs field is introduced by the theory that couples to the masses of elementary particles. The Higgs boson itself is the quantum of the field.

According to theory the Higgs boson decays in ∼60% of all cases into a bottom quark pair. However, this decay channel has not yet been observed due to the large amount of background processes with similar signatures in the detector. One

[1] *Higgs mechanism* is the usual shortened form of the Brout-Englert-Higgs (BEH) mechanism

possible way out is the investigation of the decay in distinct Higgs boson production modes. The most sensitive channel in the search for $H \to b\bar{b}$ decays is the Higgs boson production in association with either a W or a Z boson. When the W or Z bosons are required to decay leptonically, the background contributions are strongly reduced. In this production mode — and in others like the vector boson fusion or the associated production with top quark pairs — the ATLAS and CMS collaborations make huge efforts and already highly optimized analyses are published.

Another interesting production mode is the production of the Higgs boson in association with single top quarks (tHq). As the heaviest known elementary particle, the top quark supposedly holds an important position in the electroweak symmetry breaking mechanism. The tHq production provides a unique opportunity to investigate the couplings of the Higgs boson to fermions and bosons, and to test possible new physics contributions. One the one hand, the channel is sensitive to the relative phase of the couplings to fermions and bosons and thus any deviation to its prediction by the standard model can be observed. On the other hand, yet unobserved processes beyond the standard model could contribute to this channel. Both scenarios would be visible in an excess of signal-like events in data.

It is known that the standard model cannot be the Theory of Everything. With the data that is available so far, two of the main goals of the LHC era are testing the SM predictions with ever increasing precision and searching for yet unknown physics. The same goals hold for this thesis. First, it is tried to improve the search sensitivity for $H \to b\bar{b}$ decays in the WH channel by employing advanced reconstruction methods for jets. Second, the unique tHq production mode with $H \to b\bar{b}$ decays is investigated for the first time to find possible deviations from the standard model predictions. Both analyses exploit the full dataset of proton-proton collisions at a center-of-mass energy of $\sqrt{s} = 8\,\text{TeV}$ recorded by the CMS detector.

The thesis starts with the theoretical introduction to the standard model in Chapter 1. The focus lies on the properties and the production modes of the Higgs boson. A dedicated section covers the tHq production channel and the reasons for its uniqueness.

In Chapter 2 the extensive experimental setup needed to produce and identify heavy elementary particles like the Higgs boson is introduced. First, the acceleration chain at the LHC is outlined. Furthermore, the CMS experiment and the different detector parts are described.

Chapter 3 reviews the techniques used for the simulation of collision events which are compared to the recorded events in data. Moreover, the dedicated reconstruction techniques interpreting the raw electronic signals in the detector as physics objects are described. In this chapter also the different jet reconstruction algorithms are introduced that play a special role in the further analyses.

The confrontation of data and simulated events depends on several statistical tests. In addition, the use of multivariate tools is important to discriminate signal from background processes. Both facets are discussed in Chapter 4. The principles of Boosted Decision Trees and Neural Networks are described in detail.

The aim of the analysis described in Chapter 5 is to improve the search sensitivity for $H \to b\bar{b}$ decays at the CMS experiment. With this in mind the effect of including jet substructure information in the WH channel is investigated. The substructure information is extracted using a dedicated subjet/filter jet algorithm proposed by theorists. A novel filter jet energy regression technique is introduced. In order to quantify the improvements a cross check to the published CMS analysis is carried out and compared to the improved analysis using substructure information. For the first time the full statistical inference of using jet substructure in the $W(\ell\nu)H(b\bar{b})$ channel with the full 8 TeV dataset is presented.

Chapter 6 reviews the search for the associated Higgs boson production with single top quarks. The analysis is optimized for an anomalous Higgs boson coupling to fermions. Different multivariate analysis tool are used for the reconstruction of the final state and the discrimination of signal and background events. Upper limits on this exceptional production mode with $H \to b\bar{b}$ decays are evaluated for the first time at the LHC.

The findings of both analyses are discussed in the concluding chapter. Furthermore, the prospects of both production modes with the restart of the LHC in 2015 with increased energy are presented.

Contents

1. Theoretical introduction

Since many decades the standard model of particle physics is the most accurate theory describing elementary particles and their interactions. It has passed a vast amount of experimental tests with flying colors. The experimental observations of the bottom quark in 1977 [1], the top quark in 1995 [2, 3] and the tau neutrino in 2000 [4] were already major achievements of the theory. The most recent success story is the discovery of the Higgs boson by the ATLAS and CMS collaborations [5, 6].

Regardless of all these achievements, the SM cannot be the Theory of Everything; a desired hypothetical theory describing all physical aspects of the universe in one single framework. So far, all attempts to include a description of the fundamental force of gravitation[1] have failed. Moreover, the SM provides no explanation of dark matter and dark energy. For both there are strong cosmological evidences. That is why several modifications of the SM — so-called theories beyond the standard model (BSM) — exist providing possible answers to the open questions of nature.

This chapter presents an overview of the standard model, the fundamental particles and their interactions. Due to the extensive framework of the SM the mathematical introduction is left to textbooks (e.g. [8,9]) or up-to-date reviews (e.g. [10]). The second part of this chapter is dedicated to the Higgs boson, its main production modes at the LHC and the different Higgs boson decay channels. Furthermore, the discovery of the Higgs boson is discussed. A dedicated section on the unique Higgs boson production in association with single top quarks is provided at the end of the chapter.

1.1. The standard model of particle physics

The standard model of particle physics (SM) is the unified knowledge of the electroweak theory [11–14] and quantum chromodynamics (QCD) [15–18] and successfully describes the building blocks of matter, represented by fermions, and their interactions mediated by gauge bosons. The framework is formulated as a relativistic quantum field theory and is able to describe continuous systems with an infinite number of degrees of freedom. Mathematical functions known as *Lagrangian densities* constitute the dynamics of physical systems. The Lagrangians of the SM are introduced, such that they are invariant under local transformations based on the groups $\mathcal{SU}(3) \times \mathcal{SU}(2) \times \mathcal{U}(1)$. Noether's theorem [19] predicts a conserved quantity for every symmetry of a physical system. The symmetry under transformations

[1] Gravitation is satisfactorily expressed by Einstein's theory of General Relativity [7].

based on the $\mathcal{SU}(3)$ group, connected to the strong force, leads to *color charge*. Invariance under $\mathcal{SU}(2) \times \mathcal{U}(1)$ group transformations, linked to the electromagnetic and weak forces, yields the conserved quantities of *weak isospin* (T_3) and the *electric charge*.

The fundamental particles and their properties are discussed in the following. For the sake of convenience, the convention $\hbar = c = 1$ is used throughout the thesis.

1.1.1. Fundamental particles

In addition to the quantities color charge, weak isospin and electric charge, every particle has a quantum number known as *spin*. Bosons carry integer spin and follow Bose-Einstein statistics, thus an unlimited number of bosons can have the same energy state. In contrast to this fermions have half-integer spin and obey Fermi-Dirac statistics and the Pauli exclusion principle. Consequently, two fermions cannot share the same quantum state.

Gauge bosons

In the SM the quanta of the gauge fields of the electromagnetic, strong and weak forces are represented by gauge bosons all carrying a spin of $s = 1$. Table 1.1 summarizes the gauge boson and their properties.

Table 1.1.: Fundamental forces and the corresponding gauge bosons in the standard model. The electric charge and masses of the bosons are listed. For gluons the mass of zero is taken from theory predictions. All other values are taken from [10]. Furthermore, the interaction range of each force is given.

Force	Mediator	Mass	Electric charge [e]	Range
electromagnetic	photon (γ)	$< 10^{-18}$ eV	–	infinite
strong	8 gluons (g)	–	–	$\approx 10^{-15}$ m
weak	$W^{\pm}$ bosons	80.39 GeV	± 1	$\approx 10^{-18}$ m
weak	Z boson	91.19 GeV	–	

Gluons are the mediators of the strong force. The interactions between gluons and particles carrying color charge are described within the framework of QCD. The color charge can have the states *red*, *green* and *blue* or the corresponding anticolors. The gluon itself possesses a superposition of one unit of color charge and one unit of anticolor charge, and therefore is affected by the strong force as well. In total eight linearly independent kinds of color states are possible for gluons. The strong running coupling α_s increases with decreasing energy. This gives rise to two unique features of the strong force, *confinement* and *asymptotic freedom* of quarks, both discussed later in the section.

Table 1.2.: The fermions of the standard model grouped into generations. For each fermion the electric charge is given in units of e. The abbreviations r,g,b indicate the color charges red, green and blue, respectively. The heaviest fermion is the top quark with a mass of 173.2 GeV [28] (direct measurements). The bottom quark as the second heaviest fermion is ~ 40 times lighter ($m_\mathrm{b} = 4.2\,\mathrm{GeV}$).

Fermions	Generation 1	2	3	Electric charge	Color	Weak isospin (T_3)
Leptons	ν_e	ν_μ	ν_τ	0	0	$+1/2$
	e	μ	τ	-1	0	$-1/2$
Quarks	u	c	t	$+2/3$	r,g,b	$+1/2$
	d	s	b	$-1/3$	r,g,b	$-1/2$

The mediators of the electromagnetic force are photons. Photons are massless and carry no electric charge. Quantum electrodynamics (QED) [20–27] describes the interactions between photons and electrically charged particles. A consequence of the electromagnetic force is the forming of atoms, which are bound states of electrons and nuclei.

The mediators of the weak force are the electrically neutral Z boson and two charged W bosons ($\mathrm{W}^\pm$) carrying an electric charge of either $+1\,e$ or $-1\,e$, with the elementary charge of $e \approx 1.602 \cdot 10^{-19}\,\mathrm{C}$. The Z and W bosons are massive, in contrast to photons and gluons, and this restricts the range of the weak force to sub-nuclear scales. An example in which weak interactions take place is the β decay of a radioactive nuclei.

The electroweak theory accomplished the combination of QED and the theory of the weak force as a unified theory. The theory describes the mediators of the unified electroweak force as four massless gauge bosons. This seems to be in conflict with the observed masses of the Z and W bosons. The Higgs mechanism, described in the succeeding section, solves this problem by introducing the concept of electroweak symmetry breaking (EWSB).

Fermions

The fermions of the standard model can be classified into quarks and leptons. They are further ordered into three generations, each of which consists of two quarks — one up-type and one down-type quark — and two leptons – one with an electric charge and one electrically neutral neutrino. The difference between the particles of one generation compared to their partners from another generation lies in their masses. The three generations of fermions in the standard model are summarized in table 1.2. In the standard model for every fermion there is a corresponding antiparticle, which has the same mass but opposite charges.

Quarks are attributed with color charge, weak isospin and electric charge. There-

fore, they interact via the strong, the weak and the electromagnetic force. The up, charm and top quarks, summarized as up-type quarks, carry an electric charge of $+\frac{2}{3}e$. Their corresponding down-type partners, the down, strange and bottom quarks, have an electric charge of $-\frac{1}{3}e$. Quarks possess one unit of color, and antiquarks are attributed with one unit of anticolor. A bound state of quarks is *color-neutral* when it consists of three quarks with different color charges, or one quark and one antiquark attributed with a matching color-anticolor pair.

Quarks cannot exist as free particles. The energy between two quarks increases with their distance, so there are two limiting cases. For short distances quarks are quasi-free particles as the gluon field strength is very small. This phenomenon is called asymptotic freedom of quarks. As opposed to this, when two quarks are separated, the energy increases until it is large enough to produce a new quark-antiquark pair. This effect is known as confinement and is the reason why only color-neutral bound states of quarks, called *hadrons*, exist. Hadrons consisting of a quark and an antiquark, e.g. pions or kaons, are called mesons. Baryons are hadrons with three quarks as constituents, like protons or neutrons. Protons play an important role in nature, as they are the only hadrons which are considered stable. So far no experimental evidence for proton decays has been found.

Leptons on the other hand carry no color charge. The charged leptons of each generation, i.e. electrons, muons and taus, carry an electric charge of $-1e$. They interact via the electromagnetic and the weak force. Neutrinos, the weak isospin partners of the charged leptons, do not carry electric charge, and thus interact exclusively via the weak force. In the SM neutrinos are assumed to be massless. However, direct observations of neutrino oscillations (e.g. [29]) indicate that neutrinos carry mass. There are various extensions of the standard model trying to include a neutrino mass generation mechanism [30].

1.1.2. The Higgs mechanism

Within the mathematical framework of the SM all particles have to be massless, as introducing a mass term to the Lagrangians would violate local gauge symmetry. Especially, the observed high masses of the W and Z bosons, seems to be in conflict with this. Based on the work of Anderson, Brout, Englert, Guralnik, Hagen, Higgs and Kibble the *Higgs mechanism* [31–33] was developed to explain the masses of W and Z bosons via spontaneous breaking of electroweak symmetry.

The mechanism introduces a complex scalar field ϕ coupling to the mass of particles, the *Higgs field*. The field is chosen such that it only affects the $\mathcal{SU}(2)$ group symmetry from the electroweak theory, as the photons should remain massless[2]. The effective potential of the Higgs field has a local extremum at $\phi = 0$, but has an infinite number of global minima at $|\phi| > 0$ that represent the vacuum. This is often referred to as *Mexican hat potential*. At high energies the gauge bosons are located at $\phi = 0$ and the local gauge symmetry of the standard model is conserved. At

[2]The unbroken $\mathcal{U}(1)$ part – the *electric charge group* – is defined by the combination of generators $Q = T_3 + Y/2$, where T_3 is the weak isospin and Y denotes the weak hypercharge.

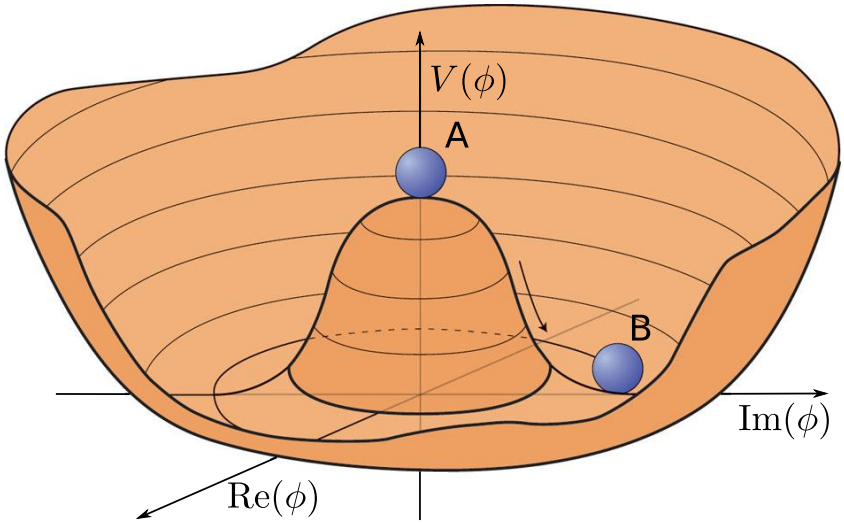


Figure 1.1.: Sketch of the effective potential of the Higgs field adopted from [34]. At high energies particles are located at $\phi = 0$ and do not interact with the Higgs field (A). The cylindrical symmetry of the system is conserved. At lower energies this symmetry is spontaneously broken, as the state of particle chooses one distinct minimum of the potential (B).

lower energies the symmetry is spontaneously broken by choosing a distinct ground state, as illustrated in Figure 1.1.

For the introduced field four degrees of freedom are postulated. According to the Goldstone theorem [35,36] for every broken symmetry there is a massless Goldstone boson, ergo four Goldstone bosons are expected. The Higgs mechanism explains how three Goldstone bosons are absorbed by the W^+, W^- and Z bosons, giving them masses and thus longitudinal polarization states. The missing forth degree of freedom predicted the existence of a massive spin-zero particle. The discovery of a massive Higgs boson announced in 2012 is proof to this theory and led to the Nobel Price in Physics 2013 for Peter Higgs and François Englert. The prize was awarded

> *"for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider"* [37].

The introduced Higgs field is also used to explain the masses of leptons and quarks. The masses of leptons are generated via *Yukawa couplings* between the Higgs field and the lepton fields. The Yukawa couplings are introduced such that only electrically charged leptons interact with the Higgs field. Neutrinos remain massless.

To explain the masses of quarks Yukawa couplings can be used in a similar but more complex way. The definition is more complicated, as the quark's weak eigenstates are not equal to their mass eigenstates. The transformations between the

weak eigenstates, denoted with q', to the mass eigenstates can be described via the Cabibbo-Kobayashi-Maskawa (CKM) matrix [38, 39]

$$\begin{pmatrix} |\mathrm{d}'\rangle \\ |\mathrm{s}'\rangle \\ |\mathrm{b}'\rangle \end{pmatrix} = \begin{pmatrix} V_{\mathrm{ud}} & V_{\mathrm{us}} & V_{\mathrm{ub}} \\ V_{\mathrm{cd}} & V_{\mathrm{cs}} & V_{\mathrm{cb}} \\ V_{\mathrm{td}} & V_{\mathrm{ts}} & V_{\mathrm{tb}} \end{pmatrix} \begin{pmatrix} |\mathrm{d}\rangle \\ |\mathrm{s}\rangle \\ |\mathrm{b}\rangle \end{pmatrix} = V_{\mathrm{CKM}} \begin{pmatrix} |\mathrm{d}\rangle \\ |\mathrm{s}\rangle \\ |\mathrm{b}\rangle \end{pmatrix} . \tag{1.1}$$

The CKM matrix has to be unitary, as all transition probabilities sum up to one. The squared absolute values of the matrix elements $|V_{\mathrm{q}_i\mathrm{q}_j}|^2$ are proportional to the electroweak transition probability from q_i into q_j. The elements are experimentally accessible via weak decay rates of mesons and baryons or cross section measurements, e.g. from single top quark production [40]. The most recent review by the Particle Data Group [10] quotes following values:

$$\tilde{V}_{\mathrm{CKM}} = \begin{pmatrix} 0.97427 \pm 0.00014 & 0.22536 \pm 0.00061 & 0.00355 \pm 0.00015 \\ 0.22522 \pm 0.00061 & 0.97343 \pm 0.00015 & 0.0414 \pm 0.0012 \\ 0.00886^{+0.00033}_{-0.00032} & 0.0405^{+0.0011}_{-0.0012} & 0.99914 \pm 0.00005 \end{pmatrix} . \tag{1.2}$$

The values are results of a global fit taking all available measurements and theoretical constraints into account. From the fact that the diagonal elements are much larger than the off-diagonal elements it can be deduced that flavor transitions inside one generation are preferred. For instance, given $|V_{\mathrm{tb}}| \approx 1$ and $|V_{\mathrm{td}}|, |V_{\mathrm{ts}}| \ll |V_{\mathrm{tb}}|$ the heaviest quark, the top quark, decays with a probability of roughly 100% into a bottom quark and a W boson.

1.1.3. Cross section calculation

The quantum field theory provides the tools to calculate the probability for the transition of an initial state $|i\rangle$ into a final state $|f\rangle$. Following the S-matrix formalism, the transition amplitude $\mathcal{A}$ is given by

$$\mathcal{A} = \langle f|S|i\rangle \, . \tag{1.3}$$

Here, the matrix S denotes the time-evolution operator in quantum mechanics. From a time-dependent perturbative analysis of $\langle f|S|i\rangle$ using the *Dyson series* [41] the matrix elements $\mathcal{M}_{fi}$ are calculated.

Following *Fermi's golden rule* [42] the cross section σ for a process $i \to f$ in a given part of the phase space Π is proportional to the square of the matrix elements

$$\mathrm{d}\sigma_{i\to f} \sim |\mathcal{M}_{fi}|^2 \cdot \mathrm{d}\Pi \, . \tag{1.4}$$

Technically, the calculation of cross section results in the calculation of the matrix elements. The order of the perturbative calculation used for the Dyson series determines the order of precision for the cross section. The more orders taken into

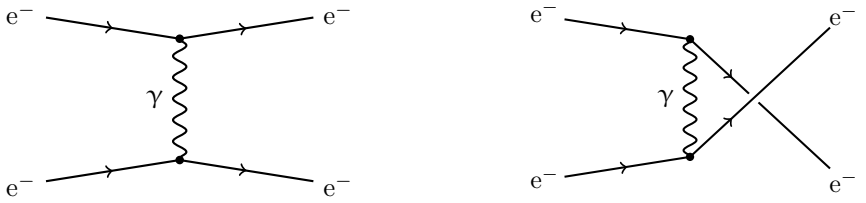


Figure 1.2.: Leading-order Feynman diagrams for Møller scattering. The lines and vertices represent mathematical terms for the calculation of the cross sections, in which all possible diagrams need to be summed. Throughout this thesis time is evolving from left to right.

account, the more precise are the theoretical predictions. Nonetheless, the cross sections applied in this thesis are mostly based on next-to-leading (NLO) or next-to-next-to-leading order (NNLO) calculations as the complexity of computation increases greatly with every order.

For the Dyson series itself all possible transitions from $|i\rangle$ to $|f\rangle$ have to be summed. Feynman diagrams provide graphical representations for these transitions. An example of electron-electron scattering, also known as Møller scattering [43], is given in Figure 1.2. Here, all possible Feynman diagrams at leading-order precision are presented. Each line and vertex represents a mathematical term following the *Feynman rules*. The vectors in Figure 1.2, for instance, represent the space-time propagation of the electrons. The internal photon is the mediator of the interaction and introduces a propagator term to the calculation. The vertices are the integration coordinates and enter the calculation with a term proportional to the corresponding coupling constants. Each vertex yields four-momentum conservation. It should be noted that a particle interaction is only symbolized by the sum of all Feynman diagrams. For the sake of convenience, usually one representative Feynman diagram at leading-order is depicted to characterize a process.

At the Large Hadron Collider, as the name suggests, hadrons are brought to collision at high energies. Hadrons are composite objects, built from valence quarks, defining their quantum numbers, a sea of virtual quarks surrounding them and gluons binding them together. When accelerated the hadron's momentum is distributed over all of its partons. The data used in this thesis was recorded in proton-proton collisions. In order to compute the theoretical predictions, the proton's substructure needs to be taken into account. A *factorization ansatz* [44] is chosen for the calculation of cross sections, via

$$\sigma_{pp\to X}(\mu_R, \mu_F) = \sum_{i,j} \int \mathrm{d}x_i \mathrm{d}x_j f_i(x_i, \mu_F) f_j(x_j, \mu_F) \hat{\sigma}_{ij\to X}(x_i, x_j, \mu_F, \mu_R)\,. \quad (1.5)$$

Here, $\hat{\sigma}_{ij\to X}$ is the cross section for the process $ij \to X$ which can be calculated perturbatively via the ansatz in Equation (1.4). The interacting partons are denoted with i and j, and f_i indicates their respective parton distribution function

(PDF). The PDF $f_i(x)$ represents the probability to find a particle of type i carrying the momentum fraction x in the proton. These functions are extracted from deep-inelastic scattering measurements. In Section 3.1 an exemplary PDF parameterization is depicted. The PDFs depend on the *factorization scale* μ_F, that is introduced to separate short and long range interactions. The partonic cross section has an explicit dependence on the *renormalization scale* μ_R, the scale at which the running coupling α_s is calculated. The choice of both parameters, μ_F and μ_R, is arbitrary to some extend. Often they are set to the typical momentum transfer Q^2 depending on the process.

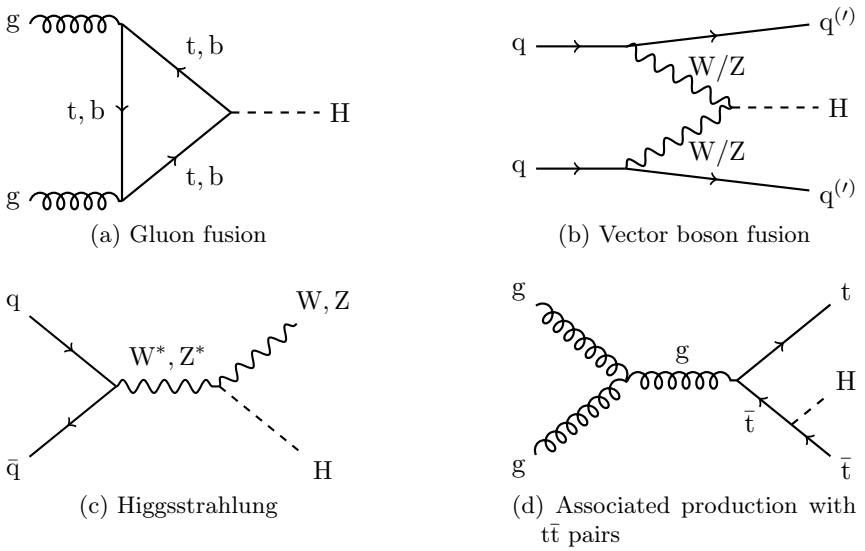
1.2. The Higgs boson

The Higgs boson has been the only left missing particle in the SM for a long time. As it is the particle associated to the Higgs field generating particle's masses, the discovery of the Higgs boson in 2012 was also proof of the Higgs mechanism.

This section covers the main Higgs boson production modes at the LHC. Furthermore, its decay channels and properties are discussed.

1.2.1. Higgs boson production channels at the LHC

There are many different production modes for Higgs bosons at the LHC. The four major channels are depicted in Figure 1.3, sorted by their cross sections. The associated Higgs boson production with single top quarks, which is investigated in Chapter 6, is discussed in a dedicated section afterwards.



(a) Gluon fusion

(b) Vector boson fusion

(c) Higgsstrahlung

(d) Associated production with $t\bar{t}$ pairs

Figure 1.3.: Representative LO Feynman diagrams for the four main Higgs boson production.

The *gluon fusion* process, as depicted in Figure 1.3(a), is the dominant production mode for the Higgs boson at the LHC. As the Higgs boson couples exclusively to massive particles a direct coupling to the massless gluons is not possible. Therefore, the production takes place via virtual quark triangle loops. The coupling is proportional to the masses of the quarks, so a top quark loop is the dominant mode. This production mode is very clean, as at leading-order no additional particles are expected in the final state.

The *vector boson fusion* shown in Figure 1.3(b) has the second largest cross section that is already one order of magnitude smaller compared to the gluon fusion. Here, the Higgs boson is produced via the fusion of two vector bosons. For that either two W bosons with opposite electromagnetic charges are radiated and change the flavor of the initial state quarks, or two Z bosons are emitted from two initial quarks. This channel has a specific topology with two forward light jets that are exploited to discriminate the signal process from background contributions.

The Higgs boson production in association with a vector boson is depicted in Figure 1.3(c). In this process, two initial state quarks produce a virtual W or Z boson, that radiates a Higgs boson. The process, often referred to as *Higgsstrahlung*, is the search channel in the analysis described in Chapter 5 and its characteristics will be discussed further there. The cross section is even lower compared two the above mentioned processes.

Another production mode is represented by the radiation of a Higgs boson from a high energetic top quark pair, as shown in Figure 1.3(d). In this interesting channel, the magnitude of the Higgs boson coupling to top quarks can be accessed. In the analysis described in Chapter 6 this production mode is considered as background.

Figure 1.4 summarizes the cross sections of the different production channels.

1.2.2. Higgs boson decay modes

With the observation of the Higgs boson and the determination of the boson's mass, it is possible to predict its decay branching ratios. In Figure 1.5 these branching ratios are shown as a function of the mass of the Higgs boson. The Higgs boson decays predominantly into a pair of bottom quarks. For the fermionic decay channels $H \to b\bar{b}$ is followed by $H \to \tau\tau$ and $H \to c\bar{c}$. The decay into two top quarks is kinematically forbidden. The dominant bosonic decay modes are $H \to WW$ and $H \to gg$. However, the decays into ZZ and $\gamma\gamma$ are due to their signatures the most sensitive channels in the search for the Higgs boson. The process $H \to \gamma\gamma$ is only possible via top quark or W boson loops. Furthermore, it should be noted that the Higgs boson decay into two vector boson requires one of the vector bosons to be virtual.

1.2.3. Higgs boson observation and properties

The most recent publications from the ATLAS and CMS collaborations [47, 48] — the culmination of many years of hard work — leave little doubt for the discovered

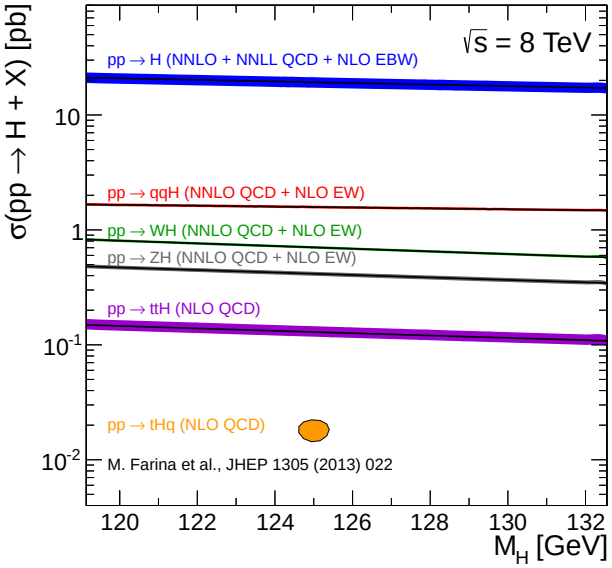


Figure 1.4.: Standard model Higgs boson production cross sections for different modes adopted from [45]. These cross section are many orders of magnitude smaller with respect to other well-known processes at the LHC, like the production of W bosons or top quark pairs.

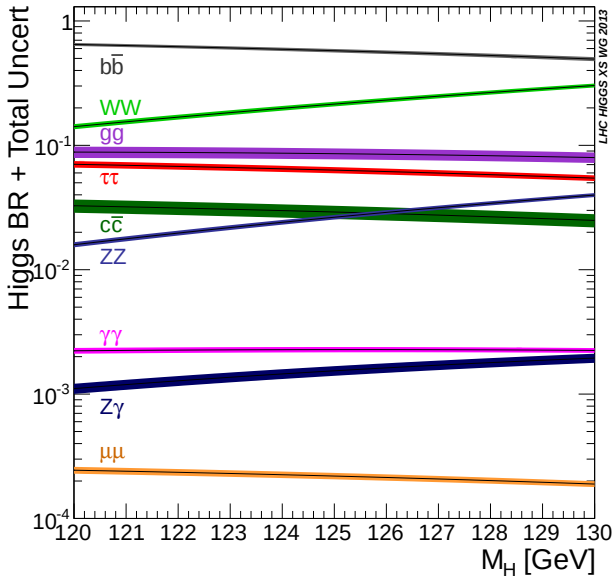


Figure 1.5.: Higgs boson branching ratios over its invariant mass, taken from [46]. The error bands account for theoretical and parametric uncertainties in the calculations. For a mass of $m_{\mathrm{H}} = 125\,\mathrm{GeV}$ the Higgs boson decays into a bottom quark pair in 58% of all cases.

Table 1.3.: Expected and observed significances for most sensitive Higgs boson decay modes at CMS [48]. The bosonic channels $H \to \gamma\gamma$, $H \to ZZ$ and $H \to WW$ are observed. For fermionic decays of the Higgs boson there is evidence in the $H \to \tau\tau$ channel.

	Significance ($m_H = 125\,\text{GeV}$)	
Channel	Expected [σ]	**Observed [σ]**
$H \to \gamma\gamma$	5.3	**5.6**
$H \to ZZ$	6.3	**6.5**
$H \to WW$	5.4	**4.7**
$H \to \tau\tau$	3.9	**3.8**
$H \to b\bar{b}$	2.6	**2.0**

boson being the searched-for Higgs boson. The updated and combined measurements all report properties which are in good agreement with the SM predictions.

Table 1.3 summarizes the search significances in the CMS effort, divided into decay channels. The table shows that the $H \to ZZ$ and $H \to \gamma\gamma$ analyses observe individual significances over 5σ, which is sufficient to claim a discovery. Furthermore, the observed significance of $H \to WW$ decays is close to 5σ, so an observation will be claimed in the near future. In the fermionic sector there is evidence ($> 3\sigma$) in the $H \to \tau\tau$ decay channel. The table also reveals that despite the large branching ratio $H \to b\bar{b}$ decays have not been observed yet. The measured cross sections of Higgs boson production are consistent with the standard model predictions. Furthermore, the spin J and parity P favor the SM expectation of $J^P = 0^+$. In Figure 1.6(a) the signal strengths are depicted for the separate decay channels.

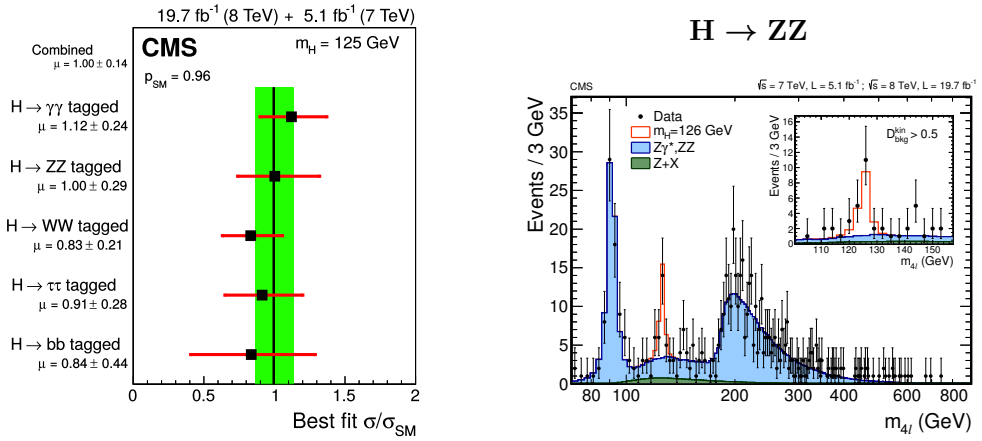
In the $H \to ZZ \to \ell\ell\ell\ell$ [49] and $H \to \gamma\gamma$ [50] channels a good invariant mass resolution can be achieved. For the former the invariant mass distribution with a visible excess in data at 125 GeV is shown in Figure 1.6(b). The combined results determine the mass of the Higgs boson to be

$$m_H = 125.02^{+0.26}_{-0.27}\,(\text{stat.})^{+0.14}_{-0.15}\,(\text{sys.}) \quad \textbf{CMS}\ [48]\,.$$

The ATLAS collaboration performed a similar measurement and reports a Higgs boson mass of

$$m_H = 125.5 \pm 0.2\ (\text{stat.})^{+0.5}_{-0.6}\ (\text{sys.}) \quad \textbf{ATLAS}\ [47]\,.$$

Within the uncertainties, the masses found in both experiments agree with each other. The combination of both measurements yields $m_H = 125.09 \pm 0.24\,\text{GeV}$ [51].



(a) (b)

Figure 1.6.: Signal strengths for the most sensitive Higgs boson decay modes [48] in (a) and m_{4l} invariant mass distribution of the H $\rightarrow$ ZZ search [49] in (b). The measured cross sections are compatible with the standard model predictions in all channels. In the left diagram the excess in data at $m_{4\ell} \approx 126\,\text{GeV}$ is covered by the Higgs boson signal (red histogram).

1.3. Higgs boson production in association with single top quarks

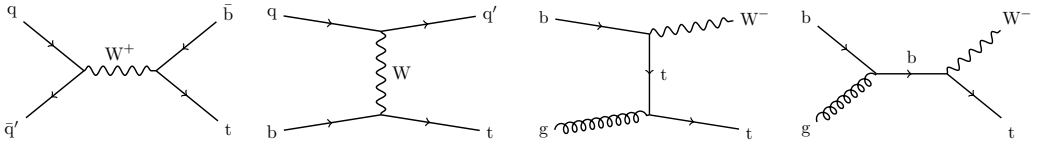
The Higgs boson production in association with a single top quark (tHq) is unique. In this channel, it is possible to study the Higgs boson couplings to fermions g_{Htt} and vector bosons g_{HWW}, and in particular their relative phase. It is convenient to normalize the investigated coupling to their SM predictions and generalize the couplings for fermions and vector bosons

$$\kappa_{\text{f}} \equiv g_{\text{Htt}}/g_{\text{Htt}}^{\text{SM}} \quad \text{and} \quad \kappa_{\text{V}} \equiv g_{\text{HWW}}/g_{\text{HWW}}^{\text{SM}}\,. \tag{1.6}$$

After a short foray into the single top quark production, the possibilities for probing these parameters of the standard model with the tHq production will be discussed.

Single top quarks are produced via the electroweak interaction. Three different production modes can be distinguished: s-channel production, t-channel production and tW-channel production. For each of the production modes in Figure 1.8 the representative LO Feynman diagrams are depicted. Top quarks decay in almost 100% of all cases into a b quark and a W boson. At the LHC, single top quark production has been observed in the t-channel [52] and the tW-channel [53]. Evidence of the s-channel production mode has only been observed at Tevatron [54].

The main channel for the tHq process is via single top t-channel production. The Higgs boson is radiated either from the single top quark or the W boson. The two representative Feynman diagrams are provided in Figure 1.8.



(a) s-channel (b) t-channel (c) tW-channel

Figure 1.7.: Representative Feynman diagrams for single top quark production. The cross section for the t-channel is the highest of the three modes.

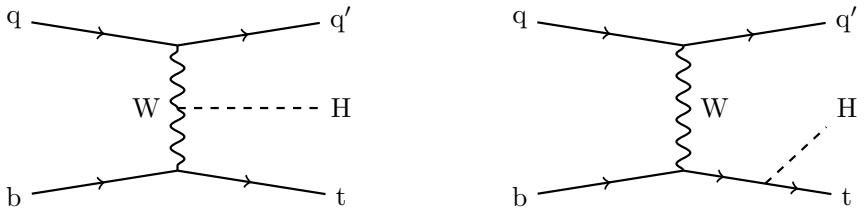


Figure 1.8.: Representative Feynman diagrams for Higgs boson production in association with single top quarks.

These two diagrams interfere with each other. The production amplitude is calculated in [55] as

$$\mathcal{A} = \frac{g}{\sqrt{2}} \left[(\kappa_\mathrm{f} - \kappa_\mathrm{V}) \frac{m_\mathrm{t}\sqrt{s}}{m_\mathrm{W} v} A\left(\frac{t}{s}, \phi; \xi_\mathrm{t}, \xi_\mathrm{b}\right) + \left(\kappa_\mathrm{V} \frac{2m_\mathrm{W}}{v} \frac{s}{t} + (2\kappa_\mathrm{f} - \kappa_\mathrm{V}) \frac{m_\mathrm{t}^2}{m_\mathrm{W} v} \right) B\left(\frac{t}{s}, \phi; \xi_\mathrm{t}, \xi_\mathrm{b}\right) \right] . \quad (1.7)$$

Here, s and t are the Mandelstam variables. A and B are functions depending on the azimuthal angle ϕ and a specific spinor basis ξ_t, ξ_b.

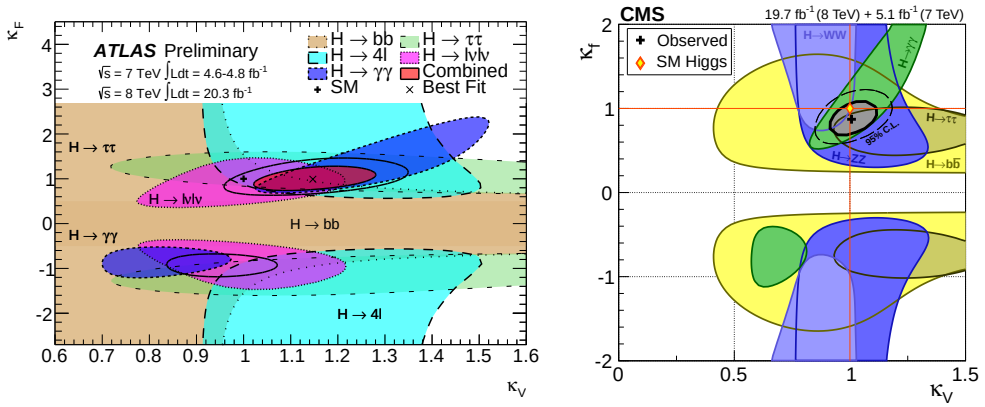
The main feature of Equation (1.7) is the term proportional to $(\kappa_\mathrm{f} - \kappa_\mathrm{V})$ highlighted in green. By construction, in the standard model κ_f and κ_V are equal to $+1$ and the term cancels out. The resulting cross section[3] of

$$\sigma(\mathrm{pp} \to \mathrm{tHq})_\mathrm{SM} = 18.28^{+0.42}_{-0.38}\,\mathrm{fb} \quad (1.8)$$

is tiny with respect to other production modes, as depicted in Figure 1.4.

However, any deviation of κ_f or κ_V with respect to SM prediction would lead to an enhanced cross section of tHq production. The values are already constrained by several measurements from the ATLAS and CMS collaborations. The allowed regions for κ_f and κ_V are shown in Figure 1.9 separately for different Higgs boson decay modes [48,56]. Most of the decay channels are only sensitive to the magnitude of the couplings. Only the decay $\mathrm{H} \to \gamma\gamma$ is sensitive to their relative phase due to the interference between the diagrams with W bosons or top quarks in the loops. While the standard model prediction is strongly favored by the measurements, the solution κ_f is not yet excluded. The constraints in Figure 1.9 assume only standard

[3] Calculated for proton-proton collisions at $\sqrt{s} = 8\,\mathrm{TeV}$ (see next chapter).



(a) (b)

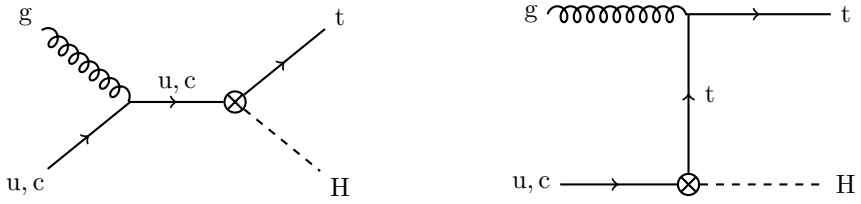
Figure 1.9.: Constraints on κ_f and κ_V from ATLAS [56] (a) and CMS measurements [48] (b). The allowed regions for the values are shown separately for the different Higgs boson decay modes.

model contribution to the total width of the Higgs boson. As shown in [57], when allowing for BSM contributions in the Higgs boson decays, the $\kappa_f = -1$ scenario is still tolerated. For $\kappa_f = -1$ the tHq production cross section would be 13 times enhanced, i.e.

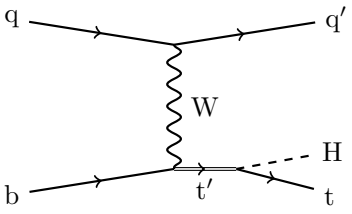
$$\sigma(pp \to tHq)_{\kappa_f=-1} = 233.8^{+4.6}_{-0.0}\,\text{fb}\,. \tag{1.9}$$

Another aspect making tHq production even more interesting is its sensitivity to new physics. At high energy scales diagrams with flavor-changing neutral currents (FCNC) involving top quarks and Higgs bosons with tHu or tHc vertices could contribute to the tHq cross section [58]. Two representative Feynman diagrams for this process, that are suppressed in the SM, are depicted in Figure 1.10(a). A possible enhancement of the tHq cross section could also arise from the production of a hypothetical heavy top partner t′. The t′ decays via t′ → tH, as depicted in Figure 1.10(b), and would mimic the standard model tHq production signature [59].

Direct searches for tHq production are carried out in the H → γγ [60], H → b$\bar{\text{b}}$ [61] and the H → WW decay channels [62] by the CMS collaboration. The H → γγ analysis observes an upper limit of 4.1 times the predicted cross section with $\kappa_f = -1$ that coincides with the expected upper limit. The H → WW analysis reports an observed (expected) upper limit of 6.7 (5.0) times the expectation with $\kappa_f = -1$. It should be noted that the H → γγ analysis exploits an additional cross section enhancement by a factor of 2.4 due to the interference of top quark and W boson loops in the decay. The search for tHq production in the H → b$\bar{\text{b}}$ decay channel is one of the main objectives of this thesis and presented in Chapter 6.



(a) FCNC processes with tHu or tHc vertices



(b) Heavy top quark partner t′

Figure 1.10.: Possible Feynman diagrams beyond the standard model contributing to tHq production. The crossed out vertices are in the two diagrams in (a) are suppressed in the standard model. The double line in (b) indicates a hypothetical top quark partner that is predicted in some new physics models.

2. Experimental setup

The nature surrounding us is primarily comprised of up quarks, down quarks and electrons. Only from these three first-generation fermions matter is built. On earth, heavier particles from the second or third generation are produced naturally only in high energetic collisions in the atmosphere. However, all second and third generation particles decay eventually to their partners from the first generation. To test the full set of elementary particles of the Standard Model and to possibly find new particles large machines are necessary providing high energetic collisions under laboratory conditions.

According to Einstein's mass-energy equivalence $E = m \cdot c^2$ [63] large center-of-mass energies $\sqrt{s}$ are needed to produce heavy particles as the Higgs boson or the top quark. In modern colliders particles are accelerated to unprecedented energies. The collider with the largest center-of-mass energy is the Large Hadron Collider (LHC) [64] at the European Organization for Nuclear Research (CERN) center in Geneva, Switzerland. For the most part of the year, the LHC is devoted to provide proton-proton collisions at a high center-of-mass energy.

To record the particles produced in such proton-proton collisions dedicated detectors are needed. The data analyzed in this thesis has been recorded by the CMS detector, one of the most complex apparatuses built by mankind.

The following sections give an overview of the main parts of the LHC acceleration chain as well as a detailed description of the CMS detector. In addition, the computing structure responsible for processing and distributing the huge amount of provided data is addressed.

2.1. The Large Hadron Collider

The LHC main ring has been installed in a 26.7 km long ring tunnel, which lies $45\,\mathrm{m} - 170\,\mathrm{m}$ below surface. For colliding protons two separate systems are needed for directing two counter rotating proton beams around the ring. To save space the two beam pipes share a common magnetic and cooling system. The ring itself is not a perfect circle, but consists of eight straight sections and eight arcs as illustrated in Figure 2.1. The protons are guided by 1232 superconducting dipole magnets providing a magnetic field up to 8.33 T through the ring. In addition, 392 quadrupole magnets govern the focusing of the beams.

Before the protons enter the LHC main ring, they are pre-accelerated step-by-step. The acceleration chain at CERN is schematically illustrated in Figure 2.1. By applying high voltage of 90 kV, protons are extracted from a hydrogen source.

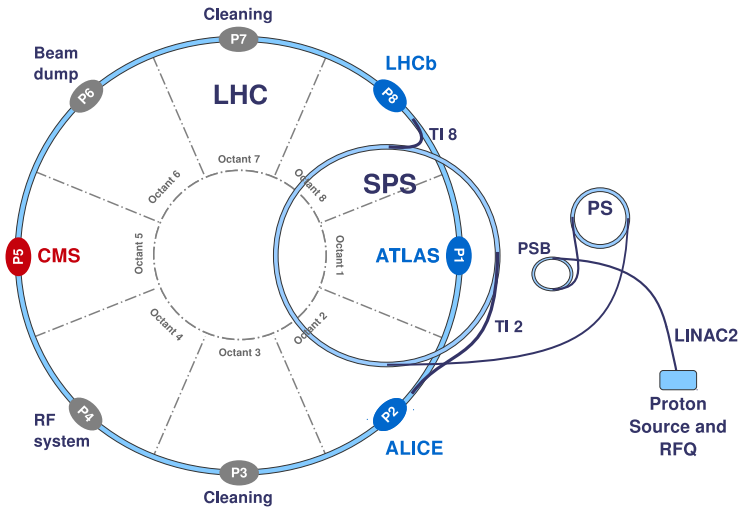


Figure 2.1.: Accelerator chain at CERN, taken from [65]. Before entering the LHC main ring, the protons provided by the proton source are pre-accelerated by the radio frequency quadrupole (RFQ), the LINAC2, the proton synchrotron booster (PSB), the proton synchrotron (PS) and the super proton synchrotron (SPS). Two counter rotating proton beams are accomplished by two different transfers lines TI2 and TI8. At eight possible collision points P1−P8 the protons beams can be crossed. The four main detectors at the LHC are ATLAS at Point 1 (P1), ALICE at P2, CMS at P5, and LHCb at P8. The drawing is not to scale.

At first the protons enter the radio frequency quadrupole, which carries out three tasks. By using resonant microwave cavities it accelerates the protons further, focuses them, and groups the protons into bunches. Subsequently, the bunches enter the LINAC2, a linear accelerator. Hereafter, the protons have an energy of 50 MeV. Their energy is further increased to 450 GeV by the proton synchrotron (PS) and the super proton synchrotron (SPS). After this step the proton bunches are divided into two beams. Via two different transfer lines the two beams are brought in opposite directions to the main ring, where the proton bunches reach their final energy. In 2012 the final energy has been 4 TeV per beam, leading to collisions with a center-of-mass energy of $\sqrt{s} = 8\,\text{TeV}$.

The LHC main ring provides eight points, where the proton beams can be brought to collisions. The instantaneous luminosity L is the measure of the rate of data that is produced. The larger L is, the more collisions can be recorded, and the greater the chance that something new is observed. For two colliding proton bunches a and b L is defined as

$$L = f \cdot \frac{N_a N_b}{4\pi\sigma_x\sigma_y}\,, \tag{2.1}$$

where N_a and N_b are the number of protons per bunch and f denotes the beam rotation frequency. The transverse sizes of both bunches σ_x and σ_y are simplified assuming Gaussian shapes. Moreover, in Equation (2.1) the crossing angle of the beams is not taken into account. Considering L, for a given process p the interaction

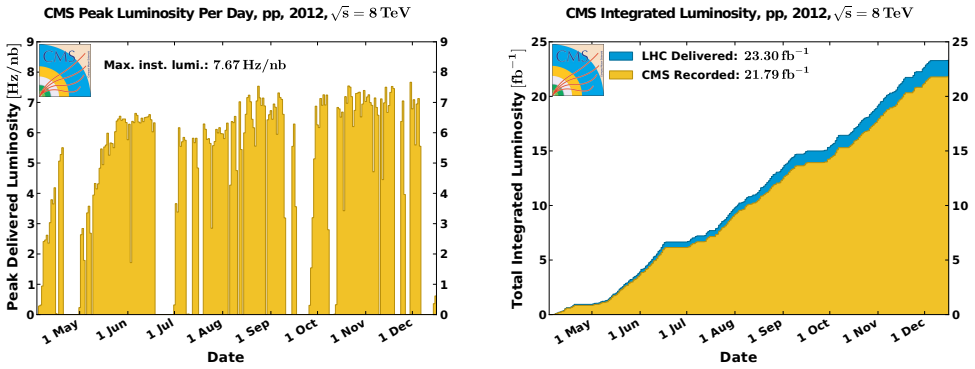


Figure 2.2.: Luminosity profile at the LHC in 2012 for proton-proton collisions at a center-of-mass energy of $\sqrt{s} = 8$ TeV, taken from [66]. The peak luminosity per day delivered from the LHC is depicted (left). Furthermore, the integrated luminosity over time (right) is given. The blue histogram shows the integrated luminosity provided by the LHC, and the yellow histogram indicates L_{int} recorded by CMS. The recorded integrated luminosity is corrected for downtime of the CMS trigger system, which is introduced in Section 2.2.4.

rate $\dot{N}_p$ is given via

$$\dot{N}_p = \sigma_p \cdot L \,. \tag{2.2}$$

Here, σ_p indicates the production cross section of the process. Figure 2.2 shows the luminosity profile of the full data taking period at $\sqrt{s} = 8\,\text{TeV}$ in the year 2012. In the diagram on the left the peak luminosity per day is shown. The maximum luminosity seen at the CMS detector of $7.7\,\text{Hz/nb}$ is the world record for hadron colliders. The diagram on the right depicts the increasing integrated luminosity $L_{\text{int}} = \int L\,\mathrm{d}t$ over time. In 2012 data corresponding to an integrated luminosity of $23.3\,\text{fb}^{-1}$ has been provided by the LHC. Due to dead time of the CMS trigger system described in Section 2.2.4 and other problems during operations the amount of stored data is slightly lower. After a two-year shutdown, the LHC will restart operation with higher energies than ever before in 2015. Proton-proton collision with center-of-mass energies of $\sqrt{s} = 13\,\text{TeV}$ and $\sqrt{s} = 14\,\text{TeV}$ are scheduled.

In the end, it is the debris from the collisions that is tracked in the four main detectors at LHC: ALICE, ATLAS, LHCb, and CMS. The ALICE (A Large Ion Collider Experiment) detector [67] is tailored towards recording heavy ion collision data and is located at P1. The LHCb detector [68] at P8 is specialized to study rare decays of hadrons containing b and c quarks. The two multipurpose detectors ATLAS (A Toroidal LHC ApparatuS) [69] and CMS (Compact Muon Solenoid), located at P1 and P5, respectively, are designed to probe the standard model with high precision and to search for new physics beyond the standard model. In the following section, the focus lies on the CMS experiment and its different subdetectors.

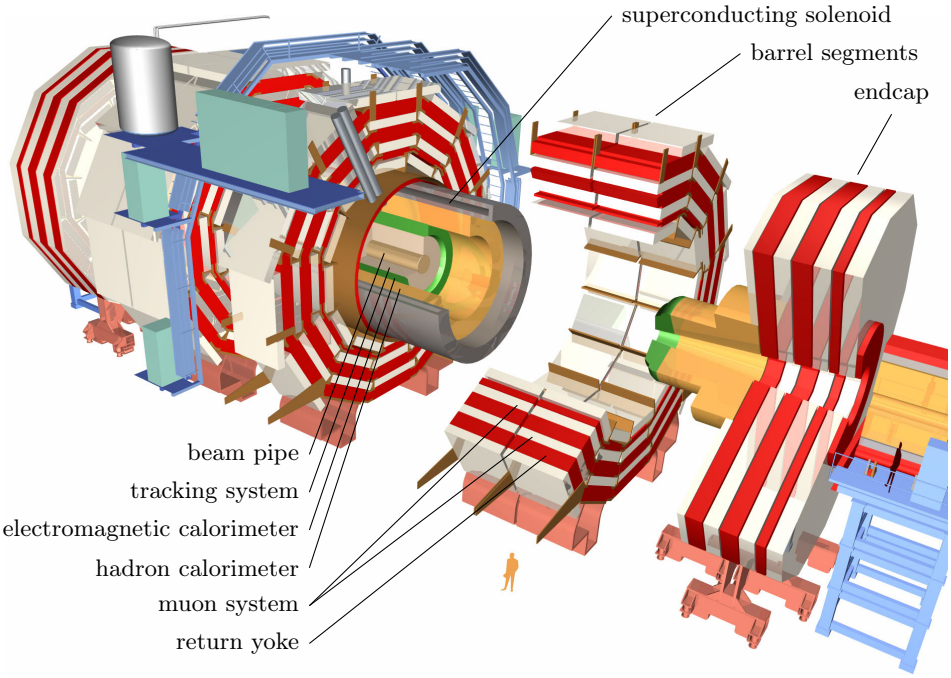


Figure 2.3.: Illustrative overview of CMS detector layout adapted from [65, 70]. The beam pipe is surrounded successively by the tracking system, the electromagnetic calorimeter and hadron calorimeter. These parts are located in the volume of the superconducting solenoid. The gas-ionizing muon chambers are found outside the solenoid embedded in the steel return yoke. The modular structure with several barrel segments and two endcaps facilitates maintenance and inspection of the detector parts.

2.2. The Compact Muon Solenoid detector

The Compact Muon Solenoid (CMS) apparatus is located in a cavern 100 m below surface at Point 5. Designed to detect the full ensemble of secondary objects arising in proton-proton collisions, the detector with a length of 21.6 m, a diameter of 14.6 m is built hermetically around the beam pipe. The dimensions make the CMS experiment more compact compared to its counterpart, the ATLAS detector. However, with a weight of about 14000 t the CMS detector is twice as heavy as ATLAS.

An overview of the detector's characteristic onion-like layout is shown in Figure 2.3. Successively, the tracking system, the electromagnetic calorimeter, the hadron calorimeter and the superconducting solenoid encompass the beam pipe. The solenoid with an internal diameter of 6 m provides a homogeneous magnetic field of 3.8 T parallel to the beam pipe. The muon system is located outside the solenoid embedded in the steel return yoke.

The CMS experiment is designed to cover large phase spaces. Furthermore, the different subdetector systems aim to identify muons with an excellent momentum

resolution, to reconstruct charged particles with an excellent momentum and position resolution allowing for b tagging (see next chapter), as well as an excellent electromagnetic energy resolution.

Conventionally, the CMS detector is described by a right-handed coordinate system centered at the nominal interaction point. The x and the y axes are directed to the center of the LHC main ring and to the sky, respectively. Consequently, the z axis points counterclockwise along the main ring. The azimuthal angle ϕ and the radius r are measured in the $x - y$ plane. The polar angle θ is given with respect to the z axis. Geometrical positions are described with z, r and ϕ. Generally, for angles with respect to the beam pipe the pseudorapidity $\eta = -\ln(\tan\theta/2)$ is used.

In the following insights into the different subdetectors are provided and the CMS trigger system and the computing structure are introduced. A much more detailed description of the different parts of the CMS experiment is given in [71].

2.2.1. Tracking system

The innermost subdetector surrounding the beam pipe is the tracking system [72] with a length of 5.8 m and a diameter of 2.5 m. Its purpose is the accurate recording of the bent trajectories of charged particles due to the magnetic field. This allows for the reconstruction of the particles' momenta as well as the sign of their electromagnetic charge. Additionally, high precision trajectories facilitate the reconstruction of vertices, as explained in Section 3.2.1.

During nominal LHC operation, in the order of 1000 charged particles per collision per bunch crossing are expected. Therefore, a high granularity and fast response time is required. On the other hand, as the tracker constitutes the innermost layer, it is subject to severe radiation. To address all these requirements, the components of choice are semiconducting silicon detectors. Traversing charged particles cause electron-hole pairs in these detectors, and the resulting electric signals can be measured. The read-out is performed by dedicated radiation hard sensors.

Figure 2.4 gives an overview of the tracking system. It consists of two subsystems, a silicon pixel and a silicon strip detector, covering in total the region with $|\eta| < 2.5$. The support tube environing the tracking system, which ensures the detector's working temperature of $-20\,^{\circ}\mathrm{C}$, is not displayed.

Silicon pixel detector

The silicon pixel tracker has an active area of $1\,\mathrm{m}^2$. It consists of three barrel layers with a length of 53 cm and two endcap disks. The 1440 modules contain 66 million pixels providing high granularity. Each pixel has a size of $100 \times 150\,\mu\mathrm{m}$. Up to three space points per charged particle are obtained at radii of 4.4 cm, 7.3 cm and 10.2 cm, with a resolution of $10\,\mu\mathrm{m}$ in the $r - \phi$ plane and $15\,\mu\mathrm{m}$ in z direction.

The high granularity allows for the reconstruction of secondary vertices that are needed for the identification of jets stemming from b quarks (see Section 3.2.5).

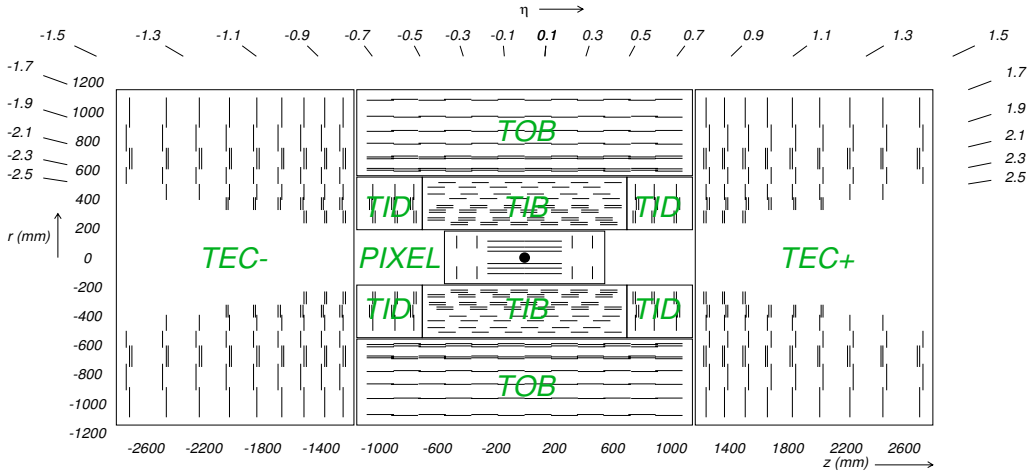


Figure 2.4.: Overview of the CMS tracking system, taken from [71]. The pixel detector environs the interaction point (black dot). The strip detector is partitioned into tracker inner barrel (TIB), tracker outer barrel (TOB), tracker inner disk (TID) and tracker endcaps (TEC). The detector modules (single line) and stereo modules (double lines) are shown.

Silicon strip detector

The strip detector system with 15148 modules encloses the pixel detector at radii between 20 cm and 116 cm from the beam pipe. In this region, the particle flux is reduced, so less expensive silicon strips are applied. Overall there are around 9.6 million readout channels yielding an active area of 200 m^2. The component itself is divided into tracker inner barrel, tracker outer barrel, tracker inner disk and tracker endcaps. The tracker inner barrel provides four layers of silicon sensors and the tracker outer barrel six layers. In total, the barrel segments equip each charged particle with up to ten $r - \phi$ measurements with a single point resolution between 30 µm and 50 µm. The tracker inner disks consist of three layers and the tracker endcaps are equipped with additional nine layers. So, the endcap part of the silicon strip detector adds up to 12 additional $z - \phi$ measurements with a resolution of 30 µm.

As depicted in Figure 2.4 stereo modules are added in all detector parts to provide measurements of the missing coordinates, i.e. z in the barrel and r in the endcap parts. These modules consist of two tilted strip sensors aligned back-to-back. The resolution for the additional coordinates ranges between 230 µm and 530 µm.

2.2.2. Calorimetry system

The second layer environing the tracker is the calorimetry system, that aims to absorb electrons, photons and hadrons in order to measure their energies. It is built up by two subdetectors: The electromagnetic calorimeter (ECAL) [73, 74] and the

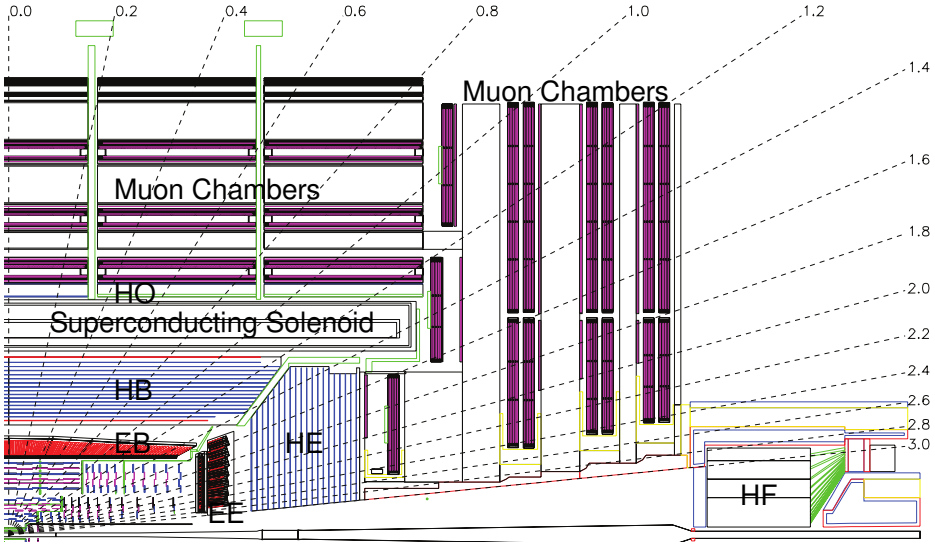


Figure 2.5.: CMS calorimetry and muon systems, adapted from [71] with modification. The sketch represents the longitudinal view in one quarter of the detector. The tracking system is encompassed by the calorimetry system, partitioned into electromagnetic barrel (EB), hadron barrel (HB) and hadron endcaps (HE). The electromagnetic endcaps (EE) are shown together with the electromagnetic preshower detector. Outside the superconduction solenoid the hadron outer calorimeter (HO) and the muon system embedded in the steel return yoke are located. The hadron forward calorimeter ensures energy measurements of particles with high pseudorapidities. For a detailed view of the muon system see Figure 2.6.

hadron calorimeter (HCAL) [75]. An overview of the layout is given in Figure 2.5. The energies of electrons, positrons and photons are measured in the ECAL, and the energies of neutral and charged hadrons in the HCAL. Precise knowledge of the particle energies is important for the reconstruction of jets and the missing transverse energy explained in the next chapter.

Typically, the length of the absorber material in the ECAL is given in units of X_0, which is the material specific radiation length of electrons. In the HCAL with the hadronic interaction length λ a similar quantity is chosen. Two different techniques are exerted. For the HCAL alternating samples of absorber material, decelerating the particles gradually to their complete absorption, and scintillator material are used. In the ECAL a material is chosen that acts as scintillator and absorber simultaneously.

Electromagnetic calorimeter

The desired homogeneous structure is reached with lead tungstate ($PbWO_4$) crystals. Though lead tungstate is very dense ($\rho = 8.3\,\mathrm{g/cm^3}$), it is still transparent for visible light. Therefore, it can act as absorber and as scintillator material simultaneously. The electromagnetic barrel provides 61200 of these crystals and the

endcaps 7324 crystals.

Traversing electromagnetically interacting particles lose energy due to bremsstrahlung and electron-positron pair production. The emitted scintillation light in the deceleration process is a direct measure for the energy of the incoming particles. The advantages of the lead tungstate crystals are the short radiation length of $X_0 = 0.89\,\text{cm}$ and the small transverse dimension of the cascades. Hence, by using these crystals a fine granularity can be achieved in the detector. Moreover, about 80% of all photons are emitted within 25 ns, thus fast measurements are possible.

The length of the crystals corresponds to $25.8 \cdot X_0$ and $24.7 \cdot X_0$ in the barrel and endcap segments, respectively. The barrel segments cover the region of $|\eta| < 1.479$, and the front face of each crystal is $22 \times 22\,\text{mm}^2$. The crystals in the endcaps, covering the forward region up to $|\eta| < 3.0$, have a bigger front face of $28.6 \times 28.6\,\text{mm}^2$. For the readout of the photons, each crystal is equipped with avalanche photodiodes (barrel) or vacuum phototriodes (endcap). Additional measurements from the preshower detector with an acceptance of $1.653 < |\eta| < 2.6$ help to distinguish photon pairs from π^0 hadron decays from prompt photons.

The ECAL's relative energy resolution $\sigma^{\text{rel}}_{\text{ECAL}}$ can be parameterized via

$$\left(\sigma^{\text{rel}}_{\text{ECAL}}\right)^2 = \left(\frac{2.8\%}{\sqrt{E}}\right)^2 + \left(\frac{12\%}{E}\right)^2 + (0.3\%)^2\,, \tag{2.3}$$

where E denotes the particle's energy measured in GeV. The first term on the right side of this formula arises due to stochastic event-by-event differences and the second summand proportional to $1/E^2$ is the noise term. The numeric values have been measured with electron test beams [76].

Hadron calorimeter

The hadron calorimeter is built up by alternating samples of non-magnetic brass, serving as absorber material, and plastic scintillator tiles. The interaction length in the brass samples is $\lambda_I = 16.42\,\text{cm}$ and around 18 times larger compared to X_0. As a consequence, hadrons deposit most of their energy in the hadron calorimeter. The effective thickness of the hadron barrel is $5.82 \cdot \lambda_I$, and the hadron endcap has a thickness of roughly $10 \cdot \lambda_I$.

Traversing hadrons cause hadronic showers due to inelastic scattering with the material. The deceleration happens mostly in the absorber material, and only a small fraction of scintillation light can be detected in the plastic tiles. This light is transported via wavelength shifting optical fibers to hybrid photo-diodes. The total energy is then estimated based on the recorded fraction. Consequently, the HCAL has a worse energy resolution compared to the homogeneous layout in the ECAL.

The hadron barrel segments cover the region $|\eta| < 1.3$ and each HCAL cell matches to 5×5 ECAL crystals. The hadron endcap system has an acceptance in the region of $1.3 < |\eta| < 3.0$. Here, fewer crystals are mapped to each HCAL cell. For the sake of completeness, it should be noted that the hadron outer calorimeter is installed outside the superconducting solenoid. It uses the solenoid as additional

active detector material and has the purpose to measure the hadrons not stopped by the hadron barrel. However, due to its worse energy resolution, information from the hadron outer detector is not used in this thesis.

In the more forward region ($|\eta| > 3.0$) the hadron forward calorimeter is installed. As in this region the particle flux is high, steel absorber and quartz fibers constitute the sampling structure. The Cherenkov light emitted by particle showers in the quartz fibers is measured via photomultipliers. Only due to the energy deposits in this forward calorimetry the proper reconstruction of forward jets is possible, which is of importance in this thesis.

Similar to the ECAL the relative energy resolution $\sigma^{\text{rel}}_{\text{HCAL}}$ for the HCAL can be parameterized via

$$\left(\sigma^{\text{rel}}_{\text{HCAL}}\right)^2 = \left(\frac{S}{\sqrt{E}}\right)^2 + C^2 \,. \tag{2.4}$$

Here, E is the particle's energy measured in GeV. The parameter S represents the stochastic term and C is a constant. The values have been calibrated with muon cosmic rays and several test beams and read $S = 0.847\,\sqrt{\text{GeV}}$ and $C = 0.074$ for the hadron endcap and barrel, and $S = 1.98\,\sqrt{\text{GeV}}$ and $C = 0.09$ for the hadron forward detector [77].

2.2.3. Muon system

As already the name of the experiment implies, the muon system [79] plays an important role at the CMS detector. The muon detection system comprises nearly 1 million electronic channels and is dedicated to identifying muons and to providing additional measurements of their kinematics. Muons are the only charged particles causing hits in the tracking system, but not being brought to halt in the calorimetry system. To detect them, three different kinds of gas detectors are embedded in the iron return yoke, that provides a magnetic field of 2 T. The layout of the muon system is depicted in Figure 2.6.

In the barrel segments covering the region with $|\eta| < 1.2$ in total 250 aluminum drift tube chambers (DT) are arranged in four layers. These chambers are filled with an Ar/CO_2 gas mixture. The endcap segments with an acceptance of up to a region of $|\eta| < 2.4$ use cathode strip chambers (CSC). The chambers contain a mixture of $\text{Ar}/\text{CO}_2/\text{CF}_4$ and are arranged in four layers. The advantage of CSC with respect to DT is that the former can cope with higher particle rates and higher magnetic fields. Additionally, the region with $|\eta| < 1.6$ is equipped with resistive plate chambers (RPC) that provide independent fast trigger information. This is necessary due to the high muon rate in this central region.

2.2.4. Trigger system, JSON files and computing structure

During nominal LHC operation over one billion proton-proton interactions occur each second. The huge amount of data produced in these collisions is impossible to store in its entirety. One of the main challenges for the CMS experiment is to

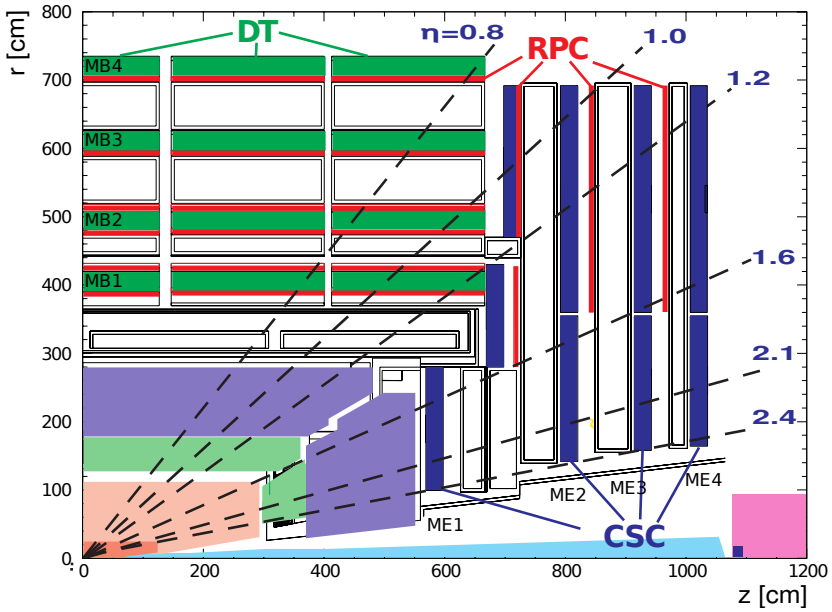


Figure 2.6.: Schematic overview of the muon system, taken from [78]. The sketch shows the longitudinal view of one quarter of the detector. Three different types of gas detectors are used. Drift tubes (DT) are installed in the four barrel muon stations, MB 1 to MB 4. The endcap muons stations, ME 1 to ME 4, are equipped with cathode strip chambers (CSC). In all stations additional resistive plate chambers (RPC) are embedded in the region $|\eta| < 1.6$, where a high muon rate is expected.

reduce the data rate of $\approx 20\,\text{MHz}$ to a reasonable level without losing interesting physics events. To attack the problem the CMS trigger system [80, 81] incorporates a two step reduction using hardware and software triggers.

The level-1 trigger (L1) of the CMS experiment consists of programmable hardware and is required to reduce the data rate to 0.1 MHz. While the full detector data is buffered, the level-1 trigger logically interprets the information from the electromagnetic and hadron calorimeters as well as from the three types of technologies of muon detectors. Events with certain signatures in the detector possibly stemming from interesting physics trigger a positive L1 decision. Only for these events the event data is read out and passed to the next level.

The second step of data reduction is accomplished with the high-level trigger (HLT) that is embedded in the computing farm at Point 5. The so-called Builder Network calipers information from about 650 data sources and reconstructs the events via dedicated algorithms. The data rate is reduced to less than 400 Hz by applying requirements on the information of the reconstructed events. The step-wise data reduction is schematically depicted in Figure 2.7.

Events passing the HLT requirements are stored to disk divided into several primary datasets. In order to provide the recorded data to analysts all over the world, the CMS collaboration adopted the structure of the LHC computing grid [82].

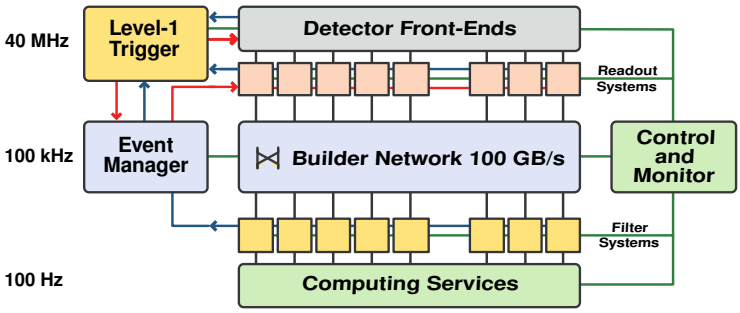


Figure 2.7.: Architecture of the CMS trigger and data acquisition system, taken from [71]. The two stages applied for event rate reduction from 40 MHz to 100 Hz are depicted. First, only events passing the hardware-driven Level-1 trigger requirements are forwarded to the readout systems. For these events the event builder combines the available detector information. In the end, the software-driven HLT filter system decides whether an event is stored or not.

The grid is organized in a tier-based manner with two Tier-0 centers, several Tier-1 sites, and numerous Tier-2 and Tier-3 facilities, as depicted in Figure 2.8. The un-worked detector information of the primary datasets (RAW datasets) is stored at the two Tier-0 facilities in Geneva and Budapest. Smaller RECO datasets are obtained after first calibration and reconstruction steps. Both, RECO and RAW datasets are transferred to at least one Tier-1 center as a backup. At the Tier-1 centers, that are national computing facilities like at the Karlsruhe Institute of Technology (KIT), AOD datasets are created that are a subset of the RAW data with sufficient information for most analyses. The AOD datasets are distributed to several Tier-2 and Tier-3 centers, where the user can access them.

The data management service responsible for transferring the huge datasets between the different CMS computing centers is PhEDEx [83]. PhEDEx also monitors and logs the data transfers, so an accurate performance is crucial. For debugging and testing the PhEDEx service, the LifeCycle agent was developed, that can simulate any request within an artificial architecture of Tier centers. As part of this thesis, modules for this LifeCycle agent have been developed, that automatically perform sanity checks and thus help scrutinizing the functionality of the PhEDEx software [84].

During data taking the conditions can change, and issues in subdetectors can spoil the recorded events. To supervise whether all components of the detector worked properly, the CMS collaborations has the centralized *Data Quality Management* group [85]. This group publishes lists of good runs, for which the detector has operated flawlessly and the conditions have been stable. The list is referred to as *JSON file* because it is stored in this format and exclusively those runs are used in this thesis.

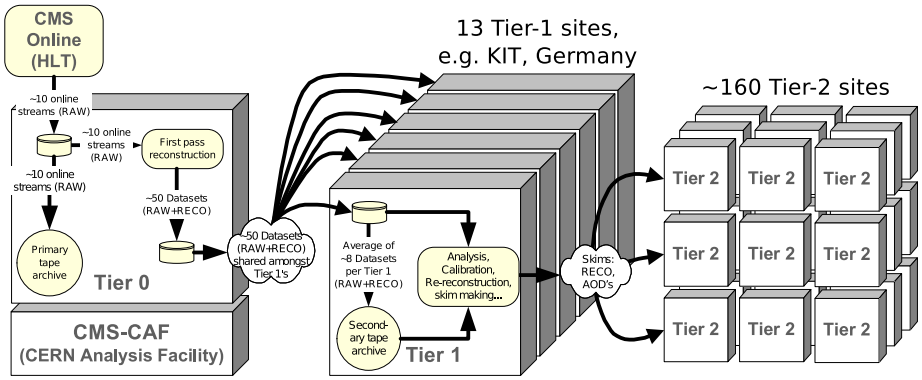


Figure 2.8.: Architecture of the CMS computing grid, taken from [71] with modifications. The RAW data format is stored at the Tier-0 sites at CERN and the Wigner research center for physics in Budapest. The 13 Tier-1 to date have large storage capacities and are responsible for save-keeping the RAW and RECO datasets. At the numerous Tier-2 sites and Tier-3 centers the AOD datasets for the analyses are saved.

3. Generation, simulation and reconstruction of events

In the high luminosity environment at the LHC, proton-proton collisions cause a vast amount of detector responses in the CMS experiment. Recording this data with the subdetectors as explained in the previous chapter is only one side of the coin. Advanced methods are needed to bring electric signals in the CMS detector and predictions from the SM down to a common denominator. Only this way, the confrontation of the experimental data with the underlying theory is possible.

Predicting the responses from particles in the complex detector environment is an analytically non-solvable problem. That is why Monte Carlo (MC) techniques are applied, which are based on random sampling. Collisions and their responses in the detector are produced stochastically according to the expected probabilities from theory. Therefore, also a precise simulation of the detector itself is needed.

To confront detector signatures from data and MC simulation, they have to be interfered with a common reconstruction of the physics objects, like electrons or jets, in each event. The hits and the energy deposits are interpreted by the Particle Flow (PF) algorithm developed within the CMS collaboration. The resulting objects serve as input for higher order physics objects, i.e. jets and missing transverse energy.

In this chapter the different steps in the simulation of collisions are described. Moreover, the several MC generators as well as the detector simulation used in this thesis are presented. Another section is dedicated to the reconstruction of physics objects via the Particle Flow algorithm. In particular, this chapter also introduces different jet clustering algorithms that are important for the further analysis.

3.1. Generation of events

The processes in proton-proton collisions obey quantum mechanics and are therefore of probabilistic nature. MC methods provide numerical solutions to non-deterministic problems, hence they can be applied for the simulation of collision events. Distinct requirements on production and decay can be enforced to generate rare physical processes with a reasonable amount of events. This is of great importance, since the investigated signal production modes as well as most of the background processes are expected to have low cross sections.

To produce such complex processes MC generators rely on a factorization of the simulation. First, the initial hard interaction usually containing only a few initial and final state particles is simulated. The resulting partons are further handed to

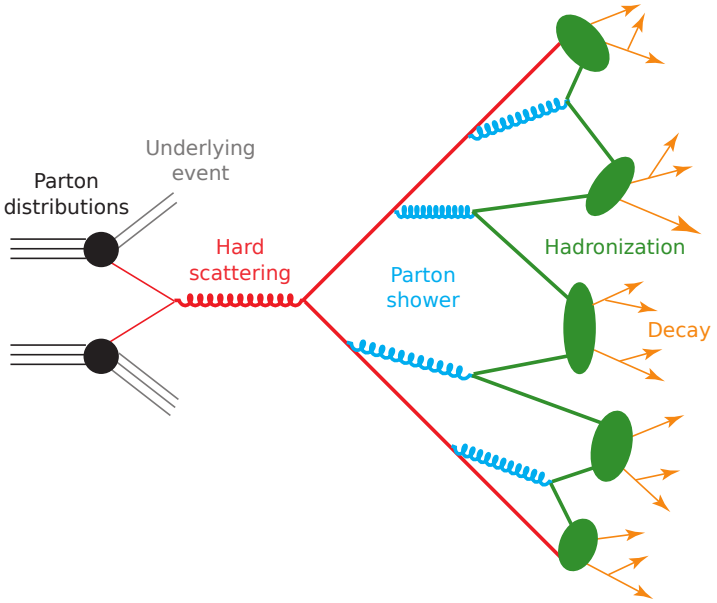


Figure 3.1.: Scheme of the Monte Carlo event generation process adopted from [86]. The proton characteristics are defined by parton distribution functions. The initial and final state partons participating in the hard scattering process are passed to the parton shower step, where soft radiations are generated. Colorless hadrons are formed in the hadronization step. Finally, the decay of unstable particles is simulated. Beyond, the remnants of the protons can interact further. The contributions from this process, known as underlying event, are simulated as well.

the parton shower step, that produces soft radiations. Eventually, the hadronization of colored objects and the decay of unstable particles are simulated.

While the details of the different steps are provided in the following, Figure 3.1 gives an illustrative overview of the event generation.

Hard scattering process

The hard scatting part of interesting processes usually occurs with high energy transitions and thus small values for α_s. Therefore, perturbative calculations are valid for computing the production cross section for a specific process.

To start with, the colliding protons are characterized with parton distribution functions (PDF), that determine the momentum fractions of the different partons (see also Section 1.1.3). These PDFs are measured in deep-inelastic scattering experiments. Different collaborations provide parameterizations for the PDFs that are used in the MC generators. Exemplarily, the CTEQ61 parameterization is shown in Figure 3.2. This configuration is adopted for the majority of MC samples used in this thesis.

Using the matrix element (ME) method, the cross sections are calculated based

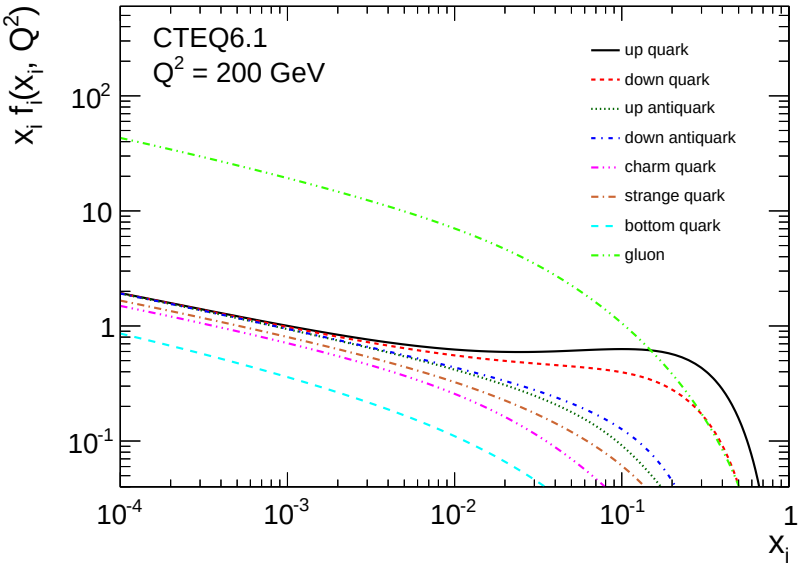


Figure 3.2.: Exemplary CTEQ61 proton PDF for gluons and quarks. This PDF is used for the majority of MC samples in thesis. The values shown at a scale of $\mathcal{Q}^2 = 200\,\text{GeV}$ are provided by the Durham HepData Project [87].

on the evaluation of all relevant Feynman diagrams. The interference between two diagrams is already taken into account. Also, ME calculations account for specific process kinematics stemming from spin and helicity effects for instance. Therefore, the decays of resonances with spin, e.g. $\mathrm{t} \to \mathrm{Wb} \to \mathrm{qqb}$, are already simulated in this step. The secondary objects are handed over to the parton shower process. The decay of scalar particles, e.g. $\mathrm{H} \to \mathrm{b\bar{b}}$, is also left to the parton shower.

Parton shower

In this step the possible radiation of accelerated color charges is simulated. Depending on which part of the process these radiations occur, they are referred to as initial state radiation (ISR) and final state radiation (FSR).

There are two popular approaches how to deal with prediction of these radiations. On the one hand, the ME method can already incorporate additional radiations in the calculations. However, this is limited by the increasing complexity of the Feynman diagrams, when taking higher orders of perturbation calculation into account. Furthermore, these calculations are only valid for small values of α_s.

Another approach is to simulate random splittings of one parton into two new particles with the parton shower (PS) method. These successive splittings are parameterized with the Altarelli-Parisi splitting functions [88]. The exponentiation using Poisson statistics leads to the *Sudakov form factor*[1], which is the probability

[1]The Sudakov form factor is closely correlated to the scale evolution of PDFs described by the Dokshitzer-Gibov-Lipatov-Altarelli-Parisi (DGLAP) equations [88–90]

that no emission takes place. In this approach simplified models are used for the kinematics in the interactions.

The two procedures are often combined in event generators. For instance, the ME is used down to a process dependent cut-off energy scale to generate high energetic radiations. The low energetic radiations are then simulated with the parton shower method. A careful combination is needed to avoid double counting, hence matching algorithms have been developed. Most prominent are the CKKW [91], with transverse momenta matching, and MLM [92], based on angular matching.

Hadronization and decay

The final products from the parton step consist of elementary particles like gluons and quarks. Objects carrying color charge obey quantum chromodynamics. Hence, they cannot occur freely due to confinement and hadronize into colorless bound states. However, a perturbative calculation is not valid anymore, as the energy transfer is very low at this point. The simulation of the hadronization of the particles into colorless bound states has to rely merely on phenomenological models.

One prominent representative is the Lund fragmentation model [93]. Here, color-flux string tubes describe the connection between colored particles depending on their distance. Iterative break-ups of the color-flux string tubes, each creating a $q\bar{q}$ pair, simulate the forming of neutral states. These break-ups continue until the energy is too low to create new quark-antiquark pairs.

An alternative approach is the cluster hadronization model [94]. This model is based on the idea that color lines connect pairs of partons after the parton shower. Each gluon emission gives rise to a new color line. In the end, all gluons are forced to decay into a $q\bar{q}$ pair. The forming of colorless bound states is realized by building *proto-hadrons* out of the connected color lines. These proto-hadrons decay into the observed final-state hadrons according to a simplified phase-space scheme.

In the end of this step the decay of the unstable hadrons according to the known branching ratios is simulated.

Pile-up and underlying event

As indicated in Figure 3.1 the proton remnants, which do not contribute to the hard process, can interact further. The products of these interactions are also recorded and assigned to the same event. This is referred to as underlying event.

On the other hand, due to the high instantaneous luminosity provided by the LHC, in each proton bunch crossing several proton-proton collisions take place. These additional interactions are called pile-up (PU) events, and can be categorized in two ways. In-time PU accounts for extra proton-proton collisions in the same bunch crossing. Due to the finite response of the detector elements, also hits from bunch crossings before or after, are recorded. This is called out-of-time PU.

Most MC generators provide methods to simulate an admixture of both, pile-up and underlying event. That way, the data in the high luminosity environment at

the LHC can be described properly.

3.1.1. Monte Carlo generators

The two analyses presented in this thesis use a wide range of different MC software packages for the simulation of signal and background processes. All of them have advantages, that are exploited for specific production modes. Thus, the different generators and their main features are presented in the following.

Pythia 6.4

The powerful PYTHIA 6.4 package [95] is a multi-purpose generator. It provides full-event simulation for a wide range of different processes for SM and BSM. The hard scattering part is calculated via the ME at LO. In the parton shower step the advantage from the parton shower method is used for soft QCD radiations, that are not possible with the ME method. For the hadronization step the Lund model is applied and many free parameters can be adjusted to allow for a solid description of data. The parameter set used in the analyses is referred to as Z2 tune [96]. In the PYTHIA 6.4 package also the generation of underlying event contributions is provided.

Due to the advantages of the parton shower, other event generators, that simulate the hard interaction at a higher order, are often interfaced with PYTHIA 6.4.

HERWIG++

HERWIG++ [97] serves as an alternative multi-purpose generator, also providing the full-event simulation for a vast number of SM and BSM processes. The cross sections are calculated at NLO. Similar to PYTHIA 6.4, the different routines of HERWIG++ can be interfaced with other generators. The main difference compared to PYTHIA 6.4 is the use of the cluster hadronization model instead of the Lund model to simulate the hadronization step.

MadGraph

The MADGRAPH [98] software, a matrix element generator, calculates all relevant LO Feynman diagrams for a given process. Also, leading-order radiation is provided. The actual event generation is then performed with the MADEVENT package, which does not cover showering or hadronization. Therefore MADGRAPH is usually interfaced with PYTHIA 6.4.

Powheg

The POWHEG package [99, 100] provides NLO precision for several processes. The main feature of POWHEG is to generate the hardest radiation first and then to apply dedicated subtraction techniques when interfacing with LO parton showers,

like HERWIG++ or PYTHIA 6.4. Therefore, the issue of over-counting Feynman diagrams does not occur.

The presented analyses resort to POWHEG for single top production [101] and the associated Higgs production with a vector boson [102].

Tauola

The TAUOLA package [103] provides a precise simulation of τ lepton decays. Especially to account for spin correlation effects TAUOLA is interfaced with MC event generators for the simulation of single top and diboson production.

3.1.2. Detector simulation

The previously summarized steps do not simulate the interactions of the resulting particles with the detector. To compare the MC events with the actual data, it is indispensable to account for the energy loss due to reactions with the detector material or the deflection within the magnetic field for instance. A full simulation of the CMS detector is provided in the GEANT 4 toolkit [104]. It includes a detailed description of the geometry and material budget of the CMS detector. The simulation of bent trajectories can be achieved with high precision. Moreover, the electric signals caused by traversing particles due to hadronic and electromagnetic showering are also modeled as well as the responses from tracker and muon systems. All MC samples used in this thesis are thus processed with the full detector simulation based on GEANT 4.

At this point data and MC are available — and comparable — in form of basic detector responses. To make comparisons of the underlying processes, for both a common reconstruction is needed, as explained in the following.

3.2. Reconstruction of events

Within the CMS collaboration a powerful approach for interpreting electric signals in the detector as physics objects such as electrons or jets has been developed: The Particle Flow algorithm. The idea is to reconstruct and identify each individual particle with an optimized combination of information from the various elements of the CMS detector. A detailed description of the Particle Flow routines is provided in [105]. The commissioning of the algorithm can be found in [106].

In a first step *PFElements* are created by reconstructing tracks from the tracking system and clusters out of the energy deposits in ECAL and HCAL. Corresponding PFElements are then clustered into *PFBlocks*. The blocks are further interpreted as *PFCandidates* in five categories: electrons, muons, photons, charged hadrons and neutral hadrons. Finally, these PFCandidates serve as input for higher level physics objects like jets and missing transverse energy.

In the following the required steps to obtain PFCandidates as well as jets and missing transverse energy are described.

3.2.1. The Particle Flow algorithm

Emitted charged particles leave hits in the tracking system. Their trajectories, or tracks, are essential ingredients to the PF algorithm. The bending radius due to the magnetic field provides information on the sign of the particle's charge as well as its transverse momentum. For the reconstruction the Combinatorial Track Finder (CTF) [107–109] is applied in four steps. First, initial candidates are formed by connecting pairs of hits in two different layers. Each of these seeds are used to propagate trajectory candidates from layer to layer with a combinatorial Kalman filter [110]. The filters take into account that tracks lose energy due to bremsstrahlung or scatter via interaction with the detector material. Additional quality criteria help to reduce the possible combinations, like a χ^2 compatibility test between the predicted trajectories and the hits. The CTF procedure is carried out more than once, and for each run the hits connected to the found tracks with a satisfying quality are cleaned from the list. This way, the optimal set of track candidates is found.

Primary vertices (PV) indicate the origin of an interaction. With a high density of protons in each bunch, there are several of these interactions in each bunch crossing, each producing dozens of tracks. The primary vertex candidates can directly be obtained from the reconstructed tracks using the Adaptive Vertex Fitting (AVF) method [111] — a modification of the Kalman filter. Tracks are weighted according to their χ^2 values from the compatibility test. After the weights are applied, the PV candidates are re-fitted to obtain best possible results. Interesting interactions including Higgs bosons and top quarks give rise to tracks with a large amount of transverse momentum. That is why, in the analysis, the PV candidates are sorted according to the squared sum of p_T of their assigned tracks. The first PV is usually selected for the analysis, while the other PVs are assigned to in-time pile-up.

The other objects acting as PFElements are PFClusters built from the energy deposits in the ECAL and HCAL systems. The clusters are formed in three steps. Initially, cells with energy deposits above twice the cell's noise level serve as seeds. Every seed leads to one cluster in the end. Secondly, topological partners are found by adding adjacent cells. These cells have to exceed the noise threshold as well, i.e. 80 MeV and 300 MeV in the ECAL barrel and endcap, respectively. In the HCAL system the threshold is up to 800 MeV. Finally, the PFClusters are built by iteratively aggregating neighboring cells weighted relatively to their distance to the seed. Here, a cell can belong to more than one cluster, and if so, its energy deposit is shared via a weighting function among the PFClusters. This way, the granularity of the calorimetry is not a limit for PF objects. Further details can be found in [105].

A charged particle traversing the CMS detector usually gives rise to both, hits in the tracker and energy deposits in the calorimetry. Therefore, the different PFElements have to be interpreted and logically connected. A dedicated linking algorithm

unifies corresponding elements into PFBlocks. Assigning a PFElement to two blocks is forbidden to avoid double-counting of energy. To link clusters to tracks, the latter are extrapolated to ECAL and HCAL. In a first step each trajectory is continued from its last measured hit in the tracker to the ECAL's pre-shower. Afterwards, the trajectory is evaluated in the ECAL to the maximum depth of energy deposits assuming a typical electron shower. A further step extrapolates the track to the HCAL at a depth of one interaction length λ, which is characteristic for hadrons. Before the linking is performed, the clusters' boundaries are enhanced by one cell to account for gaps between two cells.

If a track extrapolation lies within the boundaries of a cluster, the two are linked and tagged with a quality value depending on their distance. For connections including ECAL clusters, possible energy deposits due to bremsstrahlung are obtained by linking the tangent of the track to different ECAL clusters. When a cluster from a fine-grained region lies within a cluster in a coarse-grained area, the two are connected. Tracks in the muon system are linked to tracker tracks by a global fit and tagged with a consistency value χ^2.

Out of these blocks the algorithm starts with the final interpretation from PFBlocks to PFCandidates. First, muons are identified from blocks with links to the muon system. Subsequently, electrons as well as neutral hadrons, charged hadrons and photons are reconstructed. After each step, the corresponding PFElements are removed from the PFBlocks.

Since the reconstruction and identification of particles is possible from only a few elements, the PF algorithm turns out to be very powerful for high luminosity collisions at the LHC.

3.2.2. Muon candidates

Muons leave distinctive signatures in the detector, and are rather easy to reconstruct. As aforementioned, connections between tracks from tracker and muon system are tagged with a χ^2 value, that can be used as additional quality requirement on the muon candidate. When the tracks are compatible with each other a *global muon* is reconstructed. The momentum of each global muon candidate is compared with the measurement by only using the tracker information. If the two results coincide within three standard deviations a *PFMuon* is built.

Also the reconstruction of low energetic muons, which possibly do not give rise to hits in the muon chambers, is possible. To do so, tracks are extrapolated to the calorimeter energy deposits, that are consistent with the amount of energy deposited by a minimum ionizing particle.

Consequently, all corresponding tracks are removed and the estimated energy deposits are subtracted from the assigned clusters with an uncertainty of $\pm 100\%$. The estimate comes from cosmic muons measurements and the average deposit is equal to 3 GeV and 0.5 GeV in the HCAL and ECAL, respectively.

3.2.3. Electron candidates

Electrons traversing the detector lose a significant amount of energy already in the tracker by the emission of photons. The photons travel without further deflection to the ECAL and deposit their energies. Due to these tangential-bremsstrahlung effects, the resulting clusters have a characteristic spread in ϕ, which is used for their identification.

On the one hand, an ECAL-driven method searches for the characteristic energy deposits and links them to compatible tracks. Instead of a Kalman filter, a Gaussian sum filter (GSF) [112,113] is used for the track building. This works reasonably well for high-energetic electrons. On the other hand, a tracker-driven method dedicated to low energetic electrons is employed. Here, the blocks corresponding to electron trajectories are searched using multivariate techniques [114].

The candidates from both reconstruction methods are tagged as PFElectrons and the corresponding PFElements are removed. Further details can be found in [115].

3.2.4. Photons and hadrons

After all PFElements forming PFElectrons and PFMuons are subtracted, the remaining elements are assigned to hadrons and photons. Charged hadrons induce hits and energy deposits in the detector. Neutral hadrons and photons do not leave hits in the tracker, and only cause energy deposits in the HCAL or ECAL.

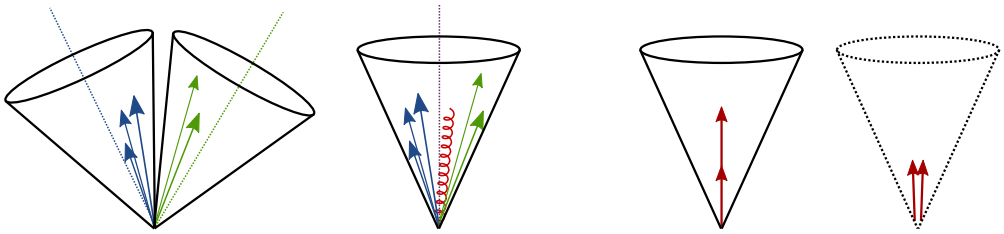
As a first step, the remaining tracks are used for the reconstruction of charged hadrons by linking them to clusters in ECAL and HCAL. The leftover clusters in the ECAL and HCAL are assigned to photons and neutral hadrons. Dedicated recalibration methods (see [105]) are performed correcting the energy of hadrons to avoid double counting, and to account for non-linearities of the calorimetry system.

At that moment, all PF objects are reconstructed and serve as input information for the clustering of jets and the calculation of missing transverse energy, as described in the following.

3.2.5. Jets

The detection of color-charged quarks and gluons plays an important role in the investigated Higgs boson production modes. However, these particles cannot be observed directly. Due to the QCD confinement radiated quarks and gluons hadronize when traversing the detector. This leads to collimated hadron tracks within the detector, so-called jets. Dedicated algorithms are needed to identify the originating particle of the jet and reconstruct its four-vector as precisely as possible.

To make meaningful comparisons between experimental data and predictions, a consistent clustering of particles to jets is very important. For a proper use in the experiments at the LHC, algorithms have to be efficient in computing time. On the other hand, two theoretical premises need to hold. Firstly, the outcome of a clustering algorithm should not fluctuate when a particle distributes its energy



(a) Infrared safety violation (b) Collinear safety violation

Figure 3.3.: Examples of violations of the two fundamental jet requirements. Additional soft emissions, e.g. by pile-up, are not supposed to change the jet definition (a). Furthermore, the number of reconstructed jets should not vary with collinear splitting (b).

among two collinear objects. Secondly, the number of reconstructed jets should not change by adding soft radiations. These two principles are illustrated by showing violations of them in Figure 3.3.

Standard jet clustering algorithms

In principle there are two types of clustering algorithms. Cone-based algorithms like SISCone [116] combine all objects in a given cone. Opposed to this, sequential clustering algorithms iteratively cluster adjacent objects. All techniques used in the analyses of this thesis rely on the latter approach: the anti-k_T jet algorithm [117] and the Cambridge/Aachen (CA) algorithm [118]. For the sake of completeness, it should be noted that the k_T algorithm [119, 120] depicts another sequential clustering technique that has been extensively used at the LEP experiments.

The sequential clustering algorithms do not fix the geometrical shape of jets. They start by calculating the distances between all pairs of objects i and j, given by

$$d_{ij} = \min\left(p_{\mathrm{T},i}^{2n}, p_{\mathrm{T},j0}^{2n}\right) \frac{\Delta R_{ij}^2}{R^2}\,, \tag{3.1}$$

where n denotes a free parameter differentiating the three algorithms. The size and resolution of the jets is determined by the size parameter R, which can also be chosen freely. In addition, for all objects the distance to the beam is calculated via

$$d_{iB} = p_{\mathrm{T},i}^{2n}\,. \tag{3.2}$$

The algorithm then searches for the smallest value in all calculated distances. If $\min(d_{ij}) < \min(d_{iB})$ the objects i and j are clustered into a new particle i'. If $\min(d_{iB}) < \min(d_{ij})$ object i is declared as jet and removed from the list of objects. Afterwards all distances are re-computed and iteratively clustered or removed until all objects are part of a jet. This type of jet reconstruction provides collinear and infrared safety by construction.

The constant re-computation of distances takes a long time for large numbers of objects. That is why a straight-forward utilization of sequential clustering algorithms would not to be applicable for the high luminosity collisions at the LHC. The implementations provided in the FastJet software package [121,122] solve this problem by using the geometrical nearest neighbor location [121].

As aforementioned, the definition of parameter n in Equations (3.1) and (3.2) is different for the sequential algorithms. The anti-k_T algorithm, being the default choice in the CMS collaboration, uses $n = -1$. This way, the resulting jets are roughly cone-shaped. In the CMS collaboration a size parameter of 0.5 is applied for anti-k_T jets, referred to as *AK5 jets*. The CA method uses $n = 0$, so only the pure geometrical distance between two objects is taken into account. This algorithm was found best for the analysis of jet substructure [123]. Both algorithms use PF objects as inputs.

Subjet/filter jet algorithm

The search for a Higgs boson in the WH production channel, presented in Chapter 5, investigates an alternative approach of reconstructing jets: The SubJet/Filter algorithm (SJF) proposed in [124]. This algorithm is designed for the reconstruction of heavily-boosted objects. In particular, the authors of [124] predicted that by implementing the SJF techniques the channel VH($\mathrm{b\bar{b}}$) could become one of the most important channels for the discovery of the Higgs boson.

Figure 3.4 provides an illustrative workflow of the SJF algorithm for the reconstruction of a $\mathrm{H} \to \mathrm{b\bar{b}}$ event. In the first step a *fat jet* j_{fat} with a large radius is clustered using the CA algorithm in order to collect all Higgs boson decay products. As a reasonable fat jet size parameter the authors suggest a value of 1.2, which is also used in this analysis.

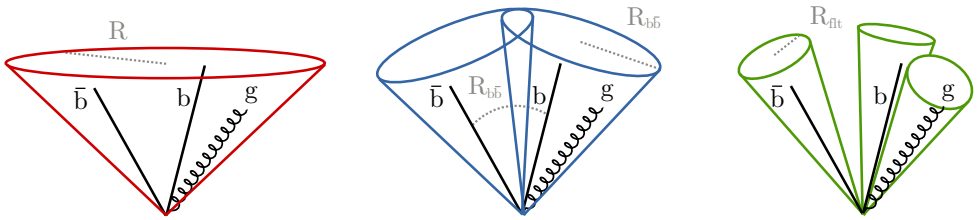
Afterwards, the clustering is undone iteratively. First, the last step of the clustering is canceled to break j_{fat} into two subjets j_1 and j_2. They are ordered such that $m_1 > m_2$. The fat jet is only assigned to the Higgs boson, if the mass of j_1 is significantly lower compared to j_{fat}, and the unclustering is not too asymmetric. Technically, the requirements

$$m_1 < \mu \cdot m_{\mathrm{fat}} \qquad \text{and} \tag{3.3}$$

$$y_{\mathrm{cut}} < y \equiv \frac{\min(p_{\mathrm{T},j_1}, p_{\mathrm{T},j_2})}{m^2_{\mathrm{fat}}} \cdot \Delta R^2(j_1, j_2) \tag{3.4}$$

have to be fulfilled. Otherwise, the subjet j_1 is taken as the fat jet from the first step and the unclustering is applied again. The two parameters μ and y_{cut} define the mass drop and the asymmetry requirement, respectively. In this thesis the explicitly suggested values of [124] are taken, i.e. $\mu = 0.67$ and $y_{\mathrm{cut}} = 0.09$.

With the fat jet assigned to the Higgs boson environment and the two corresponding subjets at hand, the CA algorithm is applied again with a much smaller jet radius of $R_{\mathrm{flt}} = \min(0.3, \Delta R(j_1, j_2)/2)$. The three hardest filter jets can then be interpreted as the two b quarks from the Higgs boson decay together with leading



(a) Fat jet (b) Subjets (c) Filter jets

Figure 3.4.: Illustrative overview of the Subjet/Filter jet algorithm for reconstructing the decay $H \to b\bar{b}$, based on [124]. In a first step fat jets are clustered with the CA algorithm using $R = 1.2$ as size parameter (a). The clustering is then withdrawn until a certain mass drop criterion is reached (b). Finally, filter jets are obtained by re-clustering inside the subjets with a size parameter of $R_{\text{filt}} = \min[0.3, R_{b\bar{b}}/2]$ (c). Therefore, filter jets have a smaller radius compared to the standard AK5 jets and are supposed to be more resilient against distortions from pile-up and underlying event.

order gluon radiation. The Higgs boson candidate built up from these three filter jets is predicted to be cleansed from contaminations due to pile-up interactions and underlying events. Therefore, the mass resolution is also expected to be improved.

Jet energy corrections

Before being able to compare the clustered jets to theory predictions, they need to be cleansed from detector influences. Saturation effects of single components or the non-linearity of the calorimetry system lead to differences in the jet response. The *jet response* is defined as the ratio of the measured jet's transverse momentum to the true transverse momentum of the generator reference particle, $p_{\text{T}}^{\text{jet}}/p_{\text{T}}^{\text{ref}}$. Factorized Jet Energy Corrections (JEC) provided by the CMS *JEC* group [125] address these different biases, as explained briefly in the following.

- *L1 FastJet:* On an event-by-event basis, the average pile-up density per unit area [126] is estimated and subtracted depending on the area of the jet.
- *L2 Relative:* Modulations of jet response depending on the pseudorapidity η are observed due to the non-linearity of the calorimetry at the CMS detector. MC simulated QCD events are used to compute η-dependent correction factors making the response flat in η.
- *L3 Absolute:* The non-linearity of calorimeters also causes a bias of the jet response in transverse momentum. The p_{T}-dependent correction factors are again evaluated using QCD MC.

In addition, for jets from data *L2L3 Residual* corrections take care of the fact that the *L2* and *L3* effects have been estimated by MC only. Unless otherwise noted,

the energies of jets used in this thesis are include all listed corrections. The energy corrections have been validated with in-situ measurements with the energy balance of dijet and γ/Z+jets events [127].

Another correction is applied to account for the differences between data and MC in the resolution of jet energies. This correction is covered in the corresponding sections in Chapters 5 and 6. For more information see [127].

Identification of b jets

The information of the origin of a jet, or more precisely whether it is originating from a b quark or not, is hugely useful in analyses dealing with multijet final states containing b quarks, as it can discriminate between signal and background processes. Dedicated methods considering the b quark decay characteristics, known as b tagging algorithms, provide such information.

When a b quark is produced in collision events, it hadronizes into a B meson. Since the meson's b quarks decay only via the weak interaction into c or u quarks, the relatively long lifetime of B mesons is on the order of $\tau = 1.6\,\text{ps}$. This delayed decay gives rise to tracks displaced with respect to the PV and forming a secondary vertex. The characteristics are illustrated in Figure 3.5.

This thesis relies on the use of the Combined Secondary Vertex (CSV) algorithm [128, 129]. This advanced method combines all available observables by applying multivariate tools. Using information of impact parameter significance of tracks, the distance between secondary and primary vertices as well as jet kinematics a likelihood discriminant is calculated. Jet stemming from b quarks get a large discriminator value, while gluon or light quark induced jets possess small values. The algorithm is very effective and even provides reliable information when no secondary vertex can be formed.

Different working points are defined according to the mistag rate, i.e. the efficiency to falsely classify a light quark or gluon induced jet as b jet. Globally provided scale factors correct for efficiency differences between data and simulation at these points. To exploit the full shape of the b tag discriminant a dedicated reshaping procedure presented in Section 5.4.1 is needed. The search for tHq final states uses a cut on the *tight* working point, corresponding to a mistag rate of 1 ‰, discussed in Section 6.4.1.

3.2.6. Missing transverse energy

The presented searches have to cope with the fact that there might be particles in each event leaving the detector without interaction. Since the colliding protons at the LHC only possess momenta longitudinal to the beam axis, the undetected particles yield an imbalance in the transverse momentum sum. This is referred to as missing transverse energy and is usually linked to the presence of neutrinos in the interactions, but could also arise from so far undiscovered particles.

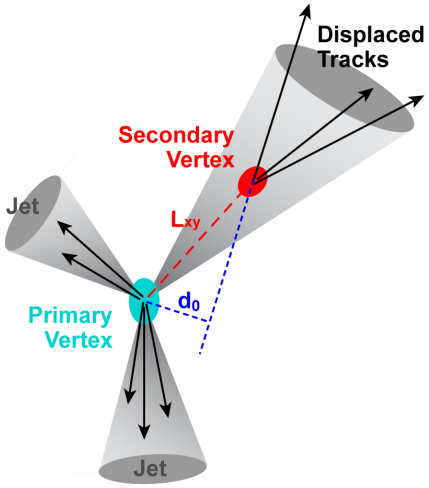


Figure 3.5.: Characteristics of collision events comprising b quarks, taken from [130]. The b quark fragments into a B meson, which typically has a lifetime on the order of $\tau = 1.6\,\mathrm{ps}$. The retarded decay is visible via displaced tracks possessing a large impact parameter d_0 with respect to the PV. In many cases out of several displaced tracks a secondary vertex with a distinctive distance to the PV ($\mathrm{L_{xy}}$) can be reconstructed.

By requiring momentum conservation the missing transverse energy is calculated from the negative sum over all reconstructed Particle Flow candidates,

$$\vec{\not{p}}_{\mathrm{T}}^{\,\mathrm{raw}} = -\sum_i \left(E_i \sin\theta_i \cos\phi_i \hat{\mathbf{x}} + E_i \sin\theta_i \sin\phi_i \hat{\mathbf{y}}\right) . \tag{3.5}$$

Here, $\hat{x}$ and $\hat{y}$ are the unit vectors in the direction of the x and y axes. This raw quantity does, in general, not represent the true transverse momentum of undetected particles due to detector and pile-up effects. That is why for both analyses presented in Chapters 5 and 6 corrections on $\vec{\not{p}}_{\mathrm{T}}^{\,\mathrm{raw}}$ are applied.

Jets are a major ingredient in Equation (3.5), thus so-called *type-I corrections* propagate the jet energy corrections to the missing energy. Furthermore, *type-0 corrections* are essential when charged hadrons are removed from pile-up interactions (see also Section 5.4). These corrections remove consistently an estimate of neutral pile-up contributions from the missing energy. For both, type-I and type-0 corrections, further information as well as the technical implementation are found in [131].

The missing transverse energy used in the analyses can be written as

$$\vec{\not{p}}_{\mathrm{T}} = \vec{\not{p}}_{\mathrm{T}}^{\,\mathrm{raw}} + \vec{C}_{\mathrm{T}}^{\,\text{type-0}} + \vec{C}_{\mathrm{T}}^{\,\text{type-I}} , \tag{3.6}$$

where $\vec{C}_{\mathrm{T}}^{\,\text{type-0}}$ and $\vec{C}_{\mathrm{T}}^{\,\text{type-I}}$ denote the specific correction terms for type-0 and type-I corrections, respectively. In events with one neutrino expected in the final state

the norm of $\vec{\not{p}}_{\mathrm{T}}$, abbreviated with $\not{E}_{\mathrm{T}}$, is usually assigned to the neutrino's transverse momentum. Another piece of information used in the following is the $\not{E}_{\mathrm{T}}$ significance, defined as $\not{E}_{\mathrm{T}}$ divided by $\sqrt{\sum_i p_{\mathrm{T},i}}$, a sum over all PF particles in the event. Its value represents the likelihood that the measured $\not{E}_{\mathrm{T}}$ is consistent with a fluctuation from zero due to imperfect detector responses.

4. Statistical methods and multivariate tools

Two cornerstones of the analyses presented in the next chapters are the use of multivariate tools and the statistical interpretation of the results. For the statistical inference the parameter estimation with the maximum likelihood estimator method as well as the construction of exclusion limits with the CL_s approach are applied. The multivariate approach allows to classify events as signal and background events and thus helps to increase the search sensitivity. In addition, multivariate methods can predict distinct parameter values, known as *regression*.

In this chapter, the concepts of maximum likelihood estimation and CL_s exclusion limits and their implementation in the statistics framework THETA [132] are described. Furthermore, an introduction to Boosted Decision Trees (BDT) and artificial Neural Networks (NN) based on their execution in the ROOT [133] TMVA package [134] are given.

4.1. Statistical methods

The following definitions assume analyses that are performed with binned histograms instead of continuous functions. This is true for the searches in Chapters 5 and 6. A more detailed description of the applied methods is given in [135].

4.1.1. Maximum likelihood parameter estimation

A common problem in high energy physics is to find an optimal parameter set $\vec{a}$ adjusting MC histograms to fit the measured data. The problem can be solved by using a maximum likelihood parameter estimation (MLE).

Consider N statistically independent measurements $X = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_N\}$, each of which is a set of values indicated by a vector. The conditional probability density functions (p.d.f.) $f(\vec{x}_i|\vec{a})$ quantify the likelihood of measuring $\vec{x}_i$ for a given set of parameters $\vec{a}$. In the calculations the different $f(\vec{x}_i|\vec{a})$ are assumed to be known, and have to be normalized for all $\vec{a}$. The joint probability of observing X given $\vec{a}$ is defined by the *likelihood function* built from the product of the individual p.d.f.s

$$\mathcal{L}(\vec{a}) = f(\vec{x}_1|\vec{a}) \cdot f(\vec{x}_2|\vec{a}) \cdots f(\vec{x}_N|\vec{a}) = \prod_{i=1}^{N} f(\vec{x}_i|\vec{a})\,. \tag{4.1}$$

To find the best set $\hat{a}$ for which the observation of the quantities $\vec{x}_i$ is most probable, $\mathcal{L}(\vec{a})$ has to be maximized.

In many scenarios, it is more convenient to use the natural logarithm of the likelihood function, known as the *log-likelihood*. The logarithm is a monotonically increasing function and thus has extrema at the same positions as the function itself. The advantage of the log-likelihood compared to the basic likelihood function is that taking derivatives is often easier, as it can be rewritten as a sum

$$\ln \mathcal{L}(\vec{a}) = \ln \prod_{i=1}^{N} f(\vec{x}_i|\vec{a}) = \sum_{i=1}^{N} \ln f(\vec{x}_i|\vec{a})\,. \tag{4.2}$$

For historical reasons the numerical methods for finding extrema are typically minimizers. Therefore the implementations in THETA and in other frameworks search for the minimum of the negative log-likelihood ($-\ln \mathcal{L}(\vec{a})$) to find the optimal set of parameters $\hat{a}$.

4.1.2. CL_s exclusion limits

As the Higgs boson mass is not predicted by theory, for a long time the Higgs boson hunt was a search for a needle in the haystack. When analyses observed no clear signal, the degree of confidence for eliminating the sensitive mass region needed to be statistically quantified. This is done by the calculation of *exclusion limits*. At the LHC the standard procedure is the computation of CL_s limits [136, 137], as explained in the following. The description is based on [138].

In general, when no clear excess predicted by a signal process with a theoretical cross section σ_{SM} is observed in data, upper limits on its cross section can be set. A signal strength multiplier $\mu \equiv \sigma/\sigma_{\mathrm{SM}}$ is introduced, to normalize the measured cross section to the standard model prediction. The exclusion limits are based on the profile likelihood ratio test statistic built from a set of nuisance parameters θ with corresponding priors π_θ, that represents their probability functions. The test statistic is calculated via

$$\tilde{q}_\mu = -2 \log \frac{\mathcal{L}(\mathrm{data}|\mu, \hat{\theta}_\mu)}{\mathcal{L}(\mathrm{data}|\hat{\mu}, \hat{\theta}_{\hat{\mu}})} \quad \text{with } 0 \leq \hat{\mu} \leq \mu\,. \tag{4.3}$$

Here, $\hat{\theta}_\mu$ is the conditional maximum for θ given a fixed value of μ and given the observed data. The values $\hat{\mu}$ and $\hat{\theta}_{\hat{\mu}}$ are the global maxima of the likelihood function. The constraint $\mu \geq 0$ is usually introduced in Equation (4.3) to achieve physically meaningful results. To obtain one-sided exclusion intervals the constraint $\hat{\mu} \leq \mu$ is introduced. The likelihood function is the product of all statistically independent bins i with n_i observed events, given by

$$\mathcal{L}(\mathrm{data}|\mu, \theta) = \prod_i \left[\frac{(\mu \cdot s_i + b_i)^{n_i}}{n_i!} e^{-(\mu \cdot s_i + b_i)} \right] \cdot \pi_\theta(\theta)\,. \tag{4.4}$$

The term in square brackets represents the Poisson distribution for one bin, i.e. $\mathrm{Poisson}(n|\mu \cdot s(\theta) + b(\theta))$. In other words, it is the likelihood to observe n events when $\mu \cdot s$ signal and b background event are expected, with a given μ.

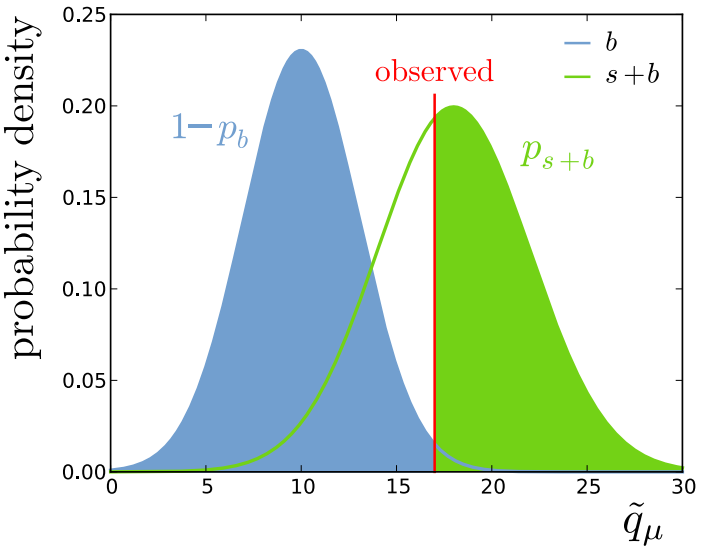


Figure 4.1.: Illustrative example of the CL_s value definition. The probability functions for the signal-plus-background $s+b$ and the background-only b hypotheses are derived from simulation. Given a measurement with an observed value for the test statistic $\tilde{q}_\mu$, the CL_s value is defined as $p_\mu/(1-p_b)$, where $p_\mu \equiv 1-p_{s+b}$.

For the derivations of a CL_s limit from the observed data first the observed test statistic from Equation (4.3) is calculated. Furthermore, the probability density functions $f(\tilde{q}_\mu|\mu,\theta)$ are constructed using Monte Carlo simulation for the signal and background processes. In particular, the scenario with $\mu = 0$, i.e. only events from background processes and no events from signal are predicted, plays a special role, and is denoted as the *background-only* or *null hypothesis*. The CL_s value opposes the *signal hypothesis* with a given μ with the null hypothesis via

$$\mathrm{CL_s}(\mu) \equiv \frac{p_\mu}{1-p_b}\,. \tag{4.5}$$

Here, the so-called p-values p_μ and p_b represent the compatibility between hypothesis and data, and are given by

$$p_\mu = P(\tilde{q}_\mu \geq \tilde{q}_\mu^{\mathrm{obs}} \mid \text{signal + background}) = \int_{\tilde{q}_\mu^{\mathrm{obs}}}^{\infty}\int_\theta f(\tilde{q}_\mu|\mu,\hat{\theta}_\mu^{\mathrm{obs}})\pi_\theta(\theta)\mathrm{d}\theta\mathrm{d}\tilde{q}_\mu\,, \tag{4.6}$$

$$1-p_b = P(\tilde{q}_\mu \geq \tilde{q}_\mu^{\mathrm{obs}} \mid \text{background-only}) = \int_{\tilde{q}_\mu^{\mathrm{obs}}}^{\infty}\int_\theta f(\tilde{q}_\mu|0,\hat{\theta}_\mu^{\mathrm{obs}})\pi_\theta(\theta)\mathrm{d}\theta\mathrm{d}\tilde{q}_\mu\,. \tag{4.7}$$

An illustrative example of the p-values and the CL_s value is shown in Figure 4.1.

The interpretation of the CL_s value depends on the tested μ. Exemplarily, for $\mu = 1$ and $CL_s = \alpha$ the interpretation reads: the scenario in question is excluded with a $(1-\alpha)$ CL_s confidence level (C.L.) at the nominal predicted signal strength. It is conventional to quote 95% C.L. upper limits in analyses. Consequently, the

value μ is adjusted until $\mathrm{CL_s} = 0.05$ is reached.

In order to set the found upper limit into perspective regarding the sensitivity of the analysis, it is common to quote the observed limits together with the *expected* limits. The expected limits are derived by generating a large number M of pseudo-datasets from background-only simulation. Each pseudo-dataset is then treated as it was real data, and so M $\mathrm{CL_s}$ upper limits at 95% C.L. are obtained. The distribution of M upper limits is normalized and integrated thereafter. The median expected limit corresponds to the point where the integral reaches 0.5. The $\pm 1\sigma$ and $\pm 2\sigma$ uncertainty bands correspond to the integral values 0.16 and 0.84, and 0.025 and 0.975, respectively.

4.1.3. Asymptotic limits

The test statistic in Equation (4.3) has a major advantage. Normally, the calculation of $\mathrm{CL_s}$ limits and in particular the expected $\mathrm{CL_s}$ limits, where the process is $\mathcal{O}(1000)$ times repeated, is very CPU intensive. Following Wilks's theorem [139] the procedure can be simplified. Assuming a large data sample size, Wilks's theorem states that the test statistic will follow asymptotically a χ^2 distribution with degrees of freedom corresponding to the difference in dimensionality between θ_μ and θ_0, i.e. equal to one when the constraint $0 < \mu$ is ignored. Therefore, the test statistic can be expressed analytically and the so-called *Asimov* dataset is introduced, defined to make estimations for all parameters equal to their true values. This Asimov dataset represents the full ensemble of pseudo datasets, and the median expected limits, as well as the $\pm 1\sigma$ and $\pm 1\sigma$ uncertainties, can be obtained easily.

The extensive mathematically derivation is given in [140].

4.1.4. Systematic uncertainties and the theta framework

For each systematic effect that influences the measurement an additional nuisance parameter θ_u is introduced in Equation (4.4). Many tools are available which provide the routines considering all nuisance parameters needed for the limit calculations (and the maximum likelihood estimation). This thesis relies mainly on the THETA framework [132] — a software package developed at KIT by Jochen Ott. THETA provides the full set of statistical methods together with a fast and stable implementation.

In the following the realization of some techniques within THETA is discussed. A more detailed description can be found in [141].

Rate uncertainties

When an uncertainty θ_u is only expected to change the overall rate of a process p all bins of its template are shifted simultaneously. Technically, a bin-independent but process-dependent factor $\xi \equiv \exp(\delta_{p,u}\theta_u)$ is introduced scaling the template in

the fit. Here, $\delta_{p,u}$ is a constant depending on the process. The prior for this *rate uncertainty* is equivalent to a log-normal distribution, given by

$$\pi_{\theta_u}(\theta_u, \delta_{p,u}) = \frac{1}{\xi \delta_{p,u} \sqrt{2\pi}} \cdot \mathrm{e}^{-\frac{(\ln \xi)^2}{2\delta_{p,u}^2}} \quad \text{with } \xi > 0\,. \tag{4.8}$$

The advantage of this kind of implementation is that by construction nonphysical values, i.e. $\xi < 0$, are not allowed in the fit.

Shape uncertainties

When systematic effects θ_s influence each event to a different degree, the resulting uncertainty presents itself in shape deviations from the nominal distributions. To account for these effects two additional histograms are introduced in the analysis. The *up* template corresponds to a $+1\sigma$ shift and is connected to $\theta_s = +1$. Consequently, the -1σ shift is assigned to $\theta_s = -1$ and referred to as *down* template. With $\theta_s = 0$ the nominal histogram is reproduced.

To introduce these templates to the statistical model, *template morphing* is performed in THETA. Technically, in the region $|\theta_s| < 1$ the template is interpolated with a cubic function, such that the process normalization as a function of θ_s and the individual bin entries is continuously differentiable at $\theta_s = \pm 1$. In the region $|\theta_s| > 1$ the template is extrapolated with the straight lines defined by the pairs $\theta_s = 0$ and $\theta_s = 1$, and $\theta_s = 0$ and $\theta_s = -1$, respectively. Furthermore, its derivative at $\theta_s = 0$ is the average of the slopes of the linear extrapolation. These constraints uniquely define the function.

MC statistical uncertainties

As the analyses introduced in Chapters 5 and 6 rely on Monte Carlo simulation as explained in the previous chapter, they face the problem that only a limited amount of events is available. The correct treatment is to introduce to the model an additional nuisance parameter $\theta_{p,i}^{\mathrm{stat}}$ following a Poisson distribution per bin and per process. This was originally proposed by Barlow and Beeston [142]. For the analyses with many analysis regions and numerous background processes, however, this leads to about $\mathcal{O}(100) - \mathcal{O}(1000)$ additional shape variations and the CPU time gets very large.

The implementation in THETA attacking this uncertainty relies on a modification of the Barlow-Beeston method, proposed in [143]. Here, only one additional nuisance parameter per bin θ_i^{stat} for all processes, is introduced. The Poisson distribution is approximated with a Gaussian. The important advantage is that the maximization of the likelihood with respect to the introduced nuisance parameters can be performed analytically. This procedure is known as *Barlow-Beeston-lite* method.

4.2. Multivariate analyses

In the analyses presented in this thesis, the signal events cannot be separated from the background processes by simply introducing requirements on a few kinematic distributions. The separation is only achieved by the simultaneous use of many variables. Dedicated algorithms are needed to make the most of the variables and their correlations.

Multivariate analysis tools (MVA) incorporate the correlations in the full set of available information and combine them to one single discriminant. The definitions of such methods are achieved in a *training* step, that needs events providing the true outcome, e.g. whether the event is signal or background, as input.

The MVAs executed in this thesis are implemented in the ROOT TMVA package [134]. In Chapter 5 Boosted Decision Trees and in Chapter 6 artificial neural networks are used for the classification of events as signal or background. BDTs are also applied in Chapter 5 to predict distinct values for a quantity, a method known as *regression*. The different techniques are introduced in the following.

4.2.1. Boosted Decision Trees

A decision tree itself is a consecutive set of yes or no questions, each of which is known as *node*. Each node depends on the answer of the former node. The final verdict — called *leaf* — is given after a fixed maximum number of nodes at most. In Figure 4.2 an exemplary decision tree that is actually used in Chapter 5 is shown.

In the training the criterion on each node is chosen such that the separation gain between successive nodes is at a maximum. Given n events with individual weights of w_i the *Gini index* of a node is defined as

$$\text{Gini} = \left(\sum_{i=1}^{n} w_i \right) \cdot P \cdot (1 - P) \,. \tag{4.9}$$

Here, P denotes the purity of the node, given by

$$P = \frac{\sum_s w_s}{\sum_s w_s + \sum_b w_b} \,, \tag{4.10}$$

where $\sum_s w_s$ and $\sum_b w_b$ are the sum of (weighted) signal and background events, respectively. The Gini index is 0 for a sample which is pure in signal or background events and has a maximum for a mixed sample, that has a purity of 0.5. Consequently, the splitting is good if the *separation gain*

$$\text{SG} = \text{Gini}_{\text{father}} - \text{Gini}_{\text{child 1}} - \text{Gini}_{\text{child 2}} \tag{4.11}$$

is maximized.

A single decision tree is easy to interpret but not very strong. Therefore, many decision trees are trained consecutively with re-weighted training datasets, known

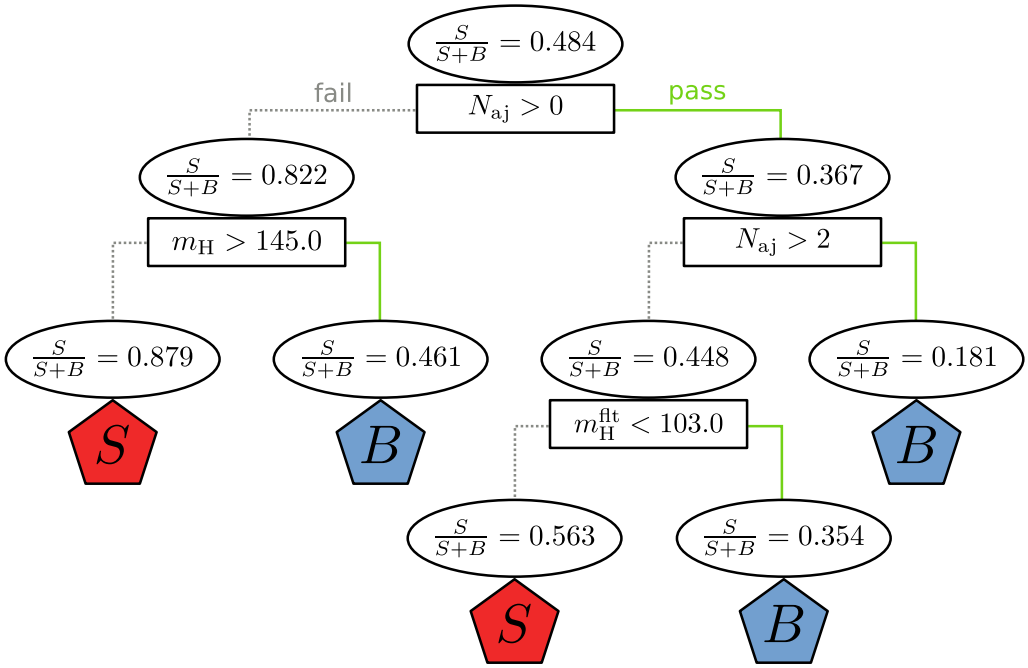


Figure 4.2.: Exemplary decision tree used in the analysis. The topmost node roughly has the same amount of signal and background events as input. Up to three subsequent requirements are applied depending on if an event passes or fails a cut. Each leaf classifies the event as signal (S) or background (B). On the order of hundreds of these trees are used simultaneously to make a majority decision.

as *boosting*. The combination of all trees is called *random forest*. It should be noted that there is an alternative randomizing approach via *bagging* [144] which is not explained here. An unknown event is put through all decision trees in the forest, and the final response is the majority vote of all trees. The idea is that a sum of weak decision trees will result in a stronger decision. This is clarified with an example. Assuming three uncorrelated decision trees are given, that are correct in 60% of all cases. To correctly classify an event as signal, only 2 out of 3 trees have to be correct. Therefore, the misclassification probability is given by

$$P_{\mathrm{mis}} = \binom{3}{2} \cdot 0.4^2 \cdot 0.6 + \binom{3}{3} \cdot 0.4^3 \cdot 0.6^0 = 0.352\,. \tag{4.12}$$

Consequently, the misclassification rate of the ensemble of trees is smaller compared to the single decision tree.

A practical tutorial on the implementations of BDTs in TMVA can be found in [145].

Boosting

The training of multiple trees can be performed in several ways. For the BDTs used in this analyses, the adaptive boost (*AdaBoost*) method [146] is applied which assigns a larger weight to misclassified events of the previous tree in the training of the subsequent tree. The weighting is implemented as follows.

Given a set of N training events, each with a weight of $w_i = 1/N$, the misclassification rate for the m^{th} tree is calculated via

$$r_m^{\text{mis}} = \frac{\sum_{i=1}^{N} w_i \delta_i(\text{wrong})}{\sum_{i=1}^{N} w_i} \,. \tag{4.13}$$

Here, $\delta_i(\text{wrong})$ is equal to 1, if the event was falsely identified, and 0 otherwise. The boost weight α_m is given by

$$\alpha_m = \beta \cdot \ln \frac{1 - r_m^{\text{mis}}}{r_m^{\text{mis}}} \,, \tag{4.14}$$

where β is a free parameter with $\beta > 0$. All event weights are changed to

$$w_i \to w_i \cdot e^{\alpha_m \delta_i(\text{wrong})} \,. \tag{4.15}$$

The criterion $\delta_i(\text{wrong})$ in the exponent ensures that only misclassified events are affected. Afterwards, the event weights are re-normalized to 1. The final response for a given event is the α_m-weighted sum of all individual trees.

Regression

In some cases not a simple yes or no decision but a distinct estimate for a quantity is wished-for. In order to attack such problems, a so-called *regression* method can be applied. The TMVA package provides the regression operation with BDTs. Regression trees are designed such that subsequent yes or no decisions lead to leaf nodes, that do not classify the events into signal or background, but give an estimate for a specified target variable. Typically, single regression trees have a larger depth ($\mathcal{O}(20)$) compared to classification trees with a depth of ≈ 3. As the Gini index from Equation (4.9) with the absence of correctly and falsely classified events is not valid anymore, the criterion for splitting a node is taken to be the *average squared error* [134, 147]

$$r_m^{\text{err}} = \frac{1}{N_m} \sum_{i}^{N_m} (y_i - \hat{y})^2 \,. \tag{4.16}$$

Here, y_i is the true value of the regression target for event i and $\hat{y}$ denotes the target's mean value over all events in the node. If r_m^{err} exceeds a given threshold the node is split.

4.2.2. Neural Networks

Neural networks are structures with artificial neurons inspired by the human brain. Similar to BDTs, for NNs the correlations between numerous input variables can be identified and used for classification and regression problems. There are many implementations available for NNs. In this analysis the multilayer perceptron (MLP) implementation of TMVA is used.

Generally, a neural network with k neurons can have k^2 connections between the nodes. The MLP is a layer-structured feed-forward neural network, where nodes are only connected to nodes from the subsequent layer. The architecture is illustrated in Figure 4.3. The first MLP layer is the input layer with one node per input variable. Thereafter a user-defined number of hidden layers is included. In a hidden layer the node i gets the weighted sum of outputs from the previous layer as input, i.e.

$$h_i(x) = \sum_i w_{ij} x_i \,. \tag{4.17}$$

Here, w_{ij} denotes the weight that is given each previous layer node's output x_i. The result, which can have values from $-\infty$ to $+\infty$, is transformed with an *activation function* to the range $[-1, +1]$. In TMVA the hyperbolic tangent

$$\tanh(x) = 1 - \frac{2}{e^{2x} + 1} \tag{4.18}$$

is used as activation function.

In classification problems, there is one node in the final output layer. The final output, exemplarily applying one hidden layer, is given by

$$y_{\mathrm{NN}} = \tanh\left(\sum_{j=1}^{N_j} w_j \cdot \tanh\left(\sum_{i=1}^{N_i} w_{ij} x_i\right)\right) \,, \tag{4.19}$$

where N_i and N_j are the number of input nodes and hidden layer nodes, respectively. The weight between each hidden node and the output layer is denoted with w_j.

All weights are subject to the training, where again a large set of events is used, for which the true output is known. For each training event a the neural network output $y_{\mathrm{NN},a}$ is calculated and compared to the target $\hat{y}_a$, which is either 1 for signal or 0 for background events. An error function is built, given by

$$E(\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_N | \vec{w}) = \sum_{a=1}^{N} \frac{1}{2} \left(y_{\mathrm{NN},a} - \hat{y}_a\right)^2 \,, \tag{4.20}$$

where $\vec{x}_a$ denotes the ensemble of input variables for event a, and $\vec{w}$ indicates the set of adjustable weights in the training. The training searches for the $\vec{w}$, for which E is minimal. TMVA applies the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm [148–151], which is an iterative method for solving non-linear optimization problems.

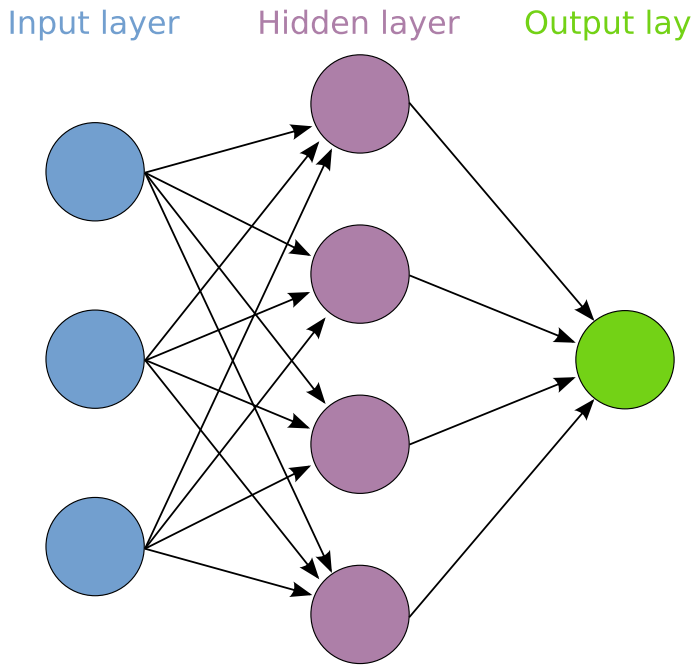


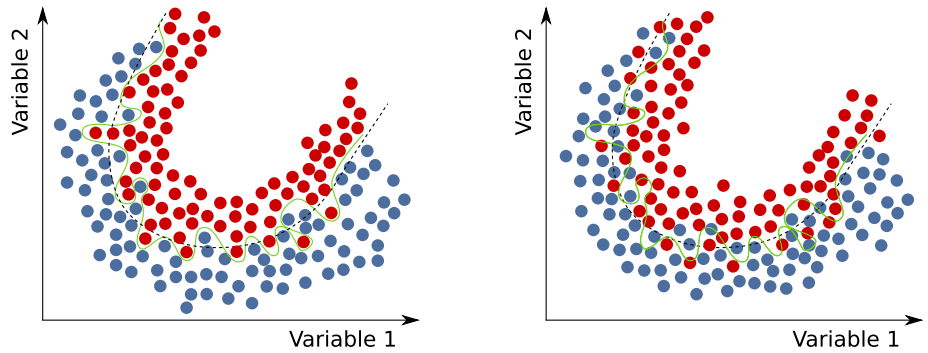
Figure 4.3.: Typical architecture of a feed-forward neural network. The shown example has three layers: input layer, hidden layer and output layer. Each node is connected to all nodes of the next layer. The number of input nodes corresponds to the number of input variables used in the training. The number of hidden layers is adjustable by the user. A bias for each neuron in the hidden layer is introduced by a *bias node* (not displayed). Eventually, the single discriminant provided by the output layer can be used as a classifier.

4.2.3. Overtraining

When a BDT or NN training uses too few training events for too many adjustable parameters it can happen that statistical fluctuations in the training sample are learned. This effect is known as *overtraining* and needs to be avoided at any price. In an independent dataset an overtrained MVA usually performs worse, as the fluctuations occur at different positions. An illustrative example is depicted in Figure 4.4. One easy way to check an MVA against overtraining, is to apply the training results on an independent simulated test sample. The comparison between the test and training distributions via a Kolmogorov-Smirnov test provides the probability that the two outputs have identical origins.

4.2.4. Ranking of variables

An helpful piece of information in a multivariate analysis is the importance ranking of the input variables. That way, it is possible to remove less important variables, or



(a) (b)

Figure 4.4.: Illustrative example for overtraining. Shown are two measurements for signal events (red) and background events (blue) in two arbitrary variables. On the left side, the training dataset is shown. The decision boundary from a well trained MVA (black dashed line) is depicted as well as from an overtrained MVA (green line) that learns from statistical fluctuations in the dataset. In a statistically independent dataset, illustrated on the right side, the fluctuations occur at different positions and the performance of the overtrained MVA is worse.

judge newly introduced variables for their benefit. For BDT and NNs the ranking is calculated in a different way.

To obtain the ranking of input variables for a BDT, it is counted how often each variable is used to split decision tree nodes, weighted by the factor in Equation (4.11) and the number of events in the node.

As opposed to this the MLP neural network ranks the input variables according to the weights between the corresponding node in the input layer and all nodes of the first hidden layer via

$$I_i = \bar{x}_i^2 \cdot \sum_{j=1}^{n_h} \left(w_{ij}^{(1)} \right)^2 . \tag{4.21}$$

Here, I_i denotes the importance of the variable i and $\bar{x}_i$ is its sample mean.

5. Search for a standard model Higgs boson in the WH production channel

The Higgs boson found at the LHC is predicted to decay predominantly into bottom quark pairs. However, in this channel a lot of effort is needed to extract the signal events whilst considering numerous background processes. Therefore, the discovery was mainly driven by bosonic decay channels into either γ, W or Z boson pairs. Using the full available dataset the ATLAS and CMS collaborations do not see evidence for $H \to b\bar{b}$ decays in their most recent results [152, 153] yet, so analysis advancements increasing the search sensitivity are desired.

The goal of the analysis presented in this chapter is to improve the search sensitivity for $H \to b\bar{b}$ decays at the CMS experiment. The analysis was developed in parallel to the CMS publication [153] and includes advanced jet reconstruction techniques based on the Fat-, Sub- and Filter Jet (SJF) algorithm explained in Section 3.2.5. A novel filter jet regression technique is presented that accounts for missing dedicated jet energy corrections. Furthermore, a cross check of the official results is carried out and for the first time the improvements of the usage of jet substructure are quantified based on the full 8 TeV dataset. These studies can help the $H \to b\bar{b}$ effort within the CMS collaboration to face the new challenges in the coming data taking period with higher center-of-mass energies.

After introducing the general search strategy in the $W(\ell\nu)H(b\bar{b})$ channel, the characteristics of signal and background processes are described in this chapter. Furthermore, the MC and data samples shared with [153] are given, as well as the reconstruction procedure and selection requirements. Finally, the computation and validation of the Boosted Decision Trees, that are employed to extract CL_s exclusion limits on the WH signal process at 95% C.L., is explained in detail.

5.1. Analysis strategy

As shown in Figure 1.5, the Higgs boson found at the LHC with a mass of $m(H) = 125\,\text{GeV}$ decays in about 60% of all cases into a bottom quark pair. Due to large background contributions with two bottom quarks in the final state, the channel $gg \to H \to b\bar{b}$ is impossible to investigate.

The first step to make the search for $H \to b\bar{b}$ events feasible is to focus on the production of a Higgs boson in association with a vector boson via Higgsstrahlung (VH). The representative Feynman diagrams are shown in Figure 5.3. This pro-

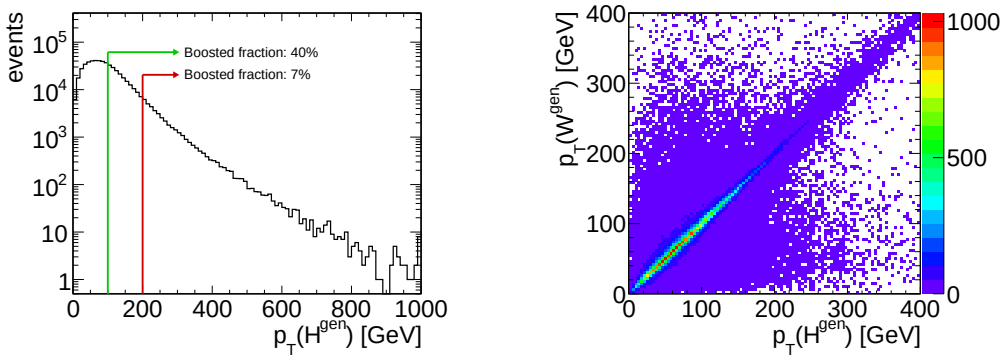


Figure 5.1.: Generated boost distributions in signal MC. The left diagram illustrates the signal efficiency for specific requirements on the transverse momenta of generated Higgs bosons. With the criterion $p_T > 100\,\text{GeV}$ 60% of the signal events are lost. In the right-hand figure the transverse momenta of W boson and H boson are compared at generator level. A clear correlation is visible.

duction mode has a lower cross section compared to $gg \to H$ (see Figure 1.4), but provides leptons and/or missing transverse energy that can be used to trigger the events. Nonetheless, further ideas were needed to suppress the dominant background processes, and for a long time the inspection of this channel was seen as futile for the Higgs boson discovery. Only after the proposal to search for $VH(b\bar{b})$ in a boosted event topology [124], this decay channel was reinvestigated. By requiring the Higgs boson and the vector boson candidates to have large transverse momenta, a significant amount of signal events is ignored. Figure 5.1 shows the expected signal efficiencies for specific requirements on the transverse momenta of generated Higgs bosons, as well as the correlation between the boosts of Higgs boson and W boson. Yet, according to the authors of [124] the advantages of this strategy predominate. In the boosted regime the multijet $b\bar{b}$ production is strongly suppressed. In addition, other background processes get indicating features that can be used to discriminate the signal against them. For instance, events from $t\bar{t}+\text{jets}$ production are likely to provide a high-energetic $b\bar{b}$ pair only with a larger jet multiplicity in the event. Another advantage is that the decay products of the boosted signal events are central in the detector and the tracking system can be used for the reconstruction. This improves the jet resolution and allows for the usage of b-tagging. Based on these ideas the layout of the analysis is constructed.

The boost requirements on the reconstructed Higgs and W boson candidates are an essential feature of the event selection defining a signal enhanced phase space. The dominant remaining background processes are $t\bar{t}$ and $W+\text{jets}$ production. The simulation of the latter is split into contributions with zero, one or two additional b quarks in the event ($W + 0b$, $W + 1b$ and $W + 2b$). Scale factors adjusting the normalization of these main background templates are estimated via a data-driven approach in dedicated control regions. To enhance the mass resolution of the re-

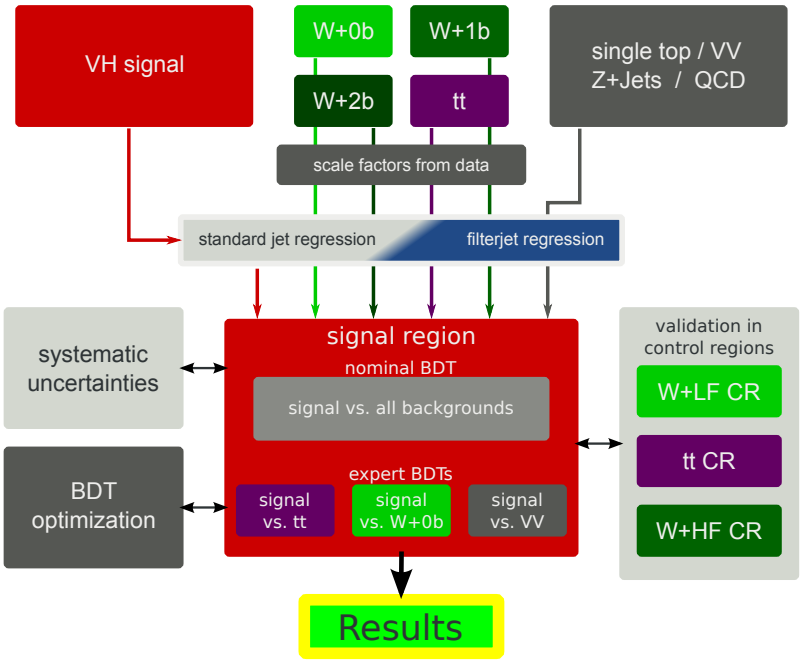


Figure 5.2.: Overview of the search strategy. Scale factors are computed in a data-driven way for $t\bar{t}$, W + 0b, W + 1b and W + 2b production. The estimates on the other processes are taken from simulation. Regression techniques are applied to correct the energies of standard and filter jets. After various validation steps and the evaluation of systematic uncertainties, four BDTs are evaluated and their combination is optimized. The exclusion limits are extracted on a template fit to the final discriminator.

constructed Higgs boson, regression techniques trained on simulated signal events are applied to both, standard jets and filter jets. The optimization and validation of the regression on filter jets is a significant part of this thesis. Furthermore, possible discriminating variables using the jet substructure are investigated to improve the search sensitivity of the CMS analysis. After the validation of all discriminating variables and the estimation of systematic influences on the results, in total four different BDTs are built. One is optimized to separate the signal events from all background processes, whereas the other three are dedicated to discriminate the signal process against $t\bar{t}$, W + 0b and diboson production separately. In the optimization procedure the best settings for all four decision trees are found. Additionally, the ideal combination of the four BDTs into one final discriminator to gain the largest search sensitivity is identified. This step is performed twice, first with the set of variables used in [153], and secondly with additional substructure information in the training. Finally, a template fit on the final discriminator is performed and CL_s exclusion limits are extracted for nine different Higgs boson mass hypotheses. The sketch in Figure 5.2 summarizes the strategy. In the following the individual parts of the analysis are described in detail.

5.2. Signal and background characteristics

5.2.1. Signal topology

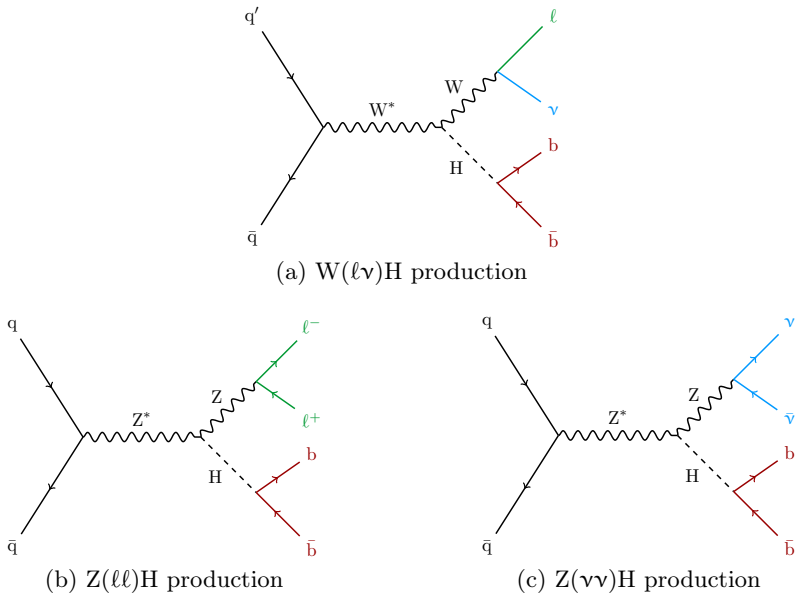
To define a signal enhanced phase space in the first place, and later to find discriminating variables that separate signal from background processes, it is important to know the topology of the investigated $W(\ell\nu)H(b\bar{b})$ process. Characteristically, the signal events include a W boson with large transverse momentum and two high energetic b jets, stemming from the Higgs boson. In the boosted regime the W and the Higgs boson are expected to be central in the detector due to the combined system's large invariant mass. The azimuthal opening angle between the two bosons $\Delta\varphi(H, W)$ is predicted to be sharply peaking at π. This means, Higgs and W bosons travel back-to-back in the majority of all cases. Smaller opening angles $\Delta\varphi(H, W) < \pi$ occur when for instance the WH system recoils from additional radiation. The system of two b jets is expected to have an invariant mass within $110\,\text{GeV} < m_{b\bar{b}} < 150\,\text{GeV}$ depending on the mass hypothesis as they originate from the Higgs boson. The transverse momentum distributions of the b jets have a maximum around $m_H/2$. Apart from the lepton stemming from the W boson, there are no further isolated leptons expected in the event, and additional activity in the detector like extra jets is predicted to be minimal. The representative Feynman diagram of the $W(\ell\nu)H(b\bar{b})$ process is shown in Figure 5.3(a). Due to misidentified leptons, contributions from the processes $Z(\ell\ell)H$ and $Z(\nu\nu)H$, depicted in Figures 5.3(b) and (c), have to be also taken into account.

5.2.2. Background topology

There are several sources creating contributions to the sample that is selected in the signal enhanced phase space in the end. Especially, $t\bar{t}$ and W+jets production are the dominant background processes. Some of the background contributions can be reduced sufficiently by enforcing selection criteria, while for others the correlations of more variables need to be incorporated. In the following, for each occurring background process the distinct difference of patterns in one or more kinematic distributions is described.

V+jets production

The production of a vector boson together with additional jets mimics the signal process. While the contributions from Z+jets can be reduced to a minimum by requiring exactly one isolated lepton, W+jets production is a very important background. Particularly, the $W+b\bar{b}$ production, shown in Figure 5.4(a), resembles the final state of the WH process. After applying b-tagging requirements this is the dominant V+jets contribution in the signal region. However, the p_T spectrum of jets in V+jets production is softer compared to the signal process. Furthermore, the invariant mass distribution of reconstructed Higgs boson candidates in these



(a) W($\ell\nu$)H production

(b) Z($\ell\ell$)H production

(c) Z($\nu\nu$)H production

Figure 5.3.: Representative LO Feynman diagrams for VH production. The analysis is optimized for W($\ell\nu$)H($b\bar{b}$) production (a). Contributions from ZH processes shown in (b) and (c) are also taken into account. To facilitate the comparisons with diagrams of the background processes a color code is introduced. In the final state b quarks are indicated in red, charged leptons are green and neutrinos that give rise to missing transverse energy are blue.

events peaks at a lower value with respect to the signal process. Moreover, decay characteristics like effective spin and color radiation can be used for the separation.

Top quark production

The production of $t\bar{t}$ pairs is a particularly challenging background in searches for the Higgs boson. Especially the topology of semi-leptonically decaying top quark pairs as depicted in Figure 5.4(b) looks much like the signal process. Typically in the highly boosted regime $t\bar{t}$ production arises with additional jets. This fact is used to discriminate this process against the signal. Additionally, the azimuthal angle between the reconstructed W boson and Higgs boson candidates is wider in $t\bar{t}$ events.

The production of single top quarks is harder to separate from the WH process. In many cases the light forward jet, characteristic for t-channel production as shown in Figure 5.4(c), escapes detection making the signature the same as for signal events. However, the b quark stemming from the initial gluon splitting is often too soft to be detected. With that said, and considering the smaller production cross section, the contribution of single top quark events is less than 10% of all backgrounds.

Diboson production

Another process that mimics the signal topology is the production of two vector bosons, i.e. WW, WZ and ZZ. Particularly the WZ production, where the W boson decays leptonically and the Z boson into a pair of b quarks as shown in Figure 5.4(d), is an irreducible background. It can only be discriminated against the signal process using the difference in the reconstructed mass of the b jet system. Therefore, a good resolution in the invariant mass distributions is crucial.

QCD multijet production

Due to the large production cross section, the influence of multijet events produced via the strong interaction has to be taken into account as well. Here, leptons in the final state can occur due to semi-leptonically decaying hadrons containing b or c quarks, or due to the misidentification of jets. Especially the boost requirement and the demand for isolated leptons in the events minimize the multijet contribution. In Section 5.6.4 a data-driven method is presented showing that QCD production can be neglected in this search.

5.3. Monte Carlo simulation and analyzed data

The vast amount of different generators adopted for the simulation of signal and background events shows the diversity of contributing processes. In Table A.1 in Appendix A the full list of the samples can be found. The cross sections applied for

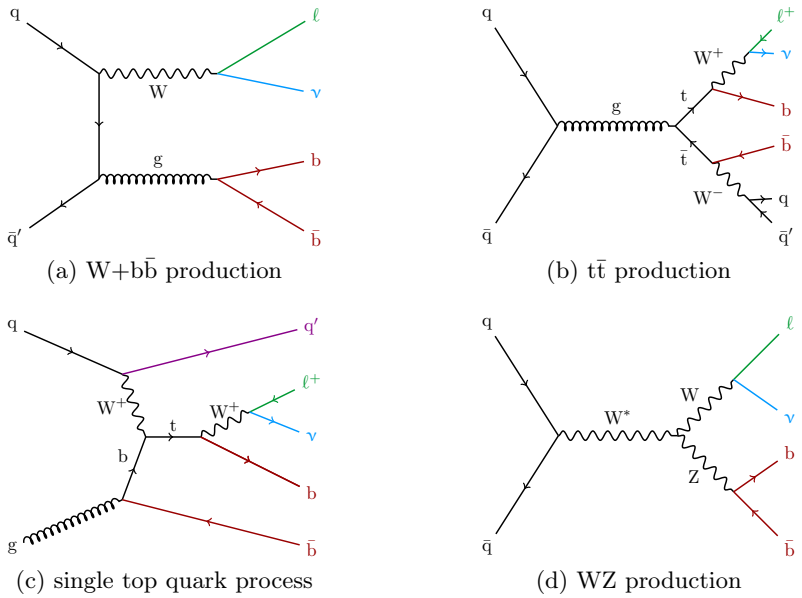


Figure 5.4.: Representative LO Feynman diagrams for important background processes to the WH search. To impart similarities in topology to the signal process, the same color code as in Figure 5.3 is introduced.

normalizing the templates and the total number of generated events are also given there. These samples are provided centrally within the CMS *VHbb* group.

For WH and ZH production modes the samples are generated with POWHEG 1.0 and showered with HERWIG++ 2.5. The templates are normalized to next-to-next-to-leading order (NNLO) cross sections [154]. Table 5.1 summarizes the cross sections for all tested mass hypotheses together with their uncertainties and shows additionally the corresponding branching ratios $\mathcal{B}(\mathrm{H} \to \mathrm{b\bar{b}})$.

The MADGRAPH 5.1 package is used to generate the $\mathrm{t\bar{t}}$ and V+jets processes. The $\mathrm{t\bar{t}}$ process is produced separately for full-hadronic, semi- and full-leptonic $\mathrm{t\bar{t}}$ decays. In the analysis the templates are scaled to the next-to-leading-order (NLO) cross section. To account for possible mismodeling in V+jets production these samples are split into W/Z with zero, one or two additional jets stemming from b quarks. All V+jets templates are scaled to inclusive NNLO cross sections.

The production of single top quark events is generated separately in t-channel, s-channel and tW-channel using POWHEG 1.0. The templates are normalized to approximate NNLO [155]. For the simulation of diboson and QCD processes PYTHIA 6.4 is used.

The simulation of the CMS detector response is performed with the GEANT 4 package [156]. All samples include additionally generated pile-up interactions for a proper description of the high luminosity environment at the LHC.

In this analysis the full dataset of proton-proton collisions at a center-of-mass

Table 5.1.: Production cross sections and branching ratios $\mathcal{B}(\mathrm{H} \to \mathrm{b\bar{b}})$ for all investigated mass hypotheses at $\sqrt{s} = 8\,\mathrm{TeV}$. All values are taken from [154]. The given uncertainties for the cross sections include contributions from uncertainties on the QCD scale, PDF and strong coupling constant α_s variation.

m_H [GeV]	σ_WH [pb]	$\Delta_{\sigma_\mathrm{WH}}$ [%]	σ_ZH [pb]	$\Delta_{\sigma_\mathrm{ZH}}$ [%]	$\mathcal{B}(\mathrm{H} \to \mathrm{b\bar{b}})$	$\Delta\mathcal{B}$ [%]
110	1.0600	+3.9, −4.4	0.5869	+5.4, −5.4	0.745	+2.1, −2.2
115	0.9165	+4.0, −4.5	0.5117	+5.6, −5.5	0.704	+2.4, −2.5
120	0.7966	+3.5, −4.0	0.4483	+5.0, −4.9	0.648	+2.8, −2.8
125	0.6966	+3.7, −4.1	0.3943	+5.1, −5.0	0.577	+3.2, −3.2
130	0.6095	+3.7, −4.1	0.3473	+5.4, −5.3	0.493	+3.7, −3.8
135	0.5351	+3.5, −4.1	0.3074	+5.4, −5.2	0.403	+4.2, −4.3
140	0.4713	+3.6, −4.2	0.2728	+5.6, −5.4	0.315	+3.4, −3.4
145	0.4164	+3.9, −4.5	0.2424	+6.0, −5.8	0.232	+3.7, −3.7
150	0.3681	+3.4, −4.0	0.2159	+5.7, −5.4	0.157	+4.0, −4.0

energy of $\sqrt{\mathrm{s}} = 8\,\mathrm{TeV}$ recorded with the CMS experiment in 2012 is investigated. After the selection of good runs according to the most recent JSON files (see Section 2.2.4) at that time [157–160] the data corresponds to an integrated luminosity of $19.0\,\mathrm{fb}^{-1}$. Here, due to pixel misalignment problems the run range 207883-208307 corresponding to an integrated luminosity of $0.6\,\mathrm{fb}^{-1}$ has already been removed. In Table A.2 the different blocks of primary datasets are listed in detail.

In the W($\mu\nu$)H channel the SingleMu and in the W(eν)H channel the SingleEle primary datasets are used. To avoid losing interesting data events, in the W($\mu\nu$)H channel different triggers with and without requirement on the muon isolation are combined to maximize the trigger efficiency over the full data-taking period. In the W(eν)H channel the HLT path `Ele27_WP80` is chosen for the analysis. This path is unrestrained for the complete data taking period. Table A.3 summarizes the triggers used in this analysis. Effects due to trigger, reconstruction and identification have been determined centrally in the CMS *VHbb* group using a tag-and-probe method [161]. This way, event weights for simulated templates are computed to correct for differences in efficiencies between data and MC.

5.4. Object selection and event reconstruction

In this section the criteria applied on physics objects and the reconstruction of W and Higgs boson candidates are described. Since this analysis shares the same processed samples for MC and data with the published CMS analysis [153], a synchronization on physics objects level is achieved by default. The reconstruction of W and Higgs boson is performed in analogy to [153] and for the regression of standard jets the same training results are applied.

5.4.1. Pre-selection criteria on physics objects

The analysis makes use of Particle Flow objects as explained in Chapter 3. To examine $\mathrm{W}(\ell\nu)\mathrm{H}(\mathrm{b\bar{b}})$ events the reconstruction of electrons, muons and jets all emerging from a common primary vertex is essential. Also the treatment of $\not{E}_{\mathrm{T}}$ assigned to neutrinos has to be discussed. The following baseline selection is applied to all events in data and MC.

Primary vertex selection and pile-up reweighting

For the analysis in each event the vertex with the largest sum of transverse momenta of all associated tracks is selected. This primary vertex (PV) is required to have a z position within 24 cm of the nominal detector center. In addition, its radial position must lie within 2 cm around the beam spot, and the vertex fit must include more than four degrees of freedom.

The PU multiplicity varies between 10 and 30 in LHC proton-proton collisions. This additional activity in the detector impairs jet momentum reconstruction and thus the reconstructed invariant mass of the Higgs boson candidates. To correct for these effects this analysis uses the Charged Hadron Subtraction (CHS) method [162], where all charged hadrons not stemming from the selected PV are filtered. The CHS only works within the tracker acceptance region. In addition, the average momentum density per unit area ρ from the FastJet package [121, 122] is used to correct for PU contaminations.

To account for potential differences between data and MC stemming from event selection bias, the standard procedure provided by the CMS collaboration is applied to reweight events in simulation. Further details can be found in [163].

Electrons

The electron candidates must lie within $|\eta| < 2.5$ and have a transverse momentum larger than 30 GeV. They are further required to pass the 80% working point (WP80) by cutting on the electron MVA ID [164]. Here, 80% is approximately the efficiency for prompt electrons to pass this criterion. The MVA ID implies additional requirements on shower shape, isolation and track cluster matching.

Muons

The muon candidates need to be identified as PF objects. Only muons with $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ and $|\eta| < 2.4$ are considered. The requirement on the normalized value $\chi^2/\mathtt{ndof} < 10$ needs to hold for the global fit of the tracks from tracker and muon systems. At least one pixel hit, at least six tracker layers with valid hits, at least one valid hit in muon chambers and in two muon stations are demanded. The impact parameter in the transverse plane of the muon's track with respect to the beam spot has to be smaller than 2 mm.

Lepton isolation

Lepton candidates produced with a large amount of hadronic activity might not originate from real W bosons. That is why for lepton candidates the isolation, defined as

$$I^{\ell}_{\mathrm{Rel}} \equiv \frac{I^{\ell}_{CH} + I^{\ell}_{NH} + I^{\ell}_{Ph}}{p^{\ell}_{\mathrm{T}}}\,, \tag{5.1}$$

is an important information. Here, I^{ℓ}_{CH}, I^{ℓ}_{NH} and I^{ℓ}_{Ph} denote the energy deposits in a cone of $\Delta R = 0.3$ around the track of the lepton stemming from charged hadrons, neutral hadrons and photons, respectively. For both, electron and muons candidates, an isolation requirement of $I^{\ell}_{\mathrm{Rel}} < 0.12$ is demanded.

Jets

This analysis uses jets reconstructed with the anti-k_{T} algorithm with a size parameter of 0.5. The standard jet-energy corrections introduced in Section 3.2.5 are applied. All jets are required to lie within $|\eta| < 2.5$ and must have a transverse momentum larger than 30 GeV. Additionally, at least two associated tracks and electromagnetic and hadronic energy fractions of at least 1% of the total energy need to be assigned to each jet. The energy resolution of jets in simulated events is smeared by 5% for $|\eta| < 1.1$ and 10% for $1.1 < |\eta| < 2.5$ to match the resolution observed in data.

Identification of b jets and reweighting

For the identification of b jets the CSV algorithm (see also Section 3.2.5) is used. Three working points are supported by the CMS *B-Tagging and Vertexing* (BTV) group, defined by the corresponding rate of falsely b-tagged jets: CSVL (CSV > 0.244, 10% mistag rate), CSVM (CSV > 0.679, 1% mistag rate), CSVL (CSV > 0.898, 0.1% mistag rate). The BTV group provides scale factors for these working points [165] to account for efficiency differences between data and simulation.

To make the full CSV distribution applicable for the analysis a more advanced procedure is needed. The aim of the following method, taken from [166], is to have a flavor dependent function correcting the original CSV value to cure the efficiency differences in the whole spectrum, i.e.

$$\mathrm{CSV}_{\mathrm{corr}} = f(\mathrm{CSV}_{\mathrm{orig}})\,. \tag{5.2}$$

Initially, in simulation a cut $\mathrm{CSV}_{\mathrm{equiv}}$ is computed at each of the n measured operating points to make the MC efficiency equal to that measured in data, i.e.

$$\epsilon^{\mathrm{data}}_{\mathrm{CSV}>\mathrm{CSV}_{\mathrm{orig}}}\Big|_n = \epsilon^{\mathrm{mc}}_{\mathrm{CSV}>\mathrm{CSV}_{\mathrm{equiv}}}\Big|_n\,. \tag{5.3}$$

By additionally fixing the function at the minimum and maximum values of the CSV range to force identity, $f(x)$ can be built with $n + 2$ constraints. To be able

to apply the same cut on $\mathrm{CSV_{corr}}$ for MC and on $\mathrm{CSV_{orig}}$ for data at all measured operating points in the analysis, the requirement $f(\mathrm{CSV}^n_{\mathrm{equiv}}) = \mathrm{CSV}^n_{\mathrm{orig}}$ must hold at each sampling point. Therefore, equation (5.2) becomes

$$\epsilon^{\mathrm{data}}_{\mathrm{CSV}>\mathrm{CSV_{orig}}}\Big|_n = \epsilon^{\mathrm{mc}}_{f(\mathrm{CSV})>f(\mathrm{CSV_{equiv}})}\Big|_n = \epsilon^{\mathrm{mc}}_{f(\mathrm{CSV})>\mathrm{CSV_{orig}}}\Big|_n \,. \tag{5.4}$$

The function $f(x)$ is linearly interpolated between the n sampling points. According to the results, event weights are computed and applied to simulation, so the full range of the CSV discriminator can be exploited.

5.4.2. Vector boson reconstruction

In this analysis the W boson is required to decay leptonically, i.e. into electrons or muons and neutrinos. Neutrinos escape from the detector unobserved, so the missing transverse energy is assigned to them. In the boosted regime the lacking z component of the missing transverse energy can be neglected and is set to zero. Consequently, the four-vector of a W boson candidate is built by the sum of the lepton's four-vector and x and y components of the missing transverse energy.

Given this reconstruction, the transverse momentum and the transverse mass, two important properties of the W boson candidates, are calculated via

$$p_{\mathrm{T}}(\mathrm{W}) = \sqrt{[\not{E}_x + p_x(\ell)]^2 + [\not{E}_y + p_y(\ell)]^2} \tag{5.5}$$

$$m_{\mathrm{T}}(\mathrm{W}) = \sqrt{[\not{E}_{\mathrm{T}} + p_{\mathrm{T}}(\ell)]^2 - p_{\mathrm{T}}(\mathrm{W})^2}. \tag{5.6}$$

The transverse momentum of the W boson is used to divide the analysis into three orthogonal regions increasing the search sensitivity. When comparing the distribution of $p_{\mathrm{T}}(\mathrm{W})$ in the highly boosted regime a mismodeling between data and MC is visible. This discrepancy is fixed by applying an event weight, described in the following.

Vector boson p_{T} reweighting

In data the $p_{\mathrm{T}}(\mathrm{W})$ spectrum is observed to be softer than in MC. This behavior is assumed to originate from higher order electroweak corrections to the vector boson simulation and needs to be corrected for. By fitting the ratio between data and MC in a W+LF dominated region a p_{T} dependent scale factor is extracted, that is applied to W + jets and $\mathrm{t\bar{t}}$ MC. The numerical values used in this analysis are taken from [153]:

$$SF = 1 - 0.0011 \cdot (p_{\mathrm{T}}(\mathrm{W})/\,\mathrm{GeV} - 170)\,. \tag{5.7}$$

In Figure 5.5 the improvements using this method in terms of data/MC agreement are shown for the W(eν)H channel.

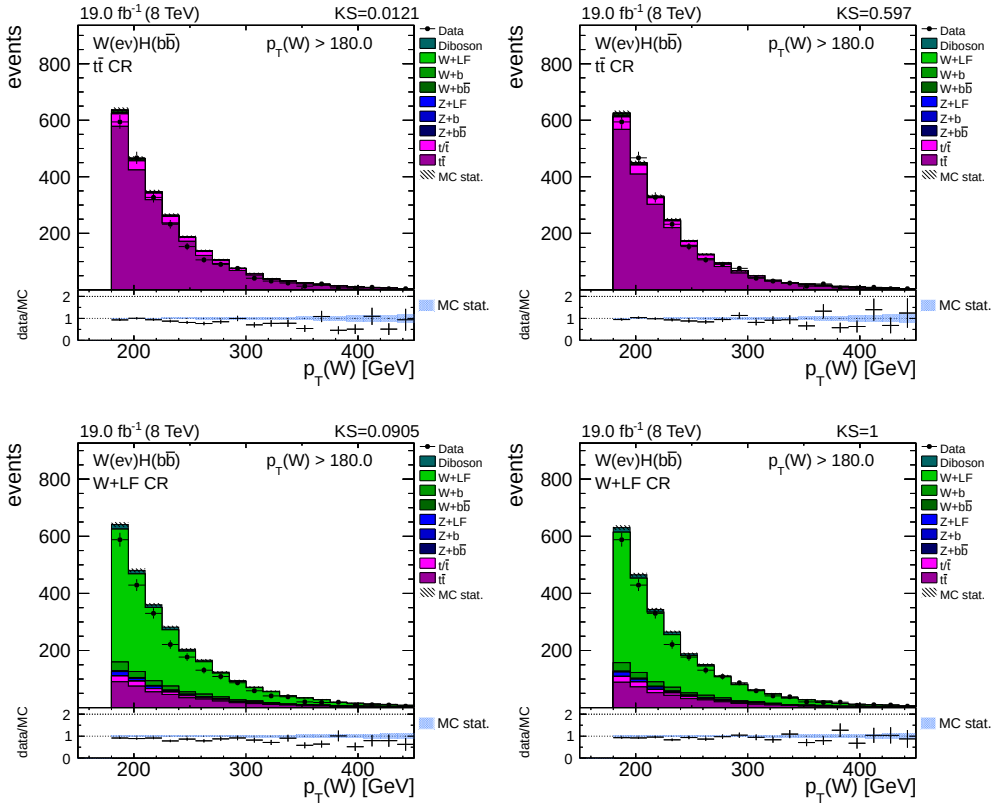


Figure 5.5.: Effect of vector boson p_{T} reweighting illustrated in the high p_{T} W(eν)H channel. Simulation is scaled to luminosity and all scale factors are applied. In the left column the data/MC agreement is displayed for the $\mathrm{t\bar{t}}$ (top row) and the W+LF control region (bottom row). The control regions will be explained in Section 5.6. For both regions the KS-tests yield much better probabilities after applying the calculated p_{T}(W)-dependent weights, as shown in the right column.

5.4.3. Higgs boson reconstruction

The reconstruction of the Higgs boson candidate is of uttermost importance in this thesis. Since the assignment of jets to the two b quarks from the Higgs boson decay is a combinatorial issue, an event-by-event criterion for making the decision is needed. In principle, the better the mass resolution is, the better signal and irreducible background processes can be separated.

In the main approach, the two central standard jets whose four-vector sum $\vec{j_1}+\vec{j_2}$ yields the highest transverse momentum are assigned to the Higgs boson candidate. Each of the jets needs to have $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and is required to lie well within the tracker region ($|\eta| < 2.4$) to be considered. The system of two jets is hereafter referred to as *dijet system* and the corresponding variables are indexed with jj.

Alternatively to the dijet construction, another Higgs boson candidate is built using substructure information via the SJF algorithm introduced in Section 3.2.5. The filter jets have a smaller cone size compared to the standard jets, so they are more robust against pile-up and supposedly lead to a better mass resolution of the Higgs boson candidate. It is tried to obtain additional information to discriminate signal from background processes out of this alternative reconstruction technique.

To simplify the reconstruction, the three hardest filter jets are assigned to the Higgs boson candidate. In cases, where only two filter jets are found, only those are assigned. The alternative Higgs boson candidate is hereafter referred to as *trifilter jet system* and indexed with flt in the variables. There is no additional event requirement introduced, since ideally for every event used in the standard jet analysis the substructure information is added. However, in about 35% of all cases in the high p_{T} region no filter jets are found. In these cases a pseudo trifilter jet system is built, where all relevant variables are set to default values.

To improve the mass resolution of Higgs boson candidates, a regression technique was adopted within the scope of the CMS analysis. First, the procedure on standard jets, provided centrally from the *VHbb* group, is outlined. A similar technique for filter jets, developed in this thesis, is explained in the next section.

Regression of standard jets

Due to the fact that in 20% of all cases neutrinos are present in B hadron decays leading to missing energy within the jets, reconstructed jets stemming from b quarks usually have a worse energy resolution compared to light flavor quark and gluon induced jets. A dedicated BDT is trained on the true transverse momenta of b jets in simulated signal events to compute correction factors for individual jets. This procedure was already established in analyses at the CDF experiment [167].

The training is set up with the TMVA package. To avoid a mass bias, signal samples generated with different Higgs boson masses from 110 to 150 GeV are used. Only jets satisfying $p_{\mathrm{T}} > 20\,\mathrm{GeV}$, $|\eta| < 2.4$ and CSV > 0 enter the training. Information such as jet properties, b-tag and soft lepton kinematics are taken into

Table 5.2.: Input variables for the regression of standard jets with explanation, as used in [153]. Unless noted otherwise, jet-energy corrections are applied.

Variable	Description
raw p_{T}	transverse momentum of the jet before jet energy corrections
p_{T}	transverse momentum of the jet
E_{T}	transverse energy of the jet
m_{T}	transverse mass of the jet
ptLeadTrk	transverse momentum of the leading track in the jet
vtx3dL	3-d flight length of the jet's secondary vertex
vtx3deL	error on the 3-d flight length of the jet's secondary vertex
vtxMass	mass of the jet's secondary vertex
vtxPt	transverse momentum of the jet's secondary vertex
JECUnc	uncertainty on the JEC
Ntot	total number of jet constituents
cef	charged EM fraction of jet
SoftLeptPtRel	relative transverse momentum of soft lepton candidate in the jet
SoftLeptPt	transverse momentum of soft lepton candidate in the jet
SoftLeptdR	distance in $\eta - \phi$ space of soft lepton candidate with respect to the jet axis

account as input to the training. Table 5.2 shows the full list of variables. The most discriminating variables are based on kinematic information, since they correct for the mismeasured b jet energy the strongest. In Figure 5.6 the left diagram illustrates the regression performance in terms of Higgs boson mass resolution for a statistically independent signal sample in the high p_{T} signal region. An improvement of 6% compared to uncorrected jets is found by applying the jet correction. The right figure shows the small difference between nominal and corrected mass distributions also for diboson and W+jets background processes.

The standard jet regression has gone through a detailed validation procedure in the *VHbb* group that lies outside the scope of this thesis. The studies can be found in [161]. However, to validate the filter jet regression similar techniques are adapted. These are presented in the following section.

5.5. Regression of filter jets

Though originally intended to correct for mismeasured jet energy in AK5 standard jets, the regression has an additional advantage for filter jets. As mentioned in

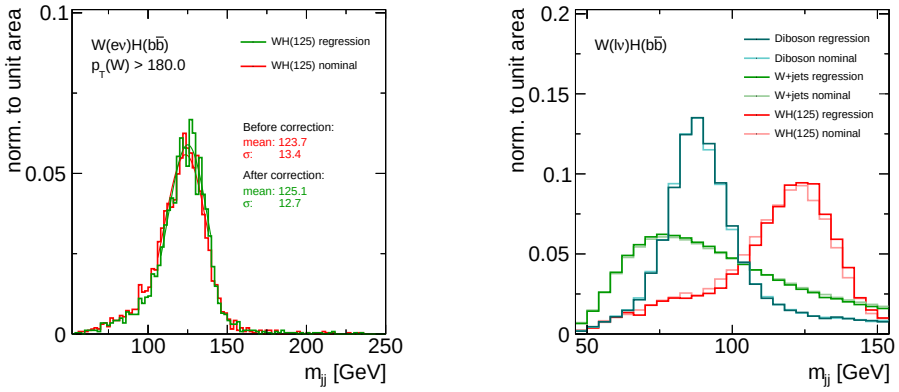


Figure 5.6.: Mass resolution improvement in signal MC (left) and mass distribution for signal, diboson and W+jets processes (right). In the left diagram the mass improvement for the 125 GeV signal template in the high p_{T} electron channel is found to be 6%. In the other regions the behavior is comparable. The right graph shows that the mass distribution of background processes are not visibly affected.

Section 3.2.5, there is a non negligible mass bias when constructing a Higgs boson candidate from filter jets, due to the lack of dedicated energy corrections. Together with improving the mass resolution of the trifilter jet system, this mass bias is removed making the application of regression for filter jets very powerful.

Again, the TMVA toolkit is used for the training. The exact configuration of the training can be found in Table A.8 in Appendix A. Similarly to the standard jet regression, signal templates with different masses are used. It is worth noting here, that the correction of the mass bias is a result of the jet-by-jet correction and not the primary target. The training target for the filter jet regression is the ratio $p_{\mathrm{T}}^{\mathrm{gen}}/p_{\mathrm{T}}^{\mathrm{rec}}$. This way the training is found to be more robust against overtraining compared to $p_{\mathrm{T}}^{\mathrm{gen}}$ as target. To obtain reasonable results it is also important to incorporate a part of the phase space in which the SJF algorithm works properly. That is why boost requirements on the trifilter jet system and on the W boson candidates are applied, i.e. $p_{\mathrm{T},jj}^{\mathrm{flt}} > 120\,\mathrm{GeV}$ and $p_{\mathrm{T}}(\mathrm{W}) > 120\,\mathrm{GeV}$. Moreover, only filter jets from events in which at least two filter jets are present, enter the training. Each filter jet is required to fulfill $p_{\mathrm{T}}(j^{\mathrm{flt}}) > 20\,\mathrm{GeV}$ and $\mathrm{CSV}(j^{\mathrm{flt}}) > 0$. The filter jet information taken as input is chosen in analogy to the standard jet regression. An improved performance is found by using additionally the filter jet area and the angle between the filter jet and the trifilter jet system. The soft lepton kinematics are not considered, due to missing dedicated studies for modeling and performance. In Table 5.3 all input variables are summarized.

The resulting correction factors for each filter jet are shown in Figure 5.7. The diagram illustrates how the regression performs for training and testing samples. Both distributions agree very well, so an overtrained regression can be excluded.

Beyond this first comparison, it is crucial to check for differences in performance

Table 5.3.: Input variables for the regression of filter jets with explanation. Unless noted otherwise, the standard jet-energy corrections are applied.

Variable	Description
raw $p_{\mathrm{T}}^{\mathrm{flt}}$	transverse momentum of the filter jet before corrections
$p_{\mathrm{T}}^{\mathrm{flt}}$	transverse momentum of the filter jet
$E_{\mathrm{T}}^{\mathrm{flt}}$	transverse energy of the filter jet
$m_{\mathrm{T}}^{\mathrm{flt}}$	transverse mass of the filter jet
ptLeadTrk$^{\mathrm{flt}}$	transverse momentum of the leading track in the filter jet
vtx3dL$^{\mathrm{flt}}$	3-d flight length of the filter jet's secondary vertex
vtx3deL$^{\mathrm{flt}}$	error on the 3-d flight length of the filter jet's secondary vertex
vtxMass$^{\mathrm{flt}}$	mass of the filter jet's secondary vertex
vtxPt$^{\mathrm{flt}}$	transverse momentum of the filter jet's secondary vertex
JECUnc$^{\mathrm{flt}}$	uncertainty on the JEC
Ntot$^{\mathrm{flt}}$	total number of filter jet constituents
cef$^{\mathrm{flt}}$	charged EM fraction of filter jets
jetArea$^{\mathrm{flt}}$	final reconstructed area of filter jet
$\cos\theta^{\mathrm{flt}}$	angle between filter jet and trifilter jet system

for data and MC, as the correction factors are applied on all jets in all samples. A first important validation is shown in Figure 5.8, namely the data/MC comparisons for the correction factors in the W+HF control region. In all analysis regions excellent agreement is found.

On a jet-by-jet basis the performance of the regression can be displayed by the relative difference between generated and reconstructed p_{T} over the reconstructed p_{T}. Figure 5.9 shows the distributions before and after applying the regression correction. In the diagrams it is visible that after applying the regression not only the bias in p_{T} is canceled out, but also the RMS value for the y-axis decreased.

One way to check the performance on an event-by-event basis is to look at the change in mass resolution for the reconstructed Higgs boson candidates. Figure 5.10 shows the invariant mass distributions with and without regression for signal MC. The mass resolution is improved by 15%. Moreover, the distribution peaks at 125 GeV after applying the regression weights, so the mass bias is removed.

A validation introduced for the standard jet regression, see e.g. [161], is adapted in this analysis for the filter jet regression. In a MC $\mathrm{t\bar{t}}$ sample the regression performance is checked by selecting semi-leptonically decaying top quark pairs. Besides requiring one isolated muon with 20 GeV, no additional isolated lepton with 15 GeV

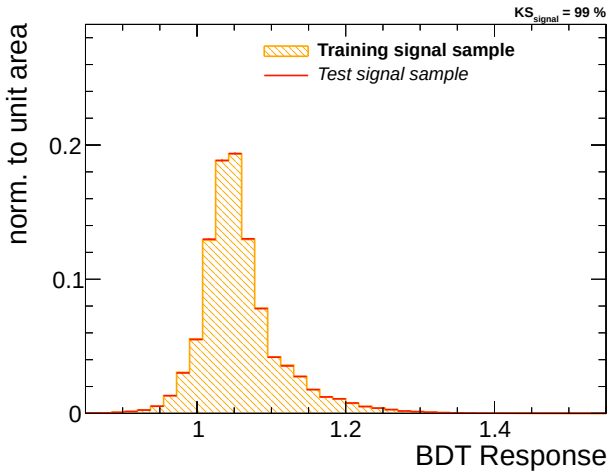


Figure 5.7.: Filter jet regression correction factors. The BDT response is shown separately for the training and an independent testing sample. When comparing the performance of the regression between the two samples, no difference can be found. The KS-test yields a probability of 1.

and $\not{E}_{\mathrm{T}} > 50\,\mathrm{GeV}$, each event must have exactly four filter jets with $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ and $|\eta| < 2.4$. Two of these four filter jets need to be b-tagged (CSV > 0.6). The two untagged filter jets are assigned to the hadronically decaying W boson ($\mathrm{W_{had}}$). Only if $|m_{\mathrm{W_{had}}} - 80.4| < 5\,\mathrm{GeV}$ the event is taken into account. The filter jet with the higher CSV value is assigned together with $\mathrm{W_{had}}$ to the hadronically decaying top quark.

The data/MC comparisons for the reconstructed mass of $\mathrm{t_{had}}$ are shown before and after applying the correction in Figure 5.11(a) and (b). Any undesirable side effects of the regression would be visible in these diagrams, but no bias is found. To quantify the effect of the regression in this scenario the mass resolution in $\mathrm{t\bar{t}}$ MC and in data events is checked. The resolution is improved by 5% in data, and by 9% in $\mathrm{t\bar{t}}$ MC. For both, the mean values get closer to the nominal top quark mass.

These thorough investigations are preformed for the first time in the scope of this thesis. They justify the use of filter jet regression in the further analysis.

5.6. Event selection and background estimation

The precise modeling of all occurring background processes is another crucial part of this analysis. After introducing the kinematic region in which the signal is extracted in the end, three control regions are defined, that are enhanced in one particular major background. In these control regions the data-driven estimation of the normalization for the major background processes is performed. Eventually, the `ABCD` method is explained, that is used to predict the contribution from QCD multijet production.

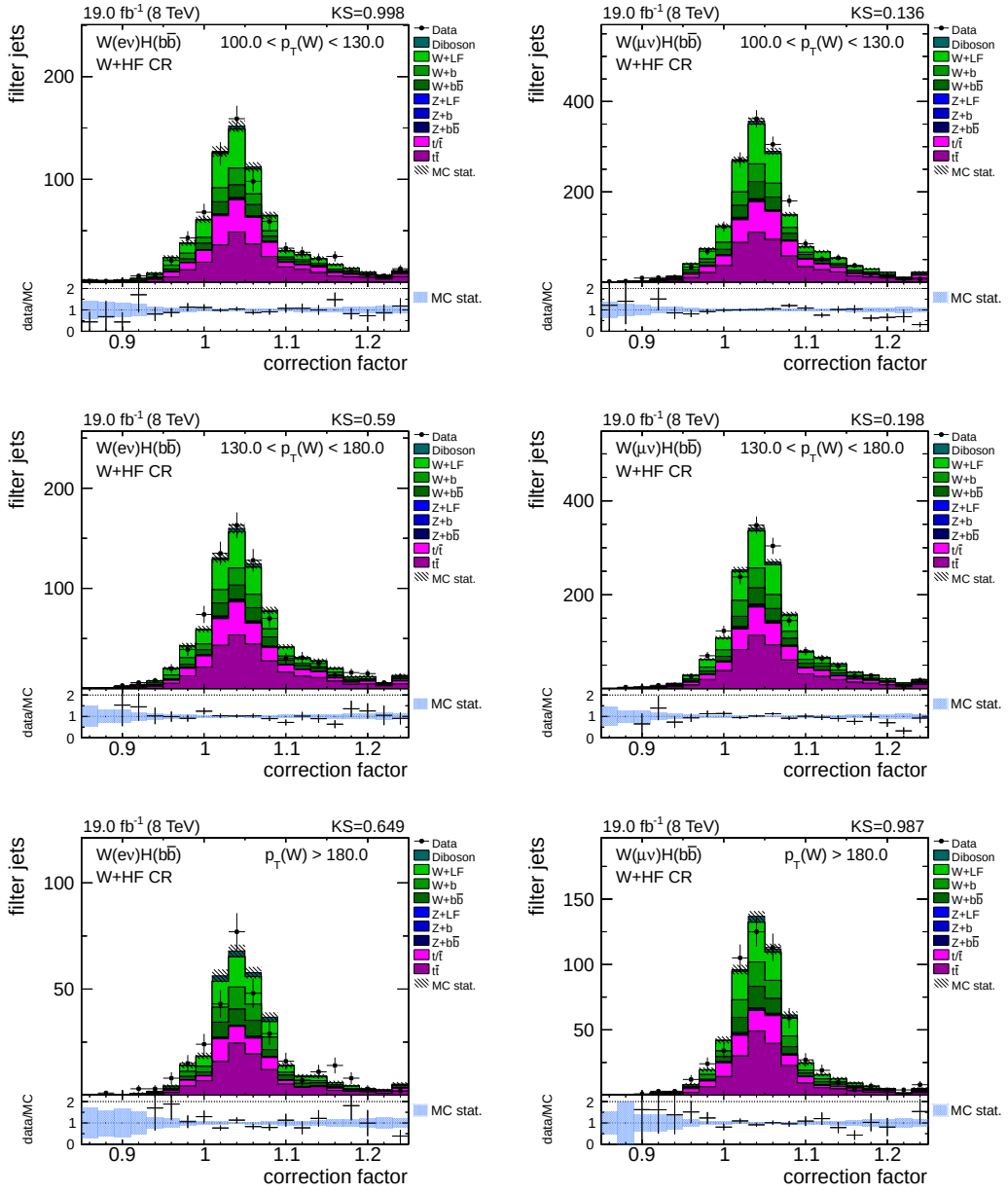


Figure 5.8.: Filter jet correction factors in the W+HF control region for the W(eν)H (left column) and the W(μν)H channel (right column). Data/MC comparisons are shown for the low p_{T} region (top row), intermediate p_{T} region (middle row) and the high p_{T} region (bottom row). Overall good agreement is found, indicated by the Kolmogorov-Smirnov-test probabilities printed on the upper right corner of each figure.

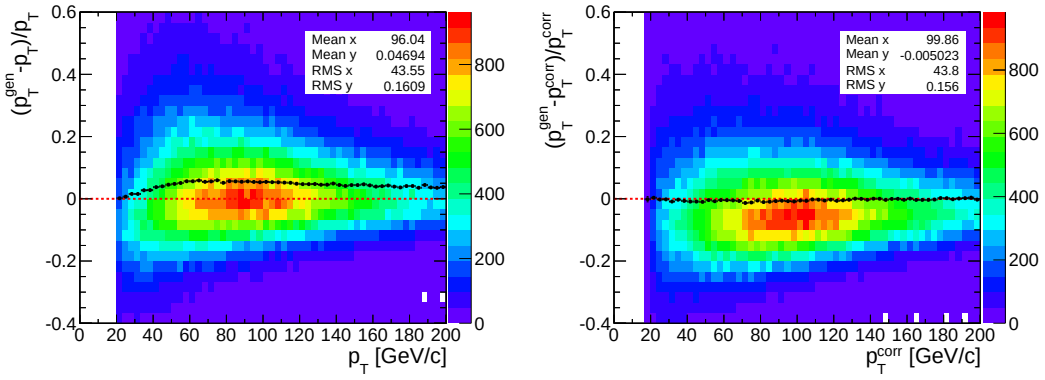


Figure 5.9.: Performance of filter jet regression on individual jet momenta. On the left side the relative difference between generated and reconstructed p_{T} over reconstructed p_{T} is shown before the correction. The black markers depict the mean values of each bin in p_{T} (x-axis). A tendency towards values greater than zero is visible. The same diagram after applying the regression weights is shown on the right side. On the one hand, the mean values are now compatible with zero, and on the other hand the RMS value for the y -axis decreased. Interestingly, the relative difference between generated and reconstructed p_{T} peaks at zero before the correction. The improvements with respect to the mean and RMS values after applying the regression weight arise mainly from the corrections of the tails of these distributions.

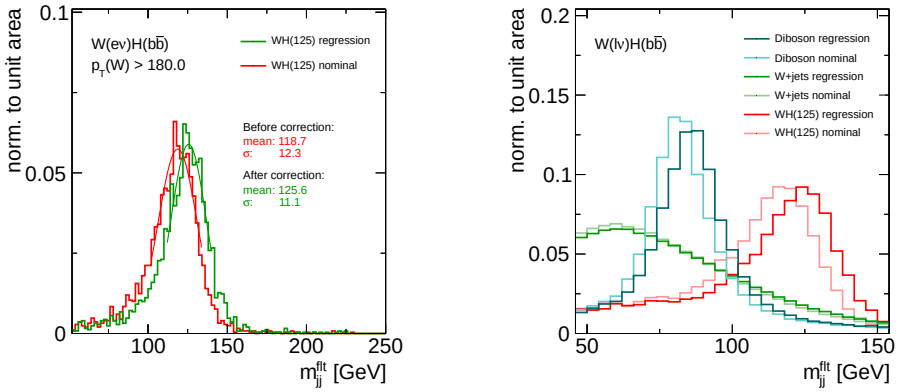
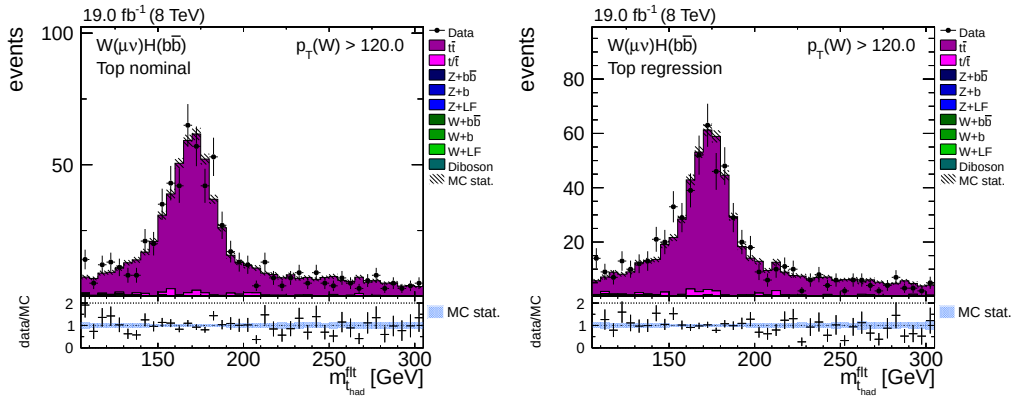


Figure 5.10.: Performance diagrams for filter jet regression. The mass resolution for signal MC (left) and mass distributions for signal, diboson and W+jets processes (right) are depicted. In the left figure the mass improvement for the 125 GeV signal template in the high p_{T} electron channel is found to be 15%. The performance is comparable in other regions. The right diagram shows that the mass bias is not only resolved for the signal process, but also for diboson production.



(a) Invariant mass before correction

(b) Invariant mass after correction

(c) Resolution change in data

(d) Resolution change in $t\bar{t}$ MC

Figure 5.11.: Different validation checks for filter jet regression in $t\bar{t}$ events as described in the text. The invariant mass of the reconstructed hadronically decaying top quark is shown before (a) and after applying the correction (b). It is visible that the regression does not introduce a bias. Furthermore, the mass resolution is compared before and after the correction for data in (c) and $t\bar{t}$ MC in (d). In both cases, the resolution is improved and the mean values are shifted more towards the nominal top quark mass.

5.6.1. Signal region

To obtain the desired boosted regime both, the Higgs boson and the W boson candidates, need to fulfill $p_{\mathrm{T}} > 100\,\mathrm{GeV}$. The two jets assigned to the Higgs boson are required to have a transverse momentum larger than 30 GeV and to be b-tagged with CSV > 0.4. To increase the signal-over-background ratio every event in the signal region needs to have exactly one isolated electron or muon and no additional isolated lepton with $p_{\mathrm{T}} > 15\,\mathrm{GeV}$ and $|\eta| < 2.5$. The QCD contribution is further suppressed by the cut $\not{E}_{\mathrm{T}} > 45\,\mathrm{GeV}$, that is motivated later. The angle between $\not{\vec{p}}_{\mathrm{T}}$ and the charged lepton should be smaller than $\pi/2$.

To enhance the search sensitivity even more the events are categorized into three orthogonal regions, determined by the transverse momentum of the W boson candidates. A low p_{T} region, $p_{\mathrm{T}}(\mathrm{W}) < 130\,\mathrm{GeV}$, an intermediate p_{T} region, $130\,\mathrm{GeV} < p_{\mathrm{T}}(\mathrm{W}) < 180\,\mathrm{GeV}$, and a high p_{T} region, $p_{\mathrm{T}}(\mathrm{W}) > 180\,\mathrm{GeV}$, are defined such that they have roughly the same amount of events in data. When extracting the upper limits, the three regions are statistically combined. The lepton isolation is tightened to 0.075 in the low p_{T} region to avoid too large contributions from QCD multijet production. Table 5.4 summarizes all selection requirements of the three signal regions.

Table 5.4.: Selection criteria for the signal regions used in this analysis. The entries in parentheses indicate the selection for the intermediate and high $p_{\mathrm{T}}(\mathrm{W})$ regions. The jets assigned to the Higgs boson candidate are labeled with j_1 and j_2 and sorted by their transverse momenta. $\mathrm{N_{al}}$ is the number of additional isolated leptons in the event. The thresholds listed for p_{T}, $\not{E}_{\mathrm{T}}$ and m are in units of GeV.

Variable	W($\ell\nu$)H
$p_{\mathrm{T}}(j_1)$	> 30
$p_{\mathrm{T}}(j_2)$	> 30
$p_{\mathrm{T},jj}$	> 100
m_{jj}	< 250
$p_{\mathrm{T}}(\mathrm{V})$	$100 - 130$ $(130 - 180, > 180)$
$\mathrm{CSV_{max}}$	> 0.40
$\mathrm{CSV_{min}}$	> 0.40
$\mathrm{N_{al}}$	$= 0$
$\not{E}_{\mathrm{T}}$	> 45
$\Delta\phi(\not{E}_{\mathrm{T}}, \ell)$	$< \pi/2$
Tightened Lepton Iso.	< 0.075 (–, –)

5.6.2. Control regions

The selection requirements of the signal regions are slightly altered to obtain orthogonal regions, that can be used for further estimations without being biased towards the final result. These control regions are chosen such that a high purity in the most important background processes — production of a W boson with jets and $t\bar{t}$ — is reached, while the kinematic characteristics are still similar to the signal region. In analogy to the signal regions, events in the control regions are categorized according to the p_T of the W boson candidate (low p_T, intermediate p_T, high p_T). Since there are in total nine control regions, only a subset of all diagrams checked for scrutinizing the analysis is shown. The background enriched phase spaces, described in the following, are exploited to validate the shape of the input variables of the BDT trainings in data and to determine the normalization factors for these processes.

- **W+LF control region:** One major background process is the production of a W boson with light jets. The b-tagging requirements for the two jets assigned to the Higgs boson are relaxed. Furthermore, the number of additional jets in the event is asked to be smaller than two to suppress $t\bar{t}$ production. Since otherwise the contamination from QCD production would be too large, an additional cut on the $\not{E}_T$ significance is introduced.
- **$t\bar{t}$ control region:** Another major background contribution arises from semileptonically decaying top quark pairs. To obtain a CR rich in $t\bar{t}$ production it is sufficient to ask for more additional jets in the events compared to the signal region.
- **W+HF control region:** When W bosons are produced in association with jets stemming from b quarks, the topology is similar to that of the signal process. A W+heavy flavor enriched control region is obtained by applying a mass veto in the range of the Higgs boson mass.

The full list of selection criteria for each CR is shown in Table 5.5. To display how well simulation describes the data, Figure 5.12 gives the data/MC comparisons for some event variables in the W($e\nu$)H channel in different control regions. The corresponding diagrams for the W($\mu\nu$)H channel can be found in Figure A.1 in the appendix.

Tables 5.6 - 5.8 show important information on the control regions. The yields for every single process are listed and compared to the number of events in data. Furthermore, the purity in the desired background process of each control region is given.

5.6.3. Scale factor determination

Scale factors for the main backgrounds, i.e. $t\bar{t}$ and W boson production with zero, one or two additional b jets, are estimated in a data-driven way. Separately for

Table 5.5.: Selection criteria for the W(eν)H and W(μν)H control regions in the low, intermediate, and high p_T(W) regions. The values in parentheses are used for the intermediate and high p_T(W) regions. LF and HF refer to light- and heavy-flavor jets. N_{al} is the number of additional isolated leptons in the event and N_{aj} denotes the number of additional jets besides the two Higgs boson daughters. $\not{E}_T$ sig. is the significance of the missing transverse energy of the event. The values listed for kinematic variables are in units of GeV.

Variable	W+LF	$t\bar{t}$	W+HF
$p_T(j_1)$	> 30	> 30	> 30
$p_T(j_2)$	> 30	> 30	> 30
$p_{T,jj}$	> 100	> 100	> 100
p_T(W)	$100 - 130\,(130 - 180, > 180)$	$100 - 130\,(130 - 180, > 180)$	$100 - 130\,(130 - 180, > 180)$
CSV_{max}	$[0.244 - 0.898]$	> 0.898	> 0.898
N_{aj}	< 2	> 1	$= 0$
N_{al}	$= 0$	$= 0$	$= 0$
$\not{E}_T$	> 45	> 45	> 45
$\not{E}_T$ sig.	$> 2.0(\mu)\ > 3.0(e)$	–	> 2.0
$\Delta\phi(\not{E}_T, \ell)$	$< \pi/2$	$< \pi/2$	$< \pi/2$
m_{jj}	< 250	< 250	veto $[90 - 150]$

Table 5.6.: Predicted yields in the low p_T control regions. MC is normalized to luminosity without additional scale factors. Only statistical uncertainties are shown. To get an estimate how pure the CR is in the desired processes, in addition the purity is given.

	W+LF		$t\bar{t}$		W+HF	
Process	W(μν)H	W(eν)H	W(μν)H	W(eν)H	W(μν)H	W(eν)H
$t/\bar{t}$	487.6 ± 12.3	112.3 ± 6.4	585.2 ± 13.4	474.4 ± 12.7	188.1 ± 7.4	146.9 ± 6.9
$t\bar{t}$	1466.7 ± 9.0	305.6 ± 4.4	8859.1 ± 23.3	7292.6 ± 22.1	330.2 ± 3.8	252.6 ± 3.5
VV	237.0 ± 3.5	54.0 ± 1.9	16.0 ± 0.7	12.3 ± 0.6	14.2 ± 0.7	11.3 ± 0.6
W+0b	12125.6 ± 40.4	2744.8 ± 20.4	63.4 ± 2.9	47.3 ± 2.6	231.4 ± 5.7	174.5 ± 5.1
W+1b	207.2 ± 5.2	44.0 ± 2.5	22.9 ± 1.7	16.9 ± 1.4	98.6 ± 3.6	66.0 ± 3.1
W+2b	98.6 ± 3.6	14.9 ± 1.4	32.0 ± 2.0	20.4 ± 1.6	97.2 ± 3.5	63.7 ± 3.0
Z+0b	472.9 ± 5.0	38.0 ± 1.6	10.2 ± 0.7	5.2 ± 0.6	5.6 ± 0.5	3.6 ± 0.5
Z+1b	20.4 ± 1.1	2.0 ± 0.4	12.8 ± 0.8	7.8 ± 0.7	7.3 ± 0.6	4.9 ± 0.6
Z+2b	6.6 ± 0.6	0.6 ± 0.2	16.8 ± 1.0	9.9 ± 0.7	3.7 ± 0.5	1.6 ± 0.3
Purity	(80.2 ± 0.9) %	(82.8 ± 0.9) %	(92.1 ± 0.9) %	(92.5 ± 0.9) %	(20.1 ± 0.6) %	(17.9 ± 0.6) %
Total MC	15123 ± 44	3316 ± 22	9618 ± 27	7887 ± 26	976 ± 11	725 ± 10
Data	14713	2941	9984	7907	1141	776

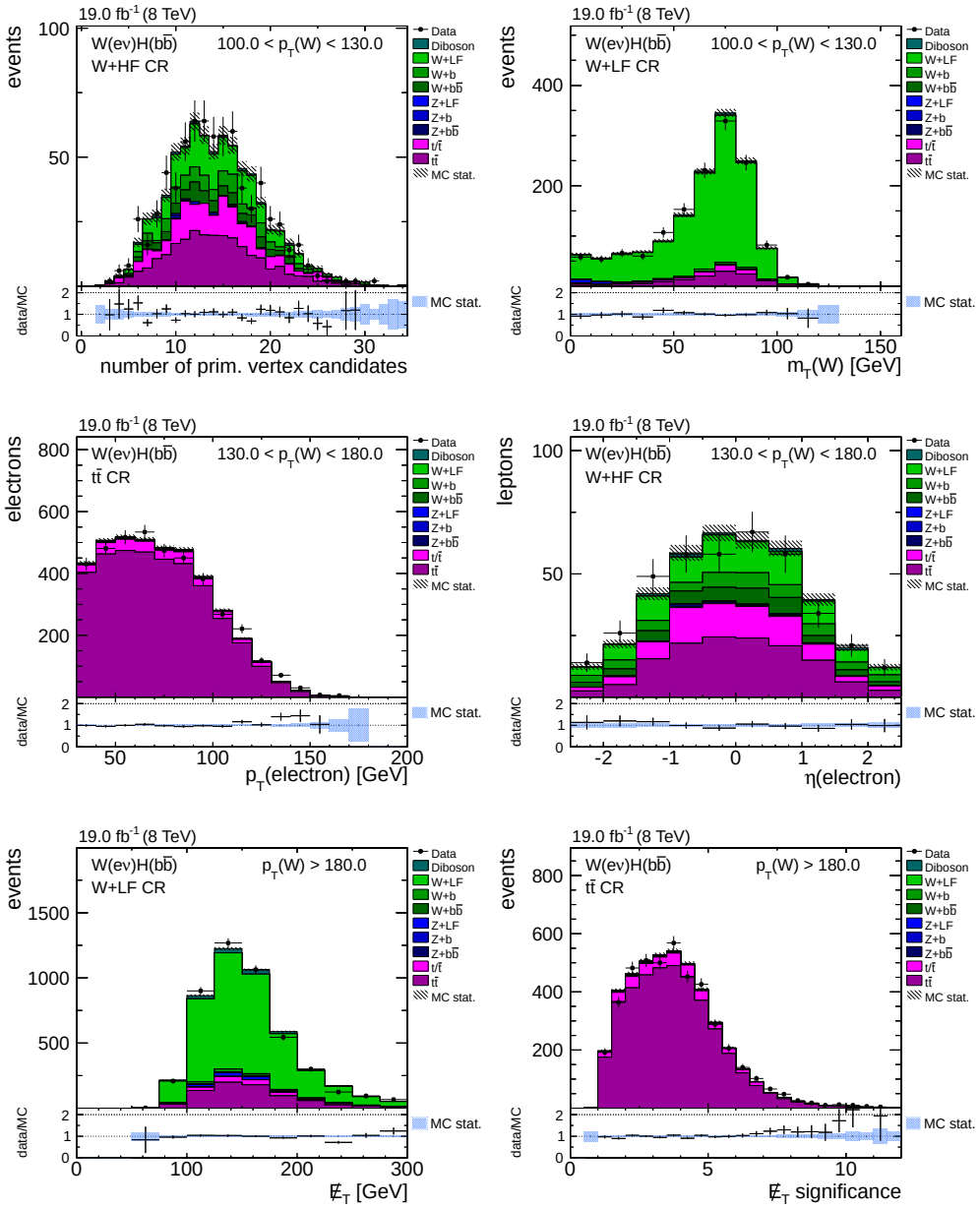


Figure 5.12.: Different event distributions in control regions for the W(eν)H channel. Data/MC comparisons are shown for the low p_{T} region (top row), intermediate p_{T} region (middle row) and the high p_{T} region (bottom row) in all three control regions as indicated in the figures. The number of events in simulation is normalized to data. The corresponding distributions for the W(μν)H channel are presented in Figure A.1 in the appendix.

Table 5.7.: Predicted yields in the intermediate p_T control regions. MC is normalized to luminosity without additional scale factors. Only statistical uncertainties are shown. To get an estimate how pure the CR is in the desired processes, in addition the purity is given.

	W+LF		$t\bar{t}$		W+HF	
Process	W(μν)H	W(eν)H	W(μν)H	W(eν)H	W(μν)H	W(eν)H
$t/\bar{t}$	552.3 ± 13.3	231.1 ± 9.1	658.0 ± 14.5	581.3 ± 14.2	167.5 ± 7.0	128.1 ± 6.3
$t\bar{t}$	2038.3 ± 10.6	897.0 ± 7.4	9610.6 ± 24.0	8127.9 ± 22.8	329.7 ± 3.8	268.5 ± 3.6
VV	340.3 ± 4.2	149.7 ± 3.0	21.6 ± 0.8	18.5 ± 0.8	16.3 ± 0.7	14.3 ± 0.7
W+0b	13543.7 ± 37.5	5890.4 ± 26.2	61.1 ± 2.4	51.7 ± 2.3	215.6 ± 4.8	182.0 ± 4.6
W+1b	278.2 ± 5.3	121.7 ± 3.7	28.0 ± 1.6	25.7 ± 1.6	106.8 ± 3.3	83.9 ± 3.0
W+2b	115.5 ± 3.4	54.9 ± 2.5	36.7 ± 1.8	27.9 ± 1.7	92.2 ± 3.0	76.3 ± 2.9
Z+0b	489.8 ± 4.9	135.6 ± 2.8	9.0 ± 0.7	5.2 ± 0.5	6.4 ± 0.5	4.6 ± 0.5
Z+1b	21.6 ± 1.0	6.1 ± 0.6	13.6 ± 0.8	8.1 ± 0.7	8.2 ± 0.7	5.9 ± 0.6
Z+2b	6.7 ± 0.6	1.8 ± 0.3	15.2 ± 0.9	10.0 ± 0.7	3.7 ± 0.4	2.1 ± 0.3
Purity	77.9 ± 0.9	78.7 ± 0.9	91.9 ± 0.8	91.8 ± 0.8	21.0 ± 0.6	20.9 ± 0.6
Total MC	17387 ± 42	7488 ± 29	10454 ± 28	8856 ± 27	946 ± 10	766 ± 10
Data	16282	6514	10674	8666	1071	864

Table 5.8.: Predicted yields in the high p_T control regions. MC is normalized to luminosity without additional scale factors. Only statistical uncertainties are shown. To get an estimate how pure the CR is in the desired processes, in addition the purity is given.

	W+LF		$t\bar{t}$		W+HF	
Process	W(μν)H	W(eν)H	W(μν)H	W(eν)H	W(μν)H	W(eν)H
$t/\bar{t}$	254.2 ± 9.2	188.8 ± 8.4	475.7 ± 12.8	407.4 ± 12.2	58.3 ± 4.2	41.8 ± 3.7
$t\bar{t}$	1284.8 ± 8.5	847.4 ± 7.1	5116.9 ± 17.2	4548.5 ± 16.7	115.7 ± 2.4	96.1 ± 2.2
VV	236.6 ± 3.5	158.0 ± 3.0	14.7 ± 0.7	16.9 ± 0.8	8.5 ± 0.5	10.6 ± 0.6
W+0b	5836.9 ± 18.4	3829.3 ± 15.4	24.4 ± 1.1	21.7 ± 1.1	69.8 ± 2.0	59.7 ± 1.9
W+1b	136.1 ± 2.8	90.6 ± 2.3	13.5 ± 0.8	13.2 ± 0.8	42.5 ± 1.6	36.7 ± 1.5
W+2b	65.6 ± 1.9	40.2 ± 1.5	17.8 ± 1.0	15.4 ± 0.9	36.0 ± 1.4	33.3 ± 1.4
Z+0b	173.5 ± 3.0	105.6 ± 2.5	4.9 ± 0.5	2.9 ± 0.4	2.6 ± 0.4	1.7 ± 0.3
Z+1b	6.7 ± 0.6	4.0 ± 0.5	4.9 ± 0.5	4.1 ± 0.5	2.4 ± 0.4	1.7 ± 0.3
Z+2b	2.3 ± 0.3	1.2 ± 0.3	6.2 ± 0.5	4.7 ± 0.5	0.9 ± 0.2	0.7 ± 0.2
Purity	73.0 ± 0.8	72.7 ± 0.8	90.1 ± 0.8	90.3 ± 0.8	23.3 ± 0.4	24.8 ± 0.5
Total MC	7997 ± 23	5265 ± 20	5679 ± 22	5035 ± 21	336 ± 6	282 ± 5
Data	8011	4921	5486	4792	380	338

each analysis bin, a maximum likelihood fit to data is performed in discriminating variables, simultaneously for W(eν)H and W(μν)H channels. The THETA package (see Section 4.1.4) is used. The priors for the scale factors of W+0b, W+1b, W+2b and $t\bar{t}$ templates are chosen flat in a range $(0, \infty)$. Priors for the other background processes are fixed to their nominal values. In the $t\bar{t}$ and the W+LF control regions the invariant dijet mass, and in the W+HF CR the CSV output of the jet with the second highest transverse momentum assigned to the Higgs boson are used as discriminating variables. These variables are found to yield the lowest uncertainties on the fit results. Figure 5.13 shows the distributions before and after applying the resulting scale factors exemplarily in the W(μν)H channel. The diagrams show already reasonable agreement before applying the scale factors, but with their inclusion the data/MC agreement improves.

The statistical uncertainty of the fit is given directly by THETA. To attain the systematic uncertainties the fit is performed again by using systematically shifted templates. This way, the influences of JER, JES and b-tagging are evaluated. The differences compared to the nominal fit are introduced in the signal extraction.

Figure 5.14 shows the resulting scale factors and their uncertainties for all regions. For the W+1b process the estimate is about two times larger than the prediction from simulation. This discrepancy is also found in other studies within the CMS collaboration and arises due to mismodeling of the generator parton shower when a gluon splits to a bottom quark pair. All other scale factors are within their uncertainties compatible with unity. The numerical values are given in Table A.4 in the appendix. Table A.5 shows the correlations between the fitted parameters separately for the three analysis bins.

5.6.4. Data-driven QCD estimation

Due to the lack of MC samples for QCD multijet production with a sufficient amount of events in the relevant kinematic regime a detailed study of this background process is ambitious. Therefore, an estimation of QCD events in the signal and control regions is performed via the `ABCD` method. The `ABCD` method requires two selection cuts which are assumed to be uncorrelated for QCD events. Here, requirements on $\not{E}_{\mathrm{T}}$ and the lepton isolation are used. In Section 6.5.2 their negligible correlation is validated. With these two cuts at hand one can define four regions as follows:

- **Region A**: ordinarily-defined signal or control region, where the amount of QCD events should be estimated.
- **Region B**: as region A, but the lepton isolation cut is inverted.
- **Region C**: as region A, but the $\not{E}_{\mathrm{T}}$ cut is inverted.
- **Region D**: both cuts described above are inverted.

An illustrative example of these four regions, for the $t\bar{t}$ CR in the electron channel, is shown in Figure 5.15. In the diagrams all simulated background process except

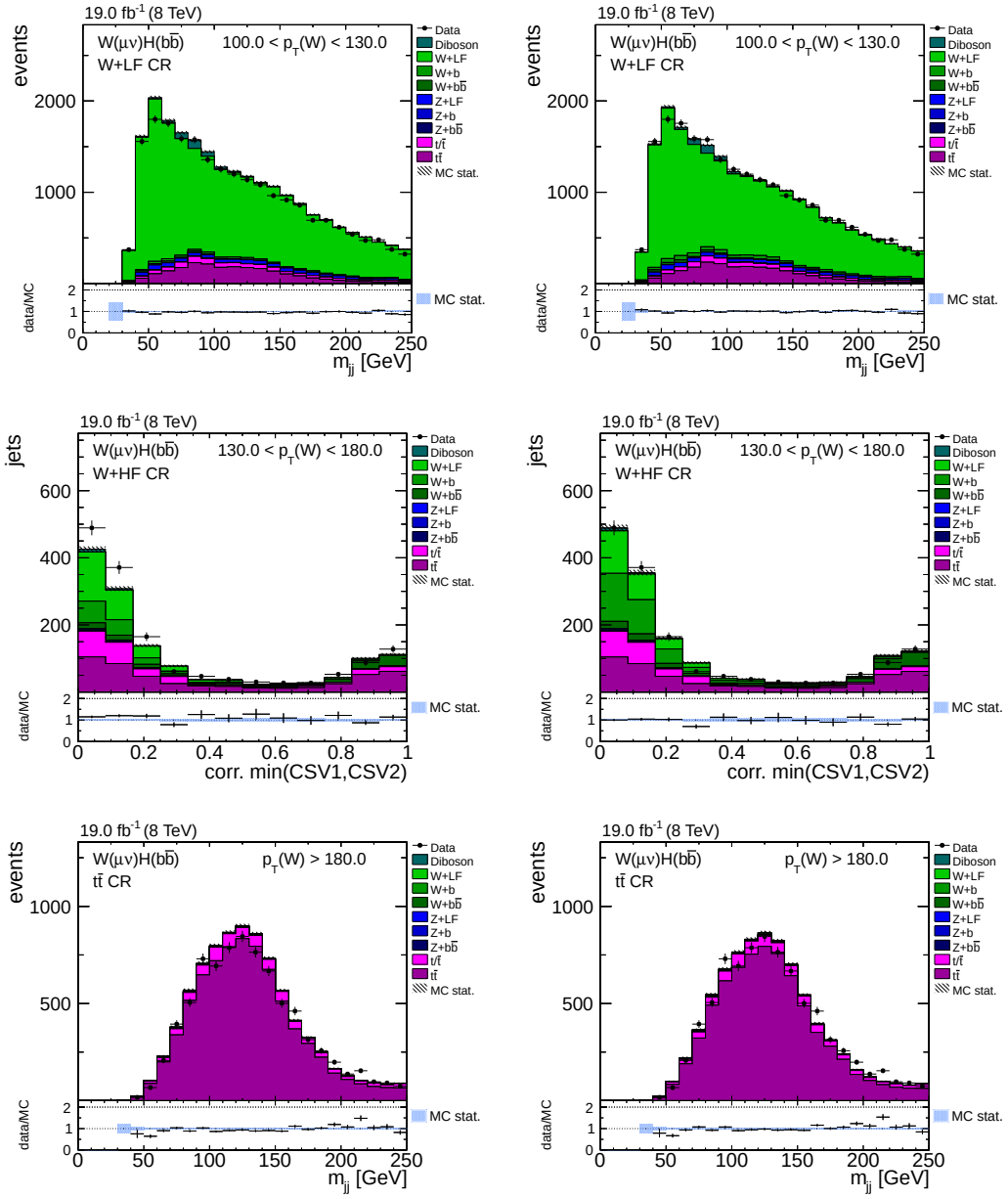


Figure 5.13.: Fitted distributions in the W(μν)H channel before (left column) and after applying the resulting scale factors (right column). Data/MC comparisons are shown for the low p_T region (top row), intermediate p_T region (middle row) and the high p_T region (bottom row) in all three control regions as indicated in the figures. The number of events in simulation is normalized to luminosity.

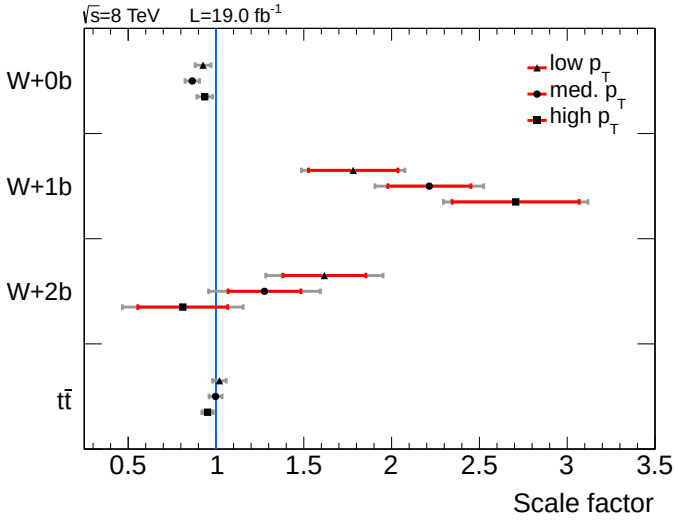


Figure 5.14.: Results of the scale factor estimation for all regions. The red and gray error bars denote the statistical and the total uncertainty on the fit, respectively. The discrepancy found in the values for W+1b arises due to generator mismodeling of the gluon splitting into a bottom quark pair.

QCD production are overlayed with data. Region D is predicted to be dominated by QCD production, and both regions B and C show a significant number of events beyond the predicted non-QCD background processes.

The basic idea is to extrapolate the amount of QCD events in region A from the amount of QCD events in regions B, C and D. The total number of QCD events in the primary region A can then be taken to be:

$$N_{\mathrm{A,QCD}} = \frac{(N_{\mathrm{B,data}} - N_{\mathrm{B,non\text{-}QCD}}) \cdot (N_{\mathrm{C,data}} - N_{\mathrm{C,non\text{-}QCD}})}{N_{\mathrm{D,data}} - N_{\mathrm{D,non\text{-}QCD}}}, \tag{5.8}$$

where $N_{\mathrm{R,data}}$ and $N_{\mathrm{R,non\text{-}QCD}}$ is the total number of data and non-QCD background events, respectively, in region R. Exemplarily, in Table 5.9 the numbers extracted in the $\mathrm{t\bar{t}}$ CR are shown. The `ABCD` method predicts a QCD contamination compatible with 0% in the signal and control regions across all channels and regions. The same result was also found in [161]. Based on these findings the contribution from QCD production is neglected in the following.

5.7. BDT analysis

The main goal of the analysis is the discrimination between signal and background processes. As described before, there are several event characteristics comprising separation power. To make the most of the correlations between these variables classification BDTs from the TMVA package are employed. In the end, the signal

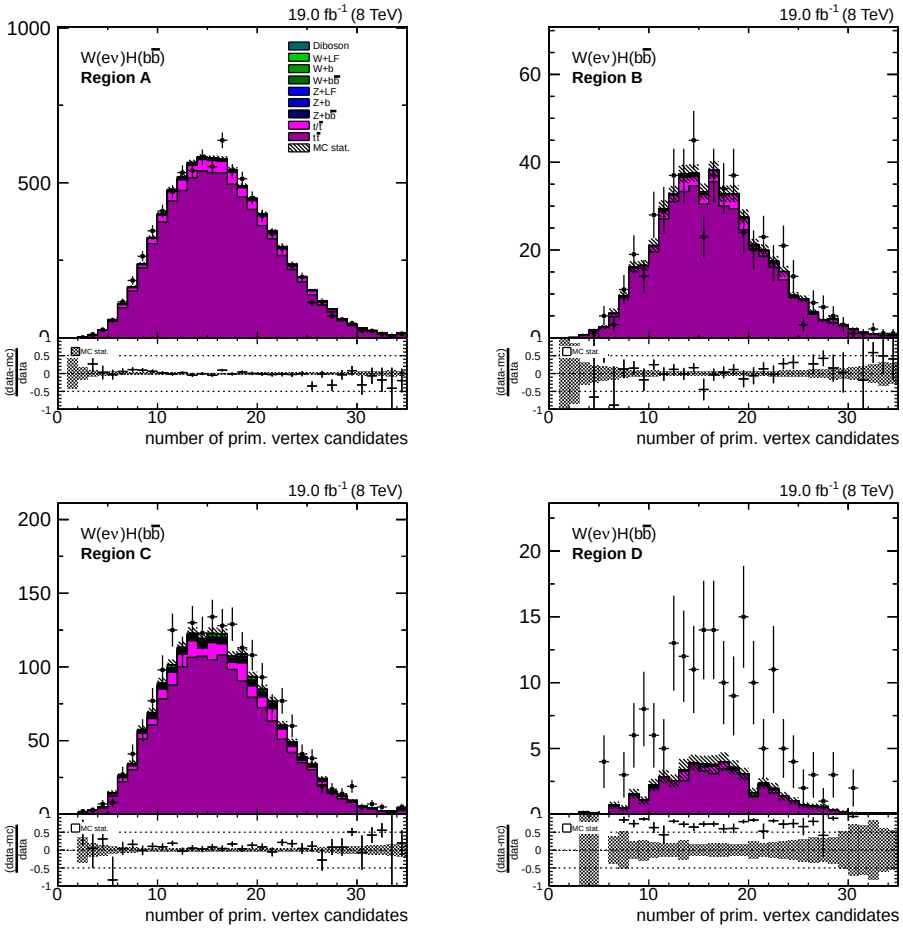


Figure 5.15.: Descriptive example of the ABCD method in the $t\bar{t}$ control region. The number of primary vertices is shown in each QCD study region for the electron channel. Region A (upper-left) is the ordinarily-defined control region. For region B (upper right) the electron isolation requirements are reversed. Region C (lower left) is obtained by inverting the $\not{E}_T$ cut and for region D (lower right) both cuts are reversed. Monte Carlo predictions are shown for all the major background processes with the exception of multi-jet QCD production. The difference between data and Monte Carlo prediction is attributed to this specific background process.

Table 5.9.: Exemplary calculation for the ABCD method. The event yields N for data, non-QCD background from Monte Carlo and their difference in the electron channel for the intermediate p_T $t\bar{t}$ control region. $N_{\text{QCD, ABCD}}$ denotes the predicted number of QCD events in region A using equation (5.8). In all signal and control regions a QCD contamination compatible with 0% is found.

$t\bar{t}$ Control region	B	C	D	A
N_{data}	511	1894	176	8155
$N_{\text{non-QCD}}$	490 ± 7	1736 ± 14	48 ± 2	
N_{diff}	21 ± 7	158 ± 14	128 ± 2	
$N_{\text{QCD, ABCD}}$				26 ± 9
$N_{\text{QCD, ABCD}}/N_{\text{data}}$				$(0.32 \pm 0.11)\%$

is extracted by a fit to the BDT response shape. This way, the event selection can be looser compared to a simple cut-and-count analysis.

With the selected events in the signal regions two analyses were performed. The purpose of the first analysis is to reproduce the results of [153], so the same BDT input variables are used. Since only information from the dijet Higgs boson candidate is used, it is labeled as *DJ analysis*. Additionally, possible improvements of the analysis by using jet substructure information in the BDT are investigated. This study is referred to as *SJF analysis*. The details of both analyses are described in the following.

5.7.1. DJ analysis

In total nine different Higgs boson mass hypotheses are tested for the WH signal. At each mass point a dedicated BDT is trained to separate the signal against all occurring background processes. The outcomes are referred to as *nominal* BDTs. The previous considerations about signal and background characteristics are taken into account to find an optimized set of discriminating variables. In this set kinematic variables of the reconstructed Higgs and W bosons are taken into account, as well as angular correlations of the decay products. The full set is given in Table 5.10.

In Figures 5.16 and 5.17 the expected separation between signal and background in these variables is illustrated. Each of the variables shows a decent amount of discrimination power, that is combined within the BDT. The outcome of the training is shown exemplarily in the high p_T region for training with $m_H = 125$ GeV in Figure 5.18. As a first validation, in this diagram the KS-test probabilities between training and test samples are shown. No hints for overtraining are found.

Every trained BDT undergoes various checks to justify its usage. For the input variables data/MC comparisons in the control regions are essential. Figures 5.19 and 5.20 show a subset of the checked diagrams. For the W(eν)H channel the comparisons can be found in Figures A.2 and A.3 in the appendix. Overall good agreement between data and simulation is found.

As an additional cross check all BDTs are evaluated in the background control

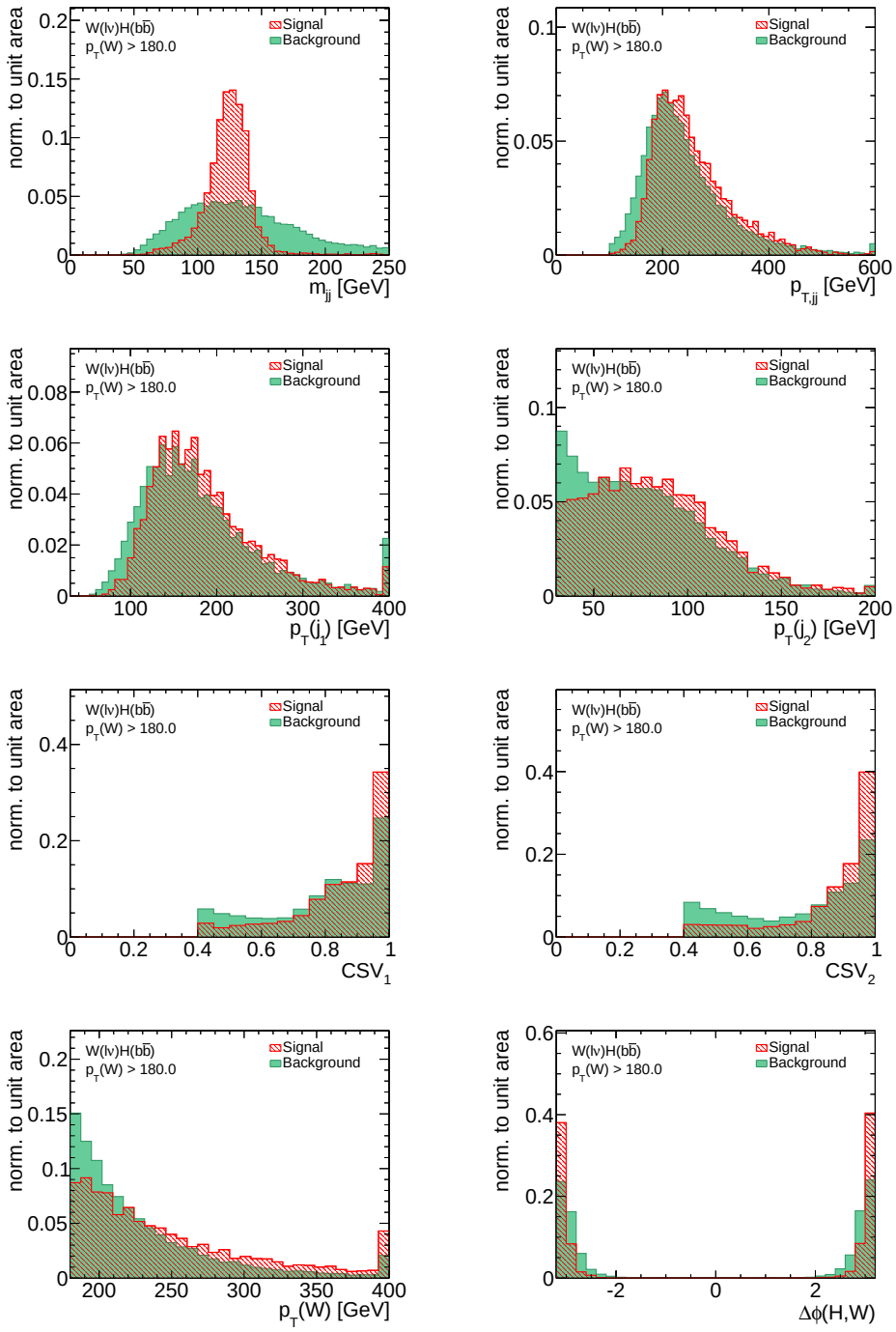


Figure 5.16.: Expected separation of BDT input variables in the DJ analysis. The signal and background distributions are normalized to unity. It should be noted that throughout this thesis the signal distributions are plotted in red.

Table 5.10.: Input variables used for the DJ analysis with description. The same set as in [153] is chosen.

Variable	Description
$p_{\mathrm{T}}(j)$	transverse momentum of each Higgs boson daughter
m_{jj}	invariant mass of dijet system
$p_{\mathrm{T},jj}$	transverse momentum of dijet system
$p_{\mathrm{T}}(\mathrm{W})$	transverse momentum of vector boson
CSV_1	value of CSV for the harder Higgs boson daughter
CSV_2	value of CSV for the softer Higgs boson daughter
$\Delta\varphi(\mathrm{H},\mathrm{W})$	azimuthal angle between W and dijet system
$\Delta\eta(j_1, j_2)$	difference in η between Higgs boson daughters
$\Delta R(j_1, j_2)$	distance in η–ϕ between Higgs boson daughters
$\mathrm{N_{aj}}$	number of additional jets
$\Delta\theta_{\mathrm{pull}}$	color pull angle [168]

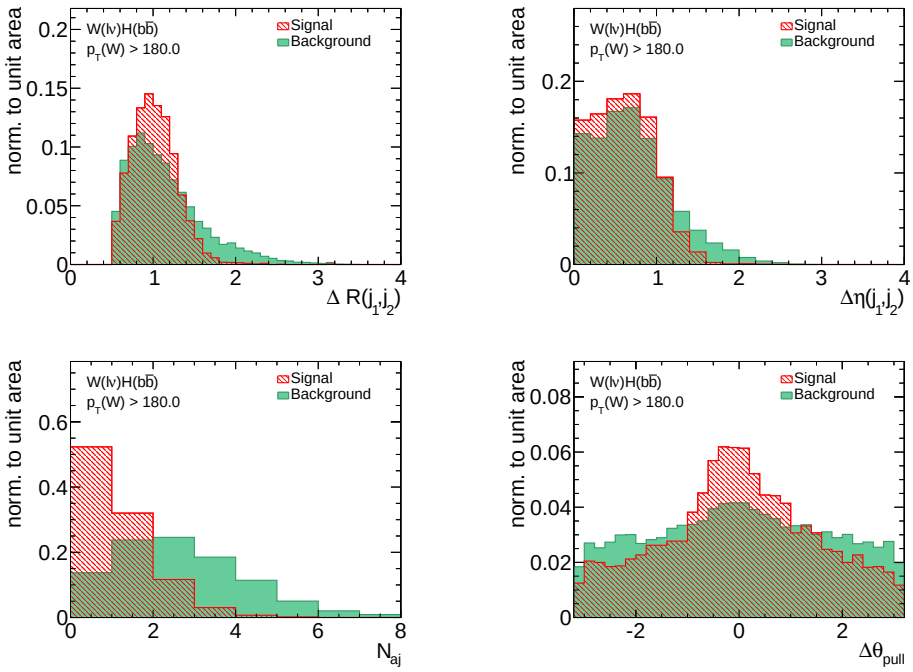


Figure 5.17.: Expected separation of BDT input variables in the DJ analysis (cont.). The signal and background distributions are normalized to unity.

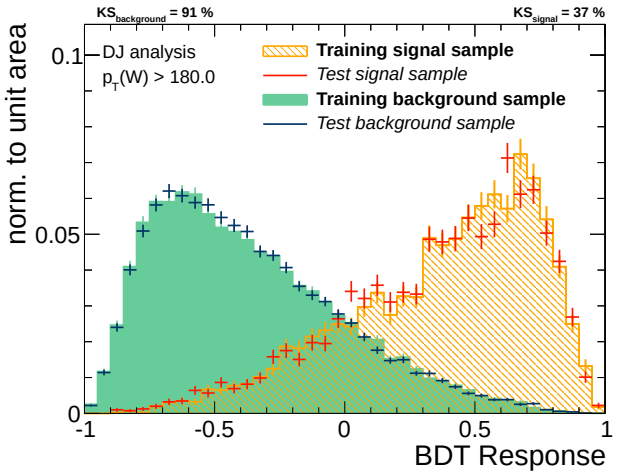


Figure 5.18.: BDT response for the DJ analysis in the high p_{T} region. The distributions are separated for signal and background events, and for training and testing samples. Adequate agreement is found by comparing the performance of the classification between the training and testing sample.

regions. Though the BDTs are evaluated in a different kinematic phase space, and thus on a different set of input information, the response in data and MC should behave in the same way. The data/MC comparisons for the $m_{\mathrm{H}}(125)$ training are explicitly shown in Figure 5.21 for different control regions. For all trainings a solid data/MC agreement is found.

In total there are 27 nominal DJ trainings resulting from nine mass points and three $p_{\mathrm{T}}(\mathrm{W})$ regions. For each training a ranking of variables according to their importance in the training exists. To evaluate the general importance of a variable, the number of occurrences in the top 5 positions is counted, as introduced in [169]. The closer this number is to 27, the more the variable helped to discriminate between signal and background. The most important variables are N_{aj}, m_{jj} and CSV_1 and CSV_2. The full list is given in Table A.6.

5.7.2. SJF analysis

The advantages in using filter jets as alternative Higgs boson reconstruction have already been studied before [169–171]. In this thesis the improvement including the substructure information in the existing, highly optimized CMS analysis is quantified. To aim for an easy implementation for other collaborating groups, additional variables are simply added on top of the nominal set in the classification training. Apart from that, exactly the same setup is chosen for the SJF analysis. From a large set of tested variables the invariant mass and transverse momentum of the trifilter jet system, as well as the absolute difference between trifilter and dijet mass, are found to be most promising. The added variables are listed in Table 5.11. To give a first impression of their discrimination power, the three added variables are

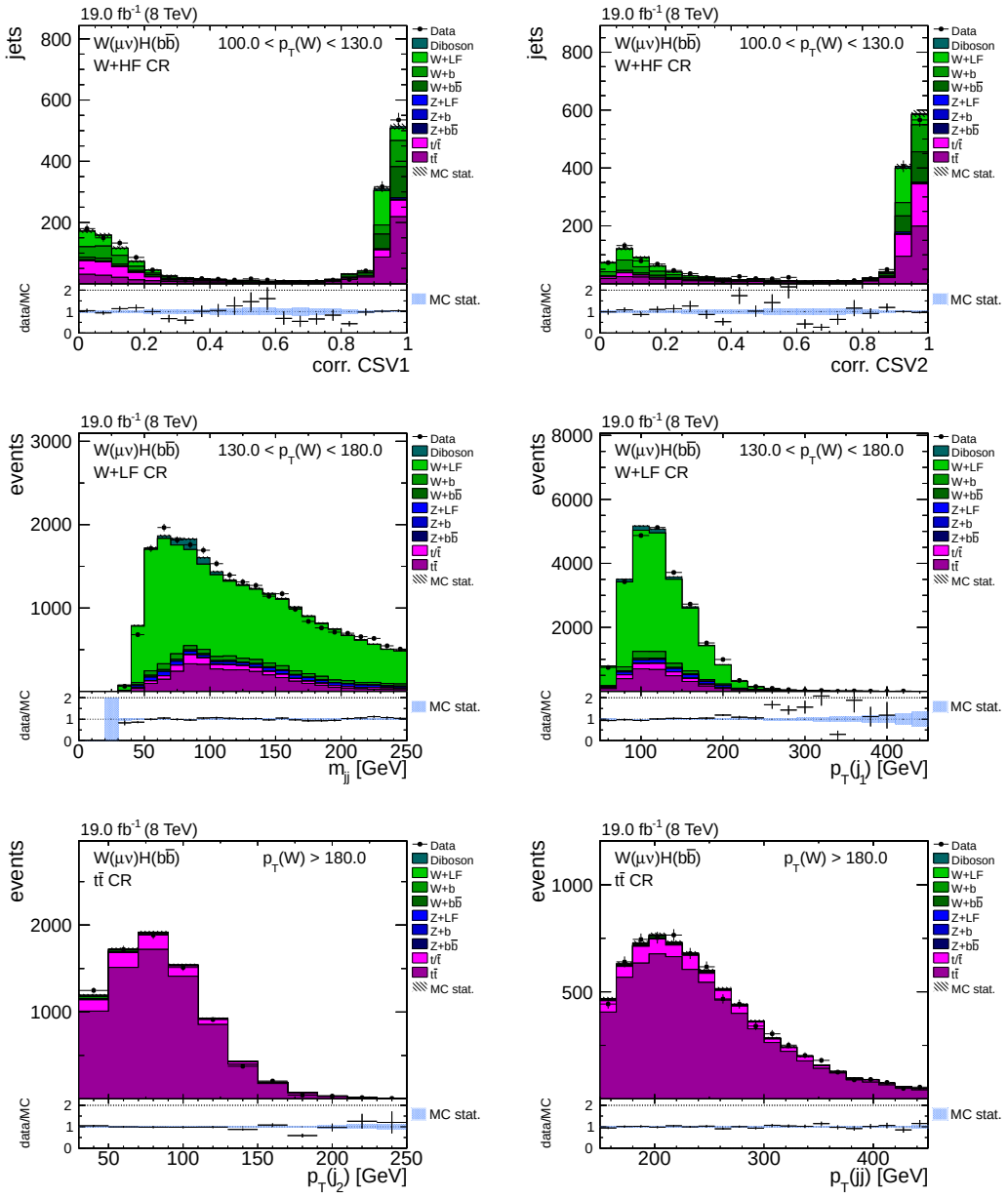


Figure 5.19.: DJ analysis input variables for the W(μν)H channel in different control regions as indicated within the diagrams. The simulation is scaled to luminosity and scale factors are applied. In all variables solid data/MC agreement is found. The corresponding distributions in the W(eν)H channel are presented in Figure A.2.

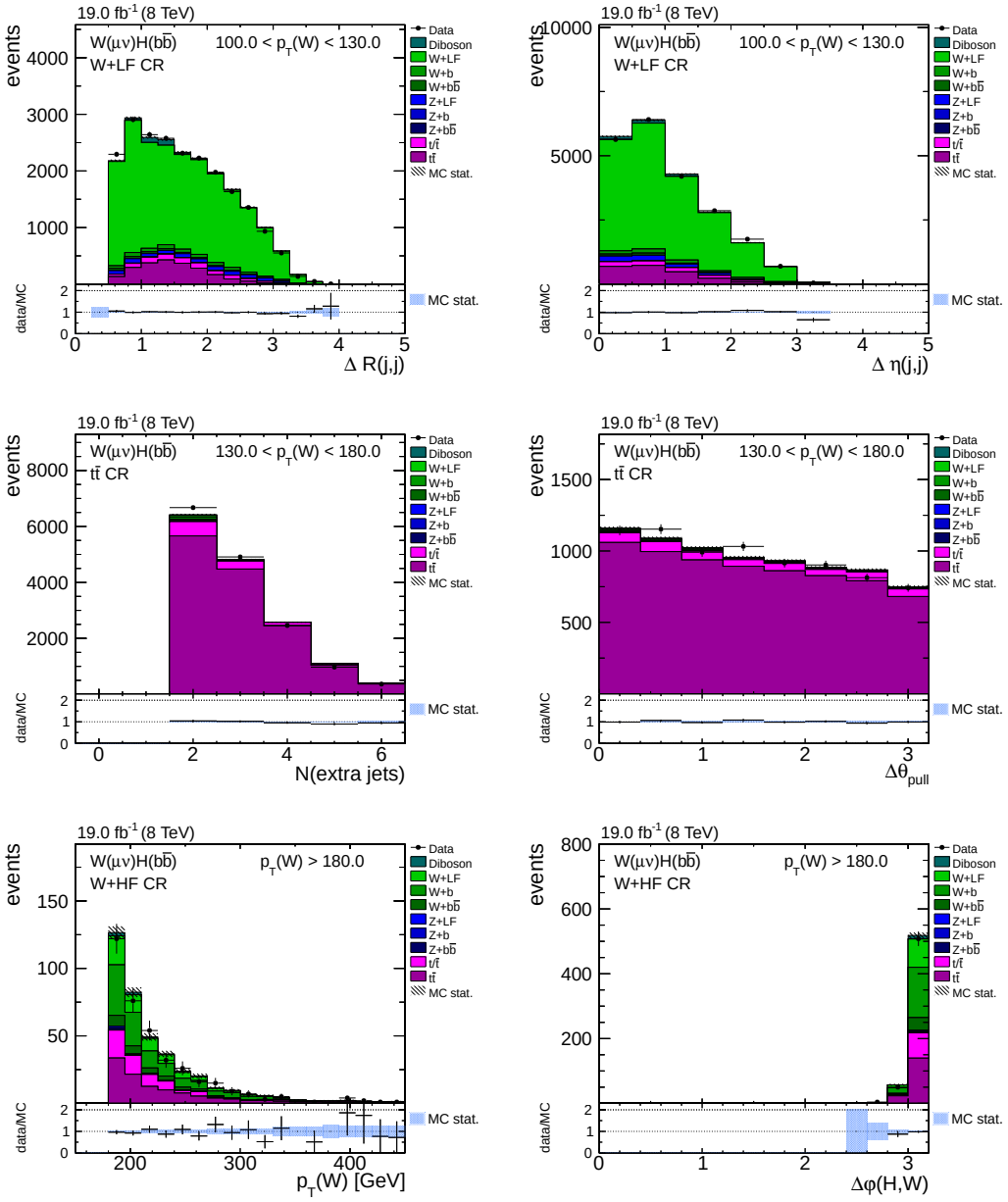


Figure 5.20.: DJ analysis input variables for the W(μν)H channel in different control regions as indicated within the diagrams (cont.). The simulation is scaled to luminosity and scale factors are applied. In all variables solid data/MC agreement is found. The corresponding distributions in the W(eν)H channel are presented in Figure A.3.

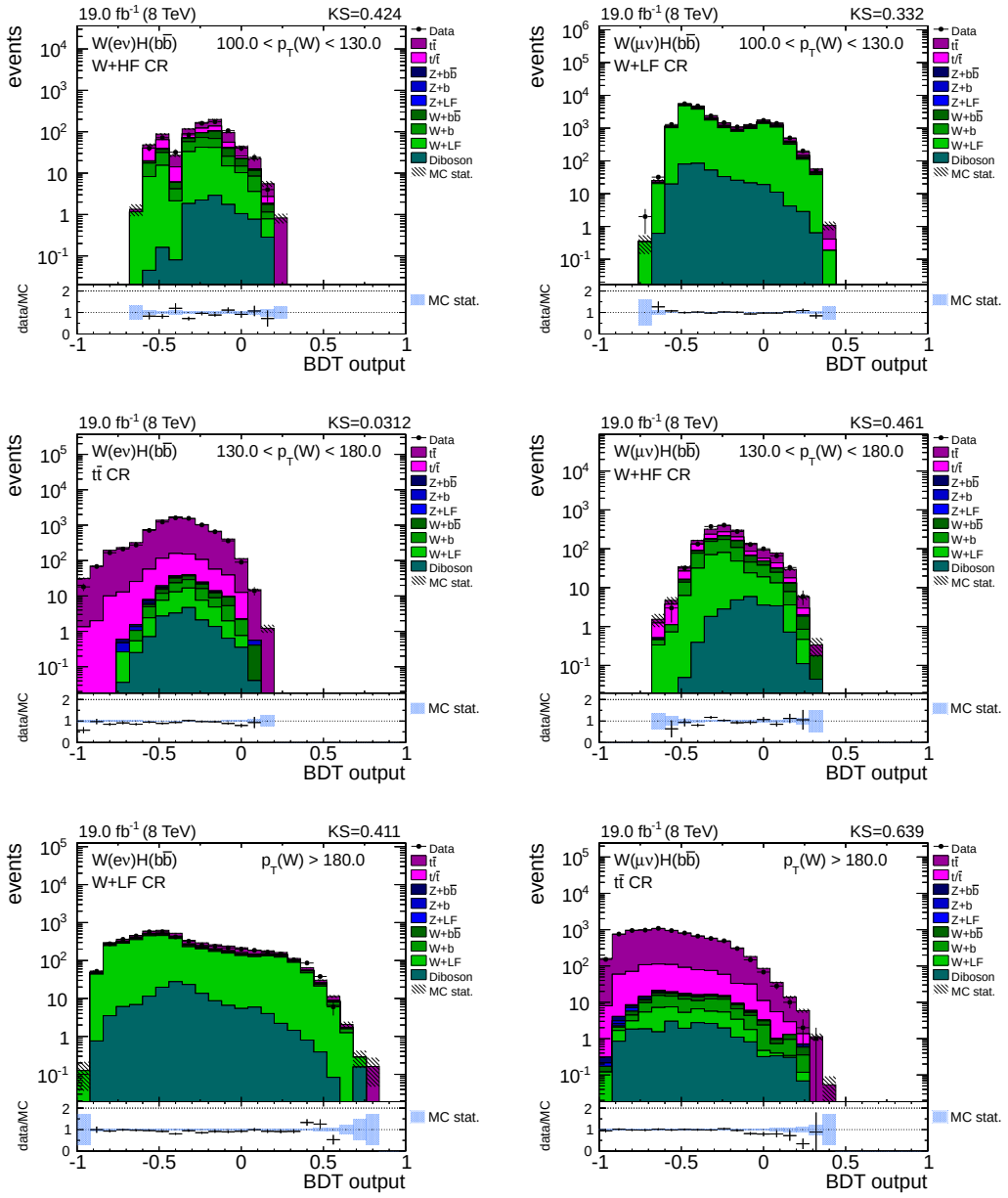


Figure 5.21.: Validation of classification BDT for the DJ analysis in control regions for the W(eν)H (left column) and the W(μν)H channel (right column). Data/MC comparisons are shown for the low p_T region (top row), intermediate p_T region (middle row) and the high p_T region (bottom row) in all three control regions as indicated in the figures. Scale factors and event weights are applied on simulation. Overall a very good agreement is found, justifying the use of the BDTs.

Table 5.11.: SJF variables introduced to classification training for analysis improvement. From a large set of tested distributions, these variables are found to be most promising in increasing the search sensitivity.

Variable	Description
m_{jj}^{flt}	invariant mass of trifilter jet system
$p_{\mathrm{T},jj}^{\mathrm{flt}}$	transverse momentum of trifilter jet system
$\lvert m_{jj} - m_{jj}^{\mathrm{flt}} \rvert$	absolute invariant mass difference between trifilter and dijet system

displayed separately for simulated signal and background events in Figure 5.22.

Again, for each mass point a BDT is trained in each region. The outcome of the training is shown exemplarily in the high p_{T} region for the 125 GeV mass point in Figure 5.23. In this diagram, also the KS-test probabilities between training and test samples are shown and no hint for overtraining can be found.

A selection of data/MC comparisons for the added substructure variables is depicted in Figure 5.24. The agreement is acceptable. Further studies show that the overall agreement for substructure information improves with cuts on $p_{\mathrm{T},jj}^{\mathrm{flt}}$ [172]. Since the event selection is kept the same as for the DJ analysis, these requirements are not introduced. As a result, the validation of the BDT responses in the control regions is, a fortiori, of great importance. Figure 5.25 shows the distributions and the data/MC agreement is found to be satisfactory.

Additionally, the correlations of the three newly introduced variables over the dijet mass are checked, as provided in Figure A.4 in the appendix. On the one hand the diagrams show good agreement between data and MC. On the other hand, the diagrams illustrate the strong correlation between the added substructure variables and the dijet mass. Therefore, the BDT cannot exploit their full discrimination power.

Table A.7 shows again the top 5 occurrences ranking. The filter jet variables are not ranked best, but could still provide information to the training. These rankings give only an estimate of the performance of a variable compared to the other ones. To judge whether an overall improvement of the search sensitivity is found, the expected upper limits need to be evaluated.

5.7.3. Expert BDTs

One major enhancement to the official analysis, improving the search sensitivity by roughly 10%, is the application of background specific BDTs [161]. These so-called expert BDTs are trained in addition to the single nominal BDTs at each mass point to separate the WH signal from a single background source, i.e. $\mathrm{t\bar{t}}$, W+0b and VV production, respectively. In the end, they are used to categorize the events into

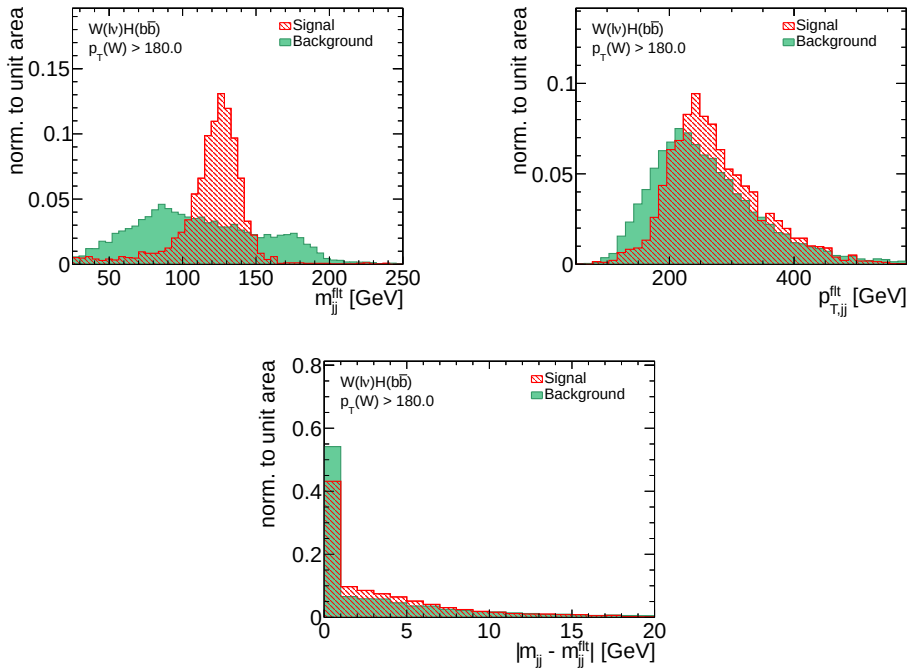


Figure 5.22.: Expected separation of added substructure variables in the SJF analysis. The signal and background distributions are normalized to unity.

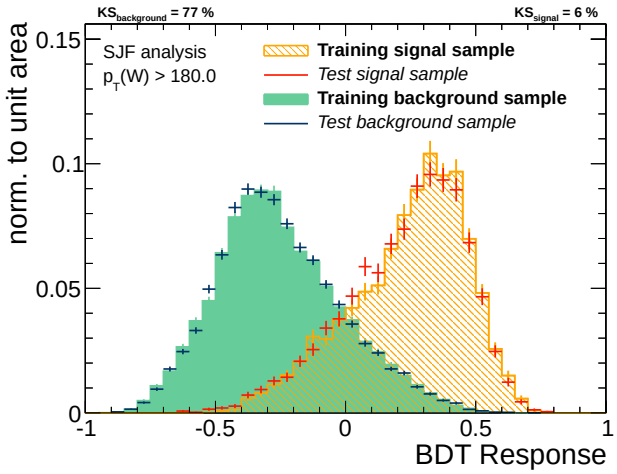


Figure 5.23.: BDT response for the SJF analysis in the high p_{T} region. The distributions are separated for signal and background events, and for training and testing samples. Good agreement is found by comparing the performance of the classification between training and testing samples.

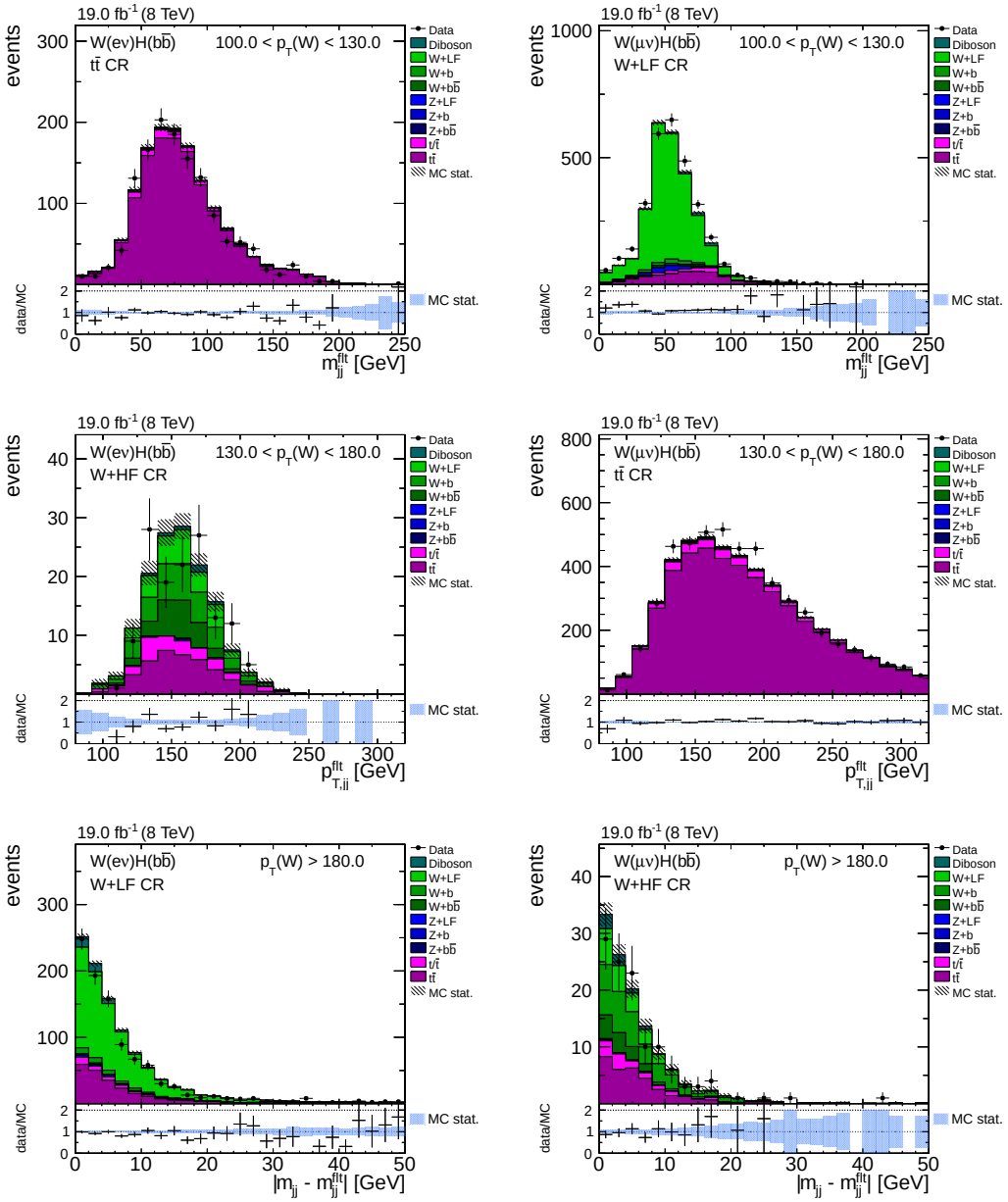


Figure 5.24.: Added filter jet variables in different control regions for the W(eν)H (left column) and the W(μν)H channel (right column). Data/MC comparisons are shown for the low p_{T} region (top row), intermediate p_{T} region (middle row) and the high p_{T} region (bottom row). Overall reasonable agreement is found. The agreement would improve, if additional requirements on $p^{\mathrm{flt}}_{\mathrm{T},jj}$ were applied.

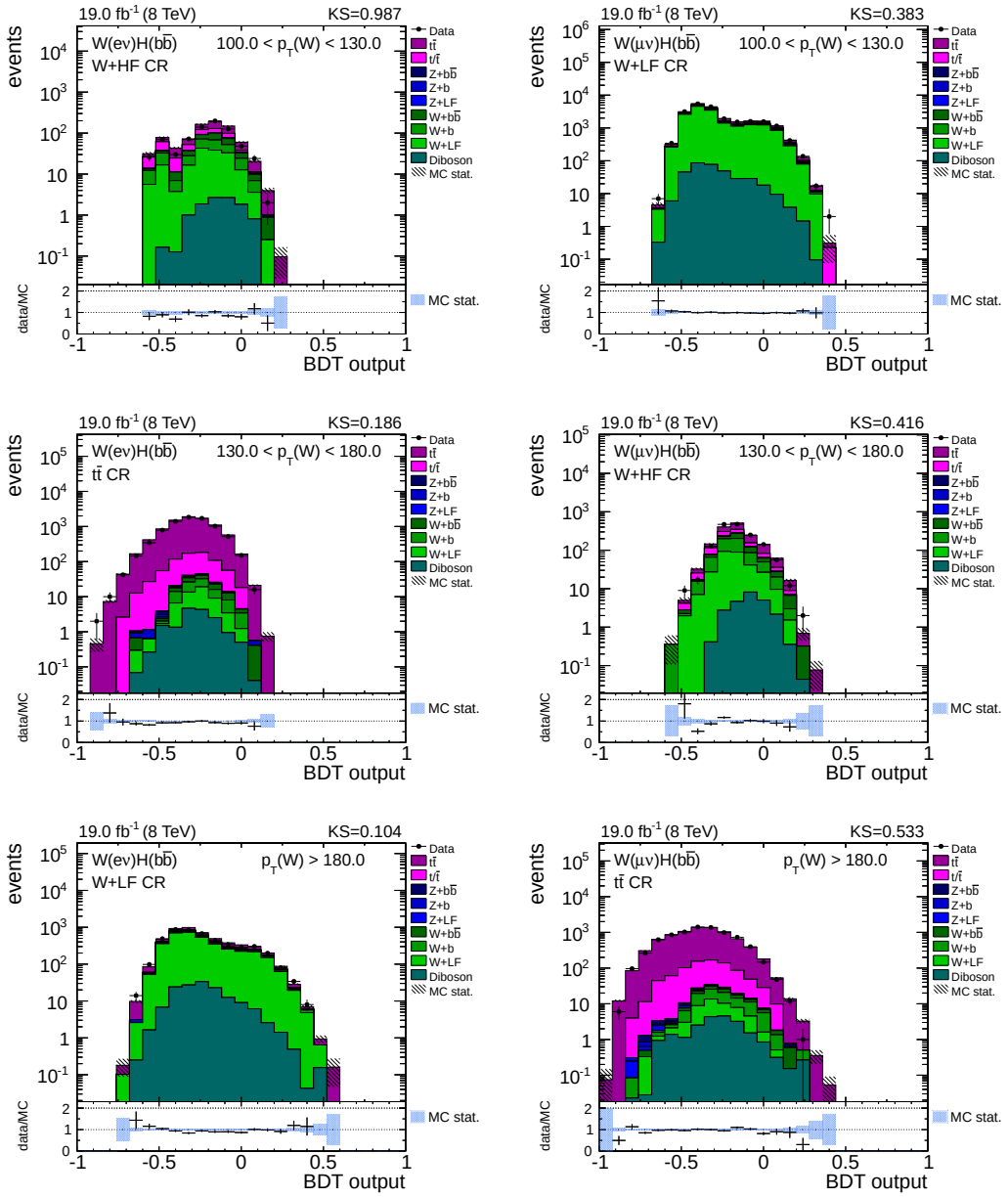


Figure 5.25.: Validation of classification BDT for the SJF analysis in control regions for the W(eν)H (left column) and the W(μν)H channel (right column). Data/MC comparisons are shown for the low p_T region (top row), intermediate p_T region (middle row) and the high p_T region (bottom row) in all three control regions as indicated in the figures. Scale factors and event weights are applied on simulation. Overall very good agreement is found, justifying the use of the BDTs.

four regions (1-4) via subsequent cuts. The separation employing the responses of $\mathrm{BDT}_{\mathrm{t\bar{t}}}$, $\mathrm{BDT}_{\mathrm{wjets}}$ and $\mathrm{BDT}_{\mathrm{vv}}$ is explained via the following pseudo code:

region 1 [-1.0, -0.5]: `if`$(\mathrm{BDT}_{\mathrm{t\bar{t}}} < c_{\mathrm{t\bar{t}}})$
region 2 [-0.5, 0.0]: `else if`$(\mathrm{BDT}_{\mathrm{wjets}} < c_{\mathrm{wjets}})$
region 3 [0.0, 0.5]: `else if`$(\mathrm{BDT}_{\mathrm{vv}} < c_{\mathrm{vv}})$
region 4 [0.5, 1.0]: `else`

The values for the three cuts c_{exp} are subject of the BDT optimization described in the next section. For all events in the regions 1-4 the response of the nominal BDT is plotted. Compared to the nominal BDT this categorized BDT, composed of four distinct sets of events, provides more discrimination power for the limit extraction.

The three types of expert BDTs are included in both, the DJ and the SJF analysis, and the corresponding set of input variables stays the same. In addition to the 27 nominal trainings another 81 expert BDTs are evaluated for each analysis. All of these BDTs are validated the same way as the nominal BDTs. Without attestation of the corresponding diagrams, all expert BDTs show no overtraining and a solid agreement between data and MC is found for all.

Tables A.6(b) - (d) show the top 5 occurrences ranking for the DJ expert trainings. The ranking for $\mathrm{BDT}_{\mathrm{t\bar{t}}}$ is similar to the nominal one. This implies that already the nominal BDT is separating the signal mainly against $\mathrm{t\bar{t}}$ production, since this is the most prominent background contribution in all channels. In the ranking for $\mathrm{BDT}_{\mathrm{wjets}}$ especially the CSV values for both jets assigned to the Higgs boson gain importance, whereas the influence of $\mathrm{N_{aj}}$ becomes minor. The most important difference in the $\mathrm{BDT}_{\mathrm{vv}}$ training compared to the nominal BDT is that m_{jj} is the dominant discriminating variable. This behavior is expected, as the diboson production can only be separated from the signal via the reconstructed mass of the Higgs boson candidate.

The corresponding rankings for the SJF expert trainings are given in Tables A.7(b) - (d). The rankings behave in a similar way to DJ analysis. The added substructure variables gain importance in the $\mathrm{BDT}_{\mathrm{t\bar{t}}}$ training, but lose importance in the other two.

5.7.4. BDT optimization

A lot of effort is put into the optimization of the BDTs to achieve the best possible search sensitivity. Different configurations of the BDT training lead to different results. Such studies are prone to overtraining, that needs to be avoided under any circumstances. The goal of the examinations described in the following is to obtain the highest expected significance for the signal process with $m_{\mathrm{H}} = 125\,\mathrm{GeV}$ by optimizing the configuration for each analysis region. The resulting parameters are applied for the training on the other mass points as well.

In a first study the TMVA default value for the allowed maximum depth of the decision trees, `maxDepth`= 3, is found to be most stable against overtraining. Two other important BDT parameters that influence overtraining and performance are

the minimal number of events required in a leaf node (`nEventsMin`) and the pruning strength (`PruneStrength`). The latter option is deactivated in all trainings. To find the training configuration with the best possible significance the parameter `nEventsMin` is scanned together with the number of trees per training (`NTrees`) in the ranges [50,1300] and [50, 1000], respectively, in steps of 50. For each tested parameter pair the expected significance is calculated on pseudo datasets including the signal and the $t\bar{t}$, VV and W+Jets background processes. To reduce computing time only *rate uncertainties* (see next section) are taken into account for the scan. The results from all trainings that show hints towards overtraining according to low KS-test probabilities are immediately dropped. Table A.8 shows the list of configurations found to yield the best significance for both the DJ and SJF analyses. The difference between best and worst expected significance is roughly 5% in all regions.

A similar scan is performed to find the best set of cuts on the expert BDT outputs described in the previous sub-section. All cuts were tested in the range $[-0.5, 0.5]$ in steps of 0.05. The resulting cuts, optimized for best significance, are $c_{t\bar{t}} = -0.3$, $c_{\text{wjets}} = 0.1$ and $c_{\text{vv}} = 0.2$ for the DJ analysis. For the SJF analysis the corresponding cut values are $c_{t\bar{t}} = -0.2$, $c_{\text{wjets}} = 0.15$, $c_{\text{vv}} = 0.15$.

Since the binning of the resulting BDT responses is arbitrary, another modification of the BDT responses is performed. Unstable fit results might occur due to statistical fluctuations in MC, for example, when the outermost bins contain only few simulated background events with a large uncertainty, but comprise a large amount of signal. That is why every evaluated BDT response is subjected to a re-binning procedure, motivated in [173]. The left- and rightmost bin edges are computed such that the relative uncertainty in those bins is less than 25%. The bin edges in between are distributed equidistantly. In this analysis the rebinning is performed separately for the DJ and SJF analyses, for all tested mass hypotheses and all analysis regions.

5.8. Systematic uncertainties

There are several sources of systematic uncertainties affecting the simulated templates, that need to be evaluated. As introduced in Section 4.1.4 an uncertainty is denoted as *rate uncertainty*, if it is expected to influence only the overall yield estimate of the concerned processes. Effects leading to an event-by-event alteration are expected to change the distribution of kinematic variables and therefore the BDT response. Hence, for these *shape uncertainties* systematically shifted templates corresponding to $\pm 1\sigma$ variations are evaluated for all processes. The re-evaluated BDT responses for these templates are included as *up* and *down* shapes in the final limit extraction.

The systematic influences can be grouped into three categories, as explained in the following. For all influences, a handling consistent with [153] is aspired.

5.8.1. Luminosity and theory uncertainties

The integrated luminosity used for normalizing the MC templates is measured centrally. The uncertainty on this measurement given in [174] is taken as a constant 5.0% on all simulated processes as rate uncertainty.

For the signal process as well as for single-top and diboson production the theoretical cross sections are needed for the normalization. The uncertainties arising from PDF variations are estimated to be 1% for single-top, VH and diboson production. In the statistical analysis the PDF uncertainties for VH and diboson production are taken into account as 100% correlated. The uncertainties from QCD scale variations are predicted to be 4% for VH and diboson production, and 6% for single-top processes. Those are assumed to be 0% correlated. Since scale factors for $t\bar{t}$ and $W + jets$ production are evaluated prior to the final fit, no additional uncertainties on the cross section are introduced there.

The kinematics of the signal region in the boosted regime, in which this analysis is performed, is challenging to simulate. Possible differences between the theoretical p_T spectrum of the MC templates and of Higgs events in the actual data could lead to systematic effects. A rate uncertainty of 10% on the signal process is assumed to account for both, the NLO electroweak [175–177] and the NNLO QCD corrections [178]. Studies in [161] showed that these uncertainties cover the difference between theory calculation and signal simulation in the veto efficiency on additional jet activity.

5.8.2. Reconstruction uncertainties

The uncertainties arising from lepton reconstruction and identification, as well as the effect of triggers are found to be 3.0% [161]. This 3.0% rate uncertainty is taken into account for the signal process and for single top and diboson production.

The systematic effects arising from reconstructed jets are split up into jet energy scale (JES) and jet energy resolution (JER) uncertainties. For both sources the CMS *JetMET* physics object group provides recommended values [179].

To account for the JES uncertainty for each MC sample two systematically shifted templates are created. In those, for every standard jet the p_T gets re-evaluated after varying the JES within its uncertainties dependent on p_T and η. The p_T of filter jets are also altered, but the uncertainties are assumed to be 2% larger compared to standard jets. All reconstructed variables depending on the jet four-vectors are re-computed. The modified BDT distributions of these templates are introduced to the limit evaluation as shape uncertainties.

The uncertainty on JER is accounted for in a similar way. Again, two templates with altered jets are created by smearing the jet energy according to the uncertainties of JER. These uncertainties are $\pm 10\%$ for standard jets within $|\eta| < 1.5$, $\pm 15\%$ within $1.5 < |\eta| < 2.5$ and $\pm 20\%$ for all others. As the difference in jet energy resolution between data and simulation is only measured for standard jets, a conservative additional uncertainty of 2% is assumed for filter jets. The values

read 12% ($|\eta| < 1.5$), 17% ($1.5 < |\eta| < 2.5$) and ±22% ($|\eta| > 2.5$). For the varied samples all affected variables and the BDT responses are re-computed and used as shape uncertainties in the limit extraction.

An uncertainty of 10% is assumed for the energy calibration of particles that are not clustered within jets. When propagating this uncertainty to the missing transverse energy a rate uncertainty of 3% is found, that is consistent with the analysis in [153].

In addition, the uncertainties on the measured scale factors used for the b-tagging reweighting procedure (Section 5.4.1) need to be accounted for. The event weight is re-evaluated by shifting either the scale factors for jets from b and c quarks or the scale factors for mistagged light and gluon induced jets within their uncertainties. Depending on the process a yield change of up to 13% is observed. The four varied BDT responses are added as shape uncertainties to the statistical evaluation.

5.8.3. Simulation uncertainties

The uncertainties on the data-driven scale factor determination are also considered. Consequently, region dependent rate uncertainties estimated in Section 5.6.3 are included for $\mathrm{t\bar{t}}$, W + 0b, W + 1b and W + 2b processes.

For the contributions from single top and diboson production the estimates are obtained from simulation only. Additional rate uncertainties for these processes are taken into account. For single top production the assumed uncertainty is 24%, which is slightly more conservative than the uncertainty quoted in the CMS measurement in [180]. The uncertainty on the expected number of events from diboson production is assumed to be on the same order as for single top processes, and is estimated to be 30%.

Within the scope of [153] a variation in BDT responses was found for W + jets production depending on the generators. The difference in the expected number of events between HERWIG++ and MadGraph was approximated to 10%, which is added as rate uncertainty.

Finally, the influence expected from the finite size of all simulated templates is accounted for by applying the *Barlow-Beeston light* method described in 4.1.4. The method introduces a shape uncertainty for each bin used in the fit. For this analysis this leads to 48 up and down variations, that are treated uncorrelated.

All delineated uncertainties are incorporated into the limit extraction procedure, described in the following. A dedicated table including the effects on limit is given in the next section.

Table 5.12.: Expected exclusion limits on WH signal with $m_{\mathrm{H}} = 125\,\mathrm{GeV}$ for the DJ analysis and the CMS analysis. The values are computed with the COMBINE package using all six channels. Additionally, the $\pm1\sigma$ and $\pm2\sigma$ errors are given. A nearly perfect consistency between both analyses is found.

$m_{\mathrm{H}} = 125\,\mathrm{GeV}$	Exp.	$+1\sigma$	-1σ	$+2\sigma$	-2σ
DJ analysis	1.84	2.72	1.28	3.92	0.94
CMS analysis	1.85	2.67	1.30	3.72	0.97
Rel. difference [%]	−0.5	1.9	−1.5	5.4	−3.1

5.9. Results

The two searches for $\mathrm{H} \to \mathrm{b\bar{b}}$ decays in the WH production mode presented in this chapter serve for quantifying the possible improvements due to the use of jet substructure information. For that reason, it is checked if the DJ analysis fulfills its purpose: reproducing the results of the official CMS analysis [153]. Based on the outcome it is judged, whether it can serve as reference for the results of the SJF analysis. Subsequently, the SJF analysis is compared with the DJ analysis. Eventually, the statistical evaluation of the SJF analysis using the full dataset is presented.

5.9.1. DJ analysis as reference

The sensitivities of two analyses are compared preferably by computing the expected exclusion limits without using the actual data to avoid statistical fluctuations. The information on the CMS analysis is taken from the central repository of the CMS *Higgs Combination* group [181]. Since the presented analyses, as well as the CMS analysis, are optimized on the signal process with $m_{\mathrm{H}} = 125\,\mathrm{GeV}$, only this mass hypothesis is used for the check. For both, all three analysis regions for the electron and muon channels are combined and asymptotic 95% C.L. $\mathrm{CL_s}$ exclusion limits are evaluated with the COMBINE package[1]. This is done by simultaneously fitting the BDT shape of simulation to the Asimov dataset in all regions. The full set of systematic uncertainties described in Section 5.8 is included. Table 5.12 shows the results including the edges of the $\pm1\sigma$ and $\pm2\sigma$ uncertainty bands.

The expected limits for DJ and CMS analyses coincide with 1% precision. As also the uncertainties agree well, the use of the DJ analysis as reference to quote possible improvements using substructure in the CMS analysis is justified.

[1]The COMBINE package is based on RooFit [182] and RooStats [183] and is the default tool in the CMS *Higgs* group. Its implementations for the statistical analysis are similar to THETA (see Section 4.1.4)

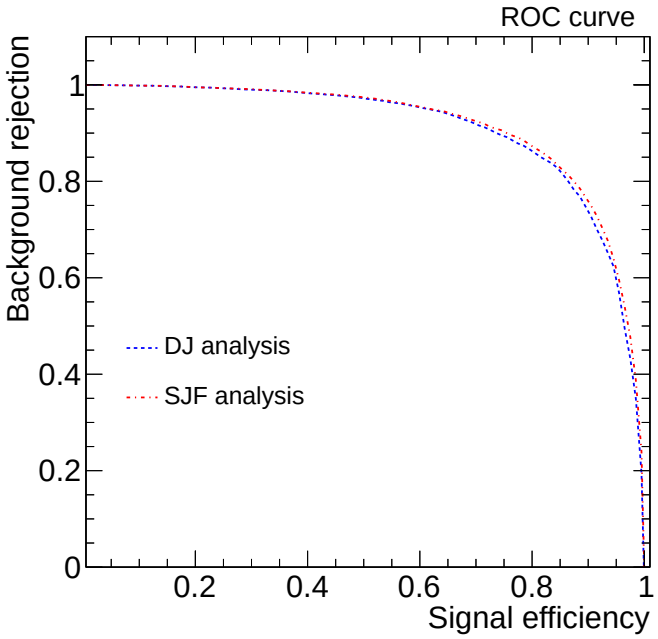


Figure 5.26.: Comparison between DJ (blue line) and SJF analyses (red line) via ROC curves. The higher curve for the SJF analysis indicates a better discrimination between signal and background events.

5.9.2. Improvements from jet substructure information

The search sensitivities of the DJ and the SJF analyses are compared in more detail. It should be noted once more that the inclusion of three variables using jet substructure information in the classification training is the only elementary difference between the two studies. All consecutive steps are performed similarly. Hence, a first comparison is performed by opposing the ROC curves[2] for the classification trainings, as depicted in Figure 5.26. A slighly better performance is found for the SJF analysis, indicated by a larger area under the ROC curve. The result is consistent with the implications of the rankings in the BDT trainings (see Section 5.7.3), which showed that no massive improvement is anticipated.

To get a deeper insight into the different performances, expected limits computed with the THETA framework serve again as measure. The comparison is done for the full set of tested mass hypotheses, as shown in Figure 5.27. In this diagram, the results for the SJF analysis (red line) are found to be generally lower than those for the DJ analysis (blue line). Hence, though the difference of the two analyses is not large, the integration of substructure information leads to an overall better search

[2]For the *receiver operating characteristic* (ROC) curves the background rejection is plotted versus the signal efficiency for several cuts on the discriminator value. The larger the area under this curve is, the more background events are rejected for a given signal efficiency, and therefore the better the training is separating signal from background events.

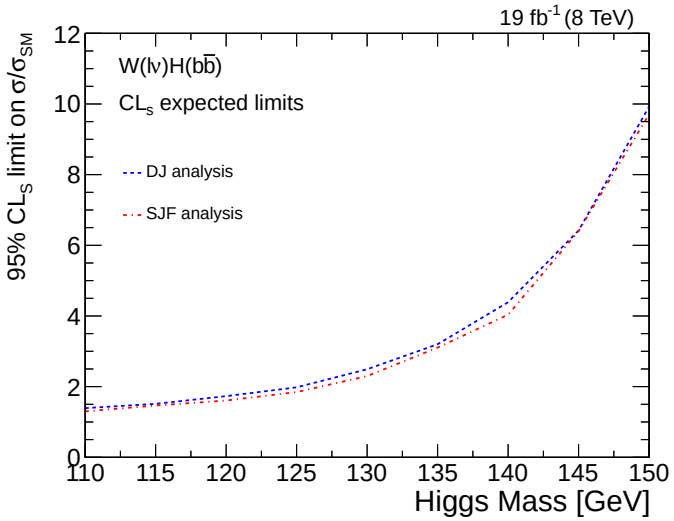


Figure 5.27.: Comparison between DJ (blue line) and SJF analyses (red line) in terms of expected limits. All analysis regions are combined. The diagram indicates that the SJF analysis yields better results in all cases.

Table 5.13.: Compared search sensitivities in terms of expected limits of DJ and SJF analyses. The relative difference is calculated with unrounded numbers. On average an improvement of 4.5% is found when adding jet substructure information.

m_{H}	110	115	120	125	130	135	140	145	150
DJ analysis	1.30	1.40	1.60	1.84	2.29	2.95	4.02	5.86	9.09
SJF analysis	1.22	1.36	1.49	1.73	2.13	2.87	3.77	5.82	8.84
Rel. improvement [%]	6.0	2.8	7.1	6.0	6.8	2.7	6.2	0.5	2.7

sensitivity.

To quantify these results, Table 5.13 breaks down the numbers of the evaluation. For the tested mass hypotheses the expected limits of the DJ analysis lie in the interval of (1.3, 9.1). The corresponding expected limits of the SJF analysis are in the interval of (1.2, 8.8). For $m_{\mathrm{H}} = 125\,\mathrm{GeV}$, an improvement of $\sim 6\%$ is found. Given the validation of the reference analysis, the average improvement of $\sim 5\%$ for all trainings is attributed to the use of three substructure variables in the classification BDTs. Especially, the background specific training against $\mathrm{t\bar{t}}$ production benefits from this inclusion.

5.9.3. Final statistical evaluation

After quantifying the advantages of using jet substructure information, the results for the SJF analysis including the actual data are evaluated. Again, a simultaneous fit in the categorized BDT distributions is performed. The THETA framework is used to obtain the asymptotic CL_s exclusion limits at 95% confidence level. Table 5.14 shows the final event yields after that fit representatively for the $m_H(125)$ training. The corresponding post-fit distributions including the combined statistical and systematical uncertainties on all simulated templates are given in Figure 5.28 separately for each of the six analysis regions. These post-fit distributions take the modulation of all nuisance parameters in the final fit into account.

Table 5.14.: Final event yields in all signal regions after the fit to data for the VH(125) SJF analysis. The given uncertainties include all systematic and statistical effects.

	Low p_T region		**Intermediate p_T region**		**High p_T region**	
Process	W(eν)H	W(μν)H	W(eν)H	W(μν)H	W(eν)H	W(μν)H
VH(125)	6.0 ± 2.4	7.6 ± 2.7	9.6 ± 3.1	11.6 ± 3.4	10.5 ± 3.2	11.6 ± 3.4
$t\bar{t}$	370.3 ± 19.2	505.3 ± 22.5	350.4 ± 18.7	473.8 ± 21.8	173.8 ± 13.2	223.4 ± 14.9
$t\bar{t}$	3184.5 ± 56.4	3873.6 ± 62.2	2749.6 ± 52.4	3270.4 ± 57.2	1074.4 ± 32.8	1188.3 ± 34.5
VV	21.8 ± 4.7	28.5 ± 5.3	30.4 ± 5.5	37.8 ± 6.1	21.7 ± 4.7	25.8 ± 5.1
W+0b	262.6 ± 16.2	331.3 ± 18.2	240.7 ± 15.5	274.5 ± 16.6	110.5 ± 10.5	121.5 ± 11.0
W+1b	51.3 ± 7.2	69.9 ± 8.4	81.1 ± 9.0	89.3 ± 9.4	40.8 ± 6.4	46.2 ± 6.8
W+2b	221.0 ± 14.9	314.3 ± 17.7	201.7 ± 14.2	238.5 ± 15.4	77.6 ± 8.8	83.6 ± 9.1
Z+0b	11.7 ± 3.4	20.2 ± 4.5	13.5 ± 3.7	18.4 ± 4.3	4.0 ± 2.0	7.1 ± 2.7
Z+1b	5.5 ± 2.3	7.1 ± 2.7	4.0 ± 2.0	6.6 ± 2.6	1.7 ± 1.3	2.6 ± 1.6
Z+2b	11.4 ± 3.4	19.6 ± 4.4	10.3 ± 3.2	15.7 ± 4.0	3.9 ± 2.0	5.9 ± 2.4
Total MC	4146 ± 64	5177 ± 72	3691 ± 61	4437 ± 67	1519 ± 39	1716 ± 41
Data	3922	5367	3660	4460	1506	1727

The exclusion limits for all tested mass hypotheses are depicted in Figure 5.29. The black solid line indicating the observed limits on data is systematically higher compared to the expected limits denoted with the black dashed line. To put these findings into perspective the dashed red line shows signal injected limits. These are received on pseudo-datasets including the $m_H = 125\,\text{GeV}$ signal scaled to the found signal strength. By construction, the signal injected limit at $m_H = 125\,\text{GeV}$ coincides with the observed one. Over the tested mass range the injected limits reveal that systematically higher observed limits compared to the expected limits are unsurprising.

To evaluate the impact of each source of systematic uncertainty another study is performed. Representatively for the $m_H = 125\,\text{GeV}$ training, the expected limit is re-computed by either allowing only one nuisance parameter to float in the fit, or all but one. Table 5.15 summarizes the results, and shows that the uncertainties on the scale factors for $t\bar{t}$ and $W + \text{jets}$ production as well as the limited MC statistics have the largest impact on the limit.

A nice way to illustrate the sensitivity of this binned analysis in numerous channels is the following recipe [153]. All events from all channels are sorted according to the expected signal-over-background ratio in their corresponding bins, as shown

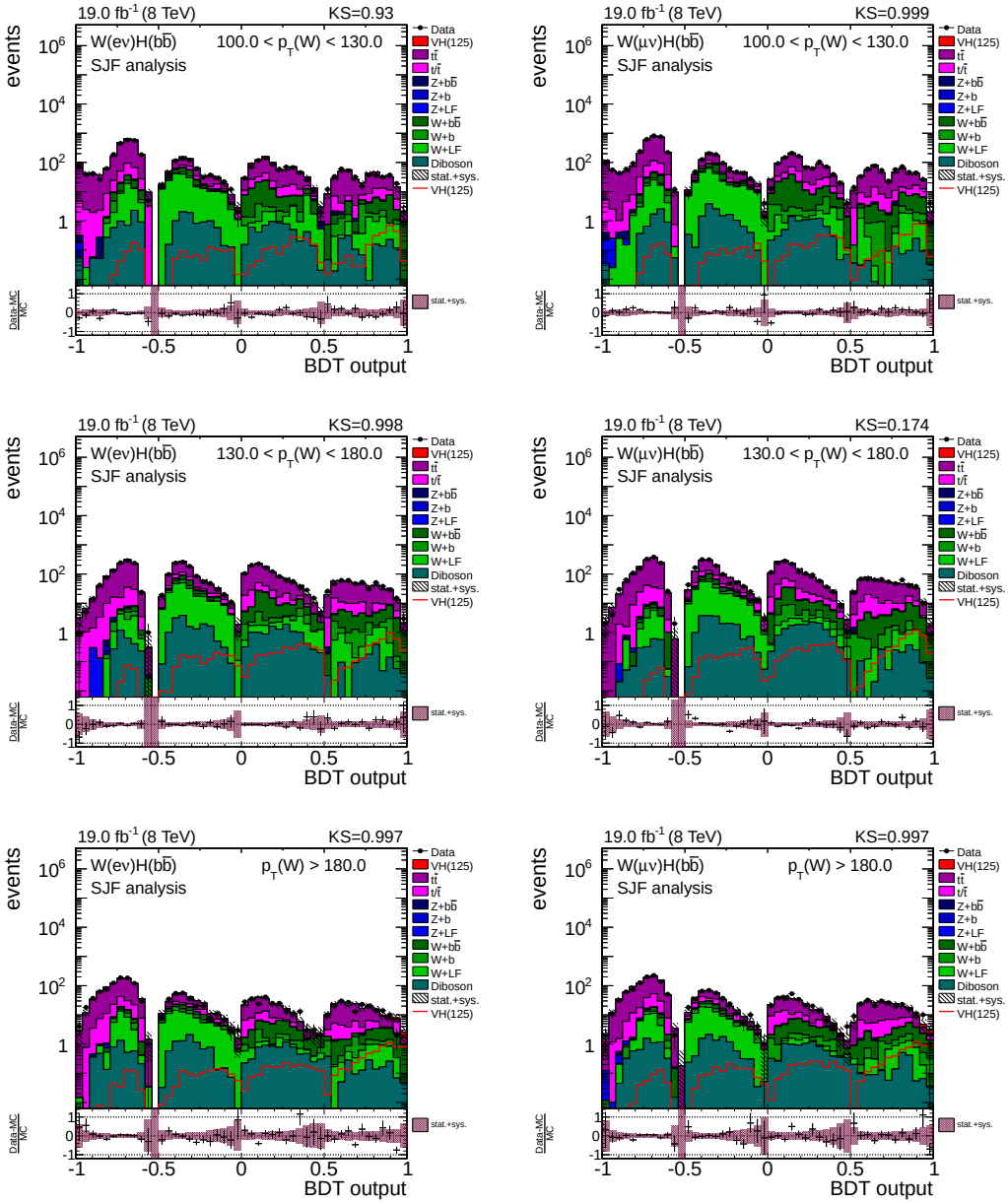


Figure 5.28.: Post-fit distributions for the $m_{\mathrm{H}}(125)$ training separately for all signal regions. The plotted uncertainties include both statistical and systematical effects.

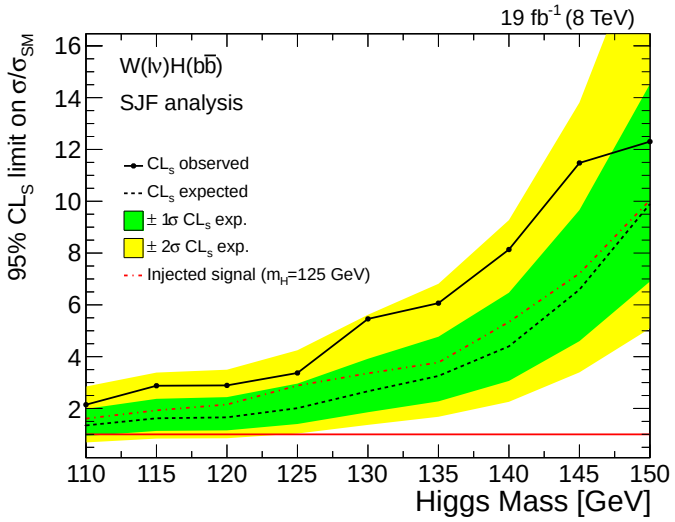


Figure 5.29.: Expected (dashed black line) and observed (solid black line) exclusion limits for the SJF analysis at all mass points. The green and yellow bands indicate $\pm 1\sigma$ and $\pm 2\sigma$ uncertainties on the expected limit. Additionally, the dashed red line shows the predicted values, when a Higgs boson with $m_{\mathrm{H}} = 125\,\mathrm{GeV}$ is present in the dataset.

Table 5.15.: Effects of systematic uncertainties in SJF analysis for the $m_{\mathrm{H}} = 125\,\mathrm{GeV}$ training. The table shows the impact of each nuisance parameter, when it enters the limit calculation as exclusive source of uncertainty and when it is removed.

Source	Type	Impact as exclusive source (%)	Removal effect (%)
Luminosity	rate	2.6	0.9
Signal cross section (scale and PDF)	rate	1.3	< 0.1
Signal cross section (p_{T} boost, EWK/QCD)	rate	3.2	3.5
Single-top (simulation estimate)	rate	3.2	< 0.1
Diboson (simulation estimate)	rate	1.1	< 0.1
Background SFs (data estimate)	rate	11.3	6.1
MC modeling (V+jets)	rate	6.1	0.9
Lepton efficiency and trigger	rate	0.5	< 0.1
Missing transverse energy	rate	3.2	0.4
Jet energy scale	shape	4.2	1.7
Jet energy resolution	shape	0.5	< 0.1
b-tagging	shape	1.6	0.9
MC statistics	shape	7.7	6.3

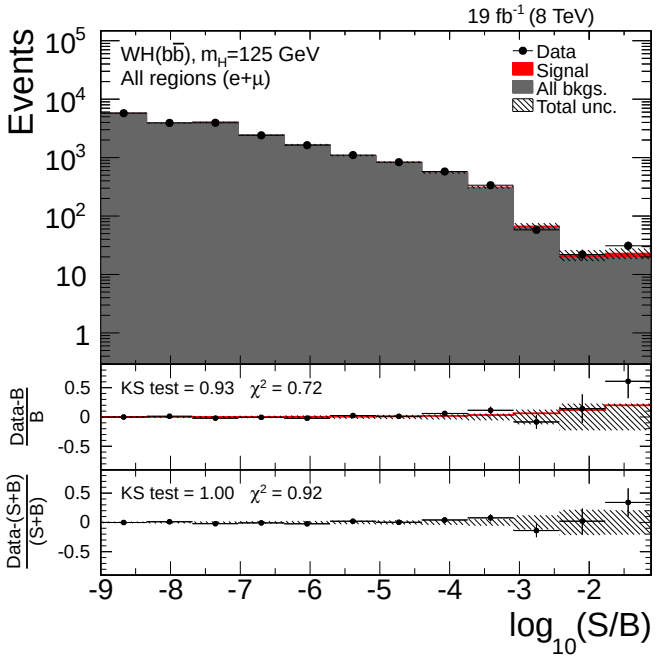


Figure 5.30.: Combination of all BDT distributions into one single diagram. Here, all events from all signal regions are sorted according to the expected signal-over-background ratio in the corresponding bin. The ratio between data and background-only and between data and signal-plus-background are also given. All statistical and systematic effects are included in the uncertainties. The data is described more accurately by the S+B hypothesis, hinting at a standard model Higgs boson.

in Figure 5.30. In the ratio at the bottom of the figure, the data is compared with the background-only and with the signal-plus-background hypotheses. Looking at the bins with the largest signal-to-background ratios, the data shows an excess that is consistent with the production of the standard model Higgs boson. The observed significance of the excess assuming $m_{\mathrm{H}} = 125\,\mathrm{GeV}$ is 1.2σ with an expected significance of 1.1σ.

In a broader perspective the combination of all 6 channels (W(eν)H, W($\mu\nu$)H, W($\tau\nu$)H, Z(ee)H, Z($\mu\mu$)H and Z($\nu\nu$)H) in CMS yields an observed (expected) significance of 2.1σ (2.5σ) [48]. The ATLAS collaboration reports a similar sensitivity [152]. In the long term an observation of $\mathrm{H} \to \mathrm{b\bar{b}}$ decays with a significance of 5σ is desired. This is only possible with more data that will be provided after the restart of the LHC with a center-of-mass energy of $\sqrt{s} = 13\,\mathrm{TeV}$. In the view of jet substructure the conditions in the WH channel might change for the new data taking period. On the one hand, the average boost of the $\mathrm{H} \to \mathrm{b\bar{b}}$ system rise, and in more cases the two b-quark jets could merge. On the other hand, the influence of pile-up interactions is assumed to increase heavily. The CMS collaboration faces this problem for instance by using anti-k_{T} jets with a size of 0.4 instead of 0.5 as default. The smaller standard jet size could decrease the improvements on the

search sensitivity expected from the SJF algorithm. Nonetheless, the inclusion of jet substructure information as presented in this chapter represents a powerful tool that can help to reach the 5σ goal.

6. Search for Higgs boson production in the tHq production channel

After the discovery of the Higgs boson, it is crucial to measure its properties with ever increasing precision. As described in Chapter 1 recent LHC measurements favor the values predicted by the standard model for the Higgs boson coupling strengths to fermions (κ_f) and bosons (κ_V).

The search for Higgs boson production in association with a single top quark (tHq) provides a great opportunity closing down the allowed range of coupling values for κ_f. Any deviation of the value for κ_f would lead to an enhanced cross section for tHq production.

In the following, the search for tHq with the Higgs boson decaying into a pair of bottom quarks is described. The chapter starts with an outline of the analysis strategy and the description of the signal characteristics and the occurring background processes. After listing the data and MC samples, the selection of events is introduced. The analysis relies on the use of three neural networks with different purposes, and the evaluation and validation of those are described in detail. Finally, the results of this first search for tHq production with $\mathrm{H} \to \mathrm{b\bar{b}}$ decays are given. The presented analysis was performed in a blind way, i.e. not analyzing the actual data in the sensitive regions to avoid biasing the results. As the different analysis steps are successfully validated, only the unblinded results are documented here.

6.1. Analysis strategy

With the data recorded by CMS so far, it is impossible to be sensitive to the standard model tHq production with a cross section of $\sigma_\mathrm{SM} = 18.3\,\mathrm{fb}$. That is why the analysis is optimized for the $\kappa_\mathrm{f} = -1$ case (see Section 1.3). From the results a general upper limit on tHq production can be derived.

The analysis strategy for the tHq search itself is straightforward. To be sensitive at all to the process, a tight selection is applied in the signal region to suppress mainly $\mathrm{t\bar{t}}$ production and obtain a reasonable signal-over-background ratio. Still, a very low amount of signal events remains under a huge amount of background events that can have a very similar signature in the detector. Hence, it is important to reconstruct the events as precisely as possible.

In the analysis each event is not only reconstructed under the hypothesis that it is a signal event, but also under the assumption that it stems from $\mathrm{t\bar{t}}$ production — the main background to tHq production. Neural networks are used in both cases for the jet-quark assignment in this challenging multijet final state. The resulting

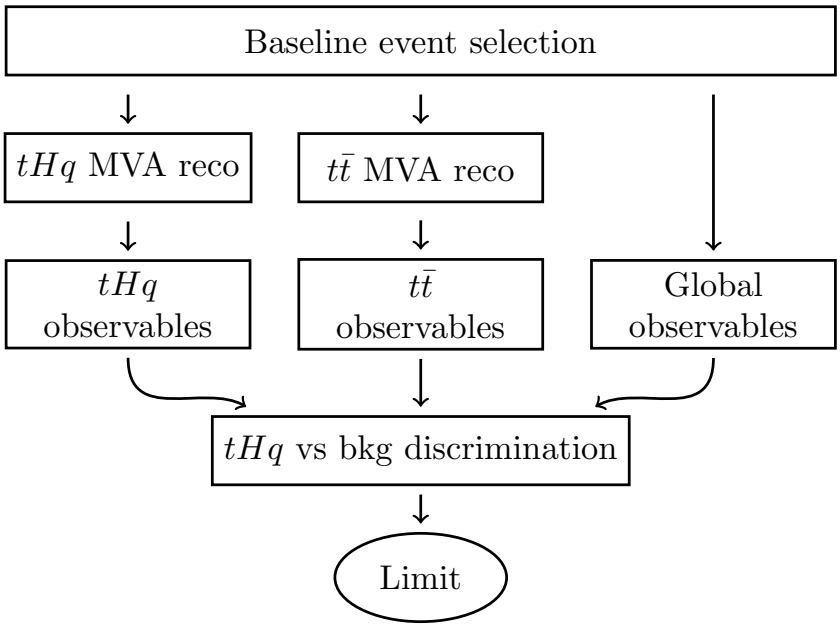
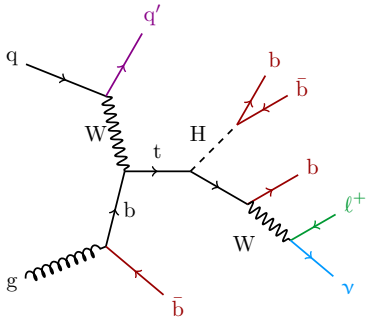


Figure 6.1.: Analysis scheme of the tHq search. After a tight event selection, every event is reconstructed under the assumption that is was a signal event and that it stems from $t\bar{t}$ production. This information is input for the classification MVA, that is used for the limit extraction. The sketch is taken from [61].

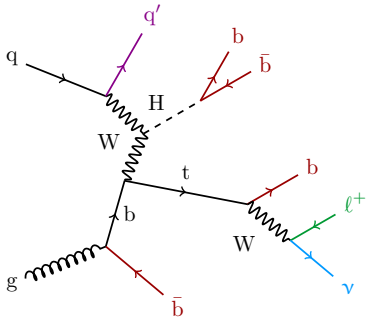
information from both reconstructions serves as input to a third neural network, which discriminates the signal events from the background processes. Finally, upper limits are set from a fit to the full shape of this classification MVA. The sketch in Figure 6.1 summarizes the analysis workflow, and the individual parts are described in more detail in the next sections.

6.2. Signal and background characteristics

The signature of tHq production events in the detector is very special. Figure 6.2(a) shows the representative Feynman diagram that is single top t-channel production, with an additional Higgs boson being radiated either from the top quark or from the W boson. The decay of the Higgs boson to a $b\bar{b}$ pair and the leptonic decay of the single top quark are also illustrated. One characteristic feature of single top production is the light forward jet. This information is used to discriminate the signal process from $t\bar{t}$ production. The single top quark is asked to decay leptonically, so there is an isolated lepton and missing transverse energy from the neutrino expected in the event. Further, there are four b quarks expected in total: two from the Higgs boson decay, one from the single top quark decay and another one from initial gluon splitting. The latter, so-called spectator b quark, lies often outside the detector's tracker acceptance due to its low transverse momentum.



(a) Higgs boson couples to top quark



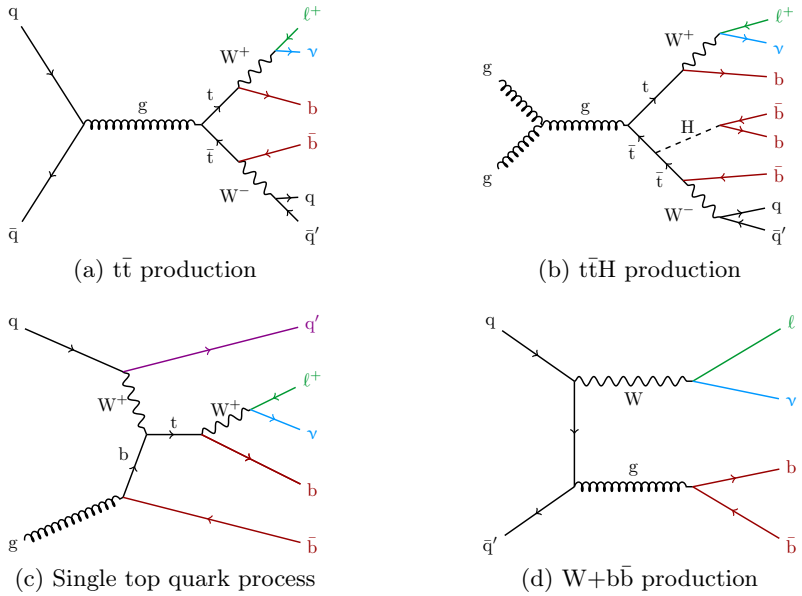
(b) Higgs boson couples to W boson

Figure 6.2.: Representative LO Feynman diagrams for tHq production. On the left side, the diagram where the Higgs boson arises from the top quark is shown. The right diagram depict the case where the Higgs boson is radiated from the W boson.

The semileptonic $t\bar{t}$ production is shown in Figure 6.3(a). When $t\bar{t}$ is produced with additional jets, with two of them being b jets or mistagged jets, the signature is similar to the signal process. Given the large cross section of $t\bar{t}+$jets compared to tHq production, it builds up by far the dominant background to this search. Hence, the crucial part is to model $t\bar{t}+$jets correctly and to discriminate the signal events mainly against this background. Another process contributing as background to this search is the Higgs boson production in association with top quark pairs. The representative Feynman diagram for the dominant production mode is given in Figure 6.3(b). The production cross section is comparable to that of the tHq production with $\kappa_f = -1$. Certainly, the t-channel single top production shown in Figure 6.3(c) needs to be considered as well. When produced with additional jets, with two of them being b jets or mistagged as such, the signature in the detector for this process mimics the tHq production. The contributions from electroweak W+jets production depicted in Figure 6.3(d), as well as from diboson and Z+jets production are almost negligible. The same statement holds for QCD multijet production. While for the electroweak processes MC samples with sufficient statistics are available, the amount of QCD events is estimated via the `ABCD` method, similar to the WH analysis.

6.3. Data, triggers and simulation

In this analysis the full amount of data from proton-proton collisions at a center-of-mass energy of $\sqrt{s} = 8\,\text{TeV}$ recorded with the CMS experiment is used. The same primary datasets as in the WH search are used, i.e. SingleMu for the muon channel and SingleEle for the electron channel. In both datasets the amount of data corresponds to an integrated luminosity of $19.7\,\text{fb}^{-1}$. Despite using the same primary datasets as in the WH search, the additional integrated luminosity is reached by



(a) $t\bar{t}$ production (b) $t\bar{t}H$ production

(c) Single top quark process (d) W+$b\bar{b}$ production

Figure 6.3.: Representative LO Feynman diagrams for the main background processes contributing to the tHq production. Similar to the WH search $t\bar{t}$ (a), single-top quark (c) and W+jets production (d) needs to be taken into account. In this analysis also the production of a Higgs boson in association with $t\bar{t}$ pairs (b) is considered.

data re-processing and an updated *golden* JSON file [184]. Table B.1 in Appendix B lists the different blocks of primary datasets with their corresponding integrated luminosity. The triggers used in the muon channel exclusively rely on the HLT path `IsoMu24_eta2p1`. For the electron channel again the path `Ele27_WP80` is chosen. Both triggers are not pre-scaled over the whole run period and both are modeled well in simulation.

All simulation samples are summarized in Table B.2 in Appendix B. The signal process tHq in the t-channel is simulated in the four-flavor scheme at leading order using MadGraph 5.1 for two different values of κ_f, -1 and $+1$. The sample only contains $H \to b\bar{b}$ decays, and the top quark is forced to decay leptonically into an electron, muon or tau. The templates extracted from these samples are scaled to next-to-leading order (NLO) predictions [55]. For this analysis possible contributions from tW-channel tHq production are neglected. The $t\bar{t}$ + jets and V+jets processes are also generated with MadGraph 5.1 and the templates are normalized to next-to-next-to-leading order (NNLO) cross sections [185, 186].

To account for possible mismodeling of $t\bar{t}$ production in association with additional b jets, the generated $t\bar{t}$ process is split into four samples depending on the quark flavors associated with the reconstructed jets in the event. The sample containing events in which at least two jets are matched to b quarks not stemming from a top quark decay is labeled as $t\bar{t}$+$b\bar{b}$. Events, in which only one jet can be assigned

to one or two extra b quarks, are grouped as $t\bar{t}$+b. This occurs, if one b quark is too soft or too forward to be detected, or if both b quarks are too close together to be resolved as two separate jets. At last, the template $t\bar{t}$+1,2c includes events, where at least one reconstructed jet is matched to a c quark. All other events are labeled as $t\bar{t}$ +LF, as they contain only jets connected to light flavor (LF) quarks or gluons. This procedure and the associated uncertainties described later in the chapter are adopted from the CMS $t\bar{t}$H analysis [187].

The single top processes, split in *t*-channel, *s*-channel and tW-channel, are simulated using POWHEG 1.0 and the templates are scaled to approximate NNLO [155]. For all of them, the top quark is forced to decay leptonically. The samples for the diboson and QCD production are produced with PYTHIA 6.4. While the diboson templates are normalized to LO or NLO prediction [186], the QCD templates are scaled using the leading order generator cross sections. The GEANT 4 package [156] was used to model the CMS detector response for all processes and an adequate admixture of additional pile-up events was added.

6.4. Physics objects and corrections

For this analysis again Particle Flow objects are used. This section covers the pre-selection criteria applied to all events. Moreover, additional corrections on the simulation are introduced. These adjustments are found to be necessary for a reasonable modeling of the data.

6.4.1. Pre-selection criteria on physics objects

The analysis relies on a trustworthy reconstruction of electrons, muons and jets stemming from a common primary vertex. Therefore, in the following the specific criteria on all PF objects are given.

Primary vertices and pile-up treatment

The selected primary vertex in each event is defined equally to the WH search. Again, Charged Hadron Subtraction (CHS) and the pile-up reweighting procedure is applied (see Section 5.4.1).

Electrons

The *loose* electron candidates in this analysis must satisfy the requirements $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ and $|\eta| < 2.5$ with a relative ρ-corrected isolation of $I_\rho < 0.15$. This isolation is defined as

$$I_\rho = \frac{I^\ell_{CH} + \max\left(I^\ell_{NH} + I^\ell_{Ph} - \rho A_{\mathrm{eff}}, 0\right)}{p_{\mathrm{T},\ell}}\,, \tag{6.1}$$

where I^ℓ_{CH}, I^ℓ_{NH} and I^ℓ_{Ph} indicate the energy generated by stable charged hadrons, neutral hadron and photons, respectively, in a cone with $\Delta R = 0.3$ around the

track of the electron. The average angular p_{T} density per event is denoted as ρ. The effective area A_{eff} is defined to compensate the components from photons and neutral hadrons in pile-up events. The CMS collaboration provides official values for it [188] depending on the pseudorapidity of the electrons. This definition varies from the one in Section 5.4.1 as the official recommendations changed.

The *tight* electron candidates are required to fulfill $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $|\eta| < 2.5$. Also the isolation requirement with $I_\rho < 0.10$ is more stringent. Candidates in the ECAL endcap-barrel transition region $1.4442 < |\eta_{\mathrm{sc}}| < 1.5660$, where $|\eta_{\mathrm{sc}}|$ is the pseudorapidity of the electron's supercluster, are rejected. Furthermore, the response from a multivariate identification technique [189], provided by the EGamma POG, needs to be larger than 0.9.

To eliminate differences between data and MC, efficiency scale factors are applied on simulated events. These factors are taken from the CMS electron efficiency measurement for top quark physics [190].

Muons

The *loose* muon candidates are required to be reconstructed as global muon and to fulfill $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ and $|\eta| < 2.5$ with a relative $\Delta\beta$-corrected isolation of $I_{\Delta\beta} < 0.2$. The corrected isolation is defined as

$$I_{\Delta\beta} = \frac{I^{\ell}_{CH} + \max\left(I^{\ell}_{NH} + I^{\ell}_{Ph} - 0.5 \cdot I^{\ell}_{CH,PU}, 0\right)}{p_{\mathrm{T},\ell}}\,, \tag{6.2}$$

where $I^{\ell}_{CH}, I^{\ell}_{NH}$ and I^{ℓ}_{Ph} are the energy generated by stable charged hadrons, charged hadron pile-up candidates, neutral hadron and photons, respectively, in a cone with $\Delta R = 0.4$ around the track of the muon.

For *tight* muon candidates the criteria on the transverse momenta, pseudorapidity and isolation are tautened, i.e. $p_{\mathrm{T}} > 26\,\mathrm{GeV}$, $\eta < 2.1$ and $I_{\Delta\beta} < 0.12$. In addition, the χ^2 value of the global fit applied in the reconstruction has to be smaller than 10. The transverse impact parameter of the muon's track with respect to the beam spot needs to lie within $|d_{xy}| < 2\,\mathrm{mm}$ and the distance between muon vertex and primary vertex in z direction is required to be smaller than 5 mm. Moreover, at least two muon stations have to be associated with the muon candidate and at least one hit in the pixel system and more than five tracker layers with measurements are needed.

Similar to electrons, efficiency scale factors are applied. These values, derived via a tag-and-probe technique in Drell-Yan events for the full 2012 dataset, are provided globally by the CMS *Muon* POG [191].

Jets

In this analysis jets clustered with the anti-k_{T} algorithm and a size parameter of 0.5 are used. The jets are cleaned from pile-up charged hadrons and isolated muons and electrons with $I_{\Delta\beta} < 0.2$ and $I_\rho < 0.15$, respectively. In addition to the standard

jet energy corrections (see Chapter 3), *L1FastJet* corrections are applied. Each jet has to pass the *loose* Particle Flow jet ID [192], that introduces requirements on the energy fractions clustered in jets. The transverse momenta of jets in simulated events are smeared according to an updated resolution measurement using dijet events [193].

Central jets with $|\eta| < 2.4$ are required to have a transverse momentum larger than 20 GeV. For forward jets with $2.4 < |\eta| < 4.7$ a tighter selection of $p_{\mathrm{T}} > 40\,\mathrm{GeV}$ is chosen. This criterion is motivated in the next section.

b Jet identification and reweighting

Similar to the WH search, the CSV algorithm [194] is used to identify jets stemming from b quarks. This analysis, however, does not exploit the whole shape of the CSV response. A simpler event-by-event weight is yet applied to account for efficiency differences between data and MC [195].

In a first step, for simulated events the b-tagging efficiencies are computed for each process independently and parameterized as functions of flavor, p_{T} and $|\eta|$ of the jet These efficiencies are then used to calculate the probability for an event with I tagged and J untagged jets to be correctly observed for both, simulation and data:

$$\mathcal{P}_{\mathrm{MC}} = \prod_{i \in \mathrm{tagged}}^{I} \epsilon_i \cdot \prod_{j \notin \mathrm{tagged}}^{J} (1 - \epsilon_j) \qquad \text{and} \tag{6.3}$$

$$\mathcal{P}_{\mathrm{Data}} = \prod_{i \in \mathrm{tagged}}^{I} s_i \epsilon_i \cdot \prod_{j \notin \mathrm{tagged}}^{J} (1 - s_j \epsilon_j)\,. \tag{6.4}$$

Here, ϵ_i indicates the computed MC b-tagging efficiencies for each b-tagged jet i and s_i denotes its scale factor measured in data. Consequently, for each event a weight $w = \mathcal{P}_{\mathrm{Data}}/\mathcal{P}_{\mathrm{MC}}$ is assigned.

Missing transverse energy and W boson reconstruction

The PF missing transverse energy including type-0 and type-1 corrections described in Section 3.2.6 is used for the analysis. In addition, xy-shift-corrections are applied, that mitigate the ϕ modulation of $\not\!\!E_{\mathrm{T}}$ and also reduce pile-up effects [131].

In contrast to the WH search, in this analysis the three-dimensional reconstruction of the leptonically decaying W boson is important. Besides assigning the x and y components of the missing transverse energy vector to $p_{x,\nu}$ and $p_{y,\nu}$, the z component of the neutrino's momentum can be obtained by the following reasoning. The neutrino and the charged lepton originate from the W boson. When neglecting the invariant masses of the neutrino and the charged lepton and assuming the W boson is produced on-shell, i.e. the invariant mass of the W boson is set to $m_{\mathrm{W}} = 80.4\,\mathrm{GeV}$,

a quadratic equation for $p_{z,\nu}$ can be derived:

$$m_{\mathrm{W}}^2 = \left(E_\ell + \sqrt{\vec{\not{p}}_{\mathrm{T}}^{\,2} + p_{z,\nu}^2} \right)^2 - \left(\vec{p}_{\mathrm{T},\ell} + \vec{\not{p}}_{\mathrm{T}} \right)^2 - \left(p_{z,\ell} + p_{z,\nu} \right)^2 , \tag{6.5}$$

where $\vec{p}_{\mathrm{T},\ell}$ and E_ℓ are the transverse momentum and the energy of the charged lepton, respectively. Further, $p_{z,\ell}$ and $p_{z,\nu}$ denote the z components of the four-momenta of the charged lepton and the neutrino, respectively. The solution of the quadratic equation (6.5) is given by

$$p_{z,\nu}^{\pm} = \frac{\mu p_{z,\ell}}{p_{\mathrm{T},\ell}^2} \pm \sqrt{\frac{\mu^2 p_{z,\ell}^2}{p_{\mathrm{T},\ell}^4} - \frac{E_\ell^2 p_{\mathrm{T},\nu}^2 - \mu^2}{p_{\mathrm{T},\ell}^2}} , \tag{6.6}$$

where $p_{\mathrm{T},\nu}^2$ denotes the transverse momentum of the neutrino and the abbreviation μ is defined as

$$\mu = \frac{m_W^2}{2} + p_{\mathrm{T},\ell} \cdot p_{\mathrm{T},\nu} \cdot \cos\Delta\phi \, . \tag{6.7}$$

Here, $\Delta\phi$ is the azimuthal angle between the charged lepton and the neutrino. When the discriminant in Equation (6.6) is positive, there are two real solutions for $p_{z,\nu}$, and the one with the smaller absolute value is taken. In the case of a negative discriminant, that arises due to the finite $\not{E}_{\mathrm{T}}$ resolution, there are two solutions with imaginary contributions. Under the assumption that the imaginary part arises from imperfect $\not{E}_{\mathrm{T}}$ measurements, in these cases $p_{x,\nu}$ and $p_{y,\nu}$ are varied such, that the discriminant vanishes and one real solution for the z-component of the neutrino vector is found [196].

6.4.2. Additional corrections to simulated events

There are two additional treatments needed for the simulated samples in the analysis due to differences between MC and data. First, the event reweighting to account for data/MC differences in the p_{T} spectrum of reconstructed top quarks is described. Whereas this top quark p_{T} reweighting has already been used in a large number of analyses within the CMS collaboration, the jet pseudorapidity handling introduced afterwards was investigated in the context of the tHq search [61] for the first time.

Top quark p_{T} reweighting

In simulated events the p_{T} spectra of top quarks, and hence of their decay products, are harder than what is observed in data. In Figure 6.4(a) the situation is shown for the $p_{\mathrm{T}}(\mathrm{W})$ distribution in the $\mathrm{t\bar{t}}$-enriched control region used in the analysis. Originally, the behavior was found in the CMS analysis measuring the normalized differential $\mathrm{t\bar{t}}$ cross section. Based on this measurement the CMS collaboration provides event weights depending on the generator-level top quark p_{T} spectrum [197]. Figure 6.4(b) shows the same distribution after the reweighting, and a clear improvement in the shape modeling is found. It should be noted that this treatment is only applied to simulated $\mathrm{t\bar{t}}$ events.

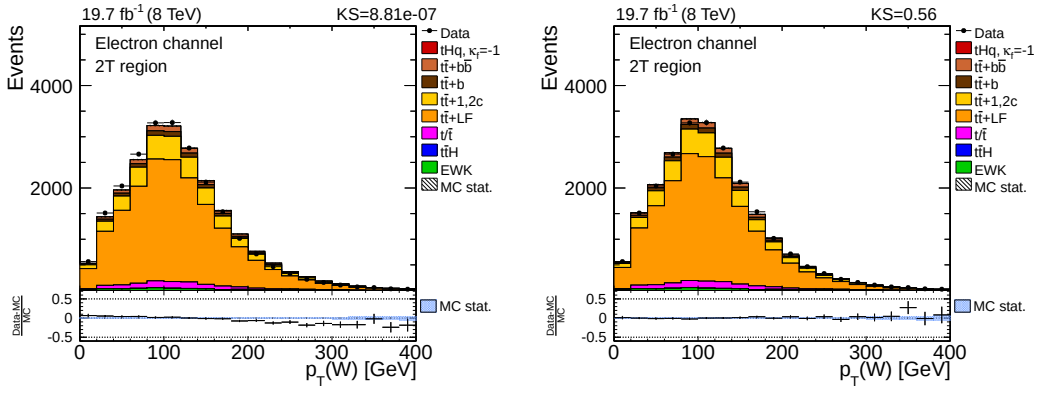


Figure 6.4.: Effect of top quark p_{T} reweighting in the 2T control region for the electron channel. The transverse momentum of the reconstructed W boson before (left) and after (right) applying the top quark p_{T} correction factors is shown. In both distributions the simulation is normalized to fit the observed events in data. The ratio in the left diagram indicates a clear slope that is cured in the right diagram. The KS-test probabilities are also vastly improved.

Jet pseudorapidity handling

As this analysis also relies on the information of low energetic jets at high pseudorapidities, it has high demands on the modeling in MC. Within the context of the tHq search, a visible discrepancy between simulation and data was found in the pseudorapidity distribution of low-energetic jets. In the following the source of the issues and their final treatments are described briefly, while more information can be found in [198, 199].

For jets with low transverse momenta ($20\,\mathrm{GeV} < p_{\mathrm{T}} < 40\,\mathrm{GeV}$) two different effects resulting in mismodeling are observed. On the one hand, there is a depression in the data/MC ratio with a peak at $|\eta| \sim 2.7$. By comparing the pseudorapidities of reconstructed jets and the corresponding jets at generator level a bias is found. Generator jets with $|\eta| \gtrsim 2.5$ or $|\eta| \lesssim 3.1$ migrate towards $|\eta| \sim 2.7$ when reconstructed. This migration effect is masked in the analysis by taking only one single bin for $2.4 < |\eta| < 3.2$ into account. On the other hand, there is a slope in the pseudorapidity distribution for the region with $|\eta| \gtrsim 3.0$. This is interpreted as a binning effect from the *L2L3Residual* jet energy corrections, since they are derived in one single bin with $|\eta| > 3.139$ due to the lack of more events in data. Hence, the adjustments are not expected to correct for any data/MC shape discrepancies within this region.

Both effects are found to be most pronounced for jets with low transverse momenta. Therefore, forward jets with $|\eta| > 2.4$ are required to have a p_{T} larger than 40 GeV and only two bins for the forward regions are taken into account. This inequidistant binning for jet pseudorapidity is showed for instance in Figure 6.7. Due to this treatment, the observed differences between data and MC vanish.

6.5. Selection of events

To be sensitive at all to the small amount of expected signal events a signal enhanced phase space needs to be defined in which the vast amount of $t\bar{t}$ production events is suppressed while keeping as many as possible signal events. In the following the selection criteria for signal and control regions are described. Moreover, the results of the data-driven QCD estimation via the `ABCD` method are presented.

6.5.1. Definition of signal and control regions

The tHq analysis focuses on events, in which the single top quarks decay leptonically into electrons or muons and neutrinos. An event needs to have exactly one *tight* electron or muon to enter the electron or muon channel, respectively. An additional *loose* lepton veto is introduced to suppress the contribution from Drell-Yan production.

The challenging multijet signature of tHq production with $H \to b\bar{b}$ decays contains 4 b quarks (see Figure 6.2(a)). As previously mentioned, the *spectator b quark* stemming from initial gluon splitting leaves the detector undetected in many cases due to its low transverse momentum. To avoid losing tHq events, two signal regions are defined asking the event to contain either three b-tagged jets (3T region) or four b-tagged jets (4T region). The *tight* working point of the CSV tagger is used (CSV > 0.898). This minimizes the misidentification of jets. To suppress the background processes even further, every event is required to have more than three jets with $p_T > 30\,\text{GeV}$ (or $p_T > 40\,\text{GeV}$ in case of forward jets). In addition, at least one untagged jet is required to account for the expected light jet. This cut is redundant in the 3T region. The QCD estimation described in the next section prompts an auxiliary cut on the missing transverse energy to suppress the multijet contribution to less than one percent. All requirements for the defined signal enhanced phase space are summarized in Table 6.1.

In total only 0.7% in the 3T and 2.1% in the 4T region of the events in data are expected to be stemming from tHq production. For the 4T region only 70 events in data are expected in total making its sensitivity suffer from large statistical uncertainties. The detailed yield comparisons between data and MC in the signal regions are only presented in Section 6.9 after the simulation is fit to data.

As already mentioned, an accurate understanding of the $t\bar{t}$ + jets background process is crucial to the analysis. Hence, control regions enriched in $t\bar{t}$ production are defined by asking for exactly two b-tagged jets in an event. The other requirements are chosen accordingly to the signal regions. The expected yields for the control regions in electron and muon channels are given in Table 6.2. These regions, in which the validation of all used variables is performed, are almost pure in $t\bar{t}$ production (above 94%). Figure 6.5 shows the data/MC agreement for some event variables in the electron channel. The corresponding diagrams for the muon channel are depicted in Figure B.1. Overall decent agreement is found.

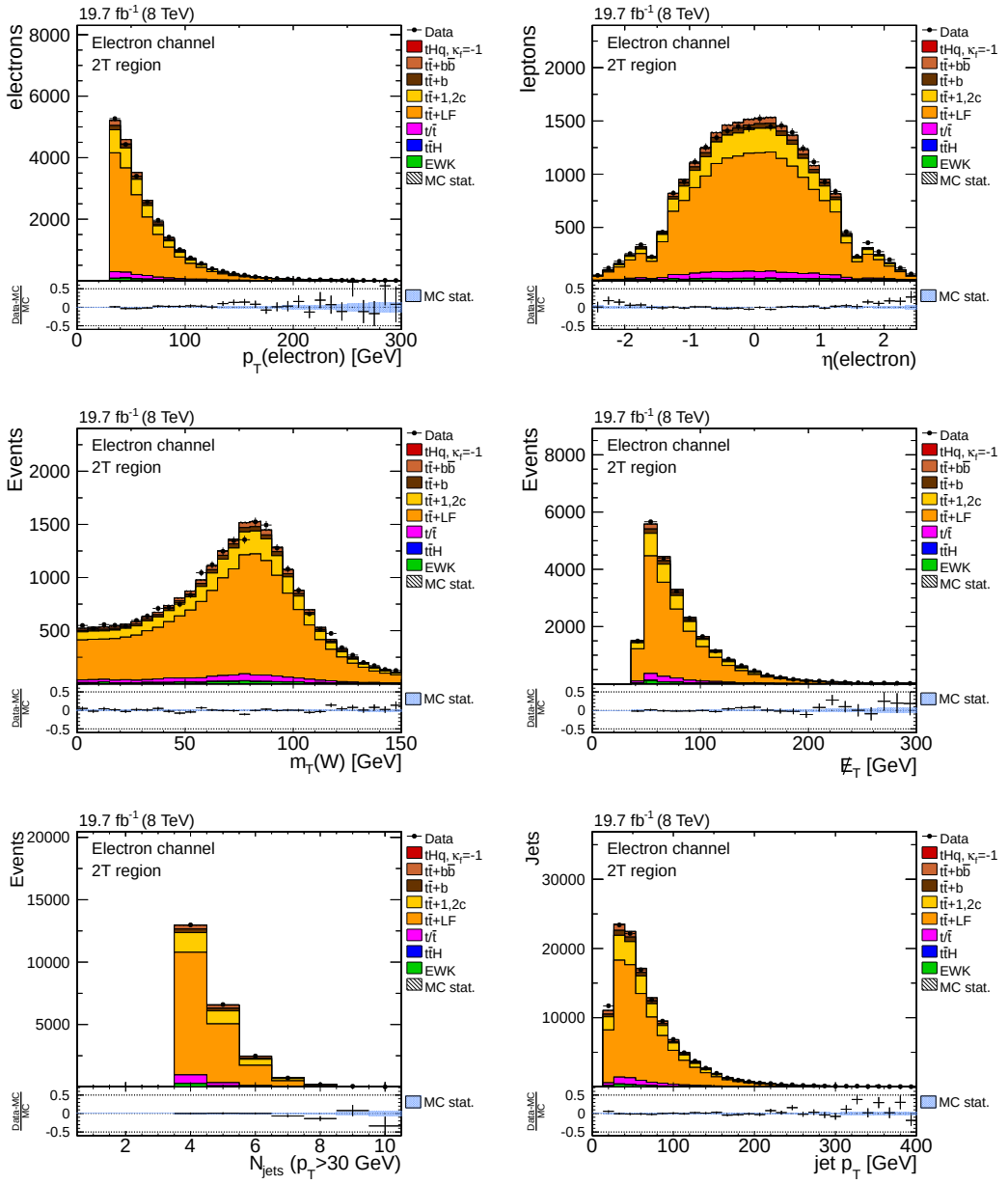


Figure 6.5.: Event information variables for the electron channel in the 2T region. The number of simulated events is normalized to data. Overall good agreement between data and simulation is found. The corresponding distributions for the muon channel are given in Figure B.1.

Table 6.1.: Selection criteria for the 3T and 4T signal regions. As indicated with parentheses there are different cuts for electron and muon channels.

	3T region	4T region
# tight leptons	1	1
# add. loose leptons	0	0
$\not{E}_{\mathrm{T}}$	$> 35/45\,\mathrm{GeV}(\mu/\mathrm{e})$	$> 35/45\,\mathrm{GeV}(\mu/\mathrm{e})$
# (central jets with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ + forward jets with $p_{\mathrm{T}} > 40\,\mathrm{GeV}$)	≥ 4	≥ 4
# additional jets	-	≥ 1
# jets with CSV > 0.898	$= 3$	$= 4$

Table 6.2.: Expected yields for signal and background processes in the 2T control regions. The number of simulated events is normalized to luminosity and all corrections are applied. Additionally, the purity in $\mathrm{t\bar{t}}$ events is given.

2T region	Electron channel	Muon channel
tHq	16.2 ± 0.1	22.5 ± 0.1
$\mathrm{t\bar{t}} + \mathrm{b\bar{b}}$	810.4 ± 6.0	1099.8 ± 7.0
$\mathrm{t\bar{t}} + \mathrm{b}$	722.7 ± 5.7	990.8 ± 6.6
$\mathrm{t\bar{t}} + 1,2\mathrm{c}$	3677.7 ± 12.8	5110.3 ± 15.0
$\mathrm{t\bar{t}} + \mathrm{LF}$	18346.6 ± 28.5	26076.5 ± 33.9
$\mathrm{t}/\bar{\mathrm{t}}$	1059.5 ± 10.2	1481.0 ± 11.9
$\mathrm{t\bar{t}H}$	35.9 ± 0.4	46.3 ± 0.4
Diboson	21.7 ± 1.2	29.5 ± 1.4
W+jets	343.0 ± 11.6	493.7 ± 14.0
Total MC	25034	35351
Purity in $\mathrm{t\bar{t}}$	$94.1 \pm 0.2\%$	$94.1 \pm 0.2\%$

6.5.2. Data-driven QCD estimation

After assigning the tight event selection criteria there are merely a few events in QCD MC samples left. Therefore, a simulation based estimation of the expected amount of QCD multijet production is critical. Similar to the WH search, the analysis avails the `ABCD` method for this purpose (compare Section 5.6.4). For the two uncorrelated requirements on the one hand the missing transverse energy and on the other hand the lepton isolation plus the MVA identification value in case of the electron channel are used. The resulting four regions read as follows.

- **Region A**: ordinarily-defined signal or control region used in this analysis.
- **Region B**: as region A, but the lepton isolation and identification cuts are inverted.

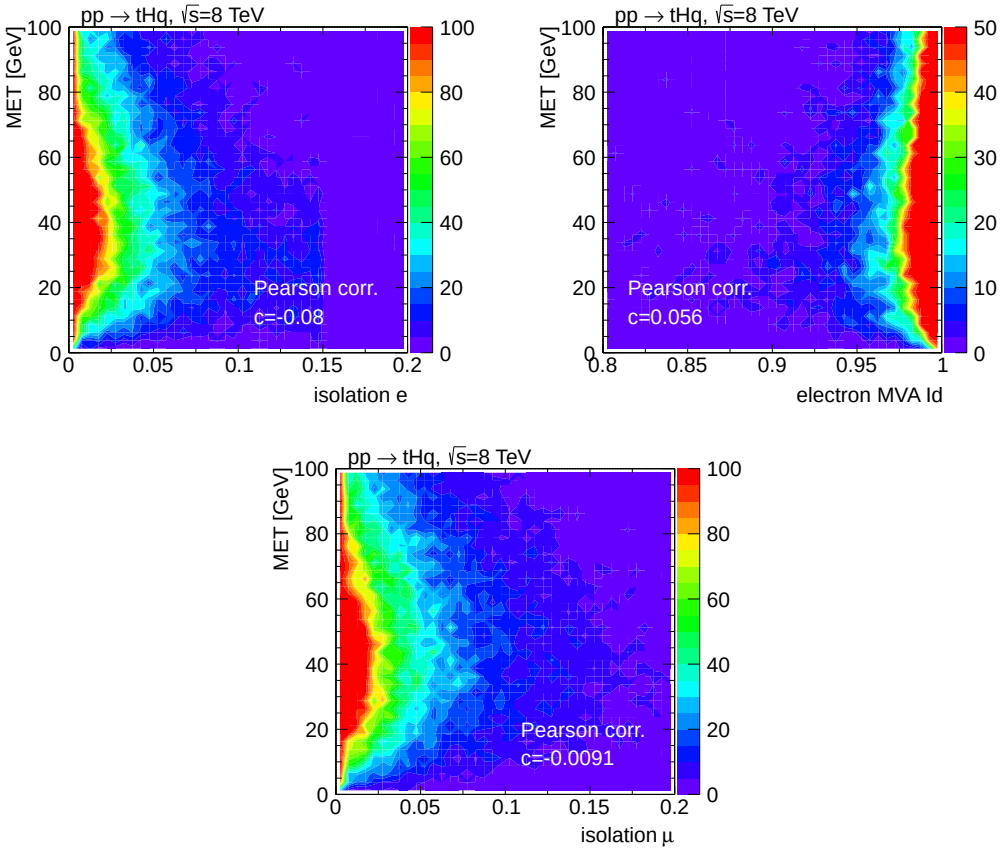


Figure 6.6.: Correlation between input variables for the ABCD method. The top row shows the correlation in data for the electron channel. Since in regions B and D both, the electron isolation and the MVA ID are inverted, two separate diagrams are shown. In the bottom row the correlation in data for the muon channel is depicted. For all three the correlation factors are given concurring in all cases with zero correlation.

- **Region C**: as region A, but the cut on $\not\!\!E_T$ is inverted.
- **Region D**: all cuts described above are inverted.

The diagrams in Figure 6.6 oppose the above variables. The assumption of negligible correlations is justified with the given Pearson coefficients. Using Equation (5.8) the expected contribution of QCD is calculated in all signal and control regions.

In this search, the ABCD method is also used to optimize additional cuts on $\not\!\!E_T$ that suppress the expected contribution of QCD events to below 1% without losing too many signal events. The requirements $\not\!\!E_T > 45\,\text{GeV}$ and $\not\!\!E_T > 35\,\text{GeV}$ for the electron and muon channels, respectively, were found this way. Table B.3 in the Appendix shows the detailed results for the signal and control regions in electron and muon channels.

Furthermore, a closure test is performed to validate the use of the ABCD method. To obtain enough generated events in the QCD samples, this test is only possible in a phase-space requiring less than two b-tagged jets in the event. In this region 50,000 pseudo experiments with randomly varied event yields according to their statistical uncertainties are drawn. Using Equation (5.8) for each pseudo dataset the result from the ABCD method N_{ABCD} is compared to the actual number of events drawn in region A, $N_{QCD,toy}$. The relative difference between the two is shown in Figure B.2 for all pseudo datasets and only a small bias is found. Hence, the ABCD method is expected to yield on average an insignificantly higher result. Based on these studies QCD production is neglegted in the further analysis.

6.6. Reconstruction of events using MVAs

To exploit the full information to separate signal from background events, the four-momenta of Higgs boson and top quark need to be reconstructed from the measurable detector objects. The reconstruction of the W boson is unequivocal as previously reported. However, the assignment between jets and final state quarks is ambiguous. An event-by-event decision criterion which combination to select needs to be defined. The strategy used in this analysis is reconstructing all possible jet-quark assignments — in the following called *interpretations* — and letting an MVA tool judge, which interpretation decribes the event best.

A peculiarity of this analysis is that the events are not only reconstructed under the assumption they were signal events. To also get a handle on how background-like an event is, every event is additionally reconstructed under the assumption it stems from $t\bar{t}$ production. In the following sections the two reconstructions are described.

6.6.1. Jet assignment under the tHq hypothesis

In tHq production events the selected jets need to be assigned to the two b quarks from the Higgs boson decay, the one b quark from the top quark decay and the light forward jet. The assignment of a jet to the spectator b quark is neglected and only considered indirectly via the selection cuts.

On signal MC for every event a *correct* interpretation, where all four quarks are matched to a jet within a cone with $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} = 0.3$, is defined. When one of the searched-for jets is not reconstructed or selected, there is no such *correct* interpretation and the event is not used for the training. All other possible interpretations are declared as *wrong* interpretations.

To reduce the combinatorics for the *wrong* interpretations loose requirements, that are met by almost every *correct* interpretation, are applied. On the one hand, only central jets ($|\eta| < 2.4$) are allowed to be assigned to the b quarks from the top quark and Higgs boson decays. This is justified, as only central jets lie within the tracker acceptance region and have valid b-tagging information. On the other hand, the jets assigned to the light quark must not have a tight b-tag with CSV > 0.898.

For the reduced set of *wrong* interpretations only one is chosen randomly per event for the training.

A neural network trained with the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, as implemented in the TMVA toolkit [134], is used to discriminate between *correct* and *wrong* interpretations. The following parameters are set in the configuration:

```
NCycles = 500; HiddenLayers = 30; NeuronType = tanh; TrainingMethod = BFGS .
```
(6.8)

Here, the option `HiddenLayers = 30` gives rise to one hidden layer with 30 neurons. For the training the 3T and 4T regions are considered together and only every fifth event is exploited. All events used in the training are removed from the succeeding analysis to avoid a training bias. Table 6.3 lists the input variables, ranked by their importance in the training. The b-tagging information from all three b quarks and

Table 6.3.: Input variables for the tHq reconstruction. The variables are ranked according to their importance in the MVA training. The jet charge is defined as the p_{T}-weighted charge of all particles in the jet, i.e. $1/p_{\mathrm{T}}(\mathrm{jet}) \sum_{j \in \mathrm{jet}} Q_j p_{\mathrm{T}}^j$ [200]

Variable	**Rank**	**Description**
bool tagged($\mathrm{b_t}$)	1.	Equals 1 if the b quark jet from the top quark decay is b-tagged, 0 otherwise
$\lvert\eta(\text{light jet})\rvert$	2.	Absolute value of the light forward jet's pseudorapidity
# b-tags of H jets	3.	Number of b-tagged jets among the two jets from the Higgs boson decay
$m(\mathrm{b_t} + \ell)$	4.	Invariant mass of the charged lepton and the b quark jet stemming from the top quark decay
$m(\mathrm{H})$	5.	Mass of the reconstructed Higgs boson
$\min\left(p_{\mathrm{T}}(\mathrm{b}_{i,\mathrm{H}})\right)$	6.	Transverse momentum of the softest jet from the Higgs boson decay
$\Delta R(\mathrm{b}_{1,\mathrm{H}}, \mathrm{b}_{2,\mathrm{H}})$	7.	ΔR between the two jets from the Higgs boson decay
$\max\lvert\eta(\mathrm{b}_{i,\mathrm{H}})\rvert$	8.	Pseudorapidity of the most forward jet from the Higgs boson decay
$\Delta R(\mathrm{b_t}, \mathrm{W})$	9.	ΔR between the b quark jet and the W boson from the top quark decay
relative H_{T}	10.	Relative H_{T}, $(p_{\mathrm{T}}(\mathrm{t}) + p_{\mathrm{T}}(\mathrm{H}))/H_{\mathrm{T}}$
$\Delta R(\mathrm{t}, \mathrm{H})$	11.	ΔR between the reconstructed top quark and Higgs boson
$Q(\mathrm{b_t}) \times Q(\ell)$	12.	Electric charge of the b quark jet from the top quark decay multiplied by the lepton's charge

the pseudorapidity of the light jet are found to be the most important variables,

followed by the reconstructed masses of top quark and Higgs boson. The shapes of those variables for *correct* and *wrong* interpretations are depicted in Figures 6.7. All other input variables can be found in Figure B.3 in the Appendix. For all variables a clear separation is visible.

The response of the resulting MVA for the training sample is provided in Figure 6.8 separately for *correct* and *wrong* interpretations. Additionally overlayed are the corresponding shapes for an independent set of events to check for possible overtraining. The shapes in training and testing samples show good agreement, and therefore no hint for overtraining is found.

When reconstructing an unknown event, for every possible interpretation meeting the requirements the MVA response is calculated and the one with the largest value is chosen for this event.

To demostrate the sanity of this reconstruction a validation is performed in the 2T region. Here, the MVA distributions when chosing a random interpretion for each event are compared for data and MC. In this way, any differences in performance would be visible. The diagram depicted in Figure 6.9 shows solid agreement. The data/MC comparisons for the best MVA responses in the 3T and 4T regions are provided in Figure B.4 in the Appendix.

After the sanity check, it is interesting to know how well the reconstruction performs. In the best case the reconstructed interpretation, i.e. the one with the largest MVA output, is the best possible interpretation, i.e. the one with the smallest sum of distances ΔR between jets and final state quarks. Therefore, in signal MC for every event all possible interpretations are ranked according to $\sum \Delta R$ and the position of the reconstructed interpretation is checked, as shown as red striped shape in Figures 6.10(a) and (b) for the 2T and 3T regions. The height of the leftmost bin gives the percentage of all cases in which the reconstructed and best possible interpretations coincide. This happens in over 30% of all cases in the 3T region, and in about 28% of all cases in the 2T region. In the latter the selection criteria are not designed for the signiture of signal events, so the lower efficiency is expected.

As cross check a basic χ^2 reconstruction is executed. For every event the interpretation with the smallest value for

$$\chi^2 = \frac{[m(\mathrm{t_{int}}) - m(\mathrm{t})]^2}{\sigma^2(\mathrm{t})} + \frac{[m(\mathrm{H_{int}}) - m(\mathrm{H})]^2}{\sigma^2(\mathrm{H})} \tag{6.9}$$

is reconstructed. Here, $m(\mathrm{t_{int}})$ and $m(\mathrm{H_{int}})$ denote the reconstructed masses of top quark and Higgs boson per interpretation and $m(\mathrm{t}) = 173\,\mathrm{GeV}$, $m(\mathrm{H}) = 125\,\mathrm{GeV}$, $\sigma(\mathrm{t}) = 30\,\mathrm{GeV}$ and $\sigma(\mathrm{H}) = 15\,\mathrm{GeV}$ are set. The outcome is depicted as solid yellow shapes in Figure 6.10. Compared to the χ^2 reconstruction, the MVA yields broader invariant mass distributions for top quark and Higgs boson, but meets the best possible interpretion in far more cases. Given these facts, the use of the presented reconstruction technique is justified.

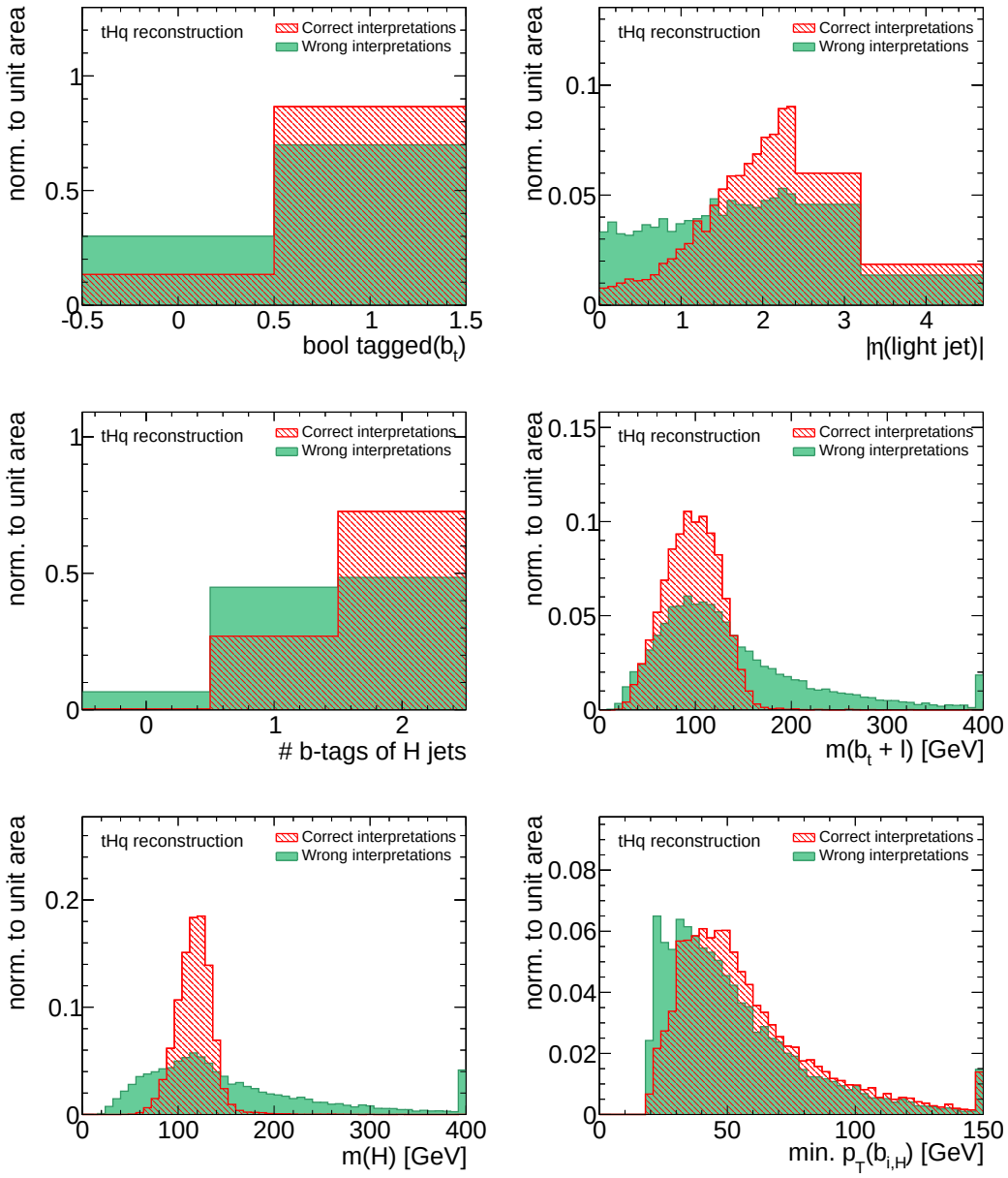


Figure 6.7.: Shapes of most important input variables for reconstruction under tHq hypothesis. The shapes are shown separately for *correct* and *wrong* interpretations. The corresponding diagrams for the remaining input variables can be found in Figure B.3. As the *correct* interpretations act as signal in the training, their distributions are consistently shown in red.

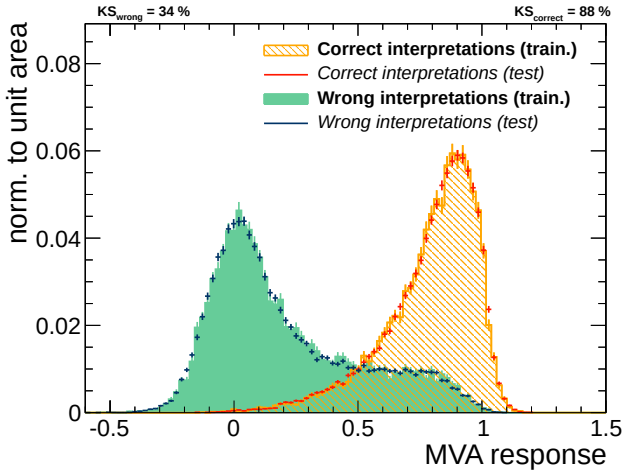


Figure 6.8.: MVA response in tHq reconstruction for *correct* and *wrong* interpretations. The shapes are shown separately for the training and test sample. By comparing the shapes a good agreement and no hint for overtraining is found, supported by the given KS-test probabilities.

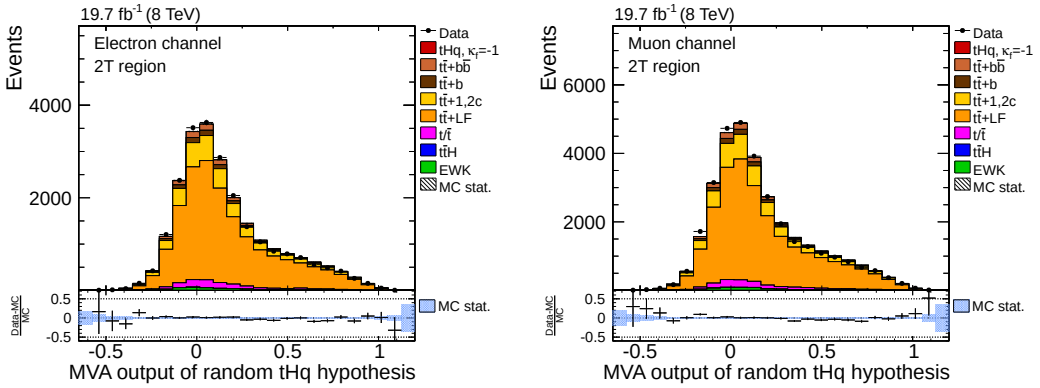


Figure 6.9.: Validation of tHq reconstruction in 2T region. The MVA response is shown for data and MC when choosing a random interpretation per event. In all distributions solid agreement between data and MC is found.

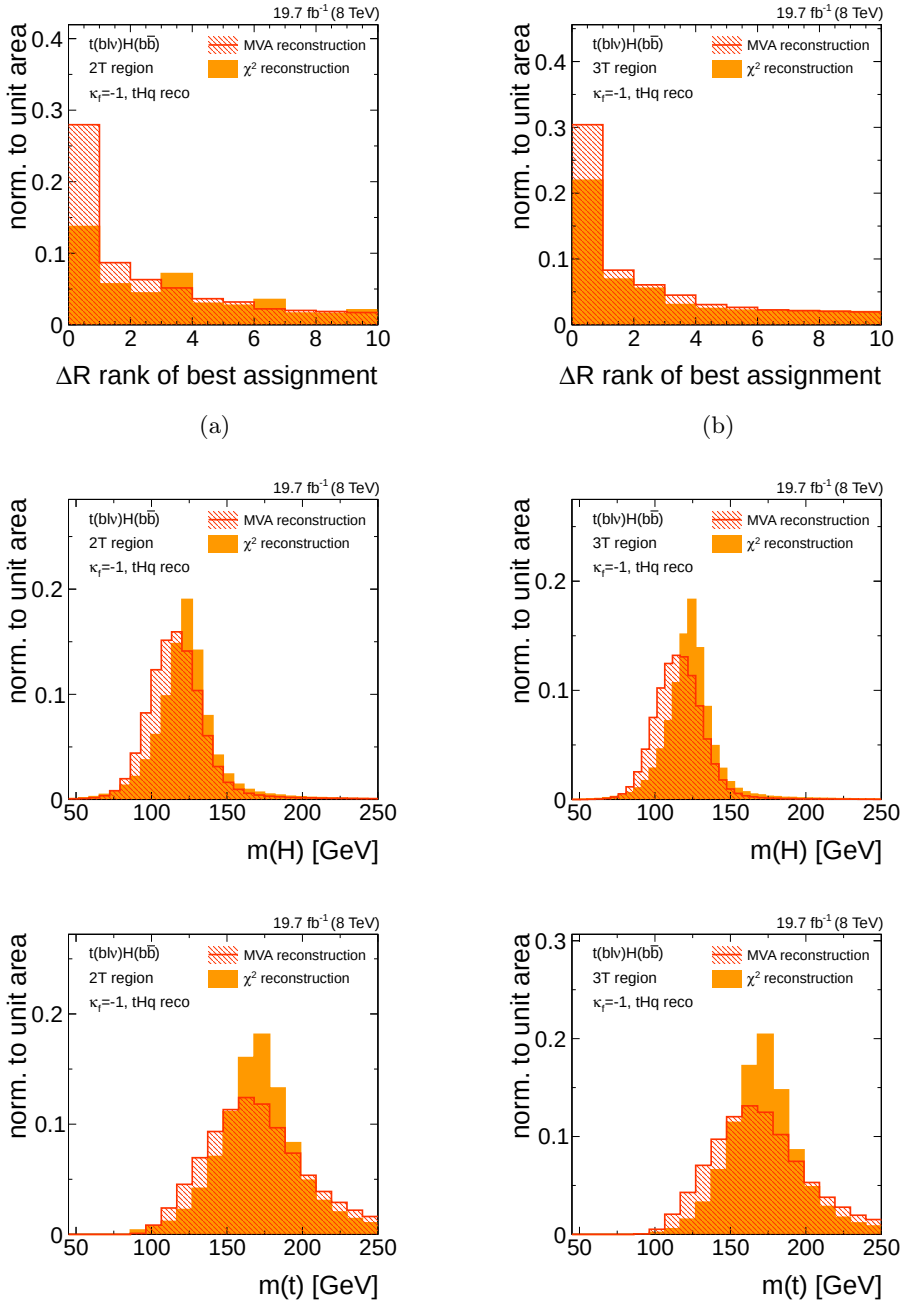


Figure 6.10.: Comparison for tHq reconstruction between MVA and χ^2 reconstruction techniques. On the left side the ΔR rank of the chosen hypothesis and the mass distributions of reconstructed Higgs boson and top quark candidates are shown in the 2T region. On the right side the same diagrams are given in the 3T region. Because the χ^2 values depend on the difference between reconstructed and true mass, the mass distributions are narrower compared to the MVA reconstruction. However, with the MVA reconstruction the correct jet assignment is found more often, leading to a better description of the kinematics of the event.

6.6.2. Jet assignment under the $t\bar{t}$ hypothesis

As mentioned before, an additional reconstruction is performed, assuming the events arise from semi-leptonic $t\bar{t}$ production. In the following, the partons stemming from the leptonically decaying top quark t_{lep} have the index *lep*, while the decay products from the hadronically decaying top quark t_{had} are labeled with *had*. To account for the signature of $t\bar{t}$ pair production, the jets need to be assigned to one b quark from each of the top quark decays and to two light quarks from the W_{had} decay.

The procedure is kept similar to the tHq reconstruction. Again, the interpretation in which all four quarks are matched to jets within $\Delta R = 0.3$ is defined as *correct*. Not every event has a *correct* interpretation. If so, this event is not used for the training. To reduce combinatorics the amount of *wrong* interpretions is restrained by requiring the jets assigned to the two b quarks to be central ($|\eta| < 2.4$) with a tight b-tag (CSV > 0.898). Another MVA with the same configurations as given in (6.8) is trained to discriminate between *correct* and *wrong* interpretations. For the latter set, only one interpretation is chosen randomly per event.

Table 6.3 lists the input variables, all possessing a satisfactory discrimination power between *correct* and *wrong* interpretations. The variables are ranked according to their importance in the training. Here, in general the kinematic information from the hadronically decaying top quark are found to be the most significant. This is due to the constrained allocation for the b quarks, so the MVA puts more weight in assigning the two light quarks from the W_{had} decay. The shapes of the most important input variables are provided in Figure 6.11 separately for *correct* and *wrong* interpretations. In Figure B.5 in the appendix the distributions of the remaining variables can be found.

The resulting response of the MVA is given in Figure 6.12. Again, the comparisons of training and test samples are provided. By applying a KS-test no hint for overtraining can be found.

The MVA response in data and MC for randomly chosen hypotheses is compared in the 2T region, to facilitate the same validation as for the tHq reconstruction. The diagrams in Figure 6.13 show the resulting distributions. It is visible that the data does not agree well with the simulation. Apparently there is a slight discrepency between data and simulation that is inflated by the combinatorics when choosing random interpretations. However, for each event the hypothesis with the largest MVA output is selected in the end. Therefore, in Figure B.6 the distributions for the actual reconstructed interpretations are shown. The provided KS-test probabilities in these diagrams encourage the further use of this reconstruction technique.

To quantify the performance for this method, again a χ^2 criterion is built from the masses of t_{lep}, t_{had} and W_{had}:

$$\chi^2 = \frac{[m(t_{lep,int}) - m(t)]^2}{\sigma^2(t)} + \frac{[m(t_{had,int}) - m(t)]^2}{\sigma^2(t)} + \frac{[m(W_{had,int}) - m(W)]^2}{\sigma^2(W)} . \quad (6.10)$$

Table 6.4.: Input variables for the $\mathrm{t\bar{t}}$ reconstruction. Information of the hadronically and leptonically decaying top quark is indexed with *had* and *lep*, respectively. The variables are ranked according to their importance in the MVA training.

Variable	**Rank**	**Description**
$m(\mathrm{W_{had}})$	1.	Mass of the W boson from the $\mathrm{t_{had}}$ decay
$m(\mathrm{t_{had}}) - m(\mathrm{W_{had}})$	2.	Difference between masses of $\mathrm{t_{had}}$ and the W boson from the $\mathrm{t_{had}}$ decay
$\Delta R(\mathrm{q_{1,had}}, \mathrm{q_{2,had}})$	3.	ΔR between the two light-flavor jets from the $\mathrm{t_{had}}$ decay
$\lvert\eta(\mathrm{t_{had}})\rvert$	4.	Absolute value of pseudorapidity of $\mathrm{t_{had}}$
$p_{\mathrm{T}}(\mathrm{t_{had}})$	5.	Transverse momentum of $\mathrm{t_{had}}$
# b-tags of $\mathrm{t_{had}}$ light jets	6.	Number of b-tagged jets among the two light-flavor jets from the $\mathrm{t_{had}}$ decay
$p_{\mathrm{T}}(\mathrm{t_{lep}})$	7.	Transverse momentum of $\mathrm{t_{lep}}$
$\Delta R(\mathrm{b_{had}}, \mathrm{W_{had}})$	8.	ΔR between b quark jet and W boson from the $\mathrm{t_{had}}$ decay
relative H_{T}	9.	Relative H_{T}, $(p_{\mathrm{T}}(\mathrm{t_{had}}) + p_{\mathrm{T}}(\mathrm{t_{lep}}))/H_{\mathrm{T}}$
$\Delta R(\mathrm{b_{lep}}, \mathrm{W_{lep}})$	10.	ΔR between b quark jet and W boson from the $\mathrm{t_{lep}}$ decay
$m(\mathrm{b_{lep}} + \ell)$	11.	Invariant mass of the charged lepton and the b quark jet from the $\mathrm{t_{lep}}$ decay
$Q(\ell) \times [Q(\mathrm{b_{had}}) - Q(\mathrm{b_{lep}})]$	12.	Difference of electric charges of the b quark jets from the $\mathrm{t_{had}}$ and $\mathrm{t_{lep}}$ decays, multiplied by the lepton's charge
$Q(\ell) \times [Q(\mathrm{q_{1,had}}) + Q(\mathrm{q_{2,had}})]$	13.	Sum of electric charges of the two light-flavor jets from the $\mathrm{t_{had}}$ decay, multiplied by the lepton's charge

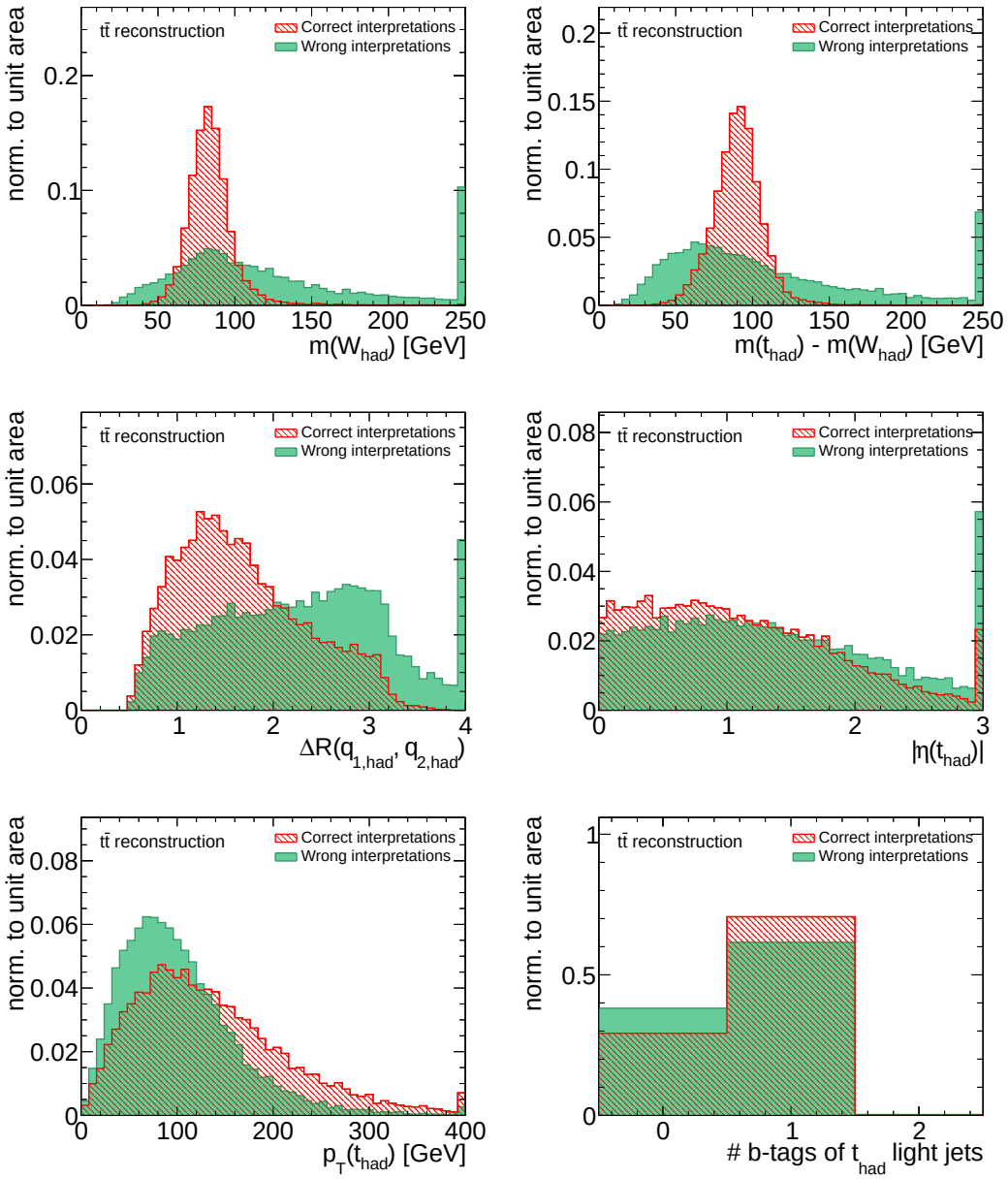


Figure 6.11.: Shapes of most important input variables for reconstruction under $t\bar{t}$ hypothesis. The distributions are shown separately for *correct* and *wrong* interpretations. The corresponding diagrams for the remaining input variables can be found in Figure B.5.

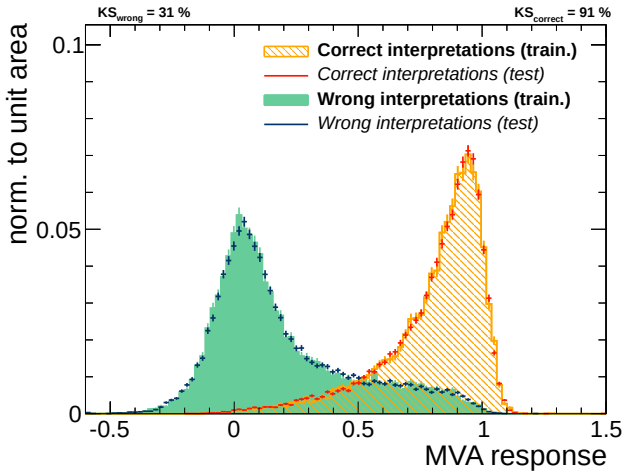


Figure 6.12.: MVA response in $t\bar{t}$ reconstruction for *correct* and *wrong* interpretations. The shapes are shown separately for the training and test samples. The KS-test yields probabilities above 30% (wrong interpretations) and above 90% (correct interpretations), so no hint for overtraining is found.

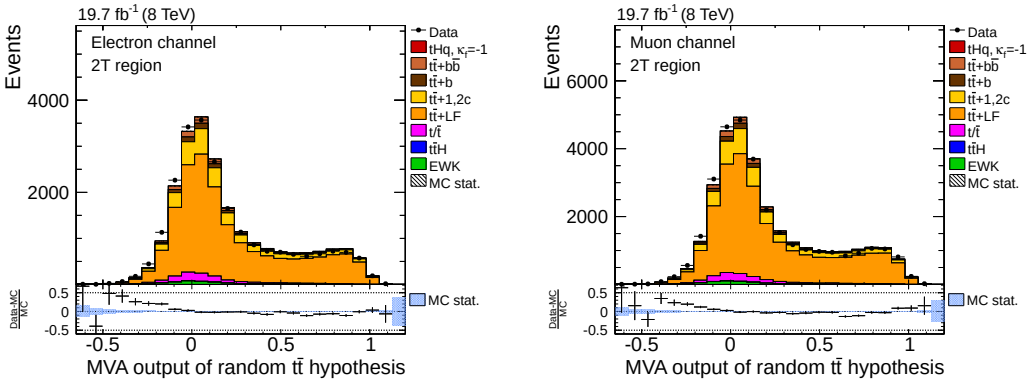


Figure 6.13.: Validation of $t\bar{t}$ reconstruction in the 2T region. The MVA response is shown for data and MC when choosing a random interpretation per event. The data does not agree well with the simulation. When selecting the hypotheses with the largest MVA outputs, the agreement in the MVA response distribution is good, as shown in Figure B.6.

In this equation $m(\mathrm{W})$ is set to 80.4 GeV and $\sigma(\mathrm{W})$ equals 11.9 GeV. The values for $m(\mathrm{t})$ and $\sigma^{(}\mathrm{t})$ are set to 173 GeV and 30 GeV, respectively. The comparisons in Figure 6.14 show that the MVA reconstruction outperforms this simple χ^2 method, especially in the 3T region.

6.7. Classification of events

The small signal-over-background ratio in the 3T and 4T regions makes a distinctive discrimination between tHq production and the background processes essential. With the two sets of information from the two reconstructions that are performed in parallel, together with global variables not depending on any interpretation, the analysis has a large pool of observables at hand.

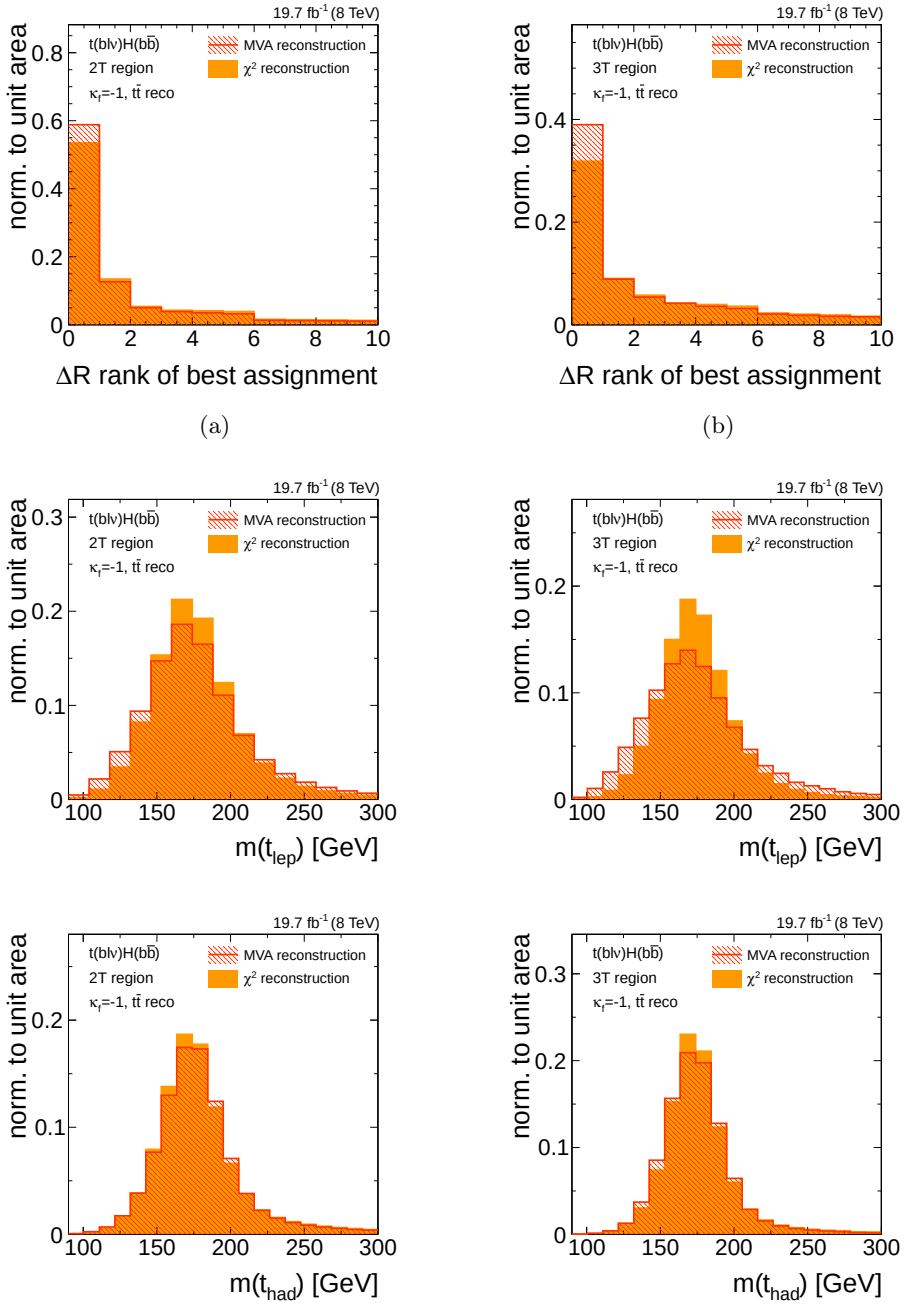
To make the most of the correlations between all chosen input variables, again an MVA method with the configuration given in (6.8) is trained. The set of background processes used for the training only constists of $\mathrm{t\bar{t}H}$ and semi-leptonic and full-leptonic $\mathrm{t\bar{t}}$ production, since all other simulated templates do not contain enough events after the tight selection. While the background templates are normalized according to their corresponding cross sections, the signal events are scaled to match the total amount of background events.

In a first step, that was performed within the CMS *tHq* group, an optimal set of variables is sought-after. The initial set of around 20 variables also re-uses information that was already input for the reconstruction trainings. Starting from this set, successively the least discriminating input variable is removed as long as the overall performance of the MVA stays approximately constant. The final list of variables is given in Table 6.5. Since the found variables add significant discrimination power to the training by construction, the ranking is only provided for the sake of completeness. It should be noted that the repeated use of $|\eta(\text{light jet})|$ reduces its impact in the classification training. However, given the separation strenght in both, reconstruction and classification, $|\eta(\text{light jet})|$ is counted among the most important variables in this analysis.

Figure 6.15 shows the shapes of all input variables split for signal and background processes. In all distributions a clear separation is found. The final MVA response is depicted in Figure 6.16, that also provides the comparisons between training and testing samples. Also here, no sign for overtraining is found.

Before using the weights of this MVA in the signal regions, first the agreement between data and simulation is checked in the $\mathrm{t\bar{t}}$ control region. In Figure 6.17 the resulting distributions are shown and good agreement is found.

To further validate the classification, the data/MC comparisons are provided for all input variables in all regions. Figures 6.18 - 6.20 show the distributions in the muon channel. The corresponding diagrams for the electron channel can be found in Figures B.7 - B.9. Overall solid agreement between simulation and data is found



(a) (b)

Figure 6.14.: Comparisons for $t\bar{t}$ reconstruction between MVA and χ^2 reconstruction techniques. On the left-hand side the ΔR rank of the chosen hypothesis and the masses of the reconstructed leptonically and hadronically decaying top quark candidates are shown in the 2T region. On the right-hand side the same diagrams are given in the 3T region. Because the χ^2 values depend on the difference between reconstructed and true mass, the mass distributions are narrower compared to the MVA reconstruction. However, with the MVA reconstruction more often the best possible jet assignment is found, leading to a better description of the kinematics of the event.

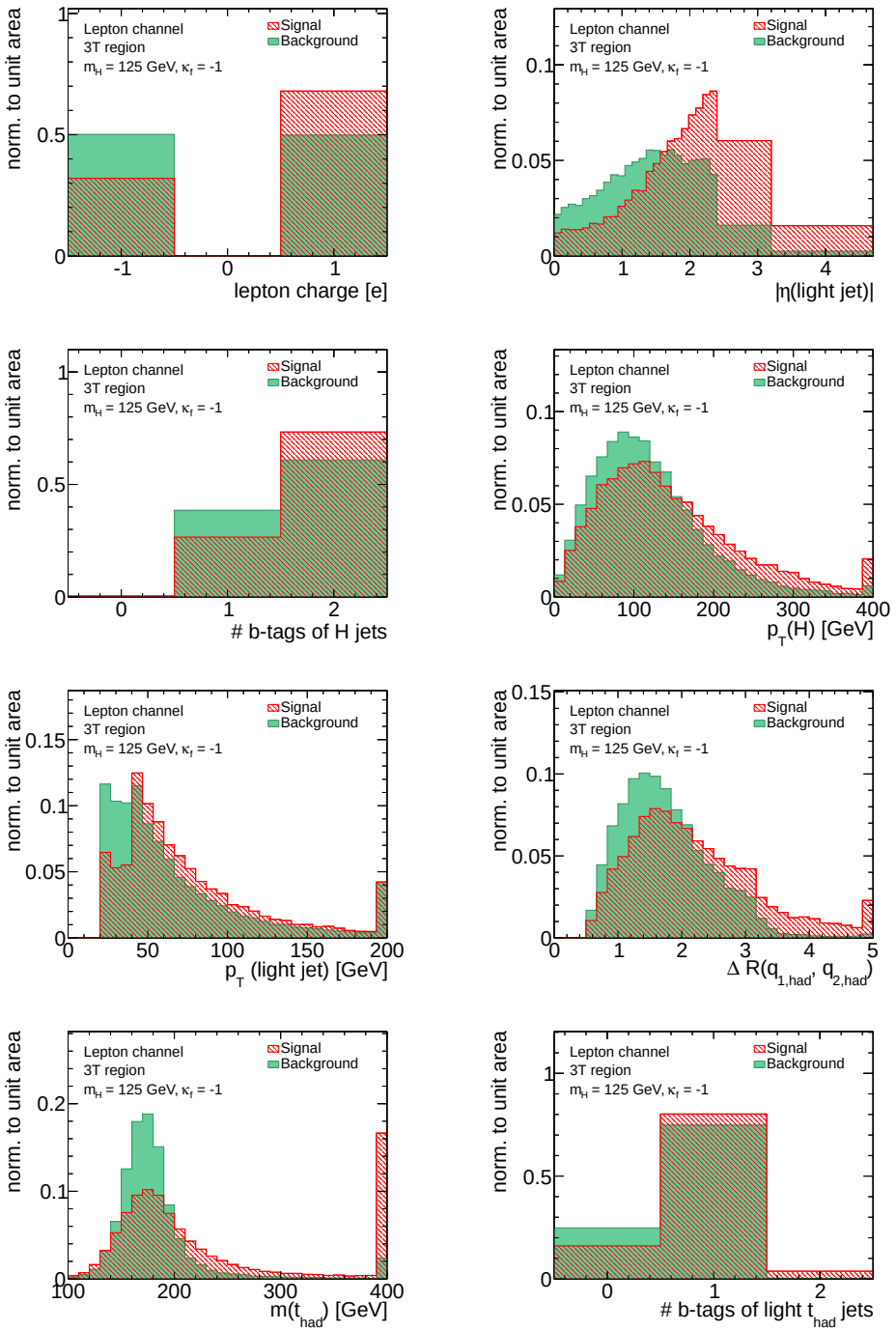


Figure 6.15.: Shapes of input variables for the final classification. The shapes are shown separately for signal and background processes and normalized to unit area.

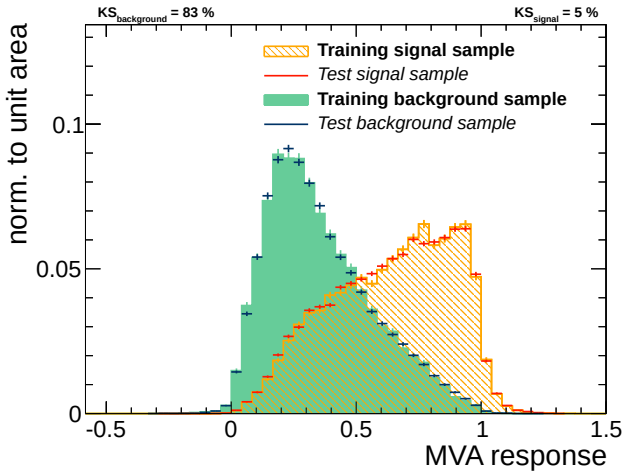


Figure 6.16.: Response of classification MVA for the signal and background processes. The shapes are shown separately for training and testing samples. A KS-test is performed and no hint for overtraining is found.

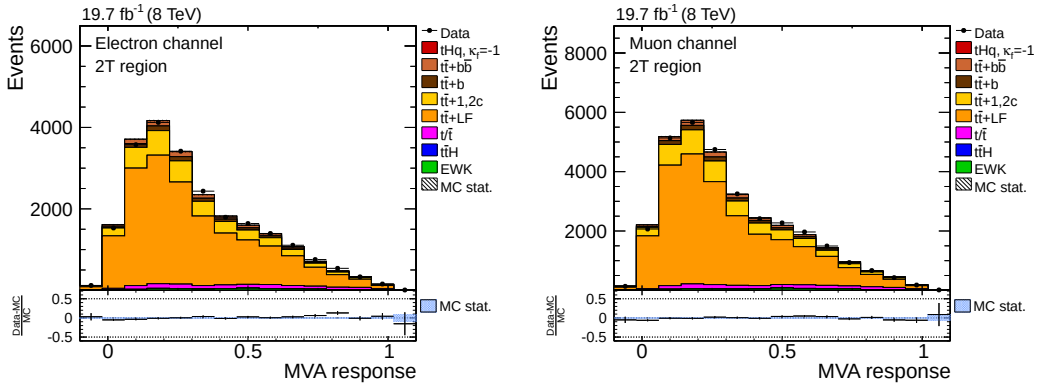


Figure 6.17.: Data/MC comparison of classification MVA response in the 2T region. The MC templates are normalized to data. The distributions agree well with each other.

Table 6.5.: Input variables of classification MVA. The variables are split into global variables, variables of the tHq reconstruction and variables of the $t\bar{t}$ reconstructions. As the set of variables is optimized, all of them add a significant amount of seperation power to the MVA, and the rank information is secondary. In the descriptions, t_{had} denotes the hadronically decaying top quark.

Variable	Rank	Description
lepton charge	1.	Electric charge of the lepton
# b-tags of H jets	3.	Number of b-tagged jets among the two jets from the Higgs boson decay
p_T(H)	4.	Transverse momentum of the Higgs boson
p_T(light jet)	7.	Transverse momentum of the light forward jet
$\|\eta$(light jet)$\|$	8.	Absolute value of the light forward jet's pseudorapidity
$m(t_{had})$	2.	Invariant mass of t_{had}
$\Delta R(q_{1,had}, q_{2,had})$	5.	ΔR between the two light-flavor jets from the t_{had} decay
# b-tags of light t_{had} jets	6.	Number of b-tagged jets among the two light-flavor jets from the t_{had} decay

giving confidence to apply the classification MVA to the signal regions.

6.8. Systematic uncertainties

In the following the evaluation of systematic effects is described. The procedure is similar to what is specified in Section 5.8 and analogous details are not repeated here. Again, the effects are grouped into theory, reconstruction and simulation uncertainties. For refining this analysis more systematic impacts are appraised as shape uncertainties.

6.8.1. Luminosity and theory uncertainties

The uncertainty on the integrated luminosity is assumed to be 2.6%. This value is based on a more recent measurement [201] with respect to the WH search and is applied to the rate of all simulated processes.

In this analysis there is no scale factor estimation for the background processes prior to the fit. The uncertainties for the theoretical cross sections used for normalizing the templates are divided into PDF and QCD scale, as shown in Table 6.6. These values are taken from [186] and introduced as rate uncertainties in the fit.

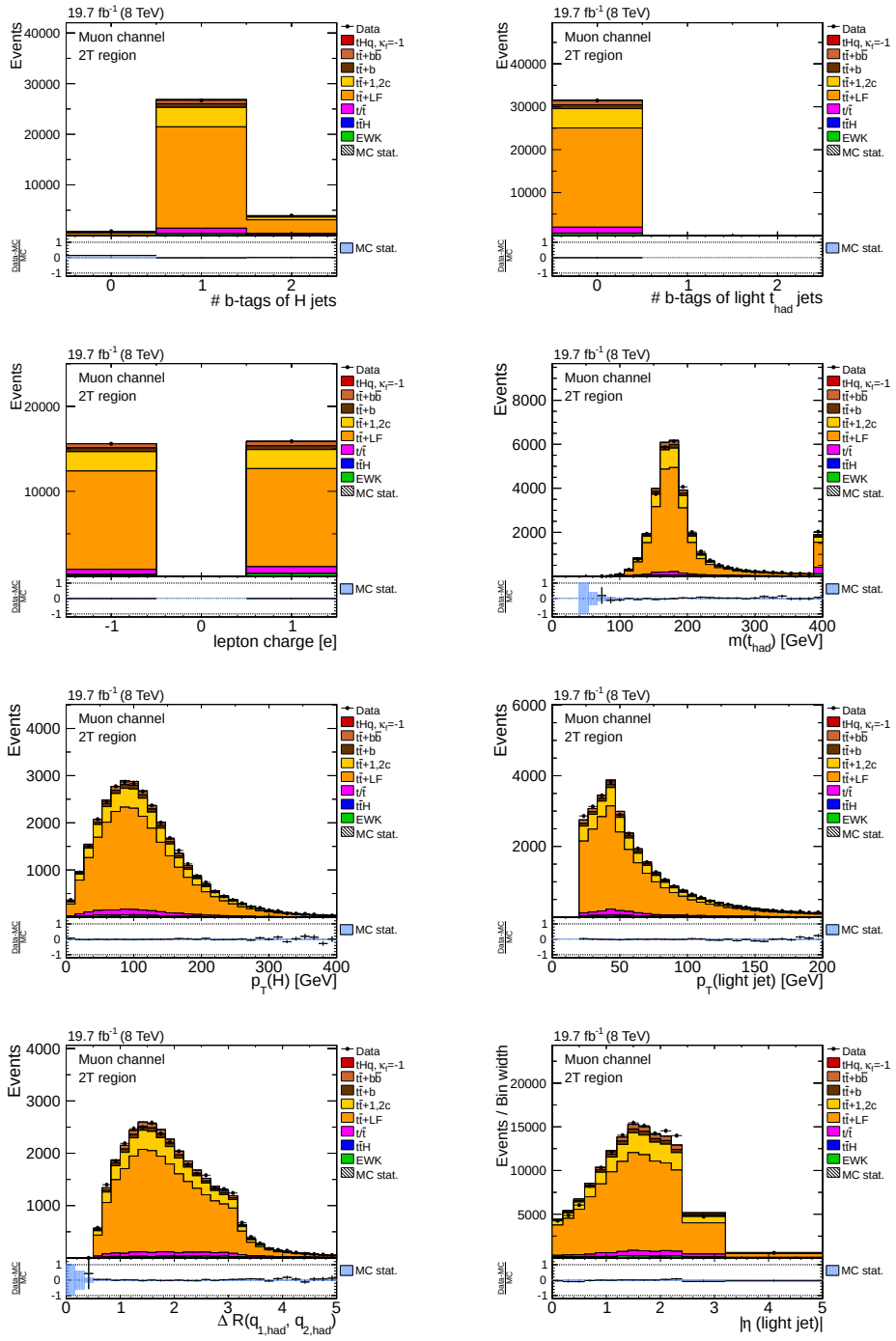


Figure 6.18.: MVA classification input variables for the muon channel in the 2T region. To facilitate shape comparisons the simulation is normalized to the number of events in data. In all variables good agreement between data and simulation is found. The corresponding distributions in the electron channel are provided in Figure B.7.

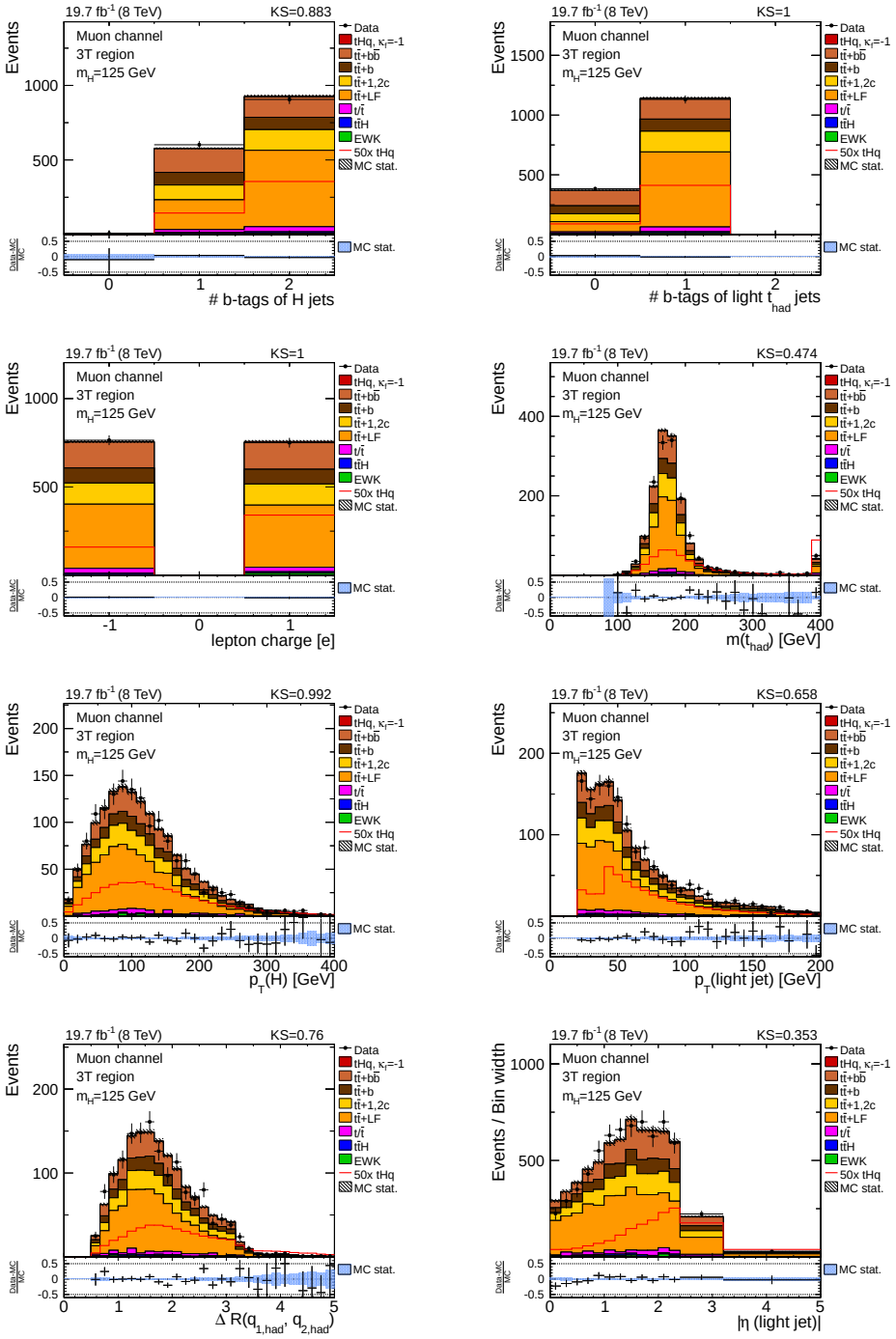


Figure 6.19.: MVA classification input variables for the muon channel in the 3T region. To facilitate shape comparisons the simulation is normalized to the number of events in data. In all variables good agreement between data and simulation is found. The corresponding distributions in the electron channel are provided in Figure B.8.

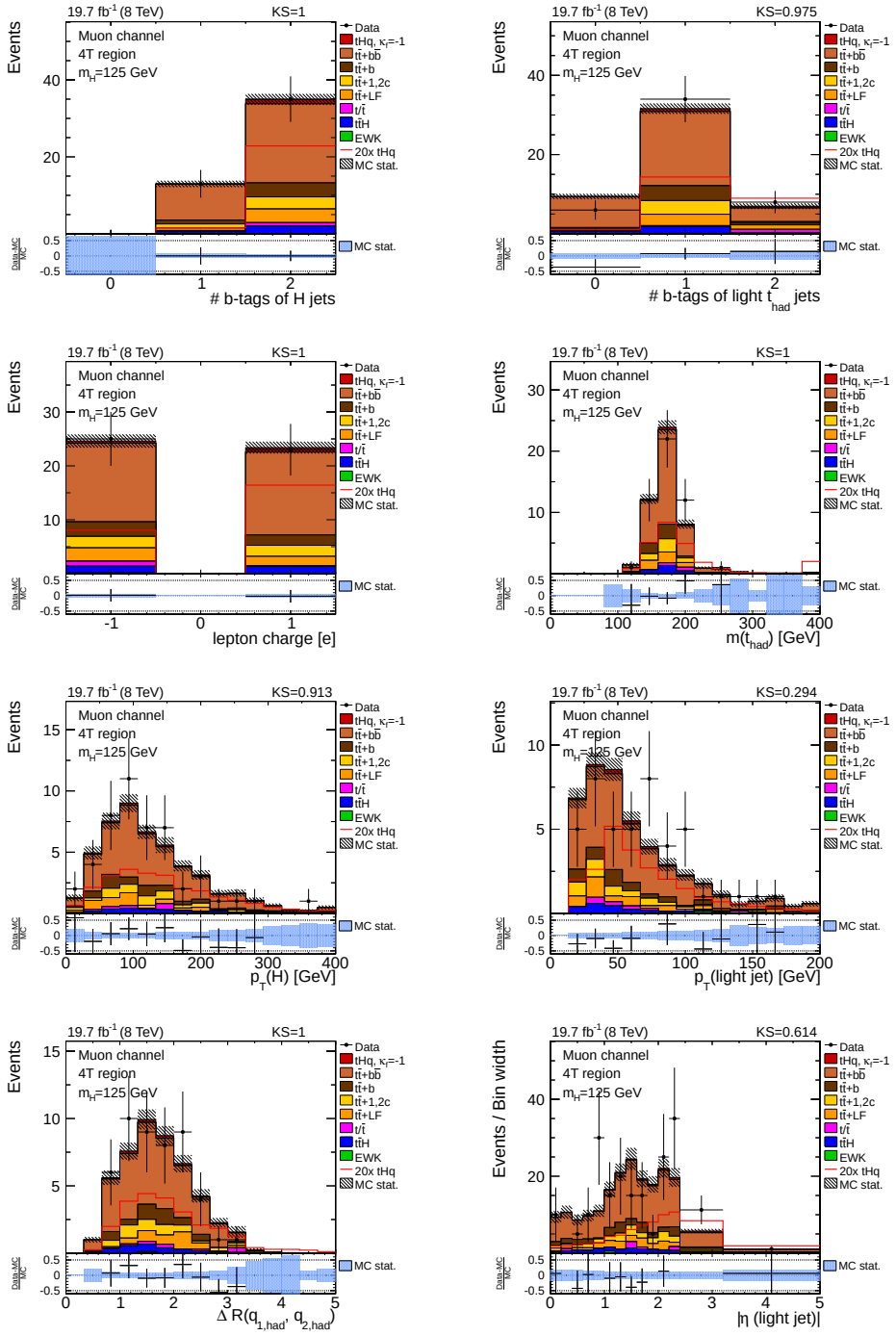


Figure 6.20.: MVA classification input variables for the muon channel in the 4T region. To facilitate shape comparisons the simulation is normalized to the number of events in data. In all variables good agreement between data and simulation is found. The corresponding distributions in the electron channel are provided in Figure B.9.

Table 6.6.: Cross section uncertainties divided into PDF and QCD scale used for the limit calculation. Each column in the table represents an independent source of uncertainty and entries in the same column are 100% correlated in the fit. Values in different columns are taken into account with 0% correlation. It should be noted that in $t\bar{t}$ and single top production the QCD scale variations have a dependency on the mass of the top quark. Therefore, the arising uncertainties are treated as correlated.

Process	PDF uncert. [%]			QCD scale uncert. [%]			
	gg	$q\bar{q}$	qg	$t/\bar{t}+t\bar{t}$	V	VV	$t\bar{t}H$
tHq			2.0				
$t\bar{t}H$	9.0						12.5
$t\bar{t}$	2.6			3.0			
Single top			4.6	2.0			
W+jets		4.8			1.3		
Z+jets		4.2			1.2		
Diboson						3.5	

6.8.2. Reconstruction uncertainties

A constant 2% uncertainty is assumed for the lepton efficiency. The value covers the effects found in [189,202], and is included for both, the electron and muon channels.

This analysis estimates the influence of the jet energy scale in an advanced way. In total 17 uncorrelated sources of uncertainties are identified and taken into account, as recommended by the CMS *JetMET* group [203]. For the uncertainty on JER an updated jet transverse momentum resolution measurement using dijet events [193] is considered in this analysis. The up and down templates are evaluated as described in Section 5.8.

Differences in the b-tagging efficiencies for simulation and data are accounted for by using scale factors as described in Section 6.4.1. The scale factors are varied within their uncertainties [204] firstly for b- and c-quark jets, and secondly for light and gluon induced jets.

To account for the uncertainty on the unclustered MET, the MVA responses are re-computed for the shifted templates and fed to the limit extraction as shape uncertainty.

6.8.3. Simulation uncertainties

For the standard pile-up reweighting procedure used in this analysis the number of pile-up interactions in data is evaluated via total inelastic cross section times measured bunch-by-bunch luminosity. To cover both effects, an uncertainty of 5% is recommended [205]. The altered reweighting according to the uncertainties is performed again and propagated through the analysis. This effect thus enters the limit calculation as shape uncertainty.

The MadGraph package is used to simulate tHq and $t\bar{t}$ events. As the LO

amplitudes for the processes depend on the chosen factorization and renormalization scale ($\mathcal{Q}^2$), two dedicated samples are generated, where $\mathcal{Q}^2$ is multiplied by a factor of 4.0 and 0.25, respectively. Using these shifted templates the MVA responses are re-evaluated and included as shape uncertainties.

For the generation of events with MadGraph another important parameter is the jet matching threshold that influences the transition between the hard process and the parton shower. To account for the choice of this matching threshold, two additional samples for $t\bar{t}$ production are generated using double or half of the nominal value. The extracted templates from these samples are different in rate and shape compared to the nominal ones, so the re-computed MVA response is used as shape uncertainty.

Another systematic uncertainty has to be assigned to the provisionally introduced top quark p_T reweighting procedure. Since the weighting functions have been calculated centrally for CMS analyses, a rather conservative approach is chosen. For the *up* templates, the event weights are applied twice, while for the *down* templates the top quark p_T weight is not used. Hence, the resulting samples are varied in rate and shape and therefore introduced as shape uncertainties.

The modeling of $t\bar{t}$ production in association with jets stemming from b quarks, is expected to be insufficient. To account for possible biases, the separation of simulated $t\bar{t}$ events according to their additional flavor content was introduced in the first place. In analogy to the $t\bar{t}H$ search [187], where this kind of splitting was introduced, a conservative 50% rate uncertainty is assigned to $t\bar{t}+b\bar{b}$, $t\bar{t}+b$ and $t\bar{t}+c\bar{c}$ templates.

The whole set of systematic uncertainties is imparted to the statistical evaluation, that is described in the following. To estimate the influence of each source on the final result, Table 6.7 shows the corresponding changes in the upper limits. The values are extracted by including either only one or all but one source of systematic uncertainty. In particular, the $\mathcal{Q}^2$ scale variations on $t\bar{t}$ production and the signal process are found to have the largest impact on the results.

6.9. Results

For the analysis presented in this chapter the statistical evalution is performed with the theta framework. The upper limits are extracted by a simultaneous fit on the shape of the classification MVA in the four signal regions. All systematic effects are taken into account. The final event yields including the modulations of the nuisance parameters are summarized in Table 6.8. Figure 6.21 shows all four distributions after the fit. The red hollow shapes peaking at the right side of each diagram illustrate the expectation if the cross section was 50 times (3T region) or 20 times (4T region) higher.

Table 6.7.: Impact of single systematic effects on the final results. The values show the relative difference on the expected limit, when either including only one exclusive source of systematic or removing one from the complete set. Excluded nuisance parameters are fixed to their post-fit values. If only a negligible influence in found, the corresponding value is labeled with < 0.1.

Source	Type	Impact as exclusive source [%]	Removal effect [%]
Cross section (PDF + scale)	rate	11.4	0.7
Luminosity	rate	10.4	0.3
Lepton efficiency	rate	5.2	0.7
$t\bar{t}$ HF rates	rate	15.2	1.4
b-tagging	shape	18.0	2.0
Pile up	shape	0.9	< 0.1
Top p_{T} reweighting	shape	19.9	2.7
Unclustered energy	shape	2.8	0.7
JER	shape	1.9	1.4
JES	shape	9.5	3.4
$\mathcal{Q}^2$ scale	shape	20.9	4.8
Matching threshold	shape	1.9	2.0
MC statistics	shape	2.3	1.7

Additional valuable information of this fit are the posterior nuisance parameters for the additional 50% uncertainties on the $\mathrm{t\bar{t}}$ + HF subprocesses. The results shown in Table B.4 range between 1.1 and 1.4 and indicate that the contribution of these heavy flavor components is underestimated in $\mathrm{t\bar{t}}$ + jets simulation with MadGraph. The same tendency is found in the CMS cross section ratio $\sigma_{\mathrm{t\bar{t}b\bar{b}}}/\sigma_{\mathrm{t\bar{t}jj}}$ measurement [206].

As no excess of signal-like events is visible in data, full $\mathrm{CL_s}$ exclusion limits at 95% C.L. are set and given in Table 6.9. The observed limit of 5.8 times the cross section predicted for the $\kappa_{\mathrm{f}} = -1$ scenario is fully covered by the 1σ uncertainties of the expected limit with a median value of 4.4. The individual results in the electron and muon channels are consistent.

Similar to the WH search, in Figure 6.22 all events from all fitted regions are put in one single distribution, by sorting them according to the expected signal-over-background ratio in their corresponding bins. The diagram shows that the data agrees well with the expected background, while the signal events are fully covered by the uncertainties on the background simulation.

The single data event in the rightmost bin in Figure 6.22 catches one's eye. This event from the 4T muon channel is the best candidate for tHq production in the full 8 TeV dataset. Therefore, Figure 6.23 shows its event display to give an impression how the best candidate for the signal process appears in the detector.

To date, the most stringent limit on tHq production can be extracted from the

Table 6.8.: Final yields in signal regions after the fit to data. The given uncertainties include all systematic and statistical effects.

	3T region		**4T region**	
Process	Electron channel	Muon channel	Electron channel	Muon channel
tHq	9.5 ± 1.5	13.6 ± 2.2	1.1 ± 0.2	1.5 ± 0.3
$t\bar{t} + b\bar{b}$	225.4 ± 8.3	323.1 ± 10.9	21.7 ± 2.0	30.8 ± 2.0
$t\bar{t} + b$	156.0 ± 6.7	205.1 ± 6.7	3.3 ± 0.4	3.5 ± 0.5
$t\bar{t} + 1, 2c$	179.5 ± 5.9	263.9 ± 9.1	4.2 ± 0.9	3.9 ± 0.6
$t\bar{t} + LF$	412.7 ± 8.3	632.3 ± 8.1	2.1 ±0.3	2.4 ± 0.4
$t/\bar{t}$	28.8 ± 3.1	43.5 ± 4.9	1.4 ± 0.3	0.8 ± 0.2
$t\bar{t}H$	10.2 ± 0.3	13.8 ± 0.6	1.6 ± 0.1	1.9 ± 0.1
EWK	4.5 ± 0.6	12.2 ± 1.8	1.0 ± 0.7	0.1 ± 0.0
Total MC	1026.5 ± 15.2	1507.5 ± 18.6	36.3 ± 2.4	44.9 ± 2.3
Data	1028	1514	32	48

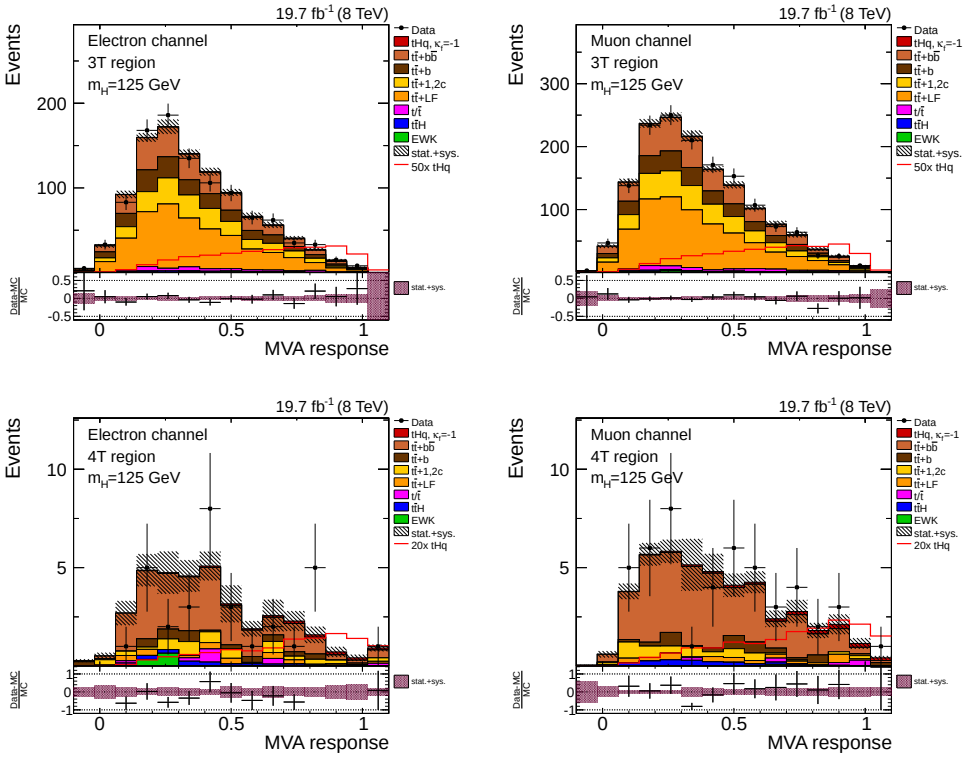


Figure 6.21.: Classification MVA output distributions after fitting to data. The MC templates are scaled according to their fit results and the displayed uncertainties include all systematic effects.

Table 6.9.: Upper limits on $\sigma/\sigma_{\kappa_f=-1}$ for tH($b\bar{b}$)q. All signal regions are combined in the fit. The observed limit coincides with the expected limit within 1σ precision.

	Limit on $\sigma/\sigma_{\kappa_f=-1}$	
	Expected	Observed
Electron channel	$6.8^{+3.3}_{-1.7}$	8.7
Muon channel	$5.6^{+2.0}_{-1.8}$	7.4
Combined	$\mathbf{4.4^{+1.8}_{-1.4}}$	**5.8**

analysis presented in this chapter, i.e. $\sigma_{tHq} < 1.36\,\text{pb}$. Similar to the analyses in the $H \to \gamma\gamma$ and $H \to WW$ decay channels (see Section 1.3) there is no hint for an excess in data, and therefore no sign for an anomalous Higgs boson coupling to the top quark. A combination of the $H \to b\bar{b}$, $H \to \gamma\gamma$ and $H \to WW$ analyses together with $H \to \tau\tau$ is in preparation and will increase the sensitivity. The upper limit of this combination is expected to lie between 2 and 3 times the predicted cross section with $\kappa_f = -1$.

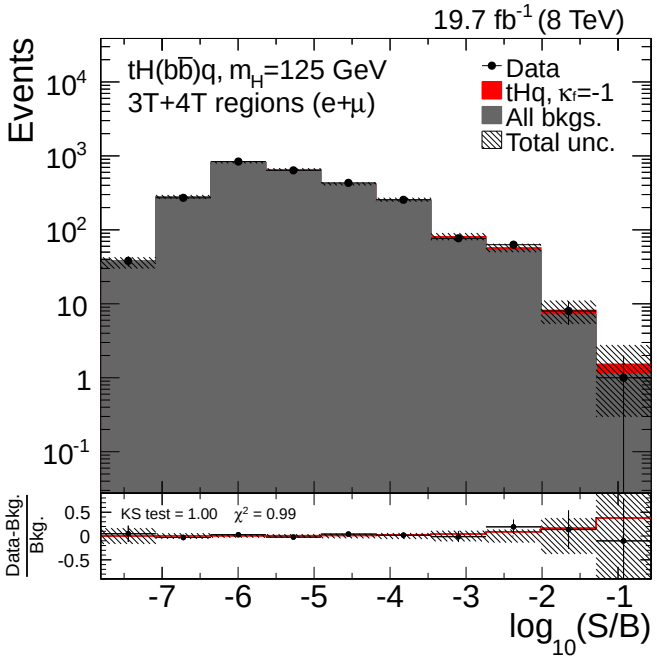


Figure 6.22.: Combination of all fitted distributions into one single diagram. Here, all events from all signal regions are sorted according to the expected signal-over-background ratios in their corresponding bins. Additionally, the ratio between data and background-only is given. The given uncertainties include all statistical and systematic effects. No excess in data with respect to the expected background is visible.

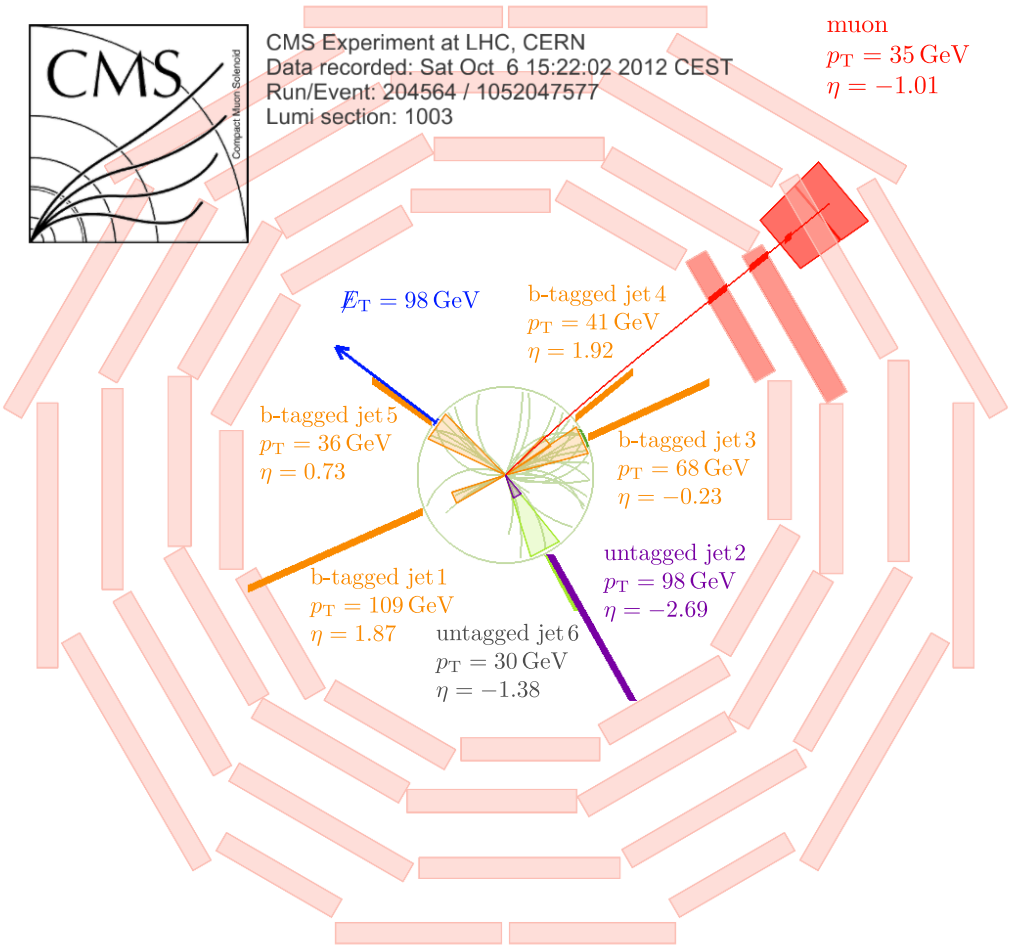


Figure 6.23.: Transverse detector view of the most signal-like event in data. The scheme shows the reconstructed objects comprising all corrections. There is exactly one isolated muon (red) and four b-tagged jets (orange bars) in the event. Furthermore, the characteristic forward jet (purple) and a large amount of missing energy is found. In this event jet 1 and jet 5 are assigned to the Higgs boson candidate. Jet 2 is the light quark candidate and jet 3 is assigned to the top quark. The reconstructed masses of the Higgs boson and top quark candidates are $m_H = 100.3\,GeV$ and $m_t = 173.3\,GeV$, respectively.

Conclusion and outlook

The last missing piece of the standard model has finally been found at the Large Hadron Collider at CERN. The discovery of the Higgs boson is the jewel in the crown of the many years of hard work for particle physicists all over the world. Many measurements from the ATLAS and CMS collaborations closed down the allowed ranges for the properties of this boson. So far there are no hints for any deviations from the predictions of the standard model. These findings are also confirmed by the latest analyses at the Tevatron collider [207].

The Higgs boson with a mass of $m_{\mathrm{H}} = 125\,\mathrm{GeV}$ is predicted to decay into bottom quark pairs in 60% of all cases. However, this decay channel has not yet been observed. This thesis investigated $\mathrm{H} \to \mathrm{b\bar{b}}$ in two different production modes. First, the Higgs boson production in association with a W boson (WH) has been studied. Second, the first analysis searching for Higgs boson production in associated with single top quarks (tHq) with $\mathrm{H} \to \mathrm{b\bar{b}}$ decays was executed. Both analyses have been performed using the full dataset corresponding to an integrated luminosity of more than $19\,\mathrm{fb}^{-1}$. This data was recorded with the CMS detector in 2012 with proton-proton collisions at a center-of-mass energy of $\sqrt{\mathrm{s}} = 8\,\mathrm{TeV}$.

The search for the Higgs boson in the WH was carried out for different mass hypotheses in the mass region from 110 to 150 GeV. The main feature of the event selection is the kinematic boost requirement on the reconstructed W boson and the Higgs boson candidates. The goal of this study was to improve the search sensitivity with the use of jet substructure information. The jet substructure information is extracted from the subjet/filter jet algorithm proposed by theorists. Filter jets have a smaller size compared to the standard jets used within the CMS collaboration, and are expected to yield a better mass resolution of the reconstructed Higgs boson. For the first time a regression technique correcting the filter jet energies was introduced and validated in this thesis. The regression accounts for undetected neutrinos in b-quark decays and for missing dedicated filter jet energy corrections. This way an improvement of the reconstructed Higgs boson mass resolution of 15% was found. For each tested mass point and for each region Boosted Decision Trees (BDT) were trained to discriminate signal events from all background processes. In addition, three expert BDTs were developed to separate the signal process from $\mathrm{t\bar{t}}$, W+0b and diboson production individually. A final discriminant was optimized using all four trainings for a categorization of the events. The shape of this discriminant was used to fit the simulation to data in order to extract upper limits on the process $\mathrm{W}(\ell\nu)\mathrm{H}(\mathrm{b\bar{b}})$. The training, validation and optimization steps were performed twice: Firstly, with the same set of variables used in the published CMS

analysis [153] (DJ analysis). Secondly, with an optimized set of information employing the jet substructure included in the classification (SJF analysis). By comparing the former with the published CMS analysis consistent results were found. Based on these findings the improvements in terms of expected limits of using substructure techniques were quantified to be ∼5% by comparing the DJ and SJF analyses. In addition, the full statistical evaluation for the SJF analysis was presented. The expected limits over the tested mass range were found in the interval of (1.2, 8.8). The observed limits lie systematically higher than the expected limits. The behavior is expected for a standard model Higgs boson. By sorting all events according to the signal-over-background ratio in the corresponding bin it was illustrated that the data is described better by signal and background simulation compared to background-only simulation. For a Higgs boson with a mass of $m_{\mathrm{H}} = 125\,\mathrm{GeV}$ the observed significance of the excess in data was evaluated to be 1.16σ.

Another analysis investigating $\mathrm{H} \to \mathrm{b\bar{b}}$ decays was performed to search for the associated Higgs boson production with single top quarks. This production mode is sensitive to the relative sign of the Higgs boson couplings to fermions and vector bosons. As the cross section at $\sqrt{\mathrm{s}} = 8\,\mathrm{TeV}$ is far too small for current analyses to be sensitive to such a rare process, the analysis was optimized for an anomalous coupling of $\kappa_{\mathrm{f}} = -1$. With this coupling the cross section is assumed to be 13 times enhanced compared to the SM prediction. Still, a huge effort is needed to extract even the increased amount of signal events. In the assigned signal-enriched part of the phase space with either three or four b-tagged jets, there are ∼15 signal events expected opposed to ∼2000 background events. The dominant background contribution is $\mathrm{t\bar{t}}$ production. To reconstruct the events and to discriminate the signal from the background process three neural networks were trained in total. As the jet-quark assignment is ambiguous in multi-jet final states, a first neural network was used to assign the jets to the expected final state particles: three bottom quarks from the Higgs boson and the top quark, and the light forward quark that is characteristic for the single top t-channel production. In a similar way each event was reconstructed under the assumption it stemmed from semi-leptonically decaying $\mathrm{t\bar{t}}$ production. Thus, a second neural network was trained to assign the jets to the two bottom quarks from the top quark decays, and to the two light quarks from one hadronically decaying W boson. With all the information from the two reconstruction techniques and a global observable an optimal set of variables was deduced to separate signal from background events and a third neural network was trained with these inputs. To extract upper limits on the production mode $\mathrm{pp} \to \mathrm{tHq}$ in the four signal regions simulation was fitted to data in the distributions of the discriminant. No excess in data with respect to the background simulation was found and full $\mathrm{CL_s}$ upper limits at 95% C.L. were computed. The observed limit was found to be 5.8 times the cross section prediction assuming $\kappa_{\mathrm{f}} = -1$. A general upper limit on tHq production of 1.36 pb was extracted. This is the most stringent upper limit on tHq production from an individual decay channel to date.

With the start of the LHC Run II an entirely new territory in particle physics will be entered. The beam energies will rise up to 7 TeV leading to proton-proton collisions with unprecedented center-of-mass energies. The higher energy will improve the signal-over-background ratio in many searches for rare processes. But the new conditions will also introduce new challenges for the analysts. In 2015 proton-proton collision data corresponding to an integrated luminosity between $10\,\text{fb}^{-1}$ and $20\,\text{fb}^{-1}$ with $\sqrt{s} = 13\,\text{TeV}$ is projected.

With the higher center-of-mass energy the cross section of WH production is increased by a factor of 2, while $\text{t}\bar{\text{t}}$ production rises by a factor of more than 3. Hence, the main challenge for the WH analysis will be to tackle the $\text{t}\bar{\text{t}}$ background even more. However, conservative projections [208] indicate that already $10\,\text{fb}^{-1}$ in combination with the results at 7 TeV and 8 TeV are sufficient to claim evidence for $\text{H} \to \text{b}\bar{\text{b}}$ in the combined $\text{WH}(\text{b}\bar{\text{b}})$ and $\text{ZH}(\text{b}\bar{\text{b}})$ channels. Further analysis improvements are needed to finally observe this channel with a significance of 5σ. Employing the jet substructure information as shown in this thesis can help to reach this goal.

The cross section of tHq production will increase by a factor of $\sim$4 and therefore the signal-over-background ratio will improve. Figure 6.24 shows the extrapolation of the results in Chapter 6 for $\sqrt{s} = 13\,\text{TeV}$ assuming an integrated luminosity between 5 to $40\,\text{fb}^{-1}$. Already with $5\,\text{fb}^{-1}$ the analysis should reach the same sensitivity as the presented one with 8 TeV and $20\,\text{fb}^{-1}$. In the figure also the projections with 10% to 30% analysis improvements are shown. These improvements are in reach, for example by exploiting the full shape of the CSV b-tagging discriminant [209]. In combination with other decay channels the analysis could be sensitive enough to exclude the $\kappa_\text{f} = -1$ scenario at 95% confidence level with the data recorded in 2015.

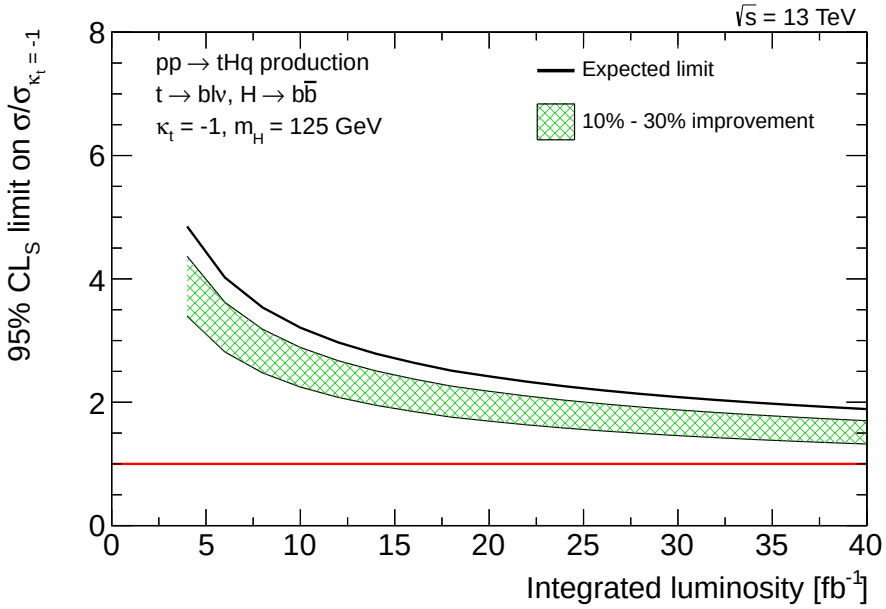


Figure 6.24.: Assumed sensitivity on tHq production with $\kappa_f = -1$ at 13 TeV. The results from Chapter 6 are extrapolated by scaling all processes according to the predicted cross sections in [210, 211]. The systematic uncertainties are simply scaled according to the higher luminosity for this study. In addition, the effect of possible 10% to 30% analysis improvement in terms of expected limits is shown.

A. Supplementary material for WH analysis

This section provides additional tables and figures for the search for $W(\ell\nu)H(b\bar{b})$ final states. These mostly technical details complement the analysis described in Chapter 5.

A.1. Technical details on data and MC samples used in the analysis

Table A.1 summarizes the MC samples for signal and background processes. The analyzed data samples are given in Table A.2 and the applied triggers are provided in Table A.3.

A.2. Additional information on BDT analysis

Auxiliary distributions for both, the DJ analysis and the SJF analysis, are given in Figures A.1 - A.4. Tables A.4 - 5.15 provide extensive information on the calculations in the scope of the analysis.

Table A.1.: List of Monte Carlo samples used for the WH search for signal and background processes. For each template the effective cross section and the total number of generated events are given. For the QCD samples additionally the filter efficiency is listed.

Process		# events	Eff. cross section [pb]	MC filter
WH	$W \to \ell\nu$, $m_H = 110\,\text{GeV}$	991083	0.2559	1
WH	$W \to \ell\nu$, $m_H = 115\,\text{GeV}$	999998	0.2090	1
WH	$W \to \ell\nu$, $m_H = 120\,\text{GeV}$	1000534	0.1672	1
WH	$W \to \ell\nu$, $m_H = 125\,\text{GeV}$	999998	0.1302	1
WH	$W \to \ell\nu$, $m_H = 130\,\text{GeV}$	999998	0.0974	1
WH	$W \to \ell\nu$, $m_H = 135\,\text{GeV}$	1000430	0.0699	1
WH	$W \to \ell\nu$, $m_H = 140\,\text{GeV}$	996353	0.0481	1
WH	$W \to \ell\nu$, $m_H = 145\,\text{GeV}$	997450	0.0313	1
WH	$W \to \ell\nu$, $m_H = 150\,\text{GeV}$	993963	0.0187	1
ZH	$Z \to \ell\ell$, $m_H = 110\,\text{GeV}$	998512	0.0441	1
ZH	$Z \to \ell\ell$, $m_H = 115\,\text{GeV}$	996597	0.0364	1
ZH	$Z \to \ell\ell$, $m_H = 120\,\text{GeV}$	990213	0.0293	1
ZH	$Z \to \ell\ell$, $m_H = 125\,\text{GeV}$	969460	0.0230	1
ZH	$Z \to \ell\ell$, $m_H = 130\,\text{GeV}$	999998	0.0173	1
ZH	$Z \to \ell\ell$, $m_H = 135\,\text{GeV}$	686398	0.0125	1
ZH	$Z \to \ell\ell$, $m_H = 140\,\text{GeV}$	982862	0.0087	1
ZH	$Z \to \ell\ell$, $m_H = 145\,\text{GeV}$	999998	0.0057	1
ZH	$Z \to \ell\ell$, $m_H = 150\,\text{GeV}$	999998	0.0034	1
ZH	$Z \to \nu\nu$, $m_H = 110\,\text{GeV}$	1000319	0.0874	1
ZH	$Z \to \nu\nu$, $m_H = 115\,\text{GeV}$	999998	0.0720	1
ZH	$Z \to \nu\nu$, $m_H = 120\,\text{GeV}$	1000382	0.0581	1
ZH	$Z \to \nu\nu$, $m_H = 125\,\text{GeV}$	999999	0.0455	1
ZH	$Z \to \nu\nu$, $m_H = 130\,\text{GeV}$	999039	0.0342	1
ZH	$Z \to \nu\nu$, $m_H = 135\,\text{GeV}$	998170	0.0248	1
ZH	$Z \to \nu\nu$, $m_H = 140\,\text{GeV}$	999411	0.0172	1
ZH	$Z \to \nu\nu$, $m_H = 145\,\text{GeV}$	999998	0.0112	1
ZH	$Z \to \nu\nu$, $m_H = 150\,\text{GeV}$	999998	0.0068	1
$t\bar{t}$ + jets	$t\bar{t} \to b\ell\nu b\ell\nu$	11684000	24.6	1
$t\bar{t}$ + jets	$t\bar{t} \to b\ell\nu bq\bar{q}$	25364796	103	1
$t\bar{t}$ + jets	$t\bar{t} \to bq\bar{q}bq\bar{q}$	24754516	106	1
t (tW-channel)	inclusive	497657	11.1	1
$\bar{t}$ (tW-channel)	inclusive	493459	11.1	1
t (t-channel)	inclusive	3158226	30.7	1
$\bar{t}$ (t-channel)	inclusive	1935071	56.4	1
t (s-channel)	inclusive	259960	1.76	1
$\bar{t}$ (s-channel)	inclusive	139973	3.79	1
W + jets	$W \to \ell\nu$, $70\,\text{GeV} < p_T(W) < 100\,\text{GeV}$	21967532	557.57	1
W + jets	$W \to \ell\nu$, $p_T(W) > 100\,\text{GeV}$	61654698	289.25	1
W + jets	$W \to \ell\nu$, $p_T(W) > 180\,\text{GeV}$	9694453	34.32	1
Z/γ^* + jets	$Z/\gamma^* \to \ell^+\ell^-$, $70\,\text{GeV} < p_T(Z) < 100\,\text{GeV}$	11734531	62.13	1
Z/γ^* + jets	$Z/\gamma^* \to \ell^+\ell^-$, $p_T(Z) > 100\,\text{GeV}$	12511319	40.50	1
WW	inclusive	10000420	56.7532	1
WZ	$WZ \to \ell\nu qq$	2848655	3.1	1
WZ	$WZ \to qq\ell\ell$	3215988	1.755	1
ZZ	inclusive	9799897	8.297	1
QCD μ-enriched	$50\,\text{GeV} < \hat{p}_T < 80\,\text{GeV}$, $p_T^\mu > 5\,\text{GeV}$	10365224	8082000.0	0.0218
QCD μ-enriched	$80\,\text{GeV} < \hat{p}_T < 120\,\text{GeV}$, $p_T^\mu > 5\,\text{GeV}$	9238636	1024000.0	0.0395
QCD μ-enriched	$120\,\text{GeV} < \hat{p}_T < 170\,\text{GeV}$, $p_T^\mu > 5\,\text{GeV}$	8291930	157800.0	0.0473
QCD μ-enriched	$170\,\text{GeV} < \hat{p}_T < 300\,\text{GeV}$, $p_T^\mu > 5\,\text{GeV}$	5839943	34020.0	0.0676
QCD μ-enriched	$300\,\text{GeV} < \hat{p}_T < 470\,\text{GeV}$, $p_T^\mu > 5\,\text{GeV}$	6482257	1757.0	0.0864
QCD μ-enriched	$470\,\text{GeV} < \hat{p}_T < 600\,\text{GeV}$, $p_T^\mu > 5\,\text{GeV}$	3513067	115.2	0.1024
QCD μ-enriched	$600\,\text{GeV} < \hat{p}_T < 800\,\text{GeV}$, $p_T^\mu > 5\,\text{GeV}$	3638997	27.01	0.0996
QCD μ-enriched	$800\,\text{GeV} < \hat{p}_T < 1000\,\text{GeV}$, $p_T^\mu > 5\,\text{GeV}$	4077850	3.57	0.1033
QCD μ-enriched	$\hat{p}_T > 1000\,\text{GeV}$, $p_T^\mu > 5\,\text{GeV}$	3247556	0.774	0.1097
QCD BCtoE	$80\,\text{GeV} < \hat{p}_T < 170\,\text{GeV}$	1945523	1191000.0	0.0109
QCD BCtoE	$170\,\text{GeV} < \hat{p}_T < 250\,\text{GeV}$	1948110	30980.0	0.0204
QCD BCtoE	$250\,\text{GeV} < \hat{p}_T < 350\,\text{GeV}$	1574884	4250.0	0.0243
QCD BCtoE	$\hat{p}_T > 350\,\text{GeV}$	1828530	811.0	0.0295

Table A.2.: List of 2012 data samples used for this analysis. In both channels, W(eν)H and W(μν)H data corresponding to an integrated luminosity of $\sim 19\,\text{fb}^{-1}$ is included in the search.

Mode	Dataset	$\mathcal{L}$ (fb^{-1})
W(μν)H	/SingleMu/Run2012A-13Jul2012-v1	0.809
	/SingleMu/Run2012A-recover-06Aug2012-v1	0.082
	/SingleMu/Run2012B-13Jul2012-v1	4.403
	/SingleMu/Run2012C-24Aug2012-v1	0.405
	/SingleMu/Run2012C-2012C-EcalRecover_11Dec2012-v1	0.090
	/SingleMu/Run2012C-PromptReco-v2	6.445
	/SingleMu/Run2012D-PromptReco-v1	6.803
	Total Lumi	19.04
W(eν)H	/SingleElectron/Run2012A-13Jul2012-v1	0.809
	/SingleElectron/Run2012A-recover-06Aug2012-v1	0.082
	/SingleElectron/Run2012B-13Jul2012-v1	4.403
	/SingleElectron/Run2012C-24Aug2012-v1	0.405
	/SingleElectron/Run2012C-2012C-EcalRecover_11Dec2012-v1	0.405
	/SingleElectron/Run2012C-PromptReco-v2	6.445
	/SingleElectron/Run2012D-PromptReco-v1	6.803
	Total Lumi	19.04

Table A.3.: List of L1 and HLT triggers used in the analysis. For the W(μν)H channel the additional η requirements on the muon was introduced in the later runs to cope with the increasing instantaneous luminosity.

Mode	L1 Seed	HLT Trigger
W(μν)H	SingleMu16(er)	IsoMu24(_eta2p1)
	SingleMu16(er)	Mu40(_eta2p1)
	SingleMu16(er)	IsoMu20(_eta2p1)_WCandPt80
W(eν)H	SingleEG20 OR 22	Ele27_WP80

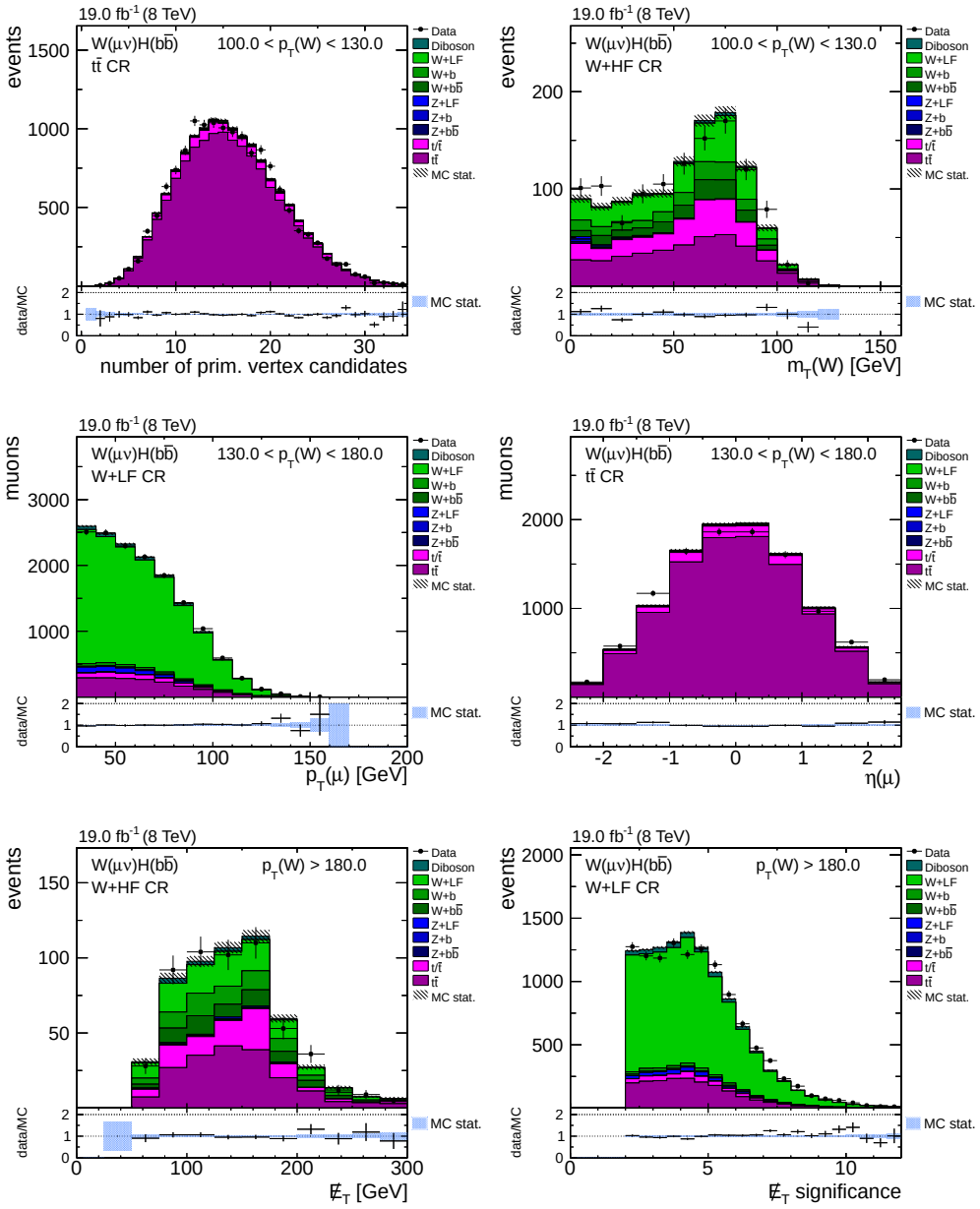


Figure A.1.: Different event distributions in control regions for the W(μν)H channel. Data/MC comparisons are shown for the low p_{T} region (top row), intermediate p_{T} region (middle row) and the high p_{T} region (bottom row) in all three control regions as indicated in the figures. The number of events in simulation is normalized to data.

Table A.4.: Data/MC scale factors for each analysis region with the statistical and systematic uncertainties.

Process	Low p_{T}	Intermediate p_{T}	High p_{T}
W+0b	$0.93 \pm 0.01 \pm 0.04$	$0.86 \pm 0.01 \pm 0.04$	$0.94 \pm 0.01 \pm 0.04$
W+1b	$1.78 \pm 0.25 \pm 0.15$	$2.21 \pm 0.24 \pm 0.20$	$2.71 \pm 0.36 \pm 0.20$
W+2b	$1.62 \pm 0.24 \pm 0.24$	$1.28 \pm 0.21 \pm 0.24$	$0.81 \pm 0.26 \pm 0.23$
$\mathrm{t\bar{t}}$	$1.02 \pm 0.01 \pm 0.04$	$1.00 \pm 0.01 \pm 0.04$	$0.95 \pm 0.01 \pm 0.03$

Table A.5.: Correlation matrix of scale factors determination. The matrices are shown for the low p_{T}, intermediate p_{T} and high p_{T} regions, where the fit is performed simultaneously for electron and muon channels.

	Low p_{T}			
	W+0b	W+1b	W+2b	$\mathrm{t\bar{t}}$
W+0b	1	-	-	-
W+1b	−0.43	1	-	-
W+2b	−0.02	−0.35	1	-
$\mathrm{t\bar{t}}$	−0.04	−0.06	−0.15	1
	Intermediate p_{T}			
	W+0b	W+1b	W+2b	$\mathrm{t\bar{t}}$
W+0b	1	-	-	-
W+1b	−0.50	1	-	-
W+2b	0.01	−0.35	1	-
$\mathrm{t\bar{t}}$	−0.05	−0.09	−0.14	1
	High p_{T}			
	W+0b	W+1b	W+2b	$\mathrm{t\bar{t}}$
W+0b	1	-	-	-
W+1b	−0.54	1	-	-
W+2b	0.01	−0.33	1	-
$\mathrm{t\bar{t}}$	−0.07	−0.11	−0.08	1

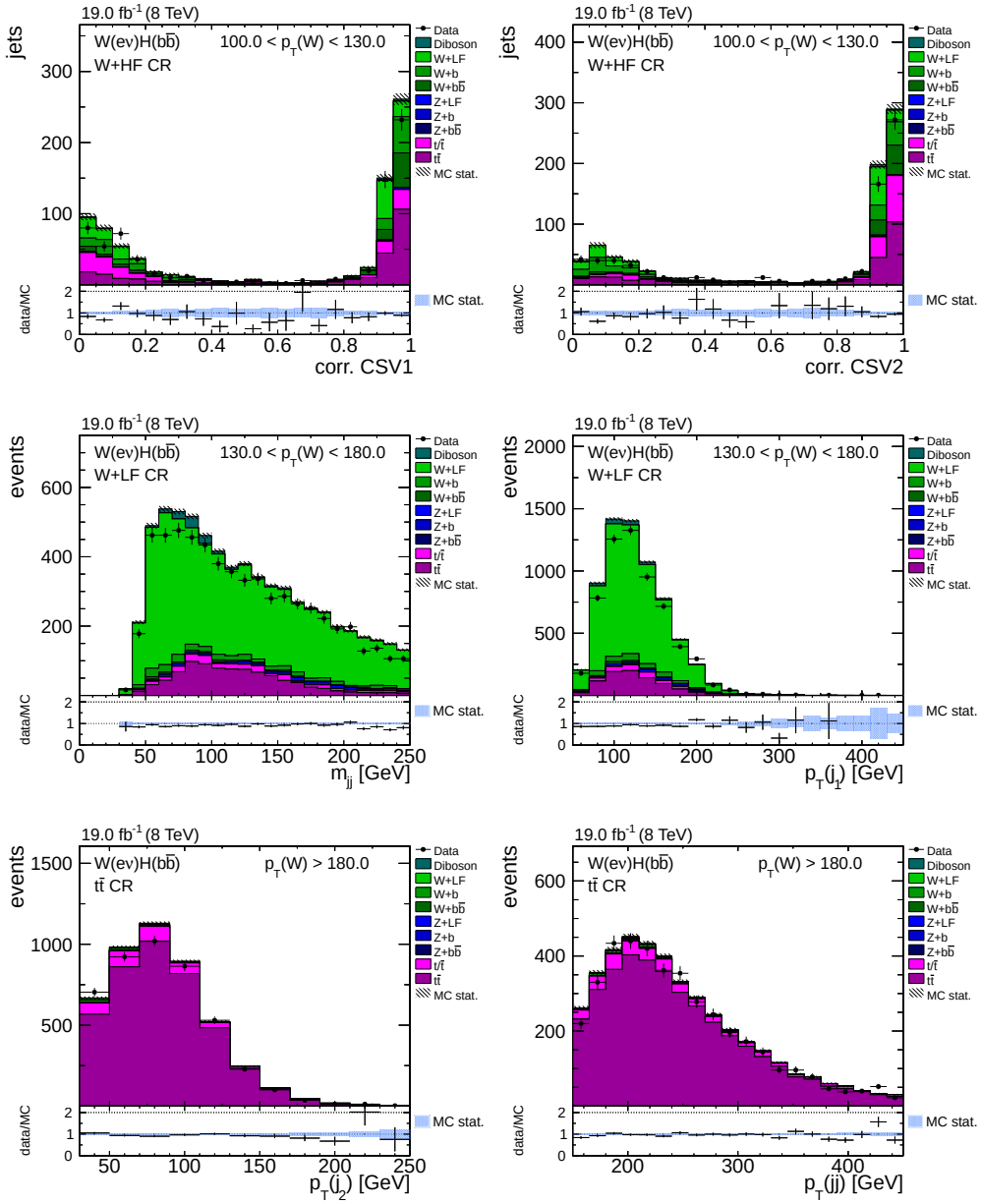


Figure A.2.: DJ analysis input variables for the W(eν)H channel in different control regions as indicated within the diagrams. The simulation is scaled to luminosity and scale factors are applied. In all variables solid data/MC agreement is found.

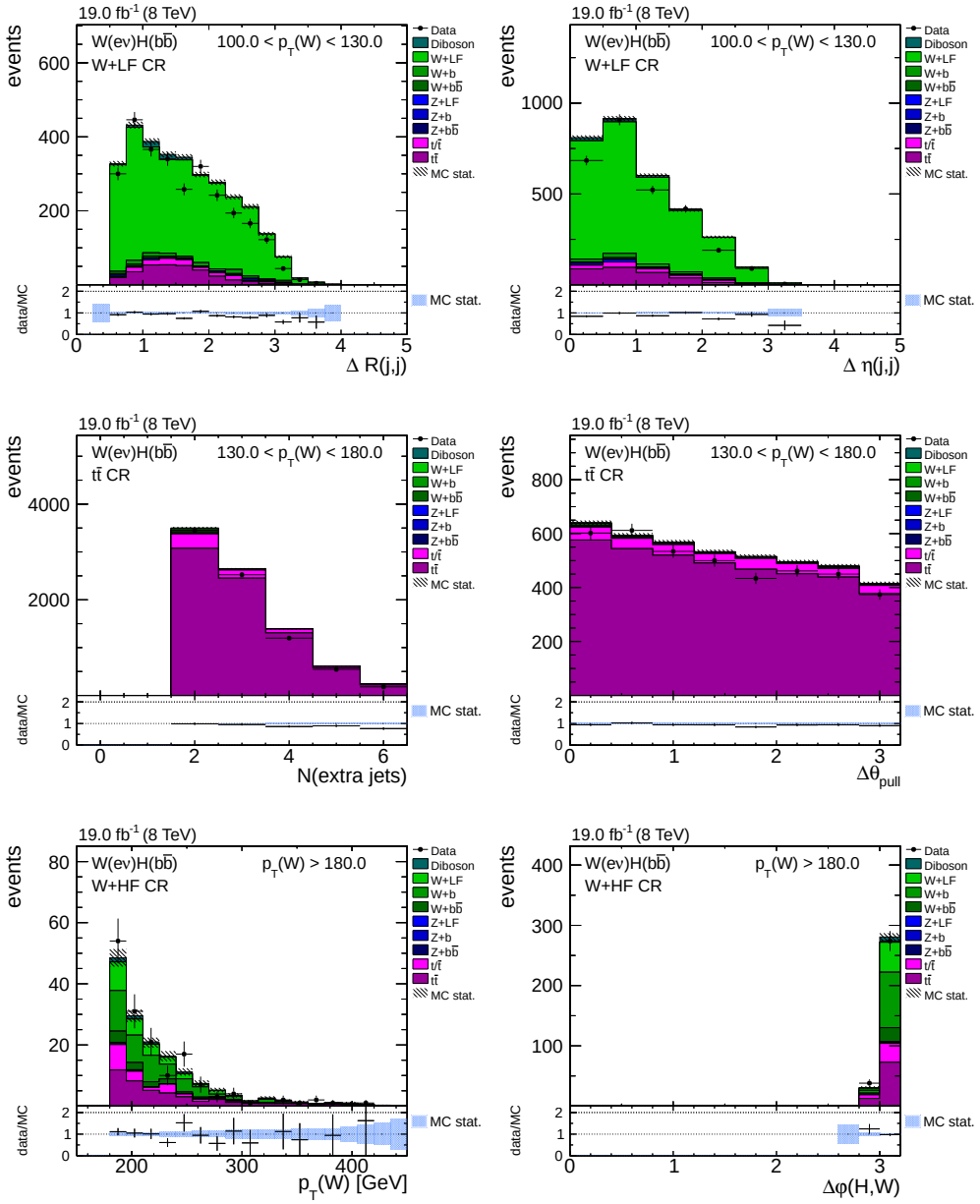


Figure A.3.: DJ analysis input variables for the W(eν)H channel in different control regions as indicated within the diagrams (cont.). The simulation is scaled to luminosity and scale factors are applied. In all variables decent data/MC agreement is found.

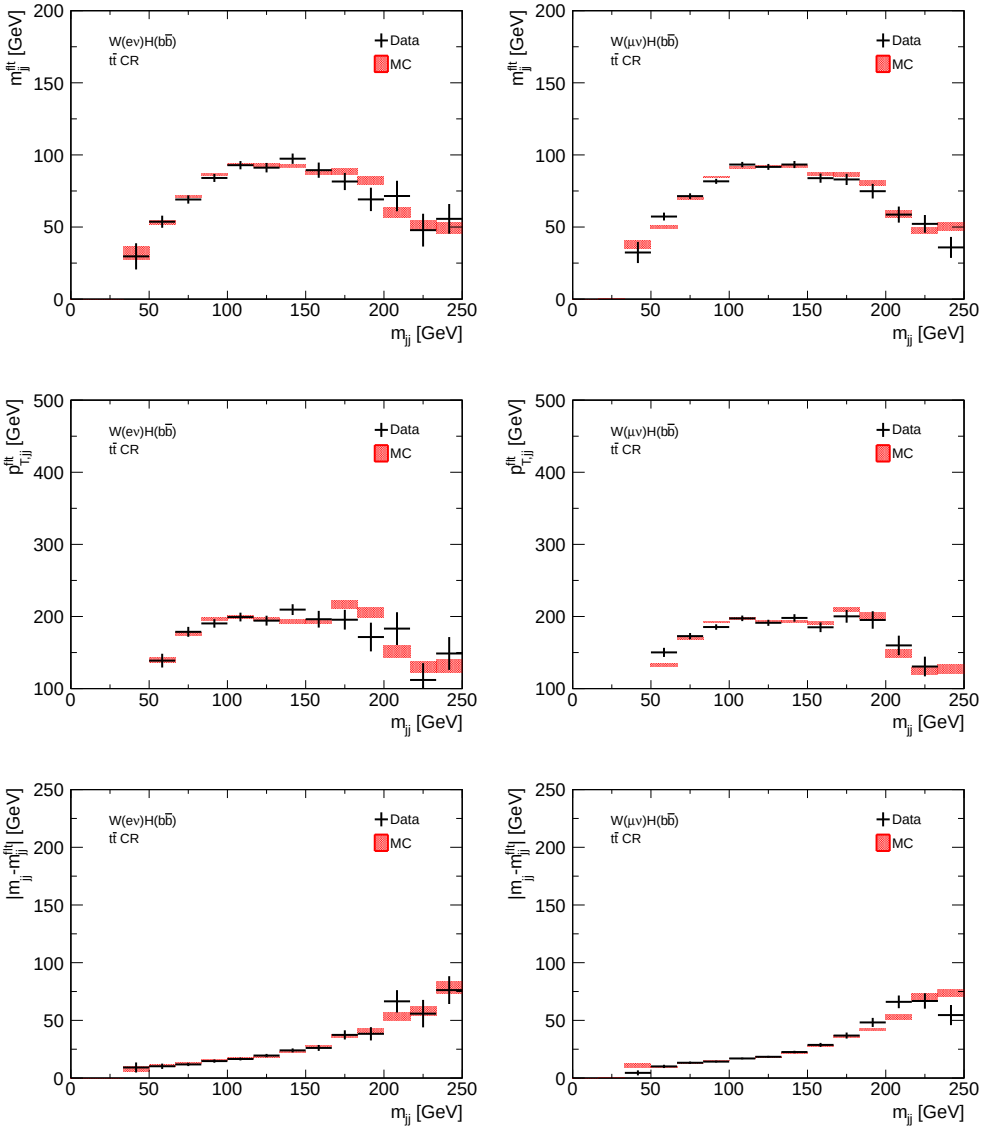


Figure A.4.: Correlation between BDT input variables in the $t\bar{t}$ CR for W(eν)H (left column) and W($\mu\nu$)H channels (right column). The diagrams show the projection along the y-axis for data and MC. The dijet mass on the x-axis is chosen, since it is one of the most discriminating variable. A good agreement is found in all three added variables.

Table A.6.: Variable rankings in the DJ analysis for nominal and expert BDTs. The number of occurrences in the top 5 positions of the individual rankings from the 27 trainings (9 mass points and 3 analysis bins) are counted.

Ranking	Variable	# Top 5 occurrences
1.	N_{aj}	27
2.	m_{jj}	23
3.	CSV_2	17
4.	$p_T(W)$	15
5.	CSV_1	12
6.	$\Delta R(j_1, j_2)$	10
7.	$\Delta\theta_{pull}$	8
8.	$\Delta\varphi(H, W)$	5
	$p_T(j_2)$	5
10.	$\Delta\eta(j_1, j_2)$	5
11.	$p_T(j_1)$	4
	$p_{T,jj}$	4

(a) DJ nominal

Ranking	Variable	# Top 5 occurrences
1.	N_{aj}	27
2.	m_{jj}	26
3.	CSV_2	17
4.	CSV_1	11
	$p_T(W)$	11
6.	$\Delta R(j_1, j_2)$	10
7.	$\Delta\varphi(H, W)$	9
8.	$\Delta\eta(j_1, j_2)$	8
9.	$\Delta\theta_{pull}$	6
10.	$p_T(j_1)$	4
	$p_{T,jj}$	4
12.	$p_T(j_2)$	2

(b) DJ $t\bar{t}$

Ranking	Variable	# Top 5 occurrences
1.	m_{jj}	27
	CSV_2	27
	CSV_1	27
4.	$\Delta R(j_1, j_2)$	17
5.	$p_T(j_2)$	16
6.	$\Delta\eta(j_1, j_2)$	7
7.	$\Delta\theta_{pull}$	4
	N_{aj}	4
	$p_T(W)$	4
10.	$p_T(j_1)$	1
	$p_{T,jj}$	1
12.	$\Delta\varphi(H, W)$	0

(c) DJ wjets

Ranking	Variable	# Top 5 occurrences
1.	m_{jj}	26
2.	$p_T(W)$	19
3.	$\Delta R(j_1, j_2)$	16
4.	$p_{T,jj}$	15
5.	$\Delta\varphi(H, W)$	14
6.	$\Delta\theta_{pull}$	13
7.	CSV_1	10
8.	CSV_2	8
9.	$\Delta\eta(j_1, j_2)$	5
10.	$p_T(j_1)$	3
	N_{aj}	3
	$p_T(j_2)$	3

(d) DJ vv

Table A.7.: Variable rankings in the SJF analysis for nominal and expert BDTs. The number of occurrences in the top 5 positions of the individual rankings from the 27 trainings (9 mass points and 3 analysis bins) are counted. The positions of the added substructure information are highlighted in red.

Ranking	Variable	# Top 5 occurrences
1.	N_{aj}	27
2.	m_{jj}	20
3.	CSV_2	14
4.	$p_T(W)$	13
5.	CSV_1	10
	m_{jj}^{flt}	10
7.	$\Delta\theta_{pull}$	9
8.	$p_{T,jj}$	8
	$\Delta R(j_1, j_2)$	8
10.	$\lvert m_{jj} - m_{jj}^{flt}\rvert$	4
11.	$\Delta\varphi(H, W)$	3
	$\Delta\eta(j_1, j_2)$	3
	$p_T(j_2)$	3
14.	$p_T(j_1)$	2
15.	$p_{T,jj}^{flt}$	1

(a) SJF nominal

Ranking	Variable	# Top 5 occurrences
1.	N_{aj}	27
2.	m_{jj}	24
3.	m_{jj}^{flt}	16
4.	CSV_2	15
	$p_T(W)$	15
6.	CSV_1	8
7.	$\Delta R(j_1, j_2)$	6
8.	$p_{T,jj}$	5
9.	$\lvert m_{jj} - m_{jj}^{flt}\rvert$	4
10.	$p_T(j_2)$	3
	$\Delta\theta_{pull}$	3
	$p_T(j_1)$	3
13.	$p_{T,jj}^{flt}$	2
	$\Delta\eta(j_1, j_2)$	2
	$\Delta\varphi(H, W)$	2

(b) SJF $t\bar{t}$

Ranking	Variable	# Top 5 occurrences
1.	m_{jj}	27
	CSV_2	27
	CSV_1	27
4.	$\Delta R(j_1, j_2)$	18
5.	$p_T(j_2)$	14
6.	$\Delta\eta(j_1, j_2)$	6
	$p_T(W)$	6
8.	m_{jj}^{flt}	4
9.	$\Delta\theta_{pull}$	2
	N_{aj}	2
11.	$\Delta\varphi(H, W)$	1
	$p_T(j_1)$	1
13.	$p_{T,jj}^{flt}$	0
	$p_{T,jj}$	0
	$\lvert m_{jj} - m_{jj}^{flt}\rvert$	0

(c) SJF wjets

Ranking	Variable	# Top 5 occurrences
1.	m_{jj}	26
2.	$\Delta\varphi(H, W)$	20
3.	$\Delta\eta(j_1, j_2)$	18
4.	$\Delta R(j_1, j_2)$	16
5.	CSV_2	11
	$p_T(W)$	11
7.	CSV_1	10
8.	m_{jj}^{flt}	9
9.	$\Delta\theta_{pull}$	6
10.	$p_T(j_1)$	2
	N_{aj}	2
	$\lvert m_{jj} - m_{jj}^{flt}\rvert$	2
13.	$p_{T,jj}$	1
	$p_T(j_2)$	1
15.	$p_{T,jj}^{flt}$	0

(d) SJF vv

Table A.8.: Configuration for all boosted decision trees trained in this analysis. For the filter jet regression BDT the boost type "Bagging" with "RegressionVariance" as separation type was used. The classification BDTs were optimized with the VH(125) signal sample in the corresponding regions and the "AdaBoost" option with AdaBoostBeta = 0.2 and "GiniIndex" as separation was employed. The number of cuts was set to 20 and no pruning was used for all BDTs.

Training		# trees	# min. events	max. depth
Filter jet regression		500	100	10
low p_{T}	DJ nominal	850	300	3
	DJ expert $\mathrm{t\bar{t}}$	150	400	3
	DJ expert W + 0b	100	450	3
	DJ expert VV	900	1200	3
med. p_{T}	DJ nominal	900	450	3
	DJ expert $\mathrm{t\bar{t}}$	100	750	3
	DJ expert W + 0b	150	450	3
	DJ expert VV	50	50	3
high p_{T}	DJ nominal	100	350	3
	DJ expert $\mathrm{t\bar{t}}$	550	1200	3
	DJ expert W + 0b	100	550	3
	DJ expert VV	150	100	3
low p_{T}	SJF nominal	900	650	3
	SJF expert $\mathrm{t\bar{t}}$	150	650	3
	SJF expert W + 0b	100	450	3
	SJF expert VV	900	1200	3
med. p_{T}	SJF nominal	1000	1050	3
	SJF expert $\mathrm{t\bar{t}}$	700	1000	3
	SJF expert W + 0b	150	400	3
	SJF expert VV	650	1250	3
high p_{T}	SJF nominal	500	600	3
	SJF expert $\mathrm{t\bar{t}}$	150	700	3
	SJF expert W + 0b	150	450	3
	SJF expert VV	100	600	3

B. Supplementary material for tHq analysis

Auxiliary information on the search for tHq final states is presented. These figures and tables supplement the analysis described in Chapter 6.

B.1. Technical details on data and MC samples used in the analysis

In Table B.1 the different analyzed datasets are given. The full list of MC samples is provided in Table B.2.

B.2. Addional information

The supplementary distributions and cross checks are given Figures B.1 - B.9. Tables B.3 - B.4 provide further results of the analysis.

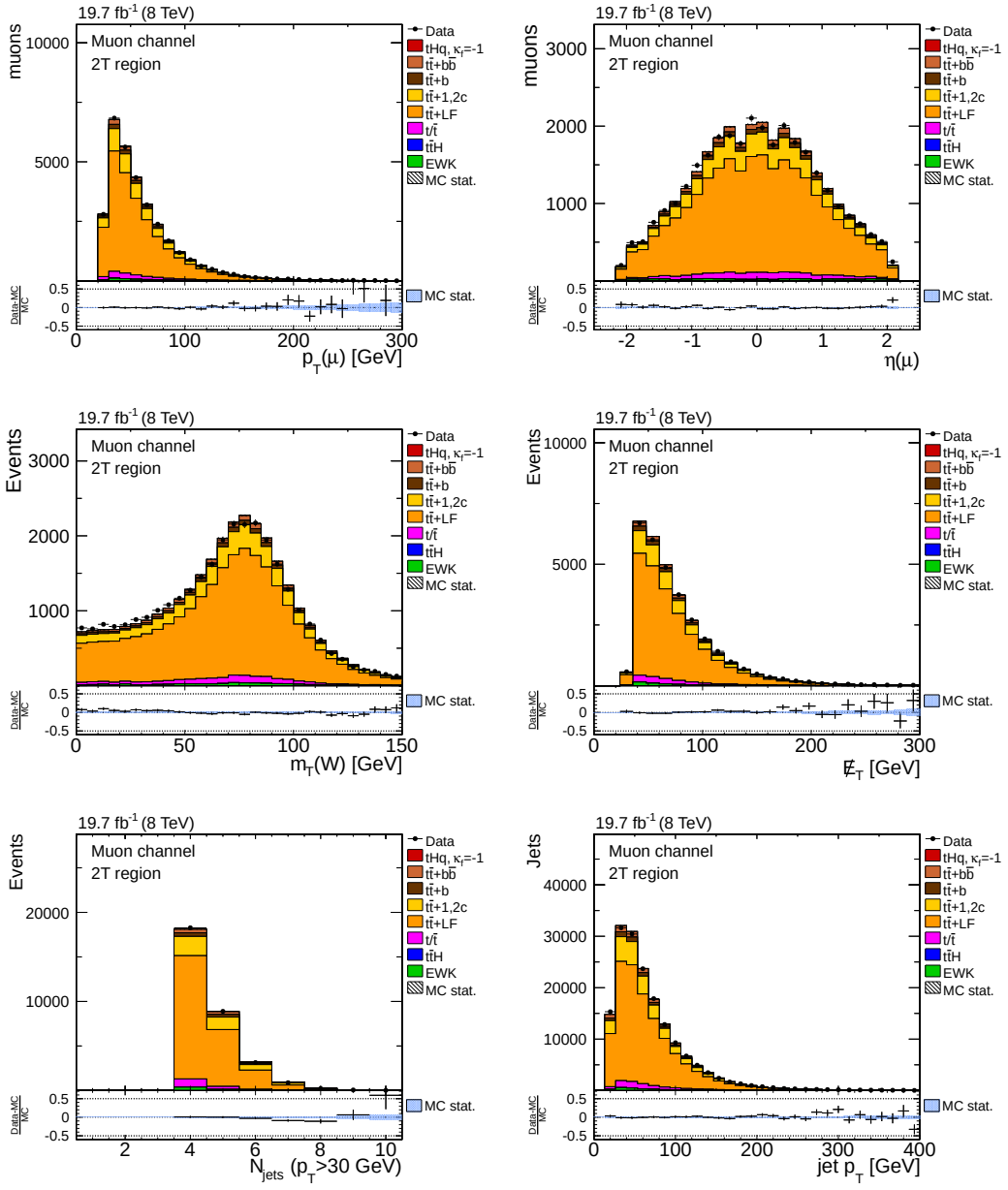


Figure B.1.: Event information variables for the muon channel in the 2T region. The number of simulated events is normalized to data. Overall good agreement between data and simulation is found.

Table B.1.: List of 2012 data samples used for the tHq search. In both, electron and muon channel, data corresponding to an integrated luminosity of $\sim 20\,\text{fb}^{-1}$ is included in the search. The third column shows the recorded integrated luminosity with only "good" runs according to the golden JSON file [184].

Channel	Dataset name	$\mathcal{L}$ (fb^{-1})
Muon	`/SingleMu/Run2012A-22Jan2013-v1/AOD`	0.876
	`/SingleMu/Run2012B-22Jan2013-v1/AOD`	4.412
	`/SingleMu/Run2012C-22Jan2013-v1/AOD`	7.055
	`/SingleMu/Run2012D-22Jan2013-v1/AOD`	7.369
	Total Lumi	19.71
Electron	`/SingleElectron/Run2012A-22Jan2013-v1/AOD`	0.876
	`/SingleElectron/Run2012B-22Jan2013-v1/AOD`	4.412
	`/SingleElectron/Run2012C-22Jan2013-v1/AOD`	7.055
	`/SingleElectron/Run2012D-22Jan2013-v1/AOD`	7.369
	Total Lumi	19.71

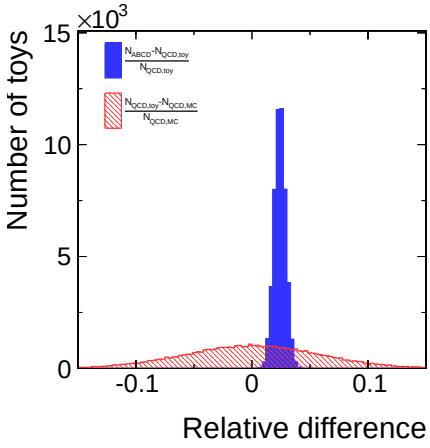


Figure B.2.: Closure test results for the `ABCD` method. In total 50,000 pseudo experiments are drawn with varied yields for each process. The blue histogram shows the relative difference between the QCD prediction calculated via the ABCD method (N_{ABCD}) and the drawn number of events in region A ($N_{\text{QCD,toy}}$). For the sake of completeness the relative difference of $N_{\text{QCD,toy}}$ and the number of events predicted by simulation in A ($N_{\text{QCD,MC}}$) is depicted in hatched red.

Table B.2.: Simulation datasets used in the tHq search. All cross sections are cited according to the generator. When only specific decays of a top quark are considered in a dataset, the inclusive cross section is scaled using $\mathcal{B}(\mathrm{W} \to \ell\nu) = 0.1080 \pm 0.0009$ [28]. If no reference is provided, the cross section is cited according to Ref. [186] or the used generator.

Process		Number of events	Eff. cross section [pb]
tHq	$\mathrm{H} \to \mathrm{b\bar{b}}$, $m_\mathrm{H} = 125\,\mathrm{GeV}$, $\kappa_\mathrm{f} = -1$	5 000 000	36.8×10^{-3} (NLO) [55]
$\mathrm{t\bar{t}H}$	inclusive, $m_\mathrm{H} = 125\,\mathrm{GeV}$	995 697	130.2×10^{-3} (NLO) [154]
$\mathrm{t\bar{t}}$ + jets	$\mathrm{t\bar{t}} \to \mathrm{b}\ell\nu\mathrm{bq\bar{q}}$	86 814 792	107.7 (NNLO) [185]
$\mathrm{t\bar{t}}$ + jets	$\mathrm{t\bar{t}} \to \mathrm{b}\ell\nu\mathrm{b}\ell\nu$	12 119 013	25.8 (NNLO) [185]
t (*t*-channel)	$\mathrm{t} \to \mathrm{b}\ell\nu$	3 915 598	18.27 (approx. NNLO) [155]
$\bar{\mathrm{t}}$ (*t*-channel)	$\bar{\mathrm{t}} \to \bar{\mathrm{b}}\ell\nu$	1 711 403	9.95 (approx. NNLO) [155]
t (tW-channel)	$\mathrm{t} \to \mathrm{b}\ell\nu$	497 658	11.1 (approx. NNLO) [155]
$\bar{\mathrm{t}}$ (tW-channel)	$\bar{\mathrm{t}} \to \bar{\mathrm{b}}\ell\nu$	493 460	11.1 (approx. NNLO) [155]
t (*s*-channel)	$\mathrm{t} \to \mathrm{b}\ell\nu$	3 932 710	1.23 (approx. NNLO) [155]
$\bar{\mathrm{t}}$ (*s*-channel)	$\bar{\mathrm{t}} \to \bar{\mathrm{b}}\ell\nu$	1 949 667	0.57 (approx. NNLO) [155]
W + jets	$\mathrm{W} \to \ell\nu$	57 509 905	35 509 (NNLO)
W + jets	$\mathrm{W} \to \ell\nu$, 2 add. jets	33 894 921	2116 (NNLO)
W + jets	$\mathrm{W} \to \ell\nu$, 3 add. jets	15 289 503	637 (NNLO)
W + jets	$\mathrm{W} \to \ell\nu$, 4 add. jets	13 382 803	262 (NNLO)
WW	inclusive	9 800 431	54.8 (NLO)
WZ	inclusive	9 950 283	12.6 (LO)
ZZ	inclusive	9 799 908	5.2 (LO)
Z/γ^* + jets	$\mathrm{Z}/\gamma^* \to \ell^+\ell^-$, $m(l^+l^-) > 50\,\mathrm{GeV}$	29 909 503	3504 (NNLO)
QCD μ-enriched	$\hat{p}_\mathrm{T} > 20\,\mathrm{GeV}$, $p_\mathrm{T}^\mu > 15\,\mathrm{GeV}$	29 013 914	135×10^3 (LO)
QCD em-enriched	$20\,\mathrm{GeV} < \hat{p}_\mathrm{T} < 30\,\mathrm{GeV}$	35 040 695	$2\,915\times10^3$ (LO)
QCD em-enriched	$30\,\mathrm{GeV} < \hat{p}_\mathrm{T} < 80\,\mathrm{GeV}$	33 088 888	$4\,616\times10^3$ (LO)
QCD em-enriched	$80\,\mathrm{GeV} < \hat{p}_\mathrm{T} < 170\,\mathrm{GeV}$	34 542 763	183×10^3 (LO)
QCD em-enriched	$170\,\mathrm{GeV} < \hat{p}_\mathrm{T} < 250\,\mathrm{GeV}$	31 697 066	4 587 (LO)
QCD em-enriched	$250\,\mathrm{GeV} < \hat{p}_\mathrm{T} < 350\,\mathrm{GeV}$	34 611 322	557 (LO)
QCD em-enriched	$\hat{p}_\mathrm{T} > 350\,\mathrm{GeV}$	34 080 562	89 (LO)
QCD BCtoE	$20\,\mathrm{GeV} < \hat{p}_\mathrm{T} < 30\,\mathrm{GeV}$	1 740 229	167×10^3 (LO)
QCD BCtoE	$30\,\mathrm{GeV} < \hat{p}_\mathrm{T} < 80\,\mathrm{GeV}$	2 048 152	167×10^3 (LO)
QCD BCtoE	$80\,\mathrm{GeV} < \hat{p}_\mathrm{T} < 170\,\mathrm{GeV}$	1 945 525	13.0×10^3 (LO)
QCD BCtoE	$170\,\mathrm{GeV} < \hat{p}_\mathrm{T} < 250\,\mathrm{GeV}$	1 948 112	632 (LO)
QCD BCtoE	$250\,\mathrm{GeV} < \hat{p}_\mathrm{T} < 350\,\mathrm{GeV}$	2 026 521	103 (LO)
QCD BCtoE	$\hat{p}_\mathrm{T} > 350\,\mathrm{GeV}$	1 948 532	24 (LO)
γ + jets	$40\,\mathrm{GeV} \leq H_\mathrm{T,had} \leq 100\,\mathrm{GeV}$	19 857 930	20 730 (LO)
γ + jets	$100\,\mathrm{GeV} \leq H_\mathrm{T,had} \leq 200\,\mathrm{GeV}$	9 612 703	5 330 (LO)
γ + jets	$200\,\mathrm{GeV} \leq H_\mathrm{T,had} \leq 400\,\mathrm{GeV}$	10 494 617	961 (LO)
γ + jets	$H_\mathrm{T,had} \geq 400\,\mathrm{GeV}$	1 611 963	103 (LO)

Table B.3.: QCD estimation results for signal and control regions via the ABCD method. Given are the event yields N for data, non-QCD background from Monte Carlo and their difference in the electron and muon channels. $N_{\text{QCD, ABCD}}$ denotes the predicted number of QCD events in Region A using equation (5.8). The calculations of the ABCD method in the signal and control regions yield a QCD contamination below 1 %.

Electron 2T region	B	C	D	A
N_{data}	1588	17808	2527	23887
$N_{\text{non-QCD}}$	1130 ± 8	16901 ± 36	696 ± 7	
N_{diff}	458 ± 8	907 ± 36	1831 ± 7	
$N_{\text{QCD, ABCD}}$				227 ± 10
$N_{\text{QCD, ABCD}}/N_{\text{data}}$				(0.95±0.04)%
Muon 2T region	B	C	D	A
N_{data}	1850	12244	806	32603
$N_{\text{non-QCD}}$	1670 ± 10	12165 ± 30	542 ± 7	
N_{diff}	180 ± 10	79 ± 30	264 ± 7	
$N_{\text{QCD, ABCD}}$				54 ± 21
$N_{\text{QCD, ABCD}}/N_{\text{data}}$				(0.17 ± 0.06)%
Electron 3T region	B	C	D	A
N_{data}	91	764	102	1063
$N_{\text{non-QCD}}$	51 ± 2	746 ± 7	30 ± 2	
N_{diff}	40 ± 2	18 ± 7	72 ± 2	
$N_{\text{QCD, ABCD}}$				10 ± 4
$N_{\text{QCD, ABCD}}/N_{\text{data}}$				(0.97 ± 0.39)%
Muon 3T region	B	C	D	A
N_{data}	105	583	50	1575
$N_{\text{non-QCD}}$	86 ± 2	578 ± 6	30 ± 3	
N_{diff}	19 ± 2	5 ± 6	20 ± 3	
$N_{\text{QCD, ABCD}}$				5^{+6}_{-5}
$N_{\text{QCD, ABCD}}/N_{\text{data}}$				(0.30±0.36)%

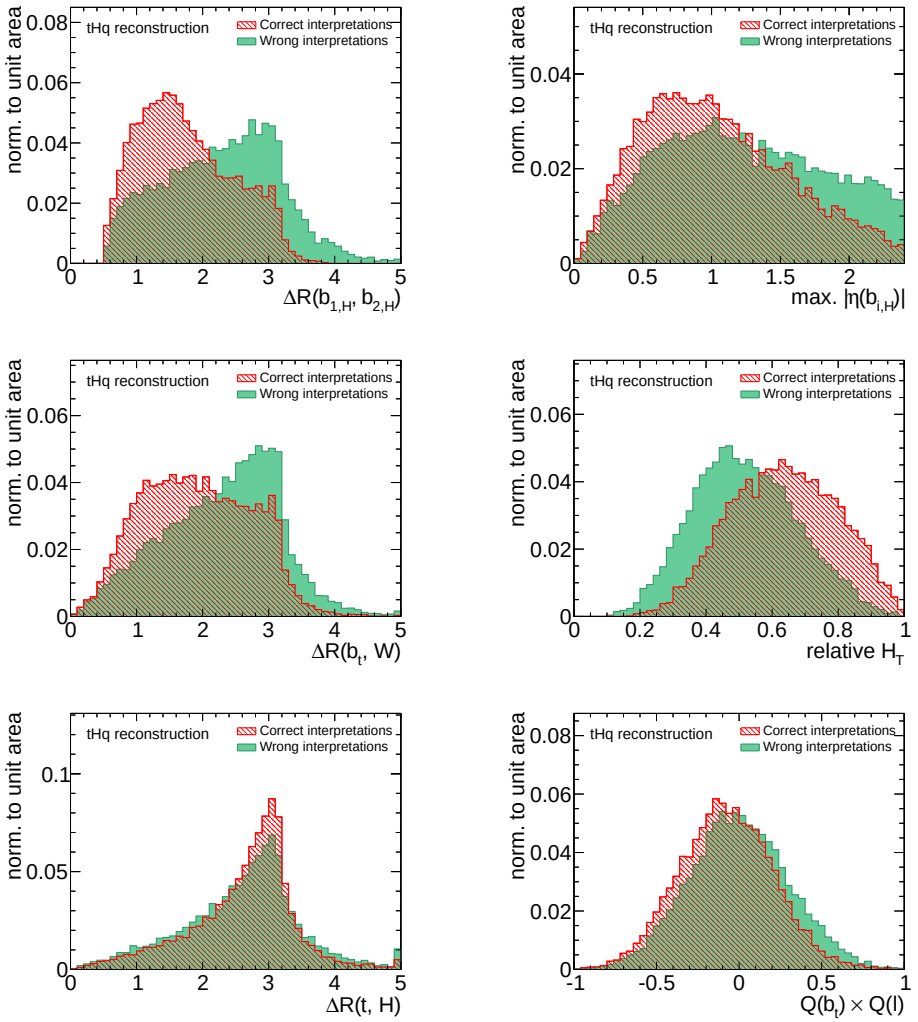


Figure B.3.: Remaining input variables for reconstruction under the tHq hypothesis.

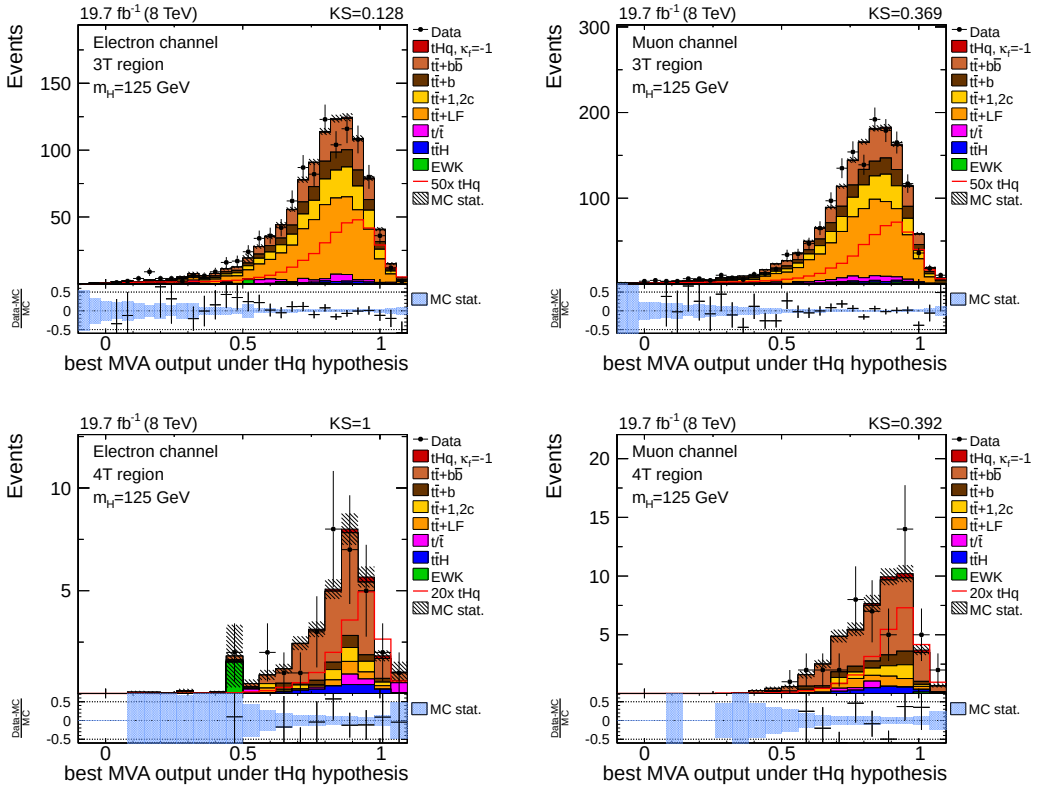


Figure B.4.: Data/MC comparisons for the tHq reconstruction when choosing the interpretion with the largest MVA output in each event. The distributions are shown for the 3T region (top row) and for the 4T region (bottom row) in both, the electron (left column) and muon channel (right column). The MC templates are normalized to data. KS-test probabilities are given and overall good agreement is found.

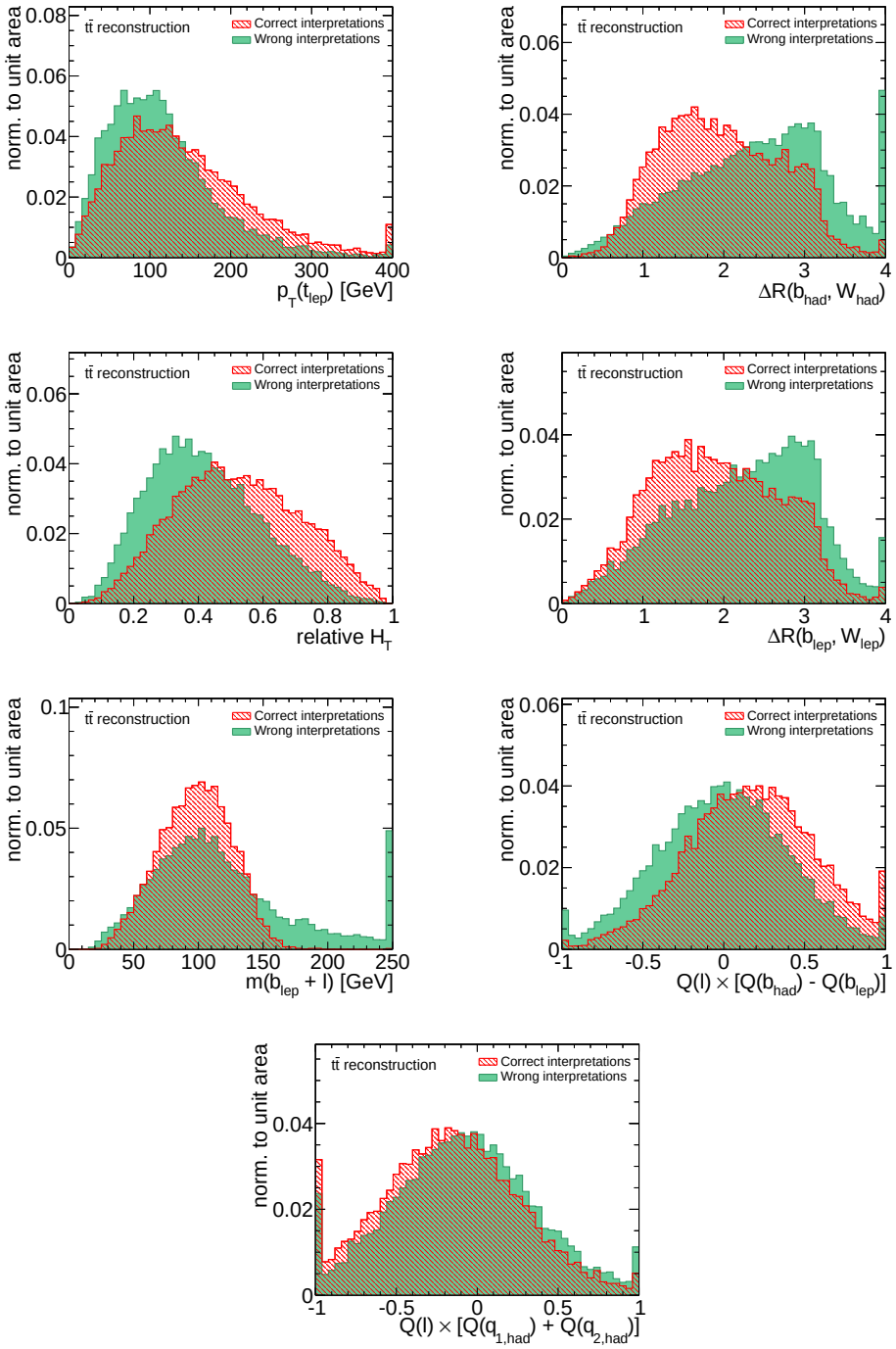


Figure B.5.: Remaining input variables for reconstruction under the $t\bar{t}$ hypothesis.

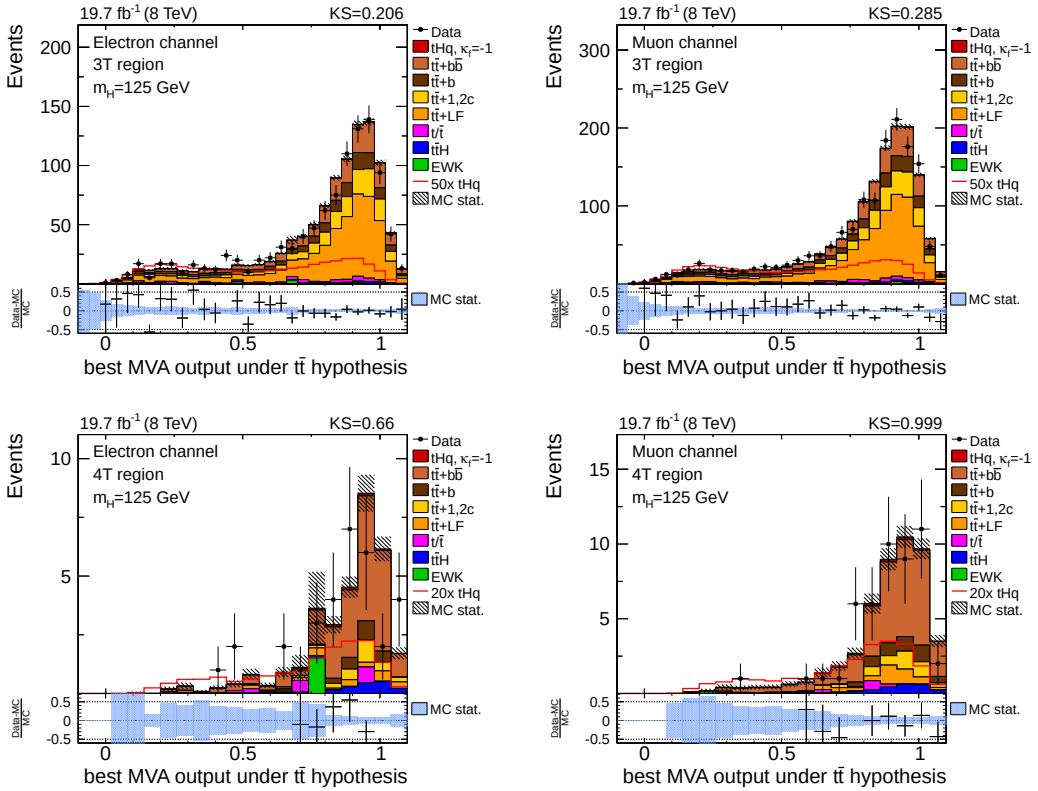


Figure B.6.: Data/MC comparisons for the tHq reconstruction when choosing the interpretion with the largest MVA output in each event. The distributions are shown for the 3T region (top row) and for the 4T region (bottom row) in both, the electron (left column) and muon channel (right column). The MC templates are normalized to data. KS-test probabilities are given and overall good agreement is found.

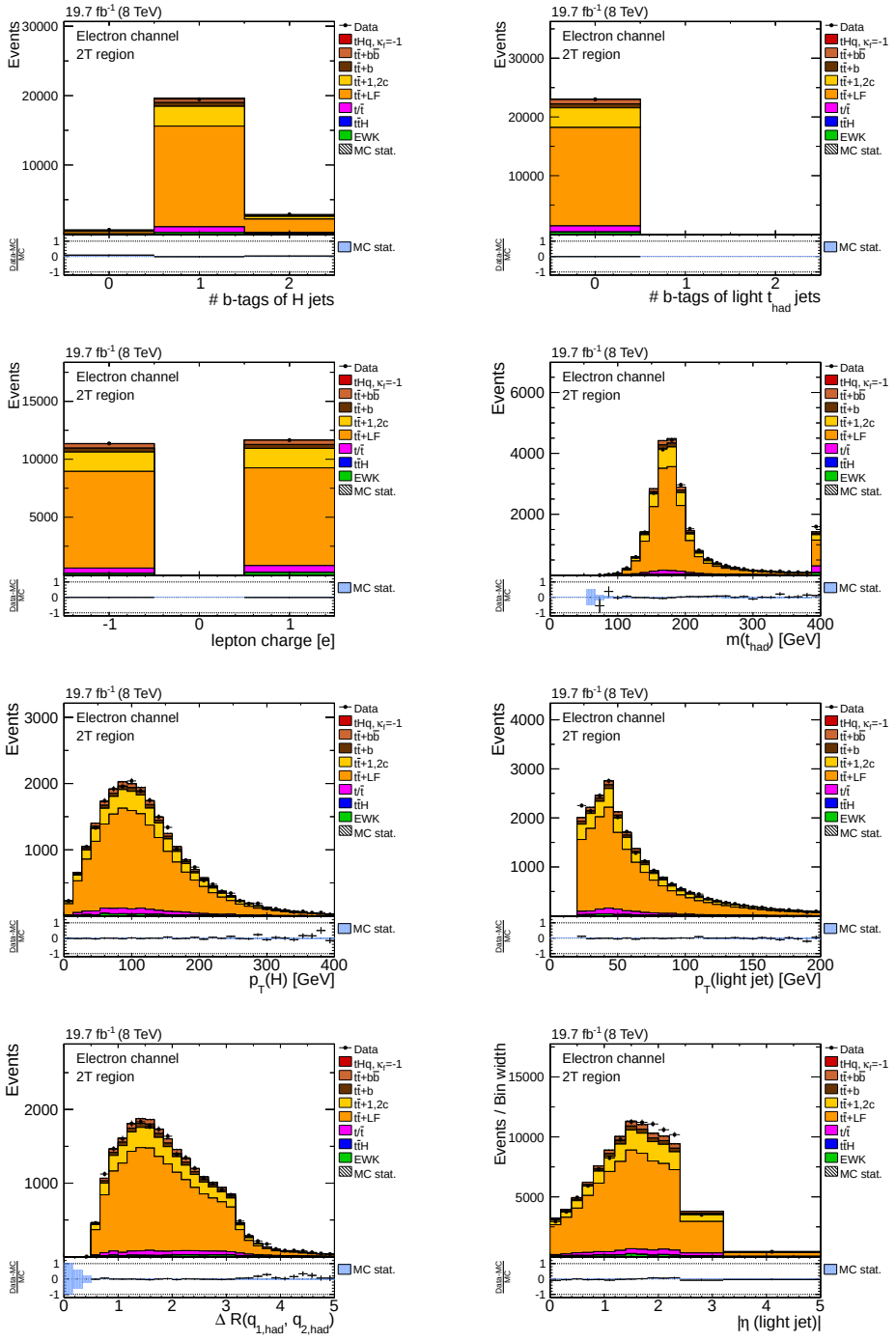


Figure B.7.: MVA classification input variables for the electron channel in the 2T region.

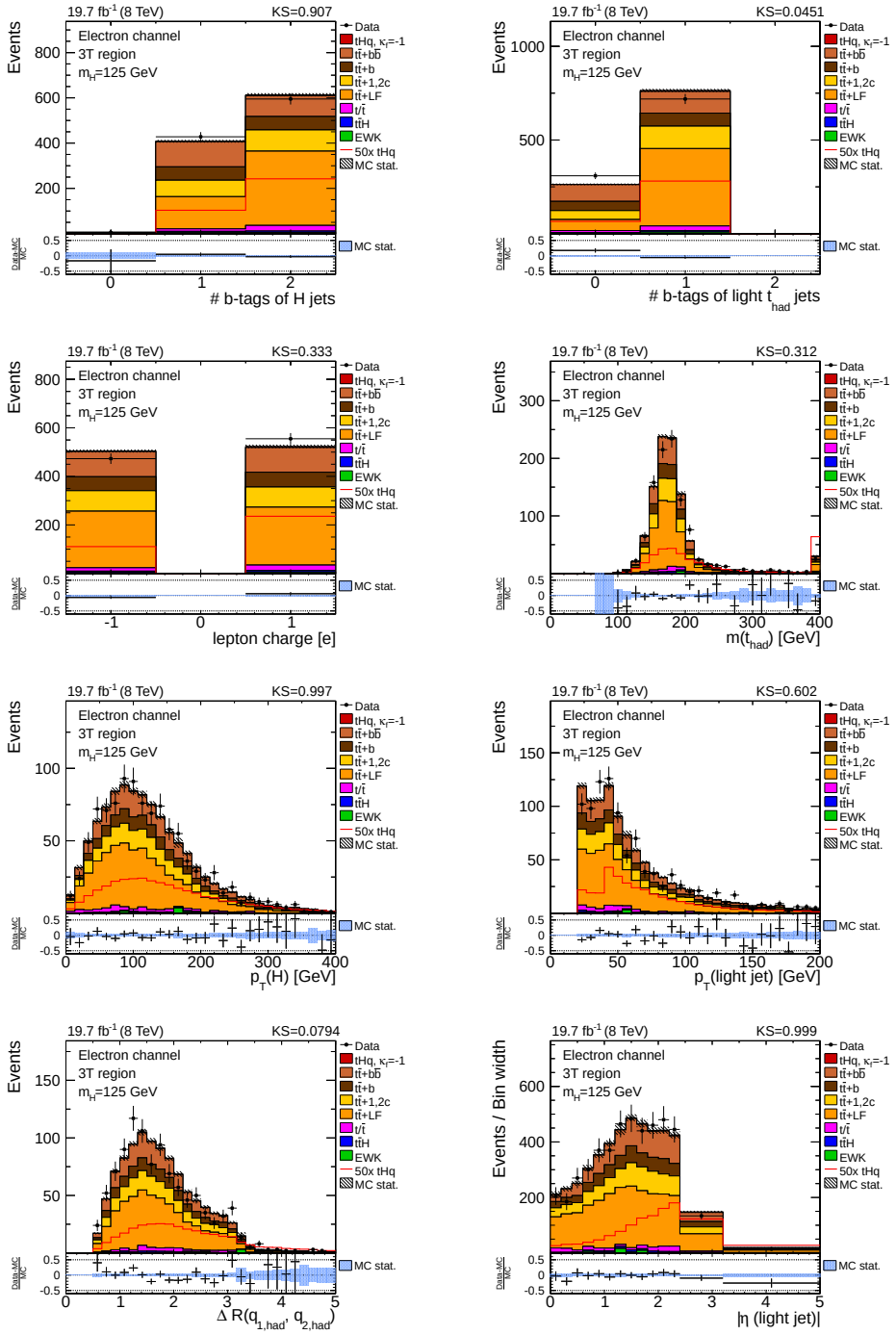


Figure B.8.: MVA classification input variables for the electron channel in the 3T region.

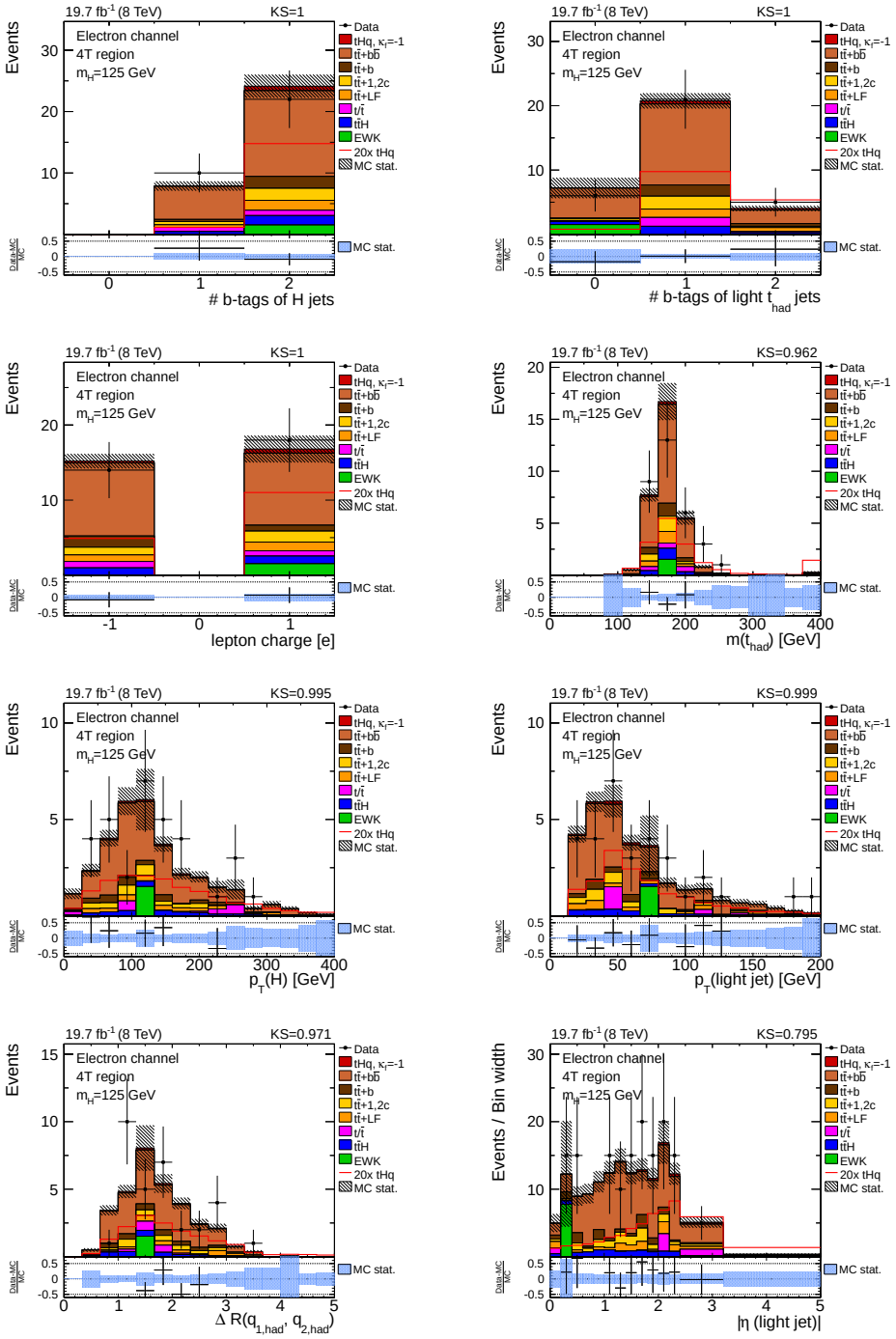


Figure B.9.: MVA classification input variables for the electron channel in the 4T region.

Table B.4.: Nuisance parameters for $t\bar{t}$ + X and their uncertainties after a simultaneous maximum-likelihood fit. All four signal regions are combined. The values indicate that the heavy-flavor content of $t\bar{t}$ + jets is under-estimated in simulation.

Process	Post-fit nuisance
$t\bar{t} + b$	1.38 ± 0.35
$t\bar{t} + b\bar{b}$	1.34 ± 0.22
$t\bar{t} + 1, 2c$	1.11 ± 0.34

List of Figures

List of Tables

Bibliography

[1] S. W. Herb et al., "Observation of a Dimuon Resonance at 9.5 GeV in 400-GeV Proton-Nucleus Collisions", Phys. Rev. Lett. 39, 252–255 (1977).

[2] CDF Collaboration, "Observation of Top Quark Production in $\overline{p}p$ Collisions with the Collider Detector at Fermilab", Phys. Rev. Lett. 74, 2626–2631 (1995).

[3] DØ Collaboration, "Observation of the Top Quark", Phys. Rev. Lett. 74, 2632–2637 (1995).

[4] K. Kodama et al., "Observation of Tau Neutrino Interactions", Physics Letters B 504, 218 – 224 (2001).

[5] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", Phys. Lett. B716, 1–29 (2012).

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

[7] A. Einstein, "Die Grundlage der Allgemeinen Relativitätstheorie", Annalen der Physik 354, 769–822 (1916).

[8] B. Povh, "Teilchen und Kerne : eine Einführung in die physikalischen Konzepte". Springer, Berlin, 8th edition, (2009).

[9] D. J. Griffiths, "Introduction to elementary particles". Wiley-VCH, Weinheim, 2nd edition, (2008).

[10] K. Olive et al., Particle Data Group Collaboration, Chin. Phys. C 38, 090001 (2014).

[11] S. L. Glashow, "Partial-Symmetries of Weak Interactions", Nuclear Physics 22, 579 – 588 (1961).

[12] A. Salam and J. Ward, "Electromagnetic and Weak Interactions", Physics Letters 13, 168 – 171 (1964).

[13] S. Weinberg, "A Model of Leptons", Phys. Rev. Lett. 19, 1264–1266 (1967).

[14] H. Georgi and S. L. Glashow, "Unified Weak and Electromagnetic Interactions without Neutral Currents", Phys. Rev. Lett. 28, 1494–1497 (1972).

[15] D. J. Gross and F. Wilczek, "Asymptotically Free Gauge Theories. 1", Phys. Rev. D8, 3633–3652 (1973).

[16] D. J. Gross and F. Wilczek, "Ultraviolet Behavior of Non-Abelian Gauge Theories", Phys. Rev. Lett. 30, 1343–1346 (1973).

[17] H. D. Politzer, "Reliable Perturbative Results for Strong Interactions?", Phys. Rev. Lett. 30, 1346–1349 (1973).

[18] H. Fritzsch, M. Gell-Mann, and H. Leutwyler, "Advantages of the Color Octet Gluon Picture", Phys. Lett. B47, 365–368 (1973).

[19] E. Noether, "Invariante Variationsprobleme", Nachr. d. Königl. Ges. d. Wiss. zu Göttingen Math–Phys. Klasse 235–257 (1918).

[20] S. Tomonaga, "On a Relativistically Invariant Formulation of the Quantum Theory of Wave Fields", Progress of Theoretical Physics 1, 27–42 (1946).

[21] J. Schwinger, "Quantum Electrodynamics. I. A Covariant Formulation", Phys. Rev. 74, 1439–1461 (1948).

[22] J. Schwinger, "Quantum Electrodynamics. II. Vacuum Polarization and Self-Energy", Phys. Rev. 75, 651–679 (1949).

[23] J. Schwinger, "Quantum Electrodynamics. III. The Electromagnetic Properties of the Electron-Radiative Corrections to Scattering", Phys. Rev. 76, 790–817 (1949).

[24] R. P. Feynman, "The Theory of Positrons", Phys. Rev. 76, 749–759 (1949).

[25] R. P. Feynman, "Space-Time Approach to Quantum Electrodynamics", Phys. Rev. 76, 769–789 (1949).

[26] R. P. Feynman, "Mathematical Formulation of the Quantum Theory of Electromagnetic Interaction", Phys. Rev. 80, 440–457 (1950).

[27] J. S. Schwinger, "Renormalisation Theory of Quantum Electrodynamics: An Individual View", The Birth of Particle Physics 329–353 (1983).

[28] J. Beringer et al., Particle Data Group Collaboration, "The Review of Particle Physics", Phys. Rev. D 86, 010001 (2012).

[29] S. Khalil and E. Torrente-Lujan, "Neutrino Mass and Oscillation as Probes of Physics Beyond the Standard Model", J. Egyptian Math. Soc. 9, 91–141 (2001).

[30] J. W. F. Valle, "Neutrino Physics Overview", Journal of Physics: Conference Series 53, 473 (2006).

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

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

[33] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, "Global Conservation Laws and Massless Particles", Phys. Rev. Lett. 13, 585–587 (1964).

[34] L. Álvarez Gaumé and J. Ellis, "Eyes on a prize particle", Nature Physics 7, 2–3 (2011).

[35] J. Goldstone, "Field Theories with Superconductor Solutions", Il Nuovo Cimento 19, 154–164 (1961).

[36] J. Goldstone, A. Salam, and S. Weinberg, "Broken Symmetries", Phys. Rev. 127, 965–970 (1962).

[37] Nobelprize.org, "The Nobel Prize in Physics 2013", `http://www.nobelprize.org/nobel_prizes/physics/laureates/2013/`, 2013.

[38] N. Cabibbo, "Unitary Symmetry and Leptonic Decays", Phys. Rev. Lett. 10, 531–533 (1963).

[39] M. Kobayashi and T. Maskawa, "CP Violation in the Renormalizable Theory of Weak Interaction", Prog. Theor. Phys. 49, 652–657 (1973).

[40] CMS Collaboration, "Measurement of the t-channel single-top-quark production cross section and of the $| V_{tb} |$ CKM matrix element in pp collisions at $\sqrt{s}$= 8 TeV", JHEP 1406, 090 (2014).

[41] F. J. Dyson, "The S Matrix in Quantum Electrodynamics", Phys. Rev. 75, 1736–1755 (Jun, 1949).

[42] P. A. M. Dirac, "The quantum theory of the emission and absorption of radiation", Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character 114, 243–265 (1927), 767.

[43] C. Møller, "Zur Theorie des Durchgangs schneller Elektronen durch Materie", Annalen der Physik 406, 531–585 (1932).

[44] J. C. Collins, D. E. Soper, and G. Sterman, "Heavy particle production in high-energy hadron collisions", Nuclear Physics B 263, 37 – 60 (1986).

[45] LHC Higgs Cross Section Working Group Collaboration, "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties", `arXiv:1307.1347` (2013).

[46] LHC Higgs Cross Section Working Group, "Picture Gallery", `https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWGCrossSectionsFigures`, March 2015.

[47] ATLAS Collaboration, "Combined measurements of the mass and signal strength of the Higgs-like boson with the ATLAS detector using up to 25 fb^{-1} of proton-proton collision data", ATLAS-CONF-2013-014, 2013.

[48] CMS Collaboration, "Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV", `arXiv:1412.8662` (2014).

[49] CMS Collaboration, "Measurement of the properties of a Higgs boson in the four-lepton final state", Phys. Rev. D89, 092007 (2014).

[50] CMS Collaboration, "Observation of the diphoton decay of the Higgs boson and measurement of its properties", Eur. Phys. J. C74, 3076 (2014), 10.

[51] ATLAS and CMS Collaborations, "Combined Measurement of the Higgs Boson Mass in pp Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments", `arXiv:1503.07589` (2015).

[52] ATLAS and CMS Collaborations, "Combination of single top-quark cross-sections measurements in the t-channel at $\sqrt{s} = 8\,\text{TeV}$ with the ATLAS and CMS experiments", ATLAS-CONF-2013-098, CMS-PAS-TOP-12-002, 2013.

[53] CMS Collaboration, "Observation of the associated production of a single top quark and a W boson in pp collisions at $\sqrt{s}$ =8 TeV", Phys. Rev. Lett. 112, 231802 (2014), 23.

[54] DØ Collaboration, "Evidence for s-channel single top quark production in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV", Phys. Lett. B726, 656–664 (2013).

[55] M. Farina, F. Grojean, F. Maltoni, E. Salvioni, and A. Thamm, "Lifting degeneracies in Higgs couplings using single top production in association with a Higgs boson", JHEP 1305, 022 (2013).

[56] ATLAS Collaboration, "Updated coupling measurements of the Higgs boson with the ATLAS detector using up to 25 fb^{-1} of proton-proton collision data", ATLAS-CONF-2014-009, 2014.

[57] J. Ellis and T. You, "Updated Global Analysis of Higgs Couplings", JHEP 1306, 103 (2013).

[58] J. Aguilar-Saavedra, "Top flavor-changing neutral interactions: Theoretical expectations and experimental detection", Acta Phys. Polon. B35, 2695–2710 (2004).

[59] J. Aguilar-Saavedra, R. Benbrik, S. Heinemeyer, and M. Pérez-Victoria, "Handbook of vectorlike quarks: Mixing and single production", Phys. Rev. D88, 094010 (2013), 9.

[60] CMS Collaboration, "Search for associated production of a single top quark and a Higgs boson in events where the Higgs boson decays to two photons at $\sqrt{s} = 8$ TeV", CMS-PAS-HIG-14-001, 2014.

[61] CMS Collaboration, "Search for H to $b\bar{b}$ in association with single top quarks as a test of Higgs couplings", CMS-PAS-HIG-14-015, 2014.

[62] CMS Collaboration, "Search for Associated Production of a Single Top Quark and a Higgs Boson in Leptonic Channels", CMS-PAS-HIG-14-026, 2015.

[63] A. Einstein, "Ist die Trägheit eines Körpers von seinem Energieinhalt abhängig?", Annalen der Physik 18, 639–643 (1905).

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

[65] J. Gruschke, "Observation of Top Quarks and First Measurement of the $t\bar{t}$ Production Cross Section at a Centre-Of-Mass Energy of 7 TeV with the CMS Experiment at the LHC", PhD Thesis, Karlsruhe Institute of Technology, CERN-THESIS-2011-030 (2011), urn:nbn:de:swb:90-223945.

[66] CMS Collaboration, "Public CMS Luminosity Information", `https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults`, March 2015.

[67] K. Aamodt et al., "The ALICE Experiment at the LHC", JINST 3, S08002 (2008).

[68] A. A. Alves et al., "The LHCb Detector at the LHC", JINST 3, S08005 (2008).

[69] G. Aad et al., "The ATLAS Experiment at the LHC", JINST 3, S08003 (2008).

[70] International Particle Physics Outreach Group Collaboration, "CMS Detector - 3D Image", `http://ippog.web.cern.ch/resources/2011/cms-detector-3d-image`, March 2015.

[71] R. Adolphi et al., CMS Collaboration, "The CMS experiment at the CERN LHC", JINST 0803, S08004 (2008).

[72] S. Chatrchyan et al., CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", JINST 9, P10009 (2014).

[73] CMS Collaboration, "The CMS Electromagnetic Calorimeter Project: Technical Design Report", CERN-LHCC-97-33 (1997).

[74] CMS Collaboration, "Changes to the CMS ECAL Electronics: Addendum to the Technical Design Report", CERN-LHCC-2002-027 (2002).

[75] CMS Collaboration, "The CMS Hadronic Calorimeter Project: Technical Design Report", CERN-LHCC-97-31 (1997).

[76] G. Della Ricca (CMS Collaboration), "Performance of the CMS Electromagnetic Calorimeter at the LHC", CMS-CR-2011-297, 2011.

[77] CMS Collaboration, "Performance of the CMS Hadron Calorimeter with Cosmic Ray Muons and LHC Beam Data", JINST 5, T03012 (2010).

[78] CMS Collaboration, "CMS Physics Technical Design Report Volume I: Detector Performance and Software", CERN-LHCC-2006-001 (2006).

[79] CMS Collaboration, "The CMS muon project: Technical Design Report", CERN-LHCC-97-032 (1997).

[80] CMS Collaboration, "The TriDAS Project Technical Design Report, Volume 1: The Trigger Systems", CERN-LHCC-2000-038 (2000).

[81] CMS Collaboration, "The TriDAS Project Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger", CERN-LHCC-2002-026 (2002).

[82] WLCG Collaboration, "Welcome to the Worldwide LHC Computing Grid", `http://wlcg.web.cern.ch`, March 2015.

[83] R. Egeland, T. Wildish, and S. Metson, "Data transfer infrastructure for CMS data taking", XII Advanced Computing and Analysis Techniques in Physics Research, PoS ACAT08, 033 (2008).

[84] C. Böser, Th. Chwalek, M. Giffels, V. Kuznetsov, and T. Wildish, "Integration and validation testing for PhEDEx, DBS and DAS with the PhEDEx LifeCycle agent", Journal of Physics: Conference Series 513, 062051 (2014).

[85] L. Tuura, A. Meyer, I. Segoni, and G. D. Ricca, "CMS data quality monitoring: Systems and experiences", Journal of Physics: Conference Series 219, 072020 (2010).

[86] M. Renz, "Erste Messung des Wirkungsquerschnitts der Top-Quark-Paarproduktion bei $\sqrt{s} = 7\,\text{TeV}$ im Elektron+Jets Kanal mit dem CMS-Experiment", PhD Thesis, Karlsruhe Institute of Technology, CERN-THESIS-2011-192 (2011), urn:nbn:de:swb:90-236748.

[87] The Durham HepData Project, "Online PDF plotting and calculation", `http://hepdata.cedar.ac.uk/pdf/pdf3.html`, March 2015.

[88] G. Altarelli and G. Parisi, "Asymptotic Freedom in Parton Language", Nucl. Phys. B126, 298 (1977).

[89] V. N. Gribov and L. N. Lipatov, "Deep Inelastic ep Scattering in Perturbation Theory", Sov. J. Nucl. Phys. 15, 438–450 (1972).

[90] Y. L. Dokshitzer, "Calculation of the Structure Functions for Deep Inelastic Scattering and e^+e^- Annihilation by Perturbation Theory in Quantum Chromodynamics", Sov. Phys. JETP 46, 641–653 (1977).

[91] S. Catani, F. Krauss, R. Kuhn, and B. Webber, "QCD matrix elements + parton showers", JHEP 0111, 063 (2001).

[92] M. L. Mangano, M. Moretti, F. Piccinini, and M. Treccani, "Matching matrix elements and shower evolution for top-quark production in hadronic collisions", JHEP 0701, 013 (2007).

[93] B. Andersson, G. Gustafson, G. Ingelman, and T. Sjöstrand, "Parton Fragmentation and String Dynamics", Phys. Rept. 97, 31–145 (1983).

[94] B. Webber, "A QCD Model for Jet Fragmentation Including Soft Gluon Interference", Nucl. Phys. B238, 492 (1984).

[95] T. Sjöstrand, S. Mrenna, and P. Z. Skands, "PYTHIA 6.4 Physics and Manual", JHEP 0605, 026 (2006).

[96] CMS Collaboration, "Measurement of the underlying event activity at the LHC with $\sqrt{s} = 7\,\text{TeV}$ and comparison with $\sqrt{s} = 0.9\,\text{TeV}$", Journal of High Energy Physics 2011 (2011), 9.

[97] M. Bähr, S. Gieseke, M. Gigg, D. Grellscheid, K. Hamilton, et al., "Herwig++ Physics and Manual", Eur. Phys. J. C58, 639–707 (2008).

[98] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, and T. Stelzer, "MadGraph 5 : Going Beyond", JHEP 1106, 128 (2011).

[99] P. Nason, "A New method for combining NLO QCD with shower Monte Carlo algorithms", JHEP 0411, 040 (2004).

[100] S. Frixione, P. Nason, and C. Oleari, "Matching NLO QCD computations with Parton Shower simulations: the POWHEG method", JHEP 0711, 070 (2007).

[101] S. Alioli, P. Nason, C. Oleari, and E. Re, "NLO single-top production matched with shower in POWHEG: s- and t-channel contributions", JHEP 09, 111 (2009).

[102] K. Hamilton, "A positive-weight next-to-leading order simulation of weak boson pair production", JHEP 1101, 009 (2011).

[103] S. Jadach, J. H. Kühn, and Z. Wąs, "TAUOLA- a library of Monte Carlo programs to simulate decays of polarized τ leptons", Computer Physics Communications 64, 275 – 299 (1991).

[104] S. Agostinelli, J. Allison, et al., "GEANT 4: A simulation toolkit", Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 506, 250 – 303 (2003).

[105] CMS Collaboration, "Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and $\not\!\!E_T$", CMS-PFT-09-001, 2009.

[106] CMS Collaboration, "Commissioning of the Particle-Flow event reconstruction with the first LHC collisions recorded in the CMS detector", CMS-PAS-PFT-10-001, 2010.

[107] W. Adam, B. Mangano, T. Speer, and T. Todorov "Track Reconstruction in the CMS tracker", CMS-NOTE-2006-041, 2006.

[108] S. Cucciarelli, M. Konecki, D. Kotlinski, and T. Todorov "Track reconstruction, primary vertex finding and seed generation with the Pixel Detector", CMS-NOTE-2006-026, 2006.

[109] T. Speer, W. Adam, R. Frühwirth, A. Strandlie, T. Todorov, and M. Winkler, "Track reconstruction in the CMS tracker", Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 559, 143 – 147 (2006).

[110] R. Frühwirth, "Application of Kalman filtering to track and vertex fitting", Nucl.Instrum.Meth. A262, 444–450 (1987).

[111] R. Frühwirth, W. Waltenberger, and P. Vanlaer "Adaptive Vertex Fitting", CMS-NOTE-2007-008, Mar, 2007.

[112] R. Frühwirth, "Track fitting with non-Gaussian noise", Comput. Phys. Commun. 100, 1–16 (1997).

[113] W. Adam, R. Frühwirth, A. Strandlie, and T. Todor "Reconstruction of Electrons with the Gaussian-Sum Filter in the CMS Tracker at the LHC", CMS-NOTE-2005-001, 2005.

[114] M. Pioppi, "A pre-identification for electron reconstruction in the CMS particle-flow algorithm", Journal of Physics: Conference Series 119, 032039 (2008).

[115] F. Beaudette, D. Benedetti, P. Janot, and M. Pioppi (CMS Collaboration), "Electron Reconstruction within the Particle Flow Algorithm", CMS Internal Note 2010/034, 2010.

[116] G. P. Salam, "A Practical seedless infrared safe cone algorithm", `arXiv:0705.2696` (2007).

[117] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k_{T} jet clustering algorithm", JHEP 04, 063 (2008).

[118] Y. L. Dokshitzer, G. Leder, S. Moretti, and B. Webber, "Better jet clustering algorithms", JHEP 9708, 001 (1997).

[119] S. Catani, Y. L. Dokshitzer, M. Seymour, and B. Webber, "Longitudinally invariant K_t clustering algorithms for hadron hadron collisions", Nucl. Phys. B406, 187–224 (1993).

[120] S. D. Ellis and D. E. Soper, "Successive combination jet algorithm for hadron collisions", Phys. Rev. D48, 3160–3166 (1993).

[121] M. Cacciari and G. P. Salam, "Dispelling the N^3 myth for the k_t jet-finder", Phys. Lett. B641, 57–61 (2006).

[122] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet User Manual", Eur. Phys. J. C72, 1896 (2012).

[123] CMS Collaboration, "Study of Jet Substructure in pp Collisions at 7 TeV in CMS", CMS-PAS-JME-10-013, 2011.

[124] J. M. Butterworth, A. R. Davison, M. Rubin, and G. P. Salam, "Jet Substructure as a New Higgs-Search Channel at the Large Hadron Collider", Phys. Rev. Lett. 100, 242001 (2008).

[125] CMS Collaboration, "Determination of jet energy calibration and transverse momentum resolution in CMS", JINST 6, 11002 (2011).

[126] M. Cacciari and G. P. Salam, "Pileup subtraction using jet areas", Phys. Lett. B659, 119–126 (2008).

[127] CMS Collaboration, "Determination of jet energy calibration and transverse momentum resolution in CMS", JINST 6, P11002 (2011).

[128] CMS Collaboration, "Algorithms for b jet identification in CMS", CMS-PAS-BTV-09-001, 2009.

[129] CMS Collaboration, "b-Jet Identification in the CMS Experiment", CMS-PAS-BTV-11-004, 2012.

[130] DØ Collaboration, "B-Jet Identification", `http://www-d0.fnal.gov/Run2Physics/top/singletop_observation/b_tagging_graphic.png`, March 2015.

[131] CMS Collaboration, "MET Analysis", `https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookMetAnalysis`, March 2015.

[132] J. Ott, T. Müller, and J. Wagner-Kuhr, "Theta – A Framework for Template-based Modeling and Interference", `http://theta-framework.org`, 2010.

[133] R. Brun and F. Rademakers, "ROOT - An Object Oriented Data Analysis Framework", `http://root.cern.ch`, Nucl. Inst. & Meth. in Phys. Res. A 389 (1997) 81-86.

[134] A. Hoecker, P. Speckmayer, J. Stelzer, J. Therhaag, E. von Toerne, H. Voss, et al., "TMVA: Toolkit for Multivariate Data Analysis", PoS ACAT, 040 (2007).

[135] V. Blobel and E. Lohrmann, "Statistische und numerische Methoden der Datenanalyse", Teubner Verlag (1998).

[136] A. L. Read, "Presentation of search results: the CL_s technique", Journal of Physics G: Nuclear and Particle Physics 28, 2693 (2002).

[137] T. Junk, "Confidence level computation for combining searches with small statistics", Nucl.Instrum.Meth. A434, 435–443 (1999).

[138] A. L. Read, "Modified frequentist analysis of search results (the CL_s method)", CERN-OPEN-2000-205 (2000).

[139] S. S. Wilks, "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses", Ann. Math. Statist. 9, 60–62 (1938).

[140] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics", The European Physical Journal C 71 (2011).

[141] J. Ott, "Search for Resonant Top Quark Pair Production in the Muon+Jets Channel with the CMS Detector", PhD Thesis, Karlsruhe Institute of Technology, CERN-THESIS-2012-262 (2012), urn:nbn:de:swb:90-315182.

[142] R. J. Barlow and C. Beeston, "Fitting using finite Monte Carlo samples", Comput. Phys. Commun. 77, 219–228 (1993).

[143] J. S. Conway, "Nuisance Parameters in Likelihoods for Multisource Spectra", Proceedings of PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding 115–120 (2011).

[144] L. Breiman, "Bagging Predictors", Machine Learning 24, 123–140 (1996).

[145] C. Böser, S. Fink, and S. Röcker, "Introduction to Boosted Decision Trees", `https://indico.scc.kit.edu/indico/contributionDisplay.py?sessionId=4&contribId=35&confId=48`, KSETA Doctoral Workshop Freudenstadt, July 2014.

[146] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting", Journal of Computer and System Sciences 55, 119 – 139 (1997).

[147] H. Drucker, "Improving Regressors Using Boosting Techniques", Proceedings of the Fourteenth International Conference on Machine Learning 107–115 (1997).

[148] C. G. Broyden, "The Convergence of a Class of Double-rank Minimization Algorithms – 1. General Considerations", IMA Journal of Applied Mathematics 6, 76–90 (1970).

[149] R. Fletcher, "A new approach to variable metric algorithms", The Computer Journal 13, 317–322 (1970).

[150] D. Goldfarb, "A family of variable-metric methods derived by variational means", Math. Comp. 24, 23–26 (1970).

[151] D. F. Shanno, "Conditioning of quasi-Newton methods for function minimization", Math. Comp. 24, 647–656 (1970).

[152] ATLAS Collaboration, "Search for the $b\bar{b}$ decay of the Standard Model Higgs boson in associated $(W/Z)H$ production with the ATLAS detector", `arXiv:1409.6212` (2014).

[153] CMS Collaboration, "Search for the standard model Higgs boson produced in association with a W or a Z boson and decaying to bottom quarks", Phys. Rev. D89, 012003 (2014).

[154] S. Heinemeyer et al., LHC Higgs Cross Section Working Group Collaboration, "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties", `arXiv:1307.1347` (2013).

[155] N. Kidonakis, "Differential and total cross sections for top pair and single top production", `arXiv:1205.3453` (2012).

[156] S. Agostinelli et al., GEANT4 Collaboration, "GEANT4 — a simulation toolkit", Nucl. Instrum. Meth. A 506, 250–303 (2003).

[157] CMS Collaboration, "Cert_190782-190949_8TeV_06Aug2012ReReco_Collisions12_JSON",

https://hypernews.cern.ch/HyperNews/CMS/get/physics-validation/1848.html, March 2015.

[158] CMS Collaboration, "Cert_190456-196531_8TeV_13Jul2012ReReco_Collisions12_JSON_v2", https://hypernews.cern.ch/HyperNews/CMS/get/physics-validation/1890.html, March 2015.

[159] CMS Collaboration, "Cert_190456-208686_8TeV_PromptReco_Collisions12_JSON", https://hypernews.cern.ch/HyperNews/CMS/get/physics-validation/1968.html, March 2015.

[160] CMS Collaboration, "Cert_201191-201191_8TeV_11Dec2012ReReco-recover_Collisions12_JSON", https://hypernews.cern.ch/HyperNews/CMS/get/physics-validation/1968/1/2/1/1.html, March 2015.

[161] VHbb Team (CMS Collaboration), "Search for the Standard Model Higgs Boson Produced in Association with a W or Z and Decaying to Bottom Quarks", CMS Internal Note 2013/069, 2013.

[162] H. Kirschenmann (CMS Collaboration), "Jet performance in CMS", CMS-CR-2013-325, 2013.

[163] CMS Collaboration, "Pileup Reweighting Utilities", https://twiki.cern.ch/twiki/bin/view/CMS/PileupMCReweightingUtilities, March 2015.

[164] CMS Collaboration, "MVA electron identification", https://twiki.cern.ch/twiki/bin/view/CMS/MultivariateElectronIdentification, 2012.

[165] CMS Collaboration, "Performance of the b-jet identification in CMS", CMS-PAS-BTV-11-001, 2011.

[166] CMS Collaboration, "Methods to apply b-tagging efficiency scale factors", https://twiki.cern.ch/twiki/bin/view/CMS/BTagSFMethods (Method 2b), March 2015.

[167] CDF and DØ Collaborations, "Improved b-jet Energy Correction for $H \to b\bar{b}$ Searches", arXiv:1107.3026 (2011).

[168] J. Gallicchio and M. D. Schwartz, "Seeing in Color: Jet Superstructure", Phys. Rev. Lett. 105, 022001 (2010).

[169] S. Fink, "Search for a Light Higgs Boson Decaying into Bottom Quarks Using Boosted Decision Trees with the CMS Experiment", Diploma Thesis, Karlsruhe Institute of Technology, IEKP-KA/2013-13 (2013).

[170] H. Held, "Measurement of the Jet Momentum Resolution and Search for a light Standard Model Higgs Boson in the $H(b\bar{b})W(\ell\nu)$ Channel with the CMS Detector at the LHC", PhD Thesis, Karlsruhe Institute of Technology, CERN-THESIS-2012-397 (2012), urn:nbn:de:swb:90-281963.

[171] B. Maier, "Search for a light Standard Model Higgs Boson in the $WH \to \ell\nu b\bar{b}$ Channel with the CMS Detector at the LHC", Diploma Thesis, Karlsruhe Institute of Technology, IEKP-KA/2012-19 (2012).

[172] C. Böser, "KIT Status Report on Substructure Studies (internal talk)", `https://indico.cern.ch/event/209767/session/1/contribution/46/material/slides/0.pdf`, September 2012.

[173] VHbb Team (CMS Collaboration), "BDT binning transformation for VHbb (internal talk)", `http://nmohr.web.cern.ch/nmohr/forVHbb/Talks/BinningTrafo.pdf`, March 2015.

[174] CMS Collaboration, "CMS Luminosity Based on Pixel Cluster Counting - Summer 2012 Update", CMS-PAS-LUM-12-001, 2012.

[175] M. Ciccolini, A. Denner, and S. Dittmaier, "Strong and electroweak corrections to the production of Higgs+2jets via weak interactions at the LHC", Phys. Rev. Lett. 99, 161803 (2007).

[176] M. Ciccolini, A. Denner, and S. Dittmaier, "Electroweak and QCD corrections to Higgs production via vector-boson fusion at the LHC", Phys. Rev. D77, 013002 (2008).

[177] A. Denner, S. Dittmaier, S. Kallweit, and A. Muck, "Electroweak corrections to Higgs-strahlung off W/Z bosons at the Tevatron and the LHC with HAWK", JHEP 1203, 075 (2012).

[178] G. Ferrera, M. Grazzini, and F. Tramontano, "Associated WH production at hadron colliders: a fully exclusive QCD calculation at NNLO", `arXiv:1107.1164` (2011).

[179] CMS Collaboration, "Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS", JINST 06, 11002 (2011).

[180] CMS Collaboration, "Measurement of the single-top t-channel cross section in pp collisions at centre-of-mass energy of 8 TeV", CMS-PAS-TOP-12-011, 2012.

[181] CMS Collaboration, "Datacard repository", `https://svnweb.cern.ch/cern/wsvn/cmshcg/trunk/summer2013/searches/vhbb`, 2014.

[182] W. Verkerke and D. Kirkby, "The RooFit toolkit for data modeling", ArXiv Physics e-prints (June, 2003).

[183] L. Moneta, K. Cranmer, G. Schott, and W. Verkerke, "The RooStats project", `arXiv:1009.1003` (2010).

[184] CMS Collaboration, "Golden JSON File", `https://hypernews.cern.ch/HyperNews/CMS/get/physics-validation/2065.html`, March 2015.

[185] M. Czakon, P. Fiedler, and A. Mitov, "The total top quark pair production cross-section at hadron colliders through $\mathcal{O}(\alpha_S^4)$", Phys. Rev. Lett. 110, 252004 (2013).

[186] CMS Collaboration, "Standard Model Cross Sections for CMS at 8 TeV", `https://twiki.cern.ch/twiki/bin/view/CMS/StandardModelCrossSectionsat8TeV`, March 2015.

[187] CMS Collaboration, "Search for Higgs Boson Production in Association with a Top-Quark Pair and Decaying to Bottom Quarks or Tau Leptons", CMS-PAS-HIG-13-019, 2013.

[188] CMS Collaboration, "Effective Areas for the particle-based isolation for electrons", `https://twiki.cern.ch/twiki/bin/view/CMS/EgammaEARhoCorrection`, March 2015.

[189] CMS Collaboration, "Multivariate Electron Identification", `http://twiki.cern.ch/twiki/bin/view/CMS/MultivariateElectronIdentification`, March 2015.

[190] CMS Collaboration, "Electron Efficiency Measurement for Top Quark Physics at $\sqrt{s} = 8$ TeV", CMS Internal Note 2012/429, 2012.

[191] CMS Collaboration, "Reference Muon ID and Isolation Efficiencies", `https://twiki.cern.ch/twiki/bin/view/CMS/MuonReferenceEffs`, March 2015.

[192] CMS Collaboration, "Jet Identification", `https://twiki.cern.ch/twiki/bin/view/CMS/JetID`, March 2015.

[193] CMS Collaboration, "Jet Transverse Momentum Resolution Measurement using Dijet Events at $\sqrt{s} = 8\,\text{TeV}$", CMS Internal Note 2013/416, 2014.

[194] CMS Collaboration, "Identification of b-quark jets with the CMS experiment", JINST 8, 04013 (2013).

[195] CMS Collaboration, "Methods to apply b-tagging efficiency scale factors", `https://twiki.cern.ch/twiki/bin/view/CMS/BTagSFMethods` (Method 1a), March 2015.

[196] Th. Chwalek, "Measurement of the W-Boson Helicity-Fractions in Top-Quark Decays with the CDF II Experiment and Prospects for an Early $t\bar{t}$ Cross-Section Measurement with the CMS Experiment", PhD Thesis, Karlsruhe Institute of Technology, CERN-THESIS-2010-255 (2010), urn:nbn:de:swb:90-168108.

[197] CMS Collaboration, "Top p_{T} Reweighting", `https://twiki.cern.ch/twiki/bin/view/CMS/TopPtReweighting`, March 2015.

[198] CMS Collaboration, "Search for $\mathrm{H} \to \mathrm{b\bar{b}}$ in association with single top quarks as a test of Higgs boson couplings", CMS Internal Note 2013/113, 2014.

[199] A. Popov, "Search for $\mathrm{H} \to \mathrm{b\bar{b}}$ in association with single top quarks as a test of Higgs boson couplings (Pre-approval)", `https://indico.cern.ch/event/320346/session/4/contribution/5/material/slides/0.pdf`. CMS Internal Talk (June 2014).

[200] D. Krohn, M. D. Schwartz, T. Lin, and W. J. Waalewijn, "Jet Charge at the LHC", Phys. Rev. Lett. 110, 212001 (2013).

[201] CMS Collaboration, "CMS Luminosity Based on Pixel Cluster Counting - Summer 2013 Update", CMS-PAS-LUM-13-001, 2013.

[202] CMS Collaboration, "Tag and Probe [Muons]", `https://twiki.cern.ch/twiki/bin/view/CMS/MuonTagAndProbe`, March 2015.

[203] CMS Collaboration, "Jet energy scale uncertainty sources", `https://twiki.cern.ch/twiki/bin/view/CMS/JECUncertaintySources`, March 2015.

[204] CMS Collaboration, "b Tag & Vertexing Physics Object Group", `https://twiki.cern.ch/twiki/bin/view/CMS/BtagPOG`, March 2015.

[205] CMS Collaboration, "Estimating Systematic Errors Due to Pileup Modeling", `https://twiki.cern.ch/twiki/bin/view/CMS/PileupSystematicErrors`, March 2015.

[206] CMS Collaboration, "Measurement of the cross section ratio $\sigma_{\mathrm{t\bar{t}b\bar{b}}}/\sigma_{\mathrm{t\bar{t}jj}}$ in pp collisions at $\sqrt{s} = 8$TeV", `arXiv:1411.5621` (2014).

[207] CDF and DØ Collaborations, "Tevatron Constraints on Models of the Higgs Boson with Exotic Spin and Parity Using Decays to Bottom-Antibottom Quark Pairs", `arXiv:1502.00967` (2015).

[208] CMS Collaboration, "VH($\mathrm{b\bar{b}}$) projections for Run2", `https://indico.cern.ch/event/370334/session/22/contribution/39`, March 2015.

[209] N. Faltermann, "Search for Standard Model Higgs boson production in association with a single top quark with the CMS experiment", Diploma Thesis, Karlsruhe Institute of Technology, IEKP-KA/2015-02 (2015).

[210] CMS Collaboration, "Standard Model Cross Sections for CMS at 13 TeV", `https://twiki.cern.ch/twiki/bin/view/CMS/StandardModelCrossSectionsat13TeV`, March 2015.

[211] LHC Higgs Cross Section Working Group, "SM Higgs production cross sections at $\sqrt{s} = 13/14\,\text{TeV}$", `https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CERNYellowReportPageAt1314TeV`, March 2015.

Danksagung

Eine Doktorarbeit ist ohne vielfältige Unterstützung von außen nicht möglich. Deshalb möchte ich zum Abschluss noch Worte des Dankes an die Personen aussprechen, die mich in dieser herausfordernden, aber auch sehr lohnenden Phase meines Lebens begleitet haben.

An erster Stelle möchte ich Herrn Prof. Dr. Thomas Müller für die tollen Jahre in seiner Arbeitsgruppe danken. Ich konnte seiner Unterstützung stets sicher sein, zum Beispiel wenn es um die Fürsprache für Konferenzen im In- und Ausland ging. Der einjährige Forschungsaufenthalt am CERN, den Herr Müller mir ermöglichte, war eines der Höhepunkte meiner Zeit als Doktorand, und dafür bedanke ich mich besonders. Auf der anderen Seite gab er mir genug Freiraum um mich frei zu entwickeln, und eigene Richtungen einzuschlagen.

Herrn Prof. Dr. Ulrich Husemann gilt mein Dank für die Übernahme des Korreferats. Über die gesamten drei Jahre meiner Promotion half er durch konstruktive Kritik die Analysen zu verbessern. Auch habe ich von seinen hilfreichen Kommentaren zur schriftlichen Arbeit profitiert.

Meinem Betreuer Dr. Thorsten Chwalek danke ich ganz herzlich für die vielen Jahre, in der er mir unermüdlich mit Rat und Tat zur Seite stand. Er wusste stets, wann der richtige Zeitpunkt zur Motivation oder zur konstruktiven Kritik ist. Vor allem fand er immer den richtigen Ton, und trug mit seinen Ideen zur Verbesserung der Analysen bei. Vielen Dank dafür.

Meinen Kollegen Dr. Hauke Held, Simon Fink und Benedikt Maier möchte ich ganz herzlich für die wunderbare Zusammenarbeit danken. Nur durch die fruchtbaren Diskussionen war es möglich die Analysen und das Software-Framework stetig zu verbessern. Auch wenn sich der Review-Prozess in die Länge gezogen hat und man sich im Stress auch manchmal auf die Nerven geht: Es hat richtig Spaß gemacht mit euch zu arbeiten.

Ich möchte auch der gesamten Arbeitsgruppe von Herrn Prof. Dr. Thomas Müller einen Dank aussprechen. Besonders hervorheben möchte ich Dr. Mathias Mozer, Dr. Jeannine Wagner-Kuhr, Steffen Röcker und Frank Roscher, die bei allen Fragen rund um die Analysen und darüber hinaus immer die passenden Antworten fanden. Auch Ehemalige wie Dr. Jochen Ott haben aus der Ferne mit Ihrem Fachwissen bei unlösbar erscheinenden Problemen geholfen. In dieser Weise war die hohe Qualität

in der Gruppe immer gewährleistet.

Das gute Klima auch außerhalb der Arbeitsgruppe hat dazu beigetragen, dass ich mich am EKP und in der gesamten Fakultät für Physik immer gut aufgehoben fühlte. Egal, zu welchem Fachgebiet Fragen auftauchten, man fand immer einen Ansprechpartner, der sich die Zeit nahm mir weiter zu helfen. Das war vor allem bei der Vorbereitung für die Doktorprüfung Gold wert. Daher möchte ich mich ganz besonders bei Johannes Bellm, Felix Frensch und Fabian Harms bedanken.

Das Admin-Team, ehemals unter der Leitung von Prof. Dr. Thomas Kuhr und nun geleitet von Dr. Thorsten Chwalek, macht einen super Job am Institut. Den Doktoranden, die neben ihren eigenen Analysen gewährleisten, dass die Infrastruktur am EKP läuft, gebührt der größte Respekt und ein großes Dankeschön. Ein besonderer Dank geht auch an Frau Bräunling und Frau Hühn, die bei der bürokratischen und organisatorischen Fragen immer freundlich und hilfsbereit zur Stelle waren.

Meinen Geschwistern mit Anhang, meinen Schwiegereltern und besonders meiner Mutter möchte ich für die Unterstützung während der gesamten Promotion danken. Es war wichtig, das Vertrauen in meine Arbeit auch außerhalb der Physik zu spüren und aufmunternde Worte zu hören. Meinen Freunden möchte ich danken, dass ich bei ihnen Ablenkung und Trost fand, wann immer ich es brauchte.

Meiner Frau Christina gebührt der größte Dank: Du bist einfach die Beste und ich konnte auch in den stressigsten Phasen immer auf dich zählen. Danke, dass du mich zum glücklichsten Mann der Welt machst!