# Measurement of the Higgs-Boson Production in Association with a Top-Quark Pair in the Fully-Hadronic Final State with the CMS Experiment

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Korbinian G. Schweiger

aus

Deutschland

CERN-THESIS-2022-391
22/02/2022

**Promotionskommission**

Prof. Dr. Florencia Canelli (Vorsitz)

Prof. Dr. Ben Kilminster

Prof. Dr. Massimiliano Grazzini

Dr. Lorenzo Bianchini

Zürich, 2022

# Contents

**Introduction** 1

**I. Theoretical foundations, experimental setup, tools and methods** 3

**1. Theoretical foundations and collider physics** 5
1.1. The standard model of particle physics . . . 5
1.2. Hadron collider phyiscs . . . 12
1.3. Higgs-boson physics . . . 18

**2. Experimental environment** 31
2.1. The Large Hadron Collider . . . 31
2.2. The CMS Detector . . . 33
2.3. Object reconstruction . . . 41
2.4. Luminosity, data certification, and pileup . . . 48

**3. Signal identification and extraction** 51
3.1. Matrix element method . . . 51
3.2. Neural Networks . . . 57
3.3. Statistical model for Higgs analyses at CMS . . . 67

**II. Analyses strategy and Results** 71

**4. Analysis overview** 73
4.1. The ttH(bb) channel . . . 73
4.2. Modeling signal and background processes . . . 77
4.3. Uncertainties in the simulation . . . 82

**5. Analysis Strategy in the fully hadronic channel** 87
5.1. Trigger . . . 88
5.2. Data-Simulation Corrections . . . 91
5.3. Object selection . . . 93
5.4. Event Selection . . . 95
5.5. Multijet background estimation . . . 98
5.6. Signal identification . . . 106

**6. Statistical Model and systematic uncertainties** 117
6.1. Statistical model . . . 117
6.2. Systematic uncertainties . . . 119

**7. Results and Combination** 125
7.1. Strategies in the semileptonic and dileptonic channels . . . 125

7.2. Run-2 result for $t\bar{t}H(b\bar{b})$ . . . . . 128

**Summary and Outlook** **149**

**Bibliography** **153**

**Appendix** **165**
A. Trigger strategy . . . . . 165
B. Additional information on object and event selections . . . . . 178
C. ANN classifier input variables . . . . . 180

**Acknowledgements** **193**

# Introduction

The standard model of particle physics (SM) is the current most complete theory of the fundamental nature of the Universe to date. When it was formulated in the 1960s and 1970s, it predicted various particles, physics processes and observables that were unknown at the time but could be experimentally confirmed later. This culminated in the observation of a new boson with a mass of roughly 125 GeV by the ATLAS [1] and CMS [2] collaborations in 2012. Since then, many measurements of the properties of this new boson have been performed and up to this date no significant deviation from the prediction of the SM has been found [3–6]. In addition to proving the predictive power of the SM once more, this is also a great achievement for CERN and the Large Hadron Collider (LHC). The LHC started its operation in 2010 with the main goal of enabling the discovery of the Higgs boson, increasing the precision of SM parameters, and searching for physics phenomena beyond the standard model (BSM). Four large detectors are located at the LHC with the CMS detector being one of them. CMS is designed as a multi-purpose experiment probing high-energy and high-intensity proton-proton collision. The CMS experiment is able to measure the energy and position of thousands of particles produced in the collisions with very high precision. At the time of writing this thesis, most production and decay processes of the Higgs boson predicted by the SM have been observed. The interaction of the Higgs boson with fermions is described by the Yukawa interaction, with a coupling strength proportional to the mass of the fermion. Among these couplings, the coupling between Higgs boson and the top quark is expected to be largest, because the top quark is by far the heaviest fermion. Investigating this coupling is also an important probe of the SM as it is important to vacuum stability, cosmology and plays a major role in many new physics scenarios. The top-Yukawa coupling can be measured indirectly in processes in which the Higgs boson is produced via a top-loop. A direct measurement is indispensable because it is more robust against phenomena that can enter through the loop unnoticed. Production modes where the Higgs boson is produced in association with top quarks, either with a top-quark pair ($t\bar{t}H$) or a single top quark (tH) are crucial for the understanding of the SM as they provide an avenue for a direct measurement of the top-Yukawa coupling, and it is the main topic of the work presented in this thesis.

The production of the Higgs boson via the $t\bar{t}H$ production mode was observed by the CMS [7] and ATLAS [8] collaborations in 2018. The observation included the decay of the Higgs boson into $WW^*$, $ZZ^*$, $\tau^+\tau^-$, $\gamma\gamma$, and $b\bar{b}$. The $t\bar{t}H$ production with the Higgs boson decaying to a pair of bottom-quarks ($t\bar{t}H(b\bar{b})$) is of particular importance given its large branching ratio.
This thesis presents the results of the measurements of $t\bar{t}H(b\bar{b})$ with data collected by the CMS experiment between 2016 and 2018 in proton-proton collisions at 13 TeV at the LHC. The main focus of the work is on the fully-hadronic final state, in which the top quarks decay exclusively to light and bottom quarks. A measurement of this final state was published by the ATLAS collaboration using data collected at 8 TeV [9]. At CMS, this channel was pioneered using $36.3\,\text{fb}^{-1}$ of data collected in 2016 [10] at 13 TeV. A refined version of this analysis was published in combination with other $t\bar{t}H(b\bar{b})$ final states using

an additional $41.5\,\mathrm{fb}^{-1}$ of data collected in 2017 [11].
The final state in the fully-hadronic channel exhibits a total of eight jets, four coming from the hadronization of light quarks from the W-boson decays and four from the bottom-quarks. Ideally, signal events would present themself in the CMS detector with eight jets, of which four could be identified as b-jets. Due to experimental reasons, events with at least seven jets are used to measure this $\mathrm{t\bar{t}H(b\bar{b})}$ process. Since no leptons are present in the final state, the signal is hidden underneath a very large and dominating background from QCD multijet events, the most common process at a hadron colliders. Additionally the production of top-quark pairs with additional jets ($\mathrm{t\bar{t}}+\mathrm{jets}$), originating from gluon splitting, contribute a sizable background. Especially $\mathrm{t\bar{t}}$ with additional bottom-quark jets ($\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$) is a background process that cannot be easily removed due to its similarity to the signal. Separating the signal from the backgrounds is among the most crucial challenges in the analysis of the fully-hadronic $\mathrm{t\bar{t}H(b\bar{b})}$ channel. Furthermore, the modeling of the QCD multijet background presents itself as a limiting factor on the sensitivity of the measurement.
This thesis presents the analysis of fully-hadronic decay channel of $\mathrm{t\bar{t}H(b\bar{b})}$. It directly builds on the foundations of the analyses presented in references [10, 11] by improving and refining the used methods: New dedicated triggers, targeting all-jets final states, were developed for the 2017 and 2018 data-taking periods of the CMS detector. The estimation method for the QCD multijet background was revised and improved. A novel neural-network-based classifier was developed to separate the $\mathrm{t\bar{t}H(b\bar{b})}$ signal from the QCD multijet and, to a lesser extent, the $\mathrm{t\bar{t}}+\mathrm{jets}$ background. An improved estimation of the QCD multijet background allowed to avoid the previous statistical limitation and use collected data events in the training of the neural-network classifier. Additionally, the analysis implemented the state-of-the-art modeling of $\mathrm{t\bar{t}}+\mathrm{b\bar{b}}$ and an improved algorithm for b-jet identification.

This thesis is split into two main parts. In part I, the theoretical foundation, the experimental setup and the tools and methods will be introduced. Chapter 1 summarizes the standard model of particle physics, describes essential concepts of collider physics, and introduces theoretical and phenomenological concepts of Higgs-boson physics with a focus on the $\mathrm{t\bar{t}H}$ production. In chapter 2 the LHC machine and the CMS experiment are described. In chapter 3, the methods for signal identification and extraction are discussed. Part II will cover the fully-hadronic $\mathrm{t\bar{t}H(b\bar{b})}$ analysis and the combined measurement of $\mathrm{t\bar{t}H(b\bar{b})}$ in all decays of the top quark pair using data collected between 2016-2018 at the LHC. The experimental aspects of the $\mathrm{t\bar{t}H(b\bar{b})}$ analysis will be introduced, including a discussion of the backgrounds and their modeling, in chapter 4. The analysis strategy of the fully-hadronic decay channel will be described in chapter 5. This chapter includes the selection of signal events, estimation of the QCD multijet background, and the signal identification. Chapter 6 will cover the statistical model and systematic uncertainties considered in the extraction of the measurement from the data. Lastly, the results are discussed in chapter 7. The main result is the measurement of the $\mathrm{t\bar{t}H(b\bar{b})}$ in the all-hadronic channel and the effect of the combination with the leptonic channels. In addition, the measurement is used to set a limit on the production of a single top quark and a Higgs boson ($\mathrm{tH(b\bar{b})}$), a measurement of the Top-Yukawa coupling, and a cross section measurement broken down into various kinematic regions (following the simplified template cross- section (STXS) framework).

# Part I.

# Theoretical foundations, experimental setup, tools and methods

# 1. Theoretical foundations and collider physics

The standard model of particle physics provides the basic framework for the measurements presented in this thesis. It is briefly summarized in this chapter with a focus on the Higgs mechanism as it plays a pivotal role in the work presented in this thesis. This chapter also includes an introduction to hadron collider physics and summarizes the phenomenological and experimental status of Higgs-boson physics.

## 1.1. The standard model of particle physics

The standard model of particle physics (SM) [12–27] is a quantum field theory [28] that describes the elementary particles and their interaction. This theory provides a mechanism to accurately predict certain observations in particle-physics experiments, like the cross section of a particle-physics process.
The SM contains the description of the fundamental particles and most fundamental forces known to date. The particles can be split into fermions and bosons. Tables 1.1 and 1.2 list all fermions and gauge bosons as they present themselves in the SM. The fermions make up matter as we know it and the gauge bosons mediate the forces and describe the interactions within the SM. The SM predicts one additional particle, that is also a boson but does not mediate a force: The Higgs boson. In the SM it is introduced through a mechanism that describes how the fermions and some of the gauge bosons obtain mass.

The SM was formulated during the 1960s and 1970s. In addition to the particles already known at that time, like the electron, it also postulated further particles, that are necessary for the internal consistency of the theory. This fact motivated a multitude of groundbreaking experiments resulting in new observations. For example the charm quark [30, 31] was found 1974, the bottom quark [32] in 1977, the W [33] and Z bosons [34] in 1983, the top quark [35, 36] in 1995, and finally the Higgs boson [1, 2] in 2012.
While the SM describes most observations in particle physics with unparalleled accuracy, it is commonly accepted, that physics beyond the SM (BSM), i.e. physics not described by the SM, must exists. The reasons for this are observations, like the neutrino oscillation phenomenon [37, 38], and strong evidences, like the recent measurements on the magnetic momentum of the muon by the Muon-g2 experiment [39], that cannot be explained within the SM framework. Searches for new particles that can be explained with extensions to the SM (like Supersymmetry) or completely new theories are an active area of research

Table 1.1.: The three generations of fermions in the standard model. The electric charge, weak isospin $T_3$ and color determine how fermions interact. The numerical values are taken from [29].

| Gen. | particle | charge/$e$ | $T_3$ | color | mass |
|---|---|---|---|---|---|
| $1^{\text{st}}$ | Up quark ($u$) | $2/3$ | $1/2$ | r,g,b | $2.16^{+0.49}_{-0.26}$ MeV |
| | Down quark ($d$) | $-1/3$ | $-1/2$ | r,g,b | $4.67^{+0.48}_{-0.17}$ MeV |
| | Electron ($e$) | $-1$ | $-1/2$ | – | $0.511 \pm (0.31 \cdot 10^{-8})$ MeV |
| | Electron neutrino ($\nu_e$) | 0 | $1/2$ | – | $< 1.1$ eV |
| $2^{\text{nd}}$ | Charm quark ($c$) | $2/3$ | $1/2$ | r,g,b | $1.27 \pm 0.02$ GeV |
| | Strange quark ($s$) | $-1/3$ | $-1/2$ | r,g,b | $93^{+11}_{-5}$ MeV |
| | Muon ($\mu$) | $-1$ | $-1/2$ | – | $105.7 \pm (2.4 \cdot 10^{-6})$ MeV |
| | Muon neutrino ($\nu_\mu$) | 0 | $1/2$ | – | $< 1.1$ eV |
| $3^{\text{rd}}$ | Top quark ($t$) | $2/3$ | $1/2$ | r,g,b | $172.76 \pm 0.3$ GeV |
| | Bottom quark ($b$) | $-1/3$ | $-1/2$ | r,g,b | $4.18^{+0.03}_{-0.02}$ GeV |
| | Tau ($\tau$) | $-1$ | $-1/2$ | – | $1776.86 \pm 0.12$ MeV |
| | Tau neutrino ($\nu_\tau$) | 0 | $1/2$ | – | $< 1.1$ eV |

Table 1.2.: Gauge bosons of the standard model. All numerical values are taken from [29].

| Particle | Interaction | charge | $T_3$ | color | mass |
|---|---|---|---|---|---|
| photon ($\gamma$) | electroweak | 0 | 0 | – | massless ($< 1 \cdot 10^{-18}$eV) |
| $W^{\pm}$ boson | electroweak | $\pm e$ | $\pm 1$ | – | $80.79 \pm 0.012$GeV |
| Z bosons | electroweak | 0 | 0 | – | $91.1876 \pm 0.0021$GeV |
| gluon ($g$) | strong | 0 | 0 | all | massless |

in particles physics. No further discussion of such theories will be included in this thesis. The measurements presented in this thesis are done in the framework of the SM phenomena. Measurements deviating from the SM prediction could hint at new physics and are a motivation to improve in order to reach the highest precision possible.

### 1.1.1. Fermions

Fermions carry a spin of 1/2 in units of $\hbar$ and follow the Fermi-Dirac statistics and subsequently the Pauli exclusion principle. An overview of the fermions and their properties in the SM is given in table 1.1. The fermions of the SM are subdivided into leptons, which do not carry color charge, and quarks, which do carry a color charge. Consequently quarks can interact via the strong force while leptons cannot. For both groups, six particles, which are subdivided into three generations or families, are part of the SM. Typically, the masses of the particles within a group increase when ascending in generations (e.g. from the first to the second generation). Because the mass increases from generation to generation, particles from higher generation typically decay to particles of lower generation. This is also the reason that particles of the first generation are stable and are typically observed in nature[1]. The two different groups, leptons and quarks, will be discussed in the following paragraphs. Additionally each of the particles has an anti-particle with opposite charge-type quantum numbers.

#### Leptons

The leptons can be further categorized in charged leptons and neutrinos. The first generation of leptons consists of the (charged) election ($e$) and the electron neutrino ($\nu_e$). The second generation out of the muon ($\mu$) and muon neutrino ($\nu_\mu$) and the third out of the tau lepton ($\tau$) and the tau neutrino ($\nu_\tau$). All charged leptons feature an electric charge of $Q = -e$ and can interact via the electromagnetic and the weak force. The neutrinos on the other hand feature a charge of $Q = 0$ and can consequently only interact via the weak interaction, which only interacts with left-handed neutrinos. Hence, right-handed neutrinos are not part of the SM and have also not been observed so far.

#### Quarks

The quarks can be further categorized into up-type and down-type quarks per generation. The first generation of quarks consists of the up quark ($u$) and the down quark ($d$). Followed by the charm quarks ($c$) and strange quark ($s$) in the second generation, and the top quark ($t$) and bottom quark ($b$) in the third generation. Up-type quarks feature an electric charge of $Q = 2/3e$, while down-type quarks feature an electric charge of $Q = -1/3e$. Furthermore, quarks carry one of the three color charges red, green, or blue. Quarks can interact via all interactions described by the SM. Due to a special property of the strong interaction, called confinement, which will be described in the next section, they cannot be observed as individual particles. Instead they are always bound into color-neutral states referred to as *hadrons*. The two main types of hadrons, which are commonly observed, are mesons and baryons. Mesons are bound states of a quark and an antiquark, carrying a color and the corresponding anticolor. Since they are comprised out of two fermions, mesons exhibit an integer-number spin in units of $\hbar$, which makes them bosons. Examples of mesons are pions, which consist of combinations of up and down quarks. $\pi^+$ consist of a up and anti-down quarks, $\pi^-$ of down and anti-up quark and $\pi^0$ of a linear combination of up and anti-up and down and anti-down quarks. Baryons are bound states of three quarks, each with a different color charge. Since their spin adds up to half-integer

[1] The electron as well as the up and down quarks are part of the firs generation: Up and down quarks comprise protons ($uud$) and neutrons ($udd$). Atoms are made from protons, neutrons and electrons.

values in units of $\hbar$, they are also fermions. Examples are the proton and neutron. Further bound states, like "pentaquarks", are predicted by the SM and are a field of active research, for example by the LHCb collaboration at the Large Hadron Collider.

### 1.1.2. Interactions

The SM defines a Lagrangian density that combines the electromagnetic, strong and weak interactions in a single relativistic quantum field theory. The interactions are mediated by vector bosons with a spin of one in units of $\hbar$. The vector bosons of the SM, also called gauge bosons, are given in table 1.2.
Because the SM is formulated in the framework of a quantum field theory, particles and forces are described by fields. The Lagrangian density $\mathcal{L}$ describes the kinematics and interactions of the fields. Applying the Euler-Lagrange equation

$$\partial_\mu \left( \frac{\partial \mathcal{L}}{\partial \partial_\mu \phi} \right) - \frac{\mathcal{L}}{\partial \phi} = 0 \tag{1.1}$$

to a given Lagrangian density $\mathcal{L}$ returns the equations of motion for the field $\phi$. The Lagrangian density of the SM is invariant under transformations of the $SU(3)_c \times SU(2)_L \times SU(1)_Y$ symmetry group. This implies the conservation of three quantities (according to Noethers's theorem [40]). In the case of the SM, the invariance under $SU(3)_c$ yields the conservation of the color charge, and the invariance under $SU(2)_L \times U(1)_Y$ the conservation of the weak isospin $I_3$ and the hypercharge $Y$. These three gauge groups of the SM are responsible for three of the four fundamental forces: the strong force, the weak force, and electromagnetism; as described below.

#### Electromagnetic Interaction

The photon $\gamma$ is the mediator of the electromagnetic interaction. The photons couple to the electric charge of particles. Since the photon itself carries no charge, no direct interaction between photons, outside loops, is possible. Additionally the photon is massless. Both properties lead to the characteristic long range of the electromagnetic interaction, because photons traveling though space continue on their path until encountering a charged particle.
Quantum electrodynamics (QED) is the formulation of the electromagnetic force as a relativistic quantum field theory. Its Lagrangian density is given by

$$\mathcal{L}_{\text{QED}} = i\bar{\psi}\gamma^\mu \partial_\mu \psi - m\bar{\psi}\psi - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} - e\bar{\psi}\gamma_\mu A^\mu \psi \ . \tag{1.2}$$

The elements of this equation are: The Dirac matrices $\gamma^\mu$. The bispinor fields $\psi$ ($\bar{\psi} = \psi^\dagger \gamma^0$) representing spin 1/2 particles. The mass $m$ of the spin 1/2 particle. The photon field potential is represented by $A^\mu$ and the electromagnetic field strength tensor is given by the $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. Each term in the equation represents separate aspects of the particles and their interaction: The first and second term describe the kinematics and the mass of the particles, respectively. The third term expresses the kinematic properties of the photon field. The last term describes the interaction between the field and the spin 1/2 particles. In the case of the SM, the spin 1/2 particles are the fermions.

#### Strong Interaction

The gluons $g$ are the mediators of the strong interaction. Gluons couple to the three color charges red, green and blue of a particle. Unlike photons, gluons also carry color charge leading to the possibility of self-interaction. They carry a superposition of color

and anti-color charges, which results in eight distinct configurations. Even though the gluons are massless, their self-interaction leads to the short range of the strong interaction. Furthermore, this leads to the so-called *confinement* phenomena, which states that the force between two strongly interacting particles does not get weaker when they are separated. Instead the energy stored in the fields grows linearly with the distance and if the energy is large enough, new strongly interacting particle-antiparticle pairs are formed. This means that strongly interacting particles can only exist in color-neutral bound states called hadrons. The counterpart to confinement is *asymptotic freedom*. According to this concept, the strong coupling becomes weaker with increasing energy and the quarks can be viewed as quasi-free particles. In this region perturbation theory can be applied. Quantum chromodynamics (QCD) [26,27] is the formulation of the strong force as a relativistic quantum field theory. Its Lagrangian density is given by

$$\mathcal{L}_{\mathrm{QCD}} = i\bar{\psi}_a\gamma^\mu\partial_\mu\delta_{ab}\psi_b - m\delta_{ab}\bar{\psi}_a\psi_b - \frac{1}{4}G^A_{\mu\nu}G^{A\mu\nu} - g\bar{\psi}_a\gamma^\mu t^C_{ab}\psi_b A^C_\mu \tag{1.3}$$

This Lagrangian density exhibits a similar structure the QED Lagrangian density, given in equation (1.2): The bispinor fields $\psi_a$ represent the quarks with a color charge $a$, that can be any of the three colors. The eight distinct gluon fields are represented by $A^C_\mu$. The gluon field strength tensor $G^A_{\mu\nu}$ is constructed using the structure constants $f^{ABC}$ of the $SU(3)$ symmetry group as $G^A_{\mu\nu} = \partial_\mu A^A_\nu - \partial_\nu A^A_\mu + g f^{ABC} A^B_\mu A^C_\nu$. $t_{ab}$ are $3 \times 3$ matrices that represent the eight generators of the $SU(3)$ symmetry group. The final element is the quark-gluon coupling constant $g = \sqrt{4\pi\alpha_s}$, with the strong-coupling constant $\alpha_s$. As for the QED Lagrangian density, each term in equation (1.3) represents aspects of the particles and their interaction: The kinematics and the mass of the quarks are described by the first two terms. The third term expresses the kinematic properties of the gluon field including the self-interaction. The last term describes the interaction between the gluons and quarks.

### Electroweak Interaction

The electroweak unification is a central pillar of the SM as it is known today. Due to this unification, the electromagnetic and weak interaction are no longer described by two separate theories but by a single theory, which is invariant under $SU(2) \times U(1)$ symmetry group. This symmetry is also known as the electroweak symmetry. Three massless boson fields $W_1$,$W_2$, and $W_3$ are introduced. They couple to a new charge-type quantum number, called the weak isospin $I_3$. Additionally a single boson field $B$, which couples to the hypercharge $Y$, is introduced. These four fields lead to four massless particles. The configuration of these four independent fields is accurate above a certain energy scale. Below this energy scale, the electroweak symmetry is spontaneously broken by the Higgs mechanism, which will be discussed in section 1.1.3, leading to a mixing between these four fields and new particles. The four particles are the photon, the Z boson and the $\mathrm{W}^\pm$ bosons, which are the commonly known particles mediating the electromagnetic and weak interactions. Mixing the fields $B$ and $W_3$ according to the Weinberg angle $\theta_W$ results in the electrically neutral bosons

$$\begin{pmatrix}\gamma\\ \mathrm{Z}\end{pmatrix} = \begin{pmatrix}\cos\theta_W & \sin\theta_W\\ -\sin\theta_W & \cos\theta_W\end{pmatrix}\begin{pmatrix}B\\ W_3\end{pmatrix} \quad . \tag{1.4}$$

Both couple to a linear combination of the weak hypercharge $Y$ and weak isospin $I_3$. In the framework of the electroweak interaction, the electrical charge is given by $Q = Y/2 + I_3$. The positively and negatively charged W bosons $\mathrm{W}^\pm$ result from a mixing of the $W_1$ and

$W_2$ fields according to:

$$\mathrm{W}^{\pm} = \frac{1}{\sqrt{2}} \left( W_1 \mp i W_2 \right) \ . \tag{1.5}$$

The $\mathrm{W}^{\pm}$ and Z bosons have masses of 80.4 GeV and 91.2 GeV [29], respectively. In the context of elementary particles, these are very massive particles, leading to the short range of the weak interaction.

Interactions mediated via W bosons are accompanied with flavor transitions because of charge conservation: Charged leptons are converted to neutrinos (charge $\pm 1e \rightarrow 0$) and vice versa, up-type quarks are converted to down-type quarks (charge $\pm 2/3e \rightarrow \mp 1/3e$), and down-type quarks are converted to up-type quarks (charge $\pm 1/3e \rightarrow \mp 2/3e$). In the SM, quark transitions can also extend beyond the generation of the incoming quarks (e.g. from a up-type $u$ quark to down-type $s$ quark, compared to $u \rightarrow d$). The reason for this is, that the quark flavor eigenstates of the weak interaction are not equal to the mass eigenstates. How these two states are mixing is given by the CKM matrix [41, 42]

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} \ , \tag{1.6}$$

with the electroweak flavor eigenstate $q'$ and the mass eigenstates $q$ of the quarks. The values $V_{xy}$ of the CKM matrix can be understood as a transition frequency for a given process. Mathematically, the elements are included in the calculation of the transition amplitude for a given vertex with quark contribution. Generally speaking, the diagonal elements $V_{ud}$, $V_{cs}$, and $V_{tb}$ are close to one and off-diagonal elements are small. Therefore, transitions within a generation are favored. For this thesis, especially the top-quark transitions are important. The currently best measurement of $V_{tb}$ is $0.999172^{+0.000024}_{-0.000035}$ [29]. This means that (in experimental practice) all top-quarks decay to bottom quarks and W bosons.

The Lagrangian density after the symmetry breaking is given by

$$\begin{aligned} \mathcal{L}_{EW} = \quad & \bar{\psi} \left( i\gamma^{\mu}\partial_{\mu} - m \right) \psi - g \sin\theta_W Q \bar{\psi}\gamma^{\mu}\psi A_{\mu} \\ & - \frac{g}{2\sqrt{2}} \bar{\psi}\gamma^{\mu}(1-\gamma^5)(T^{+}W_{\mu}^{+} + T^{-}W_{\mu}^{-})\psi \\ & - \frac{g}{2\cos\theta_W} \bar{\psi}\gamma^{\mu}(g_V - g_A\gamma^5)\psi Z_{\mu} \ . \end{aligned} \tag{1.7}$$

The fields $A_{\mu}$, $Z_{\mu}$, and $W_{\mu}^{\pm}$ for photons, neutral boson and charged bosons are defined as given by equation (1.4) and (1.5). The operators $T^{\pm}$ are the weak isospin raising and lowering operators. The factors $g_V$ and $g_A$ are vector and axial-vector couplings, respectively. The first term describes the kinematics and mass of the fermions. The second term describes the electromagnetic interaction (QED), as also given in equation (1.2). The third term represents the weak charged-current interactions and the last term describes the weak neutral-current interactions.

### 1.1.3. Higgs mechanism

As already mentioned in section 1.1.2 in the discussion of the electroweak interaction, the Higgs-mechanism (or more precisely the Anderson-Brout-Englert-Higgs-Mechanism) [14–16] is required to describe the mass of the particles within the SM Lagrangian density. Equation (1.7) already includes these mass terms. The Higgs mechanism will be be discussed in more detail in this section.

The introduction of a scalar filed $\phi$, which couples to the particle mass, is the first step in the Higgs mechanism. The filed $\phi$ is also called the *Higgs field*. It is represented by a complex $SU(2)$ doublet and therefore can be described by four real parameters. Furthermore, the corresponding potential

$$V(\phi) = -\mu\phi^\star\phi + \lambda(\phi^\star\phi) \ , \tag{1.8}$$

also called the *Mexican hat potential*, is introduced. For positive values of $\mu$ and $\lambda$, this potential features a global minimum at the vacuum expectation value $|\phi| = v = \sqrt{\mu/(2\lambda)}$ and a local maximum at $\phi = 0$. The Higgs field takes on different values depending on the energy scale. At a large energy scale, the whole potential including its symmetries are important and conserved. However, at low energy scales, the Higgs field will be confined to a valley of the potential near the global minimum. In this case the effective potential of the field can be given by approximating the potential around a single point $|\phi|$ in the valley and the symmetry of the full potential is no longer apparent; it is said to be "broken".

In the following, the general concept of the Higgs-mechanism will be explained based on a simplified Higgs field $\phi = \phi_1 + i\phi_2$ and a massless gauge field $A^\mu$, based on reference [43]. Using the potential given in equation (1.8), the Lagrangian density for this field can be formulated as

$$\mathcal{L} = \frac{1}{2}\left[\left(\partial_\mu - iqA_\mu\right)\phi^\star\right]\left[\left(\partial^\mu - iqA^\mu\right)\phi\right] + \mu\phi^\star\phi - \lambda(\phi^\star\phi) - \frac{1}{16\pi}F^{\mu\nu}F^{\nu\mu} \ , \tag{1.9}$$

with the field strength tensor $F^{\nu\mu}$ of the gauge field $A^\mu$. By introducing $\phi_1 = \eta - \sqrt{\mu/(2\lambda)}$ and $\phi_2 = \xi$, a parameter transformation is performed such that the field is described at the location of a chosen global minimum. Applying this transformation to the Lagrangian density in equation (1.9), results in terms describing a massless Goldstone boson arising from the field $\xi$ and couplings of this field to the gauge field $A^\mu$. The Goldstone theorem [44–46] predicts one massless bosons for each broken symmetry and correspondingly in this case there is one massless gauge field for the single broken symmetry. By applying further transformations that exploit gauge invariance, the contribution of $\xi$ can be eliminated and the following Lagrangian density can be formulated:

$$\begin{aligned}\mathcal{L} = \quad & \left[\frac{1}{2}(\partial_\mu\eta)(\partial^\mu\eta) - 2\mu\eta^2\right] \\ & + \left[\frac{1}{16\pi}F^{\mu\nu}F^{\nu\mu} + \frac{\mu}{4\lambda}A_\mu A^\mu\right] \\ & + \left[\sqrt{\frac{\mu}{2\lambda}}q^2A_\mu A^\mu + \frac{1}{2}q^2\eta^2A_\mu A^\mu - \sqrt{8\lambda\mu}\eta^3 - \lambda\eta^4\right] \\ & + \frac{\mu^2}{16\lambda} \ .\end{aligned} \tag{1.10}$$

This Lagrangian density exhibits the following features: The term in the first line of equation (1.10) corresponds to a new massive spin-zero particle, which we identify with the Higgs boson. The term in the second line describes the kinematics and dynamics of the gauge field and includes a mass term. The term in the third line represent the interaction of the gauge field and the Higgs field, parametrized by the coupling strength

$$\lambda_V = 2\frac{m_V^2}{v} \ , \tag{1.11}$$

as well as the self interaction of the Higgs field.

Applying analogue transformations in the SM, leads to non-zero masses for the vector bosons Z and $W^\pm$. In order to also describe massive fermions, the interaction of the Higgs

field and the fermions has to be added to the SM Lagrangian density. This is accomplished by adding terms, which transform like a singlet under the operation of the $SU(2) \times U(1)$ symmetry group, to the Lagrangian density:

$$\mathcal{L}_{\text{Yukawa}} = -y_f \left( \bar{\psi}_L \phi \psi_R + \bar{\psi}_R \bar{\phi} \psi_L \right) \ . \tag{1.12}$$

This is the Yukawa interaction with the so-called Yukawa coupling, which is given by

$$y_f = \frac{\sqrt{2} m_f}{v} \ . \tag{1.13}$$

By adding the interaction between the Higgs field and the fermion fields, the fermions acquire a non-zero mass for a non-zero vacuum expectation value. Understanding the properties of the coupling between the top quark and the Higgs fields, the so-called Top-Yukawa coupling, is one of the main motivations for pursuing measurements of the $t\bar{t}H$ and tH processes. This will be discussed further in section 1.3.

## 1.2. Hadron collider phyiscs

Analyzing the outcome of high-energy collisions of elementary particles is the main source of our knowledge of the sub-nuclear domain. Hereby, the general experimental setup did not change significantly since the formulation of the SM and the consequent first measurements in the 1960's. Particles are accelerated and brought to collision at the so-called interaction point. Because of the large energies in such collisions, new particles, like the Higgs boson, can be produced. Around the interaction point, detectors are placed with the goal of measuring all resulting particles and reconstructing the collision and physical processes. Colliding hadrons, as it is done at the LHC, instead of elementary particles, like electrons, has the advantage that much higher energies can be achieved since hadrons loose less energy while being accelerated, e.g. through synchrotron radiation. The main drawback is the more complicated theoretical and experimental description because of the composite nature of the hadrons. In the context of this thesis, only proton collisions are studied, but the following concepts apply to all hadron collisions.

The asymptotic freedom of the strong interaction and the confinement are essential concepts in the context of hadron collider physics. Asymptotic freedom enables the use of perturbation theory for the calculation of processes in the context of the SM and the confinement is the reason that only composite objects like hadrons instead of the elementary quarks can be observed at hadron colliders. Characterizing these composite objects requires the description of processes like the substructure of the proton or the formation of hadrons, which cannot be done from first principles within the SM. It is nevertheless possible to probe the SM above certain energy scales because the collisions can be factorized into mostly independent components for which a model exists or that can be calculated. The models are usually parametric and can be determined, tested and tuned using independent measurements.
This section will outline relevant hadron collider physics concepts necessary for this thesis: cross sections, Feynman diagrams, renormalization, calculation and simulation of the final states at hadron colliders, collider kinematics and jets.

### 1.2.1. Cross sections and decays

The scattering process of two beams of particles colliding head-on is usually described in terms of the cross section $\sigma$. The rate $\dot{N}$ at which a process occurs is given by the process specific cross section and the luminosity $L$, describing the beam particle flux

$$\dot{N} = L \cdot \sigma \ . \tag{1.14}$$

In practice, the cross section is usually measured using by counting the number of scattered particles produced in the collision. In a time-interval $[t_a, t_b]$, the number of expected occurrences of some process is given by:

$$N = \int_{t_a}^{t_b} L\sigma dt = \mathcal{L}\sigma \ , \tag{1.15}$$

with the integrated luminosity $\mathcal{L}$. Usually the cross section depends on the initial energy of the colliding particles. The probability of a particle $x$ to decay to a final state $y$ per unit time is expressed in terms of the decay rate $\Gamma_{x\to y}$, which is also the inverse of the particle lifetime. Often, particles can decay into multiple different final states. Therefore, the total decay rate is defined as:

$$\Gamma_{x,\text{tot}} = \sum_i \Gamma_{x\to i} \ . \tag{1.16}$$

The branching fraction

$$\mathcal{B}_{x\to y} = \frac{\Gamma_{x\to y}}{\Gamma_{x,\text{tot}}} \ , \tag{1.17}$$

is used to describe the probability of a particle (or initial state) $x$ to transition to the final state $y$.

### 1.2.2. Feynman diagrams, renormalization, and the running coupling

Feynman diagrams are used to graphically represent terms in the calculation of cross sections and decay rates. These calculations are typically performed in a perturbation framework, where the full results is approximated by a series of terms, with each term having a corresponding Feynman diagram representation. The first term is denoted as the *leading order* (LO), the second term as the *next-to-leading order* (NLO), and so forth. Leading order diagrams always represent the graph with the smallest number of vertices, i.e. points of particle interaction, possible. Since the number of vertices increases with the number of orders considered, as are the possibilities how vertices can be added to the Feynman diagram, the complexity of these calculations can increase rapidly. Approximating a process by omitting higher orders is useful, because each vertex in a diagram is associated with a multiplication by a coupling constant in the calculation. In case of hadronic collision, which are described by QCD, the perturbation expansion is given in orders of the strong coupling constant $\alpha_s$. In the regime where these perturbative calculations are performed, the couplings are small and the largest, most important, terms correspond to the diagrams having the fewest number of vertices. Using higher order predictions is still important for precise measurements, e.g. for predicting the yield of a certain simulated background or for modeling underlying phenomena that only manifest in higher orders.
Another important concept are loops: Feynman diagrams can contain particle loops, e.g. a particle pair that is emitted and absorbed within the process, that cannot be directly observed in the final state of the process. Nevertheless they represent important phenomena. For example the dominant Higgs-boson production mechanism via gluon fusion is only possible because of a loop process involving top quarks (c.f. section 1.3.1 and figure 1.4a). Figure 1.1a shows the $q\bar{q} \to q\bar{q}$ process at LO and Figure 1.1b shows a diagram contribution to higher order terms in the calculation with a fermion-pair loop.

The evaluation of diagrams at higher orders containing loops requires the calculation of divergent integrals. By redefining the fields and constants of the theory, a process called renormalization, these divergent integrals can be absorbed and regularized, in such a way that the infinities are not present, or cancel out and the final result of the calculation becomes finite. An a-priori defined energy scale, called the *renormalization scale* $\mu_R$, is




(a) (b)

Figure 1.1.: Feynman graphs for $q\bar{q} \to q\bar{q}$ in leading order (a) and next-to-leading order with a fermion loop (b).

required in this case. For the strong coupling $\alpha_s$, the dependency on the renormalization scale $\mu_R$ can be written at LO as

$$\alpha_s(\mu)^2 \approx \frac{1}{b_0 \log\left(\frac{\mu^2}{\Lambda_{\mathrm{QCD}}}\right)} \ , \tag{1.18}$$

with $b_0 > 0$ and $\Lambda_{\mathrm{QCD}} \approx 200\,\mathrm{GeV}$. This *running coupling* also implies the asymptotic freedom of QCD. The coupling $\alpha_s$ becomes small at high energies and the perturbation theory can be applied. At low energies (close to to $\Lambda_{\mathrm{QCD}}$), perturbation theory is not valid, the coupling becomes strong and the quarks cannot exists freely but are bound.

### 1.2.3. Simulation of the hard process and final states

At high energy scales, the hard processes can be calculated using perturbative QCD. The starting point is the Matrix Element $\mathcal{M}$, which encodes the underlying theory described by the SM and can be systematically calculated at fixed orders of $\alpha_s$. In reality the descriptions of these processes with such perturbative calculations reaches its limits - for a variety of reasons - quite fast. The two major reasons are: First, most final states exhibit many final state particles, leading to very complex calculations. Second, the energy scale decreases due to particles radiation, which results in stronger $\alpha_s$, making perturbation theory invalid. Additionally the proton structure, the parton shower, and the hadronization need to be modeled. Therefore, additional techniques are necessary when calculating processes, as they occur at the LHC. This section loosely follows reference [47] chapter 2 and outlines the necessary concepts and techniques used to calculate and simulate final states at hadron colliders.

The phase-space specific integrals of the scattering, which is necessary to make predictions or simulate processes, are usually calculated using Monte Carlo (MC) techniques. These techniques are used because of their superior convergence speed compared to other techniques for complex high-dimensional integrals. MC techniques are used for various applications when simulating hadronic collisions. This includes both the short distance predictions of the hard process and long-distance phenomena like the hadronization. These simulations are collectively referred to as *Monte Carlo event generation*. A general introduction on the topic of Monte Carlo event generation and their application in hadron-collider physics can for example be found in reference [47] chapter 3 or reference [48].

#### 1.2.3.1. Factorization and parton distribution functions

Since protons are not fundamental particles, their substructure needs to be considered when they collide. The quarks comprising the proton constantly interact with each other via gluon absorption and emission and the intermediate gluons do the same with quarks and gluons. Therefore, the proton is actually a "cloud" of quarks and gluons (also called

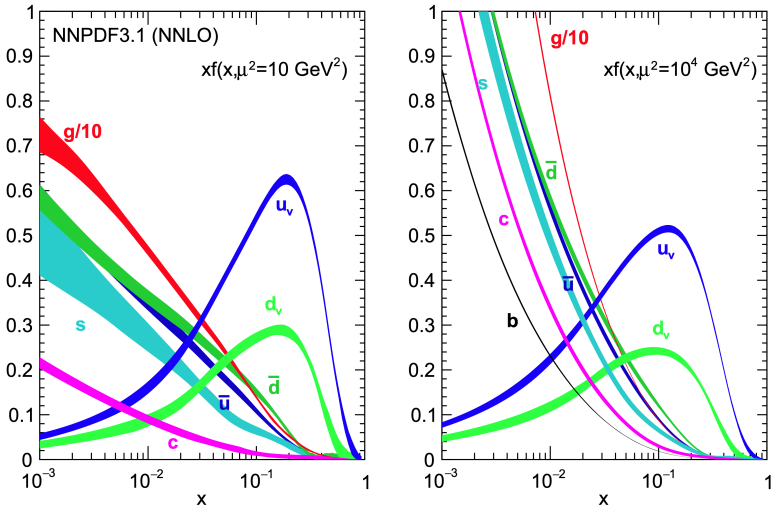



Figure 1.2.: Parton density functions for quarks and gluons at two different factorization scales as a function of the momentum fraction $x$ for the `NNPDF3.1` PDF-set [50]. At higher energies $Q$ (right compared to left), the fraction of anti-quarks, heavier quarks and gluons carrying higher fractions of the proton momentum $x$ increase.

constituents). Hereby additional quarks, or valence quarks, inside the proton can be of any flavor. According to the *factorization theorem* [49], the hard process can be separated from these low-energy scale-dynamics inside of the proton. Hereby the incoming partons (in the case of the LHC the quarks and gluons constituting the protons) are essentially free, each carrying the momentum fraction $x_i$ of the proton momentum $p_i$. The low-energy dynamics inside the proton are separated from the high-energy dynamics of the hard process by introducing a *factorization scale* $\mu_F$. The cross section of two colliding protons yielding the final state $f$ can then be calculated using:

$$\sigma(p_1, p_2 \to f) = \sum_{ji} \int_0^1 dx_i \int_0^1 dx_j f_i(x_i, \mu_F^2) f_j(x_j, \mu_F^2) \hat{\sigma}(\hat{p}_i, \hat{p}_j \to f) \ . \tag{1.19}$$

The parton distribution function (PDF) $f_i$, describes the probability for finding a certain parton $i$ with the momentum fraction $x_i$. Up to the scale $\mu_F$, gluon splittings and emissions as well as gluon to quark splittings are absorbed into the PDF. Because PDFs describe universal properties of the proton, they can be determined in independent measurements. Figure 1.2 shows examples for PDFs at different energies. The cross section $\hat{\sigma}(\hat{p}_i, \hat{p}_j \to f)$, with the momentum $\hat{p}_i$ of the parton originating from proton $p_1$ and momentum $\hat{p}_j$ of the parton originating from proton $p_2$, can be calculated using perturbation theory.

#### 1.2.3.2. Parton shower and hadronization

Quarks and gluons interact via the strong interaction and emit or collinearly split into further quarks and gluons producing a color-charged shower, more commonly referred to as parton shower (PS). Similar to the factorization into low-energy dynamics inside the proton and the high-energy hard process, this low-energy shower can be factorized from the high-energy hard process. The process is usually implemented using MC shower generators, which sequentially add lower and lower energy emissions to the final state until a cutoff is reached. Similarly initial state radiation is included.
The approximations made by the PS work very well at the soft and collinear limit but not for hard and large-angle radiation. This is modeled much better by the matrix el-

ement calculations. If one wants to combine both, the main practical issue is to avoid double counting configurations that can either originate from a higher order ME or the PS. Various methods are available to achieve this and are considered in the simulated samples used for the analysis presented in part II of the thesis. Generally, this matching (or merging) is achieved by either splitting the phase space into two regimes (jet evolution and jet production) at a certain energy scale or by generating ME's with multiple exclusive numbers of emissions and vetoing these configurations produced by the PS. In the process of simulating collisions, the PS is stopped once all partons are at the energy scale of hadronization ($\sim 1\,\mathrm{GeV}$). The process of hadronization is the evolution of a set of colored partons to a set color singlet primary hadrons. How this mapping is accomplished is encoded in models, that are motivated by the structure of QCD but still contain a number of free parameters that can be tuned. Different MC parton shower generators implement different hadronization models. For example PYTHIA [51], which is used as parton shower generator for all simulated samples in this thesis, implements the so-called string model. After the primary hadrons are formed, their decay is simulated until only stable particles remain.

#### 1.2.3.3. Underlying event and pileup

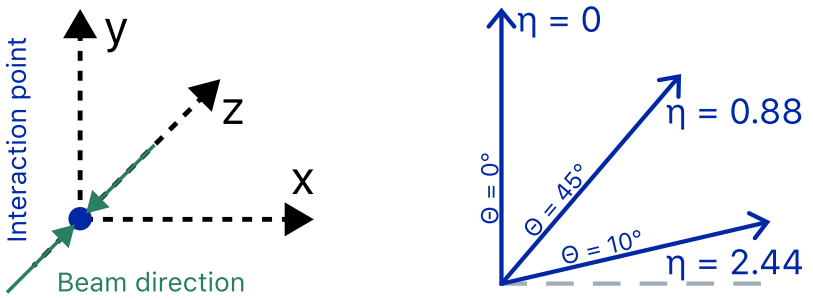
Due to the fact that composite particles collide, multi-parton interaction are always present at hadron colliders. These interactions, occurring in addition to the hard process, along with beam remnants are referred to as *underlying event*. It is usually simulated as multiple soft parton-parton interactions. Large non-perturbative effects, that can also affect the hard process, are present. Therefore, this simulation has to be tuned extensively from data measurements.
The term *pileup* refers to the effect of proton-proton collisions that occur in addition to the collision of the hard process. Two different kinds of pileup can occur at the hadron colliders: *In-time pileup* results from multiple protons colliding in the same bunch crossing. *Out-of time-pileup* occurs due to finite detector response time, which can lead to recording collision from previous and following bunch crossings. At CMS a technique called *premixing* is used to simulate both processes. The pileup is included in the event simulation by overlaying events solely reflecting the expected pileup and without the hard process.

### 1.2.4. Simulation of the detector response

The measurements in particle physics are conducted with complex machines that impart certain responses on the recorded collisions, as it will be explained in more detail in section 2.2. In order to compare the measurement to simulation, the detector response, i.e. the interaction of the final state particles with the detector, needs to be simulated as well. This includes the interaction with detector material and magnetic field as well as the simulation of the hits in the tracker and energy deposits in the calorimeters. The readout of the electronics has to also be simulated. Using such a procedure enables the reconstruction of the simulated events with the exact same procedures as are used to interpret the collected data. A commonly used toolkit is GEANT4 [52], which is also used for the CMS experiment.

### 1.2.5. Collider kinematics

For collider measurements, a coordinate system in which the z-axis is aligned with the beam axis, as illustrated in figure 1.3a, is used. At hadron colliders, the interacting partons can carry very different fractions of the proton momenta. The final state particles produced in the collisions tend to be boosted along the z-axis. While these final state particles are similar in their respective center-of-mass frame, they can seem very different when




(a) (b)

Figure 1.3.: Illustrations of the coordinate-system generally used at hadron collider experiments (a) and the relation between pseudorapidity $\eta$ and the angle in the transversal plane $\Theta$ (b).

described in terms of the total momentum and energy due to their respective boost in the lab frame. Furthermore, an unknown fraction of the beam energy in an event can escape through the beam axis. Hence, quantities invariant under longitudinal Lorentz boosts are generally preferred.
The transverse momentum is given by

$$p_{\mathrm{T}} = \sqrt{p_x^2 + p_y^2} \ , \tag{1.20}$$

and the transverse mass by

$$M_{\mathrm{T}} = \sqrt{E^2 - p_z^2} \ . \tag{1.21}$$

Both are Lorentz invariant under boosts along the beam axis. The rapidity $y$ is defined as

$$y = \frac{1}{2} \ln \left( \frac{E + p_z}{E - p_z} \right) \ . \tag{1.22}$$

Differences in rapidity are invariant under longitudinal Lorentz boosts but generally the pseudorapidity

$$\eta = -\ln \left( \tan \left( \frac{\theta}{2} \right) \right) \ , \tag{1.23}$$

with $\theta$ as the angle to the z-axis, will be used. The pseudorapidity is equal to the rapidity if the kinetic energy of the particle is large compared to its mass. Because its near invariance under longitudinal Lorentz boosts, is is common practice to use $\eta$ instead of the angle $\theta$. The relation between these two quantities is illustrated in figure 1.3b. Another commonly used observable is the angular separation between two objects (partons, particles, jets, etc.)

$$\Delta R = \sqrt{(\phi_1 - \phi_2)^2 + (\eta_1 - \eta_2)^2} \ . \tag{1.24}$$

### 1.2.6. Jets

Due to QCD confinement, quarks and gluons cannot be observed directly. Instead they generate, through the strong interaction, collimated streams of particles, which are called jets. Experimentally these jets can be reconstructed by clustering the observed particles (or more precisely the tracks and energy depositions in the detector). The resulting jets are used to draw conclusions about the kinematic properties of the quarks and gluons which originated the particle spray. The calculations of perturbative QCD can also be expressed in terms of jets. Since soft and collinear emissions cannot be reliably predicted

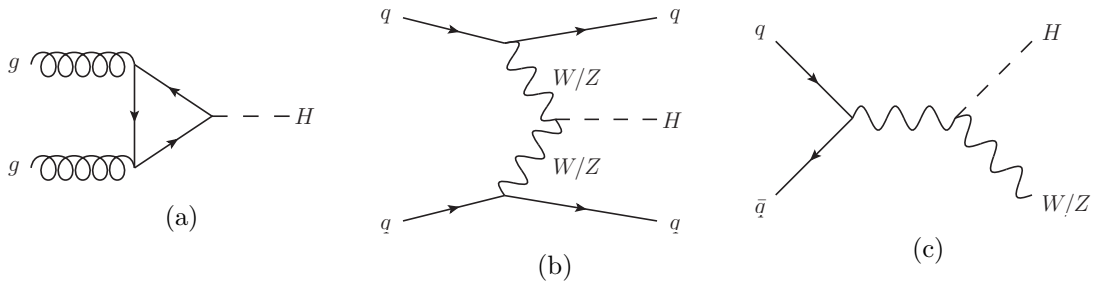



Figure 1.4.: Example Feynman diagrams for different Higgs boson production modes: Gluon fusion (a), vector boson fusion (b) and Higgs strahlung (c). Drawn using `JAXODRAW` [54].

by these calculations, the jet clustering should ideally be insensitive to both.
The process of clustering jets is determined by a set of rules describing how to group the particles and assign momenta to the resulting jets. These algorithms should be applicable to partonic calculations, experimental measurements and parton-showering Monte-Carlo simulation. This enables a common treatment of jets in theory, experiments and simulation. Jets that are derived from intermediate particles in a simulation are also referred to as generator-level or particle-level jets. An overview of jet clustering algorithms can be found in reference [53]. The algorithm used for the measurement presented in part II of this thesis will be covered in detail in section 2.3.5.

Jets play an outstanding role at hadron colliders in general, since quarks and gluons are participating in all collisions, and for the analysis presented in part II of this thesis in particular, since it aims to measure an all-jet final state.

## 1.3. Higgs-boson physics

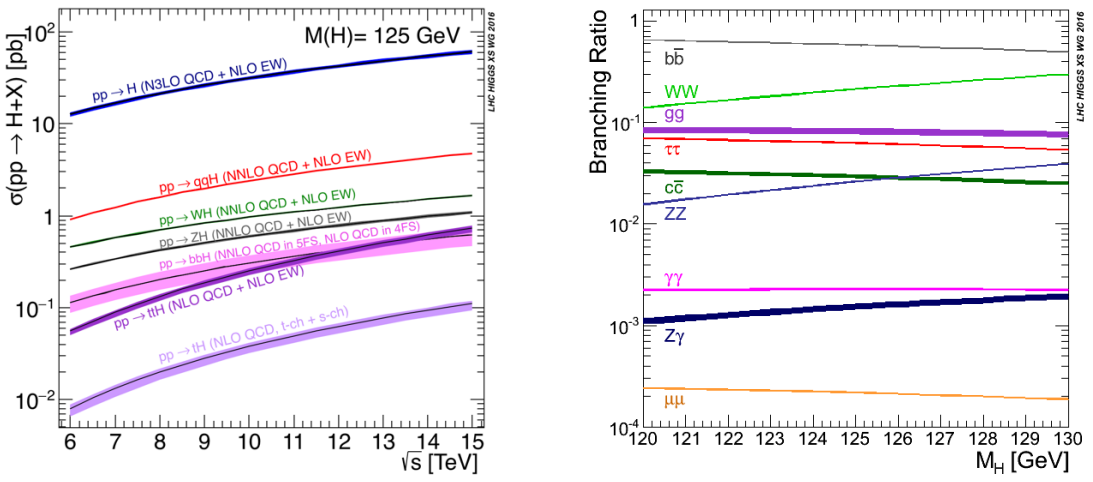
This section serves as an introduction to Higgs-boson physics, phenomenology and experimental status at the LHC. The $t\bar{t}H$ production mode, especially with the decay of the Higgs boson to bottom quarks, will be discussed in more detail, since it is the main focus of the work presented in this thesis. Details on the experimental status will be covered in section 1.3.5.

### 1.3.1. Higgs-boson properties

At the LHC, the Higgs boson is produced in four main production modes:

**Gluon fusion** The dominant production mode at the LHC is shown in figure 1.4a. In this mode, the Higgs boson is produced via a fermion-quark loop from the interaction of two gluons. Due to the virtual nature of the loop, the corresponding coupling to the Higgs boson cannot be measured directly. This mode is also referred to as ggH.

**Vector boson fusion** In this process, which is shown in figure 1.4b, the Higgs boson is produced by two vector bosons radiated from incoming quarks. This mode depends on the coupling of the W, Z and Higgs bosons and is commonly referred to as VBF production. The jets originating from the final state quarks exhibit a comparatively large distance in $\eta$ ("forward jets") and large invariant mass, which can be used to identify this production mode in the detector.




(a) (b)

Figure 1.5.: Production cross sections for a Higgs Boson, with a mass of 125 GeV, as function of center-of-mass energy at the LHC and Higgs-boson branching fractions as function of Higgs-boson mass. Taken from reference [55]

**Higgsstrahlung** The production of the Higgs boson in association with a W or Z boson is shown in figure 1.4c. This mode is also referred to as Higgs boson production in association with vector bosons (VH).

**$t\bar{t}H$** The production of a Higgs boson in association with a top-quark pair. Due to the large mass of the final-state particles it has the smallest cross section of the four main production modes. This process will be discussed in detail in section 1.3.2.

Figure 1.5a shows the production cross sections for these four processes as well as the associated production with bottom quarks ($b\bar{b}H$) and the single-top Higgs (tH) production for a Higgs boson with a mass of 125 GeV as function of the center-of-mass energy $\sqrt{s}$ . The cross-section dependency on $\sqrt{s}$ for the top-quark associated Higgs-boson production modes is large compared to the other modes.

Due to its coupling to leptons, quarks and the W and Z bosons, the Higgs boson has a wide array of possible decay channels. An overview of the branching fractions of the possible decay modes is given in figure 1.5b. Due to the mass dependency of the coupling, decays with more massive particle are preferred, if they are not restricted by phase-space factors.

The decays can be categorized into vector boson, fermion and loop induced decays:

### Vector boson decays

The Higgs boson couples strongly to the Z and $W^{\pm}$ bosons due to their large mass. Nevertheless, decays to two on-shell vector bosons are suppressed because of the mass of the Higgs boson. The branching fractions are still quite sizable with $\approx 21\,\%$ and $\approx 2.6\,\%$ for the WW and ZZ decays, respectively [55]. In combination with further subsequent leptonic decays, these decay channels can be used to identify the VH production mode. The decay to a pair of Z bosons plays an outstanding role the LHC, due to very clean experimeental signature and its role in the Higgs boson discovery [1, 2].

### Fermion decays

Since the Higgs boson couples to fermions via Yukawa couplings, direct decays are possible. Because of the direct mass dependency of this coupling, as given in equation (1.13), heavier

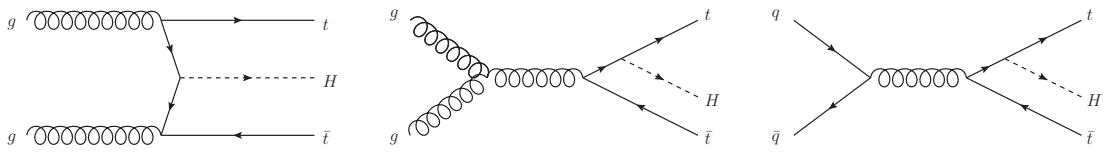



Figure 1.6.: Example LO Feynman diagrams for the $t\bar{t}H$ production process. Drawn using JAXODRAW [54].

fermions are preferred. The most common decay modes are the $b\bar{b}$ and $\tau^+\tau^-$ modes, with branching fractions of $\approx 58\,\%$ and $\approx 6\,\%$, respectively [55]. The decays to quarks are generally hard to distinguish from the typical background at hadron colliders. Therefore, they are most commonly analyzed in the context of vector bosons or top-quark associated production modes, since the distinct characteristic of the associated production modes enable separation from the background.

#### Loop induced decays

Gluons and photons do not couple to the Higgs boson directly since they are massless. Decays involving both are nevertheless possible, if loops involving massive particles are considered. The decay to a gluon pair, which is induced by quark loops, is impossible to detect at hadron colliders and therefore irrelevant in the context of the experimental program at the LHC. The decay to photons (or a photon and a Z boson) can be induced by W-boson loops as well as fermion loops. Because of its very clean experimental signature and low backgrounds it has been one of the discovery modes of the Higgs boson even though its branching fraction is very low ($\approx 0.2\,\%$ for $\gamma\gamma$ and $\approx 0.1\,\%$ for $Z\gamma$ [55]) compared to the other described decay modes.

### 1.3.2. Associated production with top quarks

The production of a Higgs boson in association with a top-quark pair was observed by the ATLAS [8] and CMS [7] collaborations at the LHC.

As described in section 1.1.3, the coupling strength of the Higgs boson to charged fermions is characterized by the mass-dependent Yukawa-coupling $y_f \sim m_f/\nu$, with $\nu = 246\,\text{GeV}$. Since the top quark is far more massive than any other SM particle, the coupling to the Higgs boson, $y_t$, is expected to be the largest Yukawa coupling. On-shell top quarks are too heavy to be produced directly in the Higgs-boson decay. Therefore, the coupling has to be obtained by measuring the modes where the Higgs boson is produced in association with top quarks. This makes the measurement of the $t\bar{t}H$ cross section the only way to extract $y_t$ directly, since other measurements extract the coupling only indirectly via loop processes. In addition, detailed understanding of the interaction between top quarks and the Higgs boson is instrumental in testing the SM and constraining models for BSM physics which predict different $y_t$ coupling strength. Example LO Feynman diagrams for the $t\bar{t}H$ production process are given in figure 1.6.

In the final state, the top quarks cannot be observed directly since the top-quark lifetime is extremely short. Top quarks decay almost exclusively to a W boson and a bottom quark, as described by the CKM matrix (c.f. section 1.1.2). The W boson couples with equal strength to all weak isospin doublets of the SM but due to the number of color charges, hadronic decays are three times as likely as leptonic decays. Decays to third generation quarks are kinematically impossible and therefore the ratio of leptonic to hadronic decays is three to six. The $t\bar{t}H$ final states can be categorized by the decays of the two top quarks. The experimentally relevant decay modes are: The dileptonic mode (both W bosons decay

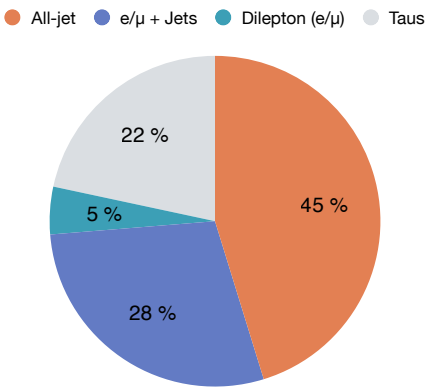



Figure 1.7.: Relative size of $t\bar{t}$ decay channels based on the decays of the W bosons of the $t \to Wb$ decay. Branching fractions are taken from reference [29].

to $(e/\mu)$ leptons), the semileptonic mode (one W boson decays to $(e/\mu)$ leptons while to other decays to quarks) and fully-hadronic mode (both W bosons decay to quarks). These three modes can be viewed as a tradeoff between purity and expected yield. With the dileptonic mode as purest mode with smallest yield and the fully-hadronic channel as impurest mode with larges yield. The relative sizes of the $t\bar{t}$ decays are visualized in figure 1.7.

All measurements, available at the time of writing the thesis, suggest the Higgs-boson coupling to top quarks is SM-like. Due to the indirect nature of most measurements (top quarks are only present in quantum loops) or the large uncertainties, as in the case of the $t\bar{t}H$ measurement, it is still possible that significant non-SM particles could be hidden in loops. Only more precise measurements, especially of the $t\bar{t}H$ process, can provide evidence for the absence or presence of such non-SM contributions.

The measurement of $t\bar{t}H$ includes most of the Higgs-boson decays shown in figure 1.5b. The multilepton final states ($\mu^+\mu^-$, $e^+e^-$, $\tau^+\tau^-$), the $H \to \gamma\gamma$ channel, and the $H \to b\bar{b}$ channel are the main channels with significant contributions to the overall result. Other channels, like $H \to c\bar{c}$ , with good prospects for later stages of the LHC are under investigation but are currently not sensitive enough due to experimental challenges, like very small production cross-sections or the requirement for better detector technology.

### 1.3.3. The $t\bar{t}H(b\bar{b})$ channel

The analysis presented in this thesis focuses on the $t\bar{t}H(b\bar{b})$ channel where both W bosons decay hadronically. This section will introduce the theoretical foundations for $t\bar{t}H(b\bar{b})$ in general and the fully-hadronic decay channel in particular.

Different combinations of Higgs-boson production and decay modes result in unique challenges and possibilities for the measurement. The predicted production cross section of the $t\bar{t}H$ process, with approximately 0.5 pb at 13 TeV[2], is small compared to the dominant Higgs-boson production channels. The cross-section of the dominant gluon-gluon fusion production mode (ggH) is approximately 100 times as large [55]. For this reason, Higgs-boson decay modes with a very pure experimental signature but small branching fraction, like the decay to four leptons or a pair of photons, are not the first choice to measure $t\bar{t}H$ production. For a Higgs boson with a mass of 125 GeV, the expected branching fraction of the $H \to b\bar{b}$ decay mode is $(0.58 \pm 0.02)$[55]. Due to this, the $H \to b\bar{b}$ decay mode is a very attractive channel for the measurement of $t\bar{t}H$. Figure 1.8a shows an example Feynman graph for the $t\bar{t}H(b\bar{b})$ process. The final state exhibits four bottom quark as well as up to

[2] This results in roughly 70000 $t\bar{t}H$ events during the four-year operation of the LHC Run-2.

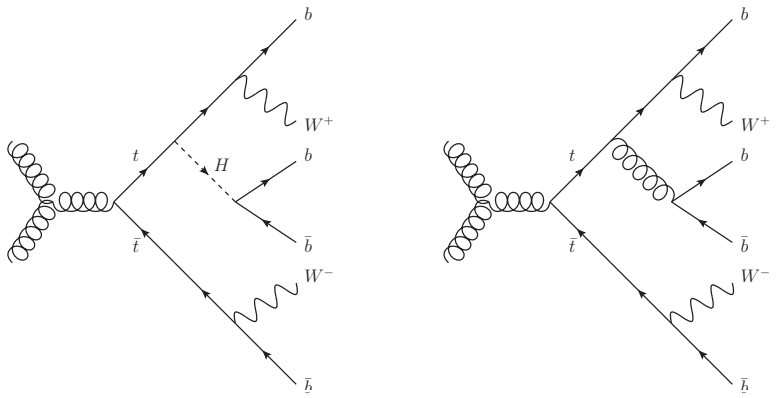



Figure 1.8.: Topology of $t\bar{t}H(b\bar{b})$ and the $t\bar{t} + b\bar{b}$ processes. Drawn using `JAXODRAW` [54]

four leptons and lighter quarks from the W-boson decays. Identifying the bottom quarks and understanding their properties is a vital part of the $t\bar{t}H(b\bar{b})$ measurement.

The main goal of the work presented in the thesis is to measure the cross section of $t\bar{t}H(b\bar{b})$ from data collected at the LHC. This is achieved by extracting the magnitude of the signal relative to the SM expectation using statistical methods, which will be introduced in section 3.3 and covered in more detail in section 6.1.

### Backgrounds

Multiple other processes can produce final states with very similar kinematic properties. Due to the rarity of the $t\bar{t}H$ process, background processes are far more likely to be observed. Some have experimental signatures that resemble the signal so closely that they can be not be distinguished. These backgrounds are also called *irreducible backgrounds*. A detailed understanding of these background processes is a vital ingredient in any $t\bar{t}H(b\bar{b})$ measurement. Processes with significant contributions are listed below:

**$t\bar{t}$ + jets** This background includes all $t\bar{t}$ processes in which initial or final-state radiation produces an additional quark-pair. The irreducible $t\bar{t} + b\bar{b}$ process, shown in figure 1.8b, poses a limiting factor to every $t\bar{t}H(b\bar{b})$ measurement. Other $t\bar{t}$ processes with additional charm or light quarks also pose a significant contribution but can be separated from the signal more easily. Since the $t\bar{t}$ process itself is very difficult to measure, large theoretical and systematic uncertainties for these background processes have to be considered in the $t\bar{t}H$ measurement. More details on this background and how it is modeled are given in section 4.2.3.

**QCD multijet** The production of multiple quarks or gluons via processes only involving the strong-interaction, collected under the term QCD multijet process, plays a non-negligible role. While it is unlikely that a QCD multijet final state exhibits as many bottom quarks as expected in the signal, the sheer scale of such processes at hadron colliders makes it likely to encounter them anyways. Requiring at least one of the W bosons to decay to a lepton-neutrino pair can be used suppress this contribution significantly.

**Single-top** The cross section of the single top-quark production (Single-top) is larger than the $t\bar{t}H(b\bar{b})$ cross section. While this should result in a

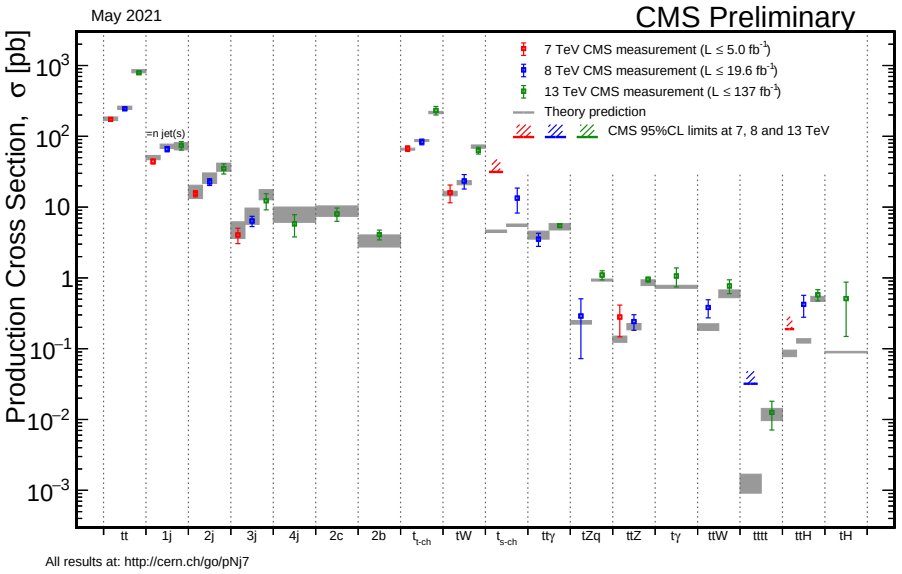



Figure 1.9.: Overview of measurements for various processes including top quarks at the LHC by the CMS collaboration. Taken from [56]. The plot includes measurements published up to May 2021.

large background, it is only a minor background in the context of this analysis. Single-top production is unlike to exhibit a signal-like final state because a large number of quarks or gluons from initial and final state radiation would be required.

**$t\bar{t}V$** The production of top-quark pairs in association with Z and W bosons ($t\bar{t}Z$ and $t\bar{t}W$ production; also collectively denoted as $t\bar{t}V$ production) have similar cross sections to the $t\bar{t}H$ processes. Both contribute very little to the overall yield since the $Z \to b\bar{b}$ branching fraction is smaller than of $H \to b\bar{b}$ and W bosons cannot decay to $b\bar{b}$ pairs. Therefore, it is also considered a minor background.

**V + jets and VV** The production of Z and W bosons with jets (Z + Jets and W + Jets or commonly referred to as V + jets) as well as the Diboson (ZZ, WZ and WW) production, have cross sections much larger than the signal but have a very small final contribution to the overall yield due to the necessary amount of initial or final-state radiation to resemble the signal final state. Both are considered minor backgrounds.

Figure 1.9 shows an overview of the measurements by the CMS collaboration of production cross sections of processes involving top-quarks, including the $t\bar{t}H$, $t\bar{t}$, $t\bar{t}$ + jets, $t\bar{t} + b\bar{b}$, Single-top, and $t\bar{t}V$ processes. The large differences in cross-sections, described above, can be clearly observed in the measurements.

### The fully-hadronic channel

The fully-hadronic decay mode of the $t\bar{t}H(b\bar{b})$ process has four bottom quarks and four lighter quarks in the final state. An example Feynman digram is shown in figure 1.10. The quarks are typically produced with large angles with respect to the beam axis, thus resulting in relatively large transverse momenta. Assigning the final state jets to the quarks is vital for the analysis since this information can be used to separate the signal from the background. One example is a highly efficient assignment which jets originated from the

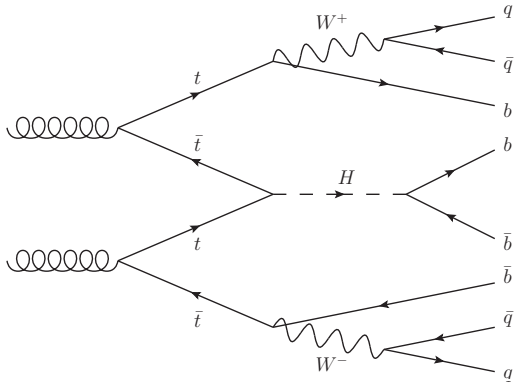



Figure 1.10.: Topology of the fully-hadronic $t\bar{t}H(b\bar{b})$ decay mode. Drawn using `JAXODRAW` [54]

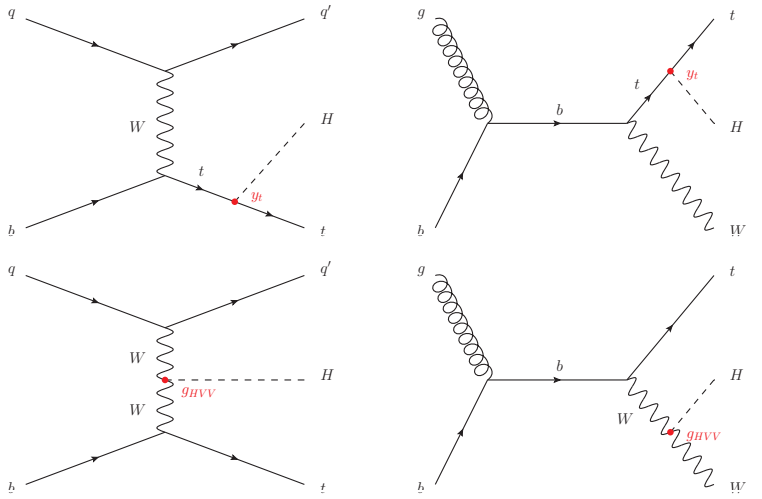
bottom quarks, making it for example possible to differentiate between $b\bar{b}$ pairs from gluon splitting or the Higgs boson. Due to the many similar final state objects, the fully-hadronic channel has to deal with a high ambiguity when making this assignment. Despite this drawback, this channel can be fully reconstructed since all final state objects can be measured in the detector. For the leptonic channels this is not the case due to the produced neutrinos. The fully-hadronic $t\bar{t}H(b\bar{b})$ process is overwhelmingly dominated by the QCD multijet background, since no lepton is present in the final state. This background, which cannot be modeled easily in simulation, requires an analysis approach that is very different to the leptonic decay modes, which solely rely on simulated backgrounds. In hadronic analyses, the multijet background is usually estimated from data in signal depleted control regions. In the case of the analysis presented in this thesis, the procedure for estimating this background will be discussed in section 5.5.
These characteristics make sophisticated methods for separating signal and background a necessity. The main approaches chosen for this analysis will be covered in chapter 3 in general and in section 5.6 specifically for the $t\bar{t}H(b\bar{b})$ fully-hadronic measurement.

### 1.3.4. Associated production of a Higgs boson in association with a single top quark

The production of the Higgs boson in association with a single top quark (tH) is very similar to the $t\bar{t}H$ production mode but exhibits unique properties that can be exploited to measure properties that are not accessible with $t\bar{t}H$ alone. Example Feynman diagrams are given in figure 1.11. Two tH production modes can be distinguished: The production in association with a light quark tHq and the production in association with a W boson tHW. Both are shown in figures 1.11a and 1.11b, respectively. The tHq production is the dominant process with roughly 5 times the cross section of tHW.
In general, the signature and backgrounds for the $tH(b\bar{b})$ process is very similar constrains as previously described for $t\bar{t}H(b\bar{b})$. Therefore, adding a measurement of $tH(b\bar{b})$ is relatively straight forward within the framework of $t\bar{t}H(b\bar{b})$ measurements.

#### 1.3.4.1. Coupling measurement

In the tH process, the Higgs boson can couple to either a top quark or a W boson. These properties can be exploited to measure the sign of the top-Higgs-Yukawa coupling $y_t$ relative to the coupling to the vector bosons $g_{HVV}$. If both exhibit the same sign, it results in constructive interference while an opposite sign results in destructive interference.
Typically, further non-SM modifications to the couplings are considered when carrying out a measurement of tH, in order to further probe the SM. One example is the measurement




(a) tHq process (b) tHW process

Figure 1.11.: Example tree-level Feynman diagrams for the production of a Higgs boson in association with a single top quark. (a) shows the tHq process and (b) the tHW process. In both cases, diagrams are shown where the Higgs boson couples to the top quark (top) and the W boson (bottom). Drawn using JAXODRAW [54]

of the Charge Parity (CP) properties of the Higgs boson. The CP transformation combines charge conjugation (interchanging particles and antiparticles by conjugating all internal quantum numbers) and parity conjugation (reversing the handedness). While most phenomena observed in the SM are invariant under CP transformation, some processes of e.g. the weak interaction violate this symmetry. In the SM, the CP symmetry is broken by complex phases in the Yukawa coupling, i.e. the coupling of the Higgs boson to quarks. At the current level of experimental and theoretical understanding, subdominant sources of CP violation from new physics are still possible.

For the top-Higgs-boson coupling measurement presented in this thesis, the SM Higgs boson coupling structure is supplemented with an additional coupling to a non-SM CP-odd Higgs boson described by the parameter $\bar{\kappa}_t$:

$$\mathcal{L} = -\frac{y_t}{\sqrt{2}}\bar{\psi}_t\left(\kappa_t + i\gamma_5\tilde{\kappa}_t\right)\psi_t X_0 \ . \tag{1.25}$$

In this framework, modifications to the SM couplings are parameterized in three coupling modifiers: $\kappa_t$ and $\kappa_V$ for the SM and $\tilde{\kappa}_t$ for a CP-mixing measurement. The couplings $\kappa$ are normalized to the expected values. In the case of $\kappa_t$ and $\kappa_V$ these are the SM expectations $y_t^{\mathrm{SM}}$ and $g_{HVV}^{\mathrm{SM}}$. All other coupling are kept at the SM values. The cross section in the SM can be parametrized as follows [57]:

$$\sigma_{\mathrm{tHq}} = (2.63\kappa_t^2 + 3.58\kappa_V^2 - 5.21\kappa_t\kappa_V)\cdot\sigma_{\mathrm{tHq}}^{\mathrm{SM}} \ , \tag{1.26}$$

$$\sigma_{\mathrm{tHW}} = (2.91\kappa_t^2 + 2.40\kappa_V^2 - 4.22\kappa_t\kappa_V)\cdot\sigma_{\mathrm{tHW}}^{\mathrm{SM}} \text{ and} \tag{1.27}$$

$$\sigma_{\mathrm{t\bar{t}H}} = \kappa_t^2\cdot\sigma_{\mathrm{t\bar{t}H}}^{\mathrm{SM}} \tag{1.28}$$

with $\kappa_t = y_t/y_t^{\mathrm{SM}}$ and $\kappa_V = g_{HVV}/g_{HVV}^{\mathrm{SM}}$. It can be seen, that $\mathrm{t\bar{t}H}$ is not sensitive to the sign of $\kappa_t$ since it only depends on $\kappa_t^2$, while this is not the case for tH.
Because of the interference of the different diagrams, the tH production cross section is sensitive to the CP state of the Higgs boson. For a CP-mixing measurement, an additional

term with quadratic dependency on $\tilde{\kappa}_t$ has to be added to equation (1.26) - (1.28). The Higgs boson has a CP quantum number of $J^{CP} = 0^+$ [29] and any violation of the CP invariance of the Higgs sector would directly indicate new physics.

### 1.3.5. Experimental Status

When the scalar (Higgs) boson was discovered in 2012 at the LHC, the exact properties were not fully known. Characterizing the properties of the Higgs boson is one of the main scientific goals of the ATLAS and CMS collaborations with the increasing luminosity (and therefore increasing expected sensitivity) delivered by the LHC machine. At the time of writing this thesis, it is mostly undisputed that the particle discovered in 2012 is the SM Higgs boson. Nevertheless, many of the very rare production and decay modes of the Higgs boson, in which new physics could hide, have not been observed yet. Moreover, many extensions of the SM that include a non-SM Higgs boson, are still not excluded. The studies of the Higgs-boson sector can be categorized in the precise understanding of the particle found in 2012 in the context of the SM and in the expectation that new physics can most likely be found in the Higgs-boson sector. The analysis presented in this thesis is mostly motivated by enhancing the knowledge of the SM Higgs boson.

The current experimental average (as given in reference [29] by combining various measurements by the ATLAS and CMS collaborations) of the Higgs-boson was determined to

$$m_{\mathrm{H}} = (125.25 \pm 0.17)\,\mathrm{GeV}$$

with ha total decay width of

$$\Gamma_{\mathrm{H}} = 0.0032^{+0.0028}_{-0.0022}\,\mathrm{GeV}\ ,$$

All measurements are compatible with the neutral, spin zero, scalar and CP-even Higgs predicted by the SM. More information on the current status can be found in reference [29]. Often, results are given relative to the SM expectation. Most commonly the signal strength

$$\mu_{\mathrm{process}} = \frac{\sigma_{\mathrm{process}}}{\sigma^{\mathrm{SM}}_{\mathrm{process}}}\ ,$$

is used to quantify the measurement. Either as absolute value, i.e. a $\mu_{\mathrm{process}} = 1$ would mean the measurement is exactly the SM expectation, or as 95 % CL exclusion limit. In the following a selection of recent results of the Higgs-boson sector will be highlighted.

#### $\mathrm{H} \to \gamma\gamma$ decays

The $\mathrm{H} \to \gamma\gamma$ decay channel is one of the most sensitive channels for measurements of the Higgs sector even though the branching fraction is very small with 0.23 % [29]. The reason for its sensitivity is that it exhibits a very clear experimental signature of two energetic photons over an exponentially falling background. The ATLAS [58] and CMS [59] collaboration published results combining various production modes using the data collected from 2016 to 2018 at the LHC. The ATLAS analysis measured the inclusive cross section times branching fraction for $\mathrm{H} \to \gamma\gamma$ (Higgs boson with $|\eta| < 2.5$) to be $(127 \pm 10)\,\mathrm{fb}$ at a SM expectation of $(116 \pm 5)\,\mathrm{fb}$. The CMS analysis determined the signal strength to $\mu = 1.12 \pm 0.09$ at a fixed value of $m_{\mathrm{H}} = 125.38\,\mathrm{GeV}$. Both publications include measurements of the CP structure of the Higgs boson, which point to SM-like Higgs boson. Figure 1.12 shows the signal strength per Higgs-boson production-mode in the $\mathrm{H} \to \gamma\gamma$ decay mode for the CMS measurement [59].
The $\mathrm{t\bar{t}H}$ production mode is included in these measurements but the sensitivity is mainly driven by the gluon fusion (ggH) production mode. Additionally, both collaborations

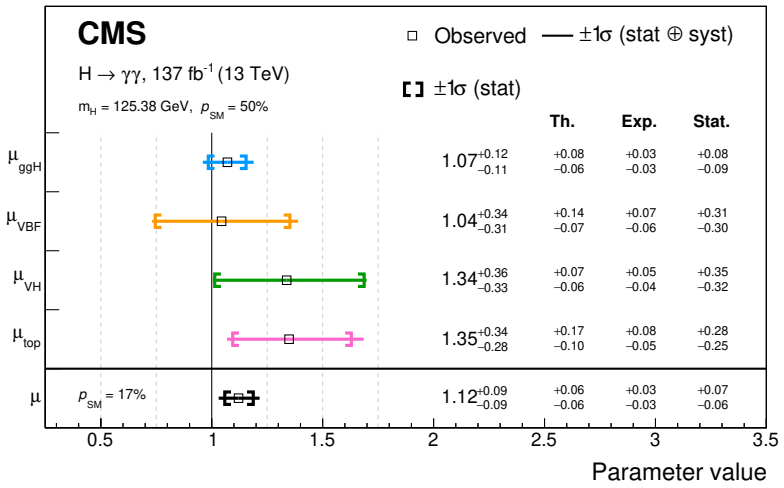



Figure 1.12.: Observed results of the fit to signal strength modifiers of the four principal production modes in the H → γγ decay mode using 137.2 fb$^{-1}$ of data collected by the CMS collaboration. Taken form reference [59].

published standalone $t\bar{t}H$ measurements that could claim the first observation of $t\bar{t}H$ in a single decay mode: The ATLAS analysis [60] determined the $t\bar{t}H$ signal strength to be $1.43^{+0.33}_{-0.31}(\text{stat})^{+0.21}_{-0.15}(\text{syst})$ times the SM expectation. The observed (expected) significance of this measurement is 5.2 (4.4)$\sigma$. The CMS analysis [61] determined the $t\bar{t}H$ signal strength to be $1.38^{+0.29}_{-0.27}(\text{stat})^{+0.21}_{-0.11}(\text{syst})$ times the SM expectation at an observed (expected) significance of 6.6 (4.7)$\sigma$.

### The H → ZZ → 4$l$ decay channel

The H → ZZ → 4$l$ decay channel is, together with the H → γγ channel, the most sensitive channel for Higgs-boson measurements at the LHC. Similarly, the branching fraction is small but the final state can be fully reconstructed and leptons can be resolved well enough to reconstruct a sharp Higgs-boson-mass peak over the background. The general strategy is to select pairs of oppositely charged lepton pairs ($\mu^+\mu^-\mu^+\mu^-$, $\mu^+\mu^-e^+e^-$, $e^+e^-e^+e^-$) in order to reconstruct the Z bosons and apply additional kinematic selection criteria. This allows for individual measurements of the ggF, VBF, VH, and $t\bar{t}H$ + tH modes and inclusive measurements.
The ATLAS [62, 63] and CMS [64] collaborations published results using the data collected from 2016 to 2018 at the LHC. The ATLAS analysis determined the inclusive fiducial cross section for H → ZZ to be $(3.28 \pm 0.32)$ fb to be in agreement with the SM expectation of $(3.41 \pm 0.18)$ pb. The CMS analysis determined the signal strength to be $0.94 \pm 0.07(\text{stat}) \pm 0.09(\text{syst})$ at a fixed value of $m_H = 125.38$ GeV. This corresponds to a inclusive fiducial cross section of $2.84^{+0.23}_{-0.22}(\text{stat})^{+0.26}_{-0.21}(\text{syst})$pb, which is in agreement with the SM expectation of $(2.84 \pm 0.15)$ pb.
The $t\bar{t}H$ production mode is included in these measurements but the sensitivity is mainly driven by the gluon fusion (ggH) production mode.

### H → b$\bar{\text{b}}$ decay channel

This channel has the largest branching fraction of all Higgs boson decay channels. However, the sensitivity suffers because the final state is difficult to distinguish from other hadronic activity and exhibits a worse energy resolution compared to the previously described channels. This means that measurements cannot rely on reconstructing a Higgs-boson mass

peak over the background and fitting the data but instead sophisticated statistical methods are required to extract the signal. Nevertheless, the $H \to b\bar{b}$ decay mode was observed by CMS [65] and ATLAS [66] in 2018 and 2019, respectively. These results do not include all of the data collected during LHC Run-2[3]. The main contribution for these results comes from the Higgs-boson production in association with vector bosons (VH), followed by $t\bar{t}H$ and ggH. Since these two measurements were published, newer results have been published in individual production modes.
The ATLAS collaboration published results for the VH production mode using $139\,\text{fb}^{-1}$ of data collected between 2015 and 2018 at 13 TeV [67]. The associated production with W and Z bosons could be established with an observed (expected) significance of 4.0 (4.1) and 5.3 (5.0) standard deviations, respectively. The corresponding signal strength are $0.95^{+0.27}_{-0.26}$ and $1.08^{+0.25}_{-0.24}$ are in accordance with the SM.
Furthermore, the ATLAS collaboration published results for the VBF production mode using $126\,\text{fb}^{-1}$ data collected between 2016 and 2018 at 13 TeV [68]. The signal strength of the VBF was determined to be $0.99^{+0.36}_{-0.34}$, corresponding to a observed (expected) significance of 2.9 (2.9) standard deviations.
A detailed overview of the $t\bar{t}H$ production mode in this decay channel will be given in section 4.1.1.
With increasing sensitivity, other measurements, like coupling or CP measurements, are also possible in this channel. The previously mentioned ATLAS publication include measurements of the CP-state of the Higgs boson. No deviations from the SM were found in both cases.

**Further measurements**

The experimental program targeting the SM Higgs sector at the LHC is vast, and fully exploring recent results is beyond the scope of this thesis. Nevertheless, a small fraction of further results should be highlighted as this point.

For $t\bar{t}H$, the Higgs to WW, ZZ and $\tau\tau$ decay modes (except for $H \to ZZ \to 4l$) are collectively published under the label of "multilepton" final states by the ATLAS and CMS collaboration. The CMS experiment published a results using $137.2\,\text{fb}^{-1}$ of data [69] with an observed (expected) significance of 4.7 (5.2) standard deviations, i.e. just short of an observation of this process, and 1.4 (0.3) standard deviations for the tH process. Furthermore, the measurement constrained the Top-Yukawa-coupling relative to the SM prediction to values of $-0.9 < \kappa_t < -0.7$ or $0.7 < \kappa_t < 1.1$, at a 95 % confidence level.

Due to its even smaller cross-section than $t\bar{t}H$, the sensitivity for the tH production is much worse. The CMS collaboration could set and upper limit on the fiducial cross-section for a SM-like tH of 1.9 pb, corresponding to 25 times the cross-section times branching fraction expected from the SM, in WW, ZZ, $\tau\tau$ and semileptonic $b\bar{b}$ decays of the Higgs boson using $36.3\,\text{fb}^{-1}$ of data [70]. As it is the case for the analysis presented in this thesis, tH measurements are included in more and more analyses and combined measurements. For example, ATLAS [58] and CMS [59] included tH measurements in their recent LHC Run-2 $H \to \gamma\gamma$ publication. In this channel alone, ATLAS could set an upper limit of 8 times the SM expectation on the tH production and CMS could set an upper limit of 12 times the SM expectation, using the full LHC Run-2 dataset.

[3] For its result, the CMS collaboration used all data collected at the LHC up to 2017. The measurement includes $\leq 5.1\,\text{fb}^{-1}$ at 7 TeV, $\leq 19.8\,\text{fb}^{-1}$ at 8 TeV, and $\leq 77.2\,\text{fb}^{-1}$ at 13 TeV. Similarly, the ATLAS collaboration combined its measurements in the same period. The measurement included $4.7\,\text{fb}^{-1}$ at 7 TeV, $20.3\,\text{fb}^{-1}$ at 8 TeV, and $24.5 - 79.8\,\text{fb}^{-1}$ at 13 TeV.
The integrated luminosity is given in ranges because not all considered measurements were conducted on the same datasets.

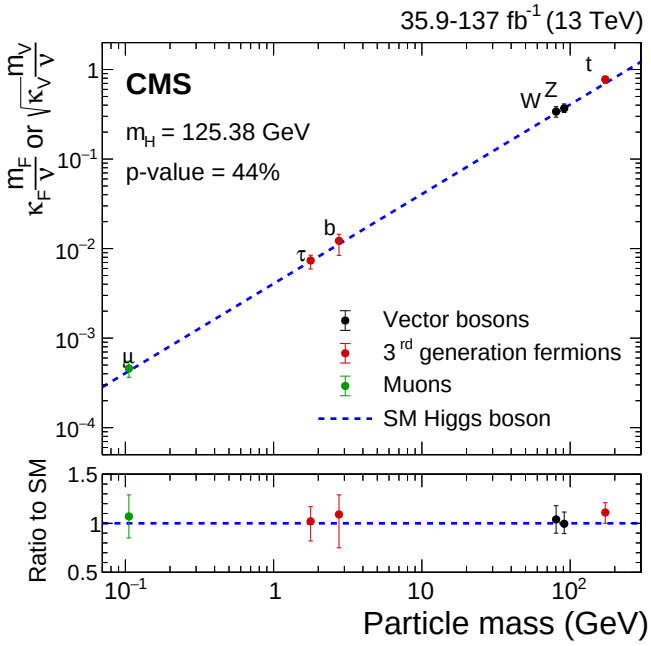



Figure 1.13.: Overview over the best fit estimates for the coupling modifiers extracted for fermions and weak bosons compared to their corresponding prediction from the SM by the CMS collaboration. The error bars represent 68% CL intervals for the measured parameters. In the lower panel, the ratios of the measured coupling modifiers values to their SM predictions are shown. Taken from reference [71]

Figure 1.13 shows an overview over the latest measurements for the coupling of the Higgs boson to fermions and weak bosons using 35.9 $\mathrm{fb}^{-1}$ to 137.2 $\mathrm{fb}^{-1}$ of data collected by the CMS Collaboration. It can be clearly seen that all currently available measurements agree with the theory prediction for a SM Higgs boson.

# 2. Experimental environment

This chapter will introduce the experimental environment used for the analysis presented in part II of this thesis. The chapter starts with an introduction of the Large Hadron Collider and the CMS detector in sections 2.1 and 2.2, respectively. Section 2.3 will introduce how the detector signals are used to reconstruct physics objects like electrons, muons, and jets. Section 2.4 will cover information on the luminosity measurement, data certification, and pileup definition.

## 2.1. The Large Hadron Collider

The Large Hadron collider (LHC) is a circular collider with a 26.7 km circumference. It is located near Geneva in the underground tunnel originally build for the LEP collider at the European Organization for Nuclear Research, commonly referred to as CERN[1]. If not otherwise indicated, the following information and details are taken from reference [73].
Four large physics experiments are located at different interaction points around the LHC ring. The two large multi-purpose detectors ATLAS [74] and CMS [75] are searching for a wide array of physics phenomena. Furthermore, two specialized detectors are in operation: The ALICE experiment [76], dedicated to heavy ion physics, and the LHCb experiment [77], with focus on B-Hadron physics, rare decays, and CP violation. The LHC is mainly operated as a proton-proton collider but configurations with heavy ions are possible. Since heavy-ion collisions are not relevant for this thesis, such collisions will not be discussed further.
In addition to the LHC, multiple smaller accelerators, that provide the LHC with protons, are in operation at CERN. Initially, the protons, which are produced by ionizing hydrogen gas, are injected as bunches into the linear accelerator LINAC2. Using the LINAC2, proton bunches are accelerated up to an energy of 50 MeV and injected into a chain of circular colliders: The Proton Synchrotron Booster (BOOSTER), the Proton Synchrotron (PS), and the Super Proton Synchrotron (SPS). At the end of this chain, the proton bunches reach the minimum energy required by the LHC. At this point, the proton bunches contain $1.1 \times 10^{11}$ protons each and exhibit an energy of 450 GeV and a temporal spacing of 25 ns. Inside the LHC, the beams are accelerated up to their final energy of 6.5 TeV using superconducting radio frequency cavities. The resulting center-of-mass energy of 13 TeV

[1] The name CERN is derived from *Conseil européen pour la recherche nucléaire*, which was the name of the temporary council established for building the laboratory.

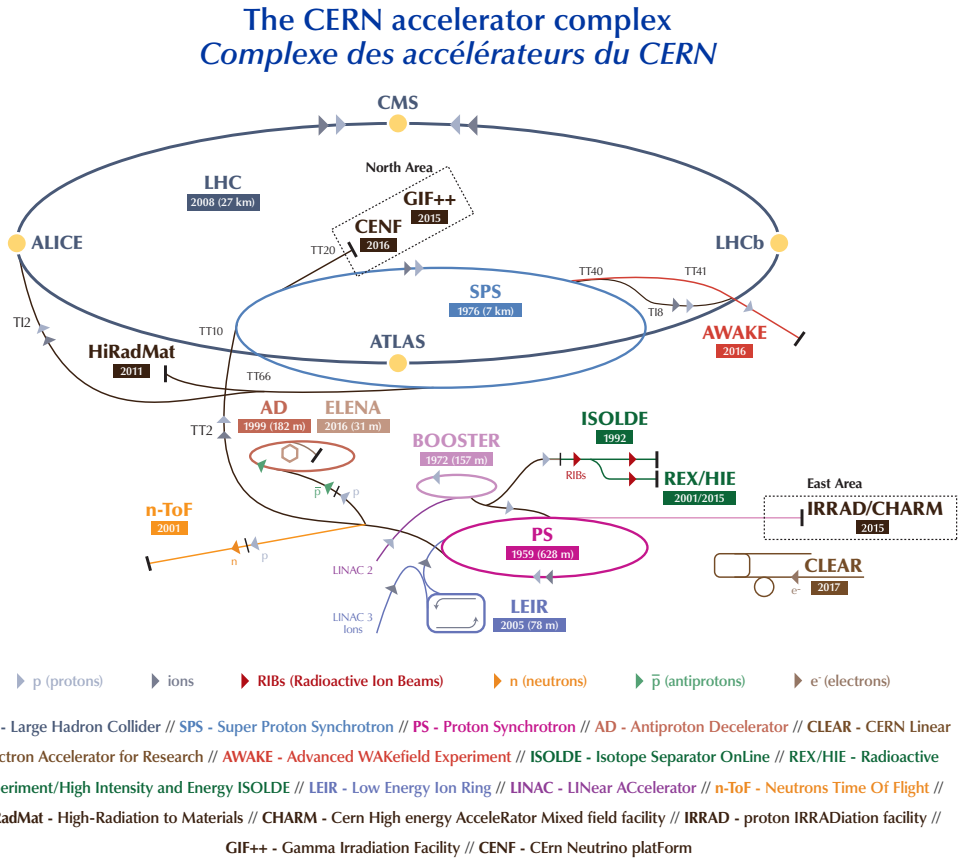



Figure 2.1.: Schematic rendering of the CERN accelerator complex [72]. In addition to the LHC, the four large experiments, the injector chain, and other related experiments, which are located at CERN, are shown.

is slightly below the initial design energy of 14 TeV[2]. Figure 2.1 shows an overview of the accelerator complex at CERN described previously.
The LHC is designed to deliver a high instantaneous luminosity and high proton-proton center-of-mass energies. The instantaneous luminosity is given by

$$L \sim \frac{f n_b N_b^2}{4\pi\sigma_x\sigma_y} \quad , \tag{2.1}$$

with the revolution frequency $f$, number of protons $n_b$ in each of the $N_b$ bunches, and the beam size at the interaction point $\sigma_x$ and $\sigma_y$ (characterized by a Gaussian profile). The design luminosity is $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ at 14 TeV. Since the rate of events directly depends on the instantaneous luminosity, it needs to be as large as possible if the goal is to find rare events. When searching for heavy particles, like the Higgs boson, high centre-of-mass energies in the collisions are necessary. This is a reason that the LHC is designed to collide protons instead of other particles like electron pairs, which loose too much energy due to synchrotron radiation, or proton-antiproton pairs, for which antiprotons cannot be produced in sufficient quantities. The two proton beams are accelerated in isolated beam pipes, which are required to be kept under an ultra-high vacuum, below $10^{-10}$ mbar, and circle to LHC ring in opposite directions. For this to be possible, opposite magnetic

[2] Collisions with a center-of-mass energy of 14 TeV are planned for the Run-3 operation starting in 2022 and beyond.

fields are required in each beam pipe. The magnetic field is generated by dipole magnets made from NbTi cables that are cooled to 2 K using superfluid helium. The magnets are designed to deliver an 8 T field, which is required for the operation with energies of up to 7 TeV per beam. Additional quadrupole magnets are used to stabilize and focus the beam. Once the beams reach the target energy, they are maintained and brought to collisions. These collisions can last up the 30 hours per fill. The instantaneous luminosity usually peaks at the beginning of the fill and then continuously decreases while the beam is maintained. This decrease does not only result from collisions but also from interactions with the detector material, gas in the beam pipe, and losses from particles escaping the magnetic field of the dipole magnets responsible for bending the beam. In addition to the events of interest, which occur with a rate of $L \cdot \sigma_{\text{proc}}$, for a process with a production cross section $\sigma_{\text{proc}}$, uninteresting, but much more common, p-p collisions occur with a cross section of $\sim 70\,\text{mb}$. Separating these uninteresting events, also referred to as pileup, from the interesting events is one of the main challenges presented by a hadron collider compared to other collider types.

The LHC started its operation in 2008 and could reach beam energies of 3.5 TeV and 4 TeV in 2011 and 2012, respectively. With such energies, first measurements probing the mechanism of electroweak symmetry breaking were possible and resulted in the discovery of the Higgs boson in 2012 [1,2]. In February 2013, this first run of the LHC machine (also called LHC Run-1) finished and the LHC was shut down to improve the machinery with the goal of increasing the beam energy to 6.5 GeV. The LHC Run-2 started in 2015 and lasted until 2018 at a target beam energy of 6.5 GeV. The design instantaneous luminosity of $10^{34}\,\text{cm}^{-2}\,\text{s}^{-1}$ could also be reached during the Run-2 operation. For practical reasons, the LHC Run-2 is further divided into per-year data-taking periods.

## 2.2. The CMS Detector

The Compact Muon Solenoid (CMS) detector is a multi-purpose experiment. It was mainly designed to study proton-proton collisions but can also be used to study other hadron collisions at the LHC. It is located near the village of Cessy in France opposite of the ATLAS detector, which is located at the main CERN site. The CMS detector is housed in a cavern 100 m underground and is built from multiple circular slices covering all directions around the LHC beam pipe. The detector slices can be moved to enable maintenance. The detector exhibits a lengths of 21.6 m, a diameter of 14.6 m, and a mass of 14 000 t. Figure 2.2 shows the detector and the placement of all subcomponents. The central design feature is the superconducting solenoid with a 6 m wide inner bore that provides an uniform magnetic field in its center. The silicon tracker, which is used to reconstruct the tracks of charged particles, and the electromagnetic and hadronic calorimeters are placed within the solenoid while the muon detector systems is placed outside within the iron yoke, which returns the magnetic flux of the solenoid. The detector exhibits two distinct elements with different design principles in all detector subcomponents: The barrel in the center of the detector, covering the central region around the interaction point, and the endcaps on both sides of the detector, covering the forward region. The orientation of the particles trajectories require different approaches in both cases.

As a multi-purpose detector, the CMS detector is designed to be able to identify a wide array of particle signatures that can be produced in hadronic collision. Therefore, it has to be able to identify and precisely measure the momenta and charge of electrons, muons and photons as well as the energy of strongly interacting particles, that can be used to identify B Hadrons or $\tau$-leptons and reconstruct jets and missing transverse energy. Furthermore, it has to be nearly hermetic. These design goals have to be reached in a very challenging environment: A fast detector response and trigger system is required by the high bunch-crossing rate. All components need to be radiation hard to survive the showers of particles

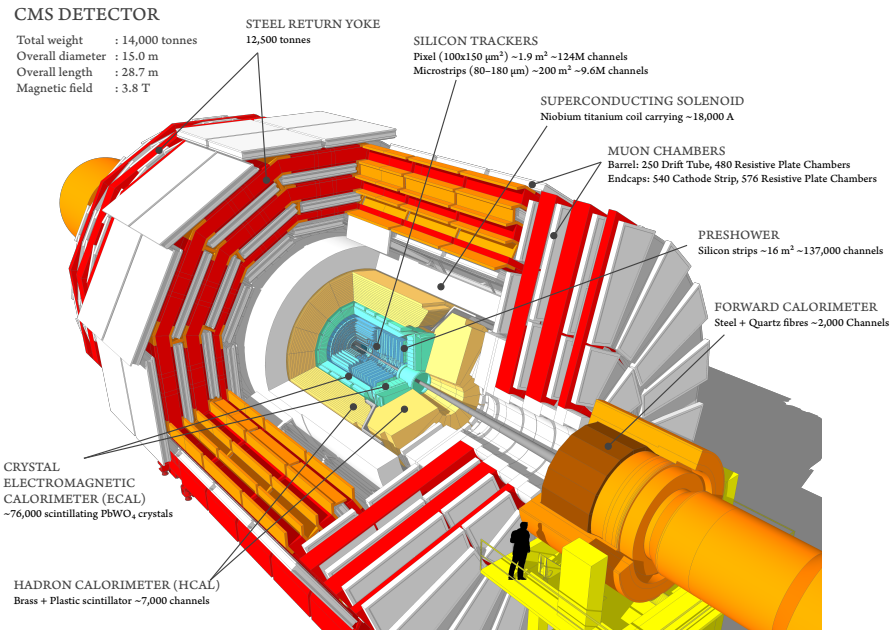



Figure 2.2.: Cutaway drawing of the CMS detector and its subsystems as configured during the 2017 and 2018 data-taking periods. [78].

produced in every collision. Multiple interactions, that occur in addition to the primary interaction, challenge the reconstruction and identification of particles.

For CMS measurements, the following coordinate system is used: The $z$-axis is aligned with beam axis and points towards the Jura mountains, the $x$-axis points towards the center of the LHC, and the $y$-axis points towards the surface. Due to the cylindrical design of the CMS detector, the angles $\phi$ ($x$-$y$-plane) and $\theta$ (measured from the $z$-axis) as well as the radial distance $r$ are more commonly used.

The subsystems of the CMS detector will be described in general in the following sections. A detailed description of the CMS detector and its subcomponents can be found in reference [75], which will also be used as main source of information for the following sections.

### 2.2.1. Solenoid

The solenoid is the central element of the CMS detector. A magnetic field is required to force charged particles on a helix shaped trajectory and measure their transverse momentum. Since the radius $r$ of the helix is given by

$$r = \frac{p}{q \cdot B} \ ,$$

with the field $B$ as well as particle momentum $p$ and charge $q$, a strong magnetic field is required to measure particles with high transverse momentum. The solenoid of the CMS detector has an inner bore of 6.3 m, a length of 12.5 m, and is built from NiTb cable. A magnetic field of 3.8 T is reached when the solenoid is in its superconducting state. For this, the solenoid is cooled with liquid helium to a temperature below 4.6 K. The iron yoke returns the magnetic flux and functions as main structural element of the CMS detector.

### 2.2.2. Tracking system

The CMS tracking system, also referred to as *the tracker*, is the subsystem closest to the beam pipe and thus the interaction point. The main purpose of the tacker is to measure

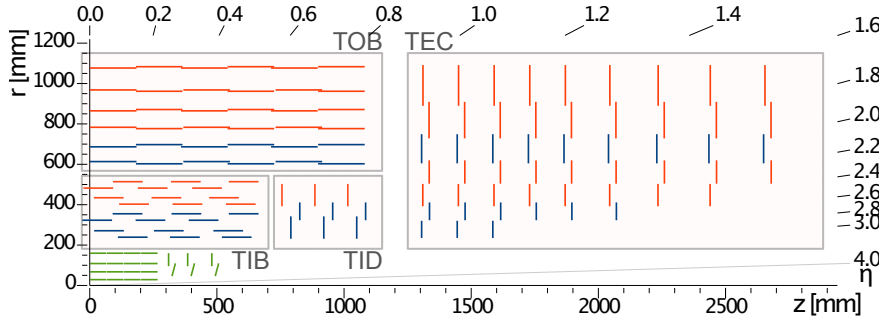
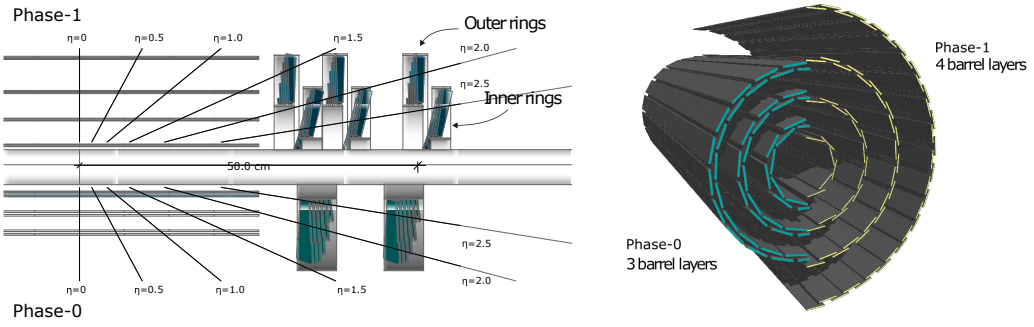



Figure 2.3.: Illustration of one quarter of the Phase-1 CMS tracking system in the $r$-$z$ plane. The pixel detector is shown in green, while single-sided and double-sided strip modules are shown as red and blue segments, respectively. For the strip detector, the tracker inner barrel (TIB), tracker inner disks (TID), tracker outer barrel (TOB), and tracker endcaps (TEC) subcomponents are shown. Adapted from reference [79].

the trajectories, also referred to as tracks, of charged particles with momenta $\gtrsim 1\,\text{GeV}$. Due to the magnetic field, the tracks are bent and can also be used to determine the charge of the particle. By reconstructing the tracks, the tracker also enables the identification of the primary vertices, i.e. the point of the p-p collision, and secondary vertices created by in-flight decays of particles, like B hadrons.
In reality, the tracker can only detect the points at which particles interacted with the active material. The reconstruction of the tracks and vertices from theses interaction points is accomplished with algorithms that will be covered in section 2.3.1. The CMS tracker has a size of 5.8 m in lengths and a diameter of 2.5 m. It contains two subsystems - the inner pixel detector and the outer strip detector - made from sensors that exploit the semiconducting properties of silicon: Charged particles traversing the depletion zone, which develops after applying a reverse bias voltage to the silicon doped with n-p-junctions, can create electron-hole pairs that can be measured. The full layout of the tracker, as it was configured during the 2017 and 2018 data-taking period, is illustrated in figure 2.3. The pixel and strip detectors both exhibit a cylindrical barrel part covering the central region of the detector and the circular endcaps covering the forward region on both sides of the detector. While the detector modules are placed parallel to the beam axis in the barrel part, they are placed roughly perpendicular to the beam-axis in the endcaps. The pixel and strip subsystems will be covered in more detail in the following:

The pixel detector is measuring the particles closest to the interactions and is vital for the reconstruction of primary and secondary vertices. The active area consists out of thousands of $100 \times 150\ \mu\text{m}^2$ sensors, also called pixels. During the shutdown between the 2016 and 2017 data-taking period, an upgraded pixel detector has been installed. The upgraded version is called the Phase-1 pixel detector. The version in use during the 2016 data-taking period is called the Phase-0 pixel detector. A direct comparison of the two layouts is given in figure 2.4. The key differences are the increased layer count in the barrel region (from 3 to 4) and forward region (from 2 to 6) as well as the placement of the innermost layers in the barrel and forward region closer to the interaction point. These changes have a big impact on tracking and vertexing performance of the CMS detector. Details on the upgrade can be found in reference [80]. In the following, the general design will be described for the Phase-1 detector with the corresponding values for the Phase-0 detector given in parenthesis:
The barrel layers are placed 3 cm, 6.8 cm, 10.9 cm, and 16 cm (4.4 cm, 7.3 cm, and 10.2 cm) from $z$-axis in radial direction and contain approximately 79 (48) million pixels. The




(a) (b)

Figure 2.4.: Layout of the layers and disks of the CMS pixel detector. The Phase-0 configuration, as used during the 2016 data-taking period, and the Phase-1 configuration, as used during the 2017 and 2018 data-taking periods, are shown. Adapted from reference [80].

layers in the pixel endcaps are placed at longitudinal positions of ±29.1 cm, ±39.6 cm, and ±51.6 cm (±34.5 cm, and ±46.5 cm) from the center of the detector and contain approximately 45 (18) million pixels. These pixelated detectors enable 3-D measurements along the trajectories of charged particles with $|\eta|$ up to 3 (2.5) and a spatial resolution of 10-20 µm, which is significantly smaller than the pixel size.

Outside the pixel detector, the strip detector extends up to 1.16 m in radial direction and ±2.82 m in longitudinal direction. The strip detector itself is further divided into subcomponents with different configurations: The tracker inner barrel (TIB), tracker inner disks (TID), tracker outer barrel (TOB), and tracker endcaps (TEC). These subcomponents are also labeled in figure 2.3. Silicon sensors with thicknesses of 320 µm and 500 µm are used to construct the layers and disks. The TIB and TOB consist out of ten and the TID and TEC out of twelve layers in total. Most layers use single-sided modules, which enables a 2-D measurement of the particle trajectories, but some layers (shown in blue in figure 2.3) are built using double-sided modules, in which the sensors on each side are placed at an stereo angle. The double sided modules, which are mostly placed at the inner most layer in each of the strip detector subcomponents, enable the measurement of an additional coordinate ($r$ in the disks or $z$ in the barrel). In total, approximately 9 million strips are distributed over 15000 modules, which results in an active area of $\sim 200\,\mathrm{m}^2$.

While working on the analysis presented in part II of this thesis, the author took over multiple pixel-detector-related monitoring responsibilities (occupancy monitoring, shifts during detector operation, and monitoring of the residual resolution) within the CMS collaboration.

### 2.2.3. Electromagnetic calorimeter

The electromagnetic calorimeter (ECAL), which is located adjoining the silicon tracking system, has the purpose of determining the energies of photons and charged particles, especially electrons, by measuring the energy deposition in the active calorimeter material made of lead tungstate crystals ($PbWO_4$). It is based on the following principle: When electrons or photons enter the transparent crystals they produce an electromagnetic shower of visible light, which is approximately proportional to the energy deposited in the crystal. The light is reflected by the polished surfaces of the crystal and recorded using photosensors, that are mounted to the far side of the crystal.

The CMS ECAL is made from more than 70000 $PbWO_4$ crystals and is divided into a central barrel ($|\eta| < 1.479$) and two endcap ($1.479 < |\eta| < 3$) components: The central barrel

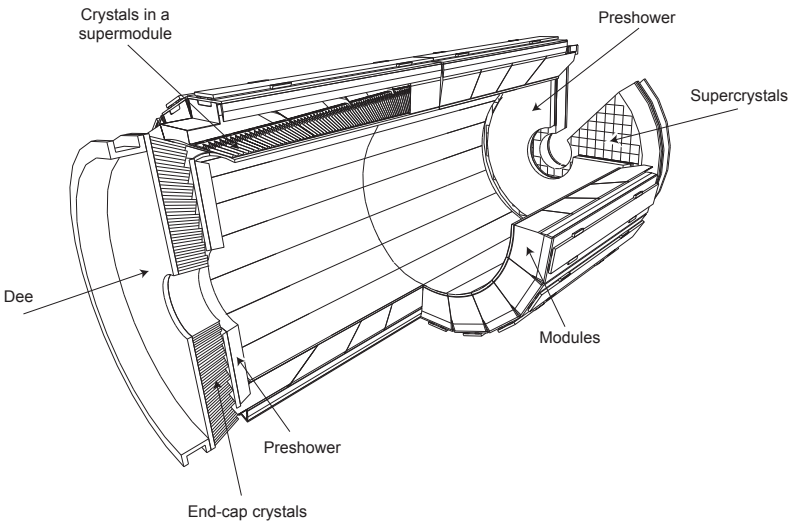



Figure 2.5.: Cutaway diagram of the CMS full electromagnetic calorimeter. Taken from reference [75].

system has an inner radius of 1.3 m and is composited of supermodules each containing 1700 crystals with a front size of 2 cm × 2 cm and a lengths of 23 cm. The endcaps are placed ±3 m from the interaction point in longitudinal direction. In the endcaps, crystals with a size of 2.86 cm × 2.86 cm × 22 cm are combined into 5 × 5 "supercrystals", that are mounted in semicircular structures. called *Dees*. Figure 2.5 shows the placement and layout of all these elements.
Because of this design, energy deposits as close as 5 cm can be resolved. The material and design choice results in a compact design and fine granularity, which were key design criteria in addition to a fast readout and high radiation hardness.

Additionally, the so-called Preshower detector is placed in front of the ECAL endcaps with the purpose of identifying photons from neutral pion decays that could otherwise be misidentified as photons. It is built as a two layer sampling calorimeter with a lead absorber and silicon strip detector.

### 2.2.4. Hadronic calorimeter

The energy measurement of strongly interacting particles, which are all stable particles except photons, neutrinos, electrons, and muons, is the task of the hadronic calorimeter (HCAL). Ultimately it enables a precise energy measurement of particle jets and missing transverse momentum. In order to reach a high precision for the latter, the HCAL has to be as hermetic as possible. Additionally, missing energy depositions in the HCAL with respect to the ECAL can be used to identify electrons and photons, since these particles are stopped in the ECAL before reaching the HCAL.

The construction follows the principles of a typical sampling calorimeter, with alternating layers of absorber- and active material: Incoming particles interact with the absorber material and produce a hadronic shower, which is measured in the active material. In the case of the CMS HCAL, the absorber layers are brass, which is non-magnetic, and the active layers are based on plastic scintillators, which exhibit a high radiation hardness. Similar to the ECAL, the HCAL consists out of multiple subsystems, which are shown in figure 2.6: The barrel part (HB) covers an area up to $|\eta| < 1.4$ and extends over the whole space from the ECAL up to solenoid in radial direction or from radii of 1.77 m to 2.95 m. The endcaps (HE) cover an area between $1.3 < |\eta| < 3$ in pseudorapidity. HB and HE consists out of towers with 17 and 19 absorbing layers, respectively. Below a

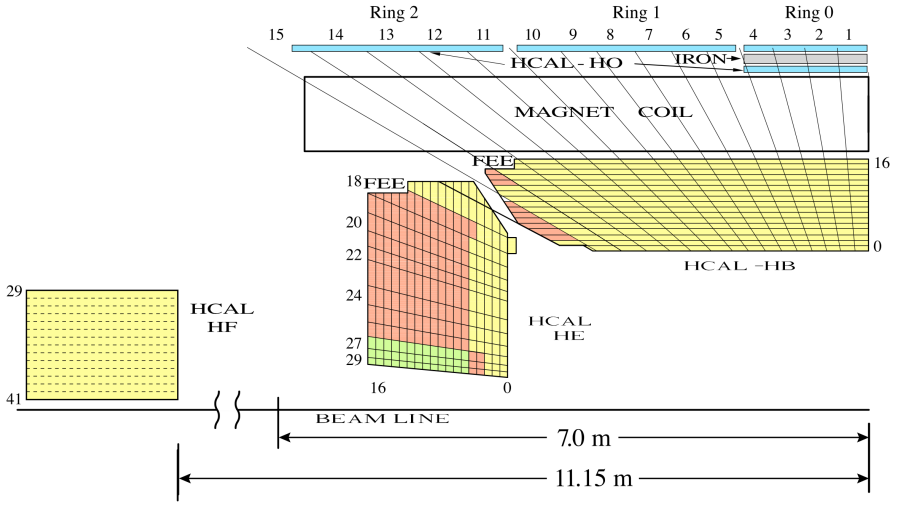



Figure 2.6.: Quarter slice of the CMS hadronic calorimeter (HCAL). The position of the barrel (HB), endcap (HE), outer (HO), and forward (HF) calorimeters is displayed. Taken from reference [81]

pseudorapidity of 1.6, the calorimeter is segmented in $0.087 \times 0.087$ cells in $\eta$ and $\phi$. For $|\eta| > 1.6$ the segments are twice as large in $\eta$ and $\phi$. In addition, an outer calorimeter (HO) is placed outside the solenoid in the barrel region. Its intent is to catch tails of hadronic showers that leak through the solenoid. The HO only consists out of a single layer of active material and extends up the $|\eta| < 1.26$. In the center - up to $|\eta| < 0.25$ - a second active layer and one absorbing layer, which is made from iron, are part of the HO. To measure the hadronic activity in the region of $3 < |\eta| < 5.2$, the forward calorimeters (HF) are placed at a distance of $\pm 11.2\,\mathrm{m}$ from the center of the detector. The HF has to be extremely radiation hard, since it very close to the beam and experiences a high flux of particles. This is achieved with a different detector technique based on Cherenkov light. Hadronic showers are created in steel absorbers and the resulting charged particles can be detected in inserted quartz fibers.

### 2.2.5. Muon detectors

The CMS muon system is another integral part of the CMS detector. Its main purpose is the identification of muons, which can transverse the detector volume without loosing much energy. Therefore, the muon system can be placed outside and solenoid, resulting in much reduced space constrains that would be presented to the design otherwise. A large instrument volume allows for a robust muon identification and precise measurement of the trajectories of muons with up to $\gtrsim 1\,\mathrm{TeV}$. Gaseous detectors are a good choice in such an environment.

Three distinct technologies of gaseous detectors are used in the muon system of the CMS detector: Drift tubes (DTs), cathode strip chambers (CSCs), and resistive plate chambers (RPCs). All three technologies operate on the same principle: A strong electromagnetic field is created in a cell filled with gas. Particles passing thought the cell ionize the gas along their path. The ions are accelerated in the field starting an avalanche of particles, which is detected as an electric signal. The layout and placement of the components of the muon system is shown in figure 2.7. In the barrel region, DTs are used due to the relatively low particle flux. In between the return yokes, four layers of DTs, also called stations, are placed at radii of 4 m, 4.9 m, 5.9 m, and 7 m from the interaction point and cover the region with $|\eta| < 1.2$. The tubes are filled with an Ar-$CO_2$ gas-mixture and contain sensitive wires, which acts as anode, and two cathode strips placed at the tube walls. In the endcaps,

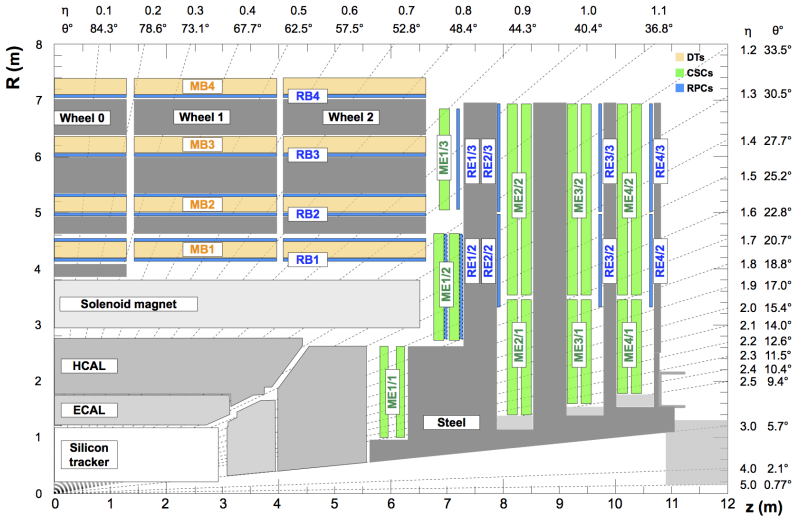



Figure 2.7.: Cross section of a quarter of the CMS detector. The origin of the coordinate systems is the center of the detector. The drift tube stations (DTs) are labeled as MB ("Muon Barrel") and the cathode strip chambers (CSCs) are labeled ME ("Muon Endcap"). The region labeled as RB and RE are the resistive plate chambers (RPCs) for the barrel and endcap regions, respectively. Taken from reference [82].

CSCs are used since they are superior in inhomogeneous magnetic fields and can cope with higher rates due to their faster response time. They are filled with a mixture of Ar-$CO_2$-$CF_4$ gas and cover the pseudorapidity region between $\pm 0.9$ and $\pm 2.4$. The drift times are shorter than in the DTs and the cathode is segmented perpendicular to the anode-wires, allowing for a three-dimensional measurement. Up to a pseudorapidity of roughly $\pm 1.6$, RPCs are placed in the barrel and endcaps. These detectors have a lower spacial resolution than DTs or CSCs but a faster response and good timing resolution. Consequently, they are important for identifying the primary bunch crossing. The RPCs are gaseous parallel plate detectors with strings between two anode-cathode Bakelite plates.

The spatial resolution of the CMS muon system is approximately 100 µm for DTs and CSCs and 1 cm for the RPCs.

### 2.2.6. Trigger and computing infrastructure

At the design luminosity of the LHC machine of $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ at 13 TeV, the bunch crossing rate is roughly 40 MHz, which corresponds to a proton-proton collision rate of 1 GHz. Most of these collisions are not relevant for the CMS physics program. The purpose of a trigger system is to select the collisions of interest and reduce the rate to the limit of the accompanying computing infrastructure. The information collected in a bunch crossing is also referred to as an *event*.

The CMS trigger system [83] has two tiers: The Level-1 (L1) trigger and the High-level trigger (HLT). Each event produces roughly 1 MB of data and the storage frequency to disk is technically limited to 1000 Hz. Therefore, the CMS trigger systems has to reduce the event rate from the bunch-crossing rate of 40 MHz to 1000 Hz. The L1 trigger and the HLT, have a rate budgets that cannot exceeded. A dedicated group within the CMS collaboration is tasked with monitoring the trigger rate and defining how the budget is split among the different analysis groups. While many analyses can trigger on straightforward signatures like an isolated lepton, complex combinations of physics objects and techniques, like jets, missing transverse energy, or b-jet identification, may be used for final states with

complex signatures. Especially for such final states, like the one expected from the fully-hadronic $t\bar{t}H(b\bar{b})$ process, the trigger can be a limiting factor if it is not considered from the beginning, e.g. by developing dedicated triggers.

The L1 trigger is implemented on custom electronics with a maximal output frequency of 100 kHz. This corresponds to a maximum processing time of 4 µs per event. It can use partial information from the muon detectors and the calorimeters to decide if an event is rejected: Ionization deposits in the DTs, CSCs, and PRCs are used to identify signatures consistent with muons and energy clusters in the ECAL, HCAL, and HF are used to identify signatures consistent with electrons, photons, jets, $\tau$-leptons or missing transverse energy. These object candidates are combined to so-called *trigger path* which are collected in a programmable *menu*. An event has to pass at least one of these paths to be considered in the next tier. The L1 trigger hardware is located near the experimental cavern in order to minimize latency.

All events passing the L1 trigger can be process by the HLT. The HLT uses the complete detector readout, i.e. it uses more sophisticated information, e.g. tracking, compared to the L1 trigger. The events are processed with similar but faster algorithms as they are used for the final reconstruction, which will be explained in section 2.3. The reconstruction in the HLT is also referred to as *online reconstruction*, while the final reconstruction used for analyses is called the *offline reconstruction*. The HLT is a software trigger that is run on a computing farm located in the building above the experimental cavern. The data is processes in a series of so-called *HLT paths*, which are sequential algorithms build from individual modules that represent various object reconstruction and selection (or filter) steps. The sequence starts with a L1 trigger *seed*, containing one or more L1 trigger paths, followed by reconstruction and selection steps of increasing complexity. The sequence is stopped as soon a step rejects an event or the event passes all steps. This approach ensures, that the computing time is kept at a minimum, since the most complex steps are only run for events with increasing similarity to the final state the trigger path is trying to identify. An example for increasing complexity would be that for jet-based HLT paths, a fast calorimeter-based jet-reconstruction is used in the beginning of the sequence to filter none-interesting events before running the slower full reconstruction, which includes tracker information. How such a sequence is built is explained in detail in appendix A.1 by the example of the trigger developed by the author of this thesis for the fully-hadronic $t\bar{t}H(b\bar{b})$ analysis.

Additionally, each L1 and HLT path includes a so-called *prescale*. If the prescale of a path is set to $n$, the path is only active for every $nth$ event. A path with $n = 1$ is also called *unprescaled*. Setting a prescale allows paths with low thresholds, which would have very high rates, to collect events that would be lost otherwise. Events collected with such paths are mainly used for testing, efficiency measurements[3], and calibration. All events passing at least one of the HLT paths, are kept for permanent storage.

To save and process this very large amount of data, the Worldwide LHC Computing Grid [84,85] (WLCG) was constructed. The data centers are organized in a tiered system. The Tier 0 data center is located a CERN and functions as a central hub, that all data passes through. The Tier 0 is responsible for safe-keeping of the unprocessed data, initial

[3] A practical example for such an efficiency measurement could be that the efficiency of a certain step in the trigger sequence should be measured. For this, an additional trigger path is defined, which is exactly the same as the path of interest but the step that should be measured is removed. Such an additional path is referred to as control path. Usually removing steps in the sequence increases the rate, which in turn would inflate the trigger budget. By running the control path with a prescale, the rate is reduced and the measurement can be carried out on events that would be lost otherwise. Since the path of interest is a subset of the control path, large prescales (of the order of $\gtrsim 100$) are unproblematic even if the rate of the path of interest is relatively small.

reconstruction, and distribution of the data to the other data centers. Thirteen data centers, located around the world, serve as the Tier 1 of the WLCG. These data centers are not only responsible for the storage and processing of the data but additionally serve as storage site for the processed data and simulations. Tier 2 and 3 data centers are usually operated by scientific institutes or universities and are used for storage, processing, and as computing clusters for researchers.

## 2.3. Object reconstruction

The process of translating the signals from the readout electronics of the CMS detector to objects that can be used for analysis of the collisions is called reconstruction. This section will describe how to get from the electrical signal to basic objects like tracks and finally to physics objects like jets.

### 2.3.1. Tracks and Vertices

Thousands of tracker hits are recorded for each collision and fitting tracks and vertices to these hits is an important step required for all types of analyses. The track reconstruction at CMS [86] employs sophisticated algorithms to achieve this. While the local position of each hit is given by the sensor, the physical location of the sensor can change due to external effects like the magnetic field of the solenoid. This is taken into account by transforming the local to a global hit position. A high precision of the transformation is ensured by monitoring the alignment [87] between the actual and expected geometry of the tracker.

For the track reconstruction at CMS, the `Combinatorial Track Finder` (CTF) algorithm, which is based on a Kalman filter [88], is used. CTF unites track reconstruction with pattern recognition and track fitting in a single framework as an iterative process. The procedure can be summarized as follows:

- In a first step, the CTF algorithm generates so-called seeds from a very limited number of hits. The seeds are used as first estimate of the trajectory.
- By extrapolating along this trajectory, additional hits are gathered. This is also called trajectory building.
- A fit is performed on the gathered hits to determine the origin, transverse momentum and direction of the trajectory.
- In a final step, quality criteria are calculated for this track. If the track passes the criteria, it will be saved and the hits are removed from the collection of hits in order to reduce complexity. Otherwise this track-hit association will be discarded and the cycle start again with the remaining hits.

The quality criteria include, among others, cuts on the minimum number of hits, layers hit, and layers skipped, and a $\chi^2$-criterion on the track fit. Reconstructing tracks of charged particles with $\gtrsim 1\,\text{GeV}$ using this algorithm has an efficiency of the order of 90 % for pions and $\sim 99\,\%$ for muons at fake rates of only a few percent. This difference can be mostly attributed to their respective nuclear interaction with the tracker material.

The points of the inelastic proton-proton scattering are called primary vertices (PV). They are located very close to the beam axis, usually within a few millimeters. This is also the main distinction to secondary vertices, which will be discussed in more detail in the context of the identification of jets originating from B hadrons in section 2.3.6. Primary vertices are reconstructed in a three step process. In a first step, tracks with low transverse distance to the beam-axis, low number of hits, and that pass a $\chi^2$ criterion are selected. In the next step, these tracks are clustered according to their z-coordinate using a deterministic

annealing [89] algorithm, which considers all possible track-vertex assignments weighted by their probability. In a final step, an adaptive vertex fitter [90] is used to determine the vertex parameters, like the $x$, $y$ and $z$ coordinates or the number of degrees of freedom of the fit. A weight is assigned to each considered track corresponding to the probability that it belongs to a given vertex. The degrees of freedom are defined as

$$n_{\text{dof}} = -3 \cdot 2 \sum_{i}^{\#\text{Tracks}} w_i \ , \tag{2.2}$$

with the track weight $w_i$. The position resolution of the primary vertices determination is of the order of 10 µm. A detailed explanation of this procedure and the resolution measurements can be found in reference [86] section 6.1.

### 2.3.2. Electrons

The electron reconstruction [91] uses energy deposits in the ECAL and tracks from the tracker. The energy of an electron is usually spread out in $\phi$-direction over multiple crystals in the ECAL. One factor for this spread is the loss of energy due to bremsstrahlung before reaching the ECAL. Depending on the direction of the electron, this energy loss is found to be between 33 % and 86 % at pseudorapidities of 0 and $\sim 1.4$, respectively.
For an accurate energy measurement, the deposits in the ECAL need to be clustered. This process starts at a seed crystal with a large energy deposit. Iteratively, $5 \times 1$ crystals in $\eta \times \phi$ are added around the seed if they pass a minimum energy threshold. This collection of clusters is combined to a so-called Supercluster (SC). The SC energy is the sum of the individual cluster energies and the position is determined by the energy-weighted mean. For the particle-flow (PF) reconstruction, which will be described in section 2.3.4, a modified clustering scheme, which aims to reconstruct individual particle showers, is used. These PF-clusters are constructed around a seed by aggregating all continuous crystal energy above two standard deviation of the electronic noise. Following this approach, can result in energy deposits in a single crystal to be shared between multiple clusters.
Instead of also using the previously described Kalman-filter-based track reconstruction for electrons, a dedicated track reconstruction was developed due to the large radiative losses of electrons. This computationally much more expensive algorithm, which is only applied to select electron seeds, takes the changes to the curvature due to bremsstrahlung into account. The considered seeds are identified by estimating electron trajectories based on the SC energy and position (ECAL-based seeding) or by extrapolating standard tracks to the ECAL and matching the SCs (tracker-based seeding). The electron seeds from both approaches are combined. This results in an overall efficiency $\geq$ 99 % (estimated from simulation). Starting from these seeds, a combinatorial Kalman filter method is used to reconstruct the tracks. Compared to the standard track reconstruction, this approach is more restrictive with respect to missing hits in order to reject converted photons.
A Gaussian sum filter (GSF), which models energy loss in the tracker using Gaussian distributions, is used to fit these reconstructed tracks. The fit generates a track from the beam spot to the ECAL and enables the estimation of the energy loss based on the curvature difference between the points. In the PF reconstruction, the GFS tracks are linked with PF clusters, which allows for recovering most bremsstrahlung photons. Finally, the electron candidates are reconstructed from GFS tracks and ECAL clusters based on the position of the ECAL cluster.
Bremsstrahlung photons decaying to electron-positron pairs can lead to ambiguities when two GSF tracks share one SC. A set of criteria, using parameters like the number of hits and the type of seed used for the fit, was developed to resolve such issues.

### 2.3.3. Muons

The muon reconstruction [82] starts from the tracker tracks, which are reconstructed as described in section 2.3.1, and, so-called, standalone-muon tracks, which only use information from the muon detector. From these two tacks, more sophisticated tracks, with different properties for the muon reconstruction, are built.

**Standalone-muon tracks** are reconstructed using a Kalman-filter-based technique with information provided by the DTs, CSCs and RPCs.

**Tracker-muon tracks** are built from the tracker tracks by extrapolating the tracks to the muon system with loose matching requirements on DT and CSC segments. Hence these tracks are built from the "inside-out".

**Global-muon tracks** are built by matching standalone-muon tracks to tracker track. This is done by comparing the track parameters and performing a combined fit. Hence these tracks are built from the "outside-in".

Overall, 99 % of muons produced within the CMS acceptance are reconstructed at least as one of tracker-muon and global-muon tracks. Tracker-muon and global-muon tracks are merged into a single candidate if they share a tracker track. For low momenta, the tracker-muon reconstruction is more efficient than the global reconstruction. But the misidentification rate is lower for global muons since they included information from the muon detectors. These muon tracks are fed to the PF algorithm.

### 2.3.4. Particle-flow event reconstruction

The particle-flow event reconstruction [92] is used to reconstruct stable particles by combining information from all subsystems of the CMS detector. The PF algorithm identifies five distinct types of particles: Charged and neutral hadrons, electrons, photons and muons. These particles, also referred to as PF candidates, can then be used to derived more complex objects like jets and the missing transverse momentum of an event. While jets can be reconstructed by only using calorimeter information, also referred to as calorimeter jets, the reconstruction from the PF candidates greatly improves the precision. The main reason for this is that many jet constituents are comparatively soft and benefit from the precise measurements of low-momentum particles in the tracker.

The PF algorithms starts from the tracks, reconstructed as described in sections 2.3.1 and 2.3.3, and calorimeter clusters. The PF ECAL clusters are described in section 2.3.2. The HCAL clustering follows a similar approach: Local maxima of energy depositions in the calorimeter are used to identify seeds. Energy depositions are combined to topological clusters by combining neighboring cells that are two standard deviations over detector noise.

One of the main features of the PF algorithm is the so-called linking procedure. Hereby the calorimeter clusters and tracks are linked to blocks of elements. From each block the PF algorithm identifies particles following predefined criteria. If a particle is identified, the blocks, e.g. energy deposits in the ECAL, are removed in order to reduce the computational complexity. The following PF particles are reconstructed in the given order:

**PF muons** are global muons, whose momentum is compatible with the momentum measurement in the tracker.

**PF electrons** are reconstructed using the procedure described in section 2.3.2.

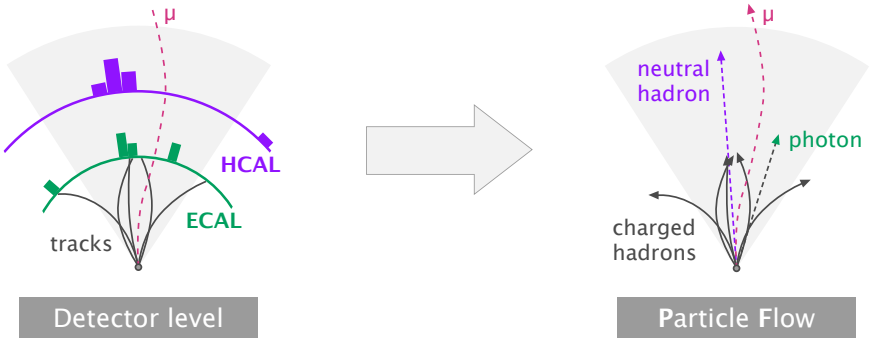



Figure 2.8.: Illustration of the particle-flow event reconstruction. A simplified sketch of the detector response is shown on the left. Reconstructed tracks in the tracker are shown as solid black lines. The dashed red line symbolizes a muon track, measured in the tracker and muon detector. Energy deposits in the calorimeters are shown in green (ECAL) and purple (HCAL). On the right side, the interpretation of these measurements by the particle-flow algorithm is shown. [93]

**PF charged hadrons** are reconstructed next. Tracks are linked to ECAL and HCAL clusters, resulting in the charged hadron candidates. If the energy determined from the track momentum is compatible with the calorimeter energy deposits, the charged hadron energy is determined from the linked track and calorimeter clusters. If the track momentum is significantly higher, an additional search for muons with looser requirements is triggered. If the track momentum is significantly lower, no charged hadron is reconstructed.

**PF photon** is required to exhibit an energy deposition in the ECAL that is much larger than the energy of the assigned tracks.

**PF neutral hadrons** is required to exhibit an energy deposition in the HCAL that is much larger than the energy of the assigned tracks.

Figure 2.8 illustrates the PF algorithm on a simplified example.

### 2.3.5. Jets and missing transverse momentum

In the CMS event reconstruction procedure, jets, as introduced in section 1.2.6, are clustered from the particle flow candidates.
The most common type of algorithms used at the LHC experiments are sequential recombination algorithms. These algorithms rely on energy and distance measures for sequentially combining particles until some exit condition is reached. Usually the $p_{\mathrm{T}}$ and $\Delta R$ are used as a proxy of the energy and distance measures, respectively. The jets resulting from these algorithms are invariant under boosts along the beam line. The procedure for the clustering progresses as follows. Initially the particle $i$ to particle $j$ distance $d_{ij}$ and beam-particle distance $d_{iB}$ is calculated for each particle pair and the minimum is computed. If it is $d_{ij}$, the two particles are merged and the four-momenta are combined. If it is $d_{iB}$, the $\geq 1$ combined particles are called a jet and the individual particles are removed from the collection of input particles. This is repeated until all particles of the input collection have been used. The distance for the typical clustering algorithms at the LHC are defined

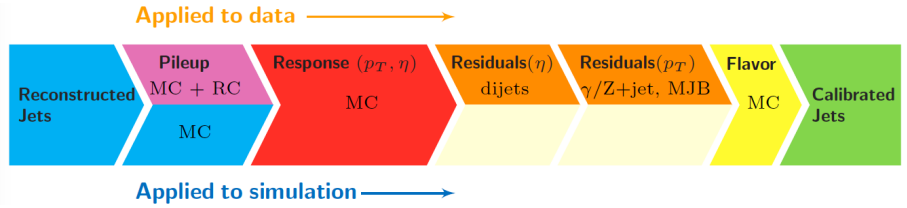



Figure 2.9.: Consecutive stages of the jet energy scale (JES) correction as it is applied to reconstructed jets in data and simulation (MC). Taken from reference [96].

as

$$d_{ij} = \min\left(p_{\mathrm{T},i}^{2p}, p_{\mathrm{T},j}^{2p}\right) \frac{(y_i - y_j)^2 + (\phi_i - \phi_j)^2}{R^2} \text{and} \tag{2.3}$$

$$d_{iB} = p_{\mathrm{T},i}^{2p}, \tag{2.4}$$

with the radius parameter $R$, the transverse momentum $p_{\mathrm{T},i}$, rapidity $y_i$ and $\phi$-direction $\phi_i$ of particle $i$. If $p = -1$ is chosen, a clustering algorithm following this procedure is called the anti-$k_{\mathrm{T}}$ algorithm [94]. The distance is defined such that high-$p_{\mathrm{T}}$ particles are clustered first. This leads to infrared and collinear safe jets[4] with a round shape in the $\eta - \phi$-plane with a radius of $R$.
In order to reduce the influence on pile-up events, candidates that do not originate from the primary vertex are usually excluded. This is called charged hadron subtraction (CHS). The jets, which are used in this thesis, are clustered using the anti-$k_{\mathrm{T}}$ algorithm with a distance parameter of 0.4 as implemented in `FastJet` [95][5].

The reconstructed jets still have to be corrected and calibrated in simulation as well as recorded data: Even after CHS, particles from pileup are left in the input collection used for the clustering because they are either neutral or could not be associated to a pileup vertex. Additionally, the transverse-momentum measurement of jets is dependent on the pseudorapidity $\eta$ and non-linear in $p_{\mathrm{T}}$. The correction and calibration used in this thesis follows the factorized approach described in reference [96] and is called the jet energy correction (JEC), which in turn includes two types of corrections: The jet energy scale (JES) and jet energy resolution (JER) correction.
The JES is a multi-step correction that aims to bring the measured jet energy closer to the true jet energy. The steps are illustrated in figure 2.9 and contain corrections for

- the energy offset due to particles from pileup,
- the detector response of hadrons,
- the residual difference between data and simulation as function of $p_{\mathrm{T}}$ and $\eta$, and
- flavor (for quarks: u and d, s, c and b; and for gluons) dependent effects.

The energy resolution observed in data is worse than the resolution predicted in simulation. Therefore, jets in simulation are smeared to better reflect the observations in data. Two smearing techniques are used: If a generator-level jet can be matched to the jet reconstructed from the PF candidates, a correction is calculated as function of the $p_{\mathrm{T}}$ and $\eta$ of the jet and $p_{\mathrm{T}}$ at generator-level. Otherwise a stochastic smearing, which aims to reproduce the average difference between data and simulation, is applied based on the jet $p_{\mathrm{T}}$ resolution and $\eta$.

[4] A jet algorithm is considered infrared and collinear safe, if radiation of soft particles or collinear splitting of partons does not change the outcome of the clustering.
[5] Analyses with specific requirements for jets, e.g. if highly boosted final states are targeted, may use other algorithms.

Uncertainties for the JES and JER corrections are considered and will be described in section 6.2.

The missing transverse momentum of an event - also referred to as missing transverse energy (MET) - is usually defined as the negative sum of the $p_T$ of all PF candidates. In a perfect measurement, this sum correspond to the transverse momentum of the initial state, which is zero. Particles, e.g. neutrinos, can escape the detector unobserved. In such cases, the MET can be identified with the momentum of these escaped particles. Since the calculation of the MET includes all PF candidates, the resolution is comparatively low. For the fully-hadronic analysis presented in part II, the MET is not used. Hence no further discussion of the MET and the corresponding corrections is presented here. More information can be found in reference [97].

### 2.3.6. Identification of jets originating from bottom quarks

The final state of the fully-hadronic $t\bar{t}H(b\bar{b})$ process features four jets that originate from bottom quarks. Therefore, the identification of such jets plays a very important role for the overall success of the measurement presented in this thesis.

Jets originating from bottom quarks are identified with algorithms collected under the term b-tagging. These algorithms exploit the characteristic properties of B hadrons (hadrons containing a valence bottom quark) which are:

- B hadrons are very heavy with a rest mass of more than 5 GeV.
- B hadrons mainly decay to hadrons containing c-quarks via the weak interaction. Because this decay is CKM suppressed, the B hadrons have a comparatively long lifetime of 0.5 mm/c.
- The decay products of B hadrons usually include multiple charged leptons, which leave tracks in the tracker.
- B hadrons carry a large fraction of the overall jet momentum because most energy of the bottom quark is passed to the B hadron in the fragmentation. Therefore, they can have large Lorentz boosts, which further increase their lifetime.
- The displaced decay of the B hadron results in a secondary vertex with a significant distance, usually referred to as the impact parameter $d_0$, to the primary interaction point. This secondary vertex can be individually reconstructed because of, among others, the high precision of the CMS pixel detector. A sketch of such a decay is shown in figure 2.10.

A detailed introduction to the algorithms and procedures used within the CMS collaboration can be found in reference [98].

In the context of the measurement present in part II of this thesis, three different algorithms are used: The combined secondary vertex tagger version 2 (`CSVv2` [98]), a further developed version of this tagger using a fully-connected neural-network (`DeepCSV` [98]), and the `DeepJet` tagger [99], which utilizes a combination fully-connected networks, convolutional networks, and recurrent networks. The performance of the algorithms improves in the given order. The first two algorithms are only used in high-level trigger as they presented the state-of-the-art algorithms at the time the triggers were developed and the data was collected. As it will be shown further down, the `DeepJet` tagger exhibits a superior performance, and is therefore used as the b-tagging algorithm in the final measurement.
The general architecture of the `DeepJet` tagger is illustrated in figure 2.11. For a given jet, the following input features are passed to the network. Information for each of the leading 25 charged and neutral PF candidates is passed to the network in 16 and 8 features, respectively. These features include information like the $\Delta R$ to the closest secondary vertex, the track fit quality, and the fraction of jet-momentum carried by the candidate.

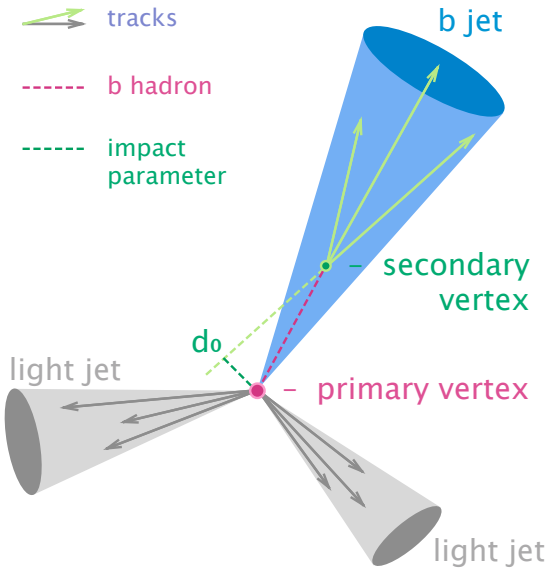



Figure 2.10.: Illustration of secondary vertex displacement from the primary vertex. The impact parameter $d_0$ is used to measure the displacement of tracks not originating from the primary vertex. [93]

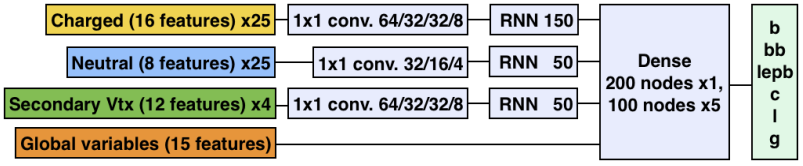
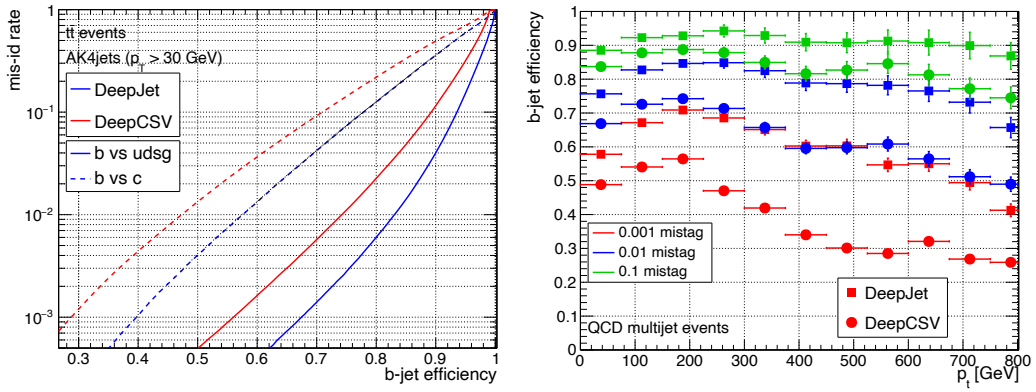



Figure 2.11.: Illustration of the `DeepJet` architecture. The input of the network is comprised of charged and neutral PF candidates, reconstructed secondary vertices and global variables. A combination of convolutional, recurrent, and feed-forward networks is used. Taken from [99].

Up to four secondary vertices (SV) are passed as 12 features each. These features include information about the SV like impact parameter value or fit quality. Furthermore, global variables of the jet, like $p_{\mathrm{T}}$ or $\eta$, and the number of primary vertices in the event are passed to the network. The network architecture is designed as a sequence of convolutional, recurrent and feed-forward networks. A description of the former two types of networks is out of the scope of this thesis, while the latter type of network will be discussed in detail in section 3.2.1. On a top-level, the different networks can be understood as follows: The features of the SV and charged as well as neutral PF candidates are processed per candidate in the convolutional networks, which perform a feature extraction, i.e. exploit correlation among the features to build new features. In the individual groups, all these "new" features a fed together to recurrent networks, which are specialized in learning sequence information. The output of the recurrent networks can be understood as combined information of all input particles and features in the individual groups. These outputs are combined with the global variables in a fully-connected feed-forward networks with multiple output nodes corresponding to the likelihood that a jet corresponding to a considered class[6]. In addition to b-tagging, the `DeepJet` tagger also provides c-tagging and gluon-tagging capabilities.

In the CMS collaboration, operating points - also called working points or WP - are defined

[6] Details about classification with neural networks with multiple output classes will be given in section 3.2.2.




(a) (b)

Figure 2.12.: Comparison of the performance of the DeepJet and DeepCSV b-tagging algorithms. Figure (a) shows the performance in simulated $t\bar{t}$ events where both top quarks decay hadronically for jets with $p_T \geq 30\,\text{GeV}$. The performance is shown for both b vs. c classification (dashed lines), and b vs. light (solid lines). Figure (b) shows the jet-$p_T$ dependency of the two taggers in simulated QCD multijet events for three fixed light-jet misidentification probabilities (mis-tag rate). Taken from [99].

corresponding to the 10 %, 1 % and 0.1 % light-flavor mis-tag rate for jets originating from bottom quarks. The three WP are usually referred to as Loose, Medium and Tight. The corresponding efficiency for tagging jets originating from bottom quarks as b-jets depends on the algorithm. In the case of the DeepJet algorithm, the Loose WP corresponds to a b-jet efficiency of approximately 92 % and the Medium WP to approximately 75 %. Other algorithms (DeepCSV and CSVv2) exhibit smaller b-jet efficiencies at the same WPs. This is equivalent to stating that those algorithms exhibit a worse performance than the DeepJet algorithm.

A direct performance comparison of the DeepJet and DeepCSV taggers is shown in figure 2.12: Figure 2.12a shows the overall mis-tag rate as a function of the b-jet efficiency in simulated $t\bar{t}$ events. The DeepJet algorithm provides a lower mis-tag rate at every b-jet efficiency than the DeepCSV algorithm. Figure 2.12b shows the b-jet efficiency as function of the jet $p_T$ in simulated QCD multijet events for the three WP defined previously. It can be seen that the DeepJet algorithm outperforms the DeepCSV algorithm at higher jet $p_T$. Especially for tighter WPs (red and blue in figure 2.12b), the difference is significant[7].

## 2.4. Luminosity, data certification, and pileup

### 2.4.1. Luminosity

A precise measurement of the luminosity is crucial for the measurements of cross sections at the CMS experiment. During data-taking the measurement of the luminosity provides real-time feedback on the LHC and CMS operation. After data-taking, the luminosity measurement is indispensable for nearly every physics analysis. For example, it is required for scaling the simulated samples to the correct yield. A variety of systems are used to provide the luminosity measurement. Details are given, for example, in reference [100]. In practice each of these systems reads out a rate $R$, which is proportional to the instantaneous luminosity $\mathcal{L}_{\text{inst}}$ with a constant called the visible cross section $\sigma_{\text{vis}}$. The visible cross

[7] In case for the $t\bar{t}H(b\bar{b})$ analysis presented in part II of this thesis, improvements of up to 130 % in signal efficiency and up to 40 % in $S/\sqrt{B}$ were observed.

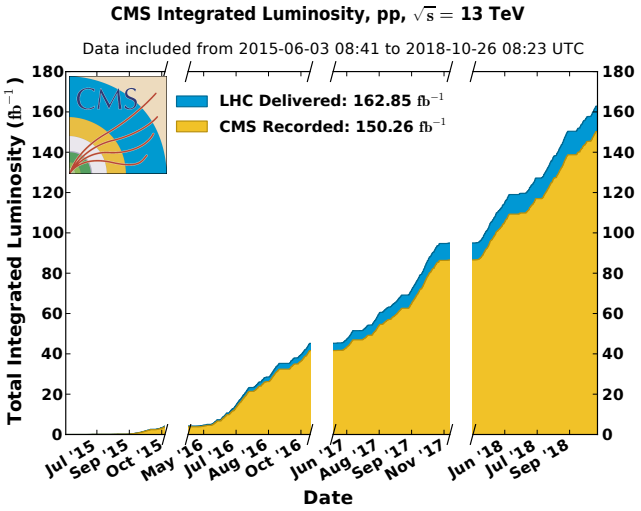



Figure 2.13.: Luminosity delivered by the LHC and recorded by the CMS detector during the Run-2 operation period from 2015 to 2018. Taken from reference [101]

section $\sigma_{\text{vis}}$ is determined using Van der Meer scans performed with a dedicated LHC machine setup. During this scan, the transverse position of the two beams is varied relative to each other resulting in a measurement of the rate as function of the beam separation. This rate can be used to determine $\sigma_{\text{vis}}$ [100].

The integrated luminosity is given by

$$\mathcal{L} = \int L\,dt \ . \tag{2.5}$$

Figure 2.13 shows the recorded and delivered luminosity during the Run-2 operation of the CMS detector.

### 2.4.2. Data certification

The LHC delivered $162.85\,\text{fb}^{-1}$ of data during the Run-2 operation period from 2015 to 2018. Not all of this data can be used for physics analyses. Firstly, the detectors cannot record every collisions for a multitude of reasons. An example would be technical malfunctions. Secondly, only high-quality data, e.g. with all critical subdetectors operational, can be used for precision measurements. In order to ensure this, the data quality is continuously monitored during data taking. Only data certified by this monitoring is used for further analysis. The analysis presented in part II of this thesis is considered the Run-2 measurement of $\text{t}\bar{\text{t}}\text{H}(\text{b}\bar{\text{b}})$ at CMS. It includes the majority of the data recorded in the Run-2 operation period. Only the short operation period during 2015 is excluded[8]. A total integrated luminosity of $137.2\,\text{fb}^{-1}$ was collected during the LHC Run-2 at 13 TeV between 2016 and 2018 [100, 102, 103].

### 2.4.3. Pileup

The number of pileup collisions is Poisson distributed with a mean value of

$$\left\langle N_{\text{data}}^{\text{PU}} \right\rangle = \sigma_{inel}\mathcal{L}_{\text{inst}} \tag{2.6}$$

with the instantaneous luminosity $\mathcal{L}_{\text{inst}}$ and the inelastic proton-proton cross section $\sigma_{inel}$. The concept of pileup was introduced in section 1.2.3.3. The distribution of the mean

[8] In the 2015 data-taking period, the CMS detector recorded $3.8\,\text{fb}^{-1}$ of collision data. Due to a malfunction in the solenoid, only $2.7\,\text{fb}^{-1}$ of data were recorded with fully operational subdetectors.

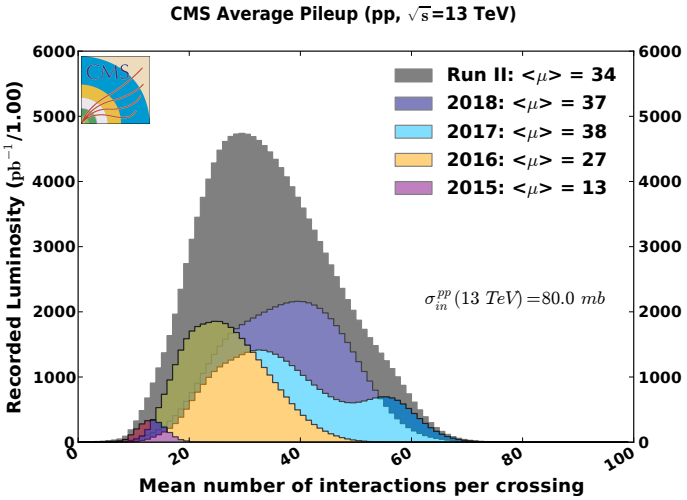



Figure 2.14.: Distribution of the average number of interactions per crossing (pileup) for p-p collisions for the Run-2 operation period and in the individual data-taking periods. Taken from reference [101].

number of interaction in Run-2 of the LHC operation is shown in figure 2.14. During Run-2, the mean number of interaction steadily increased, mainly due to increasing instantaneous luminosity. For jet-based analyses, like to one presented in part II of this thesis, this presents a huge challenge since more pileup interactions increase the chance of including pileup jets in the primary interaction, which can reduce the sensitivity. Various methods for reducing this dependency, like the previously mentioned CHS or dedicated algorithms, are considered within the CMS collaboration. Furthermore, increasing pileup effects jet-based triggers because of the increased hadronic activity and worse mitigation strategies at trigger-level compared to the fully reconstructed events. This usually present itself in the form of increased rates at constant thresholds. This generally leads to tighter trigger thresholds in later run periods of the Run-2 operation.

# 3. Signal identification and extraction

This chapter describes methods for signal identification and extraction on a general level that are used in the analysis presented in part II of this thesis. Two methods that play a major role in the signal identification are the matrix element method and fully-connected feed-forward neural networks, which are introduced in sections 3.1 and 3.2, respectively. The matrix element method provided the main discriminant in previous iterations of the fully-hadronic $t\bar{t}H(b\bar{b})$ analysis [10, 11]. In order to improve the sensitivity using the full Run-2 dataset, a neural-networks-based classifier, which combines the Matrix element discriminant with other observables, has been developed.
The signal extraction is based on the maximum likelihood method and will be discussed, as used within the ATLAS and CMS collaborations, in section 3.3. This section also introduces relevant topics like the statistical model (section 3.3.3) and upper limits (section 3.3.4).

## 3.1. Matrix element method

The matrix element method (MEM) is a likelihood-based technique resting on the leading order (LO) matrix element for the signal process and one or more background processes. The MEM was initially used by the D0 collaboration for measuring the mass of the top quark [104] and helicity of the W boson [105]. After being used by, among others, the CDF experiment for Higgs-boson measurements [106], it was introduced at CMS in the $t\bar{t}H(b\bar{b})$ analysis of the semileptonic channel at 8 TeV [107]. In the fully-hadronic $t\bar{t}H(b\bar{b})$ decay channel it was first used in the analysis of $36.3\,\mathrm{fb}^{-1}$ of data collected by the CMS experiment at 13 TeV [10]. Due to its strong foundations in the theoretical modeling of particle physics and high separation power between the considered processes, it is a popular alternative to statistics-based multivariate methods like Boosted Decision Trees (BDT) or (deep) neutral networks.

This section is intended as a summary of the method and its application in the fully-hadronic $t\bar{t}H(b\bar{b})$ analysis. An extensive introduction and review can be found in reference [108], which this section is based on.

The MEM uses the full kinematic properties of an event to calculate the probability of an event originating from a certain process. The probability density is calculated based on the cross section formulas, which include the LO matrix element amplitude, for the given process, the parton luminosity, the phase space, and a transfer function accounting for

detector effects. Densities are calculated for relevant signal and background processes and combined in a likelihood ratio. Section 3.3.2 will discuss the concept of likelihood ratios in detail.
In the fully-hadronic $t\bar{t}H(b\bar{b})$ analysis, the matrix element of the LO gluon-initiated $t\bar{t}H$ process is used for the signal hypothesis and the LO gluon-initiated $t\bar{t} + b\bar{b}$ process is used for the background hypothesis. No dedicated QCD multijet matrix element is used. Considering only gluon-initiated processes is reasonable since the majority of the events originate from such a process. In addition, only the hadronic decays of the Higgs boson and the top-quark decay-products are considered. Due to the large ambiguity in assigning the jets to the underlying quarks, all possible jet-quark assignments are considered. In order to constrain the number of possible combinations, only the four leading jets, according to the b-tagging discriminator value, are considered as candidates for the bottom quark, while all other jets are considered light-quark candidates. Hadronization and detector response effects are taken into account with transfer functions, that are derived from simulation.
One of the main limiting factors for the MEM is the computational complexity of solving the necessary integrations, that need to be processed multiple times per considered event[1].

The main ingredients for the MEM discriminant are the probability densities $\omega_i$ for the signal (S) and and background (B) hypothesis $i = \mathrm{S}, \mathrm{B}$:

$$\omega_i(\boldsymbol{y}) = \int \frac{1}{s} \frac{g(x_a; Q_i) g(x_b; Q_i)}{x_a x_b} \delta^4 \left( x_a P_a + x_b P_b - \sum_{q=1}^{8} p_q \right) |\mathcal{M}_i(\boldsymbol{x})|^2 W(\boldsymbol{y}|\boldsymbol{x}) d\Phi dx_a dx_b \ , \tag{3.1}$$

with the centre-of-mass energy $\sqrt{s}$, the gluon parton density functions $g(x; Q)$ evaluated as the factorization scale $Q$, the momentum fractions of the initial state partons (gluons) $x_a$ and $x_b$, the 4-momenta of the colliding partons $P_a$ and $P_b$, the production and decay amplitude $\mathcal{M}_i(\boldsymbol{x})$, the transfer functions $W(\boldsymbol{y}|\boldsymbol{x})$, and the phase space $d\Phi$. The gluon PDFs are calculated with `LHAPDF` [109] using the `CTEQ6.6` [110] PDF set.

### 3.1.1. Construction of the phase space

The phase space of the final-state includes three quarks per top-quark decay and two from the $b\bar{b}$ pair from the Higgs-boson or $t\bar{t} + b\bar{b}$ decay. The phases space volumes can be written as

$$d\Phi_{t_i} = \frac{1}{(2\pi)^9} \frac{d\boldsymbol{q_i}}{2E_{q_i}} \frac{d\boldsymbol{q_i'}}{2E_{q_i'}} \frac{d\boldsymbol{b_i}}{2E_b} \text{ and } d\Phi_{b\bar{b}} = \frac{1}{(2\pi)^6} \frac{d\boldsymbol{b}}{2E_b} \frac{d\bar{\boldsymbol{b}}}{2E_{\bar{b}}} \ , \tag{3.2}$$

with the 4-momenta for the quarks ($q_i$ and $q_i'$) and bottom quarks ($b_i$ and $b/\bar{b}$). The total phase-space volume is then given by

$$d\Phi = d\Phi_{t_1} d\Phi_{t_2} d\Phi_{b\bar{b}} \ . \tag{3.3}$$

### 3.1.2. Production and decay amplitudes

The decay amplitudes of the $t\bar{t}H(b\bar{b})$ and $t\bar{t} + b\bar{b}$ processes are composed of two elements: The scattering amplitude of the core $t\bar{t}H$ and $t\bar{t} + b\bar{b}$ processes and the decay amplitudes

[1] While the statement on the computational intensity is certainly true, its importance is often exaggerated when comparing the MEM to statistical methods like BDTs and deep neural networks. While the evaluation on an event-by-event basis is very fast for such methods, it has to be considered, that the necessary training and optimization, which are not necessary for the MEM, is also a computationally expensive process that is often omitted in these comparisons.

of the top quark, Higgs boson and W boson:

$$|\mathcal{M}_{\mathrm{S}}|^2 = |\mathcal{M}_{gg\to \mathrm{t\bar{t}H}}|^2 \cdot |\mathcal{M}_{t\to qqb}|^2 \cdot |\mathcal{M}_{t\to qqb}|^2 \cdot |\mathcal{M}_{\mathrm{H}\to \mathrm{b\bar{b}}}|^2 \text{ and} \tag{3.4}$$

$$|\mathcal{M}_{\mathrm{B}}|^2 = |\mathcal{M}_{gg\to \mathrm{t\bar{t}+b\bar{b}}}|^2 \cdot |\mathcal{M}_{t\to qqb}|^2 \cdot |\mathcal{M}_{t\to qqb}|^2 \ . \tag{3.5}$$

The hard scattering amplitude is calculated at LO with `OpenLoops` [111]. For each integration point, the phase space point, consisting of the mass, energy and momentum of the incoming and outgoing particles is passed and the amplitude squared is returned. For the outgoing particles the momenta and masses are given by:

$$\boldsymbol{p}_{t_i} = \boldsymbol{p}_{q_i} + \boldsymbol{p}_{q_i'} + \boldsymbol{p}_{b_i} \quad \text{and} \quad M_t = 174.3\,\mathrm{GeV} \ , \tag{3.6}$$

$$\boldsymbol{p}_H = \boldsymbol{p}_b + \boldsymbol{p}_{\bar{b}} \quad \text{and} \quad M_H = 125\,\mathrm{GeV} \text{ , and} \tag{3.7}$$

$$\boldsymbol{p}_{b,\bar{b}} = \boldsymbol{p}_{b,\bar{b}} \quad \text{and} \quad M_b = 0 \ . \tag{3.8}$$

The momenta for the incoming gluons are calculated by taking ± of the sum of the total energy and net z-momentum of the outgoing particles.

### 3.1.3. Transfer functions

The transfer functions encode the probability for observing the measurement $\boldsymbol{y}$ given the truth $\boldsymbol{x}$ in the calculation of the MEM probability density and take the detector effects and simulation into account.
The transfer function $W(\boldsymbol{y}|\boldsymbol{x})$ is defined using the jet directions $\Omega = (\cos\theta, \phi)$, the jet transverse momenta $p_{\mathrm{T}}$ and the missing transverse momentum $\not{\boldsymbol{p}}_{\mathrm{T}}$ as observables:

$$W(\boldsymbol{y}|\boldsymbol{x}) = \left( \prod_{i=0}^{8} \delta(\Omega_i - \Omega_{q_i}) T(p_{T,i}|p_{T,q_i}) \right) \cdot F\left(\not{\boldsymbol{p}}_{\mathrm{T}}|\boldsymbol{0}\right) \ , \tag{3.9}$$

where the $p_{T,q}$ and $\Omega_q$ are truth values of the quarks that produce a jet and there is zero $\not{\boldsymbol{p}}_{\mathrm{T}}$ at quark level. This further includes the assumptions

- that the directions of the quarks are perfectly measured and are given by the associated jet, and
- that the momenta of jets and quarks as well as the missing transverse momentum are independent.

The quark-$p_{\mathrm{T}}$ transfer functions $T(p_{T,i}|p_{T,q_i})$ are parametrized in terms of a double gaussian and can be viewed as a probability density function, that returns the probability of observing a jet for a given quark:

$$T(p_{T,i}|p_{T,q_i}) = p_0 \left[ 0.7 \exp \frac{-(p_T - p_1)^2}{p_2^2} + 0.3 \frac{(p_T - p_3)^2}{(p_2 + p_4)^2} \right] \ . \tag{3.10}$$

The parameters $p_1$-$p_4$ are the only parameters of interest and have been derived from signal-simulation and are functions of the $p_{\mathrm{T}}$ and $\eta$ of the quarks:

$$p_n = p_n(p_{T,q}, \eta_q), \quad n = 1, 2, 3, 4 \tag{3.11}$$

The transfer function for the missing transverse energy $\not{\boldsymbol{p}}_{\mathrm{T}} = (\not{p}_{\mathrm{x}}, \not{p}_{\mathrm{y}})$ can be expressed as

$$F\left(\not{\boldsymbol{p}}_{\mathrm{T}}|\boldsymbol{0}\right) = \frac{1}{2\pi\sigma^2} \exp(-\not{p}_{\mathrm{x}}^2/2\sigma^2) \exp(-\not{p}_{\mathrm{y}}^2/2\sigma^2) \ . \tag{3.12}$$

The parameters $p_1$ to $p_4$ (from eq. (3.10)) are determined in 58 $p_{\mathrm{T}}$ bins between 30 and 300 GeV and two $\eta$ bins from 0 to 1.0 and 1.0 to 2.5. This choice is motivated by the fact,

that the parameters have a continuous dependence on the $p_{\mathrm{T}}$ while the $\eta$ dependence is only subdominant. Furthermore, the parameters are derived independently for light ($q$= $u$,$d$,$c$,$s$) and heavy ($q$=$b$) quarks as well as for each data-taking period.

### 3.1.4. Integration, permutations and hypotheses for events with 7, 8 and $\geq$ 9 jets

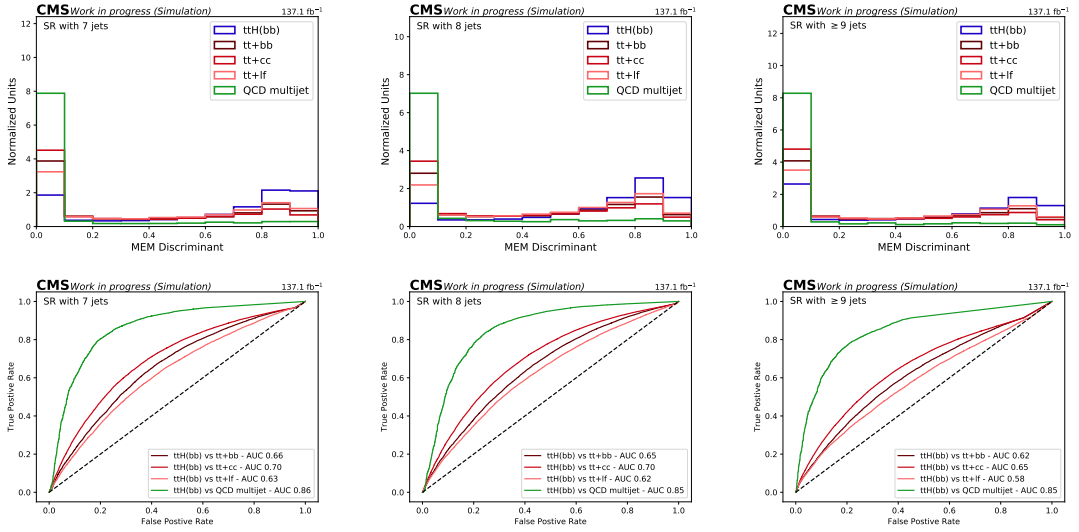
For each event in the analysis region, the probability density is calculated for the signal and background hypotheses, considering multiple permutations of quark-jet assignment. The approach and implementation will be outlined in the following:

The integration over the phase space for each permutation is done numerically with the `CUBA` [112] implementation of the `VEGAS` [113] algorithm, which is an iterative and adaptive Monte-Carlo integration method. It relies on sampling the integral in multiple regions with the largest contribution to the integral and uncertainty. For this, it processes multiple iterations of the integral. In each iteration, the multidimensional search grid is refined with respect to the previous iteration in order to concentrate the sampling. The maximum number of iterations is tuned per considered hypothesis in order to achieve the desired precision and computing time. The integration is made feasible by applying various transformations to equation (3.1). For a fully-reconstructed signal event, the dimensions are reduced to three (two over the light-quark energy from each top quark and one over the bottom-quark energy from the Higgs boson) and to four for a $\mathrm{t\bar{t}+b\bar{b}}$ event (one additional integration over the second bottom-quark energy). In case of a lost jet, an additional two-dimensional integration is necessary (over the quark directions). The details on the transformations are omitted in this summary but are extensively covered in reference [108]. For each iteration, the integrand, in a given phase-space point for each permutation, is calculated including an evaluation of the parton density function with `LHAPDF`, the matrix element calculation with `OpenLoops`, and the transfer function evaluation. Integration ranges are set to $E \pm 3\sigma_E$ for the quark energy $E$, with the jet energy resolution $\sigma_E$, $\pm\pi/2$ for the quark polar angle $\theta$, and $[0, 2\pi]$ for the quark azimuthal angle $\phi$.

In order to control computing time, the quark-jet assignment is constrained such, that only the four jets with the largest b-tagging discriminant value are considered for b-jets and all other jets are considered light-jets and are permutated among each other. Furthermore, only inequivalent permutations are considered. An example for an equivalent permutation would be interchanging the light quarks from the decay of the W boson. The maximum number of light jets considered in an event is five. In case more light jets are present, the five jets with the highest $p_{\mathrm{T}}$ values are used. Multiple configurations, of the number and jet-quark assignments and/or non-reconstructed jets, were studied. The performance in separating signal and background as well as the required computational resources were considered in these studies with the following results [108]:
For events with 7 or 8 jets (after selection), the MEM is calculated with the configuration that one light-quark from the W-boson decays was lost but all bottom quarks from the top-quark and Higgs-boson decays are reconstructed. For events with $\geq$ 9 jets, the MEM is calculated with the configuration that all four light jets from the W boson and all bottom quarks from the top-quark and Higgs-boson decays are reconstructed. In order to calculate the total event probability, the probability density for each permutation is summed. Hereby, the assumption is that unlikely jet-quark associations will result in a low probability and therefore contribute little to the overall event probability.

When used in the analysis, the signal and background hypotheses are combined in the likelihood-ratio

$$\frac{\omega_{\mathrm{S}}(\boldsymbol{y})}{\omega_{\mathrm{S}}(\boldsymbol{y}) + \kappa \cdot \omega_{\mathrm{B}}(\boldsymbol{y})} \ , \tag{3.13}$$




(a) SR with 7 jets (b) SR with 8 jets (c) SR with $\geq 9$ jets

Figure 3.1.: The top row shows the distribution of the MEM discriminant, as defined in equation (3.13) with $\kappa = 0.065$, for the three signal regions (SR) with 7, 8 and $\geq 9$ jets of the fully-hadronic $t\bar{t}H(b\bar{b})$ analysis presented in part II. The distributions are shown for the $t\bar{t}H(b\bar{b})$ signal processes as well as the largest background processes: $t\bar{t} + b\bar{b}$, $t\bar{t} + c\bar{c}$, $t\bar{t} + \text{LF}$, and QCD multijet. Additionally, the separation power between the signal and each background process provided by the MEM is shown as ROC curve in the bottom row. The ROC curve as performance indicator will be explained in detail in section 3.2.1.

which is referred to as the MEM discriminant. The factor $\kappa$ is considered a hyperparameter and subject to optimization[2]. The distribution of this discriminant is contained in the interval $[0, 1]$ and usually binned into 10 equal width bins. The MEM discriminant, as defined in equation (3.13), is shown in figure 3.1 for the three signal regions with 7, 8, and $\geq 9$ jets, that are used in the analysis presented in part II. It can be observed, that the MEM discriminant provides significant separation between the $t\bar{t}H(b\bar{b})$ signal and all large backgrounds from the $t\bar{t} + b\bar{b}$, $t\bar{t} + c\bar{c}$, $t\bar{t} + \text{LF}$, and QCD multijet processes. Especially against the dominant QCD multijet background process, a strong separation can be observed even though no dedicated hypothesis is considered for this background. Intuitively this can be understood by considering that the kinematics of the jets are very different in case of the QCD multijet background, e.g. no top quarks are involved, even though they pass the same selection criteria.

The uncertainty variations of the jet energy scale and resolution[3] in simulated events are taken into account by calculating the probability densities using the vectorial implementation of the `CUBA` integrator. In this implementation, which is extensively reviewed in reference [114], the integration is carried out over a vector encoding all the variations. The method is roughly 2-3 times as computationally expensive as the scalar integration

[2] In the previously published iterations of the fully-hadronic $t\bar{t}H(b\bar{b})$ analysis [10, 11], in which the MEM discriminant was used as distribution for the signal extraction, $\kappa = 0.065$ has been used. This value was determined by scanning the parameter and evaluating the expected significance in the signal extraction. For the analysis presented in part II, the same numerical value is used.

[3] For the jet energy scale and resolution, as introduced in section 2.3.5, multiple sources of systematic uncertainty are considered by shifting the kinematics of each jet in an event. The exact sources of uncertainty and how the shifts are implemented will be covered in section 6.2.

but calculating the probability densities sequentially for each of the $n > 20$ variations[4] in each of the many million events would not be feasible. The probability only depends on the observed kinematics of the jets via the transfer function and therefore it is the only factor in which the vectorial approach changes the calculation. In practice, the integrator expects the observed nominal kinematics $\boldsymbol{y}$ of the jets as well as the relative shifts $\delta\boldsymbol{y}_i$ for a given source of systematic uncertainty $i$. This vectorial process can be described by changing equation (3.1) as follows:

$$\begin{pmatrix} \omega(\boldsymbol{y}) \\ \omega(\boldsymbol{y}+\delta\boldsymbol{y}_1) \\ ... \\ \omega(\boldsymbol{y}+\delta\boldsymbol{y}_n) \end{pmatrix} = \int \eta(x_a,x_b)\,|\mathcal{M}_i(\boldsymbol{x})|^2 \begin{pmatrix} W(\boldsymbol{y}|\boldsymbol{x}) \\ W(\boldsymbol{y}+\delta\boldsymbol{y}_1|\boldsymbol{x}) \\ ... \\ W(\boldsymbol{y}+\delta\boldsymbol{y}_n|\boldsymbol{x}) \end{pmatrix} d\Phi dx_a dx_b \ . \tag{3.14}$$

This procedure is used when calculating the MEM probability in simulated event. Collected data events, which are not subject to the variations, are treated differently. In this case, the integration is executed with the same algorithms but the vector only represents the nominal kinematics $\boldsymbol{y}$, basically leading to a "scalar" integration. As an alternative configuration, it was studied to execute the integration with a vector of the same lengths $n+1$ as used in the simulated events, but the nominal kinematics are used for all elements of the vector, i.e. $\delta\boldsymbol{y}_i = \boldsymbol{0}$ for all $i$. This reflects the concern, that the nominal numerical result of the integration depends on the fact, that a vectorial integration was executed. Both approaches lead to the exact same numerical results, eliminating this concern. In practice, the "scalar" integration is used because it is computationally less expensive.
While it could be excluded, that the vectorial integration itself introduces a differences to the nominal numerical results of the integration, a dependency on the size of the variations $\delta\boldsymbol{y}_i$ was observed. This dependency is only relevant because the vectorial approach has to be used in simulation but the cannot be used in collected data. Since the "scalar" approach is used in the latter case and the two cases are compared in the signal extraction, the residual difference between both approaches should be as small as possible. In order to observe the dependency on $\delta\boldsymbol{y}_i$, vectorial integrations with variations, as used in simulated event, and without variations have been directly compared on a set of simulated event. This dependency can be understood by considering, that the integration search grid of the algorithm is set according to the "global" values of the kinematics, i.e. according to $\boldsymbol{y}+\delta\boldsymbol{y}_i$, and not the nominal kinematics only, i.e. according to $\boldsymbol{y}$. Hence, when variations $\delta\boldsymbol{y}_i$ are introduced, the resulting search grid can be different from the nominal case. This is especially noticeable when the variations are large compared to the nominal kinematics. The effect of this dependency on the overall distribution of the MEM discriminant in simulated $\mathrm{t\bar{t}H(b\bar{b})}$ events is shown in figure 3.2 for the regular vectorial integration and the integration where the nominal kinematics are used for all elements of the vector, i.e. $\delta\boldsymbol{y}_i = \boldsymbol{0}$ f1r all $i$. The latter is equivalent to the "scalar" integration used in the collected data events. It can be observed that the residual differences between the two integrator configurations are small and covered by the statistical uncertainty of the sample. While the statistical uncertainty itself is not necessarily meaningful in figure 3.2, since the exact same events are considered for both distributions, it nonetheless led to the following conclusion for the data. Any differences introduced by this behavior of the integrator are smaller than the statistical fluctuations expected from the data when considering the whole dataset and can therefore be neglected.

[4] The exact number of considered sources of jet energy scale and resolution uncertainty depends on the uncertainty model used in a particular analysis. In the analysis presented in II, 13 individual sources, each with an up and down variation, are considered. Therefore the MEM would need to be calculated 27 $(13 \cdot 2 + 1)$ times in each event without the vectorial implementation of the integration.

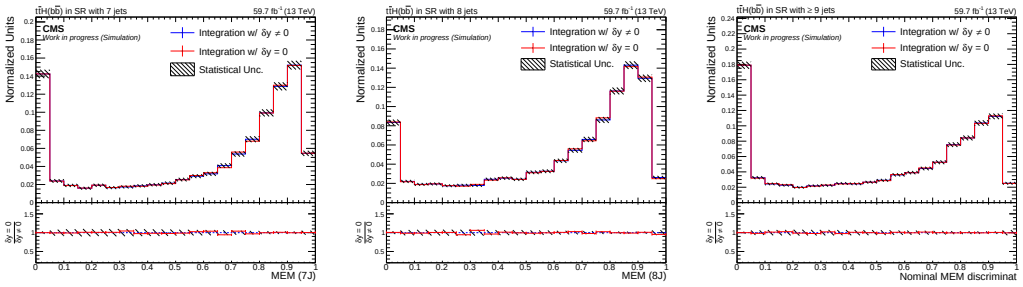



Figure 3.2.: Distribution of the nominal MEM discriminant in simulated $t\bar{t}H(b\bar{b})$ events for the 2018 data-taking period for the signal regions with 7 (left), 8 (middle), and $\geq 9$ jets (right). The results of the integration without kinematic variations (nominal kinematics are used for all elements of the vector $\leftrightarrow \delta \boldsymbol{y}_i = \boldsymbol{0}$) are shown in red and the results of the integration with kinematic variations ($\delta \boldsymbol{y}_i \neq \boldsymbol{0}$) is shown in blue. The exact same events are used for both distributions and the blue distribution corresponds to the configuration which has to be used in simulated events. The statistical uncertainty of the blue distribution is shown as hatched band.

## 3.2. Neural Networks

This section aims to summarize the machine learning methods and techniques relevant for the analysis presented in part II of this thesis. A detailed introduction to the topic can be found in [115], which will also be used as main reference for this section.

### 3.2.1. Introduction to machine learning

Many different machine learning algorithms have been developed over the last decades and with increasing amounts of data available in all areas of modern life, the applications grew. Traditional algorithms like logistic regression, neural networks, support vector machines or decision trees are still in use today but are being replaced more and more by algorithms based on deep learning approaches, that utilize the basic principles of neutral networks. A number of key circumstances led to this development: Hinton et al. demonstrated in 2006 in reference [116] that adding successive layers to the inside of a neural network, i.e. using a "deep" network, can drastically improve the performance. Advancements in GPU-based computing[5] made using these algorithms feasible since such systems are highly efficient for the necessary computations. Finally, these algorithms profit from large amounts of data, whose storage and processing was not possible only a few years ago. Deep neural networks (DNN), Convolutional neural networks (CNN), and recurrent neural networks (RNN) are just a few of the more commonly used algorithms of this type. Before diving deeper into the details of the relevant algorithm for this thesis in section 3.2.2, a brief introduction to machine learning will be given.

In general, a machine learning algorithm is an algorithm that learns from *data* (or *experience*) to fulfill a *task* by minimizing a *performance measure*.

#### The Task

Common tasks that find mainstream application in particle physics can be classified as prediction (or regression) and classification tasks. Tasks like anomaly detection (e.g. theory independent searches), denoising (e.g. improvement of calorimeter measurements) or

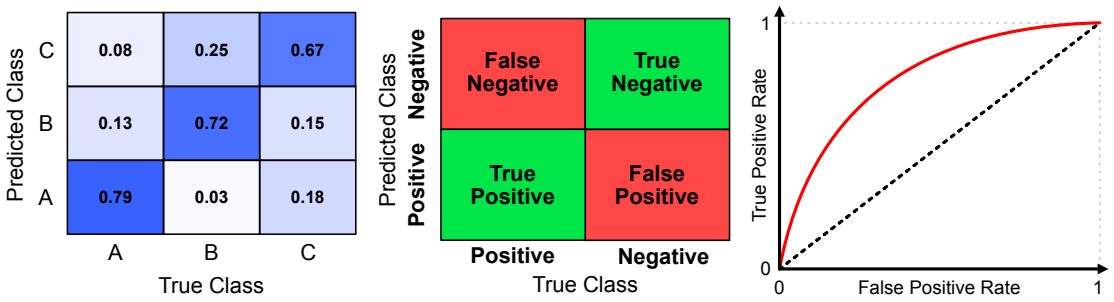
[5] Graphics Processing Unit (GPU) are specialized electronics circuits optimized for generating and processing images intended to output on a display. These processors exhibit a highly parallelized structure with fast and broad memory access. This makes them also optimal for the algorithms used in deep learning, since the main operations are matrix multiplication and manipulations as discussed in section 3.2.2.

sampling (e.g. for event generation) are becoming more and more popular within the particle-physics community. Example applications are the improvement of energy resolutions [117] (regression) or b-jet identification [99] and the construction of strongly discriminating distributions for fits to data (classification). Since the neural network, that is used in the analysis presented in part II of the thesis, is trained to classify signal and background event, these tasks will be described in more detail.
In a classification task, the algorithm is trained to assign an input to one of multiple categories $k$. The algorithm generates a function $f : \mathbb{R}^n \to \{1, ..., k\}$, with the dimensionality of the input space $n$. Since the assignment to one of the categories $k$ is ambiguous and generally the target of the algorithm, the function is represented by $\hat{y} = f(\boldsymbol{x})$ with the $n$-dimensional input vector $\boldsymbol{x}$ and the numerical output $\hat{y} \in [0, 1]$. If an actual classification into one of the categories $k$ is desired, a common approach is to define a working point in $\hat{y}$ in terms of the true-positive (fraction of inputs correctly assigned to $k$) and false-positive rates (fraction of input falsely assigned to $k$) that are acceptable, as it is done for the b-tagging described in section 2.3.6. Generally, classification algorithms fall into one of the two categories: *binary classification*, if two classes are considered, and *categorical classification*, if more than two classes are considered. The classifier for the fully-hadronic $t\bar{t}H(b\bar{b})$ analysis presented in the thesis, which will be described in section 5.6, is a binary classifier. A categorical approach was investigated for the fully-hadronic channel and is used in the dileptonic and semileptonic channels, which will be described in 7.1. The concept will be further discussed briefly in this section. In particle physics another commonly utilized approach is to use the distribution of $\hat{y}$ as final discriminant for the fit to data, which will be the topic of section 3.3. In this case, the separation between signal and background in the distribution of $\hat{y}$ is exploited to increase the sensitivity in the signal extraction.

### The Data

To train a machine learning algorithm, a dataset is necessary. A dataset is a structured collection of many instances, each characterized by a number of inputs (also called features). Generally, machine learning algorithms can be categorized in the types *supervised* and *unsupervised* depending on the dataset used for training. For supervised algorithms, the true category (or target in the case of regression) $y$ needs to be defined for every instance in the dataset. This is also called labelled data and accordingly the true category is called the label. Usually the algorithm learns to minimize the difference between the label $y$ and the prediction $\hat{y}$. This is by far the most commonly used class of algorithms in particle physics. In the case of unsupervised algorithms, no label is specified. Hence, the algorithm can only learn certain properties from the features of the dataset. These algorithms are rarely used for classification and are more commonly used for clustering, sampling or anomaly detection[6] tasks and are used much less frequently in particle physics at the time of writing this thesis. Furthermore, the lines between the two types of algorithms are blurred and there are other types like reinforced learning, semisupervised learning or

[6] Autoencoders (an introduction can, for example, be found in reference [115] chapter 14) are a form of unsupervised learning algorithm that can be used to find outliers in a dataset. In order to do this, the autoencoder firstly reduces (or encodes) the input data to a lower dimension and then expands (or decodes) the reduced representation back to the dimension of the input data. By minimizing the difference between input and the output of the decoder, the algorithm learns how the general features of the dataset usually look in the dataset. This is an interesting approach for particle physics because it can be used to find rare events without depending on an underlying theory. Furthermore it can be directly trained on the collected data. A very small contribution of the "new and rare" events to the dataset is no issue since they are statistically insignificant relative to the "regular" (or in the case of the LHC, QCD-like) events. An autoencoder-based approach was investigated for the fully-hadronic $t\bar{t}H(b\bar{b})$ analysis presented in this thesis but ultimately could not yield a competitive result within the constrains of the analysis.




(a) (b) (c)

Figure 3.3.: Illustrations of confusion matrices (a) and (b) and a Receiver-Operator Characteristic or ROC (c). The confusion matrix (a) illustrates the classification with three classes A, B and C and the number given in each box is the share of inputs assigned to this class by the classifier. The confusion matrix (b) illustrates a classification with two classes and can be used to visualize the terms *True Positive*, *False Positive*, *False Negative* and *True Negative*. The ROC curve (c) - shown in red - shows *False Positive* and *True Positive* rates for multiple threshold on the classifier output $\hat{y}$. It is usually show relative to the diagonal (dashed line), which would be expected from randomly selected one of two classes.

multi-instance learning.
The algorithms utilized in this thesis are of the supervised type and therefore the other types will not be covered in any more detail in the following sections.

### Measuring the performance

The performance of a machine learning algorithm is measured at two different stages. Firstly during the training of the algorithm and secondly the overall performance of a trained algorithm.

The performance measure during training, also called the cost-function, is used to evaluate the abilities of a machine learning algorithm to achieve the training objective and is specific to a given task. For (supervised) classification tasks, the error-rate or loss, which is the difference between the output $\hat{y}$ of the algorithm and the expected output $y$, is commonly used. During training this loss is minimized by updating the parameters of the model regularly after evaluating training data.

Two measures that are commonly used to evaluate the overall performance of a trained classification model, are the confusion matrix and the ROC curve.
A confusion matrix is a way to illustrate the prediction of the classifier relative to the true class of a dataset. Figure 3.3a is a confusion matrix of a classifier with three classes. The diagonal elements are the correct assignments while the off-diagonal elements are wrong assignments. In this case, the relative frequency of assignments to a class are given as numerical value.
If only two classes, of which one is characterized positive and one negative, are considered, the terms *true positive*, *false positive*, *false negative* and *true negative* are commonly used to describe the assignment to the classes. This is illustrated in figure 3.3b. In particle physics, signal is usually considered positive and background negative.
The Receiver-Operator Characteristic (ROC) is commonly used to compare how well various algorithms are able to discriminate between two classes. It is not limited to machine learning algorithms but can be used for any continuous distribution separating two classes,

like the MEM introduced in the previous section. It shows the rates of correctly and wrongfully assigning classes as a function of thresholds placed on the classifier output value and can be understood as the separation power of a given distribution between two classes. An illustration can be found in figure 3.3c. For this thesis, it was chosen to show the ROC curve as a function of the true-positive rate and false-positive rate. A perfect classifier would have a false-positive rate of 0 at a true positive rate 1. At a false-positive rate of 1 the true-positive rate is always 1. Since classifiers are generally not perfect, choosing the threshold for the class assignment is a trade off between both values. Another performance indicator can be derived from this curve: The Area-under-curve (AUC) of the ROC. A perfect classifier would reach an AUC of 1, while randomly drawing one of two classes classes should result in an AUC of 0.5 and corresponds to the diagonal in the ROC-curve plot. Therefore, the diagonal, shown as dashed line in figure 3.3c, is usually included as a reference when showing ROC curves. An AUC $< 0.5$ is not expected for a classifier and usually indicate that the classes have been interchanged[7].
ROC curves are calculated as follows: Given normalized distributions of the classifier for the signal and background classes, the space of the classifier output is scanned and at each threshold, the integral of each distribution below and above the threshold is calculated. For the signal distribution the integral above the threshold is the true-positive rate. For the background distribution the integral above the threshold is the false-positive rate. These rates are sorted and can then be plotted as shown in figure 3.3c. This description assumes, that larger values in the distributions point to the signal. If this is not the case the procedure must be inverted.

### Model generalization and regularization

When training a machine learning algorithm, the main goal is that it performs well on a dataset, that was not used for training. The ability to do this is called *generalization*. When training the algorithm, the training error, as defined by the cost function, is minimized. The ability to generalize is measured with the test error (or generalization error), which is evaluated on an independent test dataset. This dataset is usually randomly sampled from the full dataset before training the algorithm. This introduces a second goal, in addition to finding the function minimizing the cost-function, in the form of minimizing the difference between the error on the training and test set. If the model is not able to achieve a sufficiently small training error it is called underfitting and if it shows very different training and test errors it is called overfitting. The simplest strategy to deal with both errors is to change the model, by e.g. reducing the number of input features or changing the architecture. Another way to improve the generalization (or reduce the generalization error) of the model is *regularization*. One way to do this, is to add a penalty (also called regularizer) to the cost function, which can be understood as giving the model a preference for the function it tries estimates. One commonly used penalty is the so-called *weight decay*. Hereby the model is encouraged to use small weights (or parameters). Intuitively, this can be understood as a prior, that small weights are expected while large weights point to outliers in the data that the model tries to reach with its function estimation.

### Hyperparameters

The term hyperparameters refers to parameters of the model that are not effected by the training. This includes, among others, the model architecture, but also the utilized regularization methods, the settings of the methods or the number of input features. Tuning these hyperparameters with the goal of maximizing performance but also improving stability is a major part of training a machine learning model.

[7] In practice, this could mean that the signal distribution of given discriminant tends towards 0 while the calculation expects it to tend to 1.

### 3.2.2. Fully-connected feed-forward neural networks

Fully-connected feed-forward neural networks are the quintessential algorithm for modern deep learning applications. As discussed previously, the goal is to approximate the function $f$ given a dataset with many sets of inputs $\boldsymbol{x}$ and labels $y$. By learning a set of parameters $\boldsymbol{\theta}$ the networks approximates $y = f(\boldsymbol{x}; \boldsymbol{\theta})$. The basic components of a network like this are:

**Nodes** are the basic building block of a neural network. A node can be understood as a non-linear function

$$a\left(\boldsymbol{x}; \boldsymbol{\theta}\right) = a\left(\boldsymbol{x}; \boldsymbol{w}, b\right) = \alpha\left(\sum_k w_k x_k + b\right) = \alpha\left(\boldsymbol{w}^\mathsf{T}\boldsymbol{x} + b\right) \quad (3.15)$$

with the *activation function* $\alpha$, *weights* $\boldsymbol{w} \in \mathbb{R}^{n\times 1}$, *bias* $b \in \mathbb{R}$ and input $\boldsymbol{x} \in \mathbb{R}^{n\times 1}$. The weights $\boldsymbol{w}$ and bias $b$ are the trainable parameters $\boldsymbol{\theta}$ in the model. In a network with multiple layers and nodes, the layer $i$ is indicated by a superscript and the node number $j$ as subscript as such: $a_j^{[i]}$. $a$ is also called the *activation*.

**Layers** are a collection of nodes. Each layer is fully connected to the previous and following layer but information is only processed in direction from input to output (feed-forward). Hence, no feedback connections are present. The values in each node of a layer are usually represented as vector with a dimension of the number of nodes. The weight vector $\boldsymbol{w}$ and bias $b$ of each node can be combined to a weight matrix $\boldsymbol{W}$ ($\boldsymbol{w}^\mathsf{T}$ stacked horizontally) and bias vector $\boldsymbol{b}$ for the layer. If the input dimension of the layer is $n_\text{in}$ and contains $n$ nodes, the dimension are: $\boldsymbol{w} \in \mathbb{R}^{n_\text{in}\times 1}$, $b \in \mathbb{R}$, $\boldsymbol{W} \in \mathbb{R}^{n\times n_\text{in}}$ and $\boldsymbol{b} \in \mathbb{R}^{n\times 1}$. The output of a layer is usually denoted as $\boldsymbol{a}^{[i]}$ for the $i^\text{th}$ layer in the network with dimension $\mathbb{R}^{n\times 1}$.

**The Input Layer** passes a set of $n_x$ features from the dataset to the network. It can be viewed as the $0^\text{th}$ layer with unity activation functions, bias $\boldsymbol{b}^{[0]} = 0$ and weight matrix $\boldsymbol{W}^{[0]} = \boldsymbol{I}_{n_x}$. For ease of use, this is usually omitted. Occasionally $\boldsymbol{a}^{[0]}$ is used to denote the values in this layer instead of $\boldsymbol{x}$.

**Hidden Layers** are fully connected to each node of the previous and next layer. For each node in layer $i$, equation (3.15) is evaluated and the output $\boldsymbol{a}^{[i]}$ is passed to the next layer. The network must decide itself during training how to best use these layers (or more specifically set the weights $\boldsymbol{W}^{[i]}$ and bias $\boldsymbol{b}^{[i]}$) to approximate $y = f(\boldsymbol{x}; \boldsymbol{\theta})$. The nodes in the layer can be viewed as new features, that the network derives from the features given as input. A practical example in high-energy physics could be, that the 4-vectors of multiple jets are used as input features and the network learns to construct, for example, certain dijet masses.

**The Output Layer** collapses the network to a single node (binary classification or regression) or $n_o > 1$ nodes (categorical classification) and calculates the output $\hat{y} \in \mathbb{R}$ or $\hat{\boldsymbol{y}} \in \mathbb{R}^{n_o\times 1}$, respectively.

These general components are illustrated in figure 3.4. Additionally the terms width and depth are commonly used to describe neural networks. The width of the model refers to

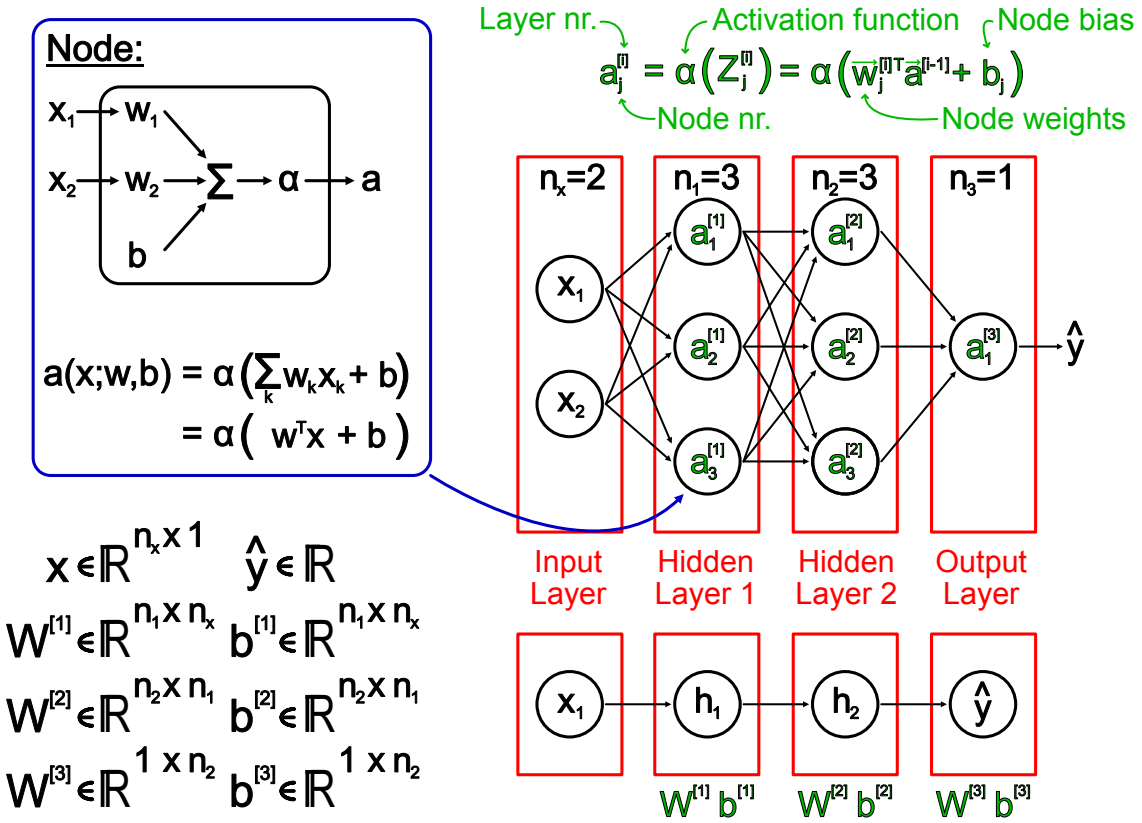
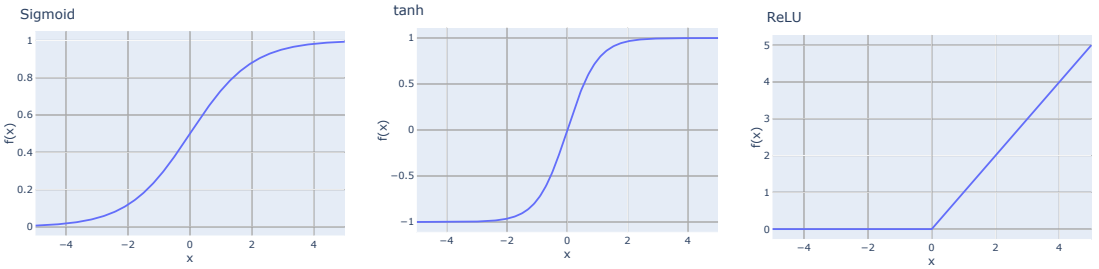



Figure 3.4.: Illustration of a fully-connected feed-forward neural networks with 3 layers, two input features and a single output node. The node output $a_j^{[i]}$ of node $j$ in layer $i$ is given as function of the weights $\boldsymbol{w}_j^{[i]}$ and bias $b_j^{[i]}$ of the node. The activation functions $\alpha$ are an important parameter when building the model and common choices are the ReLU for the hidden layers and sigmoid for the output layer. The box on the left illustrates the calculation inside a single node in the first layer. This can be identified by the use of $\boldsymbol{x}$ as node input, while other layers use the output of the previous layer as input. In this case layer 2 would use $\boldsymbol{a}^{[i]}$ as input. Two common ways of network visualization are illustrated on the right.




(a) $f(x) = \frac{1}{1+e^{-x}}$ (b) $f(x) = \tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}}$ (c) $f(x) = \max(0, x)$

Figure 3.5.: Functions commonly used as activation functions for the nodes of a neural network. The sigmoid (a), tanh (b) and rectified linear unit, or ReLU, (c) are shown.

the number of nodes in each layer. Depth refers to the length of the chained layers. The term deep learning arose for the trend (and technical possibility) to increase the lengths of this chain. Generally speaking, deeper networks excel and outperform traditional machine learning algorithms if large datasets are used.

The choice of activation functions plays an important role in the design of a neural network. While in principle any function could be used, various specific non-linear functions proved to be especially advantageous. Commonly used functions are the tanh

$$\alpha(x) = \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \ , \tag{3.16}$$

the sigmoid function

$$\alpha(x) = \sigma(x) = \frac{1}{1 + e^{-x}} \ , \tag{3.17}$$

and the rectified linear unit, or ReLU [118],

$$\alpha(x) = \max(0, x) \ . \tag{3.18}$$

These functions are depicted in figure 3.5. Generally, the ReLU function is considered a default choice for the hidden layers and sigmoid functions is commonly used in the output node of a binary classifier. These two function will also be used in the classifier developed for the analysis presented in section 5.6. For a categorical classifier, the softmax function is commonly used for the output node. The softmax function for a $K$-dimensional vector $\boldsymbol{z}$ is calculated as

$$f(\boldsymbol{z}) = \frac{\exp(\boldsymbol{z})}{\sum_{j}^{K} \exp(z_j)} \ . \tag{3.19}$$

The elements of the resulting vector (each corresponding to one of the output nodes of the network) can be interpreted as the probabilities of a set of inputs to belong to a certain class targeted by the classifier. This makes this function an interesting choice since the highest node output value can be used to categorize the input as belonging to the class represented by said output node.

### Training a neural network

When evaluating the network for a given a set of input features, only the weights $\boldsymbol{W}^{[i]}$ and biases $\boldsymbol{b}^{[i]}$ are necessary to calculate the output $\hat{y}$. This evaluation of a trained model is computationally fast compared to the training and depending on the algorithm complexity. The weights and biases are determined in the training of the network. Depending on the

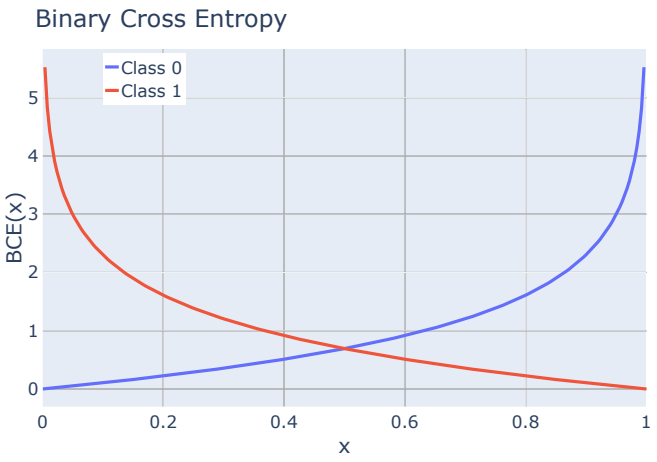



Figure 3.6.: Binary cross entropy (BCE), as defined in equation (3.21), for class 0 ($y = 0$, blue) and class 1 ($y = 1$, red). If the BCE is used as cost function, the penalty increases the further from the truth label $y$ the prediction $\hat{y}$ (or $x$ in the case of this plot) is.

architecture and size of the dataset, this can be computationally expensive. How this training procedure is implemented will be covered in the following. The major components are the *cost function*, *backward propagation* and *gradient decent*.

The choice of cost function $J(\hat{y}, y)$ is an integral aspect in the design of neural networks. The cost function can be a complex function comprised of multiple terms. One of these terms is the loss function $L$. In a simple model, $J(\hat{y}, y) = L(\hat{y}, y)$ can be used but more commonly, the cost function is defined as the average loss $L$ of $m$ input sets (also called a batch) with additional regularization term $\Gamma$:

$$J(\hat{y}, y) = \frac{1}{m} \sum_{i=0}^{m} L(\hat{y}_i, y_i) + \Gamma(\boldsymbol{\theta}) \ . \tag{3.20}$$

For classification tasks with fully-connected feed-forward networks, usual choices for the loss function are the *binary cross entropy*

$$\mathrm{BCE}(\hat{y}, y) = -\left(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right) \ , \tag{3.21}$$

with the prediction of the network $\hat{y} \in [0, 1]$ and the true label of the class $y = 0$ or 1, and the *categorical cross entropy*

$$\mathrm{CCE}(\hat{y}, y) = -\sum_{c=1}^{M} y_c \log(\hat{y}_c) \ , \tag{3.22}$$

with the number of classes $M$, the prediction $\hat{y}_c$ of a network for class $c$ and the label $y_c = 0$ or 1 specifying if the class $c$ is the true class of the current observation. As implied by the names of these two equations, the BCE is used for binary classification task and the CCE is used for categorical classification task. Figure 3.6 shows the BCE for the cases $y = 0$ and $y = 1$. That this is an excellent choice as a cost-function can be understood by considering the following property: The further from the true class $y$, the prediction of the network $\hat{y}$ is with the current paramter choice, the larger the cost function becomes.

In a feed-forward network, the information $\boldsymbol{x}$ flows forward through the network to estimate $\hat{y}$. This is also called *forward propagation*. When training the network, the calculation of

**Algorithm 1:** Forward propagation in a fully-connected feed-forward neural network with $l$ layers (each associated with a weight matrix $\boldsymbol{W}$, a bias $\boldsymbol{b}$ and an activation function $\alpha$), the single input $\boldsymbol{x}$ and true label $\boldsymbol{y}$. In this particular case, the cost function $J$ is equal to the loss function $L$. In a more realistic case, the cost function would average the loss of a sets of inputs (also called a minibatch) and include additional penalty terms for regularization.

$\boldsymbol{h}^{[0]} = \boldsymbol{x}$
**for** $i$=*1* **to** $l$ **do**
  $\boldsymbol{Z}^{[i]} = \boldsymbol{W}^{[i]}\boldsymbol{h}^{[i]} + \boldsymbol{b}^{[i]}$
  $\boldsymbol{h}^{[i]} = \alpha^{[i]}\left(\boldsymbol{Z}^{[i]}\right)$
**end**
$\hat{\boldsymbol{y}} = \boldsymbol{h}^{[l]}$ (When evaluating he network, forward propagation stops here.)
$J = L(\hat{\boldsymbol{y}}, \boldsymbol{y})$

the cost function is included in this step. The forward propagation for a simple network with $l$ layers, one input $\boldsymbol{x}$ is outlined in algorithm 1.

In order to use the information of the cost-function, it must flow backward through the network to compute the gradient used to optimize the weights. This is called backward propagation (or *backprop*). A common misconception is, that backprop refers to a training algorithm but in-fact it is only a method for calculating the gradient and other algorithms are used to perform the training based on this gradient. The backprop algorithm is also not specific to neutral networks but is a general algorithm to calculate derivates. In learning algorithms, the gradient that needs to be calculated most often is the gradient of the cost function $dJ(\boldsymbol{\theta}) = \frac{dJ(\boldsymbol{\theta})}{d\boldsymbol{\theta}}$ with respect to a set of parameters $\boldsymbol{\theta}$[8]. Since the cost function is a composite of many functions for each layers, the backprop algorithm basically implements the chain rule of calculus with a specific order of operations. The backprop algorithm for the simple network, show in algorithm 1, is outlined in algorithm 2. This example is not very realistic since no regularization is used and the cost function would be averaged over a set of input sampled from the whole sample (also called a minibatch). Therefore, it has to be noted, that this is an implementation, of both forward and backward propagation, for a simple and specific case. A detailed discussion and more general implementations can be found in reference [115] section 6.5.

The actual training of the parameters of the network, usually uses a form of gradient decent. In these algorithms, a function $f(x)$ is minimized by varying $x$ in in small iterative steps in opposite sign direction of $f'(x)$. Practically, this boils down to computing the gradient for each parameter $\boldsymbol{\theta}$ of the network with respect to the cost function and updating it according to $\boldsymbol{\theta} = \boldsymbol{\theta} - \lambda d\boldsymbol{\theta}$ with the learning rate $\lambda$. In terms of the network used to describe forward and backward propagation in algorithms 1 and 2 this means, that after each backprop step the weights $\boldsymbol{W}^{[i]}$ and biases $\boldsymbol{b}^{[i]}$ are updated to $\boldsymbol{W}^{[i]} = \boldsymbol{W}^{[i]} - \lambda \cdot d\boldsymbol{W}^{[i]}$ and $\boldsymbol{b}^{[i]} = \boldsymbol{b}^{[i]} - \lambda \cdot d\boldsymbol{b}^{[i]}$. The training is continued until a minimum in the cost function is found.

Each iteration, also called epoch, of the training consists out of the steps forward propagation → backward propagation → weight update. The training process is carried out over a predefined number of total iterations. To reduce overfitting it is common practice

[8] The previously used notation $J(\hat{y}, y)$ only implies the calculation that compares the network output to the true label. Since $\hat{y}$ depends on the parameters of the network $\boldsymbol{\theta}$ and the input $\boldsymbol{x}$, the cost function could be written as $J(\boldsymbol{x}, y; \boldsymbol{\theta})$. To improve readability in this section, the $\boldsymbol{x}$ and y parameters are omitted, since they are not relevant to backprop.

**Algorithm 2:** Backward propagation for the network of algorithm 1. The gradient of the layers $d\boldsymbol{h}^{[l]}$ flows backwards through the network from the output to the first hidden layer (the input layer does not play a role in backprop). From the gradients, which can be understood as an indicator how the output of each layer should change to reduce the overall error, the gradient on the parameters $d\boldsymbol{b}^{[i]}$ and $d\boldsymbol{W}^{[i]}$ can be derived.

Compute the gradient $d\boldsymbol{h}^{[l]}$ w.r.t. to the cost function $J$ in the output layer after forward propagation:

$d\boldsymbol{h}^{[l]} = \frac{dJ(\hat{\boldsymbol{y}},\boldsymbol{y})}{d\hat{\boldsymbol{y}}}$

**for** $i = l$ **to** *1* **do**

Compute the gradient w.r.t to the cost function to include the derivative $\beta^{[i]}$ of the activation function $\alpha^{[i]}$ for $\boldsymbol{Z}^{[i]}$ in the current layer

$d\boldsymbol{Z}^{[i]} = \frac{dJ(\hat{\boldsymbol{y}},\boldsymbol{y})}{d\boldsymbol{Z}^{[i]}} = d\boldsymbol{h}^{[i]} \odot \beta^{[i]}(\boldsymbol{Z}^{[i]})$, with an element wise multiplication $\odot$ if $\alpha^{[i]}$ is applied element-wise.

Compute the gradient on the weights and biases (in a realistic case, this would include regularization terms):

$d\boldsymbol{b}^{[i]} = \frac{dJ(\hat{\boldsymbol{y}},\boldsymbol{y})}{d\boldsymbol{b}^{[i]}} = d\boldsymbol{Z}^{[i]}$

$d\boldsymbol{W}^{[i]} = \frac{dJ(\hat{\boldsymbol{y}},\boldsymbol{y})}{d\boldsymbol{W}^{[i]}} = d\boldsymbol{Z}^{[i]} \cdot \boldsymbol{h}^{[i-1]\mathsf{T}}$

Propagate the gradients w.r.t the activation of the next lower-level layer:

$d\boldsymbol{h}^{[i-1]} = \frac{dJ(\hat{\boldsymbol{y}},\boldsymbol{y})}{d\boldsymbol{h}^{[i-1]}} = \boldsymbol{W}^{[i]\mathsf{T}} \cdot d\boldsymbol{Z}^{[i]}$

**end**

Save the derivatives of the biases $d\boldsymbol{b}^{[i]}$ and weights $d\boldsymbol{W}^{[i]}$ in order to use them in the gradient decent.

to additionally define a number of iterations after which the training is stopped if the cost functions does not decrease further (within some range). This is also called *early stopping*.

### Strategies to reduce overtraining and increase stability

Previously in this section, the topic of regularization to reduce over- and underfitting was covered in general. This section aims to introduce specific methods used in the classifier for the fully-hadronic $\mathrm{t\bar{t}H(b\bar{b})}$ analysis.

*Weight decay* is typically implemented by adding a term

$$\lambda\Omega(\boldsymbol{W}) = \lambda \sum_{k=1}^{l} \left( \sum_{i,j} \left( W_{i,j}^{[k]} \right)^2 \right) \quad , \tag{3.23}$$

with a weight decay coefficient $\lambda$ controlling the size of the penalty, to the cost function. The coefficient $\lambda$ is also referred to as the *L2 regularization* in many machine learning frameworks, for example in `TensorFlow` [119].

Another commonly used strategy is the *Dropout* [120] procedure. Hereby randomly selected nodes in a layer are temporarily removed in each epoch of the training with a rate set beforehand. This rate is also referred to as the *Dropout percentage*. The idea behind this is, that the weights of a single node represent certain features special to the dataset and neighboring node can grow reliant this behavior. While intuitively this might sound beneficial, this can lead to large generalization errors and unstable models if this is a feature solely present in the training data. By applying the dropout procedure during training, this can be effectively reduced and improve the overall performance of the model.

## 3.3. Statistical model for Higgs analyses at CMS

The ATLAS and CMS collaborations agreed upon common statistical procedures for measuring cross sections and setting limits in Higgs-boson analyses. An extensive documentation of the concepts presented in this section can be found in reference [121].
Most measurements are evaluated in terms of the signal strength $\mu$, which is defined as the ratio of the cross section measured in data and the SM expectation. Within the CMS collaboration, these methods are implemented in the `COMBINE` tool [122], which is based on `RooStats` [123], `RooFit` [124], and `ROOT` [125].
This section introduces the relevant concepts required for the measurements presented in part II. The statistical treatment of the analysis is based on the maximum likelihood method. In section 3.3.1 and 3.3.2 basic concepts are introduced. Section 3.3.3 describes the implementation of these concepts in the statistical model of the analysis and section 3.3.4 explains the limit setting procedure.

### 3.3.1. The maximum likelihood method and estimate

Assuming, the normalized probability density function $f(x|\boldsymbol{\theta})$, for parameters $\boldsymbol{\theta}$, describes the probability of observing $x$, the likelihood for a given measurement of the values $x_i$ with $i = 1, ..., n$ is defined as

$$L(\boldsymbol{x}|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(x_i|\boldsymbol{\theta}) \ , \tag{3.24}$$

as a function of the parameters $\boldsymbol{\theta}$. The values $x_i$ are independent and identically distributed random samples. It can be understood as the probability of observing $x_i$ for a given choice of parameters $\boldsymbol{\theta}$. It is a very useful tool for parameter estimation, performing hypothesis tests, or setting confidence intervals.

For a known function $f$ but unknown parameters $\boldsymbol{\theta}$, a maximum likelihood fit can be performed in order to estimate the parameters. The maximum likelihood estimates, also called best-fit values, $\hat{\boldsymbol{\theta}}$ of the parameters are calculated by maximizing the likelihood $L$ with respect to $\boldsymbol{\theta}$. Usually $L$ is not maximized but the negative logarithm of the likelihood (also called the negative log-likelihood) is minimized because of numerical considerations.

### 3.3.2. Likelihood ratio and Wilks' theorem

The likelihood ratio

$$\Lambda = \frac{L(\boldsymbol{x}|\boldsymbol{\theta_0})}{L(\boldsymbol{x}|\boldsymbol{\theta_1})} \ , \tag{3.25}$$

with parameters for the null-hypothesis $\boldsymbol{\theta_0}$ and an alternate hypothesis $\boldsymbol{\theta_1}$, is the most powerful test when comparing hypotheses according to the Neyman-Pearson lemma [126]. For the measurements in this thesis, which will be discussed in section 7.2, the hypotheses are parametrized by one or more parameters-of-interest (POI) $\boldsymbol{\mu}$ and additional parameters that are shared between the hypotheses. A fixed $\boldsymbol{\mu}$ configuration represents the null-hypothesis and the alternative hypotheses are represented by an arbitrary configuration, that is usually derived in a fit. The additional unknown parameters are called nuisance parameters and will, from now on, be denoted with $\boldsymbol{\theta}$ and describe, for example, experimental uncertainties.
The profile likelihood ratio

$$q_{\boldsymbol{\mu}} = -2\frac{L(\boldsymbol{x}|\boldsymbol{\mu}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}})}{L(\boldsymbol{x}|\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\theta}})} \ , \tag{3.26}$$

with the conditional maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\boldsymbol{\mu}}$ for a fixed $\boldsymbol{\mu}$ and the values $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\theta}}$ at the global maximum of the likelihood, is commonly used in the particle physics.

According to Wilks' theorem [127], the distribution of $q_{\boldsymbol{\mu}}$ can be approximated by a $N$-dimensional $\chi^2$ distribution in the limit of large sample sizes. This property can be used to calculate confidence intervals.

For a nuisance parameter with nominal parameter $\tilde{\theta}$ as the best estimate of $\theta$, which is usually obtained from an auxiliary measurement, a frequentist probability density function (pdf) $p(\tilde{\theta}, \theta)$ is constructed. According to Bayes' theorem, the pdf of the uncertainty $p(\theta\,|\,\tilde{\theta})$ can be expressed as a posterior probability distribution arising from a measurement of $\tilde{\theta}$:

$$p(\theta\,|\,\tilde{\theta}) \sim p(\tilde{\theta}\,|\,\theta) \cdot \pi_\theta(\theta) \ , \tag{3.27}$$

with the hyperprior $\pi_\theta(\theta)$. It expresses the probability that the true value of the parameter is $\theta$ given the best estimate $\tilde{\theta}$ from an auxiliary measurement. For the measurement presented in this thesis, the pdfs $p(\tilde{\theta}\,|\,\theta)$ are constructed such, that $\pi_\theta(\theta)$ can be kept flat. This step allows to represent all uncertainties in a frequentist context.
In the case of the analysis presented in part II of this thesis, the nuisance parameters can be either correlated 100 % (fully-correlated) or 0 % (uncorrelated). If nuisance parameters are partially correlated, and cannot be broken down into sub-components that are fully-correlated, they have to be declared to be either fully-correlated or uncorrelated, whatever is believed to be appropriate or conservative. This approach allows to include all nuisance parameters into the likelihood in a clean factorized form.

The specific choice of POIs and nuisance parameters for the main measurement presented in this thesis will be discussed in section 6.1.

### 3.3.3. Statistical model

The statistical model, which is used in the analysis discussed in part II of this thesis, builds upon the maximum likelihood method and other concepts presented in the previous section. The key detail, employed in the analysis but not considered previously, is that binned distributions are used.

The Poisson distribution is given by:

$$P(n|\lambda) = \frac{\lambda^n e^{-\lambda}}{n!} \ , \tag{3.28}$$

with the number of observed events $n$ and expected events $\lambda$ under a given selection (category or bin). For given $n$ and $\lambda$, $P(n|\lambda)$ can be interpreted as the propability to observe $n$ events if $\lambda$ events are expected. In the analysis, the distributions used in the fit are binned into $N_{\text{bins}}$. The probability to observe the distribution in data (or pseudo-data) is given by

$$P(\boldsymbol{n}|\boldsymbol{\lambda}) = \prod_{i=1}^{n_{\text{bins}}} P(n_i|\lambda_i) \ , \tag{3.29}$$

with the numbers of events observed per bin $\boldsymbol{n} = (n_1, .., n_{n_{\text{bins}}})$ and the corresponding expectation $\boldsymbol{\lambda} = (\lambda_1, .., \lambda_{n_{\text{bins}}})$. In general, the term observation or observed events implies that data collected at the LHC is used. The expectation is usually taken from simulated signal and background distributions and can be expressed as

$$\lambda_i(\mu, \boldsymbol{\theta}) = \mu \cdot s_i(\boldsymbol{\theta}) + b_i(\boldsymbol{\theta}) \ , \tag{3.30}$$

with the signal strength $\mu$ and the signal $s$ and background $b$ expectations per bin. The $s$ and $b$ distributions further depend on the one or more uncertainties parametrized as nuisance parameters $\boldsymbol{\theta}$. If $\mu > 1$ this is also called the *signal+background expectation* and *background-only expectation* if $\mu = 0$.

With these ingredients, the likelihood function can be constructed as follows:

$$L(\boldsymbol{n}|\mu,\boldsymbol{\theta}) = P(\boldsymbol{n}|\boldsymbol{\lambda}(\mu,\boldsymbol{\theta}))\cdot p(\tilde{\boldsymbol{\theta}},\boldsymbol{\theta}) \ . \tag{3.31}$$

The observation $\boldsymbol{n}$ refers to either the data that was measured with the CMS detector or pseudo-data. If the latter is used, it is usually constructed using the distribution of the background model and is commonly referred to as *Asimov-data*[9] [128]. The term $p(\tilde{\boldsymbol{\theta}},\boldsymbol{\theta})$, represents the product of the pdfs of all nuisance parameters, as defined by equation (3.27).

For the analysis presented in part II of this thesis, three types of pdfs are relevant:

**Uniform** For nuisance parameters that are unconstrained by prior measurements or considerations, a uniform pdf following the from $p(\tilde{\theta},\theta) = c$ is used. The nuisance parameters are also referred to as *freely-floating* in the fit.

**Gaussian** For nuisance parameters that can be both positive and negative a Gaussian pdf is a common choice:

$$p(\tilde{\theta},\theta) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\theta-\tilde{\theta})^2}{2\sigma^2}\right) \ . \tag{3.32}$$

This type of pdf is used for, so-called, *shape uncertainties*. As implied by the name, these uncertainties do not only change the yield of a given distribution but also the shape. Technically they are implemented by defining two histograms in addition to the nominal histogram, that correspond to one-standard-deviation shifts up and down of the uncertainty. By interpolation in each bin between the nominal and shifted histogram, a change for arbitrary shifts in the parameter is calculated. The mathematical formulation is described in reference [129].

**Log-normal** For nuisance parameters that are positively defined, the log-normal pdf is used:

$$p(\tilde{\theta},\theta) = \frac{1}{\sqrt{2\pi}\ln(\kappa)}\exp\left(-\frac{(\ln(\theta/\tilde{\theta}))^2}{2\ln(\kappa)^2}\right)\frac{1}{\theta} \ , \tag{3.33}$$

with the with $\kappa$. A value of $\kappa = 1.1$ implies, that the distribution can be larger or smaller by a factor of 1.1. Especially for large, positively defined uncertainties, this pdf is more appropriate than the Gaussian pdf, while they are asymptotically identical for small uncertainties. This type of pdf is used for, so-called, *rate uncertainties*. These uncertainties only change the overall rate of the distribution and can be defined as either symmetrical (e.g. the rate can change by 10 % up and down - $\kappa = 1.1$) or asymmetrical (e.g. the rate can change up by 8 % and down 6 % - $\kappa = 0.94/1.08$).

The definition and magnitude of the uncertainties used for the analysis in part II are given in sections 4.3 for uncertainties in the used simulation and chapter 6 for experimental uncertainties.

### 3.3.4. Upper limits

In addition to measuring the signal strength $\mu$, upper exclusion limits, with confidence level $\alpha$, may be set in certain measurements. The interpretation of these limits is that it can be excluded with a confidence level of $\alpha$, that an upward fluctuation of the background can explain the observed data (background-only hypothesis). The methods presented in

[9] Fits using this type of data are subsequently referred to as *Asimov-fits*.

the section, follow the conventions used in Higgs-boson searches at the LHC as described in reference [121].
A profile likelihood ratio, similar to equation (3.26), is used to quantify the incompatibility of the data $\boldsymbol{x}$ with a given signal strength $\mu$. The test statistic is defined as follows:

$$\tilde{q}_\mu = -2 \ln \frac{L(\boldsymbol{x}|\mu, \hat{\boldsymbol{\theta}}_\mu)}{L(\boldsymbol{x}|\hat{\mu}, \hat{\boldsymbol{\theta}})} \quad , \quad 0 \leq \hat{\mu} \leq \mu \ . \tag{3.34}$$

As previously, $\hat{\boldsymbol{\theta}}_\mu$ refers to the conditional maximum likelihood estimators given a fixed $\mu$ and the estimators $\hat{\mu}$ and $\hat{\boldsymbol{\theta}}$ refer to the global maximum of the likelihood. The constrains of $\hat{\mu}$ are motivated as follows: The bound at zero is required by physics, since negative signal strengths are not possible. The upper bound is set to ensure one-sided confidence intervals. This means that upward fluctuations in the data, i.e. $\hat{\mu} < \mu$, are not considered evidence against the signal.
In order to calculate the observed limit, the test statistic for a given signal strength $\mu$ under test needs to be calculated with respect to the observation (either data or pseudo-data). This yields $\tilde{q}_\mu^{\text{obs}}$. Furthermore the nuisance parameters $\hat{\boldsymbol{\theta}}_\mu^{\text{obs}}$ for the signal+background hypothesis and $\hat{\boldsymbol{\theta}}_0^{\text{obs}}$ for the background-only hypothesis, which are calculated by maximizing the respective likelihoods as given in equation (3.31), are required.
Using these parameters the p-values for the signal+background hypothesis

$$P(\tilde{q}_\mu \geq \tilde{q}_\mu^{\text{obs}} | \mu \cdot s + b) = \int_{\tilde{q}_\mu^{\text{obs}}}^{\infty} f(\tilde{q}_\mu | \mu, \hat{\boldsymbol{\theta}}_\mu^{\text{obs}}) d\tilde{q}_\mu \ , \tag{3.35}$$

and for the background-only hypothesis

$$P(\tilde{q}_\mu \geq \tilde{q}_\mu^{\text{obs}} | b) = \int_{\tilde{q}_\mu^{\text{obs}}}^{\infty} f(\tilde{q}_\mu | 0, \hat{\boldsymbol{\theta}}_0^{\text{obs}}) d\tilde{q}_\mu \ , \tag{3.36}$$

can be calculated. In general, the function $f(q|\eta)$ is the probability density for finding the value $q$ of the test statistic for the parameters $\eta$. In practice, the distribution $f$ can be estimated using MC methods by sampling pseudo-data from the statistical model using the given value for $\mu$ and setting the nuisance parameters to their best-fit value. As an alternative in case of a large number of observations, the distribution can be approximated using Wilks' theorem [127].
These two p-values are used to calculate the modified frequentist statistic

$$\text{CL}_s = \frac{P(\tilde{q}_\mu \geq \tilde{q}_\mu^{\text{obs}} | \mu \cdot s + b)}{P(\tilde{q}_\mu \geq \tilde{q}_\mu^{\text{obs}} | b)} \ , \tag{3.37}$$

which is used to calculate the confidence interval (CL) of the exclusion: A signal strength $\mu$ is excluded with a CL of $(1 - \alpha)$ if $\text{CL}_s \leq \alpha$. The upper limit is usually quoted as the 95 % CL upper limits. To calculate this, the $\mu$ is varied until $\text{CL}_s < 0.05$. The expected limit is commonly used to judge the sensitivity of an analysis and calculated generating MC pseudo data from the background-only hypothesis.

# Part II.

# Analyses strategy and Results

# 4. Analysis overview

The topics discussed in the previous chapters are the ingredients necessary for conducting a measurement of the $t\bar{t}H(b\bar{b})$ process with an emphasis in the fully-hadronic decay mode. In this thesis, the proton-proton collision dataset collected by the CMS experiment during 2016, 2017, and 2018 data-taking periods is considered. This dataset is referred to as the Run-2 dataset and amounts to $137.2\,\mathrm{fb}^{-1}$ of data collected by the CMS experiment that can be used for measurements.

This chapter serves as introduction to the overall analysis strategy for the $t\bar{t}H(b\bar{b})$ process and can be understood as a continuation of section 1.3.2. Section 4.1 summarizes previous results for this channel as well as the relevant backgrounds. For most of the background estimation, simulated datasets are used. What datasets are used and how they are generated will be described in section 4.2: The choice of event generators and settings for parton shower and parton distribution function (PDF) in the simulation are discussed in section 4.2.2. Especially, the treatment of the $t\bar{t}+\text{jets}$ background in simulation plays an integral part of the analysis and the details are discussed in section 4.2.3. The chapter closes with the uncertainties considered for the simulation in section 4.3.

The main topic of this thesis is the analysis of the fully-hadronic channel. Hence, it will be the main focus of this and the following chapters, but since the combined measurement of all channels is an integral part of the result, the leptonic channels will be covered as well, but in less detail.

## 4.1. The $t\bar{t}H(b\bar{b})$ channel

The $t\bar{t}H$ production mode is a very important probe of the Higgs sector. The main reason for this is that the Top-Yukawa coupling can be directly measured, while it is only possible to measure this coupling indirectly in other channels. The $H \to b\bar{b}$ decay mode in particular provides interesting features like the large branching fraction of 58 % or the circumstance that only third-generation-quark couplings are present in the final state.

Measuring the production of top-quark pairs in association with a Higgs boson decaying to a bottom-quark pair is experimentally very challenging. As discussed in the sections 1.3.2 and 1.3.3, the top quarks each decay to a bottom quark and a W boson, which in turn decays to either a quark pair or a lepton-neutrino pair, leading the a final state with four bottom quarks as well as jets and leptons. Three distinct signatures can be determined based on the decay of the W bosons:

- The fully-hadronic channel, where both W bosons decay hadronically, with a branching fraction of 45.4 %;
- the semileptonic channel, where one W boson decays to a lepton-neutrino pair and the other hadronically, with a branching fraction of 43.9 %;
- and the dileptonic channel, where both W bosons decay to lepton-neutrino pairs, with a branching fraction of 10.6 %.

The latter two are often referred to as leptonic channels since they can be analyzed with similar strategies, while the fully-hadronic channels requires a distinct approach due to the large QCD multijet background. Given the different signatures, the channels exhibit different challenges and sensitivities. For example, a given systematic uncertainty may not affect a certain channels or due to the different background composition its effect may be different. Combining all channels into a single statistical analysis allows to exploit these differences and get a more sensitive result.

The $t\bar{t}H(b\bar{b})$ analyses follow an approach that can be outline as follows. Triggers that capture the analysis phase space are used with the goal of saving as much of the signal events as possible during data-taking. The fully-hadronic channel relies on dedicated triggers, which will be discussed in detail in section 5.1, while the leptonic channels utilize standard single- and double-lepton triggers as well as a dedicated lepton+$H_{\mathrm{T}}$ trigger. All events that are selected by the channel-specific trigger collection, are required to pass certain selection criteria that aim to create multiple signal enriched regions in the dataset by removing as much of the backgrounds as possible. The criteria include among others requirements on the number of reconstructed leptons (0, 1, or 2 for the fully-hadronic, semileptonic or dileptonic channel, respectively), as well as requirements on the number of reconstructed jets and b-tags. The signal and background processes are modeled by simulation except for the multijet background in the fully-hadronic channel, which is estimated from data. Section 4.1.2 will describe the relevant backgrounds for the $t\bar{t}H(b\bar{b})$ analyses. Furthermore, the modeling of the signal and background processes with simulation will be covered in section 4.2. Uncertainties on these simulations, which are covered in section 4.3, are considered in the signal extraction. For the signal extraction, the signal and background models are used in a fit to data using a maximum-likelihood-based method, described in section 3.3, under the statistical model described in section 6. A distribution, that separates signal and background processes, is required in order to increase the sensitivity of this fit. In case of the $t\bar{t}H(b\bar{b})$ analyses, neutral-network-based classifiers are used to generates such distributions in all channels.

### 4.1.1. Previous Results in this channel

Due to the significance of $t\bar{t}H$ process, searches for this process were part of the experimental program at LHC from the very beginning of its operation. The first results of the $t\bar{t}H(b\bar{b})$ channel have been published using the Run-1 dataset collected at a center-of-mass energy of 8 TeV. CMS published an analysis using the leptonic final states [130], and ATLAS published an analysis using the leptonic [131] and fully-hadronic final states [9]. The dataset used for the CMS result [130], which utilized a machine-learning approach, was later reanalyzed with a MEM discriminant [107]. The result using the MEM discriminant [107] improved the exclusion limit by roughly 20 % with respect to the machine-learning approach [130]. In the following years, these analysis were continuously improved and refined using data collected at 13 TeV proton-proton collisions [11, 132, 133]. At CMS, the analysis in the fully-hadronic decay channel was pioneered using the dataset recorded in 2016 at $\sqrt{s}$ =13 TeV [10].

An overview of the best-fit signal strength parameters $\hat{\mu} = \hat{\sigma}/\sigma_{SM}$ for all $t\bar{t}H(b\bar{b})$ results published previous to the analysis presented in this thesis is given in table 4.1 and the available signal significance are shown in table 4.2.

Table 4.1.: Measured $t\bar{t}H(b\bar{b})$ signal strength parameters $\hat{\mu}$ ($\sigma/\sigma_{SM}$) at the LHC. If available, the uncertainty on $\hat{\mu}$ is further broken down into systematic and statistical contribution. The table is sorted by the integrated luminosity and publication date. In addition to the experiment, the considered $t\bar{t}H$ decay channels are indicated: Fully hadronic (FH), semileptonic (SL) and dileptonic (DL). If the integrated luminosity is given as sum, this indicates a statistical combination of multiple separate results.

| Int. Luminosity | $\sqrt{s}$ | Channel | Collaboration | $\hat{\mu}$ | Uncertainty | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | tot | syst | stat | |
| $19.5\,\text{fb}^{-1}$ | 8 TeV | SL, DL | CMS | 1.2 | $^{+1.6}_{-1.5}$ | - | - | [107] |
| $20.3\,\text{fb}^{-1}$ | 8 TeV | SL. DL | ATLAS | 1.5 | $^{+1.1}_{-1.1}$ | - | $^{+0.7}_{-0.7}$ | [131] |
| $20.3\,\text{fb}^{-1}$ | 8 TeV | FH | ATLAS | 1.6 | $^{+2.6}_{-2.6}$ | - | - | [9] |
| $35.9\,\text{fb}^{-1}$ | 13 TeV | FH | CMS | 0.9 | $^{+1.5}_{-1.5}$ | $^{+1.3}_{-1.3}$ | $^{+0.7}_{-0.7}$ | [10] |
| $35.9\,\text{fb}^{-1}$ | 13 TeV | SL,DL | CMS | 0.72 | $^{+0.45}_{-0.45}$ | $^{+0.38}_{-0.38}$ | $^{+0.24}_{-0.24}$ | [132] |
| $36.1\,\text{fb}^{-1}$ | 13 TeV | SL, DL | ATLAS | 0.84 | $^{+0.64}_{-0.61}$ | $^{+0.57}_{-0.54}$ | $^{+0.29}_{-0.29}$ | [133] |
| $41.5\,\text{fb}^{-1}$ | 13 TeV | SL,DL,FH | CMS | 1.49 | $^{+0.44}_{-0.40}$ | $^{+0.39}_{-0.35}$ | $^{+0.21}_{-0.20}$ | [11] |
| $35.9{+}41.5\,\text{fb}^{-1}$ | 13 TeV | SL,DL,FH | CMS | 1.15 | $^{+0.32}_{-0.29}$ | $^{+0.28}_{-0.25}$ | $^{+0.15}_{-0.15}$ | [11] |
| $139\,\text{fb}^{-1}$ | 13 TeV | SL, DL | ATLAS | 0.43 | $^{+0.36}_{-0.33}$ | $^{+0.30}_{-0.27}$ | $^{+0.20}_{-0.19}$ | [134] |

Table 4.2.: Observed and expected $t\bar{t}H(b\bar{b})$ significances at the LHC. The table is sorted by the integrated luminosity and publication date. In addition to the experiment, the considered $t\bar{t}H$ decay channels are indicated: Fully hadronic (FH), semileptonic (SL) and dileptonic (DL). If the integrated luminosity is given as sum, this indicates a statistical combination of multiple separate results. Publication listed in table 4.1 but not in this table did not provide signal significances.

| Int. Luminosity | $\sqrt{s}$ | Channel | Collaboration | Significance | | |
|---|---|---|---|---|---|---|
| | | | | Obs. | Exp. | |
| $35.9\,\text{fb}^{-1}$ | 13 TeV | SL,DL | CMS | 1.6 | 2.2 | [132] |
| $36.1\,\text{fb}^{-1}$ | 13 TeV | SL, DL | ATLAS | 1.4 | 1.6 | [133] |
| $41.5\,\text{fb}^{-1}$ | 13 TeV | SL,DL,FH | CMS | 3.7 | 2.6 | [11] |
| $35.9{+}41.5\,\text{fb}^{-1}$ | 13 TeV | SL,DL,FH | CMS | 3.9 | 3.5 | [11] |
| $139\,\text{fb}^{-1}$ | 13 TeV | SL, DL | ATLAS | 1.3 | 3.0 | [134] |

The most resent results, published previous to the results presented in this thesis, are $\hat{\mu} = 0.43^{+0.30}_{-0.27}(\text{syst})\ ^{+0.20}_{-0.19}\ (\text{stat})$ using $139\,\text{fb}^{-1}$ of data collected by the ATLAS experiment at 13 TeV and $\hat{\mu} = 1.15^{+0.32}_{-0.29}(\text{syst})\ ^{+0.15}_{-0.15}\ (\text{stat})$ using the combination of 35.9 and $41.5\,\text{fb}^{-1}$ of data collected by the CMS experiment at 13 TeV[1]. The main sources of uncertainty are the treatment of the $\text{t}\bar{\text{t}} + \text{jets}$ background, especially $\text{t}\bar{\text{t}} + \text{b}\bar{\text{b}}$, and the b-Jet identification. Considering all available data, the statistical uncertainty plays only a minor role. Methods to minimize the effect of systematic uncertainties need to be developed in order to improve the sensitivity of the $\text{t}\bar{\text{t}}\text{H}(\text{b}\bar{\text{b}})$ measurements taking advantage of the full Run-2 dataset collected by the CMS experiment.

The Run-2 CMS results, presented in this thesis, takes this into account by improving, among the most important aspects the $\text{t}\bar{\text{t}} + \text{b}\bar{\text{b}}$ description (see Section 4.2.3), the strategy to separate signal from background (see Section 5.6), the QCD multijet estimation (see Section 5.5), and the b-tagging with respect to the analysis presented in reference [11].

### 4.1.2. Background processes

Many other physics processes at a hadron collider can produce signal-like final states with a sufficiently high multiplicity to overshadow $\text{t}\bar{\text{t}}\text{H}(\text{b}\bar{\text{b}})$ with its small cross-section. Section 1.3.3 introduced all relevant background processes and this section aims to expand upon this topic. This section will focus on the backgrounds from the point of view of the fully-hadronic decay mode but the general concepts also apply to the channels involving leptonic W-boson decays.

Generally, the backgrounds can be divided into two categories: The irreducible backgrounds, that exhibit the same final state and the signal, and the reducible backgrounds, that only appear to have the same final state as the signal because they are mismeasured.

The production of multiple quarks or gluons via processes only involving the strong interaction, so-called QCD multijet production, provides the majority of events collected at a hadron collider. By requiring leptons in the final state, this process can be reduced to a negligible contribution. Therefore, this background does not play a significant role in leptonic $\text{t}\bar{\text{t}}\text{H}(\text{b}\bar{\text{b}})$ channels. In the fully-hadronic channel this is not the case. The QCD multijet background is, by far, the dominant background in this channel. Even though it is overall very unlikely that QCD events very similar to the $\text{t}\bar{\text{t}}\text{H}$ signal are produced, it contributes a large background that is difficult to suppress, due to the high production rate of QCD events. Additionally, the simulation of QCD multijet events in the relevant phase space is computationally expensive and not very precise. Therefore, this background is usually estimated from data. This estimation is a major topic of the analysis in the fully-hadronic channel and will be covered in detail in section 5.5.

The $\text{t}\bar{\text{t}} + \text{b}\bar{\text{b}}$ production exhibits a final state with the exact same particles as the signal and is therefore considered an irreducible background. Because it cannot be distinguished from the signal, interference between them would, in principle, need to be considered. It has been shown, cf. reference [135], that these effects are small and both processes can be viewed as independent until measurements reach higher precision. The main avenue for reaching some discrimination power between the signal and $\text{t}\bar{\text{t}} + \text{b}\bar{\text{b}}$, is to compare the properties of the additional bottom-quark pairs[2] in both processes. Top-quark pair

[1] The 2016 ($35.9\,\text{fb}^{-1}$) [10, 132] and 2017 ($41.5\,\text{fb}^{-1}$) [11] datasets have been analyzed separately and combined as statistically independent results. The analysis strategies, object definitions and uncertainty definitions can differ in details. This is the main difference to the Run-2 analysis presented in the thesis which reanalyzes the 2016 ($36.3\,\text{fb}^{-1}$), 2017 ($41.5\,\text{fb}^{-1}$) and 2018 ($59.7\,\text{fb}^{-1}$) datasets with a consistent strategy in order to take full advantage of the $137.2\,\text{fb}^{-1}$ collected in Run-2 by CMS.

[2] In this context, "additional bottom-quark pairs" refers to the bottom quarks not originating from the top quarks.

production in association with charm-quark pairs ($t\bar{t}+c\bar{c}$) and light-quark pairs ($t\bar{t}+LF$) can be distinguished more easily from the signal with help of b-tagging algorithms and are therefore considered reducible backgrounds. Due to the non-negligible mis-tagging rate of the b-tagging, these backgrounds cannot be completely suppressed. For $t\bar{t}+c\bar{c}$ the mis-tagging rate is comparatively large due to the relatively long lifetime and high mass of the c hadrons, similar to b hadrons. The mis-tagging rate for light quarks is very small, but because of the significant cross-section of $t\bar{t}+LF$, it still has a large contribution in the signal region.
An additional issue with the $t\bar{t}+b\bar{b}$ background is that it is only known with large uncertainties. Therefore, a precise modeling of this background is a key ingredient of the $t\bar{t}H(b\bar{b})$ measurement. The $t\bar{t}+b\bar{b}$ background will be covered in section 4.2.3. In principle, the same issue is present for the $t\bar{t}+c\bar{c}$ process, but since it can be distinguished more easily, it only plays a minor role in the analysis.
The various $t\bar{t}+$jets processes are the dominant background contribution in the leptonic $t\bar{t}H(b\bar{b})$ final states. They play a lesser role in the fully-hadronic channel because they are overshadowed by QCD multijet production. Nevertheless, the modeling of $t\bar{t}+$jets is important for the fully-hadronic channel as well, because $t\bar{t}+$jets still provides significant background in the signal region and is used for the QCD multijet estimation.

All other backgrounds (Single-top, $t\bar{t}V$, V + jets, and VV), as listed in section 1.3.3, are considered minor backgrounds and do not play a significant role in the measurement until the overall precision is greatly improved.

## 4.2. Modeling signal and background processes

Monte Carlo (MC) simulated samples are used to model the signal and some of background processes, optimize the event selection, train a classifier, and evaluate systematic uncertainties and acceptance effects. The simulation of collision events is a complex task that is computationally expensive and therefore the tuning, configuration, and production for the most common processes is done in central campaigns run by the CMS collaboration. If not otherwise indicated all simulated datasets used in this analysis are part of such campaigns. For each of the three periods (2016, 2017 and 2018) separate samples are produced in order to respect the differences in detector and machine configuration. MC generator settings, like the parton-shower tune, are set to the same value in all periods if possible.
Most datasets are shared between the channels except for some minor backgrounds ($t\bar{t}V$, V + jets) for which only the relevant final states for each channel are considered. All simulated datasets that are used in the fully-hadronic channel are listed in table 4.3.

### 4.2.1. Cross-sections for normalization of simulated datasets

In order to estimate the yield expected for a given process at a given integrated luminosity $\mathcal{L}$ as precisely as possible, the MC samples are reweighted to the best currently available cross-section calculation. The weights $w$ are calculated using the formula

$$w = \frac{\sigma \cdot \mathcal{L}}{N_{gen}} \ , \tag{4.1}$$

with the cross-section $\sigma$, the number of weighted generated events $N_{gen}$[3] and the integrated luminosity $\mathcal{L}$. If a given sample only simulates a subset of the decay modes, the branching fraction of these modes is taken into account.

[3] When calculating this number, generator weights, e.g. negative weights in certain NLO Monte Carlo generators, and pileup weights, which will be discussed in section 5.2.1, are considered.

Table 4.3.: Simulated samples for the signal and background processes used in the $t\bar{t}H(b\bar{b})$ fully-hadronic analysis. Statistically independent samples are stimulated for the 2016, 2017 and 2018 data-taking periods. References for the quoted cross sections are given in section 4.2. The exact generator setting for samples listed with the same generator can be different. Details on the generator setting for each sample can be found in section 4.2.2.

| Process | Generator | BR · | $\sigma$ [pb] | Generated Events / $10^6$ | | |
|---|---|---|---|---|---|---|
| | | | | 2016 | 2017 | 2018 |
| $t\bar{t}H, H \to b\bar{b}$ | POWHEG | 0.5824 | 0.5071 | 9.9 | 8.0 | 9.5 |
| $t\bar{t}H$, H $\to$ Other | POWHEG | 0.4176 | 0.5071 | 9.9 | 7.9 | 7.5 |
| tHq | MadGraph | | 0.07425 | 9.6 | 9.9 | 29.6 |
| tHW | MadGraph | | 0.01517 | 5.0 | 4.7 | 14.9 |
| $t\bar{t}$ + jets (FH) | POWHEG | 0.4544 | 831.76 | 68.5 | 130 | 133.6 |
| $t\bar{t}$ + jets (SL) | POWHEG | 0.4393 | 831.76 | 67.8 | 69 | 64.3 |
| $t\bar{t}$ + jets (DL) | POWHEG | 0.1062 | 831.76 | 107.3 | 110 | 101.5 |
| $t\bar{t} + b\bar{b}$ (FH) | POWHEG | 0.4544 | 831.76 | 6.1 | 6.7 | 6.7 |
| $t\bar{t} + b\bar{b}$ (SL) | POWHEG | 0.4393 | 831.76 | 3.6 | 3.6 | 3.5 |
| $t\bar{t} + b\bar{b}$ (DL) | POWHEG | 0.1062 | 831.76 | 6.2 | 5.6 | 6.1 |
| t (s-channel), hadronic decays | MadGraph | 0.6741 | 10.32 | 4.9 | 9.6 | 9.7 |
| t (s-channel), leptonic decays | MadGraph | 0.3259 | 10.32 | 9.8 | 9.9 | 19.9 |
| t (t-channel) | POWHEG | | 136.02 | 31.8 | 122.6 | 154.3 |
| $\bar{t}$ (t-channel) | POWHEG | | 80.95 | 17.8 | 63.6 | 79.1 |
| t tW | POWHEG | | 35.85 | 4.9 | 7.7 | 7.6 |
| $\bar{t}$ tW | POWHEG | | 35.85 | 4.9 | 7.9 | 9.6 |
| $t\bar{t}$W, W $\to q\bar{q}$ | MadGraph | | 0.4316 | 0.7 | 0.7 | 0.7 |
| $t\bar{t}$Z, Z $\to q\bar{q}$ | MadGraph | | 0.5104 | 0.8 | 0.8 | 0.8 |
| WW | Pythia | | 118.7 | 7.9 | 7.8 | 7.9 |
| WZ | Pythia | | 65.5443 | 4.0 | 3.9 | 3.8 |
| ZZ | Pythia | | 15.8274 | 2.0 | 1.9 | 2.0 |
| Z + Jets, Z $\to q\bar{q}$, HT $\geq$ 600 GeV | MadGraph | | 5.67 | 1.0 | - | - |
| Z + Jets, Z $\to q\bar{q}$, 400 $\leq$ HT $<$ 600 GeV | MadGraph | | 144.7 | - | 10.3 | 16.7 |
| Z + Jets, Z $\to q\bar{q}$, 600 $\leq$ HT $<$ 800 GeV | MadGraph | | 34.06 | - | 8.9 | 14.6 |
| Z + Jets, Z $\to q\bar{q}$, HT $\geq$ 800 GeV | MadGraph | | 18.57 | - | 7.8 | 10.6 |
| W + Jets, W $\to q\bar{q}$, HT $\geq$ 180 GeV | MadGraph | | 3105.0 | 22.6 | - | - |
| W + Jets, W $\to q\bar{q}$, 400 $\leq$ HT $<$ 600 GeV | MadGraph | | 314.9 | - | 9.8 | 10.1 |
| W + Jets, W $\to q\bar{q}$, 600 $\leq$HT $<$ 800 GeV | MadGraph | | 68.64 | - | 8.7 | 15.3 |
| W + Jets, W $\to q\bar{q}$, HT $\geq$ 800 GeV | MadGraph | | 34.7 | - | 8.1 | 14.6 |
| QCD, 300 $\leq$ HT $<$ 500 GeV | MadGraph | | 347700 | 17.0 | - | - |
| | | | 323600 | - | 60.3 | 54.7 |
| QCD, 500 $\leq$ HT $<$ 700 GeV | MadGraph | | 32160 | 18.5 | - | - |
| | | | 29990 | - | 56.2 | 55.2 |
| QCD, 700 $\leq$ HT $<$ 1000 GeV | MadGraph | | 6828 | 15.6 | - | - |
| | | | 6351 | - | 47.7 | 48.2 |
| QCD, 1000 $\leq$ HT $<$ 1500 GeV | MadGraph | | 1208 | 4.8 | - | - |
| | | | 1094 | - | 16.6 | 15.5 |
| QCD, 1500 $\leq$ HT $<$ 2000 GeV | MadGraph | | 120 | 4.0 | - | - |
| | | | 99.01 | - | 11.6 | 10.9 |
| QCD, HT $\geq$ 2000 GeV | MadGraph | | 25.27 | 2.0 | - | - |
| | | | 20.23 | - | 5.9 | 5.5 |

The cross-sections and, if applicable, branching fractions used for each sample are listed in table 4.3.
For the $t\bar{t}H$ signal process, the inclusive cross-section is taken from calculations at NLO QCD and NLO EWK accuracy [55]. For samples exclusive in Higgs-decay modes, the Higgs-boson branching-ratios [55] are used to calculate the correct cross-section for a given sample. For the single-top Higgs samples, the cross-sections are available at NLO accuracy in QCD (NLO EWK corrections are not available), computed in the five-flavor scheme [55, 136, 137].
The $t\bar{t}$ + jets cross-section is given in five-flavor scheme, calculated at next-to-next-to leading order (NNLO) in QCD including resummation of next-to-next-to-leading logarithmic (NNLL) soft gluon terms with `top++2.0` [138–144]. The branching fractions for samples exclusive in the W-boson decays modes are taken from [29].
The inclusive Single-top (Single-top) cross-sections are normalized to NNLO (tW channel [145]) and NLO (t- and s-channels [146, 147]).
The cross-sections for the minor backgrounds $t\bar{t}V$ (NLO+NNLL QCD and NLO EWK [148]), WW (NNLO [149]) and ZZ/WZ (NLO [150]) are also taken from cross-section calculations while W + Jets, $W \to q\bar{q}$, Z + Jets, $Z \to q\bar{q}$, and QCD are normalized to the cross-sections predicted by MadGraph.

### 4.2.2. Event generation

The simulation of the hard process, parton shower, hadronization. and detectors interaction are required when simulating particle collision events. The general approach has been described in sections 1.2.3 and 1.2.4. All samples are using a common Geant4 [52] based CMS detector simulation and pythia 8 [51] for the modeling of the parton shower and hadronization. pythia 8 uses either the tune `CUETP8M1` [151] or `CP5` [152], which describe the parameters for the underlying event description. For all samples generated for the 2017 and 2018 data-taking period, and the most important samples for the 2016 period, the more recent tune `CP5` is used. The tune `CUETP8M1` is only used in minor backgrounds simulated for the 2016 period, because a reproduction was not possible. Since the contribution from these minor backgrounds to the shape of the final distributions is negligible compared to the dominant $t\bar{t}$ + jets and QCD multijet backgrounds [4] , the effect of the difference in tune is estimated to be negligible. The `hdamp` parameter (ME-PS matching scale, powheg-only) is set to 237.9 GeV, if the tune `CP5` is used, and 172.5 GeV, if the tune `CUETP8M1` is used. The parton structure is simulated with the NNPDF3.1 [50] parton distribution function.

The generator for the hard process varies from sample to sample. The following configurations are used:

- The $t\bar{t}H$ signal samples, split into $b\bar{b}$ and non-$b\bar{b}$ Higgs-boson decay-modes, are generated with powheg [153–156] at NLO accuracy. The tH samples are generated with MadGraph5_aMC@NLO [157] at LO in the four-flavor scheme (tHq) and five-flavor scheme (tHW). The four-flavor scheme is used in tHq because it results in a better description of the additional quark from the initial-state gluon splitting. In the tHW samples, the five-flavor scheme is used for technical reason related to the coupling measurement presented in section 7.2.2. An additional inclusive $t\bar{t}H$ sample, which is exclusively used for the additional interpretations presented in section 7.2.2, is generated with MadGraph5_aMC@NLO in the four-flavor scheme. The inclusive

[4] This statement is only true if the QCD multijet background is estimated from data, which is the case for the signal extraction. The effect on the QCD simulation is not negligible (actually it is quite significant), which is one of the reasons a data-driven approach was chosen, in the phase space of the fully-hadronic $t\bar{t}H(b\bar{b})$ analysis. This topic will be covered in more detail in sections 5.4.1 and 5.5.

$t\bar{t}H$ samples and the tH samples are generated such that they include various coupling combinations following the framework specified in section 1.3.4.1. The Higgs-boson mass is set to 125 GeV and the Top-quark mass is set to 172.5 GeV. If not otherwise indicated, references to a $t\bar{t}H$ sample refer to the sample simulated with POWHEG.

- The $t\bar{t}$ + jets samples are generated with the same generator as the $t\bar{t}H$ sample, using the five-flavor scheme. The $t\bar{t} + b\bar{b}$ samples are generated with a dedicated POWHEG program [158], where the $t\bar{t} + b\bar{b}$ process is simulated in the four-flavor scheme at NLO. Further detail on these differences are given in section 4.2.3.
- The minor background samples are generated with POWHEG (Single-top t- and tW-channel), MADGRAPH5_aMC@NLO (Single-top s-channel, V + jets, $t\bar{t}V$) and PYTHIA (Diboson).
- The QCD multijet simulation is generated with MADGRAPH5_aMC@NLO. Unlike all other previously list samples, the QCD multijet simulation will not be used in the signal extraction but is only used for preliminary checks and ensure consistency with the data-driven multijet estimation.

### 4.2.3. $t\bar{t}$ + jets background model

In this analysis, $t\bar{t}$ + jets process is split into the five distinct subprocesses $t\bar{t}$ + LF, $t\bar{t} + c\bar{c}$, $t\bar{t}$ + b, $t\bar{t}$ + 2b and $t\bar{t} + b\bar{b}$ to account for different systematic uncertainties. This division is based on the flavor of the additional jets within acceptance, that do not originate from the top-quark decays. This method is referred to as heavy-flavor-hadron matching: Particle-level jets, as introduced in section 2.3.5, with $p_T > 20$ GeV and $|\eta| < 2.4$ and generator-level hadrons are matched. Depending on the number and flavor of these jet-hadron pairs, an event is categorized into one of the five processes. If at least two additional jets are generated and each of them can be matched to one or more B hadrons, the event is considered a $t\bar{t} + b\bar{b}$ event. If only one jets has a matched B hadron, it is considered a $t\bar{t}$ + b event. If two additional B hadrons are close enough in direction to produce one jets, the event is categorized as $t\bar{t}$ + 2b. These three cases are illustrated in figure 4.1a. This distinction is only relevant on a technical level in the analysis itself but is unnecessary on a theoretical level, since all three are described by the $t\bar{t} + b\bar{b}$ ME. If the event contains at least one jet matched to a C hadron and no additional B hadrons, it is categorized as $t\bar{t} + c\bar{c}$. All other events are considered as $t\bar{t}$ + light flavor ($t\bar{t}$ + LF) events and are categorized accordingly. Various combinations of these categories are used depending on the context:

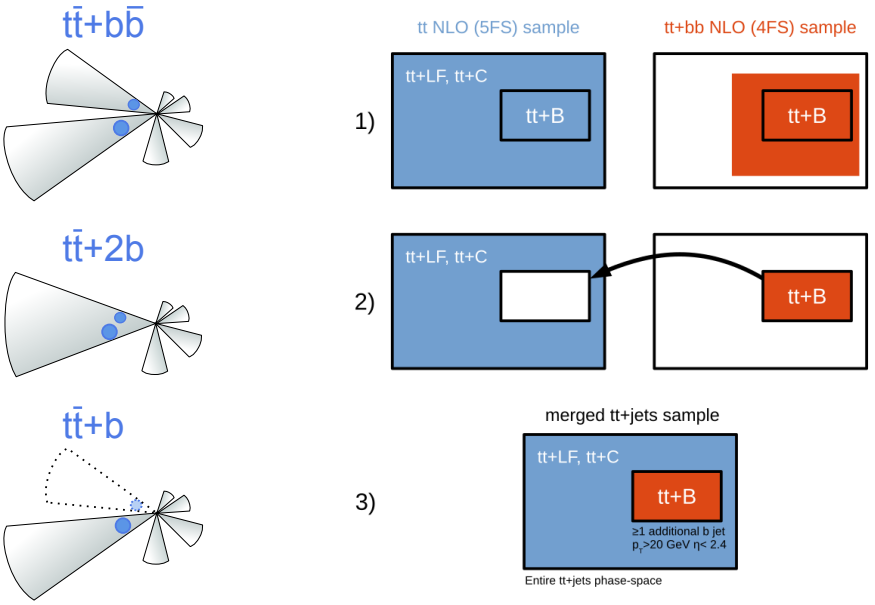
**$t\bar{t}$ + B** combines the $t\bar{t} + b\bar{b}$, $t\bar{t}$ + b as well as $t\bar{t}$ + 2b processes. This notation will be used if no distinction is necessary, as will be described below. Furthermore, it is used to separate between $t\bar{t} + b\bar{b}$ on a technically and theoretically motivated level.

**$t\bar{t}$ + HF** combines the $t\bar{t} + b\bar{b}$, $t\bar{t}$ + b, $t\bar{t}$ + 2b as well as $t\bar{t} + c\bar{c}$ processes. This notation is mostly used in connection to b-tagging because similar effects apply to jets originating from bottom and charm quarks in that context.

**$t\bar{t}$ + C** is used to separate between $t\bar{t} + c\bar{c}$ on a technically and theoretically motivated level but is in general interchangeable.

The $t\bar{t}$ + jets modeling, especially for the $t\bar{t}$ + B process, is a very important topic in all $t\bar{t}H(b\bar{b})$ analyses. In the analysis presented in reference [11], the impact due to the $t\bar{t}$ + HF modeling, which is mainly driven by the $t\bar{t}$ + B description, on the measured signal strength is $\pm 0.15$, with a total uncertainty of $\pm 0.32$[5]. Therefore, improving this

[5] The value is obtained by fixing the uncertainty to the post-fit values in the fit and subtracting the obtained result in quadrature from the result of the full fit. Details can be found Table 6 in reference [11]




(a) (b)

Figure 4.1.: Illustration of methods used for the $t\bar{t}$ + HF modeling. Figure (a) illustrates the heavy-flavor-hadron matching for $t\bar{t}$ + $b\bar{b}$, $t\bar{t}$ + 2b and $t\bar{t}$ + b. The blue circles represent heavy-flavor hadrons and the grey cones represent jets. In an event categorized as $t\bar{t}$ + $b\bar{b}$, each hadron can be matched to a jet. In $t\bar{t}$ + 2b events both hadrons are matched to a single jet and in $t\bar{t}$ + b events one jets is out of acceptance (dashed cone) and only one hadron could be matched to a jets. Figure adapted from reference [159]. Figure (b) illustrates the procedure used to merge the $t\bar{t}$ + jets sample and $t\bar{t}$ + $b\bar{b}$ sample. Step 1 shows the initial state after categorizing both samples with the heavy-flavor-hadron matching. In Step 2, $t\bar{t}$ + B is removed from the $t\bar{t}$ + jets sample and the $t\bar{t}$ + B from the $t\bar{t}$ + $b\bar{b}$ sample is scaled to the expectation in the $t\bar{t}$ + jets samples. Step 3 represent the final state of the merging process.

modeling is crucial to further improve the measurement. Different modeling approaches are available for the $t\bar{t}$ + B process. The main difference between all approaches is how the bottom-quark is treated and therefore how the bottom-quark pair is generated. The inclusive $t\bar{t}$ + jets sample is generated at NLO in the five-flavor scheme with massless bottom quarks at matrix-element level. In this approach, the additional bottom-quark pair for the $t\bar{t}$ + $b\bar{b}$ process is mostly generated by gluon splitting in the parton shower. This leads to a precision that is lower than LO but the parton shower can be accurately tuned to data. The residual uncertainties due to this behavior are large and hard to quantify.

For an improved description, the $t\bar{t}$ + $b\bar{b}$ processes can be generated at NLO in the four-flavor scheme with massive bottom quarks at matrix-element level. In this case the full phase space can be simulated in the matrix element (ME). The additional bottom quarks are therefore subject to more fundamental ME scale uncertainties, which are much smaller than the uncertainties resulting from the five-flavor scheme approach. The main practical improvements from such a samples are the improved description of event kinematics and the well defined and smaller modeling uncertainties.

In this analysis, samples simulated with both approaches are used. The $t\bar{t}$ + LF and $t\bar{t}$ + C processes are taken from the inclusive $t\bar{t}$ + jets samples while the $t\bar{t}$ + B processes are taken from $t\bar{t}$ + $b\bar{b}$ samples. For the samples to be merged successfully, no overlap in the phase space can occur. This is done via the previously described categorization

Table 4.4.: Comparison of relevant generator parameters for the $t\bar{t}$ +Jet and $t\bar{t} + b\bar{b}$ simulations. The differences reflect the different relevant scales.

| | $t\bar{t}$ + jets (five-flavor scheme) | $t\bar{t} + b\bar{b}$ (four-flavor scheme) |
|---|---|---|
| $m_t$ | 172.5 GeV | 172.5 GeV |
| $m_b$ | 0 | 4.75 GeV |
| PDF set | `NNPDF31_nnlo_hessian_pdfas` | `NNPDF31_nnlo_as_0118_nf_4` |
| $\mu_R$ | $\min(m_T(t), m_T(\bar{t}))$ | $\frac{1}{2}\sqrt[4]{m_T(b)m_T(\bar{b})m_T(t)m_T(\bar{t})}$ |
| $\mu_F$ | $\min(m_T(t), m_T(\bar{t}))$ | $\frac{1}{4}\left[m_T(b) + m_T(\bar{b}) + m_T(t) + m_T(\bar{t}) + m_T(g)\right]$ |
| `hdamp` | $1.379 m_t$ | $1.379 m_t$ |

using the heavy-flavor-hadron matching. While the $t\bar{t} + b\bar{b}$ sample is expected to improve the kinematic description of the $t\bar{t}$ + B process, this is not the case for the cross-section prediction. Therefore, the $t\bar{t}$ + B events are scaled, in addition to normalizing to the cross-section listed in table 4.3, to the yield expected by the $t\bar{t} + b\bar{b}$ events in the $t\bar{t}$ + jets sample, which is known to model the cross-section very well. The merging procedure is illustrated in figure 4.1b.
When using the $t\bar{t} + b\bar{b}$ samples, $t\bar{t} + b\bar{b}$, $t\bar{t} + 2b$ and $t\bar{t} + b$ can be considered as a single process for almost all practical purposes[6], since all three signatures are fundamentally based on the same $t\bar{t} + b\bar{b}$ ME. This is not the case when using the $t\bar{t}$ + B events from the inclusive $t\bar{t}$ + jets sample.

Various generator settings are chosen differently between the $t\bar{t}$ + jets and $t\bar{t} + b\bar{b}$ samples in order to reflect the different relevant scales. In the $t\bar{t}$ + jets sample only the top-quark scale is relevant while the bottom-quark and top-quark scales need to be considered in the $t\bar{t} + b\bar{b}$ samples. The details are given in table 4.4.

## 4.3. Uncertainties in the simulation

The uncertainties on the simulation due to the cross-section calculation and the modeling of the hard process, parton shower and hadronization are covered in this section. Further experimental uncertainties are discussed in the next chapter and in section 6.2.

Because the QCD multijet background is estimated with a data-driven approach, no uncertainties for this background will be discussed in this section. A detailed description of the method and the considered uncertainties will be given in section 5.5. In the estimation method, collected data in a control region is taken and all considered background from simulation ($t\bar{t}$ + jets, Single-top, $t\bar{t}V$, V + jets, and VV) are subtracted. The resulting shape is identified as the multijet background. Since the simulated backgrounds are used in this method, the resulting multijet estimation is subject to all uncertainties that affect a given background as all uncertainties for a given process are fully correlated between the signal region and the control region. If for example, the $t\bar{t}$ + LF background fluctuates up due to an uncertainty, the multijet shape will fluctuate down according to the relative contribution of $t\bar{t}$ + LF in the control region, which also fluctuates up. The technical implementation is described in section 6.1.

[6] An additional uncertainty for the collinear gluon-splitting, which only effects the $t\bar{t} + 2b$ subprocess, is considered in the analysis.

### 4.3.1. Statistical uncertainty

The limited size of the simulated samples leads to statistical fluctuations of the nominal prediction. The resulting uncertainty is taken into account via the Barlow-Beeston method [160]. In this method a Gaussian constrained nuisance parameter is introduced, that scales with the total yield of the sample, for each bin of the discriminant[7]. If less than 10 events are present in a given bin, a Poisson probability density function is used instead. The statistical uncertainty of the simulated signal samples is not taken into account when evaluating whether Gaussian or Poisson constrained parameter should be used.

### 4.3.2. Rate uncertainties

As discussed in section 4.2.1, the rates of the various processes are predicted by normalizing the process to the inclusive cross-section. In this analysis, the uncertainties on the cross-section calculation are considered as pure rate uncertainties. This means, that the uncertainties only effect the yield but not the shape of the distribution. The numerical values of these uncertainties are provided with the nominal value, which is used to scale the samples according to equation (4.1). The cross-section uncertainties mainly depend on the choice of renormalization and factorization scales and the uncertainties associated to the choice of the PDF. These two sources are considered separately in this analysis and are denoted as "scale" and "PDF" uncertainties.

The numerical values for uncertainties considered in the different processes are given in table 4.5. A simplified correlation scheme is used to describe how the processes interact: Firstly, different final states originating from the same initial state process (e.g. s-, t and tW channels of the Single-top process) are considered to be fully correlated for the scale- and PDF uncertainties. Scale uncertainties can be correlated between similar processes because higher order corrections have a similar effects (e.g. $t\bar{t}W$ and $t\bar{t}Z$). PDF uncertainties on the cross-sections are considered fully uncorrelated for the signals $t\bar{t}H$, tHq, tHW. All backgrounds are grouped into processes originating mostly from gg ($t\bar{t}$, $t\bar{t}Z$), gq (Single-top) or qq (V + jets, $t\bar{t}Z$, Diboson) initial states. The groups are fully uncorrelated between each other while the processes within a group are fully correlated. Since the theory prediction is independent of the detector configuration, these uncertainties are fully correlated between the three data-taking periods.

### 4.3.3. Shape uncertainties

Further uncertainties, which are introduced by the event generator, are considered to affect the shape of distributions. One a technical level, these uncertainties can be either implemented by rerunning the full simulation with a varied parameter or as event weights that are calculated during the simulation. Such weights correspond to the probability ratio of obtaining a simulated event with a variated value compared to the nominal value. Since these effect are not dependent on the LHC data-taking, these uncertainties are fully correlated between the data-taking periods.
The following uncertainties are considered:

- The variations of distributions due to the PDF set uncertainty is evaluated by reweighting to the variations provided with the NNPDF sets `NNPDF31_nnlo_hessian_pdfas` and `NNPDF31_nnlo_as_0118_nf_4` ($t\bar{t} + b\bar{b}$ sample only).
  The PDF sets are provided with a large number of replicas (103/100), which correspond to variations of various parameters. The up and down variations, used in the fit, are constructed by calculating the residual difference between a variation and the nominal

[7] It is implied that signal will be extracted with a binned maximum likelihood fit (see section 3.3.3) to a binned discriminant distribution, which will be introduced in section 5.6.

Table 4.5.: Rate uncertainties from theory prediction considered in the analysis. One process in the table can consist of multiple channels, which are considered fully correlated. Additional details are given in the right-most column.
Process uncertainties marked with the same exponent ($^{(\alpha)}$, $^{(\beta)}$, $^{(\gamma)}$, $^{(\delta)}$) are considered fully correlated. All other cases are fully uncorrelated.

| | Uncertainties | | |
|---|---|---|---|
| Process | scale [%] | PDF [%] | |
| $t\bar{t}H$ | -9.2/+5.8 | $\pm$ 3.6 | Higgs-decay modes are fully correlated |
| tHq | -14.9/+6.5 | $\pm$ 3.7 | Higgs-decay modes are fully correlated |
| tHW | -6.7/+4.9 | $\pm$ 6.3 | Higgs-decay modes are fully correlated |
| $t\bar{t}$ + B | -3.5/+2.4$^{(\alpha)}$ | $\pm$ 4.2$^{(\gamma)}$ | Additional 50% rate uncertainty for $t\bar{t}$ + B |
| $t\bar{t}$ + c$\bar{c}$ | -3.5/+2.4$^{(\alpha)}$ | $\pm$ 4.2$^{(\gamma)}$ | and $t\bar{t}$ + c$\bar{c}$ may be added for certain studies |
| $t\bar{t}$ + LF | -3.5/+2.4$^{(\alpha)}$ | $\pm$ 4.2$^{(\gamma)}$ | |
| Single-top | -2.1/+3.1 | $\pm$ 3 | All Single-top channels are fully correlated |
| $t\bar{t}Z$ | -9.3/+8.1$^{(\alpha)}$ | $\pm$ 3.5$^{(\gamma)}$ | |
| $t\bar{t}W$ | -16.4/+25.5$^{(\alpha)}$ | $\pm$ 2$^{(\delta)}$ | |
| Z + Jets | $\pm$ 2 $^{(\beta)}$ | $\pm$ 0.2$^{(\delta)}$ | |
| W + Jets | $\pm$ 3.8 $^{(\beta)}$ | -0.04/+0.08$^{(\delta)}$ | |
| Diboson | $\pm$ 3 | $\pm$ 5$^{(\delta)}$ | ZZ, WZ and WZ are fully correlated |

distribution in each bin and computing the envelope of the residuals. For `NNPDF31_nnlo_hessian_pdfas`, the envelope is computed from the quadratic sum of all residuals. For `NNPDF31_nnlo_as_0118_nf_4`, it is calculated as the RMS of all residuals. Only the shape and acceptance effects are considered for this uncertainty since the overall normalization of the PDF uncertainty is covered by the cross-section uncertainty. This is achieved by normalizing all PDF variations to the event yield before any selection.
This uncertainty is considered for the $t\bar{t}H$, tHq, tHW, $t\bar{t}$ + B, $t\bar{t}$ + C and $t\bar{t}$ + LF processes. The $t\bar{t}H$, $t\bar{t}$ + C and $t\bar{t}$ + LF processes are correlated. All other processes are considered fully uncorrelated.

- Effects of the variation of the $\mu_F$ and $\mu_R$ parameters on the shape of the distributions are considered for $t\bar{t}H$, tHq, tHW and $t\bar{t}$. Weights, modeling the variation of the scales by a factor of 0.5 and 2 independently, are provided by the ME generator. As for the PDF uncertainty, only shape and acceptance effects are considered since the uncertainty on the overall normalization is covered by the scale uncertainty on the cross-sections. Hence, the variations are normalized to the event yield before any selection.
  The uncertainty is considered fully correlated between $t\bar{t}$ + C and $t\bar{t}$ + LF. All other considered processes are fully uncorrelated.
- The impact of the choice of $\alpha_s$ on the parton shower scales is evaluated for initial-state (ISR) and final-state radiation (FSR). The scales are variated independently by factors of 0.5 and 2 using weights obtained from the parton-shower generator.
  The uncertainty is considered for the $t\bar{t}H$, $t\bar{t}$ + B, $t\bar{t}$ + C and $t\bar{t}$ + LF processes. All processes are considered fully uncorrelated.
- The effects due to the uncertainty of the underlying event tune and the ME-PS matching scale (`hdamp`) are only considered for $t\bar{t}$ + jets. These variations are not available

as weights. Instead separate samples with variated parameters were generated. Due to computing constraints, these samples cannot be generated with the same statistical power as the nominal samples used in the analysis. In practice this results in distributions with statistical fluctuations, which make it unpractical to derive shape uncertainties for these samples. Therefore, rate uncertainties are derived, by calculating the difference between the nominal and variated sample in each data-taking period per signal region category. The average of the three data-taking periods is taken as the rate uncertainty of each variation.
The ME-PS matching scale uncertainties is considered fully uncorrelated between $t\bar{t}$+B, $t\bar{t}$ + C and $t\bar{t}$ + LF since a dependency on the flavor is excepted. The underlying event, on the other hand, is not expected to depend on the flavor and is therefore correlated between all $t\bar{t}$ + Jets processes.

### 4.3.4. Heavy flavor production uncertainties

Aside from the various uncertainties on the kinematics of the $t\bar{t}$ + HF processes described in the last sections, the overall normalizations of $t\bar{t}$ + B, and to a smaller extent $t\bar{t}$ + C, are a major source of uncertainty. In previous iterations of this analysis a flat 50% rate uncertainty was used in the fit. The size of this uncertainty can be motivated by the average accuracy of current $t\bar{t} + b\bar{b}$ measurements or uncertainties of theoretical predictions (current cross section predictions exhibit uncertainties of 30% to 45% at NLO [55]). Using a 50% uncertainty can be viewed as a conservative choice.
Due to the improvements to the $t\bar{t}$ + Jets background model, described in section 4.2.3, the normalizations on $t\bar{t}$ + B and $t\bar{t}$ + C are now left freely-floating in the final fit. Using this approach is theoretically well motivated and preferred from a statistical standpoint compared to the flat rate parameter.
In measurements in the leptonic-channels and the combination of all channels, a floating normalization yields good and stable results, because of the high sensitivity to this normalization and dedicated control regions. In the fully-hadronic channel this is not the case due to the overwhelming contribution of the QCD multijet background, which is also associated with a freely floating normalization. Therefore, the approach with a 50% rate uncertainties is preferable when performing fits in the fully-hadronic channel only.

An additional uncertainty on the modeling of collinear gluon splitting in $t\bar{t}$ + B is considered. The uncertainty is constructed by varying the $t\bar{t}$ + 2b contribution in $t\bar{t}$ + B by 50% up and down.

As described in section 4.2.3, the $t\bar{t}$ + B contribution (taken from the four-flavor samples) is scaled to the expectation in the five-flavor samples. This is also taken into account for the shape uncertainties derived using weights by normalizing the variation in the four-flavor scheme samples to the same variation in the five-flavor scheme samples before any selection.

# 5. Analysis Strategy in the fully hadronic channel

This chapter will describe the analysis strategy for the $t\bar{t}H(b\bar{b})$ fully-hadronic channel in detail. The analysis strategy can be summarized as follows: Multiple consecutive selection steps (trigger, object and event selection) are applied in order to select signal and control regions for the measurement and corrections are applied to the simulation in order the improve the modeling of certain observables compared to the data. The backgrounds can be estimated from simulation except for the QCD multijet background, which is estimated from data in a control region and transferred to the signal region. In order to improve the sensitivity of the fit, a discriminant, which separates signal and background events, is constructed. If it is not specified otherwise, data refers to collision data collected by the CMS experiment.
The choice of high level triggers is discussed in section 5.1. Followed by the analysis-specific data/simulation corrections, the object selections and the event selection in sections 5.2, 5.3 and 5.4, respectively. The data-driven estimation of the QCD multijet background is discussed in section 5.5. This chapter closes with section 5.6 which describes how the discriminant for the fit is constructed using a neural-network-based classifier.

The majority of the topics discussed in this chapter represent work done by the author of the thesis: The development of dedicated HLT trigger paths for the fully-hadronic $t\bar{t}H(b\bar{b})$ analysis for the 2017 and 2018 data-taking period and corresponding measurements, the advancements in the data-driven background estimation compared to the analyses presented in [10, 108] and [11] and corresponding measurements, as well as the development of a novel neural-network-based classifier for the signal identification.
The analysis of the fully-hadronic decay channel included in reference [11] was also the primary responsibility of the author of this thesis and has to be seen as an intermediate result, motivating various choices included in the analysis presented in this section. Examples for this are the overhaul of the multijet background estimation with a novel region definition[1] and the usage of events with 3 b-tagged jets as control region instead of signal

[1] The multijet estimation method was initially developed around a likelihood-based observable for QCD rejection (called QGLR) for the analysis of the 2016 dataset [10, 108]. While it proved very effective in this analysis, it was apparent after the analysis of the 2017 dataset [11] that the modeling of this observable in the newer dataset was suboptimal. For the analysis of the full Run-2 dataset a different variable had to be found, see section 5.5 for details on this topic.

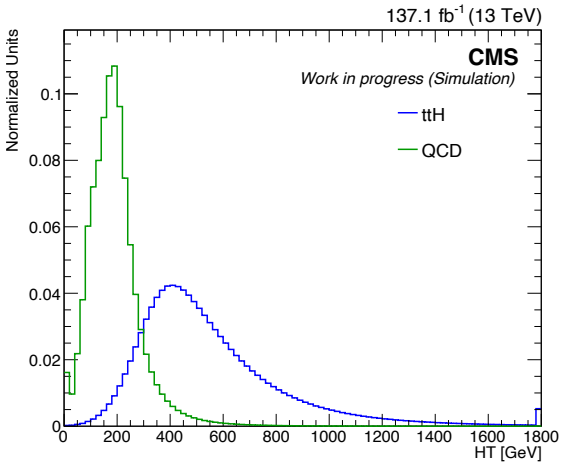



Figure 5.1.: $H_T$ (scalar sum of jet $p_T$) for fully-hadronic $t\bar{t}H(b\bar{b})$ (blue) and QCD (green) simulation before applying any event selection. Only jets with $p_T > 30\,\text{GeV}$ and $|\eta| < 2.4$ are considered.

region, as it was used previously, or the move to the neural-network based classifier[2].

## 5.1. Trigger

Dedicated high level triggers have been developed for analyses targeting the production of $t\bar{t}H(b\bar{b})$ and $t\bar{t}$ + Jets in the fully-hadronic decay channel. These trigger paths are based on particle-flow (PF) jets, see section 2.3.4, and b-tagging, see section 2.3.6. The development of new trigger paths for the 2017 and 2018 data-taking periods is a major part of the analysis presented in this thesis and a detailed discussion can be found in appendix A. This section focuses on the trigger strategy and the dedicated data/simulation correction factors derived for the phase space of this analysis.

### 5.1.1. High-level trigger paths

While analyses with leptons (e and $\mu$) in the final state can achieve very good signal efficiencies by relying on general single- or multilepton trigger paths, analyses with pure hadronic final states usually require a more phase-space specific approach. For example, in the case of the CMS trigger system, the least stringent and fully unprescaled HLT path targeting the scalar sum of the transverse momenta of all jet ($H_T$), which represents a typical path targeting hadronic final states, has a threshold of $H_T \geq 1050\,\text{GeV}$. While a cut on $H_T \geq 1050\,\text{GeV}$ would remove the vast majority of hadronic background activity, it would also remove most of the $t\bar{t}H$ signal, as it can be seen in figure 5.1. Even if more complex selections are considered, the thresholds on hadronic trigger paths tend to be rather high. For these reasons, the development of the trigger strategy plays a comparatively large role in the design of an analysis targeting fully-hadronic final states. Figure 5.1 shows the $H_T$ before any event selection in simulated fully-hadronic $t\bar{t}H(b\bar{b})$ and QCD multijet sample. From this figure, it can be seen that cuts on the $H_T$ should be as small as possible, ideally in the vicinity of 300 GeV, in order to preserve as much signal as possible signal.

[2] In reference [11], the MEM was still used as primary method for signal identification but it was augmented by an cut on an additional observable for QCD rejection $\Delta\eta_{\text{jets}}$. This approach proved to reduce sensitivity compared to initial estimations. Both observable, the MEM and $\Delta\eta_{\text{jets}}$, are now used as input features in the neural-network-based classifier.

Table 5.1.: Overview of the trigger paths used in the $t\bar{t}H(b\bar{b})$ fully-hadronic analysis in the 2016, 2017, and 2018 data-taking periods. Only the relevant cuts and values are listed in this table. E.g. the working point of the b-tagging, L1-triggers or additional cut on calorimeter jets can vary from period to period or in some cases within an period but do not contirbute to the understanding of the trigger strategy. More information on the utilized trigger paths can be found in appendix A.

| | | 2016 | 2017 | 2018 |
|---|---|---|---|---|
| `HLT_6J1T` | number of PF Jets | $\geq 6$ | $\geq 6$ | $\geq 6$ |
| | PF $H_T$ | $\geq 450\,\text{GeV}$ | $\geq 430\,\text{GeV}$ | $\geq 450$ (430) GeV |
| | $p_{T,6}$ | $\geq 40\,\text{GeV}$ | $\geq 40\,\text{GeV}$ | $\geq 36$ (40) GeV |
| | number b-tagged PF Jet | $\geq 1$ | $\geq 1$ | $\geq 1$ |
| `HLT_6J2T` | number of PF Jets | $\geq 6$ | $\geq 6$ | $\geq 6$ |
| | PF $H_T$ | $\geq 400\,\text{GeV}$ | $\geq 380\,\text{GeV}$ | $\geq 400$ (380) GeV |
| | $p_{T,6}$ | $\geq 30\,\text{GeV}$ | $\geq 32\,\text{GeV}$ | $\geq 32\,\text{GeV}$ |
| | number b-tagged PF Jet | $\geq 2$ | $\geq 2$ | $\geq 2$ |
| `HLT_4J3T` | number of PF Jets | – | $\geq 4$ | $\geq 4$ |
| | PF $H_T$ | – | $\geq 300\,\text{GeV}$ | $\geq 330\,\text{GeV}$ |
| | $p_{T,1}$ | – | $\geq 75\,\text{GeV}$ | $\geq 75\,\text{GeV}$ |
| | $p_{T,2}$ | – | $\geq 60\,\text{GeV}$ | $\geq 60\,\text{GeV}$ |
| | $p_{T,3}$ | – | $\geq 45\,\text{GeV}$ | $\geq 45\,\text{GeV}$ |
| | $p_{T,4}$ | – | $\geq 40\,\text{GeV}$ | $\geq 40\,\text{GeV}$ |
| | number b-tagged PF Jet | – | $\geq 3$ | $\geq 3$ |
| `HLT_PFJet450` | $p_{T,1}$ | $\geq 450\,\text{GeV}$ | – | – |
| `HLT_PFHT1050` | PF $H_T$ | – | $\geq 1050\,\text{GeV}$ | |

In order to keep the trigger thresholds as loose as possible and therefore keep the signal efficiency high, a phase-space specific combination of object selections is used. The trigger paths considered for the fully-hadronic $t\bar{t}H(b\bar{b})$ analysis, select events with high $H_T$ and multiple particle-flow jets in an event. Thresholds are further set on the $p_T$ of individual jets and b-tagging algorithms are utilized. Multiple trigger paths are used simultaneously in order to maximize signal efficiency. In general, the paths follow one of three templates:

1. Tightest $H_T$ cut, $\geq 6$ jets and $\geq 1$ b-tagged jets
2. Medium $H_T$ cut, $\geq 6$ jets and $\geq 2$ b-tagged jets
3. Loosest $H_T$ cut, $\geq 4$ jets and $\geq 3$ b-tagged jets (2017 and 2018 data-taking period only)

These templates will be denoted with `HLT_6J1T`, `HLT_6J2T` and `HLT_4J3T` in the following. Generally speaking, increasing the threshold on the b-tagging, e.g. going from $\geq 1$ to $\geq 2$ b-tagged jets, allows for lower threshold on other kinematic observables, like the $H_T$. Additional trigger paths may be considered in order to overcome inefficiencies that can only be observed after data-taking. Events are considered for the analysis if at least one of the considered trigger paths was fired. Table 5.1 summarizes all trigger paths used in this analysis.

During the 2016 data-taking period, HLT-paths (following the `HLT_6J1T` and `HLT_6J2T` templates) targeting events with $\geq 6$ particle-flow jets, $\geq 1$ or $\geq 2$ b-tagged jets and high $H_T$ (450 or 400 GeV), are used. The path are fed by various L1-Seeds based on L1-$H_T$.

A drop in efficiency at high $H_T$ values was observed in data. In order to recover the lost efficiency, an additional single jet trigger, targeting events with at least one jet with $p_T \geq 450\,\text{GeV}$ (`HLT_PFJet450`), is used.
During the 2017 and 2018 data-taking period, the same general trigger strategy is used. Following the increase of the instantaneous luminosity delivered by the LHC machine, the thresholds for the existing `HLT_6J1T` and `HLT_6J2T` paths, which were carried over from the 2016 data-taking, had to be increased. This was mainly accomplished via more stringent cuts on the b-tagging at trigger-level but also the $H_T$ thresholds were increased. In order to mitigate the loss in signal efficiency and add phase space with lower $H_T$, a new trigger path, which targets events with $\geq 4$ particle-flow jets and $\geq 3$ b-tagged jets (`HLT_4J3T`), was developed. This is motivated by the four b-tagged jets that are nominally expected in a $t\bar{t}H(b\bar{b})$ event. The $H_T$ threshold can be set very low[3] when requiring $\geq 3$ b-tagged jets, even though the threshold on the b-tagging is comparatively loose. Additionally, cascading $p_T$ thresholds are placed on the four jets with the highest transverse momentum (75, 60, 45, and 40 GeV) in order to maximize the efficiency and minimize the $H_T$ threshold. The same approach is utilized in the L1-triggers that are used for this HLT path. A drop in efficiency at high $H_T$ was observed in data. In order to recover the lost efficiency, the lowest fully unprescaled $H_T$-trigger path, requiring HT $\geq 1050\,\text{GeV}$ (`HLT_PFHT1050`), is used. Furthermore, the trigger-level b-tagging was improved between the 2017 and 2018 data-taking periods. Switching to a newer algorithm led to lower rates at the same mis-tag rate.

Further information on the trigger strategy, configuration and combination as well as the development of the `HLT_4J3T` trigger paths can be found in appendix A.

### 5.1.2. Trigger efficiency and corrections

Combining all trigger paths discussed in the last section, signal efficiencies[4] of 97.6 %, 93.9 % and 94.4 % are achieved for the 2016, 2017 or 2018 data-taking periods, respectively. The signal efficiency is the fraction of all signal events that pass the trigger strategy. The overall loss in signal efficiency in later periods can be mainly attributed to the tighter cuts necessary because of the increase in instantaneous luminosity and therefore increased pileup. Due to improvements in trigger-level b-tagging, the signal efficiency could be increased from the 2017 to the 2018 data-taking period even though thresholds had to be tightened further. Without the addition of the `HLT_4J3T` in the 2017 and 2018 data-taking periods, the signal efficiency would have dropped to 90.3 % and 92 %, respectively. The signal efficiencies for the individual paths as well as various other combination are listed in table A.1 in the appendix.

The trigger efficiency is further measured with respect to orthogonal reference triggers `HLT_IsoMu24` (2016) and `HLT_IsoMu27` (2017, 2018) in the single muon dataset[5]. These triggers target event with a single reconstructed and isolated muon with at least 24 GeV or 27 GeV, respectively. For the measurement, events with exactly one muon, $\geq 6$ Jets, $\geq 2$ b-tagged jets, $H_T \geq 500\,\text{GeV}$ are used. Only jets with $p_T \geq 30\,\text{GeV}$ and $|\eta| < 2.4$ are considered but six jets with $p_T$ 40 GeV are required. The efficiency is calculated by

[3] The lower bound of the $H_T$ cut is defined by the $H_T$ threshold used in the L1-trigger, which is around 300 GeV, depending on the delivered instantaneous luminosity.

[4] The signal efficiencies are estimated in $t\bar{t}H(b\bar{b})$ simulation in a region with $H_T > 500$, $\geq 6$ jets, $\geq 2$ `Medium` b-tagged jets and no e or $\mu$ with $p_T \geq 15\,\text{GeV}$ and $|\eta| < 2.4$. Only jets with $p_T \geq 30\,\text{GeV}$ and $|\eta| < 2.4$ are considered and exactly six jets with $p_T \geq 40\,\text{GeV}$ are required. This region estimates the baseline region described in section 5.4.

[5] The correlation between the signal and reference triggers was estimated to less then 1% in all data-taking periods. It is measured using the formula $(\epsilon(\text{Reference}) \cdot \epsilon(\text{Signal}))/\epsilon(\text{Reference and Signal})$ with the trigger efficiencies $\epsilon$.

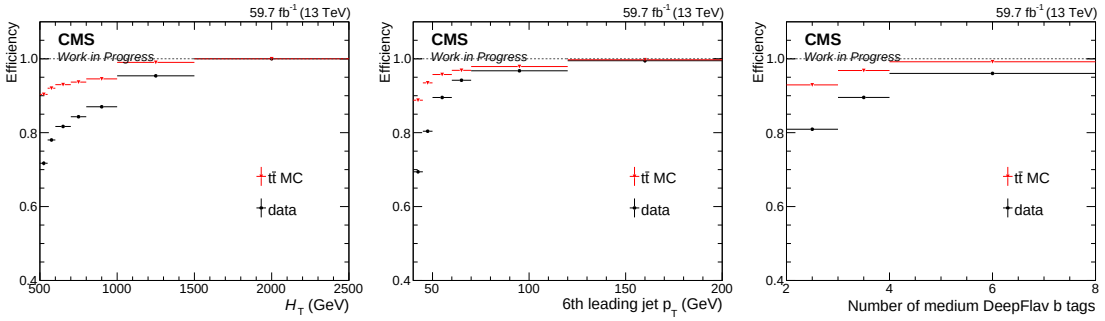



Figure 5.2.: Trigger efficiency measured in $H_T$ (left), $p_T$ of the sixth jet (middle) and number of `Medium` b-tagged jets (right) in the 2018 dataset with respect to `HLT_IsoMu27` in events with one offline lepton and passing the fully-hadronic baseline selection (excluding the usual lepton veto). The efficiency was measured in data (black) and $t\bar{t}$ + Jets simulation (red).

taking the ratio of all selected events passing the combination of the trigger paths discussed above as well as the reference single-muon trigger, and all selected event passing only the reference single-muon trigger.

Figure 5.2 shows the trigger efficiency in the 2018 data-taking period. It can be observed, that the efficiency in simulation is not modeled well compared to data. Therefore, correction factors are derived to improve the modeling in simulation. These correction factors are derived per data-taking period as a function of the $H_T$, $p_T$ of the jet with the sixth largest transverse momentum ($p_{T,6}$), and the number jets passing the `Medium` b-tagging working point[6] (nTags). The efficiencies for the 2016 and 2017 data-taking period are shown in figures A.2 and A.3. On a technical level, the correction factors are implemented as bin-by-bin rescaling factor in the $H_T/p_{T,6}$/nTags space and the statistical uncertainties on the ratio is treated as a systematic uncertainty in the final fit[7]. The majority of the correction factors are close to unity but overall ranging from 0.7 to 1.07. The binning is chosen such, that adjacent bins with similar correction factor values are grouped together in order increase the statistical significance of a given bin. The bin uncertainties are 1-2 % on average, with single bins of up to 9 %. The uncertainty on the trigger correction factors plays a negligible role in the final fit to data.

## 5.2. Data-Simulation Corrections

The simulation of physics processes can be used to make predictions with high accuracy and efforts to further improve this are made at all times, e.g. new tunes for the parton shower. Nevertheless, small discrepancies between data and simulation can be observed. By comparing the modeling of relevant processes in simulation to data in control samples, correction factors are derived. Usually this is done by taking the differences, more specifically the over or underestimation, between the data and simulation in a specific variable, or set of variables, from these control samples and applying the inverse difference as correction in the signal region. This procedure is referred to as reweighting.
Section 5.1.2 already introduced one such correction for the triggers. Further relevant corrections will be covered in this section.

[6] The scale factors are derived with respect to the b-tagging algorithm that is used in the analysis (`DeepJet`), while other algorithms are used on trigger-level (`CSVv2` in 2016, 2017 and `DeepCSV` in 2018). These differences between trigger-level and analysis-level b-tagging algorithms have an effect on the trigger efficiency turn-on that has no influence on the analysis itself. This topic will be discussed further in appendix A.

[7] The uncertainty is dominated by statistical effects. Systematic effects, like the correlation to the reference trigger, are negligible and therefore omitted. More details are given in appendix A.

### 5.2.1. Pileup correction

Due to the large effects of the pileup on object identification, e.g. lepton isolation or jet reconstruction, it is important to use the same distribution of pileup events in data and simulation. Since the average amount of pileup is unknown at the time the simulated samples are generated, a reweighting has to be applied, such that the pileup distributions match in data and simulation.
The weight is derived by taking the distribution of pileup events in data, as defined in equation (2.6), and comparing it to the corresponding distribution in simulation. For the nominal weight, a total inelastic pp cross section $\sigma_{inel}$ of 69.2 mb is used for the calculation. Variations to the distribution, used estimate the uncertainty, are separately computed by varying $\sigma_{inel}$ by 4.6 % up and down and repeating the calculation. Since the amount of pileup events depends on the machine parameters, the calculation is done separately per data-taking period. Furthermore, the weight is calculated for each simulated sample separately. This results in a per-event weight that decreases the importance of simulated events with pileup uncommon in data.

### 5.2.2. B-tag correction

Differences in the b-tagging efficiency and mis-tagging rate between data and simulation are taken into account. The correction is derived separately for light-flavor jets and b-jets with a tag-and-probe approach: Events with exactly two jets, of which one is required to pass a tagging criteria, are selected. The correction is directly derived by comparing the probe jet b-tagging value in simulation to the distribution in data. With this method, per-jet correction-factors, that are parameterized in the b-tag discriminant value, flavor (light, charm or bottom), $p_{\mathrm{T}}$ and $\eta$, are derived. Since the correction can lead to jets moving below the considered working point, this is equivalent to an efficiency correction at the considered WP. In the analysis, these correction factors are evaluated for each jet passing the selection and the product of these factors is used as the event weight. For c-jets, that correction factor is set to unity but the uncertainty is set to twice the b-jet correction. The scale factors are derived by the b-tagging experts in the CMS collaboration.

After application of the correction factors and before applying any b-tagging selection, shape changes can be observed in the jet multiplicity and other kinematic variables, e.g. leading jet $p_{\mathrm{T}}$, with respect to before applying the correction. This should not be the case for observable independent on the b-tagging. In order to mitigate this unwanted distortion effect, an additional event reweighting is applied. The correction is derived as function of jet multiplicity and the event $H_{\mathrm{T}}$, independently for $\mathrm{t\bar{t}H}$, $\mathrm{t\bar{t}} + \mathrm{B}$, $\mathrm{t\bar{t}} + \mathrm{C}$ and $\mathrm{t\bar{t}} + \mathrm{LF}$. For the minor backgrounds, the $\mathrm{t\bar{t}} + \mathrm{LF}$ correction is used. In order to derive the weights, the jet multiplicity and $H_{\mathrm{T}}$ distributions for all events passing the baseline event selection with and without applying the b-tag efficiency and mis-tag correction are compared. The inverse of the difference per $H_{\mathrm{T}}$ and jet multiplicity bin is used as weights for the simulation.
Applying these weights reduces the distortion effect in kinematic distributions without changing the shape of the b-tagging discriminant distribution.

### 5.2.3. Other data/mc corrections

During the 2016 and 2017 data-taking, a gradual timing shift of the ECAL was observed but not propagated properly to the L1-trigger primitives. This results in a non-negligible fraction of trigger primitives being wrongfully associated with previous bunch crossings in high-$|\eta|$. The L1 trigger cannot fire in two consecutive bunch crossings, which leads to events vetoed if a significant amount of ECAL energy is found in $2 < |\eta| < 3$. This issue is called the *L1 ECAL Prefiring*. A reweighting, which computes the probability of an

event not to prefire based on kinematic observable, is applied to the simulation in order to correctly model this issue.

General object corrections are considered for jets, electrons and muons and are covered in sections 2.3.5, 2.3.2 and 2.3.3 respectively.

## 5.3. Object selection

Analysis-specific selections are applied to the objects reconstructed with the particle-flow algorithm, as described in section 2.3. These selections are performed with the goal of rejecting physics objects, like leptons and jets, that are likely to be mis-identified or not well measured.

### 5.3.1. Lepton selection

In the fully-hadronic channel, leptons are solely used to veto events that are used in the context of the leptonic channels. By applying this veto, orthogonality between the channels is ensured. The objects are not used in any other capacity in the context of the fully-hadronic channel. Nevertheless the full selection will be described here since the combination of all channel will be covered in chapter 7.

Particle-flow reconstructed electrons and muons, as defined in sections 2.3.2 and 2.3.3, are selected based on the kinematic variables $p_{\mathrm{T}}$ and $\eta$. Identification and isolation criteria are defined to select high-purity prompt leptons that originate from the W-boson decays. Multiple types of electrons and muons are defined for different use cases.

#### Muon selection

The cut-based muon identification [82] criterion imposes minimum quality standards on the reconstruction to be fulfilled. Cuts are applied to the $\chi^2$/ndof of the global muon fit as well as the number of muon-chamber hits, pixel hits and tracker layers with hits. Furthermore, the tracker tracks of the muon require minimum values for the impact parameters $d_{xy}$ and $d_z$. The tracks also need to be matched to muon segments in a minimum number of muon stations. The numerical values for the cuts are listed in table B.1.
All types of muons used in the $\mathrm{t\bar{t}H(b\bar{b})}$ analyses are required to be measured in the central detector ($|\eta| < 2.4$). The relative isolation is defined as

$$\left( \sum_{\Delta R<0.4} p_T^{h\pm} + \max\left(0, \sum_{\Delta R<0.4} E_T^{h0} + \sum_{\Delta R<0.4} E_T^{\gamma} - \frac{1}{2} \sum_{\Delta R<0.4} p_T^{pu}\right)\right) / p_{T,\mu} \tag{5.1}$$

with the isolation cone distance $\Delta R$, the $p_{\mathrm{T}}$ of charged hadrons from the Primary Vertex (PV) ($p_T^{h\pm}$), the $E_T$ of neutral hadrons ($E_T^{h0}$) and photons ($E_T^{\gamma}$), the $p_{\mathrm{T}}$ of charged hadrons from pileup ($p_T^{pu}$) and the muon $p_{\mathrm{T}}$ ($p_{T,\mu}$).

A *veto muon* (or *sub-leading muon*), which is the only type used in the fully-hadronic channel, is required to pass a minimum $p_{\mathrm{T}}$ of 15 GeV, a relative isolation of 0.25 and the cuts specified above.
A *leading muon* is required to pass a minimum $p_{\mathrm{T}}$ of 26/29/26 GeV (2016/2017/2018 data-taking periods), a relative isolation of 0.15 and the cuts above in the semi-leptonic channel. In the dileptonic channel, a *leading muon* is required to pass a minimum $p_{\mathrm{T}}$ of 25 GeV, a relative isolation of 0.25 and the cuts specified above.

**Electron selection**

A cut-based election identification criterion is used. Cuts are applied to the following observables:

| | |
|---|---|
| **5x5 $\sigma_{i\eta i\eta}$** | The weighted cluster RMS inside a 5x5 crystal region of the ECAL. |
| **$\Delta\eta_{seed}$ ($\Delta\phi_{in}$)** | The difference between track and track seed $\eta$ ($\phi$). |
| **$H/E$** | The ratio of HCAL to ECAL energy. |
| **$\lvert 1/E - 1/\rho \rvert$** | The absolute difference between the inverse ECAL energy and inverse track momentum. |
| **Relative isolation** | A Isolation based on the effective-area-corrected isolation |
| **$N_{miss}$** | The Number of missing hits in the tracker. |
| **$d_{xy}$ and $d_z$** | Impact parameters in $xy$-plane and z-direction. |
| **Conversion veto** | Rejection of electrons from photon conversion. |

The numerical values for the cuts are given in table B.2.

A *veto electron* (or *sub-leading electron*), which is the only type used in the fully-hadronic channel, is required the have $p_{\mathrm{T}}$ of >15 GeV and $|\eta| < 2.4$ in addition to the cuts specified above.
A *leading electron* is required to have $p_{\mathrm{T}}$ of > 29/30/30 GeV (2016/2017/2018 data-taking periods) and $|\eta| < 2.4$ in addition to the cuts above in the semi-leptonic channel. In the dileptonic channel, for a *leading electron* a minimum $p_{\mathrm{T}}$ of 25 GeV and $|\eta| < 2.4$ is required in addition to the cuts specified above.

### 5.3.2. Jet selection

Jets are reconstructed from particle-flow candidate particles, as described in section 2.3.5. Multiple additional quality criteria are applied in order to select jets that are not clustered from very few particle or noise and are unlikely to originate from pileup.
A cut-based identification is used to further select high quality jets based on the kinematic observables:

- Number of jet constituents
- Multiplicity of charged constituents
- Contributions of neutral and charged hadronic and electromagnetic energy depositions, measured in HCAL and ECAL, to the jet energy
- Contributions of muons to the jet energy

The numerical values for this jet-id are given in table B.3. A BDT-based pileup rejection [161] is applied to jets with $p_{\mathrm{T}} < 50$ GeV utilizing tracking and jet shape variables. Loose and tight working-points are defined for this discriminant corresponding to 80 % and 99 % efficiency, respectively. See reference [162] for more information on the CMS jet algorithms. Additionally a jet is rejected if a lepton, which passes the lepton selection described above, is present within a distance of 0.4 in $\eta - \phi$ space.
In the fully-hadronic and semileptonic channels, jets with $p_{\mathrm{T}} > 30$ GeV and $|\eta| < 2.4$, that pass the jet-id and the loose pileup id, are used.
In the dileptonic channel, jets with $p_{\mathrm{T}} > 30$ GeV and $|\eta| < 2.4$, that pass the jet id and the loose/tight/tight (2016/2017/2018 data-taking periods) pileup id, are used.

## 5.4. Event Selection

In order to increase sensitivity of the measurement, multiple cleaning and selection steps are applied to remove non-physical, mis-measured and irrelevant events in simulation and data.

Criteria are developed by detector and object experts in the CMS collaboration in order to only select high-quality events for the use in physics analyses. After removing all event failing the data certification, described in section 2.4.2, further cuts are applied: Events must contain at least one primary vertex (PV) that exhibits

- a number of $> 4$ degrees of freedom in the fit used to find the PV,
- an absolute value of $\leq 4\,\text{cm}$ for the z-coordinate,
- an absolute value of $\leq 2\,\text{cm}$ for the $\rho$-coordinate,
- and in simulation the PV must be matched to a simulated vertex.

Events are also rejected based on filters targeting certain effects, like detector noise or beam halo effects. During the 2018 data-taking period, a failure in the HEM15/16 sectors of the HCAL effected the physics object reconstruction in the detector region with $-3.0 < \eta < -1.3$ and $-1.57 < \phi < -0.87$. The direct effect of this issue on the relevant distributions was found to be negligible. To account for this issue, a dedicated uncertainty is applied to jets in simulation in this area (see section 6.2 for details on this uncertainty).

### 5.4.1. Baseline selection

The previously described event selection criteria are common for all CMS analyses. The first step in the analysis-specific selection will be discussed in the following. It will be referred to as *baseline selection* throughout the rest of this thesis.

The main purpose of the baseline selection is to only select events that are at or close to the plateau of the trigger, i.e. where the triggers are fully efficient. Additionally, the region is used for various studies and validations performed in the initial stages of an analysis. For the fully-hadronic channel, events must pass the following criteria:

- No leptons ($e$ and $\mu$) passing the veto selections described in section 5.3.1
- Pass at least one of the analysis-specific HLT triggers, as discussed in section 5.1
- Scalar sum of the $p_\text{T}$ of all selected jet ($H_\text{T}$) $> 500\,\text{GeV}$
- Six or more jets passing the selection described in section 5.3.2
- Six jets with $p_\text{T} \geq 40\,\text{GeV}$
- At least 2 jets with b-tagging values greater than the `Medium` working point, as defined in section 2.3.6.

The relative efficiency of each cut (excluding the lepton veto) is given in table 5.2 for fully-hadronic events in the $\text{t}\bar{\text{t}}\text{H}(\text{b}\bar{\text{b}})$, QCD, and $\text{t}\bar{\text{t}}$ + Jets samples. After selecting events with $\geq 2$ b-tagged jets, 30.34 % of the fully-hadronic $\text{t}\bar{\text{t}}\text{H}(\text{b}\bar{\text{b}})$ events remain while only 4.17 % and 0.02 % fully-hadronic event pass the selection for the $\text{t}\bar{\text{t}}$ + Jets and QCD sample, respectively.

The dijet mass $m_\text{qq}$, defined as the mass of the dijet pair closest to the W-boson mass, is required to be within $30 > m_\text{qq} > 250\,\text{GeV}$. When calculating this observable, the four jets with the highest b-tagging discriminant values are excluded. Events with $m_\text{qq}$ values far from the nominal W-boson mass are unlikely to originate from the signal process, because it is unlikely that for each of the W bosons one jet is out of acceptance. Hence, selecting events with $m_\text{qq}$ values close the W-boson mass improves the signal purity. This is not done at this stage of the selection, because $m_\text{qq}$ is an integral component of the data-driven background estimation of the QCD multijet background, where the regions

Table 5.2.: Efficiency of the selection steps in the baseline selection, applied to the fully-hadronic $t\bar{t}H(b\bar{b})$, QCD and $t\bar{t}$ + jets simulations. Each row includes the cuts from all previous rows. The efficiency is given as average over the three data-taking periods.

| | Efficiency [%] | | |
|---|---|---|---|
| | $t\bar{t}H(b\bar{b})$ | QCD | $t\bar{t}$ + jets |
| Trigger | 54.50 | 0.29 | 12.37 |
| $\geq$ 6 Jets | 48.44 | 0.16 | 10.43 |
| 6th Jet $p_T \geq 40\,\text{GeV}$ | 38.73 | 0.11 | 7.94 |
| $H_T > 500\,\text{GeV}$ | 32.45 | 0.09 | 6.25 |
| $\geq$ 2 b-tags | 30.34 | 0.02 | 4.17 |

far from the W-boson mass are used for a validation step while the central region is used for the measurement. The 30 and 250 GeV thresholds are solely defined to remove outliers from the baseline region.

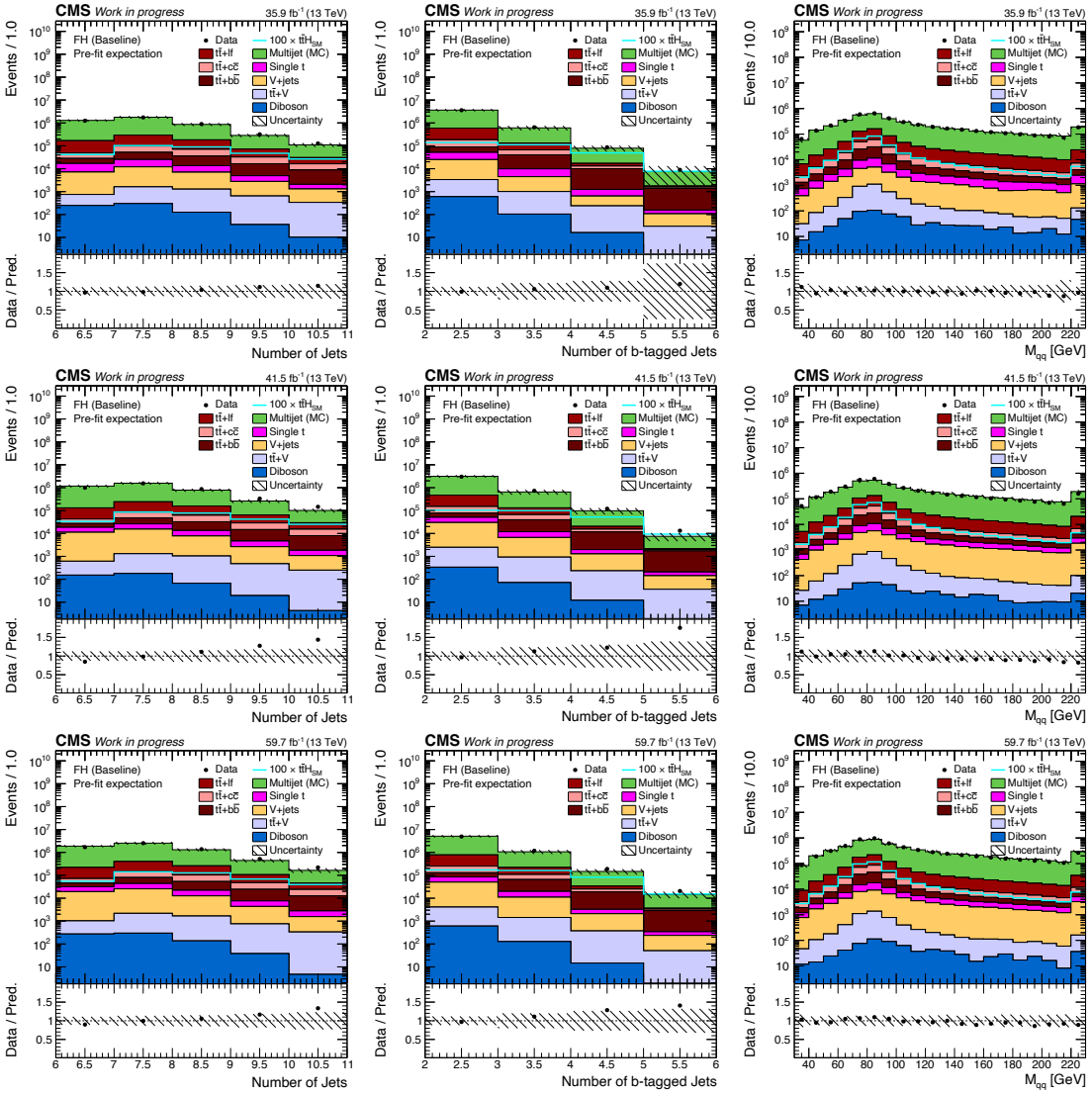
Figure 5.3 shows the number of jets, number of b-tagged jets and the $m_{qq}$ variables in the baseline region for the three data-taking periods separately. The uncertainties given in the figure correspond to the statistical and the majority of systematic uncertainties considered in the analysis. The QCD multijet background shown in these figures is taken from simulation and rescaled such that the total background yield prediction is equal to the data yield. Additionally, a reweighting is applied to the 2017 and 2018 QCD multijet simulation such that the jet multiplicity spectra match the distribution observed in the 2016 QCD multijet simulation. This step is motivated by the circumstance, that the parton shower tune `CP5` was derived with the goal of improving the modeling of the $t\bar{t}$ + Jets simulation with respect to the tune `CUETP8M1`. While the tune `CP5` achieved this goal in $t\bar{t}$ + Jets simulation, the opposite can be observed in the QCD and other[8] simulations. Especially in the phase space with a high jet multiplicity, where the QCD multijet simulation does not model the data well in the first place, this leads to large disagreements when comparing data and simulation.
Overall the simulation and data agree well within the uncertainties in all data-taking periods. The remaining disagreements in the high jet and b-tag multiplicity bins clearly motivate the need for a data-driven background estimation of the QCD multijet background, which will be covered in section 5.5.

### 5.4.2. Signal and control regions

Orthogonal categories and regions, based on the jet and b-tag multiplicity as well as the $m_{qq}$ variable, are defined. In general, the signal purity increases with increasing jet and b-tag multiplicity. This can for example be seen in the $S/\sqrt{B}$ given in figure 5.4. At the same time, the background composition changes such, that it is closer to the signal and therefore harder to separate. Hence, the sensitivity in the fit to data does not necessarily increase with higher multiplicities. Since the analysis is mostly tuned towards separating the signal from the multijet background this effect is especially noticeable in the categories with very high jet count, where the contribution of the $t\bar{t} + b\bar{b}$ background is comparably large.

In the following the notation will be used:

[8] One example is the Z + Jets simulation as described in 7.1.




(a) (b) (c)

Figure 5.3.: Jet multiplicity (a), b-tag multiplicity (b) and $m_{qq}$ (c) for the 2016 (top), 2017 (middle), and 2018 (bottom) dataset in the baseline region. The expected contribution of the background processes is (filled histograms) is stacked and the signal (line) is superimposed and scaled for better visibility. The QCD simulation is scaled such, the predicted yield matches the yield in data. In 2017 and 2018 QCD is additionally weighted such that the jet multiplicity spectra match the 2016 simulation. The hatched uncertainty bands correspond to the total statistical and systematic uncertainties added in quadrature.

**Categories** result from splitting via the jet multiplicity and

**Regions** results from splitting via the b-tag multiplicity and the $m_{qq}$ observable.

As noted previously, $m_{qq}$ is expected to be close to the actual W-Boson mass in signal events. Hence, a signal-enriched $m_{qq}$-band around the W-boson mass and signal-depleted sidebands are defined: For events with 7 or 8 jets, the central band is defined by $60 \geq m_{qq} \geq 100\,\text{GeV}$ and for events with $\geq 9$ jets by $72 \geq m_{qq} \geq 90\,\text{GeV}$. The $m_{qq}$-sidebands are limited to $> 30\,\text{GeV}$ and $< 250\,\text{GeV}$.
With this scheme, regions are defined for the signal extraction, estimation of the QCD multijet background and the training of a neural-network based classifier. Each with a

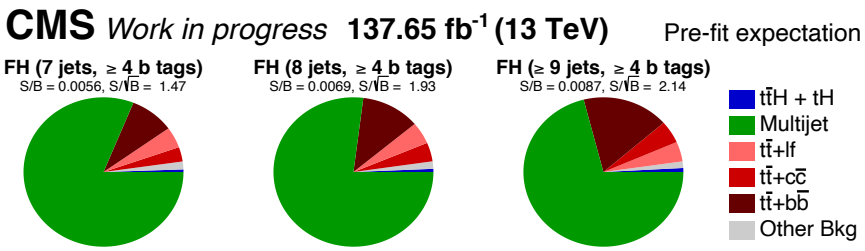



Figure 5.4.: Expected contribution of signal and main background processes in the fully-hadronic signal region.

corresponding validation region. An overview of the regions used in the analysis is given in table 5.3.

For the signal extraction, three categories with 7, 8, and $\geq 9$ jets and $\geq 4$ `Medium` b-tags in the signal-enriched $m_{\mathrm{qq}}$-band are used[9]. Collectively, these categories are also referred to as the *signal region (SR)* and will be referred to as the ($7\,\mathrm{jets}, \geq 4\,\mathrm{b\text{-}tags}$), ($8\,\mathrm{jets}, \geq 4\,\mathrm{b\text{-}tags}$), and ($\geq 9\,\mathrm{jets}, \geq 4\,\mathrm{b\text{-}tags}$) categories. Figure 5.4 shows the composition of the main backgrounds and the signal contribution in these categories for the full Run-2 dataset ($137.2\,\mathrm{fb}^{-1}$). The QCD multijet background is the dominant background in all categories with contributions $> 70$ %. The second largest background is $\mathrm{t\bar{t}} + \mathrm{B}$ followed by $\mathrm{t\bar{t}} + \mathrm{LF}$ in the 7 and 8 jets categories and $\mathrm{t\bar{t}} + \mathrm{C}$ in the $\geq 9$ jet category. All other background processes, listed in table 4.3, contribute between 1.5 % and 1.7 % to the total background yield depending on the category.

For each of the three jet multiplicity categories the following regions, in which the b-tag multiplicity and $m_{\mathrm{qq}}$ selections are replaced relative to the SR selection, are defined:
Three regions are used for the QCD multijet background estimation, which will be covered in section 5.5:

**CR** The *control region (CR)* includes events with 3 `Medium` b-tagged jets and $\geq 4$ `Loose` tagged jets[10] in the signal-enriched $m_{\mathrm{qq}}$-band.

**ValCR** The *validation control region (ValCR)* includes events in the $m_{\mathrm{qq}}$-sidebands that otherwise pass the same selection as the *CR*.

**VR** The *validation region (VR)* includes events in the $m_{\mathrm{qq}}$-sidebands that otherwise pass the same selection as the *SR*.

Two regions are used for the training and validation of the final classifier:

**TR** The *training region (TR)* includes events with 2 `Medium` b-tagged jets and $\geq 4$ `Loose` tagged jets in the signal-enriched $m_{\mathrm{qq}}$-band.

**ValTR** The *validation training region (ValTR)* includes events in the $m_{\mathrm{qq}}$-sidebands that otherwise pass the same selection as the *TR*.

## 5.5. Multijet background estimation

As noted in previous sections, in the phase space of the $\mathrm{t\bar{t}H(b\bar{b})}$ analysis the QCD multijet background is not modeled well in simulation. Because of the significant contribution from

[9] The direct predecessors to this analysis [10, 11] also used the events with 7, 8, and $\geq 9$ and `Medium` 3 b-tagged jets in the signal extraction. Studies showed, that the impact of these events to the overall result is minimal. Therefore, these categories are used as control regions in this analysis.

[10] Since all `Medium` b-tagged jets are also `Loose` tagged b-tags, this is equivalent to selecting 3 `Medium` tagged b-tags and one additional jet, that passes the `Loose` WP but not the `Medium` WP.

Table 5.3.: Definition and description of the orthogonal regions, which are defined for each of the jet multiplicity categories, in the fully-hadronic $t\bar{t}H(b\bar{b})$ analysis. The values in parenthesis for the $m_{qq}$ variable are used for event with $\geq 9$ jets. `DeepJetM` and `DeepJetL` refer to the multiplicity of jets passing the `Medium` and `Loose` working-point of the `DeepJet` b-tagging algorithm.

| | $60(72)\,\text{GeV} < m_{qq} < 100(90)\,\text{GeV}$ | $30\,\text{GeV} < m_{qq} < 60(72)\,\text{GeV}$ or $100(90)\,\text{GeV} < m_{qq} < 250\,\text{GeV}$ |
|---|---|---|
| `DeepJetM` $= 2$<br>`DeepJetL` $\geq 4$ | ***TR*: Training Region**<br>Training dataset for ANN discriminant | ***ValTR*: Validation TR**<br>ANN feature validation |
| `DeepJetM` $= 3$<br>`DeepJetL` $\geq 4$ | ***CR*: Control region**<br>Derive QCD multijet shape for SR | ***ValCR*: Validation CR**<br>Derive QCD multijet shape for VR |
| `DeepJetM` $\geq 4$ | ***SR*: Signal region**<br>Final analysis region | ***VR*: Validation region**<br>Validation of QCD multijet estimation in data before unblinding |

QCD multijet events in the signal region of the fully-hadronic channel, this background is estimated with a data-driven approach. The general strategy for such an estimation is to define a control region that is dominated by this background and has a negligible signal contribution. From this region, relevant observables are then extrapolated to the signal region.
In this analysis, the control region (*CR*) is defined by loosening the b-tagging requirement relative to the signal region (*SR*). While four or more `Medium` tagged jets are required in the *SR*, three `Medium` tagged jets and one or more additional `Loose` b-tagged jets are required in the *CR*. In order to validate the method a sideband, in which the estimation can be compared to the data before unblinding the signal region, is used. The sideband is defined via the $m_{qq}$ variable, as it was defined in section 5.4.1, which leads to a second set of control and signal regions for validation (*ValCR* and *VR*). This can be viewed as a two dimensional space, as it is also visualized in table 5.3, that is spanned by the number of `Loose` and `Medium` b-tags and the $m_{qq}$ variable. The former describing the step from *CR* to *SR* (or *ValCR* to *VR*) and the latter the step from *SR* to *VR*. Hereby it is important that these axis are correlated as little as possible in order to not bias the validation. This is achieved by constructing the $m_{qq}$ variables only from jets that are not counted towards the number of `Loose` and `Medium` b-tagged jets.

For statements applying to the central $m_{qq}$ region and the $m_{qq}$ sidebands, the *SR* + *VR* are also referred to as high b-tag regions and the *CR* + *ValCR* as low b-tag region.

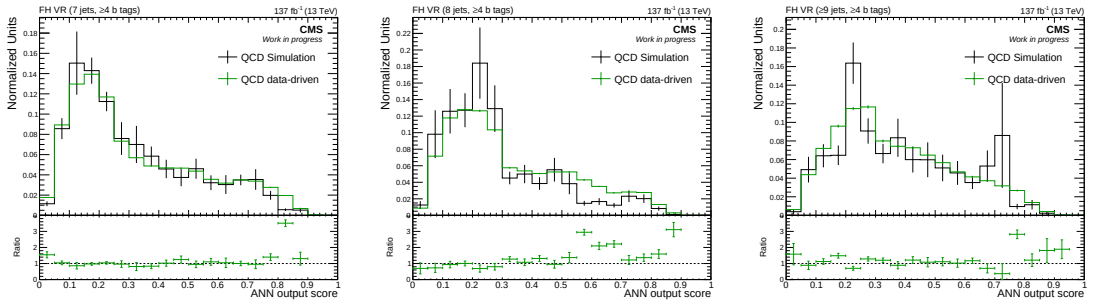



Figure 5.5.: Comparison of the final discriminant shape in the simulated QCD sample (black) and the data-driven method for QCD multijet estimation (green) in the 7 (left), 8 (middle) and $\geq 9$ jets validation regions (VR).

When comparing distribution shapes in the high and low b-tag regions[11] disagreements are expect due to kinematic differences of the additional `Loose` b-tagged jets in the low b-tag region and the corresponding `Medium` b-tagged jets in the high b-tag regions. In order to mitigate this effect, a kinematic correction for `Loose` but not `Medium` b-tagged jets was developed, that will be described in detail in section 5.5.1.

The process for deriving the QCD multijet estimation can be summarized as follows:

1. **Estimate the QCD multijet shape in the control region:** By acknowledging that the low b-tag control region is dominated by the QCD multijet background, it is a valid statement, that the shape of this background can be derived by subtracting the $\bar{\text{t}}\text{t}$ + Jets and minor background processes from the data.
2. **Correct the QCD multijet shape:** Differences in the kinematics of `Medium` b-tagged jets and `Loose` b-tagged jets can be observed in both simulation and data. A correction based on the $p_{\text{T}}$, $\eta$ and $\Delta R_{\text{min}}$ to the two jets with larges b-tagging discriminant value is applied in the low b-tag control region. This method will be described in section 5.5.1 and denoted as $\text{TF}_{\text{loose}}$ correction.
3. **Adjusting the QCD multijet yield to expectation in the signal region:** After adjusting the shape of the QCD multijet background from the control region it is assumed that it represents the shape of the QCD multijet background in the signal region. The yield cannot be predicted with this method. Therefore, it is left free-floating in the final fit. The initial value is set to the expected yield in the signal region, which is calculated by subtracting simulation from data.
4. **Validate the method:** Step 1-3 are applied to the $m_{\text{qq}}$ sidebands (*ValCR* and *VR*) in order to be able to compare the estimated QCD multijet background to data before unblinding the signal region.

A comparison between the shape of the QCD multijet simulation and the estimation derived with the method described above, is shown in figure 5.5 in the *VR*. It can be clearly observed that the estimated shape exhibits a much higher statistical power and is not fluctuating compared to the distribution of the simulated sample.

As noted in the third step of the workflow summary above, the normalization of the multijet background is a freely-floating parameter in the fit to data. The reason for this is, that no prediction on the normalization can be made when using this specific method. Setting the initial normalization to the *SR* expectation is somewhat arbitrary but also a reasonable starting point for the fit. Due the size of the multijet background and the

[11] Since the normalization is freely floating in the fit and the initial value is set by the *SR* (*VR*) expectation, only the shape agreement between the *SR* (*VR*) and *CR* (*ValCR*) regions is a relevant quantity to measure if the method is working as expected.

strong correlation between the normalization and the other uncertainties, the impact of this freely-floating parameter on the fit in the fully-hadronic channel is expected to be rather large. The normalization is independently floating in each of three categories, with (7 jets, $\geq 4$ b-tags), (8 jets, $\geq 4$ b-tags) and ($\geq 9$ jets, $\geq 4$ b-tags). This means, that, unlike other uncertainties, the multijet normalization can not be constrained better when fitting multiple categories simultaneously.

### 5.5.1. Kinematic correction to loose b-tagged jets

When comparing the shape of relevant distributions in the low and high b-tag regions, e.g. *CR* and *SR*, a disagreement is observable, due to kinematic differences introduced by the additional b-jets used in the low b-tag region. A correction, that is applied to the low b-tag region when obtaining the multijet shape, is derived to reduce these differences.

This correction is based on kinematic differences of jets passing the `Medium` b-tag working point and jets that pass the `Loose` but not the `Medium` b-tag working point and is derived for simulation and data.
The differences are parametrized in $p_{\mathrm{T}}$, $\eta$ and the minimum distance from the two jets with highes b-tag discriminant value $\Delta R_{\mathrm{min}}$ of a given jet.
These three observables are shown in figure 5.6 in $\mathrm{t\bar{t}}$ + Jet simulation for the two types of jets and similar difference can be observed in data. The correction is derived independently for each of the three data-taking periods because of small differences in the b-tagging algorithm performance[12]. It is denoted as $\mathrm{TF}_{\mathrm{loose}}$ and derived in the baseline analysis region. All jets passing at least the `Loose` working point are considered but the two jets with the largest b-tagging discriminant value are excluded, since two `Medium` b-tagged jets are present in any region and the $\Delta R_{\mathrm{min}}$ is defined with respect to these to jets.

The correction is derived by fitting one dimensional functions for $p_{\mathrm{T}}$, $\eta$ and $\Delta R_{\mathrm{min}}$ to the ratio of the kinematicc distributions between the two types of jets. The functions, with parameters $\alpha$, are defined as follows:

$$p_{\mathrm{T}} : f(x) = \alpha_0 + \mathrm{erf}(\alpha_1(x - \alpha_2))(\alpha_3 + \alpha_4 x) \tag{5.2}$$

$$\eta : g(x) = \sum_{i=0}^{16} \alpha_i x^i \tag{5.3}$$

$$\Delta R_{\mathrm{min}} : h(x) = \sum_{i=0}^{6} \alpha_i x^i \tag{5.4}$$

The correction factor is calculated using

$$\mathrm{TF}_{\mathrm{loose}} = \prod_j f(p_{\mathrm{T},j}) \cdot g(\eta_j) \cdot h(\Delta R_{\mathrm{min},j}) \tag{5.5}$$

with the previously fitted functions $f(x)$, $g(x)$ and $h(x)$, which are evaluated for all additional `Loose` but not `Medium` b-tagged jets $j$ in an events. $\mathrm{TF}_{\mathrm{loose}}$ is only calculated and applied in the low b-tag region (and the training region).

Since a slight dependency between the $\eta$ and $\Delta R_{\mathrm{min}}$ corrections was observed, a systematic uncertainty on the method is derived by applying a second $\eta$-based correction to the $p_{\mathrm{T}}$-$\eta$-$\Delta R_{\mathrm{min}}$ corrected distribution. This additional correction is considered the "up"-variation while the nominal value is used as down variation. The following function is used for the

[12] While the b-tagging algorithm is the same in all data-taking periods, the performance of the algorithm also depends on the detector performance and other machine parameters.

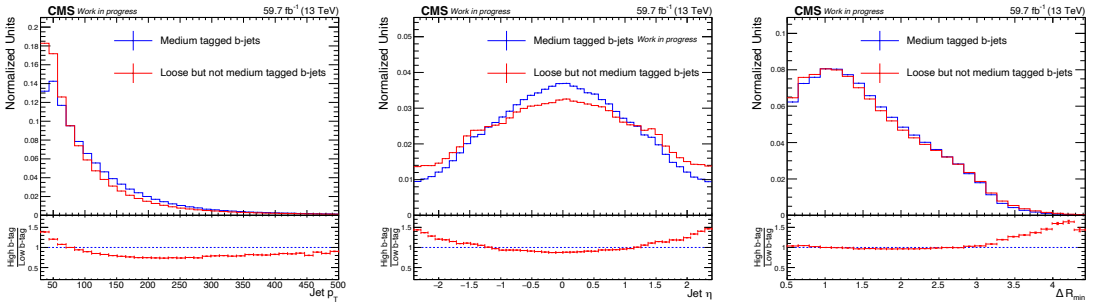



Figure 5.6.: Distributions of the $p_{\mathrm{T}}$, $\eta$ and $\Delta R_{\mathrm{min}}$ of jets passing the `Medium` b-tag working point (blue) and jets passing the `Loose` but not the `Medium` working point (red). The two leading jets, by b-tag discriminant value are not included in the blue histogram. Shown for the $\mathrm{t\bar{t}}$ + Jets simulation in the dataset for the 2018 data-taking period.

second $\eta$ correction:

$$\eta : g_2(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \alpha_3 x^3 + \alpha_4 x^4 \tag{5.6}$$

Figure 5.7 shows the ratios for data and $\mathrm{t\bar{t}}$ + Jets simulation after each correction step. The fitted functions $f(x)$, $g(x)$ and $h(x)$ are shown for the distribution they are fitted in. The correction derived in the $\mathrm{t\bar{t}}$ + Jets sample is used in all simulated background samples. This approach is justified given the small contribution of the minor backgrounds in the low b-tag regions and assumes that they are more similar to $\mathrm{t\bar{t}}$ + Jets than QCD multijet. Figure 5.8 shows how the $\mathrm{TF_{loose}}$ correction improves the agreement by example of the $p_{\mathrm{T}}$-$\eta$-plane in the $\mathrm{t\bar{t}}$ + jets dataset.

In practice the correction is used as follows: When calculating the distributions for a given observable in the *CR*, *ValCR*, *TR* and *ValTR* regions, the $\mathrm{TF_{loose}}$ correction is evaluated per events for each additional `Loose` b-tagged jet. This is done for both data and simulation with the respective parameters for the functions $f$, $g$, $h$ and $g_2$.

### 5.5.2. Validation of the background estimation method

Since the $\mathrm{t\bar{t}H(b\bar{b})}$ analysis is executed as a blinded analysis, the agreement between the background model and the data cannot be validated in the signal region before unblinding, which is the last and final step in a blinded analysis.
Simulated backgrounds are validated in regions with small signal contribution like the baseline region or other control regions. If the agreement between the simulation and the data is good, it is then assumed that the modeling of the simulation is also sufficiently good in the signal region. For the data-driven background estimation, the same principle is used but the region has to be chosen such that the transfer from the low to the high b-tag region is valid. The validation region (or $m_{\mathrm{qq}}$ sideband) is used for this step. The *VR* exhibits the same b-tag selection but can be unblinded since it will not be used for the signal extraction. The result for this validation is shown in figure 5.9 for events with 7, 8 and $\geq 9$ jets and $\geq 4$ b-tags in the $m_{\mathrm{qq}}$ sideband for each of the three data-taking periods. Figure 5.9 shows the distribution of the final discriminant, which will be described in detail in section 5.6, since ultimately, the agreement has to be good in the distributions used in the final fit to data. The method is also validated for the distributions, that are used to derive the discriminant. A selection of distributions with high impact on the final discriminant, is given in figure 5.10. A good agreement between data, simulated backgrounds and the QCD background estimation can be found in all relevant distributions.

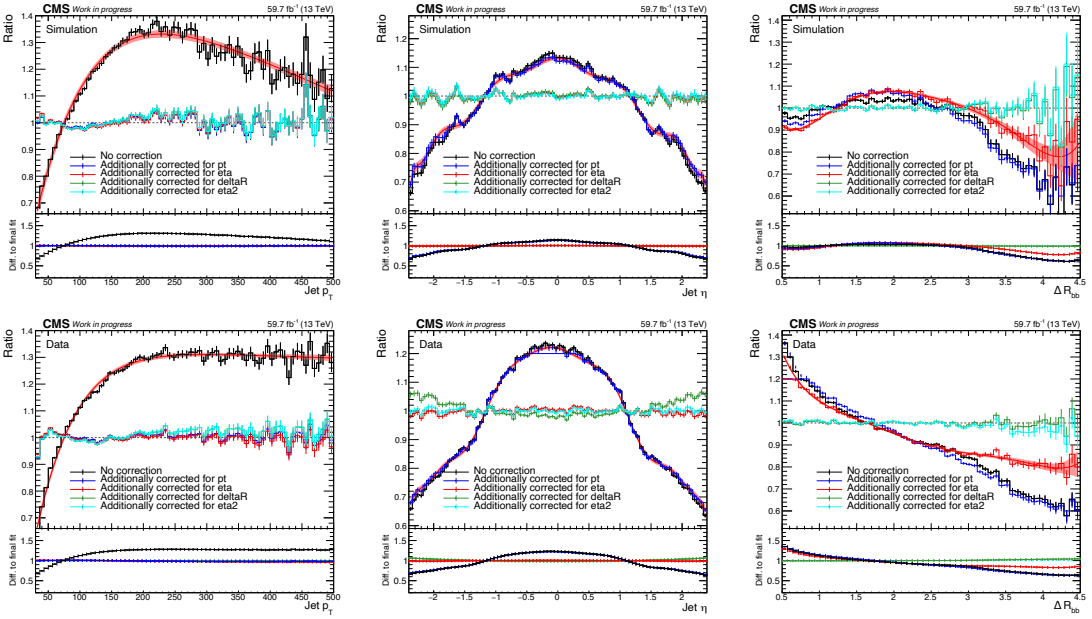



Figure 5.7.: Distribution of the ratio of jets with values greater or equal to the `Medium` b-tag working point and jets with values between the `Loose` and `Medium` working point as a function $p_{\mathrm{T}}$, $\eta$ and $\Delta R_{\mathrm{min}}$ of jets in the region after the baseline selection. The two jets with the highest b-tag value are excluded. The ratios are shown as histograms before applying any correction (black) and after sequentially applying the $p_{\mathrm{T}}$ (blue), $\eta$ (red), $\Delta R_{min}$ (green) and the second $\eta$ (cyan) corrections. The fitted functions $f(p_{\mathrm{T}})$, $g(\eta)$ and $h(\Delta R_{\mathrm{min}})$ are shown as red line with uncertainty band in the $p_{\mathrm{T}}$, $\eta$ and $\Delta R_{\mathrm{min}}$ distributions, respectively. The distributions are shown in the 2018 datasets in $\mathrm{t\bar{t}}$ + Jets simulation (top) and data (bottom).

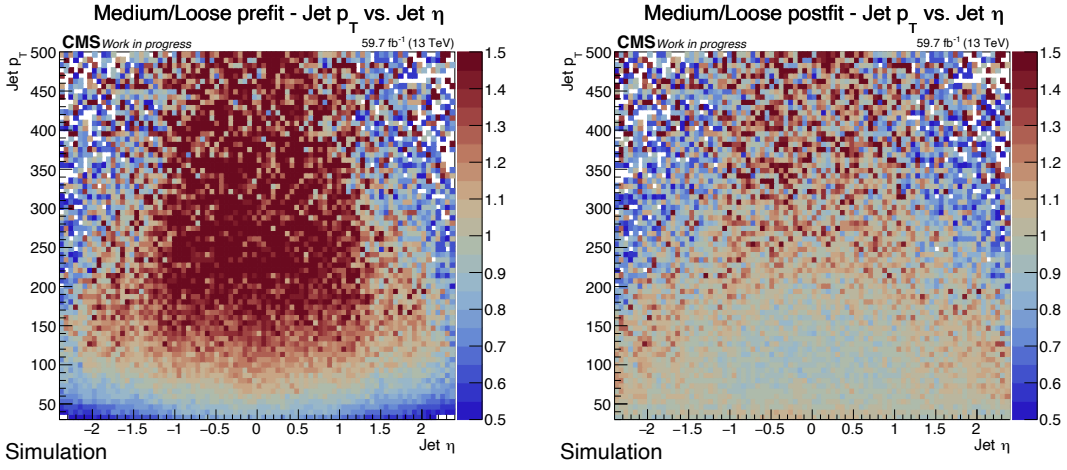



Figure 5.8.: Distribution of the ratio of jets with values greater or equal to the `Medium` b-tag working point and jets with values between the `Loose` and `Medium` working point in the $p_{\mathrm{T}}$-$\eta$ plane before (left) and after (right) applying the nominal correction. The distributions are shown in the 2018 datasets in $\mathrm{t\bar{t}}$ + Jets simulation. Ideally, a uniform distribution is targeted.

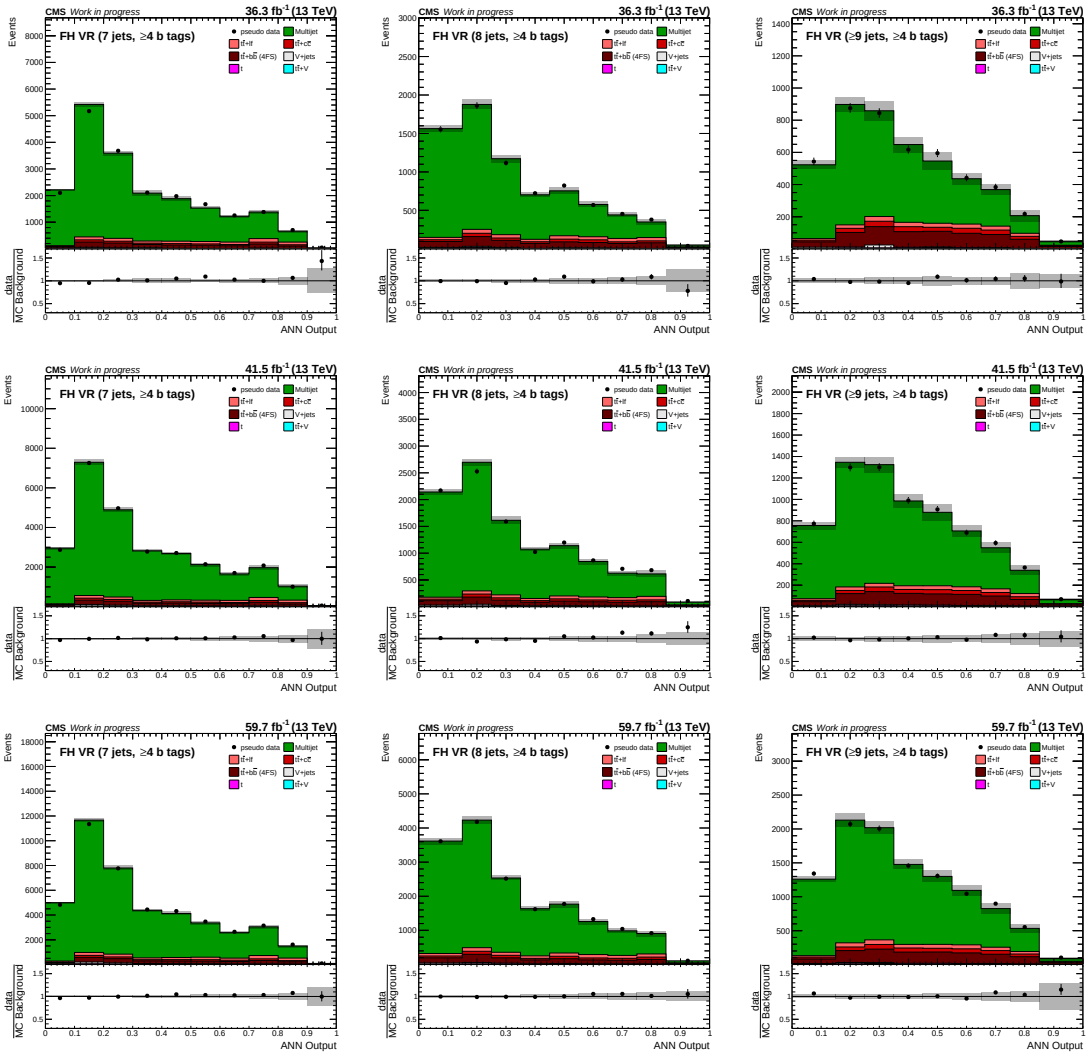
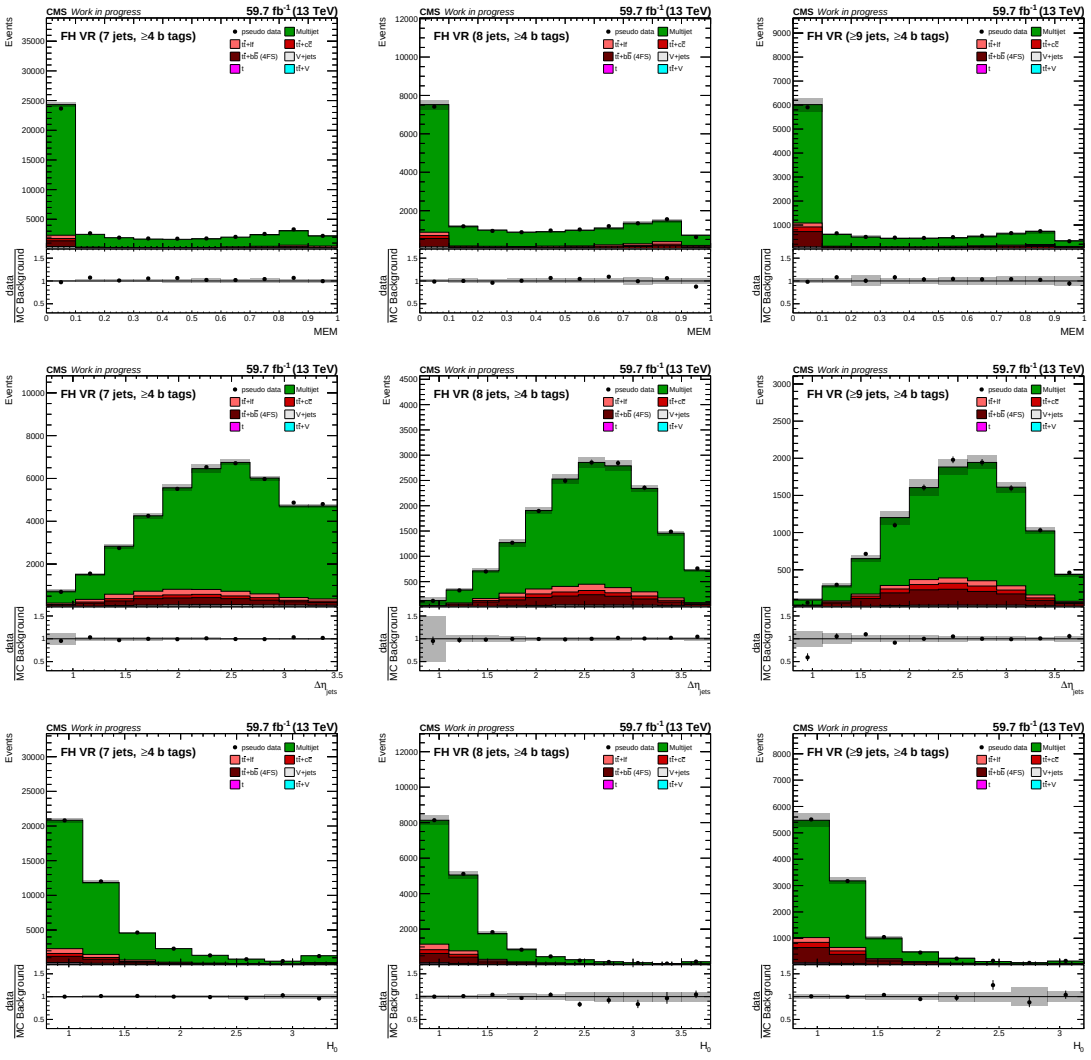



Figure 5.9.: Distribution of the final discriminant used in the fully-hadronic channel for data, simulated backgrounds and the estimated QCD background in the VR. The 7 jet category is shown on the left, the 8 jet category is shown in the middle and the $\geq 9$ jet category is shown on the right. The top row shows the 2016 dataset, the middle row the 2017 dataset and the bottom row shows the 2018 dataset. The uncertainty band in this plot shows the full uncertainty model, excluding the UE and HDAMP uncertainties, with 50% rate uncertainty on $t\bar{t} + B$ and $t\bar{t} + C$, before the fit to data (prefit).




(a) ($7\,\text{jets}, \geq 4\,\text{b-tags}$) (b) ($8\,\text{jets}, \geq 4\,\text{b-tags}$) (c) ($\geq 9\,\text{jets}, \geq 4\,\text{b-tags}$)

Figure 5.10.: Distributions with high impact on the final discriminant in the 2018 dataset in the validation region ($VR$) for each of the jet/tag categories. For events with 7 jets and $\geq$ 4 b-tags (a) the MEM discriminant, $\Delta\eta_{\text{jets}}$ and $H_0$ are shown, for events with 8 jets and $\geq$ 4 b-tags (b) the MEM discriminant, $\Delta\eta_{\text{jets}}$ and $m_{\text{b,b}}^{\text{max}\,m}$ are shown and for events with $\geq 9$ jets and $\geq 4$ b-tags (c) the MEM discriminant, $R_2$ and $H_0$ are shown (see section 5.6.2 for definition of these variables).

Table 5.4.: Sideband regions for the validation of the QCD multijet background estimation method which are defined for each of the jet multiplicity categories. The values in parenthesis for the $m_{\mathrm{qq}}$ variable are used for event with $\geq 9$ jets. `DeepJetM`, `DeepJetML` and `DeepJetL` refer to the multiplicity of jets passing the `Medium`, `MediumLoose` and `Loose` working-points of the `DeepJet` b-tagging algorithm, respectively.

| | $60(72)\,\mathrm{GeV} < m_{\mathrm{qq}} < 100(90)\,\mathrm{GeV}$ | $30\,\mathrm{GeV} < m_{\mathrm{qq}} < 60(72)\,\mathrm{GeV}$ or $100(90)\,\mathrm{GeV} < m_{\mathrm{qq}} < 250\,\mathrm{GeV}$ |
|---|---|---|
| `DeepJetM` $= 3$<br>`DeepJetML` $= 3$<br>`DeepJetL` $\geq 4$ | ***CRx*** | ***ValCRx*** |
| `DeepJetM` $= 3$<br>`DeepJetML` $\geq 4$<br>`DeepJetL` $\geq 4$ | ***SRx*** | ***VRx*** |

An additional validation, which has to goal of validating the step from the low to the high b-tag region in the central $m_{\mathrm{qq}}$ band, is performed. For this, the low b-tag region (*CR+ValCR*) is split into four regions by defining an additional b-tagging working point between `Loose` and `Medium`, that splits the low b-tag region in two equal parts.
The new b-tag working point is called `MediumLoose` with discriminator values of 0.112, 0.101 and 0.093 in the 2016, 2017 and 2018 data-taking periods, respectively. Events passing the *CR+ValCR* selection and additionally exhibit three jets passing this new working point define the low b-tag region for this validation. If the event exhibit four or more jets passing this new working point they define the high b-tag region. The new regions are denote as *CRx* and *SRx* (previously *CR*) and *ValCRx* and *VRx* (previously *ValCR*). In order to properly test the method in this sideband, a special $\mathrm{TF}_{\mathrm{loose}}$ correction has to be derived. It is used to correction jets passing the `Loose` but not the `Mediumloose` working point to approximate jets passing the `Mediumloose` but not the `Medium` working point. This correction is derived with the exact same procedure as used for the main method. The only difference are the new low and high regions. The regions and the selection criteria are also given in table 5.4.
Figure 5.11 shows the final discriminant in the sideband regions *VRx* and *SRx*, by the example of the 2018 data-taking period. Similar agreement can be observed in the other data-taking periods. The agreement between data, simulated backgrounds and the QCD background estimation in *VRx* is comparable to *VR*, as seen by comparing to figure 5.9. The agreement in the *SRx* is also good.

From these validations it can be concluded that the background estimation method, described in the section, yields a good estimate of the QCD multijet background shape, that can be used in the final fit.

## 5.6. Signal identification

Since the $\mathrm{t\bar{t}H(b\bar{b})}$ process is very rare and to increase the statistical power of the overall analysis, identifying the signal and separating it from the backgrounds plays an integral role in the design of the analysis strategy. This section covers the dedicated neural-network-based classifier used in the $\mathrm{t\bar{t}H(b\bar{b})}$ fully-hadronic analysis. The architecture and

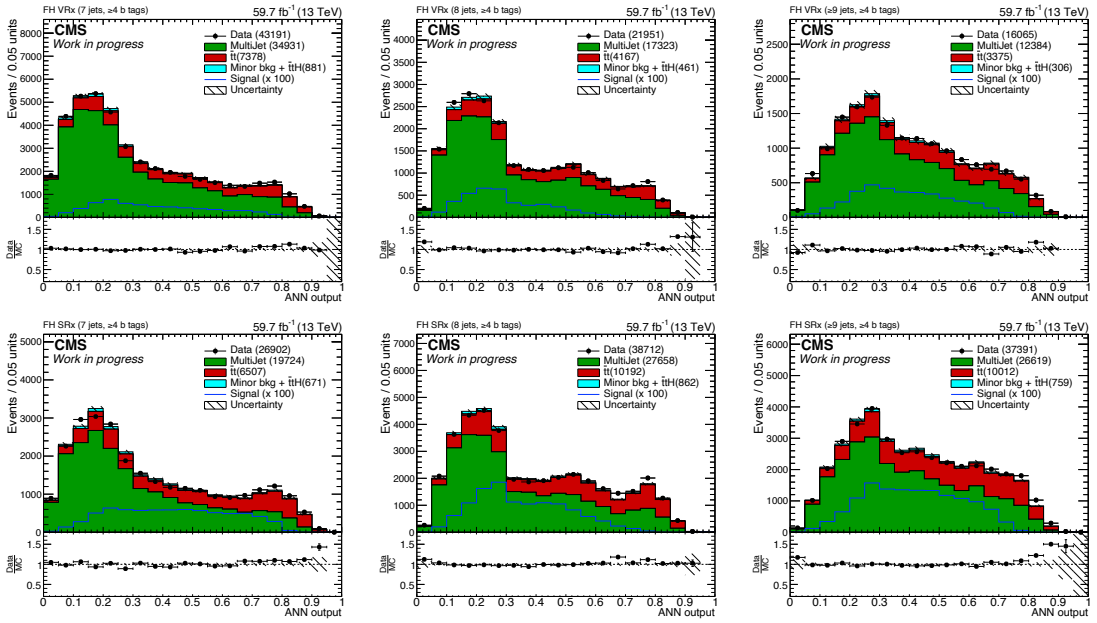



Figure 5.11.: Distribution of the final discriminant used in the fully-hadronic channel for data, simulated backgrounds and the estimated QCD background in the side-band validation regions *VRx* (top) and *SRx* (bottom) for the 2018 data-taking period in each of the jet/b-tag categories. The hashed error band shows the statistical uncertainty only.

training process of the neural network are discussed in 5.6.1. The feature selection and validation is described in section 5.6.2 and the performance of the neural-network classifier is evaluated in section 5.6.3. This neural network classifier will be referred to as Artificial Neural Network (ANN) or ANN classifier in the following.

This section builds on the foundations on neural networks established in section 3.2. The checks and performance evaluations in section 5.6.3 represent the state after developing the method and before unblinding. Hence, no data or multijet estimation in the *SR* is shown in any of the figures. Since the method has be finalized and frozen before unblinding, this plays no role in the description and development of the method.

### 5.6.1. ANN architecture and training

The ANN is implemented as a *fully-connected feed-forward* neural network with three *hidden layers* of 100 *nodes* each using `KERAS` [163] `2.2.4` with the `TensorFlow` [119] `1.13.11` backend. In the training, the *binary cross-entropy* is used as *cost function* and the *ReLu function* is used as *activation function* for all nodes except the output node, where the *sigmoid function* is used. In order to reduce overtraining and increase stability, *weight decay* as well as *dropout* is used and the training is stopped early if the loss function stays at a minimum for a number of epochs. The terms written in cursive are defined in section 3.2.

The ANN is trained to separate $t\bar{t}H(b\bar{b})$ and QCD multijet events. This means, that this ANN is a so-called binary-classifier and has a single output node. For each event an output value between zero and one is calculated. For signal events, the output node value should be high, while it should be low for background events. All values between zero and one are possible, depending how "signal-like" an event is considered by the ANN. Combing these values in a distribution, usually referred to as the node output or discriminator output,

yields a discriminant that can be used in the fit to data to extract the signal.[13]
During the development of the ANN for the fully-hadronic channel, a categorical approach, where multiple output nodes per physics process are used in the training, was investigated. For this case, an ANN was trained to separate between $t\bar{t}H(b\bar{b})$, $t\bar{t}$ + Jets and QCD multijet events. This approach would have the advantage of adding separation power against the $t\bar{t}$ + Jets background to the classifier. In the binary approach, no significant separation power against $t\bar{t}$ + Jet background is expected since these events lie between $t\bar{t}H(b\bar{b})$ and QCD multijet events in the space of the input features [14]. Using the categorical approach, an improvement, with respect to the binary approach, could be observed in the ROC curves for the $t\bar{t}$ + Jets background but no significant impact on Asimov fits (fits to pseudo data) was found. Due to these findings, the binary approach was chosen for the analysis since a better stability is expected and it is less constrained by statistics[15].

The QCD multijet sample, that is used for the training of the ANN, is estimated from data and not taken from simulation. For this, the region with 2 `Medium` b-tagged jets and four or more `Loose` b-tagged jets is used. In the central $m_{qq}$ band, this region is called *training region* (*TR*) and in the $m_{qq}$-sidebands it is called *validation training region* (*ValTR*). The $TF_{loose}$ correction, discussed in section 5.5.1, is applied since the ANN is intended to separate $t\bar{t}H(b\bar{b})$ and QCD multijet in the *SR*. The *TR* can be used to model the QCD multijet background in the SR with the same arguments presented in section 5.5 for the *CR*[16]. It was estimated that the *TR* exhibits a QCD multijet contribution of $\geq$ 80% with $t\bar{t}$ + LF as second largest contributions. The *SR* events from $t\bar{t}H(b\bar{b})$ simulation are used as the signal sample during training. To not bias the fit, the $t\bar{t}H(b\bar{b})$ sample is split into two equal parts, according to the CMS-internal event-number. One half is used for training and the other half for modeling the signal in the fit.

Unlike in all previously described procedures and methods, the three data-taking periods are not considered separately for the training of the ANN. Instead it is trained on a combination of all data-taking periods in order to increase training statistics and improve the stability of the classifier. The $t\bar{t}H(b\bar{b})$ samples are weighted according to their relative contribution to the full dataset by luminosity. This approach was validated by comparing the shapes of all input features between the datasets for the different data-taking periods and for the signal and background samples. A selection of these comparisons is shown in figure 5.12. Before training, the combined datasets are randomly shuffled in order to not bias the training towards are certain data-taking period.

The signal and background dataset are split into three subsets: A "training" set (60 %) for the training of the ANN, a "test" set (20 %) for immediate validation of the training

[13] The common use-case (outside HEP) would be to define some working point, which is then used to label the event as one of the two classes. This working point is usually calculated based on a trade off between true-positive and false-positive rates. Since the node output distribution itself is very powerful in the context of the fit to data, this approach is not chosen if the classifier is used to construct the final discriminant.

[14] Since $t\bar{t}$ + LF is closer to QCD and $t\bar{t}$ + B closer to $t\bar{t}H(b\bar{b})$ (in the space of the input features), the separation power varies between the $t\bar{t}$ + Jets subprocesses.

[15] The usual way to exploit a categorical ANN (in particle physics) is to apply a softmax activation for the output nodes and categorize events according the largest output node value. This results in multiple categories, that are enriched in the different processes. In context of the fully-hadronic analysis, this would have multiple drawbacks: While the statistical uncertainty is not dominant, it is still a significant factor in the fit (especially in the category with $\geq$ 9 jets and signal-like bins of the distribution). Therefore, splitting each category in multiple subcategory would lead to a reduced statistical power in the individual categories. In addition (and more significantly) this would increase the impact of the uncertainty on the QCD multijet estimation in the fit, since the freely floating normalization is not correlated between the categories.

[16] In previous iterations of this analysis [10, 11], the events with 2 `Medium` b-tagged and $\geq$ 4 `Loose` b-tagged jets were used as *CR* for the background estimation.

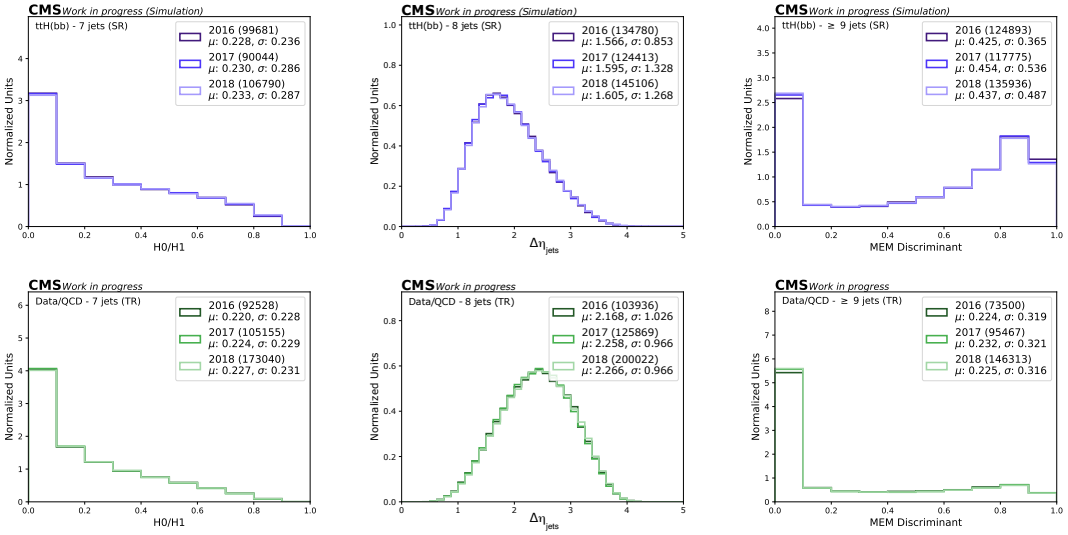



Figure 5.12.: Comparison of input features between the data-taking periods in the signal sample (top) and the background sample (bottom). The variable $R_1$ is shown for the ($7\,\text{jets}, \geq 4\,\text{b-tags}$) category (left), $\Delta\eta_{\text{jets}}$ is shown for ($8\,\text{jets}, \geq 4\,\text{b-tags}$) category (middle) and the MEM discriminant is shown for the ($\geq 9\,\text{jets}, \geq 4\,\text{b-tags}$) category (right). See section 5.6.2 for the definition of these variables. No significant shape dependency on the data-taking period is present.

during the training process, and a "validation" set (20 %) for evaluating the performance of the trained ANN. The "training" sets are weighted such that the signal and background processes exhibit the same weight. This is done in order to avoid that the network learns the relative frequency between signal and background in the training set.
Independent ANN classifiers are trained for each of the three *SR* categories. The input features, which will be covered in section 5.6.2, and the hyperparameters, a term used to describe the parameters of a neural network like number of hidden layers or the lost function, are optimized per category. The final set of hyperparameters used in this analysis are given in table 5.5. The training is done over a maximum 500 epochs and the `ADAM` [164] minimizer is used to minimize the binary cross-entropy, which is used as the loss function. The number of input features per category is determined by training the classifier and removing variables that do not contribute significantly to the overall performance of the network. Furthermore, features are removed if they do no pass a validation steps, that will be described in the next section.

### 5.6.2. Input features

The choice of input features plays a singular role in the design of neural-network-based classifier like it is used in the analysis[17]. In the case of the fully-hadronic $t\bar{t}H(b\bar{b})$ analysis, the following validation steps are performed for all features that were considered:

[17] The typical strategy for designing multivariate classifiers in high-energy physics is to use observables with a strong separation power, also called "hight-level" features in the context of multivariate classifiers. With this approach, the multivariate classifier can exploit the present separation power and use the correlation of the features to improve to overall separation power. When using techniques like BDTs for complex problems like a physics analyses, this approach is preferred due to the limited encoding power of the method. With resent developments in neural networks it can be a viable approach to instead use very basic, or "low-level", features, like the object 4-vectors and rely on the network to encode how these "low-level" features can be connected to useful "high-level" features. The latter approach especially excels with very complex architectures and large training statistics.

Table 5.5.: Hyperparamters of the ANN classifiers for the 7, 8 and $\geq 9$ *SR* categories. If only a single value is given it is common for all categories. The parameters and their function are described in section 3.2

| | $(7\,\text{jets}, \geq 4\,\text{b-tags})$ | $(8\,\text{jets}, \geq 4\,\text{b-tags})$ | $(\geq 9\,\text{jets}, \geq 4\,\text{b-tags})$ |
|---|---|---|---|
| Optimizer | | `ADAM`$(10^{-4})$ | |
| Loss function | | Binary cross-entropy | |
| Number of input variables | 19 | 21 | 21 |
| Hidden Layers | | 3 | |
| Nodes per Hidden Layer | | 100 | |
| Node Activation | | ReLu | |
| Output Node activation | | Sigmoid | |
| L2 regularisation | $10^{-4}$ | $5 \cdot 10^{-4}$ | $10^{-4}$ |
| Dropout Percentage | 0.5 | 0.6 | 0.5 |
| early stopping min epochs | 30 | 30 | 30 |
| batchsize | 5000 | 20000 | 5000 |

Since the QCD multijet dataset is taken from the *TR* with same jet multiplicity but different b-tag properties than the $t\bar{t}H(b\bar{b})$ dataset (2 `Medium` tags and $\geq 4$ `Loose` tags vs. $\geq 4$ `Medium` tags), it is important that there is as little correlation as possible between the input features and the variables used for the region definition. Hence, as a first validation step, the input features should exhibit similar shapes in *TR*/*ValTR*, *CR*/*ValCR* and *SR*/*VR* [18]. This mainly excluded features that directly depend on the b-tagging. These features were initially considered because they are generally very powerful. One example would be the average b-tag discriminator value of all jets. It is expected to be larger in $t\bar{t}H(b\bar{b})$ than in most backgrounds. The differences between the regions mainly result from the fact, that the jet with the fourth highest b-tag discriminator value always exhibits a smaller value in the low b-tag region by definition. Generally, good agreement between *TR* and *SR* shapes in central-$m_{\text{qq}}$ band and the sidebands is required for all features. Figure 5.13 shows the agreement across the regions for the signal and background samples for a selection of features. As a second validation step, the agreement between data and the background expectation under the full uncertainty model, which will be discussed in section 6.2, was considered via a $\chi^2$ criterion. All features with a small $\chi^2$/ndof were excluded from the training[19]. Figures for all input variables, which are used in the final ANN discriminator, are given in appendix C for all data-taking periods.

The following variables passed the discussed validation steps and are used for the training of the ANN classifier in the fully-hadronic $t\bar{t}H(b\bar{b})$ analysis in at least on category:

**MEM** A likelihood based MEM discriminator is derived from the MEM probability density $\omega_i$ with i = S ($t\bar{t}H$) and B ($t\bar{t}+b\bar{b}$),

[18] The *CR*/*ValCR* are important since the ANN will be evaluated in these region when estimating the QCD multijet background for the *SR*. In practice this is never a problem since this region is "between" the *TR*/*ValTR* and *SR*/*VR* and if a variable agrees between these two regions is also agrees in the *CR*/*ValCR*.

[19] No explicit value is given, since the agreement was either very good (p-value close to one) or very bad (p-values $< 0.1$) in all cases.

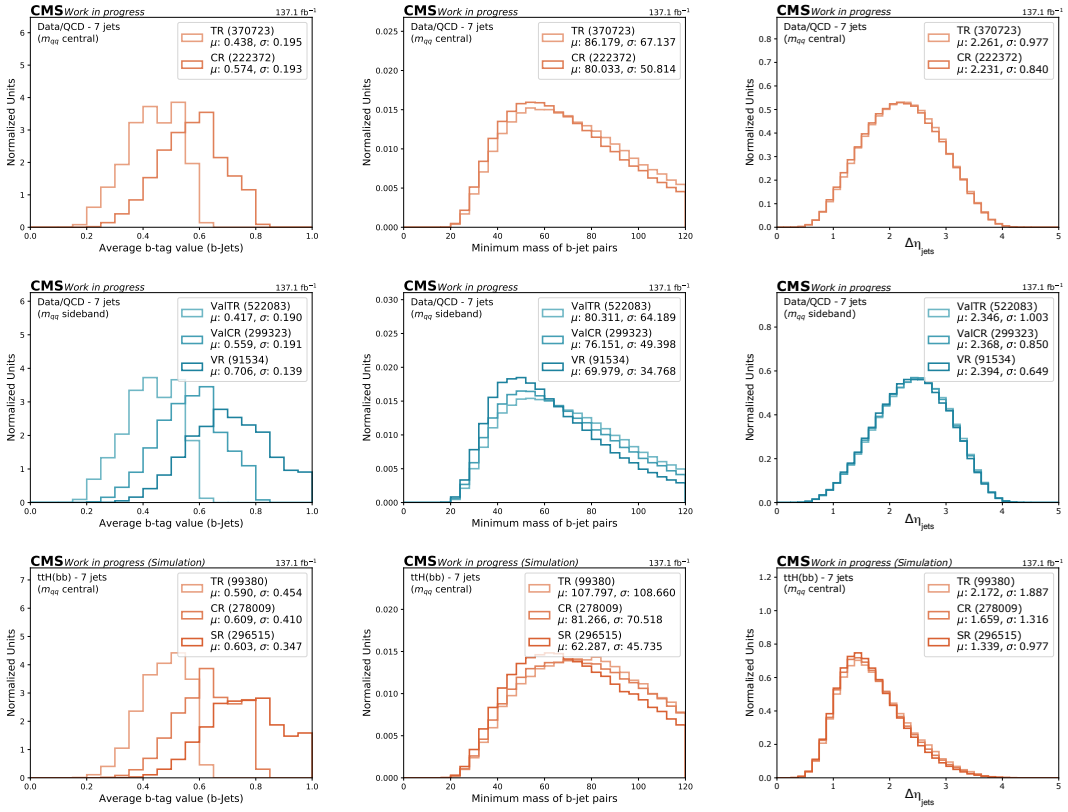



Figure 5.13.: Selection of input features for the ANN for the category with seven jets. This figure shows examples for the initial validation step in the ANN feature selection process, which aims to remove all variables that show different shapes in the analysis regions (*SR*, *TR* etc.). The average b-tag discriminator value (left), the minimum mass of b-jet pairs (middle) and the $\Delta\eta_{\text{jets}}$ (right) variables are shown. The top row shows *TR* and *CR* for the background process estimated from data. The middle row shows *ValTR*, *ValCR* and *VR* for the background process estimated from data. The bottom row shows *TR*, *CR* and *SR* for the simulated $\mathrm{t\bar{t}H(b\bar{b})}$ signal. The training will be carried out with TR Data/QCD (top row, light orange distribution) and the SR $\mathrm{t\bar{t}H(b\bar{b})}$ (bottom row, darkest orange distribution). The average b-tagging value and the minimum mass of b-jet pairs are discarded for the further analysis since the shapes do not agree well between the regions.

as described in section 3.1. It is defined as

$$\mathrm{MEM} = \frac{\omega_{\mathrm{S}}(\boldsymbol{y})}{\omega_{\mathrm{S}}(\boldsymbol{y}) + \kappa \cdot \omega_{\mathrm{B}}(\boldsymbol{y})} \tag{5.7}$$

with $\kappa = 0.065$[20] and the per-event observables $\boldsymbol{y}$.

**$\Delta\eta_{\mathrm{jets}}$** The $\Delta\eta_{\mathrm{jets}}$ variable is defined by taking the average $\Delta\eta$ between each jet and the jet furthest apart from it in $\eta$. This variable is a measure of the spacial distribution of the jets in an event and exhibits a strong discrimination power between signal and QCD multijet.

**$\mathrm{H}_i$ and $\mathrm{R}_i$** The Fox-Wolfram moments [165] $\mathrm{H}_i$ are used to describe the energy distribution in the event in a series of rotationally invariant functions using the jet momenta and energies.

$$\mathrm{H}_i = \sum_{i,j} \frac{|\boldsymbol{p}_i||\boldsymbol{p}_j|}{E_{\mathrm{tot}}^2} P_l\left(\cos(\phi(\boldsymbol{p}_i, \boldsymbol{p}_j))\right) \tag{5.8}$$

with the Legendre polynomials $P_l$ and the angle $\phi(\boldsymbol{p}_i, \boldsymbol{p}_j)$ between the jet momenta. The sum is evaluated over all jet pairs. $\mathrm{R}_i$ is defined as $\mathrm{H}_i/\mathrm{H}_0$. In this analysis $i$ = 0,...,5 is used.

**Sphericity/Aplanarity** The Sphericity $S$ and Aplanarity $A$ [166] are calculated from the eigenvalues ($\lambda_1 > \lambda_2 > \lambda_3$) of the energy-momentum tensor of the jets:

$$S = 3/2(\lambda_2 + \lambda_3) \tag{5.9}$$
$$A = 3/2\lambda_1 \tag{5.10}$$

**Jet $p_{\mathrm{T}}$ and $\eta$** Various jet $p_{\mathrm{T}}$ and $\eta$ values where tested but ultimately only the 7$^{\mathrm{th}}$ highest jet $p_{\mathrm{T}}$ in an event, and the $\eta$ of the jet with the highest as well as the $p_{\mathrm{T}}$ and $\eta$ of the jet with the second highest b-tag discriminator value are used as input features.

**Dijet observables** Multiple dijet observables are used in the ANN discriminant. Considering dijet pairs of all selected jets, the average, maximum and minimum $\Delta R$ ($\Delta R_{\mathrm{j,j}}^{\mathrm{avg}}$, $\Delta R_{\mathrm{j,j}}^{\mathrm{max}}$ and $\Delta R_{\mathrm{j,j}}^{\mathrm{min}}$) and the mass of the dijet pair closest to 125 GeV $m_{\mathrm{j,j}}^{\mathrm{closest\ to\ 125}}$ are used. Furthermore, the mass of the b-jet pair with the highest mass $m_{\mathrm{b,b}}^{\mathrm{max}\ m}$ is used. Other similar variables were considered but are not used in the ANN due to either failing validation or not contributing much to the ANN performance.

Not all features are used in each of the three jet-multiplicity categories. Which features are used in which category is listed in table 5.6.

For the training, the variables are transformed in order the make the distributions compatible with the activations functions used in the training of the ANN. From the full

[20] The $\kappa$ factor was optimized per category in previous iterations of this analysis [10, 11], in which the MEM was used as final discriminant. In the context of this analysis and the usage as input feature for the ANN, the choice of $\kappa$ plays not significant role.

Table 5.6.: ANN input features for the classifiers trained for the 7, 8 and $\geq 9$ *SR* categories. + indicates a feature is used while − indicated as is not used in a certain category. The Variables are defined in section 5.6.2.

| Input variable | (7 jets, $\geq 4$ b-tags) | (8 jets, $\geq 4$ b-tags) | ($\geq 9$ jets, $\geq 4$ b-tags) |
|---|---|---|---|
| MEM | + | + | + |
| $\Delta\eta_{\text{jets}}$ | + | + | + |
| $H_0$ | + | + | + |
| $H_1$ | + | + | + |
| $H_2$ | + | + | + |
| $H_3$ | + | + | + |
| $H_4$ | + | + | + |
| $H_5$ | + | + | + |
| $R_1$ | + | + | + |
| $R_2$ | + | + | + |
| $R_3$ | + | + | + |
| $R_4$ | + | + | + |
| Aplanarity | + | + | + |
| Sphericity | + | + | + |
| $\eta$ of $1^{st}$ jet by b-tag | + | + | + |
| $\eta$ of $2^{nd}$ jet by b-tag | + | + | + |
| $p_{\text{T}}$ of $2^{nd}$ jet by b-tag | − | + | + |
| $p_{\text{T}}$ of $7^{\text{th}}$ jet | − | − | + |
| $\Delta R_{\text{j,j}}^{\text{avg}}$ | − | + | + |
| $\Delta R_{\text{j,j}}^{\text{max}}$ | + | + | + |
| $\Delta R_{\text{j,j}}^{\text{min}}$ | + | + | − |
| $m_{\text{j,j}}^{\text{closest to 125}}$ | + | − | − |
| $m_{\text{b,b}}^{\text{max } m}$ | − | + | + |

distribution of an input variable $i$ in the trainings dataset the mean $\mu_i$ and the standard deviation $\sigma_i$ is calculated. Each value $x_i$ is then transformed according to

$$\hat{x}_i = \frac{x_i - \mu_i}{\sigma_i}. \tag{5.11}$$

The $\mu_i$ and $\sigma_i$ values, calculated during training, are saved and applied to the input features when evaluating the ANN in the analysis.

### 5.6.3. Performance

The performance of the training is evaluated in a randomly selected subset (20 %), also called the "validation" dataset, of the full dataset used for training. The separation power is evaluated in a first step by checking the ROC curves, which are shown in figure 5.14. As a reference, the MEM discriminant is shown since it was used as the final discriminant in the previous analyses of the fully-hadronic channel [11] and [10], while being used as an input feature in this analysis. The separation power against QCD multijet cannot be evaluated in the SR directly at development time since the analysis was still blinded. Therefore, the ROC curves in the *VR* are shown. In both cases it can be seen that the ANN exhibits an improved separation power compared to the MEM (larger AUC). The ROC curves in the *VR* further show, that the training with data from the *TR* results in good separation between the QCD multijet estimation from data and $t\bar{t}H(b\bar{b})$ simulation in a region close to the region which will be used in the final fit to data.
Distributions of the node output value are shown in figure 5.15 for the "training" and "validation" datasets used during the training processes of the ANNs. Additionally the distributions are shown in the *VR* for $t\bar{t}H(b\bar{b})$ and inclusive $t\bar{t}$ + jets simulations as well as QCD multijet estimated from data. These distributions clearly show, that the discrimination power can be transferred between the different analysis regions. In addition, the *VR* plots (red distributions in figure 5.15 in the bottom row) shows that no significant discrimination power, as discussed previously, can be observed for $t\bar{t}$ + jets events.

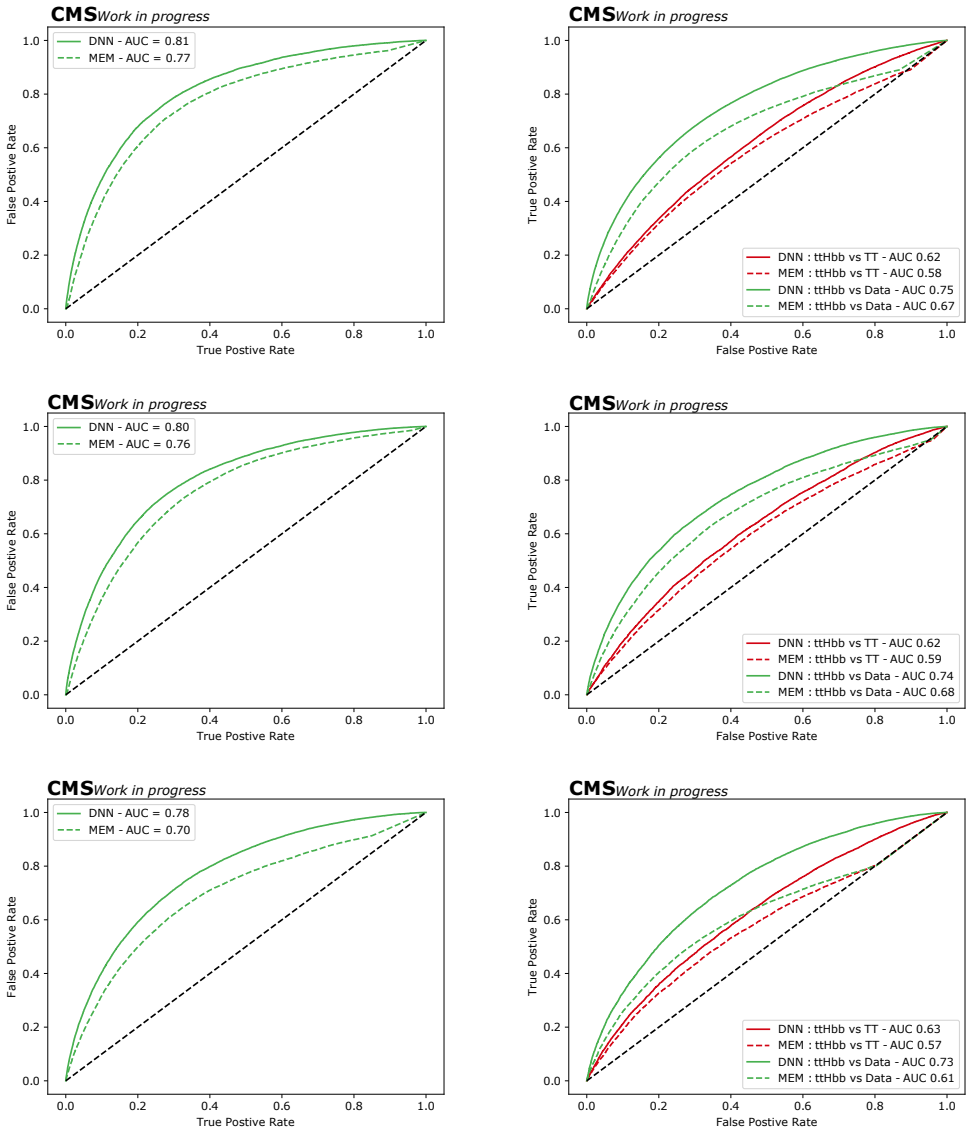
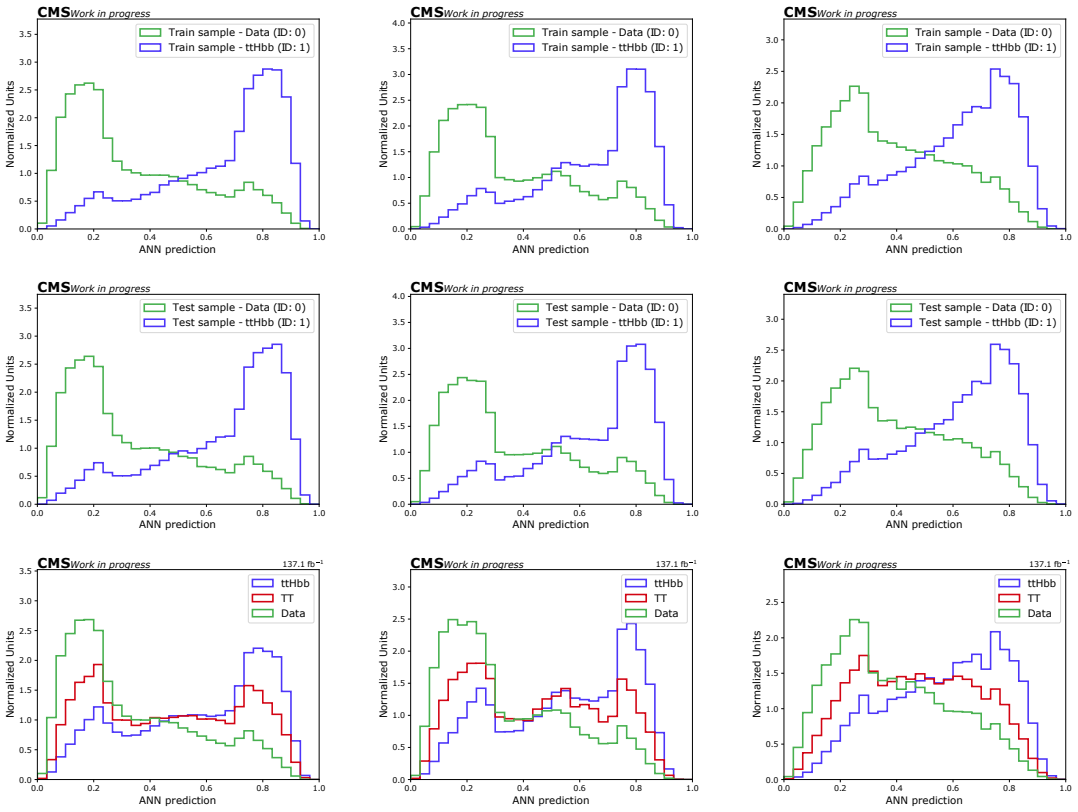



Figure 5.14.: ROC curve for the trained model for the FH categories with 7 jets (top), 8 jets (middle) and $\geq$ 9 jets (bottom). The ROCs are shown for signal vs. background in the "validation" dataset on the left as well as in the *VR* for $t\bar{t}H(b\bar{b})$ vs. QCD Multijet estimated from data (green) and $t\bar{t}H(b\bar{b})$ vs. inclusive $t\bar{t}$ + jets simulation (red) on the right. As a comparison, ROCs for the MEM (dashed lines), which was used as final discriminant in previous analyses, is shown.




(a) 7 jets (b) 8 jets (c) $\geq 9$ jets

Figure 5.15.: Node outputs for signal and background in the "training" set (top) and the "validation" set (middle) as well as Node output on the *VR* for $t\bar{t}H(b\bar{b})$, QCD Multijet estimated from data and inclusive $t\bar{t}$ + jets simulation (bottom) for the FH categories with 7 jets (a), 8 jets (b) and $\geq 9$ jets (c). The plots show how the node output distributions looks for different processes.

# 6. Statistical Model and systematic uncertainties

## 6.1. Statistical model

For the measurement of the $t\bar{t}H$ signal strength parameter

$$\mu_{t\bar{t}H} = \sigma_{t\bar{t}H}/\sigma^{SM}_{t\bar{t}H} \cdot \frac{\mathcal{B}(H \to b\bar{b})}{\mathcal{B}_{SM}(H \to b\bar{b})},$$

with the measured cross section $\sigma_{t\bar{t}H}$ and branching fraction $\mathcal{B}(H \to b\bar{b})$ and the standard model predictions $\sigma^{SM}_{t\bar{t}H}$ and $\mathcal{B}_{SM}(H \to b\bar{b})$, the observed final discriminant distribution in data is analyzed with the statistical model described in section 3.3.

A binned maximum likelihood fit, as described in section 3.3.3, is performed simultaneously in multiple categories. Various combinations are considered: The combinations of all categories for a given channel, e.g. ($7\,\text{jets}, \geq 4\,\text{b-tags}$), ($8\,\text{jets}, \geq 4\,\text{b-tags}$) and ($\geq 9\,\text{jets}, \geq 4\,\text{b-tags}$) in the case of the fully-hadronic channel, as well as the combination of all $t\bar{t}H(b\bar{b})$ decay channels are of primary interest in the context of this thesis. The combined measurement is also referred to as the $t\bar{t}H(b\bar{b})$ Run-2 combination.

The expected distributions used in the fit, the so-called templates, are determined directly from simulation for the minor and $t\bar{t}$ + Jets backgrounds as well as the signal, or are estimated from data, in the case of the multijet background as described in section 5.5. All corrections, that are described in section 5.2, are included. If not otherwise indicated, a single parameter of interest (POI) for the signal strength $\mu_{t\bar{t}H}$ shared between all channels, categories, and data-taking periods. For some additional fits, that are mainly used to set the main result into context, multiple POI are considered. One example are combined fits in all channels, in which $\mu_{t\bar{t}H}$ is determined for each $t\bar{t}H(b\bar{b})$ decay channel separately. Since the correlation of the nuisance parameters is still present in such fits, they enables to estimate how much an individual channel profits form the combined fit in all channels. The tHq and tHW processes are considered to be background processes and are kept at the SM prediction. In the fit, statistical and systematic uncertainties are taken into account in the from of nuisance parameters, that can morph or scale the expectation in order to match the data template. Two distinct types of such uncertainties are considered:

**Shape uncertainties** effect the shape and rate of a process, are implemented as Gaussian constrained (see equation (3.32)) nuisance parameters.

**Rate uncertainties** only effect the rate of a process and are implemented as log-normal constrained (see equation (3.33)) nuisance parameters.

Additionally freely-floating parameters, used to determine the normalization of a given process, are introduced for the multijet background, the $t\bar{t}$ + B and $t\bar{t}$ + C background. While $t\bar{t}$ + B and $t\bar{t}$ + C parameters are fully correlated between the data-taking periods, categories and channels, the parameter for the multijet background is uncorrelated between the categories and data-taking periods. The latter only effects the fully-hadronic channel. All considered uncertainties are summarized in section 6.2.

### Treatment of the $t\bar{t}+B$ and $t\bar{t}+C$ normalization in the fully-hadronic channel

When fitting the fully-hadronic channel on its own, the freely-floating parameters for the $t\bar{t}+B$ and $t\bar{t}+C$ processes are replaced with a fixed 50 % rate uncertainty. This is done in order to improve stability, since the fit cannot constrain these parameters well in addition to the parameter for the dominant multijet background. Choosing this approach is solely technically motivated. From a theoretical point of view, both approaches should be very similar in the case of the fully-hadronic channel. In a combined fit with the leptonic channels, which can constrain the $t\bar{t}$ + B normalization very well, this is no issue.

Further details on the technical implementation of the freely-floating normalizations for the $t\bar{t}+B$ and $t\bar{t}+C$ processes in the fully-hadronic channel are outlined in the following[1]: The total background can be written as the sum of the individual processes

$$T_{\text{Bkg,Tot.}} = n_{\text{Mutlijet}} \cdot T_{\text{Multijet}} + n_{t\bar{t}+B} \cdot T_{t\bar{t}+B} + n_{t\bar{t}+C} \cdot T_{t\bar{t}+C} + T_{t\bar{t}+LF} + \sum T_{\text{Minor}} \quad (6.1)$$

with the templates $T$ and the relevant normalization factors $n$ of certain processes. Since the multijet template is a composite of the data and simulated backgrounds in the *CR*, it can be written as

$$T_{\text{Multiejt}} = T_{\text{Data,CR}} - n_{t\bar{t}+B} \cdot T_{t\bar{t}+B,\text{CR}} - n_{t\bar{t}+C} \cdot T_{t\bar{t}+C,\text{CR}} - \sum T_{\text{Minor,CR}}. \quad (6.2)$$

This means, that the multijet template $T_{\text{Multijet}}$ is dependent on the $t\bar{t}$ + B and $t\bar{t}$ + C normalization factors.

If a fixed 50 % rate parameters are used for $t\bar{t}$ + B and $t\bar{t}$ + C, the corresponding normalization factors $n$ can be viewed as unity for the nominal case and 0.5 and 1.5 for the down and up variations, respectively. In this case, the templates for the multijet background can be precomputed[2] and treated like any other process, since all uncertainties affecting the rate and shape of the multijet background can be taken from individual templates. In practice, this would mean, that the sum, given in equation (6.2), is evaluated for the nominal case as well as all uncertainties and only the resulting templates are used in the fit. In order to properly propagate the freely-floating normalization parameters, the multijet background template has to be calculated by the fit at runtime. This is necessary, since the $n_{t\bar{t}+B}$ and $n_{t\bar{t}+C}$ parameters, which effect $t\bar{t}$ + B and $t\bar{t}$ + C in the *SR* and *CR*, can be varied to any value by the fit. On a technical level, this is achieved by considering

[1] Equations (6.1) and (6.2) should only be viewed as illustration how the normalization factors impact the templates for the overall background and the multijet background. Equation (6.1) corresponds to $b$ in equation (3.30).

[2] For a fit in `Combine` (see section 3.3), templates for all processes need to be provided as inputs. For each process, the nominal template as well as templates for the up and down variation of each shape uncertainty need to be provided. Rate uncertainties and freely-floating parameters are specified separately and included by the program at runtime. While rate uncertainties can also be added as templates (by scaling the nominal template up and down by a given rate uncertainty value and treating them like all other shape uncertainties) to the input without effecting the result, this is not possible for freely-floating parameters.

all *CR* processes with the full set of uncertainties as separate processes. At fitting time, the sum is implicitly implemented by adding a fixed scale factor of -1 to the simulated background processes in the *CR* and applying the freely-floating multijet normalization to all CR templates. Since the uncertainties of the multijet background are inverted relative to the *SR* by construction, the total prefit uncertainties seem small in comparison to the other channels. E.g. if a JES nuisance parameter scales the simulated background yields up in the *SR*, it scales the multijet background down because of the subtraction.

#### Discriminant in the fully-hadronic channel

For the fully-hadronic channel, the output distribution of the ANNs, described in 5.6, are used as final discriminant in the fit. The expected distributions before the fit (pre-fit) are shown in figure 6.1. For these figure it can be seen, that the simulation and data-driven background estimation models the data well. All residual difference are covered by the uncertainties. Since the shape over the range of the ANN output is crucial for sensitivity of the fit, the binning of the discriminant can be viewed as a parameter for optimization, where a balance between sensitivity and stability has be found. Initially an equally spaced binning was chosen with a bin width of 0.05 units. Studies using Asimov fits showed, that this leads to unstable behavior. To improve stability of the fit, bins at the lower und upper bound are merged until no empty bins remain. For the remaining inner bins[3], adjacent bins were merged. If the number of inner bins was is not even, the left-most inner bin was merged with the first bin.

The final discriminants used in the leptonic channels will be covered in section 7.1.

## 6.2. Systematic uncertainties

Table 6.1 summarizes all uncertainties, that are used in the fit to data combining all $t\bar{t}H(b\bar{b})$ channels, which will be discussed in section 7.2.1.2. Uncertainties that only apply to the additional interpretation, discussed in 7.2.2, are not included in this table. In fits to the individual decay channels certain uncertainties may be omitted. In addition to the uncertainties on the simulation, described in section 4.3, further systematic uncertainties from experimental sources play a role in the analysis and will be discussed in the following:

#### Uncertainty on the jet energy scale and resolution

Uncertainties for the jet energy scale (JES) and resolution (JER) [96], as introduced in section 2.3.5, are considered.

For the JES, the systematic uncertainty is evaluated by shifting the scale applied to a reconstructed jet up and down by one standard deviation. This can lead to jets leaving or entering the acceptance[4] and changes in jet-dependent distributions, e.g. dijet masses. To account for this, after shifting the JES up or down for every jet, events are fully reanalyzed, including the jet selection and calculation of kinematic distributions. In the final fit, this uncertainty is considered a shape uncertainty to all processes.
The total uncertainty of the JES is comprised of multiple independent, and therefore fully uncorrelated, sources due to detector, calibration and jet-flavor response effects, which

[3] In this context "inner bins" refers to all bins except for the left-most and right-most bins, which are the result of the previous step.

[4] In practice, the acceptance is only effected by jets at the lower $p_{\mathrm{T}}$-selection boundary of 30 GeV. By applying the scale shift, jets below (above) the boundary can pass (fail) the selection, which leads to a shift in the jet multiplicity and move the event to a different category or remove it from signal region. Since the direction of the jet is not changed by the JES and JER, no directional acceptance effects are considered.

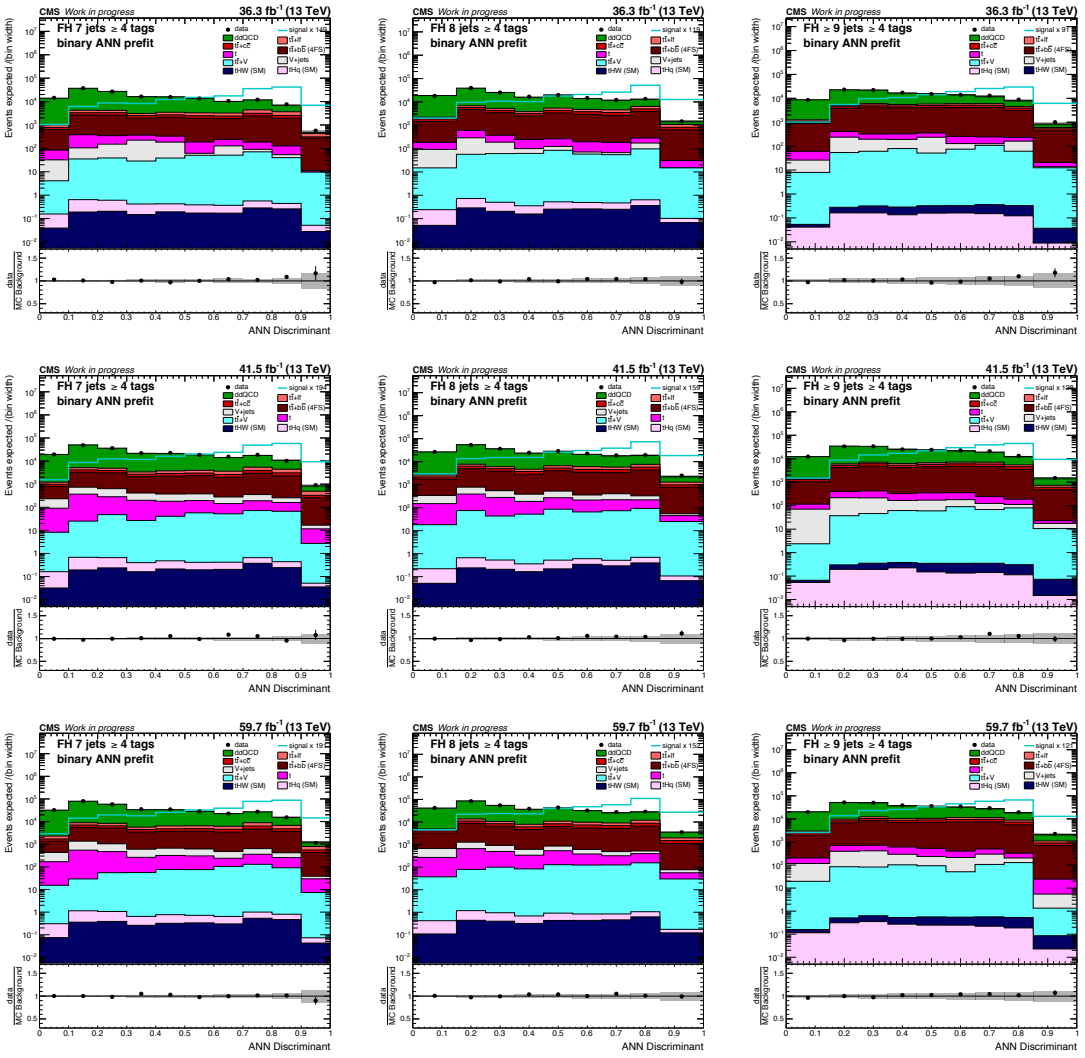



Figure 6.1.: pre-fit ANN discriminant in the fully-hadronic channel for the ($7\,\text{jets}, \geq 4\,\text{b-tags}$) (left), ($8\,\text{jets}, \geq 4\,\text{b-tags}$) (middle) and ($\geq 9\,\text{jets}, \geq 4\,\text{b-tags}$) (right) categories before the fit to data in the 2016 (top), 2017 (middle) and 2018 (bottom) data-taking periods. The filled distributions show the simulated backgrounds and the multijet estimation. The $\mathrm{t\bar{t}H}$ signal is superimposed as line and scaled for visual purposes. The gray band corresponds to the full post-fit uncertainty for the fit in the fully-hadronic $\mathrm{t\bar{t}H(b\bar{b})}$ decay channel.

Table 6.1.: Systematic uncertainties considered for the main combined analysis. Hence, uncertainties that do not apply to the fully-hadronic channel are included in this list. Additional systematics may be included for the additional interpretations (e.g. STXS interpretation). "Type" refers to rate (R) or shape (S) uncertainties and "Correlation" indicates whether the uncertainty is treated as correlated, partially correlated, or uncorrelated across the three data-taking periods. The rate uncertainties are implemented using log-normal distributed pdfs and shape uncertainties using Gaussian distributed pdfs, as described in section 3.3.3.

| Source | Type | Correlation | Remarks |
|---|---|---|---|
| Renorm./fact. scales | R | correlated | Scale uncertainty of (N)NLO prediction, independent for $t\bar{t}H$, tHq, tHW, $t\bar{t}$, t, V, VV |
| PDF+$\alpha_s$ (gg) | R | correlated | PDF uncertainty for gg initiated processes, independent for $t\bar{t}H$, tHq, tHW, and others |
| PDF+$\alpha_s$ ($q\bar{q}$) | R | correlated | PDF uncertainty of $q\bar{q}$ initiated processes ($t\bar{t}W$,W,Z) except tHq |
| PDF+$\alpha_s$ (qg) | R | correlated | PDF uncertainty of qg initiated processes (single t) except tHW |
| Collinear gluon splitting | R | correlated | Additional 50% rate uncertainty on $t\bar{t}+2b$ |
| $\mu_R$ scale | S | correlated | Renormalisation scale uncertainty of the ME generator, independent for $t\bar{t}H$, tHq, tHW, $t\bar{t}+B$ (4FS sample), other $t\bar{t}$ (5FS sample) |
| $\mu_F$ scale | S | correlated | Factorisation scale uncertainty of the ME generator, independent for $t\bar{t}H$, tHq, tHW, $t\bar{t}+B$ (4FS sample), other $t\bar{t}$ (5FS sample) |
| PDF shape | S | correlated | From NNPDF variations, independent for $t\bar{t}+B$ (4FS sample) and $t\bar{t}H$/other $t\bar{t}$ (5FS sample) |
| PS scale ISR | S | correlated | Initial state radiation uncertainty of the PS (PYTHIA), independent for $t\bar{t}H$, $t\bar{t}+B$ (4FS sample), other $t\bar{t}$ (5FS sample) |
| PS scale FSR | S | correlated | Final state radiation uncertainty of the PS (PYTHIA), independent for $t\bar{t}H$, $t\bar{t}+B$ (4FS sample), other $t\bar{t}$ (5FS sample) |
| Integrated luminosity | R | partially | Signal and all backgrounds |
| Lepton ID/Iso | S | uncorrelated | Signal and all backgrounds |
| Trigger efficiency | S | uncorrelated | Signal and all backgrounds |
| L1 prefiring correction | S | uncorrelated | Signal and all backgrounds |
| Pileup | S | correlated | Signal and all backgrounds |
| Jet energy scale (11 scheme) | S | partially | Signal and all backgrounds |
| Jet energy resolution | S | uncorrelated | Signal and all backgrounds |
| b tag HF/LF fraction | S | correlated | Signal and all backgrounds |
| b tag HF/LF stat (linear) | S | uncorrelated | Signal and all backgrounds |
| b tag HF/LF stat (quadratic) | S | uncorrelated | Signal and all backgrounds |
| b tag charm (linear) | S | uncorrelated | Signal and all backgrounds |
| b tag charm (quadratic) | S | uncorrelated | Signal and all backgrounds |
| $TF_{loose}$ correction | S | uncorrelated | QCD multijet estimate |
| N-jet correction | S | uncorrelated | Only for Z + jets in the DL-$CR_Z$ |
| Size of the MC samples | S | uncorrelated | Statistical uncertainty of signal and background prediction due to limited sample size |

in turn are split into multiple sources (see Ref. [96] for a detailed description). Since the stability of the fit suffers from splitting the JES uncertainty into many sources, each potentially with only small variations, similar sources may be merged. For this analysis a scheme with 11 sources is considered. While being fully uncorrelated among each other, some sources are correlated between the data-taking periods, if the underlying effect is independent of the period.
An additional source of uncertainty is applied to cover the small effects of the HEM-issue, as described in section 5.4, for the 2018 dataset. For jets with $-1.57 < \phi < -0.87$ and $-2.5 < \eta < -1.3$, an energy variation of 20 % is applied and treated like all other previously described sources.

The JER systematic uncertainty is evaluated by increasing and decreasing the difference between the reconstructed-level and particle-level jet energy, as described in reference [96]. As for the JES sources, the events are fully reanalyzed after applying this shift for each jet. The uncertainty is considered a shape uncertainty and is applied to all processes.

**Uncertainty on the b-tagging**

Three sources of systematic uncertainty are considered on both the heavy-flavor (HF) and light-flavor (LF) scale factors: The jet energy scale and resolution, the purity of the HF and LF contributions in the sample and the statistics of the sample. The purity describes the contamination of LF jets in b-tagged jets and HF jets in non b-tagged jets. The uncertainties on the sample refer to the samples used to derive the scale factors. The following three sources of uncertainty are fully uncorrelated among each other.

**JES and JER** The b-tagging uncertainty associated with the JES and JER uncertainties is evaluated at the same time as the corresponding source is evaluated for the jets. In practice this is done as follows: When constructing the distributions for a given JES source or the JER, instead of the nominal b-tag scale factor, the scale factor derived with the corresponding shift is used. Hence, no dedicated nuisance parameters are defined for these uncertainties but instead they are absorbed in the nuisance parameters of the JES and JER.

**Purity (HF/LF factions)** A separate nuisance parameters for light- and heavy-flavor jets control the purity uncertainty. A shift of this parameter corresponds to a change of the HF (LF) scale factor as caused by a higher or lower contamination by LF (HF) jets. Since the purity is estimated using the same simulated samples, the uncertainty is correlated between the data-taking periods.

**HF/LF statistics** Four nuisance parameters, referred to as linear and quadratic for HF and LF, are used to control the impact of the statistical uncertainties associated with the size of the samples used for the scale factor derivation. The linear parameter controls the distortion of the overall tilt of the discriminator distribution and the quadratic parameter controls higher-order distortions observable in the upper and lower ends of the distribution with respect to the center.

For charm-flavor jets, the uncertainties for the HF jets are used and doubled in size. The b-tagging purity and statistics uncertainties have been decorrelated between the 2016 and

2017+2018[5] data-taking periods, following tension observed in the semileptonic channel during initial tests of the uncertainty model with data. This is also reasonable from an experimental point of view, because the pixel detector changed significantly between the 2016 and 2017 data-taking periods, as described in section 2.2.2. Since the performance of the b-tagging algorithms is strongly dependent on the experimental performance of the pixel detector.
The b-tagging uncertainties are applied to all processes and are considered a shape uncertainty.

**Uncertainty on the multijet background estimation**

By estimating the multijet background from data, many uncertainties that would be related to this background, if predicted using simulation, are avoided.
All uncertainties affecting the simulated backgrounds used to derive the multijet shape are propagated to the prediction. This includes all theoretical and experimental uncertainties, including the freely floating normalization of $t\bar{t} + b\bar{b}$ and $t\bar{t} + c\bar{c}$ processes. See section 6.1 for technical remarks how this is implemented. The statistical uncertainty of the multijet distribution is taken from the data and the simulated backgrounds in the control region and is treated as equivalent to the statistical uncertainty on the simulated samples.
The only dedicated uncertainty, targeting the multijet background estimation, is the uncertainty on the $TF_{loose}$ correction, which is constructed using a second fit to the $\eta$ distribution, as described in section 5.5.1. This uncertainty is applied to all processes in the *CR*, including data, and is considered uncorrelated between the data-taking periods.

**Uncertainties on the trigger efficiency**

The impact of the trigger efficiency scale factors on the distributions is estimated by varying the trigger scaled factors within their uncertainties. The procedure for the fully-hadronic channel is described in section 5.1.2. The leptonic channels, for which an overview is given in section 7.1, follow a similar procedure to measure the phase-space-specific trigger efficiencies. The trigger efficiency uncertainties are considered a shape uncertainty and are applied to all processes. Since they are specific to the trigger selection in a given channel and data-taking period, the uncertainty is uncorrelated between the channels and periods.

**Uncertainty on the luminosity**

This uncertainty is considered a rate uncertainty and is applied to all processes. A partial correlation scheme, that is only used in the fits combining multiple data-taking periods, is established in order to take certain measurement- but not period-dependent effects into account. The sources and their correlation between the data-taking periods is specified blow:

| | 2016 [%] | 2017 [%] | 2018 [%] |
|---|---|---|---|
| Uncorrelated 2016 | 2.2 | – | – |
| Uncorrelated 2017 | – | 2.0 | – |
| Uncorrelated 2018 | – | – | 1.5 |
| X-Y factorization | 0.9 | 0.8 | 2.0 |
| Length scale 17-18 | – | 0.3 | 0.2 |
| Beam-beam deflection 16-17 | 0.4 | 0.4 | – |
| Dynamic beta 16-17 | 0.5 | 0.5 | – |
| Beam current calibration 17-18 | – | 0.3 | 0.2 |
| Ghosts and satellites 16-17 | 0.4 | 0.1 | – |

For the individual fits per data-taking period, rate uncertainties of 2.5 %, 2.3 % and 2.5 % are used for the 2016, 2017 and 2018 data-taking periods, respectively.

[5] The purity and statistics uncertainties are fully correlated for the 2017 and 2018 data-taking periods.

**Uncertainties on the lepton section**

Uncertainties due to the lepton selection are considered. These uncertainties originate from to data-simulation corrections for the tracking (electrons-only), reconstruction, ID, and isolation efficiencies. The uncertainties are uncorrelated between the data-taking periods and the lepton flavors ($e$ and $\mu$). These uncertainties are considered shape uncertainties and are applied to all processes (in the leptonic channels).

**Further uncertainties**

The uncertainty on the pileup reweighting is evaluated by varying the total inelastic proton-proton cross-section by $\pm 4.6\,\%$ compared to the nominal values. This variation is propagated to the final discriminant and is considered are shape uncertainty.The uncertainty is considered fully correlated between the processes and the data-taking periods.
An uncertainties on the L1 ECAL prefiring correction, described in section 5.2.3, is considered by shifting all prefiring probabilities plus or minus their uncertainties and propagating the variations to the final discriminant. It is considered a shape uncertainty and is fully correlated between the processes but uncorrelated for the 2016 and 2017 data-taking periods.

# 7. Results and Combination

This chapter presents the results of fully-hadronic channel as well as expectations for the combination of all $t\bar{t}H(b\bar{b})$ channels in the CMS $t\bar{t}H(b\bar{b})$ Run-2 analysis. At the time of writing this thesis, the final results for the combination in all channels have not yet been published. Therefore only fits to collected data in the fully-hadronic channel are presented. Due to the lack of a published result for the combination of all channels, the results are obtained from a fit to the sum of the signal and background expectation (asimov dataset). In the context of this thesis, it is mainly used to evaluate the impact and relationship of the fully-hadronic and leptonic channels. Before discussing the results for the measurements of the $t\bar{t}H(b\bar{b})$ cross section, which is also referred to as the *inclusive measurement*, the STXS measurement, the $tH(b\bar{b})$ measurement, and the coupling measurements in section 7.2, the strategies in the leptonic analysis channels are summarized in section 7.1.

Large parts of this chapter present a group effort conducted by multiple research groups, including the author of this thesis, within the CMS collaboration. In order to successfully compile the combined result, a shared infrastructure was created and the workload of running the fits and verifying the results has been shared. The individual analyses provide the final templates, as described in section 6.1, for their respective channel. In the case of the fully-hadronic channel this was the sole responsibility of the author of this thesis.
Additionally, the framework for the tH measurement and the coupling measurement was provided by groups within the CMS collaboration. In this case, the primary contribution of the author of this thesis was to provide the necessary inputs for the fully-hadronic channel.
For the measurement in the STXS framework, the development of the additionally required categorization, based on a Higgs-boson reconstruction, for the fully-hadronic channel was a sole responsibility of the author of this thesis. The additional uncertainties and procedures are provided by the CMS collaboration.

## 7.1. Strategies in the semileptonic and dileptonic channels

This sections summarizes the analysis strategies in the leptonic $t\bar{t}H(b\bar{b})$ channels. The author of this thesis did not contribute (except for minor validation studies on topics shared between all channels) to the analysis presented in the following section. This summary serves the sole purpose of providing a context for the combined measurements as they are presented in this chapter and are not intended as a complete description. The general approach for the generation, reweighting, correction, and modeling of the

simulated samples is fully synchronized between the channels. Hence, the topics covered in section 4.2 are valid in all cases. The lepton selection has been synchronized and ensures no events are double counted between the channels.

### 7.1.1. The semileptonic channel

The semileptonic decay mode of $t\bar{t}H(b\bar{b})$ exhibits, compared to the other two decay modes, the best balance of signal purity and statistical power and is therefore the channel driving the overall sensitivity of the combined analysis.

A combination of various single muon and single electron HLT triggers is used. In the 2017 and 2018 data-taking periods additional triggers based on a single electron and $H_T$ selection are used in the single electron region in addition to the pure single electron triggers.

In the semileptonic channel, events are selected that contain

- exactly one leading lepton, as defined in section 5.3.1,
- four or more jets, as defined in section 5.3.2, of which three or more are required to be b-tagged,
- and a minimum $p_T^{miss}$ of 20 GeV.

The exact values for the selection are summarized in table B.4. The signal and the full background expectation is modeled with simulated samples. The same set of simulated samples, as introduced in section 4.2, is considered for the background model[1].

The following details represent the state, which is used for the results presented in section 7.2.1.2, at the time of writing this thesis but could change with respect to the published result.
The events are split into two exclusive categories based on the number of jets and b-tagged jets, with ($\geq 4$ jets, 3 b-tags) and ($\geq 4$ jets, $\geq 4$ b-tags). In each of these categories, a categorical ANN is used to derive the discriminants for the fit. The ANN is trained to separate events into seven categories: The three signal processes $t\bar{t}H$, tHq and tHW and the four dominant background processes $t\bar{t}$ + b/$b\bar{b}$, $t\bar{t}$ + 2b, $t\bar{t}$ + $c\bar{c}$, and $t\bar{t}$ + LF. The $t\bar{t}$ + 2b process is considered separately, in order to constrain the uncertainty related to collinear gluon splitting. The ANN is implemented as fully-connected feed-forward neutral-network and uses a softmax activation for the output nodes. High-level features, similar to the ANN developed for the fully-hadronic channel, are used as inputs to the ANN. The events are categorized based on the largest output node value into one of seven categories corresponding to the seven processes used in the ANN. This leads to categories that are enriched in the given processes, which improves how well uncertainties relevant for this processes, like the $t\bar{t}$ + $b\bar{b}$ normalization in the $t\bar{t}$ + $b\bar{b}$-node, can be constrained. The expected contribution of the main background processes and the signal is shown in figure 7.1 for all categories used in the final fit to data. For events with ($\geq 4$ jets, 3 b-tags), the ANN node output of a given node is used as discriminant for the fit. For events with ($\geq 4$ jets, $\geq 4$ b-tags), the ANN node output of the tHq, tHW, $t\bar{t}$ + 2b, $t\bar{t}$ + $c\bar{c}$ and $t\bar{t}$ + LF nodes are used as discriminant. The events with largest output values in the $t\bar{t}H$ and $t\bar{t}$ + b/$b\bar{b}$ nodes, are combined into a so-called *likelihood ratio discriminant*, which is defined

[1] For some minor background only the relevant decay modes are used in all the channels. One example is the W + jets background. The fully-hadronic channel only considers simulated samples for the hadronic decay modes while the leptonic channels only consider the leptonic ones. This is valid, since the expected yield of these minor background is very small and including the samples with different decay modes to the channel, will introduce large fluctuations to the background model. For $t\bar{t}$ + jets all decay-mode exclusive samples are used and combined in each channel.

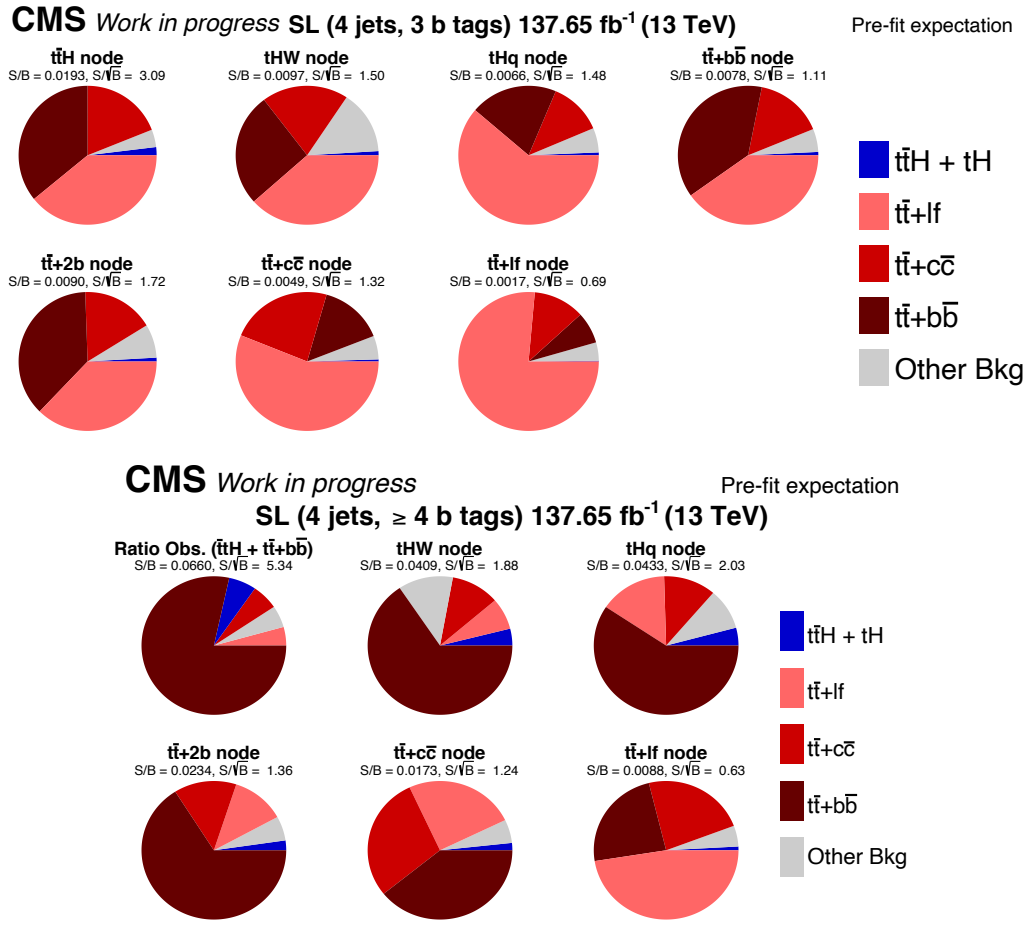



Figure 7.1.: Expected contribution of signal and main background processes in the semileptonic channel before the fit to data in the SL ($\geq 4\,\text{jets}, 3\,\text{b-tags}$) (top) and ($\geq 4\,\text{jets}, \geq 4\,\text{b-tags}$) (bottom) categories. The pie charts show the relative contribution of the relevant processes in each distribution used in the fit. The signal is scaled to the SM expectation, i.e. $\mu_{\text{t}\bar{\text{t}}\text{H}} = \mu_{\text{tH}} = 1$.

as

$$p_o = \frac{o(\text{t}\bar{\text{t}}\text{H})}{o(\text{t}\bar{\text{t}}\text{H}) + o(\text{t}\bar{\text{t}} + \text{b}/\text{b}\bar{\text{b}}) + o(\text{t}\bar{\text{t}} + 2\text{b})}\ , \tag{7.1}$$

where $o(i)$ is the ANN output value in the node $i$. For some output nodes, the distributions are merged into a single bin since the shape information is negligible to constrain the relevant uncertainties. One example is the $\text{t}\bar{\text{t}} + \text{c}\bar{\text{c}}$ node which has to main purpose of constraining the $\text{t}\bar{\text{t}} + \text{C}$ normalization parameter. In the ($\geq 4\,\text{jets}, 3\,\text{b-tags}$) category, this is the case for the $\text{t}\bar{\text{t}}\text{H}$, $\text{t}\bar{\text{t}} + \text{b}/\text{b}\bar{\text{b}}$, $\text{t}\bar{\text{t}} + 2\text{b}$, $\text{t}\bar{\text{t}} + \text{c}\bar{\text{c}}$, and $\text{t}\bar{\text{t}} + \text{LF}$ nodes. For the ($\geq 4\,\text{jets}, \geq 4\,\text{b-tags}$) category, the bins of the $\text{t}\bar{\text{t}}+2\text{b}$, $\text{t}\bar{\text{t}}+\text{c}\bar{\text{c}}$, and $\text{t}\bar{\text{t}}+\text{LF}$ node distributions are merged into a single bin. A total of 13 (7 in the ($\geq 4\,\text{jets}, 3\,\text{b-tags}$) category and 6 in the ($\geq 4\,\text{jets}, 3\,\text{b-tags}$) category) distributions, per data-taking period, are used in the final fit for the semileptonic channel.

### 7.1.2. The dileptonic channel

The dileptonic channel of $\text{t}\bar{\text{t}}\text{H}(\text{b}\bar{\text{b}})$ exhibits a very high signal purity but is limited by the low statistics compared to the other decay modes.

A combination of various two-lepton and one-lepton HLT triggers is used in all data-taking periods. The one-lepton triggers are only used to increase the signal efficiency. While some

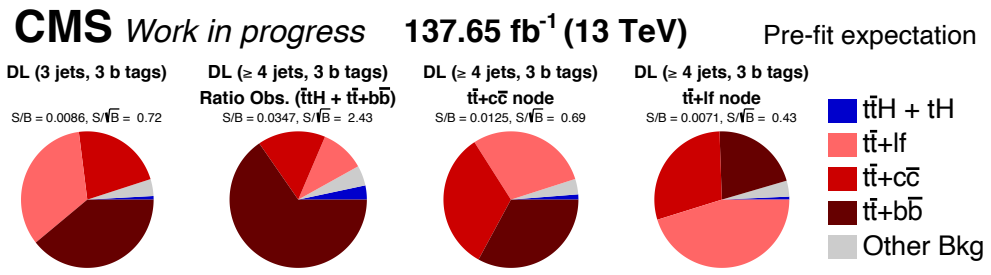



Figure 7.2.: Expected contribution of signal and main background processes in the dileptonic channel before the fit to data in the DL (3 jets, 3 b-tags) and ($\geq 4$ jets, $\geq 3$ b-tags) categories. The pie charts show the relative contribution of the relevant processes in each distribution used in the fit. The signal is scaled to the SM expectation, i.e. $\mu_{\mathrm{t\bar{t}H}} = \mu_{\mathrm{tH}} = 1$.

of these triggers are also used in the semileptonic channel, no events are double counted do to the orthogonal event selections.

In the dileptonic channel, events are selected if they contain

- exactly two charged leptons with opposite sign ($e^+e^-$, $\mu^\pm e^\mp$, $\mu^+\mu^-$), which are required to pass the leading and sub-leading lepton selection criteria described in section 5.3.1,
- two or more jets of which one is required to be b-tagged
- and a minimum $p_{\mathrm{T}}^{\mathrm{miss}}$ of 40 GeV.

The invariant mass of the lepton pair $m_{ee/\mu\mu}$ is required to be $> 20$ GeV and, in same-flavor events, it is further required to lay outside the Z-boson mass interval defined as $76 < m_{ee/\mu\mu} < 106$ GeV. The exact values for the selection are summarized in table B.4. As in the semileptonic channel, the signal and the full background expectation is modeled with simulated samples.

The following details represent the state, which is used for the results presented in section 7.2.1.2, at the time of writing this thesis but could change with respect to the published result.
The events are split into two exclusive categories based on the number of jets and b-tagged jets, with (3 jets, 3 b-tags) and ($\geq 4$ jets, $\geq 3$ b-tags). Independent ANNs, implemented as fully-connected feed-forward neural networks with three hidden layers, are trained for each of these categories. In the category with ($\geq 4$ jets, $\geq 3$ b-tags), a categorical approach, with output nodes for $\mathrm{t\bar{t}H(b\bar{b})}$, $\mathrm{t\bar{t}+b\bar{b}}$, $\mathrm{t\bar{t}+c\bar{c}}$, and $\mathrm{t\bar{t}+LF}$ and a softmax activation is used in order to exploit the high signal purity. Like in the semileptonic channel, the $\mathrm{t\bar{t}H(b\bar{b})}$ and $\mathrm{t\bar{t}+b\bar{b}}$ nodes are combined in a likelihood ratio similar to the one defined in equation (7.1). For the (3 jets, 3 b-tags) category, the ANN is implemented as binary classifier and the output node distribution is used as the final discriminant.
This approach leads to a total of four distributions, per data-taking period, that are used in the final fit for the dileptonic channel. The expected composition of signal and background processes in all categories is shown in figure 7.2. The distributions in the $\mathrm{t\bar{t}+c\bar{c}}$ and $\mathrm{t\bar{t}+LF}$ nodes of the ($\geq 4$ jets, $\geq 3$ b-tags) category are merged into a single bin.

## 7.2. Run-2 result for $\mathrm{t\bar{t}H(b\bar{b})}$

This section covers multiple measurements of the $\mathrm{t\bar{t}H(b\bar{b})}$ process in the full Run-2 dataset, corresponding to an integrated luminosity of 137.2 $\mathrm{fb}^{-1}$ collected by the CMS experiment. In addition to the inclusive measurement, which is used to extract the overall $\mathrm{t\bar{t}H}$ signal strength $\mu_{\mathrm{t\bar{t}H}}$ and is covered in section 7.2.1, additional interpretations are considered:

The simplified-template cross-section (STXS) measurement, which was mainly developed to improve various collaboration-wide combinations; a tH measurement; a SM coupling measurement; and a CP-mixing measurement.

### 7.2.1. The measurement of the $t\bar{t}H(b\bar{b})$ cross section in the fully-hadronic channel and the combination of all channels

The main result for the analysis presented in this thesis is the measurement of the $t\bar{t}H(b\bar{b})$ cross section, in form of the signal strength $\mu_{t\bar{t}H}$, in the fully-hadronic decay channel of the $t\bar{t}H(b\bar{b})$ phase space. Since the analysis was designed with the measurement in the inclusive phase space in mind, studying the impact of the fully-hadronic channel on the inclusive measurement as well as the impact of the other channels on the fully-hadronic channel, is a important part of the work presented in this thesis. As outlined previously, the combined result in all $t\bar{t}H(b\bar{b})$ channels has not been published as of the time this thesis is written. For the fully-hadronic channel, a preliminary result will be used in the following to discuss the outcome of the analysis described over the last chapters of this thesis. It is not expected that the result for fully-hadronic channel, when included in a future publication by the CMS collaboration, will be qualitatively different. While results for the fit to the collected Run-2 data in the fully-hadronic channel are included in this thesis, the combined result will be only shown for a fit to the asimov dataset.

#### 7.2.1.1. Discussion of the result in the fully hadronic channel

Binned maximum likelihood fits are performed simultaneously in the distributions of the three fully-hadronic *SR* categories in order to measure the $t\bar{t}H(b\bar{b})$ cross section in the fully-hadronic decay channel. Individual fits for the three data-taking periods as well as a simultaneous fit of all data-taking periods are considered. For the full Run-2 fit, utilizing an integrated luminosity of $137.2\,\mathrm{fb}^{-1}$ collected by the CMS experiment, 9 discriminant distributions are considered. The uncertainty model described in chapter 6 is used. Fits in the fully-hadronic categories only consider the uncertainties relevant for the fully-hadronic channel and 50 % rate uncertainties are used for $t\bar{t}+\mathrm{B}$ and $t\bar{t}+\mathrm{C}$ normalization instead of the freely-floating parameters used in the combined fit of all channels and the individual fits of the semileptonic and dileptonic channel. After studying, optimizing and validating the analysis on the asimov dataset, yieling the expected results, a step-wise unblinding procedure[2] has been performed to determine the observed results on the collected data. Tables 7.1 and 7.2 list the expected and observed signal significance and best fit values for the $t\bar{t}H$ signal strength in the fully hadronic channel for the 2016, 2017, 2018, and full Run-2 dataset, respectively. It can be seen, that the observed signal strengths in the 2017, 2018 and full Run-2 dataset was determined to a negative value by the fit. While this is unfortunate and does not make it possible to interpret this result in a physical sense, the relative size of the uncertainties is still a meaningful parameter to study. In cases in which the signal strength parameters are negative, no significances can be computed.

When studying these tables, one might expect initially that the results for the individual data-taking periods improve with the size of the dataset, i.e. the fit in the 2016 dataset would be worst and in 2018 dataset best, if the same methods and uncertainty models are used. As shown in tables 7.1 and 7.2, this is not the case for the results in the fully-hadronic channel. In fact, the fit in 2016 dataset on its own exhibits the best overall results in the

[2] Depending on the working group inside the CMS collaboration different unblinding procedures are used. In case of this analysis, the unblinding of the fit is performed in steps. In a first step the post-fit pulls of the nuisance parameters are studied and goodness-of-fit tests are performed. In the second step, the post-fit yields as well as the post-fit discriminant distributions are studies. In the final step, the full fit result is studied.

Table 7.1.: Expected best fit results for the signal strength and the signal significance for the fit in fully-hadronic decay channel. The fit is carried out on the asimov dataset, which is the sum of the signal and background expectations using $\mu_{ttH} = 1$. Two different results are given for the fit to the full Run-2 dataset. In addition the Run-2 fit only considering the fully-hadronic categories, a second result is given (marked with $^{\ddagger}$) for which all $t\bar{t}H(b\bar{b})$ decay channels are used but individual signal strength per channels are used as POIs in the fit. The latter shows how much the fully-hadronic channel profits from the combination of all channels.

| | Significance | $\mu_{ttH}$ | Uncertainty | | |
|---|---|---|---|---|---|
| | | | total | stat | sys |
| 2016 | $0.86\sigma$ | 1.0 | $^{+1.10}_{-1.15}$ | $^{+0.39}_{-0.38}$ | $^{+1.02}_{-1.08}$ |
| 2017 | $0.73\sigma$ | 1.0 | $^{+1.27}_{-1.38}$ | $^{+0.46}_{-0.46}$ | $^{+1.19}_{-1.30}$ |
| 2018 | $0.82\sigma$ | 1.0 | $^{+1.14}_{-1.22}$ | $^{+0.36}_{-0.36}$ | $^{+1.08}_{-1.17}$ |
| Run-2 | $1.26\sigma$ | 1.0 | $^{+0.78}_{-0.78}$ | $^{+0.23}_{-0.23}$ | $^{+0.74}_{-0.74}$ |
| Run-2$^{\ddagger}$ | $2.41\sigma$ | 1.0 | $^{+0.44}_{-0.41}$ | $^{+0.23}_{-0.23}$ | $^{+0.37}_{-0.34}$ |

Table 7.2.: Best fit results for the signal strength and the signal significance for the fit in fully-hadronic decay channel. The fits are carried out for the data-taking eras individually and for the full Run-2 dataset.

| | Significance | $\mu_{ttH}$ | Uncertainty | | |
|---|---|---|---|---|---|
| | | | total | stat | sys |
| 2016 | $0.8\sigma$ | $+0.93$ | $^{+1.17}_{-1.20}$ | $^{+0.38}_{-0.37}$ | $^{+1.10}_{-1.14}$ |
| 2017 | – | $-1.40$ | $^{+1.50}_{-1.62}$ | $^{+0.38}_{-0.38}$ | $^{+1.45}_{-1.57}$ |
| 2018 | – | $-2.43$ | $^{+1.37}_{-1.58}$ | $^{+0.29}_{-0.29}$ | $^{+1.33}_{-1.55}$ |
| Full Run-2 | – | $-1.28$ | $^{+0.81}_{-0.88}$ | $^{+0.20}_{-0.20}$ | $^{+0.78}_{-0.85}$ |

signal strength and significance. The main reason for this is, that the different data-taking periods cannot be compared that easily on this level. While it is true, that the 2016 data-taking period provides the smallest dataset, with an integrated luminosity of $36.3\,\mathrm{fb}^{-1}$, it is also exceptionally well understood from an experimental standpoint. The 2017 dataset on the other hand already proved hard to analyze in the analysis presented in reference [11]. One reason are hardware-related issues in the tracker that have a strong impact on the b-tagging, e.g. larger mis-tagging rates compared to the 2016 data-taking period. Because the multijet background estimation strongly depends on the performance of the b-tagging, this has an especially large impact on the fully-hadronic channel. At the time of the 2018 data-taking, these issues have been better understood and play a lesser role. Furthermore the changes to the LHC machine parameters required a retuning of the triggers used in the fully-hadronic channel, as discussed in section 5.1, and despite best efforts the signal efficiencies of the triggers in the 2017 and 2018 data-taking periods could not reach what was achieved in the 2016 data-taking period, cf. table A.1.

For the full Run-2 dataset, an expected significance of $1.26\sigma$ at a best fit signal strength parameter of $1.0 \pm 0.78$ and an observed signal strength parameter of $-1.28 \pm 0.8$ have been determined in the fit to the asimov dataset and the collected dataset, respectively.

Table 7.3.: Best fit signal strength $\mu_{ttH}$ for the three signal region categories in a fit to the 2016 data-taking period.

| | Significance | $\mu_{ttH}$ | Uncertainty | | |
|---|---|---|---|---|---|
| | | | total | stat | sys |
| $(7\,\text{jets}, \geq 4\,\text{b-tags})$ | $0.6\sigma$ | $+2.09$ | $^{+1.58}_{-1.67}$ | $^{+0.74}_{-0.73}$ | $^{+1.51}_{-1.50}$ |
| $(8\,\text{jets}, \geq 4\,\text{b-tags})$ | – | $-1.41$ | $^{+2.15}_{-2.43}$ | $^{+0.58}_{-0.57}$ | $^{+2.07}_{-2.36}$ |
| $(\geq 9\,\text{jets}, \geq 4\,\text{b-tags})$ | $0.6\sigma$ | $+2.55$ | $^{+1.74}_{-1.74}$ | $^{+0.68}_{-0.68}$ | $^{+1.60}_{-1.60}$ |
| All channels | $0.8\sigma$ | $+0.93$ | $^{+1.17}_{-1.20}$ | $^{+0.38}_{-0.37}$ | $^{+1.10}_{-1.14}$ |

The fit results in the 2016 and 2017 dataset can be compared to previous publications by the CMS collaborations. For the 2016 dataset, this analysis determines an observed best-fit signal strength parameter of $0.93^{+1.17}_{-1.20}$. Comparing this the previous result of of $0.93^{+1.5}_{-1.5}$ published in reference [10], if can be seen that the goal of overall improving the systematic description has been achieved. Since the 2016 dataset is exceptionally well understood, this is a good indicator that the same is in principle true for the full Run-2 dataset. For the 2017 dataset, this analysis determines an observed best-fit signal strength parameter of $-1.4^{+1.50}_{-1.62}$ and an expected significance of $0.73\sigma$. This is very similar to the results published in reference [11], which determined an observed signal strengths of $-1.69^{+1.43}_{-1.47}$ and an expected significance of $0.7\sigma$.

In addition to the result for the full Run-2 dataset only considering the fully-hadronic *SR* categories, a measurement of signal strength in the fully-hadronic channel is shown in which all $\mathrm{t\bar{t}H(b\bar{b})}$ decay channels are considered, but multiple parameters of interest (POIs) for $\mu_{\mathrm{t\bar{t}H}}$ in each decay channel are considered. This result in only shown for fit to the asimov dataset, since the combination with all channels has not been published at the time of writing this thesis. Fitting the full combination with multiple POIs per channels makes it possible to gauge how much an individual channels is profiting from the combination because of correlations between the channels. Due to the stronger constraints on the $\mathrm{t\bar{t}} + \text{jets}$ related uncertainties in the leptonic channels and the large freedom of the multijet background, the fully-hadronic channel is expected to profit from the combination significantly. This can be seen by comparing the two Run-2 results listed in table 7.1. When considering all channels, the expected significance almost doubles to $2.41\sigma$ and the expected systematic uncertainty on the signal strength is reduced by $\sim 40$ percentage points.

Tables 7.3, 7.4, 7.5, and 7.6 list the observed best-fit signal strength parameters per SR category for the 2016, 2017, 2018, and full Run-2 dataset, respectively. No clear order in sensitivity of the different channels can be derived form these numbers but it can be observed, that the sensitivity is rather similar over all studied datasets.

Figure 7.3 shows a graphical representation of tables 7.6 and 7.2 and can be viewed as final summary of the results obtained from the analysis presented in this thesis.

Postfit distributions of the ANNs in the fully-hadronic channel are given in figure 7.4. By comparing the post-fit shape of the signal+background model to the data, an under fluctuation can be observed in some of the signal rich bins on the right side of the distributions. This could explain the negative signal strengths parameters determined by the fit.

For the fit to the full Run-2 dataset in the fully-hadronic channel, a total 186 nuisance parameters are considered. The impacts of the 30 nuisance parameters ranked highest in impact in the fit to the asimov dataset are shown in figure 7.5. Because the sum of signal

Table 7.4.: Best fit fit signal strength $\mu_{ttH}$ for the three signal region categories in a fit to the 2017 data-taking period.

| | Significance | $\mu_{ttH}$ | Uncertainty | | |
|---|---|---|---|---|---|
| | | | total | stat | sys |
| ($7\,\text{jets}, \geq 4\,\text{b-tags}$) | – | $-1.85$ | $^{+2.41}_{-2.88}$ | $^{+0.76}_{-0.75}$ | $^{+2.29}_{-2.78}$ |
| ($8\,\text{jets}, \geq 4\,\text{b-tags}$) | $0.5\sigma$ | $+0.42$ | $^{+2.34}_{-2.45}$ | $^{+0.65}_{-0.64}$ | $^{+2.24}_{-2.36}$ |
| ($\geq 9\,\text{jets}, \geq 4\,\text{b-tags}$) | – | $-0.82$ | $^{+2.66}_{-2.68}$ | $^{+0.69}_{-0.68}$ | $^{+2.56}_{-2.59}$ |
| All channels | – | $-1.40$ | $^{+1.50}_{-1.62}$ | $^{+0.38}_{-0.38}$ | $^{+1.45}_{-1.57}$ |

Table 7.5.: Best fit fit signal strength $\mu_{ttH}$ for the three signal region categories in a fit to the 2018 data-taking period.

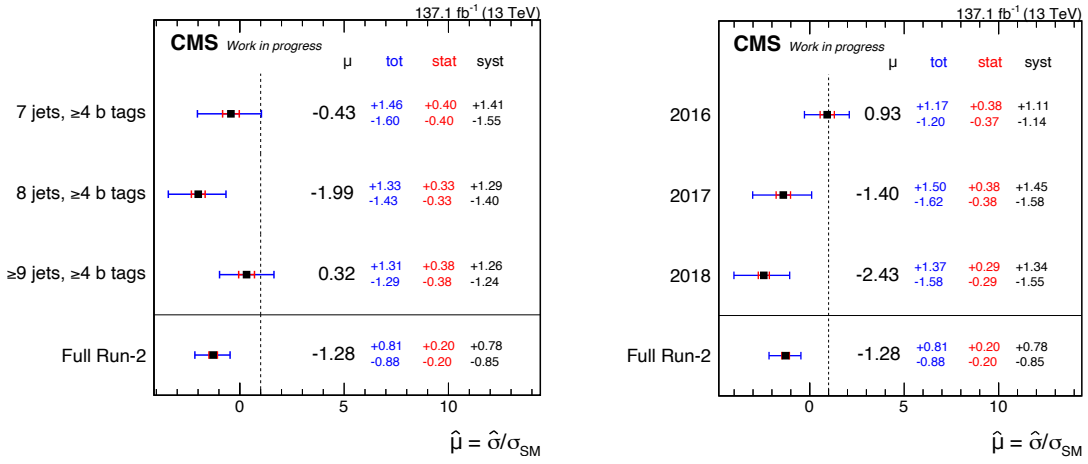
| | Significance | $\mu_{ttH}$ | Uncertainty | | |
|---|---|---|---|---|---|
| | | | total | stat | sys |
| ($7\,\text{jets}, \geq 4\,\text{b-tags}$) | – | $-1.01$ | $^{+1.78}_{-2.08}$ | $^{+0.62}_{-0.61}$ | $^{+1.98}_{-1.66}$ |
| ($8\,\text{jets}, \geq 4\,\text{b-tags}$) | – | $-2.86$ | $^{+2.06}_{-2.28}$ | $^{+0.50}_{-0.50}$ | $^{+1.99}_{-2.22}$ |
| ($\geq 9\,\text{jets}, \geq 4\,\text{b-tags}$) | – | $-0.20$ | $^{+1.99}_{-2.10}$ | $^{+0.58}_{-0.57}$ | $^{+1.90}_{-2.02}$ |
| All channels | – | $-2.43$ | $^{+1.37}_{-1.58}$ | $^{+0.29}_{-0.29}$ | $^{+1.33}_{-1.55}$ |

and background is used as proxy for the data, the pulls of the nuisance parameters are equal to zero and the freely floating rate parameters exhibit a value of one. The parameter with the largest impact is the 50 % rate uncertainty on the $\mathrm{t\bar{t}} + \mathrm{B}$ background. This is reasonable, since the fully-hadronic channel cannot control this parameter well because of the dominant multijet background and missing control regions. Since the leptonic channels can control this parameter particularly well, this also explains the large improvement observed in the fully-hadronic channel in the fit with multiples POIs. The parameters on positions two to eight, ten, and eleven are the nine freely-floating normalization parameters for the multijet background in each *SR* category and data-taking period. Together with the uncertainties on the $\mathrm{TF}_{\mathrm{loose}}$ correction, on position twelve and seventeen for the 2018 and 2017 data-taking period, respectively, eleven of the 30 highest ranking parameters are related to the multijet background estimation. Since the multijet background is the dominant background in the fully-hadronic channel and has a lot of freedom in the fit due to the method used to determine the estimation, this behavior is expected. The remaining parameters among the 30 highest ranked by their impact represent various uncertainties which are expected to influence the measurements like the uncertainty on the renormalization scale $\mu_R$ in the $\mathrm{t\bar{t}} + \mathrm{B}$ process (rank 9), multiple uncertainties on the b-tagging (ranks 13, 15, and 22), the 50 % rate uncertainty on the $\mathrm{t\bar{t}} + \mathrm{C}$ background (rank 14), and the statistical uncertainties on the signal-rich bins (ranks 16, 18-20, 24,25, and 30). The highest ranked uncertainty due to the theoretical description of the process is the parameter that describes the uncertainty on the used $\mathrm{t\bar{t}H}$ cross-section prediction due to the renormalization and factorization scales on rank 23. A high impact is reasonable since it directly affects the scale of the signal process.

The impacts of the 30 nuisance parameters ranked highest in impact after unblinding the analysis are shown in figure 7.6. Overall the same or similar nuisance parameters are ranked high in the fit to the collected data. The normalization parameters for the multijet and $\mathrm{t\bar{t}} + \mathrm{b\bar{b}}$ backgrounds also show the largest impacts. While all multijet parameters are reduced

Table 7.6.: Best fit fit signal strength $\mu_{ttH}$ for the three signal region categories in a fit to the full Run-2 data-taking period.

| | Significance | $\mu_{ttH}$ | Uncertainty | | |
|---|---|---|---|---|---|
| | | | total | stat | sys |
| (7 jets, $\geq$ 4 b-tags) | – | $-0.43$ | $^{+1.46}_{-1.60}$ | $^{+0.40}_{-0.40}$ | $^{+1.44}_{-1.55}$ |
| (8 jets, $\geq$ 4 b-tags) | – | $-1.99$ | $^{+1.33}_{-1.43}$ | $^{+0.33}_{-0.33}$ | $^{+1.28}_{-1.39}$ |
| ($\geq$ 9 jets, $\geq$ 4 b-tags) | $0.8\sigma$ | $+0.32$ | $^{+1.31}_{-1.29}$ | $^{+0.38}_{-0.38}$ | $^{+1.25}_{-1.23}$ |
| All channels | – | -1.28 | $^{+0.81}_{-0.88}$ | $^{+0.20}_{-0.20}$ | $^{+0.78}_{-0.85}$ |




(a) (b)

Figure 7.3.: Best fit values on the signal strength $\mu_{t\bar{t}H}$ in the fully-hadronic channel obtained from the fit to the Run-2 dataset per signal region category (a) and per data-taking era (b). For each measurement, the total uncertainty is shown (outer error bar) and split into statistical (inner error bars) and systematic components.

compared the expected value, the $t\bar{t} + b\bar{b}$ normalization is pulled up by $1\sigma$. Since these parameters provide significant flexibility to the fit and can only change opposite to each other by construction, this is an expected behavior. With more parameters corresponding to the statistical uncertainties on the signal-rich bins, the importance of this type of uncertainty is increased relative to the expected result. Because minor difference from the expectation can be observed in the corresponding bins in the final discriminants in figure 6.1 this behavior is reasonable. Considering these observations it can be stated that the background and uncertainty model describes the data well.

The pre- and post-fit yields are given in tables 7.7, 7.8, and 7.9 for the 2016, 2017, and 2018 data-taking period, respectively. The values are obtained from the fit in the fully-hadronic channel.

#### 7.2.1.2. Discussion of the combined result in all top-quark decay channels

For the final Run-2 results of the $t\bar{t}H(b\bar{b})$ cross-section measurement, a binned maximum likelihood fit is performed simultaneously in all three $t\bar{t}H(b\bar{b})$ decay channels and three data-taking periods. A total of 60 discriminant distributions are considered in this fit. Performing the fit with a single POI, which is the total $t\bar{t}H(b\bar{b})$ signal strength $\mu_{t\bar{t}H}$, represents the main results of the analysis and publication. Table 7.10 lists the expected

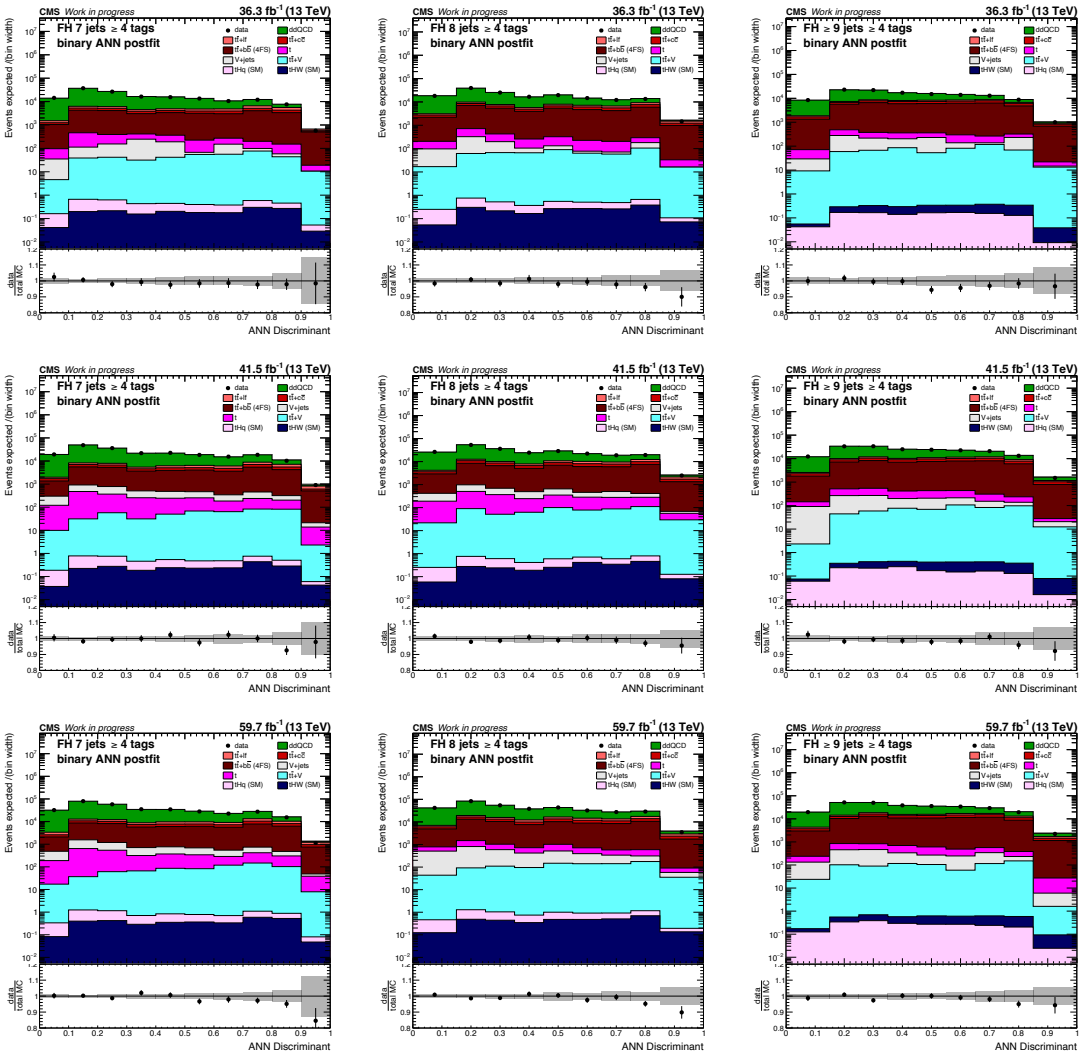



Figure 7.4.: post-fit ANN discriminant in the fully-hadronic channel for the ($7\,\text{jets}, \geq 4\,\text{b-tags}$) (left), ($8\,\text{jets}, \geq 4\,\text{b-tags}$) (middle) and ($\geq 9\,\text{jets}, \geq 4\,\text{b-tags}$) (right) categories after the fit to data in the 2016 (top), 2017 (middle) and 2018 (bottom) data-taking periods. The filled distributions show the simulated backgrounds and the multijet estimation. The gray band corresponds to the full post-fit uncertainty for the fit in the fully-hadronic $\mathrm{t\bar{t}H(b\bar{b})}$ decay channel. The corresponding pre-fit discriminants are shown in figure 6.1.

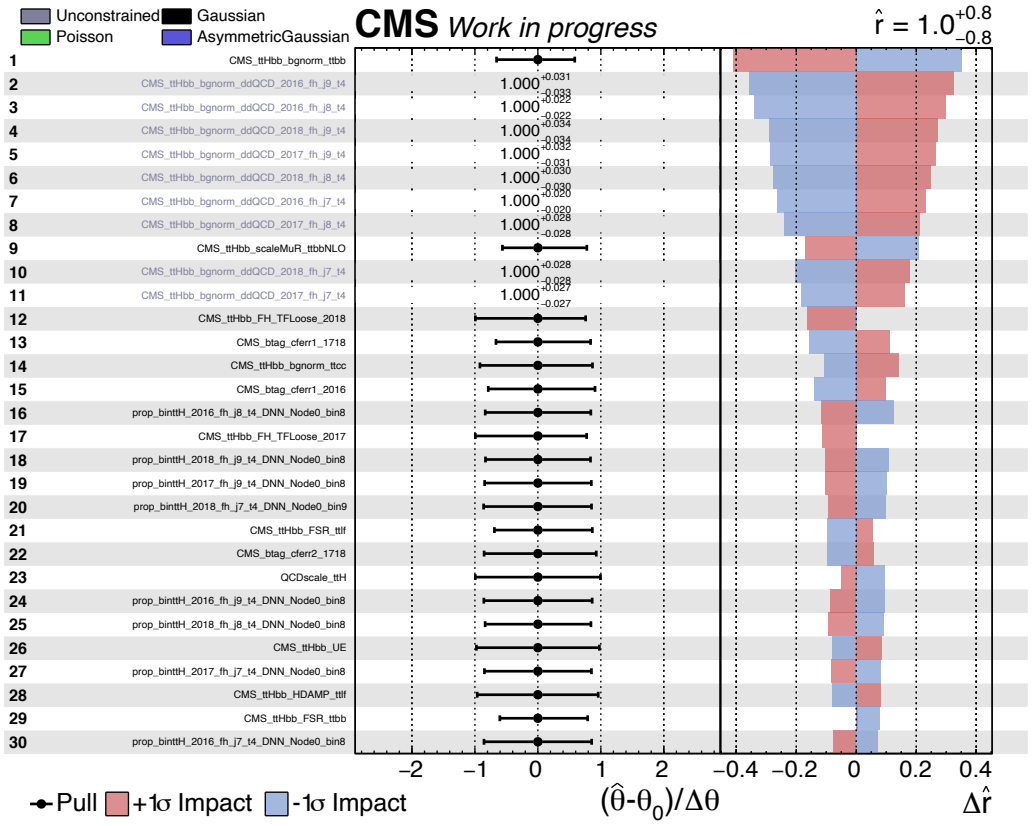



Figure 7.5.: Blinded/Expected pull of the constrained (text in black) and unconstrained/freely-flaoting (text in grey) nuisance parameters included in the fit to all three data-taking periods in the fully-hadronic channel as well as their impact on the signal strength $\mu$, ordered by their impact. Only the 30 nuisance parameters with the largest impact are shown. Since these impacts are extracted from the asimov dataset, the pulls of the nuisance parameters (black markers) are equal to zero. The impact $\Delta\mu$ is computed as the difference of the nominal best fit value of $\mu$ and the best fit value obtained when fixing the nuisance parameter under scrutiny to its best fit value $\hat{\theta}$ plus/minus its post-fit uncertainty (colored areas).

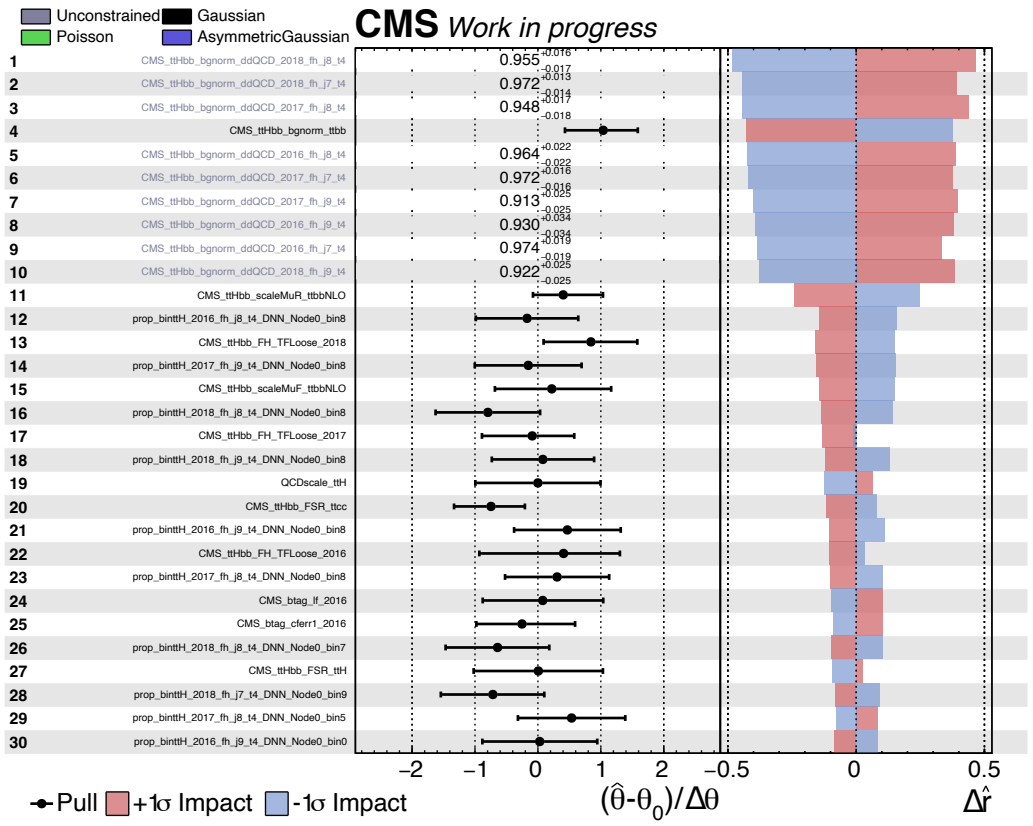



Figure 7.6.: Unblinded pull of the constrained (text in black) and unconstrained/freely-flaoting (text in grey) nuisance parameters included in the fit to all three data-taking periods in the fully-hadronic channel as well as their impact on the signal strength $\mu$, ordered by their impact. Only the 30 nuisance parameters with the largest impact are shown. The pulls of the nuisance parameters (black markers) are computed relative to their pre-fit values $\theta_0$ and uncertainties $\Delta\theta$. The impact $\Delta\mu$ is computed as the difference of the nominal best fit value of $\mu$ and the best fit value obtained when fixing the nuisance parameter under scrutiny to its best fit value $\hat{\theta}$ plus/minus its post-fit uncertainty (colored areas).

Table 7.7.: Pre- and postfit yields in the three categories of the fully-hadronic channel in the 2016 data-taking period. The values are obtained from the fit in the fully-hadronic channel.

| | pre-fit (post-fit) yields | | | | | |
|---|---|---|---|---|---|---|
| Process | $7j \geq 4t$ | | $8j \geq 4t$ | | $\geq 9j \geq 4t$ | |
| $t\bar{t}$ + LF | 751 | (833) | 823 | (914) | 557 | (622) |
| $t\bar{t}$ + C | 509 | (758) | 734 | (1075) | 669 | (991) |
| $t\bar{t}$ + B | 1625 | (2743) | 2410 | (3877) | 2720 | (4213) |
| Single-top | 122 | (147) | 130 | (154) | 92 | (110) |
| $t\bar{t}V$ | 36 | (40) | 52 | (56) | 52 | (57) |
| tH | 0.4 | (0.4) | 0.4 | (0.4) | 0.2 | (0.2) |
| Multijet Bkg | 12225 | (10975) | 12707 | (11098) | 8396 | (6826) |
| Total bkg. | 15337 | (15569) | 16921 | (17244) | 12568 | (12911) |
| ± tot unc. | ±350 | (±179) | ±566 | (±221) | ±795 | (±241) |
| $t\bar{t}H$ | 103 | (0.0) | 141 | (0.0) | 137 | (0.0) |
| ± tot unc. | ±17 | (±13.8) | ±24 | (±19.1) | ±25 | (±20.8) |
| Data | 15441 | | 17063 | | 12720 | |

significance and best fit values for the $t\bar{t}H$ signal strength in the combined fit in all $t\bar{t}H(b\bar{b})$ channels for the 2016, 2017, 2018, and full Run-2 dataset. The table shows that the performance of the combined fit in the years 2016 and 2017 is very similar with $2.49\sigma$ and $2.50\sigma$, respectively. Even though the 2017 dataset contains $\sim 13\,\%$ more data than the 2016 dataset, the contribution of the statistical uncertainty to the best fit value of $\mu_{t\bar{t}H}$ is the same. As already discussed in the previous paragraph on the fit results in the fully-hadronic channel, this most likely results from changes to the machine parameters affecting the selection efficiencies of triggers, reconstruction procedures, and identification algorithms. The result for the 2018 dataset reaches an expected significance of $2.92\sigma$ and the combination of all dataset an expected significance of $4.46\sigma$ and best fit value of $\mu_{t\bar{t}H}$ of $1^{+0.26}_{-0.23}$. Excluding the fully-hadronic channel from the calculation, yields an expected signal significance of $4.06\sigma$. Hence, an improvement of roughly $10\,\%$ can be attributed to the addition of the fully-hadronic channel to the measurement. Previously published results for measurements of the $t\bar{t}H(b\bar{b})$ cross section yield an observed (expected) significance of $3.9\sigma$ ($3.5\sigma$) and a best fit value for $\mu_{t\bar{t}H}$ of $1.15^{+0.32}_{-0.29}$ in case of the CMS experiment [11], using $35.9\,\text{fb}^{-1}$ + $41.5\,\text{fb}^{-1}$ of collision data, and an observed (expected) significance of $1.3\sigma$ ($3.0\sigma$) and best fit value for $\mu_{t\bar{t}H}$ of $0.43^{+0.36}_{-0.32}$ in case of the ATLAS experiment [134], using $139\,\text{fb}^{-1}$ of collision data. The expected results indicate that the analysis presented in this thesis is likely to set a new benchmark for the measurement of $\mu_{t\bar{t}H}$ in the $t\bar{t}H(b\bar{b})$ process.

Figure 7.7 shows the expected best fit values for $\mu_{t\bar{t}H}$ on the full Run-2 dataset for the individual channels and the combination. It is clearly visible that the sensitivity of the overall result is driven by the semileptonic decay channel of $t\bar{t}H(b\bar{b})$. The dileptonic channel is the second most sensitive channel but suffers from the higher statistical uncertainty

Table 7.8.: Pre- and postfit yields in the three categories of the fully-hadronic channel in the 2017 data-taking period. The values are obtained from the fit in the fully-hadronic channel.

| | pre-fit (post-fit) yields | | | | | |
|---|---|---|---|---|---|---|
| Process | $7j \geq 4t$ | | $8j \geq 4t$ | | $\geq 9j \geq 4t$ | |
| $t\bar{t}$ + LF | 878 | (1113) | 1053 | (1395) | 751 | (1007) |
| $t\bar{t}$ + C | 634 | (1059) | 936 | (1597) | 878 | (1537) |
| $t\bar{t}$ + B | 1769 | (3313) | 2670 | (4811) | 3057 | (5281) |
| Single-top | 143 | (189) | 152 | (202) | 118 | (154) |
| $t\bar{t}V$ | 40 | (48) | 55 | (65) | 46 | (56) |
| tH | 0.4 | (0.5) | 0.4 | (0.5) | 0.3 | (0.3) |
| Multijet Bkg | 17462 | (15422) | 19328 | (16438) | 14251 | (11407) |
| Total bkg. | 21112 | (21368) | 24378 | (24733) | 19188 | (19547) |
| ± tot unc. | ±425 | (±242) | ±746 | (±363) | ±982 | (±342) |
| $t\bar{t}H$ | 109 | (0.0) | 152 | (0.0) | 149 | (0.0) |
| ± tot unc. | ±18 | (±16.7) | ±27 | (±23.5) | ±28 | (±21.6) |
| Data | 21221 | | 24529 | | 19334 | |

due to small dileptonic branching fraction. The fully-hadronic channel is least sensitive because of the large systematic uncertainties associated with the multijet background.

In total 559 nuisance parameters are considered in the fit to the full Run-2 dataset. On a technical level, this makes this a very difficult fit to perform and many validation and testing steps have been considered in order to ensure stable and trustworthy results. The impacts of the 30 nuisance parameters ranked highest in impact are shown in figure 7.8. The parameters that describes the uncertainty on the used $t\bar{t}H$ cross-section prediction due to the renormalization and factorization scales and the PDF choice are placed on rank 1 and 13. Due to the much larger sensitivity of the combined fit, it is reasonable that these parameter play a more important role in this fit compared to the fit in the fully-hadronic channel. The second largest impact is associated with the $t\bar{t}$ + B normalization parameter, which is freely-floating in this fit unlike the fit in the fully-hadronic channel. Since the $t\bar{t}$ + B process is the most important background in the leptonic channels, which drive the sensitivity and also consider dedicated $t\bar{t}$ + B-enriched regions in the fit, a large impact is expected. Similar arguments apply to the parameter describing the $t\bar{t}$ + C normalization, which is placed on rank 22. Even though the fully-hadronic channel does provide the least sensitivity, the multijet normalization parameters still exhibit a rather larger impact in the combined fit in all channels. Studies showed that this results from a large correlation between the multijet normalization and the luminosity uncertainties, which also exhibits a high impacts on rank 21. The remaining uncertainties with large impacts in the combined fits describe effects that are expected to rank high in the impact ranking. Some important parameters, e.g. for ISR, FSR, $\mu_R$, $\mu_F$, and b-tagging, that are ranked high in the combined fit are also ranked rather high in the fit in the fully-hadronic channel. Other parameters in the thirty highest ranking parameters have no particular

Table 7.9.: Pre- and postfit yields in the three categories of the fully-hadronic channel in the 2018 data-taking period. The values are obtained from the fit in the fully-hadronic channel.

| | pre-fit (post-fit) yields | | | | | |
|---|---|---|---|---|---|---|
| Process | $7j \geq 4t$ | | $8j \geq 4t$ | | $\geq 9j \geq 4t$ | |
| $t\bar{t}$ + LF | 1541 | (1766) | 1881 | (2178) | 1308 | (1552) |
| $t\bar{t}$ + C | 1045 | (1601) | 1549 | (2335) | 1440 | (2202) |
| $t\bar{t}$ + B | 2881 | (5012) | 4326 | (7214) | 4920 | (7858) |
| Single-top | 232 | (272) | 272 | (322) | 190 | (227) |
| $t\bar{t}V$ | 63 | (72) | 89 | (101) | 67 | (77) |
| tH | 0.7 | (0.8) | 0.7 | (0.8) | 0.4 | (0.5) |
| Multijet Bkg | 27140 | (24584) | 28944 | (25478) | 20819 | (17370) |
| Total bkg. | 33238 | (33686) | 37342 | (37949) | 28899 | (29465) |
| ± tot unc. | ±711 | (±328) | ±1243 | (±455) | ±1608 | (±429) |
| $t\bar{t}H$ | 174 | (0.0) | 244 | (0.0) | 238 | (0.0) |
| ± tot unc. | ±29 | (±20.2) | ±42 | (±28.1) | ±45 | (±39.8) |
| Data | 33408 | | 37587 | | 29139 | |

Table 7.10.: Expected best fit results for the signal strength and the signal significance for the simultaneous fit in all $t\bar{t}H(b\bar{b})$ decay channels. The fit is carried out on the asimov dataset, which is the sum of the signal and background expectations using $\mu_{ttH} = 1$.

| | Significance | $\mu_{ttH}$ | Uncertainty | | |
|---|---|---|---|---|---|
| | | | total | stat | sys |
| 2016 | $2.49\sigma$ | 1.0 | $^{+0.42}_{-0.39}$ | $^{+0.29}_{-0.28}$ | $^{+0.31}_{-0.27}$ |
| 2017 | $2.50\sigma$ | 1.0 | $^{+0.44}_{-0.40}$ | $^{+0.29}_{-0.28}$ | $^{+0.33}_{-0.29}$ |
| 2018 | $2.92\sigma$ | 1.0 | $^{+0.38}_{-0.35}$ | $^{+0.24}_{-0.24}$ | $^{+0.30}_{-0.25}$ |
| Run-2 | $4.46\sigma$ | 1.0 | $^{+0.26}_{-0.23}$ | $^{+0.15}_{-0.15}$ | $^{+0.22}_{-0.18}$ |

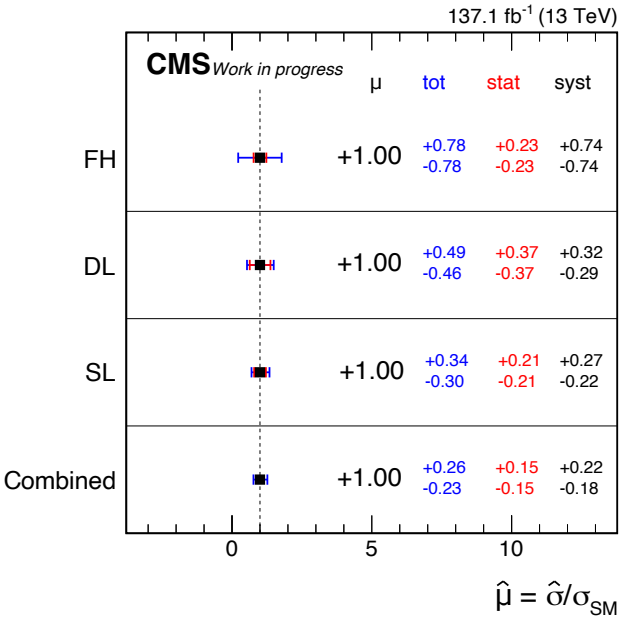



Figure 7.7.: Blinded/Expected best fit values on the signal strength $\mu_{t\bar{t}H}$ obtained from the fit to the Run-2 asimov dataset per channel and in the combination of all channels. Furthermore, the total uncertainty is shown (outer error bar) and split into statistical (inner error bars) and systematic components.

importance in the fully-hadronic channel but are important in the more sensitive leptonic channels. Examples are the uncertainty on the modeling of collinear gluon splitting in $t\bar{t} + B$ on rank 5, uncertainties related to the jet energy scale and resolution on ranks 27 and 29, and the uncertainties on the ME-PS matching (hdamp) and underlying event on ranks 18, 24, and 28. Overall, the highly ranked uncertainties and their constraints follow expected patterns and are similar to previous analyses like the CMS measurement published in reference [11].

From the asimov dataset, a relative uncertainty of 20 % is expected. In the most recent results published in reference [11] by the CMS experiment, which considers the same channels analyzed with similar strategies, the relative uncertainty on the observed signal strength was $\sim 28$ %. This indicates that the goal of improving the systematic description of the background model has been met, especially since the relative statistical uncertainty was two percentage points lower with 13 %. The expected significance of $4.46\sigma$ falls short of the $5\sigma$ threshold, that is generally considered the required level of confidence to claim the observation of a process and has not been reach yet for the $t\bar{t}H(b\bar{b})$ process. In the most recent results published in reference [11] by the CMS experiment, a significance of $3.5\sigma$ was expected (with an observation of $3.9\sigma$). Hence, the additional data and revised analysis strategies contributed significantly to improve the results. Once the analysis is unblinded, the results of the measurement of the signal strength can be discussed and compared to previous results in more detail.

### 7.2.2. Additional interpretations

Multiple other measurements, which reinterpret the data and analysis methods used for the previously discussed inclusive measurement, have been pursued. The fully-hadronic channel is included in all thesis interpretations but especially for the tH and coupling measurements sensitivity is expected to be rather limited if it is considered on its own. Since the final result has not been published at the time of writing this thesis and because the fully-hadronic channel alone only does not provide much sensitivity, no results from

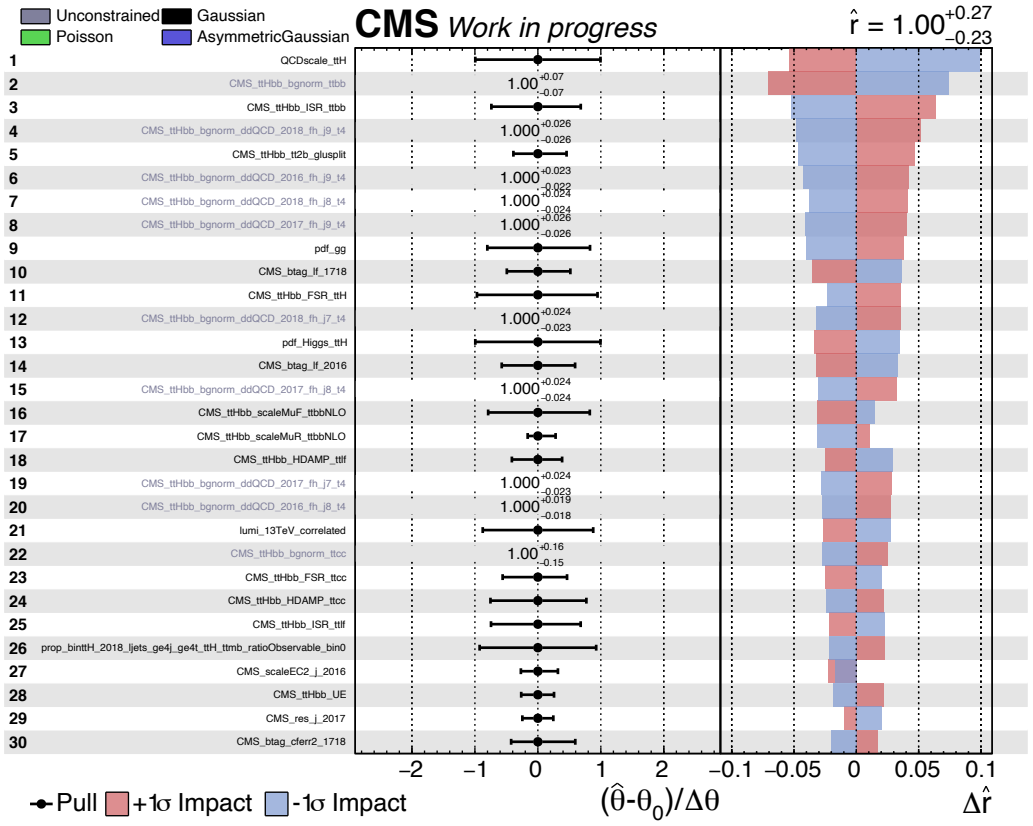



Figure 7.8.: Blinded/Expected pull of the constrained (text in black) and unconstrained/freely-flaoting (text in grey) nuisance parameters included in the fit to all three data-taking periods in the combination of all channels as well as their impact on the signal strength $\mu$, ordered by their impact. Only the 30 nuisance parameters with the largest impact are shown. Since these impacts are extracted from the asimov dataset, the pulls of the nuisance parameters (black markers) are equal to zero. The impact $\Delta\mu$ is computed as the difference of the nominal best fit value of $\mu$ and the best fit value obtained when fixing the nuisance parameter under scrutiny to its best fit value $\hat{\theta}$ plus/minus its post-fit uncertainty (colored areas).

fits to recorded data are shown in this thesis. Only the strategy and approach for the fully-hadronic channel will be presented and discussed in the following.

#### 7.2.2.1. Measurement Simplified templates cross-section framework

A measurement using the Simplified Template Cross Section (STXS) framework [167,168] was added to the Run-2 $t\bar{t}H(b\bar{b})$ analysis. The main goal of the STXS framework is to maximize the sensitivity of the measurements while minimizing the theory dependency. It is mainly aimed towards combined measurements of different Higgs-boson production and decay modes. This is also the reason it was implemented for this analysis. Categorization schemes have been developed for different Higgs-boson production-modes. In the scheme used for the $t\bar{t}H$ measurements, the phase space is split into categories, also referred to as "STXS bins", based on a reconstructed Higgs-boson $p_{\mathrm{T}}$ with the boundaries at 60, 120, 200, and 300 GeV. In practice this means, that each considered SR category is split into five categories, in which signal templates for each reconstructed Higgs-boson $p_{\mathrm{T}}$ range are fit simultaneously. Additional uncertainties for the acceptance and migration are considered and will be described below.
In order to define the five STXS bins, a Higgs Boson has to be reconstructed. Different approaches are chosen in each of the $t\bar{t}H(b\bar{b})$ analysis channels. Since this thesis focuses on the fully-hadronic channel, the approach used in this channel will be discussed in detail, while the approaches in the other channels will only be summarized briefly.

**Higgs-boson reconstruction in the fully-hadronic channel**

For the fully-hadronic channel, a $\chi^2$ reconstruction is used to reconstruct the Higgs Boson. A gaussian function

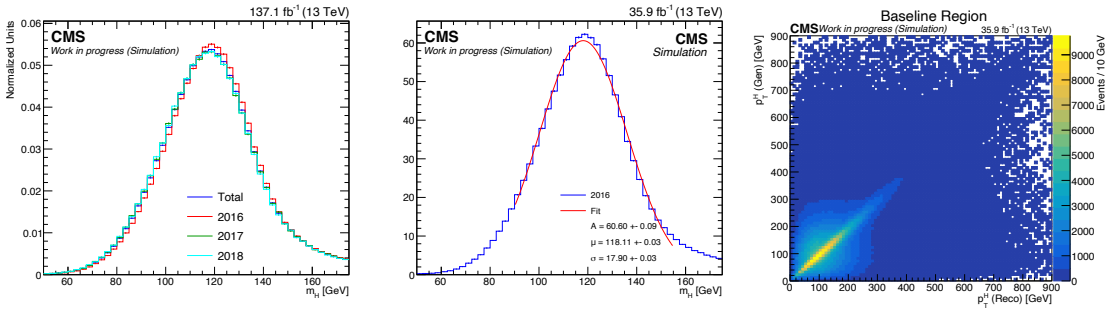
$$g(x; A, \mu, \sigma) = A \cdot exp(\frac{(x-\mu)^2}{\sigma^2}) \tag{7.2}$$

is fitted to the Higgs-boson dijet mass in $t\bar{t}H(b\bar{b})$ simulation. Only events in which a jet could be matched to each of the two bottom quarks originating from the Higgs boson are used. The distributions of this dijet mass for each of the data-taking periods and the weighted sum of all years are shown in figure 7.9a. Small differences between the distributions can be observed. Therefore, the gaussian function is fitted for each data-taking period individually. The result of this fit in the 2016 data-taking period is shown in Figure 7.9b. The relevant parameter values for each data-taking period are given in table 7.11.
In the analysis, the $\mu$ and $\sigma$ values are used to reconstruct the Higgs boson, by finding the dijet pair minimizing the exponent value in equation 7.2. Only the four jets with the highest b-tagging discriminator value are used for this calculation. Figure 7.9c shows the correlation between the reconstructed and generated Higgs-boson $p_{\mathrm{T}}$ for events, in which a jet could be matched to each generator-level bottom quark originating from the Higgs boson, in the fully-hadronic baseline region. It can be observed, that the majority of events are reconstructed very well. The $p_{\mathrm{T}}$ of the dijet pair, which has been reconstructed with the method, is used for the STXS categorization.

**Higgs boson reconstruction in the leptonic channels**

The semileptonic and dileptonic channels categorize the events based on a categorical ANNs, that are trained on signal $t\bar{t}H(b\bar{b})$ events with categories targeting the five Higgs boson $p_{\mathrm{T}}$ bins. These ANNs will be denoted as STXS-ANNs in the following.

For the semileptonic channel, the STXS-ANNs, which exhibit a similar architecture as the ANN in the inclusive analysis, are trained with high-level features separately in the two jet and b-tag multiplicity categories ($\geq 4$ jets, 3 b-tags) and ($\geq 4$ jets, $\geq 4$ b-tags) on events




(a) (b) (c)

Figure 7.9.: Various distributions used for the $\chi^2$ reconstruction of the Higgs boson in the STXS measurement in the fully-hadronic channel. (a) shows the mass of the generated Higgs boson in all three data-taking periods and the weighted sum for all samples. (b) shows the gaussian fit, that is used to derived the values for the $\chi^2$ reconstruction, to the Higgs-boson mass distribution in the 2016 dataset. (c) shows the $\chi^2$-reconstructed Higgs boson mass and the generated Higgs-boson $p_{\mathrm{T}}$ for the events in the fully-hadronic baseline region of the 2016 dataset. All plots show simulated $t\bar{t}H(b\bar{b})$ events.

Table 7.11.: Relevant values for the $\chi^2$ reconstruction of the Higgs boson used in the fully-hadronic channel obtained from the fit of the gaussian distribution, given in equation (7.2), to the Higgs-boson dijet mass in $t\bar{t}H(b\bar{b})$ events in all three data-taking periods.

| | $\mu$ [GeV] | $\sigma$ [GeV] |
|---|---|---|
| 2016 | 118.11 | 17.90 |
| 2017 | 117.02 | 18.45 |
| 2018 | 116.88 | 18.41 |

with highest output in the $t\bar{t}H$-node. This means, that the $t\bar{t}H$ node is further split into the five Higgs boson $p_{\mathrm{T}}$ bins while the other nodes are kept as in the inclusive analysis.

For the dileptonic channel, only events in the ($\geq 4$ jets, $\geq 3$ b-tags) category are used for the STXS measurement. The events in the $t\bar{t}H(b\bar{b})$ node are split into the STXS bins. Events in the other output nodes ($t\bar{t}+b\bar{b}$, $t\bar{t}+c\bar{c}$ and $t\bar{t}+\mathrm{LF}$) are not split into the STXS bins but are included in the fit. The STXS-ANN is trained on ($\geq 4$ jets, $\geq 3$ b-tags) $t\bar{t}H(b\bar{b})$ events using high-level variables and exhibits a similar general architecture as the ANN used in the inclusive analysis.

### Additional uncertainties

Two additional signal uncertainties are considered for the STXS measurement. These uncertainties replace the theory uncertainties (scale and PDF, cf. section 4.3.2) on the inclusive $t\bar{t}H$ signal cross section.
The **migration uncertainty** parametrizes migration effects between the STXS bins and follows an adaptation of the Steward-Tackmann method (see IV.6 in [167]).
The **acceptance uncertainty** parametrizes the acceptance inside a single STXS bin. and is constructed using the QCD scale weights with $\mu_F$ and $\mu_R$ scaled by factor of two for the up variation and $\mu_F$ and $\mu_R$ scale by factor of 0.5 for the down variation. The weights are normalized such that the cross-section inside a STXS bin does not change by applying

Table 7.12.: Expected best fit results for the signal strength and the signal significance for the simultaneous fit in the fully-hadronic signal region categories in all data-taking periods in the STXS bins. Since this is a fit only considering the fully-hadronic channel, 50 % rate uncertainties are used for the $t\bar{t}$ + B and $t\bar{t}$ + C normalization. The fit is carried out on the asimov dataset, which is the sum of the signal and background expectations using $\mu_{ttH} = 1$.

| | Significance | $\mu_{ttH}$ | Uncertainty | | |
|---|---|---|---|---|---|
| | | | total | stat | sys |
| $p_{T,H} \in [0, 60]\,\text{GeV}$ | $0.30\sigma$ | 1.0 | $^{+3.58}_{-3.29}$ | $^{+2.43}_{-2.43}$ | $^{+2.63}_{-2.22}$ |
| $p_{T,H} \in [60, 120]\,\text{GeV}$ | $0.42\sigma$ | 1.0 | $^{+2.43}_{-2.36}$ | $^{+1.65}_{-1.64}$ | $^{+1.78}_{-1.69}$ |
| $p_{T,H} \in [120, 200]\,\text{GeV}$ | $0.62\sigma$ | 1.0 | $^{+1.64}_{-1.59}$ | $^{+1.18}_{-1.17}$ | $^{+1.14}_{-1.08}$ |
| $p_{T,H} \in [200, 300]\,\text{GeV}$ | $0.62\sigma$ | 1.0 | $^{+1.65}_{-1.58}$ | $^{+1.20}_{-1.19}$ | $^{+1.13}_{-1.04}$ |
| $p_{T,H} > 300\,\text{GeV}$ | $0.58\sigma$ | 1.0 | $^{+1.78}_{-1.72}$ | $^{+1.32}_{-1.29}$ | $^{+1.19}_{-1.14}$ |

the weights. This uncertainty is correlated for the signal processes with low Higgs-boson $p_{\text{T}}$ ($< 300\,\text{GeV}$ and high Higgs-boson $p_{\text{T}}$ ($\geq 300\,\text{GeV}$). The same procedure is applied to derive acceptance uncertainties due to ISR and FSR effects from the corresponding generator weights.

**Results**

For the STXS measurement, the $t\bar{t}H$ signal strength is determined with a fit to data in each of the five STXS bins. A total of 33 distributions are fitted per data-taking period and combined for the $137.2\,\text{fb}^{-1}$ measurement in each bin. Aside from the different distributions, categories, multiple signal processes, and the new signal uncertainties, the fit is evaluated in the same way as in the inclusive analysis. As for uncertainty model in the main $t\bar{t}H$ results, the multijet normalization is freely floating in each distribution used in the fully-hadronic channel, i.e. it is freely floating in each category and STXS bin.

Table 7.12 shows the expected best fit value for signal strength and the expected signal significance in each of the five STXS bins for a simultaneous fit in the fully-hadronic signal region categories in all data-taking periods for a given STXS bin. The combination of all these STXS bins corresponds to the inclusive result presented previously. It can be observed, that the fit in the fully-hadronic channel is most sensitive in the STXS bins with higher $p_{T,H}$ values above 120 GeV.
Like in the inclsuive results, three distributions per data-taking period, i.e. a total of nine distributions, are fitted. But for this fit, this number of distributions is fit in each STXS bin. The ANN discriminant output for the 2018 data-taking period is shown in figure 7.10. Due to the lower number of events in the final discriminants (each *SR* category is split into five separate parts), the binning of the discriminant has been changed with respect to the inclusive result. Similar to the approach described in section 6.1, bins with low number of events have been merged such that a smooth distribution is found. Especially in the STXS bis with comparatively small yields, i.e. the lowest and highest bin in $p_{T,H}$, this resulted in much coarser binning.

Results from a fit the collected data are omitted at this point because of the limited explanatory power of the fully-hadronic channel on its own. A result including this channel and strategy will be published in the future.

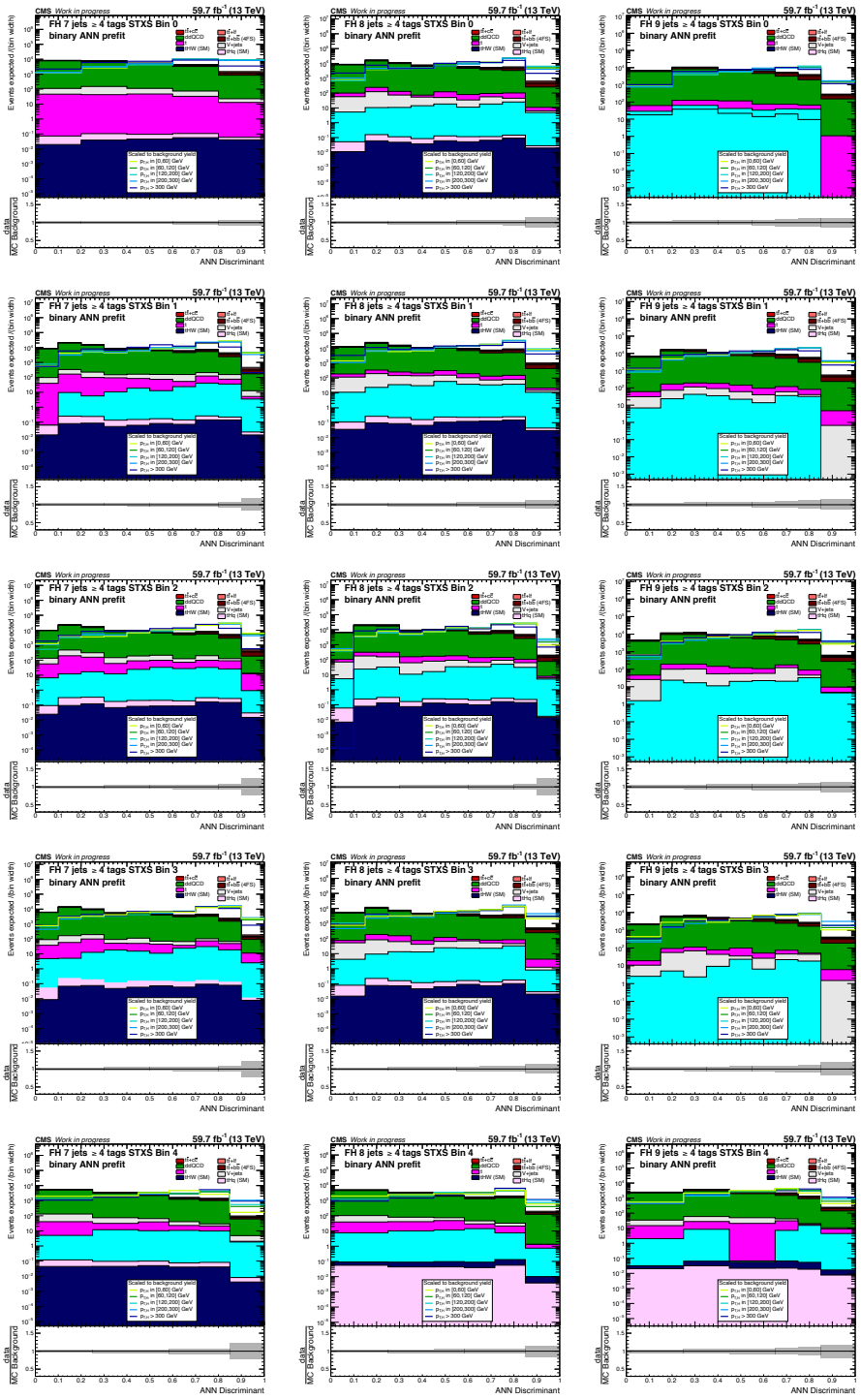



Figure 7.10.: Blinded prefit ANN discriminant in the fully-hadronic channel in the 2018 data-taking period for the ($7\,\text{jets}, \geq 4\,\text{b-tags}$) (left), ($8\,\text{jets}, \geq 4\,\text{b-tags}$) (middle) and ($\geq 9\,\text{jets}, \geq 4\,\text{b-tags}$) (right) categories before the fit to data in the five STXS bins. The STXS bins are determined according to the reconstructed Higgs-boson $p_{\text{T}}$, using the $\chi^2$ method described in section 7.2.2.1. The STXS bins are $p_{T,H}$ $[0, 60]$, $[60, 120]$, $[120, 200]$, $[200, 300]$, $\geq 300$ GeV. The filled distributions show the simulated backgrounds and the multijet estimation. The $t\bar{t}H$ signal for each STXS bin is superimposed as line and scaled for visual purposes. The gray band corresponds to the full post-fit uncertainty for the combined fit in all $t\bar{t}H(b\bar{b})$ decay channels.

#### 7.2.2.2. Measurement of the production of the Higgs boson in association with a single top-quark

The measurement of the tH production (tHq +tHW) aims to set a limit on the tH signal strength modifier

$$\mu_{\mathrm{tH}} = \frac{\sigma_{\mathrm{tH}}}{\sigma_{\mathrm{tH}}^{\mathrm{SM}}} \cdot \frac{\mathcal{B}(\mathrm{H} \to \mathrm{b\bar{b}})}{\mathcal{B}_{\mathrm{SM}}(\mathrm{H} \to \mathrm{b\bar{b}})}. \tag{7.3}$$

For this fit, the tHq and tHW processes are considered signal while the $\mathrm{t\bar{t}H}$ processes are background and kept at SM prediction. No dedicated signal region is constructed and all $\mathrm{t\bar{t}H(b\bar{b})}$ channels use the discriminators described previously for the inclusive analysis[3]. The sensitivity is driven by the semileptonic channel and its tH nodes. Using only the signal regions of the fully-hadronic final state, a

$$\text{expected } 95\,\% \text{ CL upper limit on } \mu_{\mathrm{tH}} \text{ of } 1410^{+600}_{-400}$$

has been determined. The analysis of the tH process with the Higgs-boson decay in the bottom quarks using $35.9\,\mathrm{fb}^{-1}$ of data collected by the CMS experiment [70] determined an expected limit of $41.1^{+13.3}_{-21.1}$. Compared to this result, the determined limit in the fully-hadronic channel is not very sensitive, but is included for completeness in the overall measurement, which will be part of the publication. Results for these measurements using the collected data are omitted at this point because of the limited explanatory power of the fully-hadronic channel on its own. A result including this channel will be published in the future.

#### 7.2.2.3. Coupling measurements

This measurement is parametrized with the coupling modifiers $\kappa_t$,$\tilde{\kappa}_t$ and $\kappa_V$, as defined in section 1.3.4. The coupling strength of the Higgs boson to the top quark and the vector bosons as well as the coupling of a non-SM Higgs boson with different CP state to the top quark are allowed to vary, while the rest of the SM coupling structure is fixed. In practice, changes in event kinematics are taken into account via a linear combination of templates generated for specific coupling points and not by simply scaling the template to the corresponding cross section. For the $\mathrm{t\bar{t}H}$ template, a relatively simple linear combination is considered:

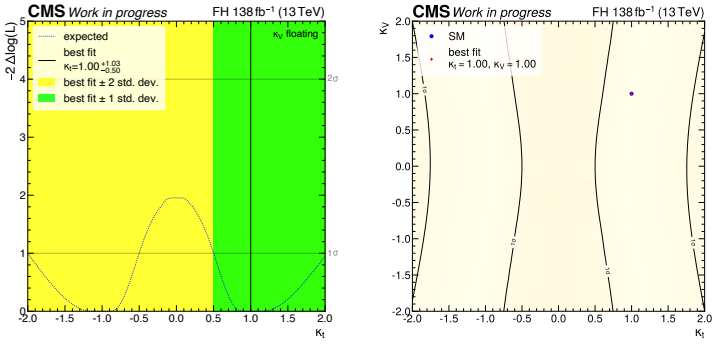
$$T_{\mathrm{t\bar{t}H}}(\kappa_t, \tilde{\kappa}_t, \kappa_V) = A \cdot \kappa_t^2 + B \cdot \tilde{\kappa}_t^2 \ , \tag{7.4}$$

with the SM template A ($\kappa_t = 1$,$\tilde{\kappa}_t = 0$ and $\kappa_V = 1$) and the pure CP-odd template B ($\kappa_t = 0$,$\tilde{\kappa}_t = 1$ and $\kappa_V = 1$). For the tH template more terms are required due to interference:

$$T_{\mathrm{tH}}(\kappa_t, \tilde{\kappa}_t, \kappa_V) = A \cdot \kappa_t(\kappa_t - \kappa_V) + B \cdot \tilde{\kappa}_t^2 C \cdot \kappa_V(\kappa_V - \kappa_t) + D \cdot \kappa_t \kappa_V \ , \tag{7.5}$$

with the templates A ($\kappa_t = 1$,$\tilde{\kappa}_t = 0$ and $\kappa_V = 0$), B ($\kappa_t = 0$,$\tilde{\kappa}_t = 1$ and $\kappa_V = 0$), C ($\kappa_t = 0$,$\tilde{\kappa}_t = 0$ and $\kappa_V = 1$), and D ($\kappa_t = 1$,$\tilde{\kappa}_t = 0$ and $\kappa_V = 1$).
The combination and scaling of the different templates as well as changes to the branching fractions are performed within the CMS `COMBINE` package in a dedicated model, which is shared with analyses of other Higgs-boson production and decay modes within the CMS collaboration. This measurement is planned to be part of the publication for the combined measurement of the $\mathrm{t\bar{t}H(b\bar{b})}$ signal strengths.

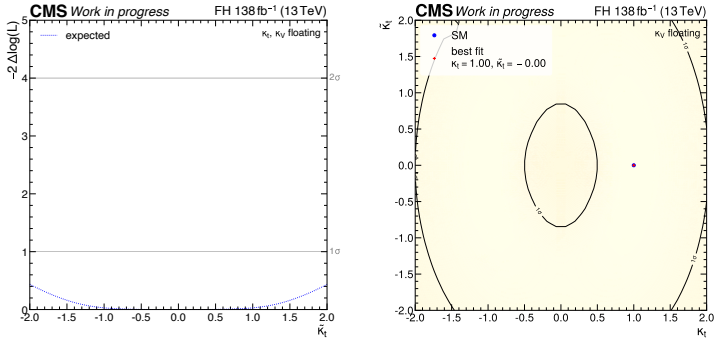
[3] In the specific case of the fully-hadronic $\mathrm{t\bar{t}H(b\bar{b})}$ analysis, which is the main work presented in this thesis, the tH measurements was only added on top in the very end of the analysis. Therefore, no additional consideration (like special observables or additional classifiers) are taken into account for the fully-hadronic channel. In practice this means that only the processed simulated tH events required for the fit are produced while the specific setups and calculations are provided by other groups within the CMS Collaboration.




(a) (b)

Figure 7.11.: Expected results for the measurement of the $\kappa_t$ and $\kappa_V$ coupling parameters in the fully-hadronic channels on the full Run-2 dataset. The plots show the negative log likelihood for the SM expectation ($\kappa_t = 1$, $\tilde{\kappa}_t = 0$, $\kappa_V = 1$). In Figure (a), the dependency on the coupling $\kappa_t$ with floating $\kappa_V$ is shown. The best fit value, as expected in the SM, as well as the one and two $\sigma$ band are shown. Figure (b) shows the negative log likelihood in the $\kappa_t$-$\kappa_V$ plane and the best fit values with the $1\sigma$ contour are shown. The SM value is shown as blue point.

The fully-hadronic channel will considered for this overall coupling measurements but due to the sensitivity of the fully-hadronic channel, when considered on its own, is limited especially for the coupling to the vector bosons $\kappa_V$, as it mainly depends on the sensitivity to tH process.
The plots in figure 7.11, show the blinded fit results in the fully-hadronic channel for the SM coupling measurement in the full Run-2 dataset. The blinded best fit value for the $\kappa_t$ is expected to be $1^{+1.03}_{-0.50}$ with $\kappa_V$ left freely floating. The scan corresponding to this measurement is shown in figure 7.11a. It shows the expected negative log likelihood scan, the best-fit value, and the one and two $\sigma$ bands for best-fit value. The sensitivity for the Top-Yukawa coupling is not very large compared to the other channels, but values below 0.5 can be excluded with more the one $\sigma$ in the fully-hadronic channel. Figure 7.11b, which shows the negative log likelihood for the SM coupling fit in the $\kappa_t$-$\kappa_V$ plane, confirms the expectation that the sensitivity to $\kappa_V$ is rather small. This can be concluded from the almost parallel direction of the $1\sigma$ contour to the $\kappa_V$ axis. As noted previously, the reason for this is the sensitivity to the tH process.
The plots in figure 7.12 show the blinded fit results in the fully-hadronic channel for the CP-mixing measurement in the full Run-2 dataset. Figure 7.12a shows the expected negative log likelihood scan for $\tilde{\kappa}_t$ with $\kappa_V$ and $\kappa_t$ left freely floating. The shape of the likelihood indicated that the fully-hadronic channel is mainly sensitivity to $\tilde{\kappa}_t$ values $> 1.5$. Figure 7.12b shows the expected negative log likelihood scan in the $\kappa_t$-$\tilde{\kappa}_t$ plane. The sensitivity to high absolute values of the $\tilde{\kappa}_t$ values can also be seen from for oval shape of the $1\sigma$ contours in the figure.
Overall, the expectation that the sensitivity of the fully-hadronic channel is rather low has been confirmed by the blinded results presented in this section. While some sensitivity is present for the Top-Yukawa couplings $\kappa_t$ and $\tilde{\kappa}_t$, almost no sensitivity has been achieved for the coupling the vector bosons $\kappa_V$.




(a) (b)

Figure 7.12.: Expected results for the measurement of the $\tilde{\kappa}_t$ coupling parameters in the fully-hadronic channels on the full Run-2 dataset. The plots show the negative log likelihood for the SM expectation ($\kappa_t = 1$, $\tilde{\kappa}_t = 0$, $\kappa_V = 1$). In Figure (a), the dependency on the coupling $\tilde{\kappa}_t$ with floating $\kappa_t$ and $\kappa_V$ is shown. Figure (b) shows the negative log likelihood in the $\kappa_t$-$\tilde{\kappa}_t$ plane and the best fit values with the $1\sigma$ contour are shown. The SM value is shown as blue point.

# Summary and Outlook

With the discovery of the Higgs boson at the LHC in 2012, the final piece of the Standard Model of Particle Physics has been found. While this was a great achievement, it was only the beginning of the experimental investigations into the Higgs sector. Since then, many different decay and production modes of the Higgs boson have been observed, evidence was found, and limits on the cross sections are set for many more. The era of precision measurements of the Higgs sector has long started and measurements of Higgs-boson properties like the Higgs-Yukawa-couplings and the CP properties of the Higgs boson are becoming a focus of the experimental program at the LHC. The $t\bar{t}H$ production mode and $H \to b\bar{b}$ decay channel have been observed individually, but the $t\bar{t}H(b\bar{b})$ final state has not been observed yet with the common level of confidence.

A new measurement of the $t\bar{t}H(b\bar{b})$ process using an integrated luminosity of $137.2\,\mathrm{fb}^{-1}$ collected by the CMS experiment between 2016 and 2018 at a centre-of-mass energy of 13 TeV in proton-proton collisions, has been presented.
While the cross-section measurement of the $t\bar{t}H(b\bar{b})$ process, considering all decay modes, as well as further connected measurements have been discussed, the main focus of this thesis is the fully-hadronic decay mode, in which both W bosons decay to quark pairs. For this decay mode, events are selected that pass dedicated all-jet triggers as well as requirements on jet kinematics and b-tagging identification. More precisely, at least 7 jets, of which at least four have to be b-tagged, and no muons or electrons are required. Events passing these criteria are further split into three orthogonal categories with 7, 8, and $\geq 9$ jets, in which the measurement of the cross section is carried out.
The modeling of backgrounds plays a key role in every physics analysis. The description of the dominant background processes has been improved decisively. The method to estimate the dominant QCD multijet background from collision data has been refined, with respect previous iterations of the analysis, in order to increase the robustness of the estimation itself and make a region in the data available to train an ANN-based classifier. The description of the $t\bar{t} + b\bar{b}$ background, which is especially crucial for the leptonic analyses and the combined fit, was revised by introducing a new simulation with a more precise description of the bottom quark kinematics.
In order to increase the sensitivity of the measurement, an ANN-based classifier has been developed and is used to create a distribution with a strong separation between the signal and dominant background process. This classifier combines various physically motivated input features, that target event characteristics like properties of the additional bottom quarks, i.e. not originating from the top-quark decay, or the spacial distribution of the jets.
Multiple different measurements on the same events have been been carried out utilizing binned maximum likelihood fits. Different parameters of interest and signal samples are used depending on the measurement. Statistical, systematical and theoretical uncertainties are considered as nuisance parameters in the fit.

The inclusive result uses the $t\bar{t}H(b\bar{b})$ signal strength as a single parameter of interest. An observed signal strength of $\mu_{t\bar{t}H} = -1.28^{+0.81}_{-0.81}$, at expected signal strength of

$\mu_{t\bar{t}H} = 1.0^{+0.78}_{-0.78}$, has been determined when only considering the three categories of the fully-hadronic channel with an integrated luminosity of $137.2\,\mathrm{fb}^{-1}$ collected by the CMS experiment between 2016 and 2018. This corresponds to an expected signal significance of $1.26\sigma$. Overall this is a clear improvement over all previously published analyses of this final state. For inclusive simultaneous fit of all $t\bar{t}H(b\bar{b})$ channels a signal strength of $\mu_{t\bar{t}H} = 1.0^{+0.26}_{-0.23}$, which corresponds to an signal significance of $4.46\sigma$, is expected.
A $t\bar{t}H(b\bar{b})$ measurement was conducted under the STXS schema, in which the phase space is further split into orthogonal regions based on the $p_T$ of a reconstructed Higgs boson. For the fully-hadronic channel this splitting is based on $\chi^2$ reconstruction of dijet pairs. On its own, this measurement does not provide much insight for the $t\bar{t}H(b\bar{b})$ process but is mostly intended for the CMS-wide combination of all Higgs boson measurements, which will be published in the future.
For the first time at the CMS collaboration, the fully-hadronic channel of the $t\bar{t}H(b\bar{b})$ and $tH(b\bar{b})$ processes is part of tH, Top-Higgs coupling, and CP-mixing measurements.
For the tH measurement, a binned maximum likelihood fit, with the tH signal strength as parameter of interest, has been performed and an upper limit on the tH cross section is extracted. In this case the SM $t\bar{t}H$ process is considered a background in the fit. For the fully-hadronic channel, an upper limit of $\mu_{tH} < 1410^{+600}_{-400}$ is expected.
The coupling and CP-mixing measurements are performed in the $\kappa$ framework, in which parameters are introduced describing the measured coupling constant relative to the predicted coupling. The parameters $\kappa_t$ (for the Top-Higgs coupling) and $\kappa_V$ (for the Vector-boson-Higgs coupling) are considered in the SM coupling measurement. For the CP-mixing measurement, an additional parameter $\tilde{\kappa}_t$, which describes the coupling to non-SM CP states of the Higgs boson, is included in the model. The expected best fit value for $\kappa_t$ has been determined to $1.0^{+1.03}_{-0.50}$ in the SM measurement with a floating $\kappa_V$ parameter. Coupling values below 0.5 can be excluded with more than $1\sigma$. In the CP-mixing measurement, sensitivity for large absolute values $\tilde{\kappa}_t$ is expected.

## Outlook

After a shutdown from 2018 to 2021, the LHC is planned to restart and deliver proton-proton collisions at 14 TeV in 2022. LHC Run-3 is planned from 2022 to 2024 and is expected to double the integrated luminosity of Run-2. A similar instantaneous luminosity to Run-2 is planned but because of a longer period of peak instantaneous luminosity in the beginning of each fill, the average pile-up of the collected data is expected to increase with respect to Run-2 while the peak pile-up will be similar. After the end of Run-3, a new era, which requires substantial upgrades to the LHC and all detectors, is planned. The *High Luminosity LHC* (HL-LHC) [169] aims to drastically increase the instantaneous luminosity with respect to the current machine configuration to a peak instantaneous luminosity of 5 to $7.5 \times 10^{34}\,\mathrm{cm}^{-1}\,\mathrm{s}^{-1}$. This would enable more accurate measurements and the observation of processes outside the sensitivity of the current LHC machine. The targeted integrated luminosity of $300\,\mathrm{fb}^{-1}$ to $400\,\mathrm{fb}^{-1}$ per year would result in a total integrated luminosity of up to $4000\,\mathrm{fb}^{-1}$ over the currently planned 12 years of operating the HL-LHC.

For the presented $t\bar{t}H(b\bar{b})$ analyses, the additional data and center-of-mass energy expected for Run-3, will not drastically improve the measurement without further deepening our understanding of the theoretical and systemic uncertainties, which are already dominant in the current dataset. For the fully-hadronic $t\bar{t}H(b\bar{b})$ analysis, the area of largest prospect of improving the measurement is the modeling of the QCD multijet background. Hereby the freely-floating normalization plays a crucial role. Revising the method such that a precise estimation of the normalization can be found outside the signal region should

be the goal. This can be either done by deriving the normalization independent of the fit or by including additional control region that can be used to constrain the parameter in the signal region. In addition to this, the development of a different discriminant, which further improves the separation between $\mathrm{t\bar{t}}$ + jets, QCD multijet, and the signal, is expected to improve the sensitivity. With the current model, reaching a significance of $5\sigma$ in the fully-hadronic channel on its own is not expected for Run-3. The tH, coupling and CP-mixing uncertainties on the other hand, should be able to profit from this new data collected during Run-3, since the statistical uncertainty plays a bigger role in these cases. On the longer time scale of the HL-LHC, the same is the case. Due to the much higher instantaneous luminosity and average pile-up, especially analysis strongly dependent on jets, like the fully-hadronic $\mathrm{t\bar{t}H(b\bar{b})}$ analysis, will need to upgrade and revise algorithms and methods related to jet reconstruction and identification.

# Bibliography


[1] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Phys. Lett. B* **716** (2012) 1–29, doi:10.1016/j.physletb.2012.08.020, arXiv:1207.7214.

[2] CMS Collaboration, "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC", *Phys. Lett. B* **716** (2012) 30–61, doi:10.1016/j.physletb.2012.08.021, arXiv:1207.7235.

[3] CMS Collaboration, "Study of the Mass and Spin-Parity of the Higgs Boson Candidate Via Its Decays to Z Boson Pairs", *Phys. Rev. Lett.* **110** (2013), no. 8, 081803, doi:10.1103/PhysRevLett.110.081803, arXiv:1212.6639.

[4] ATLAS Collaboration, "Evidence for the spin-0 nature of the Higgs boson using ATLAS data", *Phys. Lett. B* **726** (2013) 120–144, doi:10.1016/j.physletb.2013.08.026, arXiv:1307.1432.

[5] CMS Collaboration, "Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV", *Eur. Phys. J. C* **75** (2015), no. 5, 212, doi:10.1140/epjc/s10052-015-3351-7, arXiv:1412.8662.

[6] ATLAS Collaboration, "Measurements of the Higgs boson production and decay rates and coupling strengths using pp collision data at $\sqrt{s} = 7$ and 8 TeV in the ATLAS experiment", *Eur. Phys. J. C* **76** (2016), no. 1, 6, doi:10.1140/epjc/s10052-015-3769-y, arXiv:1507.04548.

[7] CMS Collaboration, "Observation of $t\bar{t}H$ production", *Phys. Rev. Lett.* **120** (2018), no. 23, 231801, doi:10.1103/PhysRevLett.120.231801, arXiv:1804.02610.

[8] ATLAS Collaboration, "Observation of Higgs boson production in association with a top quark pair at the LHC with the ATLAS detector", *Phys. Lett. B* **784** (2018) 173–191, doi:10.1016/j.physletb.2018.07.035, arXiv:1806.00425.

[9] ATLAS Collaboration, "Search for the Standard Model Higgs boson decaying into $b\bar{b}$ produced in association with top quarks decaying hadronically in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector", *JHEP* **05** (2016) 160, doi:10.1007/JHEP05(2016)160, arXiv:1604.03812.

[10] CMS Collaboration, "Search for $t\bar{t}H$ production in the all-jet final state in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **06** (2018) 101, doi:10.1007/JHEP06(2018)101, arXiv:1803.06986.

[11] CMS Collaboration, "Measurement of $t\bar{t}H$ production in the $H \to b\bar{b}$ decay channel in $41.5\,\text{fb}^{-1}$ of proton-proton collision data at $\sqrt{s} = 13\,\text{TeV}$", Technical Report CMS-PAS-HIG-18-030, CERN, Geneva, 2019.

[12] S. L. Glashow, "Partial-symmetries of weak interactions", *Nuclear Physics* **22** (February, 1961) 579–588, doi:10.1016/0029-5582(61)90469-2.

[13] M. Gell-Mann, "A schematic model of baryons and mesons", *Physics Letters* **8** (February, 1964) 214–215, doi:10.1016/S0031-9163(64)92001-3.

[14] P. W. Higgs, "Broken symmetries, massless particles and gauge fields", *Phys. Lett.* **12** (1964) 132–133, doi:10.1016/0031-9163(64)91136-9.

[15] F. Englert and R. Brout, "Broken Symmetry and the Mass of Gauge Vector Mesons", *Phys. Rev. Lett.* **13** (August, 1964) 321–323, doi:10.1103/PhysRevLett.13.321.

[16] P. W. Higgs, "Broken Symmetries and the Masses of Gauge Bosons", *Phys. Rev. Lett.* **13** (October, 1964) 508–509, doi:10.1103/PhysRevLett.13.508.

[17] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, "Global Conservation Laws and Massless Particles", *Phys. Rev. Lett.* **13** (November, 1964) 585–587, doi:10.1103/PhysRevLett.13.585.

[18] A. Salam and J. C. Ward, "Electromagnetic and weak interactions", *Physics Letters* **13** (November, 1964) 168–171, doi:10.1016/0031-9163(64)90711-5.

[19] J. D. Bjorken and E. A. Paschos, "Inelastic Electron-Proton and $\gamma$-Proton Scattering and the Structure of the Nucleon", *Physical Review* **185** (September, 1969) 1975–1982, doi:10.1103/PhysRev.185.1975.

[20] S. Weinberg, "A Model of Leptons", *Physical Review Letters* **19** (November, 1967) 1264–1266, doi:10.1103/PhysRevLett.19.1264.

[21] J. D. Bjorken, "Asymptotic Sum Rules at Infinite Momentum", *Phys. Rev.* **179** (March, 1969) 1547–1553, doi:10.1103/PhysRev.179.1547.

[22] S. L. Glashow, J. Iliopoulos, and L. Maiani, "Weak Interactions with Lepton-Hadron Symmetry", *Physical Review D* **2** (October, 1970) 1285–1292, doi:10.1103/PhysRevD.2.1285.

[23] G. 't Hooft, "Renormalizable Lagrangians for massive Yang-Mills fields", *Nuclear Physics B* **35** (1971), no. 1, 167 – 188, doi:http://dx.doi.org/10.1016/0550-3213(71)90139-8.

[24] H. Georgi and S. L. Glashow, "Unified Weak and Electromagnetic Interactions without Neutral Currents", *Phys. Rev. Lett.* **28** (May, 1972) 1494–1497, doi:10.1103/PhysRevLett.28.1494.

[25] G. 't Hooft and M. J. G. Veltman, "Regularization and Renormalization of Gauge Fields", *Nucl. Phys.* **B44** (1972) 189–213, doi:10.1016/0550-3213(72)90279-9.

[26] D. J. Gross and F. Wilczek, "Ultraviolet Behavior of Non-Abelian Gauge Theories", *Phys. Rev. Lett.* **30** (June, 1973) 1343–1346, doi:10.1103/PhysRevLett.30.1343.

[27] H. D. Politzer, "Reliable Perturbative Results for Strong Interactions?", *Phys. Rev. Lett.* **30** (1973) 1346–1349, doi:10.1103/PhysRevLett.30.1346.

[28] W. Pauli, "Relativistic Field Theories of Elementary Particles", *Rev. Mod. Phys.* **13** (1941) 203–232, doi:10.1103/RevModPhys.13.203.

[29] P. D. Group et al., "Review of Particle Physics", *Progress of Theoretical and Experimental Physics* **2020** (08, 2020) doi:10.1093/ptep/ptaa104. 083C01.

[30] J. E. Augustin et al., "Discovery of a Narrow Resonance in $e^+e^-$ Annihilation", *Phys. Rev. Lett.* **33** (December, 1974) 1406–1408, doi:10.1103/PhysRevLett.33.1406.

[31] J. J. Aubert et al., "Experimental Observation of a Heavy Particle J", *Phys. Rev. Lett.* **33** (December, 1974) 1404–1406, doi:10.1103/PhysRevLett.33.1404.

[32] S. W. Herb et al., "Observation of a Dimuon Resonance at 9.5 GeV in 400-GeV Proton-Nucleus Collisions", *Phys. Rev. Lett.* **39** (August, 1977) 252–255, doi:10.1103/PhysRevLett.39.252.

[33] UA1 Collaboration, "Experimental Observation of Isolated Large Transverse Energy Electrons with Associated Missing Energy at $\sqrt{s} = 540$ GeV", *Phys. Lett. B* **122** (1983) 103–116, doi:10.1016/0370-2693(83)91177-2.

[34] UA1 Collaboration, "Experimental Observation of Lepton Pairs of Invariant Mass Around 95-GeV/$c^2$ at the CERN SPS Collider", *Phys. Lett. B* **126** (1983) 398–410, doi:10.1016/0370-2693(83)90188-0.

[35] D0 Collaboration, "Observation of the Top Quark", *Physical Review Letters* **74** (April, 1995) 2632–2637, doi:10.1103/PhysRevLett.74.2632.

[36] CDF Collaboration, "Observation of Top Quark Production in $\bar{\text{p}}$p Collisions with the Collider Detector at Fermilab", *Physical Review Letters* **74** (April, 1995) 2626–2631, doi:10.1103/PhysRevLett.74.2626.

[37] SNO Collaboration, "Direct evidence for neutrino flavor transformation from neutral current interactions in the Sudbury Neutrino Observatory", *Phys. Rev. Lett.* **89** (2002) 011301, doi:10.1103/PhysRevLett.89.011301.

[38] Super-Kamiokande Collaboration, "Evidence for oscillation of atmospheric neutrinos", *Phys. Rev. Lett.* **81** (1998) 1562–1567, doi:10.1103/PhysRevLett.81.1562.

[39] Muon g-2 Collaboration, "Measurement of the Positive Muon Anomalous Magnetic Moment to 0.46 ppm", *Phys. Rev. Lett.* **126** (2021) 141801, doi:10.1103/PhysRevLett.126.141801, arXiv:2104.03281.

[40] E. Noether, "Invariant Variation Problems", *Gott. Nachr.* **1918** (1918) 235–257, doi:10.1080/00411457108231446, arXiv:physics/0503066.

[41] N. Cabibbo, "Unitary Symmetry and Leptonic Decays", *Phys. Rev. Lett.* **10** (1963) 531–533, doi:10.1103/PhysRevLett.10.531.

[42] M. Kobayashi and T. Maskawa, "CP Violation in the Renormalizable Theory of Weak Interaction", *Prog. Theor. Phys.* **49** (1973) 652–657, doi:10.1143/PTP.49.652.

[43] D. J. Griffiths, "Introduction to elementary particles; 2nd rev. version". Physics textbook. Wiley, New York, NY, 2008.

[44] Y. Nambu, "Quasi-particles and gauge invariance in the theory of superconductivity", *Phys. Rev.* **117** (Feb, 1960) 648–663, doi:10.1103/PhysRev.117.648.

[45] J. Goldstone, "Field Theories with Superconductor Solutions", *Nuovo Cim.* **19** (1961) 154–164, doi:10.1007/BF02812722.

[46] J. Goldstone, A. Salam, and S. Weinberg, "Broken Symmetries", *Phys. Rev.* **127** (1962) 965–970, doi:10.1103/PhysRev.127.965.

[47] P. Skands, "Introduction to QCD", in *Theoretical Advanced Study Institute in Elementary Particle Physics: Searching for New Physics at Small and Large Scales*. 7, 2012. arXiv:1207.2389. doi:10.1142/9789814525220_0008.

[48] A. Buckley et al., "General-purpose event generators for LHC physics", *Phys. Rept.* **504** (2011) 145–233, doi:10.1016/j.physrep.2011.03.005, arXiv:1101.2599.

[49] J. C. Collins and D. E. Soper, "Parton Distribution and Decay Functions", *Nucl. Phys. B* **194** (1982) 445–492, doi:10.1016/0550-3213(82)90021-9.

[50] R. D. Ball et al., "Parton distributions from high-precision collider data", *The European Physical Journal C* **77** (2017), no. 10, 663.

[51] T. Sjöstrand et al., "An introduction to PYTHIA 8.2", *Computer Physics Communications* **191** (2015) 159–177.

[52] GEANT4 Collaboration, "GEANT4—a simulation toolkit", *Nucl. Instrum. Meth. A* **506** (2003) 250, doi:10.1016/S0168-9002(03)01368-8.

[53] G. P. Salam, "Towards Jetography", *Eur. Phys. J. C* **67** (2010) 637–686, doi:10.1140/epjc/s10052-010-1314-6, arXiv:0906.1833.

[54] D. Binosi and L. Theußl, "JaxoDraw: A graphical user interface for drawing Feynman diagrams", *Computer Physics Communications* **161** (2004), no. 1, 76 – 86, doi:https://doi.org/10.1016/j.cpc.2004.05.001.

[55] LHC Higgs Cross Section Working Group Collaboration, "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector", doi:10.23731/CYRM-2017-002, arXiv:1610.07922.

[56] CMS Collaboration, "Summaries of CMS cross section measurements". Retrieved 01-Jul-2021. https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsCombined.

[57] CMS Collaboration, "Combined measurements of the Higgs boson's couplings at $\sqrt{s} = 13$ TeV", Technical Report CMS-PAS-HIG-17-031, CERN, Geneva, 2018.

[58] ATLAS Collaboration Collaboration, "Measurement of the properties of Higgs boson production at $\sqrt{s}$=13 TeV in the $H \to \gamma\gamma$ channel using 139 fb$^{-1}$ of $pp$ collision data with the ATLAS experiment", Technical Report ATLAS-CONF-2020-026, CERN, Geneva, Aug, 2020.

[59] CMS Collaboration, "Measurements of Higgs boson production cross sections and couplings in the diphoton decay channel at $\sqrt{s} = 13$ TeV", arXiv:2103.06956.

[60] ATLAS Collaboration, "$CP$ Properties of Higgs Boson Interactions with Top Quarks in the $t\bar{t}H$ and $tH$ Processes Using $H \to \gamma\gamma$ with the ATLAS Detector", *Phys. Rev. Lett.* **125** (2020), no. 6, 061802, doi:10.1103/PhysRevLett.125.061802, arXiv:2004.04545.

[61] CMS Collaboration, "Measurements of $\mathrm{t\bar{t}H}$ production and the CP structure of the Yukawa interaction between the Higgs boson and top quark in the diphoton decay channel", *Phys. Rev. Lett.* **125** (2020), no. 6, 061801, doi:10.1103/PhysRevLett.125.061801, arXiv:2003.10866.

[62] ATLAS Collaboration, "Measurements of the Higgs boson inclusive and differential fiducial cross sections in the $4\ell$ decay channel at $\sqrt{s} = 13$ TeV", *Eur. Phys. J. C* **80** (2020), no. 10, 942, doi:10.1140/epjc/s10052-020-8223-0, arXiv:2004.03969.

[63] ATLAS Collaboration, "Higgs boson production cross-section measurements and their EFT interpretation in the $4\ell$ decay channel at $\sqrt{s}$ =13 TeV with the ATLAS detector", *Eur. Phys. J. C* **80** (2020), no. 10, 957, doi:10.1140/epjc/s10052-020-8227-9, arXiv:2004.03447. [Erratum: Eur.Phys.J.C 81, 29 (2021)].

[64] CMS Collaboration, "Measurements of production cross sections of the Higgs boson in the four-lepton final state in proton-proton collisions at $\sqrt{s} = 13$ TeV", arXiv:2103.04956.

[65] CMS Collaboration, "Observation of Higgs boson decay to bottom quarks", *Phys. Rev. Lett.* **121** (2018), no. 12, 121801, doi:10.1103/PhysRevLett.121.121801, arXiv:1808.08242.

[66] ATLAS Collaboration, "Observation of the Higgs decay to beauty quarks", *J. Phys. Conf. Ser.* **1390** (2019), no. 1, 012046, doi:10.1088/1742-6596/1390/1/012046.

[67] ATLAS Collaboration, "Measurements of $WH$ and $ZH$ production in the $H \to b\bar{b}$ decay channel in $pp$ collisions at 13 TeV with the ATLAS detector", *Eur. Phys. J. C* **81** (2021), no. 2, 178, doi:10.1140/epjc/s10052-020-08677-2, arXiv:2007.02873.

[68] ATLAS Collaboration, "Measurements of Higgs Bosons Decaying to Bottom Quarks from Vector Boson Fusion Production with the ATLAS Experiment at $\sqrt{s} = 13$ TeV", arXiv:2011.08280.

[69] CMS Collaboration, "Measurement of the Higgs boson production rate in association with top quarks in final states with electrons, muons, and hadronically decaying tau leptons at $\sqrt{s} = 13$ TeV", arXiv:2011.03652.

[70] CMS Collaboration, "Search for associated production of a Higgs boson and a single top quark in proton-proton collisions at $\sqrt{s} = 13$ TeV", *Phys. Rev. D* **99** (2019), no. 9, 092005, doi:10.1103/PhysRevD.99.092005, arXiv:1811.09696.

[71] CMS Collaboration, "Evidence for Higgs boson decay to a pair of muons", *JHEP* **01** (2021) 148, doi:10.1007/JHEP01(2021)148, arXiv:2009.04363.

[72] E. Mobs, "The CERN accelerator complex - August 2018", Aug, 2018. http://cds.cern.ch/record/2636343.

[73] L. Evans and P. Bryant, "LHC Machine", *JINST* **3** (2008) S08001, doi:10.1088/1748-0221/3/08/S08001.

[74] The ATLAS Collaboration et al., "The ATLAS experiment at the CERN large hadron collider", *Journal of Instrumentation* **3** (aug, 2008) S08003–S08003, doi:10.1088/1748-0221/3/08/s08003.

[75] CMS Collaboration, "The CMS experiment at the CERN LHC", *JINST* **3** (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[76] The ALICE Collaboration et al., "The ALICE experiment at the CERN LHC", *Journal of Instrumentation* **3** (aug, 2008) S08002–S08002, doi:10.1088/1748-0221/3/08/s08002.

[77] The LHCb Collaboration et al., "The LHCb detector at the LHC", *Journal of Instrumentation* **3** (aug, 2008) S08005–S08005, doi:10.1088/1748-0221/3/08/s08005.

[78] T. Sakuma, "Cutaway diagrams of CMS detector", May, 2019. http://cds.cern.ch/record/2665537.

[79] CMS Collaboration, "Public CMS Tracker Detector Performance Results". Retrieved 11-Feb-2021. https://twiki.cern.ch/twiki/bin/view/CMSPublic/DPGResultsTRK?rev=79.

[80] CMS Collaboration, "CMS Technical Design Report for the Pixel Detector Upgrade", doi:10.2172/1151650.

[81] CMS Collaboration, "Performance of the CMS Hadron Calorimeter with Cosmic Ray Muons and LHC Beam Data", *JINST* **5** (2010) T03012, doi:10.1088/1748-0221/5/03/T03012, arXiv:0911.4991.

[82] CMS Collaboration, "Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$", *JINST* **13** (2018) P06015, doi:10.1088/1748-0221/13/06/P06015, arXiv:1804.04528.

[83] CMS Collaboration, "The CMS trigger system", *JINST* **12** (2017) P01020, doi:10.1088/1748-0221/12/01/P01020, arXiv:1609.02366.

[84] K. Bos et al., "LHC computing Grid: Technical Design Report. Version 1.06 (20 Jun 2005)". Technical design report. LCG. CERN, Geneva, 2005.

[85] I. Bird et al., "Update of the Computing Models of the WLCG and the LHC Experiments", Technical Report CERN-LHCC-2014-014. LCG-TDR-002, Apr, 2014.

[86] CMS Collaboration, "Description and performance of track and primary-vertex reconstruction with the CMS tracker", *JINST* **9** (2014) P10009, doi:10.1088/1748-0221/9/10/P10009, arXiv:1405.6569.

[87] CMS Collaboration, "Alignment of the CMS tracker with LHC and cosmic ray data", *JINST* **9** (2014) P06009, doi:10.1088/1748-0221/9/06/P06009, arXiv:1403.2286.

[88] R. Fruhwirth, "Application of Kalman filtering to track and vertex fitting", *Nucl. Instrum. Meth. A* **262** (1987) 444–450, doi:10.1016/0168-9002(87)90887-4.

[89] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems", *IEEE Proc.* **86** (1998), no. 11, 2210–2239, doi:10.1109/5.726788.

[90] W. Waltenberger, "Adaptive Vertex Reconstruction", Technical Report CMS-NOTE-2008-033, CERN, Geneva, Jul, 2008.

[91] CMS Collaboration, "Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8\,\text{TeV}$", *JINST* **10** (2015) P06005, doi:10.1088/1748-0221/10/06/P06005, arXiv:1502.02701.

[92] CMS Collaboration, "Particle-flow reconstruction and global event description with the CMS detector", *JINST* **12** (2017) P10003, doi:10.1088/1748-0221/12/10/P10003, arXiv:1706.04965.

[93] N. Bartosik, "HEP sketches". Retrieved 25-Oct-2016, Original webpage not available anymore.

[94] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-$k_t$ jet clustering algorithm", *JHEP* **04** (2008) 063, doi:10.1088/1126-6708/2008/04/063, arXiv:0802.1189.

[95] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet user manual", *Eur. Phys. J. C* **72** (2012) 1896, doi:10.1140/epjc/s10052-012-1896-2, arXiv:1111.6097.

[96] CMS Collaboration, "Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV", *JINST* **12** (2017) P02014, doi:10.1088/1748-0221/12/02/P02014, arXiv:1607.03663.

[97] CMS Collaboration, "Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$ using the CMS detector", *JINST* **14** (2019) P07004, doi:10.1088/1748-0221/14/07/P07004, arXiv:1903.06078.

[98] CMS Collaboration, "Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV", *JINST* **13** (2018), no. 05, P05011, doi:10.1088/1748-0221/13/05/P05011, arXiv:1712.07158.

[99] E. Bols et al., "Jet Flavour Classification Using DeepJet", arXiv:2008.10519.

[100] CMS Collaboration, "CMS luminosity measurement for the 2018 data-taking period at $\sqrt{s} = 13$ TeV", Technical Report CMS-PAS-LUM-18-002, CERN, Geneva, 2019.

[101] CMS Collaboration, "Public CMS Luminosity Information". Retrieved 10-Feb-2021. https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults#Multi_year_plots.

[102] CMS Collaboration, "CMS Luminosity Measurements for the 2016 Data Taking Period", Technical Report CMS-PAS-LUM-17-001, CERN, Geneva, 2017.

[103] CMS Collaboration, "CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV", Technical Report CMS-PAS-LUM-17-004, CERN, Geneva, 2018.

[104] D0 Collaboration, "A precision measurement of the mass of the top quark", *Nature* **429** (2004) 638–642, doi:10.1038/nature02589, arXiv:hep-ex/0406031.

[105] D0 Collaboration, "Helicity of the $W$ boson in lepton + jets $t\bar{t}$ events", *Phys. Lett. B* **617** (2005) 1–10, doi:10.1016/j.physletb.2005.04.069, arXiv:hep-ex/0404040.

[106] CDF Collaboration, "Search for a Higgs Boson in $WH \to \ell\nu b\bar{b}$ in $p\bar{p}$ Collisions at $\sqrt{s} = 1.96$ TeV", *Phys. Rev. Lett.* **103** (2009) 101802, doi:10.1103/PhysRevLett.103.101802, arXiv:0906.5613.

[107] CMS Collaboration, "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. C* **75** (2015), no. 6, 251, doi:10.1140/epjc/s10052-015-3454-1, arXiv:1502.02485.

[108] D. N. Salerno, "Search for the Standard Model Higgs Boson Produced in Association with Top Quarks in the Fully Hadronic Final State at the CMS Experiment", 2018. https://cds.cern.ch/record/2313961.

[109] A. Buckley et al., "Lhapdf6: parton density access in the lhc precision era", *The European Physical Journal C* **75** (2015), no. 3, 132.

[110] P. M. Nadolsky et al., "Implications of CTEQ global analysis for collider observables", *Phys. Rev. D* **78** (2008) 013004, doi:10.1103/PhysRevD.78.013004, arXiv:0802.0007.

[111] F. Cascioli, P. Maierhofer, and S. Pozzorini, "Scattering Amplitudes with Open Loops", *Phys. Rev. Lett.* **108** (2012) 111601, doi:10.1103/PhysRevLett.108.111601, arXiv:1111.5206.

[112] T. Hahn, "Cuba - A library for multidimensional numerical integration", *Computer Physics Communications* **168** (2005), no. 2, 78 – 95, doi:https://doi.org/10.1016/j.cpc.2005.01.010.

[113] G. Peter Lepage, "A new algorithm for adaptive multidimensional integration", *Journal of Computational Physics* **27** (1978), no. 2, 192 – 203, doi:https://doi.org/10.1016/0021-9991(78)90004-9.

[114] J. Pata, "Search for the production of the Higgs boson in association with a top quark pair with CMS at $\sqrt{s}$=13 TeV", 2018. http://cds.cern.ch/record/2647662.

[115] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning". MIT Press, 2016. http://www.deeplearningbook.org.

[116] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets", *Neural Computation* **18** (2006), no. 7, 1527–1554, doi:10.1162/neco.2006.18.7.1527, arXiv:https://doi.org/10.1162/neco.2006.18.7.1527. PMID: 16764513.

[117] CMS Collaboration, "A Deep Neural Network for Simultaneous Estimation of b Jet Energy and Resolution", *Comput. Softw. Big Sci.* **4** (2020), no. 1, 10, doi:10.1007/s41781-020-00041-z, arXiv:1912.06046.

[118] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines", in *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 10, pp. 807–814. Omnipress, Madison, WI, USA, 2010.

[119] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous systems", 2015. Software available from tensorflow.org. https://www.tensorflow.org/.

[120] N. Srivastava et al., "Dropout: A Simple Way to Prevent Neural Networks from Overfitting", *Journal of Machine Learning Research* **15** (2014), no. 56, 1929–1958.

[121] The ATLAS Collaboration, The CMS Collaboration, The LHC Higgs Combination Group Collaboration, "Procedure for the LHC Higgs boson search combination in Summer 2011", Technical Report CMS-NOTE-2011-005. ATL-PHYS-PUB-2011-11, CERN, Geneva, Aug, 2011.

[122] CMS Collaboration, "Combine: A RooStats / RooFit - based software tools used for statistical analysis within the CMS Higgs PAG". Retrieved 07-Jul-2021. https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/index.html.

[123] L. Moneta et al., "The RooStats Project", arXiv:1009.1003.

[124] W. Verkerke and D. Kirkby, "The RooFit toolkit for data modeling", *arXiv e-prints* (June, 2003) physics/0306116, arXiv:physics/0306116.

[125] R. Brun et al., "root-project/root: v6.18/02", August, 2019. doi:10.5281/zenodo.3895860, https://doi.org/10.5281/zenodo.3895860.

[126] J. Neyman and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses", pp. 73–108. Springer New York, New York, NY, 1992. doi:10.1007/978-1-4612-0919-5_6.

[127] S. S. Wilks, "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses", *Annals Math. Statist.* **9** (1938), no. 1, 60–62, doi:10.1214/aoms/1177732360.

[128] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics", *Eur. Phys. J. C* **71** (2011) 1554, doi:10.1140/epjc/s10052-011-1554-0, arXiv:1007.1727. [Erratum: Eur.Phys.J.C 73, 2501 (2013)].

[129] J. S. Conway, "Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra", 3, 2011. doi:10.5170/CERN-2011-006.115.

[130] CMS Collaboration, "Search for the associated production of the Higgs boson with a top-quark pair", *JHEP* **09** (2014) 087, doi:10.1007/JHEP09(2014)087, arXiv:1408.1682. [Erratum: JHEP 10, 106 (2014)].

[131] 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 $\sqrt{s} = 8$ TeV with the ATLAS detector", *Eur. Phys. J. C* **75** (2015), no. 7, 349, doi:10.1140/epjc/s10052-015-3543-1, arXiv:1503.05066.

[132] CMS Collaboration, "Search for $t\bar{t}H$ production in the $H \to b\bar{b}$ decay channel with leptonic $t\bar{t}$ decays in proton-proton collisions at $\sqrt{s} = 13$ TeV", *JHEP* **03** (2019) 026, doi:10.1007/JHEP03(2019)026, arXiv:1804.03682.

[133] ATLAS Collaboration, "Search for the standard model Higgs boson produced in association with top quarks and decaying into a $b\bar{b}$ pair in *pp* collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", *Phys. Rev. D* **97** (2018), no. 7, 072016, doi:10.1103/PhysRevD.97.072016, arXiv:1712.08895.

[134] ATLAS Collaboration, "Measurement of the Higgs boson decaying to *b*-quarks produced in association with a top-quark pair in *pp* collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector", Technical Report ATLAS-CONF-2020-058, CERN, Geneva, Nov, 2020.

[135] A. Denner, R. Feger, and A. Scharf, "Irreducible background and interference effects for Higgs-boson production in association with a top-quark pair", *JHEP* **04** (2015) 008, doi:10.1007/JHEP04(2015)008, arXiv:1412.5290.

[136] F. Demartin et al., "tWH associated production at the LHC", *Eur. Phys. J. C* **77** (2017), no. 1, 34, doi:10.1140/epjc/s10052-017-4601-7, arXiv:1607.05862.

[137] F. Demartin, F. Maltoni, K. Mawatari, and M. Zaro, "Higgs production in association with a single top quarkÂ atÂ theÂ LHC", *Eur. Phys. J. C* **75** (2015), no. 6, 267, doi:10.1140/epjc/s10052-015-3475-9, arXiv:1504.00611.

[138] M. Beneke, P. Falgari, S. Klein, and C. Schwinn, "Hadronic top-quark pair production with NNLL threshold resummation", *Nucl. Phys. B* **855** (2012) 695–741, doi:10.1016/j.nuclphysb.2011.10.021, arXiv:1109.1536.

[139] M. Cacciari et al., "Top-pair production at hadron colliders with next-to-next-to-leading logarithmic soft-gluon resummation", *Phys. Lett. B* **710** (2012) 612–622, doi:10.1016/j.physletb.2012.03.013, arXiv:1111.5869.

[140] P. Baernreuther, M. Czakon, and A. Mitov, "Percent Level Precision Physics at the Tevatron: First Genuine NNLO QCD Corrections to $q\bar{q} \to t\bar{t} + X$", *Phys. Rev. Lett.* **109** (2012) 132001, doi:10.1103/PhysRevLett.109.132001, arXiv:1204.5201.

[141] M. Czakon and A. Mitov, "NNLO corrections to top-pair production at hadron colliders: the all-fermionic scattering channels", *JHEP* **12** (2012) 054, doi:10.1007/JHEP12(2012)054, arXiv:1207.0236.

[142] M. Czakon and A. Mitov, "NNLO corrections to top pair production at hadron colliders: the quark-gluon reaction", *JHEP* **01** (2013) 080, doi:10.1007/JHEP01(2013)080, arXiv:1210.6832.

[143] M. Czakon, P. Fiedler, and A. Mitov, "Total Top-Quark Pair-Production Cross Section at Hadron Colliders Through $O(\alpha_S^4)$", *Phys. Rev. Lett.* **110** (2013) 252004, doi:10.1103/PhysRevLett.110.252004, arXiv:1303.6254.

[144] M. Czakon and A. Mitov, "Top++: A Program for the Calculation of the Top-Pair Cross-Section at Hadron Colliders", *Comput. Phys. Commun.* **185** (2014) 2930, doi:10.1016/j.cpc.2014.06.021, arXiv:1112.5675.

[145] N. Kidonakis, "Two-loop soft anomalous dimensions for single top quark associated production with a W- or H-", *Phys. Rev. D* **82** (2010) 054018, doi:10.1103/PhysRevD.82.054018, arXiv:1005.4451.

[146] P. Kant 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** (2015) 74–89, doi:10.1016/j.cpc.2015.02.001, arXiv:1406.4403.

[147] M. Aliev et al., "HATHOR: HAdronic Top and Heavy quarks crOss section calculatoR", *Comput. Phys. Commun.* **182** (2011) 1034–1046, doi:10.1016/j.cpc.2010.12.040, arXiv:1007.1327.

[148] A. Kulesza et al., "Associated production of a top quark pair with a heavy electroweak gauge boson at NLO+NNLL accuracy", *Eur. Phys. J. C* **79** (2019), no. 3, 249, doi:10.1140/epjc/s10052-019-6746-z, arXiv:1812.08622.

[149] T. Gehrmann et al., "$W^+W^-$ Production at Hadron Colliders in Next to Next to Leading Order QCD", *Phys. Rev. Lett.* **113** (2014), no. 21, 212001, doi:10.1103/PhysRevLett.113.212001, arXiv:1408.5243.

[150] J. M. Campbell, R. Ellis, and C. Williams, "Vector boson pair production at the LHC", *JHEP* **07** (2011) 018, doi:10.1007/JHEP07(2011)018, arXiv:1105.0020.

[151] V. Khachatryan et al., "Event generator tunes obtained from underlying event and multiparton scattering measurements", *The European Physical Journal C* **76** (2016), no. 3, 155.

[152] A. M. Sirunyan et al., "Extraction and validation of a new set of CMS pythia8 tunes from underlying-event measurements", *The European Physical Journal C* **80** (2020), no. 1, 4.

[153] P. Nason, "A New Method for Combining NLO QCD with Shower Monte Carlo Algorithms", *Journal of High Energy Physics* **2004** (nov, 2004) 040–040, doi:10.1088/1126-6708/2004/11/040.

[154] S. Frixione, P. Nason, and C. Oleari, "Matching NLO QCD computations with parton shower simulations: the POWHEG method", *Journal of High Energy Physics* **2007** (nov, 2007) 070–070, doi:10.1088/1126-6708/2007/11/070.

[155] S. Alioli, P. Nason, C. Oleari, and E. Re, "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX", *Journal of High Energy Physics* **2010** (2010), no. 6, 43.

[156] H. B. Hartanto, B. Jäger, L. Reina, and D. Wackeroth, "Higgs boson production in association with top quarks in the POWHEG BOX", *Phys. Rev. D* **91** (May, 2015) 094003, doi:10.1103/PhysRevD.91.094003.

[157] J. Alwall et al., "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations", *Journal of High Energy Physics* **2014** (2014), no. 7, 79.

[158] T. Ježo, J. M. Lindert, N. Moretti, and S. Pozzorini, "New NLOPS predictions for $t\bar{t} + b$ -jet production at the LHC", *Eur. Phys. J. C* **78** (2018), no. 6, 502, doi:10.1140/epjc/s10052-018-5956-0, arXiv:1802.00426.

[159] M. M. Horzela, "Merging of $t\bar{t}$ and $t\bar{t}+b\bar{b}$ simulations for an improved background modeling for $t\bar{t}H(b\bar{b})$ measurements", Master's thesis, Karlsruhe Institute of Technology (KIT), 2019.

[160] R. Barlow and C. Beeston, "Fitting using finite Monte Carlo samples", *Computer Physics Communications* **77** (1993), no. 2, 219 – 228, doi:https://doi.org/10.1016/0010-4655(93)90005-W.

[161] CMS Collaboration Collaboration, "Performance of the pile up jet identification in CMS for Run 2",.

[162] CMS Collaboration, "Jet algorithms performance in 13 TeV data", CMS Physics Analysis Summary CMS-PAS-JME-16-003, 2017.

[163] F. Chollet et al., "Keras". https://keras.io, 2015.

[164] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization", arXiv:1412.6980.

[165] G. Fox and S. Wolfram, "Event shapes in $e^+e^-$ annihilation", *Nuclear Physics B* **157** (1979), no. 3, 543 – 544, doi:10.1016/0550-3213(79)90120-2.

[166] J. D. Bjorken and S. J. Brodsky, "Statistical Model for electron-Positron Annihilation Into Hadrons", *Phys. Rev. D* **1** (1970) 1416–1420, doi:10.1103/PhysRevD.1.1416.

[167] "Les Houches 2017: Physics at TeV Colliders Standard Model Working Group Report". (3, 2018). arXiv:1803.07977.

[168] N. Berger et al., "Simplified Template Cross Sections - Stage 1.1", arXiv:1906.02754.

[169] I. Zurbano Fernandez et al., "High-Luminosity Large Hadron Collider (HL-LHC): Technical design report", doi:10.23731/CYRM-2020-0010.

# Appendix

## A. Trigger strategy

As described in section 5.1.1, trigger paths targeting leptons or other outstanding objects, e.g. MET or a single high-$p_{\mathrm{T}}$ jet, are not overly phase-space specific and have general utility in many different analyses. Since the fully-hadronic $\mathrm{t\bar{t}H(b\bar{b})}$ channel does not contain any leptons or other outstanding objects, the signal efficiency reachable with such general triggers is not very large. Especially the general trigger paths targeting hadronic final states, like $H_{\mathrm{T}}$- or single-jet paths, exhibit threshold that are too large for the phase space of $\mathrm{t\bar{t}H(b\bar{b})}$. As described in section 5.1, the $H_{\mathrm{T}}$ threshold for a fully-hadronic $\mathrm{t\bar{t}H(b\bar{b})}$ selection should realistically be as low as 300 GeV, while the least stringent trigger of this kind in the 2018 data-taking period exhibits a threshold of 1050 GeV. In the case of the single-jet trigger, the threshold of 500 GeV would similarly remove most of the signal as it can be seen in the top left plot in figure A.1. Both types of triggers would reject $> 90\,\%$ of the signal events. Therefore, the development of dedicated trigger paths plays a major role in the design of most fully-hadronic analyses. Especially for rare processes, like $\mathrm{t\bar{t}H}$ for which only roughly 70000 event are expected during LHC Run-2, this is an important step that cannot be overlooked[1]. In the case of the fully-hadronic $\mathrm{t\bar{t}H(b\bar{b})}$ analysis, maximizing the efficiency is achieved by using a combination of the jet multiplicity, $H_{\mathrm{T}}$, and b-tagging criteria at trigger level. A common issue for trigger paths targeting only hadronic activity is that their rate is strongly dependent on the instantaneous luminosity and pileup. Since all trigger paths are limited to a certain maximum rate and prescaled triggers are generally undesirable for precision measurements, controlling the rate over the fills and machine evolution of the LHC is crucial. This is done during development of a path by modeling various scenarios and during runtime by monitoring the actual rate as well as fill and run dependencies. If the path exhibits excessive rates during runtime it may has to be retuned. This means that using dedicated triggers also requires constant attention and monitoring during data-taking by the developer. The strong dependency on the instantaneous luminosity is also the reason that the thresholds of hadronic trigger paths tend to be rather high. Another common issue is that hadronic trigger paths

[1] Also leptonic analysis can profit from dedicated triggers. For example, a path targeting a single electron and $H_{\mathrm{T}}$ was developed for the semileptonic $\mathrm{t\bar{t}H(b\bar{b})}$ analysis. Since leptonic trigger paths also exhibit a (smaller) dependency on the instantaneous luminosity, the thresholds of the pure lepton triggers also had to be raised for the 2017 and 2018 data-taking periods. Targeting the $H_{\mathrm{T}}$ in addition to a single isolated electron enables the lowering of the electron thresholds.

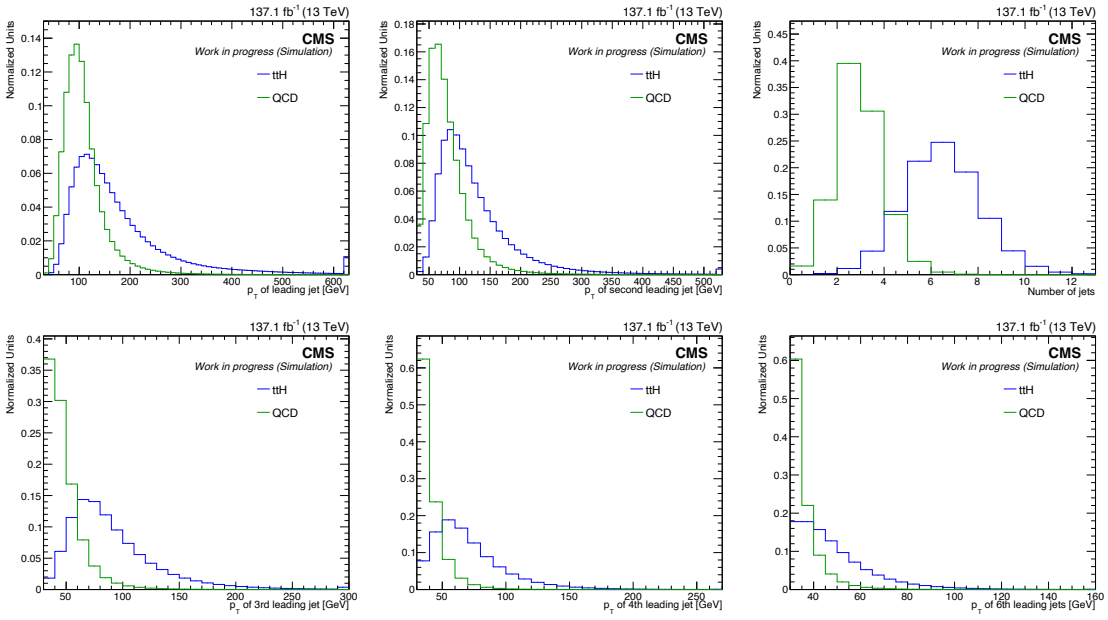



Figure A.1.: Distribution of the number of jet (top left), $p_{\mathrm{T}}$ of leading (top middle), second (top right), third (bottom left), fourth (bottom middle) and sixth (bottom right) jet, order by jet $p_{\mathrm{T}}$, for the $\mathrm{t\bar{t}H(b\bar{b})}$ (blue) and QCD (green) simulation before applying any event selection. Only jets with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $|\eta| < 2.4$ are considered.

are usually complex and computationally expensive, i.e. many objects have to be reconstructed or complex algorithms have to be run. Since the trigger systems operates under stringent technical constraints, minimizing the computational footprint of a single trigger is an important target during development. This is done on dedicated test systems that are very similar to the actual hardware used for the HLT during data-taking.

For the fully-hadronic $\mathrm{t\bar{t}H(b\bar{b})}$ analysis using $35.9\,\mathrm{fb}^{-1}$ of data collected in 2016 [10], two dedicated paths, selecting events with six or more jets, hight $H_{\mathrm{T}}$ and one or two b-tagged jets, have been developed. Paths following the design are referred to as `HLT_6J1T` and `HLT_6J2T`. While these paths are also used for the $\mathrm{t\bar{t}H(b\bar{b})}$ analysis in the 2017 and 2018 data-taking periods, the development and monitoring of these paths was taken over by other groups within the CMS collaboration since these triggers can also be used to select fully-hadronic $\mathrm{t\bar{t}} + \mathrm{b\bar{b}}$ final states. The author of this thesis consequently developed a new trigger path, which is denoted as `HLT_4J3T`, to further increase the signal efficiency at data-taking for the fully-hadronic $\mathrm{t\bar{t}H(b\bar{b})}$ analysis in the 2017 and 2018 data-taking periods. The previously described monitoring activities for this new path were also the responsibility of the author.

Section A.1 will cover details on the motivation, conception, and development of `HLT_4J3T` path. In section A.2, additional details on the trigger strategy and efficiency measurements, beyond the information given in section 5.1, will be discussed since this represent a significant portion of the work required to finish the analysis and this thesis.

## A.1. Development of a dedicated all-jet trigger for the $\mathrm{t\bar{t}H(b\bar{b})}$ analysis in the 2017 and 2018 data-taking periods.

The two paths used during the 2016 data-taking period, listed in table A.3, select events with $\geq 6$ jets with $p_{\mathrm{T}} > 30(40)$ GeV, $H_{\mathrm{T}} > 450(400)$ GeV and $\geq 1$ (2) b-tagged jets. While the combined usage results in an expected signal efficiency of 97.5 % for the 2016

Table A.1.: Signal efficiency of different trigger combination used in the fully-hadronic $t\bar{t}H(b\bar{b})$ analysis. The efficiency is estimated in $t\bar{t}H(b\bar{b})$ simulation in events with $H_T > 500$, $\geq 6$ jets, six jets with $p_T \geq 40\,\text{GeV}$, $\geq 2$ medium b-tagged jets (Only jets with $p_T \geq 30\,\text{GeV}$ and $|\eta| < 2.4$ are considered) and no e or $\mu$ with $p_T \geq 15\,\text{GeV}$ and $|\eta| < 2.4$. The naming scheme introduced in section 5.1.1 is used and `HLT_Mitigation` refers to either `HLT_PFJet450` (2016) or `HLT_PFHT1050` (2017, 2018). Full combination refers to the "logical or" of `HLT_6J1T`,`HLT_6J2T`, `HLT_4J3T`, and `HLT_Mitigation`

| Paths fired | Signal efficiency [%] | | |
|---|---|---|---|
| | 2016 | 2017 | 2018 |
| `HLT_6J1T` | 85.7 | 69.3 | 83.0 |
| `HLT_6J2T` | 93.8 | 85.4 | 86.3 |
| `HLT_4J3T` | – | 66.4 | 67.8 |
| `HLT_6J1T` or `HLT_6J2T` or `HLT_4J3T` | 97.5 | 93.6 | 95.1 |
| `HLT_Mitigation` | 6.2 | 11.5 | 11.9 |
| **Full combination** | **97.6** | **93.9** | **95.4** |
| `HLT_6J1T` or `HLT_6J2T` | 97.5 | 90.3 | 92.0 |
| `HLT_6J1T` or `HLT_Mitigation` | 86.2 | 70.4 | 83.7 |
| `HLT_6J2T` or `HLT_Mitigation` | 94.2 | 86.6 | 87.5 |
| `HLT_6J1T` or `HLT_4J3T` | – | 86.9 | 92.0 |
| `HLT_6J2T` or `HLT_4J3T` | – | 88.8 | 89.6 |
| `HLT_4J3T` or `HLT_Mitigation` | – | 69.7 | 71.2 |

data-taking period, it was foreseeable that the efficiency would drop significantly with the adaptations necessary following the updated machine configuration for the 2017 data-taking period and onward. In fact, the combination of the adapted versions of these triggers only reaches efficiencies of 90.3 % and 92 % for the 2017 and 2018 data-taking periods, respectively. Details on the efficiencies are given in table A.1. Since signal events with low numbers of online b-tagged jets[2] are covered by these two triggers, one option to further increase the yield is to add another trigger path that requires a higher number of b-tags. Such a tighter threshold on the b-tagging can be used to lower kinematic thresholds, e.g. the $H_T$ or the number of jets. Similar to the number of b-tagged jets, requirements on the jets multiplicity lower than expected from the final state drastically increases the amount signal events that can be captured. The top right plot in figure A.1 shows that a significant amount of signal events exhibit only four or five reconstructed jets and are consequently lost since the available trigger paths only select events with six or more jets. Lowering the $H_T$ threshold as far as possible is crucial since a lot of the signal events exhibit a $H_T$ significantly below the thresholds possible in the six-jet trigger paths,

[2] Generally speaking the online b-tagging performs worse than the offline b-tagging used in the analysis. Therefore, events with a low number of online b-jets can exhibit more b-jets in the analysis when the offline tagger is used. The main reasons for this are retuned or improved versions of the PF and b-tagging algorithms that have not been available at the time of data collection. In case of the b-tagging, it is also very common that completely new algorithms with superior performance are available for the analysis. In addition to improving the efficiency in the signal region, increasing the number of available events with two or three offline b-tags also benefits the analysis since these regions are used for the classifier training and the background estimation.

as it can be seen in figure 5.1.
Following these considerations, the idea for the new dedicated trigger path was to increase the number of b-tagged jets to three and reduce the $H_T$ requirement compared to the present six-jet paths. Setting the minimum number of jets to four (instead of six) further increased the signal efficiency and additionally improved the flexibility of the trigger path such that it can also be used in analyses targeting final states of processes like four-top production, HH $\rightarrow$ 4b, or $b\bar{b}$H, H $\rightarrow$ $b\bar{b}$.
The final configuration of the trigger path targets events with four or more central, i.e. $|\eta| < 2.5$, jets with $p_T \geq 30\,\text{GeV}$, the $H_T$ in the event is required to exceed 300 GeV (330 GeV in 2018) and at least three b-tagged jets are required. Additionally, thresholds are placed on the four leading, i.e. with the highest $p_T$, jets: The first jet requires $p_T \geq 75\,\text{GeV}$, the second $p_T \geq 60\,\text{GeV}$, the third $p_T \geq 45\,\text{GeV}$, and the fourth $p_T \geq 40\,\text{GeV}$. The general structure for the kinematic cuts is influenced by L1-trigger paths that were introduced around the same time. These paths select events by cutting on $H_T$ and $\geq 4$ jets. To control the rate while keeping the $H_T$ threshold comparatively low, cascading cuts on the four leading jets are used. For example the first jet could[3] require $p_T \geq 70\,\text{GeV}$, the second $p_T \geq 55\,\text{GeV}$, the third $p_T \geq 40\,\text{GeV}$ and the fourth $p_T \geq 35\,\text{GeV}$. In order to operate the `HLT_4J3T` trigger in the region where this L1 trigger is fully efficient, similar (slightly higher) cuts are required at the HLT level. In the case of the fully-hadronic $t\bar{t}$H($b\bar{b}$) final state, the amount of signal events rejected by such cuts on the $p_T$ of the four leading jets is relatively small. This can be concluded from the left and middle plots in figure A.1, which show the $p_T$ distributions of the four leading jets in simulated events before any selection. In all four cases, the majority of the $p_T$ distribution is above the values 75 GeV, 60 GeV, 45 GeV, and 40 GeV chosen in the HLT. As discussed earlier, lowering the $H_T$ threshold as much as possible was the main motivation to develop this path. The threshold is set to 300 (330) GeV in the 2017 (2018) data-taking period. This is the minimum value possible considering the point of full efficiency of the corresponding L1 triggers. Another value that can be used to control the rate is the threshold on the online b-tagging discriminator value. Initial tests showed that the threshold can be chosen comparatively loosely. For example in the 2018 data-taking period three jets are required to exhibit a PF-b-tagging discriminant value $> 0.24$, which corresponds to a mis-tag rate of $\sim 4.5\,\%$. As comparison, the path targeting events with $\geq 6$ Jets, $H_T \geq 400\,\text{GeV}$ and $\geq 2$ b-jets (`HLT_6J2T`) uses a threshold of 0.3, which corresponds to a mis-tag rate of $\sim 2.94\,\%$[4]. This means that the new trigger can also catch events that fail the other two triggers because of their tighter b-tagging threshold but still contain a third (or second and third) jet passing the looser b-tagging threshold in the new path.
The new `HLT_4J3T` trigger path reaches signal efficiencies of 66.4 % and 67.8 % in the 2017 and 2018 data-taking periods, respectively. While on its own this is arguably a low efficiency for a $t\bar{t}$H($b\bar{b}$) analysis, it was never intended to be used like that but instead should augment the combination of the `HLT_6J1T` and `HLT_6J2T` trigger paths, which already cover large parts of the phase space. The combination of the three triggers reaches signal efficiencies of 93.6 % and 95.1 % in the 2017 and 2018 data-taking periods, respectively. This corresponds to a recovery of $\sim 3$ percentage points in signal efficiency compared to not having this trigger in each year.

As discussed in section 2.2.6, trigger paths are built as a sequence of modules that either reconstruct objects in the event or filter events based on some thresholds. If an event does not pass a filtering step, the sequence is stopped immediately in order to minimize runtime. Table A.2 shows the sequence for the `HLT_4J3T` trigger path as it was configured during the 2018 data-taking period. The configuration used during the 2017 data-taking

[3] Various configuration are used depending on the instantaneous luminosity.
[4] Lower mis-tag rates (almost) always mean that less real b-jets will be b-tagged as well.

Table A.2.: Full sequence of HLT modules for the `HLT_4J3T` path as used during the 2018 data-taking period (CMS-internal name: `HLT_PFHT330PT30_QuadPFJet_75_60_45_40_TriplePFBTagDeepCSV_4p5`). The sequence runs through the listed modules sequentially from top to bottom. Modules marked as "Filter" are used to select events based on kinematic properties. If an event fails the requirement of a given filter, it is rejected and the sequence is stopped.

| HLT module name | | Explanation |
|---|---|---|
| `L1 seed` | | Combination of various L1 seeds |
| `Prescale` | | = 1; Path is not prescaled |
| `HLTAK4CaloJetsSequence` | | Reconstructs "calo-jets" only using ECAL+HCAL |
| `hltQuadCentralJet30` | Filter | Keep: $\geq 4$ calo-jets w/ $p_\mathrm{T} > 30\,\mathrm{GeV}$ and $\lvert\eta\rvert < 2.5$ |
| `hltCaloJetsQuad30ForHt` | | Produce calo-jet collection for $H_\mathrm{T}$ calculation. All calo-jets w/ $p_\mathrm{T} > 30\,\mathrm{GeV}$ and $\lvert\eta\rvert < 2.5$ are considered. |
| `hltHtMhtCaloJetsQuadC30` | | Calculate $H_\mathrm{T}$ from calo-jet collection |
| `hltCaloQuadJet30HT320` | Filter | Keep: calo-jet $H_\mathrm{T} > 320\,\mathrm{GeV}$ |
| `HLTBtagDeepCSVSequenceL3` | | Run b-tagging for calo-jets |
| `hltBTagCaloDeepCSVp17Double` | Filter | Keep: $\geq 2$ b-tagged calo jets (b-tag value $\geq 0.17$) |
| `HLTAK4PFJetsSequence` | | Reconstructs Particle flow (PF) jets |
| `hltPFCentralJetLooseIDQuad30` | Filter | Keep: $\geq 4$ PF-jets w/ $p_\mathrm{T} > 30\,\mathrm{GeV}$ and $\lvert\eta\rvert < 2.5$ |
| `hlt1PFCentralJetLooseID75` | Filter | Keep: Leading jet $p_\mathrm{T} > 75\,\mathrm{GeV}$ |
| `hlt2PFCentralJetLooseID60` | Filter | Keep: Second leading jet $p_\mathrm{T} > 60\,\mathrm{GeV}$ |
| `hlt3PFCentralJetLooseID45` | Filter | Keep: Third leading jet $p_\mathrm{T} > 45\,\mathrm{GeV}$ |
| `hlt4PFCentralJetLooseID40` | Filter | Keep: Fourth leading jet $p_\mathrm{T} > 40\,\mathrm{GeV}$ |
| `hltPFCentralJetLooseIDQuad30forHt` | | Produce PF-jet collection for $H_\mathrm{T}$ calculation. All PF-jets w/ $p_\mathrm{T} > 30\,\mathrm{GeV}$ and $\lvert\eta\rvert < 2.5$ are considered |
| `hltHtMhtPFCentralJetsLooseIDQuadC30` | | Calculate $H_\mathrm{T}$ from PF-jet collection |
| `hltPFCentralJetsLooseIDQuad30HT330` | Filter | Keep: PF-jet $H_\mathrm{T} > 330\,\mathrm{GeV}$ |
| `HLTBtagDeepCSVSequencePF` | | Run b-tagging for PF-jets |
| `hltBTagPFDeepCSV4p5Triple` | Filter | Keep: $\geq 3$ b-tagged PF-jets (b-tag value $\geq 0.24$) |

period follows an almost identical sequence. Only certain threshold and the b-tagging algorithm differ between the configurations.

This section will walk through a full HLT sequence by the example of the `HLT_4J3T` trigger path:
The sequence starts with a "logical or" of multiple $H_{\mathrm{T}}$-only and $H_{\mathrm{T}}$ + four-jets L1 trigger paths. This is followed by a prescale module, which specifies the prescale at a certain instantaneous luminosity, that every HLT sequence needs to include. In this case, the prescale is set to one over the full instantaneous luminosity range. This means that the trigger is fully unprescaled. In order to minimize computing time, a jet reconstruction and selection using only calorimeter information is used as a preselection before the slower and more complex but also more accurate PF reconstruction is run. After running the calorimeter-based reconstruction (resulting in so-called calo-jets), a filter requiring at least four calo-jets with $p_{\mathrm{T}} \geq 30\,\mathrm{GeV}$ and $|\eta| < 2.5$ is placed in the sequence. The next two modules calculate the calo-jet $H_{\mathrm{T}}$, which is used in a filter step requiring it to exceed 320 GeV. The final two modules operating on calorimeter information are calorimeter-based b-tagging modules. The b-tagging is based on the `DeepCSV` algorithm. Only events exhibiting at least two (very) loosely b-tagged calo-jets pass the corresponding filter. Since these modules are only intended as preselection and the results of the PF reconstruction can be different, the calorimeter-based filters generally operate with lower thresholds than targeted in the filters using the superior PF objects. For the remaining events, a jet reconstruction with the PF algorithm is run. In a first filtering step, events that do not exhibit at least four PF-jets with $p_{\mathrm{T}} \geq 30\,\mathrm{GeV}$ and $|\eta| < 2.5$ are rejected. The next four modules check if the first PF-jet (sorted by $p_{\mathrm{T}}$) exceeds a $p_{\mathrm{T}}$ of 75 GeV, the second exceeds 60 GeV, the third exceeds 45 GeV, and the fourth exceeds 40 GeV. For events passing these filtering steps, the $H_{\mathrm{T}}$ of the PF-jets is calculated and $H_{\mathrm{T}} > 330\,\mathrm{GeV}$ is required in a filtering step. The second to last module is running the b-tagging algorithm. As for the calorimeter-based modules, the `DeepCSV` algorithm is used. The last module rejects events that do not exhibit at least three PF-jets with a b-tagging value of 0.24 or higher. This threshold corresponds to a mis-tag rate of $\sim 4.5\,\%$. During the 2017 data-taking, the corresponding trigger path used the `CSVv2` algorithm and a mis-tag rate of $\sim 3.0\,\%$, i.e. a tighter threshold, was required in the filtering step at roughly the same rate.

Overall, adding the `HLT_4J3T` trigger path in the 2017 and 2018 data-taking periods enabled a partial recovery of the signal efficiency lost due to increased thresholds, that where necessary because of changed LHC machine parameters, in the `HLT_6J1T` and `HLT_6J2T` trigger paths. Both configurations (for the 2017 and 2018 data-taking) of the `HLT_4J3T` trigger path operated robustly without excessive rates or computing resource consumption over their operational time. No prescale or reconfiguration during the individual data-taking periods was necessary, resulting in very sable efficiencies and a minimal run dependency.

## A.2. Additional information of the trigger strategy in the fully-hadronic $\mathrm{t\bar{t}H(b\bar{b})}$ channel

This section collects various additional information and material referenced in section 5.1. The combined trigger efficiency turn-ons for the 2016 and 2017 data-taking periods are shown in figures A.2 and A.3, respectively.

### Configuration of the considered trigger paths

Tables A.3, A.4, and A.5 list the individual high-level trigger paths used during the 2016, 2017 and 2018 data-taking periods, respectively.

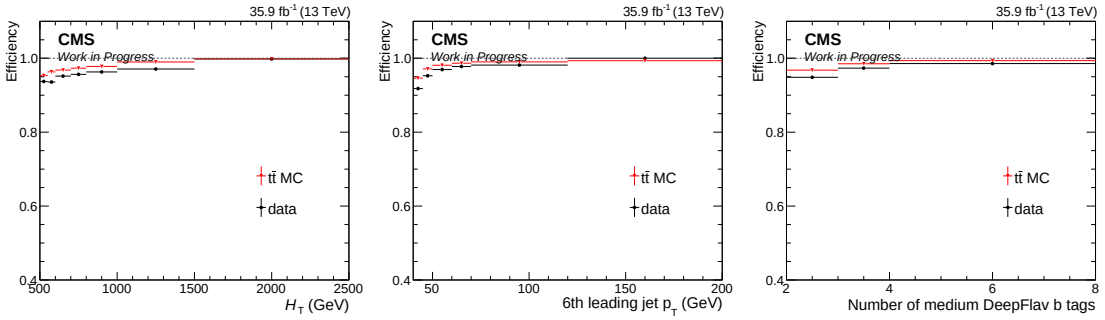



Figure A.2.: Trigger efficiency measured in $H_T$ (left), $p_T$ of the sixth jet (middle) and number of medium b-tagged jets (right) in the 2016 dataset with respect to `HLT_IsoMu24` in events with one offline lepton and passing the fully-hadronic baseline selection (excluding the lepton veto). The efficiency is measured in data (black) and $t\bar{t}$ + jets simulation (red).

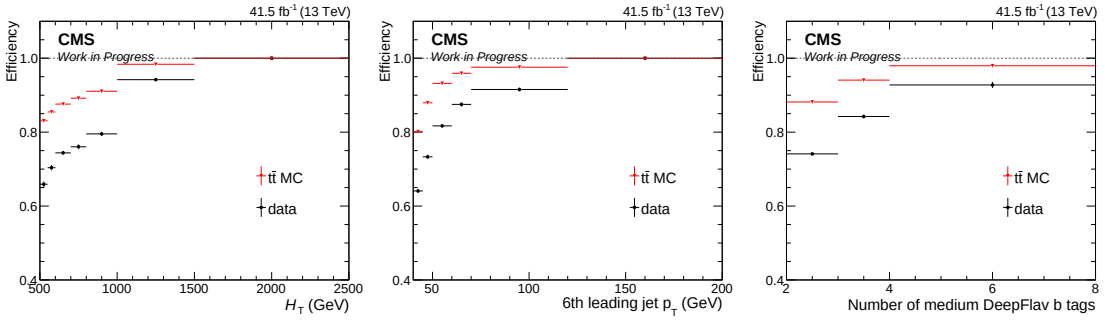



Figure A.3.: Trigger efficiency measured in $H_T$ (left), $p_T$ of the sixth jet (middle) and number of medium b-tagged jets (right) in the 2017 dataset with respect to `HLT_IsoMu27` in events with one offline lepton and passing the fully-hadronic baseline selection (excluding the lepton veto). The efficiency is measured in data (black) and $t\bar{t}$ + jets simulation (red).

Table A.3.: High level trigger paths used in the 2016 data-taking period. The name and relevant cuts are given in the table. The threshold of the b-tagging is given in the name (`BTagCSV_p*`) and is specified in terms of the discriminator value of the `CSVv2` algorithm.

| Path (CMS-internal naming scheme) | Relevant Cuts |
|---|---|
| `HLT_PFHT450_SixJet40_BTagCSV_p056` | PF $H_T \geq 450\,\text{GeV}$<br>$\geq 6$ PF jets<br>$p_{T,6} \geq 40\,\text{GeV}$<br>$\geq 1$ b-tagged PF jets |
| `HLT_PFHT400_SixJet30_DoubleBTagCSV_p056` | PF $H_T \geq 400\,\text{GeV}$<br>$\geq 6$ PF Jets<br>$p_{T,6} \geq 30\,\text{GeV}$<br>$\geq 2$ b-tagged PF jets |
| `HLT_PFJet450` | $\geq 1$ PF jets with $p_T > 450\,\text{GeV}$ |

Table A.4.: High level trigger paths used in the 2017 data-taking period. The dataset, name, and relevant cuts for all triggers are given in the table. During data-taking (between runs B and C) the names of the paths changed but the configuration stayed the same. The table lists both names. The threshold of the b-tagging is given in the path name (`BTagCSV_p*`) and is specified in terms of the mis-tag rate (or the discriminator value in the case of Run-B) of the `CSVv2` algorithm. During data-taking, events are saved in different dataset depending on the fired trigger paths. The (CMS internal) dataset is specified in the *Dataset* column. In simulation, the configuration listed first is used.

| Dataset | Path (CMS-internal naming scheme) | Relevant Cuts |
|---|---|---|
| JetHT | `HLT_PFHT430_SixPFJet40_PFBTagCSV_1p5`<br>`HLT_PFHT430_SixJet40_BTagCSV_p080` (Run-B) | PF $H_{\mathrm{T}} \geq 430\,\mathrm{GeV}$<br>$\geq 6$ PF jets<br>$p_{T,6} \geq 40\,\mathrm{GeV}$<br>$\geq 1$ b-tagged PF Jet |
| JetHT | `HLT_PFHT380_SixPFJet32_DoublePFBTagCSV_2p2`<br>`HLT_PFHT380_SixJet32_DoubleBTagCSV_p075` (Run-B) | PF $H_{\mathrm{T}} \geq 380\,\mathrm{GeV}$<br>$\geq 6$ PF jets<br>$p_{T,6} \geq 32\,\mathrm{GeV}$<br>$\geq 2$ b-tagged PF jets |
| BTagCSV | `HLT_PFHT300PT30_QuadPFJet_75_60_45_40_TriplePFBTagCSV_3p0`<br>`HLT_HT300PT30_QuadJet_75_60_45_40_TripleCSV_p07` (Run-B) | PF $H_{\mathrm{T}} \geq 300\,\mathrm{GeV}$<br>$\geq 4$ PF jets<br>$p_{T,1} \geq 75\,\mathrm{GeV}$<br>$p_{T,2} \geq 60\,\mathrm{GeV}$<br>$p_{T,3} \geq 45\,\mathrm{GeV}$<br>$p_{T,4} \geq 40\,\mathrm{GeV}$<br>$\geq 3$ b-tagged PF jets |
| JetHT | `HLT_PFHT1050` | PF $H_{\mathrm{T}} \geq 1050\,\mathrm{GeV}$ |

Table A.5.: High level trigger paths used in the 2018 data-taking period. The name and relevant cuts are given in the table. The threshold of the b-tagging is given in the path name (`BTagDeepCSV_p*`) and is specified in terms of the mis-tag rate of the `DeepCSV` algorithm. Several, slightly different, configuration of the HLT paths with $\geq 6$ PF jets have been used during data-taking. The impact on the overall trigger efficiency was estimated be be negligible. Before run 315974 (CMS internal run numbers), the `CSVv2` algorithm was used for the path with $\geq 6$ PF jets and $\geq 1$ PF b-jet. Due to prescaling of certain L1-trigger paths, both paths with $\geq 6$ PF jets had to be retuned in order to reduce their rate. In simulation, the configuration listed first is used.

| Path (CMS-internal naming scheme) | Relevant Cuts |
|---|---|
| `HLT_PFHT450_SixPFJet36_PFBTagDeepCSV_1p59`<br>`HLT_PFHT430_SixPFJet40_PFBTagCSV_1p5` (Runs 315252-315974)<br>`HLT_PFHT430_SixPFJet40_PFBTagDeepCSV_1p5` (Runs 315974-317509) | PF $H_{\mathrm{T}} \geq 450$ (430) GeV<br>$\geq 6$ PF jets<br>$p_{T,6} \geq 36$ (40) GeV<br>$\geq 1$ b-tagged PF Jet |
| `HLT_PFHT400_SixPFJet32_DoublePFBTagDeepCSV_2p94`<br>`HLT_PFHT380_SixPFJet32_DoublePFBTagDeepCSV_2p2` (Runs 315252-317509) | PF $H_{\mathrm{T}} \geq 400$ (380) GeV<br>$\geq 6$ PF jets<br>$p_{T,6} \geq 32\,\mathrm{GeV}$<br>$\geq 2$ b-tagged PF jets |
| `HLT_PFHT330PT30_QuadPFJet_75_60_45_40_TriplePFBTagDeepCSV_4p5` | PF $H_{\mathrm{T}} \geq 330\,\mathrm{GeV}$<br>$\geq 4$ PF jets<br>$p_{T,1} \geq 75\,\mathrm{GeV}$<br>$p_{T,2} \geq 60\,\mathrm{GeV}$<br>$p_{T,3} \geq 45\,\mathrm{GeV}$<br>$p_{T,4} \geq 40\,\mathrm{GeV}$<br>$\geq 3$ b-tagged PF jets |
| `HLT_PFHT1050` | PF $H_{\mathrm{T}} \geq 1050\,\mathrm{GeV}$ |

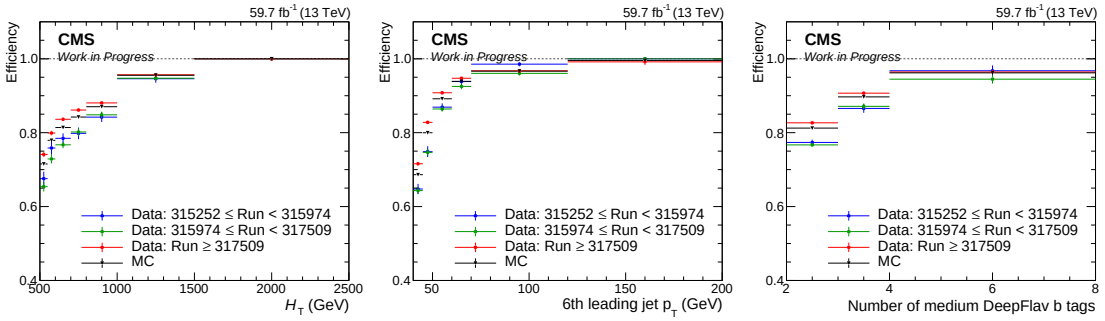



Figure A.4.: Closure tests in $H_T$ (left), $p_T$ of the sixth jet (middle) and number of medium b-tagged jets (right) for each trigger configuration used during the 2018 data-taking period in data (blue, green, and red) and the corresponding configuration in simulation (black). The scale factors derived from the efficiency measurement shown in figure 5.2 are applied to the simulation. Reasonable agreement of the corrected simulation with all run periods can be observed. The run-number given in the legend refers to the CMS-internal numbering scheme. The temporal order of the data distribution is blue → green → red. The red distribution uses the configuration that is also used in simulation (black).

The trigger names given in the tables represent the CMS-internal naming scheme and the listed cuts omit less relevant information like the exact b-tagging discriminator value (only the number of tags is shown but the value can be extracted from the name), the used L1-seeds, or the calorimeter-based criteria. If more than one configuration is listed for a given path, it is only relevant when analyzing the data. In simulation, the configuration listed first is used.
During the 2016 and 2017 data-taking periods, the thresholds of all considered paths stayed the same. Nevertheless, Table A.4 lists two distinct paths for each of the `HLT_6J1T`, `HLT_6J2T` and `HLT_4J3T` triggers, that only differ in their respective naming scheme (configuration is identical)[5], for the 2017 data-taking. During the 2018 data-taking period, the `HLT_6J1T` and `HLT_6J2T` paths required multiple configuration changes. In the beginning of the data-taking period, `HLT_6J1T` used the `CSVv2` algorithm instead of the intended `DeepCSV` algorithm. Furthermore, both paths required an increase in kinematic thresholds ($H_T$ and minimum PF-jet $p_T$ for the jet multiplicity calculation) in order to reduce the trigger rate such that prescaling was not necessary. Depending on the differences in trigger efficiency and turn-on, such changes might make it necessary to consider individual data/simulation correction for each trigger path combination active at a certain time during data-taking. Generally, it is preferable to derive data/simulation corrections for the triggers on the whole dataset since the statistical power of the measurement is a limiting factor. Furthermore, it would introduce additional complexities in context of the corresponding uncertainties considered in the signal extraction. In the case of this analysis, the correction factors are derived on the whole dataset. This is valid since the variations in the configurations are small. Closure tests were performed with the correction derived from the full dataset and the different combinations active at a certain time during data-taking. The closure tests are shown in figure A.4 and the differences are deemed small enough such that they can be neglected.

[5] In the beginning of the 2017 data-taking period it was decided, that multiple naming conventions, used in the 2016 data-taking period, should not be used anymore in order to make the naming more clear. Among others the used reconstruction method (e.g. `PF` or `Calo`) should be added to the name and the mis-tag rate used for the b-tagging should be given instead of the discriminant value.

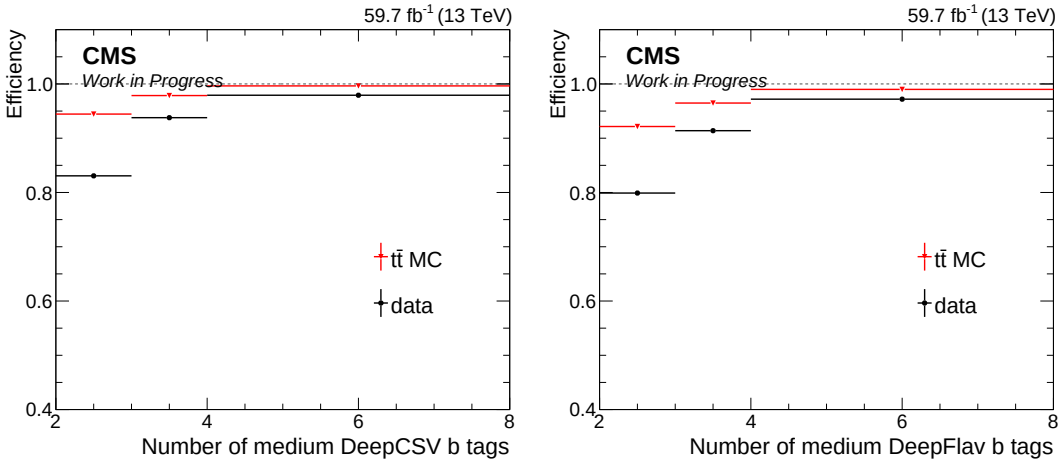



Figure A.5.: Efficiency of the b-tagging portion of the `HLT_4J3T` trigger paths during the 2018 data-taking period. Instead of measuring the trigger efficiency relative to a single muon trigger for this measurement, the efficiency is measured relative to a prescaled paths that exhibits exactly the same sequence as `HLT_4J3T` but without the calorimeter-based and PF-based b-tagging modules. Both plots show the trigger efficiency as a function of the number of medium b-tagged jets but the `DeepCSV` algorithm was used for the left plot and the `DeepJet` for the right plots. Therefore, the left plot uses the same algorithm at trigger level and analysis level.

### Effect of using different online and offline b-tagging algorithms

As mentioned in the discussion of the data/simulation correction in section 5.1.2, the online b-tagging algorithms are different to the algorithm used offline in the analysis. Generally different algorithms at these two stages should have no influence on the analysis itself. In any case, it is not possible to change the algorithm after the data is collected and using the superior tagger in the analysis, if available, is without any alternative. Nevertheless understanding the impact is necessary in order to correctly interpret the trigger efficiency measurements. The measured efficiencies are effected by the choice of the b-tagging algorithm used offline because the correction is parametrized using the b-tagging multiplicity and differences in tagging performance lead to shifts in b-tagging related observable that effect the signal and reference triggers in different ways. It can be observed that the maximum reachable efficiency gets smaller and the turn-on is slower when the relative performance between the online and offline b-tagging improves. This means that the measured efficiency may never reach 100 %, which is generally not desired. Since the offline `DeepJet` algorithm significantly outperforms both algorithms used online (`CSVv2` during the 2016 and 2017 data-taking and `DeepCSV` during the 2018 data-taking) the previously described features can be studied by comparing the efficiency measurements using different offline b-tagging algorithms. One example in which this can be seen very well is to exclusively look at the b-tagging portion of a given trigger path. Figure A.5 shows the efficiency of the b-tagging portion for the `HLT_4J3T` path in the 2018 data-taking period. The slower turn-on in the efficiency using `DeepJet`, which also does not reach 100 %, compared to the efficiency using `DeepCSV` (in the latter case the same algorithm is used online and offline) is solely an artifact of the considered algorithm at analysis-level and not the trigger itself. Similar behavior can be observed in the efficiencies of the fully strategy. This means, that a slow turn-on and low maximum efficiency, as they can be observed in the efficiency measurements with the `DeepJet` algorithm, are no concerning features.

### Systematic uncertainty in the data/simulation correction

As discussed in section 5.1.2, the uncertainty on the trigger data/simulation correction in the fit is taken from the statistical uncertainty of the correction factor measurement and sources of systematic uncertainty are neglected. The motivation for this is the understanding that the statistical uncertainty is by far the dominant source of uncertainty to the correction factors. This can be mainly attributed to the rather small overlap between single-muon and jet+b-tag triggers in the analysis phase space. Additionally, the contribution of uncertainty for the trigger data/simulation correction to the total uncertainty of the measurement is very small (the impact is essentially negligible). Nevertheless, several possible sources of systematic uncertainty have been investigated. The correlation to the reference trigger, pileup dependencies, and a dependency on the evolution of the machine parameters over the data-taking (run dependency): The correlation between the reference trigger and the combination of the signal triggers was estimated to be significantly below 1 % using the formula $(\epsilon(\text{Reference}) \cdot \epsilon(\text{Signal}))/\epsilon(\text{Reference and Signal})$ with the trigger efficiencies $\epsilon$ in all data-taking periods. As mentioned multiple times, all-jet triggers exhibit a large dependency on the pileup which can also introduce a systematic uncertainty. In the case of the data/simulation correction for the trigger strategy in the fully-hadronic $t\bar{t}H(b\bar{b})$ analysis, it is expected that this dependency is modeled well since the correction is parametrized as function of $H_T$, which is also strongly correlated to the pileup. Since the trigger configuration is very stable over the individual data-taking periods (and the pileup dependency is covered by the parametrization) run dependency was observed to be minimal. Even in the case of the 2018 data-taking period, in which the configurations changed to most, the dependency is negligible as discussed previously.

### Double counting in the 2017 data-taking period

For technical reasons, events are saved in one or more distinct datasets depending on the trigger paths a given event is passing. Events passing a single-muon trigger are saved in the so-called `SingleMuon` dataset while event passing a $H_T$ based trigger are saved in the `JetHT` dataset. If the event passes both types of trigger it will be saved in both datasets. This enables that analyses to process only the events actually needed based on the utilized type of trigger. During the 2017 data-taking period, event passing `HLT_4J3T` trigger and the `HLT_6J1T`/`HLT_6J2T` triggers were saved to different datasets. The former to a dataset called `BTagCSV` and the latter two to one called `JetHT`. Since the overlap between the three paths is quite large, a scheme was established in order to excluded the possibility of double counting events. All events passing any combination including `HLT_4J3T` are taken from `BTagCSV` and only events passing at least one of the `HLT_6J1T`/`HLT_6J2T` triggers (but not `HLT_4J3T`) are taken from `JetHT`. Further details are given in table A.6. Events that fire `HLT_PFHT1050` but none of the other three triggers are taken from the `JetHT` dataset.

Table A.6.: During the 2017 data-taking period event passing the `HLT_4J3T` path are saved in the `BTagCSV` dataset while event passing `HLT_6J1T` and `HLT_6J2T` paths are saved to a different (and completely separate) `JetHT` dataset. In order to exclude the possibility of double counting data events, the following scheme was developed. The table has to be understood as follows: The "x" denotes if a trigger path was fired. Multiple "x" in one column mean that the triggers fired on the same event.

| Trigger | Event taken from | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | `BTagCSV` Dataset | | | | `JetHT` Dataset | | | |
| 6J1T | x | | | x | x | | | x |
| 6J2T | x | | x | | x | x | | |
| 4J3T | x | x | x | x | | | | |

# B. Additional information on object and event selections

This appendix collects various tables summarizing object and event selections relevant for the $t\bar{t}H(b\bar{b})$ analyses presented in this thesis.

Table B.1.: Summary of the cut values for the tight muon id used in the $t\bar{t}H(b\bar{b})$ analysis.

| Parameter | Cut value |
|---|---|
| $\chi^2$/ndof | $< 10$ |
| muon chamber hit | $> 1$ |
| muon station with muon segments | $> 2$ |
| $d_{xy}$ | $< 2\,\mathrm{mm}$ |
| $d_z$ | $< 5\,\mathrm{mm}$ |
| number of pixel hits | $> 0$ |
| number of tracker layers with hits | $> 5$ |

Table B.2.: Summary of the cut values for the tight electron id used in the $t\bar{t}H(b\bar{b})$ analysis. Different cut values are used for the two regions in $\eta_{supercluster}$ with $\leq 1.479$ and $> 1.479$.

| Parameter | Cut value | |
|---|---|---|
| | $\lvert\eta_{supercluster}\rvert$ | |
| | $\leq 1.479$ | $> 1.479$ |
| 5x5 $\sigma_{i\eta i\eta}$ | $< 0.022$ | $< 0.0353$ |
| $\lvert\Delta\eta_{seed}\rvert$ | $< 0.00255$ | $< 0.00501$ |
| $\lvert\Delta\phi_{in}\rvert$ | $< 0.022$ | $< 0.0236$ |
| $H/E$ | $< 0.079$ | $< 0.0637$ |
| $\lvert 1/E - 1/\rho\rvert$ | $< 0.159$ | $< 0.0197$ |
| $N_{miss}$ | $< 1$ | $< 1$ |
| $d_{xy}$ | $< 0.05\,\mathrm{cm}$ | $< 0.1\,\mathrm{cm}$ |
| $d_z$ | $< 0.1\,\mathrm{cm}$ | $< 0.2\,\mathrm{cm}$ |
| conversion veto | pass | pass |
| Relative Isolation | $< 0.046$ | $< 0.0759$ |

Table B.3.: Summary of the cut values for the tight jet identification with lepton veto used in the $t\bar{t}H(b\bar{b})$ analysis.

| Parameter | Cut value |
|---|---|
| Neutral hadron fraction | $< 0.9$ |
| Neutral EM fraction | $< 0.9$ |
| Number of constituents | $> 1$ |
| Charged Hadron fraction | $> 0$ |
| Charged multiplicity | $> 0$ |
| Muon fraction | $< 0.8$ |
| Charged EM fraction | $> 0.8$ |

Table B.4.: Baseline selection for all $t\bar{t}H(b\bar{b})$ analysis channels considered in the Run-2 combination. If the criteria differ per data-taking period, they are quoted as 2016/2017/2018.

| | FH | SL | DL |
|---|---|---|---|
| Number of leptons | 0 | 1 | 2 |
| Sign and flavour of leptons | — | $e^{\pm}$, $\mu^{\pm}$ | $e^{+}e^{-}$, $\mu^{\pm}e^{\mp}$, $\mu^{+}\mu^{-}$ |
| Min. $p_{\mathrm{T}}$ of $p_{\mathrm{T}}$-leading electron [GeV] | — | 29/30/30 | 25 |
| Min. $p_{\mathrm{T}}$ of $p_{\mathrm{T}}$-leading muon [GeV] | — | 26/29/26 | 25 |
| Max. $p_{\mathrm{T}}$ of additional leptons [GeV] | 15 | 15 | 15 |
| Max. $\lvert\eta\rvert$ of leptons | 2.4 | 2.4 | 2.4 |
| Min. number of jets | 6 | 4 | 2 |
| Min. $p_{\mathrm{T}}$ of jets [GeV] | 30 | 30 | 30 |
| Min. $p_{\mathrm{T}}$ of $6^{th}$ jets [GeV] | 40 | — | — |
| Max. $\lvert\eta\rvert$ of jets | 2.4 | 2.4 | 2.4 |
| Min. number of b tagged jets | 2 | 3 | 1 |
| Min. $H_{\mathrm{T}}$ [GeV] | 500 | — | — |
| Min. $m_{e^{+}e^{-}/\mu^{+}\mu^{-}}$ [GeV] | — | — | 20 |
| $m_{ee/\mu\mu}$ [GeV] | — | — | $< 76$, $> 106$ |
| Min. $p_{\mathrm{T}}^{\mathrm{miss}}$ [GeV] | — | 20 | 40 |
| $m_{\mathrm{qq}}$ [GeV] | $> 30$, $< 250$ | — | — |

# C. ANN classifier input variables

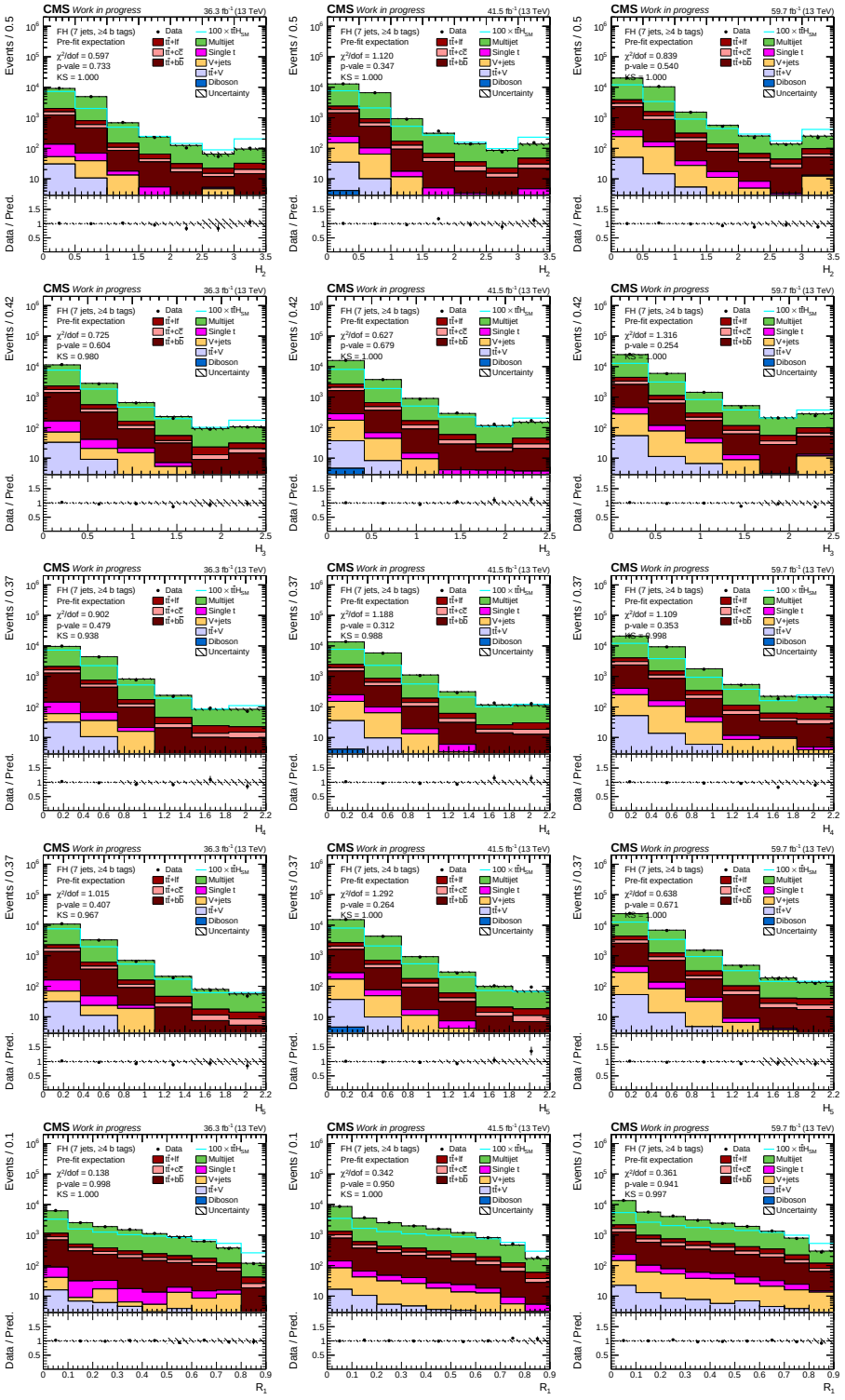



Figure C.1.: ANN input variables $H_2$, $H_3$, $H_4$, $H_5$, $R_1$ for the (7 jets, $\geq 4$ b-tags) category for 2016 (left), 2017 (middle) and 2018 (right). The hashed band visualizes the prefit uncertainty including all uncertainties used in the final fit. Additionally a 50% normalization for the $t\bar{t}+B$ and $t\bar{t}+C$ processes is considered instead of the free floating normalization used in the final fit. The QCD shown in these plots is the data driven estimation as described in section 5.5.

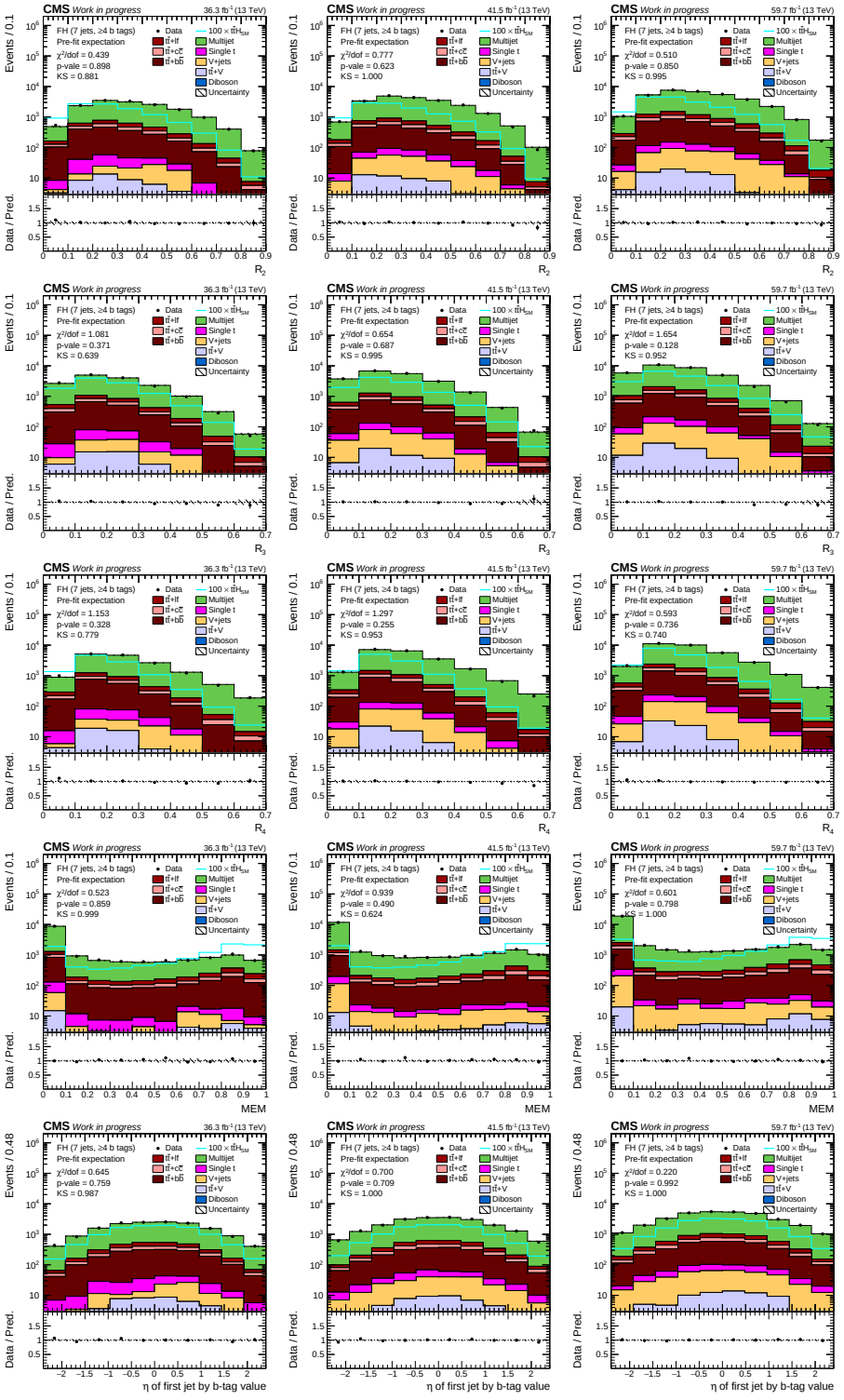



Figure C.2.: ANN input variables $R_2$, $R_3$, $R_4$, MEM and $\eta$ of the $1^{st}$ jet by b-tag for the ($7\,\text{jets}, \geq 4\,\text{b-tags}$) category for 2016 (left), 2017 (middle) and 2018 (right). The hashed band visualizes the prefit uncertainty including all uncertainties used in the final fit. Additionally a 50% normalization for the $t\bar{t} + B$ and $t\bar{t} + C$ processes is considered instead of the free floating normalization used in the final fit. The QCD shown in these plots is the data driven estimation as described in section 5.5.

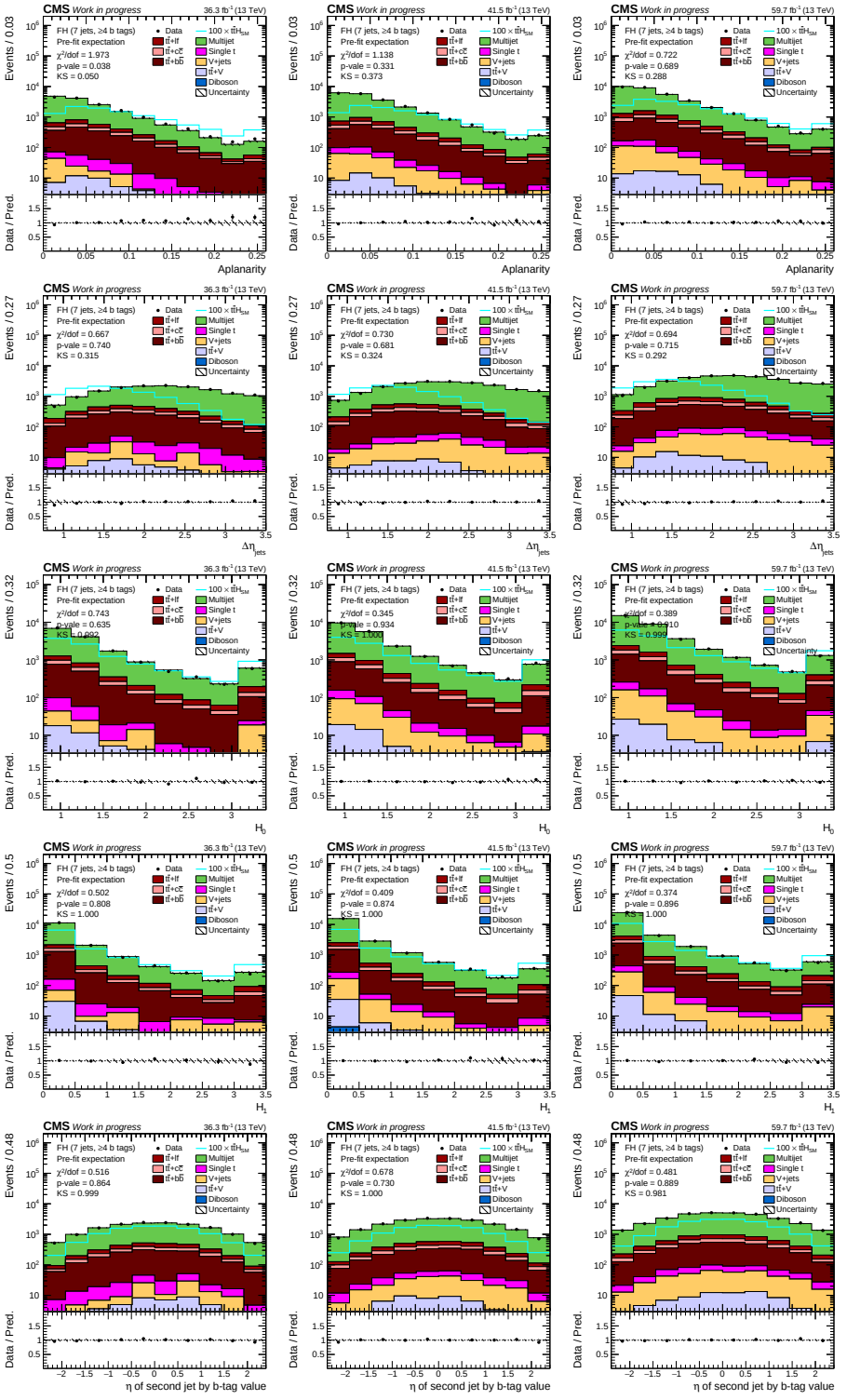



Figure C.3.: ANN input variables Aplanarity, $\Delta\eta_{\text{jets}}$, $H_0$, $H_1$ and $\eta$ of the $2^{nd}$ jet by b-tag for the ($7\,\text{jets}, \geq 4\,\text{b-tags}$) category for 2016 (left), 2017 (middle) and 2018 (right). The hashed band visualizes the prefit uncertainty including all uncertainties used in the final fit. Additionally a 50% normalization for the $\text{t}\bar{\text{t}} + \text{B}$ and $\text{t}\bar{\text{t}} + \text{C}$ processes is considered instead of the free floating normalization used in the final fit. The QCD shown in these plots is the data driven estimation as described in section 5.5.

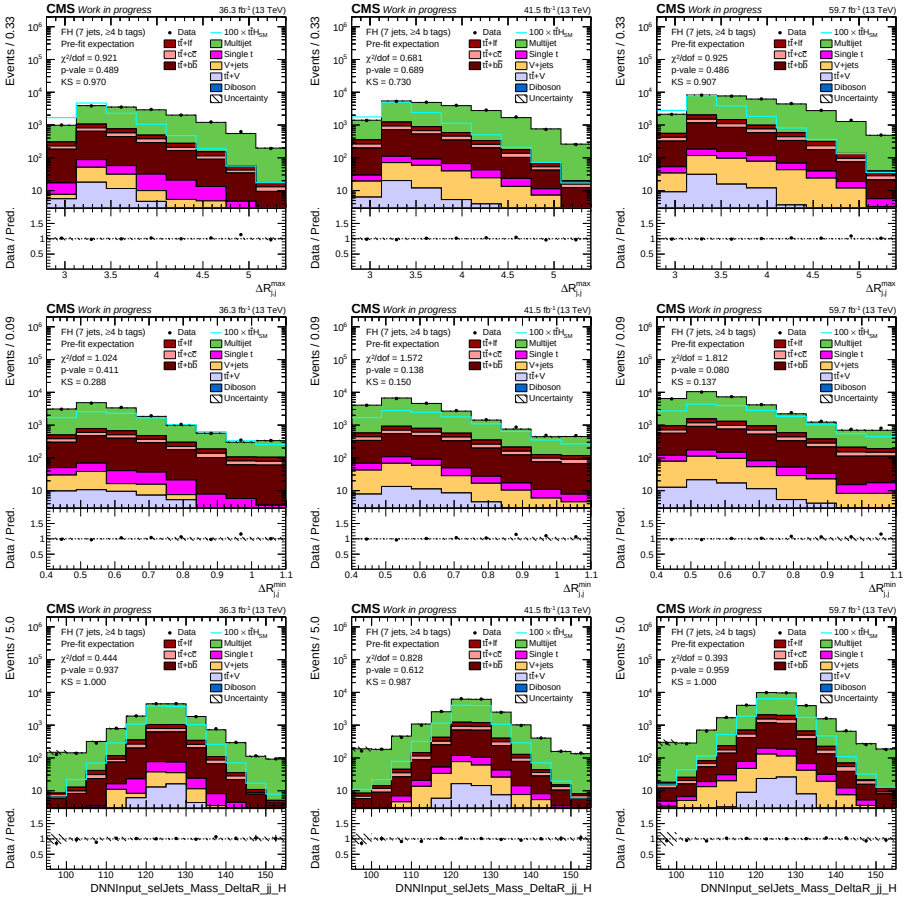



Figure C.4.: ANN input variables $\Delta R^{\text{max}}_{\text{j,j}}$, $\Delta R^{\text{min}}_{\text{j,j}}$ and $m^{\text{closest to 125}}_{\text{j,j}}$ for the ($7\,\text{jets}, \geq 4\,\text{b-tags}$) category for 2016 (left), 2017 (middle) and 2018 (right). The hashed band visualizes the prefit uncertainty including all uncertainties used in the final fit. Additionally a 50% normalization for the $\text{t}\bar{\text{t}}+\text{B}$ and $\text{t}\bar{\text{t}}+\text{C}$ processes is considered instead of the free floating normalization used in the final fit. The QCD shown in these plots is the data driven estimation as described in section 5.5.

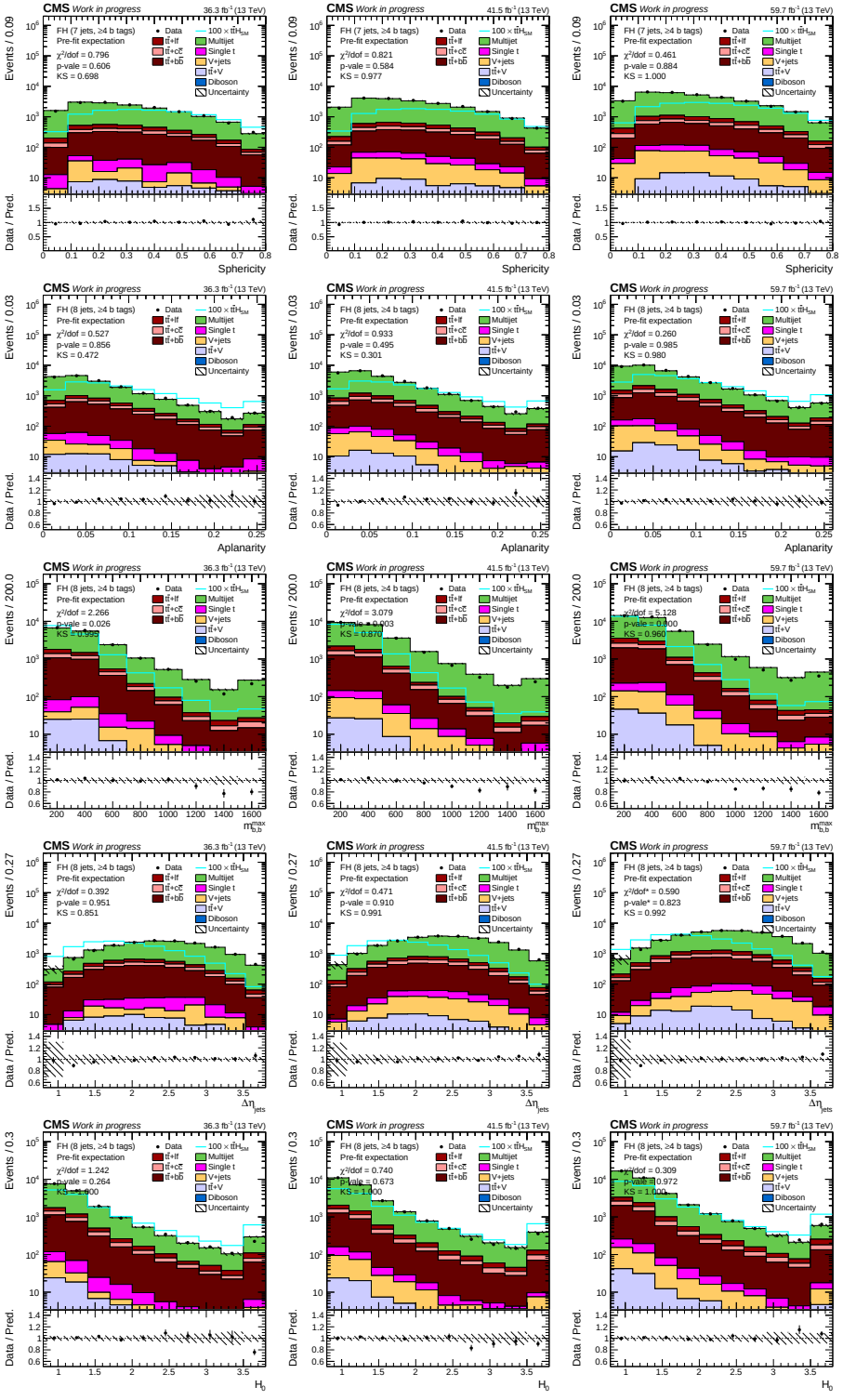



Figure C.5.: ANN input variables Sphericity for the ($7\,\text{jets}, \geq 4\,\text{b-tags}$) category as well as Aplanarity, $m_{\text{b,b}}^{\text{max}\,m}$, $\Delta\eta_{\text{jets}}$ and $\text{H}_0$ for the ($8\,\text{jets}, \geq 4\,\text{b-tags}$) category for 2016 (left), 2017 (middle) and 2018 (right). The hashed band visualizes the prefit uncertainty including all uncertainties used in the final fit. Additionally a 50% normalization for the $\text{t}\bar{\text{t}} + \text{B}$ and $\text{t}\bar{\text{t}} + \text{C}$ processes is considered instead of the free floating normalization used in the final fit. The QCD shown in these plots is the data driven estimation as described in section 5.5.

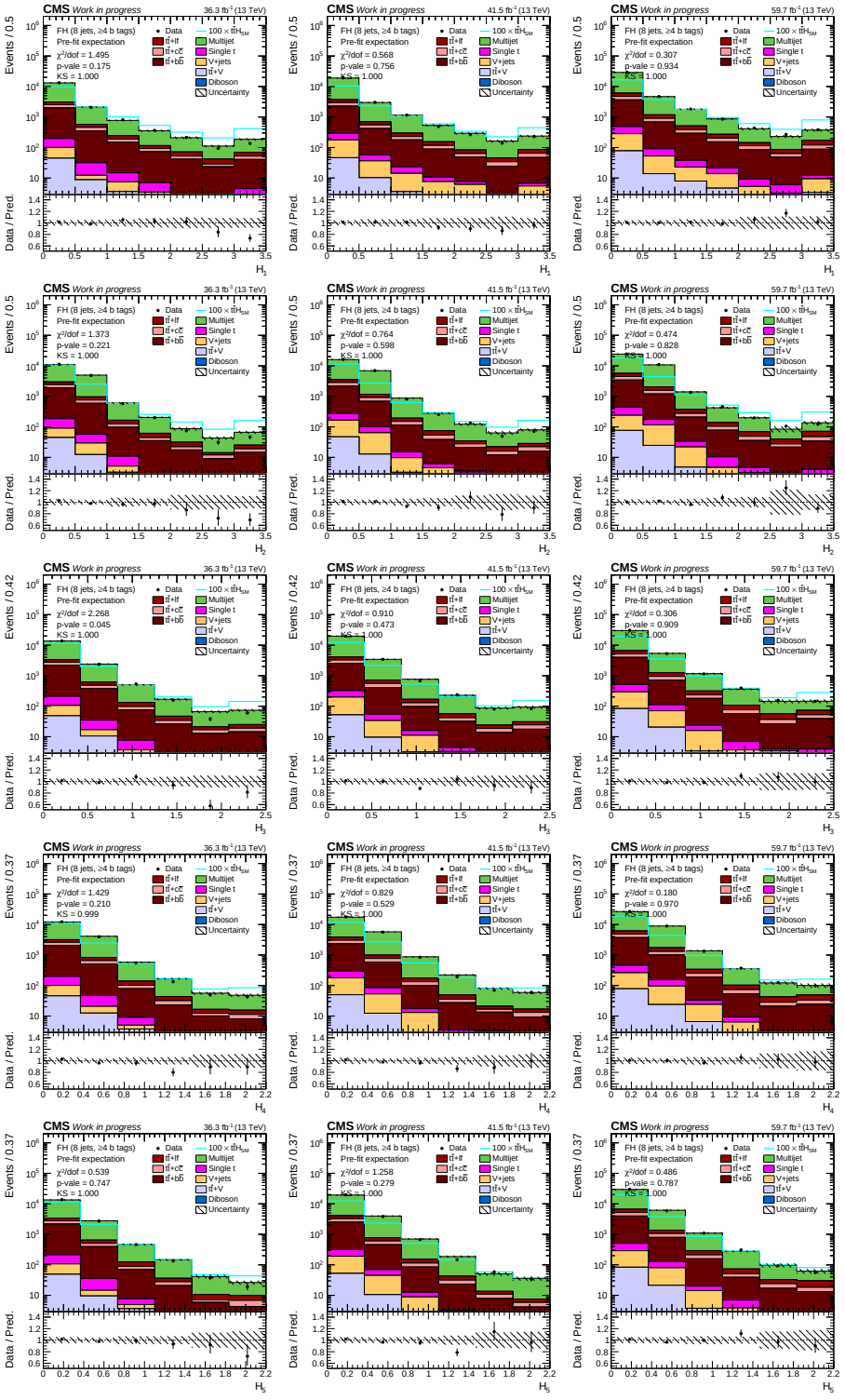



Figure C.6.: ANN input variables $H_1$, $H_2$, $H_3$, $H_4$ and $H_5$ for the ($8\,\text{jets}, \geq 4\,\text{b-tags}$) category for 2016 (left), 2017 (middle) and 2018 (right). The hashed band visualizes the prefit uncertainty including all uncertainties used in the final fit. Additionally a 50% normalization for the $\mathrm{t\bar{t}}+\mathrm{B}$ and $\mathrm{t\bar{t}}+\mathrm{C}$ processes is considered instead of the free floating normalization used in the final fit. The QCD shown in these plots is the data driven estimation as described in section 5.5.

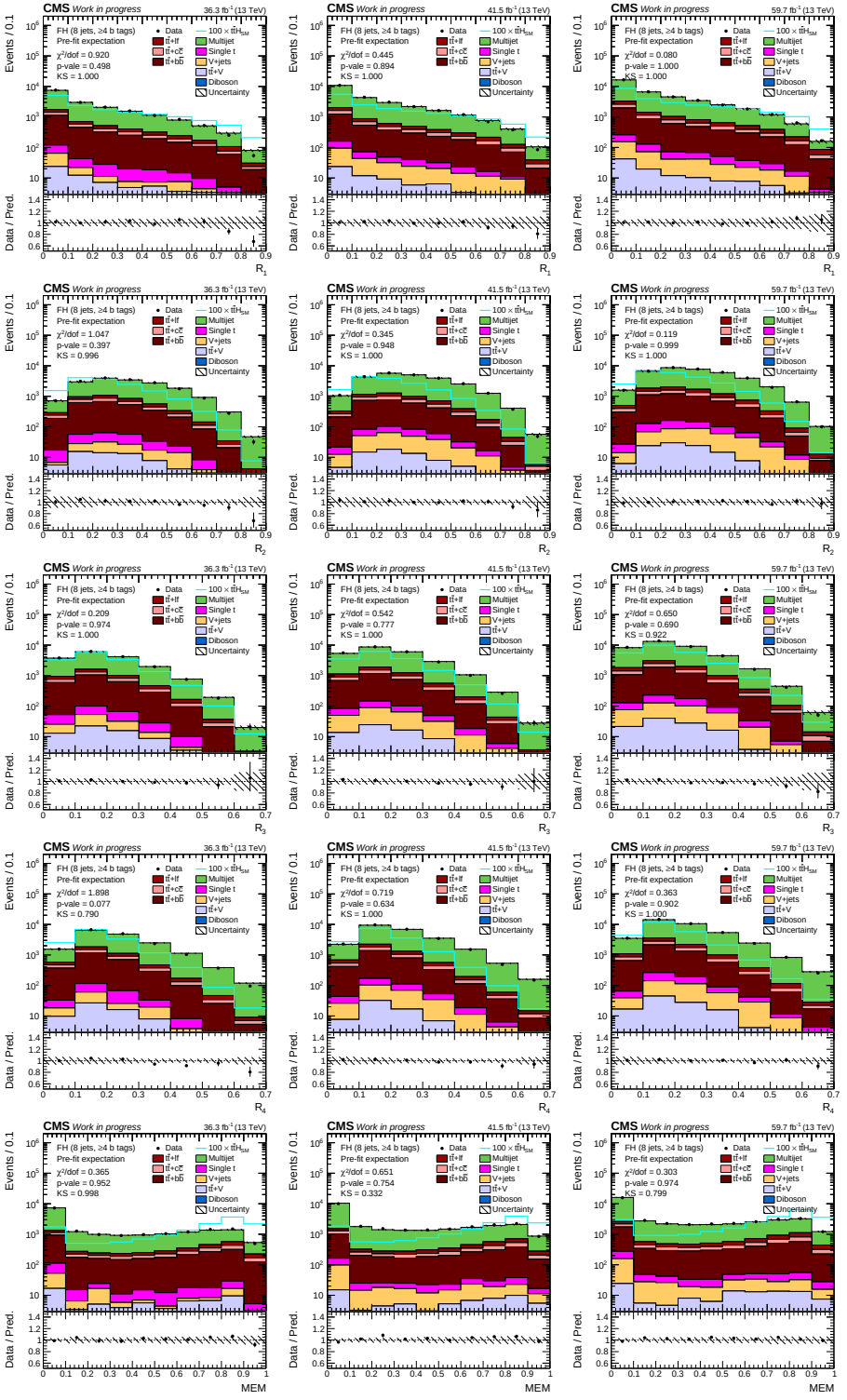



Figure C.7.: ANN input variables $R_1$, $R_2$, $R_3$, $R_4$ and MEM for the ($8\,\text{jets}, \geq 4\,\text{b-tags}$) category for 2016 (left), 2017 (middle) and 2018 (right). The hashed band visualizes the prefit uncertainty including all uncertainties used in the final fit. Additionally a 50% normalization for the $t\bar{t}+B$ and $t\bar{t}+C$ processes is considered instead of the free floating normalization used in the final fit. The QCD shown in these plots is the data driven estimation as described in section 5.5.

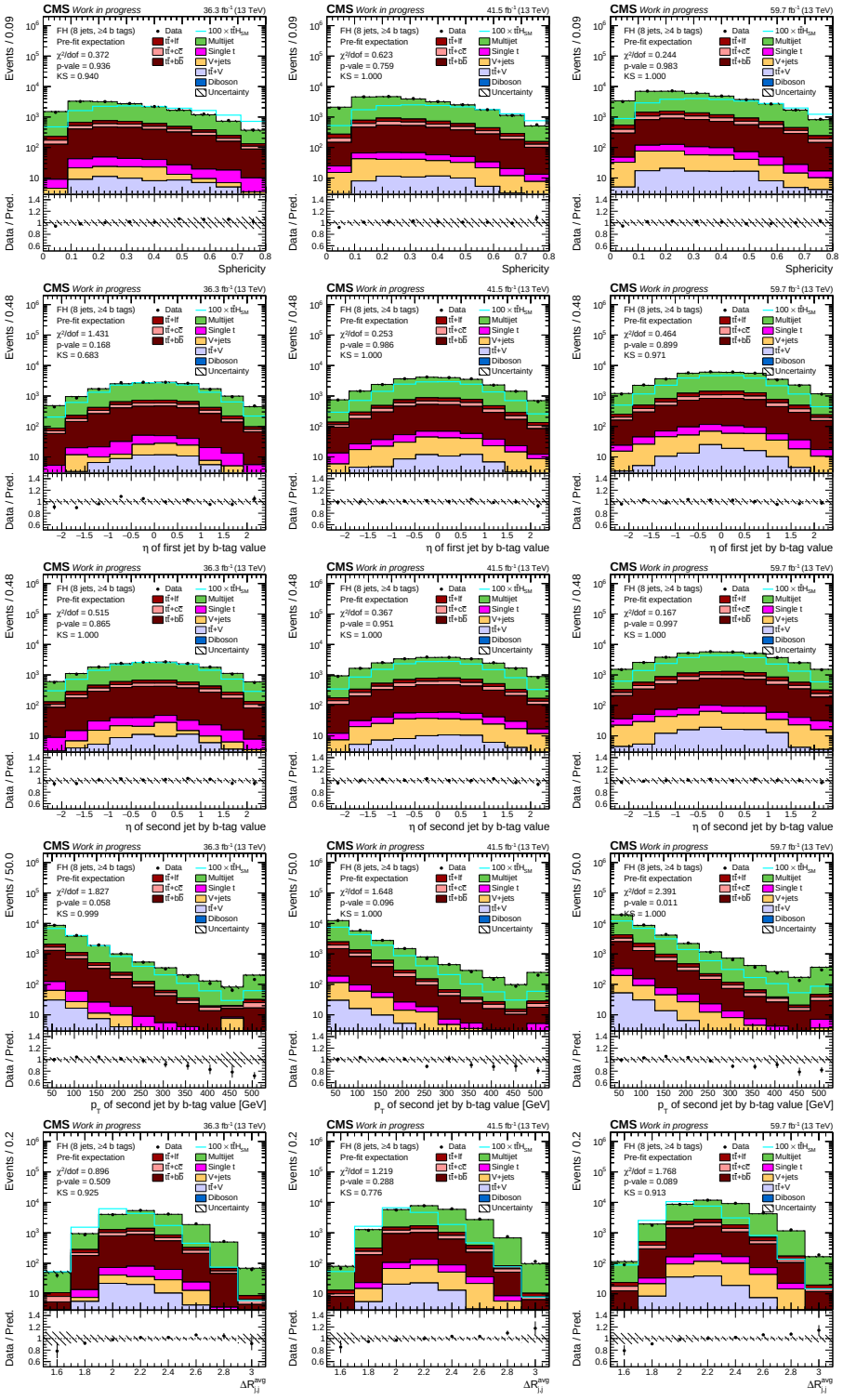



Figure C.8.: DNN input variables Sphericity, $\eta$ of the $1^{st}$ jet by b-tag, $\eta$ and $p_{\mathrm{T}}$ of the $2^{nd}$ jet by b-tag and $\Delta R^{\mathrm{avg}}_{\mathrm{j,j}}$ for the ($8\,\mathrm{jets}, \geq 4\,\mathrm{b\text{-}tags}$) category for 2016 (left), 2017 (middle) and 2018 (right). The hashed band visualizes the prefit uncertainty including all uncertainties used in the final fit. Additionally a 50% normalization for the $\mathrm{t\bar{t}} + \mathrm{B}$ and $\mathrm{t\bar{t}} + \mathrm{C}$ processes is considered instead of the free floating normalization used in the final fit. The QCD shown in these plots is the data driven estimation as described in section 5.5.

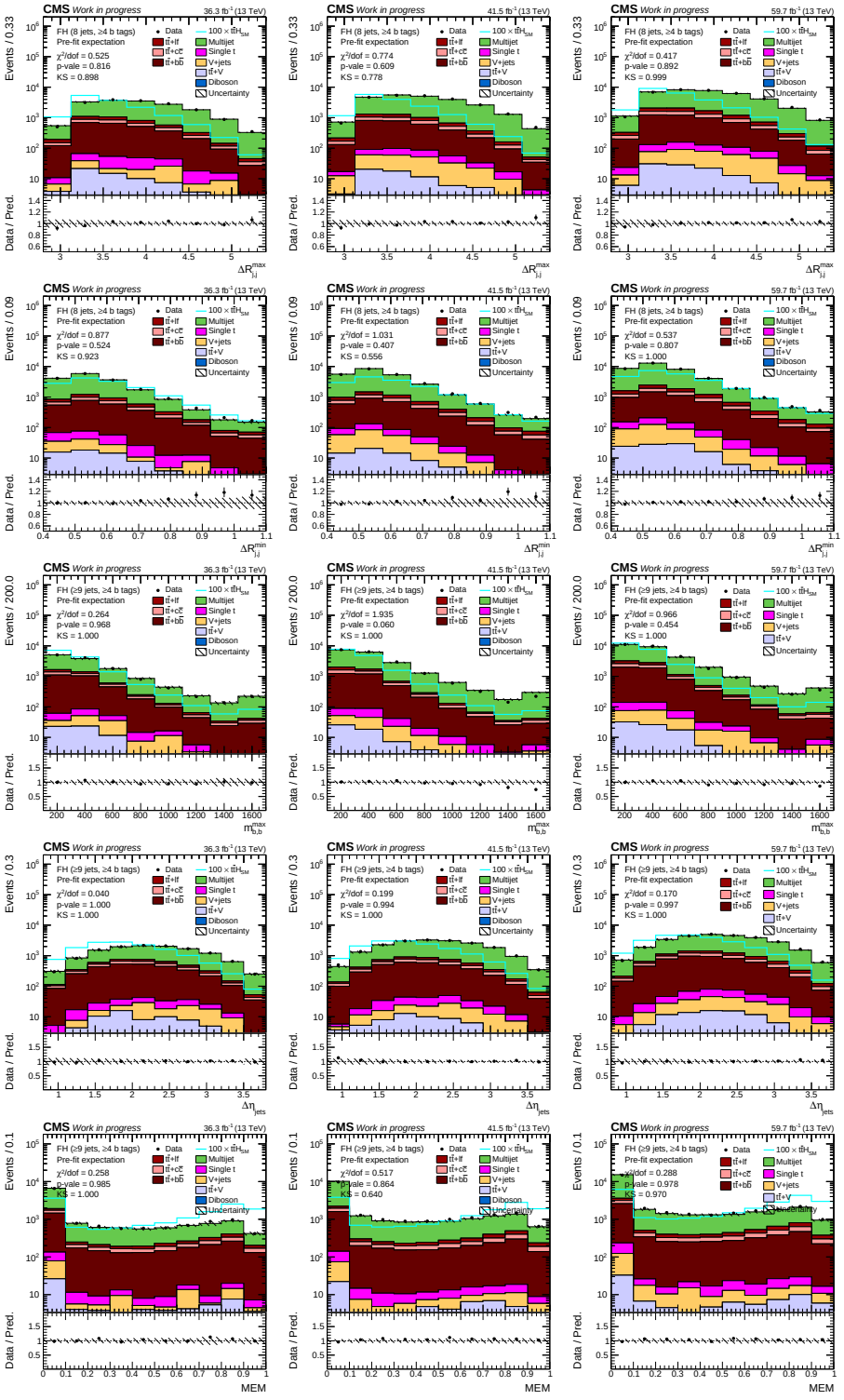



Figure C.9.: ANN input variables $\Delta R^{\text{max}}_{\text{j,j}}$,$\Delta R^{\text{min}}_{\text{j,j}}$ for the ($8\,\text{jets}, \geq 4\,\text{b-tags}$) category as well as $m^{\text{max}}_{\text{b,b}}$ $m$, $\Delta\eta_{\text{jets}}$ and MEM for the ($\geq 9\,\text{jets}, \geq 4\,\text{b-tags}$) category for 2016 (left), 2017 (middle) and 2018 (right). The hashed band visualizes the prefit uncertainty including all uncertainties used in the final fit. Additionally a 50% normalization for the $\text{t}\bar{\text{t}}+\text{B}$ and $\text{t}\bar{\text{t}}+\text{C}$ processes is considered instead of the free floating normalization used in the final fit. The QCD shown in these plots is the data driven estimation as described in section 5.5.

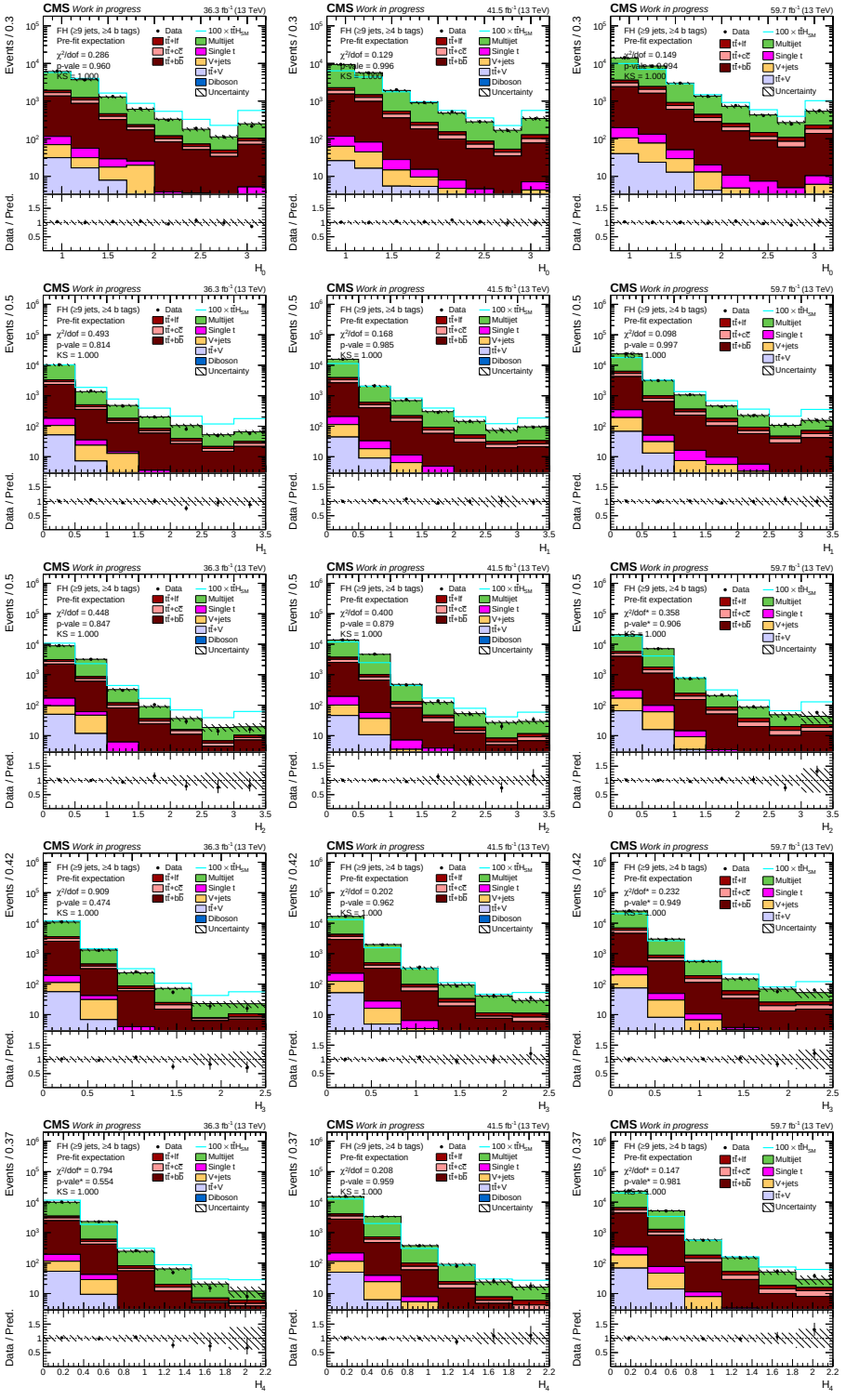



Figure C.10.: ANN input variables $H_0$, $H_1$, $H_2$, $H_3$ and $H_4$ for the ($\geq 9$ jets, $\geq 4$ b-tags) category for 2016 (left), 2017 (middle) and 2018 (right). The hashed band visualizes the prefit uncertainty including all uncertainties used in the final fit. Additionally a 50% normalization for the $t\bar{t}+B$ and $t\bar{t}+C$ processes is considered instead of the free floating normalization used in the final fit. The QCD shown in these plots is the data driven estimation as described in section 5.5.

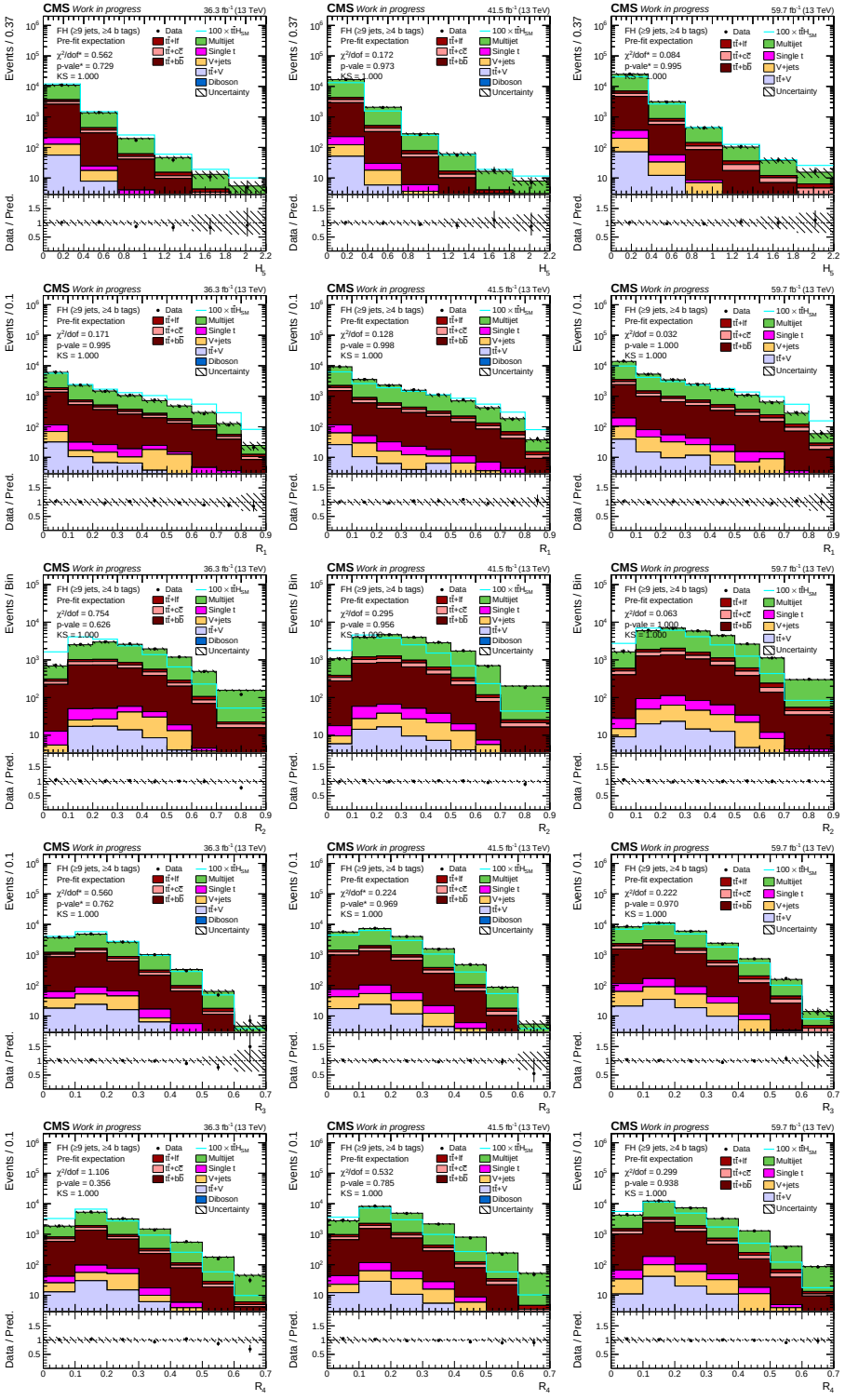



Figure C.11.: DNN input variables $H_5$, $R_1$, $R_2$, $R_3$, $R_4$ for the ($\geq 9$ jets, $\geq 4$ b-tags) category for 2016 (left), 2017 (middle) and 2018 (right). The hashed band visualizes the prefit uncertainty including all uncertainties used in the final fit. Additionally a 50% normalization for the $t\bar{t}+B$ and $t\bar{t}+C$ processes is considered instead of the free floating normalization used in the final fit. The QCD shown in these plots is the data driven estimation as described in section 5.5.

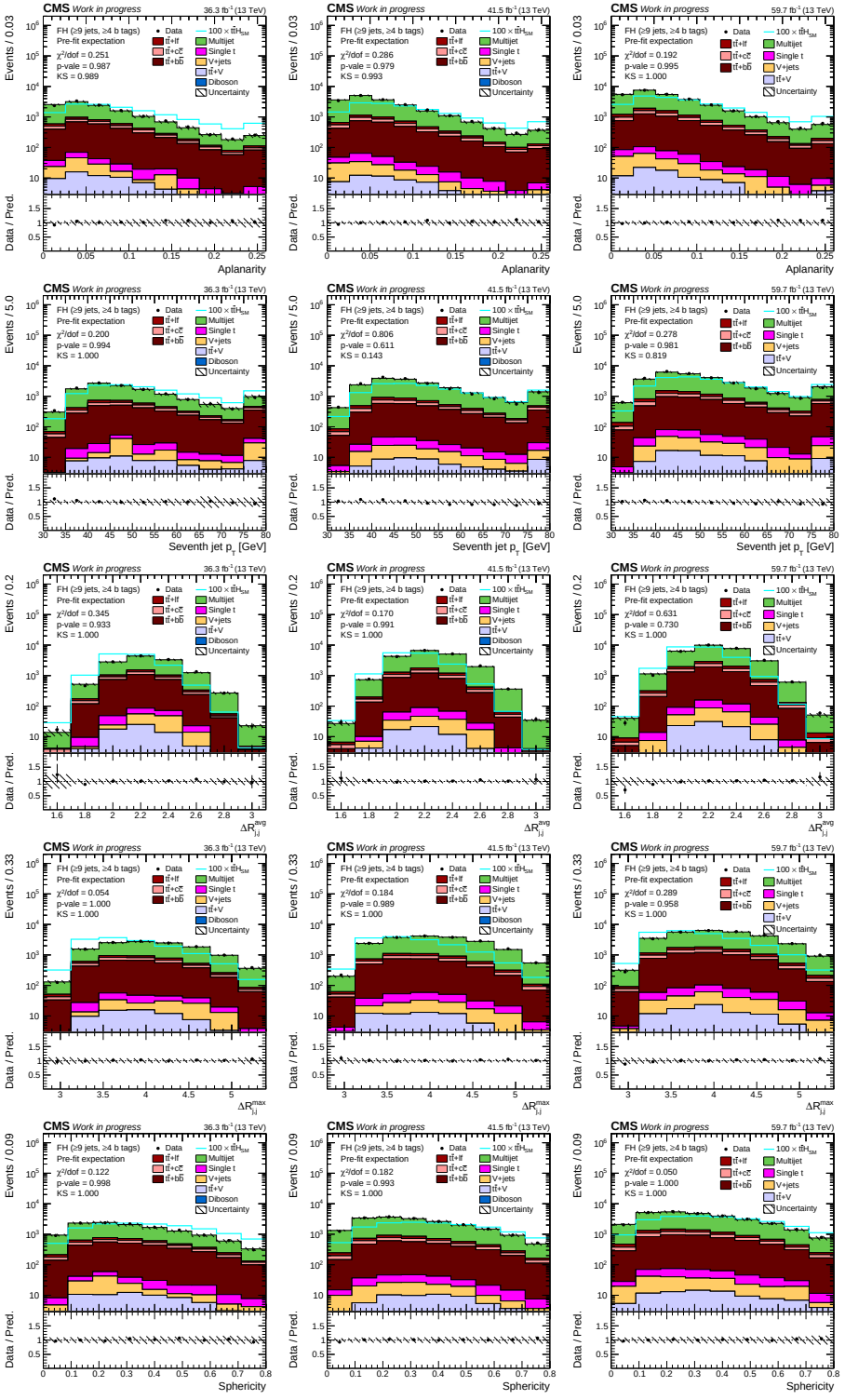



Figure C.12.: ANN input variables Aplanarity, $p_{\mathrm{T}}$ of the seventh jet, $\Delta R^{\mathrm{avg}}_{\mathrm{j,j}}$, $\Delta R^{\mathrm{max}}_{\mathrm{j,j}}$ and Sphericity for the ($\geq 9$ jets, $\geq 4$ b-tags) category for 2016 (left), 2017 (middle) and 2018 (right). The hashed band visualizes the prefit uncertainty including all uncertainties used in the final fit. Additionally a 50% normalization for the $\mathrm{t\bar{t}} + \mathrm{B}$ and $\mathrm{t\bar{t}} + \mathrm{C}$ processes is considered instead of the free floating normalization used in the final fit. The QCD shown in these plots is the data driven estimation as described in section 5.5.

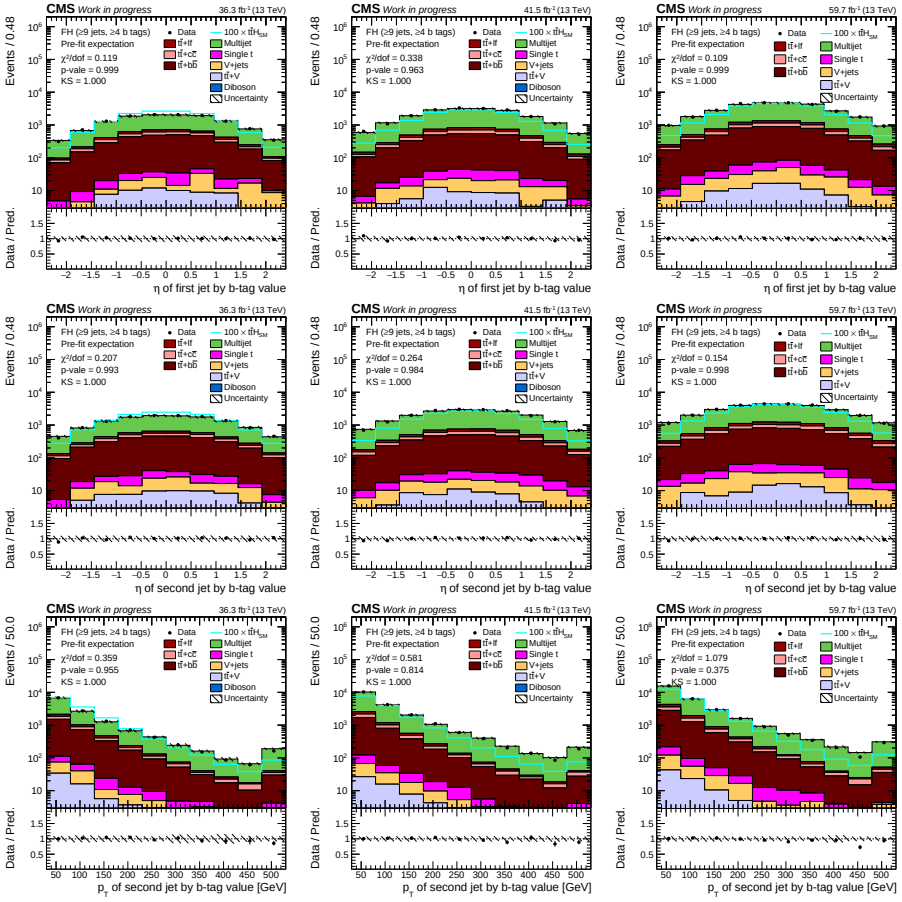



Figure C.13.: ANN input variables $\eta$ of the $1^{st}$ jet by b-tag as well as $\eta$ and $p_{\mathrm{T}}$ of the $2^{nd}$ jet for the ($\geq 9$ jets, $\geq 4$ b-tags) category for 2016 (left), 2017 (middle) and 2018 (right). The hashed band visualizes the prefit uncertainty including all uncertainties used in the final fit. Additionally a 50% normalization for the $\mathrm{t\bar{t}}+\mathrm{B}$ and $\mathrm{t\bar{t}}+\mathrm{C}$ processes is considered instead of the free floating normalization used in the final fit. The QCD shown in these plots is the data driven estimation as described in section 5.5.

# Acknowledgements

First I would like to thank my supervisor Prof. Dr. Florencia Canelli for giving me the opportunity to write this thesis in her group, her guidance, and support.
I would also like to thank the other members of my committee, Prof. Dr. Ben Kilminster, Prof. Dr. Massimiliano Grazzini and Dr. Lorenzo Bianchini for their support and suggestions.

I would also like to thank all current and former members of the CMS group at the University of Zurich. Learning and working in this environment was truly inspiring. Especially Dr. Daniel Salerno and Dr. Silvio Donato provided instrumental guidance and support, without which this thesis and all components would not have turned out as complete as they are now. Furthermore I want to express my gratitude towards Dr. Stefanos Leontsinis, Dr. Kyle Cormier and Dr. Vinicius Mikuni for helping whenever they could and thoroughly reading this thesis.
As particle physics research is largely an effort of a collective, I would also like to thank my colleagues within the CMS collaboration. In particular the ones involved in the $t\bar{t}H(b\bar{b})$ analysis, without whom this analysis would not have been possible, but also everybody else I met during my years in research at the Karlsruhe Institute of Technology and the University of Zurich. I would like to express special gratitude to Dr. Matthias Schröder, Dr. Phillip Keicher, Dr. Michael Wassmer, and Dr. Andrej Saibel, with whom I worked closely over all this time.

I would also like to thanks my parents, Martin and Monika, and the rest of my family for always supporting me. Without you, my studies and this dissertation would have not been possible.

Lastly but by no means least, I would like to deeply express my gratitude to my wife Anna for her unconditional love and support during all these years of me following my path in physics. Without you, I could not have done this!