IEKP-KA/2016-30

Karlsruhe Institute of Technology

Search for Higgs-Boson Production in Association with a Top-Quark Pair in the Boosted Regime 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. Shawn Darrell Williamson
aus Heilbronn

Tag der mündlichen Prüfung: 11.11.2016

Referent: Prof. Dr. Ulrich Husemann
Korreferent: Prof. Dr. Thomas Müller

Introduction

The classical elements fire, water, earth, air, and aether were proposed in ancient Greece, about 500 A.D.. Similar ideas arose in ancient China, Egypt, Babylonia, Japan, Tibet, and India. All of these pre-scientific concepts pursue one goal, the description of nature. Throughout history this goal has served mankind as drive for the search of the underlying rules of processes observed. A milestone developed long after the concepts of ancient eras is represented by the description of macroscopic processes by Newtonian mechanics. From there, science has moved "backwards in time" describing ever more elementary processes and constituents of matter. Via the discovery of atoms now summarized in the periodic system of elements, Rutherford's observation of the infinitesimally small atomic nucleus, and the detection of protons and neutrons as constituents of the nucleus, science approaches the description of the conditions present at the origin of everything we know, the big bang. Today, the most elementary particles and the most fundamental interactions are described by the Standard Model of particle physics.

Until 2012 one major part of the Standard Model remained unobserved, the Higgs boson. The discovery of this particle by the ATLAS [1] and CMS [2] experiments at the LHC confirms the Higgs mechanism, which was introduced to the Standard Model to assign mass to particles. However, the analysis of the Higgs boson has not stopped with its discovery. The Standard Model predicts production and decay modes of this particle that remain unobserved and properties that remain unmeasured. A production mode not discovered to date is the Higgs-boson production in association with a top-quark pair ($t\bar{t}H$). It comes with a special feature, the direct access to the top-Higgs Yukawa coupling, which characterizes the coupling of the Higgs boson to the top quark. As the strength of the coupling of the Higgs boson to all particles depends on their masses, the coupling to the top quark is especially strong. As a consequence, this coupling causes large perturbative corrections in the calculation of the Higgs-boson mass, which are linked to the hierarchy problem. The investigation of the coupling of the Higgs boson to the top quark does not come for free. The small cross section of $t\bar{t}H$ production is only one of the reasons that make the search for this process a challenging task. Especially in the search for $t\bar{t}H$ production with a Higgs-boson decay into a bottom-quark pair ($t\bar{t}(H\rightarrow b\bar{b})$), one faces large background contributions by top-quark pair ($t\bar{t}$) production and a combinatorial problem in event reconstruction. The cross section of $t\bar{t}$ production is about 1600 times larger than the one of the signal process. Further, the additional radiation of a gluon splitting makes this process an irreducible background, as it provides exactly the same final-state configuration as $t\bar{t}(H\rightarrow b\bar{b})$ production. The combinatorial problem in event reconstruction is caused by the large number of decay products in the final state. The reconstruction of events requires the assignment of reconstructed objects to the expected decay products, which is highly ambiguous in case of $t\bar{t}(H\rightarrow b\bar{b})$ production. The issues mentioned above represent some of the reasons, why $t\bar{t}H$ searches have not reached a sensitivity sufficient for the observation or exclusion of Standard Model $t\bar{t}H$ production yet. A solution to the combinatorial problem in event reconstruction is provided by the investigation of a phase space including massive particles with large transverse momenta, the boosted regime [3]. Typical use cases for this approach are searches for hypothetical particles decaying into massive particles,

like top-quarks, with large transverse momenta. In comparison, $t\bar{t}H$ production features top quarks and Higgs bosons which are only moderately boosted. Still, the background processes encountered in the analysis tend to feature softer particles, which leads to a reduction of background. The main motivation for this approach, however, is still given by the simplified reconstruction of $t\bar{t}(H\to b\bar{b})$ events. Boosted massive particles pass their momentum to their decay products, which emerge from the decays as collimated bunches of particles. Dedicated algorithms are specialized in the reconstruction and identification of these signatures.

This thesis presents the implementation of a boosted-regime analysis in the search for $t\bar{t}(H\to b\bar{b})$ production with a semileptonic decay of the top-quark pair. This is accomplished by introducing a new analysis category that targets signatures with one boosted hadronically decaying top quark and one boosted Higgs boson. From a selection of dedicated boosted reconstruction and identification methods the ones performing best for the massive particles in $t\bar{t}(H\to b\bar{b})$ events are chosen. Accordingly, the boosted top quark is reconstructed based on the HEPTopTaggerV2 algorithm [4, 5] and identified with a multivariate classification specifically developed for this analysis. The boosted Higgs boson is reconstructed with the BDRS algorithm [6] and identified using subjet b-tagging information. The event reconstruction and selection is based on the reconstructed boosted objects and optimized with respect to the selection of $t\bar{t}(H\to b\bar{b})$ signal events and the rejection of $t\bar{t}$ background. The boosted analysis category is added to a set of analysis categories investigating the non-boosted phase space. The final discrimination of signal against background is performed with a combination of boosted decision trees and the matrix element method. In the boosted analysis category, the information provided by the boosted event reconstruction is inputted into matrix element method. The semileptonic $t\bar{t}(H\to b\bar{b})$ search presented in this thesis is performed based on the first data recorded by the CMS experiment at a center-of-mass energy of $\sqrt{s} = 13$ TeV. The recorded dataset corresponds to an integrated luminosity of $\mathcal{L} = 2.7\ \text{fb}^{-1}$. The analysis has been combined with the dilepton $t\bar{t}(H\to b\bar{b})$ search and published in spring 2016 [7]. It is not only the first $t\bar{t}(H\to b\bar{b})$ analysis at a center-of-mass energy of $\sqrt{s} = 13$ TeV, but also the first $t\bar{t}H$ search specifically investigating the boosted regime. Accordingly, this analysis represents a showcase for the application of boosted analysis techniques in a moderately boosted regime and busy particle-collision events.

The first part of this thesis introduces the search for $t\bar{t}(H\to b\bar{b})$ production by covering the fundamentals necessary for understanding and performing the analysis. Chapter 1 provides a theoretical introduction to the Standard Model in general and more specifically to the physics of the top quark, the Higgs boson, and the $t\bar{t}H$ process. The proton-proton collisions analyzed in this thesis have been provided by the LHC and recorded by the CMS experiment. An overview of the experimental setup is given in Chapter 2. Next to the data recorded, predictions of signal and background processes are necessary for performing the analysis presented in this thesis. Both types of data and the procedures for obtaining predictions of signal and background processes are described in Chapter 3. Raw detector signals are not directly accessible by the analysis presented in this thesis. For this reason, several high-level analysis objects are reconstructed based on the signals provided by the detector. These objects and the corresponding reconstruction procedures are introduced in Chapter 4. Analysis techniques used throughout the $t\bar{t}(H\to b\bar{b})$ search are described in Chapter 5. These techniques include boosted decision trees (BDT), which are, for example, used for the final discrimination of signal against background, and the statistical methods applied for the determination of the final results. The search for

$t\bar{t}(H\to b\bar{b})$ production with a semileptonic decay of the top-quark pair is described in the second part of this thesis. It starts by covering the part of the analysis investigating the non-boosted phase space, which in the following will be referred to as resolved phase space. Chapter 6 introduces the selection and reconstruction of resolved events based on the reconstructed analysis objects. Further, this chapter covers the categorization of the selected events. In Chapter 7, various boosted analysis techniques are introduced. Furthermore, the reconstruction and identification of the boosted hadronically decaying top quark and the boosted Higgs boson decaying into a bottom-quark pair based on these methods are described. The reconstruction and selection of events featuring one boosted top quark and one boosted Higgs boson is described in Chapter 8. Additionally, the combination of the boosted analysis category defined by the selected boosted events with the resolved analysis categories is presented. The final discrimination of signal against background in every analysis category is performed based on a combined approach using BDTs and the matrix element method (MEM). A detailed description of the training of the BDTs, the calculation of the MEM discriminant, and the combination of both methods is given in Chapter 9. The analysis presented in this thesis includes various sources of uncertainties. An overview of the ones considered for the results is presented in Chapter 10. The final results of the semileptonic $t\bar{t}(H\to b\bar{b})$ search are presented in Chapter 11. Furthermore, the combination with the dilepton $t\bar{t}(H\to b\bar{b})$ search and all other $t\bar{t}H$ search channels performed by the CMS collaboration based on the first 13 TeV data is discussed. The results are compared to the ones provided by the analyses performed at a center-of-mass energy of $\sqrt{s} = 8$ TeV and the ones obtained by the analyses performed by the ATLAS collaboration. Prospects for future $t\bar{t}(H\to b\bar{b})$ searches are presented in Chapter 12. This chapter includes a projection of the $t\bar{t}(H\to b\bar{b})$ search presented in this thesis for larger integrated luminosities and a study introducing two new analysis categories targeting signatures with single boosted massive particles. Further, some considerations on systematic changes of the boosted analysis approach in future iterations of the $t\bar{t}(H\to b\bar{b})$ search are discussed. The conclusion in Chapter 13 gives a short recap of the analysis presented in this thesis.

Contents

Part I.

Theoretical and Technical Foundations

Chapter 1

Theoretical Foundations

The basic concept of physics is the discovery of the underlying rules of processes occurring in nature. These rules are described by theories, which are developed based on experimental observations and theoretical concepts. Numerous experiments test these theories for their general validity. Discrepancies with observations hint at a limited validity or a failure of a theory. In the former case, the regime where one theory fails is to be described by another. An example is given by Newtonian mechanics, which successfully describes macroscopic effects at velocities that are low compared to the speed of light. Nevertheless, it fails at the description of subatomic processes, where quantum mechanics takes over.

The most elementary processes and particles known to date are described by the Standard Model of particle physics (SM). An overview of the SM is provided in Section 1.1. The SM consists of set of theories that provide very accurate predictions of the observations in particle-physics experiments. An example are the cross sections of particle-physics processes, which are proportional to the probability of the occurrence of a particular process given a particular initial state. The calculation of this quantity is outlined in Section 1.2. The validity of the SM has been probed by countless particle-physics experiments. Until today the SM has held up with great success. Very popular objects of investigation today are the top quark and the Higgs boson. These two particles represent the last two massive SM particles that have been discovered. A thorough investigation of their properties provides not only a good test of the SM but also a gateway for the discovery of new physics. However, not all of the properties, production, and decay modes of the Higgs boson have been measured yet. A very interesting production mode is investigated by the analysis described in this thesis, the Higgs-boson production in association with a top-quark pair. A special characteristic of this process is the direct access to the top-Higgs Yukawa coupling. However, the small cross section and the overwhelming background turn the search for this process into an enormous challenge. In Section 1.3 an outline of the physics of the top quark and the Higgs boson is presented. Further, a more detailed discussion of the Higgs-boson production in association with a top-quark pair can be found in the same section.

1.1. Standard Model

The Standard Model of particle physics yields an accurate description of the most fundamental particles and forces known to date. It combines theories describing the electromagnetic, weak, and strong interaction based on the framework provided by quantum field theory [8]. The SM took on its current form by the unification of the electromagnetic and the weak interaction and the introduction of the Higgs mechanism in the 1970s. Yet, crucial components of the SM have been developed years before. Ever since, the SM has withstood the probing by countless particle-physics experiments with great success. This

success story has been crowned by the discovery of predicted particles, like the top quark by the CDF [9] and D0 [10] experiments at the Tevatron in 1995 and the Higgs-boson by the ATLAS [1] and CMS [2] experiments at the LHC in 2012.

Nevertheless, there are observations that are not described by the SM. One of the most prominent examples is gravity. The most accurate description of gravity to date is provided by general relativity, which is inconsistent with the SM. It is possible to add a description of gravity to the SM by introducing a corresponding mediator particle, the graviton. However, this description could not be verified by observations so far. A further prominent observation not described by the SM is given by dark matter and dark energy. Based on gravitational effects, such as the structure and motion of galaxies, known matter has been estimated to make up only about five percent of all energy in the universe. The rest is given by unknown energy contributions referred to as dark matter and dark energy. The SM provides no suitable candidates for the description of these effects. A last example is provided by the observed flavor oscillations of neutrinos. These oscillations require the neutrinos to carry a finite mass. In the SM, neutrinos are expected to be massless. In addition to the observations not described by the SM, there are some theoretical issues. One of them is the hierarchy problem expressed by huge quantum corrections to the masses of particles. Another theoretical issue is the large number of undetermined parameters in the SM. These limitations and problems of the SM leave space for new theories, so-called new physics. Some theoretical concepts that solve the issues stated above are given by supersymmetry or the introduction of extra dimensions. However, none of these new-physics theories could be experimentally verified so far.

As already mentioned, the SM includes the description of the most fundamental particles and forces known to date. The particles are divided into two classes based on their spin quantum number. Bosons are particles with integer spin in units of the reduced Planck constant $\hbar$. The most elementary bosons described by the SM are represented by the mediator particles of the different forces and the Higgs boson. The second class of particles is given by the fermions, which carry half-integer spin in units of $\hbar$. For simplicity, the convention $c = \hbar = 1$ is applied in the upcoming subsections.

1.1.1. Bosons and Interactions

Bosons follow Bose-Einstein statistics, which implies that an infinite number of these particles can take on the same quantum state. One type of bosons described by the SM is given by vector bosons, which feature a spin of one in units of $\hbar$. These particles, also referred to as gauge bosons, represent the mediators of the forces.

As SM is formulated in the framework of relativistic quantum field theory, the particles and forces are described by fields. The kinematics of these fields and the interactions between them are specified by Lagrangian densities $\mathcal{L}$. The application of the Euler-Lagrange equation,

$$\partial_\mu \left(\frac{\partial \mathcal{L}}{\partial(\partial_\mu \phi)} \right) - \frac{\mathcal{L}}{\partial \phi} \,,$$

with the respective Lagrangian densities results in the equations of motion for a given field ϕ. A special characteristic of the SM is local gauge invariance. This feature implies that the Lagrangian density of the SM is invariant under transformations by operations of the $SU(3)_\mathrm{c} \times SU(2)_\mathrm{L} \times U(1)_\mathrm{Y}$ symmetry group. According to Noether's theorem [11] every continuous symmetry brings the conservation of a certain quantity. In case of the SM, the

$SU(3)_c$ symmetry yields the conservation of the color charge, whereas the $SU(2)_\mathrm{L} \times U(1)_\mathrm{Y}$ symmetry comes with the conservation of the weak isospin I_3 and the weak hypercharge Y.

As the gauge bosons are the mediators of the interaction of the different interactions described by the SM, they are described along with them in the following.

Electromagnetic Interaction

The electromagnetic interaction is mediated by the photon, which couples to the electric charge of other particles. The photon itself carries no electric charge causing the absence of self-interaction. This feature, together with the photon being massless, explains the long range of the electromagnetic force.

A formulation of the electromagnetic force as relativistic quantum field theory is provided by quantum electrodynamics (QED) [12–17], which is an Abelian gauge theory with $U(1)$ symmetry group. The Lagrangian density of QED is given by

$$\mathcal{L}_\mathrm{QED} = i\bar{\psi}\gamma^\mu\partial_\mu\psi - m\bar{\psi}\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} - e\bar{\psi}\gamma_\mu A^\mu\psi \; . \tag{1.1}$$

In this equation, γ_μ are the Dirac matrices. Spin-1/2 particles are represented by their bispinor field ψ, whereas its Dirac adjoint $\bar{\psi}$ is the Hermitian adjoint of ψ in combination with the Dirac matrix γ^0, $\bar{\psi} = \psi^\dagger\gamma^0$. The parameter m denotes the mass of the spin-1/2 particle. The potential of the photon field is represented by A^μ and the electromagnetic field strength tensor is given by $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. Each term in Eq. (1.1) describes different aspects of particles and their interactions. The first term determines the kinematics of spin-1/2 particles, whereas the second describes their mass. The third term determines the kinematic properties of the photon field. The last term describes the interaction between the electromagnetic field and spin-1/2 particles.

Applying the QED Lagrangian density in the Euler-Lagrange equation for the fermion field ψ brings

$$i\gamma^\mu\partial_\mu\psi - m\psi = e\gamma_\mu A^\mu\psi \; .$$

The left-hand side of this equation represents the Dirac equation. The right-hand side describes the interaction of the spin-1/2 particle with the photon field. Applying the QED Lagrangian density in the Euler-Lagrange equation for the photon field A^μ results in the Maxwell equations,

$$\partial_\mu F^{\mu\nu} = e\bar{\psi}\gamma^\mu\psi \; .$$

Strong Interaction

The strong interaction is mediated by gluons, which couple to the the color charge. There are three different color charges defined, red, green, and blue. Gluons themselves carry superpositions of color and anti-color adding up to eight different configurations. The resulting effective color charge of gluons causes self-interaction. This characteristic leads to a short range of the strong interaction even though gluons are massless. A further consequence of the self-interaction of gluons is the confinement observed for the strong

interaction. This effect implies that no single strongly interacting particles can be observed. Instead, these particles are always bound to form color-neutral states, so-called hadrons. When separating two strongly interacting particles, the energy stored in the field between them increases linearly with the distance. If the energy is large enough, a new quark-antiquark pair is produced, which respectively form separate bound states with the initial particles. In case the initial particles are produced at large energies, this procedure is repeated numerous times forming a collimated shower of hadrons flying in the same directions as the initial particles. These showers of particles are referred to as jets. The counterpart to confinement is asymptotic freedom. At large energy scales or short distances, the running coupling of the strong interaction becomes asymptotically small. As a result, strongly interacting particles are only loosely bound and perturbation theory is valid.

A formulation of the strong force as relativistic quantum field theory is provided by quantum chromodynamics (QCD) [18–21], which is a non-Abelian gauge theory with an $SU(3)$ symmetry group. The Lagrangian density of QCD is given by

$$\mathcal{L}_{\mathrm{QCD}} = i\bar{\psi}_a \gamma^\mu \partial_\mu \delta_{ab} \psi_b - m\delta_{ab}\bar{\psi}_a\psi_b - \frac{1}{4} G^A_{\mu\nu} G^{A\mu\nu} + g\bar{\psi}_a \gamma_\mu t^C_{ab} \psi_b A^C_\mu \ . \tag{1.2}$$

It shows a similar structure as the QED Lagrangian density. The quark fields are described by the bispinor fields ψ_a with a color index a that runs over all three colors. Its Dirac adjoint $\bar{\psi}_a$ is defined as above. The eight different gluon fields, corresponding to the eight different color configurations, are represented by A^C_μ. The gluon field strength tensor is constructed from the gluon fields according to $G^A_{\mu\nu} = \partial_\mu A^A_\nu - \partial_\nu A^A_\mu + g f^{ABC} A^B_\mu A^C_\nu$. In this equation, f^{ABC} are the structure constants of the $SU(3)$ symmetry group. The 3×3 matrices t^C_{ab} represent the eight generators of the $SU(3)$ symmetry group. Further, g denotes the quark-gluon coupling constant. Components not described are identical to Eq. (1.1). Again, each term in Eq. (1.2) describes a different aspect of particles and their interactions. The kinematics and masses of quarks are described by the first two terms. The third term describes the kinematics of the gluon field, which include the self-interaction. The interaction of gluons with quarks is described by the last term.

The equations of motions can be derived by the application of the Euler-Lagrange equation. The concept has already been demonstrated for QED and is therefore omitted here. A remarkable difference to QED is the term caused by the self-interaction in the equations of motion of the gluon field.

Electroweak Interaction

A central part of the SM is the electroweak unification [22–28]. It achieves that the electromagnetic interaction and the weak interaction are no longer described by two separate theories but by a single one. This is accomplished by a gauge theory with $SU(2) \times U(1)$ symmetry group. Accordingly, three massless boson fields, W_1, W_2, and W_3, which couple to the weak isospin I_3, are introduced. Additionally, a single boson field B is introduced, which couples to the weak hypercharge Y. In a nutshell, these fields correspond to four massless particles. However, this configuration of independent fields is only given above a certain energy scale. Below this energy scale, electroweak symmetry is spontaneously broken by the Higgs mechanism described in Section 1.1.3. Electroweak symmetry breaking causes the original fields to mix. This produces a new set of particles, which are the

well-known mediator particles of the electromagnetic and the weak interaction. The electrically neutral bosons, the photon and the Z boson, are formed by a mixing of the fields B and W_3 according to the Weinberg angle θ_W. The mathematical expression is given by

$$\begin{pmatrix} \gamma \\ Z^0 \end{pmatrix} = \begin{pmatrix} \cos\theta_\mathrm{W} & \sin\theta_\mathrm{W} \\ -\sin\theta_\mathrm{W} & \cos\theta_\mathrm{W} \end{pmatrix} \begin{pmatrix} B \\ W_3 \end{pmatrix} .$$

The photon and the Z-boson couple to a linear combination of the weak isospin and the weak hypercharge. Accordingly, the electric charge is given by the Gell-Mann-Nishijima formula $Q = Y/2 + I_3$. The mixing of the fields W_1 and W_2 generates the electrically charged W bosons according to

$$W^\pm = \frac{1}{\sqrt{2}}(W_1 \mp iW_2) .$$

A special characteristic of the W bosons is the exclusive coupling to left-handed particles and right-handed anti-particles. As shown in Table 1.1, W and Z bosons are very massive particles, which causes the very short range of the weak interaction. The masses of the W and Z bosons are introduced by the Higgs mechanism.

Interactions including a W boson require a flavor transition. Charged leptons are converted to neutrinos, up-type quarks are converted to down-type quarks and vice versa. This effect becomes apparent if one considers the charge of the W boson. However, in case of quarks the transition does not necessarily need to happen within the same generation, as it is the case for leptons. This is due to the fact that the quark flavor eigenstates of the weak interaction are not identical to the mass eigenstates. The mixing of the states is expressed by the CKM matrix [29, 30],

$$\begin{pmatrix} \mathrm{d}' \\ \mathrm{s}' \\ \mathrm{b}' \end{pmatrix} = \begin{pmatrix} V_\mathrm{ud} & V_\mathrm{us} & V_\mathrm{ub} \\ V_\mathrm{cd} & V_\mathrm{cs} & V_\mathrm{cb} \\ V_\mathrm{td} & V_\mathrm{ts} & V_\mathrm{tb} \end{pmatrix} \begin{pmatrix} \mathrm{d} \\ \mathrm{s} \\ \mathrm{b} \end{pmatrix} ,$$

where the q' represent the electroweak eigenstates of quarks and the q the mass eigenstates. The currently best measurement of the absolute values of the CKM matrix elements [31] is given by

$$V_\mathrm{CKM} = \begin{pmatrix} 0.97427 \pm 0.00014 & 0.22536 \pm 0.00061 & 0.00355 \pm 0.00015 \\ 0.22522 \pm 0.00061 & 0.97343 \pm 0.00015 & 0.0414 \pm 0.00012 \\ 0.00886^{+0.00032}_{-0.00033} & 0.0405^{+0.0011}_{-0.0012} & 0.99914 \pm 0.00005 \end{pmatrix} .$$

In the calculation of the transition amplitude, a CKM matrix element is included corresponding to the flavors of the quarks contributing to the vertex. Due to the large diagonal elements of the CKM matrix, transitions between the same generation are favored. Transitions between generations are suppressed according to the smaller off-diagonal elements.

The description of the Lagrangian density of the electroweak interaction is omitted here. The concept is similar to the ones present for QED and QCD. The electroweak Lagrangian density includes kinetic terms describing the dynamics of the fermionic and bosonic fields. Additional terms describe the interaction between these fields. Further, there are terms describing the self-interaction and the interaction among the bosonic fields.

Table 1.1: Standard Model bosons and their properties. Next to their names and symbols, the electric charge, the spin, and the mass of these particles are listed. The bosons are grouped into gauge bosons and the Higgs boson. The stated mass of the Higgs-boson results from a combination of the measurements performed by the ALTAS and the CMS collaborations. Taken from [31].

Particle	Symbol	Electric charge [e]	Spin	Mass [GeV/c^2]
Photon	γ	0	1	0
Gluon	g	0	1	0
W boson	$W^{\pm}$	± 1	1	80.385 ± 0.002
Z boson	Z^0	0	1	91.188 ± 0.002
Higgs boson	H	0	0	125.09 ± 0.24

Higgs Boson

The Higgs boson is the second most massive particle in the SM. It is a scalar boson and accordingly carries a spin of zero in units of $\hbar$. It is not a gauge boson as it does not directly result from gauge invariance. For this reason, it is not considered a mediator particle of any interaction field. This particle results directly from the Higgs mechanism, which is based on electroweak symmetry breaking. The Higgs boson couples to the mass of particles. More details on the Higgs mechanism and the Higgs boson are given in Section 1.1.3 and Section 1.3.2.

A summary of all elementary bosons and their properties is given in Table 1.1.

1.1.2. Fermions

Fermions are particles that carry half-integer spin in units of $\hbar$. They follow Fermi-Dirac statistics and accordingly the Pauli exclusion principle. This principle states that a particular quantum state can only be occupied by exactly one fermion. Fermions are subdivided into particles that carry color charge and interact via the strong forces and the ones that do not. The former are referred to as quarks and the latter as leptons. Each of both categories includes six different particles, which are grouped into three so-called generations or families. Within each generation, two particles with different electric charge can be found. Left-handed particles of a generation form a weak isospin doublet with $I_3 = +1/2$ and $I_3 = -1/2$. Right-handed particles, on the other side, represent weak isospin singlets with $I_3 = 0$. Moving from one to the next generation, typically the masses of corresponding particles increase. An exception is given by neutrinos, for which the exact masses are unknown to date. The increase in mass between the different generations causes the particles associated to higher generations to decay into particles associated to lower generations. Accordingly, only first generation particles are stable and therefore mainly observed in nature. Additionally, each fermion features an anti-particle with opposite charge-type quantum numbers.

Leptons are further subdivided into charged leptons and neutrinos. The former are represented by electrons, muons, and tau leptons. These particles feature an electric charge of $Q = -e$ and interact via the electromagnetic and the weak interaction. Correspondingly, electron neutrinos, muon neutrinos, and tau neutrinos represent the neutrinos, which

Table 1.2: Standard Model leptons and their properties. The leptons are grouped into the different generations. Next to their names and symbols, the electric charge, the weak isospin, and the mass of all leptons are listed. The weak isospin is stated for left-handed (L) and right-handed (R) leptons. Neutrino measurements allow only upper limits on their mass. Taken from [31].

Particle	Symbol	Electric charge [e]	Weak isospin (L/R)	Mass [MeV/c^2]
Electron	e	-1	$-1/2$ / 0	0.511
Electron neutrino	ν_e	0	$+1/2$ / 0	$< 2 \cdot 10^{-6}$
Muon	μ	-1	$-1/2$ / 0	105.7
Muon neutrino	ν_μ	0	$+1/2$ / 0	$< 2 \cdot 10^{-6}$
Tau	τ	-1	$-1/2$ / 0	1776.82 ± 0.16
Tau neutrino	ν_τ	0	$+1/2$ / 0	$< 2 \cdot 10^{-6}$

carry no electric charge. Hence, they only interact via the weak interaction. The weak interaction only couples to left-handed neutrinos. Accordingly, no right-handed neutrinos have been observed so far. A summary of all leptons and their properties is given in Table 1.2.

Quarks are further subdivided into up-type quarks and down-type quarks. Up-type quarks include up quarks, charm quarks, and top quarks and feature an electric charge of $Q = 2/3\,e$. Down-type quarks include down quarks, strange quarks, and bottom quarks and feature an electric charge of $Q = -1/3\,e$. Quarks interact via all forces described by the SM. Due to the special properties of the strong interaction, quarks cannot be observed as single particles. Instead, they only occur in color-neutral bound systems referred to as hadrons. There are two kinds of hadrons mainly observed, the mesons and the baryons. Mesons are bound systems of a quark and an antiquark carrying a given color and the corresponding anti-color. As they consist of two fermions, the spin of mesons add up to an integer value in units of $\hbar$. Accordingly, mesons are bosons. Examples of mesons are pions, which consist of combinations of up and down quarks. Baryons are bound states of three quarks, each featuring a different color. Their spins add up to an half-integer value in units of $\hbar$. Accordingly, they are fermions. Examples of baryons are the nucleons, the proton and the neutron, which are also composed of up and down quarks. Further, bound states of strongly interacting particles are also predicted by the SM, however most of them have not been observed yet. An exotic bound state of quarks, the pentaquark, has been recently discovered by the LHCb collaboration [32].

1.1.3. Higgs Mechanism

The combination of the Lagrangian densities of QCD and the electroweak interaction, in order to form the SM Lagrangian density, does not include the mass terms presented in Eq. (1.1) and Eq. (1.2). A consideration of these terms would violate gauge invariance. A decomposition into the chirality states,

$$-m\bar{\psi}\psi = -m(\bar{\psi}_\mathrm{L}\psi_\mathrm{R} + \bar{\psi}_\mathrm{R}\psi_\mathrm{L}) \,, \tag{1.3}$$

makes this apparent. Left-handed fermions ψ_L, which form isospin doublets, and right-handed fermions ψ_R, which form isospin singlets, transform differently under $SU(2)\times U(1)$

Table 1.3: Standard Model quarks and their properties. The quarks are grouped into the different generations. Next to their names and symbols, the electric charge, the weak isospin, and the mass of all quarks are listed. The weak isospin is stated for left-handed (L) and right-handed (R) quarks. Taken from [31].

Particle	Symbol	Electric charge [e]	Weak isospin (L/R)	Mass [MeV/c^2]
Up	u	$+2/3$	$+1/2$ / 0	$2.3^{+0.7}_{-0.5}$
Down	d	$-1/3$	$-1/2$ / 0	$4.8^{+0.7}_{-0.3}$
Charm	c	$+2/3$	$+1/2$ / 0	$(1.275 \pm 0.025) \cdot 10^3$
Strange	s	$-1/3$	$-1/2$ / 0	95 ± 5
Top	t	$+2/3$	$+1/2$ / 0	$(173.21 \pm 0.51 \pm 0.71) \cdot 10^3$
Bottom	b	$-1/3$	$-1/2$ / 0	$(4.18 \pm 0.03) \cdot 10^3$

symmetry operations. Hence, Eq. (1.3) is not invariant under such operations. Terms of this form are omitted in the SM Lagrangian density and particles are predicted to be massless. However, this contradicts observations, which indicate that many of the particles do have mass. Particles like the top quark or W and Z bosons, for example, are indeed very massive. In order to still account for the masses of particles observed, the Higgs mechanism [33–35] is introduced, which relies on the spontaneous breaking of electroweak symmetry.

The Higgs mechanism starts by introducing a scalar field ϕ, which couples to the masses of particles, the Higgs field. This field is represented by a complex $SU(2)$ doublet and accordingly features four real parameters. Additionally, a corresponding potential is introduced, the so-called Mexican hat potential,

$$V(\phi) = -\mu\phi^*\phi + \lambda(\phi^*\phi)^2 \, . \tag{1.4}$$

As long as the parameters μ and λ are positive, this potential features a local maximum at $\phi = 0$ and global minima at the vacuum expectation value $|\phi| = v = \sqrt{\mu/(2\lambda)}$. An illustration of the Higgs potential reduced to two dimensions is displayed in Fig. 1.1. At large energies, the Higgs field takes on a value of $\phi = 0$ and symmetry is conserved. At a low energy scale, the Higgs field falls into a global minimum characterized by the vacuum expectation value. Due to the choice of the global minimum out of the infinite number of possibilities featuring the same $|\phi|$, the symmetry is broken.

In the following, the concept of the Higgs mechanism is explained based on a simplified Higgs field $\phi = \phi_1 + i\phi_2$ and a massless gauge field A^μ. The explanation is based on the one provided by [37]. Given the Higgs potential described by Eq. (1.4), the corresponding Lagrangian density can be formulated as

$$\mathcal{L} = \frac{1}{2}\left[(\partial_\mu - iqA_\mu)\,\phi^*\right]\left[(\partial^\mu - iqA^\mu)\,\phi\right] + \mu\phi^*\phi - \lambda(\phi^*\phi)^2 - \frac{1}{16\pi}F^{\mu\nu}F_{\mu\nu} \, . \tag{1.5}$$

In this equation, $F^{\mu\nu}$ denotes the field strength tensor of the gauge field A^μ. A parameter transformation is performed describing the field from the location of a chosen global minimum. This is done by introducing $\phi_1 = \eta - \sqrt{\mu/(2\lambda)}$ and $\phi_2 = \xi$. Applying this in Eq. (1.5) provides terms describing a massless Goldstone boson for the field ξ and some

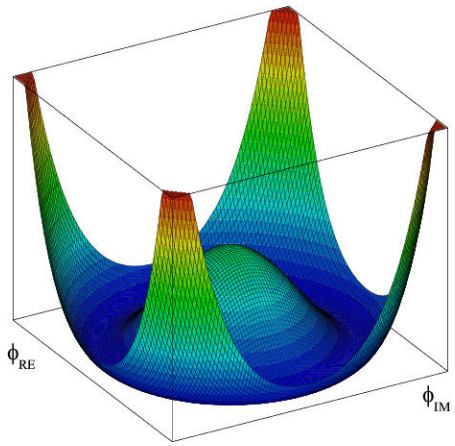


Figure 1.1: Higgs potential illustrated in two dimensions. The potential is displayed for fixed positive values of λ and μ. Taken from [36].

couplings of the field ξ to the gauge field A^μ. This is in accordance with the Goldstone theorem [38–40], which predicts one massless boson for each broken symmetry. Gauge invariance allows to completely eliminate the contributions of ξ by applying a proper transformation. Doing this results in the following Lagrangian density:

$$\begin{aligned}\mathcal{L} = &\left[\frac{1}{2}(\partial_\mu\eta)(\partial^\mu\eta) - 2\mu\eta^2\right] + \left[-\frac{1}{16\pi}F^{\mu\nu}F_{\mu\nu} + \frac{\mu}{4\lambda}A_\mu A^\mu\right] \\ &+ \left[\sqrt{\frac{\mu}{2\lambda}}q^2 A_\mu A^\mu + \frac{1}{2}q^2\eta^2 A_\mu A^\mu - \sqrt{8\lambda\mu}\eta^3 - \lambda\eta^4\right] \\ &+ \frac{\mu^2}{16\lambda}\end{aligned} \tag{1.6}$$

In this Lagrangian density, several features can be observed. First of all, the second contribution enclosed in brackets, which describes the kinematics and dynamics of the gauge field, includes a mass term. The first contribution in brackets shows terms for a new massive spin-zero particle, the Higgs boson. The terms in the second row of Eq. (1.6) describe the interaction of the Higgs field with the gauge field, which can be parametrized by the coupling strength,

$$\lambda_\mathrm{V} = 2\frac{m_\mathrm{V}^2}{v}\ ,$$

and the self-interaction of the Higgs field.

In the SM, the Z and W bosons obtain a finite mass when applying an analogue procedure as described above. However, the fermions are not yet accounted for and remain massless. In order to change this, the interaction of the fermions with the Higgs field has to be described. This is accomplished by introducing terms to the Lagrangian density that transform like a singlet under operations of the $SU(2)\times U(1)$ symmetry group. Such terms are formulated by the Yukawa interaction,

$$\mathcal{L}_\mathrm{Yukawa} = -\lambda_\mathrm{f}(\bar{\psi}_\mathrm{L}\phi\psi_\mathrm{R} + \bar{\psi}_\mathrm{R}\bar{\phi}\psi_\mathrm{L})\ .$$

In this equation, the Yukawa coupling is given by

$$\lambda_{\mathrm{f}} = \frac{\sqrt{2} m_{\mathrm{f}}}{v} \, .$$

Next to describing the interaction between the Higgs field and the fermionic field, the terms cause the fermions to acquire mass for a non-zero vacuum expectation value. The Yukawa coupling describing the coupling of the Higgs boson to the top quark is the most interesting feature of the $\mathrm{t\bar{t}H}$ production, which is the main objective of this thesis.

1.2. Cross-Section Calculation

The probability for the occurrence of a particular particle-physics processes given a particular initial state is proportional to its cross section. The calculation of this quantity is based on Fermi's golden rule [41,42], which quantifies the rate for a transition of an initial state $|i\rangle$ into a set of final states $\langle f|$,

$$\Gamma_{i \to f} = \frac{2\pi}{\hbar} |\langle f|H'|i\rangle|^2 \rho \, . \tag{1.7}$$

The transition is characterized the Hamiltonian H', which represents the particular physics process. The term $\langle f|H'|i\rangle$ is the corresponding transition-matrix element. The expression ρ represents the density of final states. For particle collisions, the transition rate is closely related to the cross section of particular particle-physics processes. In case of a proton-proton collider, the initial state is given by the incoming protons and their momentum four-vectors. Accordingly, the final state is provided by the produced particles with their corresponding momentum four-vectors. The energy scales of the QCD processes involved in the matrix element range from low scales of the bound hadron states up to large scales of the energy transfer between the interacting partons. This also causes the running strong coupling constant α_{s} to vary between a large range. Too large values of the strong coupling constant corresponding to low energy scales lead to a break-down of perturbation theory, which is applied for the calculation of QCD processes. A solution is the factorization approach, which splits the full process into subprocesses occurring at different energy scales. Hence, the processes in the bound hadron are separated from the hard interaction of the partons. Following this approach, the cross section of a particular process taking place in the collision of two protons can be calculated as

$$\begin{aligned} \sigma(ab \to n; \mu_{\mathrm{R}}^2, \mu_{\mathrm{F}}^2) = \int \frac{\mathrm{d}x_a \mathrm{d}x_b}{2x_a x_b s} \int \prod_k^n \left(\frac{\mathrm{d}^3 \vec{p}_k}{(2\pi)^3 2E_k} \right) (2\pi)^4 \delta \left(p_a + p_b - \sum_k^8 p_k \right) \\ \times f_a(x_a, \mu_{\mathrm{F}}^2) f_b(x_b, \mu_{\mathrm{F}}^2) |\mathcal{M}(ab \to n; \mu_{\mathrm{R}}^2, \mu_{\mathrm{F}}^2)|^2 \, . \end{aligned} \tag{1.8}$$

In this equation, two partons of the two incoming protons a and b interact and produce the final state n. The partons bound in the incoming protons are described by the parton distribution functions (PDF) $f(x, Q^2)$. The PDFs give the probability for a certain Bjorken scaling variable value x, which can be interpreted as the fraction of proton-momentum carried by a parton given an energy scale Q. The energy scale determines the degree to which partons are resolved. A more detailed description of the PDFs is given in Section 1.2.1. The

integration is performed over all possible momentum fractions of the incoming partons x_a and x_b. The energy scale Q is given by the factorization scale μ_F. It determines the energy scale below which initial-state radiation cannot be resolved anymore and is absorbed into the PDFs. The hard interaction between the two partons, also denoted as hard process, is accompanied by large momentum transfers. For this case, the strong coupling constant is small and perturbation theory is valid. In Eq. (1.8), this interaction is represented by the matrix element $\mathcal{M}(ab \to n; \mu_\mathrm{R}^2, \mu_\mathrm{F}^2)$ describing the transition probability from the initial partons a and b to the final state n given a certain process. The matrix element depends on the factorization scale μ_F^2 and renormalization scale μ_R^2. The renormalization scale μ_R^2 determines the absorption of high-energy physics processes into the strong coupling constant and accordingly the value of the running coupling $\alpha_\mathrm{s}(Q^2) = \alpha_\mathrm{s}(\mu_\mathrm{R}^2)$. A typical choice is $\mu_\mathrm{R} = \mu_\mathrm{F} = Q$, where Q is the energy scale of the hard process. The calculation of the matrix element is described in more detail in Section 1.2.2. Further integrations in Eq. (1.8) are performed over the considered phase space. The phase space is a multidimensional hyperspace spanned by the four-vectors of all final-state particles. In Eq. (1.8), the density of final states ρ appearing in Fermi's Golden Law (Eq. (1.7)) is represented by the Lorentz-invariant phase space expressed by the terms behind the product sign. The δ-function accounts for energy and momentum conservation for incoming and outgoing particles.

1.2.1. Parton Distribution Functions

Partons are the building bricks of hadrons and composed of quarks, antiquarks, and gluons. Partons bound inside nucleons are described by the parton distribution functions (PDF) [43], which are parametrized by the Bjorken scaling variable x and the energy scale Q. In the limit of large nucleon momenta, which is quite true for the protons accelerated by the LHC, x can be interpreted as the momentum fraction of the nucleon carried by the respective parton. The dependence on the energy scale can be explained by the resolution of radiation and virtual production of quarks and gluons at higher energy scales. The PDFs give the probability for a particular type of parton to be observed with a particular momentum fraction at a given energy scale. Different types of partons are distinguished as gluons and the quarks and antiquarks of different flavors show a different behavior in the nucleon.

The energy scale dependence of the PDFs can be calculated with perturbative QCD in a regime with $\alpha_\mathrm{s}(Q^2) \ll 1$. These calculations are represented by the DGLAP equations [44–46]. The x dependence at a given Q^2 is not predicted. Consequently, this dependence is extracted from measured cross sections of particular processes, which are linked to the PDFs via the factorization approach previously described. The procedure for deriving the PDFs starts with the parametrization of the x dependence at a low value of the energy scale. The resulting function is evolved in Q^2 by applying the DGLAP equations. The 10 to 30 free parameters are determined by a global fit of the functions to data measured for different points in the x-Q^2-space. The data used for this fit stems from deep inelastic scattering experiments, jet production in hadron collisions, dilepton production in hadron collisions, and vector-boson production in proton-antiproton collisions.

Sets of PDFs are provided by various groups. The PDFs applied for the predictions used in this thesis are mainly produced by the CT [47, 48], MSTW [49, 50], and NNPDF collaborations [51, 52]. The different sets differ in the parametrization of the PDFs, the fit procedure for extracting the PDFs, and the datasets used for the extraction. Additionally, PDF sets based on different values of the strong coupling constant, different orders of

perturbation theory, and different number of jet flavors are provided by each group.

There are two methods for the representation of the uncertainties on the PDF sets. One of them is given by the confidence interval provided by the fit, which is determined by the covariance matrix. The PDF uncertainties are represented by alternative sets of PDFs. These PDFs are varied with respect to the best-fit PDF according to the eigenvectors and eigenvalues of the Hessian matrix, which can be extracted from the covariance matrix. A second method for representing PDF uncertainties is based on a Monte Carlo approach. Random sets of pseudo data are generated based on the measured data used for the extraction of the PDFs, while taking into account the corresponding uncertainties. From these sets of pseudo data, alternative PDF sets are derived.

A common interface for all sets of PDFs is provided by the LHAPDF library [53]. A general set of recommendations for the usage of the PDFs and the evaluation of the corresponding uncertainties is provided by the LHC4PDF working group [54, 55].

1.2.2. Matrix Element

A collision event is mostly characterized by the hard process, which comes with a large momentum transfer. In the calculation of cross sections and the simulation of events, this process is represented by the matrix element. Its value squared is proportional to the probability of a transition from a initial state to a final state given a particular process. In proton-proton collisions, the initial state is given by two incoming partons. These partons interact via a QCD or electroweak process, which results in a final state with an varying number of particles.

The matrix element is determined with the help of perturbation theory. Different contributions are calculated as fixed orders of an expansion in the coupling constant. In case of QCD processes, this coupling constant is given by the strong coupling constant α_s. However, the value α_s depends on the momentum transfer at the respective vertex and can cause the breakdown of perturbation theory at too low energies. The contributions of different orders can be depicted by a set of Feynman diagrams, which give a schematic representation of the particles involved in the process and their interactions. The first contribution of the expansion denoted as leading order (LO) or tree level involves the most simple versions of the process' Feynman diagrams and provides a coarse estimation of the matrix element. The following orders denoted as next-to-leading order (NLO), next-to-next-to-leading order (NNLO), and so on (N^{x}LO) successively include real emissions and virtual loops in the Feynman diagrams. Consequently, the Feynman diagrams get more and more complicated at higher orders and their number per order increases drastically. Virtual loops introduce ultraviolet divergences in the limit of infinitely large energies. This effect is countered by renormalization, which includes the introduction of the renormalization scale μ_R^2. In this procedure, ultraviolet physics are absorbed into the the running coupling. Additionally, real emissions and virtual loops introduce singularities in the limit to infinitesimal soft energies. In an analytic calculation, both contributions typically cancel and lead to a finite result. This effect is described by the Kinoshita-Lee-Nauenberg (KLN) theorem [56, 57]. Further, the radiation and virtual loops cause divergences in the calculations. All of the mentioned go along with an increasing effort in the calculation of higher order perturbations. However, the contributions of higher orders represent corrections to the coarse estimate of the leading order calculation, which provide a more accurate description of the real process.

1.3. Top-Quark and Higgs-Boson Physics

The top quark and the Higgs boson are among the most recently discovered SM particles. Due to their large mass and their distinctive properties, they are of special interest to a large fraction of particle-physics analyses performed today. Most of the properties of the top quark are well known by now. An overview of them together with the production and decay modes of top quarks is presented in Section 1.3.1. Nevertheless, precision measurements still probe the predictions of the SM associated to this particle as discrepancies of observation and prediction may hint at contributions by new physics. With the discovery of the Higgs boson, the last missing piece of the SM has been found. An overview of its properties and the current experimental status is given in Section 1.3.2. Compared to the top quark, the entirety of all SM predictions associated to the Higgs boson, such as all properties, production, and decay modes, are not verified yet. An interesting production mode, which still remains to be discovered, is given by the Higgs-boson production in association with a top-quark pair ($t\bar{t}H$). A special characteristic of $t\bar{t}H$ production is the direct access to the coupling of the Higgs boson to the top quark. However, the observation of this process is complicated due to its small cross section and an overwhelming amount of background. A more detailed description of the $t\bar{t}H$ production process is given in Section 1.3.3. The search for $t\bar{t}H$ production represents the main subject of this thesis.

1.3.1. Top Quark

The top quark is the most massive particle in the SM. As already described in Section 1.1.2, it belongs to the quarks, the strongly interacting fermions described by the SM. Together with the bottom quark, it forms the third generation of quarks. Its large mass close to the scale of electroweak symmetry breaking makes it a popular object of investigation. This large mass further causes a very short life time, which is much smaller than the time scale of hadronization. Accordingly, the top quark is the only quark that decays before forming a color-neutral bound state.

Top-Quark Production

Due to its large mass, a large amount of energy is necessary to produce a top quark. There are two main production modes, the top-quark pair production and the single top-quark production.

The top-quark pair production ($t\bar{t}$) is a pure QCD process. This process can be initiated in two different ways, either by gluons or by a quark and an antiquark in the initial state. Both types of $t\bar{t}$ production are illustrated by Feynman diagrams in Fig. 1.2. At the LHC with a center-of-mass energy of $\sqrt{s} = 13$ TeV, about 90 % of the top quarks are produced via the gluon-initiated process. Although the top-quark pair production requires enough energy to produce two top quarks, it represents the main production mode at the LHC. This fact can be explained by the large coupling constant of the strong interaction.

Single top quarks are produced via the weak interaction. This process includes a vertex of a top quark, a W boson, and a down-type quark. The contribution of different down-type quarks to this vertex is determined by the corresponding CKM-matrix element described in Section 1.1.1. As the CKM-matrix element V_{tb} is close to one and the others negligibly small, the vertex includes a bottom quark in almost all cases. Correspondingly, the single top-quark production is well suited for the measurement of the CKM-matrix element V_{tb}. The single top-quark production is further subdivided into three production modes:

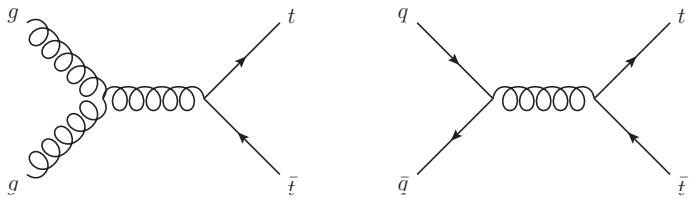


Figure 1.2: Example Feynman diagrams for the gluon-initiated (left) and the quark-initiated (right) production of a top-quark pair.

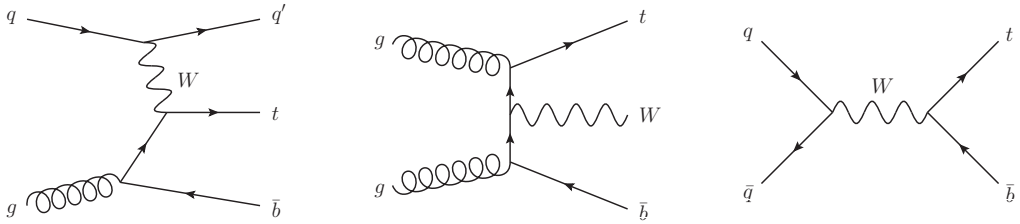


Figure 1.3: Example Feynman diagrams for the production of single top quarks via the t-channel (left), the associated production with a W boson (middle), and the s-channel (right).

- the t-channel,
- the associated production with a W boson (tW),
- and the s-channel.

They are ordered by their cross section at the LHC with the largest at the top. Example Feynman diagrams of the single top-quark production are given in Fig. 1.3. Single top-quark production features a cross section that is about five times smaller than top-quark pair production at a center-of-mass energy of $\sqrt{s} = 13$ TeV.

Top-Quark Decay

The only possibility for the decay of a top quark is via the weak interaction. Due to the large CKM-matrix element V_{tb}, it decays into a W boson and a bottom quark in almost all of the cases. One distinguishes between the leptonic and the hadronic decay of a top quark, which is characterized by the decay of the W boson. A leptonic decay of a top quark features a W-boson decay into a charged lepton and a neutrino. A hadronic decay of a top quark is indicated by a W-boson decay into an up-type and a down-type quark and antiquark. A decay of the W boson into a final state featuring a top-quark is not possible due to the large mass of the top quark. Accordingly, the hadronic W boson decay produces mainly quarks from the first and the second generation. Taking into account the three different color charges of quarks, the branching ratio for the hadronic top-quark decay occurs twice as often as the leptonic decay.

Transferring this categorization to the decay of a top-quark pair provides three different configurations:

- **Dileptonic $\mathbf{t\bar{t}}$ decay channel ($\mathbf{t\bar{t} \rightarrow l^{+}\nu bl'^{-}\bar{\nu}'\bar{b}}$):** Both top quarks decay leptonically. The dileptonic decay channel features a branching ratio of 10.5 %.

- **Semileptonic $t\bar{t}$ decay channel ($t\bar{t} \rightarrow l^+\nu bq\bar{q}'\bar{b}/q\bar{q}'bl^-\bar{\nu}\bar{b}$):** One top quark decays leptonically, while the other top quark decays hadronically. The semileptonic decay channel features a branching ratio of 43.8 %.

- **All-hadronic $t\bar{t}$ decay channel ($t\bar{t} \rightarrow q\bar{q}'bq''\bar{q}'''\bar{b}$):** Both top quarks decay hadronically. The all-hadronic decay channel features a branching ratio of 45.7 %.

1.3.2. Higgs Boson

The Higgs boson is a spin-zero particle, which results from the Higgs mechanism. Many of its properties are already covered in Section 1.1.1 and Section 1.1.3. Until its discovery in 2012, the Higgs boson has been the last missing piece of the SM. Searches for this particle strongly rely on the different production and decay modes of this particle.

Higgs-Boson Production

There are many ways for producing a Higgs boson at the LHC. In the following, the four most frequent ones are outlined.

The Higgs boson couples to the mass of particles. Nevertheless, the main production mode at the LHC, the gluon-gluon fusion (ggH), features gluons in the initial state. As the Higgs boson does not couple to massless particles, it is produced via intermediately generated particles in this process. Further, only quarks qualify as intermediate particles as gluons only couple to color charged particles. The largest contribution is provided by the top quark, due to its large mass and the resulting large coupling to the Higgs boson. The main reason for the comparably large cross section is the large number of gluons in a proton-proton collisions with an energy sufficient to enter this process. As introduced in Section 1.2.1, the probability distributions for a parton carrying a particular momentum are described by the parton distribution functions.

The second-largest Higgs-boson production mode is vector-boson fusion (VBF). This process starts with two quarks in the initial state, which produce virtual vector bosons. The vector bosons in turn produce a Higgs boson. The comparably large cross section can be explained by the large coupling of the Higgs boson to the vector bosons. A special characteristic of this process are the two outgoing quarks. The two quarks form two jets, which are directed in the forward direction of the detector. This special trait simplifies a targeted search for VBF.

A further production mode is the associated production of a Higgs boson with a vector boson (VH). This process is also known as Higgs-strahlung, which refers to bremsstrahlung as analogue process. In the VH process, a vector boson is produced by the annihilation of a quark and an antiquark. The Higgs boson is radiated by the vector boson. VH production is the Higgs-boson production mode with the third-largest cross section among all SM Higgs-boson production modes.

The associated production of a Higgs boson with a top-quark pair ($t\bar{t}H$) possesses the smallest cross section among the four main Higgs-boson production modes. In this process, a top-quark pair is produced as described in the previous section. The Higgs boson is radiated from one of the top quarks. Even though the coupling of the Higgs boson to the top quark is comparably strong, $t\bar{t}H$ production features a very small cross section. This is mainly due to the enormous amount of energy of about 500 GeV necessary to produce these three massive particles. A more detailed description of this process can be found in Section 1.3.3.

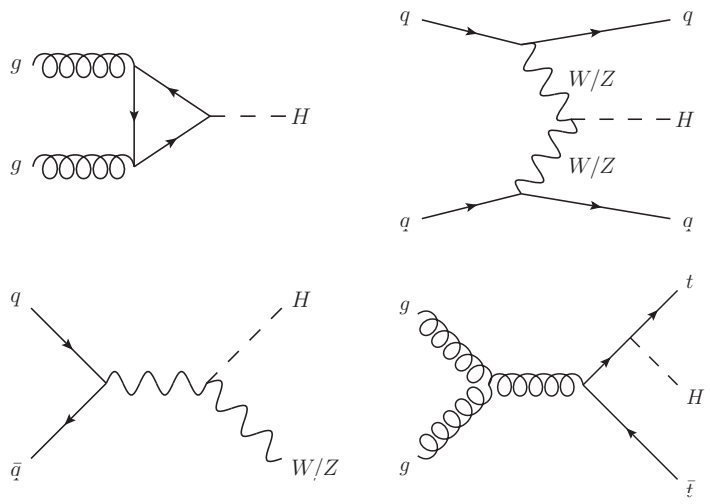


Figure 1.4: Example Feynman diagrams for the different Higgs-boson production modes: gluon-gluon fusion (top left), vector-boson fusion (top right), the associated production of a Higgs boson with a vector boson (bottom left), and the associated production of a Higgs boson with a top-quark pair (bottom right).

The cross sections of the different production channels are illustrated in Fig. 1.6. Fig. 1.4 shows example Feynman diagrams for each of the four major Higgs-boson production modes described.

Higgs-Boson Decay

The main factors determining the branching ratios of different Higgs-boson decays is the mass of the Higgs boson and the mass of the decay products. The mass of the Higgs boson provides the energy available for the decay products. Further, the coupling to the Higgs boson scales with the mass of the decay products. A decay into top quarks, for example, would be favored due to the coupling, but is not possible as the mass of two top quarks largely exceeds the mass of the Higgs boson. Instead, the by far largest branching ratio is provided by the Higgs-boson decay into two bottom quarks. This decay makes up almost 60 % of all Higgs-boson decays. However, due to the large background by QCD processes, a search for Higgs bosons decaying into a bottom-quark pair at the LHC is cumbersome. The second largest contribution with a branching ratio of about 20 % is given by the Higgs-boson decay into two W-bosons, where one W-boson is produced off-shell. In case of the W bosons decaying into leptons, this decay provides a very clean signature. One of the search channels mainly contributing to the Higgs-boson discovery in 2012 is based on the Higgs-boson decay into two Z bosons. If the Z bosons decay into charged leptons, this decay channel provides a very distinctive signature as there are hardly any backgrounds featuring four charged leptons. Due to the good momentum resolution of charged leptons, a very narrow Higgs-boson mass peak can be reconstructed in this search channel. Again, one of the bosons is produced off the mass shell as the invariant mass of two Z bosons exceeds the Higgs-boson mass. The second Higgs-boson decay mode with a major contribution to the Higgs-boson discovery is the decay into two photons. As for the gluons in the Higgs-boson production by gluon fusion, the massless photons do not couple to the Higgs boson directly. Instead, this decay proceeds via a loop.

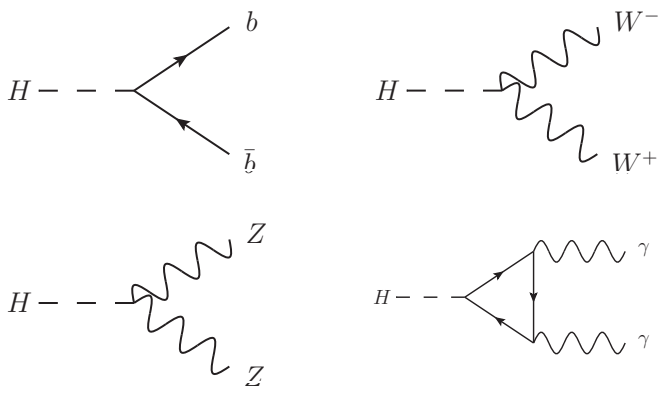


Figure 1.5: Example Feynman diagrams for different Higgs-boson decay modes. The Higgs-boson decays into bottom quarks (top left), W bosons (top right), Z bosons (bottom left), and photons (bottom right) are displayed.

Compared to gluon fusion, all electrically charged massive particles may contribute to the loop. Accordingly, a further major contribution is given by the W boson. The Higgs-boson decay into two photons also features a very clean final state with a very good Higgs-boson mass resolution. However, this decay channel has a very small branching ratio compared to the other Higgs-boson decay modes described.

Above, only a few important Higgs-boson decay modes are covered. Example Feynman diagrams of these decays are displayed in Fig. 1.5. However, there are further Higgs-boson decay modes, which will not be described in more detail. The branching ratios of many Higgs-boson decay modes as a function of the Higgs-boson mass are displayed in Fig. 1.6.

Higgs-Boson Measurements

In 2012, the Higgs boson has been independently observed by the ATLAS [1] and CMS [2] collaborations. For this discovery, various searches targeting different Higgs-boson decay channels have been combined. The convention for claiming an observation is the observation of a signal significance of five standard deviations. This corresponds to a probability of about one in 3.5 million to observe an identical excess caused by the background. Further, evidence is claimed at a signal significance of three standard deviations. Today, individual decay channels of the Higgs boson have reached an observed signal significance sufficient to claim observation [59]. In the CMS collaboration, for example, the Higgs-boson decays into two photons and two Z-bosons have been experimentally verified. This is not surprising, as these two channels have provided the largest contribution to the Higgs-boson discovery in 2012. On top of that, the Higgs-boson search targeting the decay into two W-bosons has achieved an observed signal significance of 4.8 σ, which is close to an observation. The search for a Higgs-boson decay into tau leptons achieves an observed signal significance sufficient to claim evidence. The search for Higgs-boson decays into two bottom quarks, on the other hand, only reaches a signal significance of 2.0 σ. The advantage brought by the large branching fraction in the search for this Higgs-boson decay is canceled by the enormous number of background events by QCD processes. A summary of the observed and expected signal significances for the searches performed in the different Higgs-boson decay channels by the CMS collaboration in LHC run I is presented in Table 1.4.

In 2012, the ATLAS and CMS collaboration did not claim the discovery of “the Higgs-

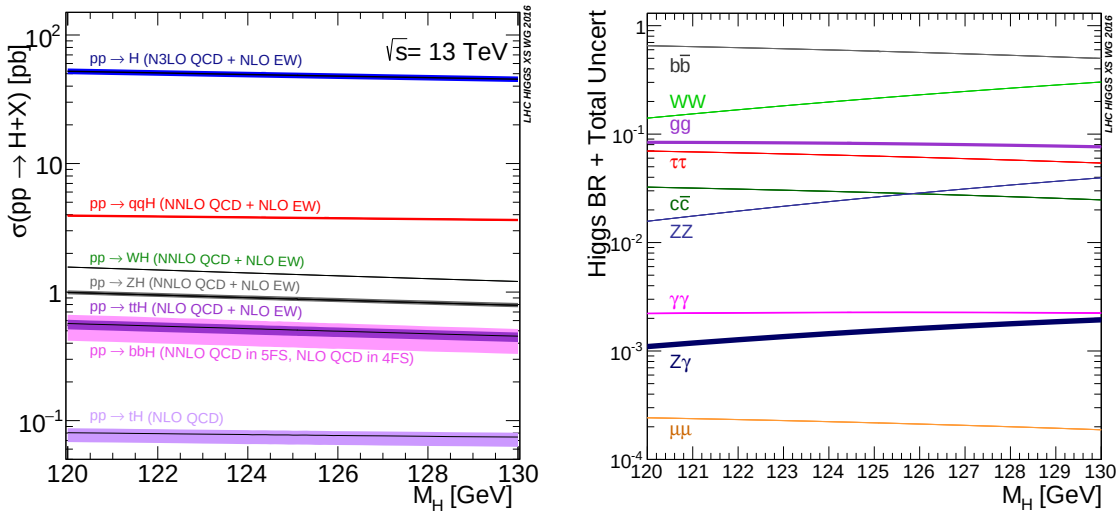


Figure 1.6: Production cross section of different Higgs-boson production channels (left) and branching ratios of different Higgs-boson-decay modes (right) both as a function of the Higgs-boson mass. The Higgs-boson production cross sections are displayed for proton-proton collisions at a center of mass energy of $\sqrt{s} = 13$ TeV. Taken from [58].

boson" but rather the discovery of "a Higgs-boson-like particle". In order to verify that the newly discovered particle is really the Higgs boson predicted by the SM, its properties are thoroughly tested. Today, the best measurement of the Higgs-boson mass is provided by the combined LHC-run-I measurements of the ATLAS and CMS collaborations. The mass found in this combination amounts to 125.09 ± 0.24 GeV/c^2 [60]. A scan of the test statistic of this measurement as function of the Higgs-boson mass is displayed in Fig. 1.7. Further, the spin and the parity of the Higgs boson have been found to match the SM prediction well. A spin of one in units of $\hbar$ is excluded by the discovery of the decay into two photons. According to the Landau-Yang theorem [61] a decay of a spin-one particle into two photons is prohibited. Various spin-two configurations have, for example, been probed by the CMS collaboration [62]. However, the SM configuration with a spin of zero is still favored over all other scenarios tested. The couplings of the Higgs boson to fermions and vector bosons have also been analyzed. A combined measurement of ATLAS

Table 1.4: Observed and expected signal significances for Higgs-boson searches targeting different decay channels performed in the CMS collaboration in LHC run I. Taken from [59].

Decay channel	Signal significance [σ]	
	Observed	Expected
H→ γγ	5.6	5.1
H→ZZ	7.0	6.8
H→WW	4.8	5.6
H→ ττ	3.4	3.7
H→bb	2.0	2.5

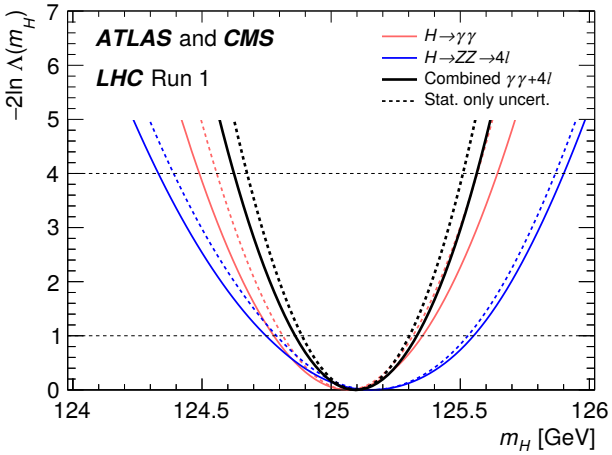


Figure 1.7: Scans of twice the negative logarithmic likelihood ratio as function of the Higgs-boson mass for the combined measurements of ATLAS and CMS in the H→ $\gamma\gamma$ channel (red), the H→ZZ channel (blue) and a combination of both (black) in LHC run I. The measurements corresponding to the dashed curves account only for the statistical uncertainties, while the nuisance parameters corresponding to the systematic uncertainties are fixed to their best fit value. The ± 1 and ± 2 standard deviation intervals are given by the intersections with the dashed lines at test statistic values of 1 and 4, respectively. Taken from [60].

and CMS based on the LHC-run-I data is shown in Fig. 1.8. In this study, the couplings to fermions and bosons have been parametrized and fit to observation. The obtained results are very well in agreement with the SM expectations.

1.3.3. Associated Production of a Higgs-Boson with a Top-Quark Pair

The analysis described in this thesis represents a search for Higgs-boson production in association with a top-quark pair ($t\bar{t}H$). At leading order in perturbation theory, the main contribution of the $t\bar{t}H$ process is the production of a high-energy top-quark pair, which radiates a Higgs boson. A minor contribution is the annihilation of two top quarks producing a Higgs boson. Both processes are illustrated by the Feynman diagrams shown in Fig. 1.9. The most accurate calculation of the cross-section of $t\bar{t}H$ production to date includes NLO QCD corrections [63–67] and NLO electro-weak corrections [68–70].

A special characteristic of this process is the interaction of the top quark and the Higgs boson. This interaction is described by the Yukawa-interaction terms introduced in Section 1.1.3. As it scales with the mass of the fermion, the corresponding top-Higgs Yukawa coupling,

$$\lambda_t = \frac{\sqrt{2} m_t}{v} \approx 1 \, ,$$

is the largest Yukawa coupling among all fermions. This coupling also contributes in the Higgs-boson production by gluon-gluon fusion and the Higgs boson decay into two photons. In these processes, the top quark may occur inside of a virtual loop, where also other particles may occur. However, due to its strong coupling to the Higgs boson, the

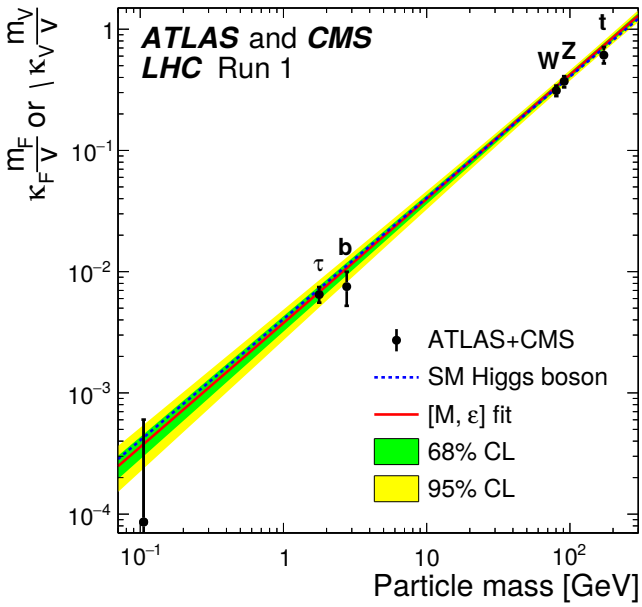


Figure 1.8: Best fit values of a fit of the Higgs coupling to different SM particles performed for the combination of the measurements by the ATLAS and the CMS collaboration as a function of the particle mass. Corresponding to the couplings, the y-axis shows $\kappa_\mathrm{f} m_\mathrm{f}/v$ for fermions and $\sqrt{\kappa_\mathrm{V}} m_\mathrm{V}/v$ for bosons. κ_f and κ_V are strength modifiers of the respective couplings and v represents the Higgs vacuum expectation value. The blue dashed line shows the prediction by the SM. The solid red line represents the best-fit values. The green and yellow areas show the confidence intervals corresponding to one and two standard deviations, respectively. Taken from [60].

top quark provides the largest contribution. In $\mathrm{t\bar{t}H}$ production, on the other hand, the top-Higgs Yukawa coupling is directly accessible. The large effect of the top-Higgs Yukawa coupling is, for example, demonstrated in the calculation of the Higgs-boson mass. There, it leads to large perturbative corrections due to contributions including a top quark. In the SM, these corrections can be remedied by renormalization. However, in expectation of new physics, these corrections at the order of the Planck scale have to be canceled by the contributions of new physics. A candidate achieving exactly this is supersymmetry. However, the parameter space of supersymmetry not excluded so far moves to ever higher energy scales. Based on the data provided by LHC run I, for example, masses of the top squark, which is relevant for the cancellations of the large corrections to the Higgs-boson mass caused by the top quark, smaller than 800 GeV could be ruled out for light neutralinos with a mass up to about 250 GeV [71]. Accordingly, the corrections to be canceled become very large compared to the observed Higgs-boson mass. The cancellation of such large corrections is considered "unnatural" and leads to the issue known as the hierarchy problem. The masses of the gauge bosons and fermions do not face such issues as they are protected from large corrections by gauge invariance and chiral symmetry, respectively.

The top-Higgs Yukawa coupling represents a gateway to new physics. Possible deviations from its predicted value can be described by an extended Higgs sector. Corresponding new theories are, for example, given by little Higgs models [72, 73] or composite Higgs

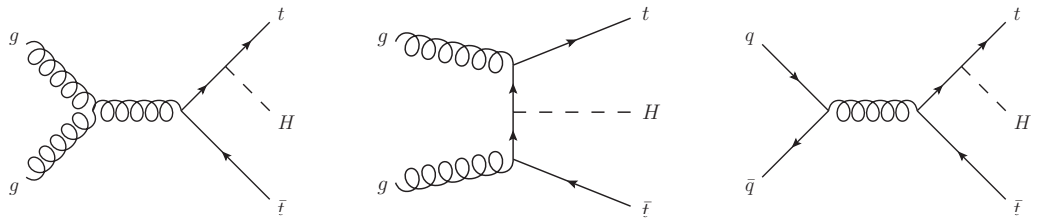


Figure 1.9: Example Feynman diagrams for two different $t\bar{t}H$ production modes. On the left side, the production of a top-quark pair with the radiation of a Higgs boson is shown. On the left side, $t\bar{t}H$ production by the annihilation of two top quarks is shown.

models [74]. A direct measurement of the top-Higgs Yukawa coupling by searching for $t\bar{t}H$ production represents an important test for the SM and whatever lies beyond.

However, the top-Higgs Yukawa coupling does not represent the only motivation for the search for $t\bar{t}H$ production. There are further new physics phenomena, which can alter the observed cross section of $t\bar{t}H$ production. An example is provided by hypothetical heavy top-quark partners, such as vector-like top quarks [75]. Vector-like fermions are hypothetical particles, for which both chiral states transform in the same way under the symmetry groups $SU(3)_c \times SU(2)_L \times U(1)_Y$. Accordingly, a charged current may involve right-handed vector-like fermions. A vector-like top quark partner can for example be produced via another hypothetical particle, the Z', in association with a SM top-quark. In case of a decay of the vector-like top-quark partner into an SM top quark and a Higgs boson, the same signature as for SM $t\bar{t}H$ production is expected.

$t\bar{t}H$ Search Channels

The search for $t\bar{t}H$ production is typically carried out in different search channels. Due to the different signatures expected, each search channels targets different decay modes of the Higgs boson. Some of the most popular and most sensitive $t\bar{t}H$ search channels are described in the following.

A typical $t\bar{t}H$ search channel targets Higgs-boson decays into two photons. As the photons are very well distinguishable from the jets and leptons usually produced by the decays of the top quarks, the reconstruction of the Higgs boson in this channel is unambiguous. As described in Section 1.3.2, this decay channel features a clean signature and a very good Higgs-boson mass resolution. Requirements on additional jets and leptons provide a handle on the top-quark part of this process. However, the small branching fraction of the Higgs-boson decay into photons and the low cross section of $t\bar{t}H$ production lead to very small event yields.

The multilepton search channel targets various Higgs-boson decay channels that feature leptons in their final state. This channel covers Higgs-boson decays into W bosons, Z bosons, tau leptons, and some more with smaller cross sections. In analyses based on this search channel, events featuring either two charged leptons with the same sign of the electric charge or at least three charged leptons are selected. The production of two same-sign leptons is only possible via a charge-independent production mechanism. In case of $t\bar{t}H$ production, this is provided if one lepton originates from the decay of the top quark and the other one from the decay of the Higgs boson. When introducing additional requirements on the jets in the event, there are hardly any background processes featuring

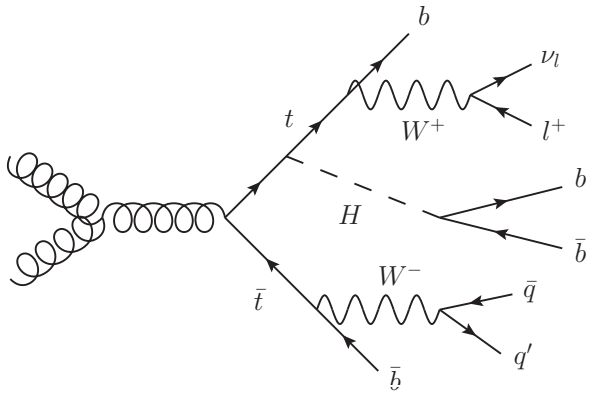


Figure 1.10: Example Feynman diagram for $t\bar{t}(H\to b\bar{b})$ production with a semileptonic decay of the top-quark pair.

a suitable signature to be selected. In the multilepton $t\bar{t}H$ search, the main background is given by misidentified leptons and incorrectly determined electric charges.

The third major search channel targets Higgs-boson decays into hadrons. This search channel includes Higgs-boson decays into bottom-quark pairs and Higgs-boson decays into tau leptons. The main advantage of this search channel is the large branching ratio. However, due to the hadronic final state, the search for these decays faces a large amount of background by top-quark pair production. Additionally, the assignment of observed jets to the decay products of the Higgs boson is highly ambiguous. This effect leads to a combinatorial problem in the event reconstruction.

The analysis presented in this thesis, targets $t\bar{t}H$ events with a Higgs-boson decay into a bottom-quark pair ($t\bar{t}(H\to b\bar{b})$) and a semileptonic decay of the top-quark pair. An example Feynman diagram of this process is displayed in Fig. 1.10. The particles expected in the final state are

- one prompt charged lepton from the leptonic W-boson decay,
- one neutrino from the leptonic W-boson decay,
- one bottom quark from the leptonic W-boson decay,
- one bottom quark from the hadronic W-boson decay,
- two light quarks from the hadronic W-boson decay,
- and two bottom quarks from the Higgs-boson decay.

This huge number of final-state particles characterizes $t\bar{t}(H\to b\bar{b})$ production and leads to a very "busy" event. The requirement of a prompt lepton in the event selection, which is expected to stem from the decay of the leptonically decaying top, achieves to reject a large fraction of background events from pure QCD processes. Yet, only electrons and muons are considered, as tau leptons are complicated reconstruct and have a less distinctive detector signature than electrons or muons. Further, an event trigger based on tau leptons is difficult to implement and less efficient.

$t\bar{t}(H\to b\bar{b})$ Search Backgrounds

The main background in the search for $t\bar{t}(H\to b\bar{b})$ production is given by $t\bar{t}$ production, whose cross section is about 1600 times larger. The signature of $t\bar{t}$ production is already very similar to the one of $t\bar{t}(H\to b\bar{b})$ production due to the presence of the top-quark pair in both processes. However, the large energy available in the proton-proton collisions at the LHC favor the radiation of additional gluons. These particles lead to the production of additional jets in the event, which promote large multiplicities of jets in the event as they are expected for $t\bar{t}(H\to b\bar{b})$ production. In case of a gluon splitting into a bottom-quark pair, $t\bar{t}$ production even provides an identical final state. This irreducible background contribution can only be discriminated from the signal process by applying information on the Higgs-boson, like its mass. However, as already mentioned, the Higgs-boson reconstruction in this search channel is complicated and accordingly information on the Higgs boson is hard to grasp.

Further, minor backgrounds are unlikely to produce final states that resemble the one of $t\bar{t}(H\to b\bar{b})$ production, feature a small cross section, or both. These backgrounds are given by

- single top-quark production (single top),
- vector-boson production in association with a top-quark pair ($t\bar{t}W$ and $t\bar{t}Z$),
- vector-boson production in association with jets (W+Jets and Z+jets),
- and diboson production (WW, WZ, and ZZ).

$t\bar{t}H$ Search Results

Searches for $t\bar{t}H$ production have already been performed by the ATLAS and CMS collaborations based on the data from the first run of the LHC at a center-of-mass energy of $\sqrt{s} = 8$ TeV. These analyses have not reached a sensitivity sufficient to observe or exclude SM $t\bar{t}H$ production. For this reason, upper limits on the signal strength of $t\bar{t}H$ production have been set. In the CMS collaboration, two separate searches for $t\bar{t}(H\to b\bar{b})$ production have been performed. The most distinctive feature of these analyses is the choice of the discriminant used for the final discrimination of signal against background.

One of the $t\bar{t}(H\to b\bar{b})$ searches relies on the discriminant provided by the matrix element method (MEM) [76]. The observed and expected upper limits on the signal-strength modifier $\mu(t\bar{t}H)$ obtained by this analysis are given by

$$95\,\%\ \mathrm{CL_s}\ \text{observed (expected) upper limit on}\ \mu(t\bar{t}H)\text{:}\ 4.2\,(3.3^{+1.6}_{-1.0})\ .$$

This result corresponds to the exclusion of a $t\bar{t}H$ cross section larger than 4.2 times the prediction provided by the SM at a confidence level of 95 %. The statistical determination and interpretation of such results is outlined in Section 5.2 and Chapter 11.

The other $t\bar{t}(H\to b\bar{b})$ search is based on a machine-learning approach [77, 78]. The observed and expected upper limits on the signal strength of $t\bar{t}H$ production obtained by this approach are

$$95\,\%\ \mathrm{CL_s}\ \text{observed (expected) upper limit on}\ \mu(t\bar{t}H)\text{:}\ 4.1\,(3.5^{+1.5}_{-1.0})\ .$$

The results of both $t\bar{t}(H\to b\bar{b})$ searches are very similar and exclude $t\bar{t}H$ production cross sections larger than 4.1-4.2 times the one predicted by the SM at a 95 % confidence level. In both cases, the observed limit exceeds the one expected. The latter analysis has been combined with the remaining $t\bar{t}H$ searches performed in the CMS collaboration at a center-of-mass energy of $\sqrt{s} = 8$ TeV [78]. The limits obtained by the combination of all search channels are

$$95\,\%\ \mathrm{CL_s}\ \text{observed (expected) upper limit on}\ \mu(t\bar{t}H)\text{: } 4.5\ (1.7^{+0.8}_{-0.5})\ .$$

This result corresponds to the exclusion of a $t\bar{t}H$ production cross section larger than 4.5 times the one predicted by the SM at a 95 % confidence level. The observed limit obtained by the combination is less restrictive than the one found for the $t\bar{t}(H\to b\bar{b})$ analyses alone. The expected upper limit, on the other hand, is much more stringent excluding $t\bar{t}H$ production cross sections larger than 1.7 times the one predicted by the SM. A large part of this discrepancy is explained by the excess that was found in the same-sign dimuon subchannel of the multilepton analysis.

A combination of all $t\bar{t}H$ searches performed by the ATLAS collaboration [79] has provided similar results

$$95\,\%\ \mathrm{CL_s}\ \text{observed (expected) upper limit on}\ \mu(t\bar{t}H)\text{: } 3.1\ (1.4^{+0.6}_{-0.4})\ .$$

The results obtained by the ATLAS collaboration also feature observed limits, which are weaker than the ones expected. As for the results provided by the CMS collaboration, this discrepancy is caused by slight upwards fluctuations in the multilepton search channel.

For the second data-taking run of the LHC, the center-of-mass energy has been increased to $\sqrt{s} = 13$ TeV. This increase in energy provides a $t\bar{t}H$ production cross section 3.9 larger than the one at a center-of-mass energy of $\sqrt{s} = 8$ TeV. This gain is larger than for all other main production modes of the Higgs-boson. The described behavior is displayed in Fig. 1.11, where the cross sections of the different Higgs-boson production modes are displayed as a function of the center-of-mass energy. The large increase of the $t\bar{t}H$ production cross section is mainly caused by gluons in the initial state of $t\bar{t}H$ production and the large invariant mass of the $t\bar{t}H$ system. By raising the center-of-mass energy, a larger portion of the gluons in the protons carry enough energy to create a top-quark pair and a Higgs boson in a collision. This behavior is determined by the parton distribution function, which describe the momentum fraction of the proton carried by the different partons. Based on the parton distribution functions, the parton-parton luminosities differential in the invariant mass of the interacting parton-parton system can be deduced. The ratio of these parton-parton luminosities for center-of-mass energies of $\sqrt{s} = 13$ TeV and $\sqrt{s} = 8$ TeV is shown in Fig. 1.12. In this distribution, a steep rise with increasing invariant mass of the parton-parton system can be observed. Accordingly, the cross section of $t\bar{t}$ production, which typically features smaller invariant masses than $t\bar{t}H$ production, only increases by a factor of 3.3 when moving from a center-of-mass energy of $\sqrt{s} = 13$ TeV to $\sqrt{s} = 8$ TeV.

Results of $t\bar{t}H$ searches based on the first data at a center-of-mass energy of $\sqrt{s} = 13$ TeV are provided by the ATLAS and CMS collaborations. The analysis searching for $t\bar{t}(H\to b\bar{b})$ production with a semileptonic decay of the top-quark pair presented in this thesis represents one of the $t\bar{t}H$ searches performed by the CMS collaboration. The results are presented in Chapter 11.

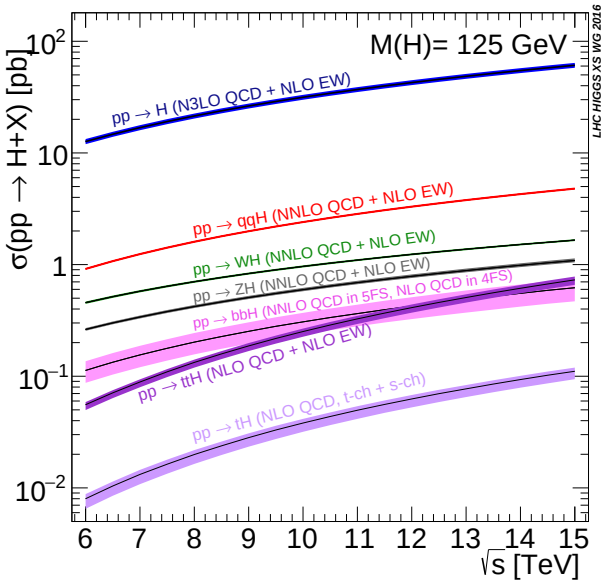


Figure 1.11: Cross section of different Higgs-boson production channels as a function of the center-of-mass energy. The Higgs-boson production cross sections are displayed for a Higgs-boson mass of $m_{\mathrm{H}} = 125\ \mathrm{GeV}/c^2$. Taken from [58].

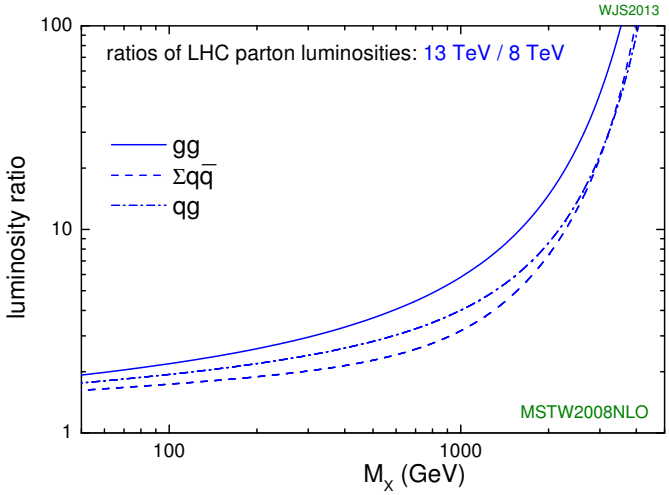


Figure 1.12: Ratio of parton-parton luminosities for a center-of-mass energy of $\sqrt{s} = 13$ TeV and a center-of-mass energy of $\sqrt{s} = 8$ TeV as a function of the invariant mass of the parton-parton system. The ratios are displayed for an initial state of two gluons, two quarks, and a quark and a gluon. Taken from [80].

Chapter 2

Experiment

The study of the most elementary particles known to mankind has proceeded to regions only accessible with huge amounts of energy. By concentrating these energies on single particles and making them collide, processes and states are induced that hardly occur in nature. Huge experimental installations are necessary to accelerate particles to such energies and to detect the outcome of the collisions.

The European Organization for Nuclear Research (CERN) is the world's largest laboratory dedicated to the exploration of particle physics. It accommodates the largest particle accelerator on earth, the Large Hadron Collider (LHC). In its almost 27 km long circular tunnel, protons are accelerated up to a center-of-mass energy of $\sqrt{s} = 13$ TeV, which is the highest energy ever achieved by a man-made accelerator. Further details on the LHC can be found in Section 2.1. The LHC provides proton-proton collisions for four main experiments. One of them is the Compact Muon Solenoid (CMS), which recorded the data analyzed in this thesis. Its onion-like structure combines different subdetectors measuring different aspects of the particles arising from the proton-proton collisions. Based on these measurements, the underlying process giving rise to these particles can be analyzed. A more detailed description of the CMS experiment is provided in Section 2.2. The processes occurring during particle collisions underlie quantum mechanics and are therefore of purely probabilistic nature. For this reason, a huge quantity of particle collisions are recorded, in order to investigate the underlying rules. The frequency of a process occurring within a set of recorded collision events is determined by its cross section and the integrated luminosity associated to this sample. The cross section is provided by theory, whereas the luminosity only depends on machine parameters. The definition of this quantity and its measurement as an interplay between the LHC and the CMS detector are outlined in Section 2.3.

2.1. Large Hadron Collider

The Large Hadron Collider (LHC) [81] provides the most energetic particle collisions under laboratory conditions. It is a synchrotron, a circular accelerator, housed in a 26.7 km long tunnel 50 to 175 meters below the ground. The LHC is situated beneath the franco-swiss border area in the north-west of Geneva, Switzerland. Before the LHC was built, the same tunnel has accommodated the Large Electron-Positron Collider (LEP), which was shut down in the year 2000.

The LHC collides protons or heavy ions. Beams composed of spatially separated bunches of these particles counter rotate in two designated beam pipes. The LHC is designed to hold 2808 bunches with each of them containing either about 10^{11} protons or about 10^8 Pb^{82+} ions. The vacuum within the beam pipes prevents interactions of the particles with gas molecules, which could lead to instabilities of the beam. The particles in the LHC are

accelerated by 16 superconducting radio-frequency cavities. They are grouped to modules including four cavities each. Within these modules, two cavities are designated for the acceleration of the particles of each beam. The cavities are built from copper coated with niobium on the inside. Using liquid helium, they are cooled down to 4.5 K in order to transfer the niobium to a superconducting state. Within the cavities, electromagnetic oscillations are stimulated at a frequency of 400 MHz. Due to the special shape, only modes longitudinal with respect to the beam direction are stimulated. Particles passing the cavities are accelerated in the oscillating field gradient ranging up to 5 MV/m. Due to the oscillations, the particles are automatically grouped to bunches. 1232 superconducting dipole magnets keep the particles on the circular path. The coils of the dipole magnets are made of niobium-titanium. They are brought to their superconducting state by cooling them down to 1.9 K with superfluid helium-4. This state allows to operate the dipole magnets with a current of 11 850 A for a maximum magnetic field of 8.33 T. More than 8000 additional superconducting magnets with higher multipole orders are installed to focus and stabilize the beam.

The acceleration of the particles in the LHC represents only the last stage in a sequence of particle-accelerations. A large complex of particle accelerators carries out the previous stages [1]. The acceleration of protons starts with a simple bottle of hydrogen gas. The atoms of this gas are ionized by an electric field in order to obtain protons. These are subsequently accelerated to 50 MeV by the linear particle accelerator LINAC2. The accelerated protons are injected into the Proton Synchrotron Booster (PSB), where they reach energies of 1.4 GeV. In a subsequent step, the protons are accelerated to 26 GeV by the Proton Synchrotron (PS). In the last step before the injection into the LHC, the protons are brought to an energy of 450 GeV by the Super Proton Synchrotron (SPS). An illustration of the entire acceleration complex is displayed in Fig 2.1.

The final energies of the protons have been continuously increased throughout the operation of the LHC. The years marking the start of the LHC operation with a particular center-of-mass energy are following:

- 2009: $\sqrt{s} = 2.36$ TeV
- 2011: $\sqrt{s} = 7$ TeV
- 2012: $\sqrt{s} = 8$ TeV
- 2015: $\sqrt{s} = 13$ TeV

The particle beams in the LHC are stored until the luminosity falls below a certain threshold or the beams show signs of instability. As long as there are stable beams, the particles are brought to collision at four points of the LHC. At these points, the four big experiments, ATLAS, CMS, ALICE, and LHCb, are located. ATLAS and CMS are multipurpose detectors designed to measure a huge variety of physics processes. Besides allowing the measurement of known processes, these experiments aim at the search of new phenomena in particle collisions, like supersymmetry or dark matter. The search for the Higgs boson is another purpose of these detectors. The other two experiments are

[1] Some of the accelerators used for the pre-acceleration of the protons have been state-of-the-art in the time they were built. The PS, for example, provided accelerated particles for the discovery of weak neutral currents with the Gargamelle bubble chamber [82]. The SPS provided the proton-antiproton collisions for the discovery of the W and Z bosons by the UA1 [83, 84] and UA2 [85, 86] experiments. Today, they are integrated in the acceleration chain of the LHC.

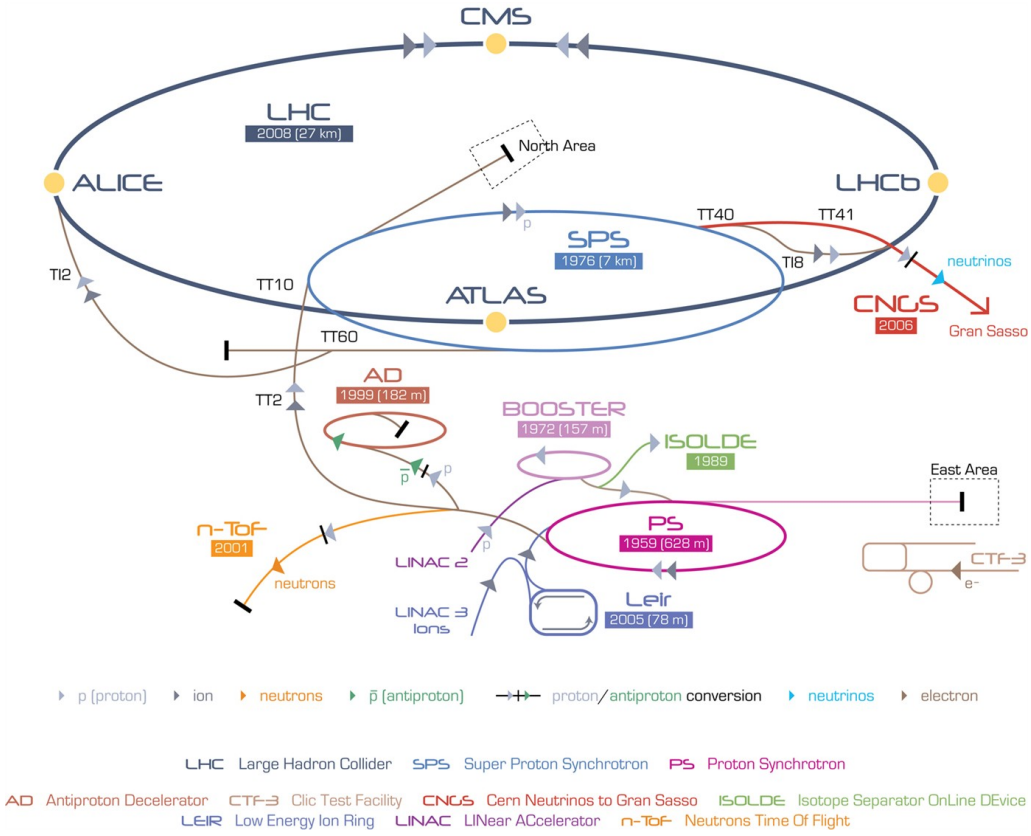


Figure 2.1: Sketch of the CERN particle acceleration complex. Taken from [87].

specialized in the investigation of particular subdomains of particle physics. The LHCb experiment targets processes associated to bottom or charm quarks. Analyses performed at this experiment include precision measurements targeting CP violation and searches for rare decays. The ALICE experiment is specialized in the investigation of heavy ion collisions. It targets the study of the quark-gluon plasma, a state present at a time shortly after the big bang.

The analysis described in this thesis is based on data recorded by the CMS experiment.

2.2. Compact Muon Solenoid Experiment

The Compact Muon Solenoid (CMS) [88] is located at LHC Point 5, which can be found in the northern part of the LHC. It is a general purpose detector designed to detect a broad range of signatures provided by interesting and new physics. It is situated in a cavern about 100 m beneath the surface. In this cavern, the CMS detector embraces the spot where the two particle beams are brought to collision. Its hermetic build aims at detecting as many of the numerous particles produced in the collisions as possible. The onion-like structure features different subdetector systems each specialized to measure the properties of different types of particles. The length of the detector adds up to 21 m and its diameter amounts to 15 m. These dimensions are necessary to ensure a proper measurement of the particles' properties. Still, compared to the ATLAS experiment, which is about double the size, the CMS indeed is quite compact. This compact build requires the application of

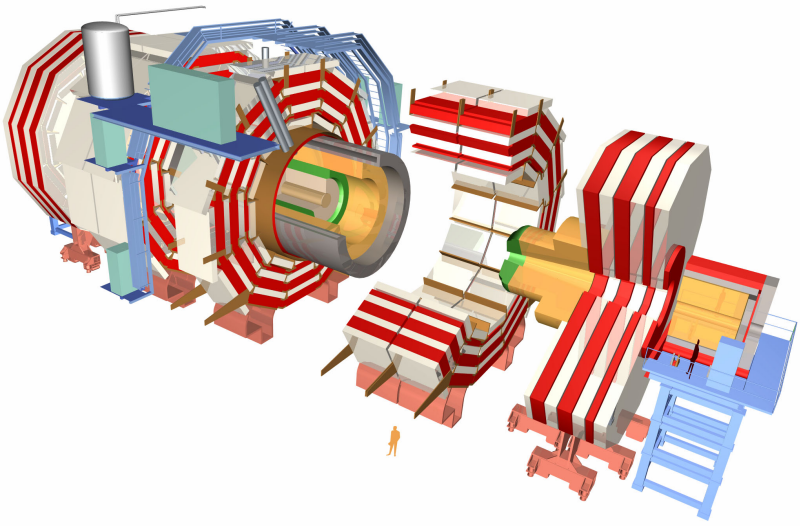


Figure 2.2: Illustration of the CMS detector. The various detector components are the tracker system in beige, the ECAL in green, the HCAL in yellow, the solenoid in gray, the return yoke in red, and the muon system in white. Taken from [89].

very dense materials in order to stop particles before they leave the detector. Accordingly, the total weight of the CMS detector adds up to 14.000 t. An illustration of the design of the CMS detector is shown in Fig. 2.2.

Starting from the collision point, the innermost subdetector is the tracker system described in Section 2.2.2. It consists of different layers of silicon detectors enclosing the interaction point. Each layer allows a precise determination of the position of passing electrically charged particles. Combining the positions of different layers, the trajectories of electrically charged particles can be determined. Together with the strong magnetic field provided by the solenoid, the trajectories allow the determination of the momentum and the electric charge of passing particles. Numerous lead-tungstate crystals, which surround tracking system, form the electromagnetic calorimeter (ECAL), which is further described in Section 2.2.3. Light electromagnetically interacting particles, like electrons, positrons, and photons, deposit all of their energy within these crystals, which is measured. The adjacent hadronic calorimeter (HCAL), which is further covered in Section 2.2.4, consists of alternating layers of absorbers and active material. Hadrons entering this subdetector interact with the absorber material and are expected to be completely stopped within the HCAL. The active medium measures the energy deposited by the initial particles. The HCAL is surrounded by the superconducting solenoid, which provides a strong magnetic field necessary for the determination of the momentum and the electric charge of particles. The return yoke is an iron structure encasing the solenoid. It provides structural support for the detector and guides the magnetic field. A more detailed description of the superconducting solenoid and the return yoke is given in Section 2.2.5. The components of the muon system described in Section 2.2.6 are embedded in the return yoke. These components are gas-ionization detectors measuring the tracks of passing electrically charged particles. As muons are the only electrically charged particles expected not to be absorbed at this part of the detector, signals in the muon system provide good identification criteria for them. The arrangement of the different subsytems in the CMS detector is illustrated in Fig. 2.3. Additionally, this figure shows examples of interactions of different types of

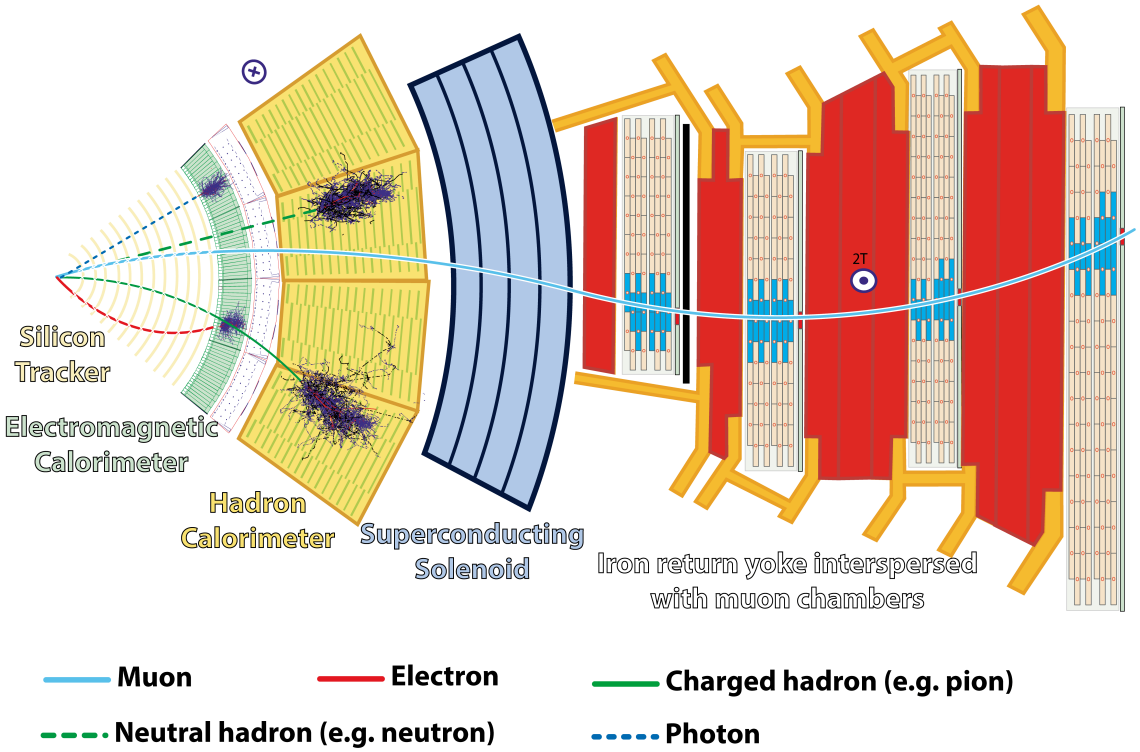


Figure 2.3: Sketch of a slice of the CMS detector. The central part of the detector, the beam line, is displayed on the left. Moving towards the outside of the detector, the tracker is depicted as beige concentric segments of a circle. Adjacent, the electromagnetic calorimeter is shown in light green. The hadronic calorimeter follows in yellow. The superconducting solenoid is displayed in blue. The outermost part of the detector is given by the return yoke in red and the embedded muon system. Further, examples of the traversal and interaction of different types of particles with the detector are shown. The trajectories of neutral particles, which cannot be measured by the detector, are represented by dashed lines. The trajectory of a muon, which passes through the detector unstopped, is shown in light blue. An electron is displayed as a red line. A photon is illustrated as a blue dashed line. Further, a charged hadron is given by the solid green line, whereas a neutral hadron is depicted as a dashed green line. Taken from [90].

particles with the subdetectors. The signals provided by the individual subdetectors are read out by the data acquisition system. However, not all events can be processed and stored, as this would exceed the capabilities of processing and storage resources. Consequently, a large fraction of events lacking interesting features is rejected by a dedicated trigger system. The data acquisition and the trigger system is covered in Section 2.2.7. The recorded data is stored and analyzed on a distributed computing infrastructure, the Worldwide LHC Computing Grid, which is described in Section 2.2.8.

Before the various subsystems are described in more detail, the special coordinate system of the detector is presented in Section 2.2.1.

2.2.1. Coordinate System

The description of positions and directions in the detector is based on a special coordinate system. It is adapted to the CMS detector, the LHC, and also to the expected particle flux. First of all, a right-handed coordinate system is defined with its origin at the designated point of collision. The z-axis points in the direction of the counter-clockwise rotating beam, which is westwards from the LHC Point 5 to the Jura mountains. The x-axis of the coordinate system points towards the center of the LHC, whereas the y-axis points vertically upwards. The most common coordinates used for the description of the detector and particles are spherical coordinates. These coordinates include the distance from collision point denoted by r and the two angles ϕ and θ. The azimuthal angle ϕ is located in the x-y-plane, which is orthogonal to beam axis. The polar angle θ is measured with respect to the z-axis.

In proton-proton collisions, a large number of interactions with small momentum transfers occur. This causes the regions with low angles of the polar angle to be highly populated. In regions with large values of the polar angle, on the other hand, comparably few particles can be found. Accordingly, the distributions of particles are not flat functions of the polar angle. Further, the interacting partons are very likely to feature different momentum fractions of the respective proton. Consequently, their system features a residual longitudinal boost. However, the polar angle is not invariant under a longitudinal boost. A Lorentz invariant variable that additionally provides flat distributions is the rapidity,

$$y = \frac{1}{2}\ln\left(\frac{E + p_z}{E - p_z}\right) .$$

In this equation, E denotes the energy of the respective particle and p_z is the z-component of the particle's momentum. If the mass of a particle is negligible compared to its momentum, the rapidity is identical to the pseudo rapidity,

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

which is a direct function of the polar angle θ. In the following, the pseudo rapidity is used for the description of positions and directions in the detector instead of the polar angle.

2.2.2. Tracker System

The CMS tracker system [91, 92] performs multiple precise position determinations of the electrically charged particles produced in collision experiments. It is the system closest to the interaction point and features a large throughput of particles produced in the collisions. In order to distinguish individual particles, a very fine granularity is required. For this reason, the CMS tracker system is subdevided into the pixel tracker with a fine granularity and the coarser segmented strip tracker. Additionally, the detector has to be very resistant against radiation damage. These are the main reasons an all-silicon configuration has been chosen for the CMS tracking system. The functional structures of the individual tracker modules are p-n junctions. A high voltage extends the depleted zone over the entire thickness of the module. Electrically charged particles passing the module cause the production of free electrons and holes. For minimum ionizing particles,

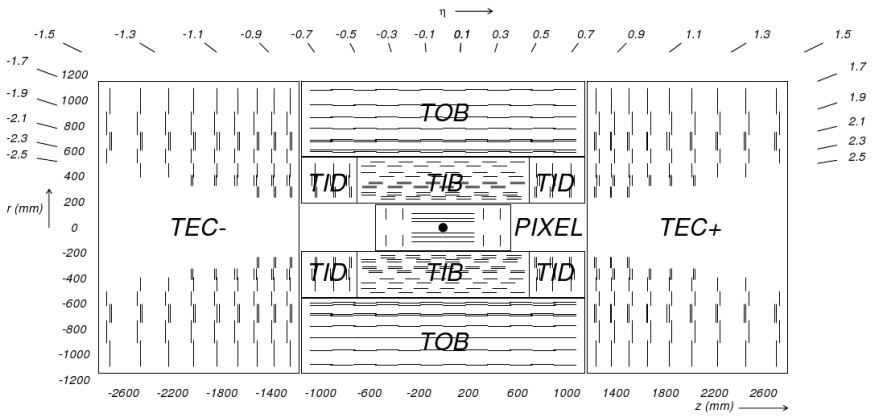


Figure 2.4: Sketch of the CMS tracker system. The different parts of the tracker system are displayed. The most inner part is the pixel tracker (PIXEL). Surrounding the pixel tracker, the first components of the strip detector can be found, the tracker inner barrel (TIB) and the tracker inner discs (TID). The most outer part of the tracker is given by the outer strip detector components, the tracker outer barrel (TOB), and the tracker endcaps (TEC). Taken from [93].

the number of electrons and holes amount to about 75 per micron thickness. The free charge carriers drift to the pixels or strips implanted into the module, where the signal is read out. The drift trajectory however is altered by the magnetic field of the CMS solenoid. This effect is quantified by the Lorentz angle and has to be accounted for in the determination of positions. Silicon detector modules are arranged in 13 to 14 layers depending on the position in the detector. Electrically charged particles passing cause hits in the different layers, which allow the reconstruction of the entire particle trajectory, which is described in Chapter 4. Such trajectories are crucial for the reconstruction and identification of particles. Further, tracks serve as input for the reconstruction of vertices and the identification of jets originating from bottom quarks. The trajectories of particles are bent by the magnetic field induced by the CMS solenoid. This effect enables the measurement of the particles' momenta and their electric charges.

The pixel tracker represents the inner part of the CMS tracker system and is described in Section 2.2.2. The strip tracker, which is described in Section 2.2.2, surrounds the pixel tracker. The CMS tracker system covers a pseudo rapidity range of $|\eta| < 2.5$. A sketch of the entire system is shown in Fig. 2.4.

Pixel tracker

The pixel tracker of the CMS experiment is the innermost part of the detector. In the barrel region covering a pseudo rapidity range of $|\eta| < 2.2$, pixel modules are arranged in three layers of cylinder barrels. These layers are positioned at radii of 4.4 cm, 7.3 cm, and 10.2 cm with each of them being 53 cm long. Two endcaps covering radii from 6 cm to 15 cm are placed at the z-coordinates of $|z| = 34.5$ cm and $|z| = 46.5$ cm. Based on this setup, two to three hits are expected for each electrically charged particle passing the pixel tracker.

The close distance to the collision point requires a fine granularity, to distinguish signals caused by different particles. Accordingly, a size of 100 µm × 150 µm was chosen for each individual pixel on a module. The pixels are given by n+-doped areas in a n-type bulk

with a p-type backside. Each module features a thickness of 285 µm. The pixel tracker consist of about 1 m^2 of active detection area. This area is populated with 1400 modules corresponding to 66 million pixels in total. The hit resolution of the pixel tracker amounts to about 10 µm in r-ϕ-direction and about 20 µm in r-z direction [88].

Strip tracker

The strip tracker surrounds the pixel tracker. It is subdivided into four parts. In the barrel region, two of these parts are given by the tracker inner barrel (TIB) and the tracker outer barrel (TOB). The shorter TIB consists of four cylindrical layers of strip modules. The TOB is composed of six cylindrical layers of strip modules. The separation into inner and outer barrel is chosen to avoid shallow crossing angles of the electrically charged particles with the detector modules. The ten layers of both parts are located at radii ranging from 25 cm to 108 cm. Another part of the strip tracker are the tracker inner disks (TID), which are three disks located at each end of the TIB. Each of these disks consists of three rings of strip modules. The last part of the strip detector are the tracker endcaps (TEC) located at each end of the TOB. The TEC consist of nine pairs of disks featuring up to seven rings of modules.

The total active detection area of the strip tracker tracker is by far larger than the one of the pixel detector. It adds up to 200 m^2 populated with silicon strip modules. However, the larger distance of the strip tracker to the collision point allows for a granularity that is coarser than the one of the pixel tracker. Accordingly, the strips are larger than the pixels and feature lengths of 9 cm for the inner parts to 21 cm for the outer parts. The pitches between the strips range between 80 µm and 120 µm. The strips themselves are given by p+-doped areas implanted into a n-type bulk with n-type backside. The total number of strips in the strip tracker amounts to about ten million. The spatial resolution of single hits in the CMS strip tracker ranges from 15 µm to 40 µm depending on the pitch between the strips [94].

2.2.3. Electromagnetic Calorimeter

The electromagnetic calorimeter (ECAL) [95] determines the energies of electrons, positrons, and photons. Its hermetic construction encloses the CMS tracker system in a pseudo rapidity range of $|\eta| < 3.0$. The main components of the ECAL are lead-tungstate crystals ($PbWO_4$) with front cross sections of about about 2 cm × 2 cm and lengths of 23 cm. About 61000 crystals populate the barrel region, while the endcap region features about 7300 crystals.

Electrons, positrons, and photons entering the crystals are expected to deposit their entire energy within the crystals. High-energy photons create electron-positron pairs in interaction with matter, whereas electrons radiate photons via bremsstrahlung. The consecutive repetition of these processes by initial and resulting particles lead to the formation of electromagnetic showers. The large atomic numbers of the elements composing the crystals promote the rate of the mentioned processes, which leads to small shower geometries. Accordingly, the energy of these particles is deposited in a small volume. The radiation length and the Molière radius, which are specific properties of materials, characterize the geometry of electromagnetic showers. The radiation length, which determines the depth of penetration of a electron until its energy has fallen to $1/e$, amounts to $\chi_0 = 0.89$ cm for lead-tungstate. Consequently, the length of an ECAL crystal adds up to 25.8 radiation lengths. The Molière radius determines the transverse extent of the electromagnetic

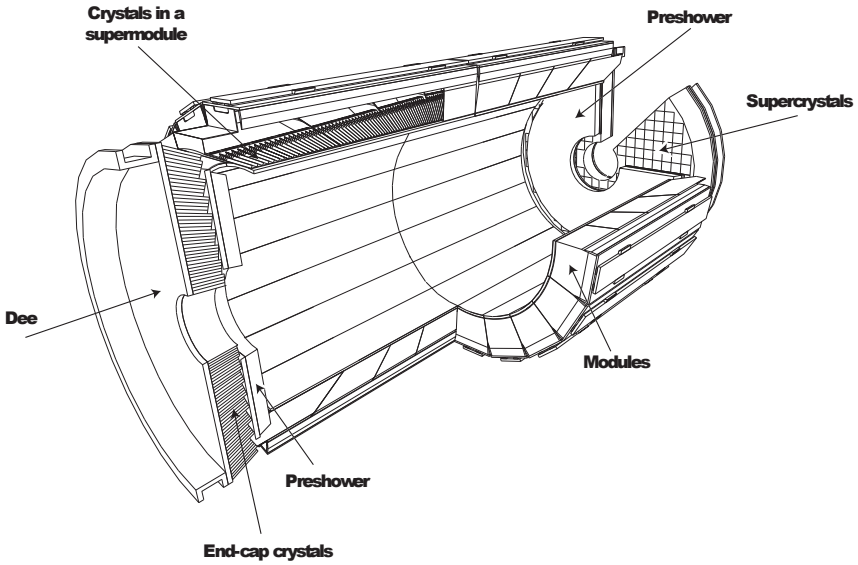


Figure 2.5: Layout of the CMS electromagnetic calorimeter showing the barrel supermodules, the two endcaps, and the preshower detectors. Taken from [96].

shower. The small value of $R_{\mathrm{M}} = 2.2\,\mathrm{cm}$ for lead-tungstate allows for a fine granularity. The lead-tungstate crystals are scintillators. The deposition of energy in the crystal stimulates the emission of photons. However, with the emission of 30 photons per MeV of energy deposited in the crystal, the photon yield is quite low. Accordingly, photodetectors with intrinsic amplification are used for the readout of the signal. Additionally, the photodetectors are required to be insensitive to the large magnetic field induced by the CMS solenoid. The photodetectors used are silicon avalanche photodiodes in the barrel region and vacuum phototriodes in the endcap region.

An additional part of the calorimeter system is the preshower (PS) attached prior to the ECAL endcaps. This detector component consists of two layers of lead and silicon strip detectors respectively. The silicon strip detectors feature a much finer granularity than the ECAL. This property allows the distinction between a single highly energetic photon and two spatially close low energetic photons stemming from the decay of a neutral pion. This distinction is crucial for the search of signatures featuring highly energetic photons, where pion decays into photons represent a large background. An important example is the search for a Higgs-boson decaying into two photons. The preshower device is only necessary in the endcap regions, where the angles between photons originating from pions are expected to be small.

2.2.4. Hadronic Calorimeter

The CMS hadron calorimeter (HCAL) [97] encloses the ECAL and represents the last subdetector inside the CMS solenoid. Its purpose is to stop strongly interacting particles and measure the energy deposited during this process. The design of the HCAL is chosen to fulfill this purpose, while still fitting in the limited space provided by the solenoid. Accordingly, as much material in terms of interaction lengths as possible is gathered inside the magnet coil. This is accomplished with a sandwich-calorimeter design, which features

alternating layers of absorber and active material. The absorber material is brass [2], which features a small interaction length and is non-magnetic. Hadronic particles passing the absorber material interact with the atomic nuclei, mainly via the strong or the electromagnetic interaction. Secondary particles are detected by the layers of active material. These layers consist of tiles of plastic scintillators emitting ultraviolet light in the interaction with particles. Embedded wavelength-shifting fibers change the ultraviolet light to visible light and direct the photons to multi-channel hybrid photodiodes. The amount of light produced is proportional to the number of particles passing the scintillator. Further, the number of particles produced in the interactions with the material is proportional to the energy of the initial particle.

The structure of the HCAL is subdivided into different parts. The hadron barrel detector (HB) consists of 2304 sandwich-calorimeter towers covering a pseudo-rapidity range of $|\eta| < 1.4$. Additionally, in the barrel region the hadronic outer detector (HO) can be found. It is made from scintillators located on the outside of the magnet coil. The HO functions as "tail-catcher" measuring the energy of particles leaking out of the HCAL and the solenoid. It covers a pseudo-rapidity range of $|\eta| < 1.26$ and extends the effective thickness of the HB to over ten interaction lengths. The hadron endcap (HE) is covering a pseudo-rapidity range of $1.3 < |\eta| < 3.0$. It consists of 2304 sandwich-calorimeter towers. The mentioned parts of the HCAL provide a similar pseudo-rapidity coverage as the ECAL. However, its granularity 25 times coarser.

The last part of the CMS calorimeter system is the hadron forward calorimeter (HF). It covers a pseudo-rapidity region of $3.0 < |\eta| < 5.0$ and is located 11 m away from the collision point. The HF covers the high rapidity region, which is highly populated by particles originating from collisions with small momentum transfers. Accordingly, a very radiation hard design was chosen. The HF is composed of steel absorbers and active material. The latter is given by quartz fibers embedded into the steel in a grid-like structure parallel to the beam line. Again, incoming particles interact with the atomic nuclei of the absorbers creating secondary particles. Electrically charged particles passing the quartz fibers cause the emission of Cerenkov light. The fibers redirect the produced light to photomultipliers, which extract the signal.

2.2.5. Superconducting Solenoid

The superconducting solenoidal coil [88] positioned after the HCAL produces a uniform axial magnetic field necessary for the determination of the momentum and the charge of particles. Its length adds up to 12.9 m, while its diameter constitutes 5.9 m. Its design is strongly influenced by the fundamental concept of the CMS experiment, which foresees the tracker, the ECAL, and major parts of the HCAL to be located within the magnet coil. At the same time, the magnet is required to provide a field that is large enough to obtain a good resolution in the measurement of the particles' momenta.

The coil of the magnet is manufactured from niobium-titanium, which is coated with aluminum. Liquid helium is used to bring it to its operating temperature of 4.5 K. At this temperature, the niobium-titanium conductors are in a superconducting state allowing a current of 19.5 kA. The current induces a magnetic field with a strength of 3.8 T and an energy of 2.7 GJ stored inside. The field further causes a hoop stress of 64 atm on the structure.

[2] The brass used for the CMS HCAL was fabricated from over a million brass shell casements from World-War-II provided by the Russian Navy.

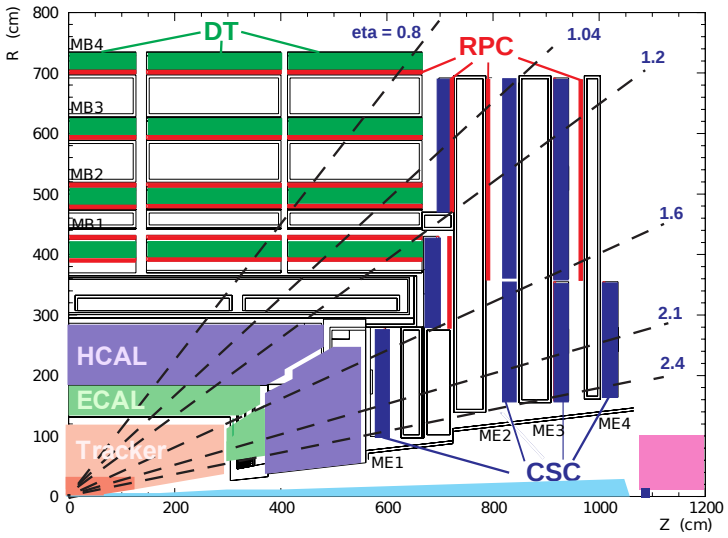


Figure 2.6: Sliced view of a quarter of the CMS detector. The various detector subsystems are highlighted in different colors. The tracker system, the ECAL and the HCAL are displayed in the lower-left corner by the areas colored in beige, light green, and purple respectively. The subsystems associated to the muon system are illustrated by the dark colors, the drift tube chambers (DT) in dark green, the cathode-strip chambers (CSC) in red, and the resistive-plate chambers (RPC) in dark blue. Taken from [99].

Surrounding the CMS solenoid, the return yoke is installed. This iron structure guides the magnetic field and provides structural support to the experiment. It consists of three layers and extends to a diameter of 14 m. Additionally, it shields the muon system from particles other than muons leaking out of the calorimeter system.

2.2.6. Muon System

The muon system [98] is embedded into the return yoke described in the previous section. Its purpose is the position determination of electrically charged particles emerging the hadron calorimeter. The measurements are performed in four layers in the barrel and the endcap respectively. As for the tracker system, these measurements can be applied to reconstruct the trajectory of electrically charged particles. In an ideal case, only muons and neutrinos are expected in this region of the detector. Accordingly, the reconstruction of a track in the muon system strongly hints at the occurrence of a muon.

The muon system provides 25.000 m^2 of active detection plane. Due to this large surface to be covered, the application of gas-ionization detectors has been chosen. Three different types of modules are installed, in order to account for the different conditions in the different regions of the detector. The modules are made up of drift tube chambers, cathode-strip chambers, and resistive-plate chambers. In total, 1400 modules are installed in the CMS detector. Fig. 2.6 shows an illustration of the arrangement of the modules in the CMS detector. The information provided by the muon system are also used in the trigger system of the CMS experiment described in Section 2.2.7.

Drift-Tube Chambers

Drift tube chambers (DT) [98,100] are installed in the barrel region of the detector covering a pseudo rapidity of $|\eta| < 1.2$. There, the muon rate as well as the neutron-induced background and the residual magnetic field is low. In total, 250 DT modules populate the four layers of the muon system in the barrel region. These four layers are located at radii of about 4.0 m, 4.9 m, 5.9 m, and 7.0 m from the beam axis. In z-direction, the muon system is divided into five parts. With respect to the azimuthal angle, these parts are further subdivided into 12 sectors. Each DT module measures $2\,\mathrm{m} \times 2.5\,\mathrm{m}$ and includes 12 layers of drift tubes. The 12 layers form three groups, of which the middle one measures the z-direction of passing electrically charged particles. The other two groups measure the coordinates in the bending plane of the magnet given by the r-ϕ plane.

Every drift tube consists of a stretched cavity bordered by aluminum, which features a width of 4 cm. The tubes are filled with a gas mixture composed of argon and carbon dioxide. On each side of the tube, a cathode is placed and an anode wire runs through the middle. The application of a high voltage leads to the formation of a uniform electric field, which is additionally shaped by electrodes installed at the top and the bottom of the drift-tube cavity. Passing electrically charged particles ionize the gas and the resulting electrons drift to the positively charged wire. In the strong field of the wire, the electrons ionize further gas atoms and cause an electric signal.

The DT modules are bordered by one or two resistive-plate chambers depending on the layer. These detectors provide the timing of a particle entering the drift tube modules. Based on this information, the drift time of the electrons can be determined, which allows a position determination more accurate than using only the position of the anode wires. The spatial resolution of a single hit in the drift tubes is about 260 µm.

Cathode-Strip Chambers

Cathode-strip chambers (CSC) [98,101] are installed in the endcaps of the detector covering a pseudo rapidity region of $1.2 < |\eta| < 2.4$. In this region, the muon rate as well as the neutron-induced background and the magnetic field is large. In total, 468 trapezoidal shaped CSC modules are distributed over the layers of the muon system in the endcap region. Within each of these modules, six gas gaps are bordered by planes of copper cathode strips and planes of anode wires. The anode wires and cathode strips are arranged perpendicular to each other. The gas gaps are filled with a mixture composed of argon, carbon dioxide, and tetrafluoromethane. A high voltage applied to the cathode strips and the anode wires induces a strong electric field. Electrically charged particles passing the gap ionize the gas atoms and molecules. In the strong electric field, the electrons produced ionize further gas atoms and molecules, which leads to an avalanche of electric charges registered by the anode wire. The signal on the wire is extremely fast and is therefore applied in the Level-1 trigger system of the CMS experiment. The ionized gas atoms and molecules induce an image charge on the cathode strips. This slower signal is used to quantify the position of the passing electrically charged particle by forming the center of gravity of the measured electric charges. The spatial resolution of a single hit in a CSC module ranges from 50 µm to 240 µm depending on the design, which is slightly different for the different layers of the muon system in the endcap region. The differences mainly concern the number of strips per chamber, the strip width, and the pitch width.

Resistive-Plate Chambers

The resistive-plate chambers (RPC) [98] are installed in both regions of the detector. 480 RPC modules can be found in the barrel region, whereas the endcaps feature 432 RPC modules. The RPCs provide coverage for a pseudo rapidity region of $|\eta| < 1.6$.

RPC modules consist of two gas gaps each bordered by an anode and a cathode plate. Each of the electrodes is covered by the high resistivity plastic material bakelite. A plane of copper readout strips is sandwiched between the two electrode-gap structures. The gas gaps are filled with a gas mixture mainly composed of tetrafluoroethane and isobutane. Electrically charged particles passing the RPCs ionize the gas molecules. The electric field induced by a high voltage applied between the electrodes causes the resulting electrons to ionize further gas molecules. This effect leads to an avalanche of electrons drifting to the positively charged electrodes. The electrodes are transparent to the electrons produced, which pass on to the readout strips and cause a signal. Based on the pattern of hits on the strips, a fast estimation of the momentum of the passing particle can be performed. This information is used in the trigger system of the CMS experiment. The RPCs feature a fast response and operate well at high rates, which allows to unambiguously identify the correct bunch crossing. The position resolution is at the order of 1 cm [102], which is much coarser than the one provided by the DTs and CSCs. The spatial resolution of hits in the RPCs mainly depend on the width of the readout strips.

2.2.7. Data Acquisition & Trigger

The LHC is designed to provide a bunch crossing rate of 40 MHz. One event recorded by the CMS experiment measures about 1 MB of zero-suppressed data. The processing and storage of all events would largely exceed the resources provided. The available storage capabilities can store data at $O(1)$ kHz and $O(100)$ MB/s. Accordingly, a huge fraction of the collision events has to be rejected at an early stage. The rejection rate necessary corresponds to a factor of about 10^6.

The CMS trigger and data acquisition system [103, 104] achieves such high rejection rates based on a two-staged approach. The front-end electronics situated in the detector receive signals from the various subdetector channels. Part of this information is passed on to the Level-1 trigger system located in the service cavern, a second cavern next to the one accommodating the CMS detector. The Level-1 trigger system selects only events with with simple signs of interesting physics. For this purpose, simple objects, so-called trigger-primitive objects, are reconstructed mainly using calorimeter and muon system information. The resulting objects are given by simple photon, electron, muon, and jet candidates. Events are selected based on the occurrence of such objects with energies or momenta above particular thresholds. Further, the total energy in the event and the missing transverse energy are considered as selection criteria. Events fulfilling none of the criteria are rejected. The Level-1 trigger system mainly consists of customized hardware, such as application specific integrated circuits (ASICs), in order to ensure a fast processing of the data. Nevertheless, also programmable hardware, like field programmable gate arrays (FPGAs), is applied. Until the response of the Level-1 trigger is returned, the entire information of the events is stored in pipelined memory given by the buffers of the front-end electronics. The time period from sending the data to the Level-1 trigger system until the response is received adds up to about 4 µs, where about 1 µs is reserved for the decision making in the Level-1 trigger system. Selected events are released for further processing, while the rejected events are dropped. At this stage, the event rate is reduced to less than

100 kHz. The data passed on by the front-end electronics is further merged, before it is transferred to the CMS computing installations on the surface. There, an event-builder network collects the data of each event and distributes them to various processing units. The second stage of data reduction is the high-level trigger (HLT) software running on each of these processing units. The software follows a strategy of rejecting events as soon as possible. This is achieved by sequentially reconstructing analysis objects using similar procedures as the ones described in Chapter 4. At different stages of this reconstruction procedure, events are checked for selection criteria. Collision events passing this selection process are transferred to the CERN Tier-0 computing facility for further processing and storage.

2.2.8. Computational Infrastructure

For the storage and analysis of recorded and simulated data, the LHC experiments make use of a tier-organized distributed computing and data storage infrastructure, the World-wide LHC Computing Grid (WLCG) [105]. The WLCG consists of over 170 centers distributed across 41 countries. The Tier-0 is located at the CERN data center at the main CERN site at Meyrin near Geneva, Switzerland. The direct connection of the Tier-0 to the LHC experiments enables a direct transfer of the data recorded to the enormous storage resources provided by the Tier-0. Accordingly, the Tier-0 represents the first contact of the data with the WLCG. An extension to the Tier-0 given by a remote Tier-0 center in Budapest, Hungary, has been put into operation rather recently. This center extends the capabilities of the primary CERN Data center in Meyrin, in order to deal with the increasing demands. Further, this center provides a backup solution in case of a failure of primary CERN Data center. The main processing of the data is carried out in the distributed Tier-1 centers. These centers are spread out all over the world and connected via high-speed networks. The Tier-1 centers additionally provide a backup for the data stored at the Tier-0 centers. The over 160 Tier-2 centers provide a platform for data analysis performed by scientists all over the world.

2.3. Luminosity

The luminosity is a quantity determining the particle flux density at the collision point. It is completely dictated by machine parameters. The instantaneous luminosity is given by

$$\mathcal{L} = fn\frac{N_1 N_2}{A} \,.$$

In this equation, f denotes the revolution frequency of the accelerator. The quantity n is the number of bunches colliding and N_i describes the number of particles per bunch. The parameter A is the geometrical cross section of the two beams at the crossing. Together with the cross section of a process, the instantaneous luminosity determines the rate for the occurrence of this particular process,

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \sigma\mathcal{L} \,.$$

Integrating this equation over time provides the total number of occurrences of the respective process as a function of the integrated luminosity.

In order to predict the number of occurrences of a particular process within a set of recorded collision events, the integrated luminosity corresponding to this dataset has to be known. The luminosity is measured by an interplay between the LHC and the CMS detector. In the CMS detector, there are five systems responsible for monitoring the instantaneous luminosity and measuring the total luminosity:

- The pixel detector,
- The muon system drift tubes,
- The forward hadronic calorimeter,
- The fast beam conditions monitor,
- The pixel luminosity telescope.

The integrated luminosity associated to the recorded data used in this analysis was measured by a method based on counting the charge clusters in the pixel detector [106]. The number of the hits in the pixel detector is closely correlated to the number of proton-proton collisions in the CMS detector. The number of proton-proton collisions N_{pp} is the product of the luminosity $\mathcal{L}$ and the proton-proton minimum-bias cross section σ_0,

$$N_{\mathrm{pp}} = \mathcal{L}\sigma_0 \ .$$

The calibration of this measurement is performed based on Van-der-Meer scans [107]. These scans rely on shifting the proton beams with respect to each other in the transverse plane at the interaction point. In this way, the size of the colliding beams can be measured and the corresponding luminosity can be determined. Based on this approach, the number of pixel hits counted can be mapped to the corresponding instantaneous luminosity.

Just like the center-of-mass energy, the instantaneous luminosity of the LHC has been increased throughout its operation. Accordingly, ever larger integrated luminosities have been achieved for subsequent data-talking periods. An exception is given by the 2015 data-taking run. During the long shutdown of the LHC before this run, some changes have been introduced in the LHC setup. These changes together with the large increase of the center-of-mass energy to $\sqrt{s} = 13$ TeV required a careful start of the operation of the LHC. The 2015 data-taking run has been started with a small instantaneous luminosity, which has been gradually increased throughout the run. Still, a peak luminosity larger than the one achieved during the 2012 data-taking run has not been not reached. An overview of the maximal instantaneous luminosities, the integrated luminosities delivered by the LHC, and integrated luminosities recorded by CMS detector for all physics data-taking runs of the LHC featuring proton-proton collisions are presented in Table 2.1. Additionally, the integrated luminosities delivered by the LHC as a function of time are displayed in Fig. 2.7 for all LHC proton runs.

Table 2.1: Luminosity values for all LHC data-taking runs featuring proton-proton collisions. The maximum instantaneous luminosity achieved in the run, the total integrated luminosity delivered by the LHC, and the total integrated luminosity recorded by the CMS experiment are listed. Taken from [108].

Year	$\sqrt{s}$	Luminosity		
		Max. instantaneous [10^{33} 1/cm^{2}s]	Delivered [fb^{-1}]	Recorded (CMS) [fb^{-1}]
2010	7 TeV	0.20	$45.0{\cdot}10^{-3}$	$41.5{\cdot}10^{-3}$
2011	7 TeV	4.02	6.10	5.55
2012	8 TeV	7.67	23.30	21.79
2015	13 TeV	5.13	4.22	3.81
2016	13 TeV	14.82	38.70	35.58

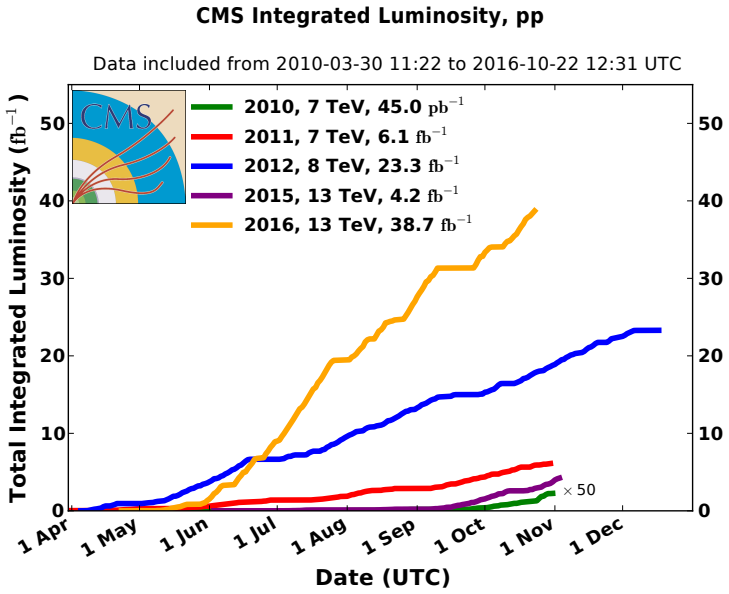


Figure 2.7: Cumulative delivered luminosity as a function of time displayed for all LHC data-taking runs featuring proton-proton collisions. Taken from [108].

Chapter 3

Measured Data and Prediction

The CMS detector described in the previous chapter detects the particles produced by a huge variety of processes in proton-proton collisions. Many of these processes are well-known by now. Nevertheless, they are still objects of investigation of high precision measurements. Processes unobserved until now, on the other hand, are mostly expected at a very low rate compared to very frequent occurrence of well-known processes. Accordingly, searches for unobserved processes are likely to face a small number of signal events hidden within a bulk of background events. In most cases, both types of analyses rely on a comparison of observation with prediction. For this reason, good knowledge of what to expect for signal and background is crucial. Predictions are derived by the calculation of single observables, such as the cross section of a particular process, or the simulation of whole proton-proton collision events based on theory. In the analysis presented in this thesis, both types of predictions are applied. It is carried out based on the first CMS data at a center-of-mass energy of $\sqrt{s} = 13$ TeV recorded in 2015, which is further specified in Section 3.1. While the calculation of observables was already discussed in Section 1.2, the procedures applied for the simulation of proton-proton collision events are described in Section 3.2. Further, the samples of simulated events used in this analysis are presented in the same section.

3.1. Measured Data

The year 2015 denotes the start of the second data-taking run of the LHC (LHC run II). The most striking feature of the second data-taking run compared to the first is the larger center-of-mass energy, which has been increased from $\sqrt{s} = 8$ TeV to $\sqrt{s} = 13$ TeV. The analysis presented in this thesis is based on the first LHC-run-II data recorded with CMS detector during the entire 2015 data-taking period.

During and after data taking, the quality of the recorded data is monitored. The data is certified as suitable for physics analysis, if a flawless detector operation and stable conditions are guaranteed [109]. This procedure sorts out data taken during runs, where issues with relevant components of the LHC or the CMS detector occurred. An example are the issues with the cryogenic system providing the CMS solenoid with liquid helium that were present throughout the 2015 data-taking period. This made an uninterrupted operation of the CMS solenoid impossible. The recorded data that was recorded with the CMS solenoid turned of corresponds to an integrated luminosity of about 800 pb^{-1}. It can only be used for a small subset of physics analyses, which does not include the analysis presented in this thesis. From the remaining data recorded while the CMS solenoid provided a magnetic field of 3.8 T, about 92 % were certified as suitable for physics analysis. This data, which is used for this analysis, amounts to recorded proton-proton collision events corresponding to an integrated luminosity of $\mathcal{L} = 2.7\ \text{fb}^{-1}$. Fig. 3.1 displays the integrated luminosity

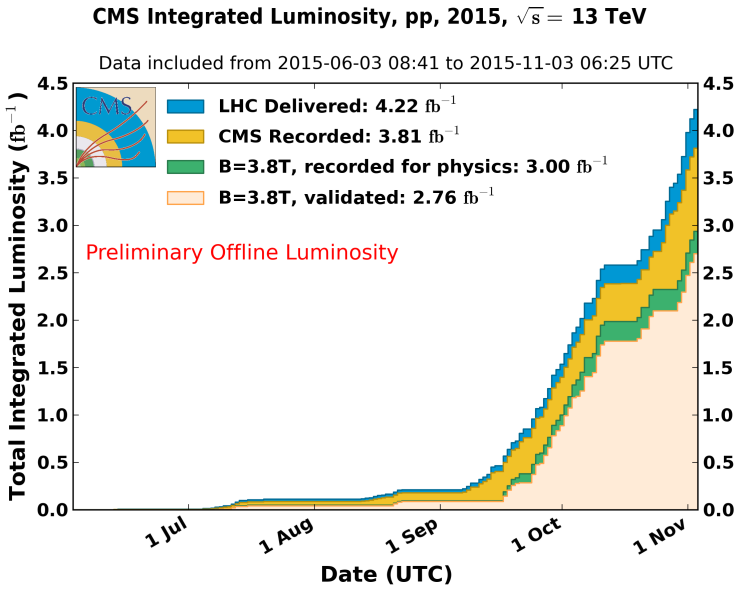


Figure 3.1: Integrated luminosity delivered by the LHC (blue) and the integrated luminosity recorded by the CMS experiment (yellow) for the 2015 data-taking period of LHC run II as a function of time. The green component shows the fraction of events recorded with the CMS solenoid providing a magnetic field of 3.8 T. The beige component shows the fraction of events certified suitable for physics analysis, with the relevant subset of subdetectors and reconstructed physics objects showing good performance. Each case includes also the area below the component with the respective color. Taken from [110].

delivered by the LHC and the fraction recorded by the CMS experiment as a function of time. Further, the same figure shows the fraction of integrated luminosity with the magnet turned on and the fraction of integrated luminosity certified as suitable for analysis. Despite the losses due to the shut down magnet, the achieved integrated luminosity is comparable to the amount of data recorded by the ATLAS experiment, which corresponds to $\mathcal{L} = 3.2\ \text{fb}^{-1}$.

The recorded data analyzed in this analysis was reprocessed, in order to incorporate the latest calibrations of detector components and analysis objects.

3.1.1. Trigger Selection

As described in Section 2.2.7, the huge rate of proton collisions is handled by only recording a subset of events relevant for physics analyses and detector studies. The selection of events is performed by triggers, which apply special requirements on the events. This analysis targets $\text{t}\bar{\text{t}}(\text{H}\rightarrow\text{b}\bar{\text{b}})$ signatures with a semileptonic decay of the top-quark pair. The most distinctive feature of the signature of this process with respect to the majority of background processes in proton-proton collisions is the prompt charged lepton originating from the leptonic top-quark decay. Accordingly, only events are analyzed that pass triggers requiring one electron or muon with large transverse momentum in the event. Events with tau leptons are not considered, as the reconstruction and the identification of these particles is rather complicated and inefficient. In the following, the two HLT triggers applied in this analysis are outlined. Both triggers rely on a fast reconstruction of lepton

objects, which is close to the one described in Section 4.4. The description of the Level-1 triggers is omitted as the HLT triggers are more restrictive.

Single-electron trigger: Events passing the single-electron trigger are required to feature one reconstructed and selected electron candidate. The selection requirements include a transverse momentum of $p_{\mathrm{T}} > 27\ \mathrm{GeV}/c$ and a pseudo rapidity of $|\eta| < 2.1$. Additionally, some identification quality criteria have to be fulfilled.

Single-muon trigger: Events passing the single-muon trigger are required to feature one reconstructed and selected muon candidate. Muon candidates have to feature a transverse momentum above $20\ \mathrm{GeV}/c$. Additional requirements on the particle-flow isolation of the muon candidates as defined in Section 4.9.2 are applied.

For consistency, the triggers are also applied to simulated events, which are described in the next section.

3.2. Event Simulation

The prediction of physics processes can be performed in two major ways, by calculating single observables and by simulating entire collision events. The results of both approaches are applied in this analysis. Simulated proton-proton collision events provide the dynamics and kinematics of the expected processes and the rate is determined by scaling the weights of these events to the respective calculated cross sections. This approach is chosen, because calculations, as they are described in Section 1.2, provide more accurate results than event simulation. Variables as they would be measured by the detector, on the other hand, are to complicated to predict by calculation and are therefore derived from simulation. The goal of event simulation is to reproduce collision events as accurately as possible. The processes taking place in collision events are rely on quantum mechanics. Hence, these processes are non-deterministic and the outcome of a single event can not be predicted in an analytic way. For this reason, the event generation is performed numerically by applying on Monte Carlo methods. Monte Carlo methods are based on a random sampling of the phase space of the initial state and the final state based on distributions predicted by theory.

The procedure for simulating a collision event is very similar to the calculation of cross sections. Processes taking part at a different energy scale and other aspects of the event simulation are factorized into different stages. The simulation of a collision event starts with the incoming hadrons. The momenta of the partons bound in these hadrons are provided by the parton distribution functions described in Section 1.2.1. The hard process, which determines the interaction between the incoming partons and the types and kinematic properties of final state particles, is represented by the transition matrix element described in Section 1.2.2 and Section 3.2.1. Both, the cross section calculation and the event simulation, have these two major ingredients in common. Also the factorization scale μ_{F} and renormalization scale μ_{R} the PDFs and the matrix element are handled in an analogue way. One of the most striking differences is the omission of the integration over the phase space in case of event simulation. As already mentioned, the initial state and the final state are randomly drawn from the distributions predicted by the PDFs and matrix element instead. Each random configuration determined in this way corresponds to one simulated event. In order for the simulated events to be comparable with recorded events, further aspects of a collision event have to be simulated. As a first step, the evolution

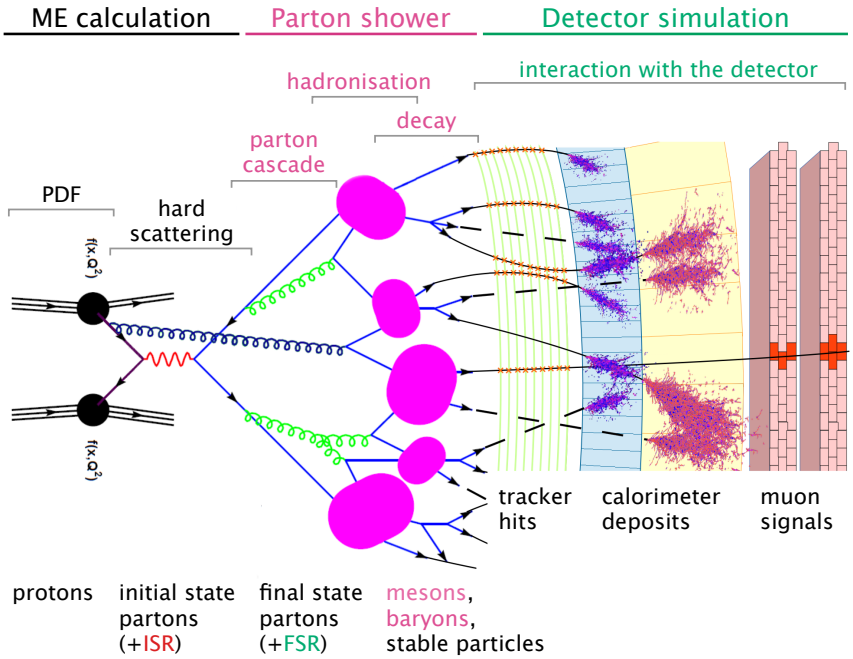


Figure 3.2: Illustration of the individual steps of the event simulation procedure. The incoming protons are depicted by the three parallel lines on the left referring to the valence quarks. The kinematics of the initial state partons are described by the PDFs. The hard process is depicted by the Feynman diagram in the middle-left part of the plot colored in red and blue. The blue lines represent the incoming and outgoing partons. The red line depicts the mediated particle carrying the transferred momentum. Further radiation and splittings (blue and green) take place in the parton shower step. In the hadronization step, the partons are merged to hadrons, which is depicted in pink. Subsequently, the decay of unstable hadrons is simulated. The last step, illustrated on the right-hand side of the graph represents the simulation of the interaction of the particles with the detector and the resulting response. Taken from [111].

of the event before and after the hard interaction is described. This evolution includes the description of initial and final-state radiation and the shower evolution of strongly interacting particles. Both effects are simulated by the parton-showering step described in Section 3.2.2. The interaction of the remaining constituents of the protons not taking part in the hard interaction is described by the underlying event covered in Section 3.2.3. As strongly interacting particles are only stable as color-neutral bound states, the quarks and gluons resulting from the previous steps are combined to hadrons in the hadronization step described in Section 3.2.4. In order to take into account detector effects and provide exactly the same output as for recorded data, the interaction of the simulated particles with the detector material and the corresponding response is determined by the detector-simulation step described in Section 3.2.5. The simulation of the effects by additional proton-proton collisions taking place during the same, previous, and following bunch crossings, so-called pile-up, is described in Section 3.2.6. A major part of the procedure for the simulation of hadron collision events is illustrated in Fig. 3.2.

3.2.1. Matrix-Element Generation

The matrix element, which is described in Section 3.2.1 in more detail, is evaluated by the application of matrix-element generators. These generators are designed for the calculation of the matrix element at a particular order of perturbation theory. For the simulation of entire collision events, matrix-element generators are interfaced to event generators. In the following, generation of the matrix elements for the first two orders in perturbation theory are outlined:

Leading-order matrix-element evaluation: As mentioned before, the leading-order contributions include only the simplest Feynman diagrams describing the process under investigation. This case does not consider any virtual loops. Hence, no singularities by such contributions arise in the calculation. Nevertheless, some LO matrix element generators support extra real emissions of partons. Singularities by these contributions are avoided by requiring an energy cutoff. The automated matrix-element generators provide a program code for the calculation of the matrix element. This code is based on the evaluation of the Feynman diagrams with the help of Feynman rules [15, 16]. There are two classes of matrix-element generators: generators for specific processes and multipurpose generators. The matrix-element generators for specific processes generate the matrix-element code based on a predefined list of partonic processes. Multipurpose generators take a chosen initial state and a chosen final state as input. The configuration can be specified by further options, for example given by the requirement for a certain processes. Based on the inputs, all compatible Feynman diagrams are generated. The matrix-element code is constructed from the Feynman diagrams obtained. An example of a multipurpose matrix-element generator used in this thesis is MadGraph5 [112].

Next-to-leading-order matrix-element evaluation: Next-to-leading-order matrix-element generators additionally take into account Feynman diagrams including additional real emissions or virtual loops. As mentioned before, additional real emissions or virtual loops cause divergences in the calculation. In the numeric evaluation of the matrix element, the KLN cancellation is not applicable, as the singularities occur in different phase spaces. In order to handle the singularities, a regularization technique is applied, which includes the extraction of poles. There are two procedures commonly used, the phase-space slicing method and the subtraction method. The former starts by separating the phase space into a region containing the singularity and region covering the remaining phase space. The latter region can be evaluated without any complications. In the singularity region, a Taylor expansion is performed in the vicinity of the pole, which is evaluated in a second step. The accuracy of the result is based on the chosen size of the singularity region. A more accurate result is obtained when the region is chosen to be small. The subtraction method is based on the construction of universally integrable terms. This is achieved by subtracting the term causing the singularity from the real emissions contribution and adding it to the virtual loops contribution. In this way, KLN cancellation is artificially induced. The NLO matrix-element generators used for the simulation of events in this thesis are MadGraph5_aMC@NLO [113] and Powheg [114–116].

3.2.2. Parton Shower

The hard process does not describe the entire event. In real particle collision events, initial and final-state particles radiate further partons, which themselves cause further radiation

and parton splittings. This effect leads to showers including a huge multiplicity of particles. The matrix elements of the corresponding processes are impossible to calculate as it would require the incorporation of all orders of perturbation theory. The parton-shower step in the simulation of collision events aims at an estimation of these higher-order corrections based on the approximation by simplified models. A simplified model for individual splittings of partons used for a long time is based on the Altarelli-Parisi splitting kernels [44]. Modern parton shower generators rely on $2 \to 3$ splitting kernels based on color dipoles [117, 118]. An exponentiation based on Poisson statistics provide the corresponding Sudakov form factor, which gives the probability for no emission taking place. The simulation procedure starts at the hard process and evolves the event in a "time" parameter until a cutoff value is reached. During the evolution, successive random splittings or radiation of particles are simulated based on the Sudakov form factor. The simulation of initial-state radiation and the corresponding showers is performed by a backwards evolution starting at the hard interaction. In order to ensure that the parton energies at the cutoff are consistent with the PDFs, the Sudakov form factors are multiplied with PDFs evaluated at the energy scale of the current iteration. The time ordering of the emissions is either determined by the transverse momentum or the angle of the simulated emission.

The simulation of the parton shower is implemented in software packages known as event generators. As indicated by the name, these programs carry out the simulation of the entire collision and therefore further include routines for the simulation of the underlying event and the hadronization, which are described in the subsequent sections. The detector simulation described in Section 3.2.5 is performed by separate processes as each experiment features distinct properties. Usually, event generators come with a built in matrix-element generator. Still, in most cases, different matrix element generators can be combined with the event generator via a standardized interface[1] [119, 120]. In this analysis, the event generator Pythia8, whose parton shower features a transverse-momentum ordering, has been used to simulate the events of the different processes described in Section 3.2.7. An event generator including an angular-ordered parton shower is Herwig++ [121]. This event generator has mainly been applied for the evaluation of uncertainties presented in Section 10.

The interface of LO matrix-element generators including additional radiation or NLO matrix element generators with parton shower generators is not trivial. Both the matrix element and the proton showers simulate the emission of additional partons. While the matrix element is more accurate for hard emissions, soft emissions are better approximated by the parton shower. In order to avoid double counting, dedicated matching algorithms are applied. These algorithms are based on a certain cutoff, which can be either be a certain angle or a particular momentum value based on the ordering of radiation. They ensure that emissions with an angle or momentum above the threshold are described by the matrix element and the remaining cases are simulated by the parton shower. Popular examples in the LO case are given by the CKKW [122] and the MLM [123] matching algorithms. An example of the matching of NLO matrix element-generators to parton showers is given by a matching approach based on the dipole formalism [117, 124, 125].

[1] The standardized interfaces between matrix-element generators and event generators are described in the Les Houches Accords. This name originates from the town, Les Houches, France, where the conference took place in 2001 during which the standardized interfaces were decided.

3.2.3. Underlying Event

The colliding protons consist of more partons than the ones taking part in the hard interaction. The remaining constituents are subject to soft interactions among each other. These processes are simulated based on the multiple parton-parton interaction approach [126]. The interactions are described by a sequence of $2 \rightarrow 2$-QCD interactions ordered in transverse momentum. Each interaction is modeled by an independent perturbative calculation with non-perturbative form factors. The number of interactions is derived by Poisson statistics based on the parton density and the parton-parton cross section.

3.2.4. Hadronization

The parton shower and the underlying event result in a huge multiplicity of single quarks and gluons. Nevertheless, the confinement dictated by the strong interaction allows only color-neutral states. Accordingly, the simulated particles are combined to color-neutral hadrons in the hadronization step. This process takes place at a very low energy scale, a regime where perturbative calculations are not valid. For this reason, the simulation of hadronization relies on phenomenological models.

One of these models is the Lund string model [127]. This model is based on the special properties of the strong-interaction field between two partons. Spatially close pairs of simulated partons are connected by so-called color-flux tubes, also called strings, with a constant energy per unit distance assigned. The hadronization is modeled by breaking up these color-flux strings with the associated creation of a new quark-antiquark pair. This strategy is pursued until the energy associated to the strings is not sufficient for the production of new quark-antiquark pairs. The Lund string model for the simulation of hadronization is implemented in the event generator Pythia.

Another method for the simulation of hadronization is the cluster hadronization model [128]. This method pursues the color structure of the parton shower. Single colored particles close in phase space are connected by color lines. Emitted gluons cause the creation of new color lines and are forced to decay into quark-antiquark pairs. Connected color lines form a set of initial clusters. Subsequently, the clusters are evolved by the decay into lighter clusters or hadrons or the radiation of photons. The cluster hadronization model is implemented in the event generator Herwig.

Many of the produced hadrons are unstable. Their decay is simulated according to the known branching fractions in a subsequent step after hadronization.

3.2.5. Detector Simulation

For a reliable comparison with data, the effects of the detector cannot be neglected. Accordingly, simulated particles are passed through a simulated version of the CMS detector. A proper detector simulation has to account for

- the geometry of the detector,
- the material of the detector components,
- the magnetic field provided by the solenoid,
- the interaction of simulated particles with detector material,
- and the electronic response of the detector.

The full simulation of the CMS detector based on the software package GEANT4 [129] provides all of these features. The geometry of the detector and its components is described by a hierarchic order of different volumes. The material corresponding to the detector components and its individual properties are associated to the respective volume. The particles traversing the detector are described by simulating their trajectory, while accounting for the magnetic field and the interaction with the detector material. A random flight length is determined based on the cross sections of the possible interaction processes. GEANT4 includes electromagnetic and hadronic processes for the interaction of particles with matter. Depending on the process, trajectories for newly produced particles are introduced. The trajectories of initial and newly produced particles are simulated in a sequential order. After the simulation of the particle trajectories and interactions, the electronic response of the various detector modules are determined.

3.2.6. Pile-Up Interactions

Next to the proton-proton collision featuring the hard process, contributions from other proton-proton collisions with small momentum transfers have to be considered. These additional contributions are known as pile-up. There are two different types of pile-up contributions: in-time pile-up and out-of-time pile-up. The former is caused by the products from additional proton-proton collisions in the same bunch crossing as the proton-proton collision featuring the hard process. Out-of-time pile-up is caused by proton-proton collision from previous or following bunch crossings. The reason for previous or following bunch crossings contributing to the event are the finite decay time of detector signals and the fact that some detectors integrate over more than one bunch crossing. The mentioned effects are described by separately simulating minimum-bias collisions including all of the steps described in the previous sections. In the CMS collaboration, these additional minimum-bias collisions are simulated using the event generator Pythia. For a proper description of out-of-time pile-up, the pulse shapes of subdetector responses are accounted for. The number of additionally simulated minimum-bias collisions for a collision event is determined based on the total inelastic proton-proton cross section and a luminosity profile. The latter is a distribution of instantaneous luminosities chosen to roughly cover the instantaneous luminosities expected for the data-taking. For each event, a random instantaneous luminosity is drawn from the luminosity profile. Together with the total inelastic proton-proton cross section, an expected number of interactions for this luminosity is determined. The number of interactions used for the simulation of pile-up is chosen randomly from a Poisson distribution with a mean value at the expected number of interactions. The contributions by pile-up are merged with the simulated hard-process collision by overlaying the detector response of the hard-interaction collision with the contributions by proton-proton collisions from the current and time-wise nearby bunch crossings.

3.2.7. Simulated Datasamples

As already mentioned, the shape of the predicted variable distributions in this thesis are derived from simulated events. Each of these events is assigned an event weight so that the sum of all event weights gives the event yield corresponding to the calculated cross section of the respective process and the integrated luminosity. The simulation of events for most processes is performed by an event generator interfaced with a matrix-element generator other than the one built in. In the following, such configurations are denoted

by the names of the matrix-element generator and the event generator connected with a plus sign.

The actual signal process of this analysis is $t\bar{t}(H\to b\bar{b})$ production. Nevertheless, contributions of $t\bar{t}H$ production with other Higgs-boson decays are also considered. Two samples of generated events for $t\bar{t}(H\to b\bar{b})$ production and $t\bar{t}H$ production covering the other Higgs-boson decays are independently generated at NLO with POWHEG+PYTHIA8. The separate $t\bar{t}(H\to b\bar{b})$ dataset ensures a sufficient number of events for the training of the multivariate methods described in Chapter 9. Both samples are scaled to the cross section recommended by the LHC Higgs cross section working group [130] multiplied with the respective branching ratio. The cross section includes NLO QCD corrections [63–67] and NLO electro-weak corrections [68–70]. In the simulation of collision events as well as in the calculation of the cross section, the factorization and renormalization scales are chosen as $\mu = m_t + m_H/2$, where m_t is the top-quark mass and m_H denotes the Higgs-boson mass.

In addition to the signal process, all background processes that give a significant contribution in the phase space considered by the analysis are simulated. The largest background contribution is given by top-quark pair production. Especially, in the case of an additional radiation of a gluon splitting into two bottom-quarks, this process shows a final state identical to the one of the signal process. Accordingly, this process is very hard to separate from signal. Nevertheless, also $t\bar{t}$ events with gluon splittings into quarks with other flavors are hard to discriminate from signal. This makes $t\bar{t}$ production the most challenging background in the search for $t\bar{t}(H\to b\bar{b})$. Just as the signal process, $t\bar{t}$ production is simulated with POWHEG+PYTHIA8. Next to an inclusive dataset including all top-quark pair decays, separate exclusive samples including only semileptonic or dileptonic top-quark pair decays are used. The exclusive samples are only applied in the analysis categories requiring at least four jets considered as originating from a bottom quark. This treatment ensures a sufficient number of events for the training of the multivariate methods in all analysis categories. For the analysis, the simulated $t\bar{t}$ events are split into different $t\bar{t}$+X contributions using a method described in Section 3.2.8. Depending on the category, the simulated events are scaled in a way that the event yield and the branching ratio matches the cross sections calculated at NNLO in QCD including resummation of next-to-next-to-leading logarithmic (NNLL) using Top++ 2.0 [131].

A contribution to the background by far smaller than the one by $t\bar{t}$ production is the single top-quark production. Besides the smaller cross section compared to $t\bar{t}$ production, the final state of this process is less similar to the one of $t\bar{t}(H\to b\bar{b})$ production. The simulation of single top-quark production is split into the three production modes, t-channel, s-channel, and the associated production of a single top quark with a W boson (tW), and is also performed with POWHEG+PYTHIA8. The generation of the samples is further subdivided in contributions including a top quark or a top antiquark. Further, t-channel and s-channel process are only simulated for leptonic top-quark decays, as the contribution of events with hadronic top-quark decays are negligible for this analysis. The t-channel and s-channel events are scaled to a next-to leading order (NLO) cross section calculated with HATHOR [132,133]. The cross section of tW production used for the scaling of the events stem from approximate NNLO QCD calulations [134].

A process with a final state very similar to the one of $t\bar{t}(H\to b\bar{b})$ but with a very small cross section is the associated production of a top-quark pair with a vector boson, $t\bar{t}Z$ and $t\bar{t}W$ production. The simulation of these processes is further subdivided into the generation of events with hadronic or leptonic vector-boson decays. The simulation is performed

using aMC@NLO+Pythia8. The events are scaled to a cross section calculated at NLO in QCD [135].

A further small background is vector-boson production in association with additional jets, W+jets and Z+jets production. The W+jets process is simulated at LO in QCD using Madgraph+Pythia8 with up to four additional jets, whereas Z+jets production is simulated at NLO in QCD using MG5_aMC@NLO+Pythia8. For the simulation of both processes, only the leptonic decay of the vector boson is considered. In order to provide smooth distributions in the phase space covered by this analysis, the processes are simulated in bins of the sum of the transverse momentum of all jets in the events for W+jets and in bins of the dilepton invariant mass in case of Z+jets production. The cross section applied for the scaling of the events are calculated at NNLO using FEWZ [136,137]. In case of the W+jets events, the cross-section fractions corresponding to the different bins of the sum of the transverse momentum of all jets are determined using Madgraph.

The smallest background process considered is diboson production. The individual contributions from WW, WZ, and ZZ production are simulated independently using Pythia8. The events are scaled to NLO cross sections derived with MCFM 6.6 [138–140].

A summary of the samples used in this analysis is given in Table 3.1. Unless otherwise noted, the simulated events are treated identically as the recorded events in the further course of the analysis.

3.2.8. $t\bar{t}$+X Flavor Splitting

The top-quark pair production process may include the production of additional gluons and quarks by radiation. Hence, the simulated samples introduced in the previous section are composed of different $t\bar{t}$+X flavor contributions. X accounts for different multiplicities of quarks with different flavors, gluons, or nothing at all. Past analyses have found that the fractions of the different $t\bar{t}$+X contributions are not properly modeled by simulation. An example is given by a measurement performed in the CMS collaboration, which determines the ratio of the $t\bar{t}$+$b\bar{b}$ cross section and the inclusive $t\bar{t}$+jets cross section [141]. For this reason, the different $t\bar{t}$+X contributions are treated separately in this analysis, which allows for an independent variation in the evaluation of the final results described in Chapter 11. The simulated $t\bar{t}$ samples introduced in the previous section are split into the different $t\bar{t}$+X contributions by finding all simulated B and D hadrons before their decay. By tracing down the event history provided by simulation, the hadrons are categorized based on their origin, which can be the decay of a top-quark or other sources like radiation. The hadrons are added to the set of simulated particles before detector simulation with their momentum four-vectors scaled down to infinitesimal small magnitudes. Using this set as input, jets are reconstructed using the method and the configuration described in Section 4.7.2. Accordingly, the collection of generated input particles is clustered using the anti-k_{T} algorithm with cone size of $R = 0.4$. The clustering process is not altered by the artificially added hadrons due to their small momentum four-vectors. The flavor of the resulting jet is determined by the flavor of the hadrons clustered inside. Jets with a B hadron clustered inside are considered bottom-flavor jets. From the remaining jets, the ones containing a D hadron are classified as charm-flavor jets. The jets left are considered light-flavor jets. Based on the flavor of the jets including hadrons not associated to the top-quark decay, simulated $t\bar{t}$ events are assigned to the following categories, where subsequent categories exclude the events selected by the categories before:

- **$t\bar{t}$+b:** Events featuring exactly one bottom-flavor jet

- **$t\bar{t}$+2b:** Events featuring a single jet with two B hadrons clustered inside
- **$t\bar{t}$+$b\bar{b}$:** Events featuring at least two bottom-flavor jet
- **$t\bar{t}$+$c\bar{c}$:** Events featuring at least one charm-flavor jet
- **$t\bar{t}$+Other/$t\bar{t}$+lf:** Remaining events

Table 3.1: Summary of the samples of simulated processes considered in this analysis. For each sample, the process, the event generator+matrix-element generator combination, and the cross section (XS) times the branching ratio (BR) used for the scaling of the events are listed.

Process	Event generator configuration	XS × BR [pb]
$t\bar{t}$(H→b$\bar{b}$)	POWHEG+PYTHIA8	0.5071 × 0.582
$t\bar{t}$(H→WW,ZZ,$\tau\tau$,$\gamma\gamma$,...)	POWHEG+PYTHIA8	0.5071 × (1-0.582)
$t\bar{t}$+jets inclusive	POWHEG+PYTHIA8	832
$t\bar{t}$+jets exclusive semileptonic	POWHEG+PYTHIA8	832 × 0.438
$t\bar{t}$+jets exclusive dileptonic	POWHEG+PYTHIA8	832 × 0.105
Single top t-channel (t)	POWHEG+PYTHIA8	45.34
Single top t-channel ($\bar{t}$)	POWHEG+PYTHIA8	27.98
Single top tW (t)	POWHEG+PYTHIA8	35.9
Single top tW ($\bar{t}$)	POWHEG+PYTHIA8	35.9
Single top s-channel	POWHEG+PYTHIA8	3.44
$t\bar{t}$+Z, Z→q$\bar{q}$	MG5_aMC@NLO+PYTHIA8	0.611
$t\bar{t}$+Z, Z→ll	MG5_aMC@NLO+PYTHIA8	0.2529
$t\bar{t}$+W, W→q$\bar{q}'$	MG5_aMC@NLO+PYTHIA8	0.435
$t\bar{t}$+W, W→lν	MG5_aMC@NLO+PYTHIA8	0.210
W+jets ($100 < H_T \leq 200$ GeV)	MADGRAPH+PYTHIA8	1345
W+jets ($200 < H_T \leq 400$ GeV)	MADGRAPH+PYTHIA8	359.7
W+jets ($400 < H_T \leq 600$ GeV)	MADGRAPH+PYTHIA8	48.91
W+jets ($600 < H_T \leq 800$ GeV)	MADGRAPH+PYTHIA8	12.05
W+jets ($800 < H_T \leq 1200$ GeV)	MADGRAPH+PYTHIA8	5.501
W+jets ($1200 < H_T \leq 2500$ GeV)	MADGRAPH+PYTHIA8	1.329
W+jets ($H_T \geq 2500$ GeV)	MADGRAPH+PYTHIA8	0.03216
Z+jets ($10 < m \leq 50$ GeV/c^2)	MG5_aMC@NLO+PYTHIA8	22635.09
Z+jets ($m > 50$ GeV/c^2)	MG5_aMC@NLO+PYTHIA8	6025.2
WW	PYTHIA8	118.7
WZ	PYTHIA8	44.9
ZZ	PYTHIA8	15.4

Chapter 4

Analysis Objects

The raw response returned by the various subsystems of the CMS detector are not directly suited for most particle-physics analyses. The detector signals stem from numerous particles produced by processes with large momentum transfers but also by other processes, like additional proton-proton collisions. In order to determine the underlying physics process, these particles are reconstructed and identified based on the response from single subsystems and their combination.

One key component for the reconstruction of stable particles are the tracks of electrically charged particles. These tracks are reconstructed from hits in the different layers of the CMS tracker and muon system as described in Section 4.1.2. The particle tracks provide information on the direction of flight at the time of production. Due to the strong magnetic field provided by the CMS solenoid, the particle tracks also bear information on the momentum and the charge of particles. Further, they enable the reconstruction of proton-proton interaction vertices as origin of the tracks, which is outlined in Section 4.2. The reconstructed vertices are identified as the primary vertex including the hard interaction process, vertices coming from additional soft proton-proton collisions, and vertices stemming from the decay of hadrons. Another key component for the reconstruction of stable particles are clusters of energy deposits in the CMS calorimeters. Electrons, positrons, and photons are expected to deposit all of their energy in the electromagnetic calorimeter, while hadrons mainly deposit their energy in the hadronic calorimeter. The methods for obtaining calorimeter clusters are described in Section 4.3. Based on the ingredients described in the previous sections, the reconstruction and identification of electrons and muons can be performed as explained in Section 4.4. Next to the standalone approach described there, the electrically charged leptons are also reconstructed and identified together with photons, electrically charged hadrons, and neutral hadrons in the particle-flow event reconstruction. This algorithm, which is described in Section 4.5, exploits the strengths of all subdetector systems of the CMS experiment, in order to provide the best possible reconstruction of the particles in the event. Due to the special properties of the strong interaction, strongly interacting particles produced at high energies form showers of collimated particles, so-called jets. The reconstruction of these objects is performed by clustering the particle objects emerging from the particle-flow event reconstruction according to dedicated algorithms, which are described in Section 4.7. The obtained jets can be identified as originating from a bottom quark based on the special properties of these particles. The procedure is called b-tagging and is described in Section 4.8. The set of analysis objects reconstructed with the procedures described contains a non-negligible number of misidentified candidates and objects not relevant to the analysis. For this reason, only reconstructed objects fulfilling the requirements presented in Section 4.9 are retained for further analysis. Finally, disagreements between measured data and simulation arising due to different behavior of the both types of data are mitigated by applying corrections as described in Section 4.10.

4.1. Particle Tracks

Electrically charged particles produce hits in the CMS tracker and muon system, when passing the different layers of silicon pixel detectors, silicon strip detectors, and muon system modules. The trajectory of the particle can be reconstructed by fitting a track to these hits, while accounting for the deflection by the magnetic field and other effects, such as multiple Coulomb scattering. The tracks of the particles provide information on their direction of flight at the time of production. The curvature caused by the magnetic field additionally enables the determination of the particle momenta and their electric charge. Tracks reconstructed in the muon system hint strongly at the occurrence of a muon in the collision event. Reconstructed particle tracks represent a keystone for the reconstruction of particles and are crucial for the performance of the particle-flow event reconstruction. Further, the reconstructed tracks serve as input for the reconstruction of vertices. Next to the identification of the primary vertex including the hard interaction, the vertex reconstruction is able to find vertices from additional proton-proton interactions and secondary vertices caused by hadron decays. The information on the latter two is combined with the information of reconstructed tracks, in order to identify contributions by pile-up interactions and jets originating from bottom quarks.

Section 4.1.1 covers the reconstruction of the hits in the tracker system and the muon system. Separate sets of tracks are reconstructed based on the tracker system hits and the muon system hits. The corresponding procedures are described in Section 4.1.2.

4.1.1. Hit Reconstruction

The hits in the tracker and muon system are reconstructed from the raw response of the respective modules. The reconstruction depends on the detector type and is therefore independently performed for the tracking system and muon system. As some modules of the muon system feature a multilayer design, the reconstructed hits in these modules are further combined to track segments.

The reconstruction of hits in the tracker system relies on the response of single pixels and strips in the respective modules. These are clustered to hits by combining adjacent pixels or strips featuring a response above a certain threshold. The positions of the hits in the pixel tracker are estimated in a local coordinate system based on two distinct approaches. The first approach relies on a projection of the clusters on the two orthogonal directions. The hit position is then extracted based on the relative charge deposited in the two pixels at each end of the respective projection. The second approach is based on templates derived from simulation. The strip-hit positions are determined by the charge-weighted average of the positions of the clustered strips. A more detailed description of the hit reconstruction in the CMS tracker can be found in [142]. The efficiency for the reconstruction of hits in the pixel and strip modules is larger than 99 %.

The reconstruction of hits in the muon system is adapted to the functionality of the three types of modules. The DT hits are given by points in the volume of the drift-tube cells. The CSC-hit positions are evaluated by a fit to the pulse heights of three adjacent strips. The positions of RPC hits are provided by the center of gravity of the signals of adjacent strips. The multilayer structures of the DTs and the CSCs allow a simple combination of the hits in each layer. Track segments are formed by a combination of aligned hits compatible with the interaction point. They are extracted by a combined linear fit to the hits associated to the candidate. A more thorough description of the muon-system hit and track-segment reconstruction can be found in [88].

4.1.2. Track Reconstruction

With the hits and track segments at hand, the tracks of electrically charged particles are reconstructed. The procedure involves collecting and fitting the hits or track segments compatible with a charged-particle track based on a technique called the combinatorial Kalman filter [143]. Mathematically, this method is a global least-squares (χ^2) minimization based on the distances of the track hypotheses to the measured hits. This method is embedded in a recursive procedure, which iteratively adds hits or segments from subsequent detector layers compatible with the track hypothesis. Charged-particle tracks are independently reconstructed in the tracker and the muon system. The individual track reconstruction procedures in the different systems are described in the following.

Tracking-System Tracks

The track reconstruction in the tracking system is subdivided into four components. In a first step, initial track candidates, so-called seeds, are determined. These are formed from any combination of two hits from different pixel detector layers that are compatible with the beam spot. Additionally, the first estimate of the transverse momentum of the seed is required to be above a certain threshold. A subsequent cleaning reduces the redundancy between the obtained seeds. The second step of the reconstruction procedure is the track finding based on the association of hits. Starting from the seeds, trajectories are extrapolated from one tracker layer to the next, while accounting for the magnetic field, potential energy loss, and deflection due to multiple Coulomb scattering. Compatible hits are selected based on their χ^2 value with respect to the track. For each compatible hit and the hypothesis of no hit in the current layer, a new trajectory with updated track parameters is created. Ambiguities concerning the sharing of hits between tracks are resolved with respect to the number of hits associated and the quality of the tracks. The track finding provides a set of compatible hits and first estimates of the track parameters. The final estimation of track parameters is performed with a two-staged fitting approach, each making use of a Kalman-filter method. In the first step, the tracks are fitted inside-out starting with the innermost hit. The algorithm proceeds by subsequently adding the associated hits analogue to the procedure used for track finding. The subsequent smoothing step reperforms the iterative fitting outside-in. Starting with the track parameters obtained in the previous fit and the hits from the outer-most layers of the tracker, the procedure refits the track by successively adding hits ever closer to the beam line. The result of this procedure still includes misidentified tracks, which are not caused by any real electrically charged particle. A reduction of misidentified tracks is achieved by selecting tracks based on the fit quality χ^2, the number of hits included in the track, and the compatibility with the beam spot. More detailed descriptions of the reconstruction of tracks in the CMS tracker system can be found in the corresponding publications [144, 145].

With this reconstruction approach, a transverse-momentum resolution better than 1 % can be achieved for tracks in the barrel region of the detector and with a transverse momentum of $p_\mathrm{T} < 10$ GeV/c. For larger transverse momenta up to 100 GeV/c, the resolution amounts to a maximum of 3.5 %. In the endcap region, the transverse-momentum resolution ranges from 2 % for low transverse momenta to about 10 % for transverse momenta up to 100 GeV/c [142].

Muon-System Tracks

The procedure for reconstructing tracks in the muon system is based on the track segments reconstructed in the drift tubes and the cathode-strip chambers, and the hits found in the resistive-plate chambers. The tracks are seeded by the track segments found in the innermost layer of the muon system. A first Kalman-filter step adds track segments or hits of subsequent layers to the seeds by extrapolating the track to the next layer and finding compatible hits and track segments. The extrapolation to subsequent layers accounts for the magnetic field, potential energy loss, and deflection due to multiple Coulomb scattering. In each iteration, the collection of tracks and their parameters are updated according to the compatible elements. Again, the case of no hit in a layer is considered. The second Kalman-filter step fits the tracks in an outside-in procedure. The track parameters are successively updated by adding information, while moving from the hits and segments of the outer-most layers towards the ones on the inside. The obtained track is refit with additional constraints on the interaction point. More thorough descriptions of the track reconstruction in the muon system can be found in the corresponding publications [88,146].

4.2. Vertices

Vertices are the origins of the observed particles and therefore hint at the occurrence of a physics interaction. Summarized, there are three types of vertices considered in this analysis: the primary-interaction vertex, additional pile-up-interaction vertices, and secondary vertices. The primary-interaction vertex represents the proton-proton collision with the largest momentum transfer. This vertex is considered to include the most interesting hard process. The pile-up vertices stem from additional proton-proton collisions in the collision event and introduce contamination in form of additional particles. A more detailed description of pile-up effects can be found in Section 3.2.6 and Section 4.7.1. Secondary vertices originate from the decays of hadrons that take place some time after the hard interaction. For this reason, these vertices are displaced with respect to the primary-collision vertex. The appearance of secondary vertices is exploited for the identification of B-hadron decays, also referred to as b-tagging. The reconstruction of secondary vertices and the b-tagging procedure are described in Section 4.8.

The collision vertices are reconstructed based on the tracker tracks resulting from the track reconstruction. In a first step, the tracks are clustered to vertex candidates. Subsequently, the best estimates on the vertex parameters are determined by a fitting procedure. The vertex finding as well as the vertex fitting rely on the adaptive vertex-fitter approach [147], which is based on a Kalman-filter technique. In contrast to the approach used for the track reconstruction, the minimization of a weighted least-square,

$$\chi'^2_{\text{min}} = \underset{\chi'^2_{\text{min}}}{\text{argmin}} \sum_{i}^{\text{tracks}} w_i r_i^2 \,, \tag{4.1}$$

is performed in an iterative way. The minimization procedure is based on a deterministic-annealing algorithm [148], a minimization algorithm that avoids getting stuck in local minima. In this formula, r_i^2 denotes the χ^2 contribution of the track i based on the distance to the respective vertex. The adaptive vertex fitter does not associate the tracks exclusively to one vertex, but rather assigns all of them to all vertices with the weight w_i,

$$w_i(r_i) = \frac{1}{1 + e^{\frac{r_i^2 - r_{\text{cutoff}}^2}{2T}}} , \tag{4.2}$$

which is also a function of the χ^2 contribution of a track. The weights can be interpreted as a probability for the track being associated to the respective vertex. The cutoff value r_{cutoff} determines the χ^2 up to which tracks still feature a considerable weight. The "temperature" T describes the smearing of the weight function. The minimization is performed iteratively, where each iteration includes the decrease of T according to an annealing schedule and the minimization of the weighted χ^2.

The vertex finding relies on the clustering of tracks based on their z-coordinate at the point of closest approach to the beam spot. The clustering is performed with the adaptive vertex reconstructor, which is able to reconstruct multiple vertices. The approach starts by considering all tracks and a single vertex candidate. The iterative minimization of the weighted least-square is performed by varying the vertex position. If a special criterion, also referred to as critical temperature, is fulfilled, an additional vertex is introduced. This minimization goes on until a termination criterion is satisfied. The final parameter estimates are obtained by a final fit of the vertices found. The procedure relies on the adaptive vertex fitter described above. The annealing in this approach is performed according to a geometric annealing schedule without the introduction of new vertices. More detailed descriptions of the vertex reconstruction procedure can be found in the corresponding publications [149–151].

4.3. Calorimeter-Energy Deposits

Most of the electromagnetic or strongly interacting particles produced in collision events deposit their energy in the calorimeter systems of the CMS detector. These deposits are mostly spread over multiple calorimeter cells. In order to draw conclusions on the initial particle, the responses of these cells are combined to calorimeter clusters. The main purposes of the calorimeter clusters are

- the reconstruction of electrons together with their emitted bremsstrahlung,
- the measurement of the energy of neutral particles,
- the energy measurement of electrically charged particles with inaccurate track information,
- the separation of energy deposits of electrically charged and neutral particles.

A special case is the reconstruction of the electron energy. Electrons produced in collision events emit bremsstrahlung photons in interaction with the tracker material. Due to the deflection of the electrons by the magnetic field, the energy of the initial electron is spread out in a wide range of the azimuthal angle. For this reason, a special clustering of the ECAL deposits is performed for the stand-alone reconstruction of electrons. Calorimeter clusters produced for the particle-flow event reconstruction are applied in the reconstruction of all particles.

Electron Energy-Deposit Clustering

The clustering of the calorimeter deposits used for the stand-alone electron reconstruction is performed with the hybrid clustering algorithm in the barrel region and the multi-5×5 clustering algorithm in the endcaps [152, 153]. These algorithms aim at clustering all contributions originating from the electron including the bremsstrahlung deposits. This is done by forming local clusters in a first step and clustering them to super-clusters in a second step. The hybrid algorithm exploits the η-ϕ-geometry of the ECAL barrel to search for a lateral electromagnetic shower caused by the electron and the additional contributions by bremsstrahlung. The clustering is seeded by crystals with locally maximum energy deposits. Arrays of azimuthally aligned crystals are added in azimuthal direction. In the endcap region, the multi-5×5 algorithm also starts by finding seeds as locally maximal ECAL deposits. To these seeds, 5×5 arrays of crystals are added based on their responses. The clusters of both algorithms are grouped to super-clusters within a fixed range in η and ϕ with respect to a seed cluster above a certain energy threshold. The energies of the super-clusters correspond to the sum of the deposits of all clustered crystals. The positions of the super-clusters are given by their energy-weighted average position.

Particle-Flow Clustering

The calorimeter clusters used for the particle-flow event reconstruction [154] are clustered separately in each subdetector component. The seeds for the clustering are given by the local maximum energy deposits. From the seeds, topological clusters are grown by successively adding adjacent cells. The topological clusters provide the seeds for the construction of particle-flow clusters. These are formed by sharing all cells among all clusters based on the distance to the respective cluster and iteratively determining the position and the energy.

4.4. Electrically charged Leptons

The reconstructed tracks and calorimeter clusters enable the reconstruction of electrically charged leptons. In the CMS collaboration, electrically charged leptons are reconstructed by combining a stand-alone approach and the particle-flow event reconstruction. The application of two complementary approaches ensures a large reconstruction efficiency. The individual signatures of electrons are mainly characterized by a tracker track altered due to bremsstrahlung and the energy deposits from the electron itself and the bremsstrahlung photons. The stand-alone reconstruction of electrons relies on the matching of specifically reconstructed tracks to the calorimeter clusters particularly clustered for this purpose. The electron reconstruction procedure is described in Section 4.4.1. The appearance of a muon is indicated by a track in the tracker system and a matching track in the muon system. The muon reconstruction described in Section 4.4.2 relies on the matching of both parts.

4.4.1. Electron Reconstruction

The trajectories of electrons are deflected due to the large radiative losses in the tracker material. This behavior is a distinctive trait of electron trajectories and lead to a decreased track-finding efficiency, when applying the standard procedure described in Section 4.1.2. For this reason a dedicated track-reconstruction procedure is applied accounting for this

effect. The electron-track reconstruction is initiated from seeds most likely originating from electrons. Such seeds are found by extrapolating the super-clusters formed in Section 4.3 to the inner layers of the tracker. Seeds are formed from two or three hits in the innermost tracker layers compatible with the extrapolation. Starting from the seeds, tracker hits are successively associated to the track in an approach based on the combinatorial Kalman filter. This approach relies on looser constraints than used for the reconstruction of tracker tracks and accounts for the energy loss by bremsstrahlung. The resulting sets of tracker hits are fit with the Gaussian-sum filter (GSF) [155], which represents a non-linear generalization of the Kalman filter. In this method, the energy loss by bremsstrahlung is approximated by a combination of Gaussian distributions. Electron candidates are build by associating the GSF tracks with the corresponding super-cluster used for seeding. More detailed descriptions of the stand-alone electron reconstruction can be found in the corresponding publications [88, 152, 153].

4.4.2. Muon Reconstruction

Muons are the only detectable stable particles effectively penetrating the detector material and emerging unstopped. Accordingly, prompt isolated muons as well as non-isolated muons from secondary particle decays in jets can be reconstructed and identified with very high efficiency and purity by exploiting the information provided by the tracker and the muon system. There are two types of muons reconstructed by the stand-alone muon reconstruction, global muons and tracker muons. The reconstruction of global muons relies on an outside-in approach, which includes the matching of muon-system tracks to tracker-system tracks. The matching is performed by extrapolating the tracks of both systems to a common surface. The track parameters are reevaluated based on a global fit of the hits and track segments of the matched tracks using a Kalman-filter technique. The inside-out approach of the tracker-muon reconstruction complements the global-muon reconstruction. In this approach, all tracker-system tracks are considered possible muon candidates. They are extrapolated to the muon system, while accounting for the magnetic field, the average expected energy loss, and multiple Coulomb scattering. If at least one muon system segment is compatible with the extrapolated track, it qualifies as tracker muon. More detailed descriptions of the stand-alone muon reconstruction can be found in the corresponding publications [88, 156].

4.5. Particle-Flow Event Reconstruction

The particle-flow (PF) event reconstruction provides a holistic reconstruction and identification of all detectable stable particles in the event: muons, electrons, photons, electrically charged hadrons, and electrically neutral hadrons. The reconstructed particles serve as inputs to further applications such as the clustering of jets (Section 4.7.2), the calculation of missing transverse energy (Section 4.6), or the calculation of charged-lepton isolation (Section 4.9.2). The PF algorithm exploits the special features of all CMS subdetector systems, in order to provide the best possible reconstruction of all types of particles and their properties. It is based on three types of input elements, the tracker system tracks, calorimeter clusters, and muon system tracks. A dedicated algorithm links these elements to blocks. The link algorithm runs over all pairs of different elements and exploits their geometrical distance as a measure of compatibility. The formed blocks serve as input for the further reconstruction of the different types of particles, which is performed in sequen-

tial order. At first, muons are reconstructed and identified from a combination of tracker and muon system tracks as described in Section 4.5.1. From the remaining elements, electrons are reconstructed with tracks and several calorimeter deposits in the ECAL. The corresponding procedure is covered in Section 4.5.2. Electrically charged hadrons are reconstructed with tracks and matching calorimeter deposits within the set of elements not used so far. The calorimeter clusters left after this procedure are associated to photons and neutral hadrons. The reconstruction of the last three types of particles is covered in Section 4.5.3.

4.5.1. Particle-Flow Muon Reconstruction

Each of the global-muon candidates found by the stand-alone reconstruction presented in Section 4.4.2 gives rise to a potential PF muon candidate. To be accepted, the candidates have to fulfill either the isolated or tight muon-identification criteria. The isolated identification is passed, if the sum of transverse momenta of tracks in a cone with $R = 0.3$ around the muon candidate is less than 10 % of its transverse momentum. The tight muon identification is composed of several quality criteria such as a minimum number of hits included in the track or the compatibility with muon segments. More details on the tight muon-identification requirements can be found in Section 4.9.2. Additional muon candidates can be accepted during the reconstruction of electrically charged hadrons in a later stage of the PF event reconstruction. The tracks of reconstructed PF muon candidates are removed from the blocks. Energy deposits in the calorimetry system expected for the PF muon candidates are subtracted from the calorimeter clusters matching the respective trajectories.

4.5.2. Particle-Flow Electron Reconstruction

Electron candidates are reconstructed from charged-particle tracks and several matching calorimeter clusters. The consideration of multiple calorimeter clusters accounts for the spread in energy deposit by bremsstrahlung photons. Another consequence of bremsstrahlung is the alteration of trajectories of electrons, which causes inefficiencies when using the standard track reconstruction described in Section 4.1.2. Potential electron tracks are reconstructed using a dedicated approach. The finding of seeds for track finding and track fitting is performed using an inside-out approach, which relies on the information of existing tracks and a matching to calorimeter clusters. In case the extrapolation of a track is geometrically compatible with a respective calorimeter clusters, the inner-layer tracker hits of the respective track are used as seeds. Further seeds are formed from the inner-layer hits of tracks, which are identified as electron tracks based on a multivariate method exploiting the distinctive features caused by bremsstrahlung. Based on the seeds, the track finding and track fitting proceeds as described in Section 4.4.1. The PF electron candidates are built by associating the GSF tracks with the PF calorimeter clusters. This is done based on a multivariate method trained on the information of the GSF tracks, the calorimeter clusters, and the combination of both. The tracks used for seeding and the calorimeter clusters associated to the electron candidate are removed from the PF blocks.

4.5.3. Particle-Flow Charged Hadron, Neutral Hadron, & Photon Reconstruction

Electrically charged hadrons, electrically neutral hadrons, and photons are reconstructed with the remaining blocks. The procedure for reconstructing and identifying electrically charged hadrons is based on matching tracks and calorimeter clusters. The tracks used for this purpose are required to fulfill tighter quality criteria than the ones used for the muon and the electron reconstruction. In a first step, redundancies are resolved by allowing every track to be linked only to the closest calorimeter cluster. Still, clusters are allowed to have more than one track assigned. After the cleaning step, the transverse momenta of the tracks are compared with the energies of the calorimeter clusters. In this comparison, three cases are considered:

- $\sum_i^{\textbf{tracks}} \boldsymbol{p}_{\mathrm{T},i} > \boldsymbol{E}$**(cluster):** In case the calorimeter cluster energy is smaller than the sum of transverse momenta of all tracks associated to it, the tracks are checked for fake tracks and muon candidates. This extended search for muons is performed by applying a set of loose muon identification criteria to global muons compatible with the tracks under investigation. The tracks found to be muons or fake tracks are removed from the block. Remaining tracks associated to the calorimeter cluster are identified as electrically charged hadrons.
- $\sum_i^{\textbf{tracks}} \boldsymbol{p}_{\mathrm{T},i} \approx \boldsymbol{E}$**(cluster):** If the calorimeter cluster energy is compatible with the total sum of transverse momenta of all tracks associated, the tracks are identified as electrically charged hadrons.
- $\sum_i^{\textbf{tracks}} \boldsymbol{p}_{\mathrm{T},i} < \boldsymbol{E}$**(cluster):** In case of a calorimeter-cluster energy larger than the sum of transverse momenta of all tracks associated to it, the tracks are identified as electrically charged hadrons. Excess energy is identified as originating from neutral particles.

Charged-hadron candidates are reconstructed by matching the tracks to the associated calorimeter cluster. The transverse momentum is reevaluated by refitting the track including the calorimeter-cluster position. The transverse momentum obtained provides a more accurate energy estimation than the energy deposit in the calorimeter. For this reason, this momentum estimation determines the momentum and the energy of the charged-hadron candidate. The tracks associated to the charged-hadron candidates are removed from the PF blocks and their energy estimates are subtracted from the corresponding calorimeter clusters.

Remaining blocks solely consist of calorimeter clusters. These blocks are reconstructed as photons and neutral hadrons depending on the size of the ECAL and HCAL energy deposits.

4.6. Missing Transverse Energy

Neutrinos and other possible weakly interacting neutral particles do not interact with the tracker material and therefore are not detected. The direction and the energy of these particles can only be estimated by an indirect approach. The incoming protons feature only longitudinal momentum. Due to momentum conservation, the transverse components of all particles produced in a collision are expected to add up to zero. In case particles are produced that have not been detected, a residual transverse momentum is obtained by this

calculation. Accordingly, the sum of the transverse momenta of all neutrinos and other possible weakly interacting particles is estimated by the negative sum of the transverse momenta of all particles reconstructed with the PF event reconstruction,

$$\vec{\not{E}}_T = -\sum_{i}^{\text{PF cand.}} \vec{p}_{\mathrm{T},i}$$

This quantity is referred to as missing transverse energy. Due to inefficiencies in the detection and reconstruction of particles and the limited transverse momentum resolution, the measured missing transverse energy does not only contain contributions of neutrinos or other possible weakly interacting particles. These contributions are partly corrected for by propagating the calibrations applied to jets to the missing transverse energy. This procedure is outlined in Section 4.7.3.

4.7. Jets

Due to the special properties of the strong interaction, quarks and gluons produced in collision events form collimated showers of particles, which are referred to as jets. In order to deduce the properties of the initially produced strongly interacting particle, all of its secondary particles are combined and analyzed. However, in most cases the grouping of all reconstructed particles is ambiguous. For this reason, clusters of particles are formed based on special rules given by jet algorithms.

The reconstruction of the original particle is complicated by additional particles in the event stemming from other sources. Especially in case of hadron colliders, where a very large fraction of interactions are based on QCD processes, a huge multiplicity of additional hadrons can be produced by the underlying event or additional proton-proton collisions. Pile-up effects, which are described in Section 3.2.6, can be partly mitigated by identifying reconstructed particles stemming from additional proton-proton collisions and removing them from the set of reconstructed particles. The corresponding procedure is described in Section 4.7.1. A set of jet algorithms relevant for the reconstruction of jets in this analysis are covered in Section 4.7.2. The energies of the reconstructed jets are biased by energy loss due to undetected particles, inefficiencies in reconstruction, non-uniformity in the detector response, and contamination by pile-up and the underlying event. The energies of the reconstructed jets are corrected for these effects as described in Section 4.7.3.

4.7.1. Pile-Up Mitigation

Particles from sources other than the hard process complicate the reconstruction of jets most interesting to this analysis and bias the measurement of their properties. A main source for additional particles in a collision event are additional proton-proton collisions, also known as pile-up interactions. As described in Section 3.2.6, there are two types of pile-up effects, in-time pile-up and out-of-time pile-up. The effect of electrically charged in-time pile-up contributions in the reconstruction of jets is mitigated by an approach called charged-hadron subtraction (CHS) [157]. This method aims at removing electrically charged hadrons stemming from additional proton-proton collisions from the set of input particles used for jet clustering. This is achieved by geometrically matching the tracks of particles to the pile-up vertices's reconstructed as described in Section 4.2. Neutral

hadrons and other pile-up effects are mitigated by the jet calibration procedure described in Section 4.7.3.

4.7.2. Jet Reconstruction

The properties of quarks and gluons produced in a collision event are deduced by analyzing the collection of particles resulting from the hadronization process. Dedicated algorithms provide a set of rules for collecting these particles and merging them into a single object. Jet algorithms can be applied on a variety of different input objects: partonic calculations, the output of parton-shower simulations, measured quantities like calorimeter deposits, or reconstructed particles. In this analysis, the particles obtained by the PF event reconstruction serve as input to the jet algorithms applied. This set of input objects is cleaned from pile-up particles using the CHS method described in the previous section.

There is a huge variety of of different jet algorithms. However, in most cases, there is no single optimal way for clustering particles to jets and the choice of a jet algorithm is ambiguous. Still, an important property determining the quality of a jet algorithm is the infrared and collinear (IRC) safety. Jet algorithms are considered infrared or collinear safe if the radiation of a soft particle or a collinear splitting of partons does not change the outcome. Jet algorithms can be grouped into two major classes: cone algorithms and sequential recombination algorithms.

Cone algorithms feature a top-down approach relying on the approximation that QCD branching and hadronization leave the energy-flow unchanged. Generally, the procedure is based on clustering all particles in a cone of a given size. However, most cone algorithms suffer from IRC unsafety, which is one of the reasons, why they are not considered in this analysis. More information on this group of algorithms is given in [158].

The second class of jet algorithms are sequential recombination algorithms. The underlying bottom-up approach relies on the iterative combination of the closest particles based on a specific distance measure. An advantage of these algorithms is their clustering sequence, which in some cases resembles QCD branching. This property is especially important for the analysis of the substructure of jets presented in Section 7.2. Another important characteristic of these algorithms is their IRC safety. Three sequential recombination algorithms represent the jet algorithms most commonly used at the LHC:

- the k_T algorithm [159, 160],
- the Cambridge/Aachen algorithm [161, 162],
- the anti-k_T algorithm [163].

The three algorithms follow the same procedure and only differ in the definition of the distance measure between two particles. The single steps of the algorithms are the following:

1. The particle-particle distances,

$$d_{ij} = \min\left(p_{\mathrm{T},i}^{2p}, p_{\mathrm{T},j}^{2p}\right) \frac{\Delta R_{ij}^2}{R^2} \,, \tag{4.3}$$

are calculated for every pair of input particles i and j with transverse momenta $p_{\mathrm{T},i}$ and $p_{\mathrm{T},j}$. Further, the "particle-beam distances"[1],

$$d_{i\mathrm{B}} = p_{\mathrm{T},i}^{2p} \,, \tag{4.4}$$

are determined for all input particles i. The value of ΔR_{ij} specifies the distance between the particles i and j in the η-ϕ-plane,

$$\Delta R_{ij}^2 = (\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2 \,. \tag{4.5}$$

The cone-size parameter R steers at which angular distance particles are still combined or declared as final jets. Accordingly, it can be interpreted as the radius of the jet in the η-ϕ-plane. The different algorithms are defined by the choice of the parameter p:

- $p = 1$: k_{T} algorithm
- $p = 0$: Cambridge/Aachen algorithm
- $p = -1$: anti-k_{T} algorithm

2. The minimum among all particle-particle and particle-beam distances is determined.

3. a) In case, the minimum is given by a particle-particle distance, the particles i and j are combined into a single object by adding their momentum four-vectors. The algorithm continues with step 1.

   b) In case the minimum is given by a particle-beam distance, particle i is declared a jet and removed from the set of particles. The algorithm continues with step 1.

4. If this step is reached, no particles remain in the set of particles and all final jets are found. Accordingly, the clustering process is stopped.

The Cambridge/Aachen algorithm is defined by $p = 0$. In this case, the particle-particle distance reduces to a term only considering the angular distance and the particle-beam distance reduces to $d_{i\mathrm{B}} = 1$. Hence, the clustering is fully independent of the momenta of the particles and only relies on their angular distances. This results in a clustering sequence resembling the QCD branching at different angular scales. Due to this property, the Cambridge/Aachen algorithm is well suited for the investigation of jet substructure. The jets resulting from the Cambridge/Aachen algorithm feature non-circular geometries as illustrated in Fig. 4.1.

The clustering procedure of the k_{T} algorithm relies on the transverse momenta of particles in addition to their angular distances. Due to its distance measure, it favors combinations that involve soft particles. The clustering sequence obtained by the k_{T} algorithm resembles the QCD branching at different energy scales. For this reason, the k_{T} algorithm

[1] The term, "particle-beam distance", stems from the first attempt of adapting the k_{T} algorithm to the application in hadron colliders [164]. In this formulation, an additional beam jet is introduced. A particles is assigned to the beam jet, if the respective particle-beam distance is smaller than all particle-particle distances. This version of the k_{T} algorithm is also known as exclusive k_{T} algorithm.

is also suited for the investigation of jet substructure. As for the Cambridge/Aachen algorithm, the geometries of jets clustered with the k_T algorithm are typically non-circular, which is displayed in Fig. 4.1.

The anti-k_T algorithm also depends on the transverse momenta of particles in addition to their angular distances. Nevertheless, the clustering behavior of the anti-k_T algorithm is contrary to the one of the k_T algorithm as it favors the combination of hard particles. The clustering sequence does not resemble QCD branching in any way. For this reason, the anti-k_T algorithm is not suited for the investigation of jet substructure. In contrast to the other two jet algorithms, the jets resulting from the anti-k_T algorithm feature circular geometries as illustrated in Fig. 4.1, which is a reason why this algorithm is in some cases preferred over the other ones.

The description of the different jet algorithms is based on the publication "Towards Jetography" [158]. There, more information on the jet algorithms presented and descriptions of further jet algorithms can be found.

The standard jets used in this analysis are clustered using the anti-k_T algorithm with a cone-size of $R = 0.4$. As already mentioned before, the input used for the clustering of the jets are the reconstructed PF candidates with charged-hadron subtraction applied, which mitigates the effects of pile-up. The choice of the cone-size parameter is based on the aim of resolving jets originating from single strongly interacting particles. For this reason, the jets clustered with this configuration are denoted as resolved jets in the following.

A special application of jets is introduced in Chapter 7. The jets described there are specialized for investigating the decays of massive particles with large transverse momenta.

4.7.3. Jet Calibration

The measured energy of jets is biased due to effects such as energy loss by inefficiencies in the detection and reconstruction of particles, non-uniformity of the detector, and contamination by pile-up and the underlying event. Additionally, simulation and measured data show differences in behavior, when applying the jet algorithms. In order to produce reliable measurements in data as well as in simulation, the energies of jets are calibrated. The calibration is separated into the correction of two different components: the jet-energy scale and the jet-energy resolution. The jet-energy scale is calibrated in data and simulation as described in Section 4.7.3. The jet-energy resolution is dictated by the detector and therefore only corrected in simulation as described in Section 4.7.3.

Jet-Energy Scale Calibration

The jet-energy scale (JES) is calibrated using an factorized approach [167]. Each part of this approach accounts for a different group of effects that cause the bias in the jet-energy measurement. The first part of the calibration accounts for the jet-energy offset due to pile-up effects not cured by the charged-hadron subtraction. The second part covers differences that are mostly due to inefficiencies in detection and reconstruction and the non-uniformity in detector response. The corresponding corrections are determined solely based on simulated events by comparing the transverse momenta of clustered jets before and after detector simulation. The third part corrects for residual differences between data and simulation and is further subdivided into corrections accounting for the relative differences in pseudo rapidity and the absolute differences. The factorization scheme and the individual parts of the jet-energy scale calibration are illustrated in Fig. 4.2. For

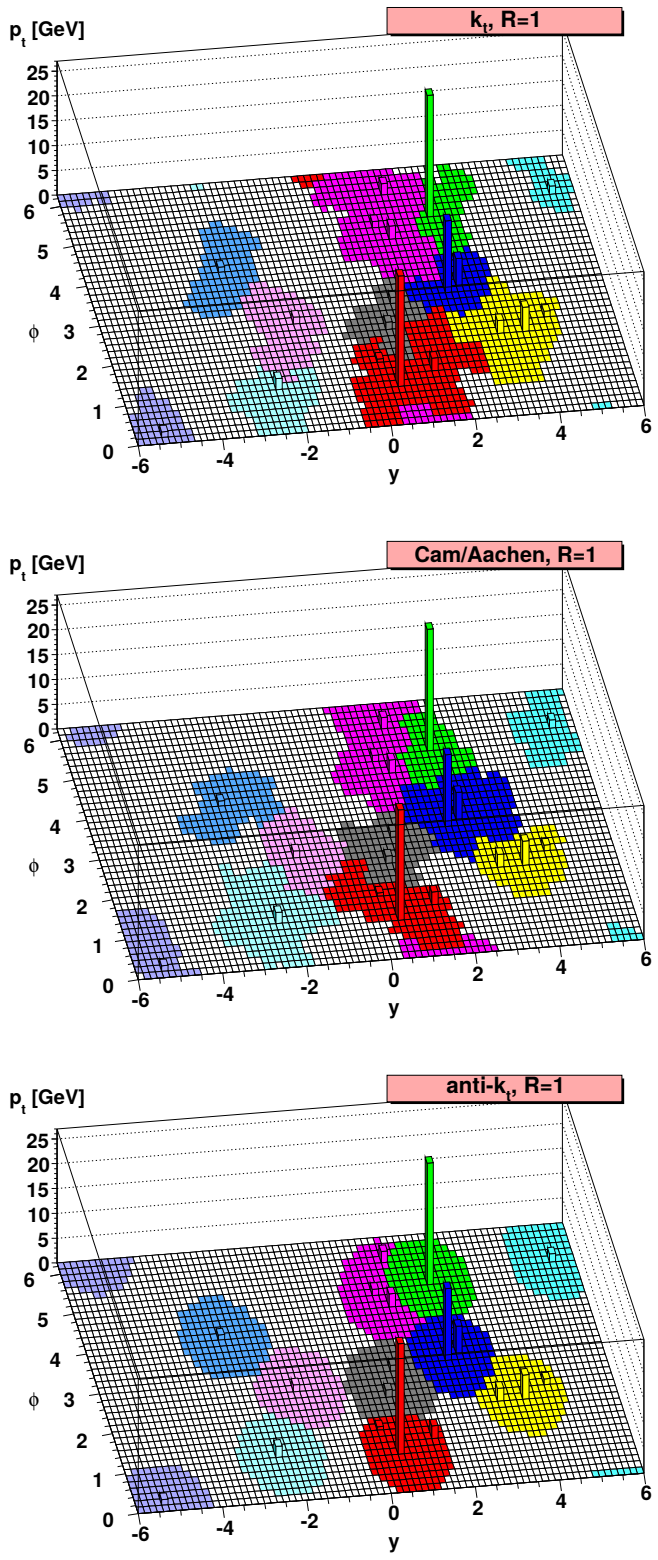


Figure 4.1: Jets resulting from the k_T algorithm (top), the Cambridge/Aachen algorithm (middle), and the anti-k_T algorithm (bottom) displayed in the η-ϕ plane. The jets are clustered from the particles of a sample event simulated with the event generator Herwig [165, 166] and artificially added infinitesimally soft particles. The space occupied by the latter particles determines the active areas of the jets. The areas of different jets are displayed in different colors. Taken from [158].

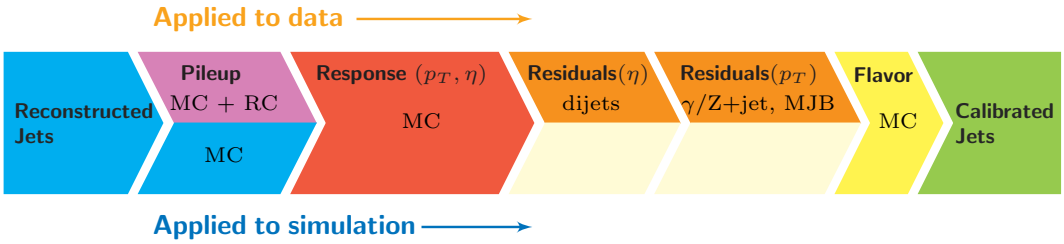


Figure 4.2: Factorization scheme of the jet-energy scale calibration applied to the reconstructed jets. Every stage of jet-energy calibration accounts for a different effect causing a bias in the reconstructed jet energy. On the top, the corrections applied to jets from measured data are illustrated. On the bottom, the corrections applied to jets from simulation are depicted. All corrections labeled with "MC" are derived from simulated events only. "RC" denotes the random cone technique used for the determination of residual differences in data and simulation for the pile-up offset corrections. "MJB" stands for the analysis of multi-jet events. The last correction accounting for differences in between different jet flavors is omitted in this analysis.

the analysis presented in this thesis, the last correction displayed in this diagram, which accounts for differences of different jet flavors, is omitted.

Pile-up offset correction: With the charged-hadron subtraction applied, the residual effects by pile-up are mainly caused by neutral pile-up contributions and out-of-time pile-up. The pile-up offset correction is determined by clustering jets in simulated dijet events with and without the simulation of pile-up applied. The correction is calculated from the differences of the jets' momenta in both cases. The pile-up offset corrections are parametrized as a function of the uncorrected transverse momentum, the pseudo rapidity, the effective jet area, and the offset energy density. The effective jet area of a jet is determined by artificially inputting infinitesimally soft particles into the jet clustering. The area occupied by the soft particles determines the effective jet area. The offset energy density is calculated as average energy density per unit area in the event. Residual differences between data and simulation caused by differences between the real detector and the detector simulation are treated with an additional correction for data. This correction is determined in zero-bias events collected with a random trigger. The corrections are determined by applying the random cone method [168], which is based on randomly placing cones with the cone size of the jets to be corrected in zero-bias events. The energies deposited within these cones correspond to the expected pile-up offset in the resolved jets. The pile-up offset corrections for the residual differences between data and simulation are given by the ratio of the random-cone energy and the energy offset determined in simulation.

Simulated jet-energy response: The simulated jet-energy response corrections account for the bias introduced by effects like inefficiencies in detection and reconstruction and non-uniformity in detector response. The corrections are derived solely based on simulation. The simulated jet-energy response corrections are determined and applied on jets already corrected for the energy bias by pile-up. They are derived by clustering jets in simulated QCD multi-jet events with and without the detector simulation applied. The corresponding jets of both types are matched in each event and a correction scale factor is calculated by forming the ratio of the transverse momenta of both jets. The scale factors

are determined as a function of the transverse momentum and the pseudo rapidity of the uncorrected jets.

Residual corrections for data: The residual corrections for data account for differences between data and simulation. The correction for these differences is subdivided into two parts. The first part accounts for the differences in jet-energy response relative to jets produced centrally in the detector. For the determination of these corrections, a tag-and-probe method is applied on a pure sample of dijet events. After the selection of such a sample, a tag jet with the pseudo rapidity requirement $|\eta| < 1.3$ is selected, whereas the other jet, also called probe jet, is unconstrained. The relative difference in transverse momentum of these jets is determined in data and simulation. The scale factors for the residual relative corrections are given by a ratio of the relative differences in transverse momentum of these jets in data and simulation. They are determined differentially in the pseudo rapidity of the probe jet. The second part of residual corrections accounts for differences in the absolute scale of transverse momentum of jets in data and simulation. The correction scale factors are determined and applied after the corrections for pile-up, simulated jet-energy response, and residual relative differences between data and simulation have been applied. The scale factors are derived by comparing the jet-energy response in data and simulation using an two staged approach. First a coarse estimate of the normalization difference is determined from $\mathrm{Z}(\rightarrow \mu\mu)$+jet events. The transverse-momentum dependence is determined by a global fit of photon+jet, $\mathrm{Z}(\rightarrow \mathrm{ll})$+jet events, and QCD multi-jet events. In each type of events, the respective jet is compared to a reference object given by photons and reconstructed Z bosons in the photon+jet and $\mathrm{Z}(\rightarrow \mathrm{ll})$+jet events and by the entire recoil system in the QCD multi-jet events. The residual corrections are only applied on jets from data.

The jets are corrected by scaling their momentum four-vector with the scale factors obtained for each correction. As mentioned in Section 4.6, the jet-energy scale corrections are also exploited for correcting the missing transverse energy for the mentioned effects. This is achieved by adding the transverse-momentum vectors of the uncorrected resolved jets to the missing transverse energy vector and subtracting the transverse-momentum vectors of the corrected jets.

Jet-Energy Resolution Calibration

The second major part of the jet-energy calibration corrects for the differences in the resolution of the jet energy in data and simulation [167]. The corrections are only applied to jets from simulated events. They are derived and applied on jets with the full jet-energy scale calibration applied. The jet-energy resolution corrections used in this analysis are extracted using the asymmetry method [168] with jets from dijet events. In this method, the transverse-momentum asymmetry of the two jets 1 and 2 in the event,

$$\mathcal{A} = \frac{p_{\mathrm{T},1} - p_{\mathrm{T},2}}{p_{\mathrm{T},1} + p_{\mathrm{T},2}} \,,$$

is calculated. The width of the asymmetry distribution is correlated to the single-jet resolution. The jet-energy resolution (JER) scale factors are derived differentially in the pseudo rapidity. They are given by the ratio of the resolution obtained for data and simulation. The jet-energy resolution corrections are applied by using a smearing approach.

The correction procedure is based on randomly shifting the four-vectors of the jets according to a Gaussian function with a mean value of one and the width of the original jet-energy resolution scaled by the scale factor.

4.8. b-Tagging

The final state of $t\bar{t}(H\to b\bar{b})$ events features four bottom quarks, two from the Higgs-boson decay and another two from the decays of the top quarks. As discussed in the previous section, these bottom quarks result in jets. However, the jets originating from bottom quarks feature distinctive properties. Bottom quarks produced in collision events form B hadrons with a mass considerably larger than the typical hadrons occurring in jets. Further, they carry a large fraction of the jet energy. However, the most distinctive characteristic of B hadrons is their long lifetime. A B hadron can only decay via the weak interaction, where the bottom quark decays into a charm or an up quark. This decay requires a transition between quark generations, which causes the presence of a small off-diagonal element of the CKM matrix in the calculation of the matrix element. The result is a small transition amplitude, which in turn is equivalent to a long lifetime of the order of 10^{-12} s. Accordingly, a B hadron travels a distance of about a few millimeters, depending on its kinetic energy, before it decays. In the detector, the decay of a B hadron leads to tracks pointing to a secondary vertex displaced from the vertex of the primary interaction. Such a decay and the corresponding secondary vertex are illustrated in Fig. 4.3. Further, the impact parameter of a track, which is given by its distance of closest approach to the primary vertex, is illustrated in the same figure. The mentioned characteristics are exploited for the identification of jets originating from bottom quarks, also referred to as b-tagging.

b-tagging is performed based on the tracks of the particles reconstructed with the PF event reconstruction with charged-hadron subtraction applied. The procedure starts by reconstructing all secondary vertices in an event by applying the inclusive vertex finder (IVF) algorithm [169]. The tracks used for the clustering of secondary vertices undergo a preselection. From the set of tracks retained, seeds are determined and secondary vertex candidates are clustered based on the properties of the corresponding tracks. The resulting clusters of tracks are fitted twice with the adaptive vertex fitter described in Section 4.2. Each fitting step is followed by the removal of vertex candidates and tracks not fulfilling particular quality criteria. The obtained secondary vertex candidates are matched to the jets reconstructed in the event based on their distance in the η-ϕ plane.

The b-tagging algorithm used in the analysis described in this thesis is the Combined Secondary Vertex v2 (CSVv2) tagger [169][2]. The algorithm combines the information on displaced tracks and secondary vertices associated to a jet in a multivariate method. It is trained in three independent categories defined by the reconstruction of vertices and their association to the respective jet:

- **Secondary-vertex category:** Jets with at least one matched secondary vertex are assigned to the secondary-vertex category. If more than one vertex is matched to the jet, the vertices are ordered by their fit uncertainty. In the training and evaluation of the multivariate method, variables based on the properties of a single vertex are determined by the vertex with the lowest fit uncertainty.

[2]The CSVv2 is an improved version of the Combined Secondary Vertex (CSV) tagger [170] commonly used in LHC run I.

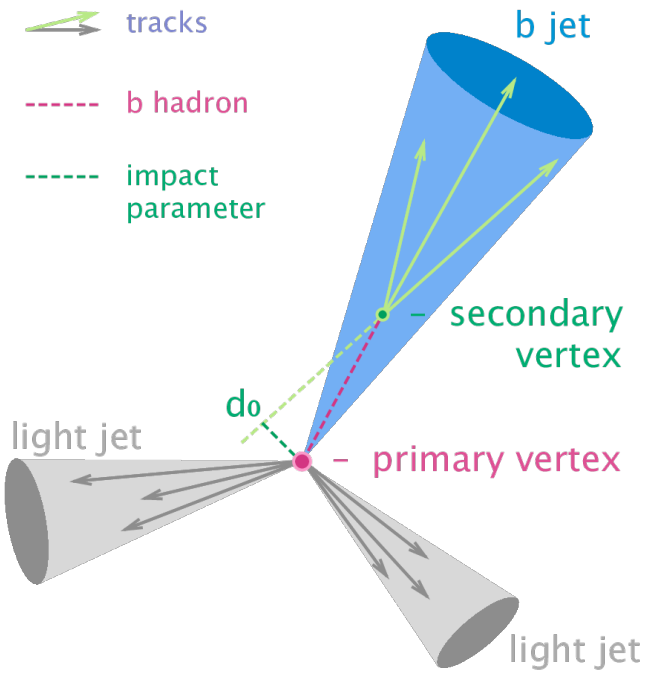


Figure 4.3: Simplified sketch of a hypothetical event featuring one jet originating from a bottom quark (colored) and two jets originating from light flavor quarks or gluons (gray). Next to the momenta of single particles depicted as arrows, the primary interaction vertex and the secondary interaction vertex from the decay of a B hadron are illustrated. The flight path of the B hadron is given by the dashed pink line. Further, an extrapolation of a particle track originating from the b-hadron decay is shown as a dashed light green line. The impact parameter determined by the distance of closest approach of this extrapolation and the primary vertex is shown as a dashed dark green line. Taken from [111].

- **Pseudo-vertex category:** Jets featuring no matched vertex are tested for the feasibility of reconstructing a so-called pseudo vertex. This case is provided if the jet includes tracks that fulfill certain quality criteria strongly hinting at the occurrence of a B-hadron decay. Jets for which this is true are assigned to the pseudo vertex category. A pseudo vertex is reconstructed by combining the respective tracks in the jet.

- **No-vertex category:** Jets featuring neither a matched secondary vertex nor a reconstructed pseudo-vertex are assigned to the no-vertex category.

In each category a neural network with one hidden layer is trained. The input variables are given by the vertex category itself, vertex properties, properties of tracks associated to vertices, and properties of tracks displaced with respect to the primary vertex. The choice of the input variables for the training in each category depend on the availability of the variables in the respective category. The outputs of the neural networks in all three categories are combined based on a likelihood ratio. A description of likelihood ratios can be found in Section 5.2.2.

4.9. Object Selections

The reconstructed objects include candidates not relevant for the analysis described in this thesis. Such objects are misidentified candidates, candidates not stemming from the hard interaction, and objects most likely originating from background processes. Their number is reduced by applying further quality criteria in form of selections, which are described in the following.

4.9.1. Primary Vertex Selection

The identification of the primary interaction vertex is crucial for identifying other objects relevant for the analysis and rejecting contributions from other proton-proton collisions. From the set of vertices reconstructed with the procedure described in Section 4.2, only the vertex with the largest sum of the squared transverse momenta of all tracks associated is considered a possible primary vertex candidate. This vertex is additionally required to feature a successful position fit and a number of degrees of freedom in the fit,

$$n_{\mathrm{dof}} = -3 + 2 \sum_{i}^{\mathrm{tracks}} w_i \ , \tag{4.6}$$

larger than four. The weights w_i in Eq. (4.6) are defined by Eq. (4.2). Accordingly, n_{dof} can be interpreted as the number of tracks that are compatible with the vertex. Additional requirements are based on the compatibility of the vertex position with the beam spot. The longitudinal distance between these two objects has to fulfill $|z| < 24$ cm. The transverse distance is required to be $|\rho| < 2$ cm for the vertex to be selected as good primary vertex.

4.9.2. Charged-Lepton Selection

In semileptonic $\mathrm{t\bar{t}}(\mathrm{H}\to\mathrm{b\bar{b}})$ events, exactly one prompt isolated lepton originating from the leptonic top-quark decay is expected. However, next to prompt leptons resulting from the hard process, the set of reconstructed leptons emerging from the reconstruction procedures described in previous sections features non-prompt electrons originating from various other sources. Main sources for non-prompt electrons are photon conversions into electron-positron pairs in interaction with the detector material, weak decays of heavy-flavor hadrons, and fake candidates by the misidentification of jets and pions. Muon candidates not relevant for the analysis are caused by hadrons reaching the muon system, muons from weak decays of heavy-flavor hadrons, and cosmic muons. Only leptons showing characteristics of prompt isolated leptons are retained for further analysis.

Electron Identification and Isolation

The identification of prompt electrons relies on a multivariate method [153]. Boosted decision trees (BDT), which are introduced in Section 5.1.1, are trained on input variables from four categories:

- track information (e.g. the transverse-momentum loss by bremsstrahlung),
- calorimeter information (e.g. the lateral extension of shower along the pseudo-rapidity direction),

- comparison of track and calorimeter information (e.g. the difference in pseudo rapidity between track and matched cluster).

Signal electrons are extracted by matching the reconstructed candidates to generated electrons in simulated Z-boson events. Fake candidates are taken from a pure sample of W+jets events in data. The BDTs are separately trained in bins of transverse momentum and pseudo rapidity. Prior to the evaluation of the multivariate-identification output, preselection cuts based on the trigger requirements and the input variables of the BDTs are applied to the electron candidates. The selection cuts applied on the multivariate-identification output are defined by working points. For this analysis, a working point corresponding to 80 % selection efficiency has been chosen.

In addition to the identification criteria, electrons are required to be isolated. The isolation of leptons is defined by the energy flow given by PF candidates with the charged-hadron subtraction applied in the vicinity of the reconstructed candidate. It is quantified by the sum over all electrically charged PF candidates (charged), electrically neutral PF hadrons (neutral), and PF photons (photon) in a cone with size $R = 0.3$ around the electron candidate divided by the transverse momentum of the electron candidate,

$$\mathrm{Iso_{PF}} = \frac{\sum_i^{\mathrm{charged}} p_{\mathrm{T},i} + \max\left(0, \sum_j^{\mathrm{neutral}} p_{\mathrm{T},j} + \sum_k^{\mathrm{photon}} p_{\mathrm{T},k} - p_{\mathrm{T,PU}}\right)}{p_{\mathrm{T,e}}} \,. \tag{4.7}$$

The pile-up contributions by neutral particles and out-of-time pile-up $p_{\mathrm{T,PU}}$ are subtracted from the sum of transverse momenta of the PF candidates. This contribution is estimated based on the average energy density and the effective area of the cone already defined in Section 4.7.3. The isolation of the electron candidates is required to be smaller than 0.15 to be selected.

Muon Identification and Isolation

The identification of muons follows a cut-based approach given by the tight PF muon identification requirements mentioned in Section 4.5.1. The identification criteria start by requiring the muon to be reconstructed as global muon, PF muon, and that its track must be matched to at least two muon stations, which is equivalent to the reconstruction as tracker muon. In order to reject hadrons reaching the muon system and muons from hadron decays, the χ^2 value from the global track fit divided by the corresponding number of degrees of freedom is required to be less than ten. Additionally, at least one hit in the muon system has to be included in the fit and the transverse impact parameter of the muon track with respect to the primary vertex has to be smaller than 2 mm. The requirement of the longitudinal impact parameter of the muon track with respect to the primary vertex to be smaller than 5 mm additionally rejects muon candidates from cosmic muons, muons from hadron decays, and tracks from pile-up. By requiring at least one pixel layer hit to be included in the track, muons from hadron decays are further rejected. Good quality of the transverse momentum measurement is ensured by requiring at least six tracker system hits to be included in the track fit.

As for the electrons, only well isolated muons are selected. The determination of the muon isolation is very similar to the one of the electron isolation given by Eq. (4.7). It is calculated based on reconstructed PF candidates with the charged-hadron subtraction applied in a cone with a size of $R = 0.4$ around the muon track. Next to the cone size,

the estimation of the pile-up contribution is altered compared to the calculation of the electron isolation. It is based on the charged-pile-up contribution, which is determined based on the charged-hadron subtraction approach: tracks matched to pile-up vertices and lying inside the isolation cone provide the charged-pile-up contribution. The halved sum of transverse momenta of these tracks represents the neutral pile-up contribution $p_{\mathrm{T,PU}}$ for muons in Eq. (4.7). The factor 0.5 accounts for the ratio of electrically charged particles and neutral particles expected in the event. The isolation of primary muon candidates is required to be $\mathrm{Iso_{PF}} < 0.15$. For veto muons, which are introduced in the next paragraph, the isolation requirement is loosened to $\mathrm{Iso_{PF}} < 0.25$.

Kinematic Requirements

Further requirements for the selection of electrically charged leptons are based on kinematic variables. The lepton with the largest transverse momentum in the event, also referred to as primary lepton, is required to feature $p_{\mathrm{T}} > 30$ GeV/c in case of an electron and $p_{\mathrm{T}} > 25$ GeV/c in case of a muon. A pseudo rapidity of $|\eta| < 2.1$ is required in both cases. In the single-lepton selection presented in Chapter 6 and Chapter 8, events with leptons additional to the primary lepton are vetoed. The definition of the veto leptons is loosened with respect to the one of the primary lepton. Additionally, the isolation is only required to be smaller than $Iso_{\mathrm{PF}} < 0.25$. Veto leptons are required to have a transverse momentum of $p_{\mathrm{T}} > 15$ GeV/c and a pseudo rapidity $|\eta| < 2.4$. The cuts on the pseudo rapidity are determined by the coverage of the ECAL and the muon system. Further, the cuts are chosen tighter than the trigger requirements presented in Section 3.1.1.

4.9.3. Jet Selection

Throughout the reconstruction of the resolved jets, hardly any quality criteria are applied. Thus, the results still feature jets consisting only of single constituents, jets without any track information, jets including or exclusively consisting of isolated leptons, and electrons and photons misidentified as jets. In order to remove such candidates, only jets fulfilling the quality criteria described in the following are retained for further analysis. First of all, jets are required to pass the PF jet identification in order to be selected. This identification requires the jets to feature at least two constituents with at least one of them being a electrically charged constituent. Further criteria target the fractions of energies by different sources clustered into the jet. Accordingly, the jet is required to feature energy fractions from

- electrically charged hadrons > 0,
- electrically neutral hadrons < 0.99,
- electrons < 0.99,
- photons < 0.99,

in order to satisfy the PF jet identification. The rejection of jets including or exclusively consisting of isolated leptons is achieved by rejecting jets within an angular distance of $\Delta R < 0.4$ to the primary leptons passing the respective selection described in Section 4.9.2. Further, only jets in a phase space relevant for this analysis are selected by applying cuts on kinematic variables. Accordingly, jets are required to have a transverse momentum of $p_{\mathrm{T}} > 30$ GeV/c and a pseudo rapidity of $|\eta| < 2.4$ in order to be selected. The cut on the pseudo rapidity is determined by the coverage of the tracker system.

4.9.4. b-Tags

The multivariate b-tagging approach described in Section 4.8 returns a continuous value between zero and one. Jets most certainly originating from bottom quarks feature rather large values of the b-tagging output, whereas jets most likely stemming from other particles are more likely to return small values. The resulting distributions are, for example, applied in the identification of boosted top-quark candidates presented in Section 7.6 or in the final discrimination of $t\bar{t}H$ events against background events as described in Chapter 9.

Nevertheless, in some cases, a classification that indicates if a jet is considered as originating from a bottom quark is desirable. An example of an application in this analysis is the categorization of selected resolved events based on the overall number of resolved jets and the number of resolved jets in the event identified as originating from bottom quarks as it is described in Section 6.1. A jet is considered as originating from a bottom quark if its b-tagging output is above a chosen threshold. This characteristic is referred to as "b-tag" in the following. The threshold applied in this analysis is defined by a working point corresponding to a misidentification efficiency of jets stemming from light quarks and gluons of about one percent. Accordingly, jets have to feature a b-tagging output larger than 0.8 in order to be considered as b-tagged.

4.10. Simulation Correction

Simulated collision events produced with the methods described in Section 3.2 achieve to describe measured data quite accurately. Still, in some cases, simulated data shows a different behavior than measured data. In order to ensure a reliable description, simulated data is corrected for such effects. The analysis described in this thesis includes the correction of three different quantities featuring discrepancies between data and simulation: pile-up (Section 4.10.1), the lepton selection efficiency (Section 4.10.2), and the b-tagging output (Section 4.10.2).

4.10.1. Pile-Up Correction

The number of proton-proton interactions in data depends on the instantaneous luminosity at the time it was recorded. At the time the simulated datasets have been produced, the information about the instantaneous luminosities at each moment of data taking has not been available. For this reason, the samples have been produced so that the distributions of number of proton-proton interactions roughly cover the conditions expected for data taking. With the exact run conditions in hand, the scenario in simulation is adapted to the one in data by the application of event weights. This reweighing of events relies on the simulated number of interactions per bunch crossing and the number of interactions per bunch crossing expected in data. The former can simply be taken from the information provided by simulation. The latter is determined based on the instantaneous luminosity per bunch crossing and the total inelastic cross section in proton-proton collisions. The instantaneous luminosity per bunch crossing is provided by the luminosity measurements described in Section 2.3. A proton-proton inelastic cross section of $\sigma_{\text{inelastic}} = 69.4\,\text{mb}$ has empirically been found to describe data well [171]. Given these two ingredients, the distribution of the number of proton-proton interactions for the whole data taking period is constructed, while only considering good data taking runs. For each number of proton-proton interactions, an event weight is calculated by forming the ratio of the number of expected and simulated events in the respective bin of the distributions. The event weights

are applied to all simulated events with respect to their generated number of proton-proton interactions.

4.10.2. Lepton and Trigger Efficiency Corrections

Data and simulation show different selection efficiencies, when applying the lepton identification criteria, the lepton isolation criteria, or the single-lepton triggers. In order to account for these effects, corrections are applied for all of the mentioned selections. The corrections are determined by measuring the efficiencies for measured and simulated events using a tag-and-probe method [172] in high purity samples of Z-boson decays into two leptons. Measured and simulated events are selected by requiring exactly two leptons, one tag lepton fulfilling tight selection criteria and one probe lepton fulfilling the criteria under investigation. In this case, the criteria under investigation are the lepton identification requirements, the lepton isolation requirements, or the single-lepton trigger selection. An additional set of events is selected for the probe lepton failing the respective criterion under investigation. The additional requirement of the invariant mass of the tag and probe lepton system lying in the Z-boson mass window 60 GeV/c^2 $< m_{\mathrm{ll}} <$ 120 GeV/c^2 provides a pure sample of Z-boson decays. The number of events passing or failing the criterion is extracted by fitting a signal and a background function to the invariant mass distribution of the dilepton system. The selection efficiencies are calculated by forming the ratio of the number of events passing the probe criterion and the number of all events either passing or failing the probe criterion. The scale factors are given by the ratio of the efficiencies derived for data and simulation. The separate scale factors for electrons and muons are derived and applied as a function of transverse momentum and pseudo rapidity of the respective lepton. The scale factors corresponding to the three selections are determined in a factorized approach in the order given above and with the leptons under investigation passing the respective previous selection. The scale factors are applied as event weights to simulated events according to the type of lepton occurring. The lepton-efficiency scale factors and the corresponding uncertainties used in this analysis are produced centrally within the CMS collaboration by the E/Gamma and the Muon physics object groups [173, 174].

4.10.3. b-Tagging Output Correction

b-tagging strongly relies on the particles and their trajectories originating from hadronization and the decay of B hadrons. For this reason, it is especially sensitive to the QCD branching and hadronization processes, which are described by the parton shower and subsequent hadronization in simulated collision events. Both procedures rely on approximations and therefore do not provide an exact representation of the original process. As a consequence, the distributions provided by simulation differ from the ones obtained from data. Corrections for these differences are derived based on a tag-and-probe method. They are determined and applied separately for heavy-flavor jets only including jets originating from bottom quarks and light-flavor jets including the jets originating from up quarks, down quarks, strange quarks, and gluons [169][3]. A control region of high purity of either heavy-flavor jets or light-flavor jets is selected. This is accomplished by requiring events

[3]The described correction of the b-tagging output has specifically been developed for the search for $\mathrm{t\bar{t}(H\to b\bar{b})}$ at CMS, as the b-tagging output plays a vital role in the identification of this process. Examples of applications of the b-tagging output are given in Chapter 6 and Chapter 9 of this thesis.

to feature exactly two electrically charged leptons and exactly two resolved jets. By applying requirements on the lepton pair, the missing transverse energy, and the tag jet, the selected sample is either enriched in $\mathrm{t\bar{t}}$ production for the determination of the corrections for heavy-flavor jets or Z+jets production for the determination of the corrections for light-flavor jets.

The corrections are derived from the b-tagging output distribution of the probe jet of events passing the selection. First the event yield of simulation is scaled to the one provided by data. The simulated events are separated into heavy-flavor and light-flavor components and the contribution of the respective other flavor is subtracted from the data. The corrections are given in form of scale factors,

$$SF_{\mathrm{HF/LF}}(\text{b-tag. output}, p_{\mathrm{T}}(,\eta)) = \frac{N(\text{data}) - N(\mathrm{MC_{LF/HF}})}{N(\mathrm{MC_{HF/LF}})} \ ,$$

determined by the ratio between the observed and the simulated event yields in each bin of the b-tagging-output distribution. These scale factors are derived as a function of the transverse momentum of the jet and for the light-flavor scale factors also as a function of the pseudo rapidity of the jet. However, these are not the final scale factors yet as there is an issue. The subtraction of the background from light-flavor jets in the scale-factor determination for heavy-flavor jets and vice versa assumes knowledge about the distributions that are subtracted. For this reason, an iterative approach of the determination of the scale factors is applied. In this approach, the obtained scale factors are applied in the next iteration of scale-factor determination, until they converge. Typically, a sufficient convergence is reached after the third iteration. In a subsequent step, the obtained scale factors are parameterized by fitting a sixth-order polynomial function to the scale factors. This smoothing step reduces the effect of statistical fluctuations.

No dedicated scale factors are obtained for jets from charm quarks due to the lack of a proper control region. As the heavy and light flavor scale factors do not properly describe jets from charm quarks, a scale factor $SF = 1$ is applied with twice the uncertainties of the heavy-flavor scale factors.

The b-tagging correction is applied in form of event weights by multiplying the scale factors obtained for all selected jets in the event,

$$SF_{\text{event}} = \prod_{i}^{N_{\text{jets}}} SF_i \ .$$

In this equation, N_{jets} represents the number of selected jets in the event and SF_i is the scale factor derived for the jet with index i.

Chapter 5

Analysis Techniques

The search for $t\bar{t}(H\to b\bar{b})$ production is a cumbersome task. The overwhelming number of background events featuring a signature very similar to the one of $t\bar{t}(H\to b\bar{b})$ production complicates the extraction of signal events. In order to isolate these events, a machine-learning approach, which exploits different kinematic properties in form of different variables and the correlations between them, is applied. The machine-learning technique applied are boosted decision trees, which are described in more detail in Section 5.1. The interpretation of the outcome of the analysis is accomplished with sophisticated statistical methods. An overview of the ones applied in the search for $t\bar{t}(H\to b\bar{b})$ production are outlined in Section 5.2.

5.1. Boosted Decision Trees

In the search for $t\bar{t}H$ production, signal events are expected to be hidden among events provided by dominating and nearly indistinguishable background processes. A cut-and-count analysis or even a shape analysis based on a single kinematic observable would by far not provide enough sensitivity to detect $t\bar{t}H$ production. A possible solution is provided by multivariate analysis techniques (MVA). MVAs combine the signal and background separation abilities of a set of variables into a single observable. The construction of these observables is based on supervised learning, which aims at an optimal signal and background separation. The supervised learning approach makes use of datasets with well known target properties, in order to train the multivariate analysis method.

The analysis presented in this thesis makes use of boosted decision trees (BDTs) for the discrimination of signal and background events and the identification of boosted hadronically decaying top quarks. BDTs are formed by an ensemble of binary-tree structured classifiers, which are called decision trees. Decision trees and the closely related regression trees, which are applied in the boosting procedure of the BDTs, are introduced in Section 5.1.1. Boosting improves the robustness and the performance of decision trees by training a large set of trees and combining them. The gradient boosting method used in this thesis is described in Section 5.1.2. A training with too many degrees of freedom bears the risk of introducing a bias and overestimating the performance of the boosted decision trees. This effect called overtraining or overfitting is specified in Section 5.1.3. The choice of input variables and configuration parameters for the training of boosted decision trees has a large impact on the discrimination power of the resulting classifier. A method for optimizing this choice, the particle-swarm algorithm, is described in Section 5.1.4.

5.1.1. Decision and Regression Trees

Decision trees and regressions trees are the basic units of the BDT and boosted regression trees (BRT) methods. Both types of trees feature a characteristic binary tree structure

given by a sequence of repeated binary decisions. Each decision is based on a single variable chosen from a set of input variables and a corresponding cut. Based on this decision, a mother node is split into two daughter nodes, which themselves become the mother nodes in the next level of decisions. This procedure results in multiple final nodes, which divide the phase space into hypercubes. The difference between the two types of trees is that decision trees aim at classifying the input, while regression trees assign values of a target variable.

The construction process of decision and regression trees is called growing or training. The procedure is started by choosing a set of training data with known classification or value of the target variable value. The entire dataset is regarded as first node of the tree, which is split into two nodes by the first decision. The optimal variable and corresponding cut for this decision are found by applying a splitting criterion. The variable ranges of all input variables are scanned with a granularity determined by N_{cuts} and each cut is rated with regard to the splitting criterion. The best performing cut is chosen to split the mother node into the two daughter nodes. For decision trees, possible splitting criteria are designed to provide the best separation between different classes of events, which are in the following referred to as signal and background. Most of the decision tree splitting criteria are based on the purity defined by sum of signal event weights divided by sum of all events,

$$P = \frac{\sum_{i \in \mathrm{S}}^{N_\mathrm{S}} w_i}{\sum_{i \in \mathrm{S}}^{N_\mathrm{S}} w_i + \sum_{i \in \mathrm{B}}^{N_\mathrm{B}} w_i} \; .$$

In this equation, N_S and N_B are the number of signal and background events respectively. The parameter w_i corresponds to the weight of the event with index i. The splitting criterion used for the growing of decision trees in this analysis is constructed from the Gini index [175],

$$\text{gini} = P(1-P) \sum_{i}^{N} w_i \; .$$

The splitting criterion is the maximum change in the Gini index, when comparing the mother node m and the daughter nodes d_1 and d_2,

$$\Delta\text{gini} = \text{gini}_\mathrm{m} - \text{gini}_{\mathrm{d}_1} - \text{gini}_{\mathrm{d}_2} \; .$$

For regression trees, splitting criteria are based on the difference between the true value of the regression target variable for each event i, y_i, and the mean of this variable in the respective node $\hat{y}$. A suited quantity for constructing a splitting criterion is given by the average squared deviation,

$$\sigma^2_{\text{avg}} = \frac{1}{N} \sum_{i}^{N} (y_i - \hat{y})^2 \; .$$

The sum runs over all training events in the respective node. As for the decision tree, the actual splitting criterion is based on the difference in the average squared error for the mother node m and the daughter nodes d_1 and d_2,

$$\Delta\sigma^2_{\text{avg}} = \sigma^2_{\text{avg,m}} - \sigma^2_{\text{avg,d}_1} - \sigma^2_{\text{avg,d}_2} \ . \tag{5.1}$$

The optimal splitting is provided by the combination of the input variable and the cut that maximizes this difference.

For the next level of decisions, the daughter nodes are split by determining best performing decisions within these nodes. Following this procedure, the tree is grown by successively finding best performing decisions and splitting nodes. The tree growing is stopped when a termination condition is fulfilled. Typical termination conditions are a maximum number of final nodes, pure final nodes, or a minimum number of events in one node. In this analysis, the tree growing is stopped, if a maximum number of decision levels, also referred to as maximum depth, is reached. In case of a decision tree, the final nodes, also called leaves, are associated to a classification category. Background nodes are assigned the value -1, while signal nodes are assigned the value $+1$. For a regression tree, the final nodes are assigned specific values of the target variable given by the mean of the target variable of the training events in this node. A major difference with respect to decision trees is the size of the trees. Regression trees are chosen to be larger, in order to describe the continuous target variable as accurately as possible.

5.1.2. Boosting

Single decision trees do not feature an outstanding performance and are prone to statistical fluctuations of the training sample. Fluctuations can influence the choice of a splitting and therefore alter the structure of the tree. In order to counter this effect, boosting methods are introduced. Instead of training a single decision tree, multiple decision trees, often referred to as a forest, are grown in an sequential order from the same input samples. The sequential training of decision trees enables the use of information from the previously trained trees. The fundamental idea underlying this approach is the creation of a slow learning procedure that combines the output of an increasing number of single decision trees. The parameters that steer the learning process are the number of trees which are grown, N_{trees}, and the shrinkage λ. The latter determines the contribution of individual trees in the combination and therefore determines the learning speed. In the following, the boosting method used in this analysis, gradient boosting, is described.

Gradient Boosting

Gradient boosting is based on the combined output $F_{M-1}(x)$ of the $M-1$ decision trees at step M of the boosting procedure. This combined output is the weighted sum of the outputs of all previously grown decision trees. Given the input dataset x of the size N with true values y, a loss function is defined that quantifies the deviation between the combined response of the $M-1$ decision trees $F_{M-1}(x_i)$ and true value y_i. For the gradient boosting implementation in the ROOT package TMVA [176, 177], which is used in this analysis, a binomial log-likelihood loss function,

$$L(F_{M-1}(x), y) = \sum_i^N \ln\left(1 + e^{-2F_{M-1}(x_i)y_i}\right) \ ,$$

has been chosen. The goal of the gradient boosting method is to minimize the loss function by adjusting the parameters. However, the influence of the tree structure on the loss

function makes an analytical solution impossible. Instead, the minimization is performed by using the steepest-descent approach. This approach starts by determining a starting point given by a constant value $F_0(x) = \gamma$ that minimizes the loss function. At each following iteration, this approach starts by deriving the gradient of the loss function, which includes the $M - 1$ decision trees of the previous iterations. The structure of the new tree for the current iteration is determined by training a regression tree using the loss function gradient as target. The values assigned to the final nodes of this tree are chosen to minimize the loss function, which includes the new tree. In the course of consecutive iterations, the loss function successively approaches its minimum. The speed of this process is defined by the shrinkage parameter λ. Hence, the quality of the outcome is determined by both the shrinkage and the termination criterion, which in this case is the maximum number of trees N_{trees}. By reducing the learning rate, the method becomes more robust against overtraining, but requires a larger number of trees to reach similar accuracy.

Another method for making gradient boosting more robust against statistical fluctuations of the training sample is bagging. This method relies on a resampling of the training sample. This is accomplished by picking a random subset of the training data for growing a decision or regression tree. The fraction of the original training sample included in the sampled subset is steered by the bagging-fraction parameter. Typical values are chosen to be between 0.5 and 0.8.

A more detailed description of the gradient-boosting method can be found in [176, 177].

5.1.3. Overtraining

For excessively complex models, like a BDT featuring a large number of nodes, the machine-learning method potentially interprets fluctuations of the training sample as features of the underlying probability density. This effect, called overtraining, especially arises if the number of adjustable parameters largely exceeds the number of training events. In this case, the fluctuations of the training sample are learned, which results in a bias. An overtrained BDT will perform worse on independent datasets based on the same probability density, as the trained fluctuation are only true for the dataset used for training. This behavior is utilized to check for overtraining. Before the training of the BDT the data sample is split into two parts, the training and the test sample. The training of the BDT is performed using the training sample. In a subsequent step, the output of the BDT is evaluated with each of the two samples independently. In case of overtraining, the BDT-output distributions provided by the two samples would differ significantly. This effect is quantified by applying the Kolmogorov-Smirnov (KS) test for two samples. The KS test returns a measure of the probability of the distributions to originate from the same underlying probability distribution. The test is based on the empirical distribution function,

$$F_n(x) = \frac{1}{n} \sum_{i}^{n} I_{[-\infty,x]}(x_i) \ ,$$

for n observations x_i. It can be interpreted as a cumulative fraction function of a distribution. In this function, $I_{[-\infty,x]}(x_i)$ is the indicator function, which returns one if x_i is smaller than x and zero otherwise. The empirical distribution functions of training and test sample, $F_{1,n}(x)$ and $F_{2,n'}(x)$, are quantitatively compared by the KS test statistic for two samples,

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)| \; . \tag{5.2}$$

The test statistic is given by the largest deviation of the empirical distribution functions of both samples. The final quantity used for judging the occurrence of overtraining is the KS probability,

$$p(z) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2z^2} \qquad \text{with} \quad z = D_{n,n'}\sqrt{\frac{nn'}{n+n'}} \; . \tag{5.3}$$

The KS probability is the probability for the test statistic to exceed the measured value z given the null hypothesis. The null hypothesis assumes that the two distributions stem from the same underlying distribution. Very small KS probabilities indicate the occurrence of overtraining. Output distributions originating from BDTs with only minor or no overtraining provide probabilities uniformly distributed between zero and one.

A cure for overtraining is the reduction of the total number of nodes by reducing the maximum number of decision trees or the maximum depth.

5.1.4. Particle-Swarm Optimization

The best choice of configuration parameters and input variables for the training of a BDT is ambiguous. Rough estimations about reasonable configurations can be made, based on plausibility arguments. Nevertheless, the optimal combination remains unknown and deviations from the ideal configuration can cause a drop in performance and an increase in overtraining.

In this analysis, the search for an optimal configuration for the training of BDTs is performed by applying the particle-swarm optimization (PSO) [178,179]. This algorithm is specialized in the optimization of non-linear functions. The PSO is an iterative procedure, which optimizes a problem with respect to a chosen measure of quality. It is based on simple entities, called particles, which move in a search space spanned by the parameters that are to be optimized. The search for an ideal set of parameter values is performed by multiple particles at the same time, a so-called swarm. Each of the particles in the swarm features a position corresponding to a certain BDT configuration and a velocity. Throughout the optimization process, the velocity of each particle is influenced by its own best position found so far, but also by the global best position found among all particles in the swarm. By these means, the search space is not just randomly scanned, the information provided by the whole swarm is used to approach the overall best position. The optimization of a BDT training is executed in a four-dimensional search space, spanned by the configuration parameters introduced in the previous subsections:

- Number of cuts N_{cuts},
- Shrinkage parameter λ,
- Number of trees N_{trees},
- bagging fraction.

The maximum depth of the decision trees is not further optimized. This value is fixed to two, which is a standard choice for BDTs with gradient boosting and a large number of trees. Additional degrees of freedom are provided by the choice of the input variables. This choice is optimized by systematically removing and adding variables to the collection of input variables during each iteration of the particle-swarm algorithm. The position in the search space and the input variable configuration is rated by the integral of the receiver operator characteristic (ROC), which compares the signal and background selection efficiency for all possible cuts on the output distribution of the BDT. In order to avoid overtraining, configurations resulting in a KS probability below a certain threshold are vetoed. The KS probability requirement chosen for the optimization of BDTs in this analysis is $p(z) > 0.1$.

The particle swarm algorithm is initialized by creating a swarm of $N_{\text{particles}}$ particles with random positions $\vec{x}_i$, random velocities $\vec{v}_i$, and random subsets of all input variables. The position of each particle is identified as its best position for the time being. The position providing the best ROC integral among all particles and satisfying the KS probability requirement is considered the global best position. After the setup procedure, the iteration starts by testing various input-variable combinations. As a first step for each iteration, the performance of the current subset of input variables is determined. To test for better input variable combinations, the current set of input variables is ranked by removing a single variable one at a time and evaluating the decrease of the ROC integral. The worst performing variable is removed from the input-variable collection. In a subsequent step, variables currently not used are tested by adding them to the set of input variables one at a time and evaluating the change in performance. If the largest increase in performance exceeds a set threshold, the corresponding variable is added to the set of used input variables. Upper and lower boundaries on the number of used input variables ensure a reasonable amount of variables in the current set. Each training for a given configuration is performed multiple times with differently split training and test samples to maintain stability against fluctuations. In this case, the smallest values of the ROC integrals and KS probabilities for the current configuration are used for rating its performance. After the performance of each particle is evaluated, the velocities and the positions are updated based on the local and the global best positions,

$$\vec{v}_i' \leftarrow w_{\text{I}}\vec{v}_i + w_{\text{P}}U(0,1)(\vec{p}_i - \vec{x}_i) + w_{\text{G}}U(0,1)(\vec{g} - \vec{x}_i) \ ,$$
$$\vec{x}_i' \leftarrow \vec{x}_i + \vec{v}_i' \ .$$

In this equation, $U(0,1)$ describes a random value uniformly distributed between zero and one. The parameter w_{I} weights the contribution of the current velocity to the newly determined velocity, hence represents the inertia of the particle. The weights w_{P} and w_{G} determine the influence of the particle's best position and the overall best position of all particles. For this analysis, the weight values $w_{\text{I}} = 0.729$, $w_{\text{P}} = 1.495$, and $w_{\text{G}} = 1.495$ have been chosen, as they have resulted in good performance of the particle swarm algorithm in past studies [180, 181]. The procedure described above is iterated until a termination criterion is fulfilled. In this analysis, the termination criterion is a maximal number of iterations.

5.2. Statistical Methods

Experimental particle physics includes the search for processes predicted by the Standard Model or beyond the Standard Model theories. The application of statistical methods for the interpretation of the outcome of the searches is crucial, in order to determine the statistical significance of an observed signal or to exclude a theoretical prediction. This section introduces the statistical methods used in the search for $t\bar{t}(H\to b\bar{b})$ production, which is presented in this thesis. Section 5.2.1 starts with describing parameter estimation based on the maximum-likelihood method. A test rating the validity of two hypotheses, the likelihood-ratio test, is introduced in Section 5.2.2. The CL_s method, a method for setting exclusion limits on the signal strength of a given process is explained in Section 5.2.3. A corresponding simplified approach for the determination of expected exclusion limits, the asymptotic method, is covered in Section 5.2.4.

5.2.1. Maximum-Liklihood Method

The inference of parameters of a theoretical model given a measured dataset is a common task in particle-physics analyses. According to the likelihood principle, all information relevant to a parameter in a given data sample is contained in the likelihood function [182, 183]. Following this principle, the maximum-likelihood method is the best approach for finding an optimal parameter set $\vec{a}$ based on a data set with N observations $\vec{x}_1, ..., \vec{x}_N$. The construction of the likelihood function requires the probability-density functions (PDF) $f(\vec{x} \mid \vec{a})$. The PDFs return the probability to measure $\vec{x}$ given the parameter set $\vec{a}$ and are by definition normalized to unity. In particle-physics analyses, these functions are mostly derived from theoretical predictions, like simulation. The likelihood function is the product of the PDFs for each observation $\vec{x}_i$,

$$L(\vec{a} \mid \vec{x}_1, ..., \vec{x}_N) = f(\vec{x}_1 \mid \vec{a}) \cdot f(\vec{x}_2 \mid \vec{a}) ... f(\vec{x}_N \mid \vec{a}) = \prod_i^N f(\vec{x}_i \mid \vec{a}) \ .$$

Contrary to the original PDFs, the likelihood can be interpreted as a measure of the likeliness of a particular parameter set $\vec{a}$ given the observations $\vec{x}_1, ..., \vec{x}_N$. Nevertheless, the likelihood function is no probability density in $\vec{a}$. In practice, often the negative logarithm of the likelihood function is used.

The likelihood function resembles a Gaussian function and the negative logarithm approximates a parabola. The optimal set of parameters, the likelihood estimate, can be found at the extrema of the likelihood functions. The limits of the confidence interval are given by the parameter values for which the likelihood function has decreased by a factor $e^{0.5}$ or the negative logarithm has increased by 0.5. The two standard-deviations interval is marked by a likelihood-function decrease with a factor of e^2 or an increase of the negative of two.

Often, likelihood functions depend on multiple parameters but only one is of major interest. In this case, the number of parameters can be reduced by expressing the parameters remaining in terms of the one of interest. This is accomplished by determining the values of the parameters remaining that maximize the likelihood function for a given value of the parameter of interest, also referred to as conditional maximum-likelihood estimators. The resulting function is called the profile likelihood function.

5.2.2. Likelihood-Ratio Test

The likelihood function does not only enable the estimation of an optimal set of parameters, but also allows to compare the validity of two hypothesis. Two hypothesis $H_0 : \vec{a} = \vec{a}_0$ and $H_1 : \vec{a} = \vec{a}_1$ are defined with a distinct set of parameters $\vec{a}_0$ and $\vec{a}_1$. Given a set of observations $\vec{x}_1, ..., \vec{x}_N$, a ratio of two likelihood functions,

$$\Lambda(\vec{x}) = \frac{\mathcal{L}(\vec{x} \mid \vec{a}_0)}{\mathcal{L}(\vec{x} \mid \vec{a}_1)} ,$$

can be formed. The likelihood-ratio test rejects hypothesis H_0 if the likelihood ratio lies below a chosen threshold η. According to the Neyman-Pearson lemma [184], the likelihood-ratio test is the most powerful hypothesis test at a significance level α for a threshold η. The significance level α is determined by calculating the p-value,

$$\alpha = P(\Lambda(\vec{x}) \leq \eta \mid H_0) = \int_{-\infty}^{\eta} f(\Lambda(\vec{x}) \mid H_0) \, \mathrm{d}\Lambda ,$$

which is the probability for measuring a likelihood ratio η or smaller given the hypothesis H_0.

5.2.3. CL_s Exclusion Limits

Some particle-physics searches provide not enough sensitivity to observe a signal excess even if one existed. A common way of dealing with such cases at the LHC is the calculation of CL_s exclusion limits [185–188]. The determination of CL_s exclusion limits is a modified frequentist approach. The short introduction of the method given in the following is based on the published description of the procedure for LHC Higgs-boson searches [189]. More details on this method can be found in this publication and in the references stated above.

The determination of CL_s exclusion limits aims at setting upper limits on the signal-strength modifier $\mu = \sigma/\sigma_{\text{pred.}}$, which varies the cross section predicted by theory $\sigma_{\text{pred.}}$, at a particular confidence level. The first step for calculating the CL_s exclusion limits is the construction of a likelihood function. The likelihood function is based on the number of observed events n and the number of predicted events $\mu s + b$, where s is the number of signal events and b is the number of background events. Depending on the application, the observed events are either given by experimentally measured data or pseudo-data. The number of predicted signal and background events underlies uncertainties caused by several sources. In the calculation of the CL_s exclusion limits, these uncertainties are considered by introducing a set of nuisance parameters θ. The number of predicted signal events $s(\theta)$ and the number of predicted background events $b(\theta)$ are functions of these nuisance parameters. The degree of belief in the default value of the nuisance parameter $\tilde{\theta}$ is parametrized by the PDFs of the systematic uncertainties $p(\theta \mid \tilde{\theta})$. Applying Bayes' theorem the PDFs can be expressed as $p(\theta \mid \tilde{\theta}) \sim p(\tilde{\theta} \mid \theta) \cdot \pi_\theta(\theta)$. The PDFs $p(\tilde{\theta} \mid \theta)$, which are mostly given by a normal or log-normal distribution, can be further re-formulated, while keeping the prior probability $\pi_\theta(\theta)$ flat.

Given these prerequisites, the likelihood function is constructed in the following way,

$$\mathcal{L}(\text{data} \mid \mu, \theta) = \text{Poisson}(\text{data}|\mu s(\theta) + b(\theta)) \cdot p(\tilde{\theta} \mid \theta) .$$

In this likelihood function, θ represents the entire set of nuisance parameters. The uncertainty PDFs $p(\tilde{\theta} \mid \theta)$ provide constraints on the measurement of μ. In case of binned data with N statistically independent bins i, the term Poisson(data$|\mu s(\theta) + b(\theta)$) is a product of Poisson probabilities,

$$\text{Poisson}(\text{data}|\mu s(\theta) + b(\theta)) = \prod_i^N \frac{(\mu s_i + b_i)^{n_i}}{n_i!} e^{-\mu s_i - b_i} \ .$$

In searches for new-physics signals, the null hypothesis H_b is the background-only scenario and the alternative hypothesis $H_\text{s+b}$ is the signal+background case. The compatibility of these hypotheses with observed data is compared using a profile-likelihood ratio test statistic,

$$\tilde{q}_\mu = -2 \ln \frac{\mathcal{L}(\text{data} \mid \mu, \hat{\theta}_\mu)}{\mathcal{L}(\text{data} \mid \hat{\mu}, \hat{\theta})} \qquad \text{with} \quad 0 \leq \hat{\mu} \leq \mu \ . \tag{5.4}$$

In this equation, $\hat{\theta}_\mu$ is a set of conditional maximum-likelihood estimators of the set of nuisance parameters θ. Hence, they are the optimal values of θ that maximize the likelihood function for a given value of μ. The values of the parameters $\hat{\mu}$ and $\hat{\theta}$ globally maximize the likelihood function. The constraint $\mu \geq 0$ ensures that only physical positive values of the signal strength are considered. The calculation of CL_s exclusion limits requires a one sided confidence interval expressed by $\hat{\mu} \geq \mu$. The confidence levels for the signal+background hypothesis, which features a variable μ parameter, and the background-only hypothesis, which features $\mu = 0$, are determined by calculating their p-values,

$$p_\mu = P(\tilde{q}_\mu \geq \tilde{q}_\mu^\text{obs} \mid H_\text{s+b}) = \int_{\tilde{q}_\mu^\text{obs}}^\infty f(\tilde{q}_\mu \mid \mu, \hat{\theta}_\mu^\text{obs}) \, \text{d}\tilde{q}_\mu \ ,$$
$$1 - p_\text{b} = P(\tilde{q}_\mu \geq \tilde{q}_\mu^\text{obs} \mid H_\text{b}) = \int_{\tilde{q}_\mu^\text{obs}}^\infty f(\tilde{q}_\mu \mid 0, \hat{\theta}_0^\text{obs}) \, \text{d}\tilde{q}_\mu \ .$$

In these equations, $\tilde{q}_\mu^\text{obs}$ is the observed test statistic for a given value of the signal strength modifier μ. The conditional maximum-likelihood estimators $\hat{\theta}_0^\text{obs}$ and $\hat{\theta}_\mu^\text{obs}$ maximize the likelihood function given the observed data for a given μ and $\mu = 0$. The probability-density functions for measuring a test statistic $\tilde{q}_\mu$, $f(\tilde{q}_\mu \mid \mu, \hat{\theta}_\mu^\text{obs})$ and $f(\tilde{q}_\mu \mid 0, \hat{\theta}_0^\text{obs})$, are computed by generating pseudo data with a Monte Carlo approach. The p-values reflect the probability for measuring the observed test statistic or a larger value given the respective hypothesis. The p-values represent the confidence level for a given value of μ in the signal+background hypothesis and for $\mu = 0$ in the background-only hypothesis. An important feature of the CL_s exclusion limit is that any conclusions about the discovery or exclusion of a signal based on fluctuations of the background are avoided. This is achieved by normalizing the confidence level of the signal+background hypothesis to the confidence level of the background-only hypothesis,

$$\text{CL}_\text{s}(\mu) = \frac{\text{CL}_\text{s+b}(\mu)}{\text{CL}_\text{b}(\mu)} = \frac{p_\mu}{1 - p_\text{b}} \ .$$

The resulting CL_s approximates the confidence level obtained in case of the complete absence of background. For the case of a predicted signal strength $\mu = 1$ and an obtained value $CL_s = \alpha$, the predicted signal would be considered as excluded at a $(1 - \alpha)$ CL_s confidence level. In particle physics searches, the standard confidence level for excluding a signal is chosen at 95 %. The corresponding 95 % CL_s confidence level upper limit on the signal-strength modifier is the value of μ corresponding to $CL_s = 0.05$.

The observed 95 % CL_s confidence level upper limit $\mu^{95\,\%\ CL_s}$ does not bring much insight without knowing the sensitivity of the analysis. The sensitivity of the analysis is quantified by the expected upper limit based on the background-only scenario. The determination of expected limits starts by generating numerous sets of pseudo data. There are three different types of expected limits, which differ in the way the sets of pseudo data are derived. The first type of expected limits is commonly used for the presentation of final results of measurements and is based on a background-only hypothesis and the observed data. For the calculation of these expected limits, the nuisance parameters and their uncertainties are determined by fitting a background-only model ($\mu = 0$) to observed data. The pseudo data is randomly generated from the fitted background-only model by taking into account the fitted uncertainties of the nuisance parameters and statistical fluctuations. Signal-injected expected limits are derived in an analogue way by adding the predicted signal contribution ($\mu = 1$) to the fitted background-only model in the generation of the pseudo data. The determination of blinded expected limits fully avoids the use of observed data by generating pseudo data from the unfitted background-only model and nuisance parameters. For each pseudo dataset, the 95 % CL_s confidence level upper limit $\mu^{95\,\%\ CL_s}$ is calculated by substituting the observed data with the generated pseudo-data. In the distribution of all $\mu^{95\,\%\ CL_s}$ obtained, the expected median upper limit is marked by the 50 % quantile. The one standard deviation boundaries can be found at the 16 % and 84 % quantiles, while the two standard deviation boundaries are located at the 2.5 % and 97.5 % quantiles.

5.2.4. Asymptotic Limits

The calculation of CL_s exclusion limits is a computationally very intensive procedure. However, if the physical requirement on the signal strength modifier $\mu \geq 0$ is dropped, the profile-likelihood test statistic can be approximated in case of large sample sizes [190]. Wilk's theorem [191] concludes that for large sample sizes the likelihood ratio is asymptotically described by a non-central χ^2 distribution with the degrees of freedom given by the difference in dimensionality of the likelihood parameters. In case of the test statistic $\tilde{q}_\mu$, the difference in dimensionality of the parameters is one. The approximation for one degree of freedom is

$$\tilde{q}_\mu = -2 \ln \frac{\mathcal{L}(\text{data} \mid \mu, \hat{\theta}_\mu)}{\mathcal{L}(\text{data} \mid \hat{\mu}, \hat{\theta})} = \frac{(\mu - \hat{\mu})^2}{\sigma^2} + \mathcal{O}\left(\frac{1}{\sqrt{N}}\right) . \tag{5.5}$$

The parameter $\hat{\mu}$ is described by a Gaussian distribution with mean μ' and standard deviation σ. The size of the data sample is represented by N. When neglecting terms of order $\mathcal{O}(1/\sqrt{N})$, the corresponding probability-density function is

$$f(\tilde{q}_\mu;\Lambda) = \frac{1}{2\sqrt{\tilde{q}_\mu}}\frac{1}{\sqrt{2\pi}}\left[\exp\left(-\frac{1}{2}\left(\sqrt{\tilde{q}_\mu}+\sqrt{\Lambda}\right)^2\right)+\exp\left(-\frac{1}{2}\left(\sqrt{\tilde{q}_\mu}-\sqrt{\Lambda}\right)^2\right)\right] ,$$
$$\Lambda = \frac{(\mu-\mu')^2}{\sigma^2} .$$

For the determination of Λ and σ, the Asimov dataset is introduced. It is the dataset that returns the true parameter values when evaluating the maximum-likelihood estimators. For this special dataset, Eq. (5.4) can be expressed as

$$\tilde{q}_{\mu,\mathrm{A}} = \frac{(\mu-\mu')^2}{\sigma_\mathrm{A}^2} = \Lambda .$$

The expected median upper limits on μ and the corresponding standard deviations are determined for the assumption of the presence of no signal. For this scenario, the true parameter value for the signal strength modifier is $\mu' = 0$ and σ_A^2 can be determined with

$$\sigma_\mathrm{A}^2 = \frac{\mu^2}{\tilde{q}_{\mu,\mathrm{A}}} .$$

With these simplifications at hand, the upper limits can be determined in an analytical way instead of using computer intensive Monte Carlo methods. However, this approach is only valid as long as the term $\mathcal{O}(1/\sqrt{N})$ in Eq. (5.5) can be neglected. Nevertheless, it has been found that the asymptotic approximation holds accurate results even for fairly small sample sizes.

Part II.

Search for $t\bar{t}(H \to b\bar{b})$ in the Single-Lepton Channel

Chapter 6

Resolved-Event Selection and Reconstruction

The first step in the search for $t\bar{t}H$ production is to reject as many events as possible that clearly originate from background processes. This is accomplished by retaining only those events for further analysis that fulfill dedicated criteria aiming at the special properties of $t\bar{t}H$ events. The $t\bar{t}H$ search presented in this thesis is based on the $t\bar{t}H$-decay channel featuring a Higgs-boson decay into a bottom-quark pair and a semileptonic decay of the top-quark pair. As described in Section 1.3.3, the resulting final state is characterized by a prompt electrically charged lepton and a large number of quarks with many of them being bottom quarks. In conventional $t\bar{t}H$ events, where top quarks and Higgs bosons feature low to moderately large transverse momenta, the large recoil caused by the large masses of the massive particles causes the decay quarks to hadronize spatially well-separated. In this case the jet reconstruction, presented in Section 4.7, results in a number of non-overlapping, resolved jets corresponding to the number of strongly interacting final-state particles. This type of events featuring low massive-particle momenta will in the following be referred to as resolved. The resolved-event selection, which is based on the described signature, selects events with a reconstructed charged lepton and a minimum number of resolved jets and b-tags. In a subsequent step, the selected events are divided into categories of different resolved jet and b-tag multiplicities. A more detailed description of the event selection is given in Section 6.1.

In the search for ttH production, a strong ingredient for the discrimination of signal against background processes is the Higgs boson. The information on the Higgs boson in $t\bar{t}(H\rightarrow b\bar{b})$ events is not easy to grasp due to the many strongly interacting particles in the final state and the resulting ambiguity in the reconstruction of a Higgs-boson candidate. A dedicated reconstruction of resolved events aiming at the extraction of the Higgs-boson information is based on the reconstruction of the $t\bar{t}$ system in a first step and the subsequent identification of the Higgs boson. The reconstruction of the $t\bar{t}$ system is approached by defining reconstruction hypotheses based on the assignment of jets to the expected decay products of the top-quark pair. The best reconstruction hypothesis is determined based on a measure of quality calculated for each hypothesis. A more detailed description of the entire resolved-event reconstruction procedure is given in Section 6.2. For a proper analysis, the data has to be well described by the simulated processes. In order to ensure that this is the case, the agreement of data and simulation is checked in well-defined control regions. The corresponding studies are presented in Section 6.3.

The treatment of events characterized by massive particle with large transverse momenta is described in Chapter 8.

6.1. Resolved-Event Selection and Categorization

The analysis presented in this thesis performs a search for $t\bar{t}H$ production with the Higgs boson decaying into a bottom-quark pair and a semileptonic decay of the top-quark pair. The Higgs-boson decay into a bottom quark features the largest branching fraction of all Higgs-boson decay channels. However, at a hadron collider, like the LHC, a search for this decay faces an overwhelming number of strongly interacting particles that stem from QCD processes and mimic the signature of the Higgs-boson decay. The semileptonic decay of the top-quark pair brings a characteristic that distinguishes the signal from a large part of the background, the prompt electrically charged lepton of the leptonic top-quark decay. Pure QCD processes are expected to feature only soft leptons originating from the decay of hadrons. The hadronic top-quark decay, on the other side, features a larger branching fraction compared to the leptonic decay. Correspondingly, the semileptonic decay of the top-quark pair represents a compromise between a clean signature with a handle on a large fraction of background events and a large branching fraction.

The selection of resolved events is performed based on the reconstructed and selected objects described in Chapter 4. In case of a fully resolved event, every strongly interacting particle in the final state gives rise to a single jet. Consequently, for the final state of a $t\bar{t}(H\rightarrow b\bar{b})$ event with a semileptonic $t\bar{t}$ decay without any additional hard radiation the following reconstructed objects are expected:

- One isolated electrically charged lepton candidate,
- Six reconstructed resolved jets,
- Four b-tags,
- Missing transverse energy.

Based on this signature, the event selection is performed by first requiring events to pass the single-lepton triggers described in Section 3.1.1. Further, a reconstructed primary vertex fulfilling the quality criteria described in Section 4.9.1 is required. Another criterion for the selection of an event is the presence of exactly one isolated electron or muon fulfilling the primary-lepton selection criteria presented in Section 4.9.2. Taus are not considered in this analysis as the reconstruction of these particles is complicated and not as efficient as for muons and electrons. Events featuring additional electrically charged leptons fulfilling the veto-lepton selection criteria are rejected. Signatures with multiple leptons are analyzed in dedicated analyses, like the dilepton channel of the $t\bar{t}(H\rightarrow b\bar{b})$ search or the multilepton $t\bar{t}H$ search. By introducing the mentioned veto on additional leptons, an unambiguous assignment of events to the different $t\bar{t}H$ searches is ensured and the double counting of events in the combination of the different search channels is avoided.

As stated above, in an optimal and fully resolved $t\bar{t}(H\rightarrow b\bar{b})$ event with a semileptonic $t\bar{t}$ decay, six jets with four of them being identified as originating from bottom quarks are expected. However, this number is altered by additional jets coming from initial and final state radiation and the loss of jets due to acceptance and selection effects. A further effect, which varies the observed numbers, is the merging of jets. A possible cause for this behavior are original massive particles with large transverse momenta, which is covered in more detail in Chapter 7 and Chapter 8. A limited b-tagging efficiency for jets stemming from bottom quarks and a non-zero b-tagging efficiency for jets not stemming from bottom quarks lead to a varying number of b-tags. In order to account for these effects and to

raise the sensitivity of this analysis, the events are divided into analysis categories based on their jet and b-tag multiplicities. In this analysis, the following seven resolved analysis categories are considered:

- $\geq$6 jets, $\geq$4 b-tags category
- 5 jets, $\geq$4 b-tags category
- 4 jets, $\geq$4 b-tags category
- $\geq$6 jets, 3 b-tags category
- 5 jets, 3 b-tags category
- 4 jets, 3 b-tags category
- $\geq$6 jets, 2 b-tags category

This set of categories was chosen based on the expected event yield for the signal process. The event yields of each category for simulated and recorded events corresponding to an integrated luminosity of $\mathcal{L} = 2.7\ \text{fb}^{-1}$ are visualized in Fig. 6.1. $\text{t}\bar{\text{t}}(\text{H}\rightarrow\text{b}\bar{\text{b}})$ events are mostly characterized by their large number of jets and b-tags. Accordingly, the $\geq$6 jets, $\geq$4 b-tags category, which requires the largest number of jets and b-tags and corresponds to an optimal resolved configuration of a $\text{t}\bar{\text{t}}(\text{H}\rightarrow\text{b}\bar{\text{b}})$ event, provides the best ratio of signal and background events. The largest background contribution is given by the process that produces a similar multiplicity of jets and b-tags, $\text{t}\bar{\text{t}}$+jets production. Especially $\text{t}\bar{\text{t}}+\text{b}\bar{\text{b}}$ events are very likely to feature a matching number of b-tags in addition to a matching jet multiplicity and therefore can hardly be further reduced by placing requirements on these values. The exact numbers of observed and simulated events in each category are presented after the introduction of the boosted analysis category in Chapter 8.

6.2. Resolved-Event Reconstruction

In this analysis, $\text{t}\bar{\text{t}}\text{H}$ production is the only relevant process featuring a real Higgs-boson. In background processes, only fake Higgs-boson candidates are reconstructed, for example, from gluon splittings into bottom-quark pairs and combinatorial background, which originates from random combinations of particles stemming from other sources in the event. For this reason, a reconstructed Higgs-boson candidate and its properties are some of the most important ingredients for the discrimination of $\text{t}\bar{\text{t}}\text{H}$ signal events from those of background processes. However, the large number of final-state particles forming jets makes the identification of the decay products of the Higgs boson a complicated task. A common strategy is the reconstruction of the decay products of the top-quark pair by means of the distinctive properties of the top-quark pair decay. The Higgs-boson decay products are identified from the objects in the event that are not used for the reconstruction of the top-quark pair. In the following, a reconstruction for resolved $\text{t}\bar{\text{t}}\text{H}$ signatures based on the minimal χ^2 method is presented.

The $\text{t}\bar{\text{t}}\text{H}$ decay with the Higgs boson decaying into a bottom-quark pair and a semileptonic top-quark decay features one charged lepton, a neutrino, four bottom quarks, and two light quarks in the final state. The electrically charged lepton originating from the leptonically decaying top quark is provided by the reconstructed electrically charged lepton required by the resolved-event selection. The neutrino stemming from the decay of

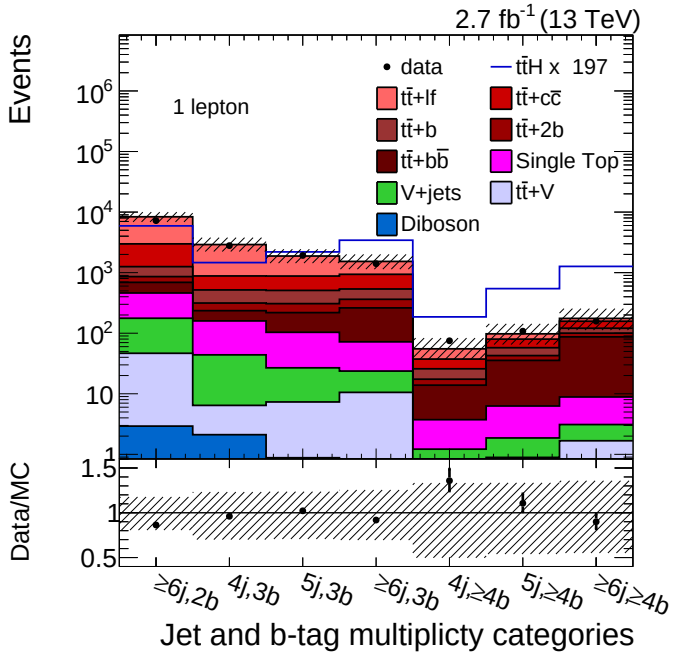


Figure 6.1: Event yields of simulated and recorded events for each resolved analysis category. Simulated background processes are displayed as stacked, filled histograms and scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7\ \mathrm{fb}^{-1}$. The $\mathrm{t\bar{t}H}$ signal is depicted as a blue line and scaled to the total predicted event yield of all backgrounds for better visibility. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions.

the leptonic top-quark decay is reconstructed based on the missing transverse energy. The longitudinal component of the four-vector is calculated by fixing the invariant mass of the system given by the neutrino and the electrically charged lepton to the W-boson mass $m(W) = 80.4\ \mathrm{GeV}/c^2$. The hadronic part of the event provides the major challenge of the reconstruction. The association of a jet to its original final-state particle is ambiguous. This problem is handled by defining hypotheses based on different assignments of jets to the strongly interacting $\mathrm{t\bar{t}}$ decay products. For an ideal and resolved $\mathrm{t\bar{t}H}$ event with six jets and four b-tags and without further constraints, this would add up to 360 hypotheses. As the exact association of two resolved jets to the two decay quarks of the hadronically decaying W boson is not relevant, this number is reduced by a factor two. By additionally requiring the bottom quarks from the top-quark decay to be only assigned b-tagged jets, the number of hypotheses becomes 72. Each of the reconstruction hypotheses is assessed by calculating a χ^2 value,

$$\begin{aligned}\chi^2 = &\frac{(m_{\mathrm{hyp}}(\text{had. top}) - m_{\mathrm{MC}}(\text{had. top}))^2}{\sigma^2_{m_{\mathrm{MC}}(\text{had. top})}} \\ &+ \frac{(m_{\mathrm{hyp}}(\text{lep. top}) - m_{\mathrm{MC}}(\text{lep. top}))^2}{\sigma^2_{m_{\mathrm{MC}}(\text{lep. top})}} \\ &+ \frac{(m_{\mathrm{hyp}}(\text{had. W}) - m_{\mathrm{MC}}(\text{had. W}))^2}{\sigma^2_{m_{\mathrm{MC}}(\text{had. W})}}\ ,\end{aligned}$$

Table 6.1: Reconstruction efficiency of the Higgs boson, the hadronically decaying top quark, or both particles in simulated $t\bar{t}H$ events selected by the resolved $\geq$6 jets, $\geq$4 b-tags category. Correct reconstructions are defined by an angular matching of the reconstructed candidates to the simulated massive particles.

Reconstruction efficiency of		
Had. top quark [%]	Higgs boson [%]	Had. top quark & Higgs boson [%]
34.3	29.2	15.8

based on the difference between the top-quark and W-boson masses reconstructed with the given hypothesis and the true values. The invariant masses of the reconstructed candidates are determined based on the vectorial sum of the momentum four-vectors of resolved jets and leptons associated to the respective decay products. The true values are derived from simulated $t\bar{t}H$ events by matching reconstructed resolved jets to the simulated decay products and determining the invariant masses of the massive particles using the resolved jets matched to their decay products. The hypothesis with the lowest χ^2 value is considered the best reconstruction. The described reconstruction procedure requires at least four resolved jets with two of them being b-tagged. A Higgs-boson candidate is reconstructed from the remaining resolved jets not used in the reconstruction of the top-quark pair. From this set, the two b-tagged resolved jets with the largest transverse momenta are assigned to the bottom quarks from the Higgs-boson decay. If not enough b-tagged resolved jets are available, missing b-tagged jets are substituted by non-b-tagged jets. The invariant mass of the Higgs-boson candidate is not used in the reconstruction in order not to bias this variable for the final discrimination of signal against background.

The efficiency for correctly reconstructing the Higgs boson, the hadronically decaying top quark, or both particles using the described procedure is studied based on simulated $t\bar{t}H$ events selected by the resolved $\geq$6 jets, $\geq$4 b-tags analysis category. In this study, correct reconstructions are defined by an angular matching of the simulated massive particles to the respective reconstructed candidates requiring an angular distance of $\Delta R < 0.5$. The results of this study are presented in Table 6.1. The observed reconstruction efficiencies, which are found to be very small, show the magnitude of the combinatorial problem arising in the reconstruction of $t\bar{t}(H\rightarrow b\bar{b})$ events. A correct reconstruction of only the hadronically decaying top quark or the Higgs boson is achieved in about $\sim$ 30 % of the selected events. Both massive particles are correctly identified at the same time in only about half of these events. As already mentioned, a major cause for the low number of correct reconstructions is the large number of final state particles, which leads to a large number of hypotheses. Additionally, the χ^2 values of different hypotheses do not hint clearly at the correct reconstruction, as the masses of correctly reconstructed massive particles are smeared by the limited resolution of the jet energies. The production of additional resolved jets by initial or final-state radiation enhances the number of wrong hypotheses and hence the possibility for choosing an incorrect reconstruction. Another effect limiting the reconstruction efficiency is the loss of jets stemming from the decay products of the massive particles by not being reconstructed correctly or not passing the resolved-jet selection. The misidentification of light jets as jets stemming from bottom quarks and vice versa are further sources of incorrect reconstructions.

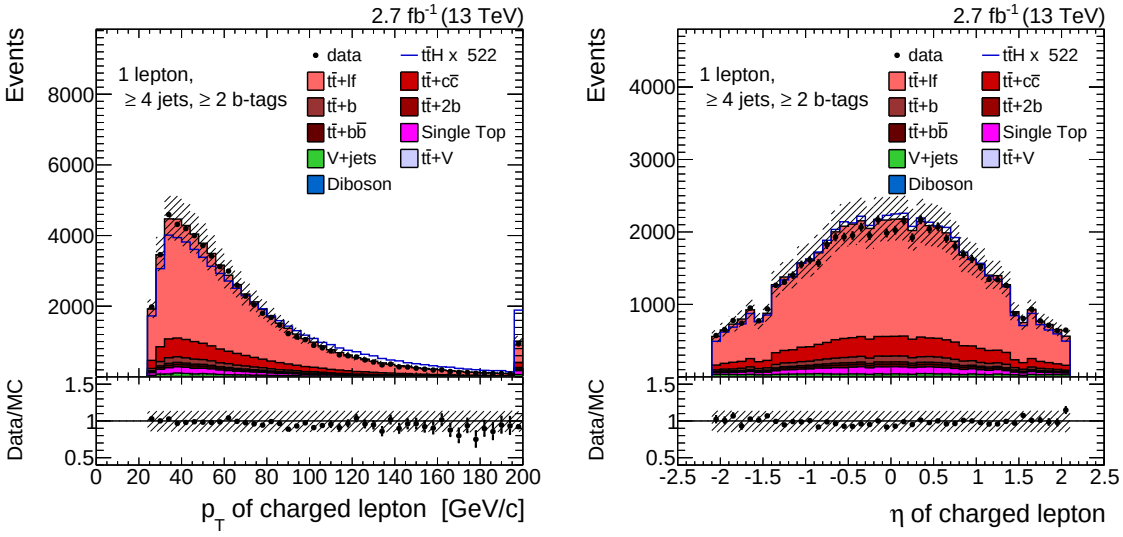


Figure 6.2: Transverse momentum (left) and pseudo rapidity (right) of selected electrically charged leptons shown for an inclusive control region requiring one selected lepton, at least four resolved jets, and at least two b-tags. Simulated background processes are displayed as stacked, filled histograms and are scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7$ fb^{-1}. The $t\bar{t}H$ signal is depicted as a blue line and scaled to the total predicted event yield of all backgrounds for better visibility. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions.

6.3. Validation

Insufficient description of the background processes by simulation potentially causes a bias in the determination of the final results. An underestimation of background, for example, introduces a difference between the event yields observed for data and simulation, which might be misinterpreted as originating from the signal process. In order to avoid such cases, the agreement between data and simulation is tested in control regions, where no major contribution of signal is expected. In the search for $t\bar{t}(H\to b\bar{b})$ production, the main background is given by $t\bar{t}$ production. Accordingly, the description of the recorded data by simulation is mainly checked in a control region enriched in events from this process. This control region is defined by the baseline selection described in Section 6.1 without the categorization being applied. Instead, at least four resolved jets and at least two b-tags are required, which corresponds to the expectation for a $t\bar{t}$ event featuring a semileptonic decay.

First, the kinematic distributions of the selected electrically charged leptons are checked. Fig. 6.2 shows the transverse momentum and the pseudo rapidity of muons and electrons representative for all kinematic variables of the leptons. In both distributions, a very good agreement between data and simulation is found for the normalization as well as for the shape.

A first comparison of data and simulation for jet variables is presented in Fig. 6.3. There, the multiplicity of resolved jets and b-tags is illustrated. Obvious differences between data and simulation can be seen for the jet multiplicity, where simulation predicts a number of jets in selected collision events that is larger than the one observed in data.

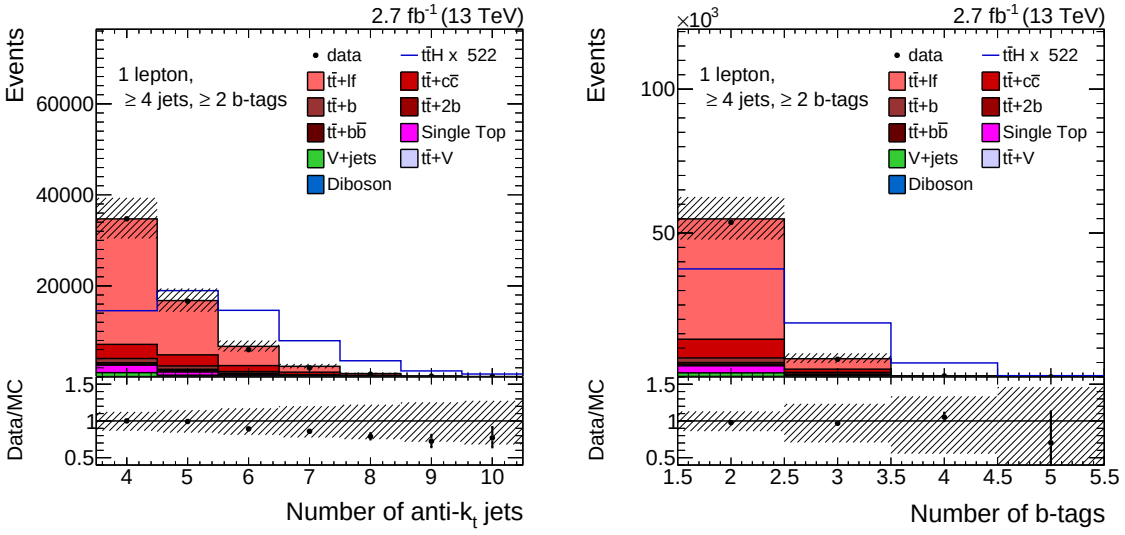


Figure 6.3: Resolved jet (left) and b-tag multiplicity (right) shown for an inclusive control region requiring one selected lepton, at least four resolved jets, and at least two b-tags. Simulated background processes are displayed as stacked, filled histograms and are scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7$ fb^{-1}. The $t\bar{t}H$ signal is depicted as a blue line and scaled to the total predicted event yield of all backgrounds for better visibility. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions.

Nevertheless, the discrepancies are still covered by the uncertainties. Accordingly, the nuisance parameters should be able to balance the discrepancies in the determination of the final results presented in Chapter 11. The distribution showing the multiplicity of b-tags shows good agreement between data simulation. Still, the slight differences observed in this distribution combined with the discrepancies in the jet multiplicity lead to the differences in event yields between data and simulation observed in the resolved analyses categories, which are shown in Fig. 6.1. Further, the kinematics of the jets are investigated. Fig. 6.4 shows the transverse momentum of all jets in the event and the pseudo rapidity of the hardest jet. Both distributions show a very good description of the data by simulation. The b-tagging output distribution of all jets in the event are displayed in Fig. 6.5. Again, good agreement between data and simulation is observed. The slight fluctuations in the center of the distribution are caused by pile-up effects.

The distributions shown in this section are only a representative selection of the distributions that have been checked for this analysis. This investigation also comprises other control regions enriched in events originating from minor backgrounds. In summary, a good description of data by simulation has been found. Further distributions of the control region presented in this section are attached in Appendix A.1.1.

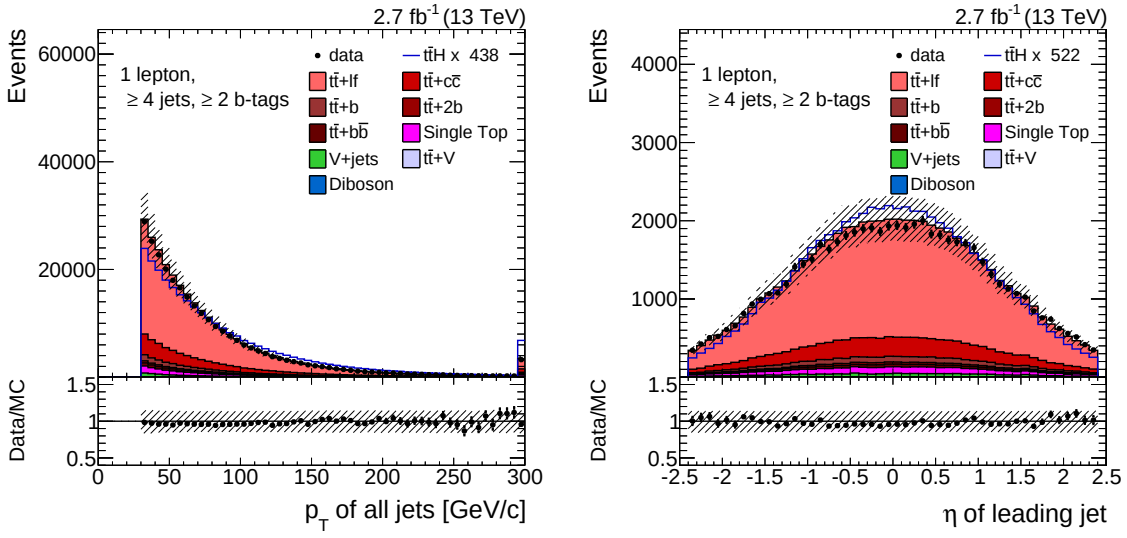


Figure 6.4: Transverse momentum of all jets in the event (left) and pseudo rapidity of of the hardest jet (right) shown for an inclusive control region requiring one selected lepton, at least four resolved jets, and at least two b-tags. Simulated background processes are displayed as stacked, filled histograms and are scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7$ fb^{-1}. The t$\bar{\text{t}}$H signal is depicted as a blue line and scaled to the total predicted event yield of all backgrounds for better visibility. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions.

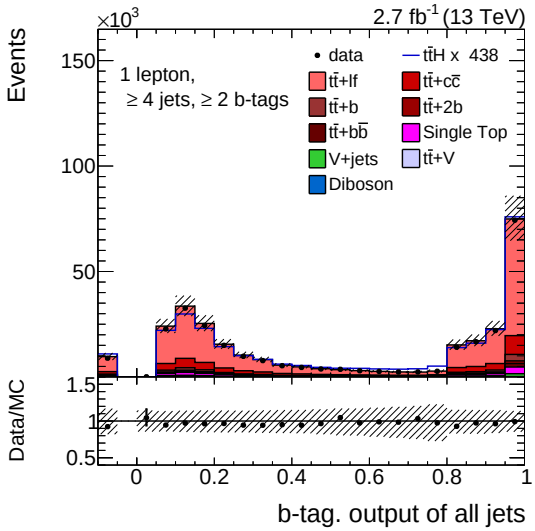


Figure 6.5: b-tagging output of all jets in the event shown for an inclusive control region requiring one selected lepton, at least four resolved jets, and at least two b-tags. Simulated background processes are displayed as stacked, filled histograms and are scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7$ fb^{-1}. The t$\bar{\text{t}}$H signal is depicted as a blue line and scaled to the total predicted event yield of all backgrounds for better visibility. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distribution.

Chapter 7

Boosted Objects

The reconstruction of boosted objects aims at massive particles with large transverse momenta decaying into strongly interacting particles. When decaying, these particles pass their momentum to the decay products, which form collimated showers of hadrons. This type of topologies is mostly beyond being resolvable with standard jet reconstruction algorithms. Yet, such configurations bear the advantage that all decay products are locally accumulated instead of being spread out in all directions. Specialized clustering and substructure algorithms do not only allow the analysis of boosted objects, these algorithms also make use of the collimated decay products in the reconstruction of the massive particles. The application of these dedicated algorithms results in large reconstruction efficiencies for massive particles with large transverse momenta. In most cases, the reconstruction efficiencies achieved exceed the ones reached in the reconstruction of fully resolved events, as it is demonstrated in this thesis. A reason for this are the combinatorics, which are reduced in the boosted-object reconstruction. In the resolved reconstruction, the ambiguous assignment of jets to the decay products of massive particles leads to a huge source of incorrect identification possibilities.

An example of a boosted reconstruction algorithm is the HEPTopTagger algorithm [3], which is designed for the reconstruction of boosted hadronically decaying top quarks. It was initially developed for the search of $t\bar{t}H$ production, but has never been applied for this purpose before the analysis described in this thesis. The algorithm was introduced to solve the mentioned combinatorial problem that arises in the reconstruction of resolved $t\bar{t}H$ events as described in Section 6.2. A further application of the HEPTopTagger was studied in the search for the supersymmetric partner of the top quark, the top squark [192]. During the first LHC run the HEPTopTagger has mainly been used in searches for heavy resonances decaying into top quarks [193]. With a large fraction of the mass of the heavy resonance being transformed into kinetic energy during its decay, searches for these processes seem like the prime example of the application of boosted analysis techniques. Compared to heavy resonance processes, $t\bar{t}H$ events feature only moderately boosted top quarks and Higgs bosons. Nevertheless, the heavy particles in $t\bar{t}H$ events tend to feature larger transverse momenta than the heavy particles produced by other SM processes. This distinctive feature is illustrated in Fig. 7.1 showing the transverse momenta of top quarks, Higgs bosons, and additional bottom-quark pairs from gluon splitting coming from $t\bar{t}H$, $t\bar{t}$, and WH events. Another advantage of using boosted analysis techniques in the search for $t\bar{t}H$ production is the reason the HEPTopTagger has initially been developed for: the solution of the combinatorial problem in the reconstruction of $t\bar{t}H$ events, which is covered in Section 8.1.

The HEPTopTagger and other boosted object reconstruction algorithms mostly feature similar procedures. The first step, which is described in Section 7.1, is the clustering of fat jets with a configuration chosen to cluster all decay products of the massive particle into one jet. Yet, the resulting fat jets do not exclusively contain the decay products

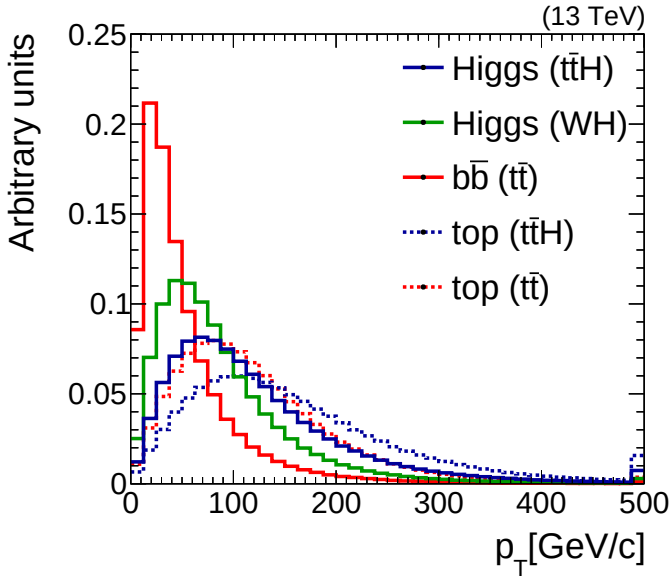


Figure 7.1: Transverse momentum of hadronically decaying top quarks (dashed), Higgs bosons (solid), and bottom-quark pairs from gluon splittings (solid). The particles are taken from simulated $t\bar{t}H$ events (blue), $t\bar{t}$ events (red), and WH events (green) generated with POWHEG+PYTHIA8.

of the massive particle, but also comprise contributions due to pile-up, the underlying event, and initial state radiation. Contamination alters the scale and the resolution of the reconstructed object's momentum and energy, which hinders the identification and analysis of boosted massive particles. The substructure algorithms described in Section 7.2 aim at the removal of contamination and the uncovering of the distinctive features of the boosted massive particle. Section 7.3 covers algorithms like the HEPTopTagger, which are specialized in the reconstruction of particular massive particles. These algorithms combine substructure algorithms to further improve the reconstruction of boosted particles.

7.1. Fat-Jet Clustering

The clustering of fat jets aims at merging all decay products of boosted massive particles into a single object. The fat jets used in this analysis are clustered with the Cambridge/Aachen algorithm. As already discussed in Section 4.7.2, this algorithm clusters objects solely based on their angular distance. The corresponding clustering sequence resembles the sequential ordering of the parton splitting process, which is a crucial feature for obtaining meaningful results by the declustering algorithms described in Section 7.2.1. A similar behavior is provided by the k_T algorithm but not by the anti-k_T algorithm. As the Cambridge/Aachen algorithm has some advantages in fat-jet clustering and substructure investigation compared to the k_T algorithm, like the fact that the fat-jet mass is less prone to soft radiation [194], it has been chosen for the clustering of the fat jets in this analysis.

An important parameter in the clustering of the fat jets is the cone-size parameter R. This parameter needs to be chosen large enough to cluster all decay products of a boosted massive particle into a single jet. The distances between the decay products depend on

the type of the decay, the mass, and the transverse momentum of the massive particle. A simple example is the two-body decay of a Higgs boson into two bottom quarks. In this case, the angular distance between the two bottom quarks is approximately given by

$$\Delta R_{\mathrm{bb}} \simeq \frac{1}{\sqrt{z(1-z)}} \frac{m_{\mathrm{H}}}{p_{\mathrm{T}}} \qquad \text{with} \quad (p_{\mathrm{T}} \gg m_{\mathrm{H}}) \; .$$

In this equation z and $1-z$ are the momentum fractions of the two bottom quarks. As $0 > z > 1$ applies, the term $1/\sqrt{z(1-z)}$ is always above 2. For fixed z, R_{bb} shows an $1/p_{\mathrm{T}}$ behavior. The combination of both characteristics can be observed in the left-hand plot of Fig. 7.2, which shows the angular distance between the bottom quarks from Higgs-boson decay in simulated $\mathrm{t\bar{t}}(\mathrm{H}\to\mathrm{b\bar{b}})$ events as a function of the transverse momentum of the Higgs boson. Strongly asymmetrical values for z and $1-z$, which correspond to large R_{bb} at a given value of p_{T}, are suppressed. The right-hand plot in this figure shows an analogue distribution for top quarks from simulated $\mathrm{t\bar{t}}(\mathrm{H}\to\mathrm{b\bar{b}})$ events. However, the distance between the decay products of hadronically decaying top quarks displayed is an effective distance based on the Cambridge/Aachen clustering. It is defined as the angular distance between the vectorial sum of the momentum vectors of the closest decay products a and b fulfilling

$$\Delta R(\vec{p}_a, \vec{p}_b) = \min(\Delta R(\vec{p}_i, \vec{p}_j)) \; ,$$

and the momentum vector of the remaining decay product c given by the expression

$$\Delta R_{\mathrm{bqq}} = \Delta R((\vec{p}_a + \vec{p}_b), \vec{p}_c) \; . \tag{7.1}$$

In these equations, the indices i, j and a, b, and c can represent any of the decay products of a hadronically decaying top quark. The characteristics of the distribution for the top quark are similar to the one observed for Higgs-bosons. However, the distribution is broadened and smeared out by the effects of the three-body decay of the top quark.

Based on Fig. 7.1, a boosted phase-space region starting at a fat-jet transverse momentum of 200 GeV/c has been chosen for this analysis.[1] Given this cut and the distributions shown in Fig. 7.2, the cone-size parameter used for the clustering of fat jets is set to $R = 1.5$. This choice ensures that the majority of boosted Higgs bosons and a large fraction of boosted top quarks in the kinematic region starting at a transverse momentum of 200 GeV/c are clustered into one fat jet. Larger cone-size-parameter values increase the number of fat jets not stemming from boosted massive particles. An increased cone size would also enhance the clustering of particles from other sources into the same fat jet. Furthermore, larger cone-size parameters would imply an expansion into the phase space, where the individual jets from the decay products would be fully resolvable. An initial approach to the implementation of the reconstruction of boosted particles in $\mathrm{t\bar{t}H}$ events has shown a significant decrease in the efficiencies of reconstructing boosted massive particles when using a smaller cone size. This effect has led to a reduction of the performance of

[1] This choice of the boosted phase-space region starting at 200 GeV/c is rather unusual for analyses performed in the boosted regime, which typically investigate boosted objects with transverse momenta starting at 400 GeV/c and above. Examples are given in [195] and [196]. Due to the fact that massive particles in $\mathrm{t\bar{t}H}$ events are only moderately boosted a relatively low transverse momentum threshold was chosen.

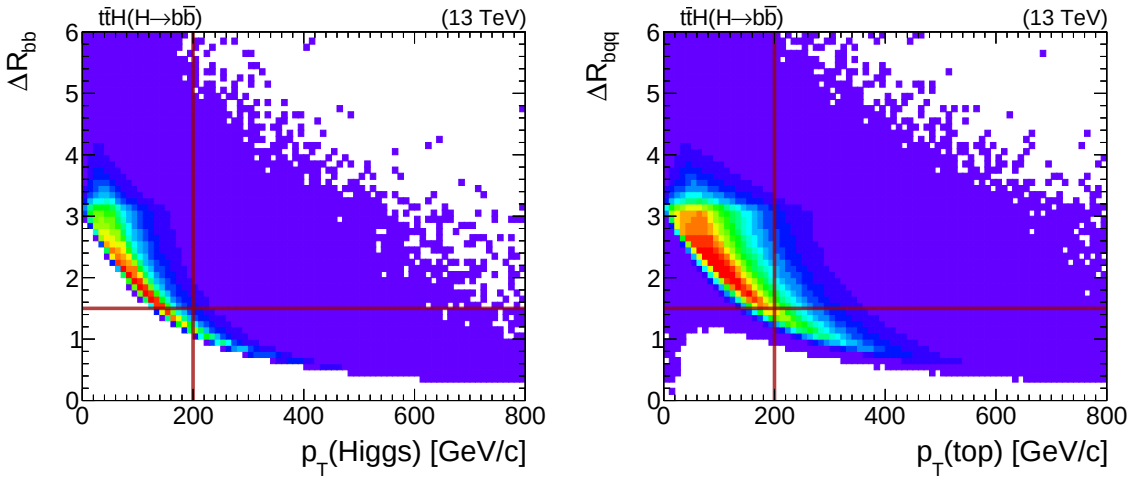


Figure 7.2: Distance of the decay products of Higgs bosons decaying into two bottom quarks (left) and hadronically decaying top quarks (right) in the η-ϕ plane differential in the transverse momentum of the decaying massive particle. Densely populated phase-space regions are displayed in red, sparsely populated phase-space regions are displayed in blue. The massive particles are taken from simulated $\mathrm{t\bar{t}H}$ events at a center-of-mass energy of 13 TeV. The effective distance ΔR_{bqq} used to describe the distance between the three decay products of the top quark is defined by Eq. (7.1) in the text.

the dedicated analysis category that makes use of the boosted object reconstruction. A detailed description of this analysis category is presented in Chapter 8.

The input for the clustering of the fat jets are the particle-flow candidates, obtained as described in Section 4.5. For the reconstruction of boosted massive particles, only hadronic decays are considered. In this analysis, prompt electrically charged leptons are only expected to originate from the leptonic top-quark decay. For this reason, isolated electrically charged leptons fulfilling the loose lepton selections described in Section 4.9.2 are omitted in the clustering.

7.2. Substructure Algorithms

The environment of proton-proton collisions and especially of $\mathrm{t\bar{t}H}$ events is very "busy". In addition to particles stemming from the hard interaction, particles originating from various other sources, like pile-up, the underlying event, and initial state radiation, can be found in the final state. Even though a major part of this contamination is removed by the selection and cleaning steps, the fat jets remain prone to these effects, which is due to their large cone size. Impurities clustered into fat jets hide the distinctive features of massive-particle decays, as the distributions of reconstructed observables are washed out. In order to obtain more information about the process underlying the particles clustered into the fat jet, substructure algorithms are applied. These algorithms aim at removing the contamination and extracting the substructure of the fat jet.

In the following sections, some of the algorithms used for the investigation of the substructure of jets are introduced. The algorithms are divided into two subgroups. In Section 7.2.1, the declustering algorithms, which undo the clustering history of the fat jets and remove soft constituents, are discussed. Jet grooming techniques, which are described

in Section 7.2.2, rely on the reclustering of the fat-jet constituents with a modified clustering configuration. Additionally, the variable N-Subjettiness, which parameterizes the substructure within a fat jet, is introduced in Section 7.2.3.

7.2.1. Declustering Algorithms

Declustering algorithms are based on the iterative decomposition of a fat jet. The first iteration starts by using the fat jet as mother jet. In each iteration, the last step of the fat-jet clustering process forming the given mother jet is undone. A criterion, which depends on the declustering algorithm applied, tests the resulting two daughter jets, also called subjets, and the mother jet and determines the subsequent step of the algorithm. If this criterion is not fulfilled, the subjet with the lower invariant mass is interpreted as contamination and discarded. In this case, the remaining daughter jet is declared as the mother jet in the next iteration of the declustering. If the criterion is fulfilled, on the other hand, there are multiple ways to proceed. For the reconstruction of two-prong decays, the algorithm mostly stops at this point and identifies the two subjets as the decay products of a massive particle. In other cases, the algorithm mainly continues with declustering both subjets until a cutoff criterion is fulfilled. Such a cutoff criterion can, for example, be a lower threshold on the mass of subjets or a certain number of subjets.

Declustering algorithms aim at removing wide-angle soft radiation and the identification of hard objects originating from the decay of a massive particle. In order to do so, they depend on the cluster history resembling the sequential ordering of the parton-splitting process. Therefore reasonable results can only be expected when applying declustering algorithms on fat jets clustered either with the Cambridge/Aachen algorithm or the k_T algorithm.

Mass-Drop Declustering

Mass-drop declustering [6] aims at the decrease of the invariant mass of the two individual subjets with respect to the mother jet, when splitting the decay products of a massive particle. Like for all declustering algorithms, the first step in each iteration is splitting the mother jet j into two daughter subjets j_1 and j_2 by undoing the last step of the clustering history. The two subjets are labeled according to their invariant mass, where the more massive subjet is denoted by j_1 and the remaining one by j_2. The second step of each iteration is to check if the mass-drop criterion,

$$m_{j_1} < \mu m_j, \tag{7.2}$$

is fulfilled. The parameter μ represents the mass-drop threshold as a fraction of the invariant mass of the mother jet j. Its value is chosen based on the mass and the decay type of the massive particle, for which the reconstruction is optimized. If Eq. (7.2) is not fulfilled, subjet j_2 is considered soft radiation, not originating from the massive-particle decay, and is discarded. In this case, subjet j_1 is declared the mother jet j for the next iteration and the declustering is continued. Based on the application of the algorithm, the declustering is continued or stopped if the criterion is fulfilled.

Soft-Drop Declustering

Soft-drop declustering [197] aims at removing wide-angle soft radiation. In the first step of each iteration the mother jet j is split into two daughter subjets j_1 and j_2 based on

the last step of the clustering history. The two subjets are labeled according to their transverse momentum where the harder one is j_1 and the softer is j_2. The second step of each iteration is to check if the soft-drop criterion,

$$\frac{p_{\mathrm{T},j_2}}{p_{\mathrm{T},j_1} + p_{\mathrm{T},j_2}} > z_{\mathrm{cut}} \left(\frac{\Delta R_{j_1,j_2}}{R_0} \right)^{\beta} , \tag{7.3}$$

is fulfilled. In this equation, z_{cut} is the soft-drop threshold, which determines the transverse momentum of particles to be removed. The soft-drop threshold plays an equivalent role as the mass-drop threshold μ in mass-drop declustering. The parameter R_0 represents the cone size used for the clustering of the fat jet. The exponent β determines the influence of the angular distance of the subjets. For $\beta \to \infty$, the last term of Eq. (7.3) becomes zero, as $\Delta R_{j_1,j_2} < R_0$, and the algorithm returns the ungroomed jet. The case $\beta = 0$ results in a behavior equivalent to mass-drop declustering. For positive values of β, wide-angle soft radiation is removed, while keeping some of the soft-collinear radiation. This configuration is called grooming mode. It is infrared and collinear safe even for jets with only one constituent. Two separated hard subjets are required to satisfy the soft-drop criterion for negative values of β. This configuration is therefore called the "tagger mode". In this mode, soft-drop declustering can remove both soft and collinear radiation.

As for the mass-drop declustering, the softer of the two subjets is discarded if the soft-drop condition is not fulfilled. In this case, the subjet j_1 is declared the mother jet j for the next iteration and the declustering is continued. If Eq. (7.3) is fulfilled, on the other hand, depending on the application, the soft-drop declustering is stopped or continued with both subjets.

7.2.2. Jet Grooming

Jet grooming represents a further way of cleaning contamination from fat jets and uncovering the underlying substructure. The algorithms in this category of substructure algorithms rely on reclustering the constituents of the fat jet with a different clustering configuration and applying additional criteria. Unlike the declustering algorithms, which are adapted to the hypothesis of a massive-particle decay, the jet grooming algorithms are completely independent of information on the massive particle. In the following, three different algorithms from this category of substructure algorithms are described.

Filtering and Trimming

Filtering [6] and trimming [198] are grooming techniques, which aim at resolving the fat jet into a finer angular scale. Both algorithms start by reclustering the constituents of the fat jet with a sequential recombination algorithm and a small cone-size parameter. A typical choice of the cone-size parameter used for filtering and trimming is $R = 0.3$.

The reclustering of the fat-jet constituents results in a number of subjets determined by its substructure. While filtering retains only the N subjets with the largest transverse momentum for further analysis, trimming discards all subjets below a chosen transverse-momentum threshold. In this way, the filtering and trimming methods remove soft radiation in form of subjets with small transverse momenta. The degree of grooming is steered by the grooming parameters, the subjet multiplicity N for filtering and the transverse-momentum threshold for trimming.

Pruning

Pruning [199,200] is a technique designed for removing soft and wide-angle radiation. Just like filtering and trimming, pruning is based on the reclustering of the fat-jet constituents. Yet, unlike these algorithms, pruning does not necessarily aim at finding subjets. Instead of discarding soft subjets, pruning removes contamination by vetoing soft and large-angle recombinations during reclustering. The requirements for vetoing a recombination of two constituents j_1 and j_2 with $p_{\mathrm{T},j_1} > p_{\mathrm{T},j_2}$ to a resulting jet j are

$$\begin{aligned} &\frac{p_{\mathrm{T},j_2}}{p_{\mathrm{T},j}} < z_{\mathrm{cut}} \qquad \text{and} \\ &\Delta R_{j_1,j_2} > D_{\mathrm{cut}} \ . \end{aligned}$$

The pruning method is steered by two parameters. The parameter z_{cut} represents a lower threshold for the transverse momentum of the softer constituent with respect to the combined jet. Hence, it determines how soft the constituents may be in order to be recombined. The parameter D_{cut} determines the minimum angular distance for a recombination to be pruned. If both requirements are fulfilled, the constituents are not combined and the softer one is discarded. In all other cases, the two constituents are merged.

If the pruning is performed with the Cambridge/Aachen algorithm, a typical choice for the transverse-momentum threshold is $z_{\mathrm{cut}} = 0.1$. The application of the k_{T}-jet-clustering algorithm requires slightly larger values, e.g. $z_{\mathrm{cut}} = 0.15$, to achieve similar performance. This fact can be explained by the transverse-momentum ordering of the recombinations in the k_{T}-clustering process. Concerning the parameter determining the minimum angular distance for pruning D_{cut}, too small values should be avoided as this would cause the pruning away of particles stemming from the original massive particle. Removing such particles would result in a degradation of the scale of the reconstructed particle's observables, as fractions of energy of the initially produced particle would be dismissed. Pruning with too large values of D_{cut}, on the other hand, would not take full advantage of the procedure, as particles from other sources would not be efficiently removed. A typical choice is $D_{\mathrm{cut}} = 0.5$.

7.2.3. N-Subjettiness

N-subjettiness [201, 202] is an inclusive jet-shape variable investigating the energy-flow properties of fat jets. Unlike the substructure algorithms presented in this section, N-subjettiness is only based on the constituents of the fat jet and does not necessarily depend on any clustering algorithm. N-subjettiness can be interpreted as a form of counting the number of hard subjets inside the fat jet by calculating the deviation of the energy flow from N subjet axes. It is calculated as the sum of the minimum angular distances of all $N_{\mathrm{particles}}$ particles i to the N subjet axes weighted by their transverse momentum,

$$\begin{aligned} \tau_N &= \frac{1}{d_0} \sum_{i}^{N_{\mathrm{particles}}} p_{\mathrm{T},i} \min\{\Delta R_{1,i}, \Delta R_{2,i}, ..., \Delta R_{N,i}\} \ , \\ d_0 &= \sum_i p_{\mathrm{T},i} R_0 \ . \end{aligned} \tag{7.4}$$

In this equation, R_0 represents the cone size used for fat jet clustering. Eq. (7.4) is linear in the particles' transverse momenta, which causes the results to be infrared and collinear safe. In cases with $\tau_N \approx 0$, all of the fat-jet constituents are aligned with the N subjet axes. Hence, the fat jet features N or fewer hard subjets. The other extreme, $\tau_N \gg 0$, implies that a large fraction of the constituents lie away from the N subjet axes. Accordingly, the fat jet features at least $N + 1$ hard subjets.

Due to varying degrees of contamination, the absolute value of N-subjettiness is biased for each fat jet individually. For this reason, the ratio of successive values of N-subjettiness τ_N/τ_{N-1} is better suited for discriminating between different hard subjet multiplicities. The ratio τ_2/τ_1, for example, is a well-performing variable for the identification of two-prong decays, as they appear in hadronic W-boson and Higgs-boson decays. The fraction τ_3/τ_2, on the other hand, is well suited to identify three-prong decays. Examples are hadronic top quarks decays.

A main issue when calculating N-subjettiness values is finding the directions of the N subjets axes. An optimal approach would be the minimization of τ_N over all possible subjet directions. In this case, the values of N-subjettiness would be strictly decreasing with increasing N. However, this approach is computationally intensive. A more practical way of finding the directions of the N subjet axes is reclustering the fat-jet constituents with the k_T-algorithm. For this approach, the clustering is stopped as soon as exactly N subjets are clustered.

7.3. Combined Algorithms

The substructure algorithms introduced in the previous section perform very well at removing contamination and investigating the underlying substructure of fat jets. Still, most of them have individual advantages and a combination may be beneficial for the overall reconstruction performance. Especially when reconstructing a particular type of boosted massive particles, the application and the configuration of the substructure algorithms can be adapted to the particles' properties and its decay.

In the following, two algorithms that represent combinations of basic substructure algorithms are introduced. The already mentioned HEPTopTagger, which is specialized in reconstructing boosted hadronically decaying top quarks, is described in Section 7.3.1. An approach that combines mass-drop declustering and filtering in order to identify the two-prong decays of Higgs bosons is described in Section 7.3.2.

7.3.1. HEPTopTagger and HEPTopTaggerV2

As already mentioned, the very first version of the HEPTopTagger was developed in order to reconstruct and identify boosted hadronically decaying top quarks in the search for $\mathrm{t\bar{t}H}$ production [3]. Over time, small modifications to the algorithm and extra features have been introduced. The best performing changes have been collected and included in the new version of the algorithm, the HEPTopTaggerV2 [4, 5].

The HEPTopTaggerV2 algorithm is applied on fat jets derived from the clustering described in Section 7.1. The first part of the algorithm is a mass-drop declustering step. In this case, the mass-drop declustering does not stop if a mass drop is found. Instead, the declustering proceeds until all subjets either feature an invariant mass below 30 GeV/c^2 or consist of only one constituent. The mass-drop threshold for declustering is set to $\mu = 0.8$. The described configuration results in a number of subjets determined by the substruc-

ture of the fat jet. From this set, the three subjets with the highest transverse momentum are retained for further analysis, while the remaining ones are discarded. Other possible criteria for choosing a triplet of subjets are described in [4, 5], but not applied in this analysis.

The second part of the subjet finding consists of a filtering step, applied to the constituents of the chosen mass-drop subjets. The cone size for the filtering,

$$R_{\text{filt}} = \min(0.3, \Delta R_{j,k}/2) \ ,$$

is determined by calculating the minimum of 0.3 and the halved distances between every subjet pair $\Delta R_{j,k}$. In the filtering step, the five hardest subjets resulting from reclustering are retained. [2] These five subjets are clustered to exactly three subjets j_1, j_2, j_3 using the Cambridge/Aachen algorithm, which are subsequently ordered by transverse momentum. An example of the complete subjet-finding procedure is illustrated in Fig. 7.3.

If the invariant masses of their pairwise combinations m_{12}, m_{23}, m_{13} and the triple combination m_{123} satisfy at least one of the criteria listed in Eq. (7.5), the top quark candidate is considered as tagged. The tagging criteria are visualized in Fig. 7.4. They represent the characteristic A-shape cuts in the plane of $\arctan(m_{13}/m_{12})$ and m_{23}/m_{123} associated with the HEPTopTagger.

$$\begin{aligned}
0.2 < \arctan\left(\frac{m_{13}}{m_{12}}\right) < 1.3 \quad \text{and} \quad R_{\text{min}} < \frac{m_{23}}{m_{123}} < R_{\text{max}} \\
R^2_{\text{min}}\left(1+\left(\frac{m_{13}}{m_{12}}\right)^2\right) < 1 - \left(\frac{m_{23}}{m_{123}}\right)^2 < R^2_{\text{max}}\left(1+\left(\frac{m_{13}}{m_{12}}\right)^2\right) \quad \text{and} \quad \frac{m_{23}}{m_{123}} < 0.35 \\
R^2_{\text{min}}\left(1+\left(\frac{m_{12}}{m_{13}}\right)^2\right) < 1 - \left(\frac{m_{23}}{m_{123}}\right)^2 < R^2_{\text{max}}\left(1+\left(\frac{m_{12}}{m_{13}}\right)^2\right) \quad \text{and} \quad \frac{m_{23}}{m_{123}} < 0.35
\end{aligned} \tag{7.5}$$

The parameters R_{min} and R_{max} are defined as

$$R_{\text{max/min}} = (1 \pm f_{\text{W}}) \ ,$$

with the width of the selection window f_{W} typically being chosen as $f_{\text{W}} = 0.15$.

The analysis presented in this thesis uses a slightly modified version of the HEPTopTaggerV2, which is optimized for this analysis. One part of the modification is the substitution of the tagging criteria by the multivariate tagging approach described in Section 7.6. Another change with respect to the original HEPTopTaggerV2 recipe is the association of the subjets to the top-quark decay products based on their b-tagger output values. [3] The

[2] In earlier versions of the HEPTopTagger, the filtering has been independently applied for all triplets of the subjets emerging from the mass-drop declustering. In this case, the algorithm proceeded with the triplet showing the invariant mass of the combination of the five filtering subjets closest to the top-quark mass. This treatment introduced a bias, which shaped the background towards the region of the true top-quark mass, and was therefore abandoned.

[3] Following the original recipe of the HEPTopTaggerV2, the three subjets are assigned to the products of the hadronic top-quark decay by comparing the invariant mass of all possible subjet pairs to the mass of the W boson. The disubjet combination with the invariant mass closest to the one of the W boson are associated to the W-boson decay products and labeled as W1 and W2 ordered by their transverse momentum. The remaining subjet is assigned to the bottom quark coming from top-quark decay and denoted as B.

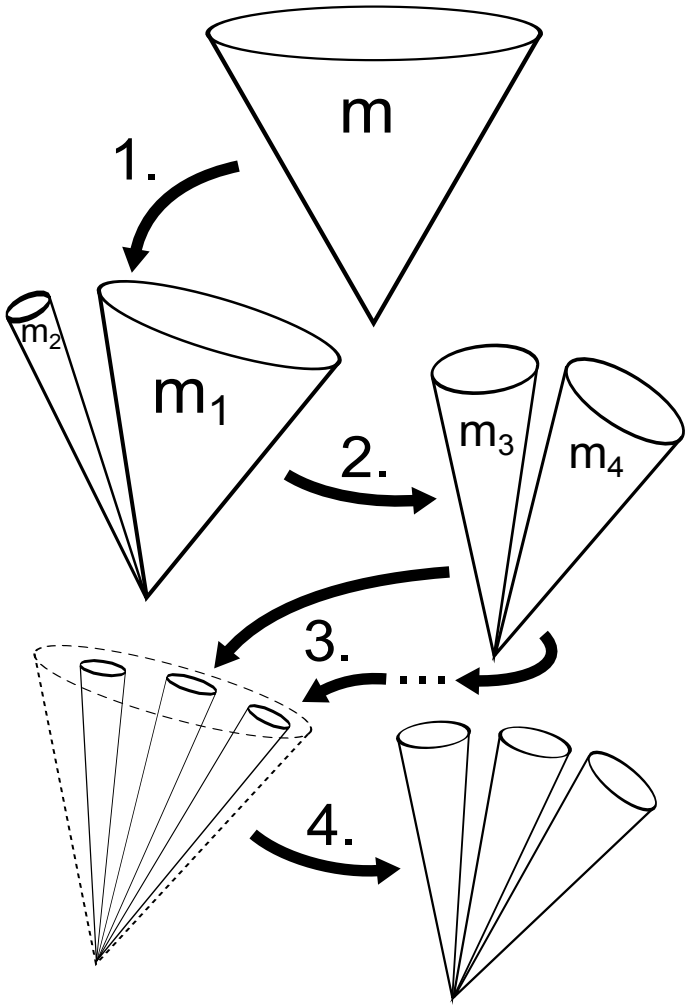


Figure 7.3: Subjet-finding procedure of the HEPTopTagger2 algorithm. The algorithm starts with declustering the fat jet j with mass m and producing the subjets j_1 and j_2 with $m_1 > m_2$ (**1.**). In this example, the mass-drop criterion fails as $m_1 > \mu m$. The algorithm goes on with discarding j_2 and declustering j_1. The two subjets j_3 and j_4 with $m_3 > m_4$ emerge from the declustering of j_1(**2.**). In this case, the mass-drop criterion $m_3 < \mu m_1$ is fulfilled and the algorithm goes on with declustering j_3 and j_4. The mass-drop declustering is stopped as soon as all of the subjets either have an invariant mass below a set threshold or consist of a single constituent. Filtering is applied to the constituents of the three hardest mass-drop subjets and the five hardest subjets resulting from the filtering are chosen (**3.**). The five filtering subjets obtained are re-clustered to exactly three subjets representing the hadronic decay products of the top quark.

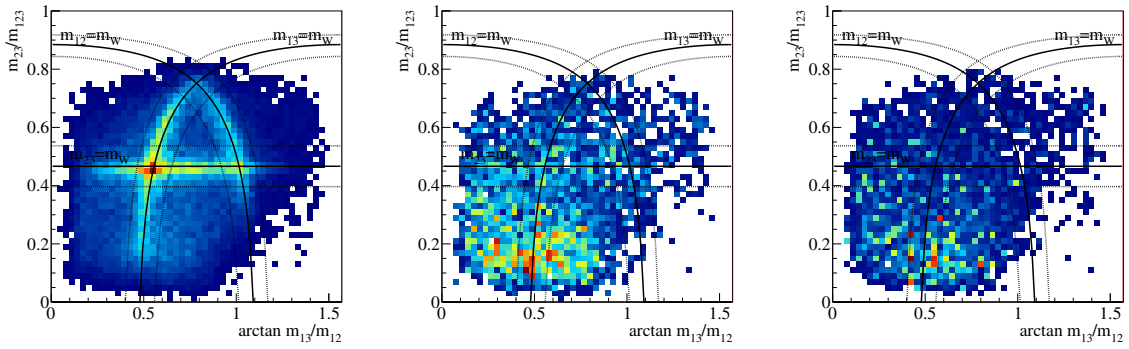


Figure 7.4: Distribution of boosted top-quark candidates reconstructed with the HEPTopTagger algorithm shown in the $\arctan(m_{13}/m_{12})$ vs m_{23}/m_{123} plane. The candidates are taken from simulated $t\bar{t}$ events (left), W+jets events (middle), and QCD-multijet events (right). Phase-space regions with a dense population are displayed in red, whereas sparsely populated regions are displayed in blue. The HEPTopTagger tagging criteria are sketched as dashed lines. Taken from [192].

subjet with the highest b-tagger output value is assigned to the bottom quark and denoted as B. The remaining two subjets are assigned to the W-boson decay products and labeled as W1 and W2 based on their transverse momentum. Both modifications are motivated and described in more detail in Section 7.6.

The boosted top-quark candidate is constructed by combining the three subjets associated to the decay products of the top quark B, W1, and W2.

OptimalR

The optimalR feature of the HEPTopTaggerV2 provides variables discriminating true boosted top quarks from fake candidates based on the expected spatial distance between the top-quark decay products. In an iterative procedure, the size of the initial fat jet is reduced in steps of $\Delta R = 0.1$. In each iteration, the full HEPTopTagger algorithm is applied to the modified fat jet and the invariant mass of the top-quark candidate is calculated. As long as the top-quark decay products are completely clustered into the fat jet, the invariant mass of the resulting top-quark candidate as a function of the cone size forms a plateau. As soon as the fat-jet cone size is too small to cluster all of the top-quark decay products, the invariant mass of the top quark drops significantly. For every iteration, a check for a drop of the invariant mass of the candidate,

$$\frac{m_{\mathrm{rec}}(1.5) - m_{\mathrm{rec}}(R)}{m_{\mathrm{rec}}(1.5)} > 0.2\,,$$

is performed. In this equation, $m_{\mathrm{rec}}(R)$ is the invariant mass of the top-quark candidate for the cone size of the current iteration and $m_{\mathrm{rec}}(1.5)$ is the invariant mass of the top candidate of the initial fat jet. If a mass drop is detected, the cone size of the previous iteration is declared as the optimal R, R_{opt}.

In order to exploit the full potential of the optimalR method, the measured cone size is compared to expectation. The expected optimal cone size $R_{\mathrm{opt(calc)}}$ is defined by

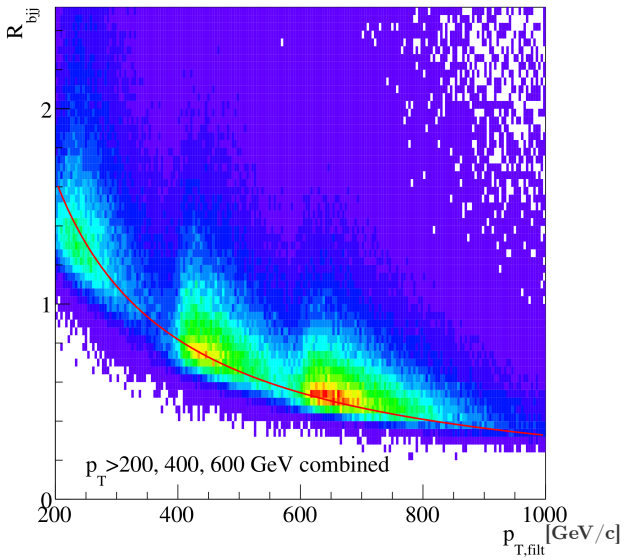


Figure 7.5: Angular distance R_{bqq} of the decay products of the hadronically decaying top quark as a function of the transverse momentum of the filtered fat jet. Filtered fat jets are derived by filtering the constituents of the fat jets with a filtering cone size of $R = 0.2$ and a retained number of subjets of $N = 10$. The top quarks are taken from $\mathrm{t\bar{t}}$ events simulated requiring top-quark transverse momenta of $p_{\mathrm{T},t} > 200, 400, 600$ GeV/c. The red line represents the result of the fit performed for the determination of the expected cone size (Eq. (7.7)). Taken from [5].

$$R_{\mathrm{opt(calc)}} = \frac{327\ \mathrm{GeV}/c}{p_{\mathrm{T,filt}}}. \tag{7.6}$$

In this equation, $p_{\mathrm{T,filt}}$ represents the transverse momentum of the filtered fat jet. In order to derive this variable, filtering is applied to the constituents of the fat jet by using a filtering cone radius of $R = 0.2$ and retaining the ten subjets with the largest transverse momentum. Eq. (7.6) is derived by plotting the angular distance of the top-quark-decay products, R_{bqq}, in simulated $\mathrm{t\bar{t}}$ events against the transverse momentum of the filtered fat jet in a first step. Subsequently, the distribution is fitted with a function,

$$R_{\mathrm{opt(calc)}} \propto \frac{1}{p_{\mathrm{T,filt}}}\ , \tag{7.7}$$

as shown in Fig. 7.5.

In this analysis, the difference between the measured and the calculated R_{opt},

$$\Delta R_{\mathrm{opt}} = R_{\mathrm{opt}} - R_{\mathrm{opt(calc)}}\ ,$$

is used as a variable for distinguishing between correctly reconstructed top quarks and fake candidates.

7.3.2. BDRS Algorithm

The combination of mass-drop declustering and filtering was initially developed for the search of Higgs-boson production in association with a vector boson (VH) [6]. The corresponding algorithm is referred to as BDRS algorithm, which is named after the authors. The algorithm is optimized to identify the two prongs of a Higgs boson decaying into a bottom-quark pair. While for the VH search the algorithm started with clustering fat jets using the Cambridge/Aachen algorithm with a cone size of 1.2, this analysis uses the fat-jet configuration described in Section 7.1. This choice has the advantage of using a single set of fat jets for the reconstruction of boosted Higgs bosons and boosted hadronically decaying top quarks. As already mentioned in Section 7.1, a larger fat-jet cone size also increases the efficiency of reconstructing boosted Higgs bosons.

Similar to the HEPTopTagger, this algorithm starts by applying a mass-drop declustering on the fat jet. The mass-drop threshold chosen in this context is $\mu = 0.67$. This choice results from an optimization. It takes into account that Higgs-boson decays into a bottom-quark pair with additional hard radiation forming a "Mercedes star" configuration in the transverse plane still pass the mass-drop criterion for a threshold value of $\mu \geq 1/\sqrt{3}$. As soon as a mass drop is found, the corresponding two subjets are associated to the two prongs of the Higgs-boson decay and the declustering is stopped.

In order to be retained, the two mass-drop subjets obtained are additionally required to pass the asymmetry condition,

$$y = \frac{\min(p_{\mathrm{T},j_1}^2, p_{\mathrm{T},j_2}^2)}{m_j^2} \Delta R_{j_1,j_2}^2 > y_{\mathrm{cut}} \, .$$

If this condition is not fulfilled, the subjets are discarded and the algorithm terminates without returning any subjets. In order to illustrate this requirement, the asymmetry expression can be approximated by

$$y \simeq \frac{\min(p_{\mathrm{T},j_1}, p_{\mathrm{T},j_2})}{\max(p_{\mathrm{T},j_1}, p_{\mathrm{T},j_2})} \, .$$

Based on this approximation, it becomes clear that the asymmetry condition removes constellations of subjets with too asymmetric transverse momenta. Hence, this criterion aims at removing asymmetric background configurations with high masses caused by soft-gluon emission. In the original BDRS publication [6], an asymmetry cut of $y_{\mathrm{cut}} \simeq 0.15$ has been proposed based on the $\mathrm{S}/\sqrt{\mathrm{B}}$ resulting from the reconstruction of Higgs bosons and light jets. For this analysis, a tighter value $y_{\mathrm{cut}} = 0.3$ was found to perform better, which is due to the large number of final-state particles, which provide a source of misreconstructed Higgs-boson candidates.

The second part of the BDRS algorithm is a filtering step applied on the constituents of the two subjets obtained by the mass-drop declustering. The cone size used for filtering is the minimum of 0.3 and the halved distance between the mass-drop subjets,

$$R_{\mathrm{filt}} = \min\left(0.3, \frac{R_{\mathrm{bb}}}{2}\right) \, .$$

Out of the reclustered jets, only the three hardest subjets are kept. For this analysis, the subjets emerging from the reclustering are first calibrated and selected, as described

in Section 7.4 and Section 7.5, before choosing the hardest three. An example of the subjet-finding procedure is illustrated in Fig. 7.6.

The three subjets are associated to the Higgs-decay products based on b-tagging information. Accordingly, all three subjets are ordered by their b-tagging output value. The two subjets with the largest b-tagging output value are assigned to the bottom quarks from the decay of the Higgs boson. They are labeled *B1* and *B2* ordered by their transverse momentum. The remaining subjet is assumed to be final-state radiation radiated of the $\mathrm{b\bar{b}}$ system and is labeled *G*. Nevertheless, in this analysis the boosted Higgs candidate is constructed by combining only the subjets *B1* and *B2* associated to the bottom-quark pair.

7.4. Boosted-Object Calibration

Similar to the anti-k_{T} jets with a cone size of $R = 0.4$, which have been introduced in Section 4.7, the energies of the reconstructed fat jets and subjets are biased. One part of this bias is caused by particles originating from pile-up interactions that have not been removed by the charged-hadron subtraction described in Section 4.7.1. Another part of the bias can be explained by energy loss due to undetected particles and as non-uniformity in detector response. Differences in the behavior of jets from measured data and simulation also result in different reconstructed energies. The reconstructed boosted-particle candidates are more robust against contamination caused by pile-up than the ones reconstructed from resolved jets. This is due to the substructure algorithms, which remove most of the soft contributions. Still, there is bias by residual pile-up contributions not removed by them.

In order to correct for these effects, jet energies are calibrated by scaling the four-momenta of the jets with specific factors depending on the transverse momentum and the pseudo rapidity of the respective jets. As already described in Section 4.7.3, these scale factors are derived in a factorized fashion for simulated and measured data. In the CMS collaboration, the jet-energy corrections are produced centrally by the JME physics-object group [167, 168]. Nevertheless, no specific jet-energy corrections for the Cambridge/Aachen fat jets with a cone-size parameter $R = 1.5$ and subjets corresponding to the various substructure algorithms have been determined so far. A solution to this issue has been studied in a search for $\mathrm{t\bar{t}}$ production by processes beyond the Standard Model at a center-of-mass energy of 7 TeV [203]. There, the application of jet-energy corrections derived for jets clustered with the anti-k_{T} algorithm and a cone-size parameter $R = 0.5$ on jets clustered with the Cambridge/Aachen algorithm and a cone-size parameter $R = 0.7$ has been studied. Despite the dissimilar clustering algorithms and the difference in cone size, the calibration has been found to perform adequately well. Related to these studies, the applicability of the available jet-energy corrections for jets clustered with the anti-k_{T} algorithm and cone-size parameters $R = 0.8$ and $R = 0.4$ on the fat jets and subjets used in this analysis has been verified. This study probes the invariant masses for true and fake boosted top-quark candidates and boosted Higgs-boson candidates reconstructed from fat jets and subjets with and without jet-energy corrections applied as described in Section 7.6 and Section 7.7. True and fake candidates are defined by an angular matching of the fat jet to the simulated massive particles and their decay products. A more detailed description of the definitions applied can be found in the sections stated above. The distributions of the invariant mass obtained are shown in Fig. 7.7 and Fig. 7.8. As a measure of the separation between true candidates and fake candidates, these figures also include receiver operator

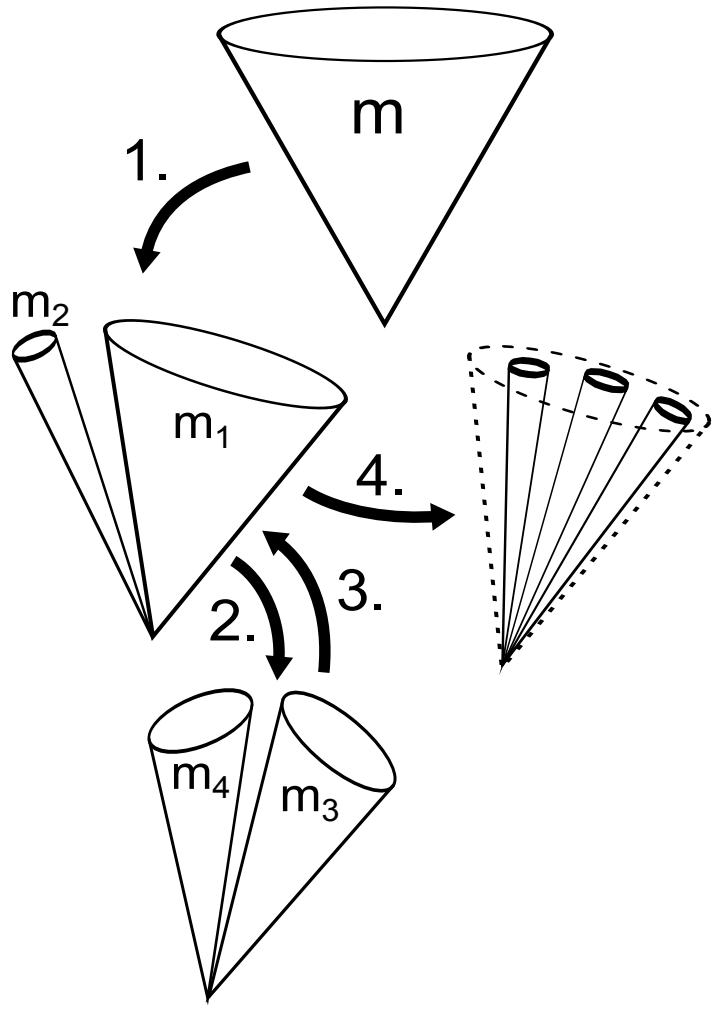


Figure 7.6: Subjet-finding procedure of the BDRS algorithm. The algorithm starts with declustering the fat jet j with mass m producing the subjets j_1 and j_2 with $m_1 > m_2$ (**1.**). In this example, the mass-drop criterion fails as $m_1 > \mu m$. The algorithm continues with discarding j_2 and declustering j_1. The two subjets j_3 and j_4 with $m_3 > m_4$ emerge from this step(**2.**). In this example, the mass-drop criterion $m_3 < \mu m_1$ is passed and j_3 and j_4 are identified as the bottom quarks from the Higgs-boson decay. The subjets j_3 and j_4 also pass the asymmetry condition (**3.**). Filtering is applied on the constituents of the Higgs-boson candidate and the three hardest filtering subjets are chosen.

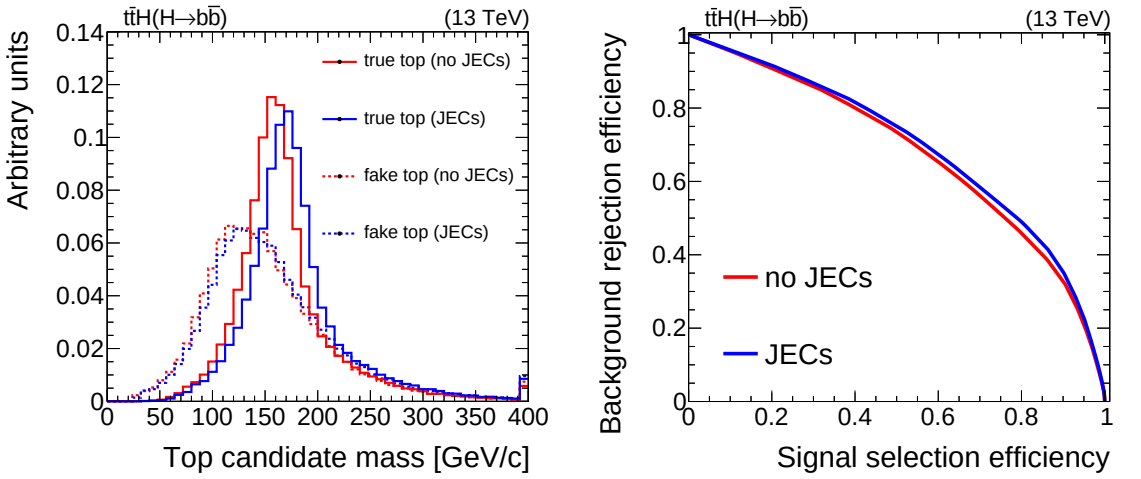


Figure 7.7: Invariant mass of true and fake boosted top-quark candidates (left) and the corresponding receiver operator characteristics (right) for candidates reconstructed with fat jets and subjets with and without jet-energy corrections (JECs) applied. True and fake candidates are extracted from simulated $\mathrm{t\bar{t}}(\mathrm{H}\to\mathrm{b\bar{b}})$ events based on an angular matching to the simulated particles. For the determination of the ROCs, the bins of the invariant mass distributions are ordered by their signal-over-background ratio.

characteristics (ROC) constructed from the invariant-mass distributions. The mean, the width, and the position of the bin with the largest number of entries of these distributions are listed in Table 7.1. When applying the jet-energy corrections, the distributions of the reconstructed top-quark mass and the reconstructed Higgs-boson mass are shifted to higher values. This effect is visible for true candidates as well as for fake candidates. However, in the former case, it is more pronounced, which leads to a better separation of true and fake candidates. This effect can also be observed in the corresponding ROCs. For true boosted Higgs-boson candidates, the application of the jet-energy corrections leads to reconstructed masses closer to the true Higgs-mass. Nevertheless, a large fraction of the reconstructed masses are still below $m_{\mathrm{H}} = 125\ \mathrm{GeV}/c^2$. The mean of the invariant-mass distribution of true boosted top-quark candidates is shifted to values above the best measurement of the top-quark mass, which is about $m_{\mathrm{t}} = 173\ \mathrm{GeV}/c^2$. However, the distribution features a long tail in the direction of large values. The position of the peak, on the other hand, is located very close to the best measurement of the top-quark mass. The widths of both distributions increase when applying the jet-energy corrections. This is due to the smearing applied in the correction of the simulated jet-energy resolution described in Section 4.7.3. Summarized, a better description of the particle masses is achieved if the jet-energy corrections are applied to the fat jets and subjets. Further, a better separation of correctly and incorrectly reconstructed boosted-particle candidates is achieved. This improves the identification of boosted top-quark candidates described in Section 7.6 and the final discrimination of signal against background described in Chapter 9. Additionally, the effect of the jet-energy corrections on the agreement between data and simulation has been tested. The differences have been found to be negligible.

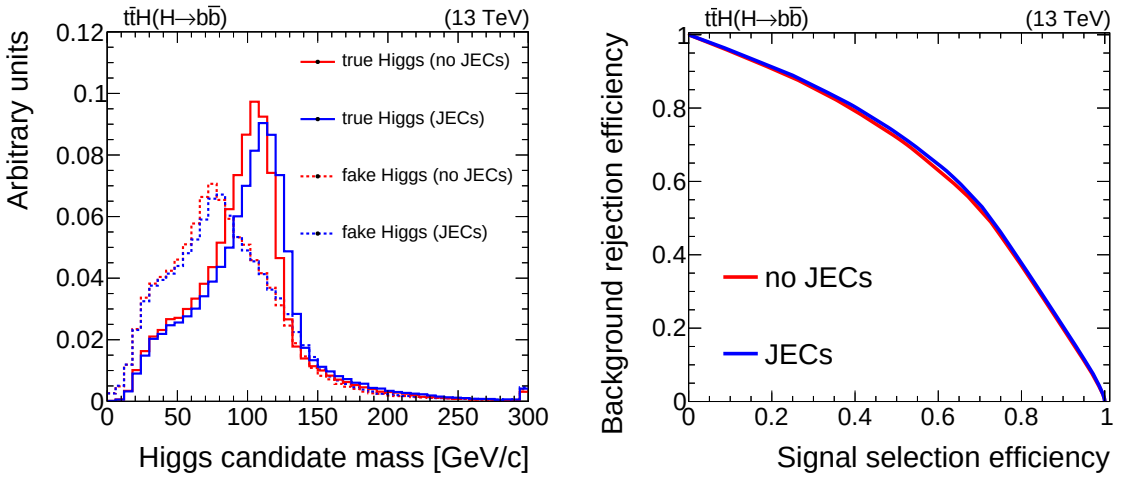


Figure 7.8: Invariant mass of true and fake boosted Higgs-boson candidates (left) and the corresponding receiver operator characteristics (right) for candidates reconstructed with fat jets and subjets with and without jet-energy corrections (JECs) applied. True and fake candidates are extracted from simulated $t\bar{t}(H\rightarrow b\bar{b})$ events based on an angular matching to the simulated particles. For the determination of the ROCs, the bins of the invariant mass distributions are ordered by their signal-over-background ratio.

7.5. Boosted-Object Selection

Particular phase-space regions of fat jets and subjets are mainly populated by reconstructed boosted-particle candidates not or only partly originating from massive particle decays. A fat jet with low transverse momentum, for example, is very unlikely to include all decay products of a massive particle. Another example of a phase-space region featuring mostly fake candidates is the forward region of the detector, which corresponds to large pseudo-rapidity values. This can be explained by the limited acceptance of the subdetectors in this region and by the fact that the massive particles are produced rather centrally in the detector.

A large fraction of fake candidates is removed by applying a kinematic selection on the fat jets and the subjets used in this analysis. As already mentioned in Section 7.1, the clustered fat jets are required to feature a transverse momentum larger than 200 GeV/c. Further, fat jets have to pass a cut on the pseudo-rapidity requiring $|\eta| < 2.0$ to be selected. This threshold was chosen so tight to ensure that the large fat jets are located within the acceptance of the CMS tracker. Subjets are required to feature a transverse momentum larger than 20 GeV/c and to pass a cut on the pseudo rapidity requiring $|\eta| < 2.4$. Additionally, subjets have to fulfill the identification criteria of resolved jets described in Section 4.9.3.

7.6. Boosted Top-Quark Reconstruction

The correct reconstruction of the hadronically decaying top quark is essential for the reconstruction of boosted $t\bar{t}H$ events. Requiring the reconstruction of a boosted hadronically decaying top quark suppresses various background processes like QCD-multijet or vector-boson+jets production. Nevertheless, the main background process, $t\bar{t}$+jets production, also features genuine hadronically decaying top quarks. A distinction between

Table 7.1: Parameters of the invariant-mass distributions of true and fake boosted top-quark candidates (top) and boosted Higgs-boson candidates (bottom) reconstructed from fat jets and subjets with and without jet-energy corrections applied. True and fake candidates are extracted from simulated $t\bar{t}(H\to b\bar{b})$ events based on an angular matching to the simulated particles. Besides the mean and the width of the distributions, also the peak position, which is the position of the histogram bin with the most entries, is stated. The corresponding distributions are shown in Fig. 7.7 and Fig. 7.8.

JECs	True top-quark mass			Fake top-quark mass		
	Mean	Peak postion	Width	Mean	Peak postion	Width
Not applied	169.02	156.00	48.89	152.00	116.00	60.56
Applied	179.91	172.00	51.45	157.75	124.00	63.06

JECs	True Higgs-boson mass			Fake Higgs-boson mass		
	Mean	Peak postion	Width	Mean	Peak postion	Width
Not applied	97.77	105.00	39.54	83.69	75.00	42.76
Applied	102.57	111.00	41.69	86.14	81.00	44.51

$t\bar{t}H$ and $t\bar{t}$ events is achieved by a correct reconstruction of the Higgs boson. However, boosted Higgs-boson candidates are easily faked by the decay products of hadronically decaying top quarks. Accordingly, the correct reconstruction of the hadronically decaying top quark prior to the identification of the boosted Higgs boson reduces the probability for reconstructing a fake boosted Higgs-boson candidate.

The identification of boosted hadronically decaying top quarks is based on their distinctive decay kinematics. The three decay products form jets, where the combination of two of them should show an invariant mass close to the W-boson mass and the combination of all three should feature a mass close to the one of the top quark. Additionally, one of the jets should show the characteristics of a B-hadron decay. In case the top quark has a large transverse momentum, the particle jets of the three decay products are collimated. The fat jets described in Section 7.1 are used to cluster this system of particle jets. The substructure originating from the individual decay products is extracted using the HEPTopTaggerV2 algorithm described in Section 7.3.1.

In order to distingush between true and fake boosted top-quark candidates, different classification methods, which are based on the kinematics of the top-quark candidate, the fat jet, and the subjets, are tested. In addition to the already existing original HEPTopTagger tagging criteria, new multivariate classification methods have been developed specifically for this analysis and benchmarked. These new methods include a modified version of the HEPTopTagger using subjet b-tagging information, an identification method based on a likelihood ratio, and a boosted decision-tree tagger. For the training of the latter two methods, true and false top-quark candidates are required to be well defined. Due to hadronization and impurities clustered into the top-quark candidates, the choice of a reasonable definition is ambiguous. In this analysis, true boosted top-quark candidates are defined by requiring that all simulated hadronic top-quark decay products are located within the fat jet. Hence, the angular distance between the simulated decay products and the fat-jet axis has to be smaller than 1.5. Fake boosted top-quark candidates are defined by their fat-jet axes featuring an angular distance of $\Delta R > 2.0$ with respect to the simulated hadronically decaying top quark. True and fake boosted top-quark candi-

dates are taken from simulated $t\bar{t}$ events. This is different from the typical choice, where fake candidates from QCD-multijet events are applied for the optimization of top taggers. However, as the most important processes investigated in this analysis, $t\bar{t}H$ and $t\bar{t}$ production, both feature a very busy final state, fake candidates mostly consist of a mixture of decay products of massive particles. By using fake candidates from simulated $t\bar{t}$ events, the identification is optimized for this scenario.

The classification methods tested for this analysis are described in the following:

- **Original HEPTopTagger selection cuts (HTT):** The original HEPTopTagger tagging criteria are represented by the characteristic A-shaped selection cuts described in Section 7.3.1. These cuts were optimized to distinguish fake boosted top-quark candidates in QCD-multijet events from true boosted top-quark candidates. By varying the window width of the A-shaped selection cuts, different selection efficiencies can be obtained.

- **Modified HEPTopTagger cuts using subjet b-tagging information (b-tag-HTT):** This classification method is a modified version of the original HEPTopTagger tagging criteria optimized for this analysis. Additional information from subjet b-tagging is added by assigning the subjets to the hadronic top-quark decay products according to their b-tagging output as described in Section 7.3.1. Using this approach, the assignment of subjets to the bottom quark and the W-boson decay products is fixed and the characteristic A-shaped selection cuts are reduced to rectangular cuts:

$$\frac{m_{\mathrm{BW1}}}{m_{\mathrm{BW2}}} > 0.2 \quad \text{and} \quad \frac{m_{\mathrm{BW2}}}{m_{\mathrm{BW1}}} > 0.2 \,,$$
$$R_{\mathrm{min}} < \frac{m_{\mathrm{W}}}{m_{\mathrm{top}}} < R_{\mathrm{max}} \,,$$
$$\text{with} \quad R_{\mathrm{min/max}} = (1 \pm f_{\mathrm{W}})$$

  The tagging criteria are based on the invariant masses m_{W}, m_{BW1}, and m_{BW2} formed by di-subjet combinations of the subjets B, W1, and W2. The mass of the top-quark candidate m_{top} is given by the invariant mass of the combination of all three subjets. The candidate is considered as boosted top-quark, if all three criteria are fulfilled. The selection efficiency of true and fake candidates can be varied by adjusting the selection-window width f_{W}.

- **Likelihood top tagger (Likelihood)**: The likelihood top tagger is based on a likelihood function constructed with the probability densities of variables applied by the original HEPTopTagger criteria. For this classification method, the subjets are associated to the decay products of the top quark with respect to their b-tagging output. The following three variables are used to form the likelihood function:
  - $\boldsymbol{m_{\mathrm{top}}}$: The invariant mass of the top-quark candidate given by the combination of all three subjets B, W1, and W2.
  - $\boldsymbol{a_m = \arctan \frac{m_{\mathrm{BW1}}}{m_{\mathrm{BW2}}}}$: The arc tangent of the ratio of the invariant masses m_{BW1} and m_{BW2}, which are given by the invariant masses of the disubjet combinations of subjet B with subjet W1 or subjet W2.
  - $\boldsymbol{r_m = \frac{m_{\mathrm{W}}}{m_{\mathrm{top}}}}$: The ratio of invariant masses of the W boson and the top quark reconstructed from the subjets B, W1, and W2.

Table 7.2: Parameter configuration used for the training of the BDT top tagger. The parameter configuration has been optimized using the particle-swarm algorithm. A more detailed explanation of the parameters can be found in Section 5.1.4.

N_{trees}	Shrinkage	Bagging Fraction	N_{cuts}	Depth
1200	0.019	0.41	20	2

The probability-density functions f corresponding to these variables are obtained from true and fake boosted top-quark candidates. As described above, the candidates are extracted from simulated $\mathrm{t\bar{t}}$ events by an angular matching of the boosted top-quark candidate fat jet to the generated particles. The events are taken from a sample that is statistically independent of the one applied for the evaluation of the final results of this analysis. Using the obtained probability-density functions, the likelihood ratio is formed as follows:

$$\mathcal{L}' = \frac{f(m_{\text{top}}|\text{true})f(a_m|\text{true})f(r_m|\text{true})}{f(m_{\text{top}}|\text{true})f(a_m|\text{true})f(r_m|\text{true}) + f(m_{\text{top}}|\text{fake})f(a_m|\text{fake})f(r_m|\text{fake})} \,.$$

Optimal performance of a likelihood ratio is only ensured if all variables used in the likelihood functions are fully uncorrelated. However, for the likelihood ratio presented here, this is not the case, as m_{top}, a_m, and r_m are indeed correlated. Correspondingly, a better performance of the likelihood top tagger can be achieved if these variables are transformed to be uncorrelated.

- **BDT top tagger (BDT)**: The BDT top tagger represents a multivariate classification method based on a machine-learning approach given by boosted decision trees. The input variables for this identification technique are provided by the output of the HEPTopTaggerV2. The BDT is optimized with the particle-swarm optimization (PSO) described in Section 5.1.4. The optimization is performed with 20 particles in the course of 50 iterations. The Kolmogorov-Smirnov probability threshold is set to 0.1 and the training of the BDT is repeated twice in each PSO iteration. The training is performed with 7000 true and 7000 fake candidates and the same number is used for testing. The candidates are taken from simulated $\mathrm{t\bar{t}}$ events from a sample statistically independent from the sample used to evaluate the final results of this analysis. The description of data by simulation for the variables used in the optimization has been tested in dedicated control regions. The corresponding approach is outlined in Section 6.3 and Section 8.3. The optimal BDT-parameter configuration obtained by the PSO is displayed in Table 7.2. The variables used for the training of the BDT top tagger are specified in Table 7.3. Further, the distributions of the input variables for data and simulation are presented in Appendix A.2. The output distribution of the BDT-top tagger for true and fake boosted top-quark candidates and the corresponding receiver operator characteristic (ROC) are displayed in Fig. 7.9.

The classification methods are benchmarked based on the selection efficiency of simulated $\mathrm{t\bar{t}H}$ and $\mathrm{t\bar{t}}$ events. For this purpose, each event is reconstructed under the hypothesis of being a $\mathrm{t\bar{t}H}$ event based on the procedure described in the upcoming Section 8.1.

Table 7.3: Input variables used in the training of the BDT top tagger. The choice of the input variables has been optimized using the particle-swarm algorithm.

m(fat jet)	Invariant mass of the fat jet
m(W)	Invariant mass of the W boson reconstructed with subjets W1 and W2
m(BW1)	Invariant mass of the combination of subjets B and W1
m(BW2)	Invariant mass of the combination of subjets B and W2
m(W)/m(top)	Ratio of the invariant masses of W boson and top quark reconstructed from subjets B, W1, and W2
b-tagger output(B)	b-tagging output of subjet B
τ_2/τ_1	Ratio of N-Subjettiness 2 and 1
τ_3/τ_2	Ratio of N-Subjettiness 3 and 2
ΔR_{opt}	Difference between measured and calculated optimalR

The boosted Higgs-boson candidate is reconstructed using the method described in Section 7.7. The boosted hadronically decaying top quark is reconstructed as described above, using the classification method under test. Besides some selection cuts also used for the resolved-event selection, the selection of boosted $t\bar{t}H$ events relies on selection cuts on the classification outputs of the boosted top-quark candidate and the boosted Higgs-boson candidate resulting from the boosted-event reconstruction. The optimal combination of cuts on the two classification outputs are defined by the cut values providing the smallest $t\bar{t}$ event selection efficiency for every given $t\bar{t}H$ selection efficiency. The ROCs showing the performance of the optimal cut combinations for all $t\bar{t}H$ selection efficiencies and all classification methods are shown in Fig. 7.10. Based on them, the different classification methods can be compared. The ROC derived with the BDT classification yields the largest background-rejection efficiency for every signal-selection efficiency value in the range. Depending on the working point, improvements with respect to the other classification methods are of the order of 10 %. Accordingly, the BDT top tagger is chosen as top-quark classification method for this analysis. There are three reasons for the better performance of the BDT classification: the training on fake candidates stemming from the combinatorial background, the complementary information provided by additional variables, and the optimized cuts. By comparing the likelihood method with the modified HTT cuts using b-tagging information, one can learn that the improvement brought by the fake candidate choice is quite small. Accordingly, a large part of the benefit in performance is caused by the additional variables and the optimized cuts in the BDT method.

7.7. Boosted Higgs-Boson Reconstruction

A reconstructed Higgs-boson candidate and its properties represents some of the most important features for discriminating $t\bar{t}H$ signal events from background. All relevant background processes occurring in the search for $t\bar{t}H$ lack a real Higgs boson. Nevertheless, these processes are able to produce fake candidates stemming from gluon splittings and combinatorial background. Compared to the three-body decay of a top quark, the two-body decay of a Higgs boson into a bottom-quark pair features a simple structure. Hence, this decay lacks some of the distinctive features exploited for the identification of true top-quark candidates and the rejection of fake candidates. Still, by placing requirements based on the large mass of the Higgs boson and the appearance of two jets originating from bottom quarks, a major fraction of fake candidates can be identified.

The boosted Higgs-boson candidates in this analysis are reconstructed based on the same fat jets also used for the reconstruction of top-quark candidates. However, the substructure

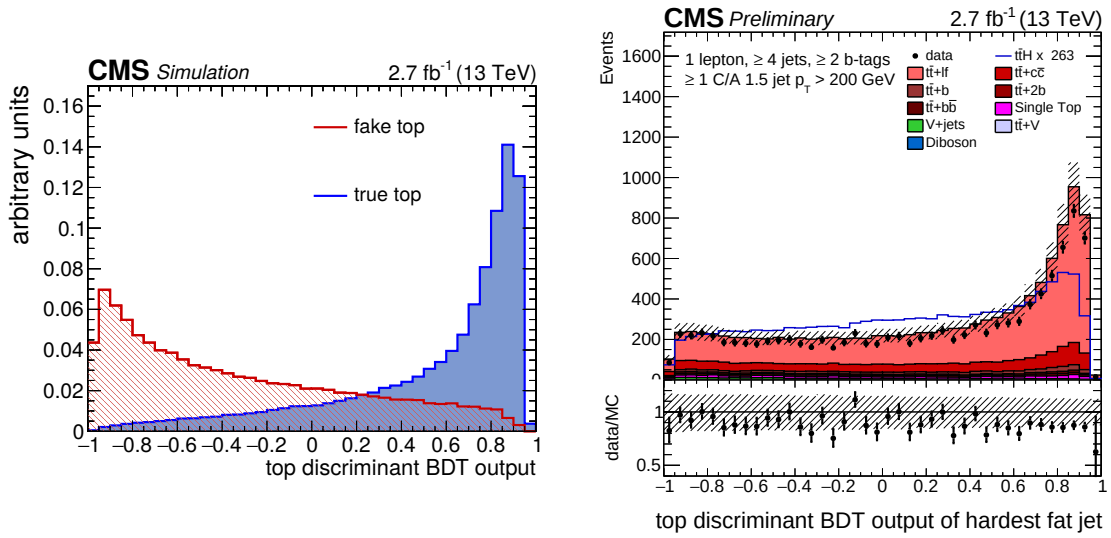


Figure 7.9: Boosted top-quark classification output derived with the BDT top tagger. On the left-hand side the distributions for true and fake boosted top-quark candidates normalized to unity are shown. True and fake boosted top-quark candidates are extracted from simulated $t\bar{t}$ events by an angular matching. The right-hand side shows a comparison of the boosted top-quark classification output of the hardest fat jet in simulated events and data. The simulated background processes are displayed by the filled and stacked histograms. The contribution of each background process is scaled to the event yield predicted for an integrated luminosity of $\mathcal{L} = 2.7\ \text{fb}^{-1}$. The distribution of the $t\bar{t}H$ signal is scaled to the total event yield of all background processes and illustrated by a blue line.

is interpreted using an algorithm specialized in reconstructing the characteristic two-prong decay of the Higgs-boson. The substructure algorithm applied for boosted Higgs-boson candidate reconstruction is chosen based the selection efficiencies of boosted $t\bar{t}H$ signal and background events and the reconstruction quality of the Higgs boson achieved with the algorithm. The following subjet algorithms already described in Section 7.2 are tested:

- **Pruning:** The constituents of the fat jet are reclustered using the pruning method until exactly two jets are formed. The resulting two subjets are associated with the bottom quarks from the Higgs-boson decay.

- **Soft-drop declustering (SD):** The fat jet is declustered until the soft-drop criterion is fulfilled. The resulting two subjets are associated with the bottom quarks stemming from the Higgs-boson decay. In this study, two soft-drop configurations SD_1 and SD_2 defined by the parameter values,

$$\begin{aligned} z_{\text{cut,SD}_1} &= 0.1 \quad \text{and} \quad \beta_{\text{SD}_1} = 0\,, \\ z_{\text{cut,SD}_2} &= 0.2 \quad \text{and} \quad \beta_{\text{SD}_2} = 1\,, \end{aligned}$$

  are tested.

- **Mass-drop declustering (MD):** A simplified version of the procedure described in Section 7.3.2 is applied. The fat jet is declustered until the mass-drop criterion

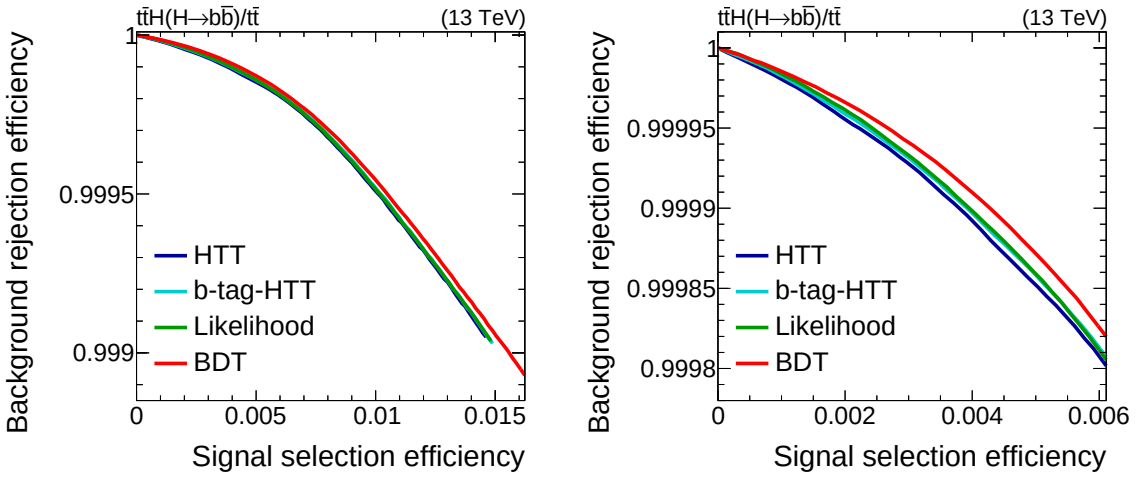


Figure 7.10: Receiver operator characteristics showing the selection efficiencies of signal events and the rejection efficiencies of background events for different boosted top-quark classification methods. The plot on the left covers the whole range of signal-selection efficiencies for events with a reconstructed boosted Higgs-boson candidate and a reconstructed boosted top-quark candidate. The plot on the right side shows a zoom into the region considered for the event selection of the boosted analysis category. The working points providing the curves are defined by the best performing combinations of cuts on the classification outputs of the reconstructed boosted Higgs-boson candidate and the reconstructed boosted top-quark candidate. Signal is given by simulated $t\bar{t}(H\to b\bar{b})$ events and background is given by simulated $t\bar{t}$ events. The selection efficiencies are calculated by forming the ratio of selected events and the total number of simulated events.

is fulfilled. The subjets are required to pass the asymmetry cut. Filtering is not applied and the resulting two subjets are associated with the bottom quarks from the Higgs-boson decay.

- **BDRS algorithm (BDRS):** The full mass-drop declustering and filtering procedure described in Section 7.3.2 is applied. The subjets are associated to the bottom quarks from the Higgs-boson decay based on their b-tagging output.

The first part of the performance study is similar to the one performed for the boosted top-quark classification. Simulated $t\bar{t}H$ and $t\bar{t}$ events are reconstructed as described in Section 7.7 using the top-quark reconstruction procedure based on the HEPTopTaggerV2 and the BDT classification. The boosted Higgs-boson candidate is reconstructed using one of the methods described above. The second highest b-tagger output among the subjets is chosen as discriminant for the classification of true and fake Higgs-boson candidates. Simulated $t\bar{t}H$ and $t\bar{t}$ events are selected by applying cuts on the top-quark and Higgs-boson classification discriminants in addition to cuts on resolved analysis objects as described in Section 8.2. The ROCs corresponding to the cut combinations that provide the highest background-rejection efficiency for a given signal-selection efficiency are used as a benchmark for the different substructure algorithms. A comparison of the ROCs derived with the different substructure algorithms is displayed in Fig. 7.11.

As the properties of the Higgs-boson candidate play a vital role in the final discrimination of signal and background events, it is necessary to study the influence of the

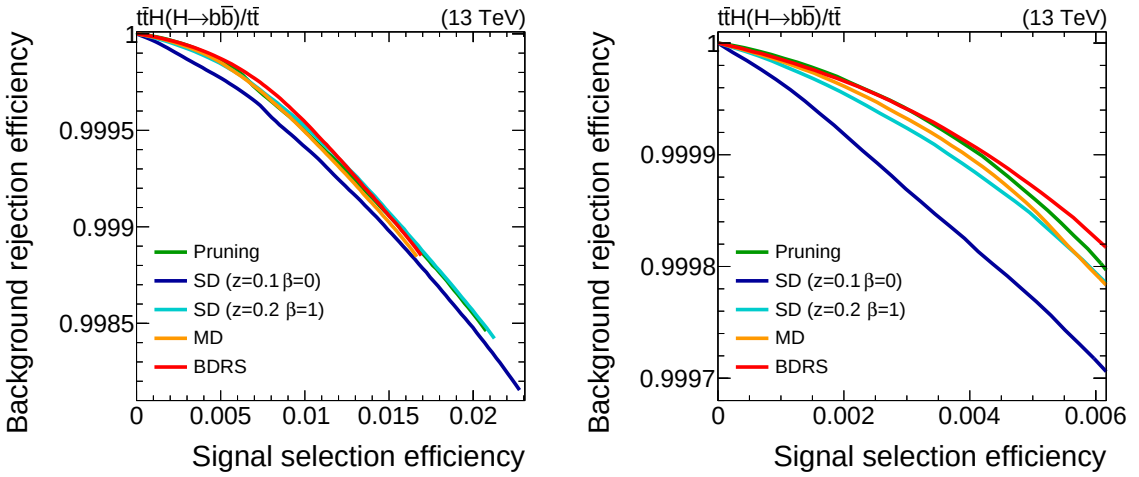


Figure 7.11: Receiver operator characteristics showing the selection efficiencies of signal events and the rejection efficiencies of background events for different boosted Higgs-boson reconstruction methods. The plot on the left covers the whole range of signal-selection efficiencies for events with a reconstructed boosted Higgs-boson candidate and a reconstructed boosted top-quark candidate. The plot on the right-hand side shows a zoom into the region considered for the event selection of the boosted analysis category. The working points providing the curves are defined by the best performing combinations of cuts on the classification outputs of the reconstructed boosted Higgs-boson candidate and the reconstructed boosted top-quark candidate. The signal is given by simulated $t\bar{t}(H\to b\bar{b})$ events and background is given by simulated $t\bar{t}$ events. The selection efficiencies are calculated by forming the ratio of selected events and the total number of simulated events.

substructure algorithms on the quality of the reconstruction of the Higgs-boson candidate. Accordingly, for the second part of the substructure-algorithm study the shape of Higgs-boson candidate observables for true and fake Higgs-boson candidates are compared.

As already mentioned for the case of the top quark, the choice of a definition for true and fake candidates is ambiguous. Due to hadronization, the underlying event, pileup, and contributions from other decay particles in the event, fat jets and subjets do not directly correspond to the Higgs boson and its decay products. Similar to the definition of true and fake top-quark candidates, a definition based on an angular matching of the fat jet to the simulated Higgs boson and its decay products has been chosen. True Higgs-boson candidates are defined by both simulated bottom quarks from the Higgs-boson decay lying within an angular distance of $\Delta R < 1.5$ with respect to the fat jet axis. The fat jets of fake candidates have to be feature an angular distance of $\Delta R > 2.0$ with respect to the simulated Higgs boson. Additionally, fake Higgs bosons are required to feature not more than one of the simulated Higgs-boson decay products within $\Delta R < 1.5$ of the fat-jet axis.

The invariant mass of the Higgs-boson candidate reconstructed with the subjets assigned to the bottom quarks is investigated as a measure of the reconstruction quality, as it represents some of the most discriminating properties. The scale and the resolution of the Higgs-mass peak of true candidates is affected by the substructure algorithms, depending on the degree of grooming away contamination and radiation from decay products. Another important aspect is the shaping of fake-candidate distributions by the substructure algorithms. Fig. 7.12 shows the distribution of the invariant mass for signal and back-

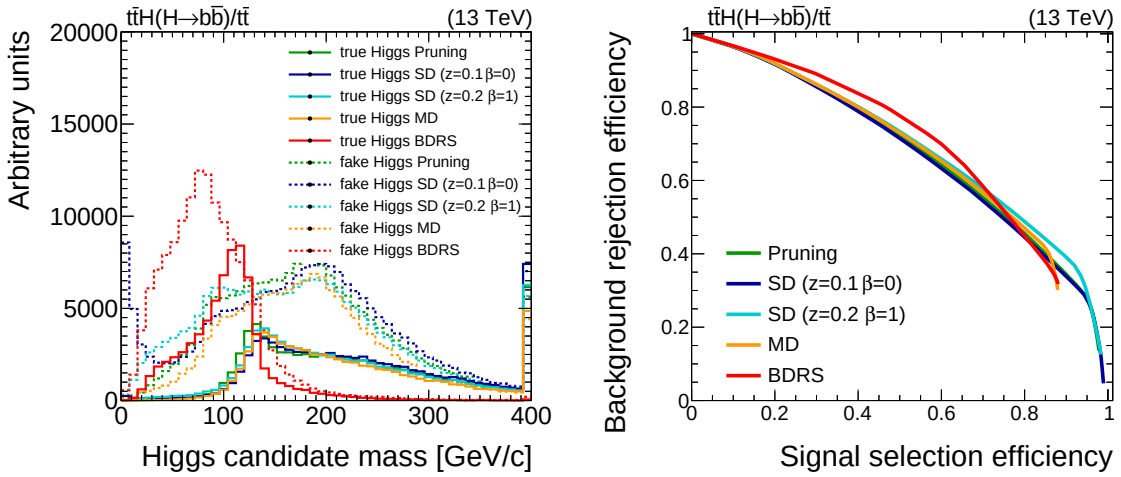


Figure 7.12: Invariant mass of true and fake boosted Higgs-boson candidates (left) and the corresponding receiver operator characteristics (ROC) (right) for different boosted Higgs-boson reconstruction methods. True boosted Higgs-boson candidates are taken from simulated $t\bar{t}(H\to b\bar{b})$ events. Fake boosted Higgs-boson candidates are taken from simulated $t\bar{t}$ events. The candidates are extracted based on an angular matching to the simulated particles. For the determination of the ROCs, the bins of the invariant-mass distributions are ordered by their signal-over-background ratio.

ground for all tested substructure algorithms. In the same figure, corresponding ROCs for the selection of true and false Higgs-boson candidates obtained by placing cuts on the signal and background distributions are displayed. For better comparability, the bins of the signal and background distributions have been ordered by their signal-over-background ratio. The distributions derived with mass-drop declustering, soft-drop declustering, and pruning provide a similar, rather broad Higgs-boson mass spectrum peaking at the invariant mass of the Higgs-boson $m(\text{Higgs cand.}) = 125$ GeV. The BDRS algorithm, on the other hand, provides a distribution with a narrow peak shifted to values slightly below the Higgs-boson mass. The background distributions derived without filtering are more similar to the shape of the signal distribution than the background distributions derived with filtering. This effect propagates to the ROCs, which show better discrimination for the BDRS reconstruction. The filtering procedure represents an aggressive grooming approach, removing not only contamination but also soft and collinear radiation from the Higgs-decay final state. Due to this missing energy the reconstructed mass of the Higgs-boson is biased to smaller values. Concerning the reconstruction of fake candidates, contributions favoring the creation of large invariant masses are removed by the filtering procedure and the resulting distribution accumulates at small values. The distributions of the algorithms without filtering are smeared out towards large values of the invariant mass, which is evidence for contamination inside the candidate.

Compared to the separation power between true and fake Higgs-boson candidates, the absolute scale of the reconstructed Higgs-boson mass is not very important for this analysis. Because of this and the good performance in the selection of $t\bar{t}H$ events and the rejection of $t\bar{t}$ events, the BDRS algorithm is chosen for the reconstruction of the Higgs-boson candidate.

As the investigated Higgs-boson decay includes two real bottom quarks, which a major part of fake candidates lack, the classification of true and fake Higgs-boson candidates

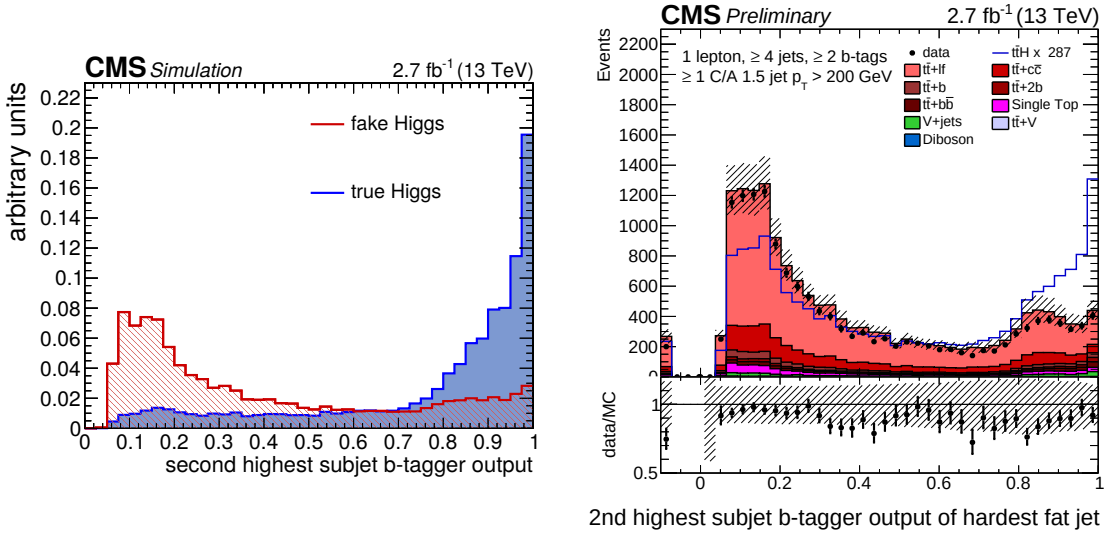


Figure 7.13: Boosted Higgs-boson classification output given by the second highest b-tagger output among subjets found. On the left-hand side, the distributions for true and fake boosted Higgs-boson candidates normalized to unity are shown. True boosted Higgs-boson candidates are taken from simulated $t\bar{t}(H\to b\bar{b})$ events. Fake boosted Higgs-boson candidates are taken from simulated $t\bar{t}$ events. The candidates are extracted based on an angular matching to the simulated particles. The right-hand side shows distributions of the boosted Higgs-boson classification output of the hardest fat jet in simulated events and data. The simulated background processes are displayed by the filled and stacked histograms. The contribution of each background process is scaled to the event yield predicted for an integrated luminosity of $\mathcal{L} = 2.7\ \text{fb}^{-1}$. The distribution of the $t\bar{t}H$ signal is scaled to the total event yield of the background processes and illustrated by a blue line.

relies on subjet b-tagging. Accordingly, the b-tagging output of the subjet *B2* is chosen as boosted Higgs-boson identification discriminant. A distribution of this variable for true and fake boosted Higgs-boson candidates is displayed in Fig. 7.13. Further, distributions of the same variable for data and simulation in a $t\bar{t}$ enriched control region is shown. Fake candidates with two real bottom quarks mainly originate from combinatorial background and gluon splittings into a bottom-quark pair. A further type of fake candidates is based on subjets incorrectly identified as stemming from bottom quarks. The number of fake candidates could be further reduced by using a multivariate method that also includes the invariant mass of the reconstructed Higgs-boson candidate and other variables like N-subjettiness. Nevertheless, this approach has not been considered, as it would bias these variables for the following steps of the analysis. The properties of reconstructed Higgs-boson candidates are strong tools for the separation of $t\bar{t}H$ events against background events and are rather used in that sense.

7.8. Boosted-Objects Summary

The boosted top-quark candidates and boosted Higgs-boson candidates used in the semileptonic $t\bar{t}(H\to b\bar{b})$ search presented in this thesis are reconstructed from one set of fat jets. These fat jets are clustered using the Cambridge/Aachen algorithm with a cone-size parameter of $R = 1.5$. As input for the clustering, all particle-flow candidates remaining

after applying the charged-hadron subtraction and removing selected charged leptons are used. To be retained for further analysis, fat jets are required to feature a transverse momentum of $p_{\mathrm{T}} > 200$ GeV/c and a pseudo rapidity of $|\eta| < 2.0$.

Boosted top-quark candidates are reconstructed and identified based on the three subjets obtained, when applying the HEPTopTaggerV2 algorithm on a selected fat jet. All subjets are required to feature a transverse momentum of $p_{\mathrm{T}} > 20$ GeV/c and a pseudo rapidity of $|\eta| < 2.4$, for the boosted top-quark candidate to be considered valid. The subjets are assigned to the decay products of a hadronically decaying top quark based on their b-tagging output. The discriminant for the classification of boosted top-quark candidates is given by a BDT trained on variables provided by the subjets and additional substructure variables.

Boosted Higgs-boson candidates are reconstructed and identified based on the subjets provided, when applying the BDRS algorithm on a selected fat jet. The subjets obtained are required to feature a transverse momentum of $p_{\mathrm{T}} > 20$ GeV/c and a pseudo rapidity of $|\eta| < 2.4$, in order to be retained for the further reconstruction of a boosted Higgs-boson candidate. Selected subjets are assigned to the decay products of a Higgs boson based on their transverse momenta and their b-tagging outputs. The discriminant for the classification of boosted Higgs-boson candidates is given by the second largest b-tagging output found among the three hardest subjets.

At this point, a given fat jet can be reconstructed as a boosted top-quark candidate as well as a boosted Higgs-boson candidate. This redundancy will be resolved in the boosted-event reconstruction presented in Section 8.1.

Chapter 8

Boosted-Event Reconstruction and Selection

Following the recipe presented in the publication "Fat Jets for a Light Higgs" [3], events are reconstructed and selected under the hypothesis of being a $t\bar{t}H$ event with a boosted hadronically decaying top quark and a boosted Higgs boson decaying into a bottom-quark pair. Accordingly, a boosted hadronic top-quark candidate and a boosted Higgs-boson candidate are identified in successive order from the collection of boosted objects reconstructed as described in Chapter 7. Further, a prompt electrically charged lepton originating from the decay of the leptonically decaying top quark is expected. A more detailed description of the boosted-event reconstruction can be found in Section 8.1. By selecting events based on the particle candidates found and their respective reconstruction quality, a large fraction of background events can be rejected. The selection achieving this is outlined in Section 8.2. Selected boosted events are included in the analysis by introducing a separate analysis category in addition to the resolved analysis categories described in Chapter 6.

8.1. Boosted-Event Reconstruction

In the publication "Fat Jets for a Light Higgs" [3], a recipe for the search of $t\bar{t}H$ events with boosted signatures featuring a Higgs boson decaying into bottom quarks and a semileptonically decaying top-quark pair is presented. The reconstruction introduced in this context aims at identifying events containing one boosted Higgs boson, one boosted top quark decaying into hadrons, and one prompt electrically charged lepton originating from the leptonic top-quark decay. Prompt electrically charged leptons are reconstructed and selected according to the procedures described in Chapter 4. Selected electrically charged leptons are removed from the set of reconstructed particle-flow objects used as input for the fat-jet clustering, which is described in Section 7.1. Over time, the methods used for substructure analysis have improved with respect to the original recipes presented in "Fat Jets for a Light Higgs". For this reason, the massive-particle reconstruction and identification has been updated with more advanced methods. Boosted-object candidates are reconstructed from fat jets using the HEPTopTaggerV2 algorithm and the BDRS algorithm as described in Section 7.6 and Section 7.7. Still, the sequential order of identifying the boosted hadronic top-quark candidate and the boosted Higgs-boson candidate has been retained, as it has proven to be crucial for good performance of the reconstruction. First, the boosted hadronic top-quark candidate is identified by choosing the fat jet with the highest top-quark discriminant value. The candidate is removed from the collection of fat jets. The boosted Higgs-boson candidate is chosen from the remaining fat jets based on the highest Higgs-boson discriminant value.

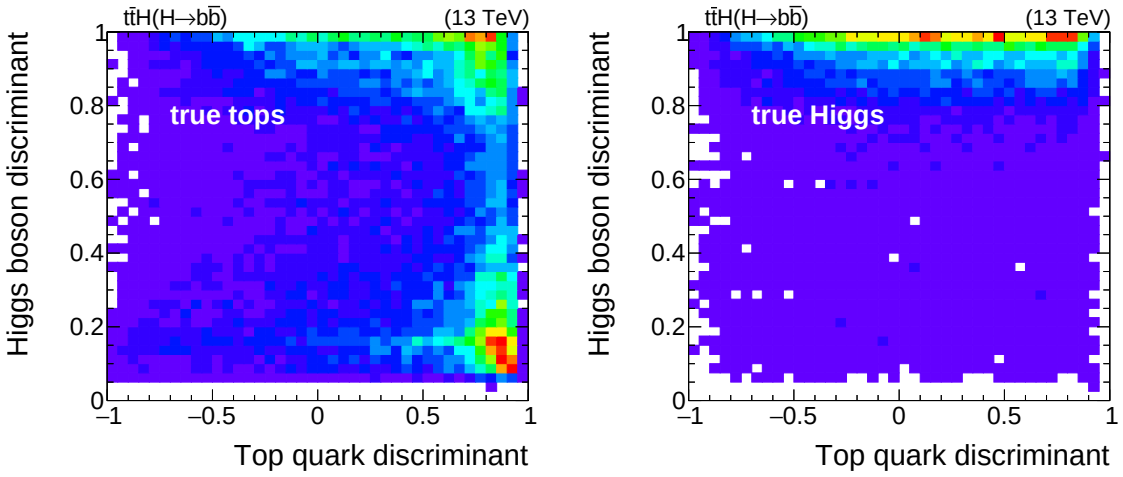


Figure 8.1: Boosted top-quark and boosted Higgs-boson discriminant values for true boosted top-quark candidates (left) and true boosted Higgs-boson candidates (right). True candidates are defined by an angular matching of simulated particles and taken from simulated $t\bar{t}H$ events. Densely populated regions are displayed in red, whereas sparsely populated regions are displayed in purple.

$t\bar{t}H$ production is the only relevant process featuring a real Higgs-boson, whereas reconstructed Higgs-bosons candidates from background processes are fake candidates originating from gluon splittings, combinatorial background, or misidentified particles. Thus, the Higgs-boson candidate and its properties are among the most important ingredients for the discrimination of $t\bar{t}H$ signal events from background processes. For a proper identification of the Higgs-boson, the order of choosing the reconstructed top-quark and Higgs-boson candidates is essential. The boosted top-quark identification is much more efficient at rejecting fake candidates than the Higgs-boson classification. As shown in Fig. 8.1, a large fraction of true boosted top-quark candidates feature both, a large top-quark classification value as well as a large Higgs-boson classification value. Consequently, boosted hadronically decaying top quarks are very likely to fake boosted Higgs-boson candidates. By correctly identifying the hadronically decaying top quark and removing it from the collection of possible Higgs-boson candidates, a large fraction of fake candidates is rejected.

8.2. Boosted-Event Selection and Categorization

As for the resolved case, a large fraction of background events can be rejected by selecting only signal-like events. The event selection is performed based on the reconstructed and selected analysis objects presented in Chapter 4 and the result of the boosted-event reconstruction described in the previous section. Targeting the signature of semileptonic $t\bar{t}(H\to b\bar{b})$ events, selected events are required to feature a prompt electrically charged lepton originating from the leptonic top-quark decay and both a boosted top-quark candidate and a boosted Higgs-boson candidate from the boosted reconstruction.

8.2.1. Boosted-Event Selection

The selection of boosted events starts with the baseline single-lepton selection also used for the selection of resolved events, which is described in Section 6.1. First of all, data

events as well as simulated events are required to pass the single-lepton triggers described in Section 3.1.1. Additionally, events have to feature a selected primary vertex fulfilling the requirements described in Section 4.9.1. Further, the boosted-event selection requires exactly one isolated electrically charged lepton fulfilling the tight identification criteria given in Section 4.9.2. Events containing additional electrically charged leptons fulfilling the loose identification criteria are vetoed. This veto is necessary to reject events from processes producing multiple leptons and to maintain a selection orthogonal to other $\mathrm{t\bar{t}H}$ searches, like searches for $\mathrm{t\bar{t}(H\to b\bar{b})}$ with a dileptonic $\mathrm{t\bar{t}}$ decay. The requirement of the lepton being isolated indeed reduces the efficiency of selecting boosted $\mathrm{t\bar{t}H}$ events. $\mathrm{t\bar{t}H}$ events with a boosted hadronically decaying top quark as well as a boosted Higgs boson are also likely to feature a leptonically decaying top quark with large transverse momentum. A top quark with large transverse momentum passes its momentum to its decay products, causing the electrically charged lepton to be in the immediate vicinity of the jet originating from the bottom quark of the same top-quark decay. Such a configuration results in large isolation values, which in turn leads to the rejection of the event due to the cut on the isolation of the electrically charged lepton. Nevertheless, the isolation criteria are applied in order to maintain an event selection that is exclusive with respect to the other $\mathrm{t\bar{t}H}$ search channels requiring electrically charged leptons. A second reason for applying the lepton-isolation requirements is the suppression of backgrounds, like QCD-multijet production, arising from the misidentification of prompt leptons.

In addition to the isolated electrically charged lepton, requirements on the resolved interpretation of the event are applied. These requirements are based on the anti-k_T jets with a cone-size parameter $R = 0.4$ described in Section 4.7. The resolved jets are clustered from the entire collection of particle-flow candidates independently from the clustering of the fat jets. Selected boosted events are required to feature at least four resolved jets, of which two have to be b-tagged based on the medium working-point requirement described in Section 4.9.4. The requirements on the resolved event interpretation are motivated by a study investigating the ratio of selected signal and background events with respect to the resolved jet and b-tag multiplicity. This study shows an enrichment in background events for events with resolved jet and b-tag multiplicities below the chosen selection cuts. The results of the study are shown in Fig. 8.2. In this figure, the resolved jet and b-tag multiplicities for all events fulfilling the boosted selection with respect to the signal selection efficiency are shown.

The most characteristic part of the boosted-event selection are the requirements based on the outcome of the boosted-event reconstruction. The event-selection requirements rely on cuts on the boosted-object classification outputs of the two reconstructed boosted massive-particle candidates. In other words, one reconstructed boosted hadronic top-quark candidate with a top-quark classification value above a particular threshold and one boosted Higgs-boson candidate with a Higgs-boson classification value above a particular threshold are required for an event to pass the boosted-event selection. In the following, a combination of cuts on the reconstructed boosted-object classification outputs will be referred to as working point. The working point chosen for this analysis is defined by cuts on the boosted-object classification discriminants at -0.49 for the boosted top-quark candidate and 0.89 for the boosted Higgs-boson candidate. The determination of the working point used in the boosted selection is described in the following. A collision event recorded by the CMS detector passing the boosted-event selection as well as the selection for the resolved category requiring at least six resolved jets and at least four b-tags is visualized in Fig. 8.3.

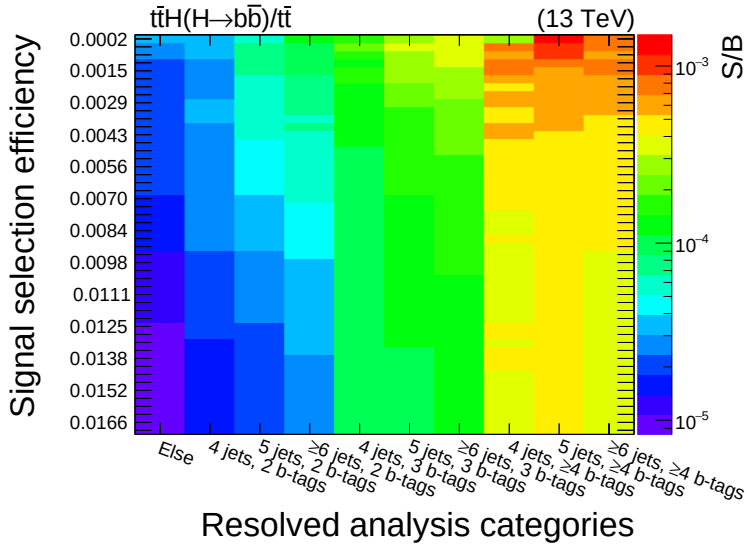


Figure 8.2: Signal-over-background ratio for all events fulfilling the boosted-event selection with respect to the resolved jet and b-tag multiplicities in the event and the signal efficiency corresponding to a particular choice of the boosted-selection working point. The ratio of selected signal and background events is determined using simulated $t\bar{t}(H\to b\bar{b})$ events as signal and simulated $t\bar{t}$ events as background.

Working Point Determination

The working point used for the boosted-event selection is chosen with respect to the efficiency for selecting signal and background events. This choice is optimized based on simulated $t\bar{t}(H\to b\bar{b})$ events as signal and simulated $t\bar{t}$ events exclusively featuring semileptonic top-quark pair decays as background. Plotting the signal-selection efficiencies and corresponding background-selection efficiencies for all possible working points, returns a broad ROC space. This space is displayed in Fig. 8.4 in a reduced form, where the selection efficiencies for certain Higgs-boson classification cuts in combination with all top-quark classification cuts are shown as black lines. In this figure, working points featuring the same signal-selection efficiency but different background-selection efficiencies can be observed. The best-performing working points are determined by finding the cut combinations with the smallest background-selection efficiency for a given signal-selection efficiency. In the ideal case of an unlimited number of simulated events, a smooth line of best-performing working points in the space spanned by the boosted-object classification discriminants is expected. However, the number of events in the mentioned signal and background samples are limited, which causes statistical fluctuations in this optimization procedure. This effect is reduced by introducing a regularization procedure based on the distance of two working points in the space spanned by the boosted-object classification discriminants. The best-performing working points found with this procedure are displayed as the black line in Fig. 8.5. Next to the best-performing working points this figure shows the signal-over-background ratio for all possible working points. Analogue graphs showing different benchmark variables can be found in Appendix A.3. The ROC corresponding to the collection of best-performing working points is shown as red line in Fig. 8.4.

Deciding on one of the best-performing working points corresponds to the decision on a signal-selection efficiency, hence to a number of signal events expected in the boosted

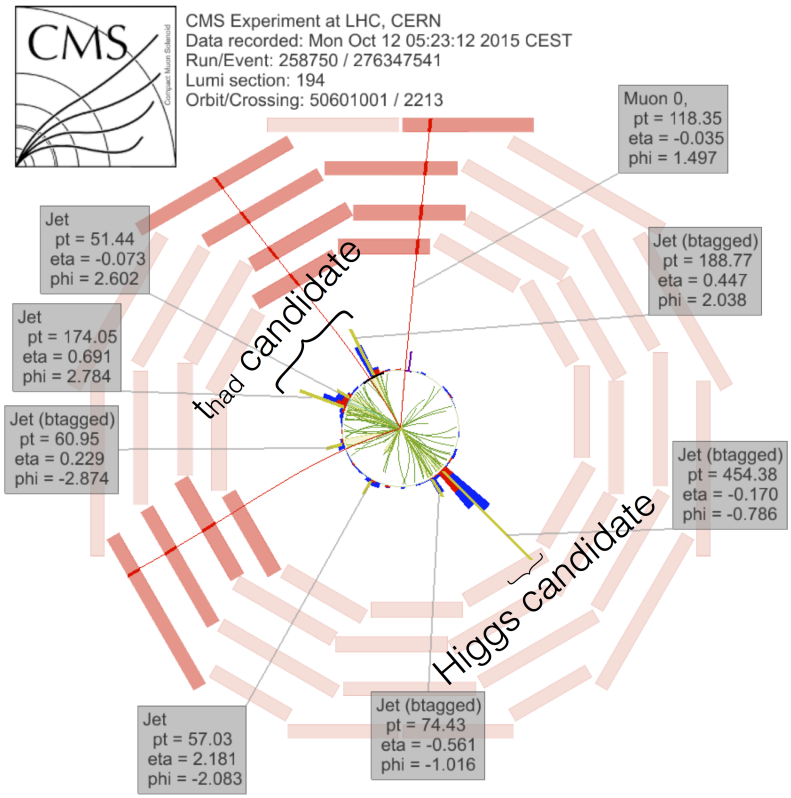


Figure 8.3: Detector signature of an event passing the boosted-event selection as well as fulfilling all requirements for the resolved ≥6 jets, ≥4 b-tags analysis category. The boxes show the properties of the selected resolved jets clustered with the anti-k_{T} algorithm and a cone-size parameter of $R = 0.4$. The boosted top-quark candidate ($\mathrm{t_{had}}$) and the boosted Higgs-boson candidate reconstructed and identified with the boosted-event reconstruction are marked. Taken from [204].

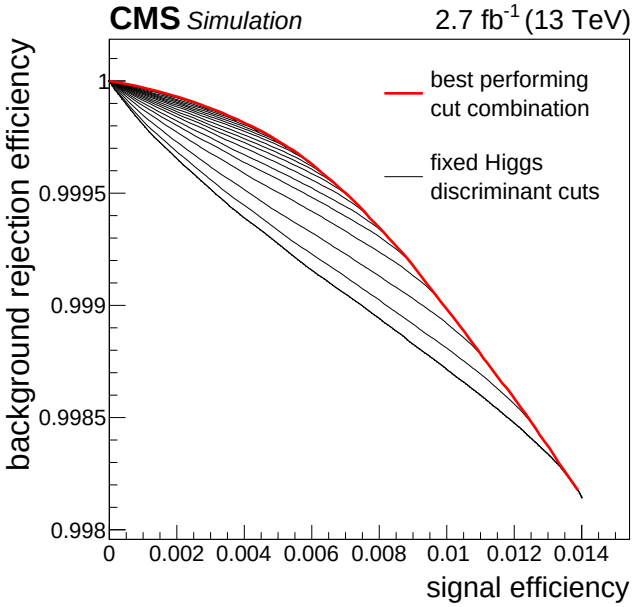


Figure 8.4: Receiver operator characteristics describing the selection efficiencies of signal and background events for different combinations of cuts on the boosted Higgs-boson and top-quark discriminant outputs. A set of receiver operator characteristics for different cuts on the boosted top-quark discriminant and for fixed cuts on the boosted Higgs-boson discriminant are shown in black. The cut combinations with the highest background rejection for a given signal efficiency are given by the top-most (red) curve. The curves were generated with simulated $t\bar{t}(H\to b\bar{b})$ events as signal and semileponically decaying $t\bar{t}$ events as background. Taken from [204].

analysis category. A tight working point with tight thresholds on the boosted-object classification outputs provides a quite large fraction of signal within the selected events, but also a low total event yield of expected signal events in the boosted analysis category. Loose working points with loose selection cuts on the boosted-object classification outputs, on the other hand, result in large signal event yields, but also large contributions by background processes. Events selected by both the resolved analysis category and the boosted analysis category, in the following denoted as overlapping events, are assigned to the boosted analysis category, which will be explained in Section 8.2.2 in more detail. Consequently, the working point also determines the number of events removed from the resolved analysis categories in favor of the boosted analysis category. A smaller number of selected signal events in an analysis category, caused by a tighter selection in case of the boosted analysis category or the removal of overlapping events in case of the resolved analysis categories, is expected to reduce the performance of the respective analysis category. The influence of the definition of the boosted category on the overall performance of the analysis is analyzed in a separate study. Three different configurations are defined based on three different working points corresponding to an expected $t\bar{t}(H\to b\bar{b})$ event yield of about one event, two events, and four events for an integrated luminosity of $\mathcal{L} = 2.7\ \text{fb}^{-1}$. For all three configurations, the blinded expected upper limits on the signal-strength modifier for the $t\bar{t}H$ process $\mu(t\bar{t}H)$ are determined according to the procedures described in Section 5.2 and Chapter 11. For each case, they are calculated for the individual analysis categories and the combination of all analysis categories. The individual analysis categories

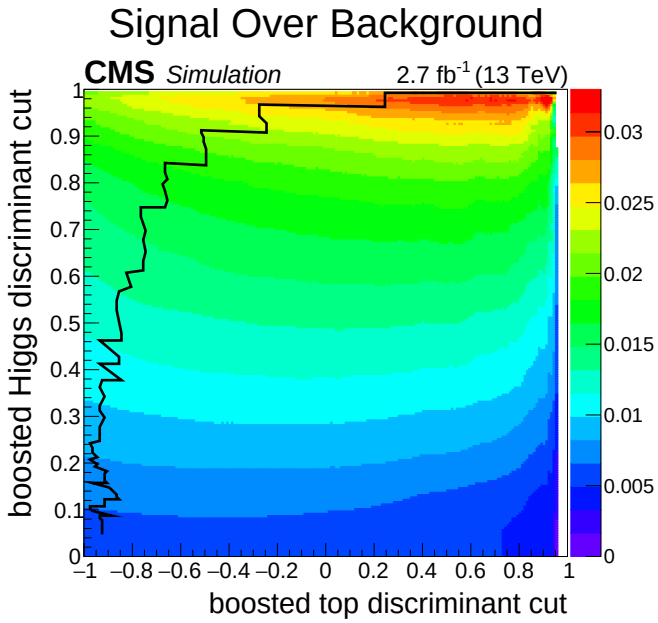


Figure 8.5: Signal-over-background ratio (S/B) resulting from the boosted-event selection as function of different requirements on the boosted Higgs-boson and top-quark discriminant outputs. The efficiencies are determined using simulated $t\bar{t}(H\to b\bar{b})$ events as signal and semileponically decaying $t\bar{t}$ events as background. The black line indicates the cut combinations with the highest background rejection for a given signal efficiency. Taken from [204].

include the boosted analysis category and the resolved $\geq$6 jets, $\geq$4 b-tags analysis category featuring a large final discriminant output, which is described in Section 9.4. The latter category is chosen representatively for all resolved analysis categories. The results of this study are presented in Table 8.1. The blinded expected upper limits on the signal-strength modifier for the individual categories show the expected behavior mentioned above. The performance of the analysis categories with reduced event yields is slightly worse compared to the case with larger event yields. As expected, the influence of the definition of the boosted analysis category on the performance of this category is larger than on the performance of the resolved $\geq$6 jets, $\geq$4 b-tags category with a high BDT output. However, this effect is not propagated to the blinded expected upper limit of the combination of all categories, which is consistent for all three cases. Consequently, the working point of the boosted-analysis category has no major impact for the configurations under investigation. Based on this result, the medium working point corresponding to about two expected signal events in the boosted analysis category is chosen for this analysis. This working point represents an intermediate solution between providing a reasonably large number of signal events, while not removing too many events from the most sensitive resolved analysis categories. As already mentioned, the corresponding cuts on the boosted object classification discriminants are -0.49 for the boosted top-quark candidate and 0.89 for the boosted Higgs-boson candidate.

In addition to the selection efficiencies for signal and background processes, the ratio of signal and background event S/B, the signal significance $S/\sqrt{B}$, and the reconstruction efficiencies of the hadronic top quark, the Higgs boson and both for the boosted analysis category are determined. In the following, the reconstruction efficiency of both massive

Table 8.1: 95 % CL_s blinded expected upper limit on the signal-strength modifier $\mu(t\bar{t}H)$ for three different definitions of the boosted analysis category. The definitions correspond to a selected $t\bar{t}(H\to b\bar{b})$ event yield of about one, two, or four events for an integrated luminosity of $\mathcal{L} = 2.7\,\text{fb}^{-1}$. Limits are shown for the resolved $\geq$6 jets, $\geq$4 b-tags analysis category with high BDT output representing the resolved analysis categories, the boosted analysis category, and the combination of all categories considered in the analysis. Further, the one standard-deviation interval is stated.

Signal efficiency	$t\bar{t}(H\to b\bar{b})$ event yield	95 % CL_s expected upper limit on $\mu(t\bar{t}H)$		
		$\geq$6 jets, $\geq$4 b-tags High BDT	Boosted	All
0.00129	1.0	$8.0^{+4.3}_{-2.6}$	$11.9^{+6.6}_{-4.0}$	$3.6^{+1.7}_{-1.0}$
0.00257	1.9	$8.2^{+4.4}_{-2.6}$	$10.1^{+5.3}_{-3.3}$	$3.6^{+1.7}_{-1.1}$
0.00514	3.8	$8.3^{+4.5}_{-2.6}$	$9.5^{+5.0}_{-3.1}$	$3.6^{+1.6}_{-1.1}$

Table 8.2: Performance benchmarks for the medium working point of the boosted-event selection. The selection efficiencies, S/B, and the signal significance are derived from simulated $t\bar{t}(H\to b\bar{b})$ events as signal and simulated $t\bar{t}$ events exclusively featuring semileptonic top-quark pair decays as background. Both samples are scaled to the event yields expected for an integrated luminosity of $\mathcal{L} = 2.7\,\text{fb}^{-1}$. The efficiencies for correctly reconstructing the top quark, the Higgs boson, or both is determined for selected simulated $t\bar{t}H$ events. Correctly reconstructed candidates are defined by an angular matching to the respective simulated particles.

Selection efficiency		S/B	$S/\sqrt{B}$
Signal	Background		
0.00257	0.00010	0.0210	0.2049

Reconstruction efficiency		
Had. top quark [%]	Higgs boson [%]	Had. top quark & Higgs boson [%]
69.9	53.1	43.0

particles will be denoted as the event-reconstruction efficiency. The selection efficiencies, S/B, and $S/\sqrt{B}$ are determined using simulated $t\bar{t}(H\to b\bar{b})$ events as signal and simulated $t\bar{t}$ events exclusively featuring semileptonic top-quark pair decays as background. Both simulated samples are scaled to the event yield expected for an integrated luminosity of $\mathcal{L} = 2.7\,\text{fb}^{-1}$. Selection efficiencies are defined by the ratio of selected events and the total number of expected events. Reconstruction efficiencies are calculated by dividing the number of selected simulated $t\bar{t}(H\to b\bar{b})$ events featuring a correct reconstruction of the respective particles by the total number of selected simulated $t\bar{t}(H\to b\bar{b})$ events. Correctly reconstructed particles are defined by the reconstructed candidate lying within an angular distance of $\Delta R < 0.5$ of the simulated particle. The benchmark values provided by the medium working point are displayed in Table 8.2.

The boosted-event selection selects about 25 times fewer $t\bar{t}$ background events than signal events. Additionally, the selected events feature a correctly reconstructed top quark in about 70 % cases, whereas the Higgs boson is correctly identified in about half of the selected events. In about 40 % of the selected events, both massive particles are correctly

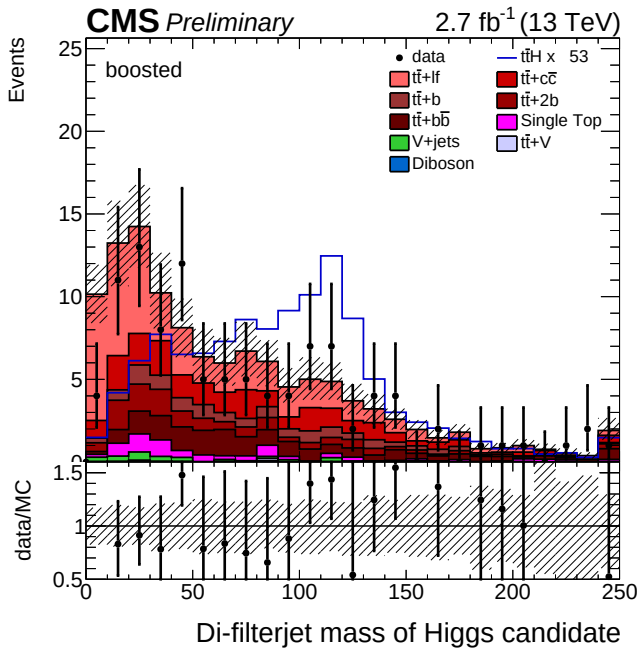


Figure 8.6: Invariant mass of the reconstructed boosted Higgs-boson candidate in the boosted analysis category displayed in units of GeV/c^2. The simulated processes are scaled to the event yield expected for an integrated luminosity of $\mathcal{L} = 2.7\,\text{fb}^{-1}$. For better visibility, the $t\bar{t}H$ signal process (blue line) is scaled to the sum of the expected event yields of all background processes. The background contributions are displayed as filled histograms. The shaded band shows the systematic uncertainties on the total event yields of all background processes for each bin of the distribution. Taken from [204].

reconstructed. This event reconstruction efficiency is three times as the one achieved by the χ^2 reconstruction applied to resolved events. As the distinctive features of the signal are more pronounced and not smeared out by combinatorial background, large reconstruction efficiencies increase the separation power the variables used for the final discrimination of signal against background, which is described in Section 9.3. An example is the clearly visible peak close to the Higgs mass in the distribution of the invariant mass of the Higgs-boson candidate for simulated $t\bar{t}H$ events displayed in Fig. 8.6.

8.2.2. Boosted-Event Categorization

The selected boosted events enter the analysis in form of an analysis category additional to the set of resolved analysis categories, which are based on the multiplicity of resolved jets and b-tags. While the resolved analysis categories are mutually exclusive by definition, the boosted analysis category contains events that are also selected by the resolved analysis categories. In order to avoid the double-counting of events and their information, events may be assigned to only a single category. Overlapping events selected by both types of analysis categories are assigned to the boosted analysis category and vetoed in the resolved analysis categories. This choice is motivated in the upcoming subsection. The event yields for recorded data and simulation corresponding to an integrated luminosity of $\mathcal{L} = 2.7\,\text{fb}^{-1}$ for all analysis categories are displayed in Table 8.3. Additionally, the

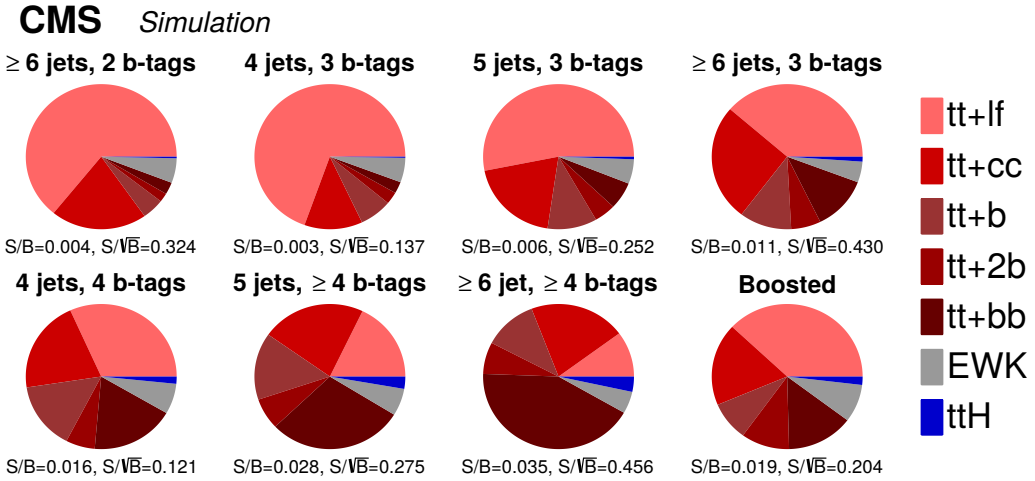


Figure 8.7: Fractions of selected signal and background events for every all analysis category. The event yields of the simulated processes are scaled to the values expected for an integrated luminosity of $\mathcal{L} = 2.7\,\mathrm{fb}^{-1}$. Single-top production, V+jets production, $\mathrm{t\bar{t}}$+V production, and diboson production are summarized by the electro-weak contribution (EWK).

fraction of signal and background events in each category is displayed in Fig. 8.7.

Based on the yields of selected events, the boosted analysis category shows a performance comparable to that of the best-performing resolved analysis categories requiring at least four b-tags or at least six resolved jets and three b-tags. Accordingly, the fraction of signal and background events and the signal significance of the boosted analysis are quite similar to those of the mentioned resolved analysis categories. The $\mathrm{t\bar{t}H}$ event yield of the boosted analysis category compares to the number of events selected in the resolved analysis categories featuring small event yields, which are the categories requiring at least 4 b-tags. The composition of background events in the boosted analysis category is most similar to the resolved $\geq$6 jets, 3 b-tags analysis category. Due to the small number of requirements on b-tagging information in the boosted-event selection, the largest fraction of background events is $\mathrm{t\bar{t}}$+light flavor production. By further exploiting the information provided by b-tagging in the final discrimination, the separation of $\mathrm{t\bar{t}}$+light-flavor background in the further course of the analysis is rather uncomplicated. The background contribution by $\mathrm{t\bar{t}b\bar{b}}$ production is much harder to separate because of the additionally radiated bottom quarks, which mimic the signature of the Higgs-boson. Nevertheless, the contribution of this process in the boosted analysis category is rather small compared to the best-performing resolved analysis categories requiring four b-tags. A characteristic feature of the boosted analysis category is the fraction of $\mathrm{t\bar{t}}$+2b background, which is enhanced compared to the resolved analysis categories. $\mathrm{t\bar{t}}$+2b events are characterized by the additionally produced bottom quarks being so close together that they are merged into a single jet. This particle configuration is very similar to the decay of a boosted Higgs boson into a bottom-quark pair and therefore rather likely to be selected by the boosted-event selection.

Table 8.3: Event yields of recorded data and simulated processes for all analysis categories considered in this analysis. The event yields of the simulated processes are scaled to the values expected for an integrated luminosity of $\mathcal{L} = 2.7\,\text{fb}^{-1}$.

Process	≥6 jets, 2 b-tags	4 jets, 3 b-tags	5 jets, 3 b-tags	≥6 jets, 3 b-tags
$t\bar{t}$+lf	5359.3 ± 1226.3	2026.1 ± 651.4	1000.2 ± 352.9	589.5 ± 199.7
$t\bar{t}$+$c\bar{c}$	1722.2 ± 849.5	363.2 ± 190.9	368.1 ± 191.3	396.6 ± 209.5
$t\bar{t}$+b	393.7 ± 188.2	203.1 ± 92.5	199.6 ± 90.8	170.8 ± 81.4
$t\bar{t}$+2b	165.2 ± 81.2	78.9 ± 38.0	87.2 ± 40.7	97.3 ± 46.8
$t\bar{t}$+$b\bar{b}$	226.4 ± 113.2	75.8 ± 35.3	114.1 ± 52.3	183.7 ± 86.7
Single Top	283.0 ± 49.0	115.3 ± 30.8	76.2 ± 19.5	47.5 ± 12.7
V+jets	130.5 ± 35.2	38.6 ± 17.8	22.8 ± 10.4	13.6 ± 6.4
$t\bar{t}$+V	43.5 ± 8.2	4.3 ± 1.2	6.4 ± 1.8	10.0 ± 2.7
Diboson	2.8 ± 1.3	2.1 ± 1.3	0.9 ± 0.5	0.2 ± 0.3
Total bkg	8326.7 ± 1788.6	2907.4 ± 836.5	1875.5 ± 534.7	1509.1 ± 423.7
$t\bar{t}$H	29.6 ± 2.1	7.4 ± 1.0	10.9 ± 1.2	16.7 ± 2.1
Data	7185	2793	1914	1386
S/B	0.0036	0.0026	0.0059	0.011
Data/B	0.9 ± 0.2	1.0 ± 0.3	1.0 ± 0.3	0.9 ± 0.3

Process	4 jets, ≥4 b-tags	5 jets, ≥4 b-tags	≥6 jets, ≥4 b-tags	boosted
$t\bar{t}$+lf	17.8 ± 10.8	17.7 ± 10.9	17.6 ± 11.3	45.1 ± 9.4
$t\bar{t}$+$c\bar{c}$	11.6 ± 8.2	22.1 ± 15.4	35.9 ± 24.9	21.8 ± 12.0
$t\bar{t}$+b	8.4 ± 4.4	14.8 ± 7.7	20.0 ± 10.9	10.3 ± 5.5
$t\bar{t}$+2b	3.5 ± 1.9	6.9 ± 3.7	12.3 ± 6.9	12.3 ± 6.6
$t\bar{t}$+$b\bar{b}$	10.1 ± 4.9	28.8 ± 13.9	73.4 ± 36.6	17.0 ± 8.4
Single Top	2.5 ± 1.1	4.3 ± 1.4	5.5 ± 2.0	7.0 ± 1.7
V+jets	1.0 ± 0.8	0.9 ± 0.8	1.4 ± 0.7	2.5 ± 0.8
$t\bar{t}$+V	0.3 ± 0.1	0.7 ± 0.3	1.6 ± 0.6	0.9 ± 0.3
Diboson	0.0 ± 0.0	0.1 ± 0.1	0.0 ± 0.0	0.1 ± 0.1
Total bkg	55.2 ± 23.0	96.5 ± 37.6	167.6 ± 65.7	117.0 ± 24.9
$t\bar{t}$H	0.9 ± 0.2	2.7 ± 0.6	5.9 ± 1.4	2.2 ± 0.3
Data	75	104	150	104
S/B	0.017	0.028	0.035	0.019
Data/B	1.4 ± 0.5	1.1 ± 0.4	0.9 ± 0.4	0.9 ± 0.2

Table 8.4: 95 % CL_s blinded expected upper limit on the signal strength $\mu(t\bar{t}H)$ for an exclusive assignment of events selected by both, the resolved ≥6 jets, ≥4 b-tags analysis category with high BDT output and the boosted analysis category, to either of them. Limits are presented for the resolved ≥6 jets, ≥4 b-tags analysis category, the boosted analysis category, and the combination of all categories considered in the analysis.

Overlap assigned	95 % CL_s expected upper limit on $\mu(t\bar{t}H)$		
	≥6 jets, ≥4 b-tags High BDT	Boosted	All
Resolved category	$7.7^{+4.1}_{-2.5}$	$13.4^{+6.6}_{-4.2}$	$3.6^{+1.7}_{-1.0}$
Boosted category	$8.2^{+4.4}_{-2.6}$	$10.1^{+5.3}_{-3.3}$	$3.6^{+1.7}_{-1.1}$

Overlap Treatment

Besides the choice of the working point, also the assignment of events selected by both, the boosted analysis category and the resolved analysis categories, to one of the analysis categories is studied. Taking away events from one category by assigning them to another, decreases the individual performance of the former category. To decide on the assignment of events that were selected by both a resolved and the boosted analysis category, a study based on the performance of the combination of all categories is carried out. The overlapping events of the boosted analysis category and the best performing resolved analysis categories, which require at least six resolved jets and at least four b-tags, are assigned to the resolved analysis category and to the boosted analysis category in turn. The overlap with the remaining resolved analysis categories is assigned to the boosted analysis category. For both configurations, the blinded expected upper limit on the signal strength modifier for the $t\bar{t}H$ process $\mu(t\bar{t}H)$ is calculated according to the procedures described in Section 5.2 and Chapter 11. The results for the individual analysis categories and the combination of all analysis categories are shown in Table 8.4. The individual categories include the boosted analysis category and the resolved ≥6 jets, ≥4 b-tags analysis category featuring a large final discriminant BDT output, which is described in Section 9.4. The latter is shown representative for the ≥6 jets, ≥4 b-tags analysis category, whereas the other resolved analysis categories are not affected by this study. The results for the individual categories show the behavior that was mentioned above. The performance of the analysis that excludes the overlapping events is reduced with respected to the case of the full event yield. The performance of the boosted analysis category decreases more by losing the overlapping events than the resolved ≥6 jets, ≥4 b-tags category. However, this effect is not propagated to the blinded expected upper limit of the combination of all categories, which is consistent for both cases. Based on the fact that the treatment of the overlap does not affect the final result, the overlapping events are assigned to the boosted analysis category.

8.3. Boosted-Object Validation

As for the resolved analysis objects, it is necessary to check the description data by simulation for the properties of boosted analysis objects. Large disagreements between the distributions in data and simulation may cause a bias in the results of the boosted analysis category, which strongly relies on the behavior of the boosted objects. The description

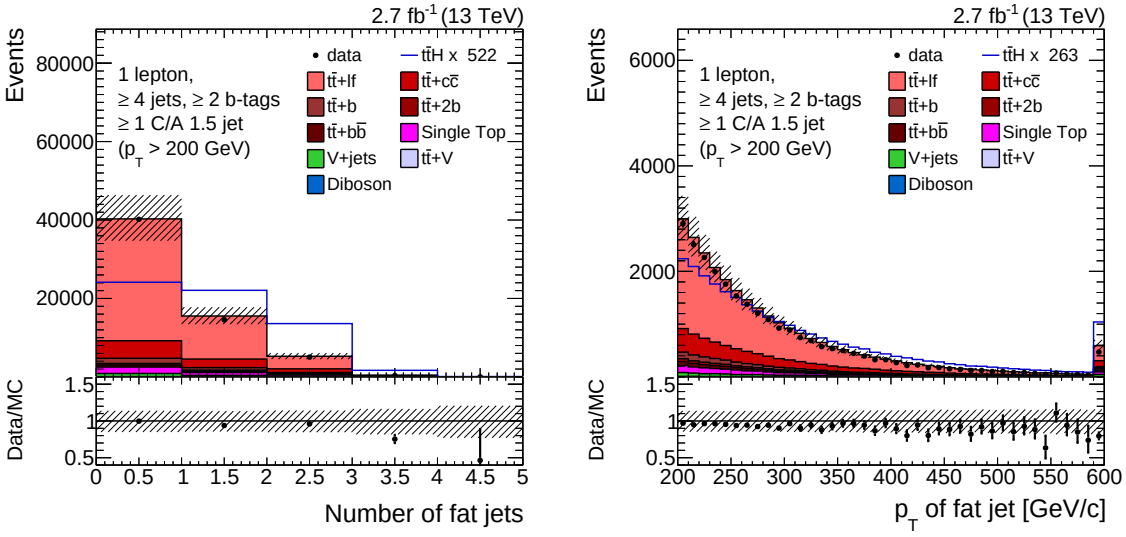


Figure 8.8: Fat-jet multiplicity (left) and transverse momentum of fat jets (right) shown in an inclusive control region requiring one selected electrically charged lepton, at least four resolved jets, and at least two b-tags. The variables are displayed only for selected fat jets with a transverse momentum of $p_{\mathrm{T}} > 200$ GeV/c. Simulated background processes are displayed as stacked filled histograms and are scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7$ fb^{-1}. The $\mathrm{t\bar{t}H}$ signal is depicted as a blue line and scaled to the total predicted event yield of all backgrounds for better visibility. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions.

of data by simulation for different boosted-variable distributions is checked in dedicated control regions enriched in background and without any major contribution by signal. Also for the boosted category, the main background is $\mathrm{t\bar{t}}$ production. Bad modeling of this process would have the largest impact on the outcome of the analysis. Accordingly, the main control region for validation is the $\mathrm{t\bar{t}}$ enriched region defined in Section 6.3. The control region is defined by the single-lepton baseline event selection requiring one good primary vertex and exactly one selected electrically charged lepton. Further, at least four resolved jets and two b-tags are required.

First the agreement between data and simulation for the properties of fat jets are checked, as fat jets represent the initial point of every boosted candidate. Fig. 8.8 shows the fat-jet multiplicity in the control-region events and the transverse momentum of the hardest fat jet. Both distributions show generally good agreement. Still, the fat-jet multiplicity distribution features a slight overestimation by simulation in the last filled histogram bin. This behavior is similar to the one observed for the resolved-jet multiplicity in Section 6.3. The transverse-momentum distribution of fat jets shows a slight constant offset. Both effects are not alarming, as they are covered by the uncertainties and can be balanced by the corresponding nuisance parameters in the evaluation of the final results.

Further, the transverse momenta and the b-tagging outputs of the subjets of reconstructed boosted top-quark candidates are displayed in Fig. 8.9. The variables are displayed for the fat jet with the largest transverse momentum. The agreement between data and simulation is worse than the one observed for the fat jets. Especially in the transverse momentum region featuring low values, the simulation overestimates the observation. Ac-

cordingly, the subjets observed are generally harder than the ones simulated. The ratio of the b-tagging distribution of the subjets between data and simulation show small wavelike disagreements, which are mainly present in the sparsely populated areas of the distributions. For all distributions, a small offset can be observed. Accordingly, simulation slightly overestimates the rate of reconstructed subjets. The mentioned disagreements are covered by the uncertainties and still acceptable. Further, distributions of variables associated to the boosted top-quark candidate are presented in Appendix A.2, where the input variables of the boosted top-quark identification BDT are shown.

Fig. 8.10 shows kinematic variables of subjets that are provided by the BDRS algorithm and used for the reconstruction of the boosted Higgs-boson candidate. It displays the transverse momentum and the b-tagging output for the three hardest subjets originating from the filtering. The observations are similar as for the boosted top-quark candidate subjets: a slight general offset due to overestimation by simulation, the subjets observed in data are harder than the ones predicted by simulation, and small wavelike discrepancies for the b-tagging output. Again, the disagreements are covered by uncertainties and therefore still acceptable.

The discrepancies observed hint at issues in the modeling of the substructure of fat jets, which is most likely due to the parton shower. In future iterations of this analysis, these discrepancies might not be acceptable anymore. In this case, the development and application of dedicated corrections are recommendable.

In this section, only a small selection of distributions of representative variables are discussed. Further distributions can be found in Appendix A.1.2.

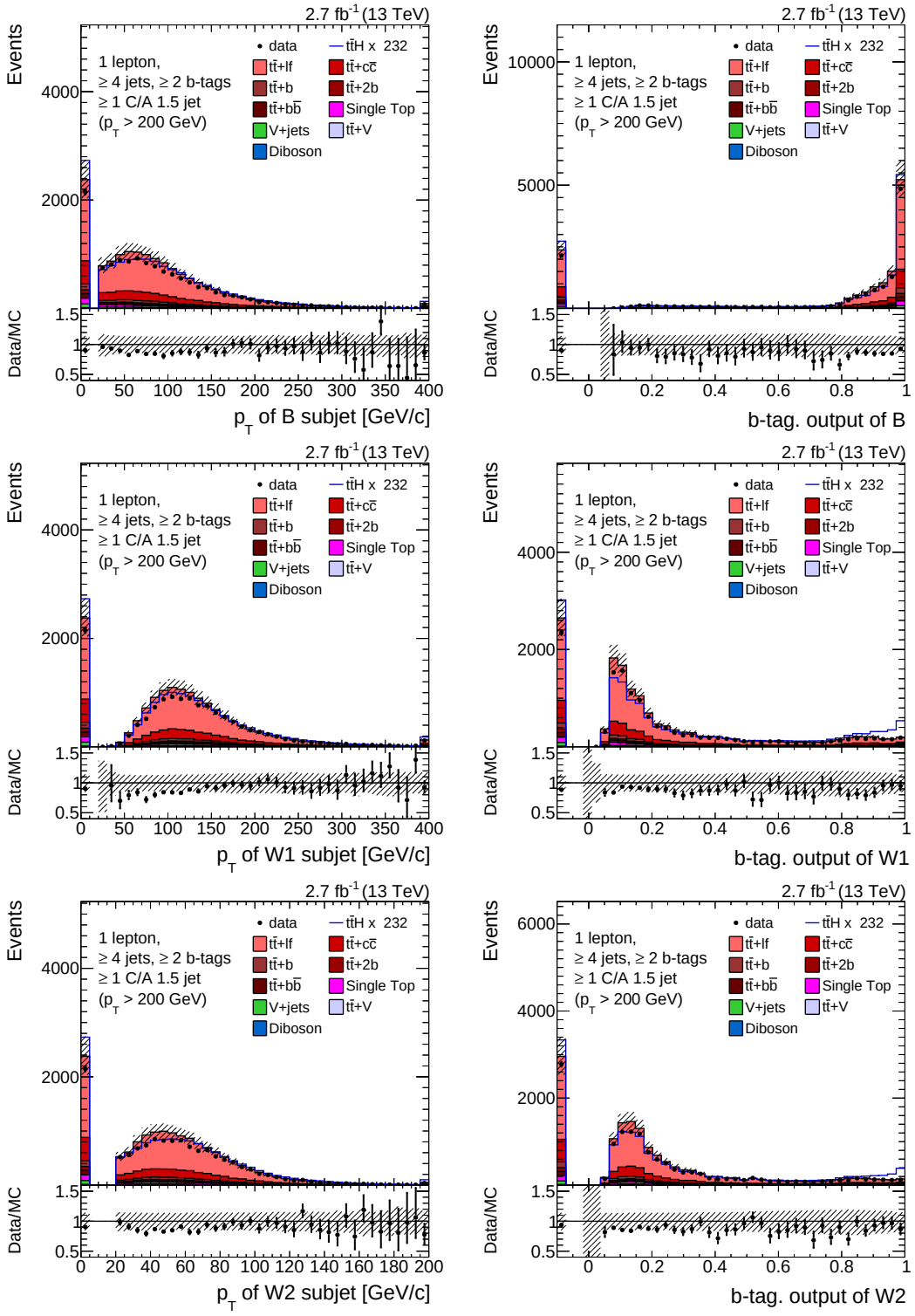


Figure 8.9: Transverse momentum (left) and b-tagging output (right) of all subjets provided by the HEPTopTaggerV2 shown in an inclusive control region requiring one selected lepton, at least four resolved jets, and at least two b-tags. The variables are displayed only for selected fat jets with a transverse momentum of $p_{\mathrm{T}} > 200$ GeV/c. Simulated background processes are displayed as stacked filled histograms and scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7$ fb^{-1}. The $\mathrm{t\bar{t}H}$ signal is depicted as a blue line and scaled to the total predicted event yield of all backgrounds for better visibility. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions. The underflow bins of the distributions are filled for fat jets for which no subjets have been found or the b-tagging algorithm has failed.

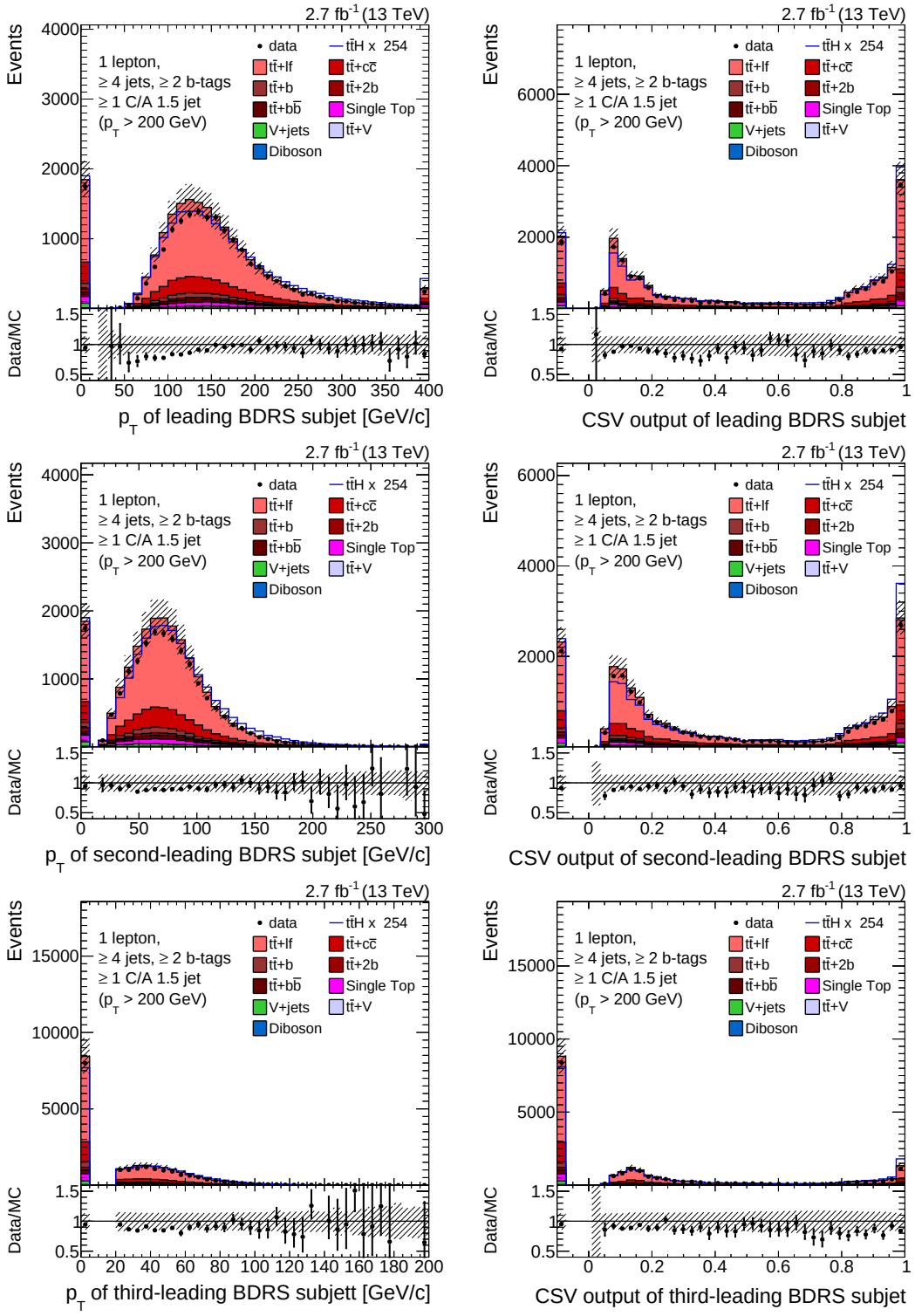


Figure 8.10: Transverse momentum (left) and b-tagging output (right) of subjets provided by the BDRS algorithm shown in an inclusive control region requiring one selected lepton, at least four resolved jets, and at least two b-tags. The variables are displayed only for selected fat jets with a transverse momentum of $p_{\mathrm{T}} > 200$ GeV/c. Simulated background processes are displayed as stacked filled histograms and scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7$ fb^{-1}. The $\mathrm{t\bar{t}H}$ signal is depicted as a blue line and scaled to the total predicted event yield of all backgrounds for better visibility. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions. The underflow bin of the distributions is filled for fat jets for which no subjets have been found or the b-tagging algorithm has failed.

Chapter 9

Final Discrimination

The event selections described in Chapter 6 and Chapter 8 reject a large fraction of the background events. For the main background, which is $t\bar{t}$ production, the fraction of rejected events makes up about 94 % of all events predicted. Still, in most analysis categories the event yields of background processes largely exceed the event yield of the signal process. By further separating signal events from background events based on the distinctive features of the processes, the sensitivity of the analysis is further improved. Such a separation can be achieved, for example, based on a single kinematic variable. Nevertheless, a combination of various characteristic variables provides a better separation of signal from background processes.

The main background process in the search for $t\bar{t}H$ production is $t\bar{t}$+jets production. As mentioned in Section 3.2.8, this process can be further subdivided into $t\bar{t}$+light-flavor and $t\bar{t}$+heavy-flavor contributions. A method that aims at separating these contributions by applying a likelihood ratio based on b-tagging information is introduced in Section 9.1. This b-tagging likelihood ratio information serves as input for the other discriminants presented in this chapter. One of the main variables used for the final discrimination of signal against background in this analysis is obtained with the matrix-element method (MEM). This physics-motivated variable uses the reconstructed objects described in Chapter 4 and Chapter 7 as input. From these inputs, probability densities are formed based on the differential cross sections of the signal and a background process as described in Section 9.2. The MEM [205–207] has been developed in 1988. The first major areas of application have been precision measurements at the Tevatron [208–210]. As mentioned in Section 1.3.3, the MEM has already successfully been used in the search for $t\bar{t}H$ production in LHC run I [76]. The other LHC-run-I analysis targeting the search for $t\bar{t}H$ production is based on a machine-learning approach [77]. The final discriminators used in this analysis are boosted decision trees (BDT), which combine kinematic variables for an optimal separation of $t\bar{t}H$ signal events against the main background, $t\bar{t}$ production. This approach provides the second main variable used for the final discrimination of signal against background processes in this analysis and is described in Section 9.3. Maximum sensitivity is reached by combining both discrimination approaches. In the analysis presented in this thesis, two types of combining the two discrimination methods have been considered: the training of the boosted decision trees with the matrix-element discriminant as input variable and a two-dimensional approach based on the splitting of categories with respect to the BDT discriminant and the application of the matrix-element discriminant for the final discrimination in each subcategory. The combination method for the evaluation of the final results has been chosen to achieve maximum performance in each category as described in Section 9.4.

9.1. b-Tagging Likelihood Ratio

The b-tagging likelihood ratio aims at separating $\mathrm{t\bar{t}}$+heavy-flavor ($\mathrm{t\bar{t}}$+hf) events including events from $\mathrm{t\bar{t}H}$ production from $\mathrm{t\bar{t}}$+light-flavor ($\mathrm{t\bar{t}}$+lf) events. This separation is achieved by constructing likelihood functions based on the b-tagging output of the resolved jets found in an collision event. $\mathrm{t\bar{t}}$+hf events are defined by featuring six resolved jets with four of them originating from bottom quarks, which corresponds to the ideal configuration of a $\mathrm{t\bar{t}}(\mathrm{H}\rightarrow\mathrm{b\bar{b}})$ event or a $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ event with a semileptonic decay of the top-quark pair. $\mathrm{t\bar{t}}$+lf events are defined analogously by requiring only two jets originating from bottom quarks instead of four. Based on these signatures, two likelihood functions,

$$f(\vec{\xi}|\mathrm{t\bar{t}}+\mathrm{hf}) = \sum_{i_1} \sum_{i_2 \neq i_1} \ldots \sum_{i_6 \neq i_1,\ldots,i_5} \left(\prod_{j \in \{i_1,i_2,i_3,i_4\}} f_{\mathrm{hf}}(\xi_j) \prod_{k \in \{i_5,i_6\}} f_{\mathrm{lf}}(\xi_k) \right) ,$$

$$f(\vec{\xi}|\mathrm{t\bar{t}}+\mathrm{lf}) = \sum_{i_1} \sum_{i_2 \neq i_1} \ldots \sum_{i_6 \neq i_1,\ldots,i_5} \left(\prod_{j \in \{i_1,i_2\}} f_{\mathrm{hf}}(\xi_j) \prod_{k \in \{i_3,i_4,i_5,i_6\}} f_{\mathrm{lf}}(\xi_k) \right) ,$$

are defined corresponding to the $\mathrm{t\bar{t}}$+hf or $\mathrm{t\bar{t}}$+lf hypothesis respectively. The vector $\vec{\xi}$ represents the b-tagging output of all jets. Correspondingly, ξ_i denotes the b-tagging output of the jet with index i. The probability density functions $f_{\mathrm{hf,lf}}$ return the probability for a jet originating from a heavy-flavor parton or a light-flavor parton to feature a particular b-tagging output value. The functions are derived by normalizing the b-tagging distributions of jets from simulated events, which are assigned a parton-flavor based on simulation information. The light-flavor probability density function is derived from jets stemming from up, down, and strange quarks. In the determination of the heavy-flavor probability density function, only jets originating from bottom quarks are considered. For simplicity, jets stemming from charm quarks are not considered. Charm-flavor jets feature b-tagging output distributions, which are more similar to the ones of bottom-flavor jets than light-flavor jets. Including charm-flavor jets in the b-tagging likelihood would lead to a degradation of the separation of light-flavor jets, which represent the largest fraction in background events. Accordingly, the b-tagging likelihood ratio is only approximate for events including W-boson decays or gluon splittings into charm quarks. A typical distribution of the b-tagging outputs for resolved jets in data and simulation is displayed in Fig. 6.5 in Section 6.3.

The b-tagging likelihood functions are constructed for the ideal number of jets expected for $\mathrm{t\bar{t}}$+hf and $\mathrm{t\bar{t}}$+lf events, which in both cases is six. The sums in the likelihood function add up the probabilities for all permutations of jets. In case more jets are found in the event, the six jets with the highest b-tagging output are used for the calculation of the likelihood function. The likelihood functions presented can also be adapted to a lower number of jets. In this special case, all of the heavy-flavor partons are assigned jets and the remaining jets are assigned to the light partons.

According to the Neyman-Pearson lemma described in Section 5.2.2, the ratio of two likelihood functions provides the most powerful discrimination between two hypothesis. Hence, the best distinction of $\mathrm{t\bar{t}}$+hf and $\mathrm{t\bar{t}}$+lf events is provided by the ratio of the two likelihood functions defined above,

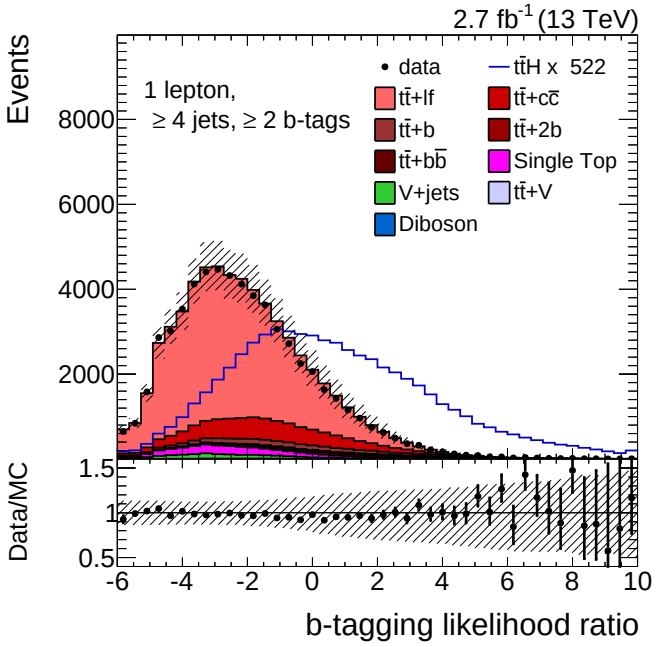


Figure 9.1: b-tagging likelihood ratio shown for an inclusive control region requiring one selected lepton, at least four resolved jets, and at least two b-tags. Simulated background processes are displayed as stacked filled histograms and are scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7\,\mathrm{fb}^{-1}$. The $\mathrm{t\bar{t}H}$ signal process is depicted as a blue line and is scaled to the total predicted event yield of all background processes for better visibility. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distribution.

$$\mathcal{F}(\vec{\xi}) = \frac{f(\vec{\xi}|\mathrm{t\bar{t}}+\mathrm{hf})}{f(\vec{\xi}|\mathrm{t\bar{t}}+\mathrm{hf}) + f(\vec{\xi}|\mathrm{t\bar{t}}+\mathrm{lf})} \,.$$

The output of the b-tagging likelihood ratio in an inclusive control region requiring one selected lepton, at least four resolved jets, and at least two b-tags is displayed in Fig. 9.1. Very good agreement between data and simulation is observed up to large values, which are dominated by statistical fluctuations.

9.2. Matrix-Element Method

The matrix-element method (MEM) provides a physics-motivated variable utilizing the distinctive kinematics and dynamics of the processes under investigation. The main constituents of this method are probability densities, which describe the probability for measuring a particular final state given a signal or a background-event hypothesis $\mathcal{H}$. Accordingly, the probability densities can be interpreted as differential cross sections of the respective signal or background process. The calculation relies on the usual concept for the derivation of cross sections of physics processes, which is described in Section 3.2. The final MEM discriminant is constructed by forming the ratio of the probability densities of the signal and the background hypotheses. In this analysis, the signal is $\mathrm{t\bar{t}}(\mathrm{H}\to\mathrm{b\bar{b}})$ production. As representation for all background processes, $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ production is chosen

for the background hypothesis. This process is very signal-like and therefore very hard to discriminate from signal. It also allows good discrimination of signal against other $\mathrm{t\bar{t}}$+jets background processes when used in the MEM. This has been as shown in the $\mathrm{t\bar{t}}(\mathrm{H}\to\mathrm{b\bar{b}})$ search using the MEM approach at a center-of-mass energy of $\sqrt{s} = 8$ TeV [76].

In the analysis presented in this thesis, the probability densities are calculated at leading order of perturbation theory. For simplicity, only gluon initiated $\mathrm{t\bar{t}}(\mathrm{H}\to\mathrm{b\bar{b}})$ and $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ production is considered, which represents the largest fraction at the LHC. A general expression for the probability densities is given by the multi-dimensional integral,

$$w(\vec{y}|\mathcal{H}) = \sum_{i}^{N} \int \frac{\mathrm{d}x_a \mathrm{d}x_b}{2x_a x_b s} \int \prod_{k}^{8} \left(\frac{\mathrm{d}^3\vec{p}_k}{(2\pi)^3 2E_k} \right) (2\pi)^4 \delta^{(E,z)} \left(p_a + p_b - \sum_{k}^{8} p_k \right) \mathcal{R}^{(x,y)} \left(\vec{\rho}, \sum_{k}^{8} p_k \right) \times g(x_a, \mu_\mathrm{F}) g(x_b, \mu_\mathrm{F}) |\mathcal{M}_\mathcal{H}(p_a, p_b, p_1, ..., p_8)|^2 W(\vec{y}_i, \vec{p}) \; . \tag{9.1}$$

In this formula, $\vec{y}$ represents the set of reconstructed particles. For the calculation of the probability density, it is necessary to associate the reconstructed particles with the final-state particles of the respective process. In this analysis, the top-quark pairs $\mathrm{t\bar{t}}(\mathrm{H}\to\mathrm{b\bar{b}})$ and $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ events are assumed to decay semileptonically. Accordingly, the reconstructed charged lepton is assigned to the charged lepton originating from the leptonically decaying top quark. The missing transverse energy is associated with the neutrino also stemming from the leptonic top-quark decay. The assignment of the jets to the final-state quarks is ambiguous. Jets have to be assigned to the bottom quarks from the top-quark decays, the light quarks from the hadronic W-boson decay, and to the bottom quarks from the decay of the Higgs boson or the additionally radiated bottom-quark pair. In order to handle these ambiguities, hypotheses, $\vec{y}_i$, corresponding to different permutations of the assignment of the jets to the expected particles are defined. The sum in the formula of the probability density runs over all N of these permutations. The four-momenta of the incoming protons in the collision event are $P_{a,b} = \left(\frac{\sqrt{s}}{2}, 0, 0, \pm\frac{\sqrt{s}}{2}\right)$. The four-momenta of the initial-state gluons are related to the proton momenta via the gluon-energy fractions $x_{a,b}$,

$$p_{a,b} = x_{a,b} P_{a,b} \; .$$

The integration is performed over the 26-dimensional phase space spanned by the four momenta of the eight final-state particles and the gluon-energy fractions by applying the VEGAS algorithm [211]. The delta function $\delta^{(E,z)}$ maintains the conservation of longitudinal momentum and absolute energy between the initial-state gluons and the final-state particles. The transverse momentum of the final-state particles is loosely constrained to the measured transverse recoil by the resolution function $\mathcal{R}^{(x,y)}$. The transverse recoil is defined as the negative vectorial sum of transverse momenta of the jets and the charged lepton added to the missing transverse energy. This aaproach is chosen in order to account for initial- and final-state radiation not accounted for by the leading-order matrix element. The functions g in Eq. (9.1) represent the PDFs of gluons in the incoming protons. For the calculation, the CTEQ65m PDF set is chosen, because of its leading order α_s parameterization, which is consistent with the one used for the calculation of the matrix element. The factorization scale μ_F is chosen as half the sum of the Higgs-boson mass and twice the top-quark mass for the signal hypothesis [212]. For the background hypothesis, μ_F is a dynamic scale equal to the quadratic sum of the transverse masses of all colored partons.

The scattering matrix $\mathcal{M}_{\mathcal{H}}$ in Eq. (9.1) is further factorized into the production of the intermediate resonances, their propagation, and their decay. The intermediate resonances are the top-quarks and the Higgs boson. The production of the resonance is the amplitude for the hard scattering. It is evaluated at leading order using OpenLoops [213]. Prior to the calculation, a Lorentz transformation corresponding to a boost into the rest frame of the considered reconstructed objects is performed. This procedure cancels the effects of possible initial-state radiation and artificially restores a leading-order configuration. The propagation of the resonance is described by the propagators of the top-quarks and the Higgs-boson given by Breit-Wigner functions. In the calculation, the narrow-width approximation [214] is applied reducing the expression to a product of delta functions. For the decay of the resonance, no spin correlations are considered. The effects of the hadronization and the detector are accounted for by the transfer functions W. These functions map the momentum four-vectors of the final-state particles to the momentum four-vectors of the reconstructed objects. In this analysis, the directions and momenta of leptons are assumed to be perfectly measured. The same is true for the direction of the quarks represented by the reconstructed jets. Based on these assumptions the transfer functions are reduced to a product of transfer functions for the energy of the quarks. The transfer functions are derived from simulated events.

In the resolved analysis categories, the calculation of the probability density functions is simplified by not assigning any jets to the decay products of the hadronically decaying W-boson. Instead, the four-momenta of these particles are integrated out in the determination of the probability density functions. This procedure is computationally more intensive, but has been found to perform equally well or better than a configuration with jets assigned to W-boson decay products. An additional advantage is the applicability of the MEM for events with only four reconstructed resolved jets. The procedure for evaluating the MEM probabilities in the resolved analysis categories considers up to eight jets in the event. If the number of resolved jets in the event exceeds this number, the eight jets with the largest transverse momenta are retained for the calculation. Out of these jets, the four jets used for the calculation of the probability densities are chosen based on the b-tagging likelihood ratio described in the previous section. This is accomplished by determining the permutation that provides the largest contribution to the b-tagging likelihood ratio. The four jets used are the ones associated to the heavy-flavor contributions for the given permutation. These four jets are assigned to the bottom quarks from the top-quark decays and the decay of the Higgs boson or the additional radiation. Taking into account that the permutation of the assignment of two jets to the two bottom quarks of the Higgs boson or additional radiation is redundant, the number of hypotheses adds up to 12.

For the calculation of the MEM probabilities in the boosted analysis category, the simplification applied in the resolved analysis categories is not considered. Hence, at least six input jets are required. One part of the input jets in the boosted category is provided by the subjets of the reconstructed boosted top-quark candidate. Distinct transfer functions for this kind of jets ensure an optimal and well-understood performance in the calculation of the probability densities. However, such transfer functions are not provided for the subjets resulting from the BDRS algorithm. For this reason, the subjets of the reconstructed boosted Higgs-boson candidate cannot be used for the calculation of the MEM probabilities. Instead, the remaining input jets are provided by resolved jets not spatially overlapping with the top-quark candidate subjets. Only resolved jets with an angular distance of $\Delta R > 0.3$ with respect to all top-quark candidate subjets are considered. From this set, the three jets with the largest b-tagging output are chosen for

the determination of the probability density functions. In the calculation of the MEM probabilities, the top-quark candidate subjet B, which is reconstructed as the bottom quark originating from the hadronic top-quark decay, is not fixed to this association, but can be also assigned to the bottom quarks stemming from the leptonic top-quark decay, the Higgs-boson decay, or additional radiation. This approach was initially chosen to add additional information by event interpretations alternative to the one found in the boosted reconstruction. However, a comparison with a configuration that fixes the subjet reconstructed as bottom quark from the boosted hadronically decaying top-quark to that association shows no significant difference. The subjets reconstructed as the light quarks from the hadronic W-boson decay W1 and W2 are assigned to the corresponding particles in the MEM-probability calculation. Considering the permutability of the association of jets to the decay products of the W-boson and the bottom quarks originating from the Higgs-boson decay or additional radiation, the number of permutation hypotheses again adds up to 12. An advantage of leaving out the simplification used for the resolved case is the omission of the computationally intensive integration performed for the W-boson decay products.

The final MEM discriminant is constructed as the ratio of the MEM probabilities of the $t\bar{t}(H\to b\bar{b})$ and $t\bar{t}+b\bar{b}$ hypotheses,

$$P_{\mathrm{MEM}} = \frac{w(\vec{y}|t\bar{t}(H\to b\bar{b}))}{w(\vec{y}|t\bar{t}(H\to b\bar{b})) + k\, w(\vec{y}|t\bar{t}+b\bar{b})} \,.$$

The parameter k adjusts the relative normalization between the signal and the background probability. In this analysis, the value $k = 0.15$ was chosen as it has shown to provide a good separation between signal and background events. The output of the MEM discriminant in an inclusive control region requiring one selected lepton, at least four resolved jets, and at least three b-tags is displayed in Fig. 9.2. Slight differences between data and simulation are observed, which are covered by the uncertainties.

9.3. Boosted Decision Tree Method

Boosted decision trees (BDTs) represent a machine learning approach to final discrimination of signal events against background. A general description of the construction, functioning, and optimization of BDTs is given in Section 5.1. In this analysis, the implementation of BDTs with gradient boosting in the software package TMVA [176] implemented in the software framework ROOT [177] is applied. The BDTs are optimized and trained using simulated $t\bar{t}(H\to b\bar{b})$ events as signal. Simulated $t\bar{t}$ events are used representative for all background processes, as they feature the largest contribution and at the same time pose the largest challenge in the separation from signal.The samples of simulated $t\bar{t}(H\to b\bar{b})$ and $t\bar{t}$ events presented in Section 3.2.7 are both split into two statistically independent samples. The events used for the training and the optimization of the BDTs are taken from one part of the samples, while the other samples are used for the evaluation of the final results.

The optimization and the training of the BDTs is carried out independently in each analysis category. An optimal set of training parameters and input variables is found using the particle-swarm optimization (PSO) described in Section 5.1.4. As input, variables associated to the following classes of observables are considered:

- MEM discriminant,

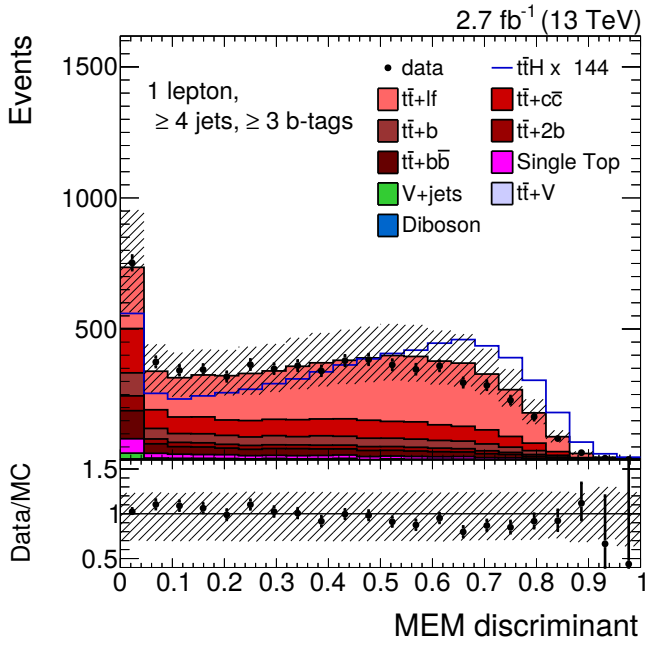


Figure 9.2: MEM discriminant shown for an inclusive control region requiring one selected lepton, at least four resolved jets, and at least three b-tags. Simulated background processes are displayed as stacked filled histograms and are scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7$ fb^{-1}. The $t\bar{t}H$ signal process is depicted as a blue line and is scaled to the total predicted event yield of all background processes for better visibility. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distribution.

- Variables using information from the resolved-event reconstruction,
- Variables using information from the boosted-event reconstruction,
- Variables based on resolved object and event kinematics,
- Event-shape variables,
- b-tagging variables.

Before feeding input variables into the particle-swarm algorithm, a rough selection is performed by hand. This selection is based on the separation abilities and the quality of the variables. The description of data by simulation for the variables as well as the correlations between them is verified in different control regions enriched with events from $t\bar{t}$ and V+jets production. Additionally, the agreement of data and simulation is checked in the signal regions represented by the various analysis categories. Variables that feature only minor separation or are not well described by simulation are not included in the optimization. Some input variables are only applied in particular analysis categories. The MEM disriminant, for example, is only used in the analysis categories requiring three b-tags and the boosted analysis category. The reasons for this are explained in Section 9.4. Other examples are boosted event-reconstruction variables, which are only available in the boosted analysis category. Besides these variables, also the ones provided by the alternative resolved-event interpretation are used as input for the training of the BDT in the

Table 9.1: Parameter configuration used for the training of the boosted decision trees in each analysis category. The values are obtained by an optimization procedure based on the particle-swarm optimization described in Section 5.1.4.

Analysis category	N_{trees}	Shrinkage	Bagging fraction	N_{cuts}	Depth
$\geq$6 jets, 2 b-tags	642	0.05	0.37	20	2
4 jets, 3 b-tags	1347	0.04	0.37	39	2
5 jets, 3 b-tags	898	0.02	0.59	46	2
$\geq$6 jets, 3 b-tags	552	0.04	0.83	57	2
4 jets, $\geq$4 b-tags	1113	0.01	0.36	25	2
5 jets, $\geq$4 b-tags	538	0.02	0.66	30	2
$\geq$6 jets, $\geq$4 b-tags	1315	0.01	0.35	64	2
boosted	737	0.03	0.63	22	2

boosted analysis category. The PSO configuration includes 20 particles, each representing a particular BDT configuration, which is optimized in the course of 100 iterations. The initial configuration for this procedure is obtained by choosing nine input variables at random for each particle. The BDT training performed during each iteration is repeated three times in order to reduce the influence of statistical fluctuations. The value chosen for the Kolmogorov-Smirnov probability threshold is 0.1. The parameter-search space is reduced by fixing the depth of the decision trees to two levels, which represents a standard choice for BDTs with gradient boosting and a large number of trees. The optimal parameter configuration obtained for each analysis category is listed in Table 9.1. The optimal set of input variables used for the training of the BDT in each category is displayed in Table 9.2. In the following, a short description of the variables used for the training of the BDTs in all categories is given. The variables are grouped based on the classes mentioned above.

MEM discriminant: The MEM discriminant presented in Section 9.2 combines the kinematics and dynamics of the given processes with the reconstructed objects. Therefore, it represents a BDT input variable with very powerful background discrimination abilities. However, the MEM discriminant is only applied as input variable for the BDTs trained in the resolved analysis categories requiring exactly three b-tags and the boosted analysis category. This due to the special combination procedure of BDT and MEM approach for the final discrimination of signal against background described in Section 9.4. In the remaining categories, where the MEM discriminant is available, its information is used in an alternative way.

Resolved-event reconstruction variables: Variables based on the objects obtained by the resolved-event reconstruction described in Section 6.2 provide further inputs for the training of the BDTs. One of them is the invariant mass of the Higgs-boson candidate $\boldsymbol{m}$**(resolved Higgs cand.)**. As $t\bar{t}H$ production is the only relevant process featuring a real Higgs boson in this analysis, this variable is well suited to discriminate between signal and background events. The second resolved reconstruction variable used in the training of the BDTs $\sqrt{\boldsymbol{\Delta\eta(t_{lep.}, b\bar{b})} \times \boldsymbol{\Delta\eta(t_{had.}, b\bar{b})}}$ probes the angular correlations between the reconstructed Higgs boson and the top quarks. The distributions of both variables for data and simulation in the resolved $\geq$6 jets, $\geq$4 b-tags and $\geq$6 jets, 3 b-tags analysis cate-

Table 9.2: Input variables used for the training of the boosted decision trees in each analysis category. The variables are determined by an optimization procedure based on the particle-swarm optimization described in Section 5.1.4.

≥4 jets, ≥2 b-tags boosted	**4 jets, 3 b-tags**	**4 jets, ≥4 b-tags**
MEM discriminant (using subjets)	MEM discriminant	p_{T}(hardest jet)
m(boosted Higgs cand.)	p_{T}(second-hardest jet)	$\sum p_{\mathrm{T}}$(jets,lepton,MET)
τ_2/τ_1(boosted Higgs cand.)	p_{T}(fourth-hardest jet)	m(closest b-tag. jets)
$\Delta\eta$(top cand.,Higgs cand.)	$\sum p_{\mathrm{T}}$(jets,leptons,MET)	avg. ΔR(b-tag. jets)
avg. ΔR(b-tag. jets)	avg. b-tag. output (b-tag. jets)	b-tagging likelihood ratio
min. ΔR(b-tag. jets)	avg. b-tag. output (all jets)	H_3
third-highest b-tag. output	b-tagging likelihood ratio	
fourth-highest b-tag. output	H_1	
avg. b-tag. output (all jets)		
b-tagging likelihood ratio		
aplanarity		
	5 jets, 3 b-tags	**5 jets, ≥4 b-tags**
	MEM discriminant	p_{T}(third-hardest jet)
	min. ΔR(lepton,jet)	$(\Sigma p_{\mathrm{T}}(\mathrm{jet}))/(\Sigma E(\mathrm{jet}))$
	avg. $\Delta\eta$(jets)	avg. $\Delta\eta$(all jets)
	max. $\Delta\lvert\eta\rvert$(b-tag. jet, all jets avg.)	b-tag. dijet mass closest to m_{H}
	avg. ΔR(b-tag. jets)	avg. ΔR(b-tag. jets)
	fourth-highest b-tag. output	fifth-highest b-tag. output
	avg. b-tag. output (all jets)	b-tagging likelihood ratio
	avg. b-tag. output (b-tag. jets)	H_1
	dev. avg. b-tag. output	
	b-tagging likelihood ratio	
	H_1	
≥6 jets, 2 b-tags	**≥6 jets, 3 b-tags**	**≥6 jets, ≥4 b-tags**
ΔR(hardest jet, second-hardest jet)	MEM discriminant	m(resolved Higgs cand.)
avg. ΔR(b-tag. jets)	$\sqrt{\Delta\eta(\mathrm{t_{lep.}}, \mathrm{b\bar{b}}) \times \Delta\eta(\mathrm{t_{had.}}, \mathrm{b\bar{b}})}$	p_{T}(fourth-hardest jet)
min. ΔR(b-tag. jets)	H_{T}	$(\Sigma p_{\mathrm{T}}(\mathrm{jet}))/(\Sigma E(\mathrm{jet}))$
avg. $\Delta\eta$(b-tag. jets)	$\sum p_{\mathrm{T}}$(jets,lepton,MET)	$\sum p_{\mathrm{T}}$(jets,leptons,MET)
max. $\Delta\lvert\eta\rvert$(b-tag. jet, all jets avg.)	max. $\Delta\lvert\eta\rvert$(b-tag. jet, all jets avg.)	b-tag. dijet mass closest to m_{H}
max. $\Delta\lvert\eta\rvert$(b-tag. jet, b-tag. jets avg.)	fourth-highest b-tag. output	max. $\Delta\lvert\eta\rvert$(b-tag. jet, b-tag. jets avg.)
third-highest b-tag. output	avg. b-tag. output (b-tag. jets)	second-highest b-tag. output
fourth-highest b-tag. output	b-tagging likelihood ratio	fifth-highest b-tag. output
b-tagging likelihood ratio	H_1	b-tagging likelihood ratio
sphericity		sphericity
		H_3

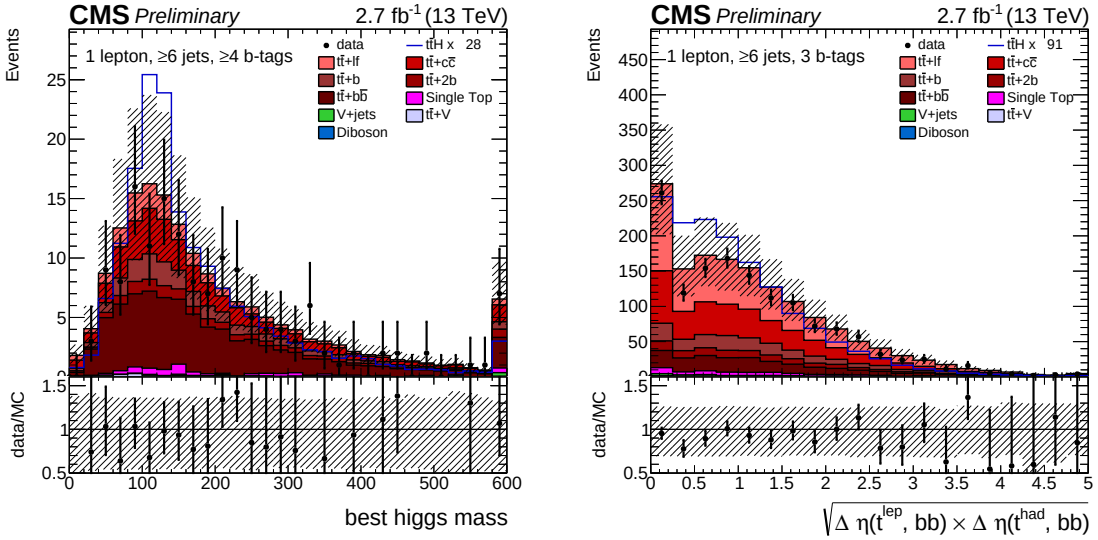


Figure 9.3: Invariant mass of the Higgs-boson candidate in units of GeV/c^2 in the resolved $\geq$6 jets, $\geq$4 b-tags analysis category (left) and square root of the product of the differences in pseudo rapidity between the Higgs-boson candidate and the leptonic or hadronic top-quark candidate in the resolved $\geq$6 jets, 3 b-tags analysis category (right). The candidates are obtained from the resolved-event reconstruction. Simulated background processes are displayed as stacked filled histograms and are scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7\,\text{fb}^{-1}$. The $\text{t}\bar{\text{t}}\text{H}$ signal process is depicted as a blue line and is scaled to the total predicted event yield of all background processes for better visibility. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions. Taken from [204].

gories are shown in Fig. 9.3.

Boosted-event reconstruction variables: The most discriminating variable derived from the boosted-event reconstruction described in Section 8.1 is the invariant mass of the Higgs-boson candidate $\boldsymbol{m}$**(boosted Higgs cand.)**. The distribution of this variable for data and simulation is shown in Fig. 8.6. There a peak slightly below the Higgs-boson mass can be observed for the signal, while background processes accumulate at lower values. The separation of events with fake Higgs-boson candidates is further supported by the ratio of 2-subjettiness and 1-subjettiness of the boosted Higgs-boson candidate $\boldsymbol{\tau_2/\tau_1}$ **(boosted Higgs cand.)**, which discriminates against candidates with fewer than two hard prongs inside the fat jet. The angular correlation between the reconstructed boosted Higgs-boson candidate and the boosted top-quark candidate is probed by the difference in pseudo rapidity between these candidates $\boldsymbol{\Delta\eta}$**(top cand.,Higgs cand.)**. Due to exclusive availability of the objects, the variables derived from the boosted-event reconstruction are solely used in the boosted analysis category. Distributions of the latter two variables for data and simulation in the boosted category are presented in Fig. 9.4.

Resolved object and event kinematics variables: The distinctive features of $\text{t}\bar{\text{t}}\text{H}$ events are not only reflected in reconstructed Higgs bosons and top quarks. A large amount of separation power can already be found in low-level observables based on reconstructed resolved jets and leptons. Their momentum, invariant masses, and angular correlations

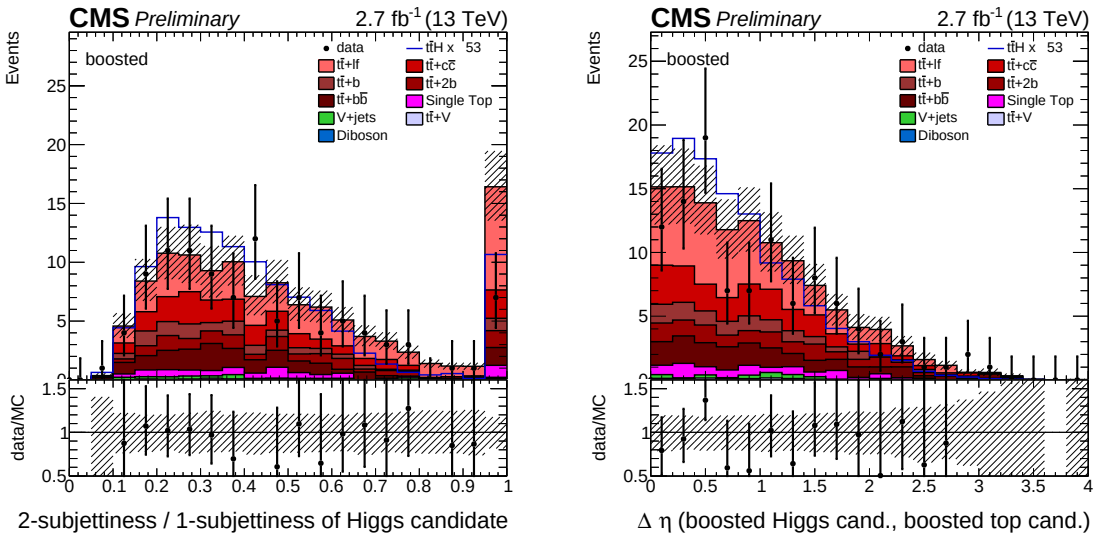


Figure 9.4: Ratio of 2-subjettiness and 1-subjettiness of the boosted Higgs-boson candidate (left) and difference in pseudo rapidity between the boosted Higgs-boson candidate and the boosted top-quark candidate (right) both shown for the boosted analysis category. The candidates are obtained from the boosted-event reconstruction. Simulated background processes are displayed as stacked filled histograms and are scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7\,\text{fb}^{-1}$. The $\text{t}\bar{\text{t}}\text{H}$ signal process is depicted as a blue line and is scaled to the total predicted event yield of all background processes for better visibility. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions. Taken from [204].

depict the kinematic properties of the underlying process. Variables associated to this class are described in the following. Some example distributions of a small selection of these variables for data and simulation are displayed in Fig. 9.5.

- Transverse-momentum variables:
  With a Higgs boson decaying into two bottom quarks and a semileptonic top-quark pair decay the final state of $\text{t}\bar{\text{t}}\text{H}$ production is dominated by objects from decays of massive particles. The large amount of energy in the event leads to final-state objects with large transverse momenta. Some input variables that are used for the training of BDTs for the final-discrimination and that reflect this characteristic are the transverse momenta of the four hardest jets $\boldsymbol{p}_\text{T}$**(*i*-th hardest jet)**. More inclusive input variables based on the total transverse momentum in the $\text{t}\bar{\text{t}}\text{H}$ event are the sum of the transverse momenta of all jets $\boldsymbol{H}_\text{T}$ and the sum of all transverse momenta of all jets, charged leptons, and the missing transverse energy $\sum \boldsymbol{p}_\text{T}$**(jets,lepton,MET)**.

- Invariant-mass variables:
  As already mentioned in Chapter 6 and Chapter 8 dealing with the reconstruction of $\text{t}\bar{\text{t}}\text{H}$ events, one of the most characteristic features of $\text{t}\bar{\text{t}}\text{H}$ production is the Higgs boson. Especially the invariant mass of the Higgs boson provides good distinction from gluon splittings into bottom-quark pairs and therefore provides one of the few handles on the discrimination of $\text{t}\bar{\text{t}}+\text{b}\bar{\text{b}}$ background. Two variables aiming at reconstructing the invariant mass of the Higgs boson are the invariant mass of the

two closest b-tagged jets $\boldsymbol{m}$**(closest b-tag. jets)** and the invariant mass of the two tagged jets closest to the Higgs-boson mass $m_{\mathrm{H}} = 125\,\mathrm{GeV}/c^2$ (**b-tag. dijet closest to** $\boldsymbol{m}_{\mathrm{H}}$).

- Angular-correlation variables:
  Another group of distinctive input variables considered for the training of the final discrimination BDTs are the angular correlations between reconstructed objects. These variables aim at the busyness of $\mathrm{t\bar{t}}(\mathrm{H}\to\mathrm{b\bar{b}})$ events and the angular features defined by the decays of the massive particles. Some of the angular correlation variables are the average or the minimum angular distance between two b-tagged jets **avg./min.** $\boldsymbol{\Delta R}$**(b-tag. jets)**. The average difference in pseudo rapidity of two (b-tagged) jets **avg.** $\boldsymbol{\Delta\eta}$**((b-tag.) jets)** represents another set of variables relying the pseudo rapidity of reconstructed objects. Complementary features are covered by the maximum difference between the pseudo rapidity of any (b-tagged) jet and the average pseudo rapidity of all (b-tagged) jets **max.** $\boldsymbol{\Delta|\eta|}$**((b-tag.) jet, all/b-tag. jets avg.)**. A further variable aiming at the features of the leptonic top-quark decay is the minimum angular distance between the reconstructed primary lepton and any jet **min.** $\boldsymbol{\Delta R}$**(lepton,jet)**.

Event shape: Event-shape variables probe the geometrical structure of the final state of a collision event. They combine different aspects like the number of objects in an event, the angular correlation between them, and the energy distribution. Main characteristics causing differences in event shape in between $\mathrm{t\bar{t}H}$ production and background processes are the massive-particle decays and the large number of particles in the final state. In the following, two major groups of event-shape variables are described, the sphericity and aplanarity and the Fox-Wolfram-Moments.

- Sphericity and aplanarity:
  Two variables probing the degree of sphericity of event topologies are given by the **sphericity** and **aplanarity** [215]. The determination of these variables is based on an eigenvector problem searching the rotation axis of a three dimensional body formed by the momentum vectors of resolved jets, charged leptons, and missing transverse energy. The starting point for calculating these variables is the sphericity tensor constructed from the momentum vectors (p_i^x, p_i^y, p_i^z),

$$S^{ab} = \frac{\sum_i^N p_i^a p_i^b}{\sum_i^N |\vec{p}_i|^2} \qquad \text{with} \quad a, b = x, y, z \; .$$

  Solving the eigenvector problem based on the sphericity tensor results in three eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$, which add up to unity. The variable sphericity is defined by the sum of the two smallest eigenvalues,

$$S = \frac{3}{2}(\lambda_2 + \lambda_3) \; .$$

  The sphericity takes on values between zero and one, in which case $S = 1$ represents a spherically symmetric distribution of objects. The other extreme $S = 0$ is given by a classic back-to-back configuration with objects being located in two narrow cones.

Aplanarity is defined by only taking into account the smallest eigenvalue,

$$A = \frac{3}{2}\lambda_3 \ .$$

This observable takes on values between zero and 1/2 and distinguishes spherical configurations from planar configurations. For final-state configurations returning an aplanarity of $A = 0$, all objects can be found in one plane. Similar to the sphericity, the upper boundary represents spherically symmetric distributions of reconstructed objects.

Due to the large number of final-state particles, $t\bar{t}H$ events tend to feature values of sphericity and aplanarity that are larger than background processes.

- Fox-Wolfram-Moments:
  Another type of event-shape variables based on the resolved jets in the event are the Fox-Wolfram-Moments [216]. These variables originate from expansions of jet-jet correlations in form of spherical harmonics and are sensitive to the number of resolved jets, their angular correlations, and the energy distribution in the event. The expansion is constructed by summing all jet-jet correlations weighted by their momentum over all $2l + 1$ directions described by spherical harmonics. A single Fox-Wolfram-Moment is defined by

$$H_l = \sum_{i,j}^{N} \frac{|\vec{p}_i||\vec{p}_j|}{E_{\text{vis}}^2} P_l(\cos\Omega_{ij}) \ .$$

  In this equation, the visible Energy E_{vis} is the sum of the energies of all jets in the event. The function $P_l(x)$ is the Legendre polynomial for a given l. The angle Ω_{ij} denotes the angle between the jets with indices i and j. In this analysis, the Fox-Wolfram-Moments $\boldsymbol{H_1}$ and $\boldsymbol{H_3}$ have proven to provide the best separation power and are therefore used in the training for the final-discrimination BDTs.

  A more detailed description of the Fox-Wolfram-Moments can be found in [217].

b-tagging variables: Due to the large number of bottom quarks produced in $t\bar{t}(H\to b\bar{b})$ events, b-tagging is a crucial part in the separation of most background processes. Corresponding to the expected number of bottom quarks in the final state of a $t\bar{t}(H\to b\bar{b})$ event, the b-tagging output values of the jets with four largest b-tagging outputs (***i*-highest b-tag. output**) are used in the training of the BDTs. Additionally, the average b-tagging output of all (b-tagged) jets (**avg. b-tag. output (all/b-tag. jets)**) tests the fraction of jets stemming from light quarks and bottom quarks. A complementary variable for testing the flavor composition of the event is the sum of the quadratic deviation of the b-tagging output of every b-tagged jet with respect to the average b-tagging output of all b-tagged jets (**dev. avg. b-tag. output**),

$$\Delta^2\text{CSV}_{\text{avg.}} = \sum_{i}^{N} \left(\text{CSV}_i - \frac{\sum_j^N \text{CSV}_j}{N} \right)^2 \ .$$

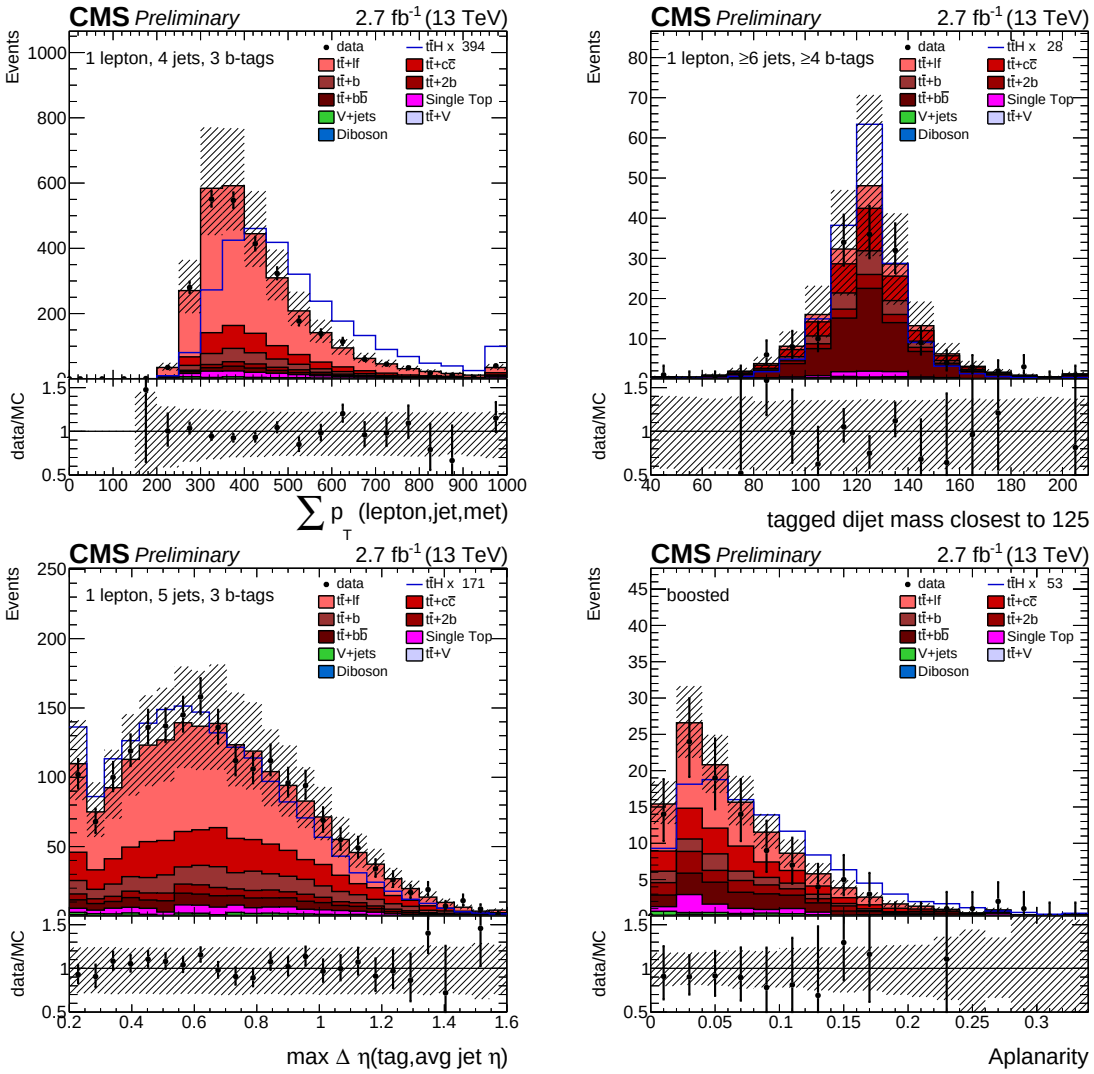


Figure 9.5: Sum of transverse momenta of all jets, charged leptons, and the missing transverse momentum in units of GeV/c in the resolved 4 jets, 3 b-tags analysis category (top left), b-tagged resolved dijet mass closest to the Higgs-boson mass $m_{\mathrm{H}} = 125$ GeV/c^2 in units of GeV/c^2 in the resolved $\geq$6 jets, $\geq$4 b-tags analysis category (top right), the maximum difference in pseudo rapidity between b-tagged resolved jets and the average pseudo rapidity of all resolved jets in the resolved 5 jets, 3 b-tags analysis category (bottom left), and aplanarity in the boosted analysis category (bottom right). Simulated background processes are displayed as stacked filled histograms and are scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7$ fb^{-1}. The t$\bar{\mathrm{t}}$H signal process is depicted as a blue line and is scaled to the total predicted event yield of all background processes for better visibility. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions. Taken from [204].

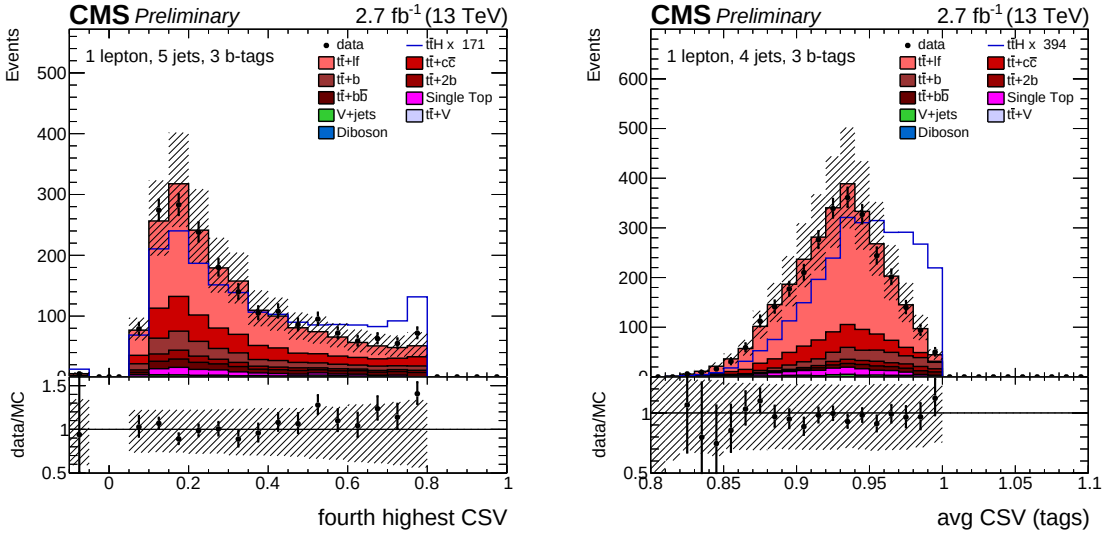


Figure 9.6: Fourth largest b-tagging output in the resolved 5 jets, 3 b-tags analysis category (left) and average b-tagging output of b-tagged jets in the resolved 4 jets, 3 b-tags analysis category (right). Simulated background processes are displayed as stacked filled histograms and are scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7\ \text{fb}^{-1}$. The $\text{t}\bar{\text{t}}\text{H}$ signal process is depicted as a blue line and is scaled to the total predicted event yield of all background processes for better visibility. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions. Taken from [204].

In this equation, the sum goes over all resolved jets in the event. The parameter CSV_i represents the b-tagging output of the resolved jet with index i. In case of a $\text{t}\bar{\text{t}}(\text{H}\to\text{b}\bar{\text{b}})$ event, the value of this variable is small, due to the large fraction of real bottom-quark jets. A further input variable of this class, the b-tagging likelihood ratio, has already been introduced in Section 9.1. It is constructed to separate $\text{t}\bar{\text{t}}$+heavy-flavor events, which also include signal, from $\text{t}\bar{\text{t}}$+light-flavor events based on b-tagging information. Examples of this category of input variables are provided by the distributions of the fourth largest b-tagging output in the resolved 5 jets, 3 b-tags analysis category and the average b-tagging output of b-tagged jets in the resolved 4 jets, 3 b-tags analysis category for data and simulation displayed in Fig. 9.6.

The output resulting from the evaluation of the BDTs on recorded data and the simulated samples also used for the determination of the final results are shown in Fig. 9.7, Fig. 9.9, and Fig. 9.11. The overall separation of signal and background events in the different categories reaches from reasonably well in the resolved $\geq$6 jets, 2 b-tags category to very well in the boosted analysis category. This result is driven by the composition of different background processes and the separation power of the variables in each analysis category. The different fractions of event yields of signal and background processes in the analysis categories are covered in Chapter 6 and Chapter 8. $\text{t}\bar{\text{t}}$+light-flavor events, for example, are typically very well distinguishable from signal events based on b-tagging information. However, in the resolved $\geq$6 jets, 2 b-tags analysis category the b-tagging information is biased by the low number of b-tagged jets. The $\text{t}\bar{\text{t}}+\text{b}\bar{\text{b}}$ contribution shows BDT outputs most similar to the ones provided by signal, which is due to their similar event signature. The BDT output in the boosted analysis category showss the best performance

in separating $t\bar{t}H$ events from background processes. The large discrimination power of the BDT in this category can be explained by the reduced combinatorics in the boosted-event reconstruction and the resulting large reconstruction efficiency. This improves the separation power of the variables provided by the boosted-event reconstruction. Concerning the resolved variables used in the boosted category, the variables based on b-tagging information are especially important as they provide a handle on the discrimination of the large fraction of $t\bar{t}$+light-flavor events.

9.4. Combination of Methods

The $t\bar{t}H$ search presented in this thesis combines the BDT and the MEM method with the target of gaining sensitivity compared to the use of a single approach. The additional information provided by the combination is tested by studying the correlations between the two discriminants. Indeed the two discriminants are found to perform differently. Part of this difference in performance stems from the fact that the MEM discriminant by construction is stronger at separating $t\bar{t}$+$b\bar{b}$ events. The BDTs, on the other hand, are more efficient at discriminating the inclusive background.

For the combination of the MEM and BDT discriminants, two approaches are considered:

- **MEM in BDT:** The MEM-in-BDT approach includes the MEM discriminant as an input variable in the training of the BDT. Accordingly, the distribution of the BDT output is used to determine the final results. More information on the training of BDTs with the MEM discriminant as input variable is presented in the previous section.

- **2D:** This approach is based on the subdivision of the analysis categories into subcategories based on the output of a BDT excluding the MEM discriminant as input. The categories are split at the median of the signal BDT distribution in order to ensure a sufficient number of events in both subcategories. In this way, subcategories with high and low BDT outputs are created, which correspond to phase-space regions enriched in signal events or background events. In each subcategory, the distribution of the MEM discriminant is used to determine the final results.

The choice of the configuration used in the analysis is based on a study investigating the impact of different combination approaches. The scenarios under investigation are defined based on the choice of the combination method in each analysis category. Exceptions are the resolved analysis category requiring exactly two b-tags and the boosted analysis category, where the final discriminants are fixed. For events with only two b-tags the calculation of the MEM discriminant is omitted due to their large number and the connected large computational effort. Accordingly, the combination of MEM and BDT is not possible in the resolved $\geq$6 jets, 2 b-tags analysis category and a BDT trained without the MEM discriminant is used for final discrimination. In order to optimally exploit the boosted reconstruction variables, which serve as input to the training of the BDT in the boosted analysis category, the MEM-in-BDT approach is used in this category. For the remaining categories, the following configurations have been tested:

- **Full MEM-in-BDT scenario:** The MEM-in-BDT approach is applied in all resolved analysis categories requiring at least three b-tags.

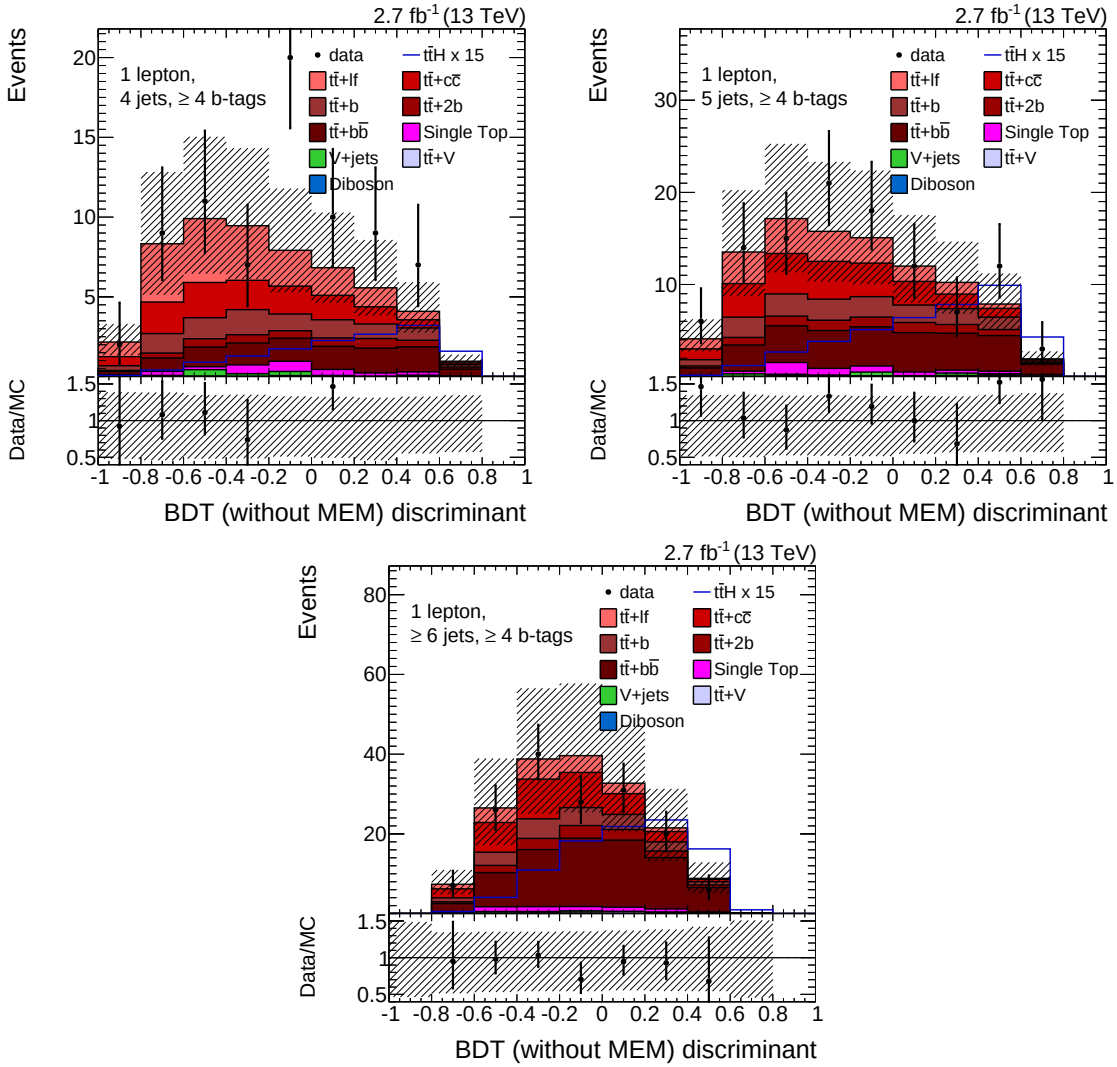


Figure 9.7: Distributions of the BDT outputs evaluated with data and simulated events in the resolved analysis categories requiring at least four b-tags. The simulated processes are scaled to the event yield expected for an integrated luminosity of $\mathcal{L} = 2.7\ \text{fb}^{-1}$. For better visibility, the $\text{t}\bar{\text{t}}\text{H}$ signal process (blue line) is additionally scaled by a factor of 15. The background contributions are displayed as stacked filled histograms. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions. The distributions of the BDT outputs of the remaining analysis categories can be found in Fig. 9.9 and Fig. 9.11.

- **Full 2D scenario:** The 2D approach is applied is applied in all resolved analysis categories requiring at least three b-tags.
- **Combined MEM-in-BDT+2D scenario:** The MEM-in-BDT approach is applied in all resolved analysis categories requiring exactly three b-tags. The 2D approach in all resolved analysis categories requiring at least four b-tags.

The scenarios are benchmarked with respect to the blinded expected upper limit on the signal-strength modifier, which is derived as described in Section 5.2 and Chapter 11. In this study, the different configurations have been found to yield very similar results, which are in good agreement with respect to their statistical precision. Based on these results, the combined MEM-in-BDT+2D approach featuring is chosen. This approach makes use of the BDTs that include the MEM discriminant in the resolved analysis categories requiring exactly three b-tagged resolved jets. The large number of events found in these categories ensures stable results in the training of the BDTs. The resolved analysis categories requiring at least four b-tags, on the other hand, are evaluated based on the discriminants provided by the 2D approach. In the resolved analysis categories requiring at least four b-tags, only a small number of simulated events are available for the training of the BDT after the event selection. By using the 2D approach and applying only a single cut on the BDT output distribution instead of using its entire shape in the evaluation of the final results, the issues caused by the small number of events in the BDT training are mitigated. The event yields of the recorded data and the simulated processes in each final category, including the subcategories provided by the 2D combination of BDT and MEM approach, are displayed in Fig. 9.8. The distributions of the final discriminants for recorded data and simulated events in each category are shown in Fig. 9.9, Fig. 9.10, and Fig. 9.11. The number of bins and the range of each histogram are chosen to ensure a minimum of two expected background events in each histogram bin. This measure guarantees a certain degree of robustness in the determination of the final results presented in Chapter 11.

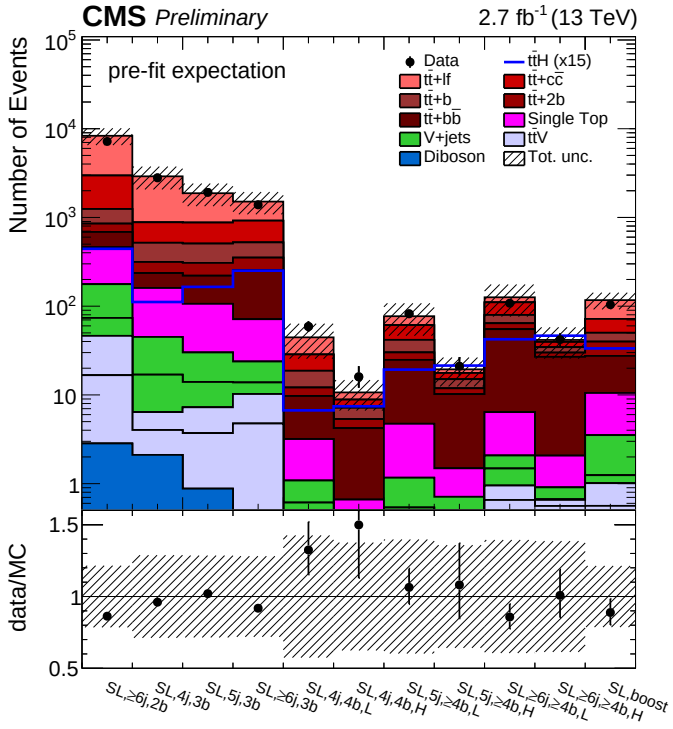


Figure 9.8: Event yields of recorded data and simulation in each analysis category, including the subcategories provided by the 2D approach. The background contributions are displayed as stacked filled histograms. The contribution by the $t\bar{t}H$ signal process is displayed as a blue line. The simulated processes are scaled to the predicted event yield expected for an integrated luminosity of $\mathcal{L} = 2.7\ \text{fb}^{-1}$. For better visibility, the $t\bar{t}H$ signal process is additionally scaled with a factor of 15. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each category. Taken from [204].

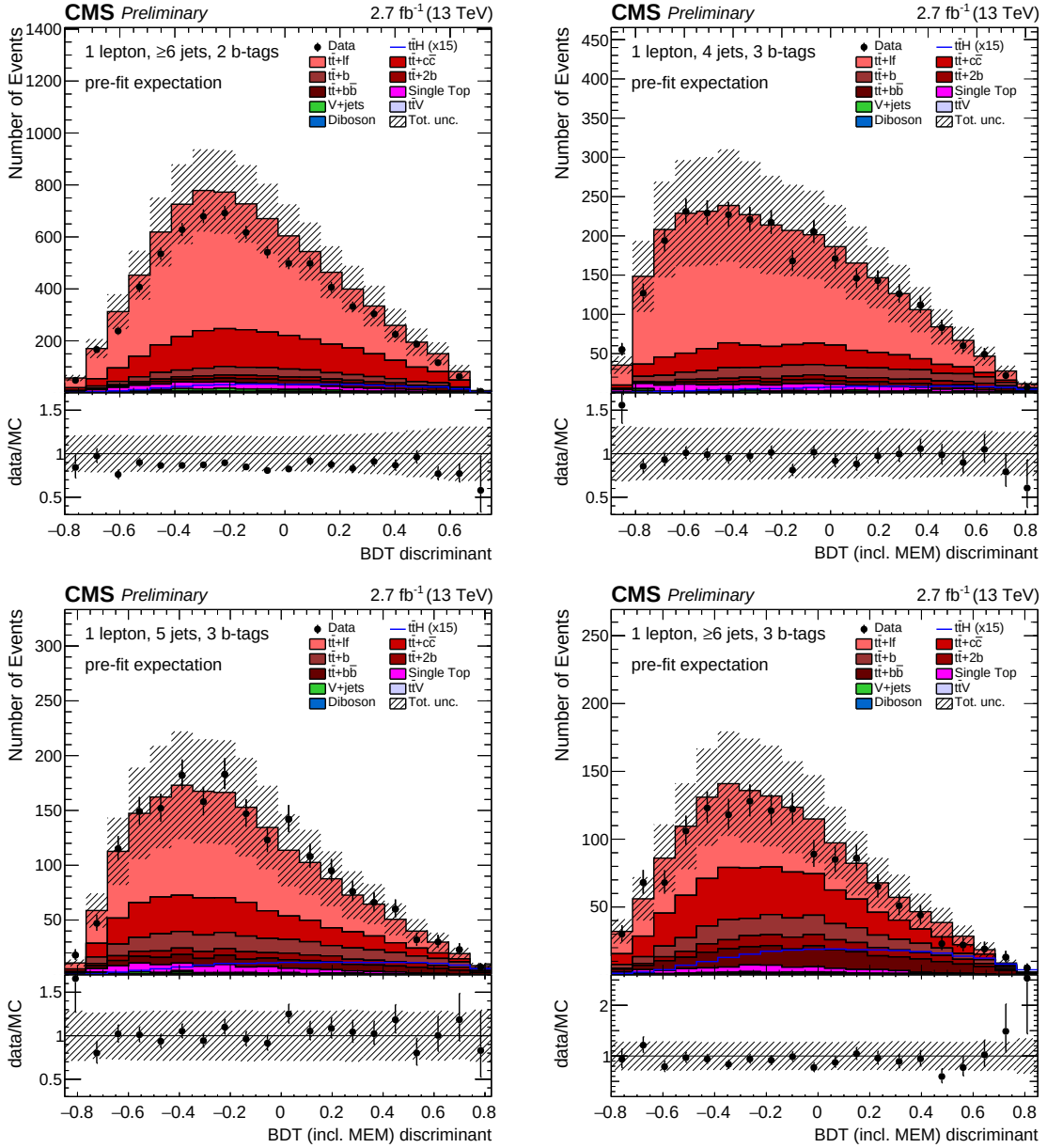


Figure 9.9: Final discriminant distributions of recorded data and simulation in each final category, including the subcategories provided by the 2D approach. The simulated processes are scaled to the event yields expected for an integrated luminosity of $\mathcal{L} = 2.7\ \mathrm{fb}^{-1}$. For better visibility, the signal process is scaled by an additional factor of 15. The background contributions are displayed as stacked filled histograms. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions. Taken from [7]. This figure is continued in Fig. 9.10.

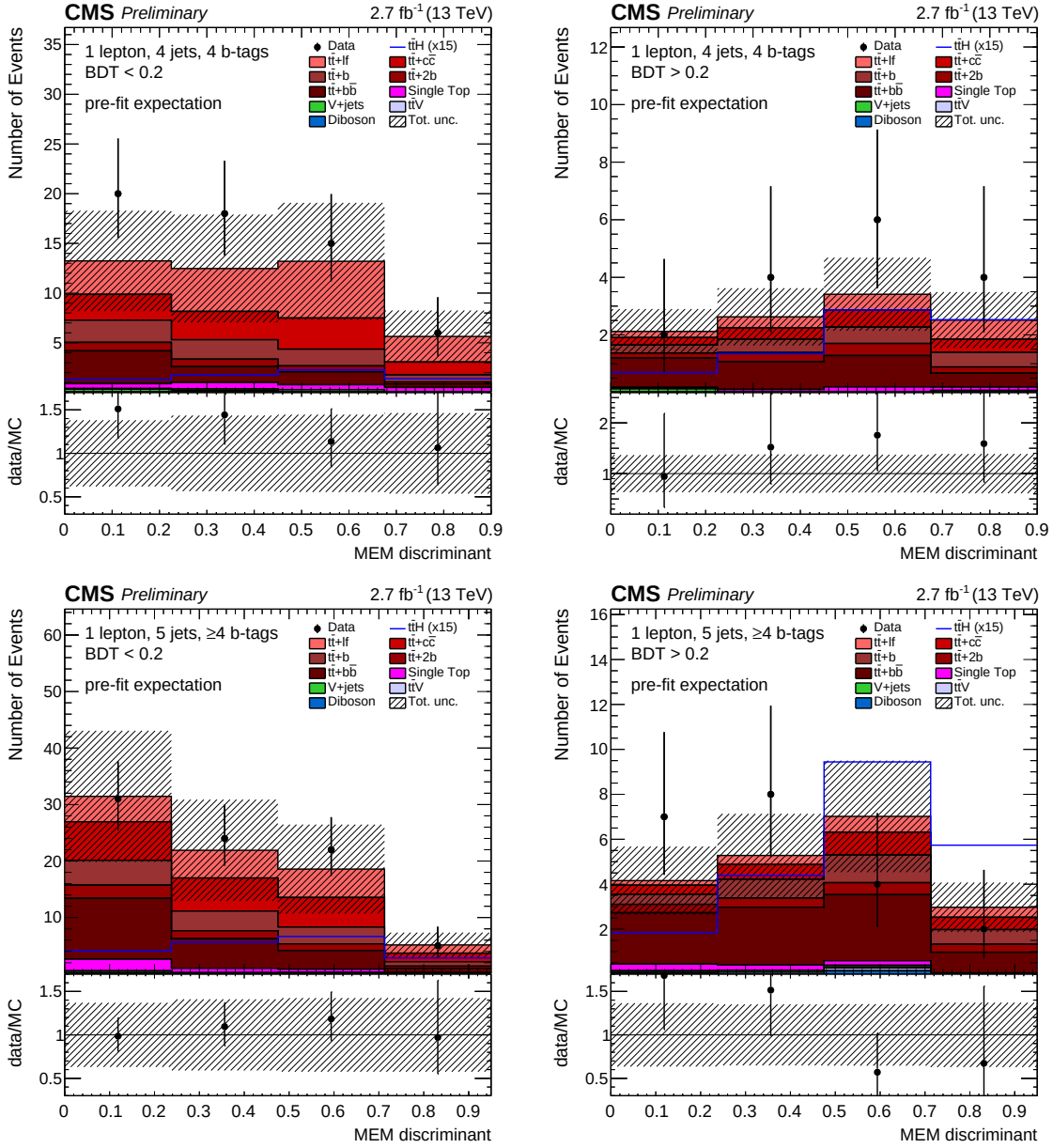


Figure 9.10: Final discriminant distributions of recorded data and simulation in each final category, including the subcategories provided by the 2D approach. The simulated processes are scaled to the event yields expected for an integrated luminosity of $\mathcal{L} = 2.7\ \mathrm{fb}^{-1}$. For better visibility, the signal process is scaled by an additional factor of 15. The background contributions are displayed as stacked filled histograms. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions. Taken from [7]. This figure is continued in Fig. 9.11.

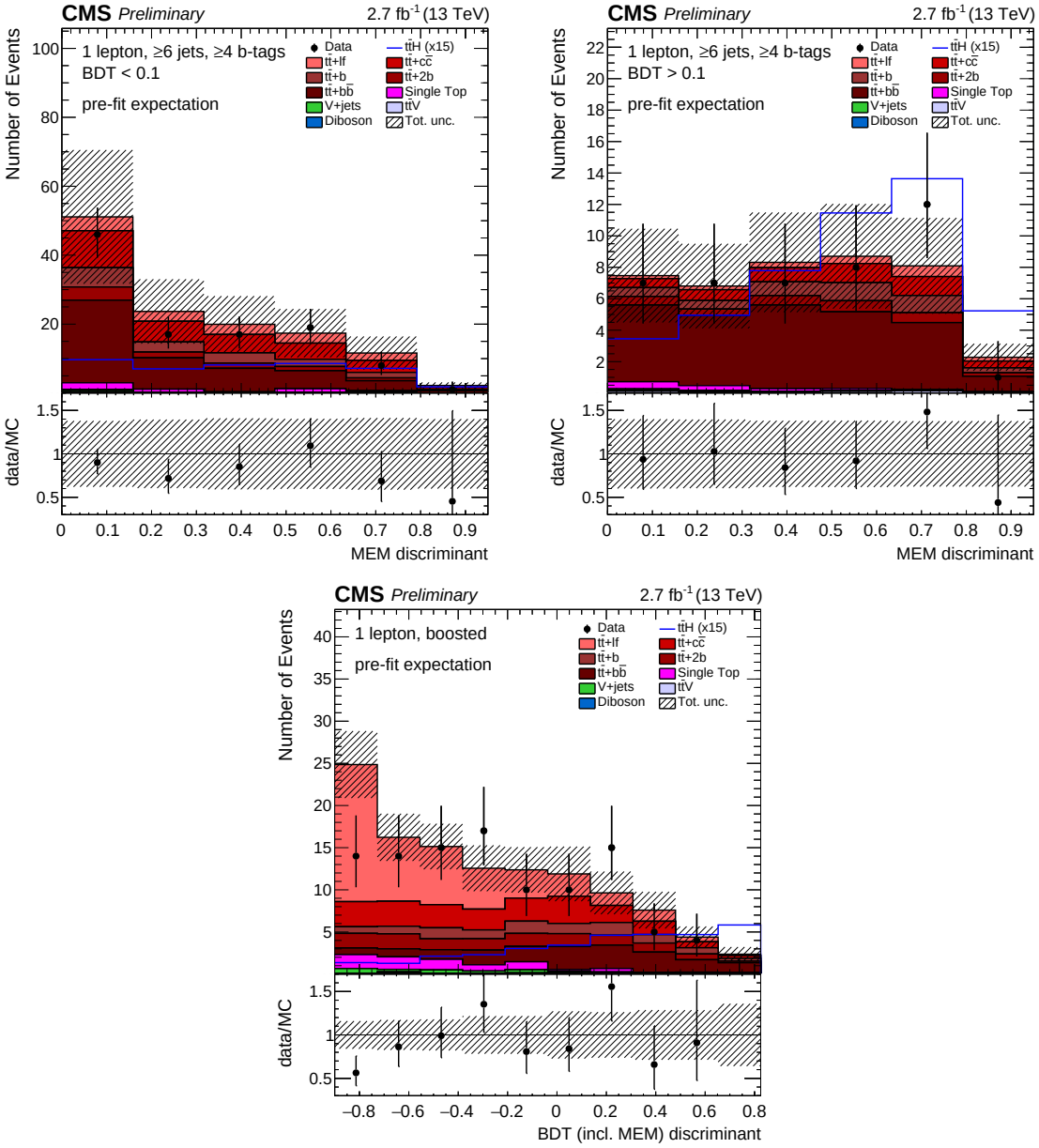


Figure 9.11: Final discriminant distributions of recorded data and simulation in each final category, including the subcategories provided by the 2D approach. The simulated processes are scaled to the event yields expected for an integrated luminosity of $\mathcal{L} = 2.7\ \mathrm{fb}^{-1}$. For better visibility, the signal process is scaled by an additional factor of 15. The background contributions are displayed as stacked filled histograms. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions. Taken from [7].

Chapter 10

Analysis Uncertainties

The analysis presented in this thesis is subject to various sources of uncertainties. Based on their effect on the processes considered, these uncertainties can be subdivided into three classes: uncertainties that affect the shape of final-discriminant distributions, uncertainties that affect the rates of processes, or uncertainties that affect both. In general, the uncertainties are evaluated by accordingly varying the rate, the shape, or both at the same time and propagating the outcome to the final results. This is achieved by introducing nuisance parameters in the limit setting procedure, as it is described in Section 5.2.3. Each uncertainty is evaluated separately in each category, but treated as fully correlated between them. In most cases, the uncertainties are also treated as fully correlated between the simulated signal and background processes. Exceptions from that case are explicitly pointed out in the subsequent sections, where the different sources of systematic uncertainties and their treatment are discussed in more detail.

The major part of uncertainties are due to systematic effects. The first uncertainty of this kind described in this chapter is the uncertainty on the measurement of the integrated luminosity corresponding to the recorded data. It is discussed in Section 10.1. Further sources of uncertainties, which are covered in Section 10.3, are associated to the reconstruction of jets and the corresponding calibrations. The calculation of cross sections and the simulation of proton-proton collision events depend on the choice of the PDFs and the renormalization and factorization scales. The uncertainties connected to these choices are discussed in Section 10.2. Differences due to the different behavior of simulated and recorded data in reconstruction, selection, and applied algorithms are accounted for by applying corrections. The systematic uncertainties coming with these corrections are described in Section 10.4. A source of statistical uncertainties is the limited number of simulated events available for this analysis. The treatment of these uncertainties is covered in Section 10.5. Finally, a summary of all systematic effects and their impact on the analysis is presented in Section 10.6.

10.1. Luminosity Uncertainty

The measurement of the luminosity corresponding to the recorded data is performed with the pixel detector of the CMS experiment and calibrated using Van-der-Meer scans. More details on the procedure can be found in Section 2.3. The main sources of systematic uncertainties on this measurement are the following:

- Hit inefficiencies in the pixel detector,
- Uncertainties on the separation of the beams,
- Approximations made in the calibration,

- Beam-beam effects during the VdM scans.

For the recorded data used in this analysis, the systematic uncertainties on the measured luminosity value caused by these effects add up to 2.7 % [106]. This uncertainty affects the predicted yields of all simulated processes. The shapes of the corresponding distributions, on the other hand, remain unchanged. For this reason, the systematic uncertainty on the measured integrated luminosity is treated as a rate uncertainty. It is evaluated by shifting the luminosity value up and down by one standard deviation.

10.2. Prediction Uncertainties

Theoretical predictions rely on the choice of three major ingredients: the PDFs determining the composition and momenta of the initial-state particles, the factorization scale determining the absorption of infrared physics into the PDFs, and the renormalization scale determining the absorption of ultraviolet physics into the strong coupling constant. The PDF sets come with uncertainties connected to their determination procedure. A major part of these uncertainties originate from the limited number of events in the data sample and from the method used for extracting the PDFs. A common choice of the renormalization and the factorization scale is the momentum transfer in the hard process. Nevertheless, the choice of the scales is arbitrary and has no effect on measurable quantities if all orders of perturbation theory are considered. The predictions used in this analysis consider only a very limited number of perturbation orders and therefore show a dependency on the choice of the scales. To account for the uncertainty in the choice of the renormalization and factorization scales, corresponding uncertainties are introduced based on the variation of the scales.

The choice of the renormalization and factorization scales and the uncertainties from the chosen PDF set affect the two types of predictions used in this analysis: cross sections and simulated proton-proton collision events. The cross sections are used to scale the rate of the processes to their expected event yields, whereas the simulated collision events determine shape of the variable distributions. The uncertainties on both types of predictions are evaluated separately and treated as uncorrelated. In Section 10.2.1, the uncertainties on the cross sections of the processes considered in this analysis are discussed. The uncertainties associated to the simulation of proton-proton collision events are covered in Section 10.2.2.

10.2.1. Cross-Section Uncertainties

The rates of the various processes considered in this analysis are predicted by inclusive cross sections of at least NLO accuracy. Since the calculated cross sections only determine the normalization of the processes in this analysis, the uncertainties on the calculated cross sections are treated as pure rate uncertainties. As mentioned above, the cross-section uncertainties originate from the choice of the renormalization and factorization scale and the uncertainties associated to the PDFs. Both types of uncertainties are treated separately and are denoted as “QCD scale” and “PDF” uncertainties in the following.

The uncertainties on the cross sections are calculated at the same time as the nominal values and are provided with the cross sections presented in Section 3.2.7. In this calculation procedure, the QCD scale uncertainties are determined by recalculating the cross sections with the standard choice of the renormalization and factorization scale multiplied

Table 10.1: Uncertainties on the cross sections of the processes considered in the analysis presented in this thesis. The uncertainties are grouped to uncertainties stemming from the PDFs and uncertainties stemming from the choice of the renormalization and factorization scale in the calculation. Each column represents an independent source of uncertainty. Uncertainties that can be found in the same column for two different processes are treated as correlated.

Process	PDF				QCD Scale				
	$gg_{t\bar{t}H}$	gg	$q\bar{q}$	qg	$t\bar{t}$	single top	V	VV	$t\bar{t}H$
$t\bar{t}H$	3.6 %								$-9.2\,\%/+5.8\,\%$
$t\bar{t}$+jets		3 %			$-4\,\%/+2\,\%$				
$t\bar{t}$+W			2 %		$-12\,\%/+13\,\%$				
$t\bar{t}$+Z		3 %			$-12\,\%/+10\,\%$				
single top				4 %		4 %			
V+jets			4 %				1 %		
diboson			2 %					2 %	

with a factor of 2 or 0.5 respectively. The procedure for evaluating the PDF uncertainties depends on the process. The PDF uncertainties on the cross sections of $t\bar{t}H$, $t\bar{t}$, and t-channel and s-channel single top-quark production have been determined based on the 2011 PDF4LHC recommendations [54]. This procedure includes the calculation of the cross section based on different PDF sets given by the MSTW2008 68 % CL NNLO set [49,50], the CT10 NNLO set [47,48], and the NNPDF2.3 5f FFN set [51]. The uncertainties are derived from the relative differences. The cross section of $t\bar{t}H$ production also considers photon-induced processes. For this reason, also the PDF set NNPDF2.3QED [218] is taken into account for the evaluation of the PDF uncertainties. For the remaining processes, the PDF uncertainties on the cross-section calculation are evaluated based on the uncertainty sets of the PDFs applied in the respective calculation. In case of $t\bar{t}Z$ production, $t\bar{t}W$ production, and the associated production of a single top quark, the error sets from the MSTW2008 NNLO PDF set corresponding to the 90 % confidence-level eigenvectors are used. For vector-boson+jets production and diboson production, the error sets provided with the CTEQ6M PDF sets [219] are applied. The uncertainties of different processes are correlated where appropriate. The values and correlations of the scale and PDF uncertainties are displayed in Table 10.1.

The associated production processes of a top-quark pair together with heavy-flavor quarks, $t\bar{t}$+$b\bar{b}$, $t\bar{t}$+2b, $t\bar{t}$+b, and $t\bar{t}$+$c\bar{c}$ production, are the main sources of irreducible background in this analysis. Nevertheless, by measuring the cross sections of these processes in control regions [141,220] or calculating them with perturbation theory [221], they can only be determined with about 50 % accuracy. For this reason, additional uncertainties of 50 % on the normalization of these processes are introduced. The treatment of these uncertainties as pure rate uncertainties has been found to be justified by a differential cross-section measurement of $t\bar{t}$+$b\bar{b}$ production [222]. The extra uncertainties on the normalization of the $t\bar{t}$+heavy-flavor processes are treated as uncorrelated with respect to each other and other uncertainties.

10.2.2. Event-Simulation Uncertainties

As the total rate of the processes are solely determined by the calculated inclusive cross sections, the event simulation only affects the shape of the distributions. Hence, corre-

sponding uncertainties are treated as pure shape uncertainties.

The uncertainties associated to the PDFs are studied using the 100 sub-PDFs provided additionally to the NNPDF3.0NLO PDF set [52], which is used for the simulation of the processes with major contributions in this analysis. The sub-PDFs are extracted in the same way as the standard PDF, but based on pseudo datasets generated from the measured data instead the measured data itself. Accordingly, the different sub-PDFs probe the statistical uncertainties and the uncertainty caused by the method for extracting the PDFs. For all 100 sub-PDFs, events are simulated, which are used to derive distributions of kinematic variables. Based on these distributions, one can observe that the variation in the PDFs mainly affect the rate of the processes, while hardly altering the shape. However, rate changes in event simulation are not relevant in this analysis, as the rate of a process is solely determined by its calculated cross section. Due to this fact, the PDF uncertainties on event simulation are not propagated to the final results.

In the procedure of event simulation, the calculation of the matrix element and the simulation of the parton shower depend on the renormalization and factorization scales. The choice of these scales can be made independently for the calculation of the matrix element and the simulation of the parton shower. Accordingly, the uncertainties on the renormalization and factorization scales in these distinct parts of simulation are treated separately and fully uncorrelated. The evaluation of the uncertainties on the scale choice in the matrix-element calculation and the parton-shower simulation are discussed in the following:

Matrix-element uncertainties: The uncertainty on the scale choice in the matrix-element calculation is evaluated by varying the renormalization and factorization scales simultaneously by factors of 0.5 or 2. The changes in the resulting matrix element are provided in form of event weights by the matrix-element generator. The uncertainties are propagated to the final results by reweighting the final-discriminant distributions using the event weights obtained with the shifted scales. The distributions describe the shape variations due to the scale uncertainties in the evaluation of the final results.

Parton-shower uncertainties: As for the evaluation of the matrix-element scale uncertainties, the parton-shower scale uncertainties are evaluated by shifting the renormalization and factorization scales simultaneously by factors of 0.5 or 2. In order to propagate the uncertainties, independent samples are generated using the shifted scales. For each sample, the event selection is applied and the final-discriminant distributions are determined. The latter are used to describe the shape variations caused by the scale uncertainties in the evaluation of the final results.

The renormalization and factorization-scale uncertainties in event simulation are only considered for the $t\bar{t}$+jets production as main background process. For other processes, these uncertainties are omitted, as their effect on the final result can be neglected. In the evaluation of the final results, independent nuisance parameters are introduced for every flavor sub-process of $t\bar{t}$+jets production: $t\bar{t}+b\bar{b}$, $t\bar{t}$+2b, $t\bar{t}$+b, $t\bar{t}+c\bar{c}$, and $t\bar{t}$+light-flavor production.

The uncertainty on the shapes of variables due to missing higher-order corrections are assumed to be small and are therefore neglected.

10.3. Reconstruction Uncertainties

Due to pile-up, undetected particles, non-uniformity in detector response, and different behavior for simulation and recorded data the reconstructed energies of jets are biased. In order to correct for these effects, the calibrations described in Section 4.7.3 and Section 7.4 are applied. The uncertainties on these corrections are propagated to the final result.

In Section 4.7.3 and Section 7.4, two different types of corrections are introduced calibrating the scale and the resolution of the reconstructed jet energies. The corresponding uncertainties are treated as fully uncorrelated. The uncertainties associated to the jet-energy scale are discussed in Section 10.3.1, whereas the uncertainties associated to the jet-energy resolution are covered in Section 10.3.2.

10.3.1. Jet-Energy Scale Uncertainties

The jet-energy scale corrections are factorized into scale factors correcting for different effects [167, 168]. These effects are the biases introduced by pile-up, the jet-energy response in simulation, and residual differences in data and simulation.

Pile-up offset: The uncertainties associated to the pile-up offset corrections can be subdivided into two main parts: the uncertainty on the pile-up offset scale factor used for the dependence on the pseudo rapidity in data and the dependence of the scale factor on the transverse momentum. The uncertainty on the scale factor used for pseudo-rapidity dependence in data is caused by the uncertainty on the measured average offset-energy density. The uncertainty on the dependence on the transverse momentum is based on the uncertainty due to the random-cone method.

Simulated jet-energy response: The determination of the corrections for the simulated jet-energy response relies solely on simulated QCD-dijet events. Accordingly, a major part of the uncertainties associated to the simulated jet-energy response is due to the modeling of the jet fragmentation and the simulation of the detector.

Residual jet-energy response: The residual jet-energy response corrections dealing with this difference are subdivided into two parts, the relative corrections accounting for the pseudo-rapidity dependence and the absolute corrections. Sources of uncertainties in the determination of the former corrections are represented by initial and final-state radiation, the jet-energy resolution, and the transverse-momentum dependence of the scale factors. For the latter correction, sources of uncertainties are the methods used for studying the jet-energy response and the fit applied for extracting the scale factors. Another major uncertainty is the fragmentation modeling. Additional sources affecting both parts of the residual corrections are the limited statistics of the samples used for the determination of the scale factors and the time dependence of measured data.

The uncertainties on the different scale factors are added in quadrature to form a single jet-energy scale uncertainty differential in transverse momentum and pseudo rapidity. However, this approach does not take into account correlations between the various sources of uncertainties of the different scale factors. A more sophisticated approach, which is recommended for future iterations of this analysis, includes a separate treatment of the different sources of uncertainties and with a proper accounting of the correlations between them.

Table 10.2: 95 % CL_s blinded expected upper limits on signal strength $\mu(t\bar{t}H)$ with and without considering the uncertainties on the jet-energy scale of subjets and fat jets. Limits are presented for the boosted analysis category.

boosted JES uncertainties	95 % CL_s expected upper limit on $\mu(t\bar{t}H)$ boosted category
considered	$10.1^{+5.3}_{-3.3}$
not considered	$10.0^{+5.3}_{-3.2}$

The jet-energy scale uncertainties are evaluated by shifting the applied jet-energy corrections by one standard deviation. The event selection and the calculation of the observables and the final discriminants are redone with the set of modified jets. The effect of the jet-energy scale uncertainties is propagated to the final results by introducing nuisance parameters based on the modified discriminant shape and event rate.

The uncertainties on the jet-energy scale are evaluated separately for the resolved jets, subjets, and fat jets, but are treated as fully correlated. As described in Section 4.7.3 and Section 7.4, resolved jets and subjets are calibrated using the jet-energy scale corrections derived for anti-k_T jets with a cone-size parameter of $R = 0.4$. For fat jets, the corrections derived for anti-k_T jets with a cone-size parameter of $R = 0.8$ are applied. The approach of fully correlating the uncertainties on the jet-energy scale is motivated by the applied jet-energy scale corrections being determined with the same method while using identical samples. However, the impact of the uncertainties on the jet-energy scale of the fat jets and the subjets on the final results has been studied. This study is based on calculating the blinded expected limits in the boosted analysis category with and without the uncertainties on the energy scale of the fat jets and the subjets. The results of this study are shown in Table 10.2. They show that the effect of the uncertainties on the jet-energy scale of subjets and fat jets are negligible. Accordingly, a fully correlated treatment of the jet-energy scale uncertainties of the different jet types is well justified.

10.3.2. Jet-Energy Resolution Uncertainties

The jet-energy resolution of simulated jets is calibrated by randomly scaling the momentum four-vectors of these jets based on given scale factors, which causes a smearing of the kinematic distributions [167, 223]. The scale factors used for this purpose are determined using the dijet-asymmetry method. The dijet-asymmetry method is based on determining the difference in transverse momentum of the two jets in dijet events. A major source of uncertainty in this procedure is the description of processes like initial and final state radiation, the underlying event, and out-of-cone showering by simulation. Further sources of uncertainties are the jet-energy scale corrections and the pile-up reweighting applied to simulation. The method for extracting the jet-energy resolution brings additional uncertainties.

The uncertainties are evaluated by shifting the scale factor used for the smearing of the jet-energy resolution by 1.5 standard deviations. The additional 50 % uncertainty have been added to account for changes in the software configurations since the scale factors have been determined in earlier stage of the LHC run II. The event selection is redone with the modified jets and the change in event yields of every process in every analysis category is determined. The jet-energy resolution uncertainties are propagated to the final results

by nuisance parameters based on the change in event yields. The final discriminants have not been re-evaluated, as the recalculation of MEM discriminant would be computationally intensive. The jet-energy resolution uncertainty has proven to have a very small effect on the final result.

The jet-energy-resolution uncertainty has only been evaluated for the resolved jets and subjets, as scale factors have only been determined for anti-k_{T} jets with a cone-size parameter of $R = 0.4$. Not considering the jet-energy resolution of the fat jets is defensible as the uncertainty on the jet-energy resolution shows an almost negligible impact on the final result.

10.4. Simulation-Correction Uncertainties

Various corrections are applied in this analysis to mitigate the different behavior of recorded and simulated data. The detailed procedures for deriving and applying them are presented in Chapter 4. Throughout their determination procedure, uncertainties that have to be accounted for in this analysis arise from various sources.

In the following, the uncertainties associated to these corrections and their propagation to the final results will be described. The uncertainties on the correction of the selection efficiencies of the lepton-identification requirements, the lepton-isolation requirements and the single-lepton triggers are described in Section 10.4.1. The handling of the uncertainties on the pile-up reweighting applied in order to adapt the simulated pile-up scenario to the one measured in data are discussed in Section 10.4.2. Section 10.4.3 covers the uncertainties on the scale factors correcting for the differences in the b-tagging output distributions of simulated and measured jets.

10.4.1. Lepton-Efficiency Uncertainties

The difference in the lepton-selection efficiency and the single-lepton trigger efficiency between data and simulation is corrected by reweighting simulated events using the three scale factors described in Section 4.10.2. These scale factors account for the different behavior of recorded and simulated events concerning the lepton-identification requirements, the lepton-isolation requirements, and the single-lepton triggers. As described in Section 4.10.2, they are derived differentially in bins of transverse momentum and pseudo rapidity by measuring the efficiencies with a tag-and-probe method in very pure samples of Z-boson decays.

The uncertainties on the scale factors are the statistical uncertainties due to the limited size of the samples used for the determination and systematic uncertainties by the tag-and-probe method. The main sources of systematic uncertainties are the tag definition, the signal and the background model, and the MC generator dependence. For muons, the uncertainties of all three scale factors add up to about 2 %. For electrons, scale factors derived with a software configuration and data from the early stages of LHC run II are used. For this reason, the combined uncertainties of all three corrections have been artificially increased to 4 %. The lepton-efficiency scale factors and the corresponding uncertainties used in this analysis are produced centrally by the E/Gamma and the Muon physics-object groups [173, 174].

The lepton-efficiency uncertainties are propagated to the final results by shifting the scale factors by one standard deviation. The uncertainties on the rate and the shape caused by the lepton-efficiency uncertainties are propagated via the altered final-discriminant

distributions. Due to the identical methods used for the determination of the scale factors and the identical sources of uncertainty, the scale-factor uncertainties of electrons and muons are treated as fully correlated. Additionally, the uncertainties caused by the lepton-identification and the lepton-isolation criteria are described by a single nuisance parameter and shifted simultaneously. The uncertainty on the single-lepton trigger scale factors are handled by a separate nuisance parameter and are thus treated as uncorrelated to the other lepton-efficiency uncertainties.

10.4.2. Pile-Up Correction Uncertainties

The pile-up scenario assumed for simulation does not match the one in measured data. In order to adapt the simulated scenario, a reweighting procedure is performed based on the number of simulated interactions and the number of interactions expected in data. A more detailed description of the procedure can be found in Section 4.10.1. The simulated number of interactions per bunch crossing is provided by simulation information, whereas the number of expected interactions per bunch crossing in measured data has to be calculated from the instantaneous luminosity and the total proton-proton inelastic cross section. The uncertainty on the calculated number of interactions is evaluated by varying the applied total proton-proton inelastic cross section by $\pm 4.6\,\%$ [171]. The final-discriminant distributions are reweighted with the event weights obtained with the modified cross section. The corresponding shape and rate uncertainties are propagated to the final results by introducing a nuisance parameter based on the altered distributions.

10.4.3. b-Tagging Scale-Factor Uncertainties

As described in Section 4.10.3, b-tagging scale factors are applied to correct for the differences observed in the distributions of the b-tagging outputs for simulated and measured jets. The scale factors are produced separately for heavy-flavor and light-flavor jets by applying a tag-and-probe method. The main uncertainties in the determination of the scale factors originate from the jet-energy scale, impurities by different-flavor jets, and the size of the applied samples. In the analysis, the uncertainties corresponding to the different sources are in most cases evaluated independently for heavy-flavor and light-flavor scale factors by varying the parameters associated to the uncertainties by one standard deviation.

The uncertainty on the scale factors stemming the jet-energy scale are determined by the same procedure outlined in Section 10.3.1. Hence, these uncertainties on the b-tagging scale factors are fully correlated to the overall jet-energy scale uncertainties in this analysis and are treated accordingly. The jet-energy scale uncertainties on the b-tagging scale factors of both flavors are shifted at the same time as the jet-energy scale scale factors. Consequently, no new nuisance parameter is introduced.

The uncertainty on the purity of the sample used for the tag-and-probe method originates from uncertainties in the modeling of the opposite jet flavor. In order to account for these uncertainties, the background contribution is shifted by $\pm 20\,\%$, which corresponds to a higher or lower degree of contamination. The uncertainties on the purity of the tag-and-probe samples are evaluated separately for both flavor contributions.

The third type of uncertainties considered are the statistical uncertainties due to the limited size of the data samples used for the determination of the scale factors. The statistical uncertainties are parametrized by two orthogonal contributions. These contributions are represented by a linear and a quadratic function, which vary the scale factors in a way

that they are still compatible with the statistical uncertainties. The linear term accounts for an overall tilt of the b-tagging output distributions. The quadratic term describes a variation of the upper and lower ends of the distribution with respect to the center. Both contributions are evaluated independently for each of jet-flavor.

As the differences in data and simulation for charm jets are not properly handled by either of the two scale factors and no suited control regions are available, charm jets are assigned a scale factor corresponding to unity. The uncertainties associated to this choice are accounted for by constructing uncertainties from the doubled values of the uncertainties of the heavy-flavor scale factors. The exact procedure starts by adding up all heavy-flavor uncertainties quadratically and doubling the result. In a subsequent step, a linear and a quadratic component is constructed based on the approach also used for the statistical uncertainties of the b-tagging scale factors.

Except for the jet-energy scale all of the uncertainties presented are treated separately. A summary of the eight b-tagging scale-factor uncertainties considered can be found in Table 10.3. In order to propagate these uncertainties to the final results, eight uncorrelated nuisance parameters are introduced, which are based on the modified final-discriminant distributions produced with the shifted scale factors.

10.5. Statistical Uncertainties

The limited number of simulated events available for modeling signal and background processes is a source of statistical uncertainties. The procedure for propagating these uncertainties to the final result is based on the approach described in [224, 225]. For every process and every histogram bin, a new nuisance parameter is introduced describing the upwards or downwards shift of the bin content according to the statistical uncertainty. In order to ensure a reasonable number of nuisance parameters in the fit, nuisance parameters that have a negligible effect on the likelihood are dropped.

10.6. Summary of Uncertainties

The various systematic uncertainties covered in the previous sections are propagated to the final results by 37 nuisance parameters. A summary of the systematic uncertainties together with their types and the processes that they are affecting are listed in Table 10.3. The nuisance parameters covering the effect of the statistical uncertainties are not listed.

The impact of each systematic uncertainty on the final result is studied based on the expected upper limit on the signal-strength modifier μ, which is determined as described in Section 5.2 and Chapter 11. For each nuisance parameter, its value is fixed to the best-fit value obtained by a maximum-likelihood fit, while all other nuisance parameters are allowed to float within the respective uncertainties. The expected upper limit on the signal-strength modifier is calculated for each fixed nuisance parameter and compared to the results of the standard configuration. A further study makes use of a similar approach, allowing the nuisance parameter under investigation to float within the uncertainties, while all other nuisance parameters are fixed to their best-fit value. The results of both studies are presented in Table 10.4. Fixing a nuisance parameter to a certain value, can be interpreted as considering this value as a perfect measurement without any uncertainties. In this study, the perfect-measurement value is given by the best-fit value returned by a maximum-likelihood fit. Unsurprisingly, the studies show that fixing the nuisance parameter associated the additional uncertainty on the cross section of the $\mathrm{t\bar{t}+b\bar{b}}$ process bears

Table 10.3: Systematic uncertainties considered in the analysis. All 37 nuisance parameters associated with the systematic uncertainties considered in this analysis are listed. Next to a short description, the type of the systematic uncertainty and the processes that are affected by the nuisance parameters are specified.

source	type	process	description
luminosity	rate	all	luminosity uncertainty
Jet-energy scale	shape	all	jet-energy scale uncertainty
Jet-energy resolution	shape	all	jet-energy resolution uncertainty
lepton efficiency	shape	all	lepton-identification efficiency unc.
trigger efficiency	shape	all	single-lepton trigger efficiency unc.
pile-up	shape	all	pile-up correction uncertainty
b-tag HF purity	shape	all	b-tag. SF purity unc. (heavy-flavor)
b-tag HF statistics (lin.)	shape	all	b-tag. SF linear stat. unc. (heavy-flavor)
b-tag HF statistics (quad.)	shape	all	b-tag. SF quad. stat. unc. (heavy-flavor)
b-tag LF purity	shape	all	b-tag. SF purity unc. (light-flavor)
b-tag LF statistics (lin.)	shape	all	b-tag. SF linear stat. unc. (light-flavor)
b-tag LF statistics (quad.)	shape	all	b-tag. SF quad. stat. unc. (light-flavor)
b-tag charm (lin.)	shape	all	b-tag. SF linear unc. (charm-flavor)
b-tag charm (quad.)	shape	all	b-tag. SF quad. unc. (charm-flavor)
QCD scale ($t\bar{t}H$)	rate	$t\bar{t}H$	scale uncertainty on NLO cross section
QCD scale ($t\bar{t}$)	rate	$t\bar{t}$	scale uncertainty on NLO cross section
QCD scale (single top)	rate	single top	scale uncertainty on NLO cross section
QCD scale (V)	rate	W,Z	scale uncertainty on NNLO cross section
QCD scale (VV)	rate	di-boson	scale uncertainty on NLO cross section
XS ($t\bar{t}$+$b\bar{b}$)	rate	$t\bar{t}$+$b\bar{b}$	extra uncertainty on NLO cross section
XS ($t\bar{t}$+2b)	rate	$t\bar{t}$+2b	extra uncertainty on NLO cross section
XS ($t\bar{t}$+b)	rate	$t\bar{t}$+b	extra uncertainty on NLO cross section
XS ($t\bar{t}$+$c\bar{c}$)	rate	$t\bar{t}$+$c\bar{c}$	extra uncertainty on NLO cross section
PDF (gg)	rate	$t\bar{t}$,$t\bar{t}$+Z	PDF uncertainty for gg initiated processes except $t\bar{t}H$
PDF (gg $t\bar{t}H$)	rate	$t\bar{t}H$	PDF uncertainty
PDF ($q\bar{q}$)	rate	$t\bar{t}$+W, W, Z	PDF uncertainty for $q\bar{q}$ initiated processes
PDF (qg)	rate	single top	PDF uncertainty of qg initiated processes
ME scale ($t\bar{t}$+$b\bar{b}$)	shape	$t\bar{t}$+$b\bar{b}$	scale uncertainties on matrix element
ME scale ($t\bar{t}$+2b)	shape	$t\bar{t}$+2b	scale uncertainties on matrix element
ME scale ($t\bar{t}$+b)	shape	$t\bar{t}$+b	scale uncertainties on matrix element
ME scale ($t\bar{t}$+$c\bar{c}$)	shape	$t\bar{t}$+$c\bar{c}$	scale uncertainties on matrix element
ME scale ($t\bar{t}$+lf)	shape	$t\bar{t}$+lf	scale uncertainties on matrix element
PS scale ($t\bar{t}$+$b\bar{b}$)	shape	$t\bar{t}$+$b\bar{b}$	scale uncertainties on parton shower
PS scale ($t\bar{t}$+2b)	shape	$t\bar{t}$+2b	scale uncertainties on parton shower
PS scale ($t\bar{t}$+b)	shape	$t\bar{t}$+b	scale uncertainties on parton shower
PS scale ($t\bar{t}$+$c\bar{c}$)	shape	$t\bar{t}$+$c\bar{c}$	scale uncertainties on parton shower
PS scale ($t\bar{t}$+lf)	shape	$t\bar{t}$+lf	scale uncertainties on parton shower

Table 10.4: Impact of individual systematic uncertainties on the final results. The values presented show the improvement of the expected upper limit on the signal-strength modifier μ with respect to the results of the standard measurement. Two cases either fixing the nuisance parameter under investigation or fixing all remaining nuisance parameters are tested. Fixed nuisance parameters are set to the best-fit value returned by a maximum-likelihood fit. Only systematic uncertainties yielding improvements of the order 1 % or larger when fixing the respective nuisance parameter are presented.

systematic uncertainty	improvement of 95 % CL_s expected upper limit on $\mu(t\bar{t}H)$ [%]	
	nuisance parameter frozen	only nuisance parameter floating
XS ($t\bar{t}+b\bar{b}$)	11.3	14.3
XS ($t\bar{t}+b$)	2.3	31.9
b-tag LF statistics (quad.)	1.5	52.3
QCD scale ($t\bar{t}H$)	1.5	51.5
b-tag HF statistics (quad.)	0.8	38.3
b-tag LF purity	0.8	37.6
PS scale ($t\bar{t}$+lf)	0.8	47.7
XS ($t\bar{t}$+2b)	0.8	37.6
XS ($t\bar{t}+c\bar{c}$)	0.8	39.1

the largest improvement in the expected upper limit on the signal strength. Due to the large uncertainty assigned to the rate, the $t\bar{t}+b\bar{b}$ process is only loosely constrained in the determination of the upper limits on the signal-strength modifier and the maximum-likelihood fit. As its shape is most similar to the one of the signal process, the $t\bar{t}+b\bar{b}$ process is able to mostly substitute the $t\bar{t}H$ contribution in the determination of the final results. Still, also the nuisance parameters associated to the cross-section uncertainties of the other $t\bar{t}$+heavy-flavor processes yield a large impact on the final results of this analysis. Further important nuisance parameters are associated to the uncertainties on the b-tagging output corrections, the signal prediction, and the $t\bar{t}$ parton-shower simulation. In summary, this study shows that the analysis can be largely improved by more accurate predictions of the various $t\bar{t}$+heavy-flavor processes. Furthermore, the reduction of the uncertainties associated to the correction of the b-tagging output would also bring a significant improvement of the analysis.

Chapter 11

Results

Based the foundation provided by the previous sections, the final results of the search for $t\bar{t}(H\to b\bar{b})$ production in the single-lepton channel can be determined. They are produced in two different forms, both expressed in terms of the signal-strength modifier $\mu_{t\bar{t}H}$ introduced in Section 5.2. It measures the deviation of the measured $t\bar{t}H$ cross section from the one predicted by the Standard Model. The first part of the results consist of the best-fit value for $\mu_{t\bar{t}H}$ derived by a maximum-likelihood fit. The likelihood function is constructed from the final-discriminant distributions for observed data and simulation taking into account the systematic uncertainties via nuisance parameters. In the procedure for maximizing the likelihood, $\mu_{t\bar{t}H}$ and the nuisance parameters are varied, in order to find optimal values. The second part of the results are the 95 % confidence-level upper limits on $\mu_{t\bar{t}H}$ calculated from the profile-likelihood ratio test statistic described in Section 5.2.3. Again, the likelihood functions applied in this procedure are constructed from the final-discriminant distributions and the nuisance parameters.

The final results for the search of $t\bar{t}H$ production with a Higgs boson decaying into a bottom-quark pair and a semileptonically decaying top-quark pair are presented in Section 11.1. The $t\bar{t}(H\to b\bar{b})$ search performed by the CMS collaboration [7] includes also analysis categories based on a dileptonic decay signature of the top-quark pair next to the ones presented in this thesis. The dilepton search channel and the results of the full $t\bar{t}(H\to b\bar{b})$ analysis are presented in Section 11.2. Further, $t\bar{t}H$ searches by the CMS collaboration using the first data provided by the CMS experiment in LHC run II have been performed based on multilepton and diphoton signatures. The combination of these searches with the the $t\bar{t}(H\to b\bar{b})$ search is presented in Section 11.3. In Section 11.4, the obtained results are compared with the results provided by the $t\bar{t}H$ searches in LHC run I and the $t\bar{t}H$ searches performed by the ATLAS collaboration in LHC run II.

11.1. Analysis Results

For the results of the single-lepton (SL) $t\bar{t}(H\to b\bar{b})$ search presented in this thesis, the likelihood functions are created using the final-discriminant distributions presented in Section 9.4 and the nuisance parameters introduced in Chapter 10. The maximum-likelihood fit for the determination of the best-fit value of the signal-strength modifier $\mu_{t\bar{t}H}$ is carried out simultaneously in all analysis categories by varying the signal-strength modifier $\mu_{t\bar{t}H}$ and the nuisance parameters at the same time in each category. The best-fit value for the signal-strength modifier obtained in this way takes on a small and negative value,

$$\mu_{\text{best-fit,SL,bb}}(t\bar{t}H) = -0.4^{+2.1}_{-2.1} \, .$$

Negative values of the signal-strength modifier are by definition unphysical. Nevertheless, such values are possible to result from downward fluctuations of the measured data.

Accordingly, the result shows no excess of $t\bar{t}H$ events in data. A reason for the negative best-fit value can be found in the signal-enriched bins of the final-discriminant distributions. An illustration of these distributions before the fit and after the fit can be found in Fig. 9.9, Fig. 9.10, and Fig. 9.11 and Appendix A.5 respectively. In most cases, the signal-enriched bins in these distributions bins feature small overall event yields and therefore are subject to large statistical fluctuations. The downwards fluctuations of data in such bins of the distributions of sensitive analysis categories, like the boosted analysis category, antagonize the observation of a positive signal strength. However, due to the comparably large uncertainties, the measured value is well in agreement with the expected Standard Model prediction of $\mu_{\mathrm{SM}}(t\bar{t}H) = 1$.

The fitted values of all nuisance parameters except the ones associated to the statistical uncertainties and their uncertainties are displayed in Fig. 11.1. In this plot, the black line represents the nuisance-parameter configuration before the fit, in which case all are set to their nominal values. The gray area shows the one sigma interval of the associated uncertainties. The colored markers with the corresponding error bars represent the nuisance parameter configuration after the maximum-likelihood fit has been performed together with the corresponding uncertainties. The red markers show the configuration after the fit of the signal+background hypothesis, whereas blue markers show the configuration after the fit of the background-only hypothesis. If the fit model and all associated uncertainties are correctly estimated, one expects the nuisance parameters to feature a value of zero and an uncertainty of one after the fit. Central values other than zero can indicate a bias in the fit model. Uncertainties different than one indicate that the corresponding systematic uncertainty is better constrained by the fit than in the determination of the systematic uncertainty in the first place. The nuisance parameters affecting the multiplicity of simulated and selected jets, like the ones associated the jet-energy scale uncertainty or the parton-shower scale uncertainties, represent one group of uncertainties with interesting post-fit features. For most of these nuisance parameters, the central values are shifted, the corresponding uncertainties are constrained, or both. An explanation for this effect are the differences in the jet multiplicity observed for data and simulation, which are balanced by the fit by shifting the corresponding nuisance parameters. A similar effect can be observed for the nuisance parameters affecting the b-tagging output, which balance the differences between data and simulation observed for the b-tagging multiplicity. The mentioned discrepancies in both distributions are displayeds in Section 6.3. A remarkable behavior can also be observed for the additional uncertainties assigned to the $t\bar{t}+b\bar{b}$ and $t\bar{t}+b$ contributions. The nuisance parameters associated to these uncertainties are shifted to values favoring a larger cross section. Concerning the $t\bar{t}+b\bar{b}$ contribution, a similar observation has been made by a separate measurement performed by the CMS collaboration, which measures the ratio of the $t\bar{t}+b\bar{b}$ cross section and the inclusive $t\bar{t}+$jets cross section [141]. Nuisance parameters affecting the overall event yield of simulated processes, like the ones associated to the luminosity or the lepton and trigger efficiencies, are shifted to smaller values. This possibly is a residual effect due to the variation of the other scale factors. The event yields of simulated and measured events in all categories do not obviously hint at a preference of a downscaling of the simulated events.

As no excess is observed, the results of this analysis are also interpreted in terms of 95 % $\mathrm{CL_s}$ upper limits on the signal-strength modifier $\mu_{t\bar{t}H}$. The observed and the expected upper limits on $\mu_{t\bar{t}H}$ are determined for the combination of all analysis categories as well as for each analysis category individually. The limits are extracted based on a profile-likelihood ratio test statistic. For the determination of the expected upper limits,

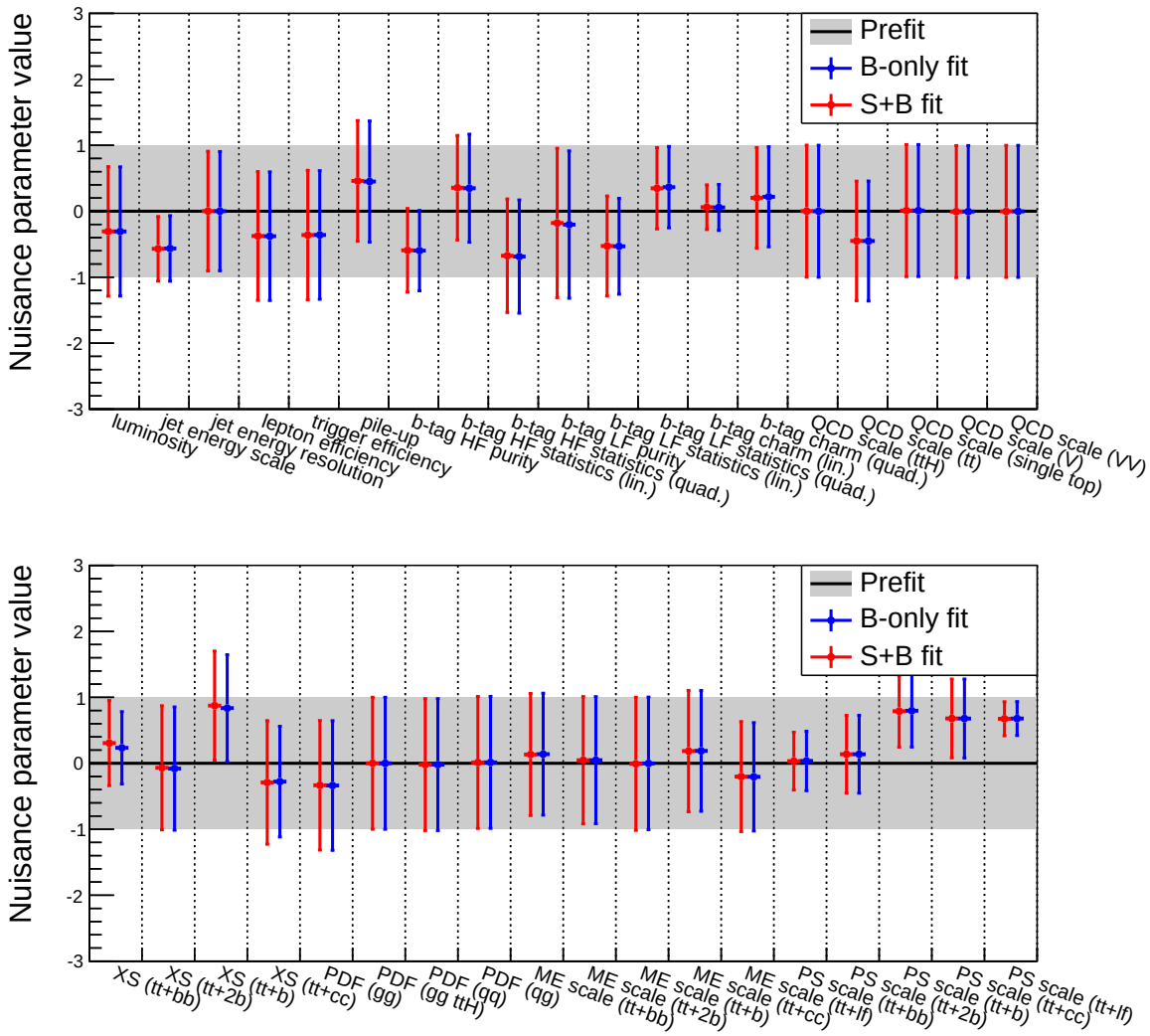


Figure 11.1: Pull distribution of the nuisance parameters before and after the maximum-likelihood fit of the signal+background (red) and the background-only (blue) hypothesis. The black line shows the configuration before the fit, where all nuisance parameters are fixed to their nominal value. The y-axis shows the relative deviation from the nominal value, where the gray area marks the one standard deviation interval.

Table 11.1: Observed and median expected 95 % CL_s upper limits on $\mu_{t\bar{t}H}$ in the $t\bar{t}(H\to b\bar{b})$ single-lepton channel. The expected limits are calculated with the asymptotic method. Further, the upper and the lower boundaries of the one standard deviation confidence interval of the median expected upper limits on $\mu_{t\bar{t}H}$ are stated.

Category	Observed upper limit	Expected upper limit
4 jets, 3 b-tags	14.5	$18.6^{+8.2}_{-5.5}$
4 jets, ≥4 b-tags high BDT output	35.7	$25.6^{+13.4}_{-8.1}$
4 jets, ≥4 b-tags low BDT output	86.6	$84.2^{+41.3}_{-25.8}$
5 jets, 3 b-tags	16.0	$12.3^{+5.5}_{-3.6}$
5 jets, ≥4 b-tags high BDT output	7.5	$10.3^{+5.6}_{-3.4}$
5 jets, ≥4 b-tags low BDT output	35.2	$31.9^{+16.1}_{-9.9}$
≥6 jets, 2 b-tags	25.4	$41.1^{+21.1}_{-13.1}$
≥6 jets, 3 b-tags	9.6	$7.6^{+3.3}_{-2.2}$
≥6 jets, ≥4 b-tags high BDT output	9.2	$8.3^{+4.4}_{-2.7}$
≥6 jets, ≥4 b-tags low BDT output	15.4	$18.3^{+9.6}_{-5.8}$
≥4 jets, ≥2 b-tags, boosted	7.5	$10.7^{+5.9}_{-3.5}$
Single-lepton combined	4.0	$4.1^{+1.8}_{-1.2}$

the asymptotic method is applied. The likelihood functions are constructed in a similar fashion as for the maximum-likelihood fit including only the relevant analysis categories. The results are presented in Table 11.1 and visualized in Fig. 11.2. The observed and the median expected limit of the combination of all analysis categories are very well in agreement. They exclude $t\bar{t}H$ cross sections larger than about four times the cross section predicted by the Standard Model at a 95 % confidence level. For the single categories, upwards and downwards shifts of the observed upper limits with respect to the median expected upper limits are observed. A cause for this behavior can be found in the signal enriched bins of the respective final-discriminant distributions. A downward fluctuation of the measured data in such bins strongly limits the signal and leads to more restrictive observed upper limits. An example of such a case is the final-discriminant distribution of the boosted-analysis category. This distribution does not feature a single observed event in the rightmost histogram bin, which shows the best signal over background ratio. This fact disfavors the appearance of $t\bar{t}H$ production and therefore leads to an observed limit which is smaller than the median expected limit. The boosted analysis category, which was newly introduced in the LHC run II analysis, yields comparably low exclusion limits. Accordingly, it is ranked among the most sensitive analysis categories in this analysis.

11.2. $t\bar{t}(H\to b\bar{b})$ Combination

Next to the single-lepton categories presented in this thesis, categories targeting a dileptonic decay of the top-quark pair have been included in the first $t\bar{t}(H\to b\bar{b})$ search performed by the CMS collaboration in LHC run II [7]. Due to the requirement of two

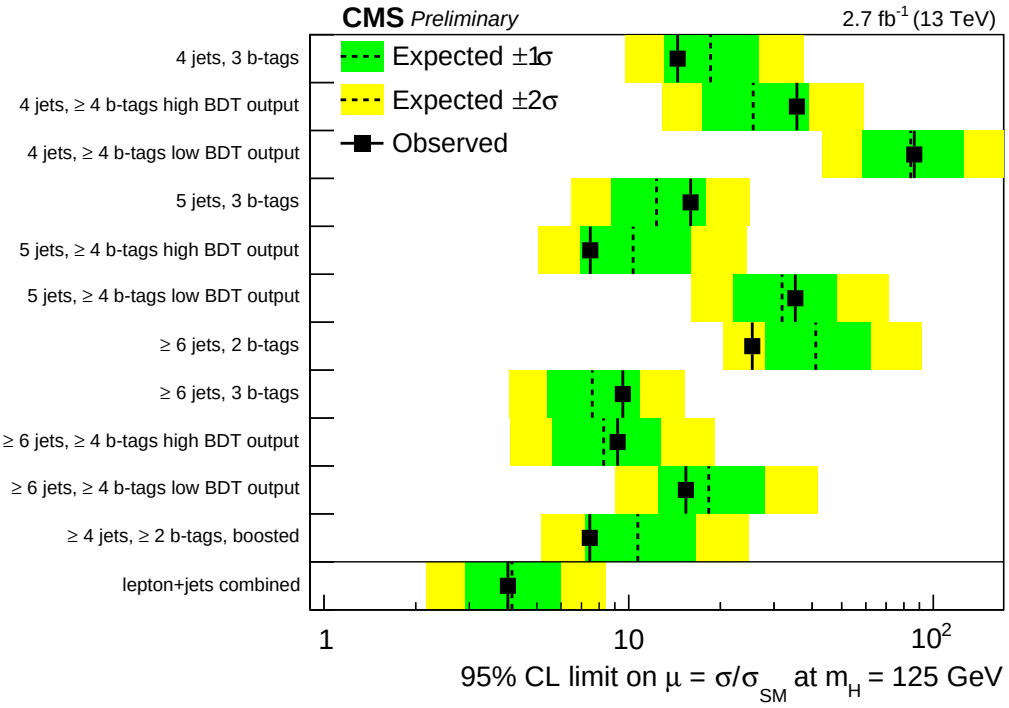


Figure 11.2: Observed and expected 95 % CL_s upper limits on $\mu_{t\bar{t}H}$ in the single-lepton channel. The observed limits are illustrated by the solid black marker and line. The expected limits are calculated with the asymptotic method and displayed by the median (black dashed line), the $\pm 1\sigma$ (green) and $\pm 2\sigma$ (yellow) confidence intervals. Taken from [7].

leptons, the dileptonic signature poses a very clean search channel. However, the branching fraction of this decay mode is smaller than the one of the semileptonic top-quark pair decay. Accordingly, the dilepton search channel features a small number of background events, but also fewer signal events than the single-lepton search channel. A more detailed description of the different $t\bar{t}$ decay modes is presented in Section 1.3.1. The analysis work flow is quite similar to the one presented in this thesis. For this reason, the analysis will be outlined only shortly, a more detailed description can be found in the published documentation of the first $t\bar{t}(H\to b\bar{b})$ search performed by the CMS collaboration in LHC run II [7].

The definitions of objects and the event selection in the dilepton search channel are very similar to the one of the single-lepton analysis described in Chapter 4 and Chapter 6. Events selected for the dilepton search channel have to pass triggers that require two isolated leptons with transverse momenta above $8-17$ GeV/c. The event selection requires exactly two oppositely charged leptons, which can be any combination of electrons and muons. The selection distinguishes between the harder leading lepton and the softer subleading lepton. The requirements on the subleading lepton are identical with the lepton definitions used for the veto of additional leptons in the single-lepton search channel. This choice ensures an orthogonal event selection with an unambiguous assignment of events to either of the two search channels. The leading lepton is required to have a transverse momentum of $p_T > 25$ GeV/c and a pseudo rapidity of $|\eta| < 2.4$. The threshold for the

isolation of the leading muon is loosened to 0.25. Further, the lepton pair has to feature an invariant mass larger than 20 GeV/c^2, in order to reject heavy-flavor resonances. Events with same-flavor leptons with an invariant mass in the Z-mass window 76 GeV/c^2 < m_{ll} < 106 GeV/c^2 are vetoed, in order to reduce the contribution by Z+jets events. Selected events are required to pass a cut on the missing transverse energy given by $\not\!\!E_T > 40$ GeV. Concerning the resolved jets, events passing the dilepton-event selection have to feature at least three jets with $p_{\mathrm{T}} > 20$ GeV/c with the leading two resolved jets featuring a transverse momentum of $p_{\mathrm{T}} > 30$ GeV/c. Similar to the single-lepton search channel, selected events are categorized according to the resolved jet and b-tag multiplicity. Based on the number of jets expected in the dilepton-event signature, the following five analysis categories are defined: the 3 jets, 2 b-tags analysis category, the 3 jets, 3 b-tags analysis category, the ≥4 jets, 2 b-tags analysis category, the ≥4 jets, 3 b-tags analysis category, and the ≥4 jets, ≥4 b-tags analysis category. In the dilepton search channel, no boosted analysis category is considered.

The final discriminants in the dilepton search channel are BDTs. These are optimized and trained following the same procedure described in Section 9.3. Yet, a different set of input variables adapted to the dilepton-event signature is chosen. The considered systematic uncertainties in the dilepton channels are almost identical to the ones presented in Chapter 10 except for the uncertainties on the lepton identification and trigger efficiencies. A maximum-likelihood fit in the combined dilepton analysis categories results in a negative best-fit value of the signal-strength modifier,

$$\mu_{\text{best-fit,bb}}(\mathrm{t\bar{t}H}) = -4.7^{+3.7}_{-3.8} \,.$$

Despite the large unphysical value, the deviations from the Standard Model prediction are not significant. As this analysis is not sensitive to the Standard Model expectation yet, the 95 % $\mathrm{CL_s}$ upper limit on the signal strength are more meaningful. The exclusion limits are determined for the combination of all dilepton analysis categories as well as for every dilepton analysis category individually. The expected upper limits are extracted using the asymptotic method. The results are shown in Table 11.2 and are additionally illustrated in Fig. 11.3. The observed upper limit on the signal-strength modifier for the combination of all dilepton analysis categories excludes a $\mathrm{t\bar{t}H}$ cross section 5.2 times larger than the one predicted by the Standard Model at a 95 % confidence level. However, the median expected limit predicts an exclusion limit, which is weaker than the one observed. The same effect can be observed for the individual most sensitive analysis categories, the ≥4 jets, 3 b-tags analysis category and the ≥4 jets, ≥4 b-tags analysis category, which explains the observed behavior in the combined case. Again, a possible cause for this behavior are statistical downward fluctuations of data in the signal-enriched bins of the final-discriminant distributions. This effect is even more pronounced in the dilepton channel, due to the small number of selected events. However, no indication for a systematic mismodelling has been found.

The 95 % $\mathrm{CL_s}$ upper limits on the signal-strength modifier $\mu_{\mathrm{t\bar{t}H}}$ obtained in the dilepton search channel are slightly weaker than the ones obtained in the single-lepton search channel. Nevertheless, the dilepton search channel adds additional sensitivity if combined with with the single-lepton search channel. The combination is based on the construction of the likelihood from the final-discriminant distributions of all single-lepton and dilepton analysis categories. The best-fit value of $\mu_{\mathrm{t\bar{t}H}}$ and the 95 % $\mathrm{CL_s}$ upper limits on $\mu_{\mathrm{t\bar{t}H}}$ derived using the combined likelihood function are stated in Table 11.3 and additionally visualized

Table 11.2: Observed and median expected 95 % CL_s upper limits on the signal-strength modifier $\mu_{t\bar{t}H}$ in the $t\bar{t}(H\to b\bar{b})$ dilepton channel. Expected limits are calculated with the asymptotic method. The upper and the lower boundaries of the one standard-deviation confidence interval of the median expected upper limits on $\mu_{t\bar{t}H}$ are also specified. Taken from [7].

Category	Observed upper Limit	Expected upper limit
3 jets, 2 b-tags	186.0	$114.8^{+52.6}_{-34.1}$
3 jets, 3 b-tags	104.9	$48.6^{+26.2}_{-15.9}$
≥4 jets, 2 b-tags	32.4	$40.1^{+16.8}_{-11.3}$
≥4 jets, 3 b-tags	7.4	$10.8^{+5.2}_{-3.3}$
≥4 jets, ≥4 b-tags	9.1	$12.2^{+7.5}_{-4.3}$
Dilepton combined	5.2	$7.7^{+3.6}_{-2.3}$

Table 11.3: Best-fit value of the signal-strength modifier $\mu_{t\bar{t}H}$ and expected and observed 95 % CL_s upper limits on the signal-strength modifier $\mu_{t\bar{t}H}$. The results are displayed for the single-lepton channel, dilepton channel, and the combination of both. Expected limits are calculated with the asymptotic method. Further, the one standard deviation ($\pm1\sigma$) intervals of the best-fit value and the expected limit are stated. Taken from [7].

Channel	Best-fit $\mu_{t\bar{t}H}$	Observed upper limit	Expected upper limit
Single lepton	$-0.4^{+2.1}_{-2.1}$	4.0	$4.1^{+1.8}_{-1.2}$
Dilepton	$-4.7^{+3.7}_{-3.8}$	5.2	$7.7^{+3.6}_{-2.3}$
Combination	$-2.0^{+1.8}_{-1.8}$	2.6	$3.6^{+1.6}_{-1.1}$

in Fig. 11.4. Again, the expected upper limits are calculated using the asymptotic method. In the combined fit, the signal-strength modifier is pulled to an intermediate solution between the best-fit values of the uncombined search channels. Even though the best-fit signal-strength obtained is negative and therefore unphysical, the large uncertainties still allow compatibility with the standard-model prediction. The upper limits on the signal-strength modifier are further reduced compared to the individual search channels and now exclude $t\bar{t}H$ cross sections larger than 2.6 times the cross section predicted by the Standard Model. A comparison of the results with the ones of other $t\bar{t}H$ searches is presented in Section 11.4.

11.3. $t\bar{t}H$ Combination

The $t\bar{t}H$ search based on a Higgs-boson decay into a bottom-quark pair presented in this thesis is not the only $t\bar{t}H$ search performed based on the first LHC-run-II data by the CMS collaboration. Further analyses have been conducted in different $t\bar{t}H$ search channels targeting different Higgs-boson decay modes. One of these analyses is the search for $t\bar{t}H$ production in the multilepton final states at a center-of-mass energy of $\sqrt{s} = 13$ TeV [226]. This analysis targets Higgs-boson decays into two Z bosons, two W bosons, or two tau leptons. $t\bar{t}H$ events with these decays include the rare signatures featuring charged leptons

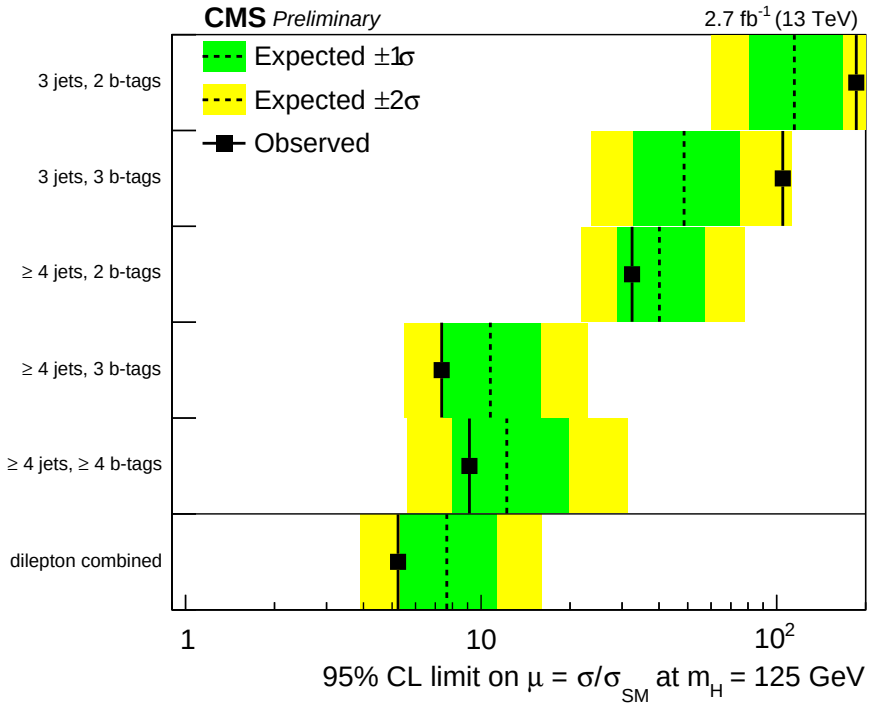


Figure 11.3: Observed and expected 95 % CL_s upper limits on $\mu_{t\bar{t}H}$ in the dilepton $t\bar{t}(H\to b\bar{b})$ search channel. The observed limits are illustrated by the solid black marker and line. The expected limits are calculated with the asymptotic method and displayed as the median (black dashed line), the $\pm1\sigma$ (green), and $\pm2\sigma$ (yellow) confidence intervals. Taken from [7].

with the same electric charge or charged-lepton multiplicities larger than two. Based on these features, two analysis categories are constructed requiring two leptons with identical electric charge or at least three electrically charged leptons. By additionally requiring two jets, there are hardly any background processes that feature a signature that is compatible with the event selection requirements. Accordingly, contributions by background processes are very small and mainly given by events with fake lepton candidates or lepton with incorrectly measured electric charges. The multilepton $t\bar{t}H$ analysis is performed on recorded data corresponding to an integrated luminosity of $\mathcal{L} = 2.3\,\text{fb}^{-1}$. The second analysis performed based on the first LHC-run-II data targets Higgs-boson decays into two photons [227]. The Higgs-boson decay into two photons has a very small cross section compared to the other Higgs-boson decay modes. Nevertheless, due to the two photons with large transverse momenta this search channel features a very clean signature, a good Higgs-boson mass resolution, and good control of the residual backgrounds. The main background processes are prompt diphoton production and events with jets that are misidentified as photons. The diphoton Higgs-boson search at a center-of-mass energy of $\sqrt{s} = 13$ TeV covers several Higgs-boson production channels, like gluon-gluon fusion and vector-boson fusion, in addition to the $t\bar{t}H$ production. The different production modes are analyzed in dedicated analysis categories. For the $t\bar{t}H$ production, two categories are introduced targeting top-quark pair decays with leptons and the full-hadronic top-quark

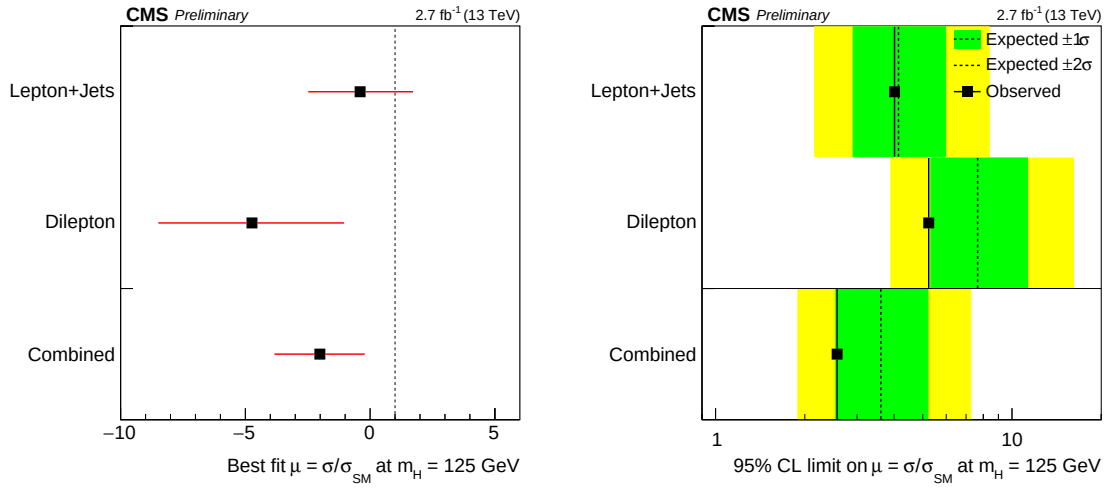


Figure 11.4: Best-fit value of the signal-strength modifier $\mu_{t\bar{t}H}$ (left) and observed and expected 95 % CL_s upper limits on $\mu_{t\bar{t}H}$ (right) in the single-lepton channel, the dilepton channel, and the combination of both. The observed limits are illustrated by the solid black marker and line. The expected limits are calculated with the asymptotic method and displayed as the median (black dashed line), the $\pm 1\sigma$ (green), and $\pm 2\sigma$ (yellow) confidence intervals. Taken from [7].

pair decay. Depending on the category, additional requirements on the number of selected jets and leptons are made. The diphoton Higgs-boson search is performed based on recorded data corresponding to an integrated luminosity of $\mathcal{L} = 2.7$ fb^{-1}.

A common result is provided by the combination of all $t\bar{t}H$ searches based on the first LHC-run-II data [228]. As for the combination of analysis categories in the $t\bar{t}(H\to b\bar{b})$ analysis, a common likelihood function based on the final discriminants of all analysis categories and the nuisance parameters of all analyses is constructed. The systematic uncertainties are correlated between different search channels as appropriate. For example, the theoretical uncertainties, which are the uncertainties on the renormalization and factorization scales and the PDFs used for the predictions, are treated as fully correlated between all $t\bar{t}H$ search channels. The experimental uncertainties on the measured luminosity and the jet energy are treated in the same way. As only the $t\bar{t}(H\to b\bar{b})$ and the multi-lepton search make use of the reweighting procedure of the b-tagging output described in Section 4.10.3, the uncertainties associated to b-tagging are only correlated between these two search channels. The branching fraction between different Higgs-boson decay modes and the small contributions by other Higgs-boson production channels are set to the standard-model expectations, but are allowed to vary with respect to their experimental and theoretical uncertainties. The best-fit value of the $t\bar{t}H$ signal-strength modifier resulting from the maximum-likelihood fit of the common likelihood function,

$$\mu_{\text{best-fit}}(t\bar{t}H) = 0.15^{+0.95}_{-0.81}\,,$$

shows no excess of signal-like events. Nevertheless, the best-fit value is still in good agreement with the standard-model expectation of $\mu(t\bar{t}H) = 1$. The negative logarithms of the likelihood values for observation and expectation are displayed in Fig. 11.5 as a function of $\mu(t\bar{t}H)$.

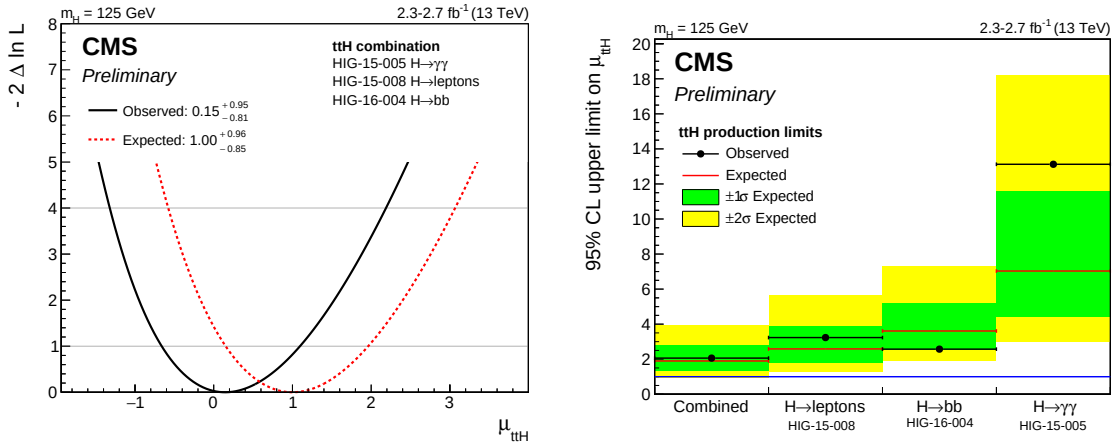


Figure 11.5: Negative logarithm of the likelihood value as function of $\mu_{t\bar{t}H}$ for the combination of all $t\bar{t}H$ searches (left) and 95 % CL_s upper limits on $\mu_{t\bar{t}H}$ for multi-lepton $t\bar{t}H$ analysis, the $t\bar{t}(H\to b\bar{b})$ analysis, the diphoton analysis, and the combination of all three (right). The plots each show the observed (black) and the expected values (red). The expected limits are calculated with the asymptotic method and displayed as the median (red), the $\pm1\sigma$ (green), and $\pm2\sigma$ (yellow) confidence intervals. Taken from [228].

As the combination has not enough sensitivity to observe Standard Model $t\bar{t}H$ production yet, the 95 % CL_s upper limits on the signal-strength modifier $\mu_{t\bar{t}H}$ are calculated. Based on the combined likelihood function, the observed(expected) upper limits on the $t\bar{t}H$ signal-strength modifier has been calculated as $\mu_{t\bar{t}H} < 2.1(1.9)$. The expected limits are determined using the asymptotic method. The observed and expected limits derived for the individual search channels and the combination of all search channels are illustrated in Fig. 11.5. The differences between observed and expected limits found in the single search channels are averaged in the combination.

After finishing the analyses presented in this section, a further iteration of the multilepton $t\bar{t}H$ search [229] and the diphoton $t\bar{t}H$ search [230] have been performed by the CMS collaboration based on the first part of the 2016 dataset corresponding to an integrated luminosity of $\mathcal{L} = 12.9\,\text{fb}^{-1}$. The multilepton $t\bar{t}H$ search has been combined with the one performed on the 2015 CMS dataset corresponding to an integrated luminosity of $\mathcal{L} = 2.3\,\text{fb}^{-1}$. In this combination, a best-fit value of the signal-strength modifier,

$$\mu_{\text{best-fit,ML}}(t\bar{t}H) = 2.0^{+0.8}_{-0.7}\,,$$

has been obtained. This value shows a slight excess in data, but is still compatible with the SM prediction within the given uncertainties. Further, upper limits on the signal-strength modifier $\mu_{t\bar{t}H}$ have been determined as

$$95\,\%\ CL_s \text{ observed (expected) upper limit on } \mu(t\bar{t}H)\text{: } 3.4\,(1.3^{+0.6}_{-0.4})\,.$$

The observed upper limit on the signal-strength modifier being weaker than the one expected is compatible with the slight excess observed for the best-fit value of $\mu_{t\bar{t}H}$.

Also a further iteration of the diphoton $t\bar{t}H$ search has been performed together with other Higgs-boson production modes. Similar to the analysis performed based on the 2015 CMS datatsets, this diphoton $t\bar{t}H$ search is represented by two dedicated analysis categories. The best-fit value of the signal-strength modifier $\mu_{t\bar{t}H}$ obtained in these categories is

$$\mu_{\text{best-fit},\gamma\gamma}(t\bar{t}H) = 1.91^{+1.5}_{-1.2} \, .$$

As for the multilepton search a slight excess is found, which is however compatible with the SM prediction within the uncertainties.

11.4. Comparison of Results

Next to the analyses presented in the previous section, further $t\bar{t}H$ searches have been performed in LHC run I and by the ATLAS collaboration. A comparison of the results provided by these analyses with the ones presented in the previous section provides a context for the interpretation of the obtained results, an important cross-check, and a measure of improvement. In Section 11.4.1, the $t\bar{t}H$ searches based on the first data of LHC run II including the one discussed in this thesis will be compared to the $t\bar{t}H$ searches performed by the CMS collaboration based on the full LHC-run-I dataset recorded at a center-of-mass energy of $\sqrt{s} = 8$ TeV. Further, in Section 11.4.2, the LHC-run-II $t\bar{t}H$ searches are compared to the analyses performed by the ATLAS collaboration based on the first LHC-run-II data.

11.4.1. 8 TeV $t\bar{t}H$ Search at CMS

The search for $t\bar{t}H$ production at a center-of-mass energy of $\sqrt{s} = 8$ TeV carried out by the CMS collaboration [78] is subdivided into a set of different analyses. Similar to the ones performed in the second data-taking run of the LHC, these analyses are based on search channels targeting particular Higgs-boson decay modes: the diphoton $t\bar{t}H$ search, the $t\bar{t}(H\to b\bar{b})$ search, the $t\bar{t}H$ search targeting hadronically decaying tau leptons, and the multilepton $t\bar{t}H$ search. The different analyses are mostly based on the full LHC-run-I dataset recorded by the CMS experiment at a center-of-mass energy of $\sqrt{s} = 8$ TeV, which corresponds to an integrated luminosity of $\mathcal{L} = 19.3 - 19.7$ fb^{-1}.

8 TeV Single-lepton $t\bar{t}(H\to b\bar{b})$ Search

The single-lepton part of the $t\bar{t}(H\to b\bar{b})$ analysis at $\sqrt{s} = 8$ TeV [77,78] follows an approach that is very similar to the one presented in this thesis. After an event selection based on the semileptonic decay signature of $t\bar{t}(H\to b\bar{b})$ events, selected events are categorized based on their resolved jet and b-tag multiplicity. The analysis categories are identical to the ones described in Section 6.1, except for the boosted analysis category, which is not considered. The final discriminants are provided by BDTs, which are individually trained in each category. The MEM approach is not included in this analysis, instead is was applied as final discriminator in an independent $t\bar{t}(H\to b\bar{b})$ search [76]. In this search, events are categorized based on their resolved jet and b-tag multiplicity, the invariant masses of dijet pairs, and the b-tagging likelihood ratio, which is described in Section 9.1. The MEM discriminator is applied as final discriminator in each category. The final results of both analyses are evaluated in the same way as for the analysis described in this thesis.

The results of the LHC-run-I $t\bar{t}(H\to b\bar{b})$ search in the single-lepton search channel based on the BDT approach are the best-fit value for the signal-strength modifier obtained by the maximum-likelihood fit,

$$\mu_{\text{best-fit,SL,bb}}(t\bar{t}H) = 0.7^{+1.9}_{-1.9} ,$$

and the 95 % CL_s exclusion limits on the signal-strength modifier,

$$95\,\%\ CL_s \text{ observed (expected) upper limit on } \mu(t\bar{t}H)\text{: } 4.1\ (3.5^{+1.5}_{-1.0}) .$$

The LHC-run-I $t\bar{t}(H\to b\bar{b})$ search in the single-lepton search channel using the MEM approach obtains a best-fit value for the signal-strength modifier of

$$\mu_{\text{best-fit,SL,bb}}(t\bar{t}H) = 1.2^{+1.6}_{-1.5} .$$

The observed and expected upper limits on the signal-strength modifier $\mu(t\bar{t}H$ derived by this analysis are

$$95\,\%\ CL_s \text{ observed (expected) upper limit on } \mu(t\bar{t}H)\text{: } 4.2\ (3.3^{+1.6}_{-1.0}) .$$

Compared to these values, the 13 TeV $t\bar{t}(H\to b\bar{b})$ search in the single-lepton channel presented in this thesis bears very similar results. The best-fit values of the signal-strength modifier are positive and closer to the Standard Model expectation of $\mu(t\bar{t}H) = 1$. With respect to the large uncertainties, both results are well in agreement. Concerning the exclusion limits, the LHC-run-I analyses yield a slightly lower expected upper limit on the signal strength than the LHC-run-II analysis. Hence, the former seem to be more sensitive. The observed upper limits, on the other hand, are nearly identical. Considering the number of expected $t\bar{t}H$ events in each of the recorded data sets, the results of the LHC-run-II search are remarkable. The cross section of $t\bar{t}H$ production predicted by the Standard Model for a center-of-mass energy of $\sqrt{s} = 13$ TeV corresponds to 3.9 times the one predicted for $\sqrt{s} = 8$ TeV. Additionally taking into account the integrated luminosities corresponding to the recorded datasets, one finds that about half as many $t\bar{t}H$ events are expected in the LHC-run-II dataset than in the LHC-run-I dataset. Still, the analyses from both LHC data-taking periods reach a comparable sensitivity. One part of this effect is caused by the increase of the cross section of the inclusive top-quark pair production from a center-of-mass energy of $\sqrt{s} = 8$ TeV to a center-of-mass energy of $\sqrt{s} = 13$ TeV. This increase is smaller than the one for $t\bar{t}H$ production. Accordingly, a smaller fraction of background events with respect to signal is expected. The improvement of the analysis performance also takes a great part in the increase of sensitivity. Main factors in that sense are the b-tagging algorithms, which are more efficient than the ones used throughout LHC run I, the combination of the MEM and BDT approach in final discrimination, and the introduction of the boosted analysis category.

8 TeV $t\bar{t}H$ Combination

All the CMS $t\bar{t}H$ searches performed in LHC run I are combined in a fashion similar to the LHC-run-II $t\bar{t}H$ searches. A common likelihood is constructed based on the final

discriminants and the nuisance parameters of all analyses. Systematic uncertainties are correlated as appropriate. The branching ratios of the different Higgs-boson decay modes and the small contributions by other Higgs-boson production channels are constrained to the SM expectations based on their experimental and theoretical uncertainties. Compared to the $t\bar{t}H$ combination performed for the analyses based on the first LHC-run-II data, the $\sqrt{s} = 8$ TeV $t\bar{t}H$ combination features an additional search channel targeting the Higgs-boson decay into a pair of hadronically decaying tau leptons. Further, the diphoton search channel has been performed in an analysis independent from the other Higgs-boson production modes and additionally includes the dataset recorded at a center-of-mass energy of $\sqrt{s} = 7$ TeV. The results are evaluated using a Higgs-boson mass of $m_{\mathrm{H}} = 125.6\ \mathrm{GeV}/c^2$, which represented the most precise measurement of the Higgs-boson mass by the CMS collaboration at that time. The best-fit value of the signal-strength modifier derived with the combined maximum-likelihood fit is

$$\mu_{\text{best-fit}}(t\bar{t}H) = 2.8^{+1.0}_{-0.9}\ .$$

The best-fit signal strength obtained exceeds the SM expectation $\mu(t\bar{t}H) = 1$, but is still compatible. The p-value for the background-only hypothesis corresponds to a combined local significance of 3.4 standard deviations. In case of the signal+background hypothesis, the expected local significance corresponds to 1.2 standard deviations. The excess found in the combination is mainly driven by the same-sign dimuon channel of multilepton $t\bar{t}H$ search. Performing the maximum-likelihood fit exclusively in this analysis category yields a best-fit signal strength of $\mu_{\text{best-fit}}(t\bar{t}H) = 8.5^{+3.3}_{-2.7}$. Nevertheless, omitting this category in the evaluation of the final results still provides a p-value for the background-only hypothesis corresponding to a local significance of 2.2 standard deviations. The best-fit values of the signal-strength modifier of the individual search channels and the combination are illustrated in Fig. 11.6. The expected (observed) upper limits on the signal-strength modifier derived for the combination are

$$95\,\%\ \mathrm{CL_s}\ \text{observed (expected) upper limit on}\ \mu(t\bar{t}H)\text{:}\ 4.5(1.7^{+0.8}_{-0.5})\ .$$

The 95 % $\mathrm{CL_s}$ exclusion limits of the single search channels and the combination are illustrated in Fig. 11.7. The observed upper limit on the signal-strength modifier being weaker than the expected upper limit is compatible with the excess found in the maximum-likelihood fit. Comparing these results to the results of the $t\bar{t}H$ combination performed for LHC run II, the excess found in the 8 TeV data can not be confirmed. Accordingly, the observed upper limit on the signal strength of 2.1 obtained for the LHC run II combination is more stringent, than the one measured in LHC run I. As for the single-lepton channel of the $t\bar{t}(H\to b\bar{b})$ analysis, the expected upper limits on the signal strength of both combinations lie very close together. Correspondingly, also for the combination of all $t\bar{t}H$ search channels in LHC run II a sensitivity similar to the one of the LHC-run-I combination has been achieved with only about half the number of expected $t\bar{t}H$ events. Again, reasons for this are the increase of the $t\bar{t}$ cross section when moving from a center-of-mass energy of $\sqrt{s} = 8$ TeV to a center-of-mass energy $\sqrt{s} = 13$ TeV, which is smaller than for $t\bar{t}H$ production, and improvements in the individual analyses.

The $t\bar{t}H$ searches performed by the ATLAS and CMS collaborations based on the data recorded at center-of-mass energies of $\sqrt{s} = 7$ TeV and $\sqrt{s} = 8$ TeV have been combined [59]. For this combination, a common best-fit value of the signal-strength modifier of

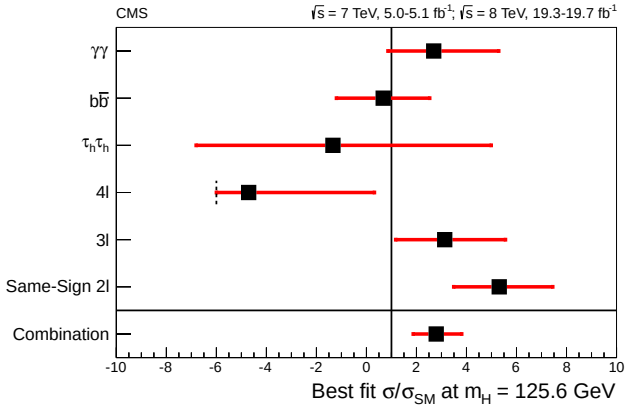


Figure 11.6: Best-fit values of $\mu_{t\bar{t}H}$ of the individual $t\bar{t}H$ search channels and their combination at a center-of-mass energy of $\sqrt{s} = 8$ TeV. The observed best-fit values of $\mu_{t\bar{t}H}$ are displayed as black markers and the corresponding uncertainties are displayed as red error bars. The Standard Model expectation of the signal strength is marked as black vertical line. Taken from [78].

$$\mu_{\text{best-fit}}(t\bar{t}H) = 2.3^{+0.7}_{-0.6} \,,$$

has been determined. The excess observed is caused by the large best-fit signal strength obtained by the combination of the $t\bar{t}H$ searches performed by the CMS collaboration and also by an excess observed in the $t\bar{t}H$ searches performed by the ATLAS collaboration [79, 231–233]. As for the CMS analyses, an especially large deviation from the SM expectation is found in the multilepton $t\bar{t}H$ search ($\mu_{\text{best-fit,ML}}(t\bar{t}H) = 2.1^{+1.4}_{-1.2}$). Nevertheless, also the searches by the ATLAS collaboration targeting $t\bar{t}H$ signatures with a Higgs-boson decay into a bottom-quark pair and into two photons have determined signal strengths above the SM expectation ($\mu_{\text{best-fit,bb}}(t\bar{t}H) = 1.4^{+1.0}_{-1.0}$ and $\mu_{\text{best-fit},\gamma\gamma}(t\bar{t}H) = 1.3^{+2.6}_{-1.7}$). The best-fit value of the signal-strength modifier derived for the combination of the ATLAS and CMS results is still in agreement with the SM prediction. The excess observed in the combination of the CMS and the ATLAS $t\bar{t}H$ searches at center-of-mass energies of $\sqrt{s} = 7\,\text{TeV}$ and $\sqrt{s} = 8\,\text{TeV}$ are not confirmed by the $\sqrt{s} = 13\,\text{TeV}$ CMS $t\bar{t}H$ combination.

Next to the signal strength, the coupling strengths of the Higgs boson to other SM particles has been measured based on the combination of all Higgs-boson analyses carried out by the ATLAS and CMS collaborations with the data recorded at center-of-mass energies of $\sqrt{s} = 7\,\text{TeV}$ and $\sqrt{s} = 8\,\text{TeV}$. This is achieved with a combined fit with coupling-strength modifiers, which are defined by the ratio of the measured coupling strength and the prediction by the SM, as parameters. For the top-Higgs coupling, a coupling-strength modifier of

$$\kappa_t = 1.43^{+0.23}_{-0.22} \,,$$

has been determined, while allowing for beyond the SM contributions to the Higgs-boson width. A similar value of

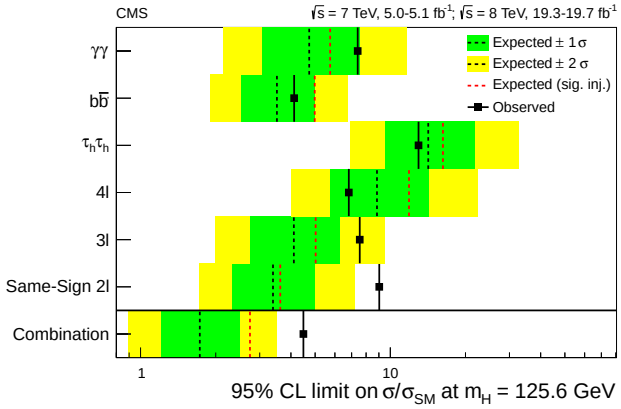


Figure 11.7: 95 % CL_s upper limits on $\mu_{t\bar{t}H}$ of the individual $t\bar{t}H$ search channels and their combination at a center-of-mass energy of $\sqrt{s} = 8\,\text{TeV}$. The observed upper limits on $\mu_{t\bar{t}H}$ are displayed as black markers. The expected limits are calculated with the asymptotic method and displayed as the median (black dashed line), the $\pm 1\sigma$ (green) and $\pm 2\sigma$ (yellow) confidence intervals. Taken from [78].

$$\kappa_t = 1.40^{+0.24}_{-0.21} \,,$$

has been derived, when not considering beyond the SM contributions to the Higgs-boson width. By definition, the SM prediction corresponds to $\kappa_{t,SM} = 1$. The deviation from this value can be explained by the excess in the signal-strength obtained for the $t\bar{t}H$ searches, as $t\bar{t}H$ production strongly contributes to the measurement of the top-Higgs coupling. Nevertheless, the measured coupling-strength modifier is still compatible with the SM prediction.

11.4.2. 13 TeV $t\bar{t}H$ Search at ATLAS

Some time after the $t\bar{t}H$ searches performed by the CMS collaborations on the first LHC-run-II data from 2015 have been published, results on the $t\bar{t}H$ searches have been released by the ATLAS collaboration. The difference in time allowed the ATLAS collaboration to include the first data recorded in the 2016 data-taking period in their first LHC-run-II $t\bar{t}H$ searches. The datasets used for these analyses correspond to an integrated luminosity of $\mathcal{L} = 13.2\,\text{fb}^{-1}$. Similar to the analyses $t\bar{t}H$ searches performed by the CMS collaboration, $t\bar{t}H$ search channels targeting Higgs-boson decays into two photons [234], multilepton final states [235], and Higgs-boson decays into a bottom-quark pair [236] have been considered. Further, a combination of all search channels [237] has been performed.

Single-lepton $t\bar{t}(H\to b\bar{b})$ Search by the ATLAS collaboration

The $t\bar{t}(H\to b\bar{b})$ search performed by the ATLAS collaboration is subdivided into two channels analyzing single-lepton and dilepton signatures similar to the $t\bar{t}(H\to b\bar{b})$ search performed by the CMS collaboration. The structure of the single-lepton $t\bar{t}(H\to b\bar{b})$ analysis performed by the ATLAS collaboration resembles the one described in this thesis. First,

events are selected based on the signature expected for $t\bar{t}(H\to b\bar{b})$ events with a semileptonic $t\bar{t}$ decay. In a subsequent step, the selected events are categorized according to their resolved jet and b-tag multiplicity. The bulk of the analysis categories are identical between the analyses performed by the CMS and the ATLAS collaboration. However, the single-lepton $t\bar{t}(H\to b\bar{b})$ search by the ATLAS collaboration includes two additional background-enriched analysis categories: the resolved 4 jets, 2 b-tags analysis category and the resolved 5 jets, 2 b-tags analysis category. The analysis by the CMS collaboration, on the other hand, additionally includes the boosted analysis category. The choice of the final discriminant in the $t\bar{t}(H\to b\bar{b})$ search by ATLAS depends on the signal fraction in the respective analysis categories. Correspondingly, the analysis categories are split into signal-enriched and background-enriched analysis categories. In the background-enriched analysis categories, a single kinematic variable, which is the sum over the transverse momenta of all resolved jets in the event, is used as final discriminant. In the signal-enriched analysis categories, a two-staged multivariate approach is chosen. The first stage is a set of reconstruction BDTs, which are trained with kinematic variables of correct and incorrect assignments of jets to the final-state quarks in simulated $t\bar{t}H$ events. Two separate BDTs are trained that include and exclude the information of the Higgs-boson in the training. The two reconstruction BDTs and additional kinematic variables provide the input for the second stage of the multivariate approach, the classification BDTs. These BDTs are trained to discriminate signal-like and background-like events and are used as final discriminants in the signal-enhanced analysis categories.

Another major difference of the analysis performed by the ATLAS collaboration with respect to the one performed by the CMS collaboration is the modeling of the $t\bar{t}$ background. Similar to the approach by the CMS collaboration, the ATLAS collaboration uses simulated $t\bar{t}$ events generated with Powheg+Pythia6. This sample of simulated events is split into different $t\bar{t}$+X contributions using an approach similar to the one described in Section 3.2.8. Afterwards, the $t\bar{t}$+$b\bar{b}$ contribution is reweighted based on an exclusive $t\bar{t}$+$b\bar{b}$ sample separately generated with Sherpa+OpenLoops [213, 238]. For the evaluation of the final results, some of the $t\bar{t}$+X contributions are merged to form a coarse set of categories: $t\bar{t}$+light-flavor, $t\bar{t}$+$\geq$ 1c, and $t\bar{t}$+$\geq$ 1b.

The results of the $t\bar{t}(H\to b\bar{b})$ search in the single-lepton channel performed by the ATLAS collaboration are the best-fit value of the signal-strength modifier,

$$\mu_{\text{best-fit}}(t\bar{t}H) = 1.6^{+1.1}_{-1.1} ,$$

and the 95 % CL_s upper limits on the signal-strength modifier,

$$95\,\%\ CL_s \text{ observed (expected) upper limit on } \mu(t\bar{t}H)\text{: } 3.6\,(2.2^{+1.0}_{-0.8}) .$$

The resulting best-fit value slightly exceeds the value expected for the SM case, but is still very well in agreement within the uncertainties. The observed upper limit on the signal-strength modifier is weaker than the expected upper limit. This effect is well compatible with the slightly increased best-fit value.

The LHC-run-II results obtained for the single-lepton $t\bar{t}(H\to b\bar{b})$ search by the ATLAS collaboration are not directly comparable to the results of the analysis presented in Section 11.1 due to the difference in the amount of data used. However, the results of the analysis performed by the ATLAS collaboration can be compared to the projections of the single-lepton $t\bar{t}(H\to b\bar{b})$ search for an integrated luminosity of $\mathcal{L} = 13.2\ \text{fb}^{-1}$ described

in Section 12.1. In the mentioned section, also the case of a reduction of the statistical and systematic uncertainties is studied. Nevertheless, for the comparison with the results provided by the ATLAS collaboration, the most conservative estimate, which does not take into account a reduction of uncertainties, is chosen. The blinded expected upper limit obtained by the projection for an integrated luminosity of $\mathcal{L} = 13.2\ \text{fb}^{-1}$ is

$$95\,\%\ \text{CL}_\text{s}\ \text{blinded expected upper limit on}\ \mu(\text{t}\bar{\text{t}}\text{H}) : 1.8^{+0.8}_{-0.5}\ .$$

The comparison of the projected result with the results of the analysis performed by the ATLAS collaboration, hints at a slightly better performance of the analysis performed by the CMS collaboration. This tendency could be explained by improvements due to the combination of BDT and MEM approach in final discrimination and the introduction of the boosted analysis category. However, the results provided in this naive way of projecting to a larger integrated luminosity have to be treated with caution. The observed difference is small and might not be significant given the uncertainty of the procedure.

$\text{t}\bar{\text{t}}\text{H}$ Combination at ATLAS

The individual $\text{t}\bar{\text{t}}\text{H}$ searches performed by the ATLAS collaboration are merged in order to provide a common result. As for the $\text{t}\bar{\text{t}}\text{H}$ combinations performed by the CMS collaboration, a common likelihood function is constructed using the final discriminants and nuisance parameters of all search channels. Systematic uncertainties are correlated as appropriate. For example uncertainties related to theoretical predictions and the detector are mainly treated as fully correlated. As the contributions of background processes are mostly different among the search channels, systematic uncertainties associated to them are treated as uncorrelated. The branching ratios of the different Higgs-boson decay modes are constrained to their SM predictions based on the respective experimental and theoretical uncertainties. Differences with respect to the combination performed by the CMS collaboration are the inclusion of signatures featuring hadronically decaying tau leptons in the multilepton search channel and the standalone diphoton $\text{t}\bar{\text{t}}\text{H}$ analysis. The results of the $\text{t}\bar{\text{t}}\text{H}$ combination performed by the ATLAS collaboration are the best-fit value of the signal-strength modifier,

$$\mu_\text{best-fit}(\text{t}\bar{\text{t}}\text{H}) = 1.8^{+0.7}_{-0.7}\ ,$$

and the $95\,\%$ CL_s upper limits on the signal-strength modifier,

$$95\,\%\ \text{CL}_\text{s}\ \text{observed (expected) upper limit on}\ \mu(\text{t}\bar{\text{t}}\text{H}):\ 3.0\,(1.2^{+0.5}_{-0.3})\ .$$

The best-fit value of the signal-strength modifier is found to lie slightly above the SM expectation $\mu(\text{t}\bar{\text{t}}\text{H}) = 1$, but is still compatible. The resulting observed upper limit on the signal-strength modifier is weaker than the expected one. This effect has also been observed for the exclusion limits of the individual $\text{t}\bar{\text{t}}(\text{H}\to\text{b}\bar{\text{b}})$ and multilepton search channels. As these search channels represent the most sensitive search channels, the difference in the exclusion limits is most likely propagated to the combination. The observed effect is well compatible with the slight excess found for the best-fit value for the signal-strength modifier. The results of the combination are illustrated in Fig. 11.8. In this figure, the

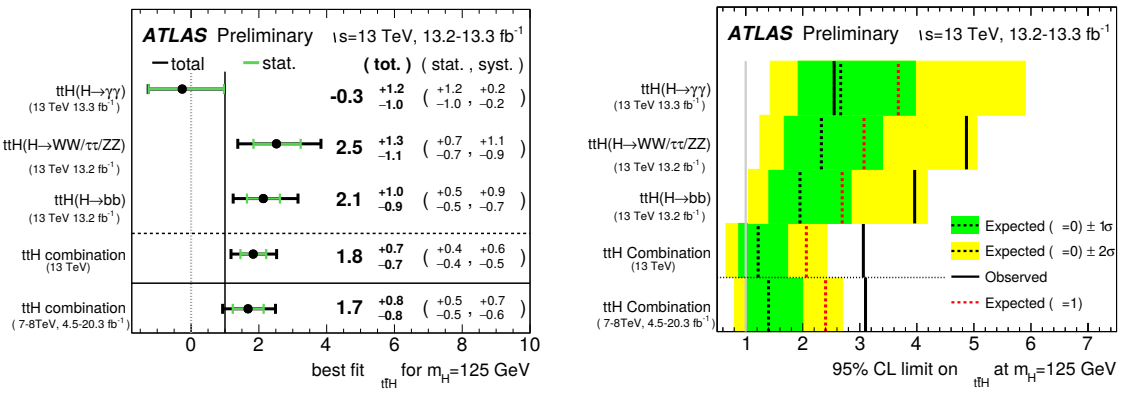


Figure 11.8: Best-fit value of $\mu_{t\bar{t}H}$ (left) for the combination and 95 % CL_s upper limits on $\mu_{t\bar{t}H}$ (right) for the individual $t\bar{t}H$ search channels and their combination performed at a center-of-mass energy of $\sqrt{s} = 8$ TeV. The best-fit values of $\mu_{t\bar{t}H}$ are displayed as black markers. The SM prediction is illustrated as black line. The observed limits are displayed as solid black lines. The expected limits are calculated with the asymptotic method and displayed as the median (dashed black line), the $\pm 1\sigma$ (green) and $\pm 2\sigma$ (yellow) confidence intervals. Taken from [237].

best-fit value of the signal-strength modifier and the observed and expected upper limits on the signal-strength modifier are shown for each individual $t\bar{t}H$ search channel, the combination of all search channels, and the $t\bar{t}H$ combination performed by the ATLAS collaboriation in LHC run I.

A rough comparison of the $t\bar{t}H$ combination performed by the ATLAS collaboration and the one performed by the CMS collaboration can be performed based on the actual results of both $t\bar{t}H$ combinations. The median expected limit of both combinations show a difference of about 40 %. Based on the signal significance $S/\sqrt{B}$, the evolution of the expected upper limit on the signal-strength modifier can be roughly estimated by $1/\sqrt{\mathcal{L}}$, which is proven by the luminosity projections performed in Section 12.1. Taking into account that the dataset available for the combination performed by the ATLAS collaboration is about four times larger than the one used for the $t\bar{t}H$ analyses performed by the CMS collaboration, one can deduce that the performance of both combinations are comparable. Relative differences in performance can be observed for the individual search channels. The multilepton search performed by the ATLAS collaboration, for example, shows a larger expected upper limit on the signal-strength modifier than the $t\bar{t}(H\to b\bar{b})$ search channel. In case of the analyses performed by the CMS collaboration on the other hand, the multilepton search channel is the most sensitive $t\bar{t}H$ search channel. Concerning the observed limits, the upper limit provided by the $t\bar{t}H$ combination of the CMS collaboration is in good agreement with the expected upper limit and consequently lower than the observed limit determined by the ATLAS collaboration. This effect can be explained by the observed limits of the two most sensitive search channels, the $t\bar{t}(H\to b\bar{b})$ analysis and the multilepton analysis, which are much weaker than the expected upper limits. The weak observed limits resulting for $t\bar{t}H$ analyses by the ATLAS collaboration are in good agreement with the corresponding best-fit values of the signal-strength modifier, which take on values above the Standard Model prediction of $\mu_{t\bar{t}H} = 1$. The best-fit value of the $t\bar{t}H$ combination performed by the CMS collaboration, on the other hand, takes on a value below the Standard Model prediction.

Chapter 12

Prospects

The search for $t\bar{t}H$ production has not reached its grand finale yet. Even though the introduction of boosted-analysis techniques and the combination of MEM and BDT approach provide systematic improvements to the $t\bar{t}(H\to b\bar{b})$ search and the combination of all $t\bar{t}H$ search channels, the searches are not sensitive enough for the observation or the exclusion of SM $t\bar{t}H$ production. The analysis presented in this thesis is based on the 2015 dataset corresponding to an integrated luminosity of $\mathcal{L} = 2.7\,\text{fb}^{-1}$ recorded with the CMS experiment at a center-of-mass energy of $\sqrt{s} = 13\,\text{TeV}$. As stated in Section 2.3, data corresponding to an integrated luminosity of about $\mathcal{L} = 38,7\,\text{fb}^{-1}$ has already been recorded throughout the course of 2016. Until the end of the run time of the current LHC setup, a total integrated luminosity of $\mathcal{L} = 300\,\text{fb}^{-1}$ is pursued. After this run period, an upgrade of the LHC to the high-luminosity LHC is planned aiming at a total integrated luminosity of about $\mathcal{L} = 3000\,\text{fb}^{-1}$. More collected data corresponds to a larger signal significance and therefore to an increased sensitivity of the search. An exact prediction about the sensitivity of $t\bar{t}H$ searches performed in the future is not possible. However, a projection to larger integrated luminosities for the analysis presented in this thesis can be performed. The corresponding procedure and the results obtained are presented in Section 12.1. In order to reach the sensitivity required for the observation or the exclusion of SM $t\bar{t}H$ production with less integrated luminosity, further systematic improvements can be introduced. A possible systematic improvement to the semileptonic $t\bar{t}(H\to b\bar{b})$ search presented in this thesis is studied in Section 12.2. This improvement includes the inclusion of additional analysis categories aiming at semileptonic $t\bar{t}(H\to b\bar{b})$ signatures with single boosted massive particles. Some more thoughts on systematic improvements concerning the application of boosted analysis techniques in the semileptonic $t\bar{t}(H\to b\bar{b})$ search are discussed in Section 12.3.

12.1. Luminosity Projection

The collecting more data will lead to increased sensitivity of the measurements, whose sensitivity is still dominated by statistical effects. An estimation of the sensitivity of such analyses provided for larger integrated luminosities can be performed by scaling the contributions of simulated processes to the prediction for a given integrated luminosity. Subsequently, the blinded expected limit and the expected significance, which solely rely on simulation and provide a measure of the sensitivity of the analysis, are calculated for this alternative setup. The blinded expected upper limits are determined based on the asymptotic method described in Section 5.2.4. The signal significances are derived based on a profile likelihood for the signal+background hypothesis. This approach represents only a projection of the results of the analysis presented in this thesis for integrated luminosities different from the one corresponding to the analyzed data. It is only a crude

estimation of the sensitivity of the results provided by future analyses, as it does not take into account the evolution of uncertainties, systematic changes in the analysis, and changes in the experimental setup. An example of the latter case is an increase of the center-of-mass energy of the LHC to its design center-of-mass energy of $\sqrt{s} = 14\,\text{TeV}$. Further changes not considered by the projection are future upgrades to the detector, which, for example, will provide a better performance of b-tagging. Larger integrated luminosities are mainly accomplished by increasing the instantaneous luminosity. Larger values of the instantaneous luminosity will lead to a larger number of proton-proton collisions, which increases the contributions by pile-up. This effect is also not accounted for by the projection presented in this section.

The blinded expected upper limits on the signal-strength modifier $\mu_{\text{t}\bar{\text{t}}\text{H}}$ and the significance for the observation of SM $\text{t}\bar{\text{t}}\text{H}$ production for the semileptonic $\text{t}\bar{\text{t}}(\text{H}\to\text{b}\bar{\text{b}})$ search as a function of the integrated luminosity is shown in Fig. 12.1. An alternative projection configuration includes an estimation of the evolution of some uncertainties affected by the change of the integrated luminosity. This is performed by scaling the one sigma interval of the respective uncertainty with a factor $1/\sqrt{f_{\mathcal{L}}}$, where $f_{\mathcal{L}}$ represents the ratio of the projected integrated luminosity and the initial integrated luminosity. For this projection, only the statistical uncertainties associated to the b-tagging scale factors are taken into account, as they have the largest impact on the results among all statistics dominated uncertainties. A third projection configuration takes into account possible improvements in the theoretical calculations of cross sections. For this configuration, all uncertainties on the calculated cross sections of the various processes considered are halved, also the ones additionally introduced to account for uncertainties on the cross sections of the different $\text{t}\bar{\text{t}}$+heavy-flavor contributions. The modification of the uncertainties on the calculated cross sections is constant for all integrated luminosities considered in the projection. Table 12.1 shows the results obtained for all three projection configurations and for various values of the integrated luminosities. The first luminosity value of $\mathcal{L} = 2.7\,\text{fb}^{-1}$ corresponds to the recored dataset analyzed by the $\text{t}\bar{\text{t}}(\text{H}\to\text{b}\bar{\text{b}})$ search presented in this thesis. The second value of $\mathcal{L} = 13.2\,\text{fb}^{-1}$ corresponds to the amount of data used for the $\text{t}\bar{\text{t}}(\text{H}\to\text{b}\bar{\text{b}})$ search performed by the ATLAS collaboration described in Section 11.4.2. An integrated luminosity of $\mathcal{L} = 30\,\text{fb}^{-1}$ and more is expected for the end of the 2016 data taking run. The integrated luminosities $\mathcal{L} = 300\,\text{fb}^{-1}$ and $\mathcal{L} = 3000\,\text{fb}^{-1}$ mark the goals of the current LHC setup and the high-luminosity LHC respectively. The evolution of the blinded expected upper limits obtained show a $1/\sqrt{f_{\mathcal{L}}}$ behavior. This is expected, as the exclusion limits rely on the signal significance $\text{S}/\sqrt{\text{B}}$, which follows the described behavior. A $\text{t}\bar{\text{t}}\text{H}$ signal-strength corresponding to the SM prediction of $\mu_{\text{t}\bar{\text{t}}\text{H}} = 1$ can be excluded at a 95 % confidence level with about $\mathcal{L} = 40\,\text{fb}^{-1}$. In case of the presence of a SM $\text{t}\bar{\text{t}}\text{H}$ signal, a significance of about 2σ is expected with the same integrated luminosity. A significance of 3σ, which corresponds to evidence for SM $\text{t}\bar{\text{t}}\text{H}$ production, is expected between $\mathcal{L} = 100\,\text{fb}^{-1}$ and $\mathcal{L} = 150\,\text{fb}^{-1}$. The projection reaches an expected significance of 5σ, which corresponds to the observation of SM $\text{t}\bar{\text{t}}\text{H}$ prediction, at an integrated luminosity of about $\mathcal{L} = 350\,\text{fb}^{-1}$. However, these results only include the semileptonic $\text{t}\bar{\text{t}}(\text{H}\to\text{b}\bar{\text{b}})$ search. A combination with other $\text{t}\bar{\text{t}}\text{H}$ search channels increases the overall sensitivity. The results obtained with the different projection methods are very similar. The consideration of the evolution of statistics dominated uncertainties has almost no effect on the final exclusion limits but a slightly larger impact on the expected significances. The reduced theoretical cross-section uncertainties have an impact on the exclusion limits, which is larger than the effect on the expected significances. However, both changes result in only minor changes of the

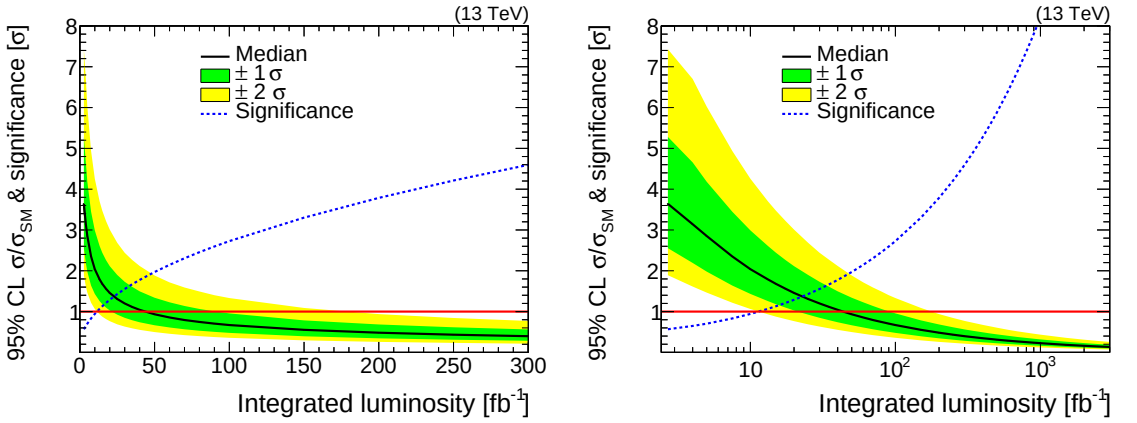


Figure 12.1: 95 % CL_s blinded expected upper limit on the signal-strength modifier $\mu_{t\bar{t}H}$ and the expected significance for the observation of SM $t\bar{t}H$ production in the semileptonic $t\bar{t}(H\to b\bar{b})$ search as a function of the integrated luminosity. For this projection, only the integrated luminosity is scaled. The projection up to an integrated luminosity of $\mathcal{L} = 300\ \text{fb}^{-1}$ is displayed on the left. An additional projection with integrated luminosities up to $\mathcal{L} = 3000\ \text{fb}^{-1}$ is shown on the right with a logarithmic x-axis scale. The blinded expected limits are calculated with the asymptotic method and displayed by the median (black), the $\pm 1\sigma$ (green) and $\pm 2\sigma$ (yellow) confidence intervals. The standard model expectation is illustrated as red line. The expected significance is displayed as dashed blue line. The starting point is the integrated luminosity of $\mathcal{L} = 2.7\ \text{fb}^{-1}$ corresponding to the dataset analyzed by the $t\bar{t}(H\to b\bar{b})$ search presented in this thesis.

exclusion limits and the expected significances.

12.2. Single Boosted Signatures

In the analysis described in the previous chapters, only boosted $t\bar{t}H$ signatures featuring a boosted hadronic top-quark as well as a boosted Higgs-boson both with a transverse momentum of $p_T > 200\ \text{GeV}/c$ are targeted. Events showing such a signature contribute about 7 % of all semileptonic $t\bar{t}(H\to b\bar{b})$ events produced at a center-of-mass energy of $\sqrt{s} = 13$ TeV. Further, about 7 % of all semileptonic $t\bar{t}(H\to b\bar{b})$ events feature only a boosted Higgs-boson, whereas 21 % feature only a boosted hadronic top-quark. The large efficiency in correctly reconstructing and identifying the massive particles in the boosted analysis category indicates that a specific analysis of events with single boosted massive particles could bring an improvement in sensitivity. Therefore, the use of the information of boosted massive particles in the search for $t\bar{t}(H\to b\bar{b})$ is extended by introducing analysis categories specifically targeting events with single boosted massive particles. This extension includes two new analysis categories requiring either a single reconstructed boosted top-quark candidate or a single boosted Higgs-boson candidate. In the following, these categories are denoted as single boosted analysis categories, while the boosted analysis category introduced in the previous chapters is referred to as double boosted analysis category. For the definition of the new categories, two different boosted-event reconstruction procedures based on the hypotheses of featuring either a single boosted top quark or a single boosted Higgs boson are performed. The reconstruction procedures are similar to the one described in Section 8.1 and also employ the objects described in Chapter 7. A

Table 12.1: 95 % CL_s blinded expected upper limits on the signal strength modifier $\mu_{t\bar{t}H}$ and the expected significance for the observation of SM $t\bar{t}H$ production for the semileptonic $t\bar{t}(H\to b\bar{b})$ search as a function of the integrated luminosity. The expected limits are displayed for different projection configurations. The configuration shown in the first row only includes the scaling of the integrated luminosity. The second row additionally considers the impact of the luminosity scaling on a set of statistics dominated uncertainties. The third row accounts for a better accuracy of the calculated cross sections by halving the corresponding uncertainties in addition to the configuration of the second row. The blinded expected upper limit displayed for an integrated luminosity of $\mathcal{L} = 2.7\ \text{fb}^{-1}$ is the exclusion limit determined by the analysis presented in this thesis.

Scaling of	95 % CL_s blinded expected upper limits for a luminosity of				
	2.7 fb^{-1}	13.2 fb^{-1}	30 fb^{-1}	300 fb^{-1}	3000 fb^{-1}
Luminosity	$3.6^{+1.6}_{-1.1}$	$1.8^{+0.8}_{-0.5}$	$1.2^{+0.5}_{-0.4}$	$0.4^{+0.2}_{-0.1}$	$0.1^{+0.1}_{-0.0}$
Lumi. & syst. unc.	$3.6^{+1.6}_{-1.1}$	$1.8^{+0.8}_{-0.5}$	$1.2^{+0.5}_{-0.4}$	$0.4^{+0.2}_{-0.1}$	$0.1^{+0.1}_{-0.0}$
Lumi., syst. unc., & XS unc.	$3.4^{+1.4}_{-1.0}$	$1.7^{+0.7}_{-0.5}$	$1.1^{+0.5}_{-0.3}$	$0.4^{+0.2}_{-0.1}$	$0.1^{+0.0}_{-0.0}$

Scaling of	Expected significance for a luminosity of				
	2.7 fb^{-1}	13.2 fb^{-1}	30 fb^{-1}	300 fb^{-1}	3000 fb^{-1}
Luminosity	0.6	1.1	1.6	4.6	14.0
Lumi. & syst. unc.	0.6	1.1	1.6	4.8	14.7
Lumi., syst. unc., & XS unc.	0.6	1.1	1.6	4.9	14.8

more detailed description is presented in Section 12.2.1. Based on the outcome of the event reconstruction, events are selected for the single boosted analysis categories. The new analysis categories are added to the set of analysis categories described in the previous chapters. The event selection and categorization for the single boosted analysis categories is described in Section 12.2.2. The remaining part of the analysis is similar to the main analysis presented in the previous chapters. The final discriminants used for the extraction of the final results are covered in Section 12.2.3. The results obtained for the analysis including single boosted analysis categories are presented in Section 12.2.4.

12.2.1. Single Boosted Event Reconstruction

Starting with the set of boosted top-quark candidates reconstructed with the procedures described in Chapter 7, the one with the highest top-quark classification output is chosen as the boosted top-quark candidate for the single boosted top-quark event interpretation. In an analogue way, the candidate for the single boosted Higgs-boson event interpretation is chosen from the set of reconstructed boosted Higgs-boson candidates. In addition to the boosted candidates, the event is further reconstructed based on resolved jets. A "cleaned" set of resolved jets is obtained by discarding resolved jets spatially overlapping with the subjets of the boosted candidate. This is achieved by applying an angular-matching procedure which identifies resolved jets as overlapping, if it features a jet axis with $\Delta R < 0.3$ with respect to any boosted candidate subjet. In the single boosted Higgs-boson analysis category, additionally a reconstruction of the $t\bar{t}$ system based on the cleaned set of resolved jets is performed. The reconstruction procedure applied is identical

to the resolved χ^2 reconstruction described in Section 6.2.

12.2.2. Single Boosted Event Selection & Categorization

The selection of events for the single boosted analysis categories makes use of the outcome of the single boosted event reconstruction and the reconstructed and selected objects described in Chapter 4. The baseline of this selection features requirements on the reconstructed resolved objects and is identical to the one of the double boosted analysis category presented in Section 8.2. Accordingly, a good reconstructed primary vertex is the first requirement for the selection of events. Further, exactly one selected charged lepton candidate is required, whereas events with additional charged leptons fulfilling the loose selection criteria described in Section 4.9.2 are vetoed. Additionally, the multiplicity of all reconstructed resolved jets in the event has to be larger than four and two resolved jets are required to be b-tagged. On top of this baseline selection, events have to fulfill criteria based on the results of the single boosted and double boosted event reconstruction. First of all, events selected by the double boosted event selection, which requires a boosted top-quark candidate and a boosted Higgs-boson candidate as described in Section 8.2, are rejected. The rejection of these events is necessary to ensure that events are not selected by both, the single boosted and the double boosted analysis categories. From the remaining events, events for the single boosted top-quark analysis category are selected. This selection requires events to feature a boosted top-quark candidate emerging from the single boosted top-quark event reconstruction with a top-quark classification output larger than -0.49. Further, events selected for this category are required to have at least three b-tagged resolved jets left after the removal of resolved jets overlapping with the boosted top-quark candidate subjets. Events not selected for the single boosted top-quark analysis category are potential candidates for the single boosted Higgs-boson analysis category. In order to be selected for this analysis category, events are required to feature a boosted Higgs-boson candidate resulting from the single boosted Higgs-boson event reconstruction with a Higgs-boson classification output above 0.89. Further, events selected for this category are required to feature at least two b-tagged resolved jets remaining after the removal of jets overlapping with the boosted Higgs-boson candidate. The requirements on the boosted-object classification outputs in the single boosted analysis categories are identical to the ones applied for the double boosted analysis category. This choice is based on the good performance achieved for the double boosted analysis category. Still, a further optimization of these requirements is not excluded and might bear an increase of the performance of the single boosted analysis categories. The requirements on the multiplicity of resolved b-tagged jets are chosen to match the number of bottom quarks expected in semileptonic $\mathrm{t\bar{t}}(\mathrm{H}\rightarrow\mathrm{b\bar{b}})$ events. With a reconstructed boosted top quark candidate, three bottom quarks remain unassociated, two bottom quarks from the Higgs-boson decay and one bottom quark from the leptonic top-quark decay. In case a boosted Higgs-boson candidate is reconstructed, there are two bottom quarks and two light quarks left to be assigned. One of the bottom quarks and the two light quarks originate from the hadronic top-quark decay and the remaining bottom quark stems from the leptonic top-quark decay. Requirements targeting the two light quarks in the single boosted Higgs-boson analysis category are not considered. The requirements have been tested in simulation and have shown to provide good performance. Requiring one b-tag less in each single boosted analysis category in order to account for acceptance effects and a limited b-tagging efficiency, enriches these analysis categories in background events. This change in configuration approximately cuts the signal over background ratios in these categories in half and weakens

the resolved analysis categories due to the increasing number of overlapping events. The latter effect is discussed in the following. The newly introduced single boosted analysis categories are added to the existing set of resolved analysis categories and the double boosted analysis category. However, similar to the double boosted analysis category, the single boosted analysis categories include events that are also selected by the resolved analysis categories. In order to avoid a double counting of events, overlapping events are assigned to the single boosted analysis category and vetoed in the resolved analysis categories.

The event yields of all analysis categories considered by the analysis configuration described in this section are displayed in Table 12.2. Additionally, the events yields of each analysis category including the splitting by the BDTs for final discrimination described in Section 9.4 are visualized in Fig. 12.2. The event yields of signal and background in the single boosted analysis categories are very promising. These categories feature 2.0 and 3.5 predicted signal events, respectively, which are not very many. However, these events face a very small number of background events. Accordingly, the event yields of the single boosted analysis categories are very similar to the ones obtained for the best performing analysis categories, the resolved analysis categories requiring four b-tags and the double boosted category. In this configuration, the single boosted analysis categories provide the best signal over background ratio (S/B) among all analysis categories. However, the effect on the resolved analysis categories caused by the vetoing of events selected by the single boosted analysis categories has to be taken into account. The extent of this effect is revealed in the comparison of the event yields of the original analysis configuration presented in Table 12.2 with the event yields provided by the analysis configuration including the single boosted analysis categories presented in Table 8.3. The comparison shows that resolved analysis categories requiring large multiplicities of jets and b-tags lose the most events by vetoing events selected by the single boosted analysis categories. A consequence of this is a significant reduction of the S/B in these analysis categories compared to the S/B obtained by the analysis configuration without single boosted analysis categories. Compared to the resolved $\geq$6 jets, $\geq$4 b-tags analysis category, the single boosted analysis categories show a slightly larger $\mathrm{t\bar{t}}$+light-flavor contribution and a slightly smaller contribution by $\mathrm{t\bar{t}}$+$\mathrm{b\bar{b}}$ production. This feature is more pronounced in the single boosted Higgs-boson analysis category. In both cases, this leads to a better separation of the overall background in the final discrimination with respect to the resolved $\geq$6 jets, $\geq$4 b-tags analysis category.

12.2.3. Final Discrimination

The final discrimination of signal against background in the single boosted analysis categories follows the same approach as the final discrimination of the other analysis categories, which is described in Chapter 9. Similar to the double boosted analysis category, BDTs trained with the MEM discriminant have been chosen as the final discriminants for the newly introduced analysis categories. The exact choice of the variables and the parameters used in the training are determined by the particle-swarm optimization (PSO) described in Section 5.1.4. A configuration of the PSO identical to the one presented in Section 9.3 is applied. The modeling of the input variables by simulation has been checked in dedicated control regions and the single boosted analysis categories. Further, the agreement of correlations between the input variables in data and simulation has been tested. The parameter configuration obtained by the optimization is shown in Table 12.3. The best performing input variables are presented in Table 12.4. Most of the input variables used for the training of the BDTs in the single boosted analysis categories have already been

Table 12.2: Event yields of recorded data and simulated processes for all analysis categories considered in the semileptonic $t\bar{t}(H\to b\bar{b})$ search including single boosted analysis categories. The event yields of simulated events are scaled to the values expected for an integrated luminosity of $\mathcal{L} = 2.7\ \text{fb}^{-1}$.

Process	≥6 jets, 2 b-tags	4 jets, 3 b-tags	5 jets, 3 b-tags	≥6 jets, 3 b-tags	4 jets, ≥4 b-tags
$t\bar{t}$+lf	5359.2 ± 1226.3	2021.4 ± 650.6	995.7 ± 351.6	578.5 ± 196.9	16.1 ± 9.8
$t\bar{t}$+$c\bar{c}$	1722.0 ± 849.4	361.6 ± 190.0	364.7 ± 189.4	384.8 ± 202.8	10.3 ± 7.4
$t\bar{t}$+b	393.7 ± 188.1	202.1 ± 92.2	198.0 ± 90.1	164.2 ± 78.2	7.6 ± 4.0
$t\bar{t}$+2b	165.2 ± 81.1	77.5 ± 37.3	84.3 ± 39.3	91.6 ± 43.9	3.0 ± 1.5
$t\bar{t}$+$b\bar{b}$	226.0 ± 112.9	74.6 ± 34.7	111.0 ± 51.0	172.7 ± 81.6	8.7 ± 4.2
Single Top	283.0 ± 49.0	114.5 ± 30.7	74.8 ± 19.3	45.8 ± 12.2	2.0 ± 0.9
V+jets	130.5 ± 35.2	37.7 ± 17.6	22.2 ± 10.2	13.2 ± 6.3	0.9 ± 0.9
$t\bar{t}$+V	43.4 ± 8.2	4.2 ± 1.2	6.2 ± 1.7	9.5 ± 1.6	0.2 ± 0.1
Diboson	2.8 ± 1.3	2.1 ± 1.3	0.9 ± 0.5	0.2 ± 0.2	0.0 ± 0.0
Total bkg	8325.8 ± 1788.4	2895.9 ± 834.1	1857.9 ± 530.6	1460.3 ± 410.9	48.8 ± 20.6
$t\bar{t}$H	29.6 ± 2.3	7.2 ± 1.0	10.7 ± 1.3	15.7 ± 1.8	0.7 ± 0.2
Data	7185	2782	1892	1343	68
S/B	0.004 ± 0.001	0.003 ± 0.001	0.006 ± 0.001	0.011 ± 0.003	0.015 ± 0.006
Data/B	0.9 ± 0.1	1.0 ± 0.2	1.0 ± 0.2	0.9 ± 0.2	1.4 ± 0.5

Process	5 jets, ≥4 b-tags	≥6 jets, ≥4 b-tags	boosted	boosted top	boosted Higgs
$t\bar{t}$+lf	15.3 ± 9.3	13.2 ± 8.4	45.1 ± 9.4	10.5 ± 5.0	18.7 ± 8.3
$t\bar{t}$+$c\bar{c}$	19.3 ± 13.5	25.3 ± 17.7	21.8 ± 12.0	14.0 ± 9.6	18.2 ± 12.8
$t\bar{t}$+b	12.9 ± 6.8	13.9 ± 7.8	10.3 ± 5.5	7.6 ± 4.2	10.7 ± 6.9
$t\bar{t}$+2b	5.6 ± 3.0	7.7 ± 4.3	12.3 ± 6.6	4.2 ± 2.7	12.8 ± 6.6
$t\bar{t}$+$b\bar{b}$	23.0 ± 11.2	44.4 ± 22.1	17.0 ± 8.4	23.0 ± 12.2	31.2 ± 17.3
Single Top	3.2 ± 1.2	3.8 ± 1.6	7.0 ± 1.7	1.6 ± 0.8	5.7 ± 1.8
V+jets	0.9 ± 0.7	0.5 ± 0.4	2.5 ± 0.8	0.6 ± 0.4	2.5 ± 1.5
$t\bar{t}$+V	0.5 ± 0.3	0.8 ± 0.3	0.9 ± 0.3	1.3 ± 0.4	1.9 ± 0.5
Diboson	0.1 ± 0.1	0.0 ± 0.0	0.1 ± 0.1	0.0 ± 0.0	0.0 ± 0.0
Total bkg	80.9 ± 25.6	109.6 ± 43.8	116.6 ± 18.7	62.7 ± 22.3	101.7 ± 33.9
$t\bar{t}$H	1.9 ± 0.4	3.2 ± 0.7	2.2 ± 0.2	2.0 ± 0.4	3.5 ± 0.7
Data	85	97	104	55	99
S/B	0.024 ± 0.009	0.029 ± 0.012	0.019 ± 0.004	0.032 ± 0.012	0.034 ± 0.011
Data/B	1.1 ± 0.4	0.9 ± 0.3	0.9 ± 0.2	0.9 ± 0.3	1.0 ± 0.3

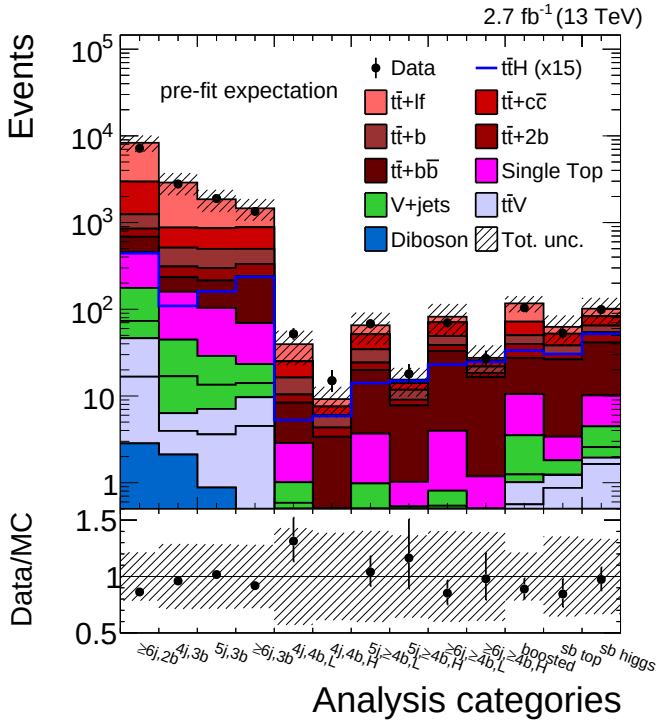


Figure 12.2: Event yields of recorded data and simulation in each final analysis category including the subcategories provided by the 2D approach for the analysis configuration and including the single boosted analysis categories. The background contributions are displayed as stacked filled histograms. The contribution of the $t\bar{t}H$ signal process is displayed as a blue line. The simulated processes are scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7\,\text{fb}^{-1}$. For better visibility, the signal process is scaled by an additional factor of 15. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distribution.

introduced in Section 9.3. The variables not covered there and special features of the single boosted analysis categories are discussed in the following:

- **Single boosted top-quark analysis category:** One of the most important variables in this category is the MEM discriminant using subjet information. As described in Section 9.2, this MEM discriminant is adapted to the boosted event reconstruction and solely relies on the subjets of the boosted top-quark candidate and additional resolved jets. No boosted Higgs-boson candidate is required for the calculation. Further information of the boosted-event reconstruction used in the training of the BDT includes the invariant mass of the boosted top-quark candidate ***m*(boosted top cand.)**. This variable discriminates signal from background events with fake boosted top-quark candidates. The lack of a reconstructed boosted Higgs-boson candidate requires the use of the Higgs-boson information provided by the resolved-event interpretation. Corresponding variables are the invariant mass of the closest b-tagged jets ***m*(closest b-tag. jets)** and the invariant dijet mass closest to the Higgs mass $m_\text{H} = 125\,\text{GeV}/c^2$ (**b-tag. dijet mass closest to m_H**). The remaining variables used in the training of the BDT in the single boosted top-quark

analysis category can be grouped into variables describing the angular correlations between jets, variables describing the energy and momenta of jets, and variables based on b-tagging information.

- **Single boosted Higgs-boson analysis category:** Due to the lack of a boosted hadronic top-quark candidate, the MEM discriminant using subjet information is not applicable in this category. Instead, the MEM discriminant solely relying on resolved jets is used. As in the double boosted analysis category, the information on the reconstructed boosted Higgs-boson candidate provides some of the most discriminating variables in the training of the BDT in the single boosted Higgs-boson analysis category. These variables include the invariant mass of the boosted Higgs-boson candidate $\boldsymbol{m}$**(boosted Higgs cand.)** and the difference in pseudo rapidity between the reconstructed top-quark candidate and the boosted Higgs-boson candidate $\boldsymbol{\Delta\eta}$**(top cand., Higgs cand.)**. In this case, the reconstructed top-quark candidate emerges from the $\mathrm{t\bar{t}}$ reconstruction described in the previous section, which relies on the resolved jets not overlapping with subjets of the boosted Higgs-boson candidate. Further, information provided by the resolved event reconstruction is used. Corresponding variables are the square root of the product of differences in pseudo rapidity between the reconstructed $\mathrm{b\bar{b}}$ system and the reconstructed leptonic and hadronic top-quark candidates $\sqrt{\boldsymbol{\Delta\eta}(\mathbf{t_{lep.}}, \mathbf{b\bar{b}}) \times \boldsymbol{\Delta\eta}(\mathbf{t_{had.}}, \mathbf{b\bar{b}})}$ and the median of the invariant mass of all pairs of b-tagged resolved jets **median** $\boldsymbol{m}$**(b-tag. dijets)**. The remaining variables used in the training of the BDT in this category are mainly variables describing the angular correlations between resolved jets in the event and variables based on b-tagging information.

The BDTs of the single boosted analysis categories are optimized and trained using simulated $\mathrm{t\bar{t}}(\mathrm{H}\rightarrow\mathrm{b\bar{b}})$ events as signal and simulated $\mathrm{t\bar{t}}$ events as background. For this purpose, the samples of simulated $\mathrm{t\bar{t}}(\mathrm{H}\rightarrow\mathrm{b\bar{b}})$ and $\mathrm{t\bar{t}}$ events presented in Section 3.2.7 are both split into two statistically independent samples. The events used for training and optimization of the BDTs are taken from one part of the samples, while the other part is used for the evaluation of the final results. This procedure is applied to avoid overtraining. The output distributions of the BDTs in the single boosted analysis categories evaluated on data and simulated events are displayed in Fig. 12.3. They show very good separation between signal and background events, which is comparable to the one observed for the double boosted category. This effect is mostly driven by the additional information brought by the boosted-event reconstruction and the composition of events in these categories. The BDTs used for the remaining categories are retrained with the BDT configurations described in Section 9.3. These BDT configurations are optimized for the analysis configuration without single boosted analysis categories. Accordingly, they do not account for the change in events caused by the vetoing of events selected by the single boosted analysis categories in the resolved analysis categories. Consequently, these BDTs do not perform optimally. However, this effect is expected to be small. For the final discrimination of signal against background in these categories, the approach presented in Section 9.4 is applied. The final discriminant distributions of the resolved analysis categories and the double boosted analysis category evaluated on data and simulated events can be found in Appendix A.6.

Table 12.3: Parameter configuration used for the training of the boosted decision trees in the single boosted analysis categories. The values are obtained by an optimization based on the particle-swarm optimization described in Section 5.1.4.

Category	N_{trees}	Shrinkage	Bagging fraction	N_{cuts}	Depth
boosted top	589	0.01	0.58	65	2
boosted Higgs	684	0.02	0.36	27	2

Table 12.4: Input variables used for the training of the boosted decision trees in the single boosted analysis categories. The variables are determined by an optimization based on the particle-swarm optimization described in Section 5.1.4.

Boosted top	**Boosted Higgs**
MEM discriminant (using subjets)	MEM discriminant
m(boosted top cand.)	m(boosted Higgs cand.)
p_{T}(third-hardest jet)	$\Delta\eta$(top cand., Higgs cand.)
p_{T}(second-hardest jet)	$\sqrt{\Delta\eta(\text{t}_{\text{lep.}}, \text{b}\bar{\text{b}}) \times \Delta\eta(\text{t}_{\text{had.}}, \text{b}\bar{\text{b}})}$
H_{T}	avg. $\Delta\eta$(jets)
avg. $\Delta\eta$(jets)	avg. ΔR(b-tag. jets)
avg. ΔR(b-tag. jets)	median m(b-tag. dijets)
m(closest b-tag. jets)	fourth-highest b-tag. output
b-tag. dijet mass closest to m_{H}	fifth-highest b-tag. output
third-highest b-tag. output	dev. avg. b-tag. output
fourth-highest b-tag. output	b-tagging likelihood ratio
fifth-highest b-tag. output	sphericity
aplanarity	

12.2.4. Results

The results for the $\text{t}\bar{\text{t}}(\text{H}\to\text{b}\bar{\text{b}})$ search including single boosted analysis categories are derived with the same procedure as for the $\text{t}\bar{\text{t}}(\text{H}\to\text{b}\bar{\text{b}})$ search presented in Chapter 11. A combined likelihood function based on the final-discriminant distributions of all analysis categories and all nuisance parameters is constructed. For this setup, exactly the same uncertainties and nuisance parameters are considered as described in Chapter 10. A first result is obtained by maximizing this likelihood function based on the variation of the signal-strength modifier $\mu_{\text{t}\bar{\text{t}}\text{H}}$ and the nuisance parameters. With this procedure, a best-fit value for the signal-strength modifier of

$$\mu_{\text{best-fit}}(\text{t}\bar{\text{t}}\text{H}) = -1.8^{+2.1}_{-2.2} \,,$$

is obtained. As already discussed in Chapter 10, negative values of the signal-strength modifier are unphysical, but not excluded by the fitting procedure. However, a similar tendency is also observed for the analysis configuration without single boosted analysis categories as shown in Chapter 11. This tendency is amplified by the inclusion of the single boosted analysis categories. A reason for this effect are the signal enriched regions in the final-discriminant distributions of these categories. There, fewer data events than simulated background events are found. This effect is especially prominent in the final-discriminant distribution of the single boosted top-quark analysis category, where no event

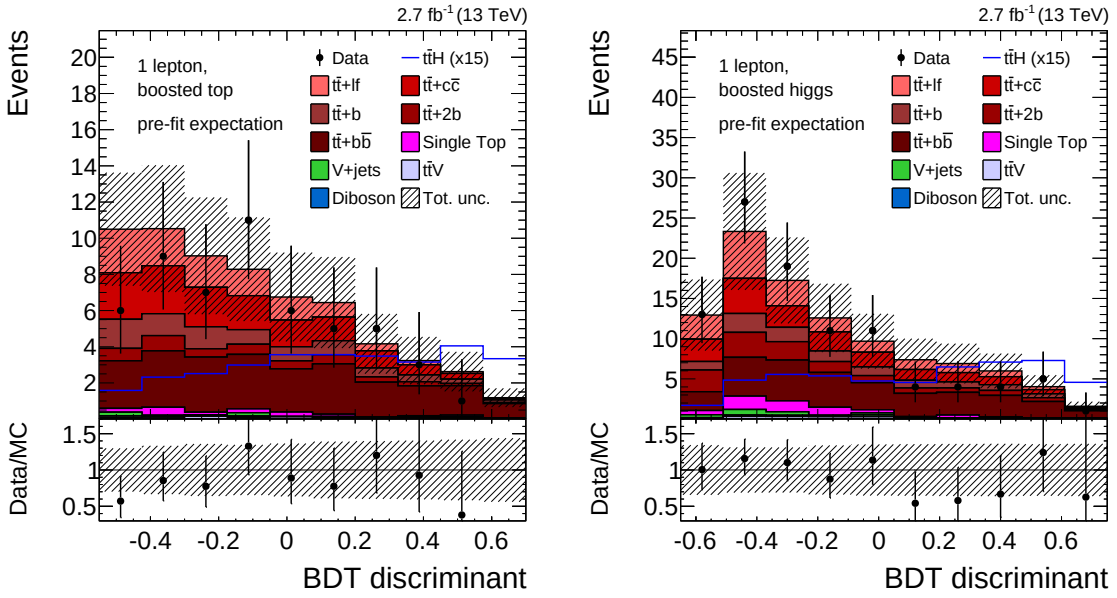


Figure 12.3: Output distributions of the BDTs evaluated on data and simulation in the single boosted analysis categories. The simulated processes are scaled to the event yields expected for an integrated luminosity of $\mathcal{L} = 2.7\,\text{fb}^{-1}$. The background contributions are displayed as stacked filled histograms. For better visibility the $\text{t}\bar{\text{t}}\text{H}$ signal process (blue line) is additionally scaled by a factor of 15.

has been observed for the last histogram bin. The variation of the nuisance parameters obtained by the maximum-likelihood fit is shown in Fig. 12.4. Compared to the configuration without single boosted analysis categories presented in Section 11.1, most of the nuisance parameters behave similar for the analysis configuration including the single boosted analysis categories. Differences can be observed for nuisance parameters affecting the $\text{t}\bar{\text{t}}$ background. Examples are the scale uncertainty on the cross section of the inclusive $\text{t}\bar{\text{t}}$ production (QCD scale($\text{t}\bar{\text{t}}$)) and the PDF uncertainty on the cross sections of gluon-induced processes (PDF(gg)), for which the corresponding nuisance parameters are very close to their pre-fit value of zero. The nuisance parameters associated to the parton-shower scale uncertainty of the different $\text{t}\bar{\text{t}}$+X contributions are constrained more tightly when including the single boosted analysis categories. Further, the nuisance parameters associated to $\text{t}\bar{\text{t}}$+b and $\text{t}\bar{\text{t}}$+$\text{c}\bar{\text{c}}$ production move closer to the expected value. The nuisance parameter associated to the uncertainty on the luminosity is also shifted closer to the pre-fit value when including the single boosted analysis categories.

More information is provided by the 95 % CL_s upper limits on the signal-strength modifier $\mu_{\text{t}\bar{\text{t}}\text{H}}$. The observed and expected upper limits are calculated for the combination of all categories as well as for each category individually. The extraction of the limits is based on a profile-likelihood ratio test statistic, which is described in Section 5.2.3. This test statistic is formed from likelihood functions similar to the one used for the maximum-likelihood fit. The respective likelihood functions are constructed from the final-discriminant distribution of all categories considered for the respective case and all nuisance parameters. For the determination of the expected upper limits, the asymptotic method is applied. The observed and expected upper limits obtained by these calculations are presented in Table 12.5. Further, an illustration of the results is presented in Fig. 12.5. Based on the

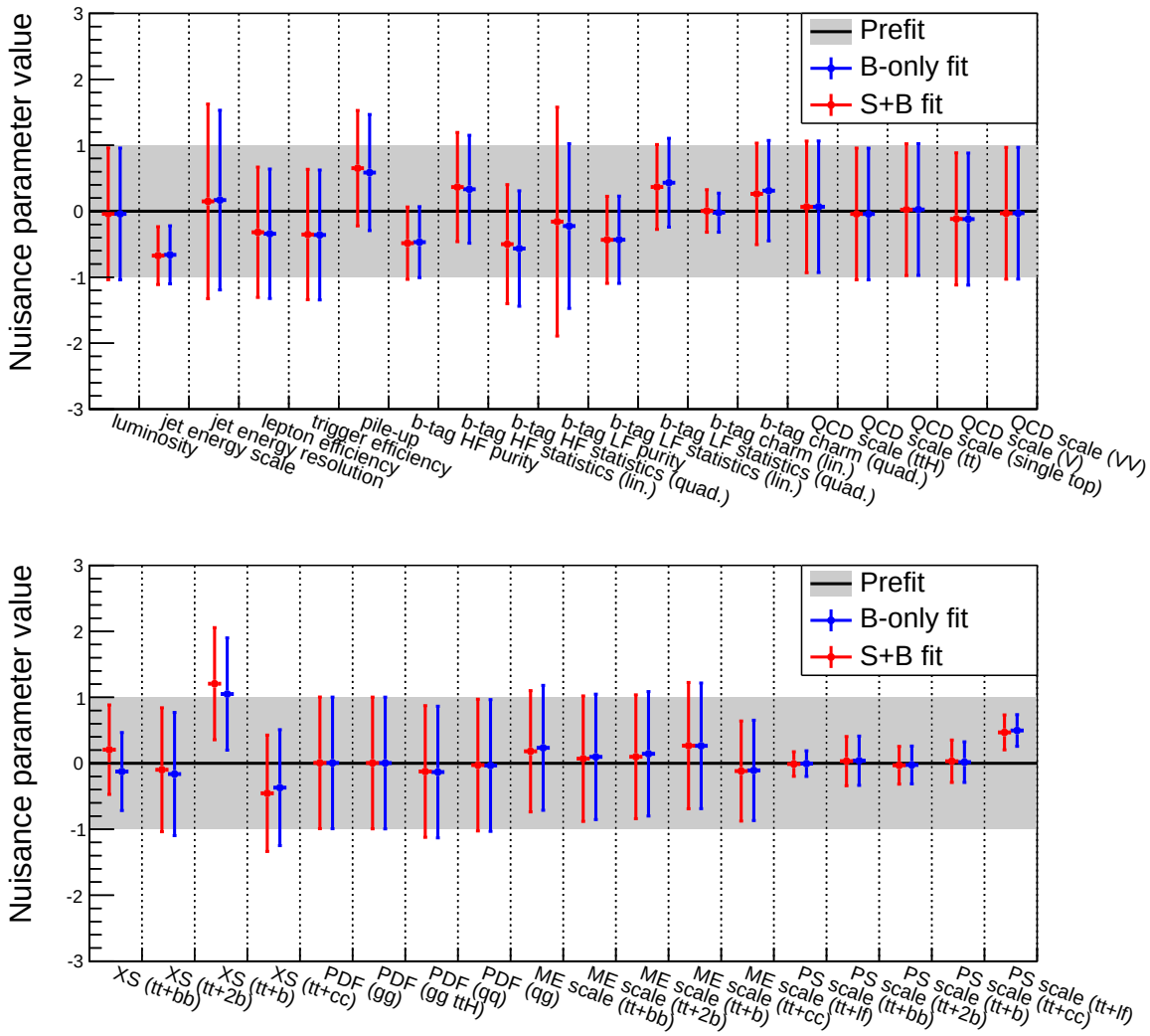


Figure 12.4: Nuisance-parameter values before and after the fit of the signal+background (red) and the background-only (blue) hypothesis for the analysis configuration including single boosted analysis categories. The black line shows the configuration before the fit, where all nuisance parameters are fixed to their initial value of zero. The y-axis shows the relative deviation from this value and the gray area marks the one standard-deviation interval.

expected limits obtained, which represent a measure of sensitivity, the newly introduced single boosted analysis categories are among the four best performing categories of this analysis configuration. This is not surprising, as the event yields of background and signal and the well discriminating BDTs indicate a good performance. However, compared to the standard configuration, the resolved analysis categories are considerably weakened. This effect is caused by the loss of events in the resolved analysis categories, which are now treated in the single boosted analysis categories. Both effects cancel each other and provide a combined expected limit very similar to the one provided by the $t\bar{t}(H\to b\bar{b})$ search without single boosted analysis categories. The slight improvement of about 2 % is not significant, when taking into account the accuracy of the result. The observed limits obtained for the resolved categories show a similar behavior as for the configuration without single boosted analysis categories. The expected and the observed upper limit on the signal-strength modifier in the double boosted analysis stays exactly when including the single boosted analysis categories. This is expected, as the definition of the boosted analysis category is not affected by the introduction of the single boosted analysis categories. The observed upper limits provided by the single boosted analysis categories are lower than the expected upper limits. Similar to the maximum-likelihood fit, this is caused by the deficit of data events in the signal-enriched regions of the final-discriminant distributions of these analysis categories. As the single boosted analysis categories are among the most sensitive categories, this effect is also propagated to the combined observed limit. Accordingly, a more stringent combined observed limit is measured for the configuration with single boosted analysis categories than for the configuration without them. The deficit of observed events in the signal-enriched regions of the final-discriminant distributions in the boosted analysis categories are possibly caused by a downwards fluctuation of data. However, the consistency of this effect among the different boosted analysis categories hints at a potential mismodelling of the variables in these analysis categories. This effect has to be studied for future iterations of this analysis.

Summarized, no significant improvement of the semileptonic $t\bar{t}(H\to b\bar{b})$ analysis is achieved by the introduction of the single boosted analysis categories presented in this section. However, with more data available, more restrictive single boosted analysis categories can be defined, while ensuring a significant amount of predicted signal events. This approach would limit the weakening of the resolved analysis categories, which might lead to a residual improvement of the overall analysis.

12.3. Future Considerations

The introduction of a new analysis category requiring a boosted hadronically decaying top quark and a boosted Higgs boson in the semileptonic $t\bar{t}(H\to b\bar{b})$ analysis already represents a milestone. Its configuration has been optimized, in order to be competitive with the resolved analysis categories and to improve the analysis. Further, the impact of introducing single boosted analysis categories on the sensitivity of the semileptonic $t\bar{t}(H\to b\bar{b})$ search has been tested. However, there are further possible systematic changes concerning the application of boosted analysis techniques that might provide an improvement in sensitivity or a better treatment of uncertainties. A detailed study of these changes exceed the scope of this thesis. Nevertheless, a selection of the most promising improvements and possible changes for future $t\bar{t}(H\to b\bar{b})$ searches are presented in the following.

Table 12.5: Observed and median expected 95 % CL_s upper limits on $\mu_{t\bar{t}H}$ obtained for the single-lepton $t\bar{t}(H\to b\bar{b})$ search with single boosted analysis categories included. Expected limits are calculated with the asymptotic method. The upper and the lower boundaries of the one standard-deviation confidence interval of the median expected upper limits on $\mu_{t\bar{t}H}$ are also stated.

Category	Observed upper limit	Expected upper limit
4 jets, 3 b-tags	13.5	$18.1^{+8.3}_{-5.3}$
4 jets, ≥4 b-tags high BDT output	42.9	$30.9^{+16.1}_{-9.8}$
4 jets, ≥4 b-tags low BDT output	116.1	$100.8^{+50.6}_{-30.8}$
5 jets, 3 b-tags	14.6	$12.6^{+5.6}_{-3.7}$
5 jets, ≥4 b-tags high BDT output	10.8	$13.3^{+7.3}_{-4.4}$
5 jets, ≥4 b-tags low BDT output	47.8	$39.9^{+20.2}_{-12.5}$
≥6 jets, 2 b-tags	26.5	$41.9^{+21.5}_{-13.1}$
≥6 jets, 3 b-tags	9.7	$8.0^{+3.4}_{-2.3}$
≥6 jets, ≥4 b-tags high BDT output	14.5	$12.3^{+6.7}_{-4.0}$
≥6 jets, ≥4 b-tags low BDT output	21.2	$25.6^{+13.6}_{-8.3}$
≥4 jets, ≥2 b-tags, boosted	7.5	$10.7^{+5.9}_{-3.5}$
≥4 jets, ≥2 b-tags, boosted top	9.0	$10.8^{+6.3}_{-3.7}$
≥4 jets, ≥2 b-tags, boosted Higgs	6.7	$8.2^{+4.6}_{-2.7}$
Single lepton combined	3.2	$4.0^{+1.8}_{-1.2}$

Categorization

The set of analysis categories presented in this thesis are physically well motivated. Still, differently defined analysis categories potentially exploit the distinctive properties of signal and background events more efficiently. An example of the exploitation of such properties is already provided by the boosted analysis category newly introduced in the search for single-lepton $t\bar{t}(H\to b\bar{b})$ production. The large amount of recorded data that will be available in the future will allow a finer separation into analysis categories, while still maintaining statistical significance. During studies targeting the boosted analysis category, it was observed that a large fraction of subjets are also described well by resolved jets. Correspondingly, a major fraction of boosted particles in the moderately boosted regime defined by the boosted analysis category can still be resolved by resolved object reconstructions. In future analyses, the boosted analysis category presented in this thesis could be further subdivided into a highly boosted analysis category requiring boosted massive particles with large transverse momenta and a semi-boosted analysis category with moderately boosted massive particles. This would allow the use of dedicated reconstruction methods adapted to the special signatures in these analysis categories. In the highly boosted analysis category, the application of the boosted analysis techniques presented in this thesis would be necessary. In the semi-boosted analysis category, a reconstruction approach based on resolved jets could be applied, while still exploiting the advantages

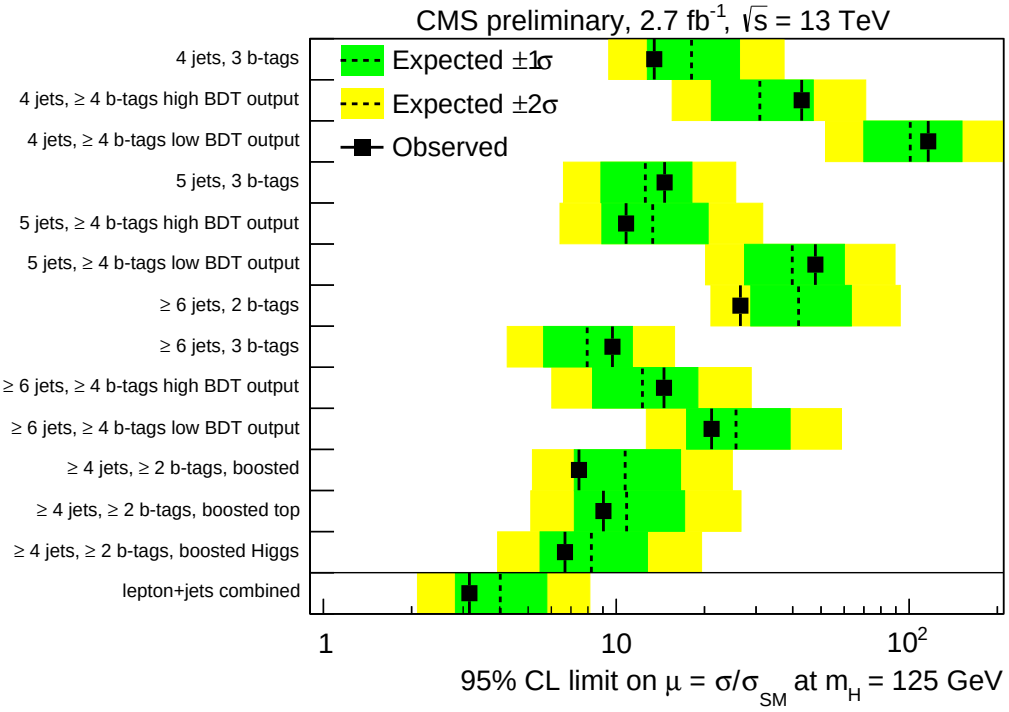


Figure 12.5: Visualization of 95 % CL_s upper limits on $\mu_{t\bar{t}H}$ obtained for the single-lepton $t\bar{t}(H\to b\bar{b})$ search with single boosted analysis categories included. The observed limits are illustrated as black markers. The expected limits are calculated with the asymptotic method and displayed by the median (dashed black line), the $\pm1\sigma$ (green) and $\pm2\sigma$ (yellow) confidence intervals.

of the boosted regime. A reconstruction method based on this idea is presented in the next subsection. Further categories could exploit the signatures featuring single boosted massive particles as presented in Section 12.2. However, the approach used there would have to be further improved in order to provide significant improvements.

Another approach for potentially improving the categorization in this analysis could be based on multivariate methods, which separate the phase space into signal-enriched regions and several control regions. Such an approach would also include the events currently covered by the resolved analysis categories.

Boosted-Object Reconstruction

The reconstruction and identification of massive particles in the boosted analysis category achieves very large efficiencies compared to the ones provided in the resolved analysis categories. Especially, the reconstruction of the hadronically decaying top quark with the combination of the HEPTopTaggerV2 and a multivariate identification is very successful. Boosted hadronic top-quark decays are reconstructed correctly in more than half of the events selected by the boosted analysis category. However, the reconstruction of the Higgs boson, which provides the most distinctive features in the separation of $t\bar{t}H$ events from background events, does not reach the same reconstruction efficiency. As presented in Section 7.7, several substructure algorithms have already been tested for the reconstruction

of a boosted Higgs boson decaying into a bottom-quark pair. In these studies, no large differences have been found for most of the approaches. As mentioned in the previous subsection, many of the subjets are also described by resolved jets. A reconstruction of moderately boosted massive particles based on resolved jets would represent an alternative to the substructure methods presented in Chapter 7. This would bring the advantage of the applicability of suited jet-energy corrections and a better treatment of the associated uncertainties. A possible boosted reconstruction approach based on resolved jets could include the clustering of the selected resolved jets in the event with the Cambridge/Aachen algorithm and a cone size of $\delta R = 1.5$. Subsequently, only clusters of jets above a particular transverse-momentum threshold are selected as possible candidates for boosted massive particles. The massive particles are reconstructed and identified based on the resolved jets associated to a cluster. For this purpose, a procedure similar to the one described in Section 7.7 could be applied. First studies of applying a boosted-object reconstruction based on resolved jets show results comparable to the ones achieved with the substructure methods.

Unlike the identification of true boosted top-quark candidates, the identification of true boosted Higgs-boson candidates relies on a single variable, the second highest b-tagging output among the three hardest subjets found in a fat jet. The reconstruction efficiency of boosted Higgs-bosons could potentially be improved by applying a multivariate approach, like BDTs, for the identification of real boosted Higgs-boson candidates. However, applying multivariate methods in the identification bears the risk of adjusting the shape of background distributions to the ones of the signal process for variables used in the final discrimination. This effect can be reduced by an elaborate choice of input variables used for the multivariate approach. Still, the impact of a multivariate boosted Higgs-boson identification on the performance of the analysis results has to be studied by a propagation to the final results.

Matrix-Element Method using Subjets

The MEM discriminant using subjets in the boosted analysis category does not make use of the full information provided by the boosted event reconstruction. Only the subjets of the boosted top-quark candidate are used as input for the calculation of the MEM discriminant in addition to a set of resolved jets. The subjets of the boosted Higgs-boson candidate are not considered so far, as suitable transfer functions are missing. However, these functions can be determined for future iterations of the $t\bar{t}(H\to b\bar{b})$ search presented in this thesis. Further, a boosted Higgs-boson reconstruction based on resolved jets, as described in the previous subsection, would not require dedicated transfer functions. However, using the full information provided by the boosted-event reconstruction would not necessarily provide better performance. Of course, the large reconstruction efficiency might be beneficial for the separation of signal and background events. However, also including resolved jets adds information by an alternative event interpretation. Which of the two effects dominates has to be tested by a dedicated study.

In the calculation of the MEM discriminant, the bottom-quark subjet of the top-quark candidate is not fixed to the assignment obtained by the reconstruction. The permutations assign the bottom-quark subjet to any bottom quark in the final state of $t\bar{t}(H\to b\bar{b})$ and $t\bar{t}+b\bar{b}$ events. The subjets associated to the light quark, on the other hand, are only assigned to the corresponding particles in the final state. Fixing the subjets to the associations obtained by the boosted-object reconstruction has hardly changed the output of the MEM-discriminant calculation. However, the time required for the calculation of the

MEM discriminant is largely reduced for this configuration, as the number of permutations decrease from twelve to three. Especially, the latter effect is a strong reason for fixing the assignment of subjets in the calculation of the MEM discriminant. For a calculation fully based on the outcome of the boosted-event reconstruction, the effect on the MEM discriminant might be larger. In this case, a repetition of this study is recommended.

Systematic Uncertainties

The uncertainties on the jet-energy corrections of subjets and fat jets have only a negligible effect on the final results of the analysis presented in this thesis. The corresponding study is described in Section 10.3.1. However, as more data is recorded and statistic as well as other systematic uncertainties become smaller, a more sophisticated treatment of these uncertainties is required. As the jet-energy corrections used for the Cambridge/Aachen subjets and fat jets are derived for resolved jets clustered with the anti-k_{T} algorithm and cone sizes of $R = 0.4$ and $R = 0.8$, the associated uncertainties should be correspondingly increased in future iterations of this analysis. Further, the dependence on the jet-energy corrections can be reduced by solely applying requirements on the subjets in the selection of boosted-object candidates, instead of also applying selection cuts on the transverse momenta of fat jets. Such requirements could target the transverse momentum of the combination of all subjets making up the boosted top-quark candidate or the boosted Higgs-boson candidate. This approach would be more restrictive than the transverse-momentum requirements on the fat jets, as the fat jets feature larger energy contributions from sources other than the hard constituents[1]. Accordingly, only boosted-object candidates with high-energetic subjets would be selected and more fat jets with a large fraction of soft contributions adding up to large transverse momenta would be rejected. Placing requirements only on subjets would lead to a slight change of the definition of the boosted analysis category and also the resolved analysis categories due to the vetoing of overlapping events. Based on the discussion above, less events with fat jets featuring no high-energetic subjets would be selected by the boosted analysis category. Instead such events would be treated in the resolved analysis category. For the determination of the impact of these effects, dedicated studies would be necessary.

An alternative to the application of the jet-energy corrections provided is the determination of dedicated jet-energy corrections for the subjets and fat jets used in this analysis. This could be accomplished by a two staged approach that first corrects for the absolute jet-energy scale in simulation and subsequently resolves residual differences for data and simulation. The former step would be based on an angular matching procedure of jets clustered from the simulated particles before the application of the detector simulation to the fat jets and subjets reconstructed based on the full simulation. The former type of jets would be clustered with the Cambridge/Aachen algorithm and a cone size of $R = 1.5$ for fat jets and a cone size of $R = 0.3$ for subjets, which should provide a reasonable approximation. From the ratios of the transverse momenta of the jets matched, scale factors differential in the transverse momentum and the pseudo rapidity would be determined. An analogue procedure is already applied by the ATLAS collaboration [239]. Scale factors correcting for differences in data and simulation could be derived by applying a tag-and-probe method in a control region enriched in $\mathrm{t\bar{t}}$ events with a semileptonic decay of the top-quark pair. This procedure is similar to the one applied for the determination

[1] There are exceptions to this if jet-energy corrections are applied to fat jets and subjets. These corrections can have a different effect on the momentum four-vectors of both jet types, what can lead to a fat-jet momentum that is smaller than the momentum of the combined subjets.

of the correction for the b-tagging output distributions of resolved jets described in Section 4.10.3. A control region enriched in $\mathrm{t\bar{t}}$ events with a semileptonic decay would be selected by requiring an isolated charged lepton, exactly four resolved jets, exactly two b-tags, and a selected fat jet with reconstructed subjets. To ensure that events are not used in the calibration as well as in the evaluation of the final results, events selected by the control region should be excluded from the definition of the boosted analysis category. In the selected events, the combination of a charged lepton and a resolved jet spatially not overlapping with the fat jet that features an invariant mass close to the top mass would serve as tag. The fat jet and the corresponding subjets, which are associated to the hadronically decaying top quark, are considered the probe. The ratios of transverse momenta of fat jets and subjets in data and simulation provide the scale factors. For the determination of scale factors for fat jets and subjets provided by the HEPTopTaggerV2 algorithm, this procedure could be applied without any alterations. The subjets provided by the BDRS algorithm could be calibrated based on the hadronically decaying W boson occurring in the top-quark decay. To ensure that the subjets describe the decay products of this particle, additional requirements are necessary. The described procedure resembles the approach used for the study of tagging efficiencies by the CMS collaboration [240].

A more sophisticated treatment of all sources contributing to the uncertainties on the jet-energy scale, which are described in Section 10.3.1, could be desirable. Such a treatment could include a correlation of the uncertainties of the different sources where appropriate and a separate handling otherwise. This change would also affect the resolved jets.

Chapter 13

Conclusion

With the discovery of the Higgs boson at the LHC in 2012, the last missing piece of the Standard Model of particle physics has been found. However, the search has not stopped at that point, as different production and decay modes of the Higgs boson remain unobserved and its detailed properties unmeasured. The years of search for this particle are now followed by years of precisely determining its properties and its behavior. Any deviations from the predictions provided by the Standard Model could hint at the existence of physics reaching beyond. One of the production modes not discovered to date is the Higgs-boson production in association with a top-quark pair ($t\bar{t}H$). This production mode is of special interest, as it comes with a particular feature, the direct access to the top-Higgs Yukawa coupling. This quantity describes the coupling of the Higgs boson to the top quark, the most massive particle of the Standard Model. As the Higgs boson couples to the mass of particles, the top-Higgs Yukawa coupling is large compared to the coupling of the Higgs boson to other particles of the Standard Model. Accordingly, this coupling plays an important role in the large perturbative corrections occurring in the calculation of the Higgs-boson mass, which are connected to the hierarchy problem.

The search for $t\bar{t}H$ production poses an enormous challenge. $t\bar{t}H$ searches have been performed based on the data recorded at the LHC at a center-of-mass energy of $\sqrt{s} = 8$ TeV. However, these analyses targeting various Higgs-boson decays are not sensitive enough to observe or exclude this process. The combination of all $t\bar{t}H$ searches performed by the CMS collaboration have achieved to exclude $t\bar{t}H$ cross sections larger than 4.5 times the Standard Model prediction. One of the main reasons for $t\bar{t}H$ production being hard to grasp is the small cross section of this process, which is due to the large amount of energy necessary to produce two top quarks and a Higgs boson. Especially, the search for $t\bar{t}H$ production with a Higgs-boson decay into a bottom-quark pair ($t\bar{t}(H\to b\bar{b})$) faces a large background from $t\bar{t}$ production, which features a cross section about 1600 times larger than the cross section of the signal process. The large number of decay products present in the final state additionally complicate the reconstruction of the event. Typical reconstruction methods require the assignment of reconstructed analysis objects to the decay products, which is ambiguous due to the large multiplicity of strongly interacting final state particles in $t\bar{t}(H\to b\bar{b})$ production. A solution to this combinatorial problem is provided by the investigation of signatures including massive particles with large transverse momenta, so-called boosted particles [3]. Typical areas of application for this approach are massive particles with very large transverse momenta originating from the decay of even more massive hypothetical particles. $t\bar{t}H$ production, on the other hand, features massive particles with only moderate transverse momenta. Still, massive particles produced by background processes tend to be softer than the ones expected in $t\bar{t}H$ events. The main reason for the application of boosted analysis techniques is the reduction of the ambiguity in the event reconstruction. Boosted massive particles pass on their momentum to their decay products, which emerge from the decays as collimated bunches of particles. Dedi-

cated analysis techniques are specialized in the reconstruction and identification of these collimated decay products. A common procedure is the clustering of all decay products into one fat jet followed by an analysis of its substructure. In case of $t\bar{t}(H\to b\bar{b})$ events, boosted hadronically decaying top quarks and Higgs bosons can be reconstructed and identified based on this approach.

This thesis describes the inclusion of this approach into the semileptonic $t\bar{t}(H\to b\bar{b})$ search based on the data recorded by the CMS experiment during the 2015 data-taking period of the LHC in form of a new analysis category. In the this context, different boosted techniques for the reconstruction and identification of boosted hadronically decaying top quarks and boosted Higgs bosons decaying into a bottom-quark pair have been tested. For the reconstruction and identification of boosted top quarks, the best performing configuration has been found to be a reconstruction method based on the HEPTopTaggerV2 combined with a multivariate identification specifically developed for this analysis. The best performing setup for the reconstruction and identification of boosted Higgs bosons is found to be the reconstruction of subjets based on the BDRS approach and the subsequent identification based on b-tagging information. The reconstructed candidates enter an event reconstruction procedure targeting signatures with one boosted hadronically decaying top quark and one boosted Higgs boson. Events for the boosted analysis category are selected based on the outcome of this boosted-event reconstruction by applying optimized selection requirements. The boosted analysis category is added to the set of analysis categories defined by standard analysis objects, like reconstructed jets and b-tags. The final discrimination of signal against background in each analysis category is performed by a combination of boosted decision trees (BDTs) and the matrix-element method (MEM). In case of the boosted analysis category, the MEM discriminant is calculated based on information provided by the boosted-event reconstruction. The results are evaluated with the data recorded by the CMS detector during the 2015 data-taking run of the LHC corresponding to an integrated luminosity of $\mathcal{L} = 2.7\,\text{fb}^{-1}$. This data represents the first data recorded at a center-of-mass energy of $\sqrt{s} = 13\,\text{TeV}$. The results show no excess of signal in data. Instead, $t\bar{t}H$ cross sections larger than four times the Standard Model expectations for $t\bar{t}H$ production are excluded at a 95 % confidence level. Compared to the analysis performed at a center-of-mass energy of $\sqrt{s} = 8\,\text{TeV}$, a similar sensitivity has been achieved with only half as many $t\bar{t}H$ events predicted. Projections to larger integrated luminosities are compatible with the results of the $t\bar{t}(H\to b\bar{b})$ search performed by the ATLAS collaboration with data corresponding to an integrated luminosity of $\mathcal{L} = 13.2\,\text{fb}^{-1}$. The analysis presented in this thesis has been combined with the dilepton $t\bar{t}H$ search and published [7]. This $t\bar{t}(H\to b\bar{b})$ search is not only the first $t\bar{t}(H\to b\bar{b})$ search at a center-of-mass energy of $\sqrt{s} = 13\,\text{TeV}$, but also the first $t\bar{t}H$ search relying on boosted analysis techniques. In this context, the analysis also represents a showcase for the application of boosted analysis techniques in a moderately boosted regime and a "busy" environment. The $t\bar{t}(H\to b\bar{b})$ search has been combined with the other $t\bar{t}H$ searches performed at a center-of-mass energy of $\sqrt{s} = 13\,\text{TeV}$ by the CMS collaboration. This combination achieves to exclude $t\bar{t}H$ cross sections larger than 2.1 times the Standard Model prediction at a 95 % confidence level. However, the sensitivity is not sufficient for the observation or exclusion of the Standard Model $t\bar{t}H$ production.

A study extending the published analysis tests the improvement brought by introducing two new analysis categories targeting $t\bar{t}(H\to b\bar{b})$ signatures with single boosted massive particles. The single boosted analysis categories are defined in a similar way as the boosted analysis category of the main analysis. Two different event-reconstruction procedures,

requiring either a boosted top quark or a boosted Higgs boson, are performed based on reconstructed boosted objects and standard resolved objects. Events are selected for the two new analysis categories based on the outcome of the respective event reconstruction. The new analysis categories are added to the set of the existing analysis categories. The final-discriminant distributions in these analysis categories are provided by a combination of the BDT and MEM discriminants. The results obtained by the setup including the single boosted analysis categories show no major improvement with respect to the analysis configuration omitting these analysis categories.

The sensitivity of the semileptonic $t\bar{t}(H\to b\bar{b})$ search for larger integrated luminosities has been studied by performing projections. They rely on scaling the simulated data samples to the prediction for a given integrated luminosity and determining blinded expected upper limits on the signal strength and the expected signal significance. According to the results, the $t\bar{t}(H\to b\bar{b})$ search presented in this thesis is able to exclude a signal-strength corresponding to the Standard Model prediction for $t\bar{t}H$ production at an integrated luminosity of about $\mathcal{L} = 40\,\text{fb}^{-1}$. Further, the projections predict an expected signal significance of three standard deviations, corresponding to evidence for SM $t\bar{t}H$ production, at an integrated luminosity of about $\mathcal{L} = 150\,\text{fb}^{-1}$ and an expected signal significance of five standard deviations, corresponding to the observation of SM $t\bar{t}H$ production, at an integrated luminosity of about $\mathcal{L} = 350\,\text{fb}^{-1}$. In the combination with other $t\bar{t}H$ searches, these milestones will be reached at even lower integrated luminosities. With an integrated luminosity of about $\mathcal{L} = 38\,\text{fb}^{-1}$ recorded by the CMS experiment in 2016, these projections indicate that exciting times for the search for $t\bar{t}H$ production are imminent.

Appendix A

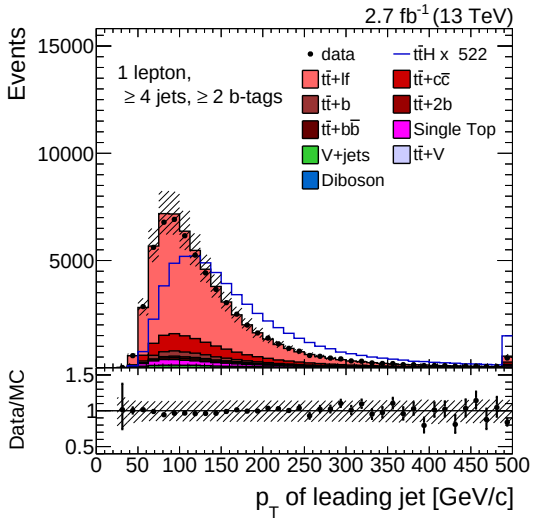
Supplementary Material for the Semileptonic $t\bar{t}(H\to b\bar{b})$ Anaylsis

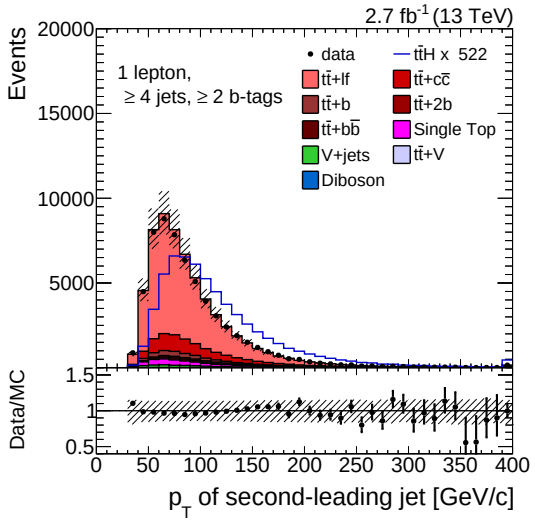
A.1. Validation

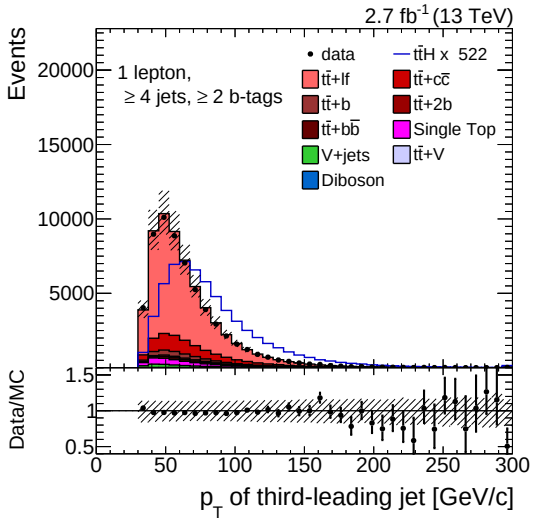
A.1.1. Resolved Validation

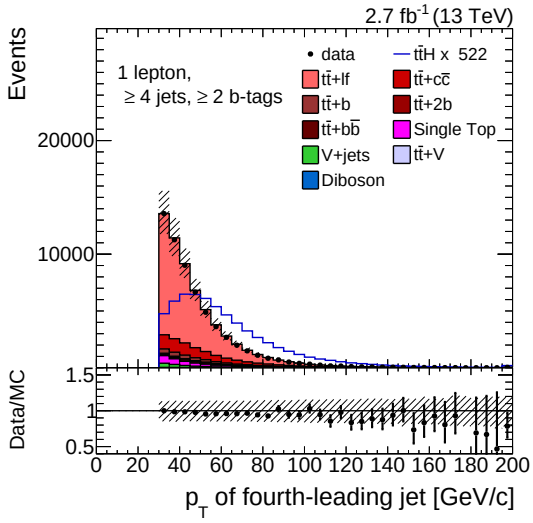
Distributions of variables are shown for an inclusive control region requiring one selected lepton, at least four resolved jets, and at least two b-tags. Simulated background processes are displayed as stacked filled histograms and are scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7$ fb^{-1}. The $t\bar{t}H$ signal is depicted as a blue line and scaled to the total predicted event yield of all background processes for better visibility. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions.

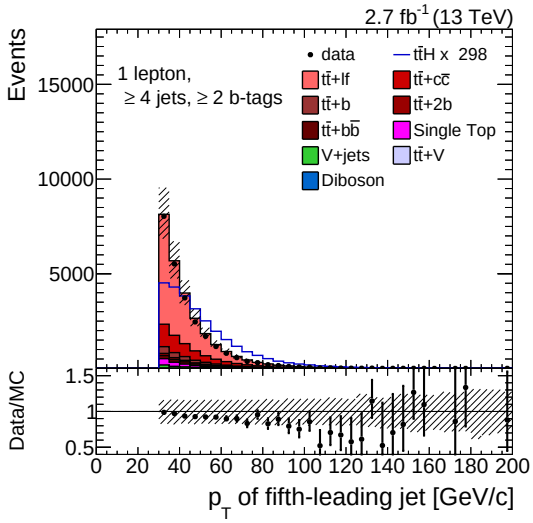
Jet Variables

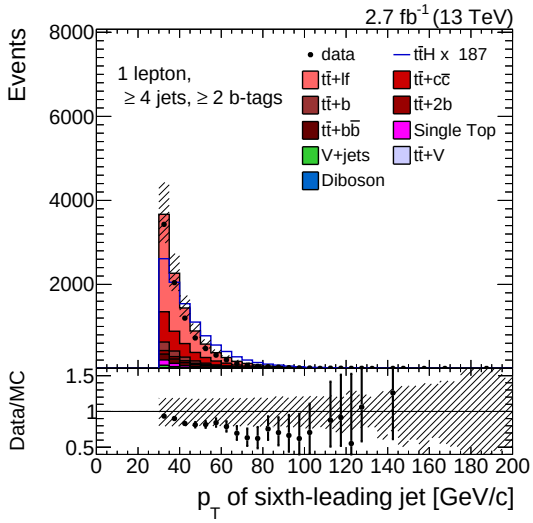


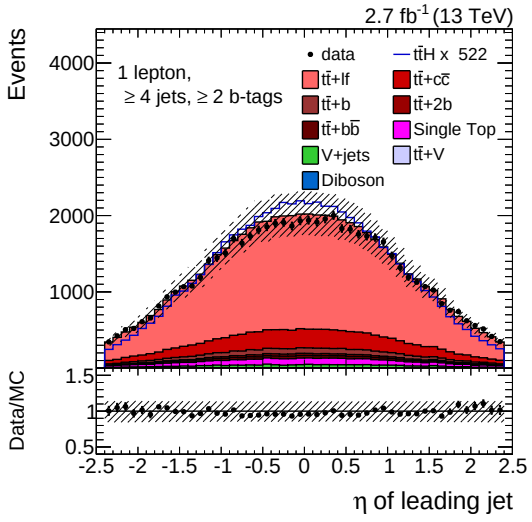


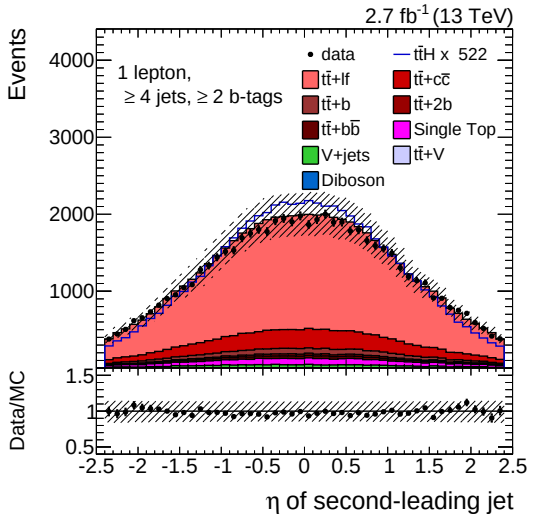

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
• data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
pT of third-leading jet [GeV/c]

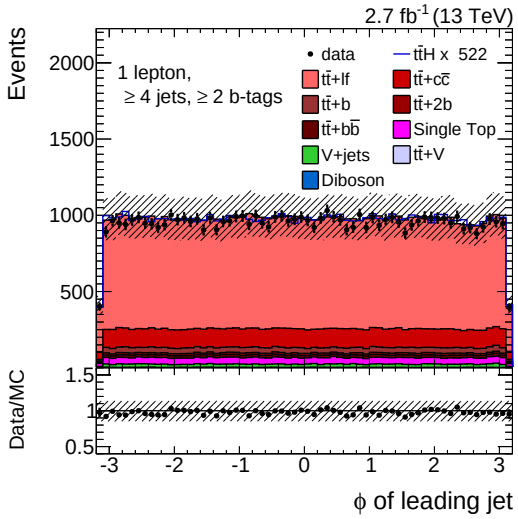

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
• data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
pT of fourth-leading jet [GeV/c]

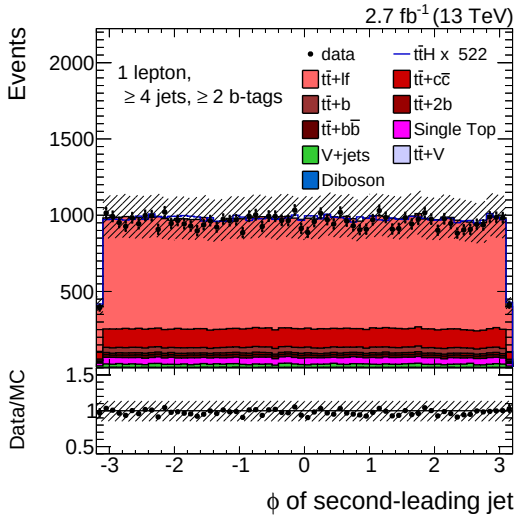

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
• data
tt̄H x 298
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
pT of fifth-leading jet [GeV/c]

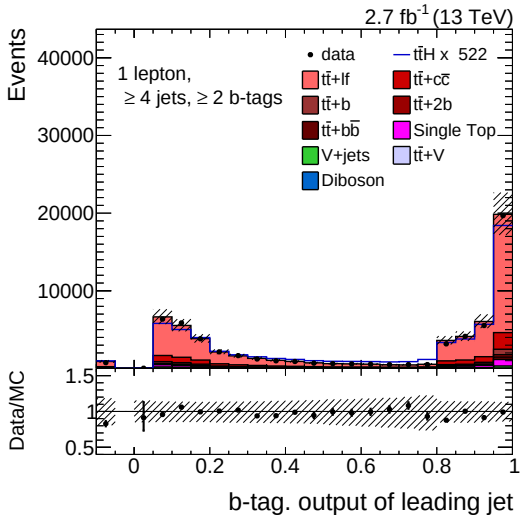

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
• data
tt̄H x 187
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
pT of sixth-leading jet [GeV/c]

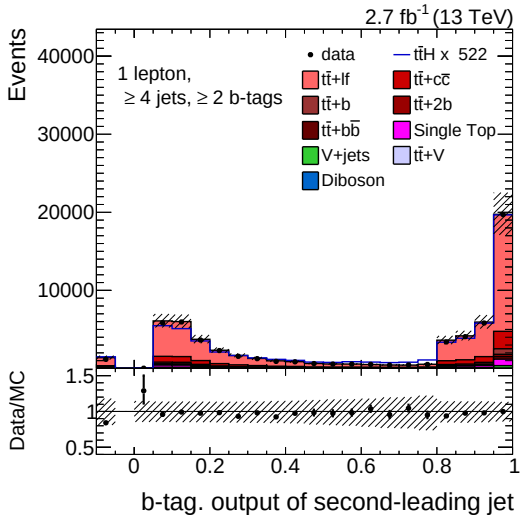

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
• data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
η of leading jet

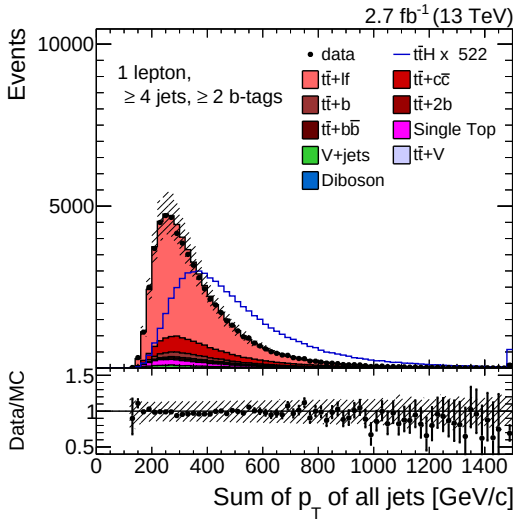

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
• data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
η of second-leading jet

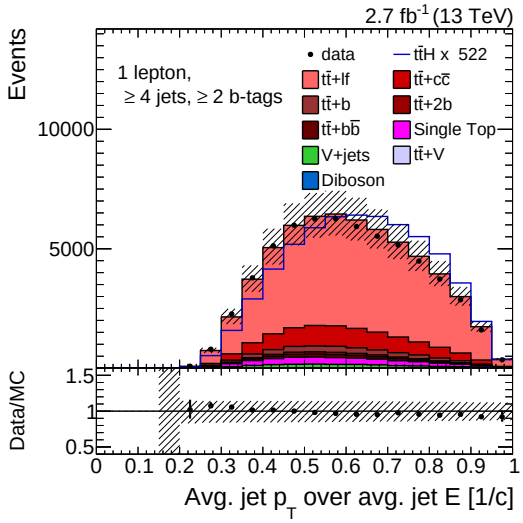

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
ϕ of leading jet

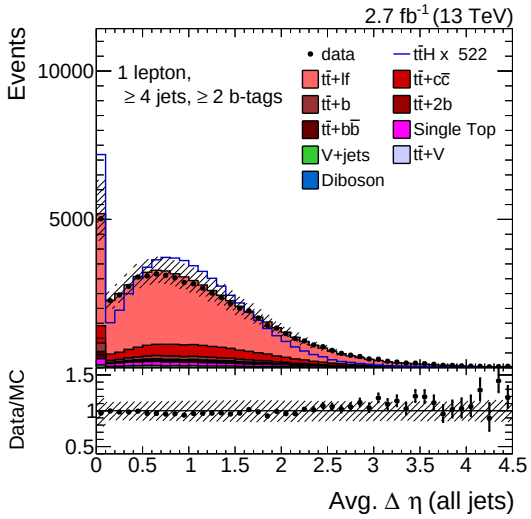

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
ϕ of second-leading jet

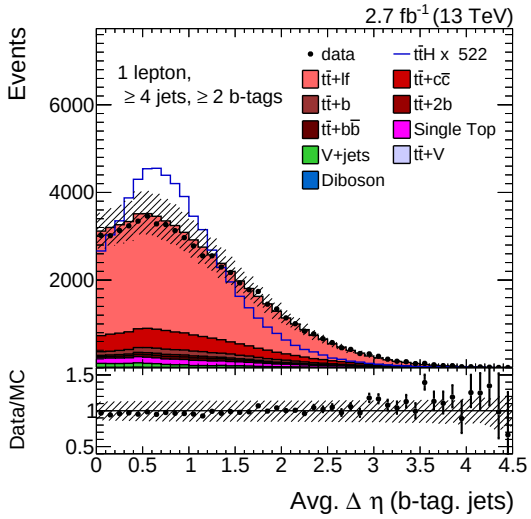

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
b-tag. output of leading jet

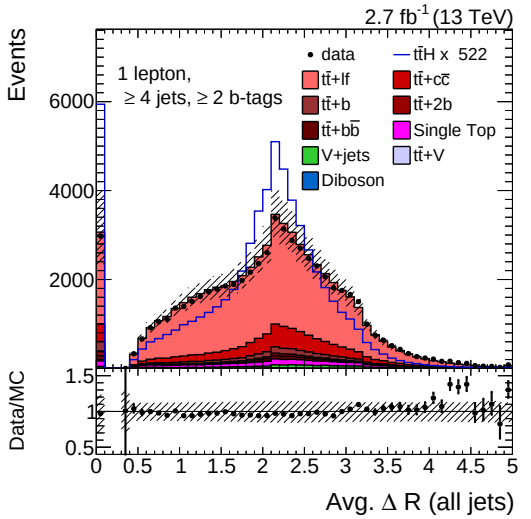

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
b-tag. output of second-leading jet

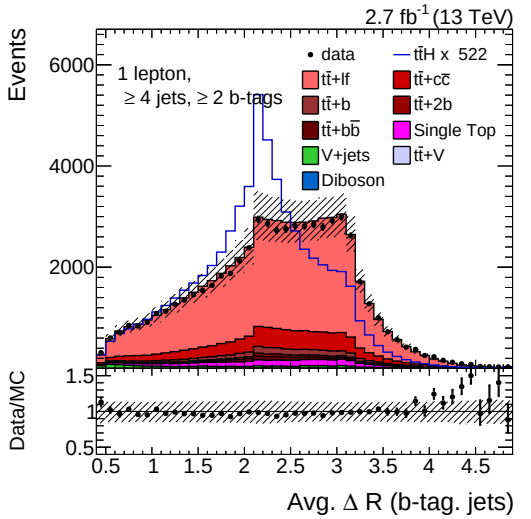

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
Sum of p_T of all jets [GeV/c]

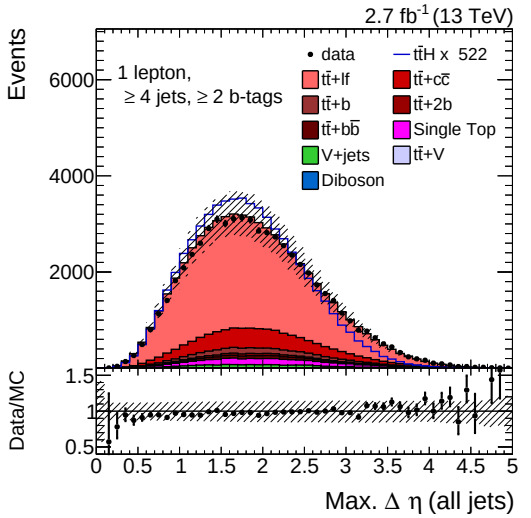

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
Avg. jet p_T over avg. jet E [1/c]

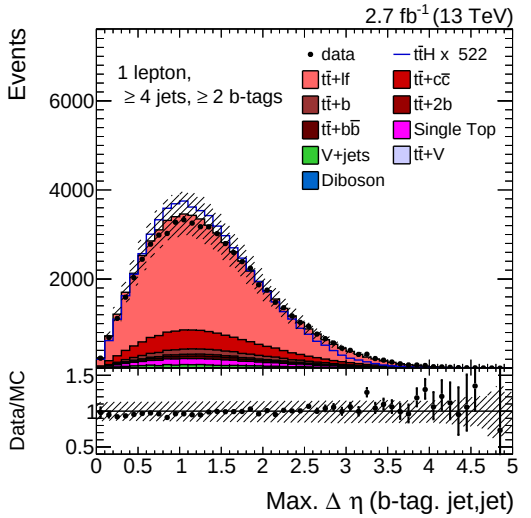

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
Avg. Δ η (all jets)

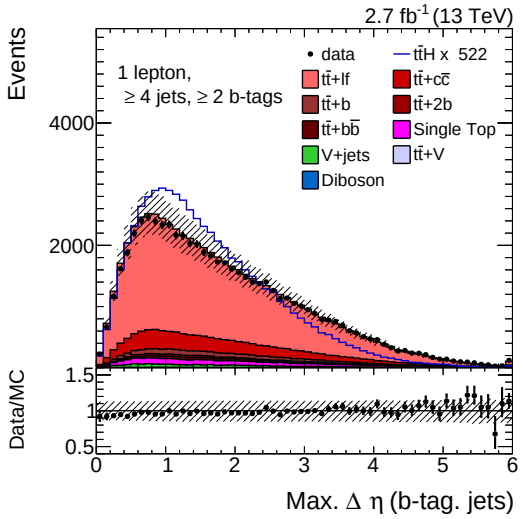

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
Avg. Δ η (b-tag. jets)

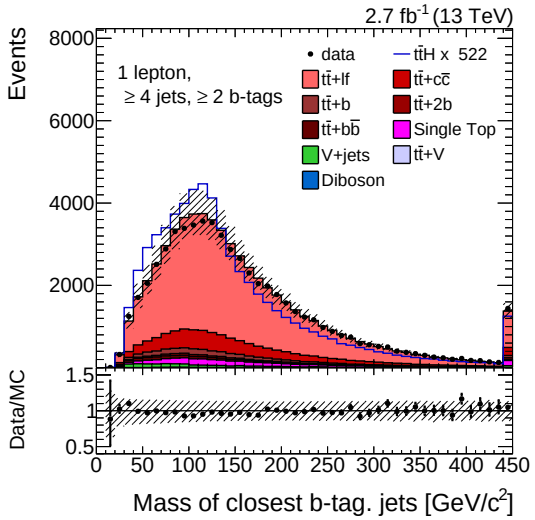

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
Avg. Δ R (all jets)

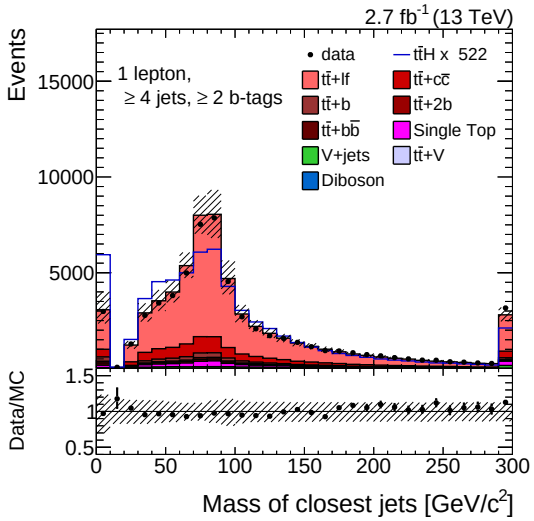

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
Avg. Δ R (b-tag. jets)

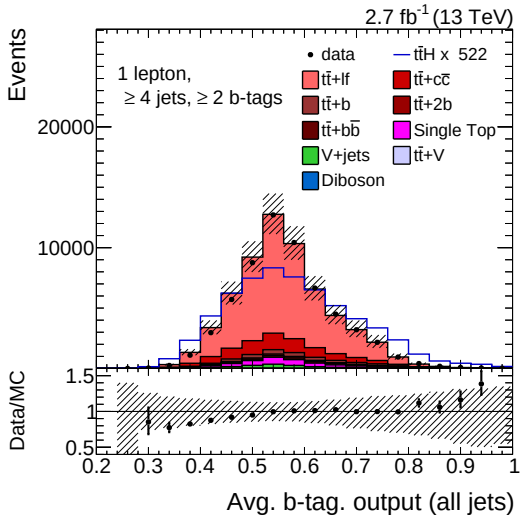

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
Max. Δ η (all jets)

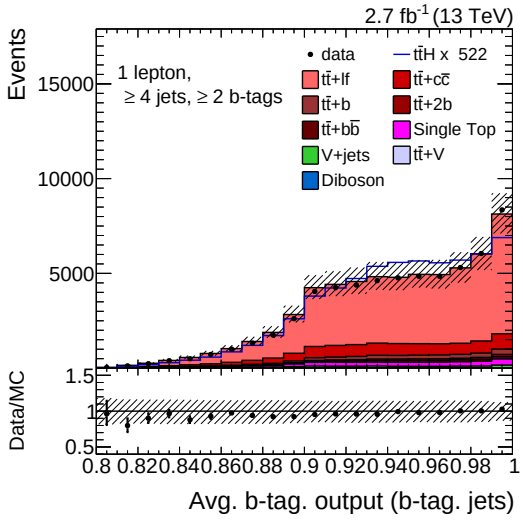

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
Max. Δ η (b-tag. jet,jet)

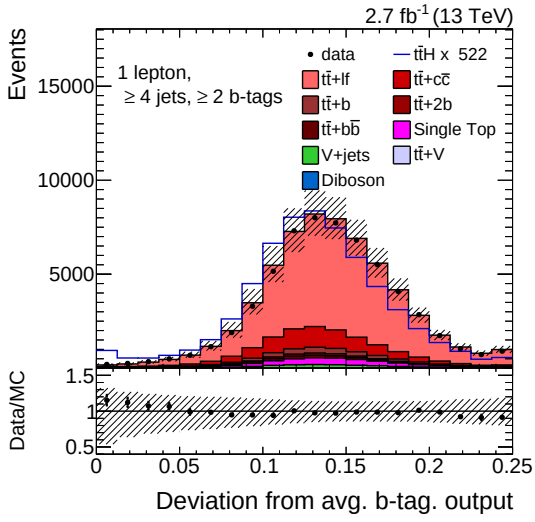

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
Max. Δ η (b-tag. jets)


2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
Mass of closest b-tag. jets [GeV/c2]

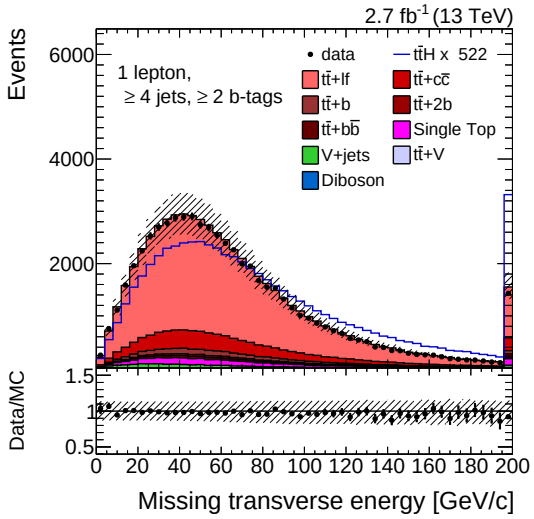

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
Mass of closest jets [GeV/c2]

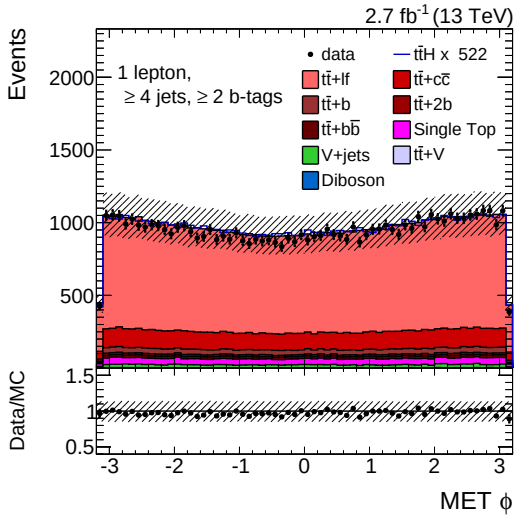

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
Avg. b-tag. output (all jets)


2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
Avg. b-tag. output (b-tag. jets)

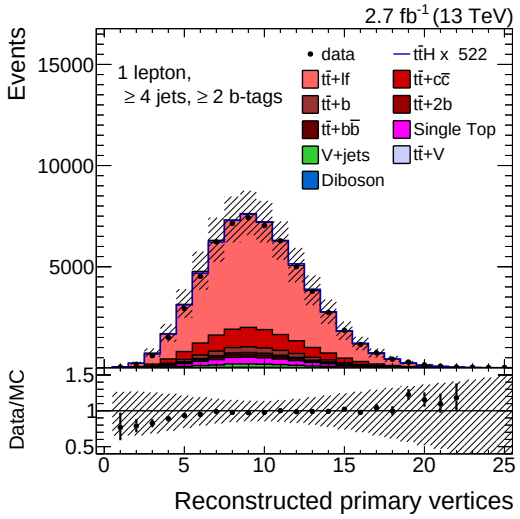

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
data
tt̄H x 522
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
Deviation from avg. b-tag. output

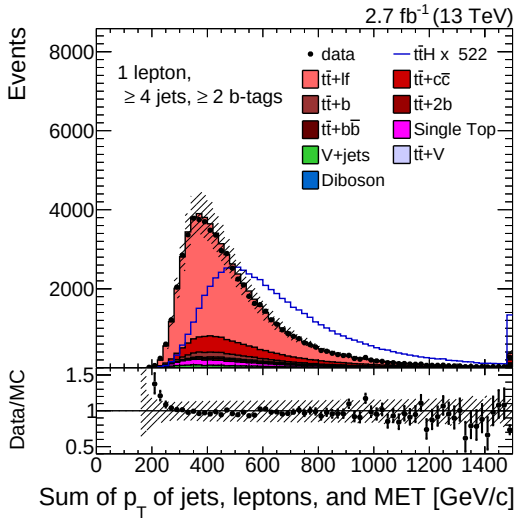
Missing Transverse Energy Variables

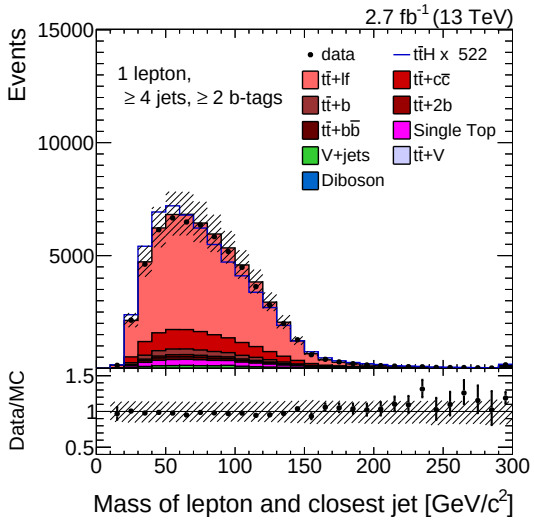




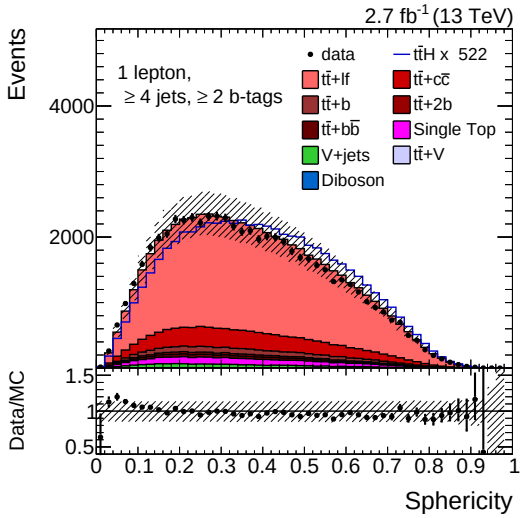
Event Variables

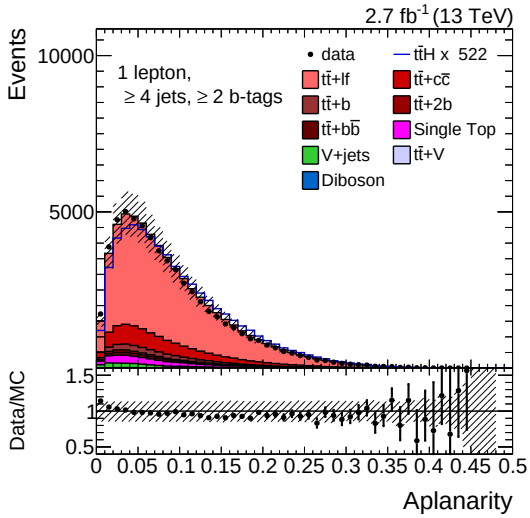


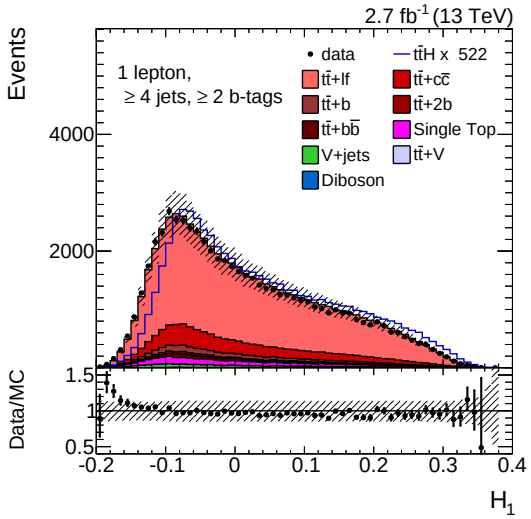


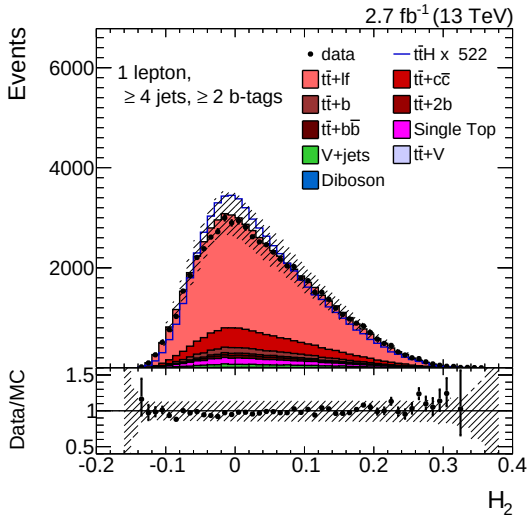


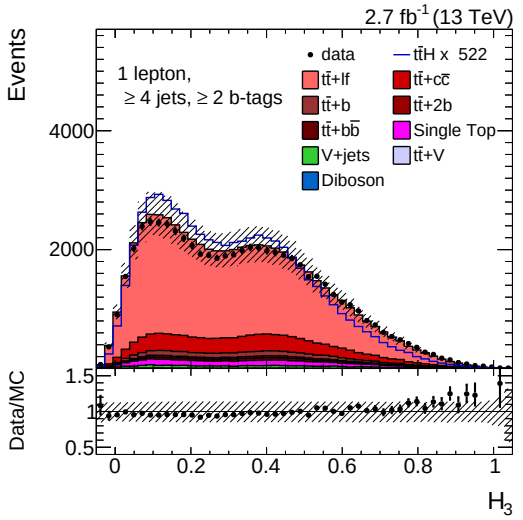
Event Shape Variables

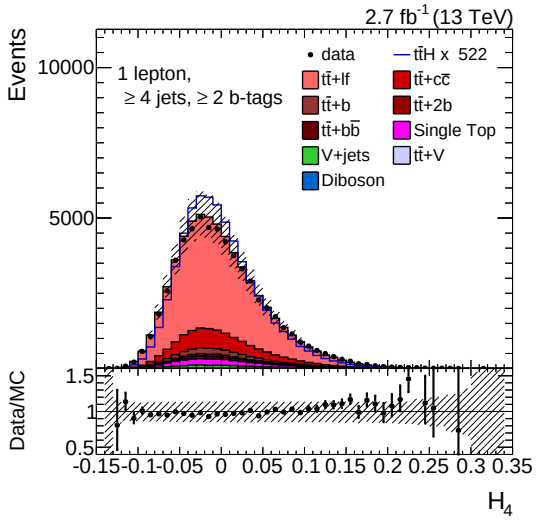






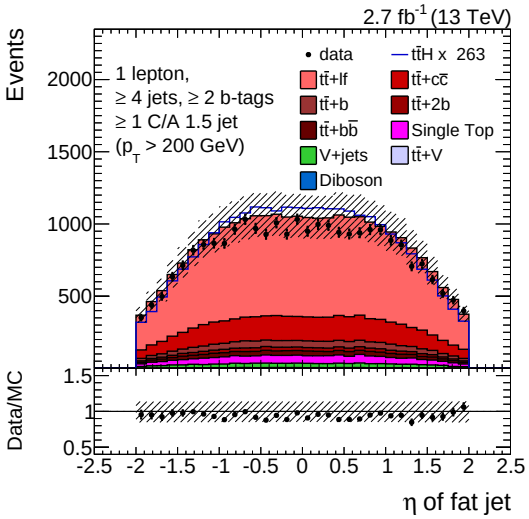




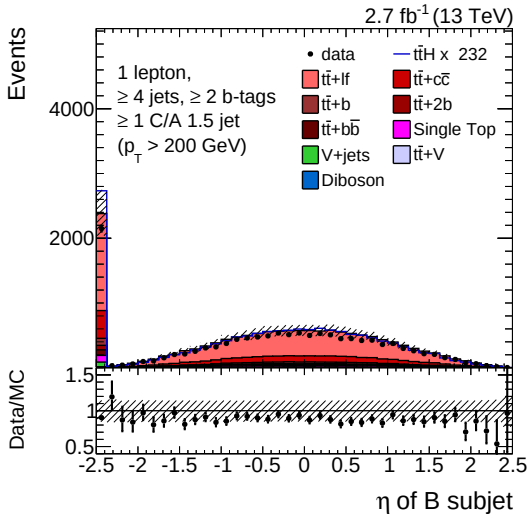


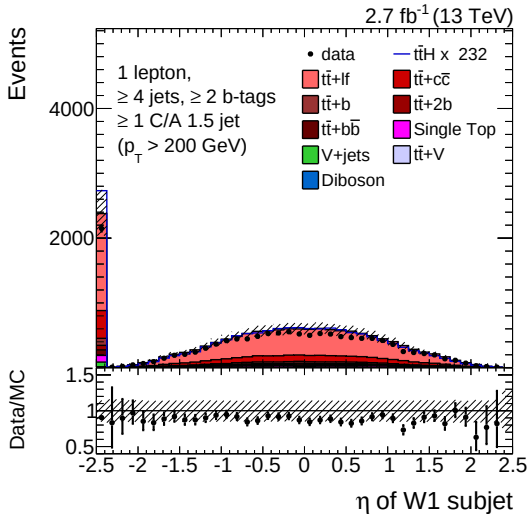
A.1.2. Boosted Validation

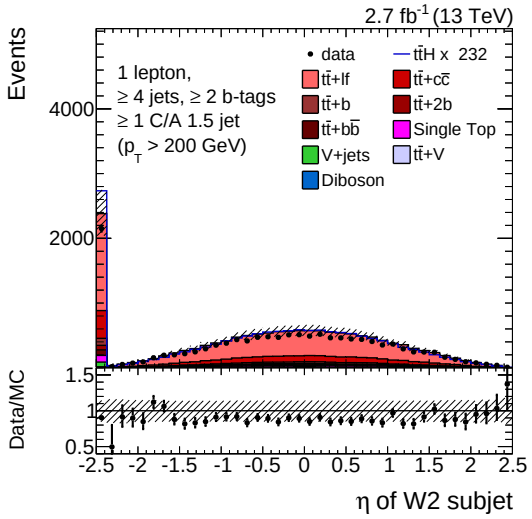
Fat Jet Variables

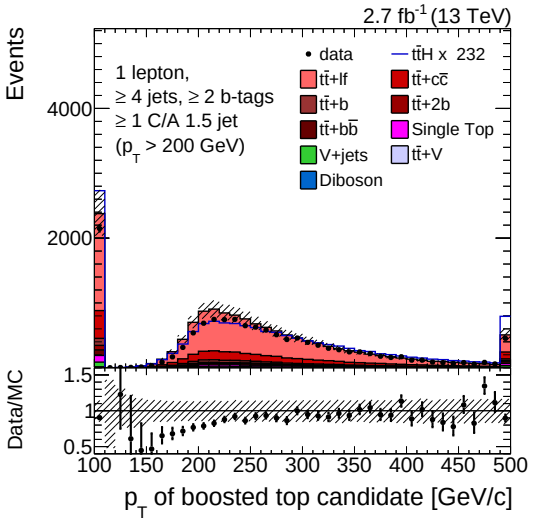


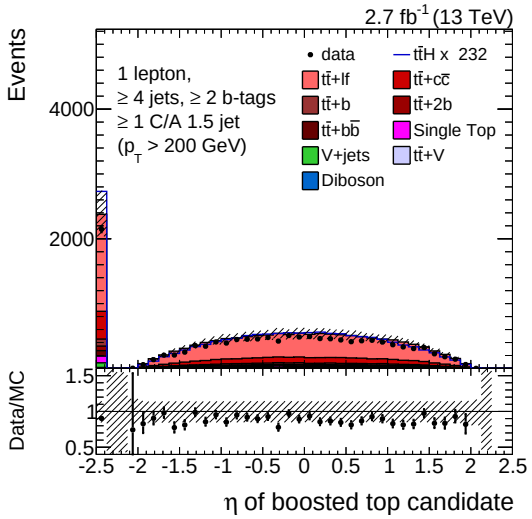
HEPTopTaggerV2 Variables

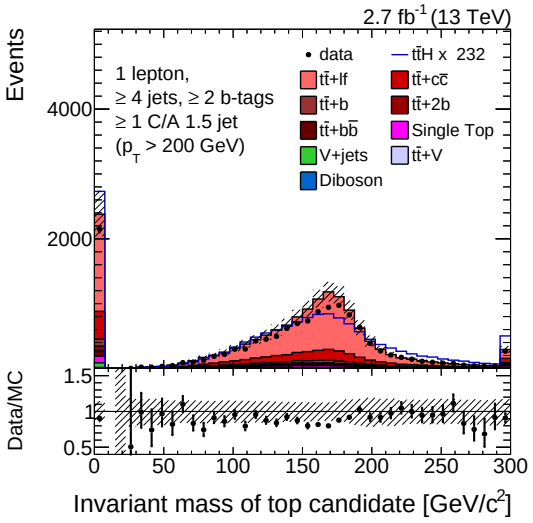


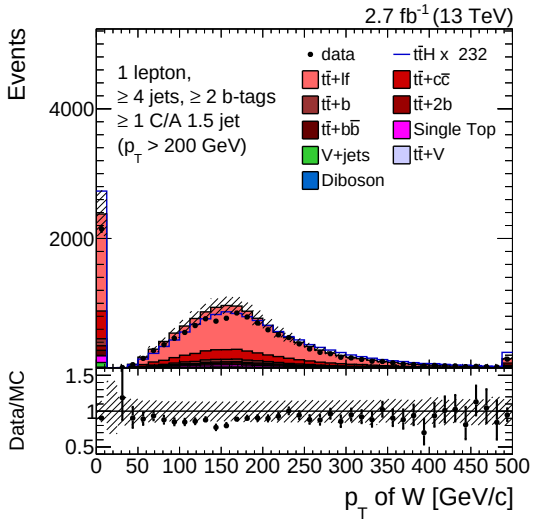


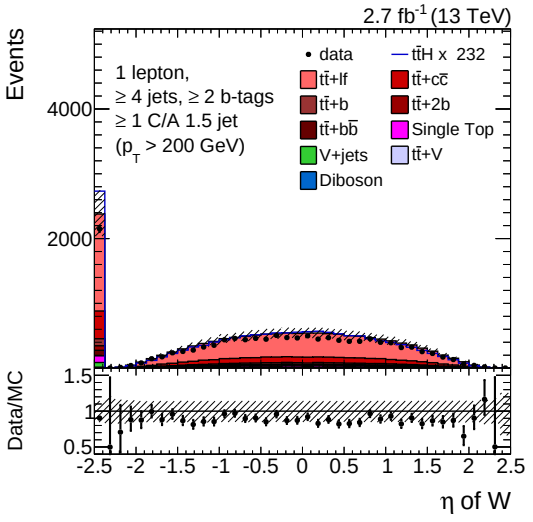

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
≥ 1 C/A 1.5 jet
(pT > 200 GeV)
data
tt̄H x 232
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
η of W2 subjet


2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
≥ 1 C/A 1.5 jet
(pT > 200 GeV)
data
tt̄H x 232
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
pT of boosted top candidate [GeV/c]

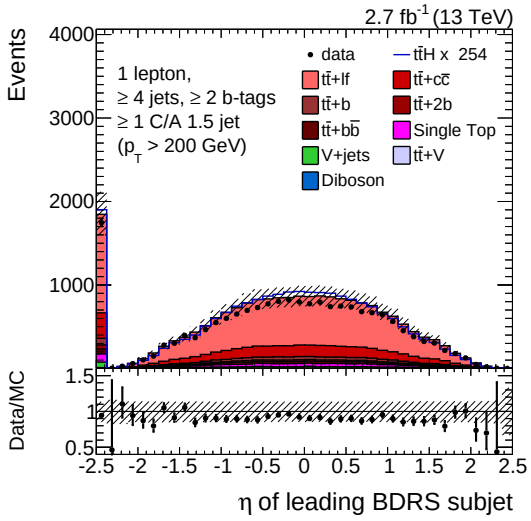

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
≥ 1 C/A 1.5 jet
(pT > 200 GeV)
data
tt̄H x 232
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
η of boosted top candidate

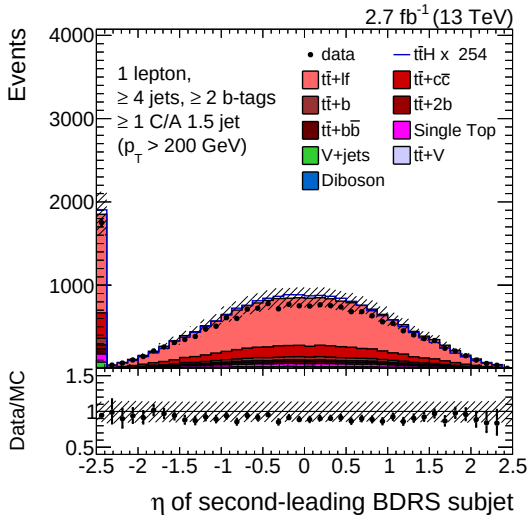

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
≥ 1 C/A 1.5 jet
(pT > 200 GeV)
data
tt̄H x 232
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
Invariant mass of top candidate [GeV/c2]

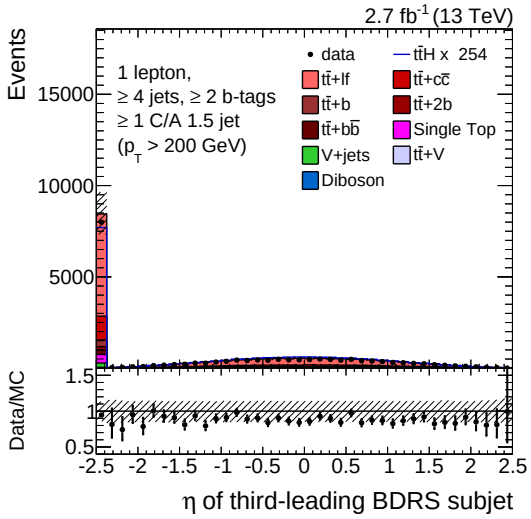

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
≥ 1 C/A 1.5 jet
(pT > 200 GeV)
data
tt̄H x 232
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
pT of W [GeV/c]

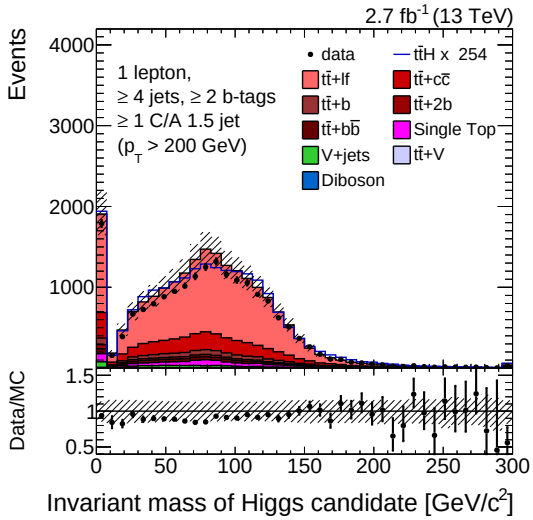

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
≥ 1 C/A 1.5 jet
(pT > 200 GeV)
data
tt̄H x 232
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
η of W

BDRS Variables



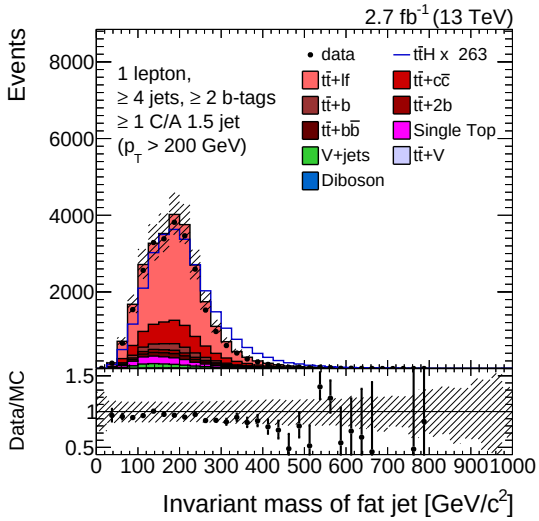


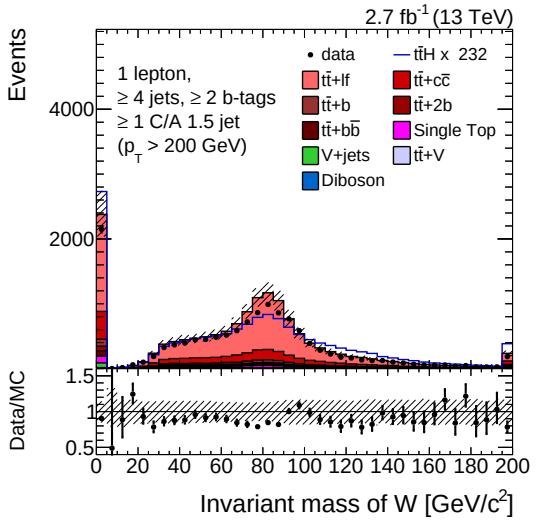


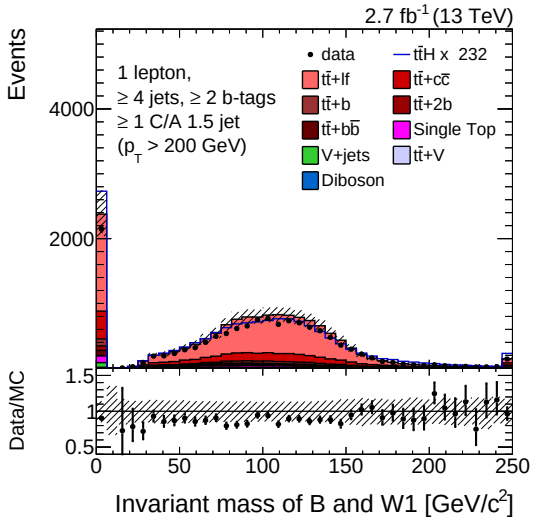


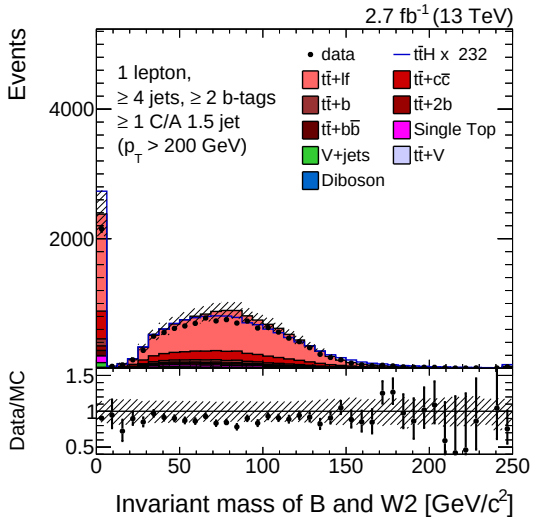
A.2. BDT Input Variables for Boosted Top-quark Identification

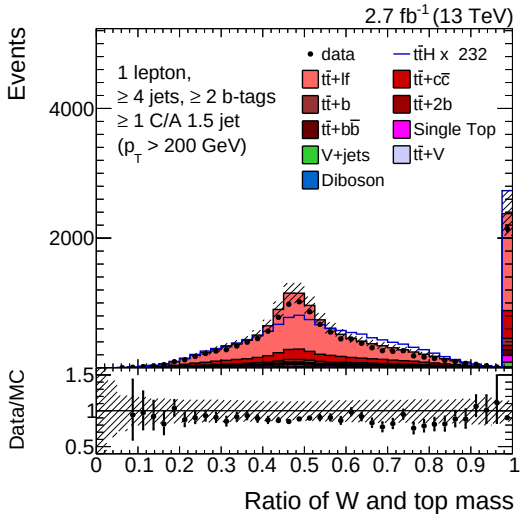
Distributions of input variables used for the training of the boosted top-quark identification BDT are displayed for recorded data and simulation. The simulated background processes are displayed as stacked filled histograms and are scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7$ fb^{-1}. The simulated signal process displayed as a blue line is scaled to the total predicted event yield expected for all background processes. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions.

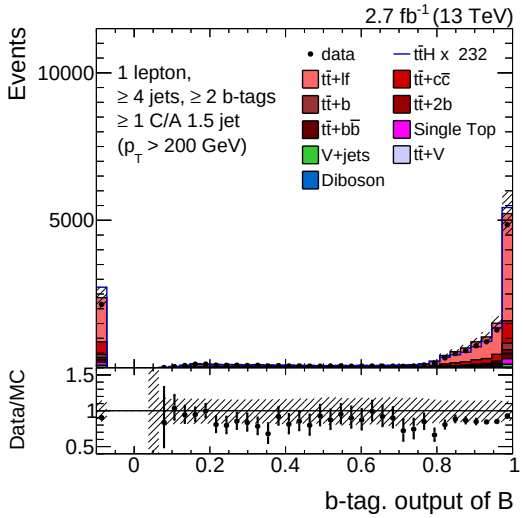


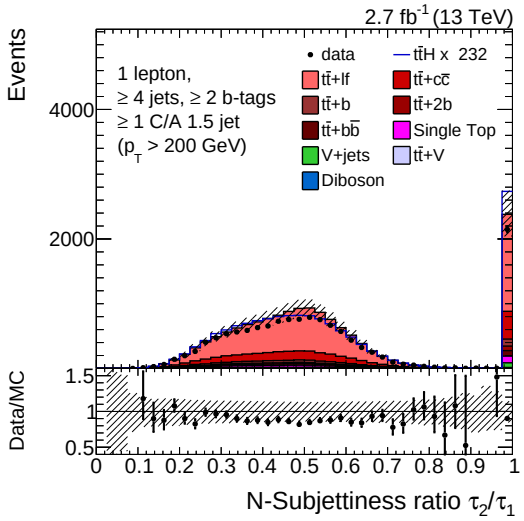


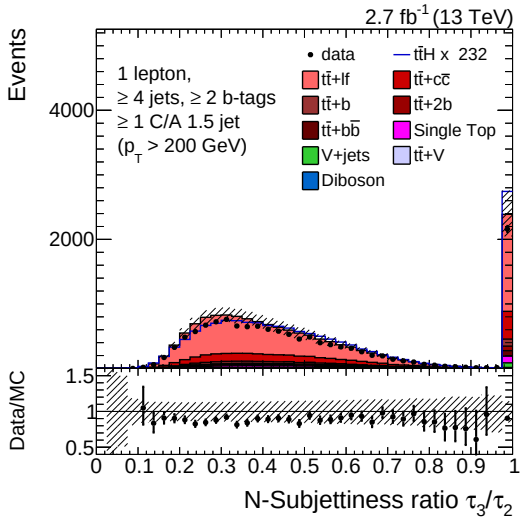


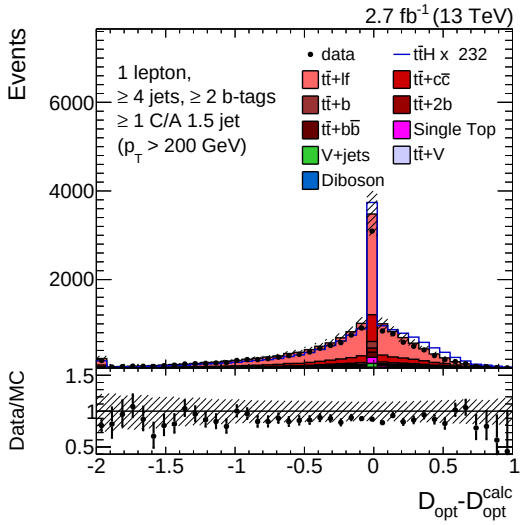



2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
≥ 1 C/A 1.5 jet
(pT > 200 GeV)
data
tt̄H x 232
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
Ratio of W and top mass


2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
≥ 1 C/A 1.5 jet
(pT > 200 GeV)
data
tt̄H x 232
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
b-tag. output of B

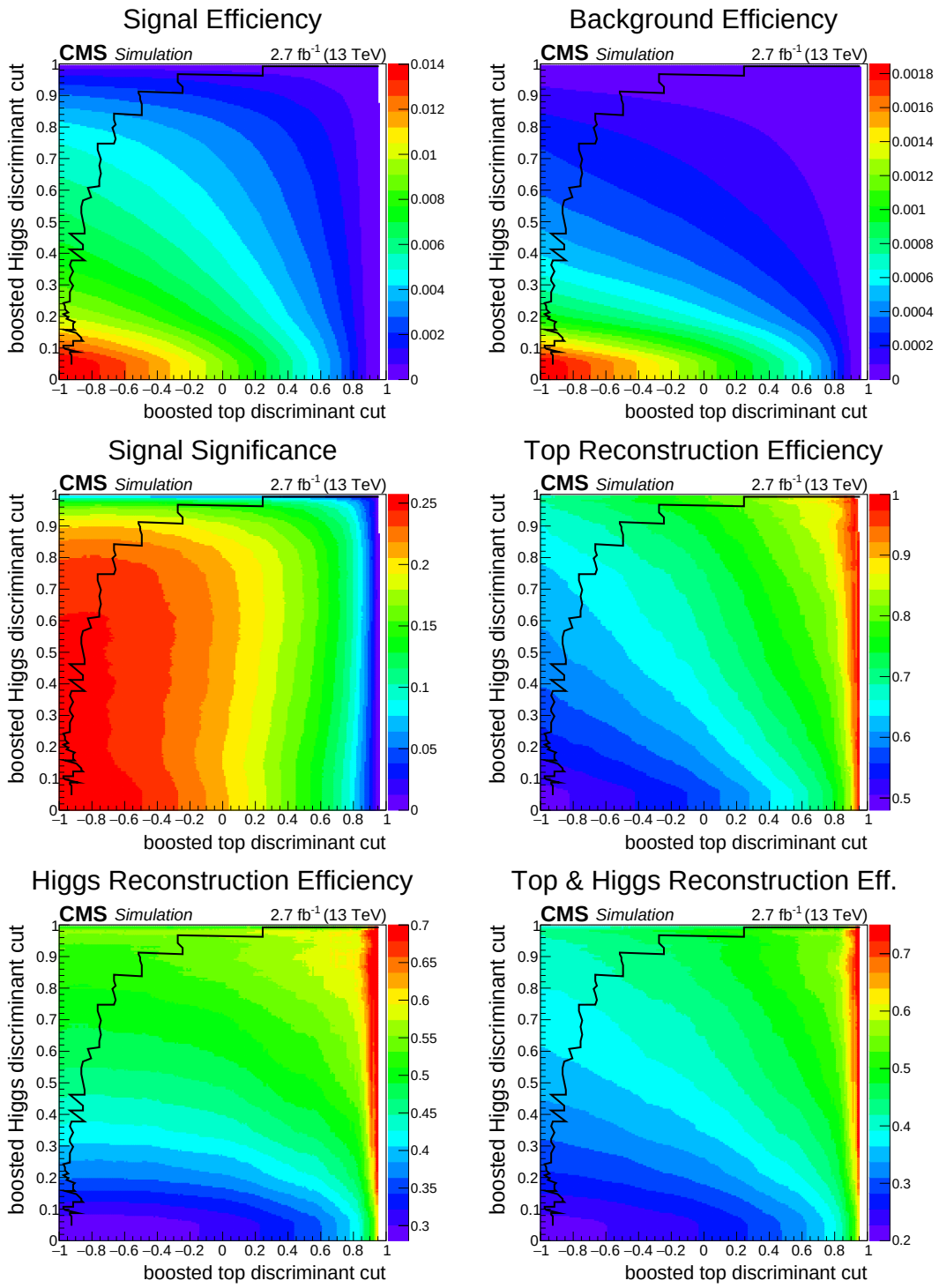

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
≥ 1 C/A 1.5 jet
(pT > 200 GeV)
data
tt̄H x 232
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
N-Subjettiness ratio τ2/τ1


2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
≥ 1 C/A 1.5 jet
(pT > 200 GeV)
data
tt̄H x 232
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
N-Subjettiness ratio τ3/τ2


2.7 fb-1 (13 TeV)
Events
1 lepton,
≥ 4 jets, ≥ 2 b-tags
≥ 1 C/A 1.5 jet
(pT > 200 GeV)
data
tt̄H x 232
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Data/MC
Dopt-Doptcalc

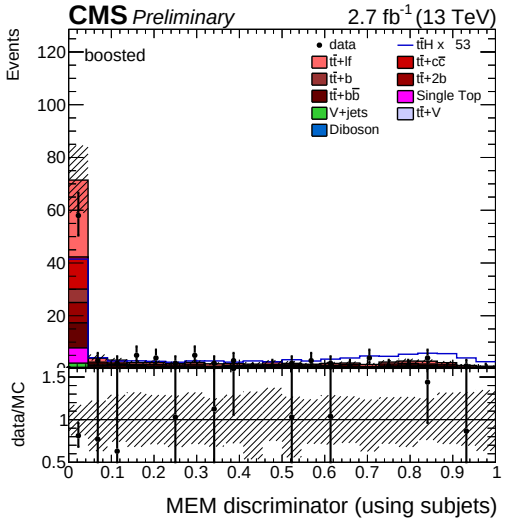
A.3. Boosted-Event Selection

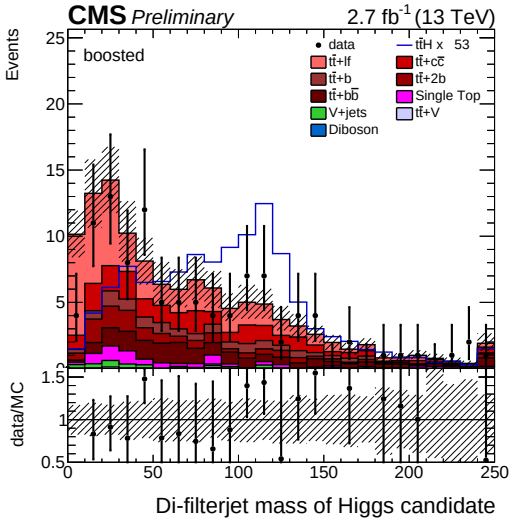
Various quantities for rating the event selection of the boosted analysis category are shown as a function of the event-selection requirements on the boosted Higgs-boson and top-quark discriminant outputs. The quantities are determined using simulated $t\bar{t}(H\to b\bar{b})$ events as signal and semileptonically decaying $t\bar{t}$ events as background. The black line indicates the cut combination with the highest background rejection efficiency for a given signal efficiency.

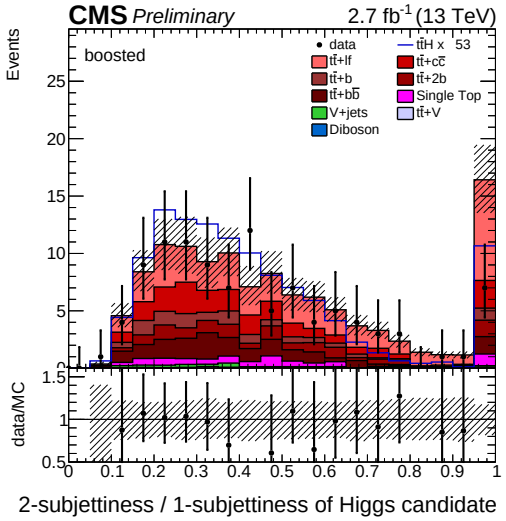


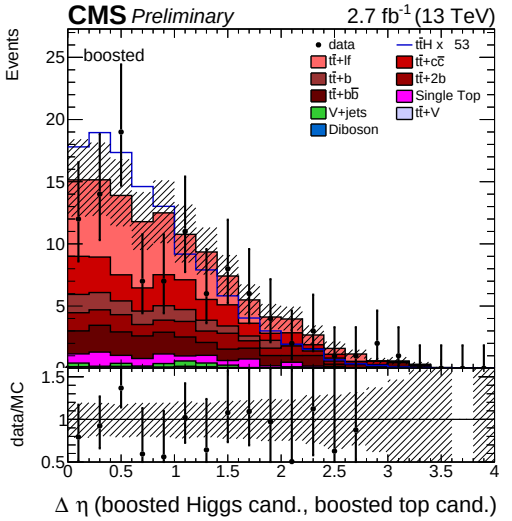
A.4. BDT Input Variables for Boosted Analysis Category

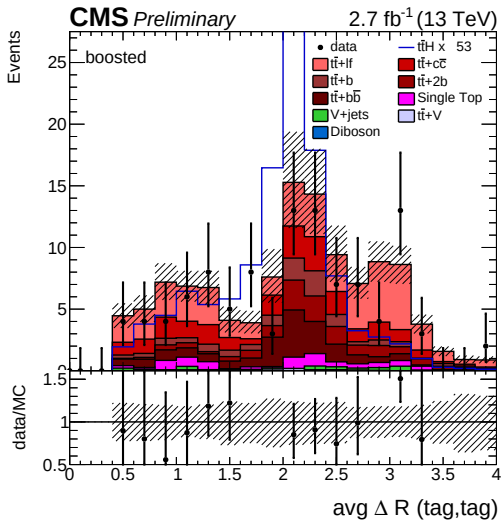
Distributions of input variables used for the training of the final-discriminant BDT in the boosted analysis category are displayed for recorded data and simulation. The simulated background processes are displayed as stacked filled histograms and are scaled to the event yields predicted for an integrated luminosity of $\mathcal{L} = 2.7\ \mathrm{fb}^{-1}$. The simulated signal process displayed as a blue line is scaled to the total predicted event yield expected for all background processes. Distributions of the input variables used for the training of the BDTs in the remaining categories can be found in the publication of the $t\bar{t}(H\to b\bar{b})$ search based on the dataset recorded with the CMS experiment in 2015 [204].

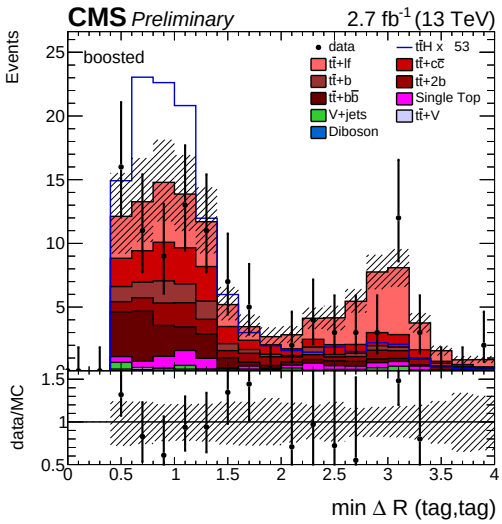


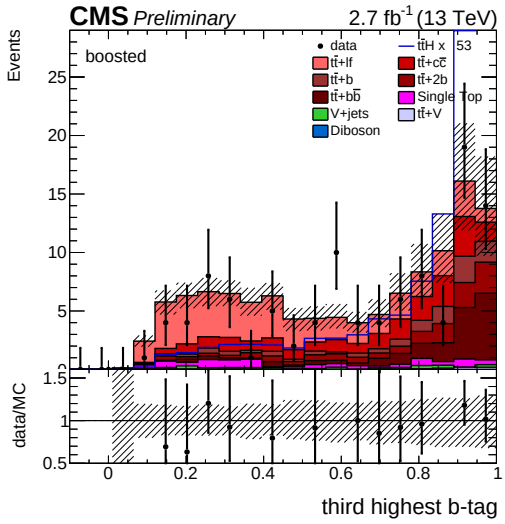


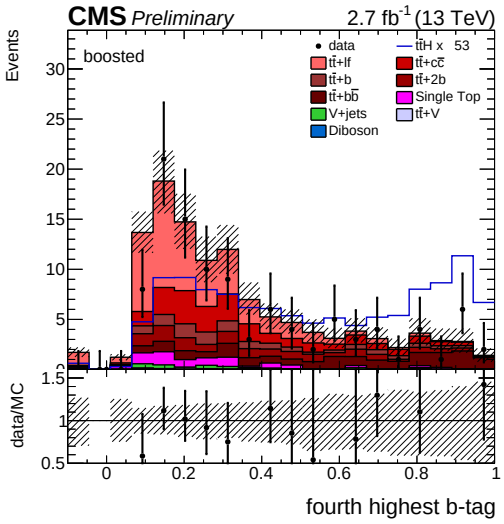


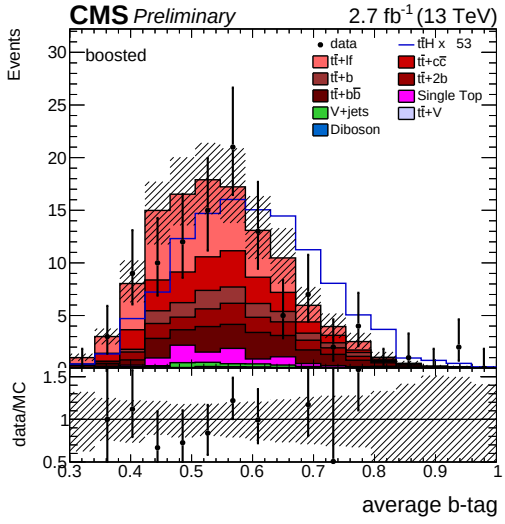


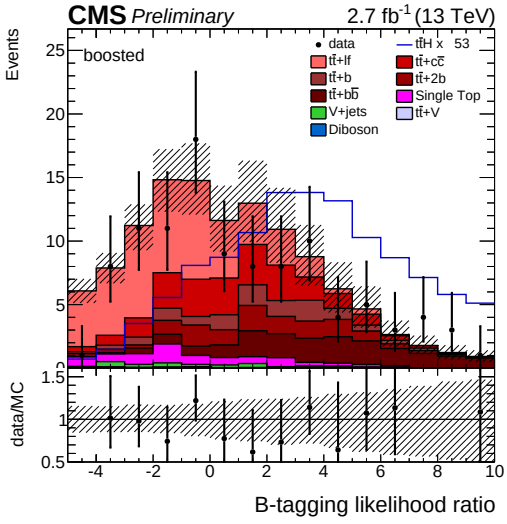

CMS Preliminary
2.7 fb-1 (13 TeV)
boosted
data
tt̄H x 53
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Events
data/MC
avg Δ R (tag,tag)

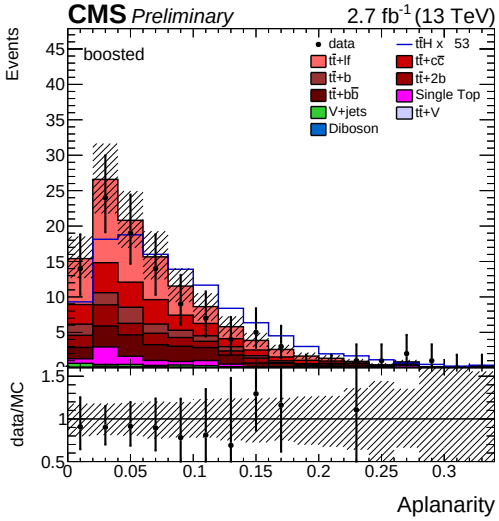

CMS Preliminary
2.7 fb-1 (13 TeV)
boosted
data
tt̄H x 53
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Events
data/MC
min Δ R (tag,tag)


CMS Preliminary
2.7 fb-1 (13 TeV)
boosted
data
tt̄H x 53
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Events
data/MC
third highest b-tag


CMS Preliminary
2.7 fb-1 (13 TeV)
boosted
data
tt̄H x 53
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Events
data/MC
fourth highest b-tag

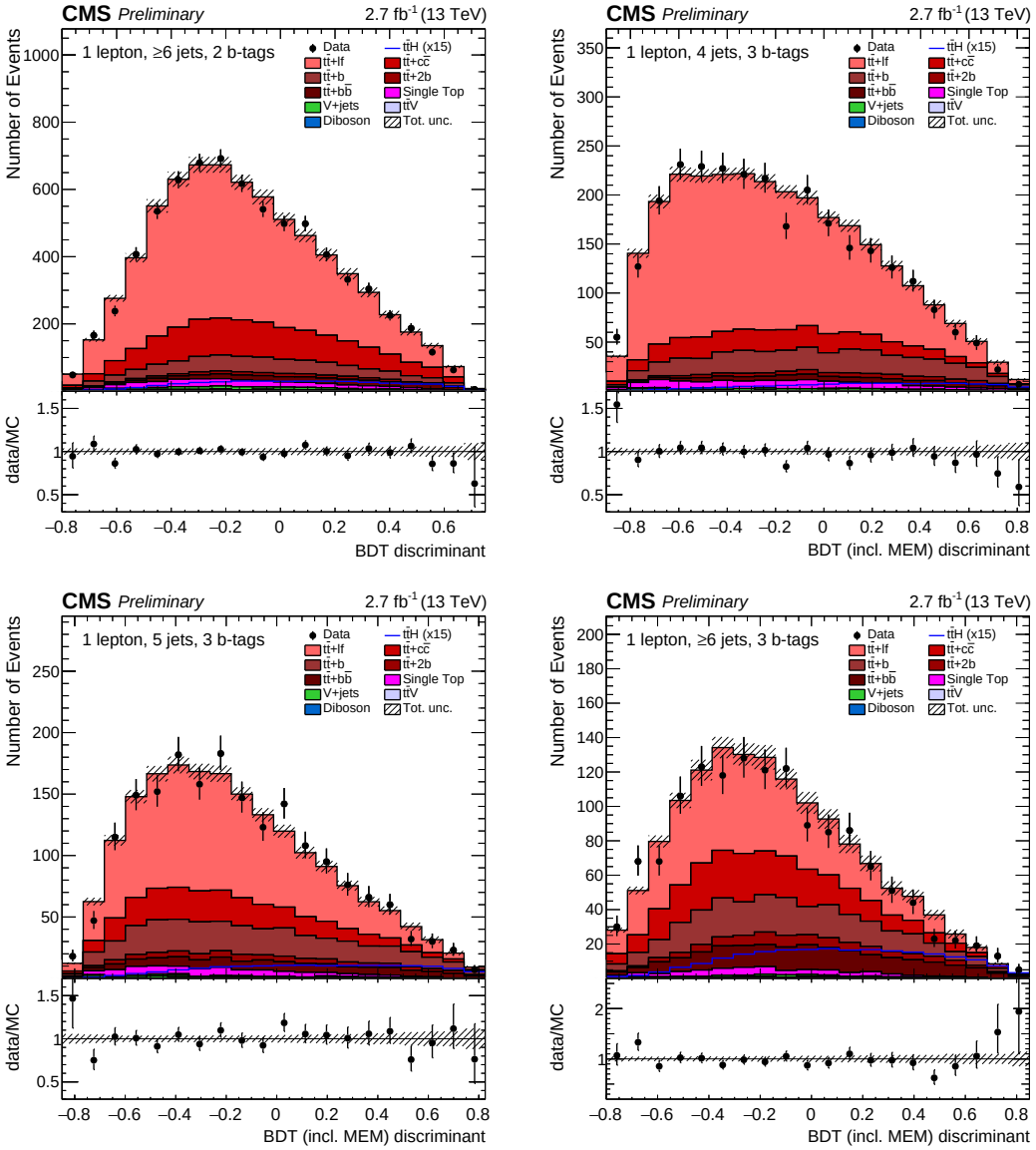

CMS Preliminary
2.7 fb-1 (13 TeV)
boosted
data
tt̄H x 53
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Events
data/MC
average b-tag

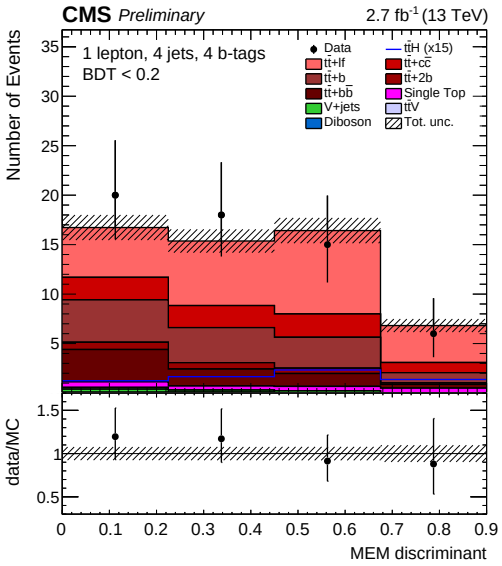

CMS Preliminary
2.7 fb-1 (13 TeV)
boosted
data
tt̄H x 53
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Events
data/MC
B-tagging likelihood ratio

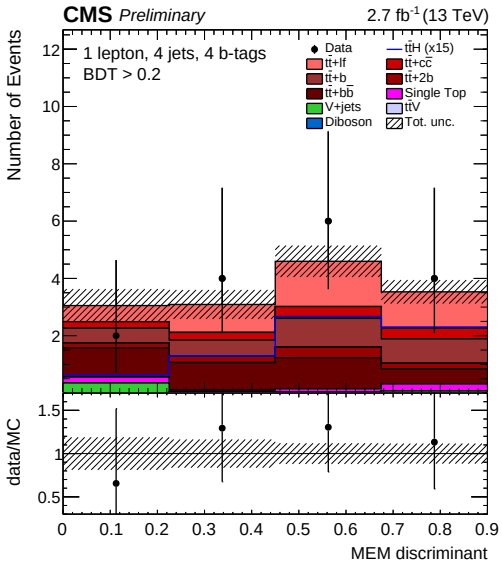

CMS Preliminary
2.7 fb-1 (13 TeV)
boosted
data
tt̄H x 53
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄+V
Diboson
Events
data/MC
Aplanarity

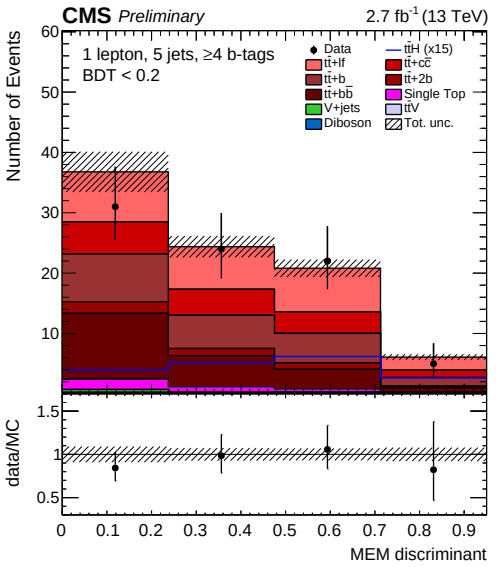
A.5. Post-Fit Final Discriminant Distributions

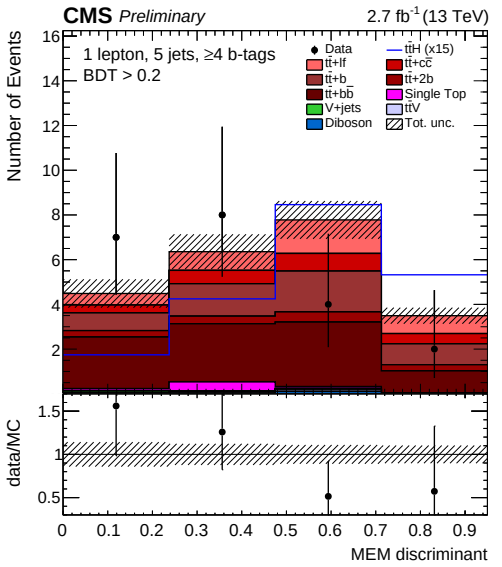
Final-discriminant distributions of recorded data and simulation are shown for each final analysis category including the subcategories provided by the 2D approach after the maximum-likelihood fit. The simulated background processes are scaled to the event yields obtained by the fit and are displayed as stacked filled histograms. The simulated signal process displayed as a blue line is scaled to the predicted event yield expected for an integrated luminosity of $\mathcal{L} = 2.7\ \mathrm{fb}^{-1}$ multiplied with a factor of 15 for better visibility. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions.

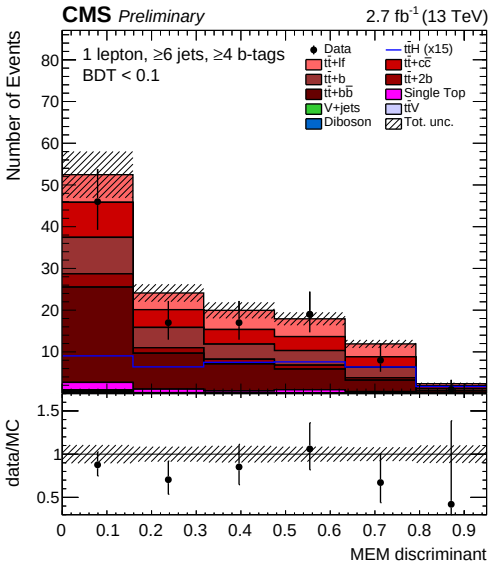


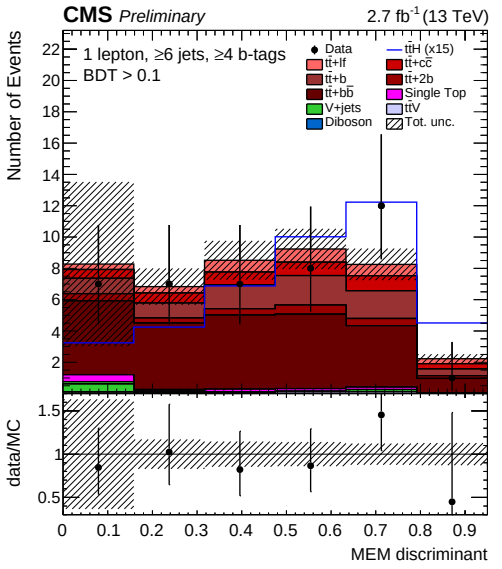

CMS Preliminary
2.7 fb-1 (13 TeV)
1 lepton, 4 jets, 4 b-tags
BDT < 0.2
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b_
tt̄+2b
tt̄+bb̄
Single Top
V+jets
ttV
Diboson
Tot. unc.
Number of Events
data/MC
MEM discriminant

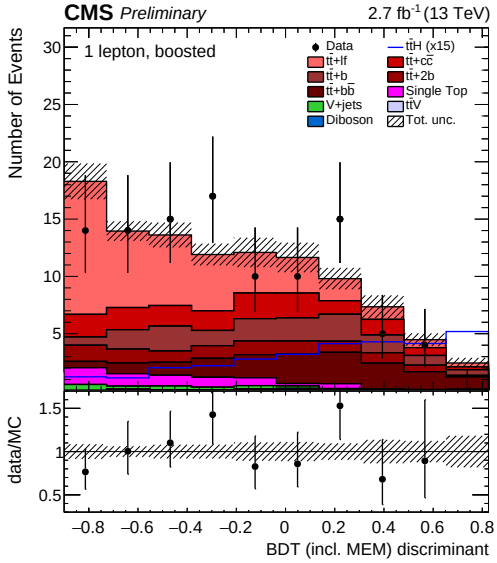

CMS Preliminary
2.7 fb-1 (13 TeV)
1 lepton, 4 jets, 4 b-tags
BDT > 0.2
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b_
tt̄+2b
tt̄+bb̄
Single Top
V+jets
ttV
Diboson
Tot. unc.
Number of Events
data/MC
MEM discriminant


CMS Preliminary
2.7 fb-1 (13 TeV)
1 lepton, 5 jets, ≥4 b-tags
BDT < 0.2
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b_
tt̄+2b
tt̄+bb̄
Single Top
V+jets
ttV
Diboson
Tot. unc.
Number of Events
data/MC
MEM discriminant


CMS Preliminary
2.7 fb-1 (13 TeV)
1 lepton, 5 jets, ≥4 b-tags
BDT > 0.2
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b_
tt̄+2b
tt̄+bb̄
Single Top
V+jets
ttV
Diboson
Tot. unc.
Number of Events
data/MC
MEM discriminant


CMS Preliminary
2.7 fb-1 (13 TeV)
1 lepton, ≥6 jets, ≥4 b-tags
BDT < 0.1
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b_
tt̄+2b
tt̄+bb̄
Single Top
V+jets
ttV
Diboson
Tot. unc.
Number of Events
data/MC
MEM discriminant

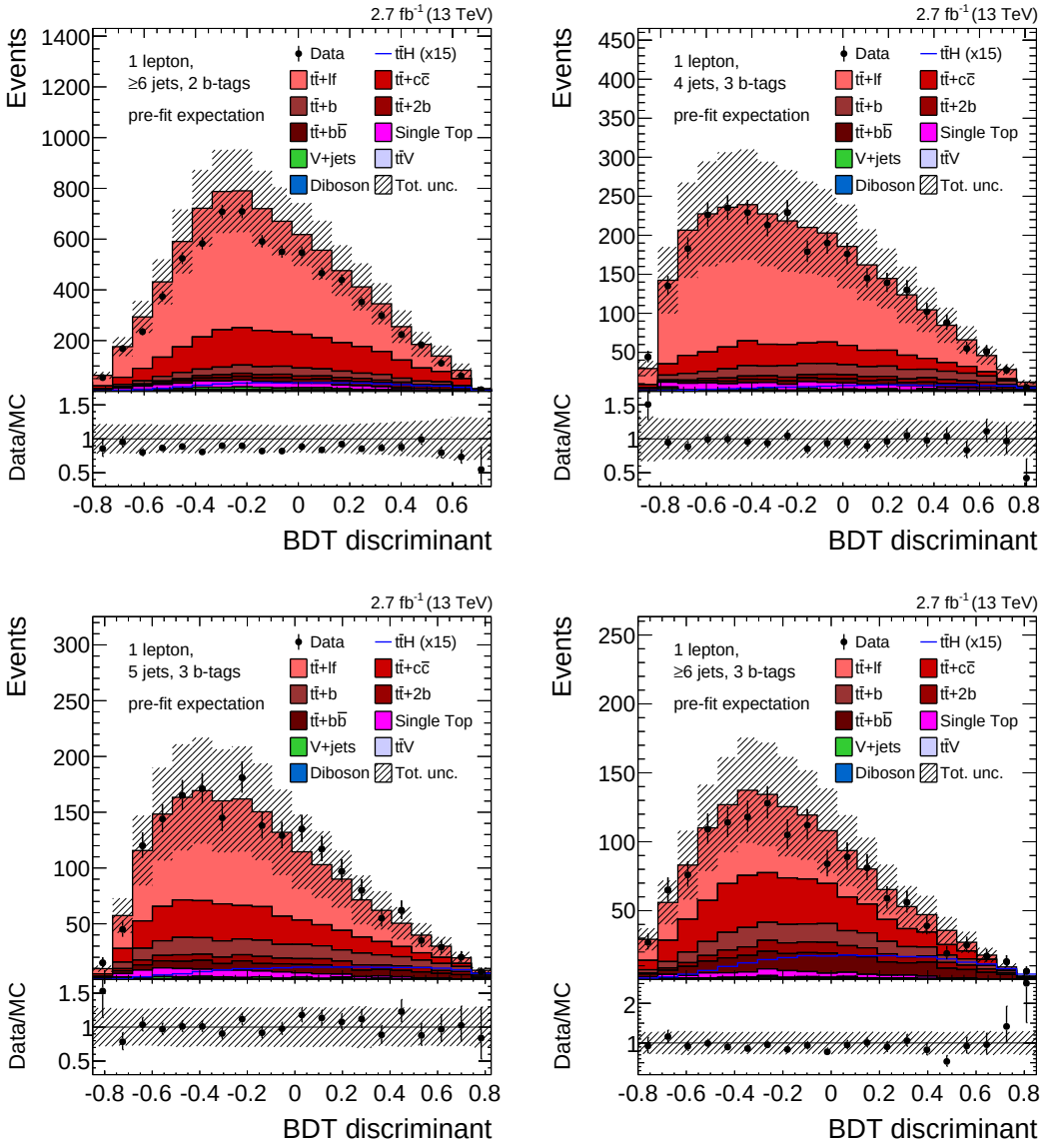

CMS Preliminary
2.7 fb-1 (13 TeV)
1 lepton, ≥6 jets, ≥4 b-tags
BDT > 0.1
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b_
tt̄+2b
tt̄+bb̄
Single Top
V+jets
ttV
Diboson
Tot. unc.
Number of Events
data/MC
MEM discriminant

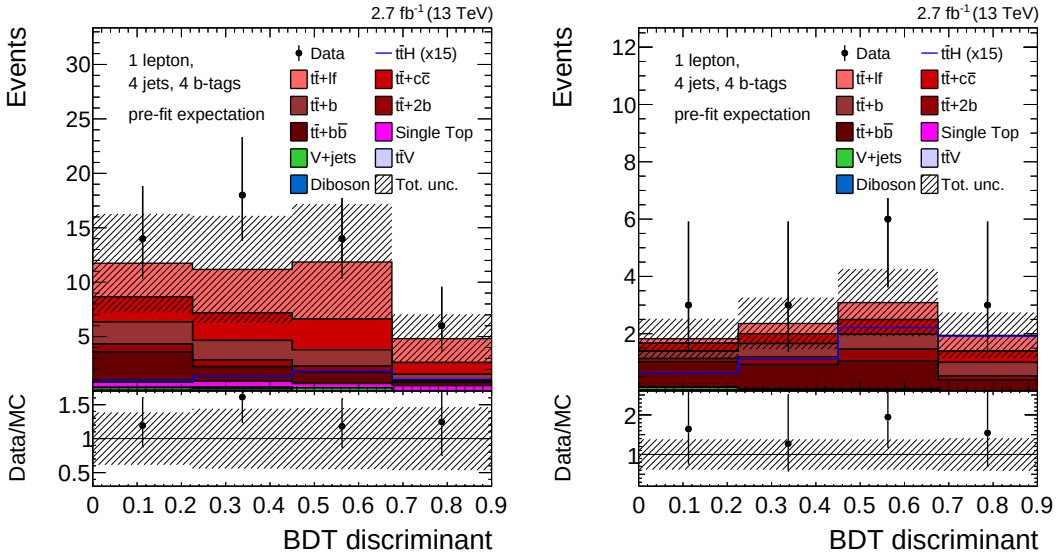

CMS Preliminary
2.7 fb-1 (13 TeV)
1 lepton, boosted
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b_
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄V
Diboson
Tot. unc.
Number of Events
data/MC
BDT (incl. MEM) discriminant

A.6. Single Boosted Signatures

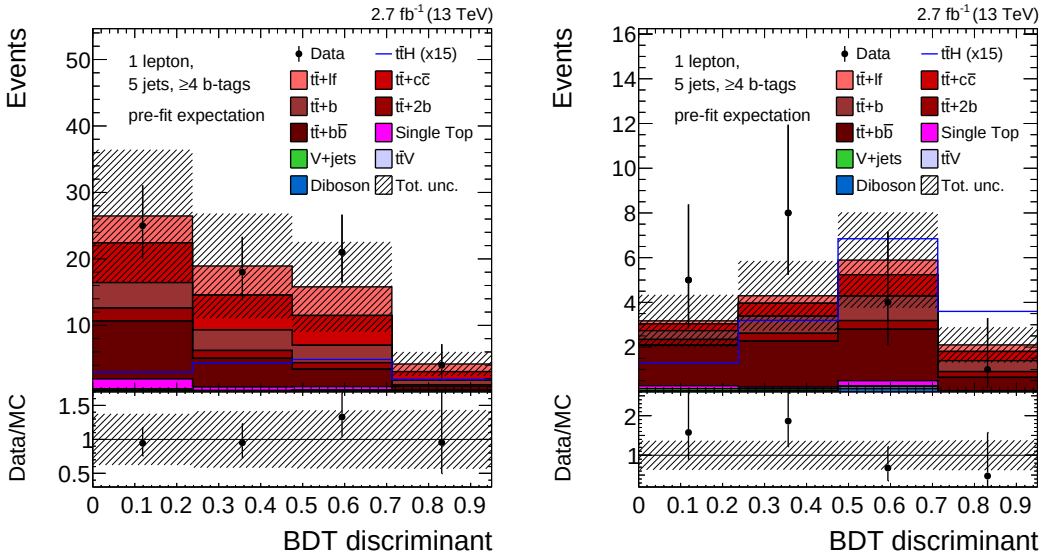
Final-discriminant distributions are displayed for recorded data and simulation in each final category in the single-lepton $t\bar{t}(H\to b\bar{b})$ search including the single boosted analysis categories before and after the maximum-likelihood fit. The simulated background processes are scaled to the event yields obtained by the fit and displayed as stacked filled histograms. The simulated signal process displayed as a blue line is scaled to the predicted event yield expected for an integrated luminosity of $\mathcal{L} = 2.7\ \text{fb}^{-1}$ multiplied with a factor of 15 for better visibility. The shaded band shows the systematic uncertainties on the total event yields of all background processes in each bin of the distributions.

A.6.1. Pre-Fit Distributions

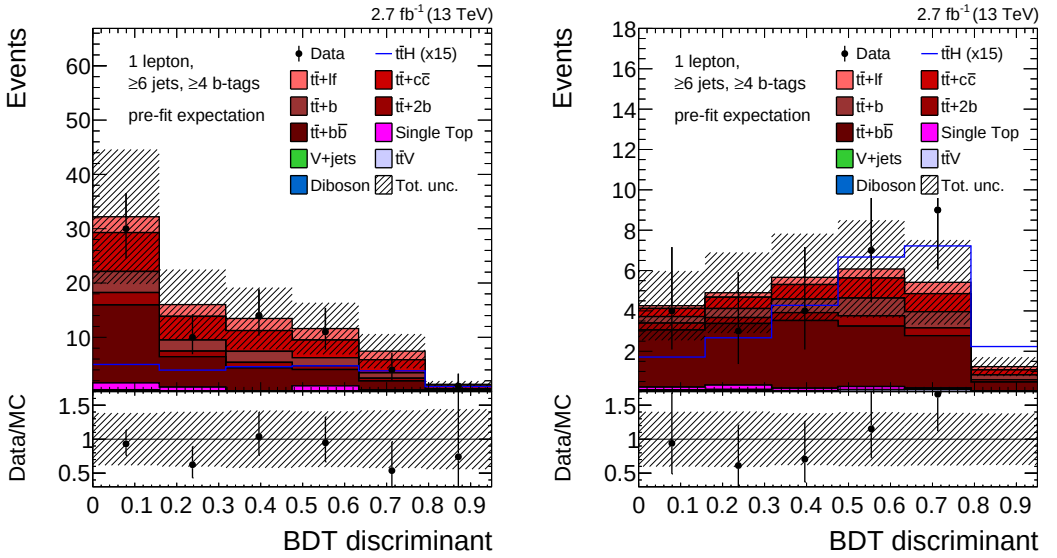



2.7 fb-1 (13 TeV)
Events
1 lepton,
4 jets, 4 b-tags
pre-fit expectation
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄V
Diboson
Tot. unc.
Data/MC
BDT discriminant

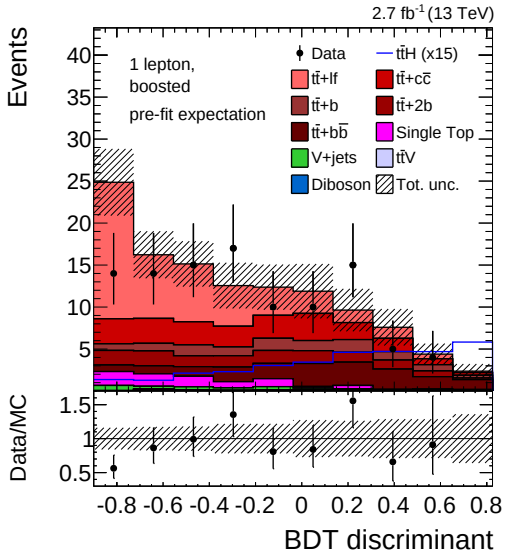
2.7 fb-1 (13 TeV)
Events
1 lepton,
4 jets, 4 b-tags
pre-fit expectation
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄V
Diboson
Tot. unc.
Data/MC
BDT discriminant


2.7 fb-1 (13 TeV)
Events
1 lepton,
5 jets, ≥4 b-tags
pre-fit expectation
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄V
Diboson
Tot. unc.
Data/MC
BDT discriminant

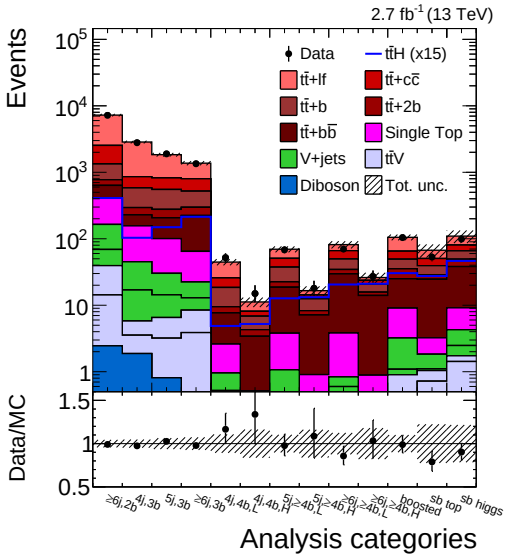
2.7 fb-1 (13 TeV)
Events
1 lepton,
5 jets, ≥4 b-tags
pre-fit expectation
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄V
Diboson
Tot. unc.
Data/MC
BDT discriminant

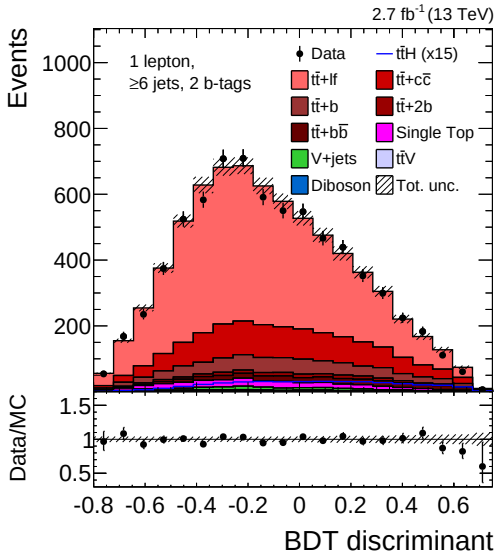

2.7 fb-1 (13 TeV)
Events
1 lepton,
≥6 jets, ≥4 b-tags
pre-fit expectation
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄V
Diboson
Tot. unc.
Data/MC
BDT discriminant

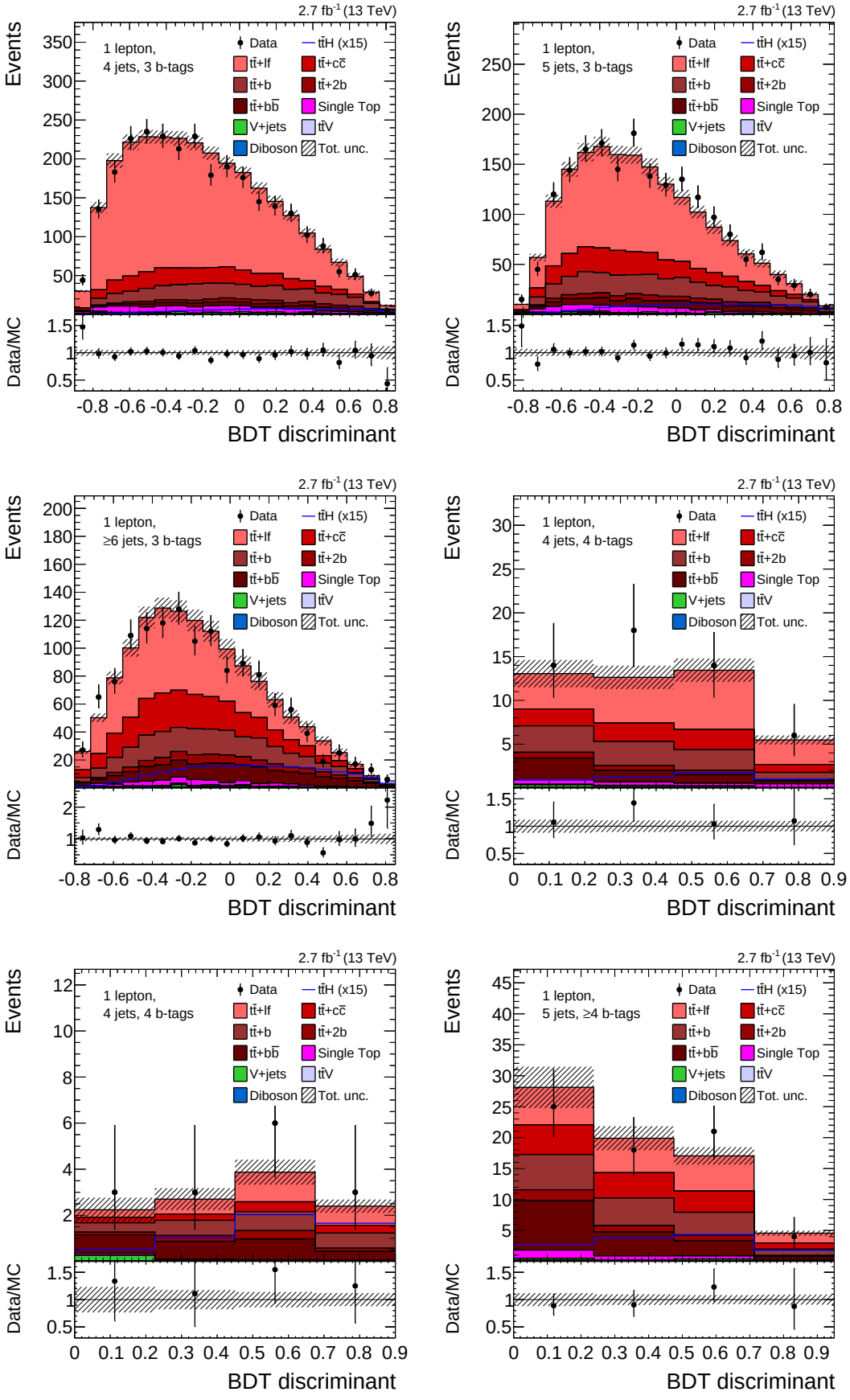
2.7 fb-1 (13 TeV)
Events
1 lepton,
≥6 jets, ≥4 b-tags
pre-fit expectation
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄V
Diboson
Tot. unc.
Data/MC
BDT discriminant

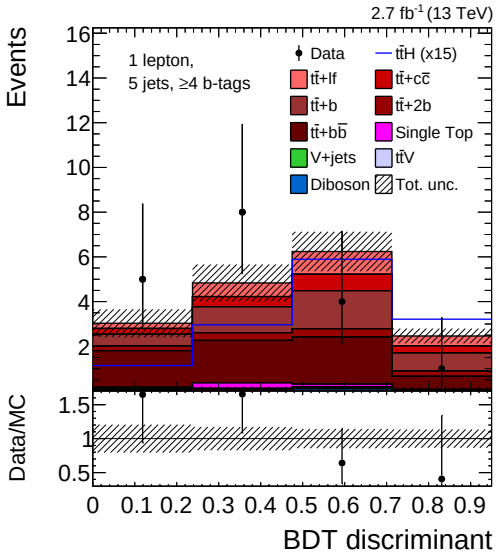


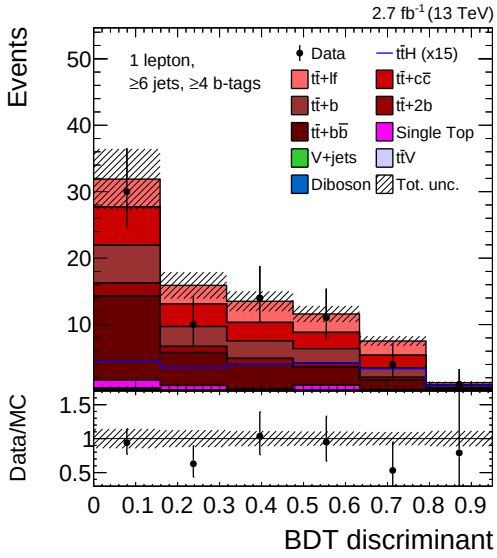
A.6.2. Post-Fit Distributions

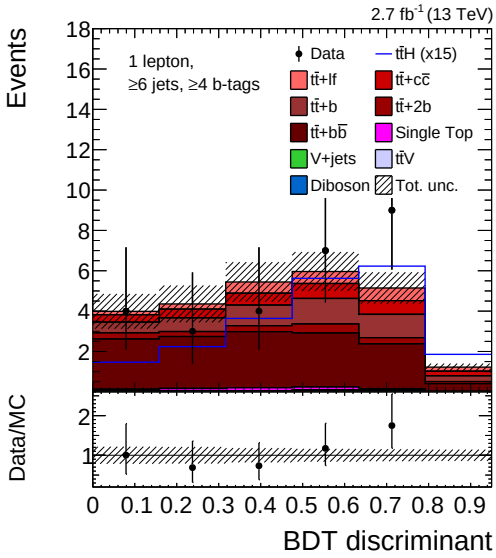


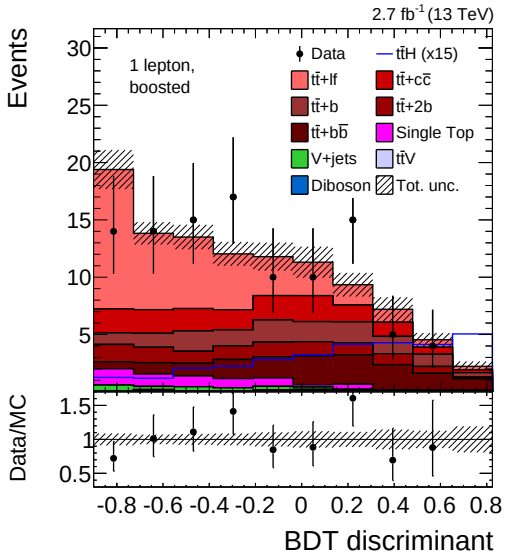


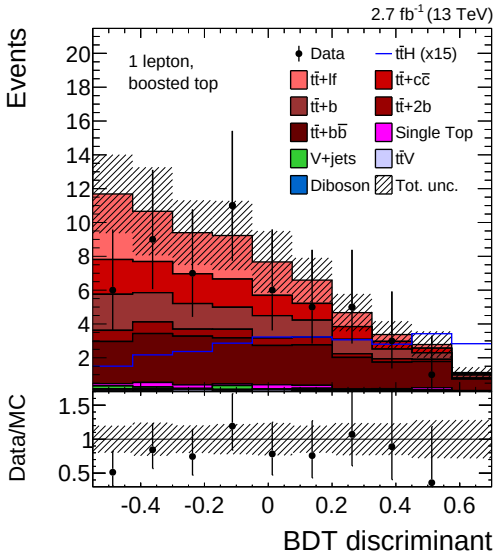

2.7 fb^{-1} (13 TeV)
1 lepton, 4 jets, 3 b-tags
1 lepton, 5 jets, 3 b-tags
1 lepton, ≥6 jets, 3 b-tags
1 lepton, 4 jets, 4 b-tags
1 lepton, 5 jets, ≥4 b-tags
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄V
Diboson
Tot. unc.
Events
Data/MC
BDT discriminant

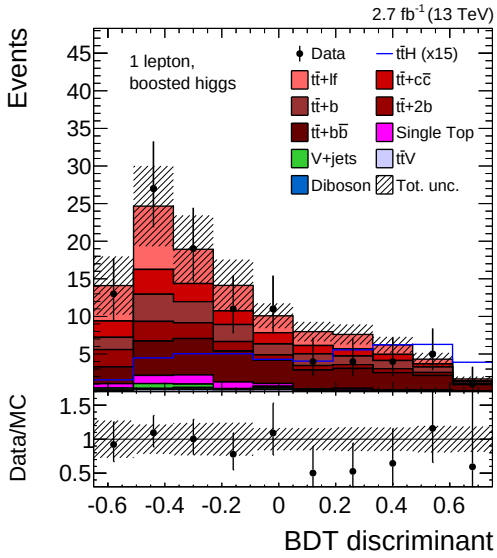

2.7 fb^{-1} (13 TeV)
1 lepton,
5 jets, ≥4 b-tags
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄V
Diboson
Tot. unc.
Events
Data/MC
BDT discriminant


2.7 fb^{-1} (13 TeV)
1 lepton,
≥6 jets, ≥4 b-tags
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄V
Diboson
Tot. unc.
Events
Data/MC
BDT discriminant


2.7 fb^{-1} (13 TeV)
1 lepton,
≥6 jets, ≥4 b-tags
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄V
Diboson
Tot. unc.
Events
Data/MC
BDT discriminant


2.7 fb^{-1} (13 TeV)
1 lepton,
boosted
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄V
Diboson
Tot. unc.
Events
Data/MC
BDT discriminant


2.7 fb^{-1} (13 TeV)
1 lepton,
boosted top
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄V
Diboson
Tot. unc.
Events
Data/MC
BDT discriminant


2.7 fb^{-1} (13 TeV)
1 lepton,
boosted higgs
Data
tt̄H (x15)
tt̄+lf
tt̄+cc̄
tt̄+b
tt̄+2b
tt̄+bb̄
Single Top
V+jets
tt̄V
Diboson
Tot. unc.
Events
Data/MC
BDT discriminant

Bibliography

[1] Aad, G. et al. "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).

[2] Chatrchyan, S. et al. "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC". *Phys. Lett.*, B716:30–61 (2012).

[3] Plehn, T., Salam, G. P., and Spannowsky, M. "Fat Jets for a Light Higgs". *Phys. Rev. Lett.*, 104:111801 (2010).

[4] Anders, C., et al. "Benchmarking an even better top tagger algorithm". *Phys. Rev.*, D89:074047 (2014).

[5] Kasieczka, G., et al. "Resonance Searches with an Updated Top Tagger". *JHEP*, 06:203 (2015).

[6] Butterworth, J. M., et al. "Jet substructure as a new Higgs search channel at the LHC". *Phys. Rev. Lett.*, 100:242001 (2008).

[7] CMS Collaboration. "Search for $t\bar{t}H$ production in the $H \to b\bar{b}$ decay channel with $\sqrt{s} = 13\,\text{TeV}$ pp collisions at the CMS experiment". *CMS-PAS-HIG-16-004* (2016).

[8] Pauli, W. "Relativistic Field Theories of Elementary Particles". *Rev. Mod. Phys.*, 13:203–232 (1941).

[9] Abe, F. et al. "Observation of top quark production in $\bar{p}p$ collisions". *Phys. Rev. Lett.*, 74:2626–2631 (1995).

[10] Abachi, S. et al. "Observation of the top quark". *Phys. Rev. Lett.*, 74:2632–2637 (1995).

[11] Noether, E. "Invariant Variation Problems". *Gott. Nachr.*, 1918:235–257 (1918). [Transp. Theory Statist. Phys., 1:186-207 (1971)].

[12] Tomonaga, S. "On a relativistically invariant formulation of the quantum theory of wave fields". *Prog. Theor. Phys.*, 1:27–42 (1946).

[13] Schwinger, J. S. "On Quantum electrodynamics and the magnetic moment of the electron". *Phys. Rev.*, 73:416–417 (1948).

[14] Schwinger, J. S. "Quantum electrodynamics. I A covariant formulation". *Phys. Rev.*, 74:1439 (1948).

[15] Feynman, R. P. "Space - time approach to quantum electrodynamics". *Phys. Rev.*, 76:769–789 (1949).

[16] Feynman, R. P. "The Theory of positrons". *Phys. Rev.*, 76:749–759 (1949).

[17] Feynman, R. P. "Mathematical formulation of the quantum theory of electromagnetic interaction". *Phys. Rev.*, 80:440–457 (1950).

[18] Fritzsch, H., Gell-Mann, M., and Leutwyler, H. "Advantages of the Color Octet Gluon Picture". *Phys. Lett.*, B47:365–368 (1973).

[19] Gross, D. J. and Wilczek, F. "Asymptotically Free Gauge Theories. 1". *Phys. Rev.*, D8:3633–3652 (1973).

[20] Gross, D. J. and Wilczek, F. "Ultraviolet Behavior of Nonabelian Gauge Theories". *Phys. Rev. Lett.*, 30:1343–1346 (1973).

[21] Politzer, H. D. "Reliable Perturbative Results for Strong Interactions?" *Phys. Rev. Lett.*, 30:1346–1349 (1973).

[22] Glashow, S. "Partial Symmetries of Weak Interactions". *Nucl. Phys.*, 22:579–588 (1961).

[23] Salam, A. "Weak and Electromagnetic Interactions". *Conf. Proc.*, C680519:367–377 (1968).

[24] Weinberg, S. "A Model of Leptons". *Phys. Rev. Lett.*, 19:1264–1266 (1967).

[25] Glashow, S., Iliopoulos, J., and Maiani, L. "Weak Interactions with Lepton-Hadron Symmetry". *Phys. Rev.*, D2:1285–1292 (1970).

[26] 't Hooft, G. "Renormalizable Lagrangians for Massive Yang-Mills Fields". *Nucl. Phys.*, B35:167–188 (1971).

[27] 't Hooft, G. and Veltman, M. "Regularization and Renormalization of Gauge Fields". *Nucl. Phys.*, B44:189–213 (1972).

[28] Georgi, H. and Glashow, S. L. "Unified weak and electromagnetic interactions without neutral currents". *Phys. Rev. Lett.*, 28:1494 (1972).

[29] Cabibbo, N. "Unitary Symmetry and Leptonic Decays". *Phys. Rev. Lett.*, 10:531–533 (1963).

[30] Kobayashi, M. and Maskawa, T. "CP Violation in the Renormalizable Theory of Weak Interaction". *Prog.Theor.Phys.*, 49:652–657 (1973).

[31] Olive, K. A. et al. "Review of Particle Physics". *Chin. Phys.*, C38:090001 (2014).

[32] Aaij, R. et al. "Observation of $J/\psi p$ Resonances Consistent with Pentaquark States in $\Lambda_b^0 \to J/\psi K^- p$ Decays". *Phys. Rev. Lett.*, 115:072001 (2015).

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

[34] Higgs, P. W. "Spontaneous Symmetry Breakdown without Massless Bosons". *Phys. Rev.*, 145:1156–1163 (1966).

[35] Guralnik, G. S., Hagen, C. R., and Kibble, T. W. B. "Global Conservation Laws and Massless Particles". *Phys. Rev. Lett.*, 13:585–587 (1964).

[36] Gonis. "Higgs Mechanism". `https://en.wikipedia.org/wiki/Higgs_mechanism`. Last visited 2016-09-11.

[37] Griffiths, D. J. "Introduction to elementary particles; 2nd rev. version". Physics textbook. Wiley, New York, NY (2008).

[38] Nambu, Y. "Quasiparticles and Gauge Invariance in the Theory of Superconductivity". *Phys. Rev.*, 117:648–663 (1960).

[39] Goldstone, J. "Field Theories with Superconductor Solutions". *Nuovo Cim.*, 19:154–164 (1961).

[40] Goldstone, J., Salam, A., and Weinberg, S. "Broken Symmetries". *Phys. Rev.*, 127:965–970 (1962).

[41] Dirac, P. A. M. "The Quantum Theory of the Emission and Absorption of Radiation". *Proceedings of the Royal Society of London A: Mathematical, Physical, and Engineering Sciences*, 114(767):243–265 (1927).

[42] Orear, J. and Fermi, E. "Nuclear Physics: A Course Given by Enrico Fermi at the University of Chicago". Midway Reprints. University of Chicago Press (1950).

[43] Forte, S. and Watt, G. "Progress in the Determination of the Partonic Structure of the Proton". *Ann. Rev. Nucl. Part. Sci.*, 63:291–328 (2013).

[44] Altarelli, G. and Parisi, G. "Asymptotic Freedom in Parton Language". *Nucl. Phys.*, B126:298–318 (1977).

[45] Gribov, V. N. and Lipatov, L. N. "Deep inelastic e p scattering in perturbation theory". *Sov. J. Nucl. Phys.*, 15:438–450 (1972). [Yad. Fiz., 15:781 (1972)].

[46] Dokshitzer, Y. L. "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). [Zh. Eksp. Teor. Fiz., 73:1216 (1977)].

[47] Lai, H.-L., et al. "New parton distributions for collider physics". *Phys. Rev.*, D82:074024 (2010).

[48] Gao, J., et al. "CT10 next-to-next-to-leading order global analysis of QCD". *Phys. Rev.*, D89(3):033009 (2014).

[49] Martin, A. D., et al. "Parton distributions for the LHC". *Eur. Phys. J.*, C63:189–285 (2009).

[50] Martin, A. D., et al. "Uncertainties on alpha(S) in global PDF analyses and implications for predicted hadronic cross sections". *Eur. Phys. J.*, C64:653–680 (2009).

[51] Ball, R. D. et al. "Parton distributions with LHC data". *Nucl. Phys.*, B867:244–289 (2013).

[52] Ball, R. D. et al. "Parton distributions for the LHC Run II". *JHEP*, 04:040 (2015).

[53] Whalley, M. R., Bourilkov, D., and Group, R. C. "The Les Houches accord PDFs (LHAPDF) and LHAGLUE". In "HERA and the LHC: A Workshop on the implications of HERA for LHC physics. Proceedings, Part B", (2005).

[54] Botje, M. et al. "The PDF4LHC Working Group Interim Recommendations". *arXiv*, 1101.0538 [hep-ph] (2011).

[55] Butterworth, J. et al. "PDF4LHC recommendations for LHC Run II". *J. Phys.*, G43:023001 (2016).

[56] Kinoshita, T. "Mass singularities of Feynman amplitudes". *J. Math. Phys.*, 3:650–677 (1962).

[57] Lee, T. D. and Nauenberg, M. "Degenerate Systems and Mass Singularities". *Phys. Rev.*, 133:B1549–B1562 (1964).

[58] LHC Higgs Cross Section Working Group. "LHC Higgs Cross Section Working Group Public TWiki". `https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWG`. Last visited 2016-09-11.

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

[60] Aad, G. et al. "Combined Measurement of the Higgs Boson Mass in *pp* Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments". *Phys. Rev. Lett.*, 114:191803 (2015).

[61] Yang, C.-N. "Selection Rules for the Dematerialization of a Particle Into Two Photons". *Phys. Rev.*, 77:242–245 (1950).

[62] Khachatryan, V. et al. "Constraints on the spin-parity and anomalous HVV couplings of the Higgs boson in proton collisions at 7 and 8 TeV". *Phys. Rev.*, D92(1):012004 (2015).

[63] Reina, L. and Dawson, S. "Next-to-leading order results for t anti-t h production at the Tevatron". *Phys. Rev. Lett.*, 87:201804 (2001).

[64] Dawson, S., et al. "Associated top quark Higgs boson production at the LHC". *Phys. Rev.*, D67:071503 (2003).

[65] Dawson, S., et al. "Associated Higgs production with top quarks at the large hadron collider: NLO QCD corrections". *Phys. Rev.*, D68:034022 (2003).

[66] Beenakker, W., et al. "Higgs radiation off top quarks at the Tevatron and the LHC". *Phys. Rev. Lett.*, 87:201805 (2001).

[67] Beenakker, W., et al. "NLO QCD corrections to t anti-t H production in hadron collisions". *Nucl. Phys.*, B653:151–203 (2003).

[68] Zhang, Y., et al. "QCD NLO and EW NLO corrections to $t\bar{t}H$ production with top quark decays at hadron collider". *Phys. Lett.*, B738:1–5 (2014).

[69] Frixione, S., et al. "Weak corrections to Higgs hadroproduction in association with a top-quark pair". *JHEP*, 09:065 (2014).

[70] Frixione, S., et al. "Electroweak and QCD corrections to top-pair hadroproduction in association with heavy bosons". *JHEP*, 06:184 (2015).

[71] Autermann, C. "Experimental status of supersymmetry after the LHC Run-I". *Prog. Part. Nucl. Phys.*, 90:125–155 (2016).

[72] Arkani-Hamed, N., Cohen, A. G., and Georgi, H. "Electroweak symmetry breaking from dimensional deconstruction". *Phys. Lett.*, B513:232–240 (2001).

[73] Arkani-Hamed, N., et al. "The Littlest Higgs". *JHEP*, 07:034 (2002).

[74] Contino, R., Da Rold, L., and Pomarol, A. "Light custodians in natural composite Higgs models". *Phys. Rev.*, D75:055014 (2007).

[75] Aguilar-Saavedra, J. A., et al. "Handbook of vectorlike quarks: Mixing and single production". *Phys. Rev.*, D88(9):094010 (2013).

[76] Khachatryan, V. et al. "Search for a Standard Model Higgs Boson Produced in Association with a Top-Quark Pair and Decaying to Bottom Quarks Using a Matrix Element Method". *Eur. Phys. J.*, C75(6):251 (2015).

[77] 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).

[78] Khachatryan, V. et al. "Search for the associated production of the Higgs boson with a top-quark pair". *JHEP*, 09:087 (2014). [Erratum: JHEP, 10:106 (2014)].

[79] Aad, G. et al. "Search for the Standard Model Higgs boson decaying into $b\bar{b}$ produced in association with top quarks decaying hadronically in pp collisions at $\sqrt{s} = 8\,\text{TeV}$ with the ATLAS detector". *JHEP*, 05:160 (2016).

[80] Stirling, J. "Parton Luminosity and cross section plots". `http://www.hep.ph.ic.ac.uk/~wstirlin/plots/plots.html`. Last visited 2016-09-11.

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

[82] Hasert, F. J. et al. "Observation of Neutrino Like Interactions Without Muon Or Electron in the Gargamelle Neutrino Experiment". *Phys. Lett.*, B46:138–140 (1973).

[83] Arnison, G. et al. "Experimental Observation of Isolated Large Transverse Energy Electrons with Associated Missing Energy at s**(1/2) = 540-GeV". *Phys. Lett.*, B122:103–116 (1983).

[84] Arnison, G. et al. "Experimental Observation of Lepton Pairs of Invariant Mass Around 95-GeV/c**2 at the CERN SPS Collider". *Phys. Lett.*, B126:398–410 (1983).

[85] Banner, M. et al. "Observation of Single Isolated Electrons of High Transverse Momentum in Events with Missing Transverse Energy at the CERN anti-p p Collider". *Phys. Lett.*, B122:476–485 (1983).

[86] Bagnaia, P. et al. "Evidence for Z0 → e+ e- at the CERN anti-p p Collider". *Phys. Lett.*, B129:130–140 (1983).

[87] DESY. "Weltmaschine". `http://www.weltmaschine.de/e92/e70729/e70770/`. Last visited 2016-09-23.

[88] Bayatian, G. et al. "CMS physics: Technical design report". *CERN-LHCC-2006-001, CMS-TDR-008-1* (2006).

[89] CMS Collaboration. "Detector Drawings", *CMS-PHO-GEN-2012-002* (2012). CMS Collection.

[90] Barney, D. "CMS Detector Slice", *CMS-PHO-GEN-2016-001* (2016). CMS Collection.

[91] CMS Collaboration. "CMS, tracker technical design report". *CERN-LHCC-98-06, CMS-TDR-5* (1998).

[92] CMS Collaboration. "The CMS tracker: addendum to the Technical Design Report". *CERN-LHCC-2000-016, CMS-TDR-5-add-1* (2000).

[93] Sprenger, D., et al. "Validation of Kalman Filter alignment algorithm with cosmic-ray data using a CMS silicon strip tracker endcap". *JINST*, 5:P06007 (2010).

[94] Agram, J. L. "CMS Silicon Strip Tracker Performance". *Phys. Procedia*, 37:844–850 (2012).

[95] CMS Collaboration. "The CMS electromagnetic calorimeter project: Technical Design Report". *CERN-LHCC-97-33, CMS-TDR-4* (1997).

[96] Chatrchyan, S. et al. "Performance and Operation of the CMS Electromagnetic Calorimeter". *JINST*, 5:T03010 (2010).

[97] CMS Collaboration. "The CMS hadron calorimeter project: Technical Design Report". *CERN-LHCC-97-31, CMS-TDR-2* (1997).

[98] CMS Collaboration. "The CMS muon project: Technical Design Report". *CERN-LHCC-97-32, CMS-TDR-3* (1997).

[99] Kim, M. S. et al. "CMS reconstruction improvement for the muon tracking by the RPC chambers". *PoS*, RPC2012:045 (2012). [JINST, 8:T03001 (2013)].

[100] Chatrchyan, S. et al. "Performance of the CMS Drift Tube Chambers with Cosmic Rays". *JINST*, 5:T03015 (2010).

[101] Chatrchyan, S. et al. "Performance of the CMS Cathode Strip Chambers with Cosmic Rays". *JINST*, 5:T03018 (2010).

[102] Thyssen, F. "Performance of the resistive plate chambers in the CMS experiment". *JINST*, 7:C01104 (2012).

[103] Dasu, S. et al. "CMS. The TriDAS project. Technical design report, Vol. 1: The trigger systems". *CERN-LHCC-2002-038, CMS-TDR-6-1* (2000).

[104] Sphicas, P. "CMS: The TriDAS project. Technical design report, Vol. 2: Data acquisition and high-level trigger". *CERN-LHCC-2002-026, CMS-TDR-6* (2002).

[105] Bird, I., et al. "LHC computing Grid. Technical design report". *CERN-LHCC-2005-024, LCG-TDR-001* (2005).

[106] CMS Collaboration. "CMS Luminosity Measurement for the 2015 Data Taking Period". *CMS-PAS-LUM-15-001* (2016).

[107] van der Meer, S. "Calibration of the Effective Beam Height in the ISR". *CERN-ISR-PO-68-31* (1968).

[108] CMS Collaboration. "Public CMS Luminosity Information". `https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults`. Last visited 2016-10-22.

[109] Batinkov, A., et al. "The CMS Data Quality Monitoring Software: Experience and future improvements". In "2013 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC 2013)", (2013).

[110] CMS Collaboration. "Public CMS Data Quality Information". `https://twiki.cern.ch/twiki/bin/view/CMSPublic/DataQuality`. Last visited 2016-09-07.

[111] Bartosik, N. "HEP Sketches". `http://bartosik.pp.ua/hep_sketches/`. Last visited 2016-10-19.

[112] Alwall, J., et al. "MadGraph 5 : Going Beyond". *JHEP*, 06:128 (2011).

[113] Alwall, J., et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". *JHEP*, 07:079 (2014).

[114] Nason, P. "A New method for combining NLO QCD with shower Monte Carlo algorithms". *JHEP*, 11:040 (2004).

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

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

[117] Catani, S. and Seymour, M. H. "A General algorithm for calculating jet cross-sections in NLO QCD". *Nucl. Phys.*, B485:291–419 (1997). [Erratum: Nucl. Phys., B510:503 (1998)].

[118] Nagy, Z. and Soper, D. E. "A New parton shower algorithm: Shower evolution, matching at leading and next-to-leading order level". In "Proceedings, Ringberg Workshop on New Trends in HERA Physics 2005: Ringberg Castle, Tegernsee, Germany, October 2-7, 2005", pages 101–123 (2006).

[119] Boos, E. et al. "Generic user process interface for event generators". In "Physics at TeV colliders. Proceedings, Euro Summer School, Les Houches, France, May 21-June 1, 2001", (2001).

[120] Alwall, J. et al. "A Standard format for Les Houches event files". *Comput. Phys. Commun.*, 176:300–304 (2007).

[121] Bahr, M. et al. "Herwig++ Physics and Manual". *Eur. Phys. J.*, C58:639–707 (2008).

[122] Catani, S., et al. "QCD matrix elements + parton showers". *JHEP*, 11:063 (2001).

[123] Mangano, M. L., et al. "Matching matrix elements and shower evolution for top-quark production in hadronic collisions". *JHEP*, 01:013 (2007).

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

[125] Alwall, J., et al. "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations". *JHEP*, 07:079 (2014).

[126] Sjostrand, T. and van Zijl, M. "Multiple Parton-parton Interactions in an Impact Parameter Picture". *Phys. Lett.*, B188:149–154 (1987).

[127] Andersson, B., et al. "Parton Fragmentation and String Dynamics". *Phys. Rept.*, 97:31–145 (1983).

[128] Webber, B. R. "A QCD Model for Jet Fragmentation Including Soft Gluon Interference". *Nucl. Phys.*, B238:492–528 (1984).

[129] Agostinelli, S. et al. "GEANT4: A Simulation toolkit". *Nucl. Instrum. Meth.*, A506:250–303 (2003).

[130] Dittmaier, S. et al. "Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables". *CERN-2011-002* (2011).

[131] Czakon, M. and Mitov, A. "Top++: A Program for the Calculation of the Top-Pair Cross-Section at Hadron Colliders". *Comput. Phys. Commun.*, 185:2930 (2014).

[132] Aliev, M., et al. "HATHOR: HAdronic Top and Heavy quarks crOss section calculatoR". *Comput. Phys. Commun.*, 182:1034–1046 (2011).

[133] Kant, P., et al. "HatHor for single top-quark production: Updated predictions and uncertainty estimates for single top-quark production in hadronic collisions". *Comput. Phys. Commun.*, 191:74–89 (2015).

[134] Kidonakis, N. "Top Quark Production". In "Proceedings, Helmholtz International Summer School on Physics of Heavy Quarks and Hadrons (HQ 2013): JINR, Dubna, Russia, July 15-28, 2013", pages 139–168 (2014).

[135] Maltoni, F., Pagani, D., and Tsinikos, I. "Associated production of a top-quark pair with vector bosons at NLO in QCD: impact on $t\bar{t}H$ searches at the LHC". *JHEP*, 02:113 (2016).

[136] Gavin, R., et al. "FEWZ 2.0: A code for hadronic Z production at next-to-next-to-leading order". *Comput. Phys. Commun.*, 182:2388–2403 (2011).

[137] Gavin, R., et al. "W Physics at the LHC with FEWZ 2.1". *Comput. Phys. Commun.*, 184:208–214 (2013).

[138] Campbell, J. M., Ellis, R. K., and Giele, W. T. "A Multi-Threaded Version of MCFM". *Eur. Phys. J.*, C75(6):246 (2015).

[139] Campbell, J. M. and Ellis, R. K. "An Update on vector boson pair production at hadron colliders". *Phys. Rev.*, D60:113006 (1999).

[140] Campbell, J. M., Ellis, R. K., and Williams, C. "Vector boson pair production at the LHC". *JHEP*, 07:018 (2011).

[141] Khachatryan, V. et al. "Measurement of the cross section ratio $\sigma_{t\bar{t}b\bar{b}}/\sigma_{t\bar{t}jj}$ in pp collisions at $\sqrt{s} = 8\,\text{TeV}$". *Phys. Lett.*, B746:132–153 (2015).

[142] Chatrchyan, S. et al. "Description and performance of track and primary-vertex reconstruction with the CMS tracker". *JINST*, 9(10):P10009 (2014).

[143] Frühwirth, R. "Application of Kalman filtering to track and vertex fitting". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 262:444–450 (1987).

[144] Speer, T., et al. "Track reconstruction in the CMS tracker". *Nucl. Instrum. Meth.*, A559:143–147 (2006).

[145] Schilling, F. P. "Track Reconstruction and Alignment with the CMS Silicon Tracker". In "Proceedings of the 33rd International Conference on High Energy Physics (ICHEP '06): Moscow, Russia, July 26-August 2, 2006", (2006).

[146] Chatrchyan, S. et al. "Performance of CMS Muon Reconstruction in Cosmic-Ray Events". *JINST*, 5:T03022 (2010).

[147] Frühwirth, R., Waltenberger, W., and Vanlaer, P. "Adaptive vertex fitting". *J. Phys.*, G34:N343 (2007).

[148] Rose, K. "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems". *Proceedings of the IEEE*, 86(11):2210–2239 (1998).

[149] Waltenberger, W. "Adaptive Vertex Reconstruction". *CMS-NOTE-2008-033* (2008).

[150] Speer, T., et al. "Vertex fitting in the CMS tracker". *CMS-NOTE-2006-032* (2006).

[151] Chatrchyan, S. et al. "Description and performance of track and primary-vertex reconstruction with the CMS tracker". *JINST*, 9(10):P10009 (2014).

[152] Baffioni, S., et al. "Electron reconstruction in CMS". *Eur. Phys. J.*, C49:1099–1116 (2007).

[153] Khachatryan, V. et al. "Performance of Electron Reconstruction and Selection with the CMS Detector in Proton-Proton Collisions at s = 8 TeV". *JINST*, 10(06):P06005 (2015).

[154] CMS Collaboration. "Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET". *CMS-PAS-PFT-09-001* (2009).

[155] Adam, W., et al. "Reconstruction of electrons with the Gaussian sum filter in the CMS tracker at LHC". *eConf*, C0303241:TULT009 (2003). [J. Phys., G31:N9 (2005)].

[156] Chatrchyan, S. et al. "Performance of CMS muon reconstruction in *pp* collision events at $\sqrt{s} = 7\,\text{TeV}$". *JINST*, 7:P10002 (2012).

[157] CMS Collaboration. "Pileup Removal Algorithms". *CMS-PAS-JME-14-001* (2014).

[158] Salam, G. P. "Towards Jetography". *Eur. Phys. J.*, C67:637–686 (2010).

[159] Catani, S., et al. "Longitudinally invariant K_t clustering algorithms for hadron hadron collisions". *Nucl. Phys.*, B406:187–224 (1993).

[160] Ellis, S. D. and Soper, D. E. "Successive combination jet algorithm for hadron collisions". *Phys. Rev.*, D48:3160–3166 (1993).

[161] Dokshitzer, Y. L., et al. "Better jet clustering algorithms". *JHEP*, 08:001 (1997).

[162] Wobisch, M. and Wengler, T. "Hadronization corrections to jet cross-sections in deep inelastic scattering". In "Monte Carlo generators for HERA physics. Proceedings, Workshop, Hamburg, Germany, 1998-1999", pages 270–279 (1998).

[163] Cacciari, M., Salam, G. P., and Soyez, G. "The Anti-k(t) jet clustering algorithm". *JHEP*, 0804:063 (2008).

[164] Catani, S., Dokshitzer, Y. L., and Webber, B. R. "The K^- perpendicular clustering algorithm for jets in deep inelastic scattering and hadron collisions". *Phys. Lett.*, B285:291–299 (1992).

[165] Marchesini, G., et al. "HERWIG: A Monte Carlo event generator for simulating hadron emission reactions with interfering gluons. Version 5.1 - April 1991". *Comput. Phys. Commun.*, 67:465–508 (1992).

[166] Corcella, G., et al. "HERWIG 6: An Event generator for hadron emission reactions with interfering gluons (including supersymmetric processes)". *JHEP*, 0101:010 (2001).

[167] Khachatryan, V. et al. "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV". *JINST*, 12(02):P02014 (2017).

[168] Chatrchyan, S. et al. "Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS". *JINST*, 6:P11002 (2011).

[169] CMS Collaboration. "Identification of b quark jets at the CMS Experiment in the LHC Run 2". *CMS-PAS-BTV-15-001* (2016).

[170] Chatrchyan, S. et al. "Identification of b-quark jets with the CMS experiment". *JINST*, 8:P04013 (2013).

[171] CMS Collaboration. "Utilities for Accessing Pileup Information for Data (internal documentation)". `https://twiki.cern.ch/twiki/bin/view/CMS/PileupJSONFileforData#Pileup_JSON_Files_For_Run_II`. Last visited 2016-10-17.

[172] Khachatryan, V. et al. "Measurements of Inclusive W and Z Cross Sections in pp Collisions at $\sqrt{s} = 7\,\text{TeV}$". *JHEP*, 01:080 (2011).

[173] CMS Collaboration. "Reference muon id, isolation and trigger efficiencies for Run-II (internal documentation)". `https://twiki.cern.ch/twiki/bin/view/CMS/MuonReferenceEffsRun2`. Last visited 2016-10-17.

[174] CMS Collaboration. "Reference electron id, isolation and trigger efficiencies for Run-II (internal documentation)". `https://twiki.cern.ch/twiki/bin/view/CMS/EgammaIDRecipesRun2`. Last visited 2016-10-17.

[175] Roe, B. P., Yang, H.-J., and Zhu, J. "Boosted decision trees, a powerful event classifier". In "Statistical problems in particle physics, astrophysics and cosmology. Proceedings, Conference, PHYSTAT05, Oxford, UK, September 12-15, 2005", pages 139–142 (2005).

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

[177] Brun, R. and Rademakers, F. "ROOT: An object oriented data analysis framework". *Nucl. Instrum. Meth.*, A389:81–86 (1997).

[178] Kennedy, J. and Eberhart, R. "Particle swarm optimization". In "Neural Networks, 1995. Proceedings., IEEE International Conference on", volume 4, pages 1942–1948 vol.4 (1995).

[179] El Morabit, K. "A study of the multivariate analysis of Higgs boson production in association with a top quark-antiquark pair in the boosted regime at the CMS experiment". Master's thesis, Karlsruhe Institute of Technology, Germany (2015).

[180] Eberhart, R. and Shi, Y. "Comparing inertia weights and constriction factors in particle swarm optimization". In "Evolutionary Computation, 2000. Proceedings of the 2000 Congress on", volume 1, pages 84–88 vol.1 (2000).

[181] Trelea, I. C. "The particle swarm optimization algorithm: convergence analysis and parameter selection ". *Inform. Process. Lett.*, 85(6):317–325 (2003).

[182] Bohm, G. and Zech, G. "Introduction to statistics and measurement analysis for physicists". DESY, Hamburg, Germany (2005).

[183] Blobel, V. and Lohrmann, E. "Statistische und numerische Methoden der Datenanalyse". Teubner Studienbücher Physik. Teubner, Stuttgar, Germany (1998).

[184] Neyman, J. and Pearson, E. S. "On the problem of the most efficient tests of statistical hypotheses". *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical, and Engineering Sciences*, 231(694-706):289–337 (1933).

[185] Read, A. L. "Presentation of search results: The CL(s) technique". *J. Phys.*, G28:2693–2704 (2002).

[186] Read, A. L. "Modified frequentist analysis of search results (The CL(s) method)". In "Workshop on confidence limits, CERN, Geneva, Switzerland, 17-18 Jan 2000: Proceedings", (2000).

[187] Junk, T. "Confidence level computation for combining searches with small statistics". *Nucl. Instrum. Meth.*, A434:435–443 (1999).

[188] Fisher, W. "Systematics and limit calculations". *D0-NOTE-5309* (2006).

[189] The ATLAS Collaboration, the CMS Collaboration, and the LHC Higgs Combination Group. "Procedure for the LHC Higgs boson search combination in Summer 2011". *CMS-NOTE-2011-005, ATL-PHYS-PUB-2011-11* (2011).

[190] Cowan, G., et al. "Asymptotic formulae for likelihood-based tests of new physics". *Eur. Phys. J.*, C71:1554 (2011). [Erratum: Eur. Phys. J., C73:2501 (2013)].

[191] Wilks, S. S. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses". *Annals Math. Statist.*, 9(1):60–62 (1938).

[192] Plehn, T., et al. "Stop Reconstruction with Tagged Tops". *JHEP*, 10:078 (2010).

[193] Aad, G. et al. "Search for resonances decaying into top-quark pairs using fully hadronic decays in *pp* collisions with ATLAS at $\sqrt(s) = 7\,\text{TeV}$". *JHEP*, 01:116 (2013).

[194] Plehn, T. and Spannowsky, M. "Top Tagging". *J. Phys.*, G39:083001 (2012).

[195] CMS Collaboration. "Search for $\text{t}\bar{\text{t}}$ resonances in boosted semileptonic final states in pp collisions at $\sqrt(s) = 13\,\text{TeV}$". *CMS-PAS-B2G-15-002* (2016).

[196] CMS Collaboration. "Search for top quark-antiquark resonances in the all-hadronic final state at $\sqrt(s) = 13\,\text{TeV}$". *CMS-PAS-B2G-15-003* (2016).

[197] Larkoski, A. J., et al. "Soft Drop". *JHEP*, 05:146 (2014).

[198] Krohn, D., Thaler, J., and Wang, L.-T. "Jet Trimming". *JHEP*, 02:084 (2010).

[199] Ellis, S. D., Vermilion, C. K., and Walsh, J. R. "Techniques for improved heavy particle searches with jet substructure". *Phys. Rev.*, D80:051501 (2009).

[200] Ellis, S. D., Vermilion, C. K., and Walsh, J. R. "Recombination Algorithms and Jet Substructure: Pruning as a Tool for Heavy Particle Searches". *Phys. Rev.*, D81:094023 (2010).

[201] Thaler, J. and Van Tilburg, K. "Identifying Boosted Objects with N-subjettiness". *JHEP*, 03:015 (2011).

[202] Thaler, J. and Van Tilburg, K. "Maximizing Boosted Top Identification by Minimizing N-subjettiness". *JHEP*, 02:093 (2012).

[203] CMS Collaboration. "Search for BSM $\text{t}\bar{\text{t}}$ Production in the Boosted All-Hadronic Final State". *CMS-PAS-EXO-11-006* (2011).

[204] CMS Collaboration. "Search for $\text{t}\bar{\text{t}}\text{H}$ production in the $\text{H}\to\text{b}\bar{\text{b}}$ decay channel with $\sqrt{s} = 13\,\text{GeV}$ pp collisions at the CMS experiment". `http://cms-results.web.cern.ch/cms-results/public-results/preliminary-results/HIG-16-004/index.html`. Last visited 2016-10-11.

[205] Kondo, K. "Dynamical Likelihood Method for Reconstruction of Events With Missing Momentum. 1: Method and Toy Models". *J. Phys. Soc. Jap.*, 57:4126–4140 (1988).

[206] Kondo, K. "Dynamical likelihood method for reconstruction of events with missing momentum. 2: Mass spectra for $2 \to 2$ processes". *J. Phys. Soc. Jap.*, 60:836–844 (1991).

[207] Kondo, K., Chikamatsu, T., and Kim, S. H. "Dynamical likelihood method for reconstruction of events with missing momentum. 3: Analysis of a CDF high p(T) e mu event as t anti-t production". *J. Phys. Soc. Jap.*, 62:1177–1182 (1993).

[208] Abbott, B. et al. "Measurement of the top quark mass in the dilepton channel". *Phys. Rev.*, D60:052001 (1999).

[209] Abazov, V. M. et al. "A precision measurement of the mass of the top quark". *Nature*, 429:638–642 (2004).

[210] Abazov, V. M. et al. "Helicity of the W boson in lepton + jets $t\bar{t}$ events". *Phys. Lett.*, B617:1–10 (2005).

[211] Lepage, G. P. "A New Algorithm for Adaptive Multidimensional Integration". *J. Comput. Phys.*, 27:192 (1978).

[212] Dawson, S., et al. "Associated top quark Higgs boson production at the LHC". *Phys. Rev.*, D67:071503 (2003).

[213] Cascioli, F., Maierhofer, P., and Pozzorini, S. "Scattering Amplitudes with Open Loops". *Phys. Rev. Lett.*, 108:111601 (2012).

[214] Kauer, N. "Narrow-width approximation limitations". *Phys. Lett.*, B649:413–416 (2007).

[215] Bjorken, J. D. and Brodsky, S. J. "Statistical Model for electron-Positron Annihilation Into Hadrons". *Phys. Rev.*, D1:1416–1420 (1970).

[216] Fox, G. C. and Wolfram, S. "Event Shapes in e+ e- Annihilation". *Nucl. Phys.*, B149:413 (1979). [Erratum: Nucl. Phys., B157:543 (1979)].

[217] Bernaciak, C., et al. "Fox-Wolfram Moments in Higgs Physics". *Phys. Rev.*, D87:073014 (2013).

[218] Ball, R. D., et al. "Parton distributions with QED corrections". *Nucl. Phys.*, B877:290–320 (2013).

[219] Pumplin, J., et al. "New generation of parton distributions with uncertainties from global QCD analysis". *JHEP*, 07:012 (2002).

[220] Chatrchyan, S. et al. "Measurement of the cross section and angular correlations for associated production of a Z boson with b hadrons in pp collisions at $\sqrt{s} = 7\,\text{TeV}$". *JHEP*, 12:039 (2013).

[221] Bredenstein, A., et al. "NLO QCD Corrections to Top Anti-Top Bottom Anti-Bottom Production at the LHC: 2. full hadronic results". *JHEP*, 03:021 (2010).

[222] Khachatryan, V. et al. "Measurement of $\mathrm{t\bar{t}}$ production with additional jet activity, including b quark jets, in the dilepton decay channel using pp collisions at $\sqrt{s} = 8\,\text{TeV}$". *Eur. Phys. J.*, C76(7):379 (2016).

[223] CMS Collaboration. "Jet Energy Resolution in CMS at $\sqrt{s} = 7\,\text{TeV}$". *CMS-PAS-JME-10-014* (2011).

[224] Barlow, R. J. and Beeston, C. "Fitting using finite Monte Carlo samples". *Comput. Phys. Commun.*, 77:219–228 (1993).

[225] Conway, J. S. "Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra". In "Proceedings, PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding, CERN,Geneva, Switzerland 17-20 January 2011", pages 115–120 (2011).

[226] CMS Collaboration. "Search for $t\bar{t}H$ production in multilepton final states at $\sqrt{s} = 13\,\text{TeV}$". *CMS-PAS-HIG-15-008* (2016).

[227] CMS Collaboration. "First results on Higgs to $\gamma\gamma$ at 13 TeV". *CMS-PAS-HIG-15-005* (2016).

[228] CMS Collaboration. "ttH Combination". `https://twiki.cern.ch/twiki/bin/view/CMSPublic/TTHCombMoriond2016`. Last visited 2016-08-25.

[229] CMS Collaboration. "Search for associated production of Higgs bosons and top quarks in multilepton final states at $\sqrt{s} = 13\,\text{TeV}$". *CMS-PAS-HIG-16-022* (2016).

[230] CMS Collaboration. "Updated measurements of Higgs boson production in the diphoton decay channel at $\sqrt{s} = 13\,\text{TeV}$ in pp collisions at CMS." *CMS-PAS-HIG-16-020* (2016).

[231] Aad, G. et al. "Search for the Standard Model Higgs boson produced in association with top quarks and decaying into $b\bar{b}$ in pp collisions at $\sqrt{s} = 8\,\text{TeV}$ with the ATLAS detector". *Eur. Phys. J.*, C75(7):349 (2015).

[232] Aad, G. et al. "Search for the associated production of the Higgs boson with a top quark pair in multilepton final states with the ATLAS detector". *Phys. Lett.*, B749:519–541 (2015).

[233] Aad, G. et al. "Search for $H \to \gamma\gamma$ produced in association with top quarks and constraints on the Yukawa coupling between the top quark and the Higgs boson using data taken at 7 TeV and 8 TeV with the ATLAS detector". *Phys. Lett.*, B740:222–242 (2015).

[234] The ATLAS collaboration. "Measurement of fiducial, differential and production cross sections in the $H \to \gamma\gamma$ decay channel with $13.3\,\text{fb}^{-1}$ of $\sqrt{s} = 13\,\text{TeV}$ proton-proton collision data with the ATLAS detector". *ATLAS-CONF-2016-067* (2016).

[235] The ATLAS collaboration. "Search for the Associated Production of a Higgs Boson and a Top Quark Pair in Multilepton Final States with the ATLAS Detector". *ATLAS-CONF-2016-058* (2016).

[236] The ATLAS collaboration. "Search for the Standard Model Higgs boson produced in association with top quarks and decaying into $b\bar{b}$ in pp collisions at $\sqrt{s} = 13\,\text{TeV}$ with the ATLAS detector". *ATLAS-CONF-2016-080* (2016).

[237] The ATLAS collaboration. "Combination of the searches for Higgs boson production in association with top quarks in the $\gamma\gamma$, multilepton, and $b\bar{b}$ decay channels at $\sqrt{s} = 13\,\text{TeV}$ with the ATLAS Detector". *ATLAS-CONF-2016-068* (2016).

[238] Gleisberg, T., et al. "Event generation with SHERPA 1.1". *JHEP*, 02:007 (2009).

[239] Aad, G. et al. "Performance of jet substructure techniques for large-R jets in proton-proton collisions at $\sqrt{s} = 7\,\text{TeV}$ using the ATLAS detector". *JHEP*, 09:076 (2013).

[240] CMS Collaboration. "Boosted Top Jet Tagging at CMS". *CMS-PAS-JME-13-007* (2014).

Acknowledgements

Zum Abschluss möchte ich noch all denjenigen meinen Dank ausdrücken, die auf die ein oder andere Weise zu dieser Doktorarbeit beigetragen haben. Sei es dadurch, dass sie mich heute und in meinem bisherigen Werdegang begleitet oder mich auf vielfältige Weise unterstützt haben. Ohne diese Menschen wäre diese Doktorarbeit niemals möglich gewesen.

Zuallererst möchte ich dem Betreuer meiner Doktorarbeit Herr Prof. Dr. Ulrich Husemann, der es mir ermöglicht hat ein Teil in seiner einzigartigen Arbeitsgruppe zu sein, herzlich danken. Bei ihm hatte ich das Privileg, frei an den Themen der experimentellen Teilchenphysik forschen und gleichzeitig jederzeit auf seine Unterstützung zählen zu können. Es war eine Freude mit ihm zu arbeiten und von ihm zu lernen.

Herr Prof. Dr. Thomas Müller gilt nicht nur für die Übernahme des Koreferats mein Dank, sondern auch für die freundliche Aufnahme in das Institut für Experimentelle Kernphysik und die Unterstützung während der Bearbeitung meiner Diplomarbeit und meiner Doktorarbeit.

Des Weiteren möchte ich allen Mitgliedern der Arbeitsgruppe von Prof. Dr. Ulrich Husemann danken. In diesem Zusammenhang, möchte ich Dr. Hannes Mildner, Dr. Matthias Schröder, Karim El Morabit und Marco Harrendorf nennen, um nur ein paar wenige dieser Personen aufzuzählen. Ihr habt mit eurer Arbeit, eurem Engagement, euren konstruktiven Anstößen und eurer Gesellschaft maßgeblich zum Erfolg dieser Doktorarbeit beigetragen. Es war eine einzige Freude mit euch zusammenzuarbeiten, egal ob wir zusammen gelacht haben, in groteske Diskussionen abgedriftet sind oder frustrierende Probleme mitten in der Nacht gelöst haben. Nicht weniger möchte ich Dr. Alexis Descroix danken, der mich zu Beginn unter seine Fittiche genommen hat und mich in die Arbeit in der experimentellen Teilchenphysik eingeführt hat.

Further, I would like to thank all those within the CMS collaboration, who contributed to the success of the $t\bar{t}(H\to b\bar{b})$ analysis. Without these people, the studies presented in this thesis would not have been possible. Representative I would like to mention Dr. Darren Puigh, who kept an overview and accomplished to combine the various contributions into an excellent result. Additionally, I would like to thank Dr. Gregor Kasieczka for his help in including the boosted analysis strategy in the $t\bar{t}(H\to b\bar{b})$ search.

Weitere Worte des Dankes möchte ich an alle weiteren Kollegen des Instituts für Experimentelle Kernphysik und der Fakultät für Physik des Karlsruher Instituts für Technologie richten. Sowohl die fachliche Unterstürzung als auch angenehme Atmosphäre die mir diese Menschen entgegengebracht haben, haben wesentlich zum Erfolg dieser Doktorarbeit beigetragen.

Ein besonderer Dank geht an all diejenigen, die diese Forschung überhaupt möglich machen, indem sie sich um die Formalitäten, die Finanzierung, die Computerinfrastruk-

tur, u.s.w kümmern. In diesem Zusammenhang möchte ich Frau Fellner, Frau Bräunling, Frau Gering, Frau Junge, Frau Lepold, Frau Hühn, die Mitglieder des Graduiertenkollegs 1694 „Elementarteilchenphysik bei höchster Energie und höchster Präzision“ und der Admingruppe des Instituts für Experimentelle Kernphysik danken.

Nicht weniger danken möchte ich den Menschen, die mich auf den verschiedensten Stationen meines Lebens begleitet und unterstützt haben: Meine Familie und meine Freunde, Doris und David Cable, Dominik Kara, Benjamin Weber, Christian Gorenflo und Andreas Reimold, um nur einige wenige zu nennen. Alle diejenigen ich ich nicht genannt habe: Bitte fühlt euch auch angesprochen. Vielen Dank!

Jemanden, den ich hier nicht vergessen möchte, ist meine Großmutter Elfriede Gäckle. Sie hat mich für einen großen Teil meines Lebens begleitet und mit die Grundsteine für meinen Werdegang gelegt. Ich danke dir für alles!

One of the biggest thank-yous goes to my Mom and my Dad, Monika and Arthur Williamson. They loved me unconditionally, supported me in everything that I did, and made me the person I am today. I owe them much more than I could ever give back and I thank them much more than I could ever say! I love you!

Zu guter Letzt möchte ich mich herzlich bei meiner Verlobten Ramona Sorg bedanken. Ich danke ihr für die Liebe und Zuwendung, die sie mir gibt, und die Zeit, die ich mit ihr verbringen darf. Sie schenkt mir Kraft und Unterstützung, die mir dabei hilft meine Ziele zu verfolgen und große Aufgaben, wie diese Doktorarbeit zu bewältigen. Ich liebe dich und freue mich auf eine gemeinsame Zukunft mit dir!

Hiermit erkläre ich, dass ich die vorliegende Dissertationsschrift selbständig und unter ausschließlicher Verwendung der angegebenen Hilfsmittel angefertigt habe.

Shawn Williamson

Karlsruhe, den 25.10.2016