Improvement of the b-jet energy measurement in a search for Higgs-boson production in association with top quarks at CMS using regression techniques

Korbinian Schweiger

Master Thesis

Korbinian Schweiger

At the Department of Physics
Institut für Experimentelle Kernphysik (IEKP)

Reviewer: Prof. Dr. Ulrich Husemann
Second reviewer: Prof. Dr. Günter Quast

Karlsruhe, 2. November 2016
Accepted by the first referee of the master thesis.

Karlsruhe, 02.11.2016

(Prof. Dr. Ulrich Husemann)
I declare that I have developed and written the enclosed thesis completely by myself, and have not used sources or means without declaration in the text.

Karlsruhe, 02.11.2016

(Korbinian Schweiger)
## Contents

1. **Introduction** ........................................... 1

2. **Theoretical foundations** ................................. 3
   2.1. Hadron collider physics ................................ 3
   2.2. The Higgs boson in association with a top-quark pair .... 7

3. **Tree-based methods for multivariate data analysis** .......... 13
   3.1. Regression trees ........................................ 13
   3.2. Decision trees .......................................... 16

4. **Experiment** ............................................. 17
   4.1. The LHC ................................................ 17
   4.2. CMS .................................................... 19
   4.3. Object reconstruction .................................. 24

5. **Event Simulation** ........................................ 29
   5.1. Event generation ....................................... 29
   5.2. Monte Carlo simulated datasets ......................... 30
   5.3. Corrections to simulated data .......................... 33

6. **Event Selection** ......................................... 35
   6.1. Physics Objects ......................................... 35
   6.2. Event Selection ........................................ 39

7. **Multivariate Analysis for the search for ttH in the H → bb final state** ...... 41
   7.1. Statistical method ...................................... 41
   7.2. Analysis strategy for ttH, H → bb .................... 44
   7.3. Uncertainties considered in this thesis .................. 46

8. **b-jet regression for the ttH search in the H → bb decay channel** .......... 49
   8.1. Jet energy correction for b jets ....................... 49
   8.2. Performance of the b-jet regression .................... 61

9. **Impact of the b-jet regression on the search for ttH in the H → bb decay channel** ...... 65
   9.1. Final classifier and limit ................................ 65
   9.2. Validation ............................................. 70

10. **Summary and outlook** .................................. 77

**Bibliography** ............................................. 79

**Appendix** .................................................. 87
   A. Performance of the b-jet regression ...................... 88
   B. Discriminating variables and the final classifier ....... 90
1. Introduction

In summer 2012, the ATLAS and CMS collaborations at the CERN LHC announced the discovery of a new particle with a mass of approximately 125 GeV/c² \[1,2\] and with properties that are in agreement with the Higgs boson of the standard model (SM) of particle physics \[3,4\]. Measuring the properties of this new boson is crucial to probe the Higgs sector in nature and gain deeper insight into the SM as well as potential new physics beyond the SM.

Because of its universal coupling to all particles, the Higgs boson can be produced and can decay in a multitude of channels. One of these is the Higgs-boson production in association with a top-quark pair (t\(\bar{t}\)H), which is not discovered yet. The t\(\bar{t}\)H cross-section depends on the coupling strength between top quark and Higgs boson, and therefore, significant deviations would indicated physics beyond the standard model. The top-Higgs coupling can also be measured indirectly via its contributions to the loop-induced Higgs-boson production in the gluon-fusion channel and decay in the \(\gamma\gamma\) channel. However, t\(\bar{t}\)H production provides a direct measurement of the top-Higgs coupling with little model dependence.

The precision of the energy measurement of jets originating from bottom quark (b jets) is often worse than for light-flavor jets. The main reason is that the jet-energy measurement is calibrated using light-flavor jet dominated data. To compensate this, dedicated corrections for b jets can be derived. One possibility is to combine information contained in observables of b jets in a multivariate regression technique, in the following denoted as b-jet regression. The b-jet regression technique presented in this thesis was developed in the context of the search for t\(\bar{t}\)H, H \(\rightarrow\) bb as summarized in \[5\]. The b-jet regression leads to an improvement of the relative energy resolution by 5\%, resulting in an improvement of the reconstructed Higgs-boson mass resolution of 5\%, when using Monte Carlo truth information. It is expected that improvements of the b-jet energy measurement can be propagated to the analysis and increase sensitivity.

The theoretical foundations necessary for t\(\bar{t}\)H searches at a hadron collider and tree-based multivariate techniques will be described in chapter 2 and 3 respectively. This is followed by an introduction to the Large Hadron Collider and the CMS experiment in chapter 4. Chapter 5 will cover the simulation of t\(\bar{t}\)H and its background processes with Monte Carlo generators. After that, in chapter 6, the selection of objects and events relevant to this thesis will be described. Chapter 7 will introduce the statistical method used for the calculation of upper limits in Higgs-boson searches at CMS and outline the multivariate analysis used for the search of t\(\bar{t}\)H, H \(\rightarrow\) bb in the semileptonic channel. The configuration
and evaluation of the b-jet regression will be discussed in chapter 8. Subsequently, this regression will be further examined in chapter 9, studying its impact on the analysis and validating its performance on recorded data.
2. Theoretical foundations

The standard model of particle physics (SM) \[6–21\] comprises the current understanding of the fundamental laws of nature. It describes particles and their interactions on a subatomic level. In this model the electromagnetic, strong and weak interactions are unified as a relativistic quantum field theory. Hence, a Lagrangian density can be determined to describe all processes possible in the SM.

The particles are separated into fermions, that have spin \(\frac{1}{2}\) in units of \(\hbar\) and behave according to the Pauli exclusion principle, and bosons with spin 1 that mediate the forces. The fermions are listed in table 2.1 and the force carrying bosons in table 2.2. When the SM was formulated in the 1960s and 1970s it summarized not only the particles known at the time but also postulated further particles, required for the consistency of the SM. This lead to a multitude of experiments searching for them. The charm quark \[22,23\] was found 1974, the bottom quark \[24\] 1977, the top quark \[25,26\] in 1995, the tau neutrino \[27\] in 2000 and the Higgs boson \[1,2\] in 2012.

Even though all data is in agreement with SM predictions it is generally accepted that physics beyond the SM must exist because some observations cannot be explained within the framework of the standard model. One example of an observed deviation is neutrino oscillations \[28,29\], which indicates that neutrinos have non-zero mass. Further unexplained phenomena, like the hierarchy problem (e.g. described in \[30\]), keep motivating particle physicists to search for extensions of the SM or completely new models.

A detailed description of the SM is not possible within this thesis, a detailed introduction can e.g. be found in \[31\]. Instead, only the areas most relevant to this thesis will outlined in sections 2.1 and 2.2.

For the sake of simplicity, the physical constants \(c\) and \(\hbar\) are set to 1 in this thesis.

2.1. Hadron collider physics

A concept that is essential for physics at hadron colliders is the concept of asymptotic freedom of the strong interaction, introduced in quantum chromodynamics \[20,21\]. This principle says that the coupling becomes weaker with increasing energy and the quarks can be viewed as quasi free particles, which means that perturbation theory can be applied. A related concept is quark confinement, stating that the force between two quarks does not get weaker when they are separated. As a consequence, quarks cannot be observed freely but are bound in hadrons.

Aside from this, further concepts are vital for physics at hadron collider and will be introduced in this section.
2. Theoretical foundations

Table 2.1.: 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 [32].

<table>
<thead>
<tr>
<th>Gen. particle</th>
<th>charge/(e)</th>
<th>(T_3)</th>
<th>color</th>
<th>mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>1\textsuperscript{st} Up quark (u)</td>
<td>2/3</td>
<td>1/2</td>
<td>r,g,b</td>
<td>2.2(^{+0.6}_{-0.4}) MeV</td>
</tr>
<tr>
<td>Down quark (d)</td>
<td>-1/3</td>
<td>-1/2</td>
<td>r,g,b</td>
<td>4.7(^{+0.5}_{-0.4}) MeV</td>
</tr>
<tr>
<td>Electron (e)</td>
<td>-1</td>
<td>-1/2</td>
<td>/</td>
<td>0.511 MeV</td>
</tr>
<tr>
<td>Electron neutrino ((\nu_e))</td>
<td>0</td>
<td>1/2</td>
<td>/</td>
<td>&lt; 2 eV</td>
</tr>
<tr>
<td>2\textsuperscript{nd} Charm quark (c)</td>
<td>2/3</td>
<td>1/2</td>
<td>r,g,b</td>
<td>1.27 ± 0.03 GeV</td>
</tr>
<tr>
<td>Strange quark (s)</td>
<td>-1/3</td>
<td>-1/2</td>
<td>r,g,b</td>
<td>96(^{+8}_{-4}) MeV</td>
</tr>
<tr>
<td>Muon ((\mu))</td>
<td>-1</td>
<td>-1/2</td>
<td>/</td>
<td>105.7 MeV</td>
</tr>
<tr>
<td>Muon neutrino ((\nu_\mu))</td>
<td>0</td>
<td>1/2</td>
<td>/</td>
<td>&lt; 2 eV</td>
</tr>
<tr>
<td>3\textsuperscript{rd} Top quark (t)</td>
<td>2/3</td>
<td>1/2</td>
<td>r,g,b</td>
<td>173.21 ± 0.51 ± 0.71 GeV</td>
</tr>
<tr>
<td>Bottom quark (b)</td>
<td>-1/3</td>
<td>-1/2</td>
<td>r,g,b</td>
<td>4.66 ± 0.04 GeV</td>
</tr>
<tr>
<td>Tau ((\tau))</td>
<td>-1</td>
<td>-1/2</td>
<td>/</td>
<td>1776.86 ± 0.12 MeV</td>
</tr>
<tr>
<td>Tau neutrino ((\nu_\tau))</td>
<td>0</td>
<td>1/2</td>
<td>/</td>
<td>&lt; 2 eV</td>
</tr>
</tbody>
</table>

Table 2.2.: Gauge bosons of the standard model. All numerical values are taken from [32].

<table>
<thead>
<tr>
<th>Particle</th>
<th>Interaction</th>
<th>charge</th>
<th>(T_3)</th>
<th>color</th>
<th>mass</th>
</tr>
</thead>
<tbody>
<tr>
<td>photon ((\gamma))</td>
<td>electroweak</td>
<td>0</td>
<td>0</td>
<td>/</td>
<td>massless</td>
</tr>
<tr>
<td>W(\pm) boson</td>
<td>electroweak</td>
<td>(\pm e)</td>
<td>(\pm 1)</td>
<td>/</td>
<td>80.385 ± 0.015 GeV</td>
</tr>
<tr>
<td>Z bosons</td>
<td>electroweak</td>
<td>0</td>
<td>0</td>
<td>/</td>
<td>91.1876 ± 0.0021 GeV</td>
</tr>
<tr>
<td>gluon ((g))</td>
<td>strong</td>
<td>0</td>
<td>0</td>
<td>all</td>
<td>massless</td>
</tr>
</tbody>
</table>
2.1. Hadron collider physics

**2.1. Decays and cross sections**

Particles decay to lighter particles and the lifetime of particles \( \tau \) is distributed following an \( \exp(-t/\tau) \) distribution, with time \( t \). The fraction of decays to a certain final state is called the branching fraction and the inverse of the lifetime \( \tau \) is the decay rate \( \Gamma \).

In scattering processes the cross section \( \sigma \) is proportional to the scattering rate \( \hat{N} \). At a collider the rate of a processes with a scattering cross section of \( \sigma \) can be expressed as

\[
\hat{N} = L \cdot \sigma ,
\]  

with the luminosity \( L \), describing the beam particle flux density, which is usually expressed in units of \( \text{fb}^{-1} \).

By using Fermi's Golden Rule (e.g. explained in [31]) scattering and decay probabilities can be calculated using the SM. These probabilities depend on the phase space available for the final state and the matrix element \( |M|^2 \) of the process, which will be explained in the next section.

**2.1.2. Feynman diagrams**

Feynman diagrams are a way to graphically represent the calculation of the matrix element \( M \) needed for the calculation of cross sections and decay rates. The matrix element can be calculated in perturbation theory using the SM Lagrangian density. The different terms in such a perturbative expansion are usually visualized by different Feynman diagrams. The first term is denoted as leading order term, the second as next-to-leading order term and so forth. In general the order increases with the number of vertices added to the diagram, starting from the lowest number of vertices possible. This is shown for \( e^+e^- \rightarrow e^+e^- \) scattering in figures 2.1a in leading order and 2.1b in higher order. Because each order additionally considered increases the number of possible diagrams the complexity of such calculations increases rapidly.

Feynman diagrams are not only used for visualizing processes but are also a recipe how to compute the matrix element. Each line and vertex represents a specific term of this formula. Every vertex contributes a factor corresponding to the coupling between the involved particles. Examples include \( e^2 \approx 1/137 \) for electromagnetic interactions or \( g_{\text{Hff}}^2 \) for the coupling of the Higgs boson to fermions (one example is the Higgs-boson vertex in figure 2.3). Because these factors are usually small, higher orders can be neglected and only using the leading order is already a viable approximation of \( M \). Because the coupling constant of the strong interaction \( g_s \) becomes large for small energies perturbation theory cannot be used in this energy regime.

**2.1.3. Factorization, parton density functions and parton shower**

By comparing the data from proton-proton collisions at the LHC with theoretical calculations, physicists try to get insights in physics at high energy scales. For this, the hard
interactions, e.g. between highly energetic protons, are of great interest. Such processes are not directly observable, one has to take several additional effects and processes into account.

The proton is not a fundamental particle. Instead it consists of two up quarks and one down quark, which are called valence quarks. They are constantly interacting with each other by absorbing and emitting gluons, which in turn absorb and emit further quarks and gluons. This “cloud” of quarks and gluons (or constituents) is what is usually called proton. The processes inside a proton occur on a low energy scale, which prohibits a perturbative treatment of these effects, while the hard interaction, initiated by a proton-proton collision, happens on a very high energy scale. This leads to the introduction of the factorization theorem (e.g. described in [33]). It states that the interaction between two protons can be factorized in the hard interaction of proton constituents and the soft interaction inside the proton. Because of the short timescale of the hard interaction only the current state of the colliding constituents is of interest, not the dynamics inside the entire protons. The soft and hard parts are divided by the energy scale $\mu_F$ also called the factorization scale.

The cross section of two colliding protons yielding the final state $f$ is calculated using:

$$\sigma(p_1, p_2 \to f) = \sum_{ji} \int_0^1 dx_i \int_0^1 dx_j f_i(x_i, \mu_F^2) f_j(x_j, \mu_F^2) \hat{\sigma}(\hat{p}_i, \hat{p}_j \to f). \quad (2.2)$$

The cross section of the hard interaction $\hat{\sigma}(\hat{p}_i, \hat{p}_j \to f)$ can be calculated with Fermi’s Golden Rule using perturbation theory. Thereby $\hat{p}_i$ is the momentum of the interacting parton $i$ from proton $p_1$ with momentum fraction $x_i$. The $\hat{p}_j$ is the same of the second proton and the sum over $i$ and $j$ takes all possible proton constituents into account. Because the soft parts of the cross section $\sigma$ cannot be calculated perturbatively, the dynamic inside the proton is described by the parton density function (PDF).

The PDF $f_j(x_j, \mu_F^2)$ describes the probability that a certain parton $i$ is found carrying the momentum fraction $x_i$ of the proton when probed at energy scale $\mu_F$. Figure 2.2 shows an example of a PDF set commonly used in high energy physics.

In the final states with strongly interacting partons similar approximation are made. Quarks and gluons decaying via the strong interaction can emit or collinearly split into further quarks and gluons, which leads to a shower of color charged particles. While parts...
2.2. The Higgs boson in association with a top-quark pair

The mechanism of electroweak symmetry breaking \cite{9,12} introduces mass terms for Z and W\textsuperscript{\pm} bosons in the SM Lagrangian density by incorporating the Higgs field with mass \( m_h = \sqrt{2\lambda v^2} \), dependent on the Higgs self-coupling \( \lambda \) as well as the vacuum expectation value \( v \) of the Higgs field. The mass of the fermions on the other hand was added ad-hoc by introducing interactions to the Lagrangian density of the fermions that are called Yukawa interactions. This lead to a fermion mass term dependent on the vacuum expectation value \( v \) of the Higgs field as well as an additional coupling term between fermions and the Higgs field, which can be expressed as \( g_{Hff} = m_f/v \). Because this (Yukawa) coupling is linearly proportional to the lepton mass \( m_l \) the production of the Higgs boson in association with a top-quark pair (usually denoted as \( t\bar{t}H \)), cf. figure 2.3, is of great interest if one wants to experimentally measure and test the coupling between the Higgs boson and fermions. The coupling between top quark and Higgs boson \( g_{Htt} \) can otherwise only be measured in loop processes, which are very model dependent. In the following sections, the properties of the Higgs boson and top quark are discussed. After this, in section 2.2.3, the \( t\bar{t}H \) process in the \( H \to b\bar{b} \) final state and the corresponding major backgrounds will be introduced.

2.2.1. Top quark

The top quark was discovered 1995 by the CDF and D\O\ collaborations \cite{25,26}. Its mass is measured to be approximately 173 GeV, and thus, it is the heaviest particle in the SM. This high mass leads to a very short lifetime, and therefore, unlike all other quarks, the top quarks decays before it undergoes hadronization. Due to the fact that the CKM matrix\footnote{The Cabibbo-Kobayashi-Maskawa (CKM) matrix \cite{35,36} describes the probability for transitions between quark flavors. It is a unitary \( 3 \times 3 \) matrix with a complex phase, whereby transition probability is \( \propto |V_{ij}|^2 \).} element \( V_{tb} \) is nearly one, the top quark almost exclusively decays into bottom quarks in association with W bosons, in the following denoted as \( t \to Wb \). Top quark decay modes are classified according to the W-boson decay, which in 1/3 of all cases decays into leptons and in the remaining 2/3 into quarks.

2.2.2. Higgs boson

Since its first prediction by P. Higgs 1964 \cite{9}, the search for the Higgs boson was subject to many different experiments. In 2012 the ATLAS and CMS collaborations announced the discovery of a massive scalar boson with a mass of approximately 125 GeV \cite{1,2}, compatible with the predictions of the SM Higgs boson \cite{3,4}. Combinations of multiple measurements by both collaborations determine the Higgs-boson mass to \( (125.09 \pm 0.24) \) GeV \cite{32}.
Production of the Higgs boson

The Higgs boson can be produced in different production modes. The direct production from interacting partons is unlikely because light quarks and gluons, which have only a weak coupling to the Higgs boson, dominate for these processes. However, the Higgs boson can be produced through virtual particles with high masses. Feynman diagrams for the the dominant, leading order Higgs-boson production modes are shown in figures 2.4 and 2.3:

Gluon fusion

A Higgs boson is produced via a top-quark loop from the interaction of two gluons. Because the contribution of the top-quark loop is strictly virtual, it cannot be measured. This mode is dominant in the Higgs-boson production at LHC and is shown in figure 2.4a.

Vector boson fusion

In this mode, the Higgs boson is produced in association with two quarks. The process is initiated by the interaction of two quarks radiating vector bosons, that fuse to a Higgs boson. This process depends on the coupling of W, Z and Higgs boson and is shown in figure 2.4b.

Higgs strahlung

The production of a Higgs boson in association with W or Z boson. This mode is shown in figure 2.4c.

ttH

The Higgs-boson production in association with a top quark pair is rather unlikely because of the high masses of the three produced particles. This process will be further discussed in section 2.2.3.

Figure 2.5 shows the cross sections for dominant Higgs-boson production modes as function of the center-of-mass energy. It can be seen that the ttH cross section profits the most from increasing center-of-mass energy of all shown processes.

Higgs boson decay channels

The Higgs boson can decay in many different decay modes because of its coupling to leptons, quarks, gluons and bosons. An overview of possible decay modes is given in figure 2.6. Because of the mass dependent coupling, decay modes with more massive particles are preferred, except phase space factors restrict them. In the following, these decays will be classified in decays to fermions or vector bosons and loop induced decays.

In leading order, the decay rate of the Higgs boson to fermions can be calculated with the formula see e.g. \[\Gamma (H \to f\bar{f}) = \frac{N_c m_h}{8\pi} \frac{m_f^2 \beta_f^2}{v^2},\] (2.3)
2.2. The Higgs boson in association with a top-quark pair

![Cross sections for the Higgs-boson production channels](image)

Figure 2.5.: Cross sections for the Higgs-boson production channels, under the assumption of a Higgs-boson mass of \( m_h = 125 \text{ GeV} \), as function of \( \sqrt{s} \), the center-of-mass energy of the proton-proton collision. [37]

with the velocity of the fermion \( \beta_f \) and a color factor \( N_c \), which is three for quarks and one for leptons. It can be observed that the decay rate increases quadratically with the fermion mass. Because the top-quark mass is larger than the Higgs-boson mass, the decay into the \( b \bar{b} \) final state has the highest rate. The leptonic decay to \( e^+e^- \) and neutrinos is negligible due to their small masses, while the \( \mu^+\mu^- \) decay, with its distinctive signature and small rate, is of interest in collider experiments.

The decays of the Higgs boson to vector-boson pairs is given by [38]

\[
\Gamma(H \rightarrow VV) = \frac{1}{32\pi} \frac{m_h^3}{v^2} \delta_v \sqrt{1 - 4x(1 - 4x + 12x^2)},
\]

with \( x = M_V^2/M_H^2 \) as well as \( \delta_Z = 1 \) and \( \delta_W = 2 \), respectively. Especially the decay of the Higgs boson into a Z-boson pair is of great interest, because of its clear signature, if both Z bosons decay leptonically.

A further possibility are decays of the Higgs boson via loops like the decay to gluons, which is the reverse of the gluon fusion process. Such decays cannot be detected at a hadron collider, because of the overwhelming background, and thus of no current experimental interest. The decay to two photons is of great interest, because only few background processes with the same signature exist. Furthermore, the energy can be measured precisely and this process can be distinguished from background photons because of its invariant mass close to \( m_H \).

2.2.3. Production of \( ttH \) in the \( H \rightarrow b\bar{b} \) final state

The production of \( ttH \) can either be induced by a quark-anti-quark pair or two gluons. The large mass of the three particles (approximately 500 GeV) is the limiting factor for the \( ttH \) production rate at LHC, which is in turn strongly dependent on the center-of-mass energy, as it is shown in figure 2.5.
2. Theoretical foundations

Figure 2.6.: Higgs-boson branching ratios as function of the Higgs-boson mass.

Figure 2.7.: Topology of \( t\bar{t}H \), \( H \rightarrow bb \) in the semileptonic channel (a) and the same final state topology from a \( t\bar{t} + bb \) process (b).

2.2.3.1. Topology

As described in section 2.2.1, the W boson from the \( t \rightarrow Wb \) decay can either decay leptonically or hadronically. As a consequence, \( ttH \) can be categorized in three decay channels. In approximately \( 1/9 \) of all events both W bosons decay into a lepton and a neutrino and thus it is called the dilepton channel. Furthermore all hadronic (both W bosons decay to quarks) and semileptonic (one W boson decays hadronically and the other leptonically) decays of the top-quark pair occur. They have a rate of \( 4/9 \) each.

The search for \( ttH \) presented in this thesis focuses on the semileptonic decay channel of the top-quark pair and the Higgs-boson decay to a bottom-quark pair. The lepton from the leptonic W-boson decay is used to suppress QCD-multijet background, while still having a comparable large branching fraction. The \( H \rightarrow bb \) channel is chosen, because it has the largest branching fraction of all Higgs-boson decays. The expected final state for this process is shown in figure 2.7a. Aside from the lepton described above, missing energy from neutrinos and six quarks of which four bottom quarks are expected in this final state. This picture is usually modified by additional QCD radiation, leading to additional quarks in the final state.
2.2. The Higgs boson in association with a top-quark pair

2.2.3.2. Backgrounds

The expected signature of $\bar{t}tH, H \rightarrow b\bar{b}$ can also be produced by other processes. Therefore, applying a selection\(^2\) corresponding to the expected final state (and slight variations), to data will inevitably lead to a contamination with non-$ttH$ events.

$tt + bb$

The process yielding the most similar final state to the expected $ttH$ signal, is the production of a top-quark pair in association with a bottom-quark pair, as it is shown in figure [2.7b]. In next-to-leading order (NLO) theory calculations it is expected that the cross section of $tt + bb$ \(^3\) is approximately eight times larger than the cross section for $ttH$ \(^4\). Because of the similar final state and the much larger rate of $tt + bb$, this background cannot be suppressed by simply making stricter cuts on single observables, but the specific properties of the processes themselves have to be studied.

The properties of the bottom-quark pair are strongly connected to the particle of its origin. This manifests itself in the invariant mass and the angular distribution of the $bb$ pair. The invariant mass should be lower for $bb$ pairs from a virtual gluon than from a Higgs boson, because of their very different masses. The Higgs boson is a color singlet and, therefore, the $bb$ pair is color connected and should be produced rather collimated \(^5\). The gluon carries spin and color charge and, therefore, the bottom quarks should be further apart. Another effect is that the Higgs bosons in $ttH$ can only couple to the top quarks, while gluons can couple to almost every particle involved in $tt$, which leads to very different kinematic distributions (and a large number of possible Feynman diagrams for $tt + bb$).

$tt + jets$

A further background induced by the $tt$ process is the production of $tt$ with additional light-flavor jets, usually denoted as $tt + jets$. This background is largely suppressed by requiring more than two jets originating from $b$ quarks. Because the algorithm used for this identification, explained in section 4.3.6, can misidentify jets, $tt + jets$ events can pass the selection despite this criterion. This process has a much larger cross section than the signal and with this, $tt + jets$ events with these misidentified jets form a non-negligible background.

Other Backgrounds

In addition to this two backgrounds, further minor background pass the $ttH$ selection:

**Single Top quark:** Single $t$ or $\bar{t}$ production exhibits a final state with two $b$ quarks, $E_T$, a lepton and additional jets, which can also look similar to a $ttH$ event. Of the three possible channels $s$, $t$- and $tW$, the $t$-channel, cf. figure 2.8a, contributes the most of all single-top events to the the total $ttH$ background.

**$tt + Vector bosons$:** Similar to $ttH$, top-quark pairs can also be produced in association with vector bosons. The $ttZ$ production with $Z \rightarrow b\bar{b}$ is most signal-like, but the contribution is rather low because of the small branching fraction.

**Vector bosons + Jets:** The production of vector bosons with additional jets can also lead to very similar final states, if leptons are produced. As an

---

\(^2\) The exact selection will be described in detail in chapter \(^6\).

\(^3\) These jets often originate from gluons emitted in the initial or final state. This is often referred to as initial-state radiation (ISR) and final-state radiation (FSR) respectively.
example, the production of a W boson with four additional jets is shown in figure 2.8b. Because the rates of these processes decrease with the jet multiplicity, the contribution to the signal region is low, but can be larger in $\bar{t}t$ control regions.

**Vector boson pairs:** Vector-boson pair production is a further small $ttH$ background, for example if two Z bosons decay to $bb$ with additional light jets.
3. Tree-based methods for multivariate data analysis

Statistical data analysis, using multivariate methods and machine learning, is a widespread subject utilized in science and economy. In multivariate analysis (MVA), a potentially large amount of inputs is analyzed simultaneously, while univariate analysis focuses on one input at a time. Commonly used MVA methods are among others (boosted) decision/regression trees, which will be the focus of the chapter, artificial neutral networks (ANN), the closely related deep neural networks (DNN) or support vector machines (SVM). A detailed overview of various methods used in statistical data analysis can be found in [43]. In the context of particle physics, multivariate methods are often used to select and distinguish events by using large numbers of observables characterizing the event. One example of such an analysis is outlined in chapter 7. Other physical problems can also profit from the flexible solutions provided by MVAs. One example, are the parton distribution functions calculated by the NNPDF collaboration utilizing neutral network regression techniques [44].

This chapter will describe tree-based methods for multivariate data analysis. In section 3.1, regression trees and the gradient boosting procedure will be introduced. Following this, the closely related decision trees are described in section 3.2.

3.1. Regression trees

In the following, input variables will be denoted as \(X\), while \(Y\) will be used for the regression target as defined below. The prediction will be written as a function \(f(X)\), dependent on a set of input variables \(X = (X_1 \ldots X_i)\).

The most general description of a regression is that a model is used to predict a quantitative output from a set of inputs and, therefore, can be viewed as a function approximation. The method presented in this section, is part of the category of supervised learning algorithms. Such algorithms use datasets that include the correct output to learn features of the dataset. The importance of the correct output, often referred to as regression target, should not be underestimated because it describes the model fitted in the training process. Subsequent to the training process, the knowledge can be used on independent datasets. The square in figure 3.1a represents the absolute response space of the regression, defined by the inputs \(X_1\) and \(X_2\), which is gradually split in smaller areas by cutting \((c_1 - c_4)\) on the inputs. In the end, discrete responses are assigned to the partitions of the response.
3. Tree-based methods for multivariate data analysis

Figure 3.1.: Visualization of the response of a regression. Figure (a) shows the 2D illustration of the subsequent division of the total response space, by cutting on the input variables. The same can be pictured in the form of a tree structure (b). The responses $A_i$ can either be interpreted as quantitative or continuous values leading to a regression scenario or as qualitative or discrete values leading to a classification scenario. For example, a regression model corresponding to (a) and (b) can be expressed as $f(X) = \sum_{m=1}^{5} r_m \mathbb{I}(X \in A_m)$, where $\mathbb{I}$ is one if $X$ is in region $A_m$ while $r_m$ is the response value.

The method used in this thesis is the tree-based regression method (as it is implemented in TMVA\textsuperscript{[45]}) and, in this context, the response regions are called end nodes. Such methods can be visualized as shown in figure 3.1b. The output is determined by starting from the root node, on top of the tree, and subsequently following the structure. At each node the given input variable is compared to the proposed cut pointing to the next node. This is repeated until an end node is reached. The building of a prediction model in an automatic, iterative process, starting from a base function, specific to the considered method, is outlined in the following:

In general, the prediction of a regression with $M$ end nodes $A_m$ can be written as

$$f(X) = \sum_{m=1}^{M} r_m \mathbb{I}(X \in A_m),$$

where $\mathbb{I}(X \in A_m)$ is one, if the input is in end node $A_m$ and is zero otherwise. The response of the regression is modeled with the constant term $r_m$, which is the average target value in the region $A_m$. The best estimator for $r_m$ is found by minimizing a separation criterion $S$ for each node leading to the according end node. For TMVA regression trees, the average squared error

$$\frac{1}{N} \sum_{i=1}^{N} (Y_i - \hat{Y})^2,$$

with the target values $Y_i$ of each event in the current node and the mean target value $\hat{Y}$ of all $N$ events in the node, is used as separation criterion. At each node splitting, the algorithm determines the best combination of input and binary cut by calculating

$$- \frac{S_P - (S_L + S_R)}{S_P},$$

the separation increase between parent node and the sum of the daughter nodes, with the separation criteria $S_P$, $S_L$ and $S_R$ for parent node as well as left and right daughter.
node. For each input variable all possible cuts, splitting the training sample in \( N_L \) and \( N_R \) events, are scanned with a granularity of \( n_{\text{Cuts}} \) set at configuration time. This is repeated until a stopping criterion is reached, which can either be a maximum tree depth, usually denoted as \( \text{maxDepth} \), or a threshold for the minimal number of events in a node.

When using multivariate analysis, one has to define a dataset used for the training of the algorithm, called training sample. Because this dataset can exhibit unique features or statistical fluctuations, the regression performance can be different on an independent dataset, usually referred to as test sample. This effect is called “overtraining” and is a big concern when working with multivariate algorithms. Different methods, like boosting and bagging, have been developed to decrease such effects.

**Gradient boosting**

To decrease overtraining and with this increase stability, a so called boosting procedure can be used. Thereby a “forest” of regression trees is trained and the majority vote of all trees is used for the prediction.

The idea is to consecutively add trees to the predictor, to compensate shortcomings of the ensemble of previously trained trees \( F(X) \). To find the next tree \( \text{TMVA} \) uses the Huber loss function

\[
L(F(X), Y) = \begin{cases} \frac{1}{2} |Y - F(X)|, & |Y - F(X)| \leq \delta \\ \delta \left( |Y - F(X)| - \delta/2 \right), & |Y - F(X)| > \delta \end{cases}
\]

with \( \delta = 0.7 \). After the first regression tree is grown, as described above, the boosting algorithm is initialized with \( F_1(X) = f(X) \), cf. equation (3.1). Before the \( t \)-th tree is trained, the negative gradients of the loss function \( L \), hence the name gradient boosting, also called pseudo residuals,

\[
r_{it} = -\left[ \frac{\partial L(F(X), Y)}{\partial F(X)} \right]_{F(X) = F_{t-1}(X)}
\]

are calculated for each event with the predictor \( F_{t-1}(X) \). For the training of the tree \( f_t(X) \) the pseudo residuals \( r_{it} \) are used as regression target. After the training is completed, the ensemble is updated according to

\[
F_t(X) = F_{t-1}(X) + \nu f_t(X)
\]

The parameter \( \nu \) is called shrinkage and is used to control the learning rate of the ensemble. Small values of \( \nu \) lead to a regression more robust to overtraining.

Regression trees with applied boosted procedure, are in the following referred to as boosted regression trees or BRTs.

**Bagging**

In addition to boosting, a bagging procedure can used, which is a resampling technique. Instead of using the whole dataset for the training of the regression, only a random subset is used. This leads to a smearing of statistical fluctuations in the data and therefore decreases overtraining. The fraction of events used for the resampling is denoted as the \textsc{BaggedSampleFraction}.

---

1 The can be easily understood by considering the note splitting at the root node in the example in figure 3.1b. The cut \( c_1 \) minimizes equation (3.3), where the parent node (in this case also the root node) includes all \( N_{\text{tot}} \) events and the daughter nodes contain \( N_L \) and \( N_R \) events with \( X_1 \leq c_1 \) and \( X_1 > c_1 \) respectively.
Performance characterization

In the context of non-linear regression models, like the tree-based method presented above, ways to characterize how good the regression fits to the model, proposed by the target, are the quadratic

$$\text{err}_{\text{quad}} = \frac{1}{N} \sum_{i}^{N} (f(X) - Y)^2$$, \hspace{1cm} (3.7)

and absolute error

$$\text{err}_{\text{abs}} = \frac{1}{N} \sum_{i}^{N} |f(X) - Y|$$ . \hspace{1cm} (3.8)

Aside from evaluating how well the regression fits the data, these errors can also be used to test the regression for overtraining. For this, a statistically independent test sample is defined and the regression is applied to these datasets. Substantial differences in the errors on this and the training sample, indicate overtraining. Furthermore, the bias of the regression prediction compared to the target can be used as an indicator for regression performance.

3.2. Decision trees

The concept of decision trees and boosted decision trees (BDT) is very similar to their regression counterparts. The main difference is that BDTs are used to predict a qualitative output, sampled from a finite set, like signal and background. In general this is achieved by assigning the end nodes to one of the possible outputs. The tree in figure [3.1b] can be interpreted as decision tree for the classification in signal and background, by only allowing the end nodes $A_1 - A_5$ to respond with ”signal”(+1) or ”background”(-1). The growing of the tree proceeds similar to the regression trees described above. The main difference is the criterion used for node splitting. Common criteria for classification are the statistical significance, defined as $S/\sqrt{S+B}$, or the Gini coefficient, defined as $p \cdot (1-p)$ with the purity $p$. The latter is the default choice in TMVA. For the training, exclusive signal and background sample are used for the truth information needed.

By using the gradient boosting procedure on decision trees, the BDT output is continuously distributed between $\pm 1$. A possible way to evaluate such output in the context of decision trees, is to define a cut for which higher values are interpreted as signal and lower as background.

The performance of multivariate classification methods is characterized by the receiver operator characteristic (ROC). It is determined by cutting on the classification output and calculating the efficiencies for selecting signal and rejecting background events. Plotting the background rejection against the signal efficiency, for multiple cut values, leads to the ROC curve. The figure of merit used when comparing classifiers is the area under the ROC curve. Higher values correspond to better discrimination power.

Furthermore, the Kolmogorov-Smirnov (KS) test applied to the BDT output shapes of the training and test sample, separated in signal and background, can be used as indicator for the overtraining of the classifier.

---

2 This is dependent on the implementation.
4. Experiment

In this chapter, the experimental foundations needed for the analysis presented in this thesis will be described. In section 4.1, the Large Hadron Collider will be presented and section 4.2 will be used to introduce the Compact Muon Solenoid. Following this, the basics of data reconstruction will be discussed in section 4.3.

4.1. The LHC

The Large Hadron Collider (LHC), located at the European Organization for Nuclear Physics (CERN), is currently the largest and most powerful particle accelerator. It is located near Geneva in an underground circular tunnel with a diameter of 26.7 km underneath the border between France and Switzerland. This tunnel was originally built for the Large Electron-Positron Collider (LEP). If it is not indicated otherwise, the following information is taken from a detailed description of the LHC that can be found in [48].

The LHC is designed to accelerate protons to energies of $7\text{ TeV}$ and let them collide at four interactions points, where the detectors of the ALICE [49], ATLAS [50], CMS [51] and LHCb [52] experiments are located. ATLAS and CMS are multi-purpose experiments designed for the study of the properties of the standard model and search for new particles. In contrast to this, the LHCb detector is dedicated to precision measurements of $CP$ violation and rare decays with $B$-Hadrons. The LHC can also be used to accelerate lead ions up to energies of $2.8\text{ TeV}$ per nucleon. This is the main focus of the ALICE experiment. Heavy-ion physics at the LHC will not be considered in this thesis but further information on this topic can be found in [53].

The overall aim of the groups working at the LHC is to reveal physics beyond the standard model. This can either be accomplished by looking for new particles and effects or measuring deviations from the standard model expectation. For both, very small cross sections are expected and therefore the number of expected events $N$ is solely limited by the integrated luminosity $L = \int L\, dt$, cf. equation (2.1). For this reason, the LHC was designed as a proton-proton collider because higher luminosities are possible compared to particle-anti-particle colliders. The luminosity is given by

$$L = \frac{f n_b N_b^2}{4\pi\sigma_x\sigma_y}, \quad (4.1)$$

where $f$ is the revolution frequency and $N_b$ the number of bunches with $n_b$ protons each. The beam size at the interaction point is characterized by a Gaussian profile of width $\sigma_x$ and $\sigma_y$. The design luminosity of the LHC is $L = 10^{34}\text{ cm}^2\text{s}^{-1}$. 

17
Protons require a certain energy before they can be accelerated by the LHC to their final energy. In order to reach this energy, the protons pass through a chain of pre-accelerators that are shown in figure 4.1.

After ionizing hydrogen gas and splitting the resulting continuous beam into bunches, the LINAC2 linear accelerator is used to accelerate them to an energy of \(50 \text{ MeV}\). After passing through the Proton Synchrotron Booster (PSB), the Proton Synchrotron (PS) and the Super Proton Synchrotron (SPS) the bunches meet the requirements to be accelerated by the LHC. At this stage the bunches have energies of \(450 \text{ GeV}\), a spacing of \(25 \text{ ns}\) and contain about \(1.1 \times 10^{11}\) protons each. The LHC injector chain is described in detail in [54]. Following this, 2808 bunches are injected into the LHC and eight \(400 \text{ MHz}\) superconducting radio frequency cavities are used to accelerate the protons to their final energy.

To keep the protons on their circular path around the LHC the magnetic field of dipol magnets is used. The dependency of the field \(B\) on the collider radius \(r\), the momentum of the particle \(p\) and its charge \(q\) can be described with

\[
B = \frac{p}{q \cdot r}. \tag{4.2}
\]

Because the collider radius is given by the existing infrastructure, the achievable energies are limited by the magnetic field. For this reason, superconducting dipol magnets are used for bending the proton beam. The dipol magnets were designed to provide a field of \(8 \text{ T}\), which is required for the operation with \(7 \text{ TeV}\) energy per beam. In the run period in 2015/2016 a beam energy of approximately \(6.5 \text{ TeV}\) were achieved. Additional superconducting quadrupole magnets are used to focus the beam. Both are cooled to a temperature below \(2 \text{ K}\) using superfluid helium.
4.2. CMS

The Compact Muon Solenoid (CMS) [51] is one of the two multi-purpose particle detectors at the LHC and is located in an underground cavern near Cessy in France. Figure 4.2 shows the design of the CMS detector and its subsystems. The dimensions are 21.6 m in length and 14.6 m in diameter. One design principle was to be able to identify the signatures of a multitude of different physics processes occurring when protons are colliding. Therefore the detector must be able to precisely identify and measure the momenta of muons, electrons and photons. Furthermore identifying B hadrons and $\tau$ decay product as well as measuring the energy of charged and strongly interacting particles is a requirement for the detector. Because particles like neutrinos can escape the detector undetected it should hermetically covered.

4.2.1. The coordinate system

This section aims to introduce the kinematic variables used in collider physics. The $z$-axis points toward the Jura mountains along the beam line, the $x$-axis towards the center of the LHC and the $y$-axis towards the surface. Because of the cylindrical detector design the angles $\phi$ ($x$-$y$-plane) and $\theta$ (measured from the $z$-axis) as well as the radial distance $r$ are used more commonly. Because the momenta of protons interacting at a hadron collider can differ widely and the colliding partons only carry a fraction of that momentum, the center of mass frame can move along the beam line. Because the $z$-components of particle momenta cannot be measured precisely enough, total momentum is not a good variable to describe particles. For this reason, only the momentum components transverse to the beam-axis are used.\footnote{The transverse momentum is only Lorentz-invariant under the assumption that the colliding proton constituent have only momentum in longitudinal direction.}

The transverse momentum is defined as

$$p_T = \sqrt{p_x^2 + p_y^2} . \quad (4.3)$$
Additionally the transverse mass is used:

\[ M_T = \sqrt{E^2 - (p_z)^2} \]  \hspace{1cm} (4.4)

The rapidity \( y \) is defined as

\[ y = \frac{1}{2} \ln \left( \frac{E + p_z}{E - p_z} \right) \]  \hspace{1cm} (4.5)

but will in general be transformed to the pseudorapidity

\[ \eta = -\ln \left( \tan \left( \frac{\theta}{2} \right) \right) \]  \hspace{1cm} (4.6)

which is equal to \( y \) if the kinetic energy of the particle is high compared to its mass.

Another variable often used in particle physics is the angular separation of two particles,

\[ \Delta R = \sqrt{(\phi_1 - \phi_2)^2 + (\eta_1 - \eta_2)^2} \]  \hspace{1cm} (4.7)

4.2.2. Solenoid

The magnet is the central element to the design of CMS. For the measurement of the transverse momentum a strong magnetic field is needed to force charged particles on a helix shaped trajectory. As seen in the relation in equation (4.2), the measurement of high momenta requires a strong magnetic field. In the CMS detector this is realized with a 12.5 m long superconducting solenoid built from NiTb cable. It has a inner bore with a diameter of 6.3 m and enough space for the tracker and calorimeters.

A current of approximately 20 kA induces a magnetic field of 3.8 T when the magnet is in its superconducting state. For this, the solenoid is cooled with liquid helium to a temperature below 4.6 K. The iron yoke, which returns the magnetic flux and functions as the main structural element of the detector, is located outside of the solenoid and houses the muon system.

4.2.3. Tracking system

The innermost part of the CMS detector is the silicon tracking system [56]. It is placed around the beam pipe with a length of 5.8 m and a diameter of 2.5 m, therefore covering a pseudorapidity \( |\eta| < 2.5 \).

The purpose of the tracker is to identify the paths of charged particles, called tracks. These can be used to identify vertices (primary as well as secondary) and measure the momenta of particles. The transverse momentum (in GeV) of a particle with charge \( q \) in units of \( e \) can be calculated using the magnetic flux density \( B \) (measured in T) and the track’s radius of curvature in the transverse plane \( \rho \) in units of m [32]:

\[ p_T = 0.3 q B \rho \]  \hspace{1cm} (4.8)

The relative \( p_T \) resolution can be approximately calculated using

\[ \frac{\sigma(p_T)}{p_T} = \frac{\sigma_x p_T}{q B L^2} \sqrt{\frac{720}{N + 4}} \oplus \sigma_{ms} \]  \hspace{1cm} (4.9)

where the first term describes effects dependent on the hit-position measurement error \( \sigma_x \) along the trajectory of length \( L \) including \( N \) hits, and the second term describes the dependence of the resolution on multiple scattering [32]. The reconstruction and vertex finding will be described in section 4.3.1.

The tracker was designed to have a low response time and high position resolution, while keeping the tracker material budget as low as possible. This is achieved by exploiting the
semiconducting properties of silicon. It is doped with n-p-junctions and when a reverse bias voltage is applied a depletion zone develops. Charged particles transversing this zone can produce electron-hole-pairs which are separated by the applied electric field and can be measured. The tracker of the CMS detector consists of silicon pixel and strip detectors that are placed in layers around the beam pipe. A total number of 66 million pixels, each with a size of \((100 \times 150)\) µm\(^2\) and its own read-out electronics, is used. Three layers are placed in the barrel, covering the region between 4.4 cm and 10.2 cm in radial distance \(r\). Additionally two layers are placed in each end cap, with a longitudinal distance of 34.5 cm and 46.5 cm. Because the 3.8 T magnetic field induces a Lorentz drift, the charge carriers in silicon can spread to adjacent pixels. These pixels can be combined, which leads to a spatial resolution of 20 µm. The strip detector is placed around the pixel detector. It consists out of 15148 strips covering an area of 200 m\(^2\). Strips are combined to modules, which can vary in size and shape, allowing a position measurement in two dimensions. Additionally tilting and rotating modules enables the strip detector to also measure the hit position in three dimensions. Overall, ten layers are installed parallel and twelve perpendicular to the beam direction.

4.2.4. Calorimeter

The relative energy resolution of calorimeters can be parametrized as

\[
\frac{\sigma(E)}{E} = \frac{N}{E} \oplus \frac{S}{\sqrt{E}} \oplus C ,
\]

with the noise term \(N\), describing the electric noise of the detector, the stochastic term \(S\), describing effects like fluctuations of the shower\(^2\) and the constant term \(C\), which describes among others miscalibrations of the calorimeter. For energies in units of eV. Calorimeters are either optimized to measure electrons and photons or hadrons. A characteristic quantity of hadron calorimeters is the ratio \(e/h\) of the detection efficiencies for electromagnetic and hadronic energy deposits in a particle shower \([32, 57]\). It can be measured using

\[
\frac{\pi}{e} = \frac{f_{\text{em}} e + (1 - f_{\text{em}}) h}{e} = \frac{1 + (e/h - 1)f_{\text{em}}}{e/h} ,
\]

the ratio of the calorimeter response to pions and electrons of the same energy. Because \(f_{\text{em}}\), the electromagnetic fraction, is dependent of the energy, \(\pi/e\) also depends on the energy leading to non-linear calorimeter response.

Electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) \([58]\) is placed after the tracking system. It is used to measure the energies of photons and charged particles by measuring the energy deposited in the calorimeter material made of lead tungstate crystals (PbWO\(_4\)). When photons with high energies are in the vicinity of a nucleus with atomic number \(Z\), they can produce pairs of electrons and positrons with a cross-section proportional to \(Z^2\). Electrons (and positrons) with high energies emit bremsstrahlung when deflected by the electromagnetic field of nuclei. These are the dominant effects for photons and electrons with energies above the order of GeV. Because these effects induce each other, electrons and photons produce a so called electromagnetic cascade. In this scenario, the energy loss is proportional to \(\exp(-x/X_0)\), with the radiation length \(X_0\). The shower stops when a

\(^2\) One example for this would be the length of the individual tracks in the cascade.

\(^3\) The radiation length is defined as the distance an electron can travel before its energy is reduced to 1/e through bremsstrahlung.
critical energy $E_c$ is reached, which is dependent on the atomic number $Z$. At energies below $E_c$, bremsstrahlung is no longer the dominant effect of energy loss for electrons so fewer photons are produced. The quantities $X_0$ and $E_c$ are connected to the Molière radius $R_M \propto X_0/E_c$, which describes the lateral width of a shower. These effects are, for example, described in [59].

The PbWO$_4$ crystals used in the ECAL of CMS have a radiation length $X_0$ of 0.89 cm as well as a Molière radius $R_M$ of 2.19 cm and the transparent crystals are also used because of their good scintillation properties. The light from the scintillation is reflected from the polished crystal surface and recorded with photo diodes. This signal is approximately proportional to the energy that was deposited in this crystal.

The ECAL in the barrel region ($|\eta| < 1.479$) consists out of 61200 crystals with a length of 23 cm (or 25.84 · $X_0$) and a front size of 2.2 cm × 2.2 cm. The endcaps contains 7324 crystals with a lengths of 22 cm and a front size of 28.62 mm × 28.62 mm, covering a range between 1.479 < $|\eta|$ < 3.0. The crystals are tilted with respect to the nominal interaction point to avoid cracks in which particles can escape undetected for both regions. Additionally the end caps house the preshower detector with the main purpose of resolving photons from neutral pion decays. It is built as a two layer sampling calorimeter with a lead absorber and silicon strip detector.

Using test-beam data [60,61], the performance of the ECAL has been measured. While the response is constant for electrons with different energies this is not the case for pions. This leads to a $e/h$ of 1.6 and a relative energy resolution of

$$\frac{\sigma(E)}{E} = \frac{0.124}{E} \oplus \frac{0.036}{\sqrt{E}} \oplus 0.0026,$$

with energy $E$ in units of GeV. For electrons with 50 GeV the relative resolution is 0.6%.

### Hadronic calorimeter

The purpose of the hadronic calorimeter (HCAL) [62] is to measure the energy of all strongly interacting particles, which includes all stable particles excluding electrons, muons, neutrinos and photons. It is a sampling calorimeter in which absorber- and active material are placed in alternating order. Incoming particles produce a hadronic shower in the absorber material, which is measured in the active material. For the HCAL of the CMS detector, brass is used as absorber because it is non-magnetic and has a short hadronic interaction length $\lambda^h$ of 16 cm. The active material consists of plastic scintillators.

Like the ECAL, the HCAL consists of multiple subsystems. The HB in the barrel covers an area of $|\eta| < 1.3$, filling the whole space between ECAL and magnet. The HE in placed in the endcap covering $1.3 < |\eta| < 3.0$. Because the hadron flux in the pseudorapidity region $3.0 < |\eta| < 5.2$ is very high, an additional forward calorimeter (HF) was installed with steel as absorber material interspersed with quartz fibers.

Using the same data as for the performance measurements of the ECAL, the ratio $e/h$ was determined to be 1.4.

The measured relative energy resolution for single hadrons of the HCAL and ECAL combined is parametrized by

$$\frac{\sigma(E)}{E} = \frac{1.2}{\sqrt{E/\text{GeV}}} \oplus 0.069,$$

which leads to 18% for pions with 50 GeV energy.

---

4 The hadronic interaction length, defined analogous to $X_0$, is the length a hadron needs in a certain type of matter to lose all but $1/e$ of its energy.
4.2.5. Muon system

Another integral part of the CMS detector design is the muon system [63]. In contrast to electrons the dominant form of energy loss of highly energetic muons is ionization, which is described by the Bethe-equation (explained e.g. in [59]). Muons with momenta above a few hundred MeV are called minimum ionizing particles (MIPs) because of their small energy loss through ionization. This is the reason that muons deposit almost no energy in the calorimeters and have to be identified by their tracks.

The muon chambers of the CMS detector are placed outside the solenoid because muons are the only detectable particles that usually get past the HCAL. The muon chambers have to cover the complete outside of the detector and to lower the costs gaseous detector designs are used. The drift tube chambers (DT) in the barrel region cover a region of $|\eta| < 1.2$. The tubes are filled with an Ar-CO$_2$ gas mixture and are positioned in a way to either measure the position in $\phi$ or $z$. In the tubes, a wire is acting as the anode with two cathode strips placed at the tube walls. Charge carriers are accelerated in the field between anode and cathode and can be measured in the form of an electrical signal. Cathode strip chambers (CSC) are used in the endcaps, covering a region of $0.9 < |\eta| < 2.4$. In the CSCs multiple cathode strips are placed vertically to the anode allowing a position measurement of the muon in two dimensions. In addition resistive plate chambers (RPC) are installed, in both the barrel and the endcaps, providing good time resolution.

4.2.6. Trigger system and computing

To cope with the high rates and luminosities at the LHC, the CMS detector is equipped with a trigger system and computing infrastructure (described in detail e.g. in [64–66]) capable of filtering the signals to a more manageable amount.

At the LHC, bunch crossings happen with a frequency of up to 40 MHz. Each bunch crossing (or event) produces approximately one MByte of data, while the storage rate is technically limited to a few 100 Hz. To reduce this a trigger with two steps is used:

**Level-1 trigger (L1)** This trigger reduces the rate to about 100 kHz. It uses coarse data from the muon and calorimeter triggers to decide if an event is stored. These triggers use data like muon chamber hits or energy deposition in calorimeter towers to generate primitive objects that can be interpreted and used to decide if an event should pass the L1 trigger. Because very fast response is needed, the L1 trigger is located in the cavern near the detector. The trigger needs 3.2 $\mu$s for a decision, during which incoming events need to be buffered.

**High level trigger** To further decrease the rate, events that pass the L1-trigger are processed by the high level trigger (HLT). This is a software trigger running on a processor farm located in the building above the CMS cavern. It uses the complete detector read-out to compute a basic reconstruction, which is used to select events with much more sophisticated criteria than in the L1 stage. This reduces the rate of events to several 100 Hz.

To process and save this very large amounts of data the Worldwide LHC Computing Grid (WLCG) [67] was developed and constructed. This grid is centered around the tier-0 computing centers at CERN and the Wigner Research Centre for Physics in Budapest, Hungary. Together with the thirteen tier-1 centers, located around the world (one is located at the KIT), the raw data is reconstructed, processed and saved. For backup and storage of the individual analysis data a large number of tier-2 centers, hosted at institutes around the world, were established.
4.3. Object reconstruction

An event recorded in a detector consists of energy depositions measured in the calorimeters and hits in the tracking systems. These quantities are used to reconstruct particles and other objects. Tracks are reconstructed from the tracker hits using an algorithmic procedure that is based on a maximum likelihood fit of the track-hit residuals and energy depositions are clustered. These information are connected using the particle flow algorithm, which is used to reconstruct leptons, photons and hadrons.

4.3.1. Track and vertex reconstruction

In each event, thousands of tracker hits are recorded and fitting tracks and vertices to all these hits is a difficult and computationally expensive task. The algorithms deployed at CMS for this task will be introduced in this section and are described in detail in [68]. During the reconstruction procedure, the hits need to be converted from their local coordinate system to the global coordinate system of the track. This procedure takes into account that the actual position of the tracker can be different from the assumed geometry. This deviation is measured using an alignment procedure [69].

At CMS, the Combinatorial Track Finder (CTF), which is based on the Kalman filter [70], is used for the reconstruction of tracks utilizing pattern recognition and track fitting inside the same framework. In multiple iterations of the CTF track reconstruction sequence tracks are found and reconstructed. This sequence includes the following steps:

- A track seed is generated providing track candidates using only two to three hits. These candidates also provide a first estimate of track momentum and uncertainty.
- Additional hits are assigned to the candidate by extrapolating the seed trajectory following the expected path of the charged particle.
- Tracks are fitted, providing the best possible estimate for the parameters of the trajectory.
- A quality measure is calculated for the trajectory. If it does not pass certain criteria, it is discarded.

After each iteration the hits assigned to a high-purity trajectory are removed from the list of hits, which decreases the complexity of the event.

Vertices can originate from proton-proton collisions (primary vertices) and from particle decays (secondary vertices). They are reconstructed by finding points in which tracks converge. Primary vertices are reconstructed using an adaptive vertex fit [71] with deterministic annealing clustering [72]. The reconstruction of secondary vertices is performed in similar way, described e.g. in [73].

4.3.2. Lepton reconstruction

Electrons

The CTF could also be used for electron reconstruction. However, since electrons lose a substantial amount of their energy through bremsstrahlung in tracker material, the hit reconstruction and parameter estimation would not work well. For this reason, a dedicated algorithm is used for electrons [74]. The flight path of an electron is a helix shaped trajectory and therefore, the energy deposition in the ECAL is expected to be spread in $\phi$ direction. The electron reconstruction starts by combining the depositions, or cluster, in the ECAL to superclusters, which takes the spread of the energy depositions into account. After these superclusters are connected to the track seed the Gaussian sum filter (GSF) [75] algorithm is used for the reconstruction of the electron tracks.
4.3. Object reconstruction

Muons

Muons are separately reconstructed in the muon system and the tracker [76]. A global muon has tracks reconstructed in both systems that can be matched. The overall fit is done using the algorithm described in section 4.3.1. The second possibility is a muon for which a track in the silicon tracker could be reconstructed. After this, the track is extrapolated considering the magnetic field, energy loss and multiple scattering and if a hit in the muon system can be matched this extrapolation, this muon is classified as tracker muon.

4.3.3. The particle-flow algorithm

The calorimeters are vital for energy measurement of photons and neutral hadrons and the energy resolution for highly energetic particles is very good. Charged particles with low energies can be measured better in the tracker, which relies on the measurement of the track’s helix radius. Because the momentum \( p \) is proportional to \( R^{-1} \) of the helix path, charged particles with low energies can be measured with high precision. In addition to this, the tracker is superior in measuring the initial flight direction of these particles.

Because most particles can be measured in multiple detector subsystems, the CMS particle-flow (PF) algorithm [77] is used for reconstructing stable particles by combining information from all subdetectors and exploiting their synergies. The particles returned by this algorithm, i.e. muons, photons, electrons and hadrons, can be used to reconstruct more complex objects.

Calorimeter clustering

The energy depositions in the calrorimeter cells of ECAL, HCAL and preshower detector are combined to clusters and used for particle reconstruction [77]. Topological clusters are grown by combining significant energy deposits in cells surrounding local maxima of calorimeter energy. When no further cells can be added to these topological clusters a PF cluster is generated. The energy of the PF clusters is determined by weighting the cells energy contribution in accordance to the distance between cluster and the associated cells.

Linking

Linking tracks and calorimeter clusters to blocks is one of the main feature of the PF algorithm. These blocks usually contain only a few elements and are identified with particles.

Reconstruction of particle flow candidates

The PF algorithm uses different definitions to find certain particles. If a particle is found, its energy depositions in HCAL and ECAL are removed, leading to lower computational complexity. The PF candidates are reconstructed in the following order:

**PF muons** are global muons, whose momentum is compatible with the momentum measurement in only the tracker within three standard deviations.

**PF electrons** are obtained from the electron reconstruction described above and can also be derived from ECAL and tracking variables.

**PF charged hadrons** are derived form the remaining tracks. For this, only tracks are considered with momentum uncertainty smaller than the calorimetric energy resolution, cf. equations (4.12) and (4.13).

**PF photon** are identified if the energy deposition in the ECAL is much larger than the energy of the assigned tracks.
PF neutral hadrons are found similar to PF photons expect that HCAL energy deposition is considered. The combination of all this information leads to better energy measurement of jets, clustered from PF candidates, compared to classical calorimeter jets\footnote{In calorimeter jets the energy of jets is only reconstructed using the clustered calorimetric energy deposition.}, which also leads to improved $E_T$ measurement.

4.3.4. Jets

The QCD confinement of quarks and gluons is the reason that both cannot be measured directly. From partons produced in a hard interaction, a collimated spray of strongly interacting particles develops in the original direction of the parton. By grouping these showers using clustering algorithms, jets are produced. Jets are also needed from a theoretical perspective, because the concept of free partons is ambiguous. In perturbative QCD the probability for soft gluon emission and splitting is divergent and therefore it is impossible to distinguish between one parton and two collinear partons.

Jets are defined by so called jet definitions, which are a set of rules describing how to group the particles and assign momentum to the resulting jet. Ideally this definition should be applicable to partonic calculations, experimental measurements, like PF candidates or calorimeter clusters (see figure 4.3), and the output of parton-showering Monte Carlo simulations, which lead to so called generator jets, or generator-level jets. This ensures that measurement, simulation and theory work with a common definition. An overview of commonly used algorithms can be found in \cite{78}. They can be grouped in cone and sequential recombination algorithms, with the latter being more robust and commonly used at the LHC. By using $p_T$ and $\Delta R$ as energy and distance measures, invariance under boosts along the beam line can be achieved. In the context of this thesis only the anti-$k_T$ algorithm \cite{79} is relevant. The distance measures of this algorithm are defined as

\begin{align}
    d_{ij} &= \min \left( \frac{p_{T,i}^2 - p_{T,j,i}^2}{\Delta R_{ij}^2} \right), \\
    d_{iB} &= p_{T,i}^2,
\end{align}

with the radius parameter $R$, the transverse momentum $p_{T,i}$ and rapidity $y_i$ of the particle $i$. The algorithm begins by calculating the particle-particle distance $d_{ij}$ and the beam-particle distance $d_{iB}$ for each particle pair and finding the minimum. If it is a $d_{ij}$, the particles are combined and the four momenta are added. If the minimum is a $d_{iB}$ the particle is called a jet and it is removed from the collection of input particles. The next step is to recalculate all distances and again find the minimum. This procedure is repeated until no particles are left in the collection. Because the distance measure is defined in a way to first cluster high $p_T$ particles, it leads to infrared and collinear safe jets with a round shape in $\eta-\phi$ space and a maximum radius of $R$.

All jet used in this thesis are clustered with a radius parameter of $R = 0.4$.

4.3.5. Missing transverse energy

Neutrinos are produced in a multitude of electroweak processes but because their interaction probability is nearly zero, they can escape the detector undetected. Only a negligible part of the momentum of the colliding protons is directed transverse to the beam direction, it is expected that the interaction products are balanced in the transverse plane. Thus, imbalances are an indication of undetected particles, called missing transverse momentum

\[ E_T = - \sum_{\text{visible}} \vec{p}_T = \sum_{\text{invisible}} \vec{p}_T. \]
4.3. Object reconstruction

Figure 4.3.: Sketch of a jet produced in a proton-proton collision [80]. Jets can be clustered on different levels dependent on the objects used for the clustering.

Because many theories beyond the standard model predict weakly interacting particles, that would (similar to neutrinos) escape the detector undetected, the $E_T$ measurement is crucial for such searches.

4.3.6. Tagging jets originating from bottom quarks

The identification of jets originating from bottom quarks is especially important in analysis targeting bottom and top quarks, like the search from $t\bar{t}H$ presented in this thesis. The B mesons, originating from the bottom quarks of the hard process, have a long lifetime (a few ps) despite their large mass. The reason for this is that bottom quarks can only decay via weak interactions to lighter quarks, which is suppressed because of the small elements of the CKM matrix describing these decays. This long lifetime leads to an observable displacement, of the order of mm, between the primary vertex and the vertex produced when the B meson decays. This is illustrated in figure 4.4.

In this thesis the combined secondary vertex algorithm (CSV) [73] is used for the b tagging of jets. This algorithm uses information about tracks, e.g. the displacement usually measured by the impact parameter $d_0$, and reconstructed secondary vertices, e.g. the distance between primary and secondary vertex in the transverse plane $L_{xy}$ or in three-dimensional space also called the 3D flight length. This information is combined in a multivariate method which is used to calculate a value, denoted as the CSV value, for each jet. The CSV value can take on values between zero and one and describes how likely it is that a jet originated from a bottom quark (indicated by higher CSV values). When used in an analysis one or more cuts on this value are specified and called working points to determine if a jet is b tagged. For the analysis presented in this thesis a loose, medium and tight working point are defined, corresponding to a mis-tag probability of approximately 10%, 1% and 0.1% respectively.

---

6 Mis-tagged jets originate from charm or lighter quarks but the b-tagging algorithm classified them as b jets. This is studied in simulation using Monte Carlo truth information to derive cuts on the b-tag discriminant corresponding to certain mis-tag efficiencies.
Figure 4.4.: Illustration of secondary vertex displacement form the primary vertex. The impact parameter $d_0$ is used to measure the displacement of tracks not originating from the primary vertex. The distance between the primary and secondary vertices (dashed magenta line) is denoted as the 3D flight length or $L_{xy}$, if only the transverse plane is considered. [81]
5. Event Simulation

Quarks and gluons are subject to QCD confinement and cannot be directly observed. In order to interpret the data recorded by the CMS detector, typically simulations of proton-proton collisions and the resulting particles are used. Furthermore reconstruction methods and the detector resolution have a big influence on the measurements. This leads to the fact that the data can only be interpreted using physical models, if it is possible to simulate the processes between the theoretical parton-picture and the observable data. For this, computationally expensive simulations, including all particles of the hard process, underlying event and pile-up collisions, their fragmentation and hadronization as well as interaction with the detector, have to be carried out. After this, the simulated and recorded data are available in an equivalent format and observables can be compared. Furthermore, the particle information from the simulation can be used to study analysis methods or train multivariate algorithms.

This chapter will introduce the basic approach for generating simulated datasets in section 5.1. Following this, in section 5.2, the specific datasets used in this thesis will be described in detail and in section 5.3 it will be described, how these datasets need to be corrected to match the recorded data.

5.1. Event generation

High-energy particle collisions are generated utilizing Monte Carlo methods and are based on the QCD factorization theorem. To generate events, first the matrix element (ME) for the hard partonic subprocess is calculated. This is simulated by dedicated Monte Carlo generators, like MG5_aMC@NLO \[82\] (in the following MadGraph will denote the leading order (LO) part and MG5_aMC@NLO the next-to-leading order (NLO) part of this software) or POWHEG \[83\] \[84\]. The type and momenta of the initial partons is chosen from a parton density function (PDF). Commonly used PDFs are calculated by the MMHT \[85\] \[2\], NNPDF \[44\] and CTEQ \[86\] groups. The final state is sampled from the probability distribution obtained by integrating, e.g. with Monte Carlo integration, over the final phase space. Further evolution of the event is simulated by parton shower (PS) programs like PYTHIA \[87\] or HERWIG \[88\] \[89\].

This includes simulation of strongly interacting particles interacting by

---

1 A detailed introduction to the Monte Carlo technique can e.g. be found in \[32\], chapter 40.
2 Previous versions of this PDF were called MSTW \[34\] but due to a change in personal, the name was altered.
3 Both, PYTHIA and HERWIG are so called multi-porpoise Monte Carlo generators, also capable of generating the matrix elements, but the PS module can be used independently, if an appropriate interface is available.
gluon splitting, gluon emission or quark-antiquark pair production. This is also called particle cascade (or shower). At this step it has to be considered that processes with radiation, e.g. $t\bar{t}+1$ jet, can be calculated in the ME but can also be generated by the PS after the $t\bar{t}$ ME. To avoid this double counting (visualized in figure 5.1) algorithms are deployed, that decide which processes are calculated by the ME or the PS generator. An overview of commonly used matching or merging algorithms can be found in [90].

The particle cascade is continued until a cutoff energy scale is reached, at which color neutral hadrons are formed. Because this cannot be described with perturbation theory (as discussed in section 2.1.3), phenomenological models are consulted for the hadronization[4]. Well established models are the Lund string model [91], used in Pythia, and the cluster hadronization model [92], used in Herwig. Subsequent to this, all unstable particles are decayed until only stable, long lived particle remain.

In addition to this, the PS is also responsible for the simulation two further effects: The hadronization of the remnants of the proton-proton collision and potentially further hard processes is called the underlying event. Additional proton-proton collisions can occur during the recording of the actual event. Collisions within the actual bunch crossing are called in-time pile-up, while the same from previous of following bunch crossings are called out-of-time pile-up. Both are simulated by superimposing additional proton-proton collisions during the event generation procedure. They are generated with a cross section, denoted minimum-bias cross section.

Following this a detector simulation is executed. For this, the complete CMS detector was implemented in the GEANT4 framework[93] to simulate among others the effects of the magnetic field or the detector material. After this, the simulated data is available in the same format as the raw data recorded with the detector. The simulation furthermore includes data on all generated particles, which is usually denoted as Monte Carlo truth information. Particles that are clustered from this particles, are called generator-level jets. Usually, neutrinos are excluded when clustering generator-level jets because such jets are more similar to jets clustered from PF candidates. If not otherwise indicated, in this thesis, generator-level jets are clustered with neutrinos to include certain effects, described in section 8.1.

5.2. Monte Carlo simulated datasets

Simulated dataset are centrally produced by the CMS collaboration. This section focuses on describing the datasets of particular interest for this thesis. The datasets correspond to signal and background processes[5] generated with a center-of-mass energy of 13 TeV. The datasets are generated with a number $N_{\text{Gen}}$ of events. This number does not correspond to the number of events expected in recorded data but should be larger to avoid statistical fluctuations in the simulation affecting the analysis. Therefore an event weight $\omega_{\sigma}$ is calculated and the events of a specific process are reweighted, to correspond to the number of events expected from the theoretical cross section $\sigma_{\text{theo}}$, using the most precise calculation, at least NLO QCD, at the integrated luminosity $\int L dt$:

$$\omega_{\sigma} = \frac{\sigma_{\text{theo}} \cdot \int L dt}{N_{\text{Gen}}}$$

(5.1)

It is also common to generate exclusive processes only considering certain decay modes for which the theoretical cross section is multiplied by the branching fraction. One example would be the separate generation of $t\bar{t}H$ in the $H \rightarrow b\bar{b}$ and $H \rightarrow \text{non } b\bar{b}$ final states. A technique, used e.g. by the generator aMC@NLO is to use negative events weights at

---

4. The cutoff scale is also phenomenologically determined and is usually chosen to be 1 GeV.

5. Which datasets are signal and background naturally depends on the specific focus of an analysis.
5.2. Monte Carlo simulated datasets

Figure 5.1.: The same processes can be generated in the matrix element (ME) and the parton shower (PS), which leads to the possibility of double counting. In vertical direction, Feynman diagrams of LO processes with increasing ISR and FSR multiplicity are shown. The additional emissions generated by the PS are shown in horizontal direction. The diagonal diagrams are equivalent and therefore would be double counted without counteracting measures. Taken from [94]

Table 5.1.: Summary of Monte Carlo datasets relevant for the analysis presented in this thesis. A detailed description of the software used for this datasets is can be found in section 5.2. Pythia8 was used for the parton shower in all samples. The sources of the listed cross sections are described in section 5.2.1.

<table>
<thead>
<tr>
<th>Process</th>
<th>Remark</th>
<th>Generator</th>
<th>(N_{\text{Gen}})</th>
<th>(\sigma_{\text{theo}}[\text{pb}])</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\bar{t}tH, H \rightarrow b\bar{b})</td>
<td>(m_H = 125) GeV</td>
<td>Powheg</td>
<td>3912212</td>
<td>0.2918</td>
</tr>
<tr>
<td>(\bar{t}tH, H \rightarrow \text{non } b\bar{b})</td>
<td>(m_H = 125) GeV</td>
<td>Powheg</td>
<td>3860872</td>
<td>0.214</td>
</tr>
<tr>
<td>(\bar{t}t)</td>
<td>(t) channel</td>
<td>Powheg</td>
<td>92925926</td>
<td>831.76</td>
</tr>
<tr>
<td>(\bar{t}\bar{t})</td>
<td>(t) channel</td>
<td>Powheg</td>
<td>3279200</td>
<td>45.34</td>
</tr>
<tr>
<td></td>
<td>(t) channel</td>
<td>Powheg</td>
<td>1682400</td>
<td>27.98</td>
</tr>
<tr>
<td></td>
<td>(tW) channel</td>
<td>Powheg</td>
<td>998400</td>
<td>35.6</td>
</tr>
<tr>
<td></td>
<td>(\bar{t}W) channel</td>
<td>Powheg</td>
<td>985000</td>
<td>35.6</td>
</tr>
<tr>
<td>(Z/\gamma* \rightarrow \ell\bar{\ell} + \text{jets})</td>
<td>(m_{\ell\bar{\ell}} \leq 50) GeV</td>
<td>MadGraph</td>
<td>91350867</td>
<td>6025.2</td>
</tr>
<tr>
<td>(W + \text{jets}, W \rightarrow \ell\nu)</td>
<td>(100 &lt; H_T \leq 200) GeV</td>
<td>MadGraph</td>
<td>275469781</td>
<td>1345</td>
</tr>
<tr>
<td>(W + \text{jets}, W \rightarrow \ell\nu)</td>
<td>(200 &lt; H_T \leq 400) GeV</td>
<td>MadGraph</td>
<td>19851624</td>
<td>359.7</td>
</tr>
<tr>
<td>(W + \text{jets}, W \rightarrow \ell\nu)</td>
<td>(400 &lt; H_T \leq 600) GeV</td>
<td>MadGraph</td>
<td>7432746</td>
<td>48.91</td>
</tr>
<tr>
<td>(W + \text{jets}, W \rightarrow \ell\nu)</td>
<td>(600 &lt; H_T \leq 800) GeV</td>
<td>MadGraph</td>
<td>3722395</td>
<td>12.05</td>
</tr>
<tr>
<td>(W + \text{jets}, W \rightarrow \ell\nu)</td>
<td>(800 &lt; H_T \leq 1200) GeV</td>
<td>MadGraph</td>
<td>6341257</td>
<td>5.501</td>
</tr>
<tr>
<td>(W + \text{jets}, W \rightarrow \ell\nu)</td>
<td>(1200 &lt; H_T \leq 2500) GeV</td>
<td>MadGraph</td>
<td>7063909</td>
<td>1.329</td>
</tr>
<tr>
<td>(W + \text{jets}, W \rightarrow \ell\nu)</td>
<td>(H_T \geq 2500) GeV</td>
<td>MadGraph</td>
<td>2507809</td>
<td>0.032</td>
</tr>
</tbody>
</table>
generation time to counteract double counting of NLO corrections. If such techniques are applied, the $N_{\text{Gen}}$ used in the denominator of (5.1) needs to be modified according to $N_{\text{Gen}} = N_{\text{Gen,positive}} - N_{\text{Gen,negative}}$.

The datasets used in this thesis are listed in table 5.1. The ttH signal sample was generated with the NLO generator POWHEG (v.2), where the Higgs-boson mass was assumed to be 125 GeV and the top-quark mass was set to 172.5 GeV. For the description of the proton substructure, the PDF NNPDF3.0 was used. Following the simulation of the hard process, the parton shower and hadronization were simulated using PYTHIA (v.8.2). The top-quark related backgrounds, $\bar{t}t$ and single top (in the t and tW channel), were also generated using POWHEG (v.2), while the remaining samples describing Z bosons and W bosons in association with jets (denoted as Z+Jets and W+Jets) were generated using the LO module of MG5_aMC@NLO (v.2.2.2) (usually abbreviated with MadGraph). The parton showering software and PDF are the same as for the ttH sample. For the PS, different tunes are available. The PS of all listed datasets was used with parameters defined by the CUETP8M1 tune [95,96], which describes the characteristics of the underlying event.

5.2.1. Calculation of the theoretical cross sections

Most of the theoretical cross sections listed in table 5.1 are available at NNLO precision with electroweak corrections. The listed datasets themselves are generated at LO or NLO precision. By utilizing the normalization procedure defined in equation (5.1) they are scaled to these more precisely calculated cross sections.

For the ttH signal processes, the cross section recommended by the LHC Higgs cross section working group [41] is used. The calculation considered include NLO QCD corrections [97–101] as well as electroweak NLO corrections [102–104]. As PDF, NNPDF2.3QED [105] was used for $\gamma$-induced processes, while the PDF4LHC recommendations [106] were used for all other processes to determine PDF uncertainty.

The cross section of $\bar{t}t$ production (the main background) is calculated at NNLO precision with the Top++2.0 software package [107] and the single top-quark backgrounds in the s- and t-channel are calculated with the software HATHOR [108,109]. PDFs were used according to the, at this time, available interim PDF4LHC recommendation [110], which is also used for the $\bar{t}t$ calculation. For W+Jets the cross sections in $H_T$ bins are calculated by using the cross section predicted by MG5_aMC@NLO in LO and multiplying it with an NNLO “k-factor”. This factor is determined by dividing the NNLO prediction, calculated with FEWZ [111,112], by the inclusive LO cross section. The FEWZ software package is also used in the calculation of the Z+Jet cross section for the dilepton [113] invariant mass ranges listed in table 5.1. For both the W- and Z+Jet cross section CTEQ6.6 [114] is used as PDF.

5.2.2. Additional jet flavor identification in $\bar{t}t$

The $\bar{t}t$ sample is further separated in processes based on the flavor of additional jets that do not originate from the top-quark decay. This is implemented with a tool described in [115], using generator information to identify jets with B or C hadrons. These hadrons are then traced back in the generator history to determine if they originated from a $t \to Wb$ decay. All events are then classified either as

$\bar{t}t + b\bar{b}$ if two additional b-jet were found,

$\bar{t}t + b$ if one additional b-jet was found.

At multiple points within this thesis, a slightly different set of simulated datasets is used, listed in the appendix in table B.1, mainly differing in minor background processes and the detector simulation. It will be clearly highlighted if this different set of simulations is used.
\[ \bar{t}t + 2b \] if one additional b-jet was found, but \( \geq 2 \) B hadrons were found in the jet,
\[ \bar{t}t + c\bar{c} \] if additional c-jets and no additional b-jet were found or
\[ \bar{t}t + \text{lf} \] in any other case.

5.3. Corrections to simulated data

The simulated data does not describe the recorded data perfectly. The reason for this are among others that approximations and assumptions are made in Monte Carlo simulations. Therefore, the simulation needs to be corrected based on observations in the recorded data. This accomplished by deriving scale factors that are applied to events or single objects like leptons or jets.

5.3.1. Jet energy resolution correction

Differences in the jet energy resolution were observed between recorded and simulated data. These corrections are only applied to simulated events and are derived using an asymmetry technique on dijet events \[116\]. The transverse momentum asymmetry

\[ A = \frac{p_{T,1} - p_{T,2}}{p_{T,1} + p_{T,2}} \] (5.2)

is defined for a dijet systems with jets 1 and 2. The width of this distribution is related to a measure of the jet energy resolution and the corresponding scale factors are derived as a function of \( \eta \). In the analysis, these scale factors \( f_{\text{JER}} \) are applied by shifting the reconstructed jets in the Monte Carlo simulation with the difference between reconstructed jet and associated particle-level jet following the formula

\[ p_{T,\text{corr}}\text{reco} = p_{T}\text{reco} + f_{\text{JER}} (p_{T}\text{reco} - p_{T}\text{gen}) \] (5.3)

5.3.2. Correction of the b-tag discriminant

The CSV algorithm, described in section \[4.3.6\], used for b tagging plays a crucial role in the search for \( t\bar{t}H \) in the \( H \to b\bar{b} \) final-state. It strongly relies on effects described by the hadronization and the parton shower in simulation. Because approximations and phenomenological models are an integral part of the simulation of such processes, it is not unexpected that the distributions important for b tagging exhibit differences, when comparing recorded data and simulation. Corrections are derived with a tag-and-probe method, separately for heavy- and light-flavor jets, where the latter denotes jets originating from up, down and strange quarks as well as gluons \[73\].

In a control region, fulfilling the requirements of exactly two leptons and two jets, a heavy-flavor enriched sample is selected by using criteria on the dilepton mass and \( p_T \) targeted at rejecting events with Z-boson decays. Additionally, one jet has to pass the medium b-tagging working point, cf. section \[6.2\] denoted as the tag jet. This selection is applied to recorded and simulated data, where the latter contains the Monte Carlo truth information about the jet flavor. The simulated dataset is either a \( \bar{t}t \) sample, for the determination of the heavy-flavor (HF) scale factor, or a Z+Jets sample, when calculating the light-flavor (LF) scale factor. In this case the cuts are inverted to include the Z boson and the tag jets have to fail the loose b-tag working point.

The scale factors are derived as a function of the jet’s b-tagging output and \( p_T \), in case of HF, and additionally of \( \eta \) for LF jets. To derive the scale factor, the according values of the probe jet are used. The first step is to scale the event yield of the simulation to the recorded data, cf. equation \[5.1\], which is subsequently divided in aN LF and aN HF.
component. For the derivation of the HF scale factor $f_{HF}$, the yields of recorded data and the LF component of the simulated data are subtracted and divided by the HF component of the simulated data. For the LF scale factor the same is done vice versa. Therefore the scale factors are given by:

$$f_{HF/LF}(\text{b-tag output}, p_T(\eta)) = \frac{N(\text{data}) - N(\text{MC}_{LF/HF})}{N(\text{MC}_{HF/LF})}.$$  \hspace{1cm} (5.4)

Because the knowledge of the LF component in the HF sample, and vice versa, is required for the subtraction from the data yields, a iterative process has to be implemented. For this, the obtained scale factors are applied and new scale factors are determined. This is repeated until the result converges. After this is the case, a sixth-order polynomial function is fitted to the scale factors, to reduce the effects of statistical fluctuations.

Because no proper control region is available for jets from charm quarks, no scale factors are derived. For such jets, a scale factor of 1 is applied, with doubled uncertainty of $f_{HF}$.

In the analysis an event weight is calculated with

$$f_{\text{event}} = \prod_{i} f_i,$$  \hspace{1cm} (5.5)

where $f_i$ is the scale factor of jet $i$.

### 5.3.3. Pile-up correction

The pile-up in Monte Carlo samples needs to be corrected because the profile of the pile-up is dieectly dependent on the instantaneous luminosity, which is only known after the data-taking. This correction is calculated as an event weight with

$$\omega_{PU} = \frac{f_{data}(\langle N_{PU}\rangle)}{\rho_{MC}(\langle N_{PU}\rangle)}.$$  \hspace{1cm} (5.6)

The fraction of events $f_{data}$, expected to contain $\langle N_{PU}\rangle$ secondary collisions can be calculated from the instantaneous luminosity, while the number of secondary interactions in the Monte Carlo simulation is sampled from the distribution $\rho_{MC}$. A value for the inelastic proton-proton cross section of 69.4 mb was determined by CMS, because of its good data description.
6. Event Selection

In this chapter, the definitions of the physics objects used in this thesis will be introduced in section 6.1. As previously discussed in section 2.2.3.2 an essential part of a \( t\bar{t}H, H \to b\bar{b} \) analysis is the selection of signal enriched samples from recorded data. This is realized via multiple selection criteria, which will be discussed in section 6.2.

6.1. Physics Objects

The objects returned by the particle flow algorithm, introduced in section 4.3.3, are not directly used for the analysis. Instead they are further processed, following the recommendations of the CMS top physics analysis group (TOP PAG)\[117\].

6.1.1. Leptons

Electrons and Muons are reconstructed and measured very differently and therefore, individual requirements were formulated.

6.1.1.1. Electrons

The electrons obtained from the PF algorithm are required to pass certain criteria, formulated by the CMS \( e\gamma \) physics object group (POG) \[118\].

**Electron isolation**

The relative electron isolation \( I_e \) is mainly used to separate prompt electrons, that originate from the hard process, from secondary electrons from the weakly decaying hadrons as well as fake electrons. The latter are often found in close proximity to the a jet and therefore a distance measure, the isolation, was introduced to effectively suppress unwanted electrons.

For the identification of isolated electrons a cone with a diameter of \( \Delta R = 0.3 \), cf. equation (4.7), is constructed for each electron candidate. After this the relative isolation

\[
I_e = \frac{\sum_{\text{charged hadrons}} p_T + \max \left( 0, \sum_{\text{neutral hadrons}} p_T + \sum_{\gamma} p_T - \rho A_{\text{eff}} \right)}{p_T,e}
\]  

is calculated. The \( p_T \) of all charged and neutral hadrons as well as photons inside the cone is summed and divided by the electron \( p_T \). The \( p_T \) of all charged hadrons is calculated,
Table 6.1.: Cut values for the trigger independent electron MVA identification, taken from [118].

<table>
<thead>
<tr>
<th>$\eta$</th>
<th>80% eff.</th>
<th>90% eff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$</td>
<td>\eta</td>
<td>&lt; 0.8$ in barrel</td>
</tr>
<tr>
<td>$</td>
<td>\eta</td>
<td>&gt; 0.8$ in barrel</td>
</tr>
<tr>
<td>in endcap</td>
<td>0.726311</td>
<td>0.358969</td>
</tr>
</tbody>
</table>

excluding hadrons not originating from the primary vertex. This exclusion is called charged hadron subtraction (CHS). The contribution of neutral hadrons and photons, from the PF algorithm, also needs to be corrected for pile-up contributions. This is estimated by subtracting $\rho A_{eff}$, with the average energy density of the event $\rho$, introduced in [119], and the $\eta$ dependent effective jet area $A_{eff}$.

**Multivariate electron ID**

A multivariate electron ID [118] is used for the analysis presented in this thesis. A boosted decision tree, cf. chapter 3, is used to classify electrons in order to suppress QCD background and secondary electrons. For this the information contained in 19 observables is used. These inputs can roughly be categorized in observables describing ECAL signals, the event clustering in the calorimeter, track measurements and photon rejection.

**Conversion rejection**

Highly energetic photons transversing through matter, e.g. inside the tracker layers, can convert to electron-positron pairs, that could be interpreted as prompt electrons. Electrons produced in the hard process should travel through all tracker layers, especially the first. Therefore, PF electrons without hits in the first tracker layer are excluded. Furthermore, electrons are excluded, if a second track with inverse curvature, indicating a positron (or electron in case the first particle was a positron), is found in close proximity.

**Loose and tight electrons**

Both loose and tight electrons have to pass the following criteria: First, electrons in the area between barrel and endcap regions of the detector are excluded, leading to the allowed region with $|\eta| < 1.444$ and $|\eta| > 1.566$. Furthermore, a relative isolation of $I_e < 0.15$ is required. Because all information usually used in cut-based electron selection are inputs for the MVA, only cuts on the PF electron candidates $p_T$, $\eta$ and the MVA value are needed. For this analysis the cuts with 80% efficiency, listed in table 6.1, are used.

Loose electrons are required to have $p_T \geq 20$ GeV and $|\eta| < 2.4$, while tight electrons need $p_T \geq 30$ GeV and $|\eta| < 2.1$.

**Soft electron**

A further independent way of defining electrons is used in this thesis to find electrons originating from hadron decays inside jets. These electrons are denoted as soft electrons, because of their very loose $p_T$ requirements. The conversion rejection is applied and the transverse and longitudinal impact parameter with respect to the primary vertex have to be smaller than 0.5 cm and 1 cm, respectively. Furthermore, a $p_T \geq 5$ GeV and $|\eta| < 2.5$ are required.

**6.1.1.2. Muons**

The muon selection recommendations at CMS are compiled by the muon POG [120].
Table 6.2.: Standard jet selection requirements. The listed variables are defined for each jet. For example the charged hadronic energy fraction denotes the fraction of the total jet energy from charged PF hadrons.

<table>
<thead>
<tr>
<th>Value</th>
<th>requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td>charged hadronic energy fraction</td>
<td>&gt; 0</td>
</tr>
<tr>
<td>neutral hadronic energy fraction</td>
<td>≤ 0.99</td>
</tr>
<tr>
<td>charged electromagnetic energy fraction</td>
<td>≤ 0.99</td>
</tr>
<tr>
<td>neutral electromagnetic energy fraction of jet</td>
<td>≤ 0.99</td>
</tr>
<tr>
<td>charged track multiplicity + neutral track multiplicity</td>
<td>&gt; 1</td>
</tr>
<tr>
<td>charged track multiplicity</td>
<td>&gt; 0</td>
</tr>
</tbody>
</table>

**Muon isolation**

Similar to the electron isolation, cf. equation (6.1), an isolation criterion for muons

\[
I_\mu = \frac{\sum_{\text{charged hadrons}} p_T + \max \left( 0, \sum_{\text{neutral hadrons}} p_T + \sum_{\gamma} p_T - 0.5 \sum_{\text{pile-up}} p_T \right)}{p_T,\mu} \tag{6.2}
\]

is constructed. The general approach is the same as for electrons. In difference to electrons, for muons a cone size of \( \Delta R = 0.4 \) is used. In this case, the neutral contribution from pile-up is corrected, using the \( \Delta \beta \) correction \[121\].

**Loose and tight muons**

The muons used in the analysis are muons reconstructed with the PF algorithm and are required to have hits in the tracker and the muon system. In the latter, at least two segments have to be hit in order to suppress, among others, punch-through hadrons and improve the \( p_T \) measurement. The isolation of the muons is required to be < 0.15 and the transverse and longitudinal impact parameters with respect to the primary vertex have to be smaller than 0.2 cm and 0.5 cm, respectively. Loose muons are required, to have a \( p_T \geq 10 \text{ GeV} \) and a \( |\eta| < 2.4 \), while tight muons need a \( p_T \geq 30 \text{ GeV} \) and a \( |\eta| < 2.1 \).

**Soft muons**

For the same reason soft electrons were introduced above, soft muons are defined. Such muons are also reconstructed with the PF algorithm but have to pass fewer selections. The transverse and longitudinal impact parameters with respect to the primary vertex have to be smaller than 0.5 cm and 1 cm, respectively and \( p_T \geq 3 \text{ GeV} \) and \( |\eta| < 2.4 \) are required.

**6.1.2. Jets**

The jets used in this thesis are obtained by clustering particle flow candidates with the anti-\( k_T \) clustering algorithm with a radius parameter of \( R = 0.4 \), cf. section 4.3.4. Charged hadrons originating from secondary collisions (pile-up) are excluded from the clustering process (CHS). Furthermore PF electrons and muons are excluded from the clustering. Jets passing the selection need to fulfill the standard requirements listed in table 6.2 to increase purity. Additionally jets with charged leptons within the distance of \( \Delta R < 0.4 \) are discarded.

Furthermore, jet energies are corrected. This is necessary because of the non-linear and \( \eta/p_T \)-dependent response of the calorimeter. The approach used at CMS is factorized in
three steps. The measurements of the individual correction is described in [116] and is in addition dependent on the state of the CMS detector at the time of data-taking, which is saved in special condition database.

The first correction is called L1 pile-up and is designed to remove any residual influence of pile-up from the jets. The second level, L2L3 MC-truth corrections\footnote{The name L2L3 MC-truth correction has historical reasons, because in the past these were two separate steps in the correction scheme. The L2 Relative correction for uniform response in \( \eta \) and the L3 Absolute correction for uniform response in \( p_T \).} are applied to make the response of the jets uniform in \( \eta \) and \( p_T \).

Furthermore, the L2L3 Residual corrections, which are only applied to data, are necessary to correct small differences that remain within the jet response in data and Monte Carlo simulation.

Jets that are corrected with these corrections, agree well with the jets from simulation.

The final cuts for reconstructed jets in this analysis are \( p_T \geq 30 \text{ GeV} \) to reduce contribution from backgrounds and pile-up and \(|\eta| < 2.4\) to only use jets from the central detector region with good reconstruction quality.

6. Event Selection

6.1.2.1. Transverse momentum response

In general, the energy measured by the detector is not equal to the energy of the physical particle. For jets this effect is described by the jet response \( \mathcal{R} \), which is defined as the ratio of the measured jet energy and the energy of the particles emerging from the initial parton before the hadronization. For hadron colliders the transverse-momentum response \( \mathcal{R} \) is used because of the special kinematic constraints. This response is defined as the ratio of the measured \( p_T \) of a detector-level jet and the \( p_T \) of an associated particle-level object, which can be, in the context of this thesis, either a particle or a generator-level jet:

\[
\mathcal{R} = \frac{p_T^{\text{Jet}}}{p_T^{\text{particle}}} \quad \text{or} \quad \mathcal{R} = \frac{p_T^{\text{Jet}}}{p_T^{\text{generator-level jet}}}. \tag{6.3}
\]

The association between these two objects is usually accomplished by geometrically matching them via a \( \Delta R \) criterion. The average response \( \langle \mathcal{R} \rangle \) should be at a value of one after calibration and will be denoted as the jet energy scale. Most deviations from this value, can be explained by systematically mismeasured quantities. The relative jet \( p_T \) resolution corresponds to the width of \( \mathcal{R} \), which e.g. gets broader because of unmeasured energy. These effects will be discussed in greater detail, for b jets, in section 8.1.

The response distribution is dominated by a central Gaussian region, and therefore, in this thesis the jet \( p_T \) response will be approximated with a Gaussian distribution

\[
G(x, \mu_{\mathcal{R}}, \sigma_{\mathcal{R}}) = \frac{1}{\sqrt{2\pi}\sigma_{\mathcal{R}}} \exp\left(-\frac{(x - \mu_{\mathcal{R}})^2}{2\sigma_{\mathcal{R}}^2}\right) \tag{6.5}
\]

where \( \mathcal{R} \) identifies the Gaussian approximation of the jet-\( p_T \) response with the mean value \( \mu_{\mathcal{R}} \) and standard deviation \( \sigma_{\mathcal{R}} \). To retrieve this information from a measured distribution, a Gaussian distribution in a symmetrical region around \( \mu_{\mathcal{R}} \) is fitted, where the range was determined “by eye”\footnote{In principle, an iterative algorithm performing the determination of the range should be used, but developing such an algorithm was not possible within the time constraints of this thesis.}.

1 The name L2L3 MC-truth correction has historical reasons, because in the past these were two separate steps in the correction scheme. The L2 Relative correction for uniform response in \( \eta \) and the L3 Absolute correction for uniform response in \( p_T \).
6.1.3. Missing transverse energy

The $E_T$ is calculated as described in section 4.3.4 and because it suffers from similar effect as jets, it is corrected with the corrections described above and the resolution corrections, described in section 5.3.1.

6.2. Event Selection

To select $t\bar{t}H$ event candidates from a data sample, a chain of selections is applied to enrich the sample with signal-like events. The different links of this chain are described in the following paragraphs.

The recorded data used, if not otherwise indicated, was collected from proton-proton collisions recorded in the time span from April 22th to July 15th, 2016, by the CMS experiment, corresponding to 12.9 fb$^{-1}$ of integrated luminosity.

6.2.1. Leptons+Jets selection

This selection chain is used for the $t\bar{t}H, H \rightarrow b\bar{b}$ search in the semileptonic channel and is designed to enrich the data with $t\bar{t}H$ events.

**Trigger selection**

As a first criterion, all events have to pass certain high level trigger (HLT) paths. For this selection the path starting with HLT$_{\text{Ele27}_{\text{eta2p1}}_{\text{WPTight}}_{\text{Gsf}}}$ was used for electron events. As indicated by the name events with electrons having $p_T > 27$ GeV, $|\eta| < 2.1$ pass this trigger. For events with muons either the HLT path starting with HLT$_{\text{IsoMu22}}$ or HLT$_{\text{IsoTkMu22}}$ were required, both selecting events with muons that have $p_T \geq 22$ GeV.

**Vertex selection**

Each event is required to have a well reconstructed primary vertex because it is needed for precise measurements. Therefore, the vertex needs to be closer than 24 cm in longitudinal and 2 cm in transverse direction to the nominal primary interaction point and the degrees of freedom of the vertex fit should be larger than four.

**Lepton selection**

Because the semileptonic channel is investigated in this analysis exactly one tight lepton and no loose leptons (electron or muon) are required for all events. This reduces the QCD multijet background greatly. This criterion is used to suppress Drell-Yan events and to separate the semileptonic from the dilepton final state.

**Jet selection**

Events are discarded, if they contain fewer than four jets. While the expected final state has six jets, the number of required jets is lessened to allow jets outside the acceptance or overlapping jets. Furthermore, this selection reduces the background from vector-boson pair production, vector bosons in association with jets and $t\bar{t}$ events.

**b-Tag selection**

The final applied selection is the requirement of at least two b-tagged jets with a CSV value passing the medium working point (CSV value $\geq 0.8$). After this, the background is dominated by $t\bar{t}$ and $t\bar{t} +$Jets events.
6.2.2. Z+Jets selection

In chapter 9 a selection, not designed for the \(t\bar{t}H\) search, will be applied for a validation technique. Events containing a Z boson decaying into a lepton pair and one or two jets with b tags should be selected. The chain for this selection is similar to the one presented above, and therefore, only the changing parts will be described.

**Trigger selection**

Because the desired events are dilepton events, the following triggers were used:

**Dimuon:**
- HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ
- HLT_Mu17_TrkIsoVVL_TkMu8_TrkIsoVVL_DZ

**Dielectron**
- HLT_Ele23_Ele12_CaloIdL_TrackIdL_IsoVL_DZ

**Lepton selection**

To select events with leptonically decaying Z bosons two tight electrons or muons with \(|\eta| \leq 2.4\) are required, where the leading lepton is required to have a \(p_T\) of more than 25 GeV, while the subleading lepton is only required have \(p_T \geq 15\) GeV.

**Z-boson window selection**

To select Z-boson decays dilepton masses between 76 GeV and 106 GeV are explicitly required to decrease non-Z-boson background.

**Jet-tag selection**

Next, events are selected with one or two jets and more than one b tag, passing the medium working point (CSV value \(\geq 0.8\)). This selection is chosen because of the requirements of the validation method described in section 9.2.1.
7. Multivariate Analysis for the search for $t\bar{t}H$ in the $H \rightarrow b\bar{b}$ final state

In this chapter, the $t\bar{t}H$, $H \rightarrow b\bar{b}$ analysis used for further studies in chapters 8 and 9 will be outlined. The statistical methods will be introduced in section 7.1. This is followed by an outline of the analysis strategy in section 7.2 and the considered uncertainties in section 7.3.

7.1. Statistical method

To test different hypotheses of the $t\bar{t}H$ cross section the ATLAS and CMS collaborations agreed upon a procedure [122] to set an upper limit on the ratio between the production cross section and the prediction of the standard model, also called signal strength modifier $\mu$. This prescription is implemented in the CMS Higgs combination software using RooStats [123], RooFit [124] and ROOT [125]. An overview of the combination of all Higgs-boson decay channels investigated at the CMS experiment with the full dataset taken in proton-proton collisions at $\sqrt{s} = 7$ TeV can be found in [126]. The following section will provide an overview of the methods used in this thesis.

The upper limit on the signal strength modifier is calculated by comparing the observed distribution of a discriminating variable — like the output of the final classifier — to the distributions obtained from signal+background and background-only predictions. The upper limit is calculated instead of the actual cross section, because based on simulations the required sensitivity cannot be reached with the current analyses and the integrated luminosity recorded so far.

7.1.1. Model and uncertainties

Statistical Model

In the limit of large number of events the Poisson distribution

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

with the expected number of events under a certain selection (category or histogram bin) $\lambda$ and the number of observed events $n$ under this selection, can be used. This is interpreted as the probability $P(n|\lambda)$ to observe $n$ events, if $\lambda$ were expected. For the $t\bar{t}H$ analysis
the relevant case is that events are distributed over \( n_{\text{bins}} \) of the output distribution of the final classifier. The probability to observe the measured distribution in data is then given by

\[
P(\mathbf{n}|\mathbf{\lambda}) = \prod_{i=0}^{n_{\text{bins}}} P(n_i|\lambda_i)
\] (7.2)

with \( \mathbf{n} = (n_1, \ldots, n_{n_{\text{bins}}}) \) and \( \mathbf{\lambda} = (\lambda_1, \ldots, \lambda_{n_{\text{bins}}}) \) as observed and expected events in each bin.

The expected events \( \lambda_i \) can furthermore be expressed as the sum of the signal and background predictions

\[
\lambda_i = \mu \cdot s_i + \sum_{\text{background}} b_i.
\] (7.3)

The signal prediction is multiplied by the signal strength modifier \( \mu \) to include possible deviation from the standard model prediction.

**Treatment of systematic uncertainties**

The signal and background expectation are themselves subject to certain uncertainties. This is handled by introducing nuisance parameters \( \theta \) into the signal and background expectations \[122\]. With the nominal parameter \( \bar{\theta} \) as the best estimate of \( \theta \), obtained from auxiliary measurements, the frequentist probability density function (pdf) \( p(\bar{\theta}, \theta) \) can be constructed. Using Bayes’s theorem the systematic uncertainty pdfs \( \rho(\theta, \bar{\theta}) \), interpreted as posterior arising from a measurement of \( \bar{\theta} \), can be expressed in the form of

\[
\rho(\theta, \bar{\theta}) \sim p(\bar{\theta}, \theta) \cdot \pi(\theta)
\] (7.4)

with the prior probability \( \pi(\theta) \), which is considered flat in the context of this analysis.

Systematic rate uncertainties describe changes in the overall normalization of processes. Examples of this class of uncertainties are theoretical cross section uncertainties for specific processes, or the uncertainty on the value of the integrated luminosity, affecting all processes in the same way. These uncertainties are modeled with the log-normal distribution

\[
\rho(\theta, \bar{\theta}) = \frac{1}{\sqrt{2\pi} \ln \kappa} \exp \left( -\frac{\left( \ln \left( \theta / \bar{\theta} \right) \right)^2}{2 \left( \ln \kappa \right)^2} \right) \frac{1}{\theta}
\] (7.5)

with the nominal value \( \bar{\theta} \) and the width \( \kappa \).

Systematic shape uncertainties on the other hand describe the shape changes of the distributions. They are modeled with a unit Gaussian centered around zero. For each shape uncertainty new distributions with the parameter shifted up and down one standard deviation are created and propagated to the final discriminant. By interpolation between the nominal and shifted distributions in each bin a change for arbitrary shifts in the parameter is calculated. If this change affects the overall normalization, it is treated as a rate uncertainty.

**Binned likelihood**

From the probability distribution of the expected events, given in equation \[7.2\], the binned likelihood

\[
\mathcal{L}(\mathbf{n}|\mu, \theta) = P(\mathbf{n}|\mu, \theta) \cdot p(\bar{\theta}, \theta)
\] (7.6)

can be formulated with the ensemble of all nuisance parameters \( \theta \) and the product of the pdfs of all nuisance parameters \( p(\bar{\theta}, \theta) \). By maximizing this likelihood function, for the observed data \( \mathbf{n} \) an estimate of the parameters \( \mu \) and \( \theta \) can be determined.
7.1.2. Limit

With the likelihood function defined in equation (7.6), an upper limit can be calculated on the signal strength modifier $\mu$ with confidence level $\alpha$. The usual interpretation is that the background-only hypothesis (observed data can be explained by upward fluctuation of the background) can be excluded with a confidence level of $\alpha$. The Neyman-Pearson lemma [127] states that the most powerful test-statistic for a hypothesis test at the confidence level $\alpha$ is the likelihood ratio, which is defined as

$$\tilde{q}_\mu = -2 \ln \frac{L(n|\mu, \hat{\theta}_\mu)}{L(n|\hat{\mu}, \hat{\theta})},$$

with the constraint of $0 \leq \hat{\mu} \leq \mu$. In this case $\hat{\theta}_\mu$ refers to the conditional maximum likelihood estimate for $\theta$, given a signal strength modifier $\mu$. The denominator refers to the value at the global maximum of the likelihood function. The observed events $n$ can either be real data or from toy experiments.

**Observed limit**

The upper limit is calculated by scanning a range of positive values for $\mu$. For each, a set of calculations is executed. First the observed test statistic $\tilde{q}_\mu^{obs}$ for the given signal strength modifier is calculated by maximizing the likelihoods in equation (7.7). This yields the nuisance parameters $\hat{\theta}_\mu^{obs}$ best describing the observed data. Additionally the nuisance parameters $\hat{\theta}_0^{obs}$ need to be calculated once for the background-only hypothesis. Next the pdfs $f(\tilde{q}_\mu|\mu, \hat{\theta}_\mu^{obs})$ and $f(\tilde{q}_\mu|0, \hat{\theta}_0^{obs})$ are constructed by generating toy Monte Carlo pseudo-data using the current value for $\mu$ and the previously calculated nuisance parameters $\hat{\theta}_\mu^{obs}$ and $\hat{\theta}_0^{obs}$. With these pdfs the p-values for the signal+background hypothesis

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

and the background-only hypothesis

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

are defined. These p-values describe the probability to obtain values of the test statistic $\tilde{q}_\mu$ greater the observed value and are usually denoted as

$$\text{CL}_{s+b} = P \left( \tilde{q}_\mu \geq \tilde{q}_\mu^{obs} | \mu \cdot s + b \right) \quad \text{and} \quad \text{CL}_b = P \left( \tilde{q}_\mu \geq \tilde{q}_\mu^{obs} | b \right).$$

In the so-called modified frequentist approached [128][129] the quantity

$$\text{CL}_s = \frac{\text{CL}_{s+b}}{\text{CL}_b}$$

is used to calculate a confidence level for the exclusion of signal strength modifiers greater than $\mu$. This means, if $\text{CL}_s = 1 - \alpha$ for $\mu^*$, values of $\mu$ greater than $\mu^*$ can be excluded at a confidence level of $\alpha$. To quote the upper limit on the signal strength modifier $\mu$ is adjusted until $\text{CL}_s = 0.05$. 
7. Multivariate Analysis for the search for $tt\bar{H}$ in the $H \rightarrow b\bar{b}$ final state

Expected limit

The sensitivity of a search for unobserved particles can be tested without using the measured data by calculating the median expected upper limit. For this, large sets of toy Monte Carlo data following the background-only hypothesis are generated and the 95\% confidence level upper limit is calculated for each dataset. The median of the distribution obtained from this procedure is quoted as expected upper limit. The $\pm 1\sigma$ and $\pm 2\sigma$ uncertainties are obtained by finding the 16\% and 84\% respectively 2.5\% and 97.5\% quantiles of this distribution. Because a procedure like this is computationally intensive an estimation for the median upper limit is calculated with an asymptotic formula [130]. This method uses the fact that the distributions $f(\tilde{q}_\mu|\mu, \tilde{\theta}_\mu^{\text{obs}})$ and $f(\tilde{q}_\mu|0, \tilde{\theta}_0^{\text{obs}})$, in the limit of large datasets, can be analytically described using a single representative dataset, referred to as the Asimov dataset. This dataset is defined in a way that for a given signal strength modifier $\mu$, the nuisance parameters and background expectations are at their nominal value. Using the analytical description, the $\mu$ value with $\text{CL}_S = 0.05$ can be found by solving the formula and therefore without generating pseudo data. The limit obtained in this way is called the asymptotic expected upper limit.

7.2. Analysis strategy for $tt\bar{H}$, $H \rightarrow b\bar{b}$

The analysis presented in this section, is based on the latest $tt\bar{H}$, $H \rightarrow b\bar{b}$ analysis of the CMS collaboration summarized and published in [5].

7.2.1. Event categorization

After the event selection was applied, as described in section 6.2.1, the events are further categorized by their jet and b-tag multiplicity. In this analysis, the categories

- 4 jets, 3 b tags and 4 jets, $\geq 4$ b tags,
- 5 jets, 3 b tags and 5 jets, $\geq 4$ b tags as well as
- $\geq 6$ jets, 2 b tags, $\geq 6$ jets, 3 b tags and $\geq 6$ jets, $\geq 4$ b tags,

are used. Figure 7.1 shows the contributions of simulated $tt\bar{H}$, $\bar{t}t$ (split as described in section 5.2.1) and other background events to each analysis category. The categories with 4 b tags have, as expected, the highest signal to background ratio (S/B) because of the close resemblance to the expected $tt\bar{H}$, $H \rightarrow b\bar{b}$ topology. The remaining categories have a very large background contributions and are mainly used for constraining the background.

7.2.2. Input variables

The input variables for the $tt\bar{H}$ classifier used in this thesis are taken from [5]. In this section, only the variables relevant to the studies discussed in chapters 8 and 9 are described:

$p_T$ of the n-th jet

The transverse momentum of the n-th jet in the event is expected to be larger for jets originating from particles with high masses. A commonly used term is calling the jet with the highest $p_T$ in an event “hardest jets”.

$p_T$ sums

The scalar sum of all jets in an event is called $H_T$ and should be higher on average in a $tt\bar{H}$ event, because of the the expected decay products for the Higgs boson. The same is expected for the scalar sum of the transverse momenta of jets, lepton and $E_T$.
7.2. Analysis strategy for $t\bar{t}H$, $H \rightarrow bb$

![Figure 7.1.](image)

Figure 7.1.: Fraction of processes contributing to the different analysis categories in the search for $t\bar{t}H$, $H \rightarrow bb$ in the semileptonic channel. Adapted from [5].

\[ \sum p_{T,jet} / \sum E_{jet} \]

For this variable, the scalar sum of all jets is divided by the energy of all jets.

**Aplanarity and sphericity**

The eigenvalues $\lambda_i$ of the sphericity tensor

\[ S_{ij} = \frac{\sum_k p_{k,i} p_{k,j}}{\sum_k (p_k)^2} \]  \hspace{1cm} (7.13)

describing the direction of momentum flow in an event are used to calculate the aplanarity and sphericity. The aplanarity, defined as $2/3 \cdot \lambda_3$, describes the “flatness” of an event. Comparing $t\bar{t}$ and $t\bar{t}H$, the third particle in the latter leads to momentum flow in its direction and therefore increased aplanarity. The sphericity $3/2 \cdot (\lambda_2 + \lambda_3)$ describes the isotropy of the event and should increase in $t\bar{t}H$ in comparison to $t\bar{t}$, due to the same reasons as the aplanarity.

**Dijet mass closest to 125 GeV**

The mass of the pair of b-tagged jets with a mass closest to 125 GeV peaks at this value for signal and background processes, but should be narrower for $t\bar{t}H$, $H \rightarrow bb$ events.

**Closest tagged dijet mass**

As described in section 2.2.3.2 the b-tagged jets from the Higgs-boson decay tend to be closer together and therefore this variable separates signal and background very well. This is reflected by the mass of the two jets closest to each other, measured in $\Delta R$.

**Best Higgs mass**

The mass of the reconstructed Higgs boson should exhibit a strong peak at the Higgs-boson mass for correctly reconstructed events.

---

1. The method studied and described in [131] is used for the reconstruction of the $t\bar{t}H$ system. The pair of top quarks is reconstructed by calculating the mass most compatible with W boson and top quarks by considering every permutation of jets, lepton and $E_T$. Only b-tagged jets are considered for the b-quark candidates from the top quark. In the $\geq 6$ jets, $\geq 4$ b tags category the remaining two b jets are used for the Higgs boson mass. In the $\geq 6$ jets, 3 b tags category the jet with the fourth highest CSV value is used and in the 5 jets, $\geq 4$ b tags category, the minimum $p_T$ of the sixth jet is lowered to 10 GeV.
7. Multivariate Analysis for the search for $t\bar{t}H$ in the $H \rightarrow b\bar{b}$ final state

For each analysis category a dedicated BDT is constructed, that combines the information of a subset of all input variables into a final discriminant. This is realized via an optimization procedure \cite{132} based on the particle swarm optimization \cite{133}. The output of such optimized BDTs (also referred to as final discriminant) in the analysis categories with six jets. The signal contributions increase when comparing the categories with increasing number of b tags. Furthermore the shapes of the BDT outputs of signal and background are becoming more distinctive.

The shapes of the final discriminants of each analysis category are used to compute the limit on the signal strength modifier $\mu$.

7.3. Uncertainties considered in this thesis

The systematic uncertainties considered in this thesis are listed in table 7.1. The list was adopted from \cite{5}. These uncertainties either affect the yield of signal and background processes (denoted as rate uncertainty), the distribution shape \cite{2}, or both.

The uncertainty on the luminosity estimate is determined to be 2.7% \cite{134}. The effects on the expected number of pile-up interactions by the pile-up uncertainty is estimated by varying the expected cross section in simulation for such interactions, cf. section 5.3.3, by $\pm 4.6\%$. The uncertainty of the jet energy scale \cite{116} is estimated by scaling the jet energy scale correction by one standard deviation upwards (downwards) and propagating these scaled jets through the analysis. This leads, among others, to an harder (softer) jet-$p_T$ spectrum. The uncertainties on the CSV b-tagging scale factors, introduced in section 5.3.2, are evaluated for the different jet flavors: Heavy (HF), light (LF) and charm flavor. This is done by applying alternative scale factors obtained from varying the contamination in the background processes in the control sample, the jet energy scale uncertainty and the statistical uncertainties in the scale factor evaluation, by one standard deviation. The latter is parametrized as the sum of a linear and a quadratic term.

The uncertainties denoted as “QCD scale” describe the theoretical uncertainty of the corresponding cross sections, arising primarily from the parton distribution function.

\footnote{These affect the shape of the observables used as input for the classifier as well as the discriminant.}
### Table 7.1: Systematic uncertainties considered for this analysis, adopted from [5]. Type denotes the influence on the shape of the uncertainties, as defined in section 7.1.1.

<table>
<thead>
<tr>
<th>Source</th>
<th>Type</th>
<th>Remark</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luminosity</td>
<td>rate</td>
<td>Signal and all backgrounds</td>
</tr>
<tr>
<td>pile-up</td>
<td>shape</td>
<td>Signal and all backgrounds</td>
</tr>
<tr>
<td>Jet energy scale</td>
<td>shape</td>
<td>Signal and all backgrounds</td>
</tr>
<tr>
<td>b-tag HF fraction</td>
<td>shape</td>
<td>Signal and all backgrounds</td>
</tr>
<tr>
<td>b-tag HF (linear)</td>
<td>shape</td>
<td>Signal and all backgrounds</td>
</tr>
<tr>
<td>b-tag HF (quadratic)</td>
<td>shape</td>
<td>Signal and all backgrounds</td>
</tr>
<tr>
<td>b-tag LF fraction</td>
<td>shape</td>
<td>Signal and all backgrounds</td>
</tr>
<tr>
<td>b-tag LF (linear)</td>
<td>shape</td>
<td>Signal and all backgrounds</td>
</tr>
<tr>
<td>b-tag LF (quadratic)</td>
<td>shape</td>
<td>Signal and all backgrounds</td>
</tr>
<tr>
<td>b-tag charm (linear)</td>
<td>shape</td>
<td>Signal and all backgrounds</td>
</tr>
<tr>
<td>b-tag charm (quadratic)</td>
<td>shape</td>
<td>Signal and all backgrounds</td>
</tr>
<tr>
<td>QCD scale (t(\bar{t})H)</td>
<td>rate</td>
<td>Scale uncertainty of NLO t(\bar{t})H prediction</td>
</tr>
<tr>
<td>QCD scale (tt)</td>
<td>rate</td>
<td>Scale uncertainty of NLO tt prediction</td>
</tr>
<tr>
<td>QCD scale (t(\bar{t})+hf)</td>
<td>rate</td>
<td>Additional scale uncertainty of NLO t(\bar{t})+hf prediction</td>
</tr>
<tr>
<td>QCD scale (t)</td>
<td>rate</td>
<td>Scale uncertainty of NLO single top prediction</td>
</tr>
<tr>
<td>QCD scale (V)</td>
<td>rate</td>
<td>Scale uncertainty of NNLO W and Z prediction</td>
</tr>
<tr>
<td>QCD scale (VV)</td>
<td>rate</td>
<td>Scale uncertainty of NLO diboson prediction</td>
</tr>
</tbody>
</table>
8. b-jet regression for the \( \ttbar \H \) search in the \( \H \to \bb \) decay channel

The aim of the b-jet regression is to derive a correction factor for the transverse momentum of b jet with the tree-based regression technique implemented in \texttt{TMVA} \cite{45}. Using particle information from the Monte Carlo generator, the aim of the regression is to improve the b-jet \( p_T \) response (see section 6.1.2.1), which plays an important role in many analysis investigating \( \H \to \bb \) decays. A similar procedure was deployed successfully in the \( \H \to \bb \) search at the CDF detector \cite{135}.

This chapter will introduce a b-jet regression for the application in a \( \ttbar \H \) search in the \( \H \to \bb \) final state performed by CMS (the latest results were published in \cite{5}). Section 8.1 will focus on the presentation of the problem and outline considerations, like the choice of a regression target and inputs as well as other challenges that have to be taken into account. Following this, in section 8.2 the performance of the regression will be evaluated on simulated data.

If not indicated otherwise all jets in this section are clustered from PF candidates, using a the anti-\( k_T \) algorithm with a distance parameter \( R = 0.4 \), with all corrections applied (see section 6.1.2) and b tags are evaluated for the medium working point described in section 4.3.6. Jets that pass this working point are denoted as b-jets. In this section, Monte Carlo truth particles from the hard process respectively parton shower are denoted as partons. Events are selected with the semileptonic selection. The matching of jets clustered from PF candidates to Monte Carlo truth particles is realized by finding the closest parton, measured in \( \Delta R \), cf. equation (4.7), to the jet also fulfilling the criteria of a distance less than 0.4. If not otherwise indicated this procedure is used in this section, if matching is mentioned.

8.1. Jet energy correction for b jets

As already discussed in section 4.3.6 b jets feature special characteristics that separate them from jets that originate from lighter quarks or gluons. The jet corrections used at CMS, cf. section 5.3 are derived using simulated QCD-multijet data \cite{116} and therefore mostly jets from light-flavor quarks or gluon are used. Because of this, the jet-\( p_T \) response is expected to be worse for b jets than light-flavor jets. This is shown in figure 8.1 by the jet-\( p_T \) response with respect to a parton that is closer than 0.4 in \( \Delta R \), cf. equation (4.7), defined in equation (6.3). As described in section 6.1.2.1 the jet-\( p_T \) response \( \mathcal{R} \) should be
Figure 8.1.: Jet-$p_T$ response with respect to a parton (see equation (6.3)) that is closer than 0.4 in $\Delta R$ (see equation (4.7)) to the jet. Distributions are shown for light- and c-jets (blue) and b-jets (green) in a simulated $t\bar{t}$ dataset.

Figure 8.2.: Feynman diagram for the weak decay of a $B^-$ to a $D^0$ by radiating a $W^-$ boson. The leptonic decay of the $W$ boson with a neutrino leads to an incomplete energy measurement of b jets.
Gaussian distributed with a mean value $\mu_R$ of one, after all corrections are applied. The deterioration of the jet-$p_T$ resolution depends on two independent effects:

For track-based measurements the relative $p_T$ resolution, cf. equation (4.9), for particles with $p_T$ up to 100 GeV is 2% and degrades approximately proportional to $p_T$ for higher transverse momenta [51].

For calorimeter measurements, the relative energy resolution, defined in equation (4.10), improves with increasing $p_T$ [32]. It is dominated by miscalibrations and non-uniformities of the calorimeter for high $p_T$, stochastic fluctuations of the shower development in the medium $p_T$ regime and by electronic noise of the readout systems at low $p_T$. On the one hand, the relative $p_T$ resolution of the ECAL is very good, e.g. 0.6% at 50 GeV, cf. equation (4.12), for electrons and photons [136]. The resolution of the HCAL, on the other hand, is worse with approximately 18% for pions with 50 GeV, cf. equation (4.13). This behavior has its origin in effects of the hadronic shower. The response of the HCAL is different for electrons than for hadrons and the electromagnetic fraction $f_{em}$, introduced in equation (4.11), fluctuates from shower to shower. Furthermore, the amount of invisible energy in the shower fluctuates. The influence of such effects is also jet-flavor dependent. Particles, that live long and lose their energy in the calorimeters, lose on average less energy, due to mismeasurements, than particles that have a short lifetime and a high probability to decay leptonically with neutrinos. One example of the latter are B mesons that are formed from b quarks during the shower. They decay weakly with neutrinos involved in approximately 30% of all cases [32]. A example of this is shown in figure 8.2 for the decay of a $B^-$ into a $D^0$, with the W bosons decaying leptonically. Because energy of neutrinos cannot be measured, and varies statistically from decay to decay, the jet-$p_T$ resolution is affected by the invisible energy. Usually jets contain multiple hadrons with overlapping showers, which leads to further deterioration of the jet-$p_T$ response. The PF algorithm, introduced in section 4.3.3 also has an influence on the jet-$p_T$ resolution [77], which can be reduced by the tracking efficiency in the regime of low transverse momenta. A similar effect can observed for high transverse momenta because of the uncertainties of the single-hadron response corrections. Studies on different PF algorithms [137] for linear colliders have shown that wrong matching between calorimetric energy depositions to reconstructed constituents can impact also the jet $p_T$ response.

The usual jet corrections should remove most of these effects. A good approximation the case for light- and c-jets (blue distribution in figure 8.1), but not for b jets (green distribution in figure 8.1) and therefore further corrections, specially dedicated to b-jets are required. One possibility to derive such corrections is using multivariate regression techniques.

8.1.1. Figure of merit

To evaluate the performance of the regression certain quantities and distributions, called figure of merit (FoM), are to be compared and interpreted. For the specific application of the regression presented in this thesis, the following FoMs are considered.

**Quadratic error of the regression**

The prediction error can be calculated by comparing regression target and output for a large enough number of events. The quadratic error

$$\text{err}_{\text{quad}} = \frac{1}{N} \sum_{i} (f(X_i) - Y)^2$$

1 Otherwise the resolution would not be affected. It would only cause an changed scale.

2 In principle jets containing D mesons suffer from similar effects, but because of the D meson’s slightly longer lifetime and smaller branching ratio for decays with neutrinos, the error is smaller.
with the number of events $N$, the target $Y$ and the prediction $f$, which is a function of its input parameters $X_i$, can be used to quantify the quality of the regression and will be used in the following studies.

**Higgs-boson mass**

The Higgs-boson mass distribution plays an important role for the final classifier in categories with four b-tags, but effects that cause the worse b-jet $p_T$ response are also leading to a mean mass diverging from the mass used in the MC generator. Because these categories are signal enriched the analysis could particularly benefit from an improved b jet and consequential Higgs-boson mass resolution. Therefore applying corrections from the regression to b jets should also improve the relative resolution $\sigma/\mu$, where $\sigma$ denotes the width und $\mu$ the mean of the distribution.

Hence, the relative resolution of the Higgs-boson mass can be used to evaluate and tune the performance of the regression. The MC Higgs-boson mass in simulation is determined in events with exactly two b quarks from the Higgs boson on particle level that could be matched to jets, which are then used to calculate the mass.

**Regression bias**

Small biases in a regression regarding a input variable are acceptable, because in principle it would be possible that for example jets with low $p_T$ are corrected differently than jets with high $p_T$. Nonetheless the bias should not be too strong. The regression will not be trained on the signal sample and therefore strong biases could lead to a regression not applicable when changing from one process to another one. Additionally the performance in extreme regions of the phase space could be worse than expected. One example of such a behavior would be the boosted region with very high transverse momenta, where a biased regression could correct the jet because of its bias and not the properties of the jet. Therefore, the $p_T$-dependent bias is part of the consideration, when evaluation the performance of the regression, with the aim of finding a regression with minimal bias.

**8.1.2. Choice of regression target**

As described in section 3.1 the choice of a target plays an important role for the overall performance of the regression. Two possibilities are explored in order to derive b-jet corrections, both relying on particle information from the MC generator, because they are not affected by most of the detector effects identified as the main reasons for the deterioration of the jet-$p_T$ response.

The parton $p_T$ is an obvious choice: It should reflect the properties of the b quark best and is not affected by the parton showering. On the other hand, processes like the weak decay of B mesons are described by the parton shower. Therefore, the $p_T$ of the generator-level jet, as defined in section 5.1, is a possible target, because it contains this informations without including a detector simulation, which would lead to further noise. Both contain the information about the jets quark flavor from the MC generator and therefore can be used the identify b jets. In the following studies partons and generator-level jet are geometrically matched to jets to relate particle-level information with variables that can be reconstructed from measurements in the detector. Both possible targets are highly correlated (see figure 8.3a) and therefore the differences should be minor.

Because multivariate tree-based techniques can only respond in certain discrete values a

---

3 In principle also the MC mass of the hadronic top could be used as FoM, but because improving the signal background separation is an important goal of the regression, it is not used for this regression.

4 Regression algorithms with multiple targets are not included in this consideration.

5 The same procedure defined for matching partons in the introduction is used for both: generator-level jets and partons.
8.1. Jet energy correction for b jets

Figure 8.3.: Distribution of the generator-level jet and parton \( p_T \) (a) and parton respectively generator-level jet \( p_T \) response, cf. (6.3) respectively (6.4), (b), for the same jet. The particle level objects are matched to the jets if they are in \( \Delta R \) closer than 0.4.

Table 8.1.: Fits for the MC Higgs-boson mass, cf. figure 8.4a, under the approximation of Gaussian distributed mass peaks. The regressions only differ in the choice of regression target, with all other settings and input variables identical. Additionally the regression error, cf. equation (8.1) and figure 8.4b with respect to the target-jet \( p_T \) response is given.

<table>
<thead>
<tr>
<th>Target</th>
<th>Mean ( \mu ) [GeV]</th>
<th>Width ( \sigma ) [GeV]</th>
<th>rel. resolution ( \frac{\sigma}{\mu} ) ( /10^{-2} )</th>
<th>( \text{err}_{\text{quad}} ) ( /10^{-1} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>parton ( p_T )</td>
<td>124.91</td>
<td>14.73</td>
<td>11.79</td>
<td>0.67</td>
</tr>
<tr>
<td>parton-( p_T ) ratio</td>
<td>124.64</td>
<td>14.54</td>
<td>11.67</td>
<td>0.67</td>
</tr>
<tr>
<td>generator-level jet ( p_T )</td>
<td>122.05</td>
<td>14.34</td>
<td>11.75</td>
<td>0.54</td>
</tr>
<tr>
<td>generator-level jet-( p_T ) ratio</td>
<td>121.94</td>
<td>14.48</td>
<td>11.87</td>
<td>0.53</td>
</tr>
</tbody>
</table>

compact response-space should improve the prediction of a tree. The target \( p_T \) range is usually between 25 GeV and 600 GeV and therefore very broad. By using the ratio between target \( p_T \) and jet \( p_T \) as new target the response space of the regression can be compressed. The correlation between these ratios for parton and generator-level jet targets is shown in 8.3b.

For the b-jet regression presented in this thesis, four possible targets were explored. The \( p_T \) of partons and generator-level jets as well as both divided by the jet \( p_T \), which will henceforth be denoted as parton-\( p_T \) and generator-level jet-\( p_T \) response, cf. equations (6.3) and (6.4), respectively. For this comparison, the FoMs described in 8.1.1 were compared on a \( t\bar{t}H, H \rightarrow b\bar{b} \) sample and the jets were required to have a CSV value > 0.8. Each regression was trained on equal train and test samples of a simulated \( t\bar{t} \) sample\(^6\). The quoted uncertainty, is computed for the parton/generator-level jet \( p_T \) response\(^7\).

The MC Higgs-boson mass presented in figure 8.4a shows minimal dependency on the choice of target. Even though the peaks differ, the relative resolution only shows minimal deviations (see table 8.1).

---

\(^6\) All Monte Carlo samples are described in section 5.2.  
\(^7\) Analytically this is accomplished by substituting the prediction \( f(x_i) \) in equation (8.1) with \( f(x_i)/p_T^{\text{jet}} \) when computing the regression error for trainings with parton or generator-level jet \( p_T \) as target.
8. b-jet regression for the $t\bar{t}H$ search in the $H \to b\bar{b}$ decay channel

Figure 8.4.: Different figures of merit (FoM) for choosing a target for the b-jet regression. Each regression was trained with identical events, events-selection, settings and input variables on a simulated $t\bar{t}H$ sample. Each regression was evaluated on a $t\bar{t}H$ sample for all b-tagged jets. For the MC Higgs-boson mass (a) only events where used that contain two jets with matching b quarks ($\Delta R > 0.4$) from the Higgs boson. The distributions where fitted under the assumption of a Gaussian distributed peak and the results are listed in table 8.1. The prediction error (b) is given with respect to the parton/generator-level jet-$p_T$ response, with the minimal error highlighted by a dashed, horizontal line. The $p_T$-dependent bias of the prediction for trainings with parton $p_T$ response (c) and generator-level jet $p_T$ response (d), cf. (6.3) and (6.4) respectively, as target are shown.
8.1. Jet energy correction for $b$ jets

In figure 8.4b the regression error with respect to the parton/generator-level jet response is displayed and the errors are listed in table 8.1. While the difference in errors between parton-based (blue tones) and generator-level jet-based regressions (red tones) is large in relation to each other, the absolute difference is minor compared to the mean correction value, which is approx. one. Therefore the relative difference in error between parton-based and generator-level jet-based regressions is approximately 1% of the total correction and therefore negligible compared to the other FoMs.

The $p_T$ dependent bias of the regression prediction is shown for parton-$p_T$ response (figure 8.4c) and generator-level jet-$p_T$ response (figure 8.4d). As discussed before, strongly biased regressions are undesirable, because of their bad generalizability. Comparing the two stacked distributions, the overall shapes are similar to each other. The bias is clearly visible, when separately plotting the regression output for different $p_T$ regions. For both targets, jets with low $p_T$ values, up to 100 GeV, are mostly corrected upwards from the nominal value. For jets with transverse momenta greater than 200 GeV, the opposite is the case. The strongest bias can be observed for jet with transverse momenta of less than 50 GeV or more than 300 GeV. Comparing these regions for parton-based and generator-level jet-based trainings leads to the conclusion that the $p_T$-dependent prediction is less biased for the latter. In addition to this, the same distributions are shown unstacked in figure A.1 in appendix A. They show that the prediction shapes in the different $p_T$ regions are similarly and less shifted with respect to each other, in the case of the generator-level jet-based regression.

In conclusion, the generator-level jet-$p_T$ response is considered the most promising option for further studies and a final implementation of the regression in the analysis. The $p_T$ response was chosen over the transverse momentum because of its advantages related to the response space of the regression. While the Higgs-boson mass and the regression errors were very similar for parton-based and generator-level jet-based trainings leads to the conclusion that the $p_T$-dependent prediction is less biased for the latter. In addition to this, the same distributions are shown unstacked in figure A.1 in appendix A. They show that the prediction shapes in the different $p_T$ regions are similarly and less shifted with respect to each other, in the case of the generator-level jet-based regression.

8.1.3. Choice of regression input

Choosing a good set of input variables is an important part of setting up a regression, or multivariate algorithms in general. Since the algorithm uses the relation between input variable and target to gain information for its prediction, variables containing beneficial information for the regression need to be found. Additionally consistent behavior between data and Monte Carlo samples, e.g. training sample and the sample the regression will be applied to, needs to be verified, which will be discussed in section 9.2. For the regression presented in this thesis, a physics-based approach was chosen, with input variables that contain information correlated to specific physical phenomena in regards to $b$ jets. All variables used in the regression can be assigned to one of the following categories:

**Jet kinematics**

These variables contribute information about the kinematic properties of $b$ jets to the regression. The jet and leading track $p_T$ are used because the aim of the regression is to decrease the difference between the measured and the actual bottom-quark $p_T$, while the $\eta$ can show possible directional dependencies of the response. The transverse mass, cf (4.4), contains directional information about the jet constituent and therefore the hadronization. This is useful, because $b$ jets should contain
fewer collinear constituents while the opposite is the case for light-quark jets. As described in section 5.3 all jets are corrected to compensate effects like pile up as well as $p_T$ and $\eta$ dependencies, cf. section 6.1.2. For easy incorporation of the regression into the analysis work flow, the jets used for training and evaluation have jet energy scale (JES) corrections$^8$ applied. Therefore the inverse of this correction factor is used as an input for the regression to get an estimate of the actual measured energy.

Secondary vertex

Variables describing the reconstructed secondary vertex are used because of their direct connection to properties of bottom quarks, as described in section 4.3.6. For this regression the $p_T$, mass and number of tracks assigned to the secondary vertex are used to measure the energy of all charged particle in a jet. Because the distance between primary interaction point and the vertex depends on the bottom-quark lifetime and therefore the energy, the 3D flight distance and its significance are also used as an input for the regression.

Energy loss from $\nu$

To retrieve information about missing energy in the form of neutrinos, leptons are matched to the jet if they are closer than 0.4 in $\Delta R$. These leptons are required to pass the soft selections described in section 6.1.1 and 6.1.1.2 with the requirements of $p_T > 5$ GeV and $|\eta| < 2.5$ for electrons as well as $p_T > 3$ GeV and $|\eta| < 2.4$ for muons. The isolated lepton used as the decay candidate from the leptonically decaying W boson, in $t\bar{t}$ events, is excluded from this matching. The lepton $p_T$ as well as the relative $p_T$ and $\Delta R$ to the jet are used in the regression.

Particle composition

Information about the particle composition of the jet and the energy measurement in the calorimeters is added to the regression by using the neutral electromagnetic as well as the total hadronic energy fraction of the jet, defined as the fraction of energy measured in the ECAL and HCAL.

Event kinematics

Because random energy deposits in the calorimeter from secondary proton collisions can still be present in the jet after the L1 pile-up correction, the number of primary vertices is used to include information about the kinematic properties of the whole event.

All used input variables are listed in table 8.2. Initial testing showed that these variables yield a good prediction of the target distribution. Excluding weakly correlated variables leads to a deterioration of the prediction, while no improvement was observable by adding further variables. In addition to this, good modeling in data played an important role for the input variables configuration of the regression, see section 9.2 for further discussions on this topic. When using TMVA, the variables are ranked by their absolute correlation to the target, which corresponds to the frequency the variable is used for node splitting in the tree. The correlation values in table 8.2 are given for using the generator-level jet-$p_T$ response as target. The strong (and similar) correlation of the jet $p_T$ and $M_T$ is reasonable considering

\[ f_{\text{JES}} = f_{\text{L1 pile-up}} \cdot f_{\text{L2L3 MC-truth}}. \]

If the regression is used on data, the L2L3 Residual correction is removed, to get consistent input variables.

---

$^8$ The JES correction $f_{\text{JES}}$ can be expressed in terms of the corrections introduced in section 6.1.2 with $f_{\text{JES}} = f_{\text{L1 pile-up}} \cdot f_{\text{L2L3 MC-truth}}$. If the regression is used on data, the L2L3 Residual correction is removed, to get consistent input variables.
that both are strongly correlated among themselves and the construction of the target. Strongly correlated are also the total hadronic energy fraction, the distance of jet and lepton as well as the mass of the vertex. This shows that the regression uses variables related to the mismeasured b jets, to find the corresponding correction.

8.1.4. Training Dataset

When using machine learning techniques a dataset is needed for the training of the algorithm. Because the purpose of the regression presented in the thesis is to be used in the search for $t\bar{t}H$ with the $H \rightarrow b\bar{b}$ final state, it would be best to train the regression on a simulated dataset of this process. While the description of the phase space would be as close as possible to the expected signal, the limited number of currently available simulated $t\bar{t}H$ events and the fact that part of these events are used for the training of the final classifier in the search for $t\bar{t}H$, using simulated events of this process is no option. Because most of the backgrounds are various $\bar{t}t$ processes (see 2.2.3.2) and the similarity between these and the expected final state, using the $\bar{t}t$ dataset is the best choice for this regression. Another possibility would be to use simulated $Z \rightarrow b\bar{b}$ events because of the expected similarities between this process and $H \rightarrow b\bar{b}$, which is not considered in this thesis. Additionally it is expected that the dependence of the training on the exact trainings dataset is rather small, because the input variables, except for the number of primary vertices, are jet quantities and should therefore be, in approximation, not dependent on the exact event topology.
8.1.5. Boosted regression tree settings

The BDT/BRT performance can also depend to a large extent on the configuration. While the classification of events is a problem that can be accomplished with high efficiency, the b jet regression is a problem that can not be solved nearly that well. As it will be shown in section 8.2, the performance of the b jet regression is measurable, but the agreement between target and output is not good compared to the usual examples described in text books like [43]. This means that the b-jet regression is not limited by the BDT/BRT settings, but by the information contained in the input variables and the unknown and unmeasurable factors occurring in the detector.

Nevertheless small optimizations, mainly for the stability of the regression and its prediction, can be achieved by investigating settings like the number of trees or cuts. The basic BDT/BRT configuration used for this regression can be found in [45]. For the b-jet regression only the number of trees \( n_{\text{Trees}} \), the number of cuts \( n_{\text{Cuts}} \), the shrinkage and maximum tree depth \( \text{maxDepth} \) were changed from this standard settings and gradient boosting was activated. The main goal of modifying \( n_{\text{Trees}} \), \( n_{\text{Cuts}} \) and \( \text{maxDepth} \) was to increase prediction stability. The shrinkage is mainly used to decrease overtraining effects. Figure 8.5 shows the squared prediction error, as defined in equation (8.1), for different BRT settings. The displayed settings are an excerpt of a study in which the four parameters where varied systematically. Values lower than 1200 for \( n_{\text{Trees}} \) were discarded, because of their inferior stability. The standard TMVA settings are displayed in brown for reference. The difference between the regression error on the training and test samples can be interpreted as a measure of overtraining and stability. Based on this, the regression with 1200 Trees, each with a depth of three and 30 cuts, and a shrinkage of 0.1 (green points) performs best, while the regression trained with the standard TMVA settings preforms worst. Comparing this best setting to the other settings shows that a tradeoff between prediction power, or small error, and overtraining has to be made. One example is the training with an increased depth of five (teal points), which has a decreased error but a increased gap between test and training sample compared to the training with a \( \text{maxDepth} \) of three (green points). Similar considerations can be made for the other configurations shown in figure 8.5. Following this study, the settings \( n_{\text{Trees}} = 1200, n_{\text{Cuts}} = 30, \text{maxDepth} = 3 \) and shrinkage = 0.1 are chosen modified from the default TMVA settings for the b-jet regression presented in this thesis.

8.1.6. Influence of randomness

The energy measurement of jets depends on many different effects that cannot be easily reduced or predicted. Examples include the calorimeter resolution or the noise in the calorimeter from pile-up and, on a smaller scale, the underlying event. From the standpoint of a regression these effects are purely random and to study such effects a toy experiment was conducted.

For this regression model, a Gaussian distribution was used as base function. By adding two random uniformly distributed numbers, the distribution was smeared and shifted as is would be the case, when simulating resolution or noise effects. The morphed distribution \( f_m \) depends on the targeted Gaussian distribution \( g(\mu, \sigma) \) as given in equation (8.2). The distribution \( U_s \) is defined in a way to widen \( g \) without shifting the mean while \( U_o \) can be chosen arbitrarily. The factors \( \tilde{x} \) and \( \tilde{y} \) are used to control the strength of the morphing.

\[
f_m(\mu, \sigma, \tilde{x}, x_0, \tilde{y}, y_0, y_1) = [g(\mu, \sigma) + \tilde{x} \cdot U_s(-x_0, x_0)] + \tilde{y} \cdot U_o(y_0, y_1)
\]

9 This is best visible when using e.g. the parton-\( p_T \) target.
10 This is only the case on a per-jet basis. On average, this effects can be modeled rather well.
11 In reality this smearing would probably correspond to a Gaussian distribution, but the influence of this should have a negligible effect on the conclusion of this toy experiment.
To study the influence of random effects on the regression performance the morphing parameters of $f_m$ were chosen to be $\mu = \bar{y} = y_1 = 1$, $\sigma = x_0 = y_0 = 0.5$ and $\bar{x} = 1.5$. A dataset was generated containing the distribution $f_m$ as well as its subdistributions $g$, $U_s$ and $U_o$. To test how the prediction of the regression is influenced by the random changes to the target distribution $g$, two separate regressions are trained. For both the tree settings are fixed and the same training and test datasets are used. In an ideal case all information available ($f_m$, $U_s$ and $U_o$) are given to the regression. This is shown in figure 8.6a, where the regression output is given in relation to the target on a statistically independent test sample. It is clearly visible that the algorithm was able to learn and predict the morphing procedure. To test the influence of the randomness, a more realistic case was used for the second training. For this, only the distributions $f_m$ and $U_o$ where used as regression inputs. Figure 8.6b shows the outcome of this training. The output distribution shows a clear peak at the mean of the target distribution. It follows that the regression does not draw wrong conclusions from the unpredictable behavior but instead adjusts its prediction model to be more conservative. This behavior also translates to the correlation between target and regression output, which decreases by approximately one fourth, cf. figure A.2 in the appendix.

This toy study shows that the regression can be assumed to be robust against biases due to random effects but will correct conservatively if necessary.

### 8.1.7. Uncertainties on input variables

The goal of this section is to check how the regression output is influenced by uncertainties on the regression input variables.

---

$^{12}$ The settings different from the standard TMVA configuration are: $nTrees = 1200$, $nCuts = 20$, $maxDepth = 3$ and $\text{shrinkage} = 0.09$. 
Figure 8.6.: Simple regression model for a Gaussian distribution that was transformed using two random uniformly distributed numbers, cf. equation (8.2). The plots show the target (blue) and output (red) of the regression. One regression was trained with all information (a), while another regression was trained without the knowledge of the random distribution smearing the Gaussian distribution (b). The inferior prediction power is clearly visible, but the regression with less information is also not biased, because it peaks at the mean of the target distribution.

Figure 8.7.: Dependency of the regression output on pile-up (a) and jet energy corrections (b), both introduced in section 7.3. In the case of the pile-up uncertainty, the distributions are reweighed corresponding to the expected pile-up cross section $\sigma_{PU}$ (Up: $\sigma_{PU} \cdot 1.05$ and Down: $\sigma_{PU} \cdot 0.95$), while for the jet energy correction uncertainty, the evaluated jets are rescaled according to the uncertainty $\sigma_{JEC}$ of the correction factor $f_{JEC}$ (Up: $f_{JEC} + \sigma_{JEC}$ and Down: $f_{JEC} - \sigma_{JEC}$) before any selection was applied and then used as regression input.
8.2. Performance of the b-jet regression

The b-jet regression is directly dependent on the pile-up and jet energy correction (JEC) uncertainties, because these influence the input variables the most. The plots in figure 8.7 show this dependency. The CSV value, used to choose which jets are regressed, is no direct input variable and therefore the uncertainty on the b-tagging should be covered by the CSV uncertainty assumed for the analysis. All considered uncertainties are described in section 7.3. It is recognizable in figure 8.7a that the influence of the 4.6% pile-up uncertainty on the regression output is negligible. Because the jet energy correction uncertainty is added to the JEC in the “up” case and subtracted in the “down” case, the jet-$p_T$ spectrum is shifted on average either upwards or downwards and therefore the regression output is expected to be shifted in the opposite direction. The reason for this is that a downward shift of the jet-$p_T$ spectrum results in more low $p_T$ jets, which are in general corrected upward by the correction and vice versa. This expected influence of the jet energy correction uncertainty is visible in figure 8.7b.

8.2. Performance of the b-jet regression

Evaluating the general performance of the b-jet regression can be done on simulated $t\bar{t}$ and $t\bar{t}H$ datasets by directly comparing distributions with strong b-jet response dependency. For this, variables on three levels are examined. The first level is formed from variables using kinematic parton-level information like the jet response relative to the parton of its origin. At the second level are variables directly connected to particles at generator level, like the mass of the two jets originating form the Higgs boson. The third level are quantities reconstructed from jets not relying on particle or generator information, e.g. measurable quantities that are also available for data.

Improving the jet-$p_T$ response relative to the quark of its origin is an important motivation behind the b-jet regression. Figure 8.8 shows the influence of the regression on this response. The dashed line indicates the expected jet energy scale after calibration,
Figure 8.9.: Mass of the MC Higgs boson (a) and the MC hadronic top quark (b) with and without applied regression. Both masses are calculated with jets that could be matched to particle level bottom quarks, as described in the introduction of this chapter, from the Higgs boson respectively the hadronically decaying top quark. The Higgs boson was generated with a mass of 125 GeV and the top quark with 173 GeV, which are indicated with a dashed line. The distributions were fitted under the assumption of approximately Gaussian distributed peaks, by choosing a symmetrically distributed range around the mean maximum. The regression improves the relative resolution \( \sigma/\mu \) of both distributions. Furthermore, the mean of both distributions shift closer to the expected value.

Figure 8.10.: Impact of the regression on reconstructed quantities. The mass of the dijet pair with the mass closest to 125 GeV is displayed in (a) and the mass of the two closest tagged jets in (b). See section 7.2.2 for definition of these quantities. On a \( t\bar{t}H \) sample both should show peaks in the area surrounding the generator Higgs-boson mass of 125 GeV, indicated by the dashed line. For both distributions this behavior is increased, when using the regression.
8.2. Performance of the b-jet regression

Before the regression is applied the jet energy scale differs from one by approximately 5%. The regression decreases this to 2%, which is a clear sign that the regression is working as intended. Also the relative resolution $\sigma/\mu$ improves by 5%, which is calculated under the assumption of a approximately Gaussian distributed peak, cf. (6.5). The mean and width values used for this and all following calculation of the relative resolution are given in the legends of the corresponding figures.

For $t\bar{t}H$ with the $H \rightarrow b\bar{b}$ final state, two particle-level variables are particularly interesting: The mass of the Higgs boson and the mass of the hadronically decaying top quark, shown in figure 8.9a and 8.9b respectively. The Higgs-boson mass is calculated as explained in section 8.1.1. Using the standard jet corrections the mean of this Higgs-boson mass distribution differs about 7% from the mass value at generation time. By applying the regression, the mean is shifted in the direction of the mass used in the event generator, which reduces the underestimation to 3%, while also the width of the distribution decreases. This results in an improved relative resolution of approximately 5%. The mass of the hadronically decaying top quark is calculated from three jets, with two of them matched to quarks form the hadronic W boson and one matched to the b quark form the corresponding top quark decay. In contrast to the Higgs-boson mass, the measurement is much better even without regression applied, because of the better measurement of the jet originating from light and c jets assigned to the W boson. Nevertheless applying a regression improves the relative resolution by 3%. Following from this, the improvements from the regression are still present after the transition form parton-level to particle-level quantities.

The final step is to evaluate the influence of the regression on reconstructed variables. The influence of such quantities with applied regression on the analysis will be studied in more detail in section 9.1. For this performance evaluation, two dijet mass variables are used. These variables are found by calculating all combinations of two jets and keeping the one combination that fulfills certain criteria corresponding to the Higgs-boson mass. Figure 8.10a shows the combined mass of the two b-tagged jet closest to 125 GeV in the event. Because of its construction, the corresponding distribution should show a strong peak at 125 GeV for a $t\bar{t}H$, or signal, event and a broader peak for a $t\bar{t}$, or background, event. By applying the regression this behavior is reinforced. Because the bottom quarks from a Higgs-boson decay are usually produced in close proximity to each other, the mass of the closest tagged jets (shown in figure 8.10b) should exhibit a peak at the Higgs-boson mass on a $t\bar{t}$H dataset. For the closest tagged dijet mass, an improvement through the regression is much more obvious. The broad peak between 100 GeV and 130 GeV gets moved to a clear peak near the Higgs-boson mass indicated by the dashed line. This shows that the b-jet regression can also improve reconstructed quantities, that are also used to determine the final discriminant of $t\bar{t}H$. These distributions are shown on a $t\bar{t}$ dataset in figure A.3 in the appendix.

This study shows that the b-jet regression, derived from particle-level quantities, can improve the features used for discriminating signal and background and could therefore improve the performance of the final classifier for $t\bar{t}H$. 
9. Impact of the b-jet regression on the search for $\bar{t}tH$ in the $H \rightarrow b\bar{b}$ decay channel

In chapter 8 it was shown that a dedicated correction for b jets, determined with a tree-based regression, can improve b-jet energy measurements and response and thus improve separation power of input variables for the final classifier for the ttH search, like reconstructed Higgs-boson masses. Therefore, in this chapter, the influence of the b-jet regression on the analysis, which was described, including the final classifier and discriminating variables, in chapter 7 will be examined.

Section 9.1.1 will focus on evaluating the impact of the b-jet regression on the signal-background separation of the final BDT based on its input variables in Monte Carlo datasets. Following this, a study on the impact of the regression on the expected signal strength modifier is shown in section 9.1.2. Furthermore, different validation methods for the b-jet regression will be explored in section 9.2.

9.1. Final classifier and limit

9.1.1. Discriminating variables

If not indicated otherwise, the signal and background samples used for this study are simulated $ttH$, $H \rightarrow b\bar{b}$ and $tt$ samples (see section 5.2 for detailed information on the Monte Carlo samples). The full simulated datasets were split, with one half used for the training of the BDTs and the other half to study and test them. This assures that properties specific to the training events cannot bias the evaluation. A small subset of the same training and test sample was used to train the regression. Hence, a correlation between the two methods is possible. No event selection on the jet or b-tag multiplicity is required for the regression and therefore the influence of this correlation is expected to be small, because most of the events have two b tags but are not used in any analysis category. Two collections of jets are used in this section. The unregressed jets clustered from PF candidates and corrected with the usual CMS corrections introduced in section 6.1.2 and the regressed jets, which are corrected with the b-jet correction from the regression, introduced in chapter 8 after all other corrections are applied.

The variables used for the final BDT are chosen because of their features discriminating signal from background. The separation power can be measured by comparing the area
9. Impact of the b-jet regression on the search for \( \bar{t}tH \) in the \( H \rightarrow \bar{b}b \) decay channel

Figure 9.1.: Input variables for the BDT, whose signal-background separation is improved by the regression. The values given in the upper left corner correspond to the signal background separation power of the variable. In each plot, the signal (red) and background (blue) distributions are shown without (dashed) and with (solid) regression applied. For the category with \( \geq 6 \) Jets and \( \geq 4 \) b tags, the masses of the dijet pairs closest to 125 GeV (a) and in closest proximity to each other (b), the best Higgs-boson mass (c) and the sphericity (d) receive the biggest improvement from the regression.

Figure 9.2.: BDT outputs and ROC curves, in the category with \( \geq 6 \) Jets and \( \geq 4 \) b tags. The BDT output (a) for training with regressed (solid) and unregressed (dashed) jets is shown for signal \( t\bar{t}H \) (red) and background \( t\bar{t} \) (blue) sample. The ROC curves (b) are calculated for the BDTs trained/evaluated with regressed (red) and unregressed (blue) input variables.
Figure 9.3.: Input variables for the BDT, whose signal-background separation is improved by the regression. The values given in the upper left corner correspond to the signal background separation power of the variable. In each plot, the signal (red) and background (blue) distributions are shown without (dashed) and with (solid) regression applied. For the category with 5 Jets and \( \geq 4 \) b tags, the masses of the dijet pairs closest to 125 GeV (a) and in closest proximity to each other (b), the \( p_T \) of the fourth hardest jet (c) and \( \sum p_T/\sum E \) (d) receive the biggest improvement from the regression.

Figure 9.4.: BDT outputs and ROC curves, in the category with 5 Jets and \( \geq 4 \) b tags. The BDT output (a) for training with regressed (solid) and unregressed (dashed) jets is shown for signal tt\( H \) (red) and background tt (blue) sample. The ROC curves (b) are calculated for the classifiers trained/evaluated with regressed (red) and unregressed (blue) input variables.
under the ROC curve, cf. section 3.2, for this variables. In this section these values are
given in each plot in the upper right corner with and without regression applied. To
compare these two cases, all input variables (see section 7.2.2 for detailed definitions) were
recomputed with the regressed jets. The BDTs used for this chapter are optimized, as
outlined in section 7.2.3, and retrained with these optimized settings for the regressed
and unregressed variables on the same signal and background events to assure consistency.
In the following, variables that are calculated with jets corrected according to the b-jet
regression output are denoted as regressed variables, while the term unregressed will be
used for variables without this correction.

Input variables for the final classifier

The analysis category with events that have \( \geq 6 \) jets, \( \geq 4 \) b tags corresponds to the
expected \( t\bar{t}H \) topology, at LO with all jets in the acceptance, and is therefore of great
interest for studies on improving the analysis. Four input variables for which the regression
had the largest impact are shown in figure 9.1. As already discussed in section 8.2, the
dijet mass closest to 125 GeV and of the two jets closest to each other in \( \Delta R \) can be used
to differentiate between signal and background. For the \( \geq 6 \) jets, \( \geq 4 \) b tags category,
these variables are given in figure 9.1a and 9.1b for signal (red) and background (blue)
with regression applied (solid) and without (dashed). For the tagged dijet mass closest
to 125 GeV the signal distribution gets narrower, while the changes of the background
are very minor. The closest tagged dijet mass exhibits similar features with a stronger
peak at the Higgs-boson mass and no big change in the background. The reconstructed
Higgs-boson mass in figure 9.1c is expected to be very sensitive to improvements in the
b-jet response. However, because the features of both signal and background change in a
similar way, the improvements of the area under the ROC curve, and therefore separation,
are minor. The sphericity, as a measure of the \( p_T \) isotropy of the event, in figure 9.1d
also shows an improvement, because the additional objects that increase the isotropy in
ttH events are b jets.

Similar conclusions can be drawn from the other analysis categories. An excerpt of im-
proved input variables in the category with 5 jets, \( \geq 4 \) b tags is given in figure 9.3. Because
the events in this category are still signal-like, the tagged dijet mass closest to 125 GeV
(figure 9.3a) and the closest tagged dijet mass (figure 9.3b) are good variables for separat-
ing signal and background. The improvements from applying the b-jet regression are
therefore very similar to the \( \geq 6 \) jets, \( \geq 4 \) b tags category. Because the \( p_T \) of the forth
hardest jet in figure 9.3c is usually on the lower end of the \( p_T \) spectrum, the changes
through the regression are expected to be rather large. This follows from the observation
that the regression corrects low-\( p_T \) jets stronger than jets with high \( p_T \), cf. section 8.1.2.
Because more jets are expected in a \( t\bar{t}H, H \to b\bar{b} \) decay, the fourth jet is softer in the
background sample and therefore more strongly corrected, which leads to an improved
separation of signal and background. The variable \( \frac{\sum p_{T,jet}}{\sum E_{jet}} \) (see figure 9.3d), is a
measure of the centrality of the event. It is expected that its value is higher on average
for \( t\bar{t}H, H \to b\bar{b} \) than for \( tt \). This behavior is enhanced by the regression.

Output of the final classifier

Figure 9.2 shows the BDT output, which follows the same coloring and style scheme as
the plots before, and the associated ROC curves for the unregressed (blue) and regressed
(red) BDT outputs. Comparing the output shapes, for the \( \geq 6 \) jets, \( \geq 4 \) b tags category
in figure 9.2a it can be observed that the output for the signal sample shifts to more
signal-like values when using regressed input variables, while the opposite happens for the
background sample. The area under the ROC curve of the BDT output with regressed
input variables improves from 0.721 to 0.725, which can also be seen in figure 9.2b. These
9.1. Final classifier and limit

Table 9.1.: Area under the ROC curve for the BDT output in the different analysis categories with and without applied regression.

<table>
<thead>
<tr>
<th>Category</th>
<th>ROC integral w/o regression</th>
<th>ROC integral w/ reg</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 jets, 3 b tags</td>
<td>0.783</td>
<td>0.785</td>
</tr>
<tr>
<td>4 jets, ( \geq 4 ) b tags</td>
<td>0.753</td>
<td>0.754</td>
</tr>
<tr>
<td>5 jets, 3 b tags</td>
<td>0.755</td>
<td>0.757</td>
</tr>
<tr>
<td>5 jets, ( \geq 4 ) b tags</td>
<td>0.756</td>
<td>0.758</td>
</tr>
<tr>
<td>( \geq 6 ) jets, 2 b tags</td>
<td>0.723</td>
<td>0.724</td>
</tr>
<tr>
<td>( \geq 6 ) jets, 3 b tags</td>
<td>0.735</td>
<td>0.736</td>
</tr>
<tr>
<td>( \geq 6 ) jets, ( \geq 4 ) b tags</td>
<td>0.721</td>
<td>0.725</td>
</tr>
</tbody>
</table>

Two observations show that the improvements by the b-jet regression can be propagated to the BDT and improve its signal-background separation. For the 5 jets, \( \geq 4 \) b tags category, the behavior of the BDT output in figure 9.4a is similar to the \( \geq 6 \) jets, \( \geq 4 \) b tags category, and the shifted shapes lead to an increase in the area under the ROC curves shown in figure 9.4b.

The integrals under the ROC curve for this and all other categories are listed in table 9.1. The results presented in this section indicate that the b-jet regression can be used to improve the discrimination power of the final classifier. The area under the ROC curves of the BDT outputs in the different analysis categories, which are trained/evaluated with regressed respectively unregressed input variables, are summarized in table 9.1. Small improvements can be observed in all categories, with the BDTs for the categories with \( \geq 6 \) jets, \( \geq 4 \) b tags and 5 jets, \( \geq 4 \) b tags profiting the most from the regression. While the individual improvements are not very significant, a positive trend can be observed.

Excerpts of BDT input variables, improved by the regressions, as well as the BDT outputs and associated ROC curves can be reviewed in figures B.1 to B.10 in the appendix.

9.1.2. Limit on the signal strength

With the signal strength modifier \( \mu \), defined in equation (7.3), as the most important parameter in analyses searching for ttH, new features have to be benchmarked against the limits established previously, e.g. in [5]. To do this, the BDTs trained for regressed and unregressed variables, which were described before, are used to compute the asymptotic expected upper limit (described in section 7.1). The uncertainties considered for the limit computation are listed in table 7.1 and are only the most important uncertainties for the analysis, as stated in section 7.3.

The expected limits, listed in table 9.2, are calculated using the BDT output shapes of the final classifier computed with the Monte Carlo samples listed in table 5.1. The combined limit improves by approximately 4%. Because the individual categories are changed in a similar way, this indicates a small but systematic improvement through the regression. At this point, it has to be noted that these improvements are covered by the uncertainties, calculated with the 84% respectively 16% quantile. Furthermore, a slight instability of the limit calculation procedure can cause changes of this order. Nevertheless, this is evidence that the b-jet regression can influence and improve the limit on the signal strength modifier \( \mu \). In addition to this, new optimization of the final classifier could lead to bigger improvements because of the changed separation power of the input variables.
Table 9.2.: Asymptotic expected upper limits at 95 % CL on the signal strength modifier $\mu$ using regressed and unregressed BDTs with the $\pm \sigma$ uncertainties, as described in section 7.1.2. To achieve a consistent comparison for the unregressed and the regressed case, the BDTs were trained with equal settings and events.

<table>
<thead>
<tr>
<th>Category</th>
<th>Limit</th>
<th>w/o regression</th>
<th>w/ regression</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 jets, 3 b tags</td>
<td>11.16$^{+4.67}_{-3.14}$</td>
<td>11.03$^{+4.62}_{-3.11}$</td>
<td></td>
</tr>
<tr>
<td>4 jets, $\geq$ 4 b tags</td>
<td>12.22$^{+5.99}_{-3.78}$</td>
<td>12.22$^{+5.99}_{-3.77}$</td>
<td></td>
</tr>
<tr>
<td>5 jets, 3 b tags</td>
<td>7.28$^{+2.99}_{-2.07}$</td>
<td>7.16$^{+2.99}_{-2.01}$</td>
<td></td>
</tr>
<tr>
<td>5 jets, $\geq$ 4 b tags</td>
<td>6.02$^{+2.81}_{-1.81}$</td>
<td>5.89$^{+2.79}_{-1.77}$</td>
<td></td>
</tr>
<tr>
<td>$\geq$ 6 jets, 2 b tags</td>
<td>13.95$^{+6.51}_{-4.19}$</td>
<td>13.69$^{+6.38}_{-4.07}$</td>
<td></td>
</tr>
<tr>
<td>$\geq$ 6 jets, 3 b tags</td>
<td>5.39$^{+2.21}_{-1.51}$</td>
<td>5.19$^{+2.18}_{-1.47}$</td>
<td></td>
</tr>
<tr>
<td>$\geq$ 6 jets, $\geq$ 4 b tags</td>
<td>4.05$^{+1.89}_{-1.12}$</td>
<td>3.95$^{+1.84}_{-1.19}$</td>
<td></td>
</tr>
<tr>
<td>Combination</td>
<td>2.90$^{+1.24}_{-0.83}$</td>
<td>2.79$^{+1.19}_{-0.81}$</td>
<td></td>
</tr>
</tbody>
</table>

9.2. Validation

In chapter 8 and section 9.1 it was shown that the b-jet regression can improve measurements in simulated data. Because the regression should also perform on real data, this section will explore different validation methods to evaluate the regression performance on data.

The data used for these studies was collected from proton-proton collisions recorded in the time span from April 22th to July 15th, 2016, by the CMS experiment, corresponding to 12.9 fb$^{-1}$ of integrated luminosity. The associated Monte Carlo datasets are listed in table 5.1 whereas the $\bar{t}t$ sample is split into subprocesses based on the flavor of additional jets (see section 5.2.1 for detailed information).

9.2.1. Regression input and output

A big concern with multivariate methods used in particle physics is to assert the behavior of the input variables as well as the output on simulation and data. The main reason for this is that, while the algorithm is trained on the simulation, the data is where the physics can be found and therefore good agreement between both is essential for all input variables.

Comparisons between real and simulated data for an excerpt of the input variables for the b-jet regression are shown in figure 9.5. This can be viewed as a representative set, displaying the observable agreements between simulation and data. For this validation method, the single lepton selection was chosen because the aim is to investigate the data/simulation discrepancy in the context of the search for $t\bar{t}H$, $H \to b\bar{b}$ in the semileptonic channel. Only the behavior of jets used for the regression is of interest for this validation and therefore, only jets with a CSV value $\geq 0.8$ are included. Figure 9.5a shows the $p_T$ of the jets prior to the regression. The overall shape agreement is good, and

\[1\] The remaining variables can be found in appendix C in figures C.1, C.2 and C.3.
9.2. Validation

Figure 9.5.: Representative set of b-jet regression input variables. The uncertainties considered for the error band are listed in table 7.1. The modeling/calibration of the input variables is good for most kinematic variables, like the jet $p_T$ (a). Other variables are not well modeled/calibrated like the vertex variables, e.g. the mass of the secondary vertex associated to the jet (b).

Figure 9.6.: Output of the b-jet regression. The uncertainties considered for the error band are listed in table 7.1. While perfect agreement between real and simulated data is not necessary, because differences in the jet description are possible, the strong disagreement observable in the ratio plot can very likely not only be explained by these expected differences. An explanation could be the bad modeling/calibration of the input variables, cf. figure 9.5.
Figure 9.7.: Comparison of the agreement between data and simulation for different periods of data taking. The uncertainties considered for the error band are listed in table 7.1. The plots on the left side (a,c,e) are generated with the MC dataset listed in table B.1 and the latest data, as used before. The plots on the right side (b,d,f) are generated with the MC dataset listed in table B.1 and 2.7 fb⁻¹ of data collected in 2015 from the July 3rd to November 3rd, also used in ttH, H → bb analysis described in 5. Because these datasets where reprocessed using dedicated alignments and calibrations, better modeling on the variables is expected and was also observed.
the slight rate difference is within the uncertainties and can probably be attributed to missing/incomplete calibrations. Other variables show large disagreements in shape and rate. Many vertex variables, e.g. the mass shown in figure 9.5b have a different shape in data, which could be explained by the non-final tracker alignment and should therefore be improvable by further calibrations. Other highly ranked variables (see table 8.2), like the total hadronic energy fraction (fig. 9.7c), show comparable shape discrepancies.

The regression output shape in figure 9.6 shows the consequences of these disagreements and therefore is a very good example of the difficulties encountered in multivariate analysis. However perfect agreement for the regression output is not expected, because the description of the input variables, especially in extreme regions of the phase space, is not perfect either. One reason for this is that some differences between recorded and simulated data for b jets are not covered by the residual correction\(^2\). Because only b-tagged jets are compared in this study and the regression was trained on simulated data, the regression could be sensitive to such effects.

At the time this thesis is composed, the dataset was collected very recently and therefore it is not well calibrated, e.g. because of increased instantaneous luminosity due to better performance of the LHC and the fact that some corrections are derived from data and therefore can only be correctly computed after data was taken. To prove this, a b-jet regression trained for the datasets, used in [5], associated with 2.7 fb\(^{-1}\) of data collected in 2015 from July 3\(^{rd}\) to November 3\(^{rd}\), is compared to the most recent datasets. Because the older datasets are better understood and well calibrated, the agreement between data and simulation for the regression input variables should be better. This is presented in figure 9.7. Each row shows the same variables, on the left with the latest and on the right with the well understood datasets. The mass of the reconstructed secondary vertex is shown in 9.7a and 9.7b respectively. Because the calibration, e.g. the tracker alignment, for the older dataset is much better, the description of the vertex variables, which are particularly dependent on precise track reconstruction, is superior. Figures 9.7c and 9.7d show the total hadronic fractions of the jet energy, which are among others measured using the HCAL. Because in the newer runs of the LHC, the instantaneous luminosity increased,

\(^2\) Similarly to JES the residual correction is derived using samples with only small numbers of b-jets.
more underlying events are recorded and therefore the pile-up increases, which leads to
more noise in the calorimeters. These effects are not fully understood for the new dataset
and therefore, the descriptions of such variables is not good. Comparisons for all other
regression input variables can be found in appendix C, figures C.4 to C.8.
Because the agreement for the old datasets is better, the same is expected for the re-
gression output, which is shown in figures 9.7c and 9.7f. In fact, this is the case. The
agreement between data and simulation for the 2.7 fb$^{-1}$ datasets is much better than for
the recent datasets, but as described before, perfect agreement for the regression output
is not expected.
Therefore it should be aim for good description of variables, like the jet $p_T$ before and
after the regression was applied. Such variables show, if the regression deteriorated the
"over all" agreement between recorded and simulated data for variables used in analysis
steps after the regression was applied (e.g. in the final classifier of the tH analysis). This
is shown for the $p_T$ of the hardest jet of each event before (figure 9.8a) and after (fig-
ure 9.8b) the regression was applied. By comparing the ratio plots, it can be seen that
the regression does not worsen the agreement between data and simulation for higher level
variables, used in parts of the analysis after the regression. However, this is no indication
that the bad modeling of the input variables is negligible.

9.2.2. b-jet $p_T$ resolution

Methods that are used for the calibration and validation of the general jet energy correc-
tions at CMS [116,138] can also be applied to validate the b-jet regression. The $\gamma/Z+$jets
$p_T$-balancing technique is used to measure the jet energy response or the jet-$p_T$ resolution
with respect to a well-reconstructed reference object. The reference object has a much
better $p_T$ resolution than the jet and therefore can be used to measure the absolute re-
sponse

$$R_{\text{abs}} = \frac{p_{\text{Jet}}}{p_{\gamma/Z}},$$

(9.1)

which will from now on be denoted as the $p_T$ balance of the system. For this study, only
decays of Z bosons to lepton pairs were considered and therefore the Z+jets selection
described in section 6.2.2 was used. Furthermore, $1 \leq N_{\text{jets}} \leq 2$ and $1 \leq N_{\text{bTags}} \leq 2$
were demanded to get a good measure of the relative resolution of b jets with respect to
leptonic Z-boson decay products. Because most of the jets originate from initial or final
state radiation, the appearance of b-jets is rather unlikely. Therefore, not many events
pass this selection, which is the reason to include events with one jet. Similarly to the
jet-$p_T$ response, described in section 6.1.2.1, the $p_T$ balance of the system is expected to be
symmetrically distributed around one. In the case of $N_{\text{jets}} = 2$, the $p_{\text{Jet}}$ in equation (9.1)
is the vectorial sum of the $p_T$ of two jets.

Figure 9.9a shows the $p_T$ balance without the regression applied. As described in sec-
tion 6.1.2.1 responses are expected to be approximately Gaussian distributed, cf. equa-
tion (6.3), with a mean of one. In figure 9.9a it can be observed that this is not the case for b jets, because the distribution is slightly asymmetrical and the mean is below
one. This changes when applying the regression, which is shown in figure 9.9b. The
overall distribution gets more symmetric and the mean is shifted closer to one, both on
simulation and data. To illustrate this further, the simulated Z+Jets sample is plotted
in figures 9.10a for $N_{\text{jets}} = 1$ and 9.10b for $N_{\text{jets}} = 2$. The fits for all distributions are
performed for the peak, under the assumption that it follows a Gaussian distribution and
the range was chosen to represent a symmetrical region around the peak. The plots show

\footnote{The input variables and output of the regression with this selection can be examined in appendix C, figures C.9 to C.11}
Figure 9.9.: Validation of the b-jet regression with the $\gamma/Z+\text{Jets}$ $p_T$-balancing technique (see equation (9.1)). The $Z+\text{jets}$ selection described in section 6.2 was used on the samples listed in table 7.1, with the additional requirement of $1 \leq N_{\text{Jets}} \leq 2$ and at least one b-tagged jet. The relative resolution of the b-jets with respect to the $Z$-boson decay products improves when the b-jet regression is applied.

Figure 9.10.: The $Z+\text{jets}$ selection described in section 6.2 with at least one b-tagged jet was used to compute the $p_T$ balance in a $Z+\text{Jets}$ single (a) and dijet (b) system. The peak was fitted under the assumption that it is approximately Gaussian distributed, where the fit region was chosen to represent a symmetrical region around the mean value.
that the relative resolution $\sigma/\mu$ improved in both cases. For one jet this improvement is approximately 10% and for two jets approximately 3%.

In conclusion, the jet-$p_T$ response improvements found in simulation can be confirmed using data-based methods.

9.2.3. Conclusion

For the current CMS dataset, containing 12.9 fb$^{-1}$ of data, agreement with the associated simulations for regression input variables is not good in several cases. Since using a regression in the context of a physics analysis requires reliable performance on simulation as well as data, it is not advisable to use it for the current datasets. It is expected that the descriptions improve, e.g. with better tracker alignment, after the data are reprocessed, the regression should be ready to be used in future analysis. This is backed by the fact that the description of input variables is better in well understood/calibrated datasets. Additionally figures 9.8 and 9.9 show, that the response improvements can be measured and validated on independent data samples.

Furthermore an uncertainty for the b-jet regression should be determined. While the actual calculation is outside the scope of this thesis, the foundations are laid in the preceding sections. The uncertainties with the larges impact on the regression input variables and their influence on the output were investigated in section 8.1.7. No large or unexpected behavior could be observed. However, detailed studies are needed to determine a uncertainty on the output of the b-jet regression. Using the $\gamma/Z$+Jets $p_T$-balancing technique described above or the missing transverse energy projection fraction method, usually abbreviated as MPF, described in [138], the error on the response can be calculated, similarly to the methods used in [116], which could be used as uncertainty of the regression. Thereby it should be ensured that already covered uncertainties, like the JES uncertainty, are not considered a second time.

The influence of different Monte Carlo generators for the hard process as well as the parton shower should also be included in this considerations, because the processes corrected by the b-jet regression depend on the corresponding modeling.
10. Summary and outlook

In 2012, a standard model like Higgs boson was discovered at the LHC. To determine if this is the Higgs boson predicted by the standard model its properties need to be measured. Because not all decay and production modes are discovered yet, one possibility to contribute to the understanding of the Higgs boson is to search for these modes. The search for the Higgs-boson production in association with a top-quark pair (t\(\bar{t}\)H), which can be used to measure the nature of the coupling between the Higgs boson and fermions, is one example of such an analysis. Furthermore this production channel could yield insights into physics beyond the standard model.

The production of t\(\bar{t}\)H has a small cross section as well as a complicated topology with very difficult signal-background separation. Current analyses are not sensitive enough to discover t\(\bar{t}\)H with the available data recorded by the CMS experiment. Therefore, new methods are required to further improve sensitivity. One such method, the b-jet regression, was introduced and evaluated in this thesis:

The main goal of the b-jet regression was to compensate effects worsening the energy measurement of jets originating from bottom quarks, by deriving a new dedicated correction for such jets. First the configuration and performance of the b-jet regression was studied. By incorporating the b-jet regression into the analysis the energy measurement of b jet was improved. This led to improved jet-\(p_T\) response for b-jets and better mass reconstruction performance.

The sensitivity of the t\(\bar{t}\)H analysis can be improved by increasing the separation power of the final classifier for t\(\bar{t}\)H. By applying the b-jet regression, many variables with strong dependence on the b-jet response, like reconstructed dijet masses related to the Higgs boson, showed promising improvements in signal background separation power. Propagating these to the final discriminant and with this to the expected limit on the signal strength modifier \(\mu = \sigma_{t\bar{t}H}/\sigma_{t\bar{t}H,SM}\) resulted in small improvements, whose significance could not be proven in this thesis. Various validation methods were investigated using the latest only recently collected, at the time this thesis was written, data recorded by the CMS experiment. The very different behavior of the regression on simulated and recorded data indicates that it is not advisable to utilize the regression until the final detector calibration is available, because the regression is highly sensitive to miscalibrations of primary and secondary vertex related quantities. Comparing this difference in behavior to a regression trained and evaluated on an older but very well calibrated dataset showed that similar performance on simulated and recorded data is achievable. Furthermore, it was shown that
the performance of the b-jet regression can be validated and measured, using methods commonly utilized at CMS, like $Z + \text{jets} \ p_T$ balancing.

To evaluate the performance of the b-jet regression on the search for $t\bar{t}H$ and determine an uncertainty, further studies are required after the final calibration is finished and the datasets are reprocessed. Even though the influence of the b-jet regression on the expected limit is small, $t\bar{t}H$ analyses, especially in the $H \rightarrow b\bar{b}$ final state, could nevertheless profit from including a technique like this. Improving the b-jet response is expected to improve not only reconstructed observables, as shown in section 8.2 for some variables, but also reconstruction efficiencies or other procedures strongly dependent on b-jet energy measurements. One example is the search for $t\bar{t}H$ using objects with high transverse momentum, as published in [139]. Further studies on input variables for the b-jet regression as well as advanced techniques, like multi-target regression algorithms, could yield additional improvements.

Since these studies were conducted, more data was recorded by the CMS experiment and with this the discovery or exclusion of $t\bar{t}H$ is getting closer. But even after the discovery the analysis techniques need to be refined in order to improve the precision of the top-Higgs coupling measurement. The b-jet regression can be used for such refinements not only for $t\bar{t}H$ but also for other searches and analyses relying on b-jets.
Bibliography

Standard Model Higgs boson with the ATLAS detector at the LHC,” *Physics Letters

2012.


date Via Its Decays to Z Boson Pairs,” *Phys. Rev. Lett.*, vol. 110, no. 8, p. 081803,
2013.

[5] CMS Collaboration, “Search for $t\bar{t}H$ production in the $H \rightarrow b\bar{b}$ decay channel with
$\sqrt{s} = 13$ TeV pp collisions at the cms experiment,” Tech. Rep. CMS-PAS-HIG-16-
004, 2016.


and the Structure of the Nucleon,” *Physical Review*, vol. 185, no. 5, pp. 1975–1982,


Bibliography


[102] Z. Yu *et al.*, “QCD NLO and EW NLO corrections to $t\bar{t}H$ production with top quark decays at hadron collider,” *Physics Letters B*, vol. 738, pp. 1–5, Nov. 2014.


Appendix
A. Performance of the b-jet regression

Figure A.1.: $p_T$-dependent bias of the prediction for trainings with parton $p_T$ response (a) and generator-level jet $p_T$ response (b), cf. (6.3) respectively (6.4), as target. These distributions are also shown in figures 8.4c and 8.4d with stacked distributions.

Figure A.2.: Simple regression model for a Gaussian distribution that was transformed using two random uniformly distributed numbers, cf. (8.2). The plots show the correlation between target and output of the regression. One regression was trained with all information (a), while another regression was trained without the knowledge of the random distribution smearing the Gaussian distribution (b).
Figure A.3.: The mass of the dijet pair with the mass closest to 125 GeV is displayed in (a) and the mass of the two closest tagged jets in (b). On a $t\bar{t}$ sample (a) should peak at the generator Higgs-boson mass of 125 GeV but with a less pronounced peak compared to a $H \to b\bar{b}$ sample (see figure 8.10a). Figure (b) on the other hand should exhibit no special peak around 125 GeV in comparison to figure 8.10b. The regression does not create features that could lead to a decrease discriminability between signal and background.
B. Discriminating variables and the final classifier

Figure B.1.: Input variables for the BDT, whose signal-background separation is improved by the regression. The values given in the upper left corner correspond to the signal background separation power of the variable. In each plot, the signal (red) and background (blue) distributions are shown without (dashed) and with (solid) regression applied. For the category with \( \geq 6 \) Jets and 3 b tags, e.g. the masses of the dijet pairs closest to 125 GeV (a) and in closest proximity to each other (b) receive an improvement from the regression.

Figure B.2.: BDT outputs and ROC curves, in the category with \( \geq 6 \) Jets and 3 b tags. The BDT output (a) for training with regressed (solid) and unregressed (dashed) jets is shown for signal ttH (red) and background tt (blue) sample. The ROC curves (b) are calculated for the classifiers trained/evaluated with regressed (red) and unregressed (blue) input variables.
B. Discriminating variables and the final classifier

Figure B.3.: Input variables for the BDT, whose signal-background separation is improved by the regression. The values given in the upper left corner correspond to the signal background separation power of the variable. In each plot, the signal (red) and background (blue) distributions are shown without (dashed) and with (solid) regression applied. For the category with \( \geq 6 \) Jets and 2 b tags, e.g. the \( p_T \) of all objects in the event (a) and the mass of the closest tagged dijet pair (b) receive an improvement from the regression.

Figure B.4.: BDT outputs and ROC curves, in the category with \( \geq 6 \) Jets and 2 b tags. The BDT output (a) for training with regressed (solid) and unregressed (dashed) jets is shown for signal t\( t\bar{H} \) (red) and background \( \bar{t}t \) (blue) sample. The ROC curves (b) are calculated for the classifiers trained/evaluated with regressed (red) and unregressed (blue) input variables.
Figure B.5.: Input variables for the BDT, whose signal-background separation is improved by the regression. The values given in the upper left corner correspond to the signal background separation power of the variable. In each plot, the signal (red) and background (blue) distributions are shown without (dashed) and with (solid) regression applied. For the category with 5 Jets and 3 b tags, e.g. the masses of the dijet pairs closest to 125 GeV (a) and the sphericity (b) receive an biggest improvement from the regression.

Figure B.6.: BDT outputs and ROC curves, in the category with 5 Jets and 3 b tags. The BDT output (a) for training with regressed (solid) and unregressed (dashed) jets is shown for signal ttH (red) and background tt (blue) sample. The ROC curves (b) are calculated for the classifiers trained/evaluated with regressed (red) and unregressed (blue) input variables.
B. Discriminating variables and the final classifier

Figure B.7.: Input variables for the BDT, whose signal-background separation is improved by the regression. The values given in the upper left corner correspond to the signal-background separation power of the variable. In each plot, the signal (red) and background (blue) distributions are shown without (dashed) and with (solid) regression applied. For the category with 4 Jets and \( \geq 4 \) b tags, e.g. the sum of lepton, jet and \( E_T \) (a) and the mass of the dijet pair closest to 125 GeV (b) receive an improvement from the regression.

Figure B.8.: BDT outputs and ROC curves, in the category with 4 Jets and \( \geq 4 \) b tags. The BDT output (a) for training with regressed (solid) and unregressed (dashed) jets is shown for signal ttH (red) and background tt (blue) sample. The ROC curves (b) are calculated for the classifiers trained/evaluated with regressed (red) and unregressed (blue) input variables.
Figure B.9.: Input variables for the BDT, whose signal-background separation is improved by the regression. The values given in the upper left corner correspond to the signal background separation power of the variable. In each plot, the signal (red) and background (blue) distributions are shown without (dashed) and with (solid) regression applied. For the category with 4 Jets and 3 b tags, e.g. the mass of the dijet pair closest to 125 GeV (a) an HT (b) receive an improvement from the regression.

![Figure B.9.](bdt_input_variables.png)

Figure B.10.: BDT outputs and ROC curves, in the category with 4 Jets and 3 b tags. The BDT output (a) for training with regressed (solid) and unregressed (dashed) jets is shown for signal t\(\bar{t}\)H (red) and background \(\bar{t}\)t (blue) sample. The ROC curves (b) are calculated for the classifiers trained/evaluated with regressed (red) and unregressed (blue) input variables.

![Figure B.10.](bdt_outputs_and_roc_curves.png)
Table B.1.: Summary of the samples of simulated processes considered in the analysis, described in [5]. For each sample, the process, the event generator+matrix-element generator combination, and the cross section used for the scaling of the events are listed. The datasets are generated as described in section 5.2 and the listed cross section are calculated, if not otherwise indicated, as outlined in section 5.2.1.

<table>
<thead>
<tr>
<th>Process</th>
<th>Remark</th>
<th>Event Generator Configuration</th>
<th>$\sigma_{\text{theo}}$ [pb]</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t\bar{t}H, H \to b\bar{b}$</td>
<td>$m_H = 125$ GeV</td>
<td>Powheg+Pythia8</td>
<td>0.2918</td>
</tr>
<tr>
<td>$t\bar{t}H, H \to$ non $b\bar{b}$</td>
<td>$m_H = 125$ GeV</td>
<td>Powheg+Pythia8</td>
<td>0.214</td>
</tr>
<tr>
<td>$t\bar{t}+\text{jets}$</td>
<td>inclusive</td>
<td>Powheg+Pythia8</td>
<td>831.76</td>
</tr>
<tr>
<td>$t\bar{t}+\text{jets}$</td>
<td>exclusive semi-leptonic</td>
<td>Powheg+Pythia8</td>
<td>364.416</td>
</tr>
<tr>
<td>$t\bar{t}+\text{jets}$</td>
<td>exclusive di-leptonic</td>
<td>Powheg+Pythia8</td>
<td>87.36</td>
</tr>
<tr>
<td>$t$</td>
<td>t-channel</td>
<td>Powheg+Pythia8</td>
<td>45.34</td>
</tr>
<tr>
<td>$\bar{t}$</td>
<td>t-channel</td>
<td>Powheg+Pythia8</td>
<td>27.98</td>
</tr>
<tr>
<td>$t$</td>
<td>tW-channel</td>
<td>Powheg+Pythia8</td>
<td>35.9</td>
</tr>
<tr>
<td>$\bar{t}$</td>
<td>tW-channel</td>
<td>Powheg+Pythia8</td>
<td>35.9</td>
</tr>
<tr>
<td>$t$</td>
<td>s-channel</td>
<td>Powheg+Pythia8</td>
<td>3.44</td>
</tr>
<tr>
<td>$t\bar{t}+Z, Z \to q\bar{q}$</td>
<td></td>
<td>MG5_aMC@NLO+Pythia8</td>
<td>0.611 [104]</td>
</tr>
<tr>
<td>$t\bar{t}+Z, Z \to ll$</td>
<td></td>
<td>MG5_aMC@NLO+Pythia8</td>
<td>0.2529 [104]</td>
</tr>
<tr>
<td>$t\bar{t}+W, W \to q\bar{q}$</td>
<td></td>
<td>MG5_aMC@NLO+Pythia8</td>
<td>0.435 [104]</td>
</tr>
<tr>
<td>$t\bar{t}+W, W \to ll$</td>
<td></td>
<td>MG5_aMC@NLO+Pythia8</td>
<td>0.210 [104]</td>
</tr>
<tr>
<td>$W + \text{Jets}$, $W \to ll$</td>
<td>$100 &lt; H_T \leq 200$ GeV</td>
<td>Madgraph+Pythia8</td>
<td>1345</td>
</tr>
<tr>
<td>$W + \text{Jets}$, $W \to ll$</td>
<td>$200 &lt; H_T \leq 400$ GeV</td>
<td>Madgraph+Pythia8</td>
<td>359.7</td>
</tr>
<tr>
<td>$W + \text{Jets}$, $W \to ll$</td>
<td>$400 &lt; H_T \leq 600$ GeV</td>
<td>Madgraph+Pythia8</td>
<td>48.91</td>
</tr>
<tr>
<td>$W + \text{Jets}$, $W \to ll$</td>
<td>$600 &lt; H_T \leq 800$ GeV</td>
<td>Madgraph+Pythia8</td>
<td>12.05</td>
</tr>
<tr>
<td>$W + \text{Jets}$, $W \to ll$</td>
<td>$800 &lt; H_T \leq 1200$ GeV</td>
<td>Madgraph+Pythia8</td>
<td>5.501</td>
</tr>
<tr>
<td>$W + \text{Jets}$, $W \to ll$</td>
<td>$1200 &lt; H_T \leq 2500$ GeV</td>
<td>Madgraph+Pythia8</td>
<td>1.329</td>
</tr>
<tr>
<td>$W + \text{Jets}$, $W \to ll$</td>
<td>$H_T \geq 2500$ GeV</td>
<td>Madgraph+Pythia8</td>
<td>0.03216</td>
</tr>
<tr>
<td>$Z/\gamma* \to \ell\bar{\ell} + \text{Jets}$</td>
<td>$10 &lt; m \leq 50$ GeV</td>
<td>MG5_aMC@NLO+Pythia8</td>
<td>22635.09</td>
</tr>
<tr>
<td>$Z/\gamma* \to \ell\bar{\ell} + \text{Jets}$</td>
<td>$m &gt; 50$ GeV</td>
<td>MG5_aMC@NLO+Pythia8</td>
<td>6025.2</td>
</tr>
<tr>
<td>$WW$</td>
<td></td>
<td>Pythia8</td>
<td>118.7 [140]</td>
</tr>
<tr>
<td>$WZ$</td>
<td></td>
<td>Pythia8</td>
<td>44.9 [141]</td>
</tr>
<tr>
<td>$ZZ$</td>
<td></td>
<td>Pythia8</td>
<td>15.4 [141]</td>
</tr>
</tbody>
</table>
C. Validation

Figure C.1.: Input variables for the b-jet regression containing information about the jet kinematics. The uncertainties considered for the error band are listed in table 7.1.
Figure C.2.: Input variables for the b-jet regression containing information about the secondary vertex associated with the jet. The uncertainties considered for the error band are listed in table 7.1.
Figure C.3.: Input variables for the b-jet regression containing information about the particle composition (a), the matched leptons (b)-(d) and the number of primary vertices of the event (e). The uncertainties considered for the error band are listed in table 7.1.
Figure C.4.: Influence of the modeling and calibration of different datasets on the b-jet regression. The uncertainties considered for the error band are listed in table 7.1. The plots on the right side (a,c,e) are generated with the MC dataset listed in table B.1 and the data used before. The plots in the right side (b,d,f) are generated with the MC dataset listed in table B.1 and 2.7 fb⁻¹ collected in 2015 from the July 3rd to November 3rd.
Figure C.5.: Influence of the modeling and calibration of different datasets on the b-jet regression. The uncertainties considered for the error band are listed in table 7.1. The plots on the right side (a,c,e) are generated with the MC dataset listed in table 5.1 and the data used before. The plots in the right side (b,d,f) are generated with the MC dataset listed in table B.1 and 2.7 fb$^{-1}$ collected in 2015 from the July 3rd to November 3rd.
Figure C.6.: Influence of the modeling and calibration of different datasets on the b-jet regression. The uncertainties considered for the error band are listed in table 7.1. The plots on the right side (a,c,e) are generated with the MC dataset listed in table 5.1 and the data used before. The plots in the right side (b,d,f) are generated with the MC dataset listed in table B.1 and 2.7 fb$^{-1}$ collected in 2015 from the July 3rd to November 3rd.
Figure C.7.: Influence of the modeling and calibration of different datasets on the b-jet regression. The uncertainties considered for the error band are listed in table 7.1. The plots on the right side (a,c,e) are generated with the MC dataset listed in table 5.1 and the data used before. The plots in the right side (b,d,f) are generated with the MC dataset listed in table B.1 and 2.7 fb⁻¹ collected in 2015 from the July 3rd to November 3rd.
Figure C.8.: Influence of the modeling and calibration of different datasets on the b-jet regression. The uncertainties considered for the error band are listed in table 7.1. The plots on the right side (a,c,e) are generated with the MC dataset listed in table 5.1 and the data used before. The plots in the right side (b,d,f) are generated with the MC dataset listed in table B.1 and 2.7 fb$^{-1}$ collected in 2015 from the July 3rd to November 3rd.
Figure C.9.: Input variables of the b-jet regression for samples with applied Z+jets selection, cf. section 6.2.2. The regression was trained on a simulated tt Sample.
Figure C.10.: Input variables of the b-jet regression for samples with applied Z+jets selection, cf. section 6.2.2. The regression was trained on a simulated tt Sample.
Figure C.11.: Input variables of the b-jet regression (a)-(d) and regression output (e) for samples with applied Z+jets selection, cf. section 6.2.2. The regression was trained on a simulated tt Sample.
Danksagungen

Zuerst möchte ich mich bei Prof. Dr. Ulrich Husemann für die Aufnahme in seine Arbeitsgruppe, die Ermöglichung dieser Masterarbeit und die Unterstützung bedanken.

Des Weiteren möchte ich mich bei Prof. Dr. Günter Quast für die Übernahme des Korreferat bedanken.

Großer Dank gilt außerdem Dr. Matthias Schröder und Karim El Morabit für die geduldige Unterstützung sowie das Korrekturlesen dieser Arbeit. Außerdem möchte ich mich an dieser Stelle bei Hannes Mildner für die Beantwortung aller meiner fachlichen Fragen und für die Diskussionen rund um das Thema meiner Arbeit bedanken.


Besonderer Dank gilt meiner Frau Anna, die mich über das gesamte Physikstudium begleitete, mir den Rücken frei hielt, als es nötig war, und die letzten fünf Jahre insgesamt zu einer wunderschönen Zeit gemacht hat.

Herzlichen Dank euch allen!