Performance of Jet-Substructure Techniques in Higgs-Boson Production in Association with tt and Study of First LHC Run-II Data at CMS

Tobias Christian Pfotzer

Master Thesis

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

Referee: Prof. Dr. Ulrich Husemann
Second referee: Prof. Dr. Günter Quast

Karlsruhe, 08 October 2015
This page intentionally left blank.
Accepted by the first referee of the master thesis.
Karlsruhe, 08 October 2015

(Prof. Dr. Ulrich Husemann)
This page intentionally left blank.
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, 08 October 2015

(Tobias Christian Pfotzer)
This page intentionally left blank.
Introduction

The new boson, whose discovery was announced in Summer 2012 at both the CMS and ATLAS experiment at the LHC is compatible with the Higgs-Boson that was predicted by the theories of Brout, Englert and Higgs. The Higgs mechanism, which uses spontaneous symmetry breaking to allow for massive vector bosons, introduces the Higgs-boson as a by-product of the theory. The Higgs-Boson couples to all massive particles, the strength of the coupling to fermions is proportional to the size of the mass. Among others, the LHC was built to produce the Higgs-boson in the mass window that was predicted by the combined results of theoretical physicists and previous particle accelerators.

By now, the Higgs boson has been discovered in several direct production processes. Another very promising process is the associated production of the Higgs boson and a top-quark anti-quark pair ($t\bar{t}H$), which has already been studied in a variety of decay channels [1]. This process is especially interesting, as it allows for a direct measurement of the coupling of the Higgs to the top quark, which could so far only be measured indirectly, using additional assumptions.

The process that is studied in the context of this thesis is the decay of the Higgs boson into bottom quarks and the decay of a top-quark with one hadronically and leptonically decaying W-boson, which leads to a single lepton (muon or electron) and several jets in the final state. A large challenge is posed by the many possible reconstruction hypotheses due to the large amount of jets. As explained in [2], events were selected in which the Higgs boson and the hadronically decaying top quark have large transverse momenta. This leads to a collimation of their decay objects into large jets, called fat jets. The elements of these fat jets are reconstructed using different jet-substructure algorithms. This procedure also allows to recover events that were otherwise not selected.

The results of this first part are based on proton-proton-collisions at the LHC with a center of mass energy of 8 TeV, corresponding to an integrated luminosity of 19.8 fb$^{-1}$. The performance of these subjet-algorithms and the bottom-quark jet identification, when applied on these subjets, have a performance equal to that of standard jets.

In the second part of the thesis, recent proton-proton collision data collected at a center of mass energy of 13 TeV are studied. The data represent an integrated luminosity of 42 pb$^{-1}$. The distributions, which contain events with at least one lepton (electron or muon) aim to show the agreement between data and Monte Carlo and the effects of the first residual corrections based on this data. Generally, good description of the data in the new energy regime is found within the statistical precision.

In chapter 1, an overview of the standard model of particle physics, the Top quark and the Higgs boson is given. Chapters 2 presents the properties of the Large Hadron Collider (LHC) and the Compact Muon Solenoid (CMS) experiment and its subsystems. In chapter 3, the methods of reconstructing physical objects from the information gathered at the CMS experiment are shown. To simulate interactions at particle level, special programs, called Monte Carlo generators, are employed. The mechanics of these generators together with a selection of generators is given in chapter 4. In chapter 5, The selection criteria which are applied on the measured and simulated data are presented. Chapter 6 contains
the studies on the two subjet algorithms that aim to reconstruct the Higgs boson and the hadronically decaying top-quark. The studies of the performance of b-tagging algorithms when applied on subjets are shown in chapter 7. Finally, the distributions based on the first LHC Run-II data are presented in chapter 8.
## Contents

1 Theory of Particle Physics ........................................... 1  
1.1 Standard Model .................................................... 1  
1.1.1 Elementary Particles of the Standard Model ............... 1  
1.1.2 Quantum Field Theory and Gauge Theory ................. 3  
1.1.3 Elementary Forces of the Standard Model ............... 4  
1.1.4 Higgs Mechanism and Symmetry Breaking ................. 8  
1.1.5 Cross Section Calculation .................................. 10  
1.2 Top Quark .......................................................... 12  
1.2.1 Properties of the Top Quark ............................... 12  
1.2.2 Top Quark Production ....................................... 12  
1.2.3 Top Quark Decay Channels ................................ 13  
1.3 Higgs Boson ........................................................ 13  
1.3.1 Higgs Boson Production .................................... 14  
1.3.2 Higgs Decay .................................................. 15  

2 Experimental Setup at CERN ..................................... 17  
2.1 Large Hadron Collider .......................................... 17  
2.2 The CMS Experiment ............................................ 20  
2.2.1 Detector Geometry and Variables ......................... 21  
2.2.2 Tracking System ............................................ 22  
2.2.3 Electromagnetic Calorimeter ............................ 24  
2.2.4 Hadronic Calorimeter ..................................... 25  
2.2.5 Magnet Yoke and Muon System ......................... 27  
2.2.6 Data Taking and Trigger ................................ 29  

3 Object Reconstruction at the CMS Experiment ............. 31  
3.1 Track and Primary Vertex Reconstruction .................. 32  
3.1.1 Primary Vertex Reconstruction .......................... 32  
3.2 Electron and Photon Reconstruction ......................... 33  
3.3 Muon Reconstruction ........................................... 33  
3.4 Missing Transverse Energy .................................... 34  
3.5 Particle Flow Algorithm ...................................... 34  
3.6 Jet Reconstruction and Algorithms .......................... 35  
3.6.1 Cone Algorithms ........................................ 36  
3.6.2 Sequential Recombination Algorithms .................. 36  
3.6.3 Jet Energy Correction ................................... 37  
3.7 Identification of Bottom-Quark Jets ......................... 38  
3.7.1 Impact Parameter Algorithms ........................... 38  
3.7.2 Secondary Vertex Algorithms ............................ 39  
3.8 Subjet Reconstruction with Fat Jet Algorithms ........... 39  
3.8.1 HEP-Toptagger ........................................... 40  
3.8.2 Subjet-Filterjet-Algorithm ............................... 42
4 Data Generation and Monte Carlo Generators
4.1 Process of Event Generation ................................. 45
4.1.1 Hard Process ........................................... 46
4.1.2 Showering and Hadronization ............................ 46
4.1.3 Underlying Event and Pile Up ............................ 47
4.1.4 Hadron Decay ........................................... 47
4.1.5 Detector Simulation and Unfolding ....................... 47
4.2 Variants of Monte Carlo Generators ....................... 48
4.2.1 General Purpose Generators .............................. 48
4.2.2 Shower Monte Carlo ..................................... 49

5 Data and Event Selection ................................. 51
5.1 Signal and Background Processes ......................... 52
5.1.1 Event Signature ........................................ 52
5.1.2 Top-Quark Pair with Additional Partons ................ 52
5.1.3 Single Top Background .................................. 52
5.1.4 Vector Boson with Additional Jets ..................... 53
5.1.5 Diboson .................................................. 53
5.1.6 Not included: QCD ...................................... 54
5.2 Data Samples ............................................... 54
5.3 Monte Carlo Samples ....................................... 54
5.4 Event Selection .............................................. 55
5.4.1 Lepton Selection ........................................ 56
5.4.2 Jet Selection ............................................ 57
5.4.3 B-Tagging Selection .................................... 57
5.4.4 Fat Jet and Subjet Selection ............................. 57
5.5 Monte Carlo Event Corrections ......................... 58
5.5.1 Pileup Correction ....................................... 58
5.5.2 B-Tag Scale Factors ................................. 59
5.6 Event Yields ............................................... 60

6 Validation of Jet Substructure Reconstruction Techniques Employed in the ttH Analysis in the LHC Run I 63
6.1 Introduction ................................................. 63
6.1.1 Event Preselection ..................................... 64
6.1.2 Event Variables ....................................... 65
6.2 Performance of the HEP-Toptagger Algorithm ............ 66
6.2.1 Kinematics of CA 1.5 Jets ............................. 67
6.2.2 Performance of Different Methods of Subjet Assignment .... 69
6.2.3 Hadronic Top Tagging Method and Efficiency .......... 70
6.3 Studies on the Subjet-Filterjet Algorithm .................. 75
6.3.1 Kinematics of Higgsjets and Filterjets ................. 75
6.3.2 Tagging Methods and Efficiency ..................... 77
6.4 Validation of BDT Input Parameters ...................... 78

7 Correction of B-Tag Distribution of Filtered Subjets 81
7.1 Motivation .................................................. 81
7.2 Matching of Subjets to ak5-jets ......................... 82
7.2.1 Selection of Subjets ................................... 82
7.2.2 Distance Based Jet Matching ......................... 83
7.2.3 Comparison between Matched and Unmatched Subjets .... 84
## Contents

7.3 Comparison of Matched Subjects and Anti-kt Jets ............................................. 86
7.3.1 Jet Flavour of Matched Jets .............................................................................. 86
7.3.2 Kinematic Variables of Matched Jets ............................................................ 86
7.3.3 B-Tag Output of Matched Jets ....................................................................... 88
7.4 Correction of the B-tag Output of Filterjets ...................................................... 88
7.4.1 Introduction ........................................................................................................ 88
7.4.2 Comparison Studies between Unmatched Jets and Anti-kt Jets ................. 89
7.4.3 Comparison of the Different B-Tag Event Weights ........................................ 91
7.4.4 Event Yields Based on Different B-Tag Scale Factors .................................. 91

8 First Look at the LHC Run-II Data ........................................................................... 93
8.1 Introduction ........................................................................................................... 93
8.2 Samples ................................................................................................................ 94
8.2.1 Data .................................................................................................................. 94
8.2.2 Simulation ......................................................................................................... 94
8.3 Event Selection ...................................................................................................... 95
8.3.1 Lepton Selection ............................................................................................... 95
8.3.2 Jet Selection ..................................................................................................... 95
8.3.3 B-Tag Selection ............................................................................................... 96
8.3.4 Pileup Reweighting ........................................................................................ 96
8.3.5 QCD-Multijet Normalization Estimation ........................................................ 96
8.3.6 Event Yields ..................................................................................................... 97
8.4 Validation of LHC Run-II Data ............................................................................. 97
8.4.1 Event Variables ............................................................................................... 98
8.4.2 Jet Variables .................................................................................................... 99
8.4.3 Lepton Variables and Z-Boson Reconstruction .......................................... 102
8.4.4 Top Quark Reconstruction ............................................................................. 106
8.5 Conclusion and Outlook ...................................................................................... 107

9 Conclusion and Outlook .......................................................................................... 109

Bibliography ............................................................................................................... 111
1. Theory of Particle Physics

1.1 Standard Model

1.1.1 Elementary Particles of the Standard Model

The standard model contains 16 elementary particles which can be split into two groups based on the nature of their spin. Particles with integer spin are called bosons, while particles with half-integer spin are called fermions.

The half integer spin of fermions is equivalent to an antisymmetric wave state. Fermions can be described by a distribution function, the Fermi-Dirac distribution that describes the probability of finding a fermion with a given energy. The antisymmetric nature of the wave states leads to the so-called "Pauli principle", e.g. [4]: Since exchanging two identical fermions does not lead to the same wave state, each quantum state can only be occupied by one fermion. Fermions carry charges which are connected to the elementary interactions of the standard model: the electromagnetic, strong and weak force.

Based on the colour charge, that is associated with the strong force, fermions can be separated into two groups: Leptons, that have no colour, and quarks, which carry one of three different colours. Leptons can be further separated into charged and neutral leptons, the latter being called neutrinos.

The concept of quarks was introduced to explain the underlying symmetry that was found in strongly interacting objects[5]. The name quark was later introduced by Gell-Mann as a tribute to James Joyce[6]. Due to the nature of the strong interaction, quarks cannot be observed in isolation, but must form composite objects that consist of several quarks. The six quarks that are known today can be split into two groups. They are named up-type and down-type after the lightest quarks in each group and have an electric charge of 2/3 and -1/3, respectively. The combination of an up-type and a down-type quark forms a doublet with regard to the weak isospin, which is the charge of the weak interaction.

Leptons carry weak isospin, and in the case of the charged lepton, also an integer electric charge. The combination of a charged lepton and its neutrino forms a weak isospin doublet.

Based on their mass, fermions are separated into three so-called families. Each family contains two leptons and two quarks. The first family consists of the fermions that are the building blocks of our world: While electrons had already been discovered in electromagnetic experiments, up and down quarks were introduced as the components of protons.
Table 1.1: The fermions in the standard model separated in leptons and quarks. The presented values are the particle mass, the electric charge, the colour charge and the third component of the weak isospin, $T_3$. Masses taken from [3].

<table>
<thead>
<tr>
<th>particle</th>
<th>mass</th>
<th>charge/e</th>
<th>colour</th>
<th>$T_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electron ($e^-$)</td>
<td>0.511 MeV/c^2</td>
<td>-1</td>
<td>no</td>
<td>-1/2</td>
</tr>
<tr>
<td>Electron neutrino ($\nu_e$)</td>
<td>&lt; 0.002 MeV/c^2</td>
<td>0</td>
<td>no</td>
<td>1/2</td>
</tr>
<tr>
<td>Muon ($\mu^-$)</td>
<td>105.7 MeV/c^2</td>
<td>-1</td>
<td>no</td>
<td>-1/2</td>
</tr>
<tr>
<td>Muon neutrino ($\nu_\mu$)</td>
<td>&lt; 0.002 MeV/c^2</td>
<td>0</td>
<td>no</td>
<td>1/2</td>
</tr>
<tr>
<td>Tau ($\tau^-$)</td>
<td>1.777 GeV/c^2</td>
<td>-1</td>
<td>no</td>
<td>-1/2</td>
</tr>
<tr>
<td>Tau neutrino ($\nu_\tau$)</td>
<td>&lt; 0.002 MeV/c^2</td>
<td>0</td>
<td>no</td>
<td>1/2</td>
</tr>
<tr>
<td>Up quark (u)</td>
<td>$2.3^{+0.7}_{-0.5}$ MeV/c^2</td>
<td>1/2</td>
<td>r,g,b</td>
<td>1</td>
</tr>
<tr>
<td>Down quark (d)</td>
<td>$4.8^{+0.7}_{-0.5}$ MeV/c^2</td>
<td>1/2</td>
<td>r,g,b</td>
<td>-1/2</td>
</tr>
<tr>
<td>Charm quark (c)</td>
<td>$1.275 \pm 0.025$ GeV/c^2</td>
<td>2/3</td>
<td>r,g,b</td>
<td>1/2</td>
</tr>
<tr>
<td>Strange quark (s)</td>
<td>$95 \pm 5$ MeV/c^2</td>
<td>1/2</td>
<td>r,g,b</td>
<td>-1/2</td>
</tr>
<tr>
<td>Top quark (t)</td>
<td>$173.5 \pm 1.0$ GeV/c^2</td>
<td>2/3</td>
<td>r,g,b</td>
<td>1/2</td>
</tr>
<tr>
<td>Bottom quark (b)</td>
<td>$4.18 \pm 0.03$ GeV/c^2</td>
<td>2/3</td>
<td>r,g,b</td>
<td>-1/2</td>
</tr>
</tbody>
</table>

Table 1.2: The vector and scalar bosons that are part of the standard model. The presented values are the particle mass, the electric charge, the colour charge and the third component of the weak isospin, $T_3$. Masses taken from [3].

<table>
<thead>
<tr>
<th>particle</th>
<th>mass</th>
<th>charge/e</th>
<th>colour</th>
<th>$T_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Photon ($\gamma$)</td>
<td>$&lt; 10^{-18}$ eV/c^2</td>
<td>0</td>
<td>no</td>
<td>0</td>
</tr>
<tr>
<td>Gluon ($\nu_\mu$)</td>
<td>0</td>
<td>0</td>
<td>r,g,b</td>
<td>0</td>
</tr>
<tr>
<td>Z Boson ($\nu_e$)</td>
<td>91.188 $\pm$ 0.002 GeV/c^2</td>
<td>0</td>
<td>no</td>
<td>0</td>
</tr>
<tr>
<td>W Boson ($\mu^-$)</td>
<td>80.385 $\pm$ 0.015 GeV/c^2</td>
<td>$\pm$ 1</td>
<td>no</td>
<td>$\pm$ 1</td>
</tr>
<tr>
<td>Higgs Boson ($\tau^-$)</td>
<td>$125.09 \pm 0.21 \pm 0.11$ GeV/c^2</td>
<td>-1</td>
<td>no</td>
<td>0</td>
</tr>
</tbody>
</table>

and neutrons. The electron neutrino finally was postulated to explain the methodology of the weak interaction[7].

The second family consists of the muon and its neutrino and the strange and charm quarks. While the muon was already detected in cosmic radiation experiments in 1936[8], it took another 26 years to find its neutrino partner at Brookhaven National Laboratory (BNL)[9]. The measurement of long-living particles led to the postulation of a new quantum number, later called strangeness[10] and eventually to the SU(3) flavour symmetry. The charm quark was nearly simultaneously detected at the Stanford Linear Accelerator Center (SLAC)[11] and BNL[12] in the form of the q\bar{q} resonance in November 1974.

The remaining four particles, the tau lepton, its neutrino and the top and bottom quark form the third family of fermions. The tau and its neutrino where discovered in 1975[13] and 2001[14], respectively. Like the charm quark, the bottom quark was detected in the form of its diquark resonance, called Υ meson, in 1977[15] at Fermilab. Finally, the top quark was detected in 1995 by the DØ[16] and CDF collaboration[17] at Fermilab.

To each fermion, there exists an anti-fermion with similar properties. While masses and energies of fermions and anti-fermions are equal, the quantum numbers are reversed, for instance, a top anti-quark has a charge of $-\frac{2}{3}$, while the positron, which is the antiparticle to the electron, has a charge of +1.

Bosons have integer spin corresponding to a symmetric wave state. The distribution function of bosons is the Bose-Einstein distribution. Depending on their absolute spin,
1.1. Standard Model

Bosons are further divided into scalar bosons with spin 0 and vector bosons with spin 1. The known vector bosons in the standard model are the photon, the gluon, the $W^+$, $W^-$ and $Z$-bosons as mediators of electromagnetic, strong and weak force, respectively. The recently discovered Higgs boson is presumed to be a scalar boson with positive parity [18, 19]. While photon and gluon are massless, $W$, $Z$ and Higgs boson have a distinctive mass. The Higgs mechanism has been introduced as the theoretical foundation of the otherwise impossible: By coupling to the $W$ and $Z$ boson, the declared Higgs field "gives" mass to the bosons, giving rise to the Higgs-boson in the process. The Higgs mechanism via spontaneous symmetry breaking and the properties of the Higgs boson are explained in 1.1.4.

1.1.2 Quantum Field Theory and Gauge Theory

Quantum field theory (QFT) was introduced to expand the formalism of relativistic quantum mechanics on systems with large amounts of particles.

The mathematical description of QFT is based on the Lagrangian formalism that is used in classical physics. By solving the Euler-Lagrange equations for the Lagrange function $L$, the classical equations of motion can be derived:

$$\frac{d}{dt} \frac{\partial L}{\partial \dot{q}} - \frac{\partial L}{\partial q} = 0. \quad (1.1)$$

Here, $\frac{d}{dt}$ stands for the time derivative and $q$ and $\dot{q}$ represent the generalized coordinates of space and momentum.

In quantum field theory, a comparable approach is used. Instead of the Lagrange function $L$, the Lagrangian density $\mathcal{L}$ is used. In its simplest case, the Lagrangian describes free particles. Therefore it only contains a mass term and a term describing the kinetic energy of the particles. Similar to the classical case, there also exist the relativistic Euler-Lagrange equations that use the field $\phi$ and the covariant derivative $\partial_{\mu}$:

$$\frac{\partial \mathcal{L}}{\partial \phi} - \partial_{\mu} \left( \frac{\partial \mathcal{L}}{\partial \partial_{\mu} \phi} \right) = 0. \quad (1.2)$$

Again, the formula returns the equations of motion for the system that is described by the Lagrangian. In the case of a fermion, it is the Fermi-Dirac equation that describes a single free fermion:

$$(i\gamma^\mu \partial_{\mu} - m)\psi = 0. \quad (1.3)$$

Gauge theory is based on Noether’s theorem[20], which states that an invariance with respect to a global transformation is equivalent to the existence of a conserved quantity. Examples are the conservation of energy and time invariance or the conservation of momentum and translational invariance. In gauge theory, the symmetry with respect to a local gauge transformation leads to the conservation of charge and to the appearance of interactions based on these charges. In QFT, this charge corresponds to the charge of the elementary interaction and the gauge transformation to the transformation of the symmetry group that is associated to the charge. For example, the orientation of the charge should have no effect on the underlying physics of the associated interaction.

By demanding local gauge invariance, the Lagrangian of the elementary interactions can be derived. Starting from the Lagrangian describing the free particles, the terms are corrected...
to obey local gauge symmetry. This introduces new terms that can be associated to a term that represents the interaction between free particle and the mediator particle of the respective gauge field. Finally, field terms are added that represent the self energy of the mediator particles.

1.1.3 Elementary Forces of the Standard Model

In addition to the aforementioned elementary particles, the standard model contains three of the four known interactions: The electromagnetic force, the weak and strong nuclear interaction. Gravity is disregarded in this thesis since on microscopic scales the strength of this force is negligible compared to the other three. Further, there has been no complete formulation of quantum gravity, whose goal it is to describe gravity as a renormalizable QFT. The following section contains an overview of the interactions. Additionally, a short description of the electroweak unification and the Higgs mechanism is given.

QED

The electromagnetic force acts between particles that carry electric charge. The respective mediator particle is the photon, a vector boson. While in classical physics, light is described as an electromagnetic wave that connects the phenomena of light and electromagnetism, in particle physics the photon is a quantized particle that is exchanged or emitted. While a virtual photon is emitted every time an electromagnetic interaction occurs, photons themselves do not carry electric charge and therefore cannot interact with each other.

The gauge theory that corresponds of the electromagnetic force is Quantum Electrodynamics (QED). QED describes the electromagnetic interaction between photons and spin 1/2 particles. Historically, QED was the first theory to describe an elementary interaction using group theory. The first complete formulation of the theory was given by Tomonaga[21], Schwinger[22] and Feynman[23] in the late 1940s.

Starting with the Lagrangian of a free fermion, the demand for local gauge invariance under U(1) transformation adds terms that represent the QED vertex and the self energy of the photon:

\[ L_{QED} = \bar{\psi}(i\gamma^\mu D_\mu - m)\psi - \frac{1}{4} F_{\mu\nu} F^{\mu\nu}. \] (1.4)

In this equation, the fermion bispinor that contains the two spin states for both particle and antiparticle is represented by the object \( \psi \). The derivative has been adjusted to the gauge covariant derivative

\[ D_\mu = \partial_\mu + ieA_\mu \] (1.5)

and contains the photon interaction term. Finally, the self energy is represented by the field terms

\[ F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu. \] (1.6)

A massive photon is not allowed since an additional mass term would not be gauge invariant. Further, as the photon has no charge, there are also no terms describing photon-photon interactions.

The coupling coefficient of the electromagnetic force is the fine structure constant \( \alpha \). It is connected to the electric charge and at low energies is defined by the following formula:

\[ \alpha = \frac{e^2}{4\pi\hbar c} \approx \frac{1}{137} \] (1.7)
Measurements at different energies have shown that the value of $\alpha$ is depending on the scale of observation. An energetic scale is equivalent to an inverse length scale, where higher energies correspond to a greater resolution of the observed object. In the case of a charged fermion, a greater resolution shows that the charge is shielded by virtual lepton pairs. These leptons are created by virtual photons that are emitted by the charge. By resolving this shielding, the value of the charge increases. Increasing the distance to the charge to zero results in a diverging charge, a phenomenon called the Landau pole. Using the method of renormalization, the electric coupling can be redefined as a quantity depending on an energy scale $\Lambda$, see e.g. [24].

**QCD**

The strong force acts between particles carrying colour charge. Like the photon, the gluon is a vector boson that has no mass. However, the gluon also carries colour charge like the quark, which allows interactions between gluons and both quarks and other gluons.

The strong force gets its name by the fact that it is the strongest of the three elementary interactions. This means that for particles that are able to interact via multiple interactions, like quarks, the strong interaction in general is the dominant type.

The classical potential describing the strong force consists of two parts. Like the electromagnetic force, it has a short-range potential that reduces in strength as the distance is increased. However, it also contains a second term that increases with distance. This part of the potential, which resembles the behaviour of a spring leads to large stored energies when two quarks are pulled apart. As soon as a critical distance is reached, a new quark anti-quark pair is created from the energy, which means that it is impossible to observe a singular quark or gluon. This effect is called quark confinement and is caused by the gluon self-interaction.

The gluon self-interaction and the quark-gluon coupling lead to a change of the colour charge with energy comparable to that in QED. The corresponding coupling strength in QCD, called $\alpha_s$, also shows a scale-depending behaviour. The effect is caused by the emission of virtual gluons that split into quark-anti-quark pairs. However, the emitted gluons are also able to split into pairs of gluons, causing an anti-shielding effect. The combination of these two effects leads to a decrease of $\alpha_s$ with an increasing energy scale.

In the limit of very high energies, the coupling between quarks is very small. In this state, called asymptotic freedom, the quarks move as free particles without the QCD potential.

Since the size of the colour charge and therefore the coupling $\alpha_s$ is scale-dependant, it is necessary to define a scale $\mu$ at which $\alpha_s$ is calculated. Via the methodology of renormalization, the divergent factors that appear in the calculation of physical quantities like the coupling strength can be redefined at this scale as a function of a different scale.

Colour charge as a vector like quantity had been introduced to explain the existence of quantum states that are forbidden by Pauli’s law like the $\Delta^{++}$, having otherwise symmetric spin and flavour states, see e.g. [25]. Quarks and anti-quarks carry one colour or anti-colour each, while gluons carry a combination of colour and anti-colour. While impossible to observe directly, the number of colour charges has indirectly been confirmed in several experiments. Due to the nature of the charge, only colourless objects made up of several quarks can be observed. Mesons consist of one quark and one anti-quark, while baryons are formed by three quarks. While in some cases mesons are their own antiparticles, the antiparticles of baryons are made out of three anti-quarks. In a baryon such as a proton, the three main quarks are called valence quarks. These can emit gluons which in turn split into pairs of quarks and anti-quarks, called sea quarks. The forming of larger quarks
combinations, such as pentaquarks, which are combinations of four quarks and one anti-quark, have been researched for several years [26]. At the time of writing this thesis, an observation of a pentaquark candidate has been made [27].

Producing baryons and mesons from single quarks is done via fragmentation and hadronization: Radiated gluons create quark-anti-quark pairs which are built into colour-neutral objects. Since this process mostly happens in energy ranges where pertubation theory is not applicable, empiric hadronization models have to be applied. For instance, the Lund-String model [28] uses the strong quark-quark coupling to create strings of quark-anti-quark pairs, while the cluster-model [29] creates colour-neutral clusters using available quarks.

QCD corresponds to the SU(3) symmetry group, since the colour has the dimension three. The Lagrangian of the QCD can be written similar to that of the QED:

$$\mathcal{L}_{QCD} = \bar{\psi} (i \gamma^\mu D_\mu - m) \psi - \frac{1}{4} G^a_{\mu
u} G^{a\mu\nu}. \quad (1.8)$$

Like in QED, the covariant derivative that represents the kinematic energy of the fermion has to be replaced by the gauge invariant derivative

$$D_\mu = \partial_\mu - ig T^a G^a_\mu. \quad (1.9)$$

Here, $g$ stands for the colour charge and $T^a$ represents the generator of the SU(3) rotation. The index $a$ stands for the vector of colour eigenstates. Where in QED charge is a scalar quantity, in QCD colour charge has 8 different eigenstates.

The QCD field strength tensor is given by:

$$F^{a\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu + g f^{abc} G^b_\mu G^c_\nu. \quad (1.10)$$

This term describes the self-energy of the gluons in the colour-field. Compared to QED, additional terms appear that represent the self-interaction of gluons in addition to the quark-gluon coupling. The factor $f^{abc}$, which is called the structure constant of the SU(3) symmetry group, is non-zero and corresponds to non-commuting generators of the symmetry transformation. Symmetry groups that show this behaviour are called non-abelian, while groups with disappearing structure constants, such as QED, are called abelian groups.

Like in QED, gluon mass terms are not permitted since they are not gauge invariant.

**Weak Force**

The weak force causes radioactive beta decay. To explain the transition of a neutron into a proton while emitting an electron-neutrino pair, Fermi introduced the weak force as a contact interaction[30]. However, this theory was discarded, as it was not renormalizable.

In the standard model, the weak force is an interaction with the W boson as mediator particles. The W boson is a vector bosons with non-zero masses. A second vector boson, the Z boson, connects the weak force with the electromagnetic force, as it interacts both with neutral particles, such as the neutrino, but also mimics the photon by coupling to charged objects. The Z boson is massive and therefore the range of its interaction is also very small. The origin of the short range of the weak force lies in the short lifetime of the bosons.

In charged currents, an electrically charged W boson is emitted. It can be created either by emission or annihilation of a fermion with its isospin partner. Emitting a W boson will
change the isospin of a fermion, transforming a charged lepton into a neutrino or a quark into its isospin partners. This type of process is called a flavour changing current. In the neutral current, a neutral Z boson is exchanged either via pair-annihilation, comparable to the photon, or as a radiated particle. Z bosons leave the flavour of the interacting particle unchanged. W and Z bosons have a mean lifetime of \(<10^{-24}\) s.

At the early stages of particle physics, only three flavours of quarks were known: u, d and s quark, forming the SU(3) flavour symmetry. However, when the cross section of the process \(K \rightarrow \mu \mu\) was measured to be much smaller than expected, the physicists Sheldon Lee Glashow, John Iliopoulos and Luciano Maiani (GIM) predicted a fourth quark\([31]\). They argued that the reason for the small cross section was the destructive interference of the two different up-type quarks.

At the same time, the physicist Cabbibo worked on the unitarity of the weak force. To explain the coupling of up-type quarks to both down and strange quarks, he introduced a new set of quark eigenstates. They are defined as a superposition of the quark mass eigenstates. He introduced the flavour mixing matrix that explains the transition between up-type and down-type quarks. The Cabibbo angle \(\theta_c\)[32] represents the rotation of the down quark eigenstates of the weak interaction with regard to the mass eigenstates. The elements of the matrix represent the square root of the transition probability between the quark flavours.

Based on the GIM mechanism, the two physicists Kobayashi and Maskawa proclaimed a third generation of quarks in 1973 to explain the process of CP violation that had already been observed in the Kaon system\([33]\). They proved that only a three dimensional mixing matrix would possess enough degrees of freedom to mathematically allow CP violation. This new matrix, labelled CKM matrix, contains the transition probabilities of the three known quark families. While the lighter quarks represented by the SU(3) flavour symmetry have a higher chance of transition between quark families, the top quark decays into a bottom quark with a probability of nearly 100 percent.

There are several different ways of representing the CKM matrix. One of them shows the transformation of the quark mass eigenstates \(|d\rangle, |s\rangle\) and \(|b\rangle\) into the eigenstates of the weak force \(|d'\rangle, |s'\rangle\) and \(|b'\rangle\). These transformed eigenstates couple to the up-type quark eigenstates \(|u\rangle, |c\rangle\) and \(|t\rangle\), therefore allowing a transition between different quark mass eigenstates

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

(1.11)

The transition probabilities are given by the squared matrix entries\([3]\)

$$
(|V_{ij}|) = \begin{bmatrix}
0.97427 \pm 0.00015 & 0.22534 \pm 0.00065 & 0.00351 \pm 0.00015 \\
0.22520 \pm 0.00065 & 0.97344 \pm 0.00016 & 0.04124 \pm 0.00001 \\
0.00867 \pm 0.00029 & 0.04044 \pm 0.00011 & 0.999146 \pm 0.000046
\end{bmatrix}.
$$

(1.12)

Formulating the Lagrangian of the SU(2) leads to several problems that have to be solved in order to make the Lagrangian both globally and locally gauge invariant. The first problem is caused by the fermion mass terms. While these are also present in the Lagrangians of the QED and QCD, the Parity violating coupling of the W boson to left-handed particles requires setting the fermion mass \(m\) to zero, a conclusion that contradicts experimental results. Furthermore, while it is possible to introduce terms representing the field energy as in QED and QCD, the mediator particles would also have to be massless in order to
guarantee local gauge symmetry. Again, the measurement of a non-zero mass of both W and Z boson contradicts this conclusion. The existence of these mass terms is only possible by breaking the gauge symmetry, a method that is explained in the following section.

Electroweak Unification

The electroweak unification combines the symmetry groups of U(1) of QED and SU(2) of the weak force to form the \( U(1)_Y \times SU(2)_L \) symmetry group. The theories behind the electroweak unification have been developed by Glashow\[34\], Salam and Ward\[35\] and Weinberg\[36\]. The charge of the U(1) has been expanded to form the hypercharge \( Y = 2(Q - T_3) \), with \( T_3 \) being the third component of the weak isospin. However, the unification does not replace the original symmetry groups or interactions, but it is rather an alternative equivalent representation of both interactions. In addition to the combination of the symmetry groups, the coupling strengths of the two forces converge at energies greater than 100 GeV. At this point, both couplings can be replaced by a unified electroweak coupling.

The Lagrangian of the electroweak interaction contains two fields: The triplet of the weak isospin current \( J_\mu \) that is coupled to the three vector bosons \( W_\mu \) and the singlet of the hypercharge current together with its gauge boson \( B_\mu \). Corresponding to the value of the hypercharge, the QED current can be written as the combination of the hypercharge and weak current:

\[
j_{em}^\mu = J_3^\mu + \frac{1}{2} j_Y^\mu.
\]

While the physical W bosons represent the creation and annihilation operators of the weak isospin, the Z boson and the photon field A are represented as linear combination of \( B_\mu \) and \( W_3^\mu \). Therefore, the electromagnetic charge can be associated to the weak charge \( g \) and the hypercharge \( g' \) via the weak mixing angle \( \theta_W = 28.74^\circ \):

\[
e = g \sin \theta_W = g' \cos \theta_W.
\]

1.1.4 Higgs Mechanism and Symmetry Breaking

The information in the following section are taken from \[37\].

As stated, the measurement of non-zero masses for both fermions and W and Z bosons contradict the local gauge invariance of the SU(2). Implementing the mass terms in the Lagrangian of the weak force would result in a broken gauge symmetry. However, the gauge symmetry is necessary for the system to be renormalizable, which is a necessary property of the fundamental forces. The Higgs mechanism was introduced to solve these problems by means of spontaneous symmetry breaking.

In general, a physical system is described by expanding its potential around the ground state. In particle physics, this is done via perturbative calculations. This makes the choice of the minimum very important, since a perturbative calculation around a local, unstable minimum does not converge. By choosing a different ground state, additional terms are introduced. These can be interpreted as virtual particles, so-called Goldstone bosons. The choice of a specific parametrization of the field replaces these Goldstone bosons with new degrees of freedom in the system, such as mass.

In order to generate the needed mass terms in the SU(2), the Higgs field and its potential shown in figure 1.1 is introduced. It is given by the following Lagrangian:

\[
\mathcal{L} = (\partial_\mu \phi \bar{\phi}) (\partial^\mu \phi) - \mu^2 \phi^\dagger \phi - \lambda (\phi^\dagger \phi)^2.
\]
In this potential, $\phi$ represents a SU(2) doublet of scalar fields:

$$\phi = \begin{pmatrix} \phi_a \\ \phi_b \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}$$  \hspace{1cm} (1.16)$$

To ensure local gauge symmetry, the gauge invariant derivative is introduced:

$$D_\mu = \partial_\mu + ig \tau^a_2 W_\mu^a + ig' Y^a_2 B_\mu.$$  \hspace{1cm} (1.17)$$

This introduces the three gauge fields $W_\mu^a$. Adding the self energy terms of these fields results in the now gauge invariant Lagrangian

$$\mathcal{L} = (\partial_\mu \phi + ig \tau^a_2 W_\mu^a \phi)^\dagger (\partial^\mu \phi) \phi + \frac{1}{2} W_{\mu \nu} W^{\mu \nu} - V(\phi) - \frac{1}{4} W_{\mu \nu} W^{\mu \nu}.$$  \hspace{1cm} (1.18)$$

When choosing both $\mu^2$ and $\lambda$ greater zero, the aforementioned potential

$$V(\phi) = \mu^2 \phi^\dagger \phi - \lambda (\phi^\dagger \phi)^2,$$  \hspace{1cm} (1.19)$$

is symmetric around the local minimum defined by $\phi_1 = \phi_2 = \phi_3 = \phi_4 = 0$. However, in the case of $\mu^2 < 0$ and $\lambda > 0$, the form of the potential changes to the so-called sombrero-potential, depicted in figure 1.1. This potential has a second, global minimum that is defined by its equal distance to the center. By choosing any of the infinitely many possible minima as the new ground state, the symmetry of the system is broken. These ground states are characterized by the equation:

$$\phi^\dagger \phi = \frac{1}{2} (\phi_1^2 + \phi_2^2 + \phi_3^2 + \phi_4^2) = -\frac{\mu^2}{2\lambda}.$$  \hspace{1cm} (1.20)$$

A possible parametrisation is the following:

$$\phi_1 = \phi_2 = \phi_4 = 0, \phi_3^2 = -\frac{\mu^2}{\lambda} = v^2.$$  \hspace{1cm} (1.21)$$

This leads to a potential $\phi$ of the form

$$\phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}. \hspace{1cm} (1.22)$$
By shifting the ground state to its new minimum and choosing a certain parametrization, new terms appear that can be interpreted as the Higgs field of a scalar particle. Inserting this potential into the original Lagrangian results in four additional degrees of freedom. These terms can be associated with mass terms of the fields $W^\pm$ and $Z$ and that of an additional scalar boson. This result is important since it allows the existence of massive vector bosons while keeping the (now hidden) local gauge invariance.

**Massive Electroweak Theory**

By applying the Higgs methodology on the massless electroweak symmetry group of $U(1)_Y \times SU(2)_L$, mass terms are introduced that are associated with the two $W$ bosons and the $Z$ boson. In this case, while the parametrization of the potential breaks the symmetry of both $U(1)_Y$ and $SU(2)_L$, the $U(1)$ group that represents the QED remains unbroken. This explains why the photon remains massless even after the symmetry breaking.

The Higgs mechanism introduces mass terms for the heavy gauge bosons:

\[ M_W = \frac{1}{2} v g, \]  
\[ M_Z = \frac{1}{2} v \sqrt{g^2 + g'^2}. \]

Rewriting the mass terms as a function of the weak mixing angle $\theta_W$ gives the mass relation for $W$ and $Z$ boson: \[ \frac{M_W}{M_Z} = \cos \theta_W. \]

**Fermion Masses**

The symmetry breaking of the electroweak theory also allows the existence of massive fermions. The coupling between the Higgs boson and the fermion, called the Yukawa coupling, grants mass to the fermions. In this case, the strength of the coupling constant is proportional to the fermion mass. However, the coupling strength has to be measured separately for each fermion. This model is consistent with experimental results since the Yukawa coupling does not demand massive neutrinos.

In the case of the top quark, which is the heaviest object in the standard model, the coupling has its maximal value. However, while the positive sign of the coupling is the expected sign, the negative sign has not been completely excluded. This case would have significant effects on the associated production of a Higgs boson with a single top quark, since the different sign would lead to constructive interference in the calculation of the cross section.

**1.1.5 Cross Section Calculation**

The cross section $\sigma$ is an important process dependant quantity measurable in particle accelerator experiments. It describes the probability of a specific process based on the parameter of luminosity $L$, an accelerator specific quantity described in more detail in 2.1. The following formula shows the connection between cross section, luminosity and expected number of observable events:

\[ N_{\text{Events}} = \sigma \cdot L. \]

The cross section can be calculated with "Fermi’s golden rule", a formula derived by Dirac. It consists of the squared matrix element, which can be interpreted as a transition probability between two states, and the density of states of the final state. The former can be calculated using the theory of perturbation, given the coupling coefficient is sufficiently small. One possible way of calculation is via the methodology of path integrals introduced by Richard Feynman.
Feynman Graphs and Rules

Feynman's goal was to both calculate and visualize all possible contributions of a process. To do so, he developed a type of visual diagrams called Feynman graphs. They show the initial and final states together with the interactions involved. The external lines describe the initial and final-state particles and their respective four-momenta, the internal lines describe the four-momenta of virtual particles, so-called propagators. The connecting points, or vertices, represent points of interaction.

To calculate the contribution of each of these diagrams, Feynman also developed the Feynman rules for each elementary interaction. These rules contain the mathematical contributions for each initial state and final state particle, each vertex and propagator inside the diagram. To calculate the contribution, the so-created formula has to be integrated over all of the four-vectors of all participating particles. However, this is simplified by the additions of the conservations of the four-vectors for both initial and final state particles and at every vertex.

The number of vertices for a given process defines its "order" in perturbation theory. The diagram with the lowest number of vertices is called the leading order or Born level. Adding further vertices increases the order, for instance an additional initial or final state radiation increases the order from leading order to next-to-leading order. In theory, this would mean that only a calculation of every order can completely describe a process. However, if the coupling strength which is introduced as a factor with every new vertex is small, then contribution of the higher order diagrams also becomes small.

Factorization and Parton Distribution Functions

Factorization is the separation of a large, complex equation into a product of several factors that are multiplied. This has the advantage that each of the parts can be calculated on their own. In QCD, factorization is used to separate the hard scattering process of the partons that is described by the matrix element and the behaviour of the partons in the proton, which is modelled by the PDF. In both cases, it is possible for the parton to exchange energy by means of emitting or absorbing gluons: In the proton, there is quark-gluon-interaction, which is included in the parton distribution function (PDF), while the hard process includes emissions of gluons in the form of initial state radiation. To avoid double-counting between these processes, the factorization scale \( \mu_F^2 \) is introduced at which these processes are separated.

When colliding elementary particles such as electrons in a particle accelerator, the available energy for the process is the sum of the energy of both particles. In a proton accelerator however, the energy of the proton is divided between its constituents, the valence quarks, sea quarks and gluons. Because of this, the particles interacting in the collision only carry a fraction of the total proton energy. To calculate the energy of the interacting parton, the PDFs or parton distribution functions \( f \) are used. They represent the probability of finding a certain parton in a proton with a given momentum fraction. The integral over the total distribution is used as an input parameter to calculate the cross section \( \sigma(p_1, p_2) \) as a function of the parton momentum \( p_i \):

\[
s(p_1, p_2) = \int_0^1 dx_i \int_0^1 dx_j f_i(x_i, \mu_F^2) f_j(x_j, \mu_F^2) \hat{\sigma}(x_i p_1, x_j p_2).
\]

PDFs are created for both quarks and gluons. For quarks, it can be measured using data from different particle accelerator sources, such as the TEVATRON and the LHC or deep inelastic scattering at the HERA detector experiment. Gluon PDFs can be extracted
using jet data gathered in particle accelerator experiments. The DGLAP equations, which contain splitting functions to calculate the probability of a gluon splitting off a quark or gluon can be used to evolve the PDFs to different energies $Q^2$.

1.2 Top Quark

This thesis covers the production and decay of top quarks both as pair production and in association with a Higgs boson. While the top quark and its basic properties have been introduced in section 1.1.1, a more detailed overview of it will be given in section 1.2.1. Since the top quark has a very short lifetime, it has to be produced in particle accelerator experiments. The processes involving the production of top quarks, for instance at the LHC particle accelerator, will be discussed in section 1.2.1. Finally, a measurement of top particles involves studying the different decay channels of the top quark and their decay products, the details of which will be given in section 1.2.3.

1.2.1 Properties of the Top Quark

After the discovery of the bottom quarks at FERMILAB in 1977, the search for the top quark begun. However, it would take another 18 years and several new accelerators and detectors to finally observe a first candidate for topquark-antiquark pair-production at the DØ detector experiment at FERMILAB in 1992. Since then, the top quark has become an important tool in particle physics. At ATLAS and CMS the production of top quarks is involved in a multitude of analyses either as a signal or one of the main background processes. Especially the top mass is a very important parameter in Higgs and high-precision electroweak physics. Current studies involve, among others, the production of top quarks in combination with additional objects, for instance in association with a Higgs boson, a subject also covered in this thesis.

The top quark has a mass of $173.5 \text{ GeV/}c^2[3]$ which is roughly equivalent to that of a gold atom. It is the heaviest known elementary particle. Due to its high mass, the lifetime of the top quark is smaller than the time needed to begin the hadronization. Therefore it is not possible to find mesons or baryons involving top quarks which means it can only be found as a "bare quark".

1.2.2 Top Quark Production

At particle accelerators, top quarks can be produced either as single quarks or as a quark-antiquark-pair. Top-quark-pairs are produced in the $s$-channel via exchange of a photon or Z-boson or in the case of a proton collider via exchange of a gluon. Since the strong force has a stronger coupling to quarks than the electromagnetic force, the production via exchange of gluons is the dominant process. The proton splitting into the quark-antiquark-pair can be produced either by fusion of two other gluons or by annihilation of a quark and an antiquark. The TEVATRON accelerator at FERMILAB collided protons and antiprotons, which means that both quarks and antiquarks are present at the collision and therefore favours the process of pair-annihilation. At the LHC accelerator, which is a proton-proton-accelerator, anti-quarks only appear as sea quarks. This makes pair-annihilation unlikely, which is why gluon fusion is the dominant process at the LHC.

A single top quark can only be produced via the weak interaction between a bottom quark and a W boson. The respective channels for this process are the $s$- and $t$-channels and the associated W production. In the $s$-channel, a quark and antiquark annihilate to form a W boson which will decay into a top quark and bottom antiquark. In the $t$-channel, a quark will emit a W boson that will interact with a bottom quark, forming a top quark. Finally, in the process of associated W top production, a bottom quark emits a W boson and become a top quark.
1.2.3 Top Quark Decay Channels

As stated in section 1.1.3, the top quark dominantly decays into a b quark and a W boson. While the b quark will always decay hadronically, forming a jet, the W boson can decay either hadronically or leptonically. In the case of the leptonic W boson, both a lepton and neutrino are created, in the hadronic channel the quarks will hadronize and form two jets. In the case of a top pair decay, each of the W bosons can decay either leptonically or hadronically, resulting in three different decay channels in leading order. In the fully hadronic decay, both W bosons decay hadronically in addition to the two jets coming from the bottom quarks. This results in a total amount of six jets. Due to the large hadronic branching ratio of W bosons, this is the process with the highest cross section. However, compared to channels with at least one lepton, this channel contains a high rate of QCD multijet background events. The dileptonic channel contains two W bosons in the leptonic decay channel. The objects measurable in the event are two jets coming from bottom quarks, two leptons and the combined missing transverse energy from two neutrinos. The semileptonic or semihadronic channel features both a leptonically and hadronically decaying W boson. The event contains a total of four jets of which two are from b-quarks, a lepton and a neutrino.

The branching fraction of the different top-quark-pair decay channels are depicted in 1.2.

1.3 Higgs Boson

The Higgs boson is the latest member of the particle zoo in the standard model. While it had already been proclaimed in the theories of Brout, Higgs and Englert in 1964, it would
take another 48 years until the discovery of the Higgs boson was officially announced in July 2012. Before, other detector experiments at accelerators with lower energies had already put constraints on the mass of the Higgs boson without discovering it. The Higgs boson was introduced to explain the non-zero mass of the two heavy vector bosons of the weak force, the W and Z boson. By applying the model of spontaneous symmetry breaking that had been already used to explain the theory of superconductivity, Higgs et al. introduced the Higgs boson as a byproduct of the transformation of the electroweak field to a new minimum. Additionally, the theory of Yukawa coupling is used to explain the mass of fermions in the standard model. Further information about the Higgs boson origin and the mechanism of symmetry breaking are given in section 1.1.4.

As the Higgs boson is a spin 0 scalar boson, it has a different coupling than vector bosons and therefore different production and decay channels. Especially the analysis of the associated production of a Higgs boson and a top-quark pair is important due to the possibility of a direct measurement of the Yukawa coupling.

1.3.1 Higgs Boson Production

As the coupling of the Higgs boson to a particle is proportional to its mass, the it will more likely couple to heavy particles such as the W boson or the bottom and top quark. However, in a proton proton accelerator it is mostly light partons that interact. While these gluons and light quarks have high energies, their small masses results in a weak coupling and small cross sections for the production processes. However, there are still several processes that enable Higgs production. The production cross sections are presented in Fig. 1.3, while the corresponding Feynman graphs are shown in Fig. 1.4.

Figure 1.3: Higgs boson production cross sections as a function of Higgs mass. Taken from [40].

While gluons do not directly couple to the Higgs boson, they can still produce a Higgs in a virtual top quark loop via gluon fusion. While this quark loop can consist of any quark flavour, the top quark is the most likely since it has the highest mass and therefore also has the highest coupling coefficient.

Due to their high mass, the Higgs also couples strongly to the W and Z boson. Here, the Higgs can be produced in two different channels. The first is called vector boson fusion and involves the coupling of two W bosons to form a Higgs boson. The W bosons are here produced by weak interaction of the initial state quarks. The second channel that involves vector bosons is called the Higgs strahlung channel. Here, the Higgs is emitted by a W or Z boson that is created by quark antiquark annihilation.
Finally, the Higgs can also be produced in association with a topquark-antiquark pair. In this case, the Higgs is emitted from one of the two top quarks. This channel is especially interesting since it allows the direct measurement of the Yukawa coupling between the top quark and the Higgs boson. As stated before, in the gluon fusion process the quark loop can contain both the lighter quark flavours but also heavier flavours of additional quark families and other beyond standard model particles. In the case of the associated production however, the Higgs only couples to the top quark since it is emitted by it.

### 1.3.2 Higgs Decay

The branching fraction of the Higgs boson heavily depends on the Higgs mass. Since a higher mass allows the production of heavy particles such as W or Z bosons with a decreased virtuality, the cross section of a W boson or Z boson pair production increases for higher Higgs boson masses. At the measured Higgs mass, these processes are still possible, in this case however one of the two vector bosons is produced as a virtual particle.

![Higgs boson branching fractions as a function of mass](image)

The decay into pairs of quarks or leptons is also dependent on the Higgs mass. For a potential mass of more than 350 GeV/c², the production of two real top quarks becomes possible. At the current mass point however, this process is heavily suppressed due to the high virtuality of the top quarks. The decay into lighter quarks however is still possible. From figure 1.5, it can be seen that at the measured Higgs boson mass, the branching fraction of the bottom quark pair production amounts to about 57 percent. Bottom quarks also have the benefit of a high detection and reconstruction quality, which is why
this decay channel is used in this analysis. While not as high, the decay of lighter fermions, such as tau leptons or charm quarks are also used in other analyses.

Finally, the decay into a gluon or photon pair is possible, despite the fact that these vector bosons are both massless. In this case, the Higgs boson again couples to a quark loop which decays into either boson. The di-photon decay channel, while having a comparatively small branching ration, has been used in the Higgs discovery, since the high photon energy resolution allows for a precise measurement of the di-photon mass.
2. Experimental Setup at CERN

The goal of particle physics is to understand the fundamental building blocks and interactions between them. These particles are mostly short-lived and must therefore be artificially created, for example in high energy particle collider experiments. To create them in sufficient numbers, high amounts of stable particles are brought to velocities close to the speed of light and are eventually collided. Detector experiments are built around the points of collision, called vertices, to measure the particles and their decay products.

The Centre Européen pour la Recherche Nucléaire near Geneva in Switzerland is the location of a variety of particle accelerators and detector experiments, among them the Large Hadron Collider and the Compact Muon Solenoid experiment. These facilities are used to create and study the particle collisions that provide the data used in this thesis. The general setup of the LHC with the dedicated detector experiments is contained in section 2.1. This section also provides information about general terminology used to describe particle accelerators. The second part, section 2.2, will outline the components of the CMS experiment and their functions. It also contains a short overview of the data processing used at the CMS experiment.

2.1 Large Hadron Collider

The LHC [42] is a particle accelerator and collider located at CERN. Placed at a depth of 50-175 meters below the surface and with a circumference of 26.659 km it is the world’s largest particle accelerator. Its primary use is to accelerate protons, but it can also be used to accelerate different ions. The structure consists of two parallel beam pipes forming a near-circular shape combined into a single tube. The LHC was designed to fit into the cavern that was originally built to house LEP[43], an electron positron accelerator. Some of the technical limitations of the LHC have their origin in the design of the LEP accelerator.

At straight segments of the LHC, a total of 16 superconducting cavity resonators are used to accelerate the protons. They work at a frequency of 400 MHz and accelerate the protons with 5 MeV/m. At bent segments, a total of 1,232 superconducting dipole magnets is used to bend the particle beam. For a nominal energy of 7 TeV per beam, a nominal field of 8.33 Tesla is needed. An additional 392 quadrupole magnets in combination to hexapole and octapole magnets are used to stabilize the particle beam and focus it at the points of collision. The total number of magnets used in the accelerator amounts to over 9,500.

Due to an accident in the earliest stage of the first collisions, the nominal center-of-mass energy of 14 TeV was never used. After the subsequent repair phase, a beam energy of
3.5 TeV was used in the first half of the first run that lasted from 2010 to 2011. In 2012, a beam energy of 4 TeV was employed. The second run of data taking starting officially in June 2015 uses a near maximum beam energy of 6.5 TeV per beam. However, in the LHC the protons are not injected at rest. Rather, the protons need to be accelerated to an energy of 450 MeV using a chain of smaller particle accelerators. At the linear accelerator LINAC2 protons are produced at an energy of 50 MeV and brought into the Proton Synchrotron Booster, where they are accelerated up to 1.4 GeV. From there, they are injected into the Proton Synchrotron and later the Super Proton Synchrotron, which increase the energy to 26 and 450 GeV, respectively. Finally the protons are injected into the LHC over the course of several minutes and there accelerated to the nominal energy used in the collisions.

In the beam pipes, protons are not accelerated as a continuous stream, but as series of bunches. This is a result of the acceleration process that uses a radio-frequency scheme in which the acceleration is based on the frequency of the electric field in the cavities. In total, up to 2,808 bunches traverse the pipes at any given moment. The time difference between the arrival of pairs of colliding bunches, called bunch spacing, amounts to 50 ns in Run I. This bunch crossing was also used in the first weeks of data taking in 2015. For the remainder of Run II, a bunch crossing of 25 ns is being used. The bunches are typically of Gaussian shape and contain about $10^{11}$ protons or $10^8$ ions each.

The instantaneous luminosity is a quantity that characterizes a particle accelerator and determines the achieved event rate of a given process. It is proportional to the potential number of particle collisions per second at the intersection point. For the LHC, the instantaneous luminosity can be described by the following formula:

$$L = \frac{N_1 N_2 f N_b}{4\pi \sigma_x \sigma_y}$$  \hspace{1cm} (2.1)
Here, $N_1$ and $N_2$ are the number of particles in the two bunches while $f$ and $N_b$ correspond to the bunch frequency and number of bunches. The product $4\pi \sigma_x \sigma_y$ represents the cross section of the beams transverse to the beam axis. The goal of the accelerator design is to maximize the instantaneous luminosity. This can be achieved by either increasing the number of bunches, the number of particles per bunch or the bunch frequency or by a higher focus of the beam resulting in a smaller cross section.

To calculate the total amount of collisions in a given time frame, the luminosity has to be integrated over the time of the measurement:

$$L = \int_{t_1}^{t_2} L(t) dt$$

(2.2)

As shown before, the integrated luminosity can be used to calculate the expected number of events for a certain process given its cross section:

$$N = L \cdot \sigma$$

(2.3)

The instantaneous luminosity is at its maximum shortly after the injection of protons into the ring. Over the course of the fill, collisions and beam losses will decrease the luminosity. Finally, at the end of a cycle, or in the case of an emergency, the bunches are dumped into a target, which is a graphite block for the LHC.

Chapters 6 and 7 include the recorded data of the year 2012 which represent all events with a center of mass energy of $\sqrt{s} = 8 \text{ TeV}$. The total recorded luminosity over the course of the year is depicted in Fig. 2.2. This figure also contains the integrated luminosity per day.

Detector experiments are located at the intersection points of the beam axes to measure the collision of protons. At LHC, there are four detector experiments at different positions around the ring. The CMS[45] and ATLAS[46] detectors are general purpose detectors and can detect a wide variety of particles and energies. The ALICE[47] detector was built to study QCD under extreme conditions by colliding lead ions or protons. An additional goal is the creation of environments similar to the state of the early universe. Finally, the LHCb[48] experiment consists of an asymmetric detector and was built to do further studies concerning CP violation and rare processes with heavy quarks.
2.2 The CMS Experiment

The Compact Muon Selenoid experiment\[45\] is one of two general purpose detectors built for the LHC, meant to detect a wide variety of particles. It is positioned at point 5 of the LHC ring, at the opposite site of the ATLAS location. It has a mass of 14,000 tons, a length of 21 metres and a width and height of both 15 metres. The detector consists of a central, cylindrical part called the barrel and two endcaps that cap off the detector at both sides. Both barrel and endcaps have an onion-shaped design and consist of several layers of different detector types. The overall goal of the detector design is to avoid dead zones in which particles cannot be registered.

The innermost layer is a two-part silicon-based tracking detector and is installed directly around the beam axis. It consists of three layers of pixel detectors followed by layers of strip detectors. The tracking system is used to detect charged particles and to reconstruct their trajectories based on a combination of interaction points, called hits. The magnetic field created by the surrounding solenoid bends trajectory of passing charged particles. The curvature can then be used to calculate the momentum of the particle. In order to improve the energy resolution of objects such as jets, tracks detected in the tracking system are combined with calorimeter information in an algorithm called particle flow. This algorithm is explained in more detail in section 3.5.

Following the tracking detector are the two calorimeters of the detector, the electromagnetic and hadronic calorimeter. They consist of absorber materials that absorb particles by inducing particle showers and active materials that are used to detect the shower particles. The electromagnetic calorimeter is used primarily to detect electrons and photons via electromagnetic interactions. The light created by these interactions is measured using photodiodes and triodes. The hadronic calorimeter absorbs all heavier objects that interact via strong and electromagnetic interactions. It is a sampling calorimeter that uses alternating layers of absorbing and active materials.

The two calorimeters are surrounded by a superconducting solenoid magnet. This solenoid creates a magnetic field of 3.8 Tesla on the inside. The strong magnetic field bends the
trajectories of charged particles based on their momentum, which is used in the aforementioned tracking system. Outside the solenoid is the outer hadronic calorimeter which is used to measure the energy of particles that have not yet been completely absorbed.

Following the solenoid and the outer calorimeter is the steel yoke. It is used to confine the magnetic field on its inside and also to absorb particles that are unwanted in the outer parts of the detector. Additionally, the gaps in the yoke contain the muon chambers. These drift chambers are used to detect and measure muons, since they only deposit small amounts of energy in the calorimeter system.

2.2.1 Detector Geometry and Variables

The CMS detector uses a cartesian coordinate system which is centered around the collision point inside the detector. The $z$-axis is the longitudinal axis parallel to the beam with its positive side pointing to the west, representing the counter-clockwise beam. The $x$-axis points towards the center of the LHC. Finally, the $y$-axis points upwards, the three axes forming a right-handed coordinate system. In order to better accommodate the cylindrical shape of the detector, spherical coordinates are used for spatial parameters.

Even though the LHC collides protons with balanced energies, the interacting partons have an unknown fraction of the proton momentum, which leads to an unknown effective longitudinal momentum of the center of the proton-proton system. While the transverse component of the momentum can be measured with the help of the magnetic field, there is no magnetic field in longitudinal direction. This complicates the calculation of the longitudinal momentum further. Therefore, the longitudinal part of the momentum is often disregarded, leaving only its transverse component:

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

(2.4)

The azimuth angle $\phi$ is unchanged, since the detector is symmetric with regard to the longitudinal axis. The polar angle $\theta$ however is replaced by the rapidity:

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

(2.5)

Here, $E$ and $p_z$ represent the energy and the longitudinal momentum of the particle, respectively. Using the rapidity in favour of the angle $\theta$ has several reasons. First, the particle production as a function of the rapidity is approximately constant, which means that distributions of the rapidity are more evenly distributed than those of the angle $\theta$. Further, differences in rapidity are Lorentz invariant, making a measurement of $\Delta y$ independent of the frame, unlike a measurement of $\Delta \theta$.

In particle accelerator experiments, the pseudo-rapidity is often used to approximate the rapidity, as they converge in the limit of massless particles. The pseudo-rapidity is given by the formula

$$\eta = -\ln \left( \cos \frac{\theta}{2} \right).$$  

(2.6)

and is often preferred to the rapidity, as it is only a function of $\theta$. 
To calculate the spatial distance between two objects, the parameter

\[
\Delta R = \sqrt{(\Delta \eta)^2 + (\Delta \phi)^2}
\]  

is used. The value \(\Delta R\) also corresponds to the opening angle of a cone in the spherical coordinate system of the detector.

### 2.2.2 Tracking System

The tracking system is a semiconductor-based detector located closest to the beam axis. It consists of several layers of pixel and strip modules. Passing charged particles create hits in these modules on which tracks are fitted which represent the trajectory of the particles inside the tracking system. Based on the high granularity of the system, the path of the particles can be resolved with high precision even in the case of large particle densities.

The pixels and strips in the track system are based on semiconductor p-n-junctions. In these junctions, the combination of p- and n-doped parts leads to the recombination of free charge carriers. This creates a depletion region with a minimum of free charge carriers, resulting in a very small electric conductivity. The depletion region can be further increased by applying a reverse bias that further pairs electrons with holes. By interacting with high-energy charged particles, electrons will be excited from the valence to the conduction band, creating pairs of electrons and holes. These pairs of electrons and holes will drift towards electrodes in the E field, which creates a small electric current inside the material that is measured by the electronics.

The accurate track reconstruction allows to measure whether a particle originated from the primary vertex of the event or an alternative, secondary vertex. This information can be used in the process of b-tagging and is explained in more detail in 3.5.

The choice of silicon was made since it was both cheaply available and proved to be more radiation hard than alternative materials.

### Pixel Detector

The inner part of the track system consist of pixel modules. The structure of the pixel system can be seen in Fig. 2.5. In the barrel region, the pixel cells are arranged in three layers of modules, each module carrying 66,000 pixels. The layers themselves have a length...
Figure 2.5: Components of the pixel tracking detector at the CMS experiment. While the barrel region consists of three layers of modules, the endcaps are built to form turbine-like constructions to guarantee an overlap between the modules. Taken from [51].

of 53 cm and are positioned at a radial distance of 4.4 cm, 7.3 cm and 10.2 cm from the beam axis. In total, there are 768 modules in this region of the pixel detector.

The pixels have a size of 100 µm × 150 µm. This results both in a high granularity and a high spatial resolution. The small thickness of these pixels further guarantees small energy distortions from effects as multiple scatterings and shower processes. The effective hit position resolution amounts to 20 µm in longitudinal and 10 µm in transverse direction.

The endcaps of the pixel detector consist of two discs of pixel modules at either side of the barrel. Each disc contains 24 modules which in turn are controlled by 7 readout chips. The modules themselves are arranged to form a turbine-like geometry while overlapping each other. They are further turned by 20 degrees around their radial symmetry axis.

**Strip Detector**

The silicon strip tracker is the second part of the tracker system. Following the pixel detector, it is placed between 22 cm and 60 cm in radial direction. Its barrel region is split radially into two sections. The first section contains the Tracker Inner Barrel (TIB) and the Tracker Inner Disks (TID). The outer section is called the Tracker Outer Barrel (TOB). At both sides, the strip tracking system is capped off by the Tracker End Caps (TEC).

The four cylindrical segments of the TIB are placed equidistantly and contain rectangular modules. Rings one and two are double-sided and have modules with a strip pitch of 80 µm, layers three and four consist of single-sided modules with a strip pitch of 120 µm. The aforementioned TIDs replace the outer part of the innermost cylinder, forming a mini end-cap. Tracks with |η| < 1.2 are fully contained in the cylinder segments, tracks with 1.2 < |η| < 2.0 cross both cylinder and end-cap segments. The modules themselves have an area of 64 mm × 64 mm of which 62.5 mm × 62.5 mm are active. The number of channels per module increases from the outside, going from 512 with rings four and five up to 1024 with ring one. To be able to measure the z-direction, adjoining modules are tilted by 100 mrad with respect to the beam line. They further overlap to avoid dead regions.

The TIDs consist of three rings in radial direction and contain 36 modules each. The two inner rings host back-to-back modules while the outer ring is only single-sided. As
with the barrel cylinders, the modules in the TIDs overlap. The three rings are held by a supporting structure that is positioned between the rings. The modules each contain two wedge-shaped detectors, of which one is placed at either side of the module. While the radial strips on the front side point towards the beam axis, they are again tilted by 100 mrad on the back side to allow for a longitudinal measurement.

The TOB consists of a single structure (wheel) which supports 688 sub-modules, called rods. The structure is separated into four disks with a total of 334 holes. The aforementioned rods holds either six or twelve detector modules which are placed at either side of the rod.

The TECs of the strip detector consist of nine parallel rings at either side, covering an $\eta$ range of $1.2 < \eta < 2.5$. To allow for an easier construction of the silicon strip modules in this section of the detector, they have been arranged in autonomous structures called petals. Each ring holds 16 petals, of which 8 are placed on either side of each ring. To allow a overlap in $\phi$ direction, strip detectors are placed alternately on side side of each ring. Since only a coverage up to $|\eta|=2.5$ is necessary, the rings 4 to 9 lack some of the inner modules of each petal.

The design of the strip detectors allows for an $r-\phi$ measurement with a resolution between 23 $\mu$m and 34 $\mu$m and a resolution of 230 $\mu$m in the $r-z$ direction.

### 2.2.3 Electromagnetic Calorimeter

![Cut-away view of the electromagnetic calorimeter used in the CMS experiment.](image)

The electromagnetic calorimeter (ECAL)\cite{52} is a homogeneous calorimeter built out of scintillator material and is primarily used to detect electrons, positrons and photons. It consists of 75,848 lead tungstate crystals (PbWO$_4$) of which 61200 are placed in the barrel region and 7324 in the two endcaps. A view of the ECAL is given in 2.6. Particles that pass the crystals will interact electromagnetically, emitting additional particles. While electrons and positrons emit photons via the process of bremsstrahlung, photons decay into pairs of electrons and positrons. This process, called "showering", distributes the total energy among the particles and continues until their energy falls below a critical value, after which they are absorbed in the material. The shower particles that pass the scintillator crystal will excite electrons of the crystal, which then relax and emit light of a certain wavelength. The quantity of emitted light gives information about the number of particles in the shower and thus about the total shower energy.
To quantify the interacting qualities of a material, the material specific radiation length $X_0$ is introduced. It is both defined as the length after which the energy of an electron decreases to $1/e$ of its original value and as $79$ of the typical photon conversion length [3]. In order to confine the shower in the calorimeter, it is therefore necessary to choose a material with a small radiation length. A second quantity, the Molière radius $R_m$[53] quantifies the transverse expansion of a shower.

Lead tungstate is an efficient detector material[54], having a short radiation length $X_0 = 0.89$ cm and a small Molière radius $R_m = 2.19$ cm. The good transverse resolution based on the small Molière radius helps separating close showers, while the small interaction length limits the total length of a shower. This is especially important since the solenoid limits the space available for the calorimeters. Further aspects of lead tungstate are the fast response, with an emission of 80% of the total light in 25 ns and its radiation hardness of up to 10 Mrad. A disadvantage of the material is the low photon yield of the scintillation. However, this problem can be solved by using electronic read-out with intrinsic gain properties. Additionally, the photon yield has a strong dependency on the temperature of the system. It is therefore necessary to keep the crystals at a stable temperature of $(18 \pm 0.05)^\circ$C.

The crystals in the ECAL have a geometrical range up to $|\eta| = 3.0$ and also separate into the barrel and the endcap region.

In the barrel region ($0 < |\eta| < 1.479$), crystals with a length of 23 cm are used. The length is equivalent to $25.7X_0$, the estimated distance after which most high-energy objects have deposited their complete energy in the material. The singular crystals have a front area of $22 \text{ mm} \times 22 \text{ mm}$ which corresponds to a transverse granularity of $\Delta \eta \cdot \Delta \phi = 0.0175 \times 0.0175$.

In the endcaps ($1.479 < |\eta| < 3.0$), the front area increases to $24.7 \text{ mm} \times 24.7 \text{ mm}$, while the granularity increases up to $\Delta \eta \cdot \Delta \phi = 0.05 \times 0.05$. The crystals are ordered into superstructures that are placed in semicircular frames. A digital model of these superstructures is shown in 2.7. In this area, crystals with a length of 22 cm are used. To guarantee a complete energy resolution despite shorter length, a lead-based pre-showering is implemented.

2.2.4 Hadronic Calorimeter

The hadronic calorimeter (HCAL)[56] is a sampling calorimeter that is used to detect strongly interacting objects that are not absorbed in the electromagnetic calorimeter. A
2. Experimental Setup at CERN

Figure 2.8: Sliced view of the hadronic calorimeter. The HCAL consists of the Hadronic Barrel, Hadronic Endcap, Hadronic Outer and Hadronic Forward. Taken from [45].

sampling calorimeter uses alternating layers of absorbing material and an active material. High energy hadrons interact with the nucleons of the absorber material via the strong interaction. This process creates a shower of hadrons and other light particles that disperse inside the material. These particles are observed in the material by ionization of absorber atoms. The total number of ionized atoms is proportional to the energy of the primary particle. In order to measure the number of ionized atoms, the light that is emitted by the ionized atoms is measured in the active material.

Similar to the radiation length in the electromagnetic calorimeter, the nuclear interaction length $\lambda_I$ is a material dependant parameter that describes the interaction probability between a hadron and nucleon. On average, the energy of an interacting object inside the material has decreased to $\frac{1}{e}$ after an interaction length. In order to measure the complete energy of passing hadrons, the interaction length has to be small to allow for a high amount of interactions. Brass has been chosen as the primary absorber material, since it has an interaction length of only 16 cm. Further, it is non-magnetic, making it insusceptible to the strong magnetic field created by the solenoid. While the deposited energy in the HCAL comes both from hadronic and EM showering, the calorimeter response is not equal for these processes. Therefore, a precise balancing of the thickness of the active medium is necessary to allow for a similar response and therefore an exact energy measurement[57].

In the case of the hadronic calorimeter, the active material consists of thin layers of plastic scintillators. Wavelength shifters bring the emitted light into the optical spectrum. Finally, it is absorbed using multi channel hybrid photo diodes with 19 channels each.

Like the electromagnetic calorimeter, the hadronic calorimeter is split into a barrel ($0 < \vert \eta \vert < 1.3$) region and the two endcaps ($1.3 < \vert \eta \vert < 3.0$). The barrel consists of 17 layers of absorber material which are equivalent to $5.82 \lambda_I$ at $\eta = 0$ and $10.6 \lambda_I$ at $\vert \eta \vert = 1.3$. Of these, 15 are made out of brass with a thickness of 5 cm while the inner- and outermost layers consist of stainless steel. The gaps are filled with layers of scintillators with a thickness of 9 mm for the first and 3.7 mm for all following layers. The barrel is segmented into 2304 towers which results in a granularity of $\Delta \eta \times \Delta \phi = 0.087 \times 0.087$.

The endcaps also consist of alternating layers of absorber material and scintillators. In total there are 19 layers, divided into 17 layers of brass with two encapsulating steel plates. The total length of all absorbing layers corresponds to 10 hadronic interaction lengths.

The outer hadronic calorimeter or Hadronic Outer (HO) is an additional part of the calorimeter behind the solenoid coil used to absorb the energy residues of high energy
hadrons. This extension of the barrel calorimeter is necessary since a parton shower in the central $\eta$ region may not deposit the complete energy. The HO is positioned between the solenoid and the muon system and consists mainly of two absorber plates made of scintillator plates with a thickness of 30 cm. They are placed on the inside and outside of the first steel yoke layer.

To measure hadrons with even the highest energies, the forward hadronic calorimeter or Hadronic Forward is used. It covers the range of $3.0 < |\eta| < 5.0$. This calorimeter has two main purposes: To further improve the measurement of missing transverse energy and to reconstruct very forward jets. The materials used in the forward calorimeter are chosen to be very radiation hard, since the particle flux in this region is extremely high ($6.0 \times 10^6$ cm$^{-2}$ s$^{-1}$ which is equivalent to 100 Mrad per year).

The forward calorimeter measures shower particles via the Cerenkov effect. A charged particle that passes a material with a velocity higher than the speed of light in the material will emit electromagnetic radiation, called Cerenkov radiation. The energy of the radiation depends on parameters like the particle energy, particle trajectory and incident angle. To measure this radiation, quartz fibres are used which are embedded in a block of copper. The usage of a Cerenkov detector has both advantages and disadvantages compared to an absorber calorimeter. The showers that develop in the fibre detector are both narrow and short. Further, the Cerenkov effect can be assumed to be almost instantaneous, improving the detection time of showers. However, the Cerenkov detector cannot detect neutral particles such as neutrons, or particles below the $\beta$ threshold of the material.

### 2.2.5 Magnet Yoke and Muon System

![Figure 2.9: Illustration of the muon system based on rapidity. The depicted components are the Drift Tubes, Cathode Strip Chambers and Resistive Plate Chambers. Taken from [58].](image)

The outer compartments of the detector consist of the yoke and the muon system. The yoke is made out of steel and used to confine the magnetic field created by the solenoid. This results in an opposite field direction inside the yoke with a strength of 2 Tesla.

While the electromagnetic calorimeter can be used very efficiently to detect both electrons and photons, it is not efficient at detecting muons. As explained before, calorimeters measure energy by absorbing the particles created in showers. The amount of absorbed energy per distance in the absorber material can be calculated using the Bethe-Bloch formula. Muons deposit only small fractions of their energy in the absorber materials,
making them "minimal ionising particles". The weak absorption of muons is primarily based on their high masses, which is about 200 times larger than that of electrons. Since the muons will not be absorbed in the electromagnetic and hadronic calorimeter, the CMS detector needs a special muon system.

The muon chamber[59] is the outermost part of the CMS detector. It is divided into a barrel and endcap region, each using their own means of muon detection. In the barrel, spanning from $0.0 < |\eta| < 1.3$, drift tubes are implemented, while in the endcap region, going from $0.9 < |\eta| < 2.4$, cathodic strip chambers are used. Further, resistive plate chambers are installed in both barrel and endcap, going from $0.9 < |\eta| < 2.1$.

**Drift Tubes and Barrel Geometry**

The barrel area of the muon system is characterized by a magnetic field of 2 T that is dampened by the surrounding iron yoke. Further, there is a small expected muon rate and a small expected neutron background. These factors benefit the usage of drift tubes. Drift tubes consist of a sealed, gas-filled volume with an anode wire spanned in the centre. Charged particles passing the tube will ionize the gas atoms, a high electric field then separates and amplifies the charges, creating a measurable current. The position of the passing charge can be calculated from the drift time. In this case, the tubes are filled with a mixture of Ar and CO$_2$. The choice of geometry and the hull width of 2 mm is used to partially decouple adjoining tubes.

The complete barrel region is separated in five rings along the longitudinal direction, that have a length of 2.5 m. The rings are further split into twelve sectors, each monitoring 30° of the total $\phi$-space. A further separation exists in radial direction. Here, the rings are separated in four stations, of which two are inside the iron, while the other two are at the outer and inner face of the yoke. The four stations consist of a total of 250 chambers.

Inside each chamber, the twelve planes of drift tubes are segmented into three super layers. Each super layer consists of three layers with parallel wires, of which two are used to measure the $\phi$- and one to measure the longitudinal coordinate. This setup allows for a resolution of 200 $\mu$m.

**Cathode Strip Chambers**

Since the endcaps have both a significantly higher magnetic field and a higher rate of neutrons and muons, a different type of drift chambers is used. CSCs detect charged particles with the drift current mechanism that is used in the drift tubes. They consist of a wedge-shaped geometry with six layers of anode wires and cathode panels each, positioned perpendicularly towards each other. The layers of wires create measurements in the $r$-direction of the object, while the panel strips create measurements in the $\phi$-coordinate. The width of the strips ranges from 3-16 mm. While the CSCs have a very high resolution in $\phi$, they are much coarser in radial direction.

The layout of the two endcaps in the muon system consist of four muon stations of trapezoidal shape. The stations are arranged in concentric rings around the beam and separated by the massive iron parts of the return yoke. They are divided in subrings made of the aforementioned chambers that also overlap in $\phi$ to prevent dead areas. The rings either separate into 36 chambers at the outer radius and 18 chambers at the inner radius.

While the first two chambers (ME1/1 and ME1/2) operate in a significant magnetic field, the fields surrounding the other rings are much smaller. This results in a large bending of the muon trajectory followed by a reversal of the curvature of the muon trajectory after the muons have passed the yoke.
Resistive Plate Chambers

RPCs are the third type of drift chambers and are placed complementary both in the barrel region and the endcaps. While they have a good level of spatial resolution, their high time resolution makes them very effective in the muon trigger system. With the trigger signals from all three types of drift chambers arriving at the trigger logic in parallel, the RPC can be used to improve the trigger efficiency and to provide redundant information for efficiency tests.

The resistive plate chambers consist of two Bakelite plates with a gap of several millimetres, forming a gas-tight container. Bakelite has been chosen due to its high bulk resistivity of $10^{10} - 10^{11} \Omega \text{cm}$. While the outer surfaces are coated with conductive graphite as a means of grounding, the cell is filled with a mixture of mostly tetraflourethane ($C_2 H_2 F_4$). A total number of 610 of these modules are installed in both barrel and endcap.

2.2.6 Data Taking and Trigger

The protons that travel separated in up to 2808 bunches at nearly the speed of light with a distance of 50 ns (25 ns) result in 40 million bunch crossings per second. The high data rate created by this large number of collisions is impossible to save on any type of current storage system. It is thus necessary to preselect the events using data that is acquired in the collision process. This so-called triggering[60] is separated in several levels with separate computing frequencies.

The first step is called Level 1 Trigger and uses low level information gathered in the calorimeter and muon system, for instance the measured transverse momenta or energy in the subsystem. These information are gathered using systems with lower granularity and resolution to accelerate the evaluation process. During the triggering, the full event information are stored in pipelined memory. With an evaluation time of about 3.2 $\mu s$, the effective trigger frequency equals to about 100 kHz.

In the second triggering phase, called High Level Trigger, the event frequency is further reduced to the order of 400 Hz. The data is transported to front-end readout buffers that can temporarily save the event while at the same time enable partial reconstruction of more complex parts of the event. Based on this information, the second line of triggers checks for thresholds in parameters such as the amount and quality of jets, muons or electrons. In the case of triggers with a low threshold, the rate would still be too high for the computers to write all events. Therefore, a prescaling is installed, which only selects a fraction of the events that activate the trigger. Finally, the selected events are saved on the disks at the available computing centres.
In particle collisions, many short-lived objects are created and decay. They are detected and their kinematic properties such as the energy are measured in the several layers of the CMS detector, as explained in Chapter 2. The algorithms to reconstruct higher-level objects such as jets from these are described in this chapter.

The reconstruction of the charged particle trajectories in the tracking system and the muon chamber is explained in section 3.1. The tracks can be further used to calculate the position of the primary vertex, which is explained in section 3.1.1.

Electrons and photons are reconstructed using primarily information coming from the electromagnetic calorimeter. Further details are given in section 3.2. Due to their nature as minimally ionizing particles muons have a separate, track based reconstruction which combines information from the tracking system and the muon system. Further information are given in section 3.3.

Due to the disappearing transverse momentum of the protons, a net zero transverse energy in the event is expected. However, due to the appearance of neutrinos and certain detector effects, the negative vector sum of the transverse energies, called missing transverse energy (MET), is often non-zero. The details of the calculation and the importance of a correct calibration of the MET are presented in section 3.4.

The particle flow algorithm combines information from the tracking system and the two calorimeters to particle flow candidates, therefore profiting from the benefits of both systems and thus improving the reconstruction performance. The particle flow algorithm is explained in further detail in section 3.5.

The so-created particle flow candidates are used in the following to cluster jets. Jets represent the decay objects of high-energy partons and are reconstructed using a variety of algorithms which are presented in section 3.6.

As explained in chapter 2, bottom quarks have an increased importance in many analyses. It is therefore interesting to identify jets coming from bottom quarks using the process of b-tagging. While there is a wide variety of b-tag algorithms, many exploit the long lifetime of b-hadrons. The algorithm used in this thesis constructs the secondary vertex of the bottom quark based on the tracks in the tracking system.
The reconstruction of final states with many hadronically decaying objects requires a new approach to reconstructing jets. Instead of many smaller jets, few, larger jets are clustered which are then separated into subjets. The methods of clustering and reclustering these large jets is done via two subjet algorithms which are presented in section 3.8.

### 3.1 Track and Primary Vertex Reconstruction

The method of track reconstruction is used both in the inner tracking system and the muon system and is explained in more detail in [61]. While the hits coming from the tracking system are used to reconstruct tracks of all charged particles passing the detector, the muon system is used to primarily reconstruct tracks created by muons. Tracks can be reconstructed starting both at the inside and the outside of the detector. The combination of both tracker tracks and muon system tracks allows for a precise muon reconstruction. The information from the tracks is further used when defining quality criteria for leptons and jets, compare section 5.4. As stated before, the particle momentum can be measured using the tracks curvature.

Several neighboring pixels or strips in the detector are combined to form a pixel or strip hit. This step is called hit recognition and is used to measure the particles position with high precision. In the next step, the track seeds are created. These serve as starting points for the tracks and are formed by two collateral hits. These seeds are the candidates for potential tracks and contain preliminary information coming from the containing hits. In the next step, further hits are added to the track based on the Kalman filter algorithms[62]. Here, further parameters like energy loss of the particle due to material scattering are included. This tightens the parameter set of the starting seed that represents the parameters of the track that is based on the seed.

In a final step, double counting in tracks is removed. This includes cases like two tracks from different seeds that represent the same particle or two different tracks coming from a single seed. Here, the tracks are chosen by measuring the fraction of shared hits between the tracks, called the ambiguity resolution. If this value exceeds a threshold value, the tracks with the smaller amount of hits is discarded. In a final step, the tracks are fitted again to calculate the final set of track parameters and to smoothen out the track. The Kalman method[62] is used twice, both starting from the inner end of the muon chamber and once starting from the outer end, going through all layers of the track system. While the first fit is used to reduce the bias introduced by the seed finding algorithm, the second fit is used to smoothen the trajectory, forming the final form of the track.

The muon system is able to reconstruct tracks from passing charged particles just as the track system. However, due to the calorimeters and the surrounding iron yoke, only muons reach the muon chambers. In general, the track reconstruction in the muon chamber is comparable to that in the track system. While the hits in the track system mostly measure the position of the passing particle, the modules in the muon chamber measure both position and direction. Both parameters are later used in the track fitting. Again, two Kalman fits are used to create the final track. The first fit starts from the inner end of the muon chamber using a generated seed. The second and final fit starts from the outer end of the detector going inside. It is later extrapolated towards the primary vertex. The position of the vertex is also used as an input parameter for a refit. Muon tracks that are generated this way are called standalone muons. Further information about muon tracks is given in section 3.3.

#### 3.1.1 Primary Vertex Reconstruction

The primary vertex is the collision point of the two interacting protons. Its exact position is used to separate the particles in the event from those coming from pileup. It is further
3.2 Electron and Photon Reconstruction

Photons and electrons deposit most of their energy in the electromagnetic calorimeter via showering. While electrons shower via bremsstrahlung, photons create pairs of electrons and positrons. Using the aforementioned particle flow algorithm, electrons can be reconstructed using both the calorimeter energy and their tracks, while photons are only measured through their energy deposits in the ECAL.

Electrons are reconstructed using information coming from the tracker system and the electromagnetic calorimeter [63]. In general, electrons deposit their complete energy in the ECAL by showering. Since these showers are compact, the majority of the electron energy is stored in a compact area of ECAL crystals. However, electrons already emit energy via bremsstrahlung before entering the calorimeter. These emitted photons are radiated in a wide $\phi$ range due to the bending of the electron in the magnetic field. To reconstruct the complete electron with its full energy, this photon energy has to be included. Supercluster algorithms scan the ECAL in $\phi$ direction for small energy deposits in adjoining crystals, starting from isolated energy clusters. The ECAL information is further combined with the electrons tracker track to form a particle flow electron. Based on the geometry and resolution of the CMS detector, the reconstruction of electrons is limited on $|\eta| < 2.5$ and $p_T > 5\text{ GeV}/c$.

It is further important to notice that both electrons and photons can already interact with the silicon in the tracking system. The whole length of the tracking system represents between 0.4 and 1.8 interaction lengths depending on $\eta$, which leads to a 20-60% chance of photon conversion due to pair-production and an average energy loss in the order of GeV of electrons due to bremsstrahlung effects.

The photon reconstruction is similar to that of electrons [64]. Photons create pairs of leptons which in turn also emit new photons and other tertiary objects. Due to the magnetic field, they are also widely spread in $\phi$. Therefore a similar supercluster procedure is applied that searches for energy deposits in the ECAL based on a primary energy seed. A second challenge is the high number of photons created as decay products. However, these decay photons can be suppressed by applying an isolation filter on the prompt photon.

3.3 Muon Reconstruction

As explained in section 3.1, muons are reconstructed in two different ways[65] using tracks in the tracking system and the muon chamber. These tracks are combined using a global fit which can start either at the inside or the outside of the detector.

The first muon variant is called the global muon. Here, the standalone muon that is constructed in the muon system is extrapolated into the track system. If a matching track can be found, a global muon track is fitted including the aforementioned Kalman filtering. Combining tracks from both sides of the detector greatly improves muon resolution compared to using only a single track, especially in the case of muons with high transverse momenta.
3. Object Reconstruction at the CMS Experiment

The second muon type is called a tracker muon. In this case, muon candidates are chosen among the tracker tracks that have a transverse momentum of at least 0.5 GeV/c. The tracks are extrapolated towards the muon chamber and are further corrected for deviation by magnetic field and scattering in the material. Finally, the track is matched to a hit in a muon segment. By regarding only a single segment instead of a whole muon track, this muon reconstruction allows for a high efficiency for muons with low momenta.

In the analysis, global muons are required, compare 5.4.1.

3.4 Missing Transverse Energy

In the LHC, protons are collided at certain center of mass energies, creating a multitude of particles in the process. While the absolute longitudinal momentum is unknown due to the Bjorken factor, the total transverse momentum in the center of mass system is nearly zero, since the incoming protons have very little transverse momentum. Therefore, due to four-momentum conservation, the total transverse energy in the event should be also zero. However, there are still effects that cause a nonzero transverse energy. First, neutrinos are not measured directly in the detector due to their very weak interaction with detector material. This means that all processes involving leptonically decaying W bosons automatically have missing transverse energy. Another source can be energy coming from processes other than the primary process, for instance pileup. Since the detector is not built hermetically, particles can go unnoticed if they are in the dead zones of the detector parts. Finally, the limited resolution of the calorimeters, especially the hadronic calorimeter, can also lead to missing energy.

The total missing transverse energy, or $E_T$, is defined as the negative sum of the transverse momenta of all particle flow candidates in the event[66]:

$$E_T = - \sum p_{T,PF}.$$  \hspace{1cm} (3.1)

Using particle flow candidates for the calculation instead of calorimeter information only greatly improves the resolution of the missing energy.

3.5 Particle Flow Algorithm

In the CMS experiment, the information coming from the different subsystems is used to calculate energy and momentum of charged and neutral particles. However, the subsystems show different performance depending on the energy of the particle that is measured.

In the calorimeters, the uncertainty $\sigma_E$ of the measured energy is given by the following formula:

$$\left(\frac{\sigma_E}{E}\right)^2 = \left(\frac{N}{E}\right)^2 + \left(\frac{S}{\sqrt{E}}\right)^2 + C^2.$$  \hspace{1cm} (3.2)

The uncertainty is divided into systematic uncertainties $C$, stochastic effects $S$ and noise effects $N$. As most of these uncertainties scale inversely with the energy of the object, the calorimeter is especially precise when measuring high energy particles. Measuring particles by calorimeter interaction is possible for both charged and neutral particles.

In the case of the track reconstruction in the tracking system, a different formula applies:

$$\Delta p_T \propto \frac{1}{p_T} \Delta \frac{1}{R} \propto \frac{1}{p_T} \frac{\Delta R}{R^2} \propto p_T \Delta R,$$  \hspace{1cm} (3.3)
where $\Delta R$ is determined by the position resolution.

In this case, the errors are smaller if the $p_T$ of the object is also small, as this increases the curvature of the track in the tracking system and makes a track reconstruction easier. However, this type of reconstruction is only applicable for charged particles.

The particle flow algorithm was introduced to combine the benefits of both systems to particle flow objects.

The algorithm begins by forming energy clusters from the deposited energy. These clusters are started at seeds that represent local maxima of energy deposition. From there, all energy from neighbouring cells is added. These particle flow clusters are then linked to tracks.

In the linking phase, the tracks that are fitted in the tracking system are extrapolated to the calorimeter system and combined with the calorimeter clusters and preshower hits. This creates particle flow candidates, which are then identified based on the information from the tracking system and the calorimeter. After the reconstruction of an object is finished, all tracks and calorimeter towers that are associated with it are removed from the event.

For particle flow muons, global muons are used. They are identified using the combined tracks from the muon and tracking system and the fact that they only deposit small amounts of energy in the calorimeter.

Next, electrons are identified by combining tracks with bremsstrahlung energy that is deposited in the ECAL. It is further important to correctly assign the photonic energy in order to avoid double-counting.

The remaining tracks are assumed to be charged hadrons. Of these, a large fraction comes from charged pions. Using this information, the energy that is associated to the tracks is recalibrated for the pion mass.

Particle flow photons are reconstructed from the clusters in the ECAL in areas where no charged tracks exist or where the associated energy is much smaller than the energy in the cluster. These photons are so-called prompt photons in comparison to photons that are emitted in the ECAL showering process.

The remaining energy in the HCAL that has not yet been associated to any objects is interpreted as neutral hadronic energy. While this energy can only be measured with the high uncertainty of the hadronic calorimeter, it only amounts to 10% of the average total energy.

3.6 Jet Reconstruction and Algorithms

Due to the confinement in QCD, free partons cannot be observed. Singular partons will undergo fragmentation soon after creation and deposit their energy by emitting gluons which in turn split into pairs of quarks and anti-quarks. These partons will then hadronize and form stable states by combining to mesons and baryons. Based on the momentum of the originating particle, these decay objects will collimate and form cone-shaped streams.

Due to energy conservation, the sum of all objects in the cone represents the total energy of the originating object. It is therefore necessary to correctly allocate all decay particles to calculate the correct energy. This is done by combining the four-vectors of all particles inside the cone. This object is then called a jet.

In the event reconstruction, all particle flow candidates are combined to form the jets. While it is possible to allocate these objects by hand, the large number of collision events...
3. Object Reconstruction at the CMS Experiment

creates the necessity for automated clustering algorithms\[67\]. These algorithms can be split into two categories: Cone algorithms use cone-shaped templates that are usually placed around a seed or otherwise calculated starting point, while clustering algorithms sequentially combine objects based on energy and distance.

3.6.1 Cone Algorithms

Historically, cone algorithms are the first jet algorithms used since they need less computing power compared to iterative algorithms. The basic version of this algorithm is called the Iterative Cone. A cone with a fixed radius is placed around a seed point and all the energy inside the cone is added up. In most cases, the seed is the highest energy deposition. Depending on the type of the algorithm, the objects inside the cone are simply removed from the event (Iterative Removal) or split between several cones (Split and Merge) depending on their overlap. The shape of the cone is mostly round, except for the case of overlapping cones.

However, these cone algorithms are mostly neither collinear safe nor infrared safe. These effects potentially change the amount of jets in an event based on the splitting of high-energy clusters (collinear) or based on the emission of low energy objects (infrared). The seedless cone algorithm solves this problem by calculating a new set of possible seeds in each clustering step. While this algorithm is both infrared and collinear safe, its computing time rises proportional to \(2^N\) where \(N\) is the number of objects in the event. Further developments in this area resulted in an optimized version, called SISCONE\[68\] algorithm.

3.6.2 Sequential Recombination Algorithms

Sequential recombination algorithms are designed to solve the problem of having a fast algorithm that is both collinear and infrared safe. Instead of starting the jet at a seed point, the algorithm forms jets by distance-based object combination. At the CMS experiment, several algorithms based on the \(k_T\) algorithm are used.

This algorithm begins by calculating the distance between all objects \(i\) and \(j\) and further weighting them with the transverse momentum of the two objects. The following general formula shows this parameter for the \(k_T\) algorithm with \(n = 1\):

\[
d_{i,j} = \min \left( p_{T,i}, p_{T,j} \right) \cdot \Delta R_{i,j} \tag{3.4}
\]

Compared to lepton-lepton accelerators, where there are no remains of the colliding particle, in a proton-proton-collider there are always the remaining parts of the protons that do not interact and therefore are separated from the interacting parton. This object, called beam remnant, will also undergo fragmentation and hadronisation. Therefore, a second distance parameter is defined:

\[
d_{i,B} = p_{T,i} \tag{3.5}
\]

which is the distance to the beam.

In the following, the two objects with the smallest distance are combined to form a proto jet. If the beam distance happens to be the smallest distance, the object is added to the beam remnant. This recombination step is repeated until a fixed distance parameter is reached. After that, all remaining proto-jets are declared jets.

The value of \(n\) can be chosen to be 1, 0 and \(-1\), resulting in three different clustering algorithms. The algorithms with \(n = 1\) and \(n = -1\) are called the \(k_T\) and anti-\(k_T\) algorithm, setting \(n = 0\) results in the Cambridge-Aachen algorithm.
The $k_T$ algorithm starts by clustering close objects with small momenta. This has the benefit that the decay process of the event can be inverted and can be used to allocate the decay objects to their origin objects. However, the shape of the $k_T$ algorithm is often not round, making the distinction of jets more difficult. Later on, the anti-$k_T$ [69] algorithm was introduced. By starting the clustering with the objects with the highest $p_T$, the algorithm creates very round jets that have shapes similar to cone jets. In typical CMS analyses anti-$k_T$ jets with distance parameters 0.5 (LHC Run I) and 0.4 (LHC Run II) are used. In figure 3.1, the jet areas of the algorithms that were introduced in the section are shown.

### 3.6.3 Jet Energy Correction

The detector resolution and geometry lead to effects that smear and affect the kinematic properties of reconstructed jets. The measured jet four vector is affected by effects that appear due to the different detector response in different areas of the detector and by missing jet energy caused by dead zones in the detector. The jet energy correction[70] aims at correcting the measured jet energy on average to the energy of the underlying particle level jet. The corrections are determined to large parts from simulation. Small residual differences to the response in data are measured and corrected for with data driven techniques. The jet energy corrections are split into several steps which are applied consecutively. Factorizing the corrections has the advantage of optimizing each step on its own and being able to independently determine the uncertainties of each correction. The first three corrections are mandatory, the remaining corrections are optional.

**Required Corrections:**

**Level 1: Pileup:** In the first step, the jet is corrected for energy coming from pileup and detector noise.
3. Object Reconstruction at the CMS Experiment

Level 2: Relative ($\eta$): The jet energy is calculated using the information from the hadronic calorimeter. Based on the pseudorapidity of the jet, they can be separated into central jets ($|\eta| < 1.3$) and forward and very forward jets ($|\eta| > 1.3$). To correct the forward jets, they are balanced against central jets using data containing two jets.

Level 3: Absolute ($p_T$): In this step, the absolute transverse momentum of the jet is rescaled. This is done by either comparing the momentum of the reconstructed jet to the corresponding parton jet or by balancing the jet energy against an object with high energy resolution (photon or Z boson).

In addition to these obligatory corrections, there are a number of optional corrections which are not applied in the context of this analysis. At first, the jet energy is corrected for energy that is deposited in the ECAL. Next, flavour-specific corrections are applied. In a further step, the jet is corrected for energy coming from the underlying event. In the final correction, the reconstructed jet is compared to the generator level parton.

3.7 Identification of Bottom-Quark Jets

Nearly all top quarks that are created in a collider experiment will decay into a bottom quark the via emission of a W boson. Furthermore, the Higgs boson has a high branching ratio into bottom quarks. While these two factors make the identification of bottom quarks in the context of this thesis very important, there is a multitude of other analyses that involve studying bottom quarks, among them measuring of CP violation in B hadron systems or the search for new physics, like additional quark families. In the context of particle physics, the identification of bottom quarks is called b-tagging, the output of the algorithms is called the b-tag.

The identification of b-jets is possible since the bottom quark has physical properties that distinguishes them from the four lighter quarks. First, it has a much higher mass than the lighter quarks. Second, hadrons that contain bottom quarks have a significant longer lifetime than hadrons with light flavour quarks, resulting in a measurable flight distance with a decay point that is discernible from the primary vertex of the event. When looking at the reconstructed jets, it is further noticeable that jets coming from B hadrons produce objects with high momenta (compared to light hadrons).

To identify bottom quarks, a variety of algorithms is used. They employ a set of input parameters to construct a discriminator value which is optimized to separate bottom quarks from light flavour quarks. A threshold cut that is applied on the output separates these two groups. The possible thresholds are called working points and are defined by their probability of wrongly tagging a light flavour quark as a bottom quark. At the loose working point this probability amounts to 10%, at the medium working point to 1% and at the tight working point to 0.1%. However, due to the more restrictive selection at the medium and tight working point, the identification rate of bottom quark jets also decreases. In most cases, the medium working point is chosen, since it poses a compromise between discrimination and efficiency.

In the following, several different b-tagging algorithms are presented. While different detector setups allow for a wide variety of tagging algorithms, these are the prevalent methods used by the CMS experiment[71].

3.7.1 Impact Parameter Algorithms

As noticed in the last section, hadrons containing bottom quarks have significantly higher lifetimes than hadrons containing only light flavours. Therefore, these hadrons will propagate away from the primary vertex of the event and decay with a significant distance...
to it. Since the primary vertex is the decay point of most light objects, their tracks can be extrapolated to it. However, the decay products of B hadrons have a different decay point, which is why their tracks will not extrapolate back to the primary vertex but pass it with a measurable distance. The impact parameter defines the minimum of this distance. By dividing the impact parameter (IP) by its uncertainty $\sigma(IP)$, one can define the impact parameter significance $S_{IP}$. The impact parameter as discriminating variable has been chosen since the high three-dimensional resolution of the tracker system allows for a precise measurement of the primary vertex and the track direction.

While the IP significance is no tagging algorithm by itself, it is a quantity that separates tracks coming from B hadrons and light flavour hadrons. After sorting the tracks of a jet by this significance, the second- or third-highest value is used to discriminate between bottom and non-bottom jets. A different variety of tagging algorithms, called Jet Probability, uses a weighted combination of significances to calculate the probability for the jet to come from a bottom quark.

### 3.7.2 Secondary Vertex Algorithms

The second approach tries to calculate the decay vertex of the B hadron, called secondary vertex. To enhance the purity of the secondary vertex candidates, the fraction of shared tracks between primary and secondary vertex may not exceed 65%. Further cuts on the maximal distance and mass of the vertex are applied to reject long-living kaons.

The first algorithm, called Simple Secondary Vertex (SSC) evaluates a secondary vertex based on the flight distance significance. The two subalgorithms, called High Efficiency and High Purity require two and three associated tracks, respectively. However, the tagging efficiency is limited by the reconstruction efficiency of the secondary vertices itself.

To improve on this first approach, the Combined Secondary Vertex (CSV) algorithm is used. In the case where no secondary vertex can be found, a different evaluation is taken into action, which uses a reduced set of variables. By combining all tracks that exceed a impact parameter significance threshold, a pseudo vertex is created. If this is not possible, an evaluation based on general jet track parameters is employed that is comparable to the jet probability algorithm based on the impact parameter. The set of variables included in the CSV algorithm consist mostly of variables describing the jet tracks or the secondary vertex, such as the track multiplicity or energy fraction of tracks coming from the vertex or the vertex mass.

### 3.8 Subjet Reconstruction with Fat Jet Algorithms

The production of a Higgs boson decaying into two bottom quarks in association with a semileptonic top quark pair is the main process investigated in this thesis. In leading order, the event consists of one lepton, six jets, of which four are coming from bottom quarks and missing transverse energy from the neutrino. The main background is represented by semileptonic top-pair production with an emitted gluon splitting into a pair of bottom quarks, which has an identical event signature. To separate these two processes from the rest of the background processes and from each other, it is necessary to correctly assign the reconstructed jets. However, the high jet multiplicity results in a high number of possible jet assignments.

The main reason for the difficult jet assignment is the fact that the top quarks and the Higgs boson have comparably small momenta due to their high masses. In the center of mass frame of the decaying particle, this high mass results in decay products with high momenta, which are distributed isotropically. This leads to a high spread of the jets in
the \(\eta-\phi\) space in the laboratory frame of the detector. By putting a lower boundary on the transverse momenta of the top quarks and the Higgs boson, the jets receive a large boost in the direction of the decayed object. If this boost momentum is larger than the center of mass jet momentum, they get collimated and form large, overlapping structures, called fat jets. These jets offer both advantages and disadvantages when reconstructing the complete event. The advantage is that all objects coming from the primary object are collimated and can therefore be assigned more easily. However, since they are collimated into a single jet, it becomes more difficult to reconstruct the individual objects. Further, these large jets often contain energy deposits that do not originate from the decay object. They consist of gluon emissions with small energies coming from pile-up and the underlying event or of initial and final state radiation coming from other sub-processes in the event.

Subjet algorithms aim to correctly assign the energy and tracks inside the jet radius and also to remove the additional radiation. At first, traditional jet algorithms like CA are used to create jets with large radii. Afterwards, the algorithms analyse the components of the fat jets and create a set of so-called subjets which are candidates for the decay products of the primary particles. In this analysis, two kinds of subjet algorithms are employed. The reconstruction of the top quark with a hadronically decaying W boson is done via the HEP-Toptagger subjet algorithm which is more closely explained in section 3.8.1. This algorithm separates the fat jet in a set of smaller subjets that are cleaned from additional radiation and reclustered into exactly three jets that correspond to the bottom quark and the two jets coming from the hadronic W boson. The reconstruction of the Higgs boson happens via the subjet-filterjet algorithm, which is detailed in section 3.8.2. This second algorithm splits the fat jet into a variable amount of subjets which are also cleaned and reclustered. This algorithm has been optimized for the use in the Higgs boson to bottom quark decay channel, which is the channel used in this analysis.

### 3.8.1 HEP-Toptagger

The reconstruction of the hadronic top is achieved via the HEP-Toptagger, an algorithm that has been developed at the ITT institute at the University of Heidelberg intended for the reconstruction of top quarks as decay objects from supersymmetric top squarks[3].

In a laboratory system with small transverse momenta, the leading order of the hadronic decay channel of the top quark results in three jets. Two of them originate in the W boson, while the third comes from the bottom quark. In higher orders, additional jets can be created by the emission of gluons from the hadrons in the initial or final state. In the boosted state that is observed in parts of this thesis, the top components are collimated into a single cone of particles. The following task is to correctly assign these objects to the three decay objects of the top quark. The procedure of the algorithm is visualized in Fig. 3.2.

The HEP-Toptagger algorithm begins by clustering the complete event into Cambridge Aachen jets with distance parameter \(R = 1.5\). In the following, these jets are referred to as "fat jets". A \(p_T\) threshold of at least 200 GeV/c and a radius of 1.5 guarantees that in the majority of the cases all the three partons coming from the top quark are inside the fat jet. Further, the high cut on the transverse momentum revokes a large fraction of the combinatorial background.

The next step of the process is presented by step 1. in Fig. 3.2. By reversing the iterative jet clustering, an object is split into two objects which can be either a smaller composite object or a single particle flow candidate. To remove objects that come either from the underlying event or pileup, a mass drop criterion is employed, which is visualised after step 1. Here, the lighter of the two objects is discarded if the harder object has 80% or
3.8. Subjet Reconstruction with Fat Jet Algorithms

Figure 3.2: Visualisation of the HEP-Toptagger algorithm [72]. The goal of the algorithm is the reconstruction of the top quark with the hadronically decaying W boson. The reconstruction is separated into several stages. In 1., the original jet is declustered until the exit condition for every object is met. In 2.-4., a combination of selection and reclustering steps is taken which leads to the clustering of three jets, called subjets which are then assigned to the three decay objects of the top quark, the bottom quark and the two jets from the W boson decay. Taken from [73].

more of the mass of the combined object. The declustering is continued (2.) until one of the exit conditions is met:

- The element is a single particle flow object;
- The mass of the element is less than 30 GeV/c².

The declustering of the fat jets is continued, until the exit condition is met for every emerging object. The final set of objects is called subjets. The original fat jet is discarded if the fat jet contains less than three subjets.

Following the declustering, the subjets are reclustered. From the set of subjets, combinations of three are formed (3.). For each combination of three subjets, the elements are reclustered to form Cambridge Aachen jets with variable distance parameters. Of these new jets, the five jets with highest $p_T$ are kept. These so-called filtered subjets are combined and their mass is referred to as filtered mass.
In the next step (4.), the combination of five filtered subjets with the mass closest to the top mass is chosen. As a last step, these five jets are again reclustered to form exactly three CA jets.

The three final jets are assigned as the bottom quark jet and the two jets coming from the W boson. The bottom quark jet is assigned to the jet with the highest b tagging output. The other two jets are combined to form the hadronic W boson. In the end, each event can contain several fat jets where each can be the real hadronic top. Subsequent selection steps are employed to select the fat jet which most likely represents the top quark decay products.

At the time of writing this thesis at the beginning of LHC run 2, a new iteration of the HEP-Toptagger has been released[74]. This new version replaces the fixed fat jet radius by an optimized variable value that is calculated using parameters of the event. Further, the top mass is no longer used in the jet selection to not falsely improve the combinatorial background. However, since this thesis mostly relies on data registered of LHC run I, the fat jets are still clustered using the older version of this algorithm.

### 3.8.2 Subjet-Filterjet-Algorithm

The second subjet algorithm is the subjet-filterjet algorithm [76] and is visualised in Fig. 3.3. This algorithm was developed especially for the reconstruction of Higgs bosons decaying into a pair of bottom quarks. The general procedure is comparable to that of the HEP-Toptagger algorithm. It can be split into the clustering and declustering stage. In the end, each fat jet will consist of a variable amount of small jets that can be combined to form the Higgs boson. As mentioned before, hadrons coming from bottom quarks hold a large fraction of the total mass and momentum. This has to be regarded in the declustering of the fat jet.

The first stage consists of the CA clustering. Again, a $p_T$ threshold is used in combination with a radius larger than 1. The distribution of the Higgs transverse momentum in combination with the distance between the two bottom quarks show that a lower boundary of 150 GeV/c for the transverse momentum and a radius of 1.2 will result in a high purity of the fat jets.

Since both algorithms use CA jets as the basic jet, the declustering phase (1.) is comparable to that of the HEP-Toptagger. However, due to the special properties of B hadrons, this algorithm features different exit conditions in the declustering. Here, a mass drop criterion is employed (2.) which stops the declustering if the heavier of the two objects has more than 66% of the combined mass.

In the reclustering step, the final declustering step is reversed (3.). Afterwards, the set of objects in the CA jet is reclustered again using the CA algorithm with a radius of 0.3. After a cleaning step, that removes radiation from the underlying event and other parts of the event, the fat jet now consists of a variable amount of filtered subjets (4.). In order to count as Higgs jet candidate, two of them have to fulfil the b tagging threshold. To reconstruct the four-vector of the Higgs boson, different combinations of the filterjets are possible. The first possibility is the combination of the two jets with the highest b tag output. Since it is possible for any of the two jets to radiate gluons, a second combinations adds the filterjet with the highest $p_T$ of the remaining jets. In the analysis, both the two- and three-jet combinations are studied.
The goal of the algorithm is the reconstruction of the Higgs boson which decays into two bottom quarks. The reconstruction is separated into several stages. In 1., the original jet is declustered until the exit condition for every object is met. In 2.-4., a combination of selection and reclustering steps is taken which leads to the clustering of a number of jets, called subjets, which are then assigned to the two bottom quarks and additional radiated particles. Taken from [75].
4. Data Generation and Monte Carlo Generators

One of the most important parts in an analysis based on a particle accelerator experiment is the validation of the simulated data that is generated using Monte Carlo generators. This validation has several goals. For once, agreement in both data sources indicates that the theoretical calculations are valid and that they are able to effectively describe the experimental data. Furthermore, deviations between the experimental and simulated data especially in high precision measurements can provide input for future revisions and additions to the theoretical base that is the standard model. It is therefore necessary to create simulated data in such a way that it describes the experimental data in the best way possible.

Event generators are software that simulate data using a multistage process. They can be split into parton-based generators, multi-purpose generators and others, which for instance simulate neutrino interaction. While each generator contains the basic tools to simulate the matrix element of the event, each generator is further specialized for a certain task. While some generators are more efficient in simulating a wide variety of processes, others are very specialized and contain additional tools to calculate rare subprocesses. Further, the choice of input parameters and physical models that are used as additional input can also lead to different results. Further, data from different generators is often used to estimate the uncertainties that come from the event generation, as different generators implement different physics models.

The chapter is structured as follows: In section 4.1, the different steps in the event generation are explained. They consist of the hard process, the fragmentation and hadronization, the addition of the underlying event and the subsequent simulation of the detector. In section 4.2, the different types of Monte Carlo event generators are presented. They are split into event generators, which are able to simulate the complete event including all steps, and shower Monte Carlo generators, which are able to generate the matrix element of more complex processes and for higher orders.

4.1 Process of Event Generation

The creation and adjustment of the simulated data consists usually of several consecutive processes. The first step is called the event generation and involves a series of calculations and simulations that can either be done via a single program or via several adjointed
programs. While some parts of the simulation come from analytic calculations, others are provided through empiric modelling in cases where the calculations cannot be applied, for instance where perturbation theory cannot be applied. The central part of the event generation is the hard process. Here, the processes in the matrix element are simulated, that result in the final state given by the calculation. After the decay of the very short-lived objects, fragmentation and hadronization simulate the steps that form physical objects from the unphysical final state. In the end, the interaction between the particles with the detector material is simulated, which makes the generated event comparable to that measured in the experiment.

4.1.1 Hard Process

The central part of the event simulation is the calculation of the hard process. At first, the cross section of the parton interaction at fixed energies, also known as hard process, is calculated. However, as the LHC is a proton accelerator, the momentum fraction of the parton must be calculated. This is done with the help of the PDF as explained in section 1.1.5

As the process is happening in a high energy frame, the calculation via perturbation theory is used. Here, all possible variations of a fixed order of the process are calculated and combined. As stated before, to get the exact matrix element, it has to be calculated up to infinite orders. However, since the amount of diagrams that are included in higher orders rises exponentially, a complete simulation of all orders would strongly increase the computation time of an event. To circumvent this, the so-called K-factor is introduced. The K-factor is the ratio of the cross section in which the event can be simulated and the value of the cross section which can be calculated up to several orders higher. It is then possible to scale the production cross section of the simulated event to the calculated value, increasing the accuracy of the simulation in theory. However, especially at low orders the differences in the shape of the cross section differ when going towards higher orders, which means that a simple scaling does not necessary improve the agreement.

To calculate the phase space integral, often the principle of Monte Carlo integration is applied. The phase space is a multidimensional space that includes all the final state energy distributions that are allowed with the given boundary conditions. While it is possible to calculate such an integral by approximating the space with classical integration methods such as the trapezoidal rule, it can be shown that for a number of dimensions larger than three, the Monte Carlo method is faster[77]. The algorithm creates a reference area that contains the phase space. By filling the reference area with randomly placed points and comparing the ratio of points inside the target area with the total number of points, the target area can be approximated. The method of Monte Carlo integration is also used in other calculation steps in the event generation.

4.1.2 Showering and Hadronization

The quarks that are created in the hard process are isolated objects with nonzero colour charge. In reality, however, only colourless objects that are combinations of several quarks can be observed. Fragmentation and hadronization are two steps that are simulated in event generators. Fragmentation simulates the splitting of quarks and gluons to form new partons, while hadronization combines these partons to form colourless hadrons.

In QCD, both quarks and gluons carry colour charge, therefore an interaction between both quarks and gluons and between gluons and gluons is possible. In practice, quarks emit gluons under shallow angles, while the gluons afterwards will split into pairs of quarks and antiquarks. This splitting is simulated by splitting evolution equations such as the
DGLAP equation in a process similar to the evolution of the PDF. They calculate the
chance of emitting a parton based on the energy of the emitting particle and the emission
angle.

As the showering progresses, the energy of the partons and the radiation will decrease.
At energies below 1 GeV[3], perturbative calculations no longer are valid, since the strong
coupling diverges. At this point, phenomenological models are used to describe the process
of hadronization. The two most common models are the Lund string model and the cluster
model.

The Lund string model is based on the long range properties of the strong interaction.
At small distances, the strong coupling is small, which leads to the state of asymptotic
freedom similar as in QED. However, at long distances the linear contribution in QCD that
represents the self-interaction of gluons becomes dominant. This leads to a compression
of the gluon field lines into a tube, called string. The process of hadronization happens
via the creation of additional quark antiquark pairs. As the distance between two quarks
increases, the energy stored in the coupling increases as well until the connecting gluon is
split, forming a new quark pair.

The second model is called the cluster model. This model is based on the preconfinement
property of QCD. As shown by Amati and Veneziano, partons coming from fragmentation
showers will form clusters. These clusters have mass distributions that are depending only
on properties of the showering, like the absolute energy scale $q$ and the QCD scale $\Lambda$.
Clusters above a certain mass threshold are decayed using a simple phase space model.
The decayed clusters are assigned as proto-hadrons.

### 4.1.3 Underlying Event and Pile Up

The underlying event describes all processes simulated in the event generation that are
not directly connected to the hard process. While it is assumed that the primary parton
coming from the proton carries the largest fraction of the proton energy, it is also possible
for the rest of the proton, called beam remnant to take part in the event. This includes
the emission of soft gluons that form additional hadrons, the interaction with partons
coming from the hard process and the hadronization of the beam remnant. There are two
types of pileup: In-time pileup is the simultaneous collision of several proton pairs inside
a single bunch crossing, and is simulated by adding additional soft proton collisions to the
event. The second type of pileup, called out-of-time pileup, comes from the superposition
of several adjacent bunch crossings. It is simulated in the same way as the in-time pileup.

### 4.1.4 Hadron Decay

In the next step, the decay of the unstable hadrons that are created in the hadronization
is simulated. Partons decay until only long-lived or stable particles are left. This step
is especially important in high-energy particle accelerators since previous measurements
have shown that a large fraction of the particles found in the final state of the event come
from excited hadron states.

### 4.1.5 Detector Simulation and Unfolding

In the previous steps, the simulation of the event via simulating the hard process and
the following hadronization has been done. However, these events cannot yet be used to
compare to real data. Experimental events are measured using the detector, which results
in inefficiencies and other effects based on the method of measurement.

To be comparable to experimental data, the simulated events have to undergo the same
experimental reconstruction in the detector. This is achieved by measuring the event in a
digital model of the detector. Further, for these Monte Carlo events, the same reconstruction algorithms are applied as for the data events. The process of digitally measuring the event in the simulated detector is called detector simulation and is provided by the Geant software. While Geant provides a complete, but slow simulation of the detector, faster detector simulations can be used for preliminary results.

In some situations it is necessary to reverse the effects of the detector on the real data. In a process called unfolding, simulated data is used to create the correlation of parameters before and after the detector simulation. This correlation is later applied to experimental data. Unfolding is used to make the experimental results independent of the detector experiment and allows for a direct comparison to results of other analysis groups.

4.2 Variants of Monte Carlo Generators

Event generators can be broadly split in two groups. General purpose generators are developed to simulate the complete event. They include a matrix element generator with a parton shower and the subsequent hadronization. While they are standalone and support a wide range of processes, they have a limited final-state multiplicity and mostly only support leading order (LO) matrix elements. Shower Monte Carlo generators disregard the hadronization in favor of a broader matrix-element generation and showering. Modern shower generators support next-to-leading order (NLO) processes with in principle arbitrary numbers of final-state particles and the subsequent showering. Since these generators have no hadronization mechanics themselves, they use the standalone hadronization parts of other general purpose generators.

4.2.1 General Purpose Generators

An event generator that is able to simulate the complete event it is called a general purpose event generator. They include the hard process in addition to the fragmentation and hadronization. However, due to the complexity of both the showering process and the matrix element generation, only leading order processes with one or two objects in the final state are supported. The reason for this is that the implemented parton showering is only able to simulate either hard emissions under small angles (collinear) or soft emissions under large angles (infrared). The inability to simulate hard emissions with large angles prohibits many-body final states. To be able to match the showering to the hard process, only general kinematic parameters such as the energy and flavour of the incoming and outgoing particles are used. General purpose generators mostly differ in their choice of hadronization and particle decay models or in the choice of input parameters such as the evolution scale of the showering process.

Herwig

Herwig\textsuperscript{[78]} is a general purpose event generator originally programmed in Fortran. The first version of Herwig was published in 1986. Its latest implementation, called Herwig++, is completely rewritten in C++. While Herwig++ is basically a completely rewritten generator based on the code of the classic Herwig program, it also contains several new modules that allow additional functionality without external support. Herwig is able to simulate collisions between leptons and hadrons. An important feature is the good simulation of QCD parton showers with angular ordered evolution both in the initial state and the fragmentation. The hadronization is done via the cluster model.
4.2. Variants of Monte Carlo Generators

Pythia

Pythia[79] is a general purpose event generator that is used both for the complete event generation and for its showering and hadronization models. It is developed by the theoretical particle physics group in Lund. Its current version, Pythia 8, is the successor to the previously used version, Pythia 6, which was like Herwig still programmed in Fortran. The generator is able to simulate collisions of protons, electrons and muons in an energy range from 10 GeV to 100 TeV. While the matrix-element generator is only able to simulate 2 → 1 and 2 → 2 processes, it is nevertheless able to generate a wide range of processes, going from QCD, electroweak, quarkonia, top and a wide variety of beyond standard model processes. The high number of decay objects from decays of the top quark or the W boson are simulated using the included parton showering.

The parton showering is done via a $p_T$-ordered evolution, which combines the initial state shower with the final state fragmentation. While this simulation is designed to model collinear and infrared emissions, it is also able to create hard additional jets with LO accuracy. The simulation of the underlying event allows for multiple scatterings and the production of prompt photons, quarkonia and the t-channel exchange of vector bosons.

The process of hadronization is done via the Lund string model. This model simulates the confinement between partons as a colour-tubes with a potential energy of $\approx 1$ GeV/fm. If the stored energy is large enough, the tube will break, resulting in new pairs of partons. This process continues until all partons are on-shell.

4.2.2 Shower Monte Carlo

Shower Monte Carlo generators are generators that only simulate the hard process and the following fragmentation. They were developed to support processes with an arbitrary number of final state particles. By implementing advanced matrix element techniques, it is further possible to generate the hard process both in leading order and in higher orders. The difference in the calculation of leading and next-to-leading order are the addition of virtual corrections coming from interference effects with loop diagrams and the correction terms coming from real emissions in the initial or final state.

The good description of hard emissions allows for the modelling of additional partons in the matrix element. These additional jets are equivalent to jets which are simulated in the fragmentation stage. However, the accuracy of the modelling is improved, especially at larger angles and higher $p_T$, since matrix element jets come from exact calculations, while parton shower jets are simulated using showering models.

As shower Monte Carlo generators do not include the hadronization step, the generated events are used as input for the showering and hadronization models that are included in general purpose generators such as Phythia or Herwig. To avoid double-counting events, the processes of matching and merging are used. Merging is used in the case of combining exclusive samples. Even in the case of leading order processes, double counting is possible if the matrix elements contain different numbers of final state partons. Matching combines events with equal jet multiplicities that are created either in the matrix element or the showering. These events appear when comparing next-to-leading order processes, where events featuring the real emission may be identical to those with the virtual corrections and an additional parton shower jet. In recent years, several approaches have been developed that apply matching and merging in different ways, mostly depending on the type of process that is being simulated.

MadGraph and aMC@NLO

The matrix-element generator MadGraph[80] is a program to create leading-order matrix elements. MadGraph supports both particle decays and $2 \rightarrow n$ processes. The matrix-
element generator takes the initial- and final-state particles as input and computes all possible Feynman diagrams. The calculation of the matrix element is then done via helicity wave-functions. This process is also applied for all potential subprocesses. The subsequent simulation of the kinematic event is done with the MadEvent event generator.

The aMC@NLO\cite{81} approach is an addition to the leading order matrix element generation done in MadGraph. It allows both the simulation of next-to-leading processes and leading-order processes with additional QCD radiation. Further, a merging procedure is included to combine processes with different final state multiplicities. The corrections based on next-to-leading are applied by subtracting the leading-order Monte Carlo shower corrections from the next-to-leading calculations. This is done by creating events with negative weights that represent events which appear both in the corrected leading order and the next-to-leading order calculation.

**Powheg**

The Powheg\cite{82} method is an alternative approach to generate processes in next-to-leading order precision. It is used to combine events generated at matrix-element level with shower Monte Carlo generators such as Pythia. One of the problems of combining these generators is the double-counting of events, since many SMCs already use corrections on next-to-leading order level. In Powheg and especially in Powheg-Box, the hardest emission after the matrix element process is also simulated using next-to-leading order accuracy, which removes the necessity of negative weight events like in Madgraph. A further speciality of Powheg is that simulations in next-to-leading order are independent of the parton shower and can therefore be interfaced with Monte Carlo shower generators.
5. Data and Event Selection

A major topic of this thesis is the comparison of the data from the CMS detector experiment with equivalent simulated data. In general, these studies are performed both in a signal enriched region of the phase space and a perpendicular sideband. This is achieved by subjecting the data to a combination of selections that aim at suppressing the majority of the background processes while only rejecting a small part of the signal event fraction. The actual verification is done by comparing the distributions of the data with expectations coming from Monte Carlo samples. These samples consist of the signal process and a variety of background processes. The choice of these processes and a short description of their respective event signature is given in section 5.1.

The following sections provide information about the data and Monte Carlo samples. While section 5.2 states the analysed data sets, the applied trigger and the luminosity, the section 5.3 contains a list of the samples in addition to the Monte Carlo event generator, the number of generated events and the production cross section that was used in the event generation.

The selection of an event consists of several steps. At first, events are preselected by the choice of trigger conditions. While generated events do not have to activate a trigger to be selected, it is still necessary to ensure an equal selection and therefore, the same trigger selection is applied. In the next step, the events are further selected offline based mostly on the multiplicity of objects like jets or leptons.

While the event generation with Monte Carlo generators allows for a simulation of events very similar to that of experimental data, the exact simulation of events is a difficult task, as additional corrections from higher orders are not always available and also lead to a rapid increase in computational time. Further, the reconstruction of simulated events in digital model of the detector is not exact and leads to further differences in data and Monte Carlo samples.

It is therefore necessary to correct the simulated events, e.g. with a combination of reweighting procedures. These corrections are partially based on the comparison of generated and measured events and on comparison between parton level information and the completely reconstructed event. In this analysis, a correction is applied on the distribution of the number of simultaneous proton-proton-collisions, called pileup, that is represented by the number of primary vertices. Further corrections are applied on the output of the b-tagging algorithm and the kinematic distributions of the top quark. These corrections and their
effect on the matching between experimental and simulated data are discussed in section 5.5.

5.1 Signal and Background Processes

5.1.1 Event Signature

The signal process of this analysis is the associated production of a Higgs boson and a top-antitop-quark pair. In this case, the Higgs boson is radiated from one of the top quarks and decays into a pair of bottom quarks. The $t\bar{t}$ lepton plus jets decay channel is chosen. In this channel, one of the $W$ bosons decays into a lepton and a neutrino, while the other decays into jets. In an event in which the Higgs boson and the top quarks have small transverse momenta compared to their rest mass, the event is composed of a total of six jets, four of which come from bottom quarks, an isolated lepton coming from the leptonic $W$ boson and additional missing energy coming from the neutrino.

The semileptonic decay channel is chosen since it provides a single, isolated lepton with high transverse momentum that can be efficiently detected. This lepton in the event is especially useful as a selection criterion, as it allows a lepton-based trigger that suppresses QCD background processes. While the dileptonic top-quark-pair decay channel has a similar reconstruction efficiency, the branching ratio of this channel is about four times smaller.

However, in case of high transverse momenta of the Higgs boson and the hadronic top quark, the decay products of different objects can overlap. These large bundles of particles are then clustered into jets with radii in $R=(\eta, \phi)$ of 1 or more. However, while the jets constructed with the aforementioned subjet algorithms are used to perform the analysis in events with high transverse momenta, the construction of standard anti-kt jets is still performed. They are used in the categorization of the event and as means of selecting without using information from the top-quark or Higgs boson system.

5.1.2 Top-Quark Pair with Additional Partons

The dominant background is the top-quark pair production. The additional radiation of a gluon splitting into a pair of bottom quarks, makes this process nearly indistinguishable from the signal process. A Feynman graph of this process is shown in Fig. 5.1a. Furthermore, the top-antitop pair-production process has a nearly 2000 times larger production cross section than $t\bar{t}H$. While the bottom quarks coming from the Higgs boson can be reconstructed to form the resonance at the known Higgs mass, the number of expected events for this process makes it possible to fake such a peak in this background.

The separation of the signal process and the top-quark pair production is done in a multi-step selection. In the first step, a pre-selection is applied to suppress all processes but the signal process and the semileptonic top-quark-pair production. After that, a more specialized selection is applied to separate these two processes.

5.1.3 Single Top Background

While the production of a single top quark has a two times smaller production cross section than top-quark pair production, this process still contains one of the three primary objects of the signal process. This background can be split further into the three different production channels that are available at the LHC accelerator.

The $t$-channel has the highest production cross sections of the three single top channels. In this channel, a bottom quark transforms into a top quark by interacting with a $W$ boson.
that has come from a different flavour changing process. The final state contains a top quark and an additional bottom quark, in case the original bottom quark was produced in a gluon splitting.

In the s-channel the top quark is produced from a W boson that is created from a quark-antiquark annihilation. Since the LHC is a proton proton collider, the antiquark is a sea quark. Since sea quarks have softer transverse momenta than the valence quarks, it is unlikely for a sea quark to have enough momentum to form a real W boson that can split into a top and bottom quark, resulting in a small production cross section. The importance of this channel results from the fact that it has an additional bottom quark in the final state, such that the signal is mimicked.

In the associated top W boson channel a bottom quark splits into a top quark and a W boson after an interaction with a gluon. At the LHC, this process has a similar production cross section if the bottom quark is again created by gluon splitting. In that case the final state contains a top quark, a W boson and an additional bottom quark, which makes the event signature identical to that of the top-quark pair production.

5.1.4 Vector Boson with Additional Jets

Another important background process is the direct production of vector bosons in association with additional jets. Especially the production of W bosons that decay into leptons and neutrinos is an important channel due to its large production cross section and the presence of an isolated lepton. This process can be suppressed by demanding at least one b-tagged jet. The second process, the decay of a Z boson into two leptons with additional jets has also a similarity to the signal process in case only one of the two leptons of the Z boson is reconstructed. However, Z boson events can be suppressed by vetoing events that contain a second isolated lepton.

5.1.5 Diboson

The production of pairs of vector bosons is the last of the background processes considered in this thesis. Possible combinations are either a pair of W or Z bosons or a combination of both. Since the final state of the signal process contains one lepton, only the leptonic final states are included. Due to their small production cross sections, these processes are not as important as the production of a single W or Z boson. Again, these backgrounds can be suppressed by demanding at least one b-tagged jet and a minimal number of jets.
5.1 Not included: QCD

Processes that solely involve QCD interaction and that do not create a final state with a characteristic mass resonance are called QCD multijet background. Due to the high relative strength of the strong force, the cross section of this process is several orders of magnitude larger than any process that involves the production of massive particles. Since these interactions happen at low scales at which the strong coupling constant is large, it is difficult to simulate the processes using perturbation theory. However, these QCD processes are efficiently suppressed by the selection of isolated leptons and b-tagged jets. Therefore, this background process is negligible.

5.2 Data Samples

This thesis uses the complete recorded data of the 2012 run with a center of mass energy of 8 TeV corresponding to an integrated luminosity of 19.8 fb\(^{-1}\). The data sets consist of the re-reconstructed data that was released at January 22 2013 and contains the final data and Monte Carlo corrections for the data of 2012.

The data is split into two groups based on the lepton that is triggered. In this case, a muon and an electron trigger are employed. The muon trigger selects isolated muons with a \(p_T\) of at least 24 GeV/c and \(|\eta| < 2.1\). The electron trigger selects electrons with a \(p_T\) of at least 27 GeV/c.

The recorded events are split into Runs. A Run contains all events of a certain proton injection into the collider. A selection based on the so-called JSON file only selects data of Runs for which an overall good detector condition is certified. In this case, the golden JSON that is associated with the re-reconstructed data sets is used. A golden JSON only contains the Runs for which all components of the detector are functional. The complete list of data samples is given in table 5.1.

5.3 Monte Carlo Samples

The samples of generated events are chosen based on the aforementioned signal and background processes. In all processes, the number of generated events is larger than the number of expected events based on the production cross section and the integrated luminosity of the data. Therefore, they are reweighted using the ratio of generated and expected events which are calculated via the luminosity of data and the production cross section.

For the signal process, a sample has been chosen that includes all decay channels of the Higgs boson and the top-quark pair.

The top-quark anti-quark pair-production is represented by samples of semileptonic and dileptonic decay channels. The dilepton sample was added since it is possible for one of the two leptons not to fulfil the lepton selection.
Table 5.2: The list of Monte Carlo samples included in this analysis. The table lists process, generator, number of generated events and the production cross section, which are taken from the internal Analysis Note associated to [83].

<table>
<thead>
<tr>
<th>Process</th>
<th>Generator</th>
<th>$N_{\text{generated}}$</th>
<th>$\sigma$ / pb</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t\bar{t}H$ inclusive</td>
<td>Pythia</td>
<td>995697</td>
<td>0.13</td>
</tr>
<tr>
<td>$t\bar{t}$ semileptonic</td>
<td>MadGraph</td>
<td>24953451</td>
<td>107.66</td>
</tr>
<tr>
<td>$t\bar{t}$ dileptonic</td>
<td>MadGraph</td>
<td>12011428</td>
<td>25.81</td>
</tr>
<tr>
<td>$t$ (s-channel)</td>
<td>Powheg</td>
<td>239963</td>
<td>3.79</td>
</tr>
<tr>
<td>$t$ (s-channel)</td>
<td>Powheg</td>
<td>139974</td>
<td>1.76</td>
</tr>
<tr>
<td>$t$ (t-channel)</td>
<td>Powheg</td>
<td>3758227</td>
<td>56.40</td>
</tr>
<tr>
<td>$t$ (t-channel)</td>
<td>Powheg</td>
<td>1935072</td>
<td>30.70</td>
</tr>
<tr>
<td>$t$ (tW-channel)</td>
<td>Powheg</td>
<td>490480</td>
<td>11.10</td>
</tr>
<tr>
<td>$t$ (tW-channel)</td>
<td>Powheg</td>
<td>493460</td>
<td>11.10</td>
</tr>
<tr>
<td>$W \rightarrow l^+\nu + 1$ Jet</td>
<td>MadGraph</td>
<td>52779746</td>
<td>6440.40</td>
</tr>
<tr>
<td>$W \rightarrow l^+\nu + 2$ Jets</td>
<td>MadGraph</td>
<td>52084368</td>
<td>2087.20</td>
</tr>
<tr>
<td>$W \rightarrow l^+\nu + 3$ Jets</td>
<td>MadGraph</td>
<td>30773032</td>
<td>619.0</td>
</tr>
<tr>
<td>$W \rightarrow l^+\nu + 4$ Jets</td>
<td>MadGraph</td>
<td>13382803</td>
<td>255.20</td>
</tr>
<tr>
<td>$Z \rightarrow l^+l^-$ + 3 Jets</td>
<td>MadGraph</td>
<td>11053357</td>
<td>66.07</td>
</tr>
<tr>
<td>$Z \rightarrow l^+l^-$ + 4 Jets</td>
<td>MadGraph</td>
<td>6404237</td>
<td>27.38</td>
</tr>
<tr>
<td>WW</td>
<td>Pythia</td>
<td>9988485</td>
<td>54.80</td>
</tr>
<tr>
<td>WZ</td>
<td>Pythia</td>
<td>9988254</td>
<td>32.30</td>
</tr>
<tr>
<td>ZZ</td>
<td>Pythia</td>
<td>9799908</td>
<td>7.70</td>
</tr>
</tbody>
</table>

In the single top process, samples of the s-channel, t-channel and associated tW production are employed. These samples contain both the leptonic and the hadronic decay channels of the top quark and the W-boson in the case of the associated production.

The processes of vector bosons with additional jets are represented by samples that include all leptonic final states. For the process of W bosons with jets, samples with one, two three and four additional jets are included, while for the production of Z bosons with additional jets the samples with three and four jets are included. These samples were chosen in favour of samples that include all decay channels to allow for a higher statistical precision in the selected phase space.

Finally, the three diboson processes are each represented by samples that contain both leptonic and hadronic final states.

The Monte Carlo samples were produced in the context of the Summer 2012 campaign with the version 5.3 of the CMS software which is also used in the context of this analysis. The following versions of the event generators are used: MadGraph v5.1.3.30 with PDF set CTEQ6L1 for the top-quark pair production and the vector boson samples in next-to-leading order. Powheg-Box v1.0 with PDF CTEQ6M was used for the single top samples, also in next-to-leading order. The process of associated Higgs-Boson production was produced with Pythia v6.426 in leading order. The processes of fragmentation and hadronization were also done with Pythia v6.426 with the underlying event tune Z2 star. The decay of $\tau$-leptons was simulated using Tauola.

The list of included Monte Carlo samples is given in table 5.2.

### 5.4 Event Selection

After applying the trigger preselection, events are further selected based on the primary vertex. The position of the vertex has to have a distance less than 24 cm in longitudinal and
less than 2 cm in radial direction from the nominal interaction point. A precisely calculated primary vertex is important since several of the algorithms in the event reconstruction such as the b-tagging are based on the position of the primary vertex. After this step, a kinematic selection is applied to further reduce the amount of background process events. This selection generally requires a lepton to suppress the QCD-multijet background and in addition a minimal number of jets, some of which are tagged as bottom quark jets, and is based on the aforementioned lepton and jet IDs.

5.4.1 Lepton Selection

The lepton selection is based on a combination of selection steps that can be separated into a kinematic selection, a cut on the relative isolation of the lepton and additional step based on the lepton flavour. For each lepton a loose and tight selection are defined. Tight leptons have tighter cuts and are used in the analysis. Loose leptons are mostly used to reject further leptons in the event selection. The details to each step are given in the respective subsections and are provided by the respective physics analysis groups (PAGs), in this case by the top PAG[].

Electron ID

The electron ID contains requirements on the isolation, a kinematic selection and a multivariate selection:

- The **Relative electron isolation** evaluates the energy in a cone with radius 0.3 around the electron. The energy coming from photons, charged and neutral hadrons is added, while the pileup energy is subtracted. This sum is divided by the total transverse momentum of the electron. This variable is effective to separate prompt electrons from electrons that are created in showering processes. The formula for the isolation is given as follows:

  \[
  I_e = \frac{1}{p_T,e} \cdot \sum_{\text{charged}} p_T + \max \left( \sum_{\text{neutral}} p_T + \sum_{\gamma} p_T - \rho A_{\text{eff}} \right).
  \]  

- The **kinematic selection** is based on the rapidity and transverse momentum of the electron. Electrons are vetoed if \(|\eta| > 2.5\), since the electromagnetic calorimeter and the tracking system only reaches to this rapidity. Electrons are further vetoed if they are in the pseudorapidity range \(1.4442 < |\eta| < 1.5660\), since no ECAL crystals are placed in this area. For loose electrons, \(p_T \geq 10\) GeV/c, for tight electrons \(p_T \geq 30\) GeV/c is demanded.

- The **multivariate selection** contains parameters that describe the reconstruction quality of the electron, such as the number of hits in the track system or the amount of hadronic energy.

Muon ID

The muon selection also contains cuts on isolation and kinematics. Further selection steps are applied that consider the reconstruction quality of the muon track.

- The **Relative muon isolation** calculation is similar to that of electrons, except that the radius of the isolation is chosen to be 0.4.

- The **Kinematic selection** demands a \(p_T \geq 10\) GeV/c for loose and \(p_T \geq 30\) GeV/c for tight muons. Further, loose muons have to have \(|\eta| < 2.5\) and tight muons \(|\eta| < 2.1\).
Table 5.3: The list of different lepton IDs and the kinematic selections that are applied to them.

| lepton ID     | max. Iso | min $p_T$ / GeV/c | max $|\eta|$ |
|---------------|----------|-------------------|-------------|
| loose electron | 0.200    | 10                | 2.5         |
| tight electron | 0.100    | 30                | 2.5         |
| loose muon    | 0.200    | 10                | 2.5         |
| tight muon    | 0.120    | 30                | 2.1         |

- The **Reconstruction selection** evaluates the quality of the muon track reconstruction. For the muon to be selected it has to be a global muon, which guarantees a good track as the muon is reconstructed from both sides of the detector. For the tight muons, additional cuts regarding the muon system are applied. The muon has to have at least one hit in the muon system. Further, it has to have been measured in the pixel system and in at least five tracker layers. The muon track has to be matched to at least two stations of the muon system. Finally, the impact parameter of the muon with respect to the primary vertex has to be less than 2 mm in transverse and less than 5 mm in longitudinal direction.

The kinematic cuts on the respective lepton IDs are listed in table 5.3. In the semileptonic selection, each event has to have both a loose and tight lepton of the same flavour. Events with additional loose leptons are rejected. By using a $\tau$ veto, events that contain a decay of a W boson into a $\tau$ lepton are rejected. The lepton $p_T$ threshold is chosen to be at $30 \text{ GeV}/c$ since there exists a turn-on phase in which the trigger efficiency is less than 100%.

### 5.4.2 Jet Selection

The standard jets that are used in the analysis are anti-kt jets with a distance parameter $R = 0.5$. At first, charged hadron subtraction is applied to remove charged objects from pileup. Since leptons can be identified using the particle flow algorithm, they are removed from the set of clustered objects. The jets are then clustered using all remaining particle flow candidates in the event.

After the clustering step, the following quality criteria are applied to the jets: They must contain more than one particle, of which at least one must be a charged object. Further, the fraction of neutral, hadronic and charged electromagnetic energy must be less than 99%. Finally, the charged hadronic energy fraction must be larger than 1%.

Next, the jet energy corrections are applied. Finally, the kinematic selection of the jets requires $p_T \geq 30 \text{ GeV}/c$ and $|\eta| < 2.5$.

### 5.4.3 B-Tagging Selection

The selection of bottom quark initiated jets is based on the output of the Combined Secondary Vertex algorithm which is explained in more detail in section 3.7.2. The semileptonic selection that is applied in the first two chapters requires at least two tagged jets at the medium working point which is at a b-tag output of 0.679 and corresponds to a $\approx 1\%$ mistag rate.

### 5.4.4 Fat Jet and Subjet Selection

CA 1.2 and 1.5 jets are clustered in the event without a lower kinematic threshold, a cut on the transverse momentum has to be applied to guarantee that the decay objects of the
hadronic top and the Higgs boson are sufficiently close and therefore contained in the fat jet. Therefore, the 1.5 CA jets have to have at least 200 GeV/c transverse momentum, while CA 1.2 jets have to have a $p_T$ of at least 180 GeV/c. A similar selection is applied on the subjets, which must have a transverse momentum of at least 20 GeV/c. An additional cut on the pseudorapidity is applied: The CA 1.2 and 1.5 jets have to have an $|\eta| < 2.0$, while for the subjets $|\eta| < 2.5$. For both fat jets and subjets, no jet energy corrections are applied.

5.5 Monte Carlo Event Corrections

While the simulated events that are generated by the Monte Carlo event generators generally describe the data well, further correction and rescaling steps are necessary to guarantee an optimal agreement. While most of these steps are performed during the reconstruction of the event that happens before the analysis, some additional corrections have to be applied during the actual analysis.

An important correction that is applied on the generated events is the reshaping of the pileup distribution, represented by the reconstructed number of primary vertices in the event. In the experiment, the pileup heavily depends on time-depending quantities such as the instantaneous luminosity. The Monte Carlo events however are created following a given pileup distribution and are later rescaled to match the data distribution. The correct description of the pileup is important since it has a sizeable effect on the missing transverse energy of the event.

Identification of bottom quarks is a vital component in this analysis. It is therefore necessary that the methods that identify the bottom quarks, called b-tags, have the same discrimination powers for both measured and simulated data. However, jets in data and Monte Carlo have different b-tagging efficiencies, which leads to a different number of b-tagged jets in the reconstruction. It is therefore necessary to rescale the distribution of the b-tag output in order to better match the experimental to the simulated data. In the case of this analysis, the scale factors are based on a tag-and-probe method.

The kinematic distributions of Monte Carlo events are simulated using leading-order or next-to-leading calculations. In general, these calculations provide the correct shape of the distributions and are later rescaled using the K-factor, which is the ratio of LO or NLO and NNLO production cross sections. However, in the case of the top quark, the effect of higher orders on the distribution shapes is non-negligible. It is therefore important to reweight the Monte Carlo events based on the experimental data.

5.5.1 Pileup Correction

The simultaneous inelastic scattering of proton pairs, also known as pileup, has a significant impact on the reconstruction quality of the event. In the event reconstruction, this quantity is represented by the number of primary vertices. To allow for a consistent event reconstruction, it is therefore necessary to ensure a good agreement of this variable for data and Monte Carlo events. As both the pileup and the instantaneous luminosity are dependent on the number and time distance of the proton bunches, these quantities are strongly correlated.

However, when producing the Monte Carlo samples, the distribution of the number of primary vertices has to be calculated without any knowledge of the luminosity. Therefore, it is generated using a predefined template. However, to ensure optimal agreement, a reweighting of the number of primary vertices in Monte Carlo is done. The scale factors are calculated by comparing the distribution of the number of vertices which is used for the
Monte Carlo production and the number of expected vertices in data which is calculated with the minimal bias cross section $\sigma_{mb} = 69$ mb.

The effect of the pileup correction is shown in figure 5.2. Here, the distribution of the reconstructed number of vertices is shown. The uncorrected distribution shows a large difference of the shape of data and Monte Carlo. The corrected distribution shows good agreement both in normalization and shape. However, the agreement of the shapes is not absolute, as it was the number of expected, not reconstructed primary vertices which was reweighted.

![Figure 5.2: Distribution of the number of primary vertices before (a) and after (b) the reweighting.](image)

The effect of the pileup reweighting can be seen in variables such as the missing transverse energy, which is presented in Fig. 5.3. While the non-reweighted distribution of MET already shows a good agreement, the reweighting improves this agreement even further, especially for low amounts of missing energy, and the difference in event numbers decreases roughly by half the amount.

![Figure 5.3: Distribution of the missing transverse energy without (a) and with (a) pileup reweighting.](image)

### 5.5.2 B-Tag Scale Factors

The measured differences in the distribution of the b-tag algorithm output make it necessary to reweight the distribution in generated events. The reweighting is based on the
calculation of an additional event weight based on the flavour and the kinematic properties in each event. This event weight is defined as the ratio of the tagging efficiency in recorded and generated events.

In a first procedure, the event weights were calculated to guarantee the same number of b-tagged jets for each working point. However, this somewhat rigid reweighting was susceptible in case of a change of the tagging algorithm or the position of the working points. Therefore, a second reweighting was executed which had the goal of matching the whole distribution of the tagging algorithm output.

The methods that were used to calculate the scale factors are described in more detail in the analysis note associated to [83].

This reweighting was done using tag-and-probe method in a dijet selection. By applying a Z boson mass veto, a phase space with high numbers of bottom quark jet pairs was selected. After tagging one jet as bottom quark using parton level information, the b-tag algorithm was extracted from the probe jet. This method was repeated for charm quark and light quark jets by applying the Z boson mask to enrich a sample with light jets. Finally, the method was repeated using recorded data. The scaling factors as a function of jet $p_T$ and $\eta$ were calculated by dividing the tag output distributions in the respective $p_T$ and $\eta$ bins.

![Figure 5.4: Distribution of b-tag output of the anti-kt jets with radius parameter 0.5 before (a) and after (b) the reweighting.](image)

The effects on the matching of data and Monte Carlo events can be seen in figure 5.4. While the differences in normalization are unchanged, there is a better agreement of the distribution shape.

The b-tag reweighting is also applied on the subjets that are clustered in the two subjet-algorithms. The effects of this reweighting are presented in the following chapter.

### 5.6 Event Yields

The following table shows the number of events after selecting events with exactly one lepton and at least four jets, of which two are b-tagged. This selection aims to select events which contain the $t\bar{t}$ process. The Monte Carlo event yields are calculated using the calculated cross section, the number of generated events and the integrated luminosity of the data sets. The yields correspond to a selection with at least four jets of which are two b-tagged.
### 5.6. Event Yields

<table>
<thead>
<tr>
<th>Category</th>
<th>Muon</th>
<th>Electron</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>71092</td>
<td>70210</td>
<td>141302</td>
</tr>
<tr>
<td>$t\bar{t}$ semileptonic</td>
<td>56429</td>
<td>54350</td>
<td>110779</td>
</tr>
<tr>
<td>$t\bar{t}$ dileptonic</td>
<td>5017</td>
<td>4827</td>
<td>9844</td>
</tr>
<tr>
<td>Single Top</td>
<td>3900</td>
<td>3791</td>
<td>7691</td>
</tr>
<tr>
<td>$W \rightarrow l+\nu +$ Jets</td>
<td>2041</td>
<td>1600</td>
<td>3641</td>
</tr>
<tr>
<td>$Z \rightarrow ll +$ Jets</td>
<td>250</td>
<td>457</td>
<td>707</td>
</tr>
<tr>
<td>Diboson</td>
<td>77</td>
<td>78</td>
<td>155</td>
</tr>
<tr>
<td>Total Bkg</td>
<td>67714</td>
<td>65102</td>
<td>132816</td>
</tr>
<tr>
<td>$t\bar{t}H$ inclusive</td>
<td>104</td>
<td>102</td>
<td>206</td>
</tr>
</tbody>
</table>
6. Validation of Jet Substructure Reconstruction Techniques Employed in the ttH Analysis in the LHC Run I

6.1 Introduction

In experimental particle physics, many analyses rely on good agreement between the data measured with detector experiments and the Monte Carlo simulation. However, good agreement is not guaranteed by default. To guarantee that the distributions of the relevant variables match not only in the absolute number of entries, but also in the shape of the distribution, often a variety of corrections, calibrations and reweighting steps of the simulation are necessary. These corrections are typically obtained by comparing the Monte Carlo simulation data to real data, e.g. in signal depleted control regions. For the measured data, it is necessary to understand the specific detector properties to be able to calibrate the measured attributes of the particles in the event. Data-Monte Carlo comparisons can often also help spotting effects like mis-calibration or instrumental noise. The knowledge of the detector properties is also vital for a correct simulation of the detector in the detector simulation.

In this chapter, the agreement between the measured data and the relevant simulated signal and background processes is studied. In the first section a selection of general distributions of event properties is presented. Since parameters such as jet multiplicity and number of b-tagged jets are used in the event selection and categorization in later analysis steps, it is necessary that the selection have an equal effect on the measured data and the generated data.

The main subject of this chapter is the validation of the subjet algorithms that are used to reconstruct both the Higgs boson and the top quark that decays into a hadronically decaying W boson. In the following sections, these objects will be referred to as Higgs and hadronic top. Each is reconstructed using an optimized subjet algorithm that first clusters all particle-flow candidates into Cambridge Aachen jets, c.f. section 3.6.2, with distance parameter $R = 1.2$ (Higgs) and $R = 1.5$ (hadronic top). These jets are further referred to as ‘fat jets’. The constituents of the fat jets are then declustered and reclustered into smaller Cambridge Aachen jets, further referred to as ‘subjets’, that represent the decay products of the hadronic top and Higgs. After the clustering of the fat jets and the subjets, tagging algorithms are used to select the fat jet which has the highest probability of containing the hadronic top and Higgs, respectively.
The algorithm used to reconstruct the hadronic top is the HEP-Toptagger [72]. In section 6.2.1, the distribution of kinematic quantities of the clustered fat jets are presented. This section also lists the selections that are applied on the fat jets. After the declustering and reclustering of the subjets, the jets are assigned to the decay objects of the top. Two different assignment methods and their efficiencies are compared. In the last section, two tagging algorithms are presented and the dependence of their performance on different parameters is studied.

To reconstruct the Higgs candidate, a similar approach is applied. The subjet-filterjet algorithm [76] is used to reconstruct Higgs boson candidates in the event. Again, the final choice of the Higgs candidate is done by a specialized Higgs tagger based on the specific kinematic properties of the filterjets in the CA 1.2 jet, c.f. section 6.3.2.

After the reconstruction of the hadronic top and Higgs jets and the selection of the respective candidates in the event, a boosted decision tree (BDT) is constructed. BDTs use an iterative algorithm that optimizes cuts on variables aiming at the optimal separation into a signal and background regime. This BDT uses variables of both subjet algorithms and the general event to separate the signal process from the main background, the production of a top quark pair with additional jets. Since the event signatures of these processes are very similar, it is both important to find discriminative distributions and to ensure that the modelling in measured and generated data is equivalent, since the decision trees are trained with the Monte Carlo datasets.

The Monte Carlo simulated physics processes that are discussed in this chapter are categorized as follows:

- **t\bar{t}H**: Represents the associated production of a Higgs boson and a top-quark pair.
- **EWK**: Represents production of W bosons and Z bosons with additional jets and the combined di-boson production channels (WW, WZ and ZZ).
- **Single Top**: Represents the s-channel, t-channel and associated top-W boson production for top quark and anti-quark.
- **t\bar{t} DL**: Represents the di-leptonic decay channel for top-quark pair-production.
- **t\bar{t} SL**: Represents the semi-leptonic decay channel for top-quark pair-production.

### 6.1.1 Event Preselection

The data that are studied in this analysis have been collected by either a single muon or single electron trigger. To reduce the absolute number of events and to reduce a large fraction of the known background events, a pre-selection was introduced which is applied on both measured and generated data. This selection, called *skimming*, requires:

- Exactly one lepton that fulfils the lepton ID *tight*, c.f. section 5.4.1. A veto on additional leptons that fulfil the lepton ID *loose* is applied.
- At least one jet that is b-tagged with the CSV algorithm at the medium working point.
- At least one CA jet with radius parameter $R = 1.2$, $p_T > 180 \text{ GeV}/c$ and $|\eta| < 2.0$.
- At least one CA jet with radius parameter $R = 1.5$, $p_T > 200 \text{ GeV}/c$ and $|\eta| < 2.0$.

Of the following, at least one has to be fulfilled for the event to be selected:

- At least four anti-kt jets with radius $R = 0.5$ that fulfil the tight jet ID.
- At least one Cambridge-Aachen (CA) jet with radius parameter $R = 1.2$, $p_T > 180 \text{ GeV}/c$ and $|\eta| < 2.0$.
- At least one CA jet with radius parameter $R = 1.5$, $p_T > 200 \text{ GeV}/c$ and $|\eta| < 2.0$. 

64
This selection suppresses processes like W-boson or Z-boson production as they rarely contain b-tagged jets. The selection of one isolated lepton with high \( p_T \) is effective at reducing QCD-multijet events. Events are selected if they contain at least one of the two fat jet types that fulfil the stated \( p_T \) threshold. If, however, neither of the two fat jets is present, a four-jet selection is applied.

### 6.1.2 Event Variables

The selection that is applied in all further sections of this chapter targets at suppressing the subdominant background processes besides the dominant semileptonic top-quark pair-production. The selected events are then used to evaluate the performance of the subjet algorithms, especially since the production of two top quarks with additional radiated jets has an event topology that is very similar to that of a signal event. The selection is an extension of the preselection that is used in the skimming process. By demanding at least four \( R = 0.5 \) anti-kt jets and two b-tagged jets, the fraction of W- and Z-boson production events and also single top processes is further suppressed.

The jet multiplicity is shown in Fig. 6.1a. The distribution shows good agreement over several orders of magnitude, although it is noticeable that there are generally \( \approx 4\% \) more events in measured than in simulated data. While there is a significant difference in the category of 8 and more jets, the statistical uncertainty of the Monte Carlo is large enough to cover the discrepancy.

The number of b-tagged jets, which is shown in Fig. 6.1b, also shows good agreement. The differences seen in the region with two tagged jets is likely to come from additional background QCD multijet processes that are not included.

Figure 6.1: Number of anti-kt jets (a) and b-tagged jets (b) in the combined electron and muon channel. The ratio between data and the sum of simulated events is shown at the bottom. Error bars denote the statistical uncertainty of the data.

Since the reconstruction of the hadronic top quark relies on a correct reconstruction of the associated W boson, good modelling of the missing transverse energy is important. The distribution of MET, which is shown in Fig. 6.2a, shows overall good agreement. The visible differences in regions of small missing transverse energy can to some extent be attributed to the missing contribution of the QCD multijet process. Further differences may come from differences in jet energy.

The scalar sum of the transverse energy of all jets, leptons and MET is presented in Fig. 6.2b. While the distribution shows good agreement, there is a shift of the data towards higher energies, which may be a result of differences in the jet energy.
6. Validation of Jet Substructure Reconstruction Techniques Employed in the ttH Analysis in the LHC Run I

(a) Distribution of the missing transverse energy.

(b) Scalar sum of the transverse momentum of jets, leptons and the missing transverse energy.

Figure 6.2: Distribution of MET (a) and the total transverse energy (b) in the combined electron and muon channel.

6.2 Performance of the HEP-Toptagger Algorithm

The HEP-Toptagger clusters Cambridge Aachen jets with a radius parameter of 1.5 from all particle candidates in the event. In the following, these jets are referred to as Topjets. From each of these jets, a set of subjets is clustered. The subjets are reclustered and additional radiation is removed. The final set of filtered subjets is called filterjets.

To guarantee that the initially clustered Topjets contain all decay products of the hadronic top quark, the radius of the jet clustering has to be large enough to allow for a wide spread. However, by increasing the jet radius, the amount of energy deposits inside the fat jet that originate from pileup or the underlying event also increases. The spread of the decay products can be decreased by setting a lower threshold on the transverse momentum of the hadronic top which is approximately the transverse momentum of the fat jet. Studies have shown that the aforementioned radius of 1.5 in combination with a minimal $p_T$ of 200 GeV/c for the fat jet yield a large fraction of jets coming from a top-quark decay. The properties of the raw fat jets, including multiplicity and kinematic properties are presented in section 6.2.1.

After the clustering, a hypothesis is constructed that assigns the three filterjets to the three partons coming from the hadronic top: The bottom quark and the two quarks coming from the hadronically decaying W boson. This assignment is based on one of two methods that chooses the filterjets based on certain kinematic properties. The four-vectors of the two W-boson jets are added to form the W boson and the sum of all three four-vectors is defined as the top-quark. In section 6.2.2, the two possible assignment schemes are presented and compared.

Finally, a tagging method is used to select the Topjet that is likeliest to originate from the top-quark. The top-tag is calculated using a likelihood method and the Topjet with the highest toptag is defined as the hadronic top candidate.

In order to test the reconstruction efficiency of the subjet algorithm and the assignment of the filterjets, generator level information is used in the Monte Carlo samples. The Topjet that corresponds to the hadronically decaying top is identified by calculating the distance in $R = (\eta, \phi)$ between the fat jet axis and the three identified partons from the hadronically decaying top quark. In the context of this thesis, a reconstructed top quark is called a true top if all three partons have a distance of $\Delta R = 1.5$ or less from the axis.
of the reconstructed top quark. If this is the case, the reconstruction of the hadronic top is theoretically possible, as the sub-jet algorithm only uses particles in the clustering that are inside the fat jet. All remaining top quarks are called false top quarks that come from combinatorial background. While this is a simplified approach to quantify the potential reconstruction, this information can be used to visualize the performance of the following processes. Since the identification of the hadronically decaying top quark is both best visible and most important in the top-quark pair-production channel, the separation into true and false tops is only applied to this process.

### 6.2.1 Kinematics of CA 1.5 Jets

The initial clustering of the CA 1.5 jets creates a variable number of Topjets which are used in the analysis. In Fig. 6.3, the number of Topjets per event is shown with a selection of four jets and two b-tagged jets applied. As these jets have a $p_T$ threshold of $200\text{ GeV/c}$, the majority of the events have no Topjets. While the shape of the distribution shows good agreement with data, there is a difference of 5.1% between measured and simulated events. A possible reason for this lack of events is that a QCD-multijet background sample was neglected. Further evidence for this assumption is that the normalisation improves for tighter selections as they are applied in the later stages of the reconstruction, which further suppresses the relative QCD contribution.

Figure 6.3: Number of CA 1.5 jets, called Topjets, with $p_T \geq 200\text{ GeV/c}$. A selection of four jets of which two are b-tagged is applied.

Several algorithms of the reconstruction of the hadronically decaying top quark, such as the b-tagging and the jet clustering, rely on information of the tracking system. Therefore, only Topjets with $|\eta| < 2.0$ are used. Furthermore, only those Topjets are selected in which the three subjets have been reconstructed. Finally, similar to the anti-kt selection, only those Topjets are used for which all subjets have $p_T \geq 20\text{ GeV/c}$ and $|\eta| < 2.5$.

The transverse momentum of the so selected Topjets is presented in Fig. 6.4a. The overall performance of the distribution is similar to that of the Topjet multiplicity: While the overall shape is similar, there are about 4% more measured events than simulated events. A possible reason for the observed differences is the lack of jet energy corrections for the fat jets and the subjets.

The distribution of the pseudorapidity $\eta$ is presented in Fig. 6.4b. In this case, the shape of the distribution is different in data and Monte Carlo based on the $\eta$-region. In the region $\eta < 0.9$, the agreement between data and Monte Carlo is good both in shape and
normalization. However, for larger rapidities, there is a higher number of events in data which also presents a different shape than the simulated data. Since jets that are produced in QCD-multijet processes are usually softer than processes like top-quark production, it is likely that a potential missing QCD background would appear in regions of higher $\eta$. Further, there are more areas in this region of the pseudorapidity in which jet energy cannot be measured due to non-instrumented zones in the detector, which could lead to a miscalculation of the Topjet $p_T$ and therefore a migration of the event.

Figure 6.4: Distribution of the $p_T$ (a) and $\eta$ (b) of Topjets. The Topjets are selected to have a $p_T > 200$ GeV/$c$ and $\eta < 2.0$. Further, only Topjets are selected for which all subjets have a $p_T \geq 20$ GeV/$c$ and $\eta < 2.5$.

The distribution of the energy of the Topjets, which is depicted in Fig. 6.5a, supports the assumption that the lack of jet energy corrections has an impact on the kinematic distributions. While the overall normalization is similar, the peak of the energy is shifted towards higher energies in measured data.

The mass distribution of the Topjets in Fig. 6.5b shows a peak at 200 GeV/$c^2$. Relating this peak to the mass resonance of the top quark suggests that there is a substantial amount of energy coming from other sources, such as pileup. While the overall shape of the distribution is similar in data and Monte Carlo, the distribution is slightly wider in data.

Figure 6.5: Distribution of the energy of Topjets (a) and their invariant mass (b). For selection details see Fig. 6.4.
6.2.2 Performance of Different Methods of Subjet Assignment

In the case of the hadronically decaying top quark, the final state consists of three jets, which is why the elements of the Topjet are clustered into exactly three subjets. However, the assignment of the jets to the three partons coming from the top quark is done after the clustering. In the context of this analysis, two different assignment approaches have been used which are presented in the following.

To estimate the assignment efficiency Monte Carlo information is used. To verify the jet assignment, the distance between the assigned bottom quark jet and the generator level bottom quark coming from the hadronic top is calculated. A jet assignment is considered correct, if the distance in $R(\eta, \phi)$ is less than 0.5.

The first method calculates all possible dijet masses and combines the two jets that are closest to the W-boson mass. By doing this, the reconstructed W-boson mass is strongly biased towards the true value, which can be seen in Fig. 6.6a. The overall distribution is centred around the mass resonance. This bias is also visible for false tops, as the selection of the two masses artificially improves the distribution of these Topjets. This makes it harder for the later separation steps, as the distribution of the combinatorial W-boson mass is very similar to that of the background.

The bias that is introduced via the mass selection is also visible in other distributions of the so-reconstructed W boson. The transverse momentum of the dijet combination which is shown in Fig. 6.6b shows a peaking behaviour when comparing signal and background and can also be seen in the combinatorial background in the top-pair process.

The second assignment method begins by sorting the three filterjets by their b-tag discrim-inator output. While the bottom quark jet that comes from the hadronic top usually has a high b-tag output, the two jets coming from the W boson in most cases have a light flavour and thus a lower b-tag output. Following this, the jet with the highest b-tag is assigned as the bottom quark jet. This b-tag output is shown in Fig. 6.7a. The distribution shows that there is a high probability that this jet stems from a bottom quark, as most of the entries have a discriminator output close to 1. The distribution shows good agreement,
the deviations between data and Monte Carlo in regions of low b-tag output values are covered by the statistical uncertainty.

The effect of this assignment on the diboson mass is presented in Fig. 6.7b. While the mass of the true tops is centered around the expected mass, the W-boson mass of the combinatorial background has a peak at a lower value. The distribution shows very good agreement which is supported by the flat ratio between data and Monte Carlo.

The width of the W-boson mass distribution is much larger in the cases where all partons are in the fat jet. This widening of the distribution has its origin in the declustering and reclustering of the filterjets. In both steps, objects are removed from the jets if they do not fulfill the required mass drop or if they appear to come from pileup or the underlying event. Finally, due to the high boost of the decay products of the hadronic top, it is possible that the distance between them is smaller than the clustering radius of the subjets inside the Topjet. This means that particles can be clustered into objects which they do not originate from. These effects can lead to a distortion of the kinematic properties. By applying the mass-biased combination, it is therefore possible to disregard the right combination in favour of a combination with a better mass.

The absolute number of correctly assigned bottom quark jets in both assignment options show that the assignment based on the highest b-tag output of the three subjets is more effective compared to the method based on the dijet W-boson mass. Further benefits of this method are that an artificial bias on the W-boson mass is avoided, which is introduced when choosing the dijet mass closest to the W-boson mass. This bias can lead to a wrong assignment, since a combinatorial dijet with a good W-boson mass can be closer than a correct assignment when its kinematic distributions are affected by clustering effects.

The benefits of a correct assignment only apply to the fat jets for which a correct reconstruction is already possible. As the clustering of the fat jet and the subjets is not varied, fat jets for which one or more of the partons or a fraction of the parton energy are not included in the fat jet, will still see kinematic effects based on the clustering method.

6.2.3 Hadronic Top Tagging Method and Efficiency

After the clustering step and the assignment of the filterjets, the event now contains one or more reconstructed hadronic tops. In the tagging step, the top candidate is selected.
This candidate is regarded as the true hadronic top for the remaining analysis. For the tagging, a likelihood method based on histograms is employed.

The following distributions again use Monte Carlo information to illustrate the separation power of the input parameters. In this case, the semileptonic top pair production sample is split into two categories that verify whether all partons from the hadronically decaying top quark have $\Delta R < 1.5$ to the CA 1.5 fat jet. If this is the case, the jet is called a true top. In the case where this condition is not met, the jet is labelled combinatorial background.

The likelihood method is based on the evaluation of three kinematic variables of the hadronic top-system:

- $M_3$: The first variable is the invariant mass of the three subjets. Due to its resonant behaviour the variable $M_3$ is effective at separating true from false tops. However, since the filterjets are chosen by their closeness to the top mass and due to the filtering and reclustering of the jets, it is possible for combinatorial jets to form a mass close to the top mass.

- $M_W/M_3$: The second variable is the ratio of the W mass to the aforementioned three-jet mass. Like the top mass, the distribution of the W mass shows a distinct resonant behaviour which can be used as an effective discriminator. The mass ratio has been chosen in favour of using the W-boson mass by itself, as it is possible that the W mass spectrum is distorted. This can happen by adding objects that do not come from the W boson in the reclustering or by removing soft radiated objects in the jet filtering. Since the three jet system as a whole has undergone the same procedure, it is reasonable to assume that the top mass has been shifted in a similar way, leading to a constant ratio.

- $\arctan\left(\frac{M_{B+W_1}}{M_{B+W_2}}\right)$: The third variable is the arctangent of the ratio of the two dijet masses which combine the four-vector of the jet with highest b-tag output and one of the two remaining $p_T$ sorted jets.

The distributions of the trijet mass and the dijet W-boson mass are shown in Fig. 6.8. The category that represents the true tops shows a clear peak that is centered at a value slightly lower than the known top mass. Compared to that, the combinatorial background has a much broader distribution with a wide peak at around 115 GeV/$c^2$. While the overall accordance between measured and generated events is good, there are more events in data for lower tri-jet masses, while for higher masses there is a surplus of Monte Carlo events.

A similar separation effect can be seen in the distribution of the reconstructed dijet mass that is associated with the hadronic W boson which is presented in Fig. 6.7b. Again, the category that represents the true tops shows a peaking behaviour slightly below the W-boson mass, where the combinatorial background is broadened over a large spectrum. While the slope in the ratio is again visible, it is not as distinct as in the top mass.

The separation of the two categories is also visible in the ratio of the two masses which is presented in Fig. 6.9a.

It is further noticeable that there is good agreement in the shape of the distribution, which makes it likely that the effects which lead to a slight gradient in the hadronic top and W-boson mass have their origin in the subjet clustering process.

The final parameter which is used in the likelihood is the arctangent of the two alternative dijet combinations and is depicted in Fig. 6.9b. Like the previous parameters, the distribution shows very good agreement between data and Monte Carlo.
6. Validation of Jet Substructure Reconstruction Techniques Employed in the $ttH$ Analysis in the LHC Run I

(a) Invariant mass of the combined four-vector of all three subjets.

(b) Reconstructed $W$-boson mass. The subjets are assigned using the b-tag oriented method.

Figure 6.8: Reconstructed hadronic top and $W$-boson mass. To emphasize the separation power of the parameters, the semileptonic top-pair production channel is split into two categories based on Monte Carlo information.

(a) Ratio of the reconstructed $W$ boson and hadronic top mass.

(b) Arctangent of the mass-ratio of the bottom quark jet and either of the two hadronic $W$ boson jets.

Figure 6.9: Additional top-likelihood variables. To emphasize the separation power of the parameters, the semileptonic top-pair production channel is split into two categories based on Monte Carlo information.
The likelihood is calculated with a formula that uses the probability of the Topjet to belong to the category of true tops or the combinatorial background, a separation which is defined based on the distance in $R = (\eta, \phi)$ between the top quark parton from the hadronically decaying top quark on the generator level and the particular Topjet. If the distance between these objects is smaller than 1.5, a reconstruction within the Topjet is theoretically possible. In case the parton has a larger distance, the Topjet is regarded as background.

For these two categories, a normalized histogram for each of the three input variables is constructed. In the analysis, for each Topjet, the value of the three variables $V_i$ is interpolated in the histogram of signal ($S$) and background ($B$) to find the histogram bin in each category. The relative fraction $P(x, S/B)$ of this bin is then used.

The discriminator is defined as follows:

$$D = 2L - 1, \quad \text{with}$$

$$L = \frac{P(V_1, S) \cdot P(V_2, S) \cdot P(V_3, S)}{P(V_1, S) \cdot P(V_2, S) \cdot P(V_3, S) + P(V_1, B) \cdot P(V_2, B) \cdot P(V_3, B)} = \frac{L(S)}{L(S + B)} \quad (6.2)$$

The Topjet with the highest tag is selected as the hadronic top candidate. The distribution of the highest tag in each event is presented in Fig. 6.10. It shows overall good agreement between measured and simulated data. The two categories, in which the semileptonic top-quark anti-quark process is separated, show that the majority of the presented Topjets do not come from combinatorial background, which shows that the likelihood is able to effectively select the Topjets for which a reconstruction is possible.

![Figure 6.10: Output of the likelihood tag for the hadronic top candidate, which is the Topjet with the highest likelihood output in each event. To emphasize the separation power of the parameters, the semileptonic top-pair production channel is split into two categories which represent combinatorial background and Topjets that contain the decay objects from the top quark.](image)

To quantify the performance of the likelihood tag, a comparison between this likelihood based tag and a simple mass-based likelihood was done. The mass-based discriminator is calculated using the following formula:

$$D = \left(1 - \frac{|M_3 - 173.2\, \text{GeV}/c^2|}{M_3 + 173.2\, \text{GeV}/c^2}\right) \cdot \left(1 - \frac{|M_W - 82.4\, \text{GeV}/c^2|}{M_W + 82.4\, \text{GeV}/c^2}\right) \quad (6.3)$$
The likelihoods are compared for all Topjets that fulfil the kinematic selection given at the beginning of the section. All Monte Carlo samples are used and are scaled for the luminosity used in data. The signal and background regions are split using the distance between the top quark and the Topjet in \( R = (\eta, \phi) \): If the distance is smaller than 1.5, the reconstruction is theoretically possible. All Topjets that fulfil this selection are considered signal, all others are considered background.

The distribution of the two likelihoods is shown in Fig. 6.11. Both distributions show good agreement in shape and normalization. Further, both likelihoods are efficient at separating the signal and background parts into different regions. However, the shapes of both signal and background are different. Where in the Toptagger the majority of the signal is focused towards the signal region, the background is very uniformly distributed. In the alternative likelihood, the signal is not as strongly pushed towards the signal region, but the distribution of the background allows for good discrimination while keeping a majority of the signal events.

To quantify the separation power of the two discriminators, the receiver-operator characteristics (ROC) curve is calculated. This curve plots the signal efficiency, defined as the fraction of remaining signal events as a function of the likelihood threshold against the background rejection, which is calculated as \( 1 - \text{background efficiency} \). The performance of a discriminator can be estimated using these two different quantities. It is important to have a good background rejection while having the best signal efficiency possible. Further, the behaviour at very low signal efficiencies and high background rejection is important.

The ROC curve is calculated by using all Monte Carlo samples which are scaled to represent their relative contribution based on production cross section, as it is shown in Fig. 6.11. The resulting curves are presented in Fig. 6.12. When comparing the quality criteria mentioned above, it is apparent that the two likelihoods have very similar integrals and similar distributions, as they use very similar input variables. It is further interesting to note that the comparison likelihood which is not as optimized as the current Toptagger likelihood already shows a similar signal efficiency and background rejection in the relevant
area. While the Toptagger has since been further improved upon, it is still interesting to further study the performance of the second likelihood.

6.3 Studies on the Subjet-Filterjet Algorithm

The reconstruction of the Higgs boson using the subjet-filterjet algorithm is the second task in the analysis of the signal process. However, compared to the reconstruction of the hadronic top quark, there are additional challenges that arise in the reconstruction of the Higgs boson. The most important is the number of background Higgsjets. As the distributions in the previous sections have shown, it is possible to enrich the data with top-quark events by applying a simple selection based on jet and b-tag multiplicity. However, due to the very small production cross section and the large similarities to the top-quark pair-production process, such an enriching process is not applicable for the Higgs decay.

The reconstruction process begins with the clustering of CA 1.2 jets which are referred to as Higgsjets in the following. Of these, only those with a transverse momentum of at least 180 GeV/c are used in the following selections. On these jets, a declustering process similar to the hadronic top reconstruction is applied. However, in this case no mass-drop based removal is applied in the declustering. The declustering phase is terminated when a significant mass drop has occurred. In the context of a Higgs decay, this mass drop signifies the splitting of the Higgs boson into two bottom quark jets. After the declustering is done, a variable number of CA 0.3 jets is clustered from the particle flow candidates which are then cleaned from particles that come from the underlying event. The resulting jets, called filterjets, are assigned to the decay products of the Higgs boson. The two filterjets with the highest b-tag output assigned as the bottom quark jets, while a potential third filterjet or the filterjet with the highest transverse momentum in the case of four or more is assigned to additional gluon radiation.

6.3.1 Kinematics of Higgsjets and Filterjets

Similar to the HEP-Toptagger, the subjet-filterjet algorithm starts by clustering CA 1.2 jets of all particle flow candidates in the event. The number of these Higgsjets per event is shown in Fig. 6.13. While the shape of the distribution is very similar which can be seen by the flat ratio, there is an offset of 6% in event numbers.

The signal process has nearly as many events in the category of one or two Higgsjets as in the category with no Higgsjets. In this process, the inclusion of all decay channels of the
top-quark pair and the Higgs boson can lead to all objects decaying hadronically, which leads to a large number of Higgsjets which fulfil the kinematic selection. The top-quark channels (dileptonic and semileptonic topquark pair-production, single top) show a falling spectrum. While there are few events in the electroweak channel due to the restrictive selection, most of them are in the category without any Higgsjets, as only the leptonic decay channels are included and therefore the objects which are clustered into jets have to come from additional jet radiations.

The distribution of the $p_T$ of Higgsjets which is presented in Fig 6.14a shows a falling spectrum which is similar to that of anti-kt jets. The majority of the Higgsjets is from the semileptonic decaying top-quark pair-production. The pseudorapidity $\eta$, shown in Fig. 6.14b, shows that most of these jets are produced very centrally, which is expected due to their high transverse momentum. While the shape of the distribution agrees well for $p_T$, the distribution of $\eta$ shows a comparable behaviour to that of Topjets, c.f. section 6.2.1. The normalization has an offset of $\approx 5\%$.

Figure 6.14: Distribution of the transverse momentum (a) and pseudorapidity (b) of Higgsjets. The Higgsjets are selected to have a $p_T \geq 180\text{ GeV}/c$ and $|\eta| < 2.0$. Further, only Higgsjets are selected for which all subjets have a $p_T \geq 20\text{ GeV}/c$ and $|\eta| < 2.5$. 
The decay of a Higgs boson into two bottom quarks leads to two reconstructable jets. Therefore, an obvious variable in the reconstruction of the Higgs boson with the subjet algorithm is the dijet mass of the two subjets with highest \( p_T \). This variable is presented in Fig. 6.15a. The distribution shows a peak near the W-boson mass, with the majority of entries from the semileptonic top-quark pair-production. Since Higgsjets and Topjets are clustered with similar conditions (same clustering algorithm, similar radius parameter), it is likely that the hadronically decaying top quark is clustered as a Higgsjet which can lead to the reconstruction of the hadronically decaying W boson inside the Higgsjet. However, this false reconstruction is mitigated, as the candidate for the hadronically decaying top quark is removed from the event before choosing the Higgs boson candidate. While the distribution above the W-boson mass shows good agreement in shape with a difference in entries of \( \approx 5\% \), the agreement below the W-boson mass is worse. Whether this difference comes from different jet energies scales in data and Monte Carlo simulation or whether it is an effect of the jet clustering is not clear.

The decay of the Higgs boson into two bottom quarks leads to two jets with expected high b-tag output. Additional jets that are reconstructed are likely to come from gluon radiation in the final state. The second distribution, Fig. 6.15b, shows the third highest b-tag output of all subjets in the case where such a jet exists. The distribution is expected to be focused towards very low values of the b-tag output, which is confirmed in the shown distribution. However, as most of these jets come from the topquark pair-production, most of these jets are probably radiated from the hadronically decaying W boson as final state radiation. However, in this case the same argument holds.

![Distribution of dijet mass of the two leading subjets (a) and the third highest subjet b-tag output (b).](image)

(a) Dijet invariant mass of hardest two subjets.  
(b) Third highest subjet b-tag output.

Figure 6.15: Distribution of dijet mass of the two leading subjets (a) and the third highest subjet b-tag output (b). For selection details see Fig. 6.15.

### 6.3.2 Tagging Methods and Efficiency

Like in the case of the hadronic top, a tagging method is applied to choose the Higgs candidate out of the set of CA 1.2 jets. Similar to the likelihood tag that was applied on the CA 1.5 jets, the Higgs tagging algorithm tries to exploit the kinematic properties of the decay products of the Higgs boson, in this case the two bottom quark jets. Here, the distribution of the second highest b-tag of the reconstructed filterjets is used as the tag output. In each event, the CA 1.2 with the highest tag output is chosen as the Higgs candidate. The distribution of the second highest b-tag of the so-chosen Higgs event candidate is shown in Fig. 6.16.

Similar to the reconstruction of the hadronically decaying top quark, again two different likelihood discriminators are compared. The first is the above presented second highest
b-tag discriminator output of the subjets of all Higgsjets, the second is a mass-based discriminator which uses the reconstructed di-jet mass of the Higgsjet subjets:

\[ D = 1 - \frac{|M_2 - 125 \text{GeV}/c^2|}{M_2 + 125 \text{GeV}/c^2} \]  

(6.4)

The distribution of the two likelihoods is presented in Fig. 6.17. As the production cross section of the signal process is much smaller than that of the top-pair production, the number of events which are in the signal category is much smaller. It is interesting to note, that while the mass-based likelihood has a better focus of the signal Higgsjets towards the signal region, the background Higgsjets are also to a large fraction in the signal region.

In the b-tag based likelihood, both signal and background jets are distributed both in signal and background region, however in this case the number of background Higgsjets in the background region is much higher. It is therefore to be expected that this discriminator has a better separation ability.

The comparison of the two discriminators is again done by comparing the receiver-operator curves, which is shown in Fig. 6.18. In this case, the differences between the two likelihoods are much more visible. The b-tag output based likelihood shows an overall better performance.

### 6.4 Validation of BDT Input Parameters

The final step in the analysis employs a boosted decision tree (BDT) to separate the signal process from the top-quark pair-production backgrounds. A boosted decision tree is a multivariate analysis tool that uses a series of selections to separate a signal and several background processes. BDTs are trained by iteratively changing the thresholds of each selection variable to separate a predefined signal and background sample. Afterwards, the BDT is tested on a second sample.

In the final analysis, the BDT is applied on the measured data events. The validation of the input variables that are used in the BDT is therefore very important, as both the selection of the input variables and the thresholds that are chosen in the training steps should have equal results.
6.4. Validation of BDT Input Parameters

(a) Second highest b-tag output of Higgsjet (b) Output of the alternate likelihood based on the Higgs boson mass.

Figure 6.17: Distribution of the discriminator output based on the b-tag output (a) and the mass-based discriminator (b). The Monte Carlo consists of the sum of all processes weighted with the respective scale factors. The split into a signal and background region is based on the distance in $R = (\eta, \phi)$ between the Higgs boson and the Higgsjet. A Higgsjet is put into the signal region if the distance is less than 1.2, which makes a reconstruction of the Higgs boson theoretically possible.

Figure 6.18: Receiver-Operator characteristics curve for both likelihood discriminators. The curves refer to the B-tag output based likelihood and the Higgs mass based reference likelihood, respectively.
The most important background process is top-pair production with semileptonic decay with additional pairs of jets coming from gluon radiation. Especially the case, in which the additional jets are bottom quark jets, is a nearly indistinguishable background, which is why it is very important to train the BDT for optimal separation power.

The BDT that was trained with the Monte Carlo samples of LHC Run I uses the following variables:

- General event variables such as the sphericity[3], aplanarity[3] or the total transverse energy.
- Variables of the anti-kt jets in the event, among them the mass or dijet mass of b-tagged jets, the dijet mass of the two closest tagged jets or the distance in $\eta$ or in $R = (\eta, \phi)$ between jets or jets and leptons.
- Variables of the Higgs event candidate, such as the two highest b-tag discriminator outputs or the dijet mass of the subjets with highest $p_T$.
- Variables of the hadronic top event candidate, among them the masses of the di-subjet combinations, the $p_T$ of the subjets or the trijet mass.

The distribution of the resulting BDT is presented in Fig. 6.19. The overall agreement between data and Monte Carlo is good, both in shape and normalization. As it can be seen, even after a specific optimization, there is no clear separation in signal and background, as the reconstructed objects, such as the hadronically decaying top, are very similar to those of the signal process. However, the further optimization of the hadronic top candidate and the Higgs candidate, for instance by applying a selection on the respective likelihood output, can improve the separation at the cost of signal efficiency.

![Figure 6.19: Distribution of the output of the BDT that is used to separate the signal process from the dominant background processes such as the top-quark pair production with semileptonic decay channel and additional bottom quarks.](image-url)
7. Correction of B-Tag Distribution of Filtered Subjets

7.1 Motivation

In order to optimize the agreement between measured and simulated data, several correction steps have been applied to the Monte Carlo events. These corrections include the reweighting of Monte Carlo events based on the output of the b-tagging algorithm for anti-kt jets. The correction factors of the b-tag reweighting are based on anti-kt jets with a radius of 0.5 which corresponds to the standard jet algorithm used at the CMS experiment. The goal of the reweighting is correct the MC simulation such that the distribution of the b-tag output matches the one observed in data. This is necessary the b-tagging output is often used as a selection criteria.

The implementation of the two fat jet algorithms leads to a re-clustering of the event into different jet objects. The jet clustering is inclusive, which means that each set of jets coming from either algorithm is a representation of the event since they contain all reconstructed particles (except, in the case of fat jets, the charged leptons). For the b-tagging of both anti-kt and subjets, the aforementioned combined secondary vertex algorithm is used. This algorithm combines all tracks within a specified a distance around the jet center and, based on a multivariate analysis, calculates the probability that the jet stems from a bottom quark.

In this analysis, the aforementioned anti-kt jets and the subjets are used in combination. It is therefore studied which effect the jet radius or the clustering algorithm has on the resulting properties and the b-tagging output. This effects can be observed when comparing anti-kt jets and subjets that are clustered using a similar set of tracks and particle flow candidates, respectively. To do so, a distance based matching procedure between anti-kt jets and subjets is performed. The so-created pairs of jets are compared and studied for correlation.

The event weight that is used in the b-tagging reweighting is calculated using the selected anti-kt jets. However, in this analysis it is both the b-tag output of subjets and anti-kt jets that needs to be corrected via these correction factors. In this case, it was assumed that the correction factors correct the b-tag output distribution of the subjets the same way as the standard anti-kt b-tag output. The questions is now whether this assumption is true and whether an inclusion of additional subjets in the calculation is necessary.
It is therefore further studied whether there are subjets in the event which have to be used in addition to the anti-kt jets in the calculation of the scale factors. However, as the b-tag output of these jets has to be uncorrelated to the jets already included in the calculation, additional jets would have to be non-overlapping to these. These jets come from the set of subjets that were not matched to anti-kt jets and are also studied for correlated behaviour.

The structure of this chapter is as follows:

The method of jet matching is introduced in section 7.2. Further, the properties of matched and unmatched jets are presented.

Comparing the properties of subjets to their respectively matched anti-kt jets is done in section 7.2.3.

In the last section, the properties of unmatched subjets are compared to that of anti-kt jets based on the distance between the jets, to test whether they are really uncorrelated. Further, the b-tag scale factors based on the inclusion of these additional subjets in the calculation is presented. Finally, the event yields for the standard and alternative scale factors are compared.

As the reweighting is only applied to Monte Carlo events, measured events are not included in the jet matching. Furthermore, only samples of processes with several high-mass objects have enough energy to consistently produce fat jets. Therefore only the signal sample and the sample representing the semileptonic decay of the top-quark pair production were included in the studies. In the figures, the Monte Carlo processes are represented by the same labels as in the last chapter. Further, all subjets have to have $p_T \geq 30$ GeV/c as the scale factors use this value as the lower threshold for the calculation.

7.2 Matching of Subjets to ak5-jets

This section describes the method by which the anti-kt jets are matched to the subjets. The employment of the anti-kt algorithm in combination to the two subjet-algorithms leads to three sets of jets. While the subjet-algorithms both cluster the particles into Cambridge-Aachen jets, the preceding declustering and reclustering results in subjets with different kinematic properties. It is therefore possible to individually compare the subjets to the anti-kt jets, however this is not done as the final analysis also uses these two sets of jets in parallel.

The subjet matching consists of two steps. At first, the set of subjets that are used in the matching is reduced by discarding overlapping subjets to avoid double-counting and correlation between the subjets. Next, the subjets are matched to anti-kt jets based on a variable maximal distance. The resulting comparisons between matched jets in terms of b-tagging and kinematic distributions is shown in the last section.

7.2.1 Selection of Subjets

The calculation of the b-tagging likelihood takes into account all tracks within a distance in $R(\eta, \phi)$ of 0.3. For anti-kt jets with a radius of 0.5 this usually means that all tracks are only assigned once and that there is no double-counting that could resolve in a correlation of b-tags. In the case of the subjets however, the Cambridge-Aachen algorithm in combination with the smaller radius of the jets can lead to a correlation of the b-tag of nearby jets. The fact that subjets from both subjet-algorithms are studied, which are likely be in the same $(\eta, \phi)$ space due to similar clustering premises, makes it necessary to remove subjets so that they no longer overlap.

Therefore, in the first step all overlapping subjets are removed from the set of studied subjets. The subjets are sorted by $p_T$ and afterwards, for each remaining jet, all other subjets are discarded that have a $\Delta R$ of 0.5 or less.
7.2.2 Distance Based Jet Matching

After the isolation of the subjets, they are assigned to the anti-kt jets. The matching is based on the $\Delta R$ between the two jets and uses the axis of the anti-kt jet as starting point.

The first idea that was studied was whether to match only a single subjet to each anti-kt jet or to match all subjets with a distance smaller than the defined maximal distance. This study was based on the fact that due to the small jet radius of the subjets it is possible to fit several subjets into a single anti-kt jet. However, the matching of all possible subjets leads to the question how to compare the b-tag of the jets. Since it is impossible to simply add the b-tagging results, the b-tags of all matched subjets would have to be compared separately. This would result in potential partial overlaps between matched jets which would distort the overall correlation. Therefore, it was decided to match only one subjet to every anti-kt jet.

The second parameter that is outlined further is the maximal matching distance between the jets. Since the b-tagging algorithm only uses objects within a cone of 0.5, the maximal radius of matching that is sensible is also 0.5. Nevertheless, it is also interesting to study the effects of smaller maximal matching distances on the correlation of the jets. In this thesis, the values of 0.1 and 0.3 are studied in addition to the value of 0.5.

In the jet matching, for each anti-kt jet in the event, sorted by $p_T$, the subjet closest to it is chosen out of the possible subjets that are closer than the aforementioned maximal matching distance. To avoid matching a single subjet to several anti-kt jets, all matched subjets are removed from the list of possible matches.

To determine the maximal matching distance $\Delta R_{max}$, the distance between the remaining unmatched subjets and the closest anti-kt jet was examined. The results are shown in Fig. 7.1. It can be seen that for the closest $\Delta R_{max} = 0.1$, there are still many subjets with a $\Delta R$ of less than 0.5, which means that these jets at least partially overlap. For the values of 0.3 and 0.5, the number of unmatched subjets which fall in this category is already much smaller. The minimal distance between an unmatched subjet and an anti-kt jet is also higher, however it is still smaller than the maximal distance at which the jet matching was performed. This is due to the fact that there are unmatched subjets with a distance smaller than $\Delta R_{max}$ to an anti-kt jet which has already been matched.
Since the matching with a maximal matching distance of 0.1 resulted in a large amount of unmatched jets with large overlap to anti-kt jets, this option was discarded. Further, the difference between $\Delta R_{\text{max}} = 0.3$ and $\Delta R_{\text{max}} = 0.5$ are marginal, which leads to the decision of using $\Delta R_{\text{max}} = 0.3$ in the following stages. This intermediate value is a compromise between matching jets with a decent overlap while avoiding unmatched subjets that are close to anti-kt jets.

### 7.2.3 Comparison between Matched and Unmatched Subjets

The number of subjets that were matched is shown in 7.2a. While there are generally very few events without matched jets, the multiplicity shows a clear difference between the two samples. In the semileptonic top-pair process, most events have between one and three matched jets, since in this case there is usually only the top quark which can lead to a fat jet with such a high transverse momentum. In the signal sample, there are two objects that decay into jets and therefore one observes a shifted distribution towards more matched jets.

The number of unmatched subjets, which is presented in Fig. 7.2b, shows that between 70 and 80% of events, depending on the process, have no unmatched jets. However, in the case of the signal sample, there is a higher fraction of events that have at least one unmatched jet. This is again due to the higher number of objects that decay hadronically, as the Higgs boson can decay into pairs of jets.

The distribution of the transverse momentum for matched and unmatched jets is shown in Fig. 7.3. While the majority of the unmatched jets have transverse momenta just above the threshold, the distribution of matched jets is shifted more towards higher momenta. However, it is important to note that the jets with the highest $p_T$ were selected for the matching, as described in section 7.2.1. One of the reasons that these jets are not matched is that the anti-kt jet that was matched from the particle flow candidates has a $p_T < 30$ GeV/c due to the applied jet energy corrections and is therefore not included in the analysis.

The distribution of the jet flavour is distinct for both matched and unmatched jets. Further differences can be seen when comparing the two different processes. In the case of the signal sample, about 50% of all matched jets are bottom quark jets. Bottom quark jets are the
7.2. Matching of Subjets to ak5-jets

The majority of the matched jets, as there are up to four bottom quark jets in the leading order process. A further 17% are gluon jets. These come from either initial state or final state radiations, for instance from the top quarks. In the case of the semileptonic top-quark pair production sample, the fraction of bottom quark jets is smaller, whereas the fraction of light or gluon flavour jets increases. This is also expected since the signal process contains all decay channels of the top quark, while in the case of the $t\bar{t}$ process, one of the top quarks always decays hadronically into light flavour jets.

Unmatched jets mostly consist of gluon jets. As with the matched gluon jets, these also represent initial and final state radiations. It is further noticeable that there are about twice as many unmatched bottom quark jets in the signal sample due to the generally higher number of bottom quark jets in the final state.

The distribution of the b-tag output reflects the behaviour of the flavour and $p_T$ of both matched and unmatched jets. The b-tag output of matched jets is very well discriminated between light and heavy flavour. A large fraction of the jets have a b-tag output close to one, corresponding to the flavour distribution that is shown above.
Unmatched jets have mostly small b-tag outputs, the latter is expected as these jets are mostly gluon jets.

Figure 7.5: B-Tag output of matched subjets (a) and unmatched subjets (b). For further details compare 7.2.

The result of this comparison is that the jet matching manages to match jets which have mostly high $p_T$ and heavy flavour. This behaviour is expected, since the anti-kt algorithm most likely also clusters jets in these areas. Unmatched jets have mostly low $p_T$ and light flavour and therefore are likely to represent additionally radiated jets.

### 7.3 Comparison of Matched Subjets and Anti-kt Jets

In this section, the properties of the matched jets, which were presented in the previous section, are now compared with those of the associated anti-kt jets. Each parameter is compared for two cases: In the first case, a maximal distance for the jet matching in $R(\eta, \phi)$ of 0.3 is chosen. In the second case, this distance is reduced to 0.1 to compare the properties in the case of a nearly complete overlap.

#### 7.3.1 Jet Flavour of Matched Jets

In Fig. 7.6a, the distribution of the flavour of the matched subjet is plotted against the flavour of the anti-kt jet it is matched with for a distance of 0.3 or less. It can be seen that in most cases, these flavours coincide, which is to be expected since the process of assigning the flavour to the jet is also done via a distance based method. In the second case, presented in Fig. 7.6b, the maximal distance is reduced to 0.1, which means that the two jets now overlap to an even larger degree. In this case, the fraction of cases in which the flavour of the two jets do not agree is even smaller, which shows that the flavour identification is effective at finding the jet flavour independent of clustering algorithm and jet size. This agreement has large relevance since the jets which are used for the calculation of the b-tag scale factors are separated by flavour, which makes a good flavour-matching very important.

#### 7.3.2 Kinematic Variables of Matched Jets

The distribution of $p_T$ of matched jets is shown in Fig. 7.7. In the first case, which corresponds to the larger maximal matching distance, the majority of jets have a similar $p_T$ when below a value of $\approx$200 GeV/c. However, there are also combinations in which the anti-kt jet has a higher $p_T$ than the matched subjets. There are two reasons for this behaviour: First, the distance between the jet centers leads to different clustered objects.
Second, the larger jet radius of the anti-kt jets can lead to additional objects being clustered into the jet. However, the opposite can also happen, since there are also subjets with a larger $p_T$ than the matched anti-kt jet.

When decreasing the maximal distance between the jets to 0.1, the correlation between the $p_T$ becomes even more apparent. In this case, the majority of jet pairs have a very similar $p_T$. If there is a deviation of the $p_T$, it is mostly towards higher $p_T$ of the matched anti-kt jet which hints that the additional energy is coming from the larger jet area.

The distribution of the energy of the two matched jets, which is shown in Fig 7.8, shows a very similar shape compared to that of the $p_T$. A linear correlation between the two quantities is visible, although it is not as strong as for the $p_T$. This similarity is also visible when decreasing the maximal matching distance to 0.1. Again, there is a fraction of entries where the anti-kt jet has a larger energy than the matched subject. For these cases it is once again logical to assume that the energy came from additional particle flow candidates outside the radius of the subjet.
7.3.3 B-Tag Output of Matched Jets

The distribution of the b-tag output of both the subjet and the associated anti-kt jet are shown in Fig. 7.9. Overall, the majority of entries are in the lower left and upper right corner. That’s because, as stated in the previous section, the majority of these jets are either gluon or bottom quark flavour jets, which have an expected very low or very high b-tag output, respectively. It is also noticeable that there is a visible positive correlation between the two b-tag outputs with a small widening at the corners of the distribution. Considering the fact that the comparison of the jet flavours showed that there is a very good agreement between the flavour, it can be stated that the b-tag algorithm is able to discriminate between different jet flavours independent of the clustering algorithm and jet size parameter.

In the second figure, the distance between the jets is reduced even further. The effect of this reduction can be seen at several points in the distribution. First, the widening at the lower left and upper right corner was reduced. In these cases, the b-tag output of jets which had a distance between 0.1 and 0.3 had a slight difference due to the different tracks which were used for the calculation of the discriminator output. Second, the number of entries at the upper left and lower right corner which represents completely different b-tag output values has decreased even further, even though this region already had a negligible contribution. Most importantly, though, the relative number of pairs with very similar b-tag output values has stayed almost unchanged. All in all, the b-tag algorithm has a very similar discrimination power independent of the jet algorithm which means that the probability of identifying a jet as coming from a bottom quark is the same. Further, the assumptions which were made at several stages in the reconstruction of the hadronically decaying top quark and the Higgs boson which involve the b-tag output of subjets are applicable.

7.4 Correction of the B-tag Output of Filterjets

7.4.1 Introduction

The set of the unmatched jets, whose properties were presented in section 7.2.3, are the candidates for the addition to the set of jets which are used in the scale factor calculation. As the scale factor is based on the b-tag output of the jet, it is important to show that all jets have a b-tag output that is uncorrelated to any other jet. For the anti-kt jets, this is true in such way as the calculation of the b-tag output only takes into account all tracks
within a distance of 0.5, which is the radius of these jets. Therefore, no sharing of tracks between these jets is expected.

However, as the subjets have a smaller radius than the anti-kt jets and overlap with them, there is a possibility of track sharing and therefore of a correlation of the b-tag (and other variables). To reduce this correlation, only these subjets are included which have not already been selected in the jet matching process. As the comparison in section 7.3 has shown, there is a clear correlation between the matched jets both in terms of the b-tag output and kinematic variables.

To be able to test the correlation between unmatched subjets and anti-kt jets, the unmatched jets are compared to their closest anti-kt jet. The goal is to show that there is a negligible correlation between the subjet and the anti-kt in the case of non-overlapping jets. The results of these comparisons are shown in section 7.4.2.

As it is impossible to include a jet only partially in the calculation of the b-tag scale factors, the jets that are then chosen are presumed to be totally uncorrelated to the already implemented anti-kt jets. These jets are then included in the calculation of the scale factors, the resulting scale factors are presented in section 7.4.3.

Finally, the new b-tag scale factors are calculated for all Monte Carlo samples. The resulting event yields for Monte Carlo are presented and the changes in the agreement between data and Monte Carlo are evaluated.

### 7.4.2 Comparison Studies between Unmatched Jets and Anti-kt Jets

After studying the the properties of the unmatched subjets, in the following they are compared to their closest anti-kt jet. To further evaluate the effect of overlapping jets, the subjets are put into two categories. In the first category, all subjets are presented which have a $\Delta R \leq 0.5$ to the nearest anti-kt jet, which is equivalent to an overlap. In the second category, the remaining subjets with $\Delta R > 0.5$ are shown. For these, only a weak correlation is expected.

The comparison of the $p_T$ shows that there are two areas in which the entries are positioned. In both areas, a certain correlation between the $p_T$ of the two jets can be seen. While there is an area in which either of the two jets has a higher $p_T$ than the other jet, there are only very few cases in which the $p_T$ of the two jets is identical.

In the category of jets with a $\Delta R > 0.5$, the entries are much more focused on the left side of the distribution, which corresponds to a very low $p_T$ of the unmatched subjets and an
anti-kt $p_T$ which is largely independent to this. The majority of the entries is around an anti-kt $p_T$ of $\approx 150$ GeV/$c$, which can be explained by the fact that these jets are close to collinear to a fat jet with a $p_T$ of at least 180 GeV/$c$. Taking this into account, there is no obvious correlation between the unmatched subjet and the anti-kt jet.

The second variable which is compared between the two jets is the b-tag output. In Fig. 7.11, the comparison is shown for the two above mentioned categories. In the case of the overlapping jets, no apparent correlation between the b-tag output of both jets is visible. The majority of entries is in the lower left corner and represents pairs of jets where both jets have a low b-tag output, which means that they are probably light flavour or gluon flavour jets. The fact that these jets have not been matched means that there are other subjets closer to this jet. The second accumulation of entries is in the upper left corner. Here, entries are found for which the anti-kt jet has a high b-tag output, but the subjet has not. In these cases, the subjet likely represents a gluon radiation from a bottom quark jet.

In the second category, the distribution is similar to that in category one. Again, no apparent correlation is visible. However, in this case there is a larger number of entries in the upper left corner than in the lower left. As these jets are non-overlapping, it is expected that there is a larger fraction of radiated jets than in the case of overlapping unmatched jets.
In conclusion, it can be said that there is still a certain degree of correlation between unmatched jets which have a $\Delta R < 0.5$ to any anti-kt jet. However, as there is no way of partially including a jet in the event weight calculation, the decision has been made not to include these jets in the calculation. For the remaining jets, the distributions have shown that the combinations of unmatched subjets and anti-kt jets are sufficiently uncorrelated to allow for an inclusion in the calculation of the b-tag scale factors.

### 7.4.3 Comparison of the Different B-Tag Event Weights

In this section, the effects of additional jets in the calculation of the event weight are examined. As already stated, the event weight is a product of weights that are calculated for each included jet based on flavour, $p_T$ and $\eta$ of the jet. While the event weights are centred around the value 1, they are not normalized, which can lead to different event yields based on the b-tag weight.

Due to the isolation selection of the subjets and the subsequent jet matching, the majority of events has no or only one unmatched subjet. This means that the expected effect on the overall distributions of the event weights will be limited. However, this behaviour is expected, since the standard clustering with anti-kt jets usually contains all particle flow objects in the event, making it difficult to find additional objects to cluster into jets.

To demonstrate the difference between the different calculated event weights, the distributions of the b-tag output event weight are presented in Fig. 7.12. In both samples, the weights are centered around a value of 0.9 and have tails going in both directions. In the center region, the standard weight has a higher fraction of events, while the recalculated weight has a higher fraction in areas of smaller and larger event weights. This effect comes from the additional jets, which contribute a multiplicative factor on the event weight. The factor will be likely smaller or larger than 1 which leads to the broadening effect which is visible in the distribution.

![Figure 7.12](https://via.placeholder.com/150)

(a)  
(b)

Figure 7.12: Comparison of the b-tag event weight for the signal sample (a) and the process of top-quark pair-production with semileptonic decay (b). The dots represent the standard event weight, while the bars represent the b-tag event weight based on the combination of ak5 jets and the set of unmatched subjets. Error bars represent the statistical uncertainty of the standard weight distribution.

### 7.4.4 Event Yields Based on Different B-Tag Scale Factors

In this section, the effect of the different b-tag weights is propagated to the comparison between the measured and simulated data. The standard b-tag event weight is replaced
7. Correction of B-Tag Distribution of Filtered Subjets

Table 7.1: Event Yields for the distribution of the second highest b-tag output in Topjets after selecting events that have ≥ 4 Jets of which ≥ 2 are b-tagged.

<table>
<thead>
<tr>
<th></th>
<th>≥ 4 Jets, ≥ 2 B-Tags</th>
<th>Standard Weight</th>
<th>New Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>26618</td>
<td>26618</td>
<td></td>
</tr>
<tr>
<td>$t\bar{t}$ semileptonic</td>
<td>22362</td>
<td>22400</td>
<td></td>
</tr>
<tr>
<td>$t\bar{t}$ dileptonic</td>
<td>1247</td>
<td>1251</td>
<td></td>
</tr>
<tr>
<td>Single Top</td>
<td>1190</td>
<td>1191</td>
<td></td>
</tr>
<tr>
<td>$W \rightarrow l+\nu +$ Jets</td>
<td>414</td>
<td>414</td>
<td></td>
</tr>
<tr>
<td>$Z \rightarrow ll +$ Jets</td>
<td>68</td>
<td>68</td>
<td></td>
</tr>
<tr>
<td>Diboson</td>
<td>21</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>Total Bkg</td>
<td>25388</td>
<td>25433</td>
<td></td>
</tr>
<tr>
<td>$t\bar{t}H$ inclusive</td>
<td>87</td>
<td>87</td>
<td></td>
</tr>
</tbody>
</table>

with the new weight and applied to all samples. To interpret the changes, the distribution of second highest b-tag output in selected Topjets is presented in events with ≥ 4 jets of which ≥ 2 are b-tagged. As the difference between the distributions is less than 1%, the event yields for these distributions are shown in table 7.1. It can be seen that the total effect that results from the reweighting is very small, however this is somewhat expected as the combination of jet matching and the removing of overlapping jets removes the majority of subjets in each event. Further, in the electroweak processes, the lack of fat jets leads to an unchanged behaviour, as there are no additional jets that contribute in the event weight.

Figure 7.13: The pseudorapidity $\eta$ of matched subjets (left) and unmatched subjets (right). For further details compare Fig. 7.2.
8. First Look at the LHC Run-II Data

8.1 Introduction

After a shutdown which lasted from February 2013 to April 2015, a time in which both the LHC accelerator and the CMS experiment underwent a series of repairs and upgrades, operation started again in June 2015 with an increased center of mass energy of 13 TeV compared to the 8 TeV used in Run I in 2012. This increase in energy has several effects when looking at production cross sections and event topologies.

First, the increased energy results in a general increase of the production cross section. The relative increase is particularly large for the production of particles with high masses, such as the top quark and the Higgs-boson. This is important for the analysis of the associated Higgs boson and top-quark. At an energy of 8 TeV in Run I, the process had a production cross section of 0.13 pb, which increases by a factor of four to a value of 0.5 pb for an energy of 13 TeV.

Another change that is introduced with the second LHC Run is an increased instantaneous luminosity. Among other factors, the bunch spacing of the proton bunches will be decreased to 25 ns, which effectively increases the number of bunch crossings by a factor of four. Further changes in the composition of the proton bunch diminish the increase of the instantaneous luminosity while greatly reducing the in-time pileup. However, the data that is presented in this chapter was recorded using a bunch spacing of 50 ns.

The combined effect of these two changes is an increase in the absolute number of expected signal events, as it scales both linearly with the integrated luminosity and the cross section.

A further advantage when comparing the beginnings of Run II to those of Run I is an increased knowledge of the detector and the sophisticated calibration techniques of all of its subsytems which in a large part is built on studies that were executed both on measured and on simulated data.

In this chapter, a first look at the kinematic quantities relevant to the ttH analysis of the first 13 TeV data with the associated Monte Carlo samples is presented. The goal of these studies is to verify the Monte Carlo simulation for the higher center of mass energy, study the data for signs of mis-reconstructions and also study the effects of the early residual corrections which are calculated using Run II data.
Table 8.1: Data sets analysed in this thesis. The samples represent all data that was recorded using a single-lepton trigger in the 2015 LHC Run II with a bunch spacing of 50 ns and a center of mass energy of 13 TeV. Listed are the data set and trigger name and the integrated luminosity $L$.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Trigger</th>
<th>$L$ / pb$^{-1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Electron</td>
<td>HLT_Ele27_eta2p1_WPTight</td>
<td>42</td>
</tr>
<tr>
<td>Single Muon Run</td>
<td>HLT_IsoMu24_eta2p1</td>
<td>42</td>
</tr>
</tbody>
</table>

8.2 Samples

8.2.1 Data

The analysed data were collected in July and August 2015, at a center of mass energy of 13 TeV, at a bunch spacing of 50 ns with the magnetic field at 3.8 T. It corresponds to an integrated luminosity of 42 pb$^{-1}$. The datasets were selected using single-muon and single-electron triggers. The samples and their properties are listed in Table 8.1.

The electron trigger selects electrons with a $p_T \geq 27$ GeV/c and $|\eta| < 2.1$. Further, a set of identification cuts at the tight working point are applied, that correspond to a signal efficiency of 85%.

The muon trigger selects isolated muons with a $p_T \geq 24$ GeV/c and $|\eta| < 2.1$.

8.2.2 Simulation

For the first measured data in Run-II, a selection of Monte Carlo samples similar to Run-I was chosen. All presented samples are simulated using a bunch spacing of 50 ns. The Monte Carlo samples are normalized to the expected number of expected events, the latter being calculated as the product of integrated luminosity and production cross section at LO or N(N)LO depending on the process.

A dedicated signal sample, the associated production of a Higgs boson and a top-quark pair, is not considered here due to a expected lack of events. For a cross section of 590 fb, an integrated luminosity of 42 pb$^{-1}$ and a reconstruction efficiency of the order of $10^{-2}$, less than one event in this channel are expected to be selected after the reconstruction.

The combination of the top-pair production samples with semileptonic and dileptonic decay channels that were included in the Run I analysis is replaced by a sample which contains all decay channels.

The production of a single top quark is represented by the t-channel and the associated production of a W boson and a top quark. For the t-channel, a sample with the leptonic top decay in the four flavour scheme is included. In case of the associated top W-boson production, a sample in the five flavour scheme is included that contains all decay channels of the top quark and the W-boson. An s-channel production sample is not included since such a sample was not produced for a bunch spacing of 50 ns.

For the production of both W and Z bosons with additional jets, samples with leptonic decay channels and inclusive jet multiplicity are included.

Diboson production is taken into account with a di-W boson production sample that contains all W-boson decay channels.
8.3 Event Selection

The object selection criteria which are applied on the LHC Run-II events follow those employed for LHC Run-I, which are given in detail in section 5.4. In the following, only the differences to these selection criteria are presented.

8.3.1 Lepton Selection

The lepton selection criteria are nearly identical to those used in Run I which are presented in section 5.4.1.

For the first part of the chapter, events are selected which contain at least one lepton that fulfills the tight lepton ID. For the reconstruction of the di-lepton mass resonance, a di-lepton selection is applied.

8.3.2 Jet Selection

The clustering of jets in Run II events has only a few differences compared to the reconstruction and subsequent selection that was presented in section 5.4.2. The main difference is a decrease in the size parameter from 0.5 to 0.4. The reason for this change is that due to the higher average energy in the event with a center of mass energy of 13 TeV, particles that are produced in the hadronization have a greater Lorentz boost than at a center of
mass energy of 8 TeV. Therefore, the cone which contains the particles is smaller, which is reflected by the smaller distance parameter in the reconstruction. A beneficial side-effect of this reduction is that the fraction of energy coming from pileup and the underlying event is reduced. Because of this, both ATLAS and CMS decided to use this jet radius parameter, which was already used in the ATLAS collaboration in LHC Run-I.

The kinematic thresholds on jets are $p_T \geq 30 \text{ GeV/c}$ and $|\eta| < 2.5$. The jet energy corrections provided by the CMS collaboration for the anti-kt jets in Run II are applied in the same way as for Run I 3.6.3. For both data and Monte Carlo simulation, level 1, 2 and 3 corrections are applied. For data, an additional residual correction is applied.

8.3.3 B-Tag Selection

Similar to Run I, the combined secondary vertex algorithm is used. However, due to a retrained discriminator specific to the Run II conditions, e.g. smaller jet radius parameter, the medium working point has changed to a value of 0.89.

The implemented scale factors that were used to reweight the distribution of the b-tag output in Monte Carlo events in LHC Run I are based on tagging efficiency differences in data and Monte Carlo simulation. As the scale factors are based not only on jet flavour, but also on jet $p_T$ and $\eta$, the amount of data that was recorded for a bunch spacing of 50 ns is not sufficient yet to derive scale factors at the time of the thesis. Thus, no selection based on the number of b-tagged jets is applied yet.

8.3.4 Pileup Reweighting

To improve the agreement between data and Monte Carlo simulation data, the distribution of the number of primary vertices in the Monte Carlo samples is scaled to match the shape of the data distribution. The scale factors are derived from a comparison of the distributions of the first measured Run-II data with an integrated luminosity of 5.59 pb$^{-1}$ and the sum of the background Monte Carlo events which is listed in 8.2. No QCD-multijet samples were included, to mitigate the effects of a wrong QCD normalization, e.g. luminosity uncertainty, jet energy corrections or cross section. Therefore, the normalized data and Monte Carlo distributions are compared, as the distribution of the number of primary vertices is mostly independent of the process. The scaling factors are calculated bin-by-bin by calculating the ratio of the relative fraction of events for data and Monte Carlo.

A comparison of the distribution of the number of primary vertices before and after the reweighting is given in Fig. 8.1. Before the reweighting, the distribution of data events has a lower peak position and a smaller width than in the Monte Carlo simulation. After the reweighting, the data distribution is well described by the Monte Carlo. The difference in number of events is less than 2% for most bins.

8.3.5 QCD-Multijet Normalization Estimation

The reweighted distribution of the number of primary vertices shows a flat data to Monte Carlo ratio, however there is a offset which can be explained by a lacking QCD-multijet sample. To describe the QCD background contribution, a muon-enriched QCD sample has been used. The cross section of the process was estimated using a fit on the distribution of the number of jets in data. The following function was minimized to find the optimal value:

$$X = \sum_i \left( N_{\text{data},i} - N_{\text{MC},i} - k \cdot N_{\text{QCD},i} \right). \quad (8.1)$$
8.4. Validation of LHC Run-II Data

Figure 8.1: Distribution of the number of reconstructed primary vertices before (a) and after (b) the reweighting. Events are selected to contain at least one muon with $p_T \geq 30$ GeV/c and $|\eta| < 2.5$.

Table 8.3: Event Yields for the electron and muon channel. In both channels, events are selected which have at least one electron or muon with $p_T \geq 30$ GeV and $|\eta| < 2.5$.

<table>
<thead>
<tr>
<th>$\geq 1$ tight lepton</th>
<th>Muon</th>
<th>Electron</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>275842</td>
<td>246908</td>
</tr>
<tr>
<td>WW</td>
<td>401</td>
<td>327</td>
</tr>
<tr>
<td>$Z \rightarrow ll +$ Jets</td>
<td>33013</td>
<td>26213</td>
</tr>
<tr>
<td>$W \rightarrow l+\nu +$ Jets</td>
<td>238032</td>
<td>174564</td>
</tr>
<tr>
<td>Single Top</td>
<td>714</td>
<td>604</td>
</tr>
<tr>
<td>$t\bar{t}$</td>
<td>3532</td>
<td>3091</td>
</tr>
<tr>
<td>QCD ($\mu$-enriched)</td>
<td>6418</td>
<td>0</td>
</tr>
<tr>
<td>Total MC Sim</td>
<td>282110</td>
<td>204799</td>
</tr>
</tbody>
</table>

Here, $i$ is the bin of the distribution, $N_{\text{data},i}$ is the number of events in bin $i$ and $N_{\text{QCD},i}$ and $N_{\text{MC},i}$ are the number of weighted events per bin for the QCD-multijet process and all remaining Monte Carlo processes, respectively. The fit resulted in a production cross section $\sigma$ of 109.3 nb. This value is used in the following.

Since a similar sample was not provided for an electron enriched QCD-multijet process, this cross section estimation has not been done in the electron channel.

8.3.6 Event Yields

Table 8.3 shows the number of events after a selection of at least one tight lepton which has $p_T \geq 30$ GeV and $|\eta| < 2.5$. The Monte Carlo event yields are calculated using the production cross section, the number of generated events and the integrated luminosity of the data sets.

8.4 Validation of LHC Run-II Data

In the following sections, event and object based variables relevant in the ttH analysis are investigated. The goal of these studies is to verify the general agreement between measured
and simulated data. In the case of insufficient agreement, the reasons for this behaviour are investigated.

In section 8.4.1, event variables such as the number of jets and leptons and different kinematic variables, such as the total transverse energy are presented.

Section 8.4.2 presents further distributions regarding the properties of jets, such as $p_T$ and the distribution of the b-tag output. The agreement of these variables can hint at the quality of the applied jet energy corrections.

In section 8.4.3, distributions of the different kinematic properties of leptons, including the di-lepton mass, are presented. Further, differences in the normalization between data and Monte Carlo based on the lepton trigger are shown and discussed.

In section 8.4.4, the kinematic properties of the reconstructed hadronically decayed top-quark are presented.

8.4.1 Event Variables

The analysed data sets contain events that were collected with one of the triggers which are sensitive to leptons. Since at trigger level only a loose selection with at least one muon with $p_T \geq 30\text{ GeV}$ and $|\eta| < 2.5$ is applied, a large fraction of QCD-multijet events are still present in the data. This is simulated in events that contain muons, and a respective sample is included. However, a similar sample is not yet provided for electrons by the central CMS production.

The production cross section of QCD-multijet events that contain at least one muon which fulfils the tight lepton selection is much smaller than the cross section of the same process with an electron. These leptons, which are misreconstructed non-prompt leptons, are produced as decay objects from the multijet process. However, it is easier to misreconstruct an electron than a muon, which is reflected by the larger difference in normalization when comparing electron and muon data. To ensure good agreement between data and Monte Carlo simulation in the electron channel it would therefore require a QCD-multijet sample, which is, however, not yet provided for the bunch spacing of 50 ns in LHC Run-II. Therefore, all variables which are presented for the aforementioned loose selection are only shown in the muon channel.

In Fig. 8.2a, the number of muons that fulfil the tight muon ID is presented. These are referred to as tight muons in the following. Due to the good reconstruction quality of muons, the agreement between data and Monte Carlo simulation is within 2% and also shows good agreement in the shape of the distribution.

In the single muon channel, the main contribution comes from W-boson production, as it has the largest production cross section besides the QCD-multijet process. The latter also has a significant contribution in this bin. The process of Z-boson production is present both in the single-muon and the di-muon bin. Especially in the case where the Z boson has low energy, it can happen that one of the two muons does not have enough transverse momentum to be selected by the trigger selection or the subsequent muon ID selection.

Based on the decay channel, the production of a single top quark or a pair of top quarks leads to one or two muons in the final state, where the former has a more significant contribution due to the higher branching fraction based on the branching ratio of the W boson which is more likely to decay hadronically.

The distribution of the number of anti-kt jets per event is presented in Fig. 8.2b. This distribution again shows very good agreement in shape, with a difference of less than 2%.
As expected, the W- and Z-Boson production has a low jet multiplicity, since in this case typically the leptonic final states are included. In the leading order of these processes, jets are mostly produced in the initial state after gluon splitting. However, due to the higher cross section of the W boson production, the contribution even in higher jet multiplicity bins is still significant. As the WW sample contains both the leptonic and hadronic decays, events with up to four jets and more are possible.

The QCD-multijet events contribute to low jet multiplicities, with most events having only one or two jets with $p_T \geq 30$ GeV. The jet multiplicity spectrum falls steeply due to the behaviour of the propagator which is non-resonant. For top-quark production, the leading order already has between two and six jets. The possibility of both initial- and final-state radiation allows for an overall high number of jets. It is noticeable that due to the simulation of the process in next-to-leading order, the agreement in the jet multiplicity is within the statistical uncertainty for up to 8 jets, of which many are modelled by the parton shower simulation.

![Graphs showing distribution of muons and anti-kt jets](image)

Figure 8.2: Distribution of the number of muons that fulfil the tight ID (a) and number of anti-kt jets with distance parameter 0.4, $p_T \geq 30$ GeV and $|\eta| < 2.5$ (b). Events are selected to contain at least one muon with $p_T \geq 30$ GeV and $|\eta| < 2.5$.

The distribution of the missing transverse energy MET is presented in Fig. 8.3a. Given the overall similar normalization, the two distributions have very different shapes. Based on the information given by the Jet/MET analysis group, the difference between data and simulation is mostly based on a mis-calibration of data coming from neutral hadronic energy in the barrel region of the hadronic calorimeter and from the forward hadronic calorimeter.

The aforementioned quantities are combined to a variable which measures the total amount of energy which can be found in an event. This variable, called $H_T$, is the scalar sum of the transverse energy of all jets, leptons and the missing transverse energy. The distribution is depicted in Fig. 8.3b. While the distribution of the number of jets and leptons shows good agreement, MET has a distinctively different shape when comparing data and Monte Carlo, which is propagated to the shape differences of $H_T$ as well.

### 8.4.2 Jet Variables

In Fig. 8.4, the distribution of the $p_T$ and $\eta$ of all jets with $p_T \geq 30$ GeV and $|\eta| < 2.5$ are presented. The transverse momentum of the jets is very well modelled, both in shape and normalization. This agreement is largely the result of the jet energy corrections derived from the 2015 data by the CMS collaboration which are applied for both data and Monte Carlo. Further, residual corrections based on the first data are applied.
100 8. First Look at the LHC Run-II Data

(a) Missing Transverse Energy

Figure 8.3: Distribution of missing transverse energy (a) and $H_T$ (b). Events are selected to contain at least one muon with $p_T \geq 30 \text{ GeV}$ and $|\eta| < 2.5$, while the included jets have $p_T \geq 30 \text{ GeV/c}$ and $|\eta| < 2.5$ as well. The last bin contains the overflow events of the distributions.

The pseudorapidity distribution shows a similar agreement. In general, the jets that come from W- or Z-boson production are concentrated on towards the center of the distribution, as they are often produced by initial state radiation from gluon splitting.

(b) $H_T$

Figure 8.4: Distribution of jet $p_T$ (a) and $\eta$ (b). The jets are the anti-kt jets with a distance parameter of 0.4, $p_T \geq 30 \text{ GeV}$ and $|\eta| < 2.5$. Events are selected to contain at least one muon with $p_T \geq 30 \text{ GeV}$ and $|\eta| < 2.5$.

The distribution of the transverse momentum for the first four $p_T$-sorted jets is presented in Fig. 8.5. Each of the distributions shows good agreement both in shape and normalization. The W- and Z-boson production processes have larger production cross-sections than processes that involve top quarks and are therefore the dominant process in the distribution of the first and second jet $p_T$. However, the number of jets in top production is larger in all orders of perturbation theory, which is why the top-quark pair-production has a much larger relative contribution in the third and fourth jet $p_T$, as this process in leading order already has at least two jets in the dileptonic and four jets in the semileptonic decay channel.

The distribution of the output of the b-tagging discriminator is shown in Fig. 8.6a. Overall, while the normalization shows good agreement, the shape differs significantly between data and Monte Carlo simulation, as there is a visible gradient in the ratio plot, which means
Figure 8.5: Jet $p_T$ for the jet with the first (a), second (b), third (c), and fourth highest $p_T$ per event (d). The jets are anti-kt jets with a distance parameter of 0.4. Events are selected to contain at least one muon with $p_T \geq 30$ GeV and $|\eta| < 2.5$. 
that Monte Carlo simulation predicts more jets with high b-tag output than data. Potential reasons are a different modelling of the variables in data and Monte Carlo that are used as input for the b-tag output. As they have to be calculated using both measured and simulated data which is further split into categories of \( p_T \) and \( \eta \), the scale factors cannot be calculated to the same accuracy as in Run I. The agreement of the b-tag output is expected to improve with the implementation of data-driven scale factors for the 25 ns bunch spacing data.

The different shape of the b-tagging output distribution leads to a different number of b-tagged jets above a given threshold. Above an output value of 0.89 Monte Carlo is larger than data, which leads to a migration effect in the distribution of the number of b-tagged jets per event. This effect can be seen in Fig. 8.6b. While the number of events with no b-tagged jets is nearly identical, the differences in event numbers increases as the number of b-tagged jets per event increases. It is therefore to be expected that a selection that includes the number of b-tags results in different event yields for data and Monte Carlo simulation.

![Figure 8.6: Distribution of the b-tagging output (a) and the number of b-tagged jets per event (b).](image)

8.4.3 Lepton Variables and Z-Boson Reconstruction

The distribution of the number of tight muons, which is presented in Fig. 8.2a, shows a small, but significant difference between data and Monte Carlo simulation. The distributions of \( p_T \) and \( \eta \) of these muons are presented in Fig. 8.7. The normalization differs as there is a difference of 2.5% in the total number of events and there is a small downwards slope in the ratio between data and Monte Carlo in the \( p_T \) distribution. However, this difference can be explained with the uncertainty of the Monte Carlo production cross sections. Further reasons for a different overall normalization can come from data uncertainties such as the luminosity which has a estimated uncertainty of 12\%[88].

The pseudorapidity spectrum shows comparable agreement in shape, however, there are fewer data events in areas of large |\( \eta \)|. There is a good overall agreement in the normalization, with a difference of \( \approx 2\% \).

Using the two tight leptons with the highest \( p_T \), the di-lepton resonance is reconstructed. For this object, both the case of two electrons and two muons is studied without requiring two opposite charge leptons. The selection of two tight leptons greatly reduces the
8.4. Validation of LHC Run-II Data

(a) Muon $p_T$

(b) Muon $\eta$

Figure 8.7: Distribution of the $p_T$ (a) and $\eta$ (b) of muons. Events are selected to contain at least one muon with $p_T \geq 30$ GeV and $|\eta| < 2.5$.

contribution of QCD-multijet background, which is not simulated in the case where the lepton is an electron. Since only very few events contain more than two tight leptons, the reconstruction uses two hardest leptons.

For the di-muon reconstruction, the distributions of $p_T$ and energy of these muons are presented in Fig. 8.8. The distribution shows good overall agreement. It is further noticeable that this loose selection already returns a very pure sample, as all other processes combined amount to less than 1% of events. It is noticeable that most di-muon objects have very small $p_T$ with a peaking behaviour around 5 GeV/c, which means that the Z boson in most cases is produced at small transverse velocities.

This observation is also visible in the di-muon energy distribution. The peak of the distribution is slightly above the mass of the Z boson, which confirms that most Z bosons are only produced with small momenta.

The distribution of the pseudorapidity of the dimuon system is shown in Fig. 8.8c. The distribution shows a distinct structure with two distinct peaks at $|\eta| \approx 3$. This overall symmetric behaviour which coincides with the small energy and momentum is due to the production process of the Z boson. In the simplest case, the Z boson is produced by quark-antiquark annihilation. The antiquark either comes from the proton as a seaquark or is created by gluon splitting. In both cases, it has a small energy fraction which leads to the overall small energy of the di-muon.

The invariant mass of the dimuon system is presented in Fig. 8.8d. While the overall normalization of the distribution is good, the shape of the resonance shows differences in the width of the peak. While for data, the distribution has a standard deviation of 6.1 GeV/$c^2$, the Monte Carlo distribution has a standard deviation of 5.5 GeV/$c^2$. This suggests that the Monte Carlo production is not yet finally tuned.

The general agreement in the di-electron case is very similar to that of the di-muon object. However, while for energy, $p_T$ and $\eta$ the shape is in good agreement, there are generally $\approx 5\%$ fewer measured than simulated events. Finally, the di-electron invariant mass shows the same behaviour in that the peak has a different width for data and Monte Carlo. In data, the distribution has a standard deviation of 6.1 GeV/$c^2$, the Monte Carlo distribution has a standard deviation of 5.6 GeV/$c^2$.

In conclusion, the overall agreement of leptons is good, however the distribution of the di-lepton mass has shown that there is still a necessity for fine-tuning the Monte Carlo simulation.
Figure 8.8: Distribution of the di-muon $p_T$ (a), energy (b), $\eta$ (c) and invariant mass (d).
8.4. Validation of LHC Run-II Data

Figure 8.9: Distribution of the di-electron $p_T$ (a), energy (b), $\eta$ (c) and mass (d).
8.4.4 Top Quark Reconstruction

To reduce the number of non-top-quark pair-production events, at least four jets are required in the event. However, no selection on the number of b-tagged jets is applied, as the different shape of the distribution of the b-tag output leads to a different number of b-tagged jets in data and Monte Carlo, c.f. Fig. 8.6.

As a final analysis, the hadronically decaying top quark is reconstructed. The reconstruction is done via a simple mass-based scheme: First, of all jets that are not tagged as b-quarks, those two are combined which have a dijet mass closest to the known W-boson mass of 80.4 GeV/c². Then, the object is combined with that of the b-tagged jets so to form the trijet mass closest to the top quark mass of 172.4 GeV/c².

The \( p_T \) and invariant mass of the so-reconstructed W boson are displayed in Fig. 8.10a. The dijet has a larger transverse momentum compared to a prompt W boson as it stems from a top quark decay. A directly produced W boson usually has a smaller transverse momentum, since it is produced from a quark and anti-quark combination, the latter having usually a small absolute momentum. The majority of the Monte Carlo events come from top quark production processes, however there are still some QCD and W boson events.

Due to the mass-based reconstruction, the dijet mass has a distinct peak at the known W-boson mass. However, there are also some events with a much lower reconstructed mass. This is a combinatorial effect which, among other reasons, may happen if an additional radiated jet or a jet from the second hadronically decaying W boson is chosen for the dijet combination.

The overall agreement of these distributions is good. While there are about 10% fewer events in data than in Monte Carlo, the shapes of the distributions are very similar. The normalization differences are small, since at this stage of the reconstruction no b-tagged jets are demanded.

The lack of events in the W-boson production is due to the selection of an isolated muon which reduces the contribution of the hadronically decaying W boson, a process that is also not included in the set of Monte Carlo samples.

![Figure 8.10: Distribution of the \( p_T \) (a) and invariant mass (b) of the dijet combination that was selected as the W boson of the reconstructed top quark.](image)

The \( p_T \) and invariant mass of the reconstructed top quark are presented in Fig. 8.11. In these distributions, the differences in normalization are larger, since at least one b-tagged jet is needed for the reconstruction. However, there are fewer events in data that fulfil
this requirement. However, when comparing the mass and $p_T$ distributions, the overall agreement of the shapes is still good. The agreement of the normalization is expected to improve with the implementation of b-tag scale factors.

Figure 8.11: Distribution of the $p_T$ (a) and invariant mass (b) of the combined dijet and b-tagged jet that was chosen as the reconstructed top quark.

8.5 Conclusion and Outlook

In the variables presented in this chapter, the data that was taken from the first proton-proton-collisions in the LHC Run-II shows a good agreement with the Monte Carlo simulation. This result comes in part from the implemented pileup reweighting and the estimation of the QCD-multijet contribution in the muon-channel, but also shows the quality of the Monte Carlo event generation. The observed differences between data and Monte Carlo simulation, for instance in the output of the b-tagging algorithm or the di-lepton system, are expected to decrease as new jet energy corrections, detector calibrations and b-tagging scaling factors are introduced.
9. Conclusion and Outlook

The process of Higgs-boson production in association with a top-quark-antiquark pair (t¯tH) is a promising Higgs-boson production channel, as it allows for a direct measurement of the coupling between the Higgs-boson and the top-quark. This process has been measured by CMS [1] and ATLAS [89, 90] in a variety of decay channels at a center-of-mass energy of 8 TeV. However, the sensitivity of the analyses was not sufficient to declare a discovery of this process. To improve on this result, alternate methods of event reconstruction were introduced. Based on the idea of reconstructing the Higgs boson via jet-substructure methods [2], the subjet-filterjet algorithm [76] was implemented for the CMS analysis. A similar method has also been applied in the reconstruction of the hadronically decaying top-quark, leading to the HEP-Toptagger algorithm [72]. These substructure algorithms were implemented in the analysis and studied in the time of the LHC shut-down from 2013 to June of 2015.

The analysed process was the decay of a Higgs boson into bottom quarks in association with the decay of a top-quark antiquark-pair into jets and leptons, as this process combines high branching fractions of the decays and final state particles that can be efficiently reconstructed. In this thesis, the impact of the aforementioned substructure algorithms was investigated. The studies were performed using data recorded from proton-proton collisions at the CMS experiment at a center-of-mass energy of 8 TeV in proton-proton collisions at the LHC.

The distributions of the kinematic variables of the fat jets of both subjet-algorithms show good agreement between data and simulation. The studies have shown further that the HEP-Toptagger algorithm is effective at reconstructing the three jets that represent the decay objects of the hadronic top-quark. Furthermore, the selection of the hadronic top-quark candidate was analysed, which uses a likelihood method to select fat jets that contain the decay objects from the hadronically decaying top quark. The likelihood proved to be effective at selecting these jets while at the same time rejecting a large fraction of the combinatorial background in the t¯t process. An alternate likelihood method was introduced and compared, which resulted in a performance similar to the already implemented method. In the case of the Higgs-boson reconstruction, the irreducible background in the form of top-quark pair-production makes the reconstruction of the Higgs-boson difficult. However, many parameters of the chosen Higgs-boson candidate jet are used as input variables in subsequent multivariate analyses.

As the studied process contains a large number of bottom quark jets, the kinematic properties and bottom-quark selection efficiency for anti-kt jets and subjets are compared. These
studies are performed for pairs of subjets and anti-kt jets with large or complete overlap, so that the objects which are used for the clustering and the calculation of the b-tag output are similar for both types of jets. There is a significant correlation of these parameters, which signifies that the identification efficiency of bottom quarks that are clustered into subjets is similar to that of anti-kt jets. With that knowledge, corrections for the b-tags of subjets were implemented using the correction factors that were derived for anti-kt jets. After selecting only the subjets which show no correlation to the already included anti-kt jets, the effect of these additional jets is insignificant.

After a shutdown period of two and a half years, the LHC Run-II started, providing proton-proton collisions at a center-of-mass energy of 13 TeV. Especially for the $t\bar{t}H$ process, this increase is of significant importance, as the production cross section increases by a factor of four. This, in addition with the increased integrated luminosity, makes it likely that an analysis of this process will eventually lead to a discovery of $t\bar{t}H$, as the analysis was mostly limited by a lack of recorded events.

In the second part of the thesis, the first data that was taken from proton-proton collisions at a center-of-mass energy of 13 TeV was studied. While the statistical significance of this data was not enough to study the aforementioned jet-substructure-algorithms, they were still used to test the updated Monte Carlo simulation. After a reweighting of the pileup distribution and an estimation of the QCD-multijet background in the muon-channel, the distributions showed a generally good agreement between data and Monte Carlo. These studies also revealed a number of differences, for instance in the distribution of the b-tag discriminator or of the missing transverse energy. However, as the sources for these differences are known and the methods to correct them will be implemented for future data, an overall good agreement between data and Monte Carlo is expected for the remainder of the LHC Run-II. This result is important, as many of the analysis techniques that are applied for the $t\bar{t}H$ process rely on an accurate description in Monte Carlo.
Bibliography


[64] V. Khachatryan et al., “Performance of photon reconstruction and identification with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ TeV,” *JINST*, vol. 10, no. 08, p. P08010, 2015.


*JHEP*, vol. 05, p. 026, 2006.

[80] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, and T. Stelzer, “MadGraph 5 : 

[81] R. Frederix and S. Frixione, “Merging meets matching in MC@NLO,” *JHEP*, vol. 12, 

2010.

[83] C. Collaboration, “Search for Higgs Boson Production in Association with a Top- 
Quark Pair and Decaying to Bottom Quarks or Tau Leptons,” 2013.


differential cross section of Z bosons produced in association with jets in pp collisions 
at $\sqrt{s} = 8$ tev.” 2013.

[87] T. Gehrmann, M. Grazzini, S. Kallweit, P. Maierhöfer, A. von Manteuffel, S. Poz- 
zorini, D. Rathlev, and L. Tancredi, “$W^+W^-$ Production at Hadron Colliders in Next 

[88] “Measurement of the inclusive and differential $tt$ production cross sections in lepton 
2015.

with top quarks and decaying into $b\bar{b}$ in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS 

[90] G. Aad *et al.*, “Search for the associated production of the Higgs boson with a 
top quark pair in multilepton final states with the ATLAS detector,” *Phys. Lett.*, 