Associated top-quark-pair and b-jet production in the dilepton channel at $\sqrt{s} = 8$ TeV as test of QCD and background to tt+Higgs production

Dissertation

with the aim of achieving a doctoral degree

at the Department of Physics

of Universität Hamburg

submitted by

Nazar Bartosik

from Ivan'ky, Ukraine

Hamburg

2015

Abstract

Two technical and two physics analyses are reported in this thesis, which were carried out using data recorded by CMS at $\sqrt{s} = 8$ TeV. The first technical contribution is related to the improvement of the performance of the CMS tracking detector, which provides measurements of trajectories of charged particles, and is relevant for nearly every physics analysis at CMS. A new method was studied for Lorentz-angle estimation and correction for the lost signal due to the short readout time between subsequent proton-bunch crossings. Such parameters are determined simultaneously with the tracker geometry, which allows to include correlations between different effects and to even absorb unknown or mismodelled ones. The obtained results show an improvement of hit-position resolution, which is more stable in time compared to standalone measurements. Possible ways of further improvement of the method are also proposed.

In the second technical contribution, a standalone software package was developed for the identification of the origin of jets initiated by bottom quarks (b jets) in Monte Carlo (MC) simulations. It allows to differentiate at generator level between b jets originating from decays of e.g. a top-quark, or a Higgs boson, or a Z boson, based on the actual particle relations stored in the simulated event. This functionality is crucial for the presented physics analyses, and it did not exist in CMS before. Its generic implementation provides equally high performance with all parton-shower generators that are used at CMS, and was chosen as a recommended way of heavy-flavour-jet definition in future CMS analyses.

As one of the primary results of this thesis, the cross sections of top-quark-pair production in association with at least one ($t\bar{t}b(\bar{b})$) or at least two ($t\bar{t}b\bar{b}$) b jets with $p_\mathrm{T} > 20$ GeV were measured differentially as functions of transverse momentum (p_T) and absolute pseudorapidity ($|\eta|$) of the first or the second additional b jet respectively. The $t\bar{t}b\bar{b}$ cross section was also measured as a function of the angular distance ($\Delta R_{b\bar{b}}$) and of the invariant mass ($m_{b\bar{b}}$) of the two additional b jets. The latter provides substantial sensitivity to the $t\bar{t}H$ process, to which $t\bar{t}b\bar{b}$ is an almost irreducible background, therefore the data is blinded in the $t\bar{t}H$-enhanced invariant-mass region. The dileptonic final state of the $t\bar{t}$ process is considered to minimise the probability of possible jet misassignments. The measured $t\bar{t}b(\bar{b})$ ($t\bar{t}b\bar{b}$) cross section is about 1.9 (1.3) times higher than predicted by MC simulations, which is in agreement with the previous inclusive $t\bar{t}b\bar{b}$ cross-section measurement by CMS. The shapes of the measured cross sections are well modelled by MC predictions, except for the $\Delta R_{b\bar{b}}$ spectrum, which shows a tendency towards smaller simulated separation between additional b jets than measured in the data.

The measured $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ cross sections were used to estimate the sensitivity to the $t\bar{t}H(b\bar{b})$ process based on MC simulations, without systematic-uncertainties treatment. Several approaches of candidate jet-pair selection were studied, which aim at the extraction of the $t\bar{t}H$ signal from the $m_{b\bar{b}}$ distribution, which is a well-understood physical quantity. The highest signal-over-background ratio reached is 1/9, at which the $t\bar{t}H$-signal significance is 0.18σ for the available 19.7 fb^{-1} of data. A rough estimate of the sensitivity with future data recorded at $\sqrt{s} = 14$ TeV showed that about 130 fb^{-1} of data should be sufficient to see a mass peak from $H \to b\bar{b}$ decays with a significance of 2σ.

Zusammenfassung

Diese Arbeit beschreibt zwei technische und zwei Physikanalysen, durchgeführt mit von CMS bei $\sqrt{s} = 8\,\text{TeV}$ aufgezeichneten Daten. Der erste technische Beitrag optimiert die Signalinterpretation des CMS-Spurdetektors, der Trajektorien geladener Teilchen misst, und für nahezu alle CMS-Physikanalysen relevant ist. Eine neue Methode zur Bestimmung des Lorentzwinkels, und einer Korrektur auf Grund unvollständiger Signalauslese bedingt durch die kurzen Zeitabstände zwischen den Protonkollisionen, wurde untersucht. Hierbei werden die Parameter zusammen mit der Spurdetektorgeometrie ermittelt, und somit Korrelationen zwischen verschiedenen Effekten berücksichtigt, und sogar unbekannte oder schlecht modellierte absorbiert. Die Ergebnisse zeigen im Vergleich zu unabhängig bestimmten Parametern eine Verbesserung der Ortsauflösung, und eine geringere Zeitabhängigkeit dieser. Weitere mögliche Verbesserungen der Methode werden auch vorgeschlagen.

Im zweiten technischen Beitrag wurde ein Softwarepaket entwickelt, welches von b-Quarks herrührende Jets (b-Jets) in Monte-Carlo-Simulationen identifziert. Es verwendet die in simulierten Ereignissen gespeicherten Teilchenrelationen, um den Ursprung der auf Generatorebene definierten b-Jets, z. B. aus Zerfällen von Top-Quarks, Higgs-Boson oder Z-Boson, zu finden. Dies ist essentiell für die vorliegenden Physikanalysen, und war bisher in CMS nicht vorhanden. Die generische Implementierung führt zu gleichmäßig hoher Effizienz bei allen in CMS verwendeten Partonschauergeneratoren, und ist neuer Standard der Identifikation von Jets aus schweren Quarks für kommende CMS-Analysen.

Als eines der Hauptresultate dieser Arbeit wurden Wirkungsquerschnitte von assoziierter Top-Quark-Paarproduktion und mindestens einem ($t\bar{t}b(\bar{b})$) oder zwei ($t\bar{t}b\bar{b}$) b-Jets mit $p_\text{T} > 20\,\text{GeV}$ differentiell als Funktion des Transversalimpulses (p_T) und dem Betrag der Rapidität ($|\eta|$) des ersten beziehungsweise zweiten zusätzlichen b-Jets gemessen. Des weiteren wurde der $t\bar{t}b\bar{b}$-Wirkungsquerschnitt als Funktion des Abstandswinkels ($\Delta R_{b\bar{b}}$) und der invarianten Masse ($m_{b\bar{b}}$) der zwei zusätzlichen b-Jets gemessen. Letztere ist sensitiv auf den $t\bar{t}H$-Prozess, für den $t\bar{t}b\bar{b}$ einen nahezu irreduziblen Untergrund darstellt, weshalb im relevanten $t\bar{t}H$-Massenbereich der Ansatz der einer Blindanalyse gewählt wird. Der dileptonische Endzustand des $t\bar{t}$-Prozesses minimiert die Wahrscheinlichkeit falscher Jet-Zuweisungen. Der gemessene Wirkungsquerschnitt von $t\bar{t}b(\bar{b})$ ($t\bar{t}b\bar{b}$) ist etwa einen Faktor 1.9 (1.3) höher als in der Monte-Carlo-Simulation, in Einklang mit der vorigen inklusiven Messung von CMS. Die Formen der gemessenen Wirkungsquerschnitte sind gut von den Vorhersagen beschrieben, lediglich die $\Delta R_{b\bar{b}}$-Verteilung tendiert zu kleineren Abständen zwischen den zusätzlichen b-Jets.

Die $t\bar{t}H$-Sensitivität wurde unter Verwendung der gemessenen $t\bar{t}b(\bar{b})$- und $t\bar{t}b\bar{b}$-Wirkungsquerschnitte bei Vernachlässigung systematischer Unsicherheiten abgeschätzt. Mehrere Methoden der Jet-Paarselektion zur Extrahierung des $t\bar{t}H$-Signals aus der $m_{b\bar{b}}$-Verteilung, einer gut verstandenen physikalischen Größe, wurden untersucht. Das höchste gefundene Signal-zu-Untergrund-Verhältnis beträgt $1/9$, die entsprechende Signifikanz für die verfügbaren $19.7\,\text{fb}^{-1}$ Daten ist 0.18σ. Eine Näherung für Datennahmen bei $\sqrt{s} = 14\,\text{TeV}$ ergibt, dass grob $180\,\text{fb}^{-1}$ ausreichen sollten, um einen Massenpeak mit 2σ-Signifikanz sehen zu können.

Contents

Chapter 1

Introduction

One of the distinctive features of the human kind is permanent curiosity about itself and the world around it. It is this curiosity that has led to great advances in the understanding of nature and to enormous technological achievements of the modern civilisation, which would look like a miracle to anyone a few decades ago.

It would be fair to say that one of the key prerequisites to this development is the simple idea that all matter that surrounds us is made of basic building blocks at a very small scale. Although seeming obvious nowadays, this idea has been just one of the philosophical hypotheses some 2500–3000 years ago. The concept of atomism is present in works of some Greek philosophers such as Leucippus, Democritus, Anaxagoras, Aristotle, Epicurus, dating back to fifth century B.C. [1, 2]. Similar ideas have been found in Indian philosophical texts of the Mimasma, Nyaya-Vaisesika and Jaina schools dating back to sixth century B.C. [3].

The main reason why the concept of atomism has been abandoned for thousands of years is that there was no way to prove it by means of pure logics. There must be an experimental proof or evidence supporting this idea to make it widely adopted by the society. Exactly such evidence has been presented to the scientific community by John Dalton in 1803, as a result of his studies on the absorption of gases by water and other liquids [4, 5]. He calculated the first table of relative weights of different substances, defining a clear path for future studies of the structure of matter.

Advancing further in the understanding of nature and collecting more experimental observations, purely qualitative explanations were not enough. Therefore, the ideas had to be put on a mathematical basis, allowing to precisely quantify observables that can be measured experimentally. This has formed the basic pattern of fundamental research on the structure of matter:

1. formulate a theory in mathematical language;

2. calculate predictions of the theory in certain circumstances;

3. test the predictions by an experiment;

4. in case of inconsistent results improve the theory and return to step 2.

In this way some theories have been abandoned, and new ones have been created that were able to properly describe all the phenomena that have been observed via countless experiments.

Of particular interest is a series of Geiger-Marsden experiments, carried out during 1908–1913, in which golden foil has been bombarded by bunches of α particles [6, 7]. From those experiments Ernest Rutherford has concluded the existence of a tiny positively charged nucleus inside the atom [8], which was a big step in understanding the nature of subatomic processes. At that time of relatively wide acceptance of the Joseph Thomson's *plum pudding* model [9], positively charged α particles were supposed to penetrate atoms with very small scattering angles. Therefore, the observation of results inconsistent with this model during the experiment, in particular the deflection of a tiny fraction of the α particles by angles larger than $90°$, produced hints for a new model of the atom that would describe such a phenomenon. This example shows the vital importance of experimental results that can not be described by an existing model, no matter how successful it was before. Only results that contradict predictions of the model can lead to better models that are closer to the truth.

During the last century, many different aspects of nature have been studied and successfully described by corresponding theories, which nowadays are assembled together into the Standard Model (SM) of particle physics, which describes different types of matter, its structure and interactions. This theory is currently in exactly this state where some experimental observations that can not be described by it are highly demanded. This is because the Standard Model has been in great agreement with data from particle-physics experiments during the last 40 years.

While this is a truly remarkable achievement for such a sophisticated theory, it is known that the Standard Model does not yet describe all phenomena known to modern science. In particular, it does not incorporate gravitational interactions, has no description of the dark matter constituting about 27% of the mass-energy content of the observable universe, and no explanation of its accelerating expansion. Some of these aspects going beyond the Standard Model are addressed by other theories, either with the help of additional exotic particles or additional space dimensions.

In order to find particular phenomena that disagree with SM predictions, or to search for possible signatures of theories beyond the Standard Model, there is hardly a better place than the Large Hadron Collider (LHC) [10] — the most complicated and powerful particle accelerator, built at the European Organisation for Nuclear Research, CERN. Having started its operation on November 20, 2009, it has been accelerating and colliding proton bunches to study what happens to elementary particles at very small scale, in a somewhat similar way as Ernest Rutherford was studying atoms by colliding α particles with atoms of gold.

The work presented in this thesis has been performed with data from proton-proton collisions provided by the LHC at a centre-of-mass energy of $\sqrt{s} = 8$ TeV, recorded with the Compact Muon Solenoid (CMS) experiment [11]. Two big constituents of the CMS experiment are the detector, which is responsible for measuring particles produced from the proton-proton collisions, and the analysis software needed to produce a scientific result from a combination of objects measured by the detector. Luckily, it was possible to contribute to both these technical parts, which are presented in Chapters 5 and 8 of this thesis. In addition, physics analyses carried out in the scope of this thesis are performed as a test of the Standard Model with the available data and as a projection for future running of the LHC. An overview of the theoretical basis of the Standard Model relevant for the presented analyses is given in Chapter 2, while the actual physics analyses are described in Chapters 9 and 10. Each technical contribution and physics analysis is briefly described below.

1.1   Technical contribution

The heart of the CMS experiment is a multi-purpose detector with an inner structure typical for this kind of experiments. A distinctive feature of the detector is its compact size and a superconducting solenoid creating a very strong magnetic field, inside of which most of the subdetectors are located.

Typically, the part closest to the collision point is a tracking detector (also called tracker). The combination of a large number of detecting modules of the tracker with a strong magnetic field around it allows to measure trajectories of charged particles (also called tracks), which penetrate multiple silicon modules of the detector. These measured trajectories are then used to estimate the momenta of the charged particles — an important quantity used in nearly any high-energy-physics analysis. Trajectories are reconstructed from sets of local penetration-point coordinates on the surface of the modules. Since coordinates are measured relative to the module surface, the position of the modules themselves has to be known with a precision higher than their own spatial resolution, in order to achieve the precision needed for physics analyses. The possibly incorrect assumption of the tracker geometry can have an impact on nearly every CMS measurement that uses tracker information. In particular, this is important for the identification of jets produced from the hadronisation of bottom or charm quarks. Precise track trajectories are needed to reconstruct secondary vertices from such hadron decays, as used in physics analyses described later.

The design geometry of the tracker can not be used directly for analysis, since the mounting precision of the tracker and its modules is not high enough. Furthermore, it can change with time, which requires continuous monitoring of the tracker geometry. Therefore, it is used only as a starting point for the estimation of the real tracker geometry with much higher precision, using the large statistics of recorded tracks and sophisticated stochastic methods, which is generally called track-based alignment.

Several effects taking place in the tracking detector reduce the performance of the track-based alignment procedure. The first one is the Lorentz-angle effect, which leads to a shift of the measured hit positions in the strong magnetic field in different directions depending on the module orientation. This effect is present in all modules that have signal charge carriers moving not in the plane of the magnetic field direction, and has been incorporated in the track-based alignment procedure as a Lorentz-angle calibration. The second effect is a result of the very short period between subsequent proton-bunch crossings, which is not enough to collect the full charge induced in the module. This leads to a shift in the measured hit position, which depends on the track angle relative to the module surface. It was included in the alignment procedure as an additional backplane correction. Details about these effects, exact implementation of the described calibrations and results of the studies performed in the scope of this thesis are described in Chapter 5.

On the analysis software side, a contribution was made to the tool-set for the identification of b jets and of their origin in Monte Carlo simulations of high-energy-physics processes. This was an important ingredient for the physics analyses presented in this thesis, and existing tools could not provide sufficient information and performance. Mainly due to the lack of a tool with similar performance within CMS, and due to its importance for a variety of other analyses, it has become a part of the official CMS software framework. Furthermore, it was chosen as a baseline for

process definitions in future analyses, which will use data recorded by CMS starting from 2015. Details about the technical implementation and performance of the developed software are given in Chapter 8.

1.2 Physics contribution

One of the major milestones for the LHC has been the observation of the Higgs boson, which is the last elementary particle of the Standard Model that had not yet been observed at the time of its construction. This milestone was successfully achieved when two independent experiments, ATLAS [12] and CMS [11], have announced on July 4, 2012 the observation of a particle with a mass around 125 GeV [13, 14]. Later the Physics Nobel Prize 2013 was awarded to François Englert [15] and Peter Higgs [16] "for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery of the predicted fundamental particle, by the ATLAS and CMS experiments at CERNs Large Hadron Collider" [17].

Now, since the Higgs boson has been discovered, its properties have to be measured and probed for consistency with the Standard Model. One of its important properties is the coupling to fermions, in particular to the top quark, which is the heaviest fermion of the Standard Model and has an important role in the electroweak-symmetry-breaking mechanism. The main way for a direct measurement of this coupling is through the associated top-quark-pair and Higgs-boson ($t\bar{t}H$) production.

In order to observe a sufficient number of such events in the limited amount of collisions, which have been recorded by the CMS detector, the search is performed for the most probable final state of the Higgs boson: $H \rightarrow b\bar{b}$. The $t\bar{t}H$ signal in this final state is overwhelmed by a several orders of magnitude larger background, mainly represented by top-quark-pair production in association with two additional b jets ($t\bar{t}b\bar{b}$). Therefore, the amount of this background and the shape of its distributions of interest have to be well known before extracting a signal from the recorded experimental data. The estimation of this dominant background process has led to a separate measurement of differential $t\bar{t}b\bar{b}$ cross section as a function of properties of additional b-jets, which is the primary topic of this thesis. The corresponding analysis of top-quark-pair production in association with one or two additional b jets is described with all details in Chapter 9.

Having properly estimated the $t\bar{t}b\bar{b}$ background, sensitivity to associated top-quark-pair and Higgs-boson production was estimated using the distribution of the invariant mass of additional b jets. Also a projection for future runs of the LHC was done to estimate the amount of data required to observe a mass peak from $H \rightarrow b\bar{b}$ decays. The developed analysis techniques will be of high relevance in future measurements with a higher centre-of-mass energy and with larger amount of the recorded data, and are being prepared for the new data from the LHC in 2015. Details about the developed analysis strategy, as well as the estimated sensitivity to the $t\bar{t}H(b\bar{b})$ process, are documented in Chapter 10.

Chapter 2

Theoretical basis

This chapter provides a description of the essential theoretical background needed for the understanding of the processes that are the subject of research described in this thesis. Since all the described measurements are carried out within the Standard Model of particle physics, the main concepts of this theory are introduced in Section 2.1. The nature of the proton-proton collisions at the Large Hadron Collider brings additional complications to the measurements of elementary particle interactions. In particular, the parameterization of the proton structure and its translation to the measured processes are described in Section 2.2. Finally theoretical aspects of the specific processes that are presented in this thesis, in particular top-quark-pair production and Higgs boson production, are addressed in Section 2.3.

2.1 The Standard Model of particle physics

The Standard Model of particle physics (SM) summarizes the current knowledge of the scientific community about the structure of matter and the known interactions, which has been obtained through experimental research and theoretical developments in different fields of physics during the last century. According to this model, all observable matter in the universe is a manifestation of 12 types of elementary matter particles — fermions. They have received their name from Paul Dirac as an indication that fermions are characterised by the Fermi-Dirac statistics [18, 19] and, therefore, obey the Pauli exclusion principle [20]. Interactions of the fermions are realized through the exchange of force mediating particles — bosons, which have obtained their name from Paul Dirac due to obeying the Bose-Einstein statistics [21, 22].

Fundamental interactions are represented in the Standard Model by the electromagnetic, the weak and the strong forces, each of which is mediated by corresponding bosons. The fourth known force, which is most obvious from the day-to-day experience — gravity, is not included in the model. Nevertheless, its strength is many orders of magnitude smaller than the other three, therefore it is absolutely negligible in the processes described by the Standard Model, and does not affect calculations at all.

In this chapter only those aspects of the model are described that are most important for the analyses presented in the thesis. A more complete overview of the theoretical basics of the Standard Model can be found e.g. in [23, 24].

2.1.1   Elementary particles

According to the Standard Model there are two conceptually different groups of elementary particles: fermions and bosons. Fermions constitute all the observable matter in the universe and, having fractional spin, comply with the Pauli exclusion principle [20], which, according to the spin-statistics theorem, means that the wave-function of a system of identical fermions is antisymmetric with respect to the exchange of positions of any two particles [25, 26]. Bosons mediate fundamental interactions between fermions and between the bosons themselves. Having integer spin, bosons follow the Bose-Einstein statistics [21, 22], which means that the wave-function of a system of identical bosons is symmetric with respect to swapping the positions of any two particles [25, 26]. The general classification of the elementary particles together with their basic properties is visualised in Figure 2.1.

Figure 2.1: Elementary particles of the Standard Model, listed with the indication of the spin quantum number (S), electric charge (Q) and mass in the $\overline{\text{MS}}$ scheme (m) for each of the particles. The type of experienced interactions is specified for each particle by coloured circles. Numeric values are taken from the PDG 2014 Particle Listing [27]).

The properties of the twelve fundamental fermions are categorized by the types of interaction that they experience. As can be seen from the overview diagram Fermions come in two groups: quarks and leptons. All of them undergo weak interactions. With the exception of the neutrinos, which are electrically neutral, the

other nine fermions are electrically charged and participate in the electromagnetic interaction of Quantum Electrodynamics (QED) — the theory describing electromagnetic interactions of electrically charged particles. Only the quarks carry colour charge, which is an equivalent of the electric charge in Quantum Chromodynamics (QCD) — the theory describing strong interactions between colour-charged particles. Consequently, only the quarks feel the strong force. Because of the nature of the QCD interaction, quarks are never observed as free particles, but are always confined to bound states called hadrons, such as the proton and neutron. Because the quarks undergo the strong interaction, their properties are very different from those of the electron, muon, tau-lepton and the neutrinos, which are collectively referred to as the leptons [24].

From the diagram in Figure 2.1 it can also be seen that both quarks and leptons come in three generations (from left to right). In each next generation, the masses of the particles increase, while all the other basic properties of the fermions do not change. Furthermore, only matter particles are shown in the diagram, while each listed fermion has a corresponding antimatter counterpart with all the same properties except of charge-like quantities (e.g. electric charge, colour, isospin, helicity, ...), which have the same magnitude but the opposite sign.

The particles on the right side of the diagram, except the Higgs boson, are the gauge bosons of the Standard Model — force mediating particles. The photon is the gauge boson of the electromagnetic interaction. Its zero mass implies the infinite range of the electromagnetic force. The gluon is the mediator of the strong QCD interaction and is also massless. Three gauge bosons: Z^0, W^+ and W^- are responsible for the weak interactions. In particular, neutral weak interactions are mediated by the Z^0 boson, and charged weak interactions are mediated by the W^+ and W^- bosons. More details about the actual interactions being mediated by the mentioned bosons are given in Section 2.1.2.

The recently discovered H (Higgs) boson does not explicitly mediate any force, unlike the rest of the bosons in the Standard Model. It is rather the smallest possible excitation of the Higgs field, via which Z and W bosons acquire their masses. The properties of the Higgs boson largely depend on its mass, which has been measured up to now with a precision of better than $0.5\,\%$. The agreement of its measured properties with the Standard Model predictions has yet to be checked with higher precision. More details about the Higgs mechanism of spontaneous electroweak symmetry breaking [28, 16, 15] are given in Section 2.1.3.

2.1.2   Fundamental interactions

The mathematical principles of the Standard Model are expressed in the form of relativistic Quantum Field Theories (QFT). Within this formalism, all fermions are represented by quantised spinor fields, and all gauge bosons by quantised vector fields. In QFT each fermion and interaction is described by a term in the Lagrangian density ($\mathcal{L}$), which is a scalar function of all particle fields. Interactions in the SM appear as a solution to make $\mathcal{L}$ invariant under specific local gauge transformations — unitary local transformations (U) for the fermion spinors (Ψ):

$$\Psi \rightarrow \Psi' = U\Psi. \tag{2.1}$$

According to the Noether theorem [29], each invariance leads to a conserved physical quantity, which can be identified as a specific charge g of the interacting particle. In order to fulfil the invariance criteria, additional vector fields are introduced that couple to Ψ with coupling constants α, which are related to the corresponding charge g:

$$\alpha \propto g^2. \tag{2.2}$$

Different interactions correspond to different types of unitary transformations. In general the dynamics of the Standard Model arises from the postulation of local gauge invariance under transformations of the direct product of Lie groups:

$$SU(3)_C \otimes SU(2)_L \otimes U(1)_Y. \tag{2.3}$$

Each of the three fundamental forces (strong, electromagnetic and weak) is the result of a separate symmetry group. $SU(3)_C$ represents the strong interaction, while $SU(2)_L \otimes U(1)_Y$ represents the unified electroweak interaction, which at low energy scale is reduced to the electromagnetic force represented by $U(1)_{EM}$.

Electromagnetic interaction

The oldest, the simplest and the most precisely verified gauge field theory (see e.g. [30, 31]) is the theory of Quantum Electrodynamics (QED) [32]. It describes the interaction between electrically charged particles via the exchange of the corresponding gauge boson, the virtual photon. The photon is a massless particle with no electric charge, therefore photons do not directly interact with each other.

In QED the electromagnetic field is described as a field that couples to a charged Dirac particle, fulfilling the condition of gauge invariance of the fermionic field Lagrangian under arbitrary time and space dependent unitary transformations in accordance with the Abelian symmetry group $U(1)_{EM}$. The strength of the interaction between charged particles and photons is specified by the small dimensionless electromagnetic coupling (fine-structure) constant α. With the electron charge e the coupling constant is given as:

$$\alpha = \frac{e^2}{4\pi\hbar c} \approx \frac{1}{137} \ll 1. \tag{2.4}$$

One of the simplest examples of a QED process is electron-positron scattering (Bhabha scattering) shown in Figure 2.2. There are two ways how an electron and a positron can interact: either via annihilation into a photon, which then decays into an electron-positron pair (s-channel), or via a scattering process by the exchange of a photon (t-channel). In principle there is no limitation on the number of photons that can participate in the interaction, with additional photons adding powers of α to factors by which the corresponding contribution is multiplied. Nevertheless, due to the small value of α, the simplest processes involving a single photon are dominant compared to the higher order radiative corrections, which can be calculated in perturbation theory.

For the calculation of these higher order corrections, all possible processes involving specific powers of α have to be calculated. One of the greatest inventions in

QED is the technique of Feynman diagrams, which has been developed by Richard Feynman (presented in 1949 [33]), and was heavily popularized by his younger associate Freeman Dyson [34]. Feynman diagrams provide an elegant way of visualizing any QED process and greatly simplify bookkeeping of countless higher order perturbative terms. Diagrams depicting the leading order (LO) Bhabha scattering as well as several examples of processes involving exchange of two photons are shown in Figure 2.2. Due to the great success of the Feynman diagrams in QED calculations, they have become a standard method for the visualization of all kinds of processes in particle physics.

As can be seen from the diagram in Figure 2.2g, a photon can split into a pair of e.g. electron and positron for a short time. Since this happens in the field of an electric charge, the pair becomes polarised. Such an effect is called *vacuum polarisation* and can change the effective charge, leading to the dependence of α on the momentum transfer Q^2. In particular, the value stated in equation 2.4 is valid at zero momentum transfer and becomes slightly higher at the scale of the Z-boson mass ($Q = m_Z = 91.2\,\text{GeV}$) [36]:

$$\alpha(m_Z^2) \approx \frac{1}{128.957} \ll 1. \tag{2.5}$$

The running of $\alpha(Q^2)$ is estimated by dedicated calculations and measurements e.g. [37] and [38].

Strong interaction

The strong interaction is characterised by the theory of Quantum Chromodynamics (QCD), which is a gauge theory based on the local colour transformations from the non-Abelian Lie group $SU(3)_C$.

In the group name, the notation C refers to the colour space and 3 refers to the number of possible colour states of the quarks (colour charge). From the structure of the $SU(3)$ group it can be deduced that as an outcome of combining the 3 and $\bar{3}$ representations of $SU(3)$:

$$3 \otimes \bar{3} = 8 \oplus 1\,, \tag{2.6}$$

there is an octet of eight gauge bosons, carrying colour and anti-colour charge, which are called *gluons*.

The QCD Lagrangian density is therefore given by:

$$\mathcal{L}_{QCD} = \sum_f \bar{\Psi}_f^i (i\gamma_\mu D^\mu - m_f)_{ij} \Psi_f^j - \frac{1}{4} F_a^{\mu\nu} F_{\mu\nu}^a\,, \tag{2.7}$$

with the gluon field strength tensor:

$$F_a^{\mu\nu} = \partial^\mu A_a^\nu - \partial^\nu A_\mu^a + g f^{abc} A_b^\mu A_c^\nu\,, \tag{2.8}$$

where f^{abc} are the $SU(3)$ structure constants and $A_a (a = 1, \ldots, 8)$ are the gluon fields [39].

The last term in (2.8) represents the interaction of the gluons with each other, as they carry colour charge themselves. This is the main difference of QCD with

(a) s-channel

(b) t-channel

(c) s-channel with one additional photon

(d) t-channel with one additional photon

(e) t-channel with positron self-energy correction

(f) t-channel with vertex correction

(g) s-channel with virtual fermion loop

(h) box diagram
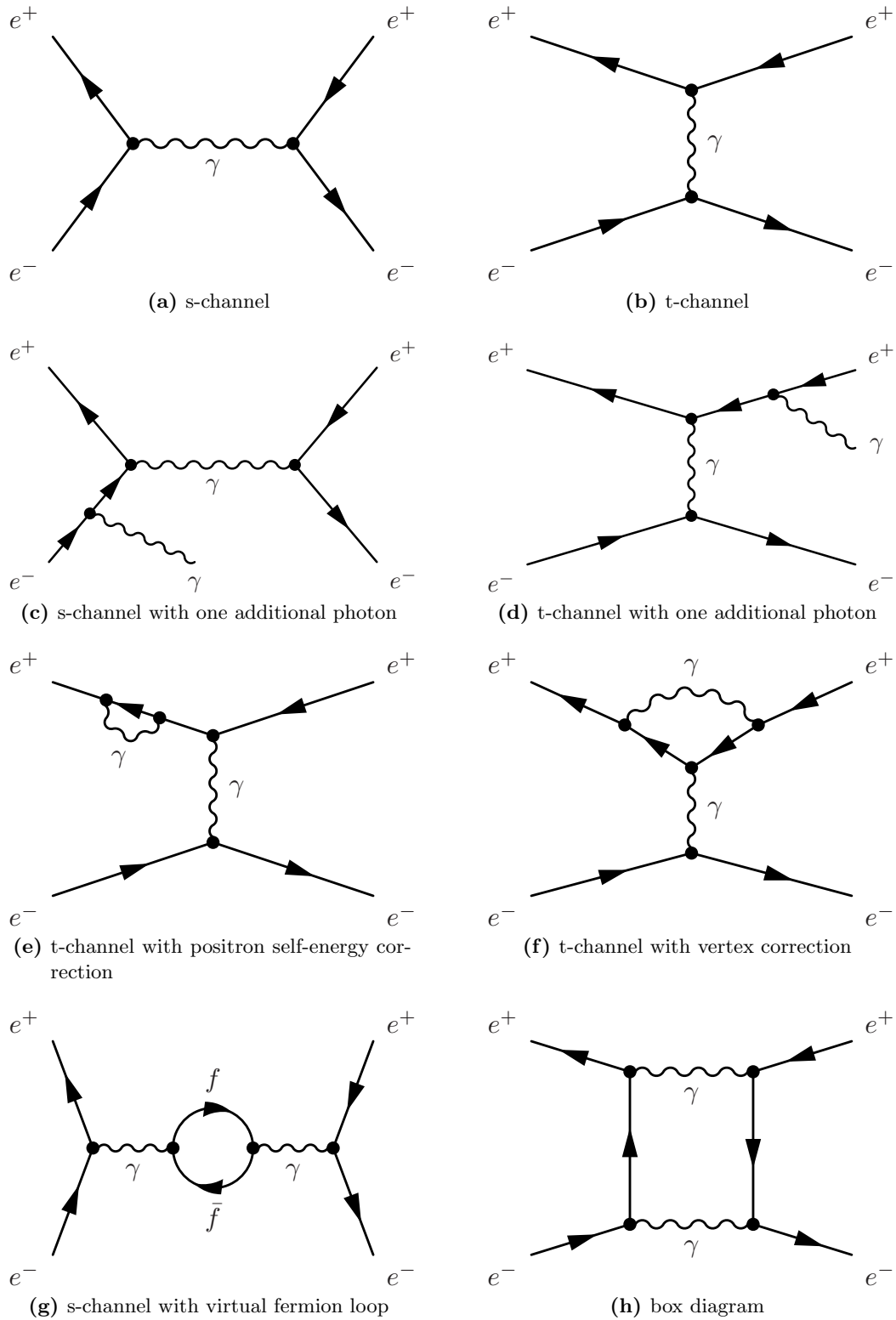
Figure 2.2: Exemplary Feynman diagrams for Bhabha (electron-positron) scattering [35]. Shown are leading order (LO) diagrams (a, b) and some next-to-leading order (NLO) diagrams (c,d). Additional NLO contributions in a perturbative expansion come from interference terms between the (a,b) and (e,f,g,h) diagrams.

respect to QED, leading to a higher complexity of the theory and a greater variety of QCD processes.

Confinement. Similarly to the running of the fine-structure constant α, the strong coupling constant α_s also depends on the momentum transfer Q^2, and accordingly, on the distance between two quarks. At distances of the order of atomic nuclei ($10^{-15}m$) and larger, which correspond to low values of Q^2, α_s and therefore the energy of the gluon field become so high that it is energetically favoured to create an additional quark-antiquark (q$\bar{q}$) pair rather than to have a single q$\bar{q}$ pair with large spatial separation. Therefore, quarks are never observed as free particles, but rather inside colour-neutral hadrons. This phenomenon is called *confinement*, and the process of creation of colour-neutral hadrons initiated by single quarks is called *hadronisation.*

The hadrons, in which quarks are confined, can be of two types: *mesons* (quark-antiquark pair) and *baryons* (three quarks or three antiquarks). Also exotic mesons that consist of four quarks (*tetraquarks*) are allowed in QCD. Their existence is supported by several strong experimental evidences. In particular, a tetraquark candidate $Z(4430)$ has been measured by the Belle [40] and LHCb [41] experiments with a significance up to 13.9σ.

Of special importance in this thesis are protons, which are the initial particles that are scattered in the described experiment. The proton is a baryon consisting of three valence quarks *uud* together with gluons and fluctuating sea quarks, which appear and disappear as q$\bar{q}$ pairs. All the constituents of the hadron are usually referred to as *partons.* More details on the parton model of hadron structure will follow in Section 2.2.

Asymptotic freedom. In contrast to the confinement at large distances, at low distances between quarks and at high values of Q^2, the strength of the strong interaction α_s becomes small: $\mathcal{O}(0.1)$ at the scale $Q = m_Z$. This feature of QCD is called *asymptotic freedom* and is crucial for the calculation of QCD processes at higher orders. Only with small values of α_s strong interactions can be calculated in perturbation theory, as such calculations are expanded in orders of α_s. The single terms of the expansion can be illustrated as distinct Feynman diagrams, some of which are shown in Figure 2.3. Otherwise, when the characteristic scale of a QCD process is of the order of $\Lambda_{QCD} = \mathcal{O}(200-300 \text{ MeV})$, the coupling reaches the order of unity and perturbative QCD cannot be applied.

Jet fragmentation. If the energy of a quark and its lifetime are large enough, whole cascades of q$\bar{q}$ pairs can be produced followed by subsequent additional soft radiation, hadronisation and decay, leading to a narrow cone of hadrons, usually called a *jet*. If a jet was initiated by a heavy quark, it is referred to as a *heavy-flavour jet*. More specifically, a jet initiated by a bottom quark is called a *b jet*, while a jet produced from a charm quark is called a *c jet*. Jets produced from light quarks (u,d,s) or gluons are collectively referred to as *light jets*.

The process of formation of jets from high-energetic quarks or gluons involves both perturbative and non-perturbative effects. In particular, the hadronisation of partons into colourless hadrons is parameterized by *fragmentation functions*, which are determined phenomenologically and are tuned in simulations to describe the observed data best.

(a) LO

(b) NLO (additional gluon radiation)

(c) NLO (virtual q$\bar{\text{q}}$ loop correction) interference term with the LO diagram

(d) NNLO (virtual loop + additional gluon radiation) interference term with the NLO diagram (b)
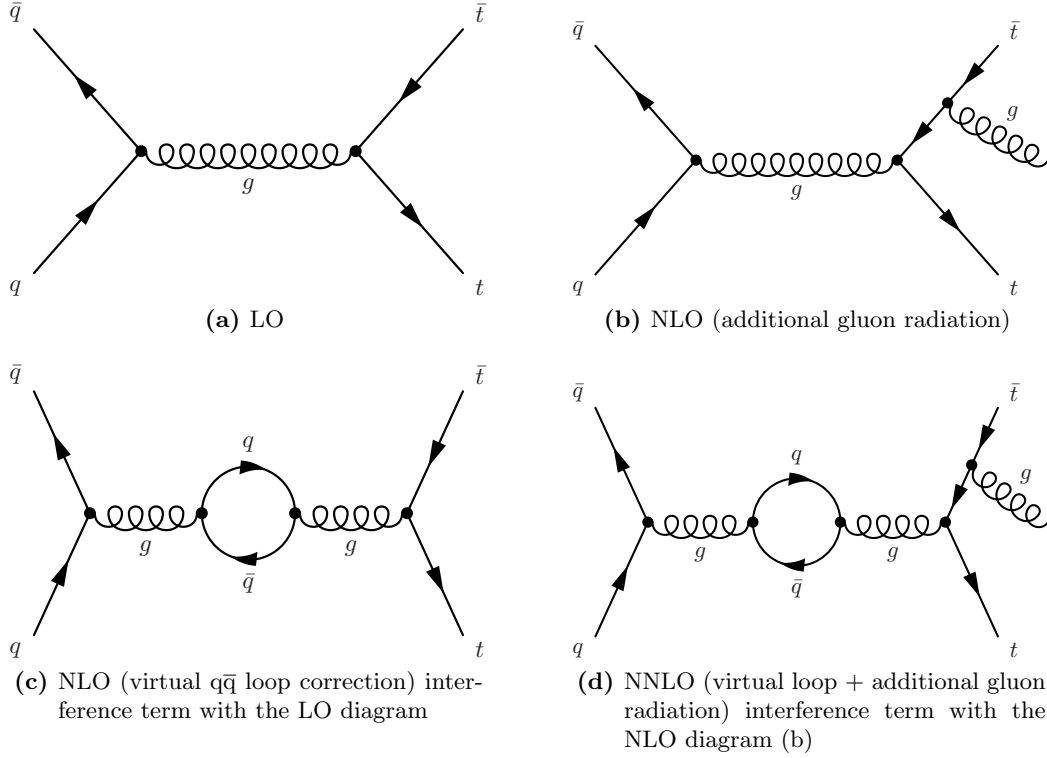
Figure 2.3: Exemplary Feynman diagrams for $t\bar{t}$ production via the quark-antiquark annihilation process. Shown are diagrams entering into different terms of a perturbative expansion.

Jets, especially b jets, are particularly important for the studies described in this thesis due to the presence of multiple bottom quarks in final states of the analysed processes.

Weak interaction

The weak interaction can involve all types of fermions — quarks and leptons. One of the most prominent effects of the weak interaction is the radioactive β decay. Within the Standard Model, weak interactions happen by the exchange of the electroweak gauge bosons $W^{\pm}$ and Z^0. The β decay is an interaction with charged currents, involving the $W^{\pm}$ boson, while neutral-current weak interactions via the exchange of the Z^0 boson are also possible. Exemplary Feynman diagrams of charged- and neutral-current weak interactions are shown in Figure 2.4.

The large mass of the weak gauge bosons implies a strong restriction on the range of the weak interaction. Therefore, at the nuclear scale, the strength of the weak interaction is very small compared to that of the electromagnetic interaction. Nevertheless, at the scale of $\mathcal{O}(100\,\text{GeV})$ and higher, the weak interaction becomes close in strength to the electromagnetic force.

The charged weak gauge bosons ($W^{\pm}$) couple only to left-handed fermions and right-handed antifermions. Therefore, weak interactions violate parity P and charge conjugation C. The combined operation CP is almost conserved with small violations that have been measured i.e. in [42] and [43]. Since CPT has to be conserved in any field theory, time inversion T is violated as well.

$\bar{u}$ $\bar{\nu}_e$ $\bar{q}$ e^+

W^- Z^0

d e^- q e^-

(a) charged current   **(b)** neutral current
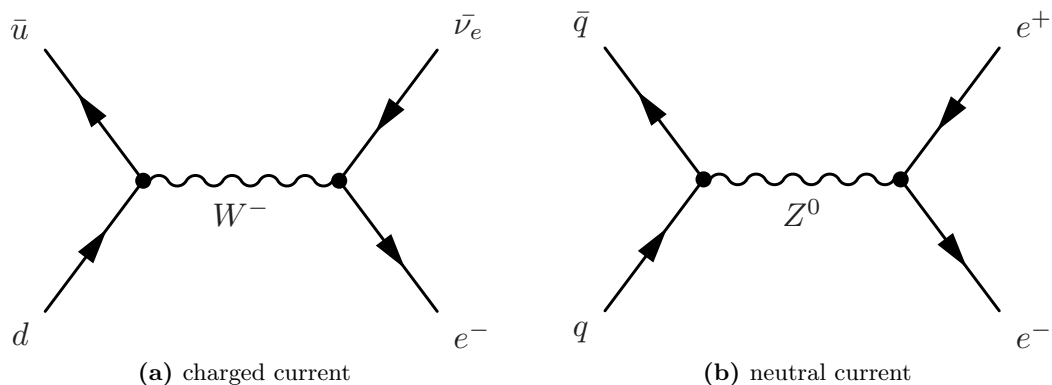
Figure 2.4: Exemplary Feynman diagrams for the weak interaction.

The weak eigenstates of the quarks are different from the mass eigenstates, leading to possible transformations of the quark flavour under the weak interactions. For example, an up-type quark converts into a down-type quark via flavour-changing charged currents depending on its coupling to the $W^\pm$ boson. Flavour-changing neutral currents by the exchange of the Z^0 boson have not been observed up to now. The transition probability between different quark flavours is described by the Cabibbo-Kobayashi-Maskawa matrix (CKM-matrix) [44]:

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb} \end{pmatrix} \times \begin{pmatrix} d \\ s \\ b \end{pmatrix} . \tag{2.9}$$

The matrix elements of the bottom row, which involve the top quark, are not yet directly measured very precisely. In particular, $|V_{td}|$ and $|V_{ts}|$ have only been indirectly measured from $B\bar{B}$ oscillations or loop-mediated K and B decays. In the Standard Model, the CKM-matrix is assumed to be unitary. Absolute values of some of the CKM-matrix elements, like V_{ud}, V_{us}, are known from experimental measurements of specific weak decays with high precision $\mathcal{O}(0.02\,\% - -0.3\,\%)$, while some others, e.g. V_{td}, V_{ub}, have fairly large uncertainties $\mathcal{O}(7\,\% - 12\,\%)$ [27]. Direct measurements of $|V_{tb}|$ from single-top-quark production cross sections yield a value close to unity: $|V_{tb}| = 0.998 \pm 0.041$ [45]. This fact has important implications for the analyses described in the thesis, which involve top-quark production. Due to $|V_{tb}| \approx 1$, the top quark almost exclusively decays into a W boson and a b quark. In some simulations of top-quark decays this element is even set to one for simplicity: $|V_{tb}| = 1.0$, which eliminates all other decay modes of the top quarks except $t \to bW$.

A similar mixing has been observed between neutrinos, whose mass eigenstates are not the same as the flavour eigenstates. This leads to neutrino oscillations between different generations and is a strong evidence for their non-zero masses. Currently neutrinos in the Standard Model are defined as massless, but their masses can be introduced as an extension.

2.1.3 Electroweak unification and symmetry breaking

As mentioned before, at high interaction energy scales, the electromagnetic and weak interactions are unified into a combined electroweak interaction. It is formulated

into a gauge theory developed by S. L. Glashow [46], A. Salam [47] and S. Weinberg [48], which requires the Lagrangian to be invariant under transformations from the $SU(2)_L \otimes U(1)_Y$ symmetry group.

In the electroweak theory, four vector fields are introduced, three of which are associated with the adjoint representation of the $SU(2)$ group (W_μ^1, W_μ^2, W_μ^3) and one with the $U(1)$ group (B_μ). Gauge invariance of the Lagrangian is fulfilled by replacing δ_μ in the fermion kinetic energy terms by the covariant derivative D_μ:

$$\partial_\mu \to D_\mu \equiv \partial_\mu + igI_jW_\mu^j + ig'\frac{Y}{2}B_\mu \ , \tag{2.10}$$

where g and I_j are the coupling and group generator of the $SU(2)$ group respectively, while g' and $\frac{1}{2}Y$ are the coupling and group generator of the $U(1)$ group, respectively.

The experimentally observed parity violation is incorporated by assigning the left- and right-handed components of the fermions to different group representations. All the left handed fermions are taken to transform as doublets under $SU(2)$, while the right-handed fermions are singlets, as shown in Table 2.1. The generators act on fermions as follows:

$$I_j\psi_\mathrm{L} = \frac{1}{2}\tau_j\psi_\mathrm{L} \ , \qquad I_j\psi_\mathrm{R} = 0 \ , \tag{2.11}$$

where the τ_j are the 2×2 Pauli matrices [49].

Table 2.1: Classification of the three generations of SM fermions by their handedness, according to the Glashow-Salam-Weinberg theory. The quarks d', s', b' denote superpositions of the mass eigenstates d, s, b in accordance with the CKM-matrix (2.9).

	I		**II**		**III**	
quarks:	$\begin{pmatrix} u \\ d' \end{pmatrix}_L$	u_R, d_R	$\begin{pmatrix} c \\ s' \end{pmatrix}_L$	c_R, s_R	$\begin{pmatrix} t \\ b' \end{pmatrix}_L$	t_R, b_R
leptons:	$\begin{pmatrix} \nu_e \\ e \end{pmatrix}_L$	e_R	$\begin{pmatrix} \nu_\mu \\ \mu \end{pmatrix}_L$	μ_R	$\begin{pmatrix} \nu_\tau \\ \tau \end{pmatrix}_L$	τ_R

The electroweak hypercharge Y is related to the electromagnetic charge Q via its linear combination with the weak isospin I_3, as given by the Gell-Mann-Nishijima relation:

$$Y = 2(Q - I_3) \ . \tag{2.12}$$

The weak isospin quantum number $I_3 = \frac{1}{2}$ is assigned to neutrinos and left-handed up-type quarks, while charged leptons and down-type quarks have $I_3 = -\frac{1}{2}$. Corresponding antiparticles have the same value of the weak isospin, but with the opposite sign. A weak isospin $I_3 = 0$ is assigned to all right-handed particles. An overview of the weak isospin and hypercharge assignments is given in Table 2.2.

The strength of the $SU(2)$ and $U(1)$ gauge fields is given by the tensors:

$$W_{\mu\nu}^i = \partial_\mu W_\nu^i - \partial_\nu W_\mu^i - g\varepsilon_{ijk}W_\mu^j W_\nu^k \ , \tag{2.13}$$

$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu \ . \tag{2.14}$$

Table 2.2: An overview of weak isospin and hypercharge assignments to the SM fermions according to the Glashow-Salam-Weinberg theory.

Fermions			I_3^L	Y_L	I_3^R	Y_R	Q
u	c	t	$+\frac{1}{2}$	$+\frac{1}{6}$	0	$+\frac{2}{3}$	$+\frac{2}{3}$
d	s	b	$-\frac{1}{2}$	$+\frac{1}{6}$	0	$-\frac{1}{3}$	$-\frac{1}{3}$
e	μ	τ	$-\frac{1}{2}$	$-\frac{1}{2}$	0	-1	-1
ν_e	ν_μ	ν_τ	$+\frac{1}{2}$	$-\frac{1}{2}$	$-$	$-$	0

The last bilinear term $W_{\mu\nu}$ in (2.13) generates triple and quartic self-coupling of the W_μ fields, which is a characteristic of non-Abelian gauge theories and implies that W bosons carry electroweak charge. In contrast, the B boson has no self-coupling, therefore the electroweak hypercharge $Y = 0$ and the weak isospin $I_3 = 0$ are assigned to it.

The physical fields of the weak bosons W^+ and W^- are identified as a superposition of the mentioned fields [39]:

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \pm iW_\mu^2) \tag{2.15}$$

The remaining fields W_μ^3 and B_μ^3 couple to the neutrinos and cannot represent the electromagnetic field. Instead, the electromagnetic field A_μ is defined as a linear combination of the two, and is orthogonal to the Z_μ term responsible for coupling to the neutrinos. These two fields can be written in matrix form as follows:

$$\begin{pmatrix} A_\mu \\ Z_\mu \end{pmatrix} = \begin{pmatrix} \cos\theta_W & \sin\theta_W \\ -\sin\theta_W & \cos\theta_W \end{pmatrix} \times \begin{pmatrix} B_\mu \\ W_\mu^3 \end{pmatrix}, \tag{2.16}$$

where θ_W denotes the *weak mixing angle* (or *Weinberg angle*), which is defined by:

$$\cos\theta_W = \frac{g}{\sqrt{g^2 + g'^2}}, \qquad \sin\theta_W = \frac{g'}{\sqrt{g^2 + g'^2}} \tag{2.17}$$

The mixing angle is a free parameter of the Standard Model, and is measured experimentally: $\sin^2\theta_W(Q) = 0.2397 \pm 0.5\,\%$ at $Q^2 = 0.026\,\text{GeV}^2$ [50].

The Higgs mechanism

Invariance of the electroweak Lagrangian under $SU(2)_L \otimes U(1)_Y$ transformations implies that electroweak gauge bosons are massless and fermions in each isospin doublet have the same mass. Nevertheless, experimental observations show that among electroweak bosons only the photon is massless while the others are massive, i.e. $m_{W^\pm} = 80.385 \pm 0.015\,\text{GeV}$ and $m_{Z^0} = 91.1876 \pm 0.0021\,\text{GeV}$ [27]. Therefore, the electroweak symmetry must be spontaneously broken.

In the Standard Model the breaking of the electroweak symmetry is achieved by the *Higgs mechanism* — the approach initially proposed and developed by Peter Higgs [16], Robert Brout and François Englert [15]. According to this method, complex scalar fields are introduced that couple gauge invariantly to the gauge

bosons and through the Yukawa coupling to the fermions. The requirement of being gauge-invariant implies an $SU(2)$ doublet ($I = \frac{1}{2}$) with hypercharge $Y = 1$, which can be written as:

$$\Phi = \begin{pmatrix} \Phi^+ \\ \Phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \Phi_1 + i\Phi_2 \\ \Phi_3 + i\Phi_4 \end{pmatrix} . \tag{2.18}$$

The Lagrangian density of the Higgs doublet is given as:

$$\mathcal{L}_H = (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(\Phi) , \tag{2.19}$$

where the potential term can be expressed as:

$$V(\Phi) = \mu^2 (\Phi^\dagger \Phi) + \lambda (\Phi^\dagger \Phi)^2 = \mu^2 \Phi^2 + \lambda \Phi^4 \tag{2.20}$$

(a) $\mu^2 > 0$ **(b)** $\mu^2 < 0$
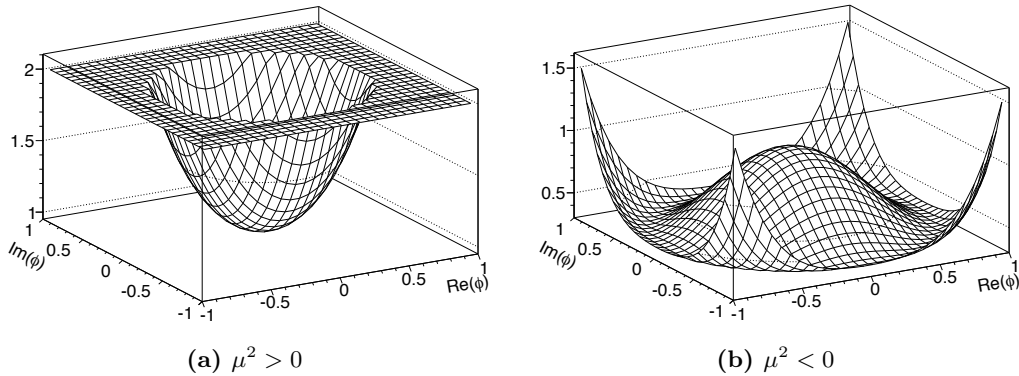
Figure 2.5: Higgs field potential $V(\Phi)$ for $\lambda > 0$ and different signs of μ^2 [51].

The potential $V(\Phi)$ defined in (2.20) has a vacuum state with a finite lower bound when $\lambda > 0$. Depending on the sign of μ^2 this vacuum state can be either unique ($\mu^2 > 0$, with minimum in $\Phi = 0$, as in Figure 2.5a) or degenerate ($\mu^2 < 0$, with minimum in a set of points, as in Figure 2.5b). For the symmetry to be broken μ^2 in the Higgs potential $V(\Phi)$ must be negative, which leads to a non-zero vacuum expectation value:

$$\langle 0 | \Phi | 0 \rangle = \sqrt{\frac{-\mu^2}{2\lambda}} = \frac{v}{\sqrt{2}} \neq 0 . \tag{2.21}$$

Once a particular ground state is chosen, the $SU(2)_L \otimes U(1)_Y$ symmetry gets spontaneously broken to $U(1)_{EM}$. Expanding Φ around its ground state the Higgs doublet can be redefined as:

$$\Phi = e^{\frac{i}{2}\sigma_j \theta^j} \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H \end{pmatrix} , \tag{2.22}$$

with four real fields θ^j ($j = 1, 2, 3$) and H. The θ^j describe massless *Goldstone bosons* [52], while H describes the Higgs boson, which obtains a mass of $m_H = \sqrt{2}\mu$. The mass of the Higgs boson is a free parameter of the Standard Model, with the most recent measured value $m_H = 125.7 \pm 0.4\,\text{GeV}$ [27].

The local invariance of the Lagrangian under $SU(2)_L$ tranformations allows to rotate the fields into the physical states with $\theta^i = 0$, acquiring masses for the $W^\pm$ and Z^0 bosons:

$$m_{W^\pm} = \frac{1}{2}vg \,, \tag{2.23}$$

$$m_{Z^0} = \frac{m_{W^\pm}}{\cos\theta_W} \,, \tag{2.24}$$

while the photon as a gauge boson of the electromagnetic field stays massless ($m_\gamma = 0$).

The scalar Higgs field couples fermion states of opposite helicity via Yukawa coupling in terms of the form:

$$\mathcal{L}_f = -\frac{g_f v}{\sqrt{2}}(\bar{f}_L f_R + \bar{f}_R f_L) + \frac{g_f h}{\sqrt{2}}(\bar{f}_L f_R + \bar{f}_R f_L) \,, \tag{2.25}$$

where the first term represents the fermion mass and the second term is the fermion coupling to the Higgs field. Therefore, spontaneous symmetry breaking generates fermion masses:

$$m_f = \frac{g_f v}{\sqrt{2}} \,, \tag{2.26}$$

with Yukawa coupling constant proportional to the fermion mass ($g_f \propto m_f$). This leads to the Higgs boson predominantly coupling to the top quark, bottom quark, charm quark and tau lepton. This theoretically predicted property is in good agreement with the experimental measurements [53], as demonstrated in Figure 2.6. This is why the measurement of associated production of top quark pairs with the Higgs boson decaying into a pair of bottom quarks is an important test of the Standard Model, as described in Chapter 10.
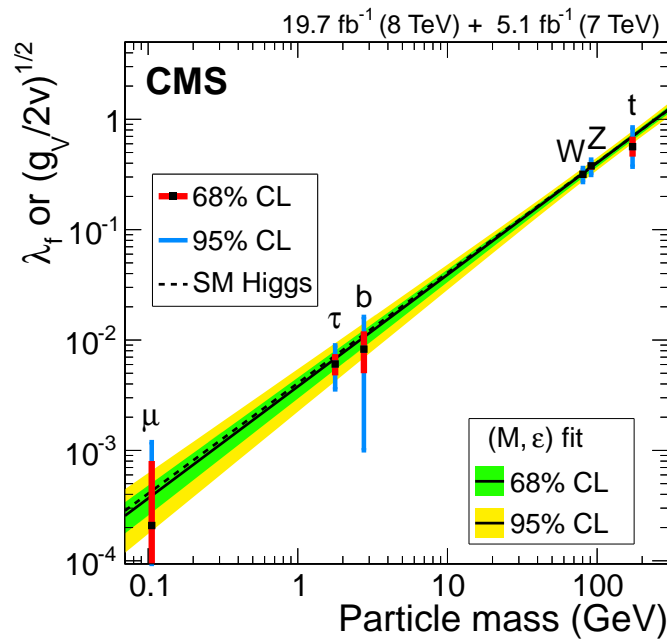
Figure 2.6: Summary of fits for deviations in the Higgs-boson coupling from the Standard Model, expressed as a function of the particle mass. Taken from [53].

2.1.4   Perturbative calculations

The total Lagrangian of the Standard Model is defined as:

$$\mathcal{L}_{SM} = \sum_{flavours} \bar{\Psi}(i\gamma_\mu D^\mu)\Psi + (D^\mu\Phi)^\dagger(D_\mu\Phi) - V(\Phi) - \frac{1}{4}F^a_{\mu\nu}F^{\mu\nu}_a + \mathcal{L}_{Yukawa} \ , \quad (2.27)$$

where $\mathcal{L}_{Yukawa}$ denotes the Lagrangian density of the Yukawa coupling between the fermions and the Higgs boson, $F^a_{\mu\nu}$ is the gauge field tensors, and D_μ is the covariant derivative.

Although the field equations derived from the SM Lagrangian (2.27) can not be solved analytically, it is possible to expand their solutions in perturbation series by orders of the coupling constant. In particular, cross sections can be calculated using matrix elements (ME) that represent the transition probability amplitude from the intial state Φ_i into a final state Ψ_f:

$$\mathcal{M}_{fi} = \int d^3x \Psi^\dagger_f S_{fi} \Psi_i \ , \quad (2.28)$$

where S_{fi} corresponds to the *scattering matrix*. Thus, the probability amplitude can be expanded in orders of the strong coupling strength g_s:

$$\mathcal{M}(g_s) = g^k_s \mathcal{M}_0 + g^{k+1}_s \mathcal{M}_1 + g^{k+2}_s \mathcal{M}_2 + \dots \ , \quad (2.29)$$

with each term representing a specific process with a corresponding number of strong interaction vertices, as shown in Figure 2.7.

(a) g^k_s **(b)** g^{k+1}_s: one leg (real)   **(c)** g^{k+2}_s: one loop (virtual)
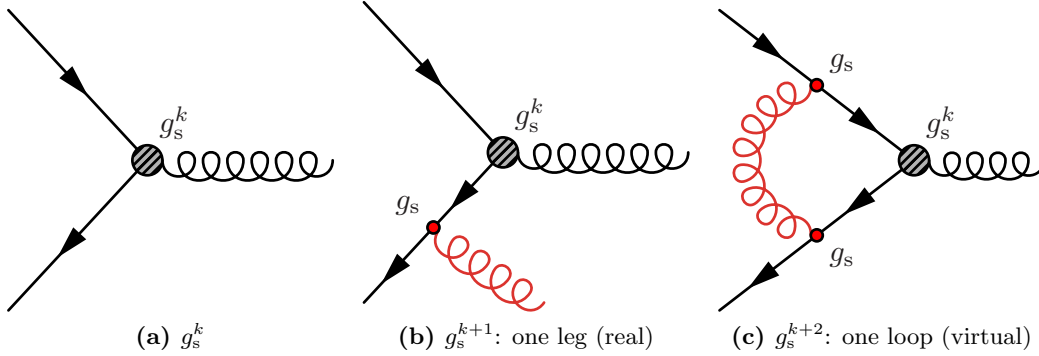
Figure 2.7: Schematic Feynman diagrams contributing to different terms in the expansion of the matrix element from (2.29) in orders of the strong coupling strength parameter g_s. Radiation of an additional real gluon or quark is called a *leg*, while the exchange of a virtual gluon or quark, which yields a closed curve, is called a *loop*.

The actual probability is then given by the square of the amplitude $\mathcal{M}$, which is expanded in powers of the strong coupling constant $\alpha_s = g^2_s$ under condition that $\alpha_s \ll 1$:

$$\mathcal{P} = |\mathcal{M}(g_s)|^2 = \underbrace{\alpha^k_s|\mathcal{M}_0|^2}_{\text{LO}} + \underbrace{\alpha^{k+1}_s\left[|\mathcal{M}_1|^2 + (\mathcal{M}^\dagger_2\mathcal{M}_0 + \mathcal{M}^\dagger_0\mathcal{M}_2)\right]}_{\text{NLO}} + \dots \quad (2.30)$$

Each term of the expansion would, therefore, correspond to a set of unique processes contributing to the total interaction with a factor of fixed power of the coupling constant, and can be represented by a corresponding set of Feynman diagrams, like those shown in Figure 2.3. If the production of a $t\bar{t}$ pair is defined as the process of interest, Figure 2.3a represents a diagram entering into the leading-order (LO) term of (2.30). The diagrams in Figures 2.3b and 2.3c have one additional leg and loop respectively, making them a part of the next-to-leading-order (NLO) term. Similarly the diagram shown in Figure 2.3d enters into an even higher-order NNLO term.

In principle, higher-order diagrams are always defined with respect to the LO diagrams by adding gluon radiation or loops. In particular, if the process of interest is $t\bar{t}$ production with an additional gluon, the diagram from Figure 2.3b would represent a LO contribution, while the interference between diagrams from Figures 2.3d and 2.3b would be a part of the one-loop NLO correction term.

To calculate a cross section of some process in any particular order of perturbation theory, all diagrams entering with a corresponding power of α_s have to be considered. Of course, with every higher order in the coupling constant more diagrams are possible, which significantly complicates the calculations. Therefore, cross sections are usually calculated up to NLO or NNLO, while some processes with more complicated final states are yet calculated only at LO.

Processes involving self-energy loop corrections for the gluon propagator, like the ones shown in Figures 2.3c and 2.3d, lead to so called ultraviolet (UV) divergencies, if the particle in the loop has momentum approaching infinity. Such divergencies can be removed by a *renormalisation* procedure, which is performed in two basic steps. Starting with regularisation, a cut-off parameter is introduced to split the integral into finite and infinite terms. In the renormalisation step, the divergent terms are absorbed by adding counterterms, which replace the dependence of the result on the cut-off parameters by a dependence on the *renormalisation scale* (μ_R). The counterterms are generated by redefining the "bare" coupling constants and particle masses by the "effective" ones. These effective parameters are therefore referred to as *running* constants and are functions of the renormalisation scale.

The applicability of this technique is the definition of renormalisable theories, and it was proven that locally invariant gauge theories are renormalisable [54]. The most widely used renormalisation scheme for QCD calculations is the *modified minimal subtraction* ($\overline{\text{MS}}$) scheme, developed by G. 't Hooft and S. Weinberg [54, 55].

2.1.5   Extensions of the Standard Model

The Standard Model successfully describes most of the experimental observations in high energy physics. Nevertheless, some observations and theoretical aspects of the model give evidence of its incompleteness.

One of the most obvious shortcomings of the SM is the missing incorporation of the gravitational interaction and the incompatibility of the theory of general relativity with field theory. Furthermore, studies of galaxy movements show that the visible matter can not describe the observed gravitational effects, suggesting that most of the energy density in the universe comes from *dark matter* [56], for which there is no suitable description within the Standard Model. Furthermore,

the observed accelerating expansion of the universe [57, 58] is typically attributed to an unknown form of energy, which is referred to as *dark energy*. In addition, the neutrinos are assumed to be massless in the Standard Model, while numerous experimental observations suggest that they have very small non-zero mass [59] [60].

The so-called hierarchy problem also exists between the scale of electroweak unification and the grand unification scale, at which electroweak and strong interactions are assumed to become of equal strength. The quantum effects of the gravitational interaction are assumed to become relevant at the even larger Planck scale. Due to radiative corrections of the same order as the mentioned scales, the Higgs boson mass should be much higher than that of other SM particles. The observed small mass can be explained only by a fine-tuning of SM parameters to make the quantum corrections cancel each other.

A number of theoretical models have been developed that solve some of these problems. Therefore, in addition to testing the Standard Model, physics analyses also allow to find signatures for or disprove other existing models.

One such model is Supersymmetry (SUSY) [61], signatures of which are heavily searched for at the LHC. Within this theory a symmetry between bosonic and fermionic fields is proposed, which implies that each fermion has a corresponding bosonic partner and vice versa. If supersymmetry would be an exact symmetry, all supersymmetric partners would have the same properties as the ordinary particles except of the spin. Corrections from such superpartners would cancel the divergencies caused by the SM particles to the Higgs boson mass. Another important feature of supersymmetric models is that many of them can provide candidates for the dark matter. Nevertheless, if superpartner particles would have the same properties as ordinary particles, they should have been observed, which is not the case. Therefore, if SUSY exists, it must be broken, i.e. the masses of the superpartner particles have to be much larger. More details about theoretical aspects and experimental implications of Supersymmetry can be found in the corresponding literature, e.g. [62, 63].

Another approach to solving some of the problems is taken by models that postulate the existence of extra space dimensions, which allow to include the description of gravity and to solve the hierarchy problem [64]. In most of them there is a finite number of additional dimensions, which are curled up and are of finite size. Such models can be distinguished by the universality of these dimensions. In some models only gravity can propagate into large extra dimensions, explaining why it is so weak [65, 66], while others, like Kaluza-Klein theory, suggest the existence of very small but universal extra dimensions [67].

Unlike Supersymmetry, which extends the Standard Model by external symmetries, the SM can be extended by additional internal gauge symmetries. Examples of such extensions are the Grand Unified Theories (GUTs) [68], according to which the Standard Model gauge group is embedded into a higher order gauge group (e.g. $SO(10), SU(5), E(6)$). Such symmetries imply that above some unification scale all interactions have the same strength, but at a lower energy scale the group breaks to subgroups. New observed resonances could be indications of some particular GUT models.

Also the Standard Model can be extended by a fourth generation of quarks [69], which might predominantly decay into quarks of the third generation and massive

gauge bosons. Some particular mass resonances can be used to verify the existence of such particles.

Searches for signatures of these models are an important part of the physics program of the LHC together with model independent searches for new phenomena beyond the Standard Model.

2.2   Proton-proton collisions

While the underlying theory of particle interactions can be elegantly formulated using the Lagrangian formalism, it can not be directly compared to experimental measurements due to the nature of accessing the measured observables.

Since most of the measured parameters of any theory in high-energy physics are accessed via particle scattering experiments, the deduced verifiable predictions are usually formulated as cross sections of the analysed scattering process. Therefore, two types of processes can be distinguished at high-energy hadron colliders: the *hard scattering* (rates and properties can be precisely predicted using perturbation theory) and the *soft interactions* (characterised mainly by non-perturbative effects and described by phenomenological models [70]). Processes studied within the scope of this thesis represent the hard scattering, nevertheless they are usually accompanied by additional soft processes, the understanding of which is also important.

On one hand, the production of top quarks, bottom quarks or the Higgs boson happens during hard interactions between initial state quarks or gluons. On the other hand, as the name of the LHC tells, this initial state is produced from hadronic collisions – proton-proton collisions, to be precise. In order to measure a cross section of a particular process, the number of such events has to be known, as well as the number of recorded proton-proton collisions. The latter has to be related to the number of events with a particular initial state configuration based on the internal hadronic structure of the proton [71], i.e. the distribution of partons (quarks, gluons) inside the proton [72].

An example of a proton-proton collision event is graphically shown in Figure 2.8. Perturbative QCD calculations are used for the parameterization of the hard processes. Soft processes, including hadronisation/fragmentation, parton distribution (density) functions (PDF), multiple parton interactions (MPI), as well as initial-state (ISR) and final-state (FSR) radiation are parameterized based on empirical models.

2.2.1   Proton structure

Protons, being hadrons, have inner structure that is described by the *QCD improved parton model*. It is based on the original *parton model* proposed by Richard Feynman [74] in 1969, and was first applied to deep inelastic electron-proton scattering by J. Bjorken and E. Paschos [75]. The original parton model, describing the hadrons as bound states of point-like constituents (partons) has been developed before the formulation of the QCD theory. Later it has been rethought, taking into account theoretical outcomes of the QCD theory and assumptions from the phenomenology of hadronic interactions, leading to the association of partons to quarks and gluons, which are known from QCD. Therefore, later in the text *parton model* always
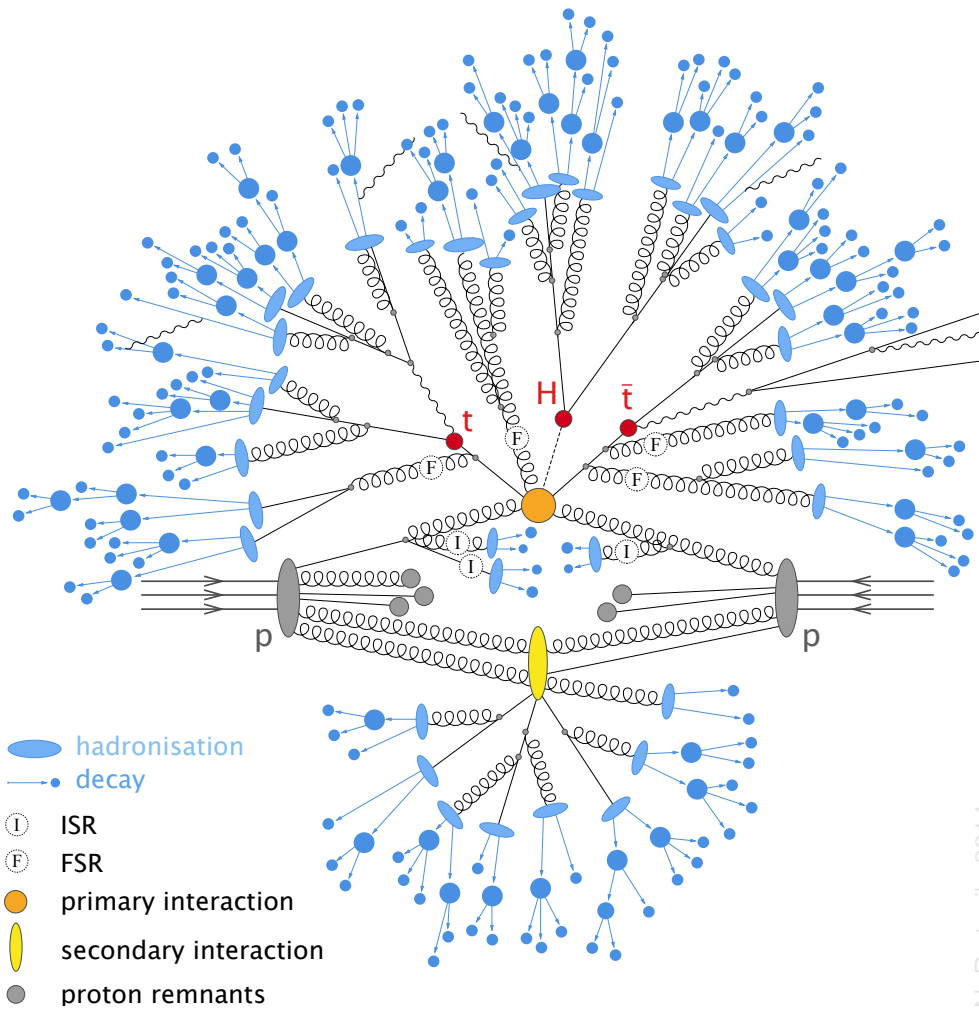
Figure 2.8: Graphical representation of a $t\bar{t}H$ event as produced in proton-proton collision simulation. It includes the hard interaction followed by the decay of the top quarks and the Higgs boson, additional QCD radiation, a secondary interaction as well as hadronisation of the final-state partons and hadron decays. Based on the diagram by F.Krauss from [73].

refers to the *QCD improved parton model* as the one dominantly used in high-energy physics.

According to the parton model, a proton is a bound state of three valence quarks: two up quarks and one down quark. The valence quarks constantly radiate and absorb gluons, which can split into $q\bar{q}$ pairs (sea quarks). Therefore, high energy collisions of protons can lead to not only scattering between valence quarks, but also between gluons and quarks of other flavours. In order to calculate the cross section of a process with any particular initial state, the probability of finding corresponding partons that carry the fraction x of the proton's momentum has to be known.

The momentum fraction distribution is parameterized by the *parton distribution functions* (PDFs) $f_{a|p}(x, Q^2)$ and depends on the energy scale Q^2. Once the PDF is defined at some starting scale, it can be derived at any other scale according to the DGLAP equations [76][77][78][79], which describe the Q^2-evolution of the PDF:

$$\frac{\partial f_{a|p}(x, Q^2)}{\partial \log Q^2} = \frac{\alpha_s}{2\pi} \int_x^1 \frac{dz}{z} \left\{ P_{aa'}(z, \alpha_s) f_{a'|p}(\frac{x}{z}, Q^2) \right\} , \qquad (2.31)$$

where $P_{aa'}(z, \alpha_s)$ denotes the splitting function – the transition probability of the parton a into a parton a' by emitting a quark or gluon, where the parton a' carries the momentum fraction z of the parton a momentum. This function is known as the Altarelli-Parisi splitting function [79] and can be calculated perturbatively in orders of α_s:

$$P_{aa'}(z, \alpha_s) = P_{aa'}(z) + \frac{\alpha_s}{2\pi} P_{aa'}^{\text{LO}}(z) + \left(\frac{\alpha_s}{2\pi}\right)^2 P_{aa'}^{\text{NLO}}(z) + \dots . \qquad (2.32)$$
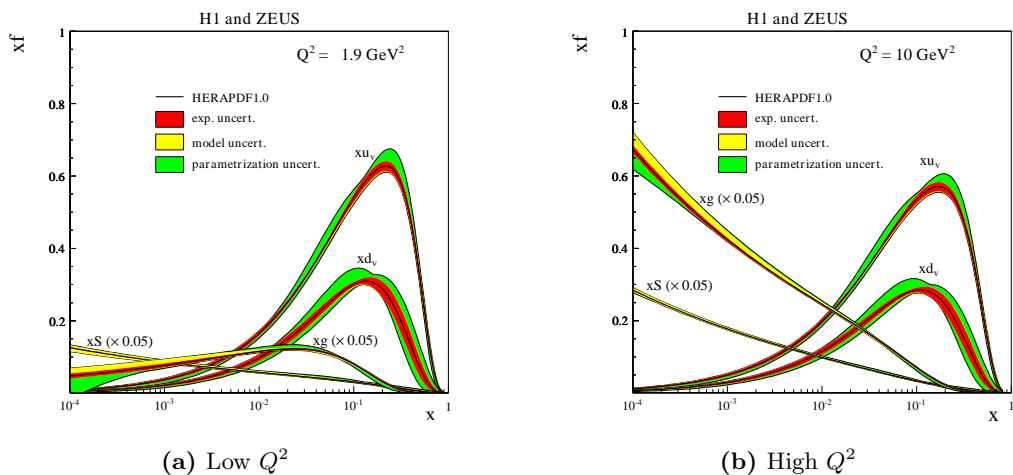
Figure 2.9: Distribution of the valence quarks (u_v, d_v), gluons (g) and sea quarks (S) multiplied by the carried fraction x of the proton momentum as functions of x for two momentum-transfer scales: $Q^2 = 1.9\,\text{GeV}^2$ and $Q^2 = 10\,\text{GeV}^2$. The distribution for gluons and sea quarks is scaled down by factor 20 in order to fit on the Y axis. Obtained from the HERAPDF1.0 set [80].

The actual PDFs are extracted from global fits of hard QCD measurements from different experiments. The most widely used PDFs at the LHC are those provided by the CTEQ [81], MSTW [82], NNPDF [83] and the HERAPDF project [80]. For

instance, CTEQ is using data from numerous experiments: lepton-hadron collider (ZEUS [84] and H1 [85]), hadron collider (D0 [86] and CDF [87]), and some fixed target experiments (BCDMS [88], NMC [89], CCFR [90]). HERAPDF is restricted only to data from the ZEUS and H1 experiments at the HERA accelerator.

An example of PDFs from the HERAPDF1.0 set [80] is shown in Figure 2.9. As can be seen from the distributions, at $Q^2 = 1.9\,\mathrm{GeV}^2$, valence quarks carry the largest fraction of the proton's momentum, while with decreasing x, the probability of finding a gluon or a sea quark increases. Since the scale at which the PDFs are evaluated also depends on the four-momentum transfer in the hard scattering, at higher Q^2 the relative contribution of gluons and sea quarks increases.

A distinction is usually made between two types of schemes: fixed-flavour- and variable-flavour-number scheme. In the *fixed-flavour-number scheme* (FFNS), the number of quark flavours in the hadron is fixed — it does not depend on the energy scale of the process. Three-flavour (u, d, s), four-flavour (u, d, s, c) and five-flavour (u, d, s, c, b) schemes are distinguished.

In the *variable-flavour-number scheme* (VFNS), the number of quark flavours in the hadron is defined by a step function that depends on the energy scale. Since *heavy flavour* (c, b) quarks are significantly heavier than *light flavour* (u, d, s) quarks, their contribution is suppressed at low Q^2 scales. Therefore, with higher Q^2 scale heavy flavour quarks are gradually added to the PDF.

2.2.2 Factorisation theorem

As stated before, protons are composite particles, consisting of valence quarks, sea quarks and gluons. Therefore, when calculating a cross section for any process that takes place in proton-proton collisions, such a complex initial state has to be taken into account.

According to the *factorisation theorem*, a cross-section for processes in a hadron-hadron collision can be represented as a convolution of the partonic cross section $\sigma_{ab\to X}$ with the corresponding PDF $f_{a|p}(x, Q^2)$. The partonic cross-section strongly depends on Q^2 at leading order in perturbative QCD, and depends less strongly on it at higher orders. Similarly to the renormalisation scale, which was described in Section 2.1.4, when calculated to all orders in perturbative QCD, the total hadronic cross section is completely independent of the factorisation scale Q^2. However, at any finite order, the calculated hadronic cross section depends on Q^2. In order to obtain a reliable prediction, higher order corrections have to be calculated up to the order at which the factorisation-scale dependence becomes negligible [91]. In case this is not possible, the remaining scale dependence has to be included as an uncertainty due to the corresponding variation of the Q^2.

The factorisation scale has to be carefully chosen such that the short-distance physics of the hard-scattering cross section is separated from the long-distance soft hadronic processes [92]. In this way the partonic cross section $\sigma_{ab\to X}$ is calculated using matrix elements in perturbation theory, as described in Section 2.1.4. Long-distance effects, like initial state collinear gluon radiation, are included in the PDF $f_{a|p}(x, Q^2)$, which is determined phenomenologically from experimental measurements.

2.2.3   Underlying event, pileup, non-collision background

As can be seen in Figure 2.8, beside the hard interaction of interest (primary inter-action, e.g. $t\bar{t}H$ production), several other processes can take place in the proton-proton collisions. Everything apart from the primary hard interaction is collectively called *underlying event*, which includes all secondary (usually softer) interactions constituting multiparton interaction (MPI), proton remnants. If the process is de-fined at leading order, initial-state (ISR) is also treated as underlying event. At higher orders, ISR and FSR become a part of the radiative corrections to the cross section, and are therefore included in the definition of the process. Most of the processes constituting the underlying event are not perturbatively calculable at the moment, therefore its description in simulations is based on phenomenological mod-els, which are tuned for any particular initial state to obtain best description of the measured data [93, 94].

What is not included in Figure 2.8 are real life conditions at which such proton-proton collisions occur. In particular, to achieve high collision rates, colliding protons are organised in bunches containing $\sim 10^{11}$ protons each. Therefore, the crossing of two bunches can lead to multiple proton-proton collisions taking place almost simultaneously. Such multiple collisions are referred to as *pileup*, and can create additional energy flow close to the primary interaction.

In addition to the underlying event and pileup, other contributions can appear that do not stem from proton-proton collisions. In particular, high energetic muons originating from cosmic showers can leave a signal in the detector or protons can collide with remaining gas in the beampipe, along which protons are accelerated, or with boundaries of the beampipe itself. Most of such events are identified and discarded, nevertheless some of them can survive and have to be considered in the measurement.

2.3   Processes relevant for the thesis

The physics analyses performed and described in this thesis, namely $t\bar{t}b\bar{b}$ (Chapter 9) and $t\bar{t}H$ (Chapter 10) production, are characterised by complex final states, and involve the heaviest fermion (top quark) and the heaviest boson (Higgs boson) of the Standard Model. In order to better understand the motivation and ideas behind the analyses, details of these processes are described in the following sections.

In particular, all the analyses involve the dileptonic final state of the $t\bar{t}$ system, therefore an overview of production and consequent decay of top-quark pairs is given in Section 2.3.1, with emphasis on the dileptonic final state. In the $t\bar{t}H$ analysis, events are studied with a $t\bar{t}$ pair produced in association with the Higgs boson, which decays into a pair of bottom quarks ($H \rightarrow b\bar{b}$). Therefore, production and decay of the Higgs boson is described in Section 2.3.2, and explicit details of the associated top-quark-pair and Higgs-boson production are given in Section 2.3.3.

2.3.1   Top quark production and decay

The top quark, the heaviest known fundamental particle of the Standard Model, was predicted in 1973 after the postulation of the existence of the third generation of

quarks in order to explain the experimentally observed CP violation [95]. Due to its very high mass ($m_t = 173\,\text{GeV}$ [27]), the top quark was experimentally discovered only in 1994 by the DØ [96] and CDF [97] experiments at the Tevatron proton-antiproton ($p\bar{p}$) collider. Many properties of the top quark have been measured with high precision by the ATLAS and CMS experiments at the LHC thanks to the high production rates of the top quark at energies that can be provided only by the LHC, which is the reason why it is often called a "top factory". An extensive overview of top quark physics can be found in e.g. [98] [99] [100].

Production mechanisms

The top quark can be produced via the electroweak interaction involving a Wtb vertex in a Feynman diagram. Leading-order diagrams for single-top-quark production are shown in Figure 2.10. Cross sections for single-top production at the LHC with the centre-of-mass energy $\sqrt{s} = 8\,\text{TeV}$ have been calculated up to approximate NNLO [101], with values shown in Table 2.3. As can be seen from the table, single top quarks are mainly produced through the t-channel processes. Such processes allow the investigation of properties of the electroweak interaction, in particular the structure of the tWb coupling. Since its cross section is directly proportional to $|V_{tb}|^2$, this is the only way to directly measure the CKM matrix element V_{tb} (2.9). Furthermore, single top production can be sensitive to some processes beyond the Standard Model.

(a) t-channel ($\sigma \propto \alpha_\text{s}\alpha_\text{W}^2$)   **(b)** s-channel ($\sigma \propto \alpha_\text{W}^2$)

(c) tW-channel ($\sigma \propto \alpha_\text{s}\alpha_\text{W}$)   **(d)** tW-channel ($\sigma \propto \alpha_\text{s}\alpha_\text{W}$)
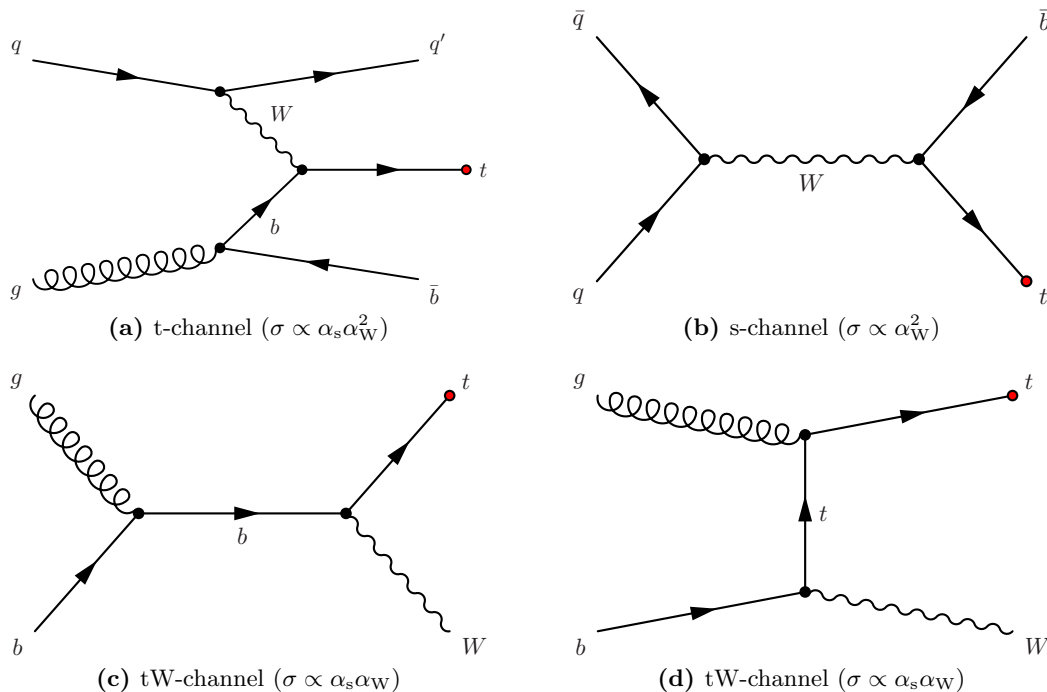
Figure 2.10: Exemplary Feynman diagrams for the single top quark production via the electroweak interaction.

Pairs of top and antitop quarks ($t\bar{t}$) are produced at the LHC via the strong interaction at much higher rates than single top quarks. At leading order, a $t\bar{t}$ pair can be produced either via quark-antiquark annihilation ($q\bar{q} \to t\bar{t}$) or via gluon-gluon

Table 2.3: Single-top-quark (t or $\bar{t}$) production cross sections from pp collisions at $\sqrt{s} = 8\,\mathrm{TeV}$ as calculated at approximate NNLO [101] for the top quark mass $m_t = 173$ GeV. The uncertainties account for renormalisation/factorisation scales and the PDF variation.

Channel	Cross section, σ [pb]
t-channel	$87.2^{+3.4}_{-2.4}$
s-channel	5.55 ± 0.23
tW-channel	22.2 ± 1.5

fusion ($gg \rightarrow t\bar{t}$), as shown by the Feynman diagrams in Figure 2.11. Quark-gluon scattering ($qg/\bar{q}g \rightarrow t\bar{t}q/\bar{q}$) contributes to $t\bar{t}$ production only at next-to-leading order.

Unlike the Tevatron, where $t\bar{t}$ pairs are mainly produced via q$\bar{\mathrm{q}}$ annihilation, at the LHC top-quark pairs are dominantly produced via gluon-gluon fusion. This is a consequence of the higher center-of-mass energy of the collisions at the LHC. At the Tevatron, with proton-antiproton collisions at $\sqrt{s} = 1.96\,\mathrm{TeV}$, the production of a $t\bar{t}$ pair requires the initial state partons to carry a large fraction of the colliding hadrons' momentum:

$$x \geq \frac{2m_t}{\sqrt{s}} = \mathcal{O}(2 \times 10^{-1}) \,, \tag{2.33}$$

while at the LHC, with $\sqrt{s} = 8\,\mathrm{TeV}$, this fraction is smaller: $x = \mathcal{O}(4 \times 10^{-2})$. These two values of x correspond to different ranges of the PDF shown in Figure 2.9b, from which it can be seen that, indeed, in the region of x typical for $t\bar{t}$ production at the Tevatron, valence quarks make the dominant contribution, while for the LHC, at $x = \mathcal{O}(4 \times 10^{-2})$, the dominant contribution comes from gluons.

The total $t\bar{t}$-production cross section at $\sqrt{s} = 8\,\mathrm{TeV}$ has been calculated up to NNLO accuracy [102] including next-to-next-to-leading logarithmic (NNLL) soft-gluon resummation [103, 104, 105]: $\sigma_{t\bar{t}} = 245.8^{+8.8}_{-10.6}$pb for the top quark mass $m_t = 173.3$ GeV. However, the dependence of the cross section on kinematic properties of the top quarks or their decay products (differential cross section), which is more sensitive to possible new physics or problems in the SM, is currently calculated only up to approximate NNLO accuracy [101].

Decay modes

A unique property of the top quark is its very large mass, $m_t = 173.21 \pm 0.87$ GeV [27], which leads to a large value of the decay width (Γ_t):

$$\Gamma(t) = 2.00^{+0.47}_{-0.43} \text{ GeV} \,, \tag{2.34}$$

resulting in a very short lifetime (τ_t):

$$\tau_\mathrm{t} = \frac{1}{\Gamma(t)} = 5 \times 10^{25}\,\mathrm{s} \,, \tag{2.35}$$

(a) s-channel

(b) t-channel

(c) u-channel

(d) q$\bar{q}$ annihilation
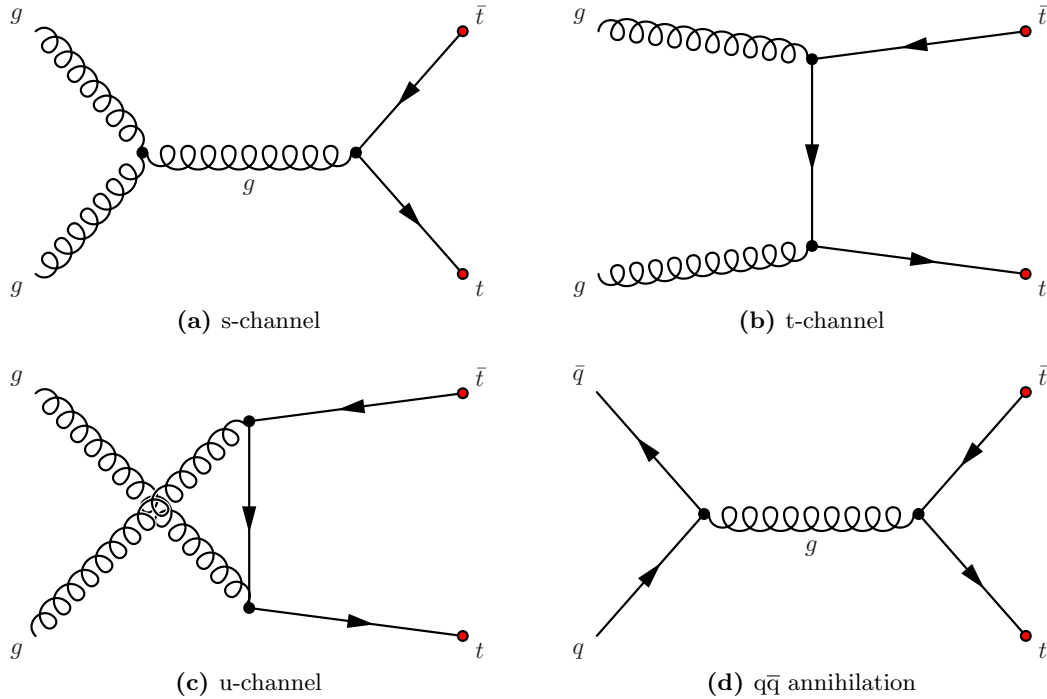
Figure 2.11: Exemplary Feynman diagrams for $t\bar{t}$ production via strong interaction at leading order.

which is about one order of magnitude smaller than the hadronisation time scale:

$$\tau_{\text{had}} \propto \frac{1}{\Lambda_{\text{QCD}}} = 3 \times 10^{-24}\text{s}. \tag{2.36}$$

Consequently, top quarks can not form hadrons, which allows to directly access properties of the *bare* quark. Nevertheless, this advantage is nearly cancelled by the fact that the top quark always decays into a down-type quark and a W boson, which, in its turn, decays further into quarks or leptons. Therefore, a study of the *bare* quark properties implies the measurement of its decay products, which do not always have such unique properties.

Since the top quark decays through the weak interaction, it produces a W boson and a down-type quark (down, strange, bottom). The probability of a particle to decay into a specific final state with respect to the probability of decaying to any final state is referred to as *branching ratio* or *branching fraction*:

$$\text{BR}(X) = \frac{\Gamma(X)}{\Gamma_{total}}, \tag{2.37}$$

and for the top quark is defined by elements of the CKM matrix (2.9). As already mentioned in Section 2.1.2, the element V_{tb} is close to one, which leads to the fact that the top quark almost exclusively decays into a W boson and a b quark ($t \rightarrow Wb$):

$$\text{BR}(t \rightarrow Wb) = \frac{\Gamma(t \rightarrow Wb)}{\Gamma(t \rightarrow Wb|s|d)} = |V_{tb}|^2 = 0.91 \pm 0.04 \quad [27]. \tag{2.38}$$

Taking the branching ratio $\text{BR}(t \rightarrow Wb) \approx 1$, the final state of the top quark decay is usually defined exclusively in terms of the W boson decay products — a

fermion-antifermion pair. Two basic types of the W boson decay are distinguished: *leptonic* ($W \to l\bar{\nu}_l$, where $l = e, \mu, \tau$), and *hadronic* ($W \to \bar{q}_\uparrow q_\downarrow$, where $\bar{q}_\uparrow = \bar{u}, \bar{c}$ and $q_\downarrow = d, s, b$), which is dominated by the $W \to \bar{u}d$ and $W \to \bar{c}s$ modes. Decays of the W boson with the top and bottom quarks in the final state are kinematically forbidden since their combined mass is higher than that of the W boson.

In total, there are 9 possible decay modes for a W boson: 3 leptonic (one for each lepton combination) and 6 hadronic (2 quark-antiquark combinations $\times$ 3 colour states). Due to the small masses of the fermions from the W decay compared to the W boson mass $m_W = 80.4 \, \mathrm{GeV}$ [27], branching fractions for each particular decay mode are approximately multiples of $\frac{1}{9}$:

$$\mathrm{BR}(W \to e\bar{\nu}_e) \approx \mathrm{BR}(W \to \mu\bar{\nu}_\mu) \approx \mathrm{BR}(W \to \tau\bar{\nu}_\tau) \approx \frac{1}{9}, \qquad (2.39)$$

$$\mathrm{BR}(W \to \bar{u}d) \approx \mathrm{BR}(W \to \bar{c}s) \approx \frac{3}{9}. \qquad (2.40)$$

This assumption is made in some simulations and is usually a good approximation for measurements that are not sensitive to the relative contribution of individual decay modes. Nevertheless, these branching fractions have been measured with very high precision, $\mathcal{O}(1.5\,\%)$, and there is no reason not to use the exact values, which are listed in Table 2.4a.
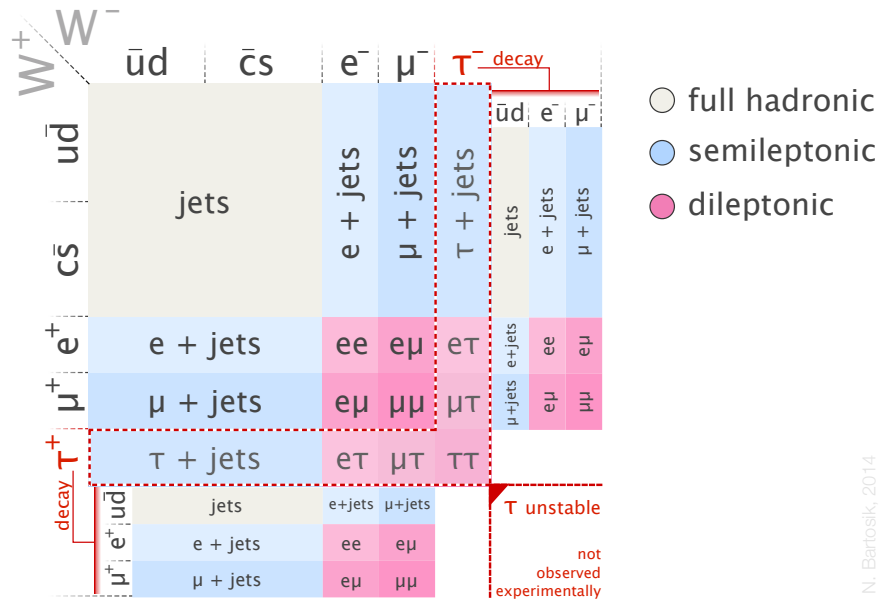
Figure 2.12: Possible decay channels of a $t\bar{t}$ pair in terms of decay modes of each W boson, including further decays of the unstable τ leptons. Three basic colours (gray, blue, magenta) represent the three usually distinguished decay channels (full hadronic, semileptonic, dileptonic) based on the number of leptons in the final state.

When a pair of top quarks is considered, its final state is determined by the decay mode of each individual W boson. An overview of different possible final states of the $t\bar{t}$ system is given by the diagram in Figure 2.12. Usually three basic decay modes (channels) are distinguished when a $t\bar{t}$ pair is concerned:

- **full hadronic:** both W bosons decay hadronically;

- **semileptonic:** one W boson decays hadronically and one leptonically;

- **dileptonic:** both W bosons decay leptonically.

Nevertheless, as can be seen from the overview diagram, there is an ambiguity in the definition of the $t\bar{t}$ decay channel if it is based purely on the W boson decays. This ambiguity comes from the τ leptons, which have the mean lifetime $\tau_\tau = 2.9 \times 10^{-13}$ s, and decay either leptonically or hadronically. Since only stable particles can be detected experimentally, decays of the τ leptons have to be taken into account with the corresponding branching ratios $\mathrm{BR}(\tau \to X)$, as shown in Table 2.4b. The final branching ratios for final states with stable particles including τ leptons can be easily calculated using the values in Tables 2.4a and 2.4b, and are listed in Table 2.5. In particular, if a $t\bar{t}$ system in the "dileptonic" channel is defined in terms of the W boson decays as $t\bar{t} \to bW^+\bar{b}W^- \to b(e^+\nu_e)\bar{b}(\tau^-\bar{\nu}_\tau)$ and the τ lepton decays hadronically $(\tau^- \to \bar{u}d)$, the experimentally accessible final state would be semileptonic instead of dileptonic. Therefore, it is necessary to explicitly state how τ leptons are treated when defining any particular $t\bar{t}$ decay channel to avoid confusion.

Table 2.4: Branching fractions of the W boson and the τ lepton. Values taken from [27]. $\mathrm{BR}(\tau \to hadrons)$ obtained by $1 - \big[\, \mathrm{BR}(\tau \to e\bar{\nu}_e\nu_\tau) + \mathrm{BR}(\tau \to \mu\bar{\nu}_\mu\nu_\tau) \,\big]$.

(a) W boson

Decay	B, [%]
$W \to had.$	67.41 ± 0.27
$W \to e\bar{\nu}_e$	10.71 ± 0.16
$W \to \mu\bar{\nu}_\mu$	10.63 ± 0.15
$W \to \tau\bar{\nu}_\tau$	11.38 ± 0.21

(b) τ lepton

Decay	B, [%]
$\tau \to had.$	64.76 ± 0.08
$\tau \to e\bar{\nu}_e\nu_\tau$	17.83 ± 0.04
$\tau \to \mu\bar{\nu}_\mu\nu_\tau$	17.41 ± 0.04

Table 2.5: Branching fractions of a W boson into stable particles, including τ decays

Decay	Formula	B, [%]
$W \to had.$	$\mathrm{BR}(W \to had.) + \mathrm{BR}(W \to \tau\bar{\nu}_\tau) \times \mathrm{BR}(\tau \to had.)$	74.78 ± 0.41
$W \to e$	$\mathrm{BR}(W \to e\bar{\nu}_e) + \mathrm{BR}(W \to \tau\bar{\nu}_\tau) \times \mathrm{BR}(\tau \to e\bar{\nu}_e\nu_\tau)$	12.74 ± 0.20
$W \to \mu$	$\mathrm{BR}(W \to \mu\bar{\nu}_\mu) + \mathrm{BR}(W \to \tau\bar{\nu}_\tau) \times \mathrm{BR}(\tau \to \mu\bar{\nu}_\mu\nu_\tau)$	12.61 ± 0.19

Dileptonic decay channels, which are the subject of this thesis, are defined in terms of the final state after potential τ lepton decay, and are denoted by the combination of charged leptons from the W and τ decays in the final state, namely ee, $e\mu$, $\mu\mu$. The branching fractions of each final state can be easily calculated from the values in Table 2.5, and are listed in Table 2.6.

As can be seen from Table 2.6, the combination of all dileptonic channels (marked in magenta) gives the branching fraction $\mathrm{BR}(t\bar{t} \to X\ell^+\ell^-) = 6.43\,\%$, which is significantly smaller than the semileptonic or full hadronic channels. Nevertheless, the lower cross section is compensated by fewer background rates. Moreover, in order to access the properties of the top quarks, the final state particles from each

Table 2.6: Branching fractions of a $t\bar{t}$ pair into different stable final states with the colour coding as in the diagram in Figure 2.12.

	$jets$	e^-	μ^-
$jets$	55.92 %	9.53 %	9.43 %
e^+	9.53 %	1.62 %	1.61 %
μ^+	9.43 %	1.61 %	1.59 %

top quark have to be combined individually, leading to a reconstructed $t\bar{t}$ system. Since the charge of the leptons can be detected with much better precision than that of hadronised quarks, it is easier to reconstruct the $t\bar{t}$ system in the dileptonic channel compared to the semileptonic and full-hadronic channels.

2.3.2 Higgs-boson production and decay

The Higgs boson, the heaviest boson of the Standard Model, was predicted in the 1960s as a result of the spontaneous electroweak symmetry breaking mechanism, as described in Section 2.1.3. Since the mass of the Higgs boson is a free parameter in the Standard Model, it took about 50 years of continuous research at particle colliders with ever increasing collision energies and precision to pin down the range of masses of the Higgs boson that would be consistent with the Standard Model. This search has finished when the H boson was finally discovered at the LHC [14, 13].

Production mechanism

It took so long to discover the Higgs boson due to its large mass, $m_H = 125.7\,\text{GeV}$ and small couplings to light particles. The proton-antiproton collision center-of-mass energy $\sqrt{s} = 1.96\,\text{TeV}$ at the Tevatron was enough to produce the Higgs boson, but sensitivity was not high enough to claim a discovery. Nevertheless, the combination of search results from the CDF and DØ experiments showed an excess of observed events with respect to the background estimation corresponding to a significance of 2.5σ (standard deviations) in the mass region of $115 < m_H < 140\,\text{GeV}$ [106]. This excess, shown in Figure 2.13, has been confirmed at the LHC with the higher center-of-mass energy $\sqrt{s} = 7\,\text{TeV}$ and $\sqrt{s} = 8\,\text{TeV}$, as shown in Figure 2.14.

At the LHC, four major production mechanisms of the Higgs boson are predicted by the Standard Model: gluon-gluon fusion, vector-boson fusion, production in association with a vector boson ($VH, V = W, Z$) via Bremsstrahlung (Higgsstrahlung), and production in association with top quarks ($t\bar{t}H$). Examples of leading-order Feynman diagrams for each of the processes are shown in Figure 2.15.

As in case of $t\bar{t}$ production, a dominant contribution to the Higgs-boson production is the *gluon-gluon fusion* process: $pp \to gg \to H$. It can be seen from the set of cross sections for different Higgs-boson production channels at the LHC at $\sqrt{s} = 8\,\text{TeV}$ as functions of m_H, shown in Figure 2.16. The corresponding diagram at the lowest order in perturbation theory is shown in Figure 2.15a, and involves a quark loop. Since the coupling between the Higgs field and a fermion is proportional
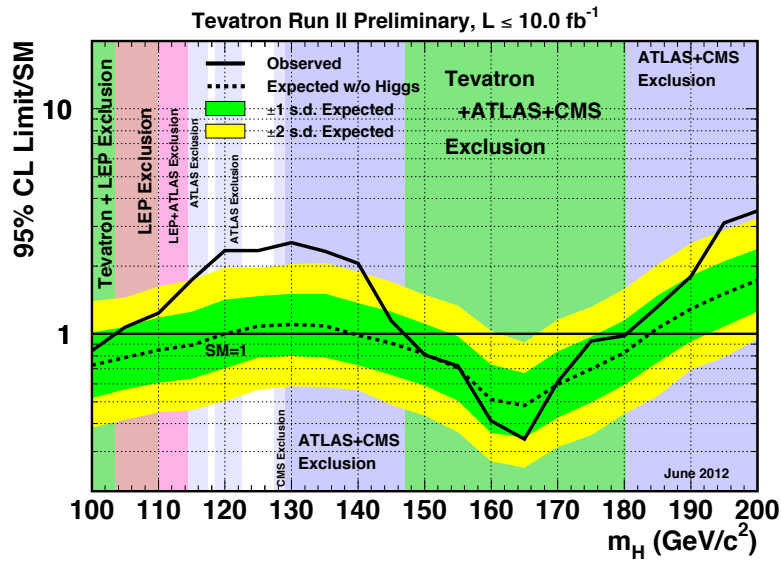
Figure 2.13: Observed and expected 95 % C.L. upper limits on the ratios to the SM cross section, as functions of the Higgs boson mass for the combined CDF and D0 analyses [106].
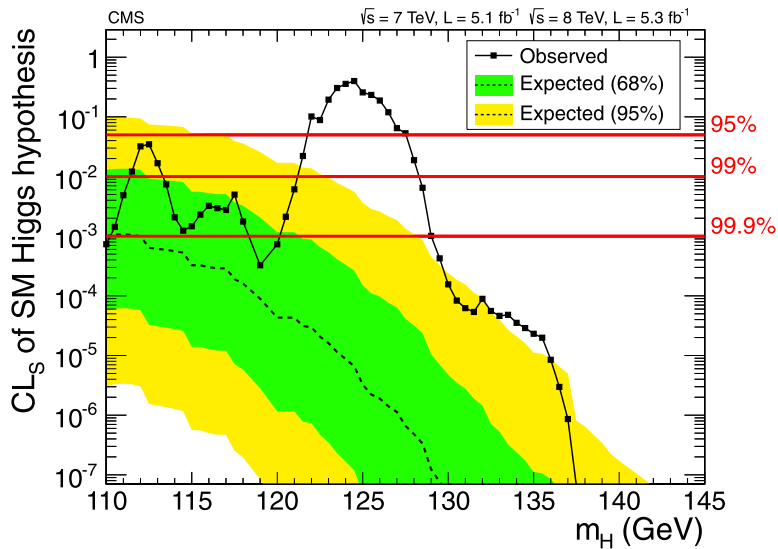
Figure 2.14: The CLs values for the SM Higgs boson hypothesis as a function of the Higgs boson mass in the range 110-145 GeV. The background-only expectations are represented by their median (dashed line) and by the 68 % and 95 % CL bands [14].

(a) gluon-gluon fusion

(b) vector-boson fusion

(c) vector-boson   associated   production
     (Higgsstrahlung)

(d) top-quark associated production
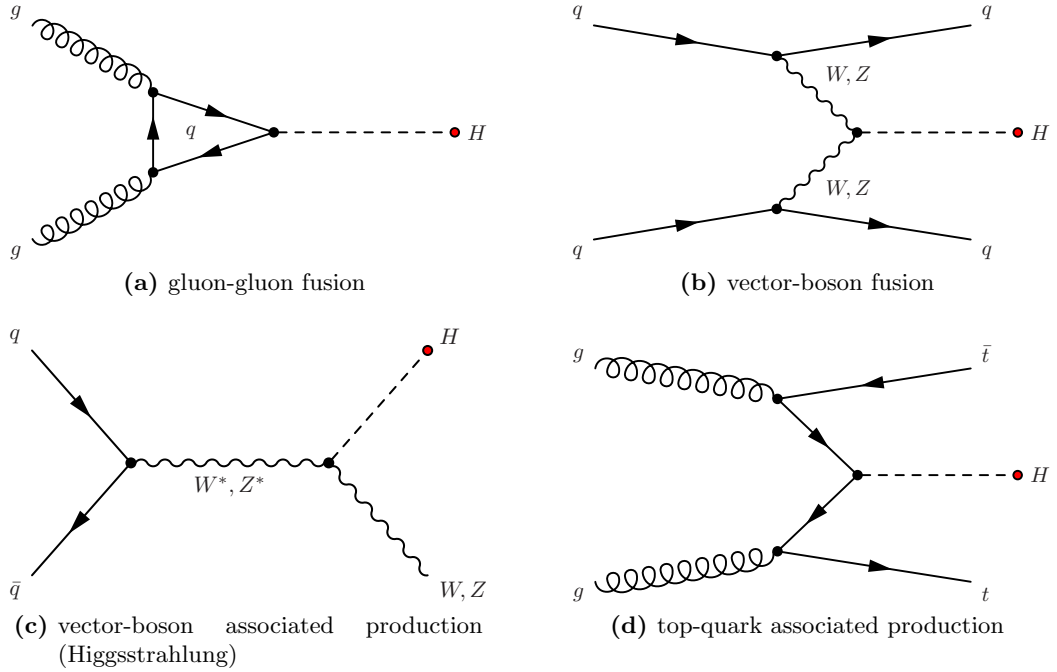
Figure 2.15: Exemplary leading-order Feynman diagrams for different Higgs boson pro-
duction mechanisms.

to its mass, the dominant contribution to this diagram comes from the top-quark
loop.

The second largest contribution to the Higgs-production cross section is the
vector-boson fusion, where the vector bosons (W or Z) are radiated from quarks
inside protons: $pp \rightarrow qq \rightarrow qqV^*V^* \rightarrow qqH$, as shown in Figure 2.15b. For the
observed mass of the Higgs boson, $m_H = 125.7\,\text{GeV}$, the cross section of this process
is one order of magnitude smaller compared to that of the gluon-gluon fusion. The
absence of colour exchange between the incoming quarks leads to a suppression of
additional gluon radiation between the outgoing quarks. Therefore, this process has
the distinctive feature of two high energetic jets from hadronisation of the outgoing
quarks with very little hadronic activity between them (rapidity gap). Processes
with e.g. leptonic final states of the Higgs boson can be well distinguished from
background processes.

Higgs-boson production in association with vector bosons, $pp \rightarrow q\bar{q} \rightarrow V^* \rightarrow$
VH, takes place when a vector boson (W or Z) is produced from q$\bar{\text{q}}$ annihilation
and radiates a Higgs boson, as shown in Figure 2.15c. This production channel is a
good candidate for the measurement of Higgs bosons decaying into bottom quarks,
by the requirement of the vector boson to decay leptonically, making a distinctive
final state.

Production of the Higgs boson in association with top quarks has the lowest
cross section, and is usually initiated by gluons at the LHC, $pp \rightarrow gg \rightarrow t\bar{t}H$, as
shown in Figure 2.15d. The low cross section is compensated by a rather unique
final state that has manageable backgrounds. Since $t\bar{t}H$ production is one of the
primary subjects of this thesis, it is described in more detail in the next section.
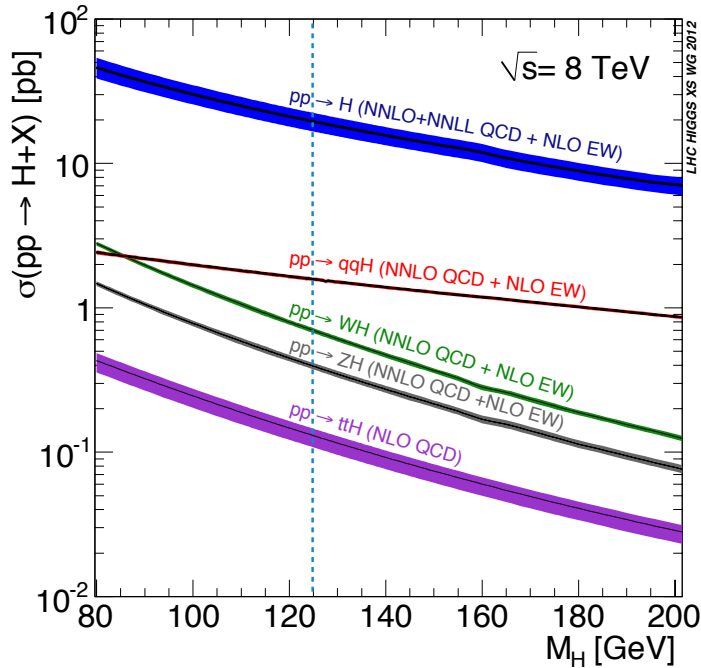
Figure 2.16: Predicted cross-sections for $pp \rightarrow H + X$ processes at the LHC at $\sqrt{s} = 8\,\mathrm{TeV}$ calculated up to NNLO accuracy depending on the process [107]. Coloured bands represent theoretical uncertainties coming from factorisation scale, renormalisation scale and PDF variation.

Decay modes

It is not expected to directly observe the total decay width of the Higgs boson with mass $m_H = 125.7\,\mathrm{GeV}$ at the LHC. The Standard Model predicts the decay width to be approximately $\Gamma(H) \approx 4\,\mathrm{MeV}$ [27], as shown in Figure 2.17, which leads to the mean lifetime $\tau_H \approx 1.56 \times 10^{-22}$ s. Therefore, the Higgs boson immediately decays to fermions or bosons.

The partial decay width to fermions is proportional to the fermion masses, leading to the highest branching ratio for decays to a pair of bottom quarks: $H \rightarrow b\bar{b}$ (the heaviest fermions satisfying the condition: $2m_f < m_H$), as can be seen from Figure 2.18. Due to very high production rate of bottom quarks from standard QCD interactions ($\sigma(bb) \approx 10^7 \times \sigma(H \rightarrow bb)$), it is impossible to obtain a considerable contribution from Higgs-boson decays. Decays to τ leptons ($H \rightarrow \tau\tau$) or a pair of charm quarks ($H \rightarrow c\bar{c}$) also make a sizeable contribution to the total decay width.

Decays to a pair of vector bosons, $H \rightarrow WW$ or $H \rightarrow ZZ$, have sufficient branching fraction for measurements with available data. In particular, $H \rightarrow ZZ \rightarrow l^+l^-l^+l^-$ is one of the most sensitive channels, due to the precise measurement of charged leptons by the tracking detectors and good lepton identification. In fact, this is one of the channels in which the Higgs boson was discovered [14, 13].

Decays to massless bosons (gluons, photons) can not occur at leading order since the Higgs field does not couple to massless particles. Therefore, such decays must involve loops of massive fermions or bosons. Thanks to the excellent precision of
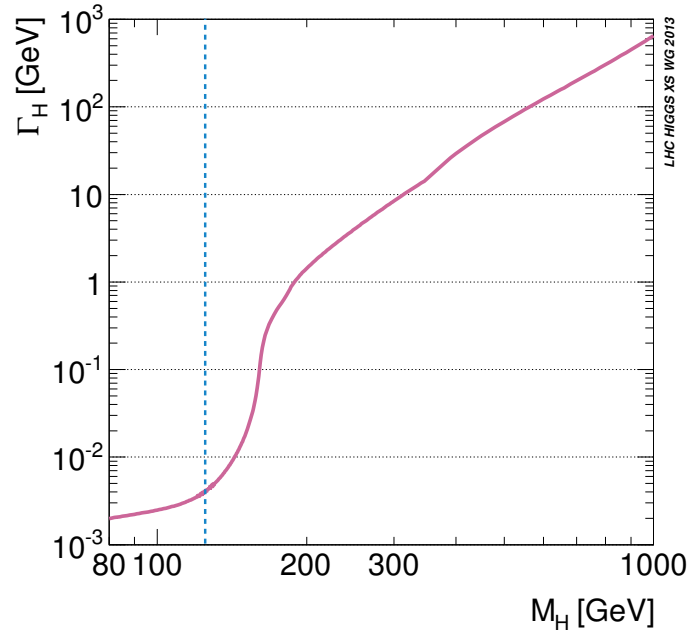
Figure 2.17: Total decay width of the Standard Model Higgs boson, $\Gamma(H)$, as a function of its mass m_H. The observed Higgs-boson mass is marked by the dashed line.
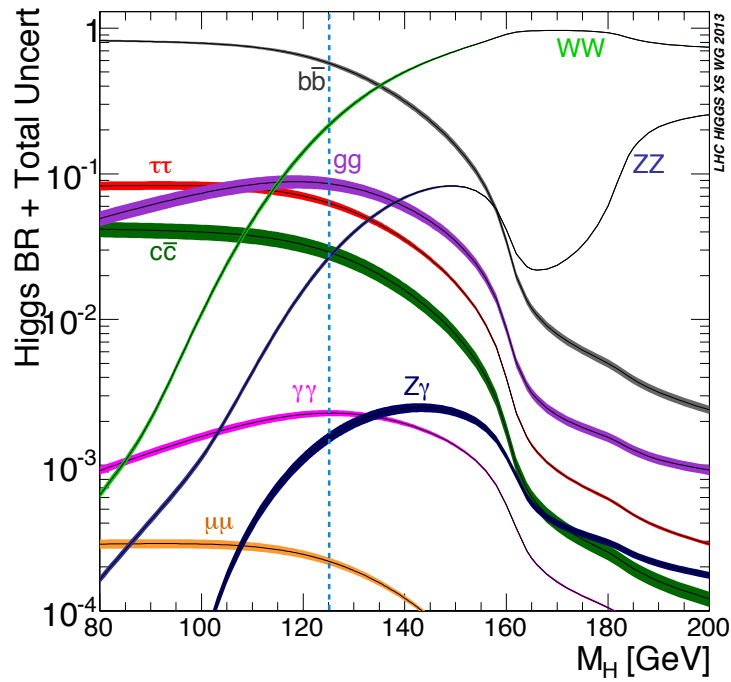
Figure 2.18: Branching fractions (BR) for different Higgs boson decay modes as functions of its mass m_H. The observed Higgs-boson mass is marked by the dashed line.

photon measurements by the electromagnetic calorimeters, the decay $H \to \gamma\gamma$ is also one of the Higgs-boson discovery channels [14, 13].

2.3.3   Associated top-pair and Higgs-boson production

The coupling of the the top quark to the Higgs field (top Yukawa coupling) plays an important role in the electroweak symmetry breaking mechanism. An indirect test of this coupling can be provided by measuring the production rates of the Higgs boson in one of the discovery channels (e.g. $H \to \gamma\gamma, H \to ZZ \to 4l$), produced mainly through the gluon-gluon fusion via a top quark loop (Figure 2.15a). The main direct way of its estimation is through associated top-quark pair and Higgs boson ($t\bar{t}H$) production.

Furthermore, several BSM (beyond the Standard Model) physics scenarios predict the existence of a heavier top quark partner with a sizeable decay rate to a top quark and a Higgs boson [108, 109, 110, 111]. If such a particle exists, it could increase the observed $t\bar{t}H$ production cross section by a factor $\mathcal{O}(2)$ while being still consistent with the current constraints from the measurements by the CMS [112] and ATLAS [113] experiments.

Moreover, this production channel is especially interesting from the perspetive of the higher centre-of-mass energy (13/14 TeV) that will be available during the future running of the LHC. As can be seen from the distributions in Figure 2.19, the $t\bar{t}H$-production cross section grows fastest with the increasing centre-of-mass energy compared to all the other production channels. Therefore, it is important to study the $t\bar{t}H$ process even with the low statistical power available at $\sqrt{s} = 8$ TeV, since this is the only way to be prepared for future data with the increased cross section.

Such a complex process can be separated into different decay channels according to the decay of the Higgs boson, as described in Section 2.3.2, and according to the final state of the $t\bar{t}$ system, as described in Section 2.3.1. Since in the analysis described in Chapter 10 a dileptonic final state of the $t\bar{t}$ system is considered, together with the Higgs boson decaying to a b$\bar{\text{b}}$ pair, exactly this decay channel is introduced in more detail.

$t\bar{t}H(b\bar{b})$ in the dileptonic decay channel

As already mentioned, $H \to b\bar{b}$ is the dominant decay channel of the Standard Model Higgs boson and has a branching ratio:

$$\text{BR}(H \to b\bar{b}) = \frac{\Gamma(H \to b\bar{b})}{\Gamma(H)} = 56.6\% \pm 1.9\% \tag{2.41}$$

for the mass of the Higgs boson $m_H = 125.7$ GeV according to [114]. It was also stated that observation of the Higgs boson alone, decaying to a pair of bottom quarks, is nearly impossible due to an overwhelming QCD background, whose production rate exceeds that of the $H \to b\bar{b}$ signal by seven orders of magnitude. Nevertheless, if such a Higgs boson is produced in association with a pair of top quarks, the final state is rather unique, leading to a greatly reduced production rate for background processes.
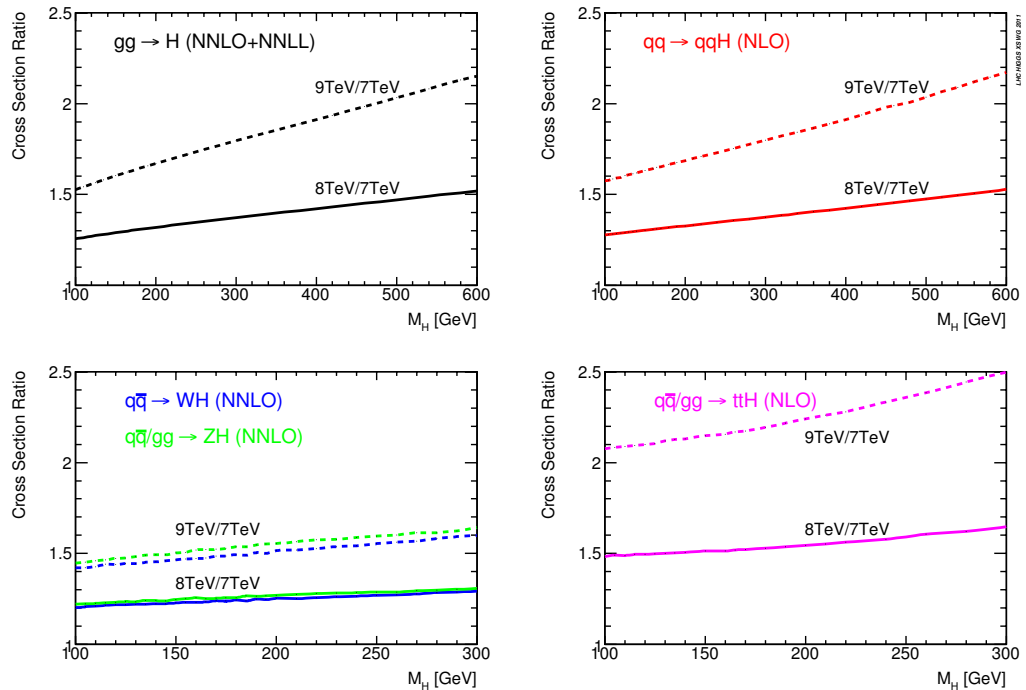
Figure 2.19: Ratios of the Higgs boson production cross-sections at $\sqrt{s} = 8/9$ TeV with respect to $\sqrt{s} = 7$ TeV at the LHC for the main four Higgs production channels [107].

A basic Feynman diagram depicting the final state of interest is shown in Figure 2.20a. One of the most important features of such a final state is the presence of two bottom quarks from the Higgs boson decay and two bottom quarks from the decays of the two top quarks. Each of these b quarks form B hadrons, which have a lifetime $\tau_B = \mathcal{O}(10^{-12}s)$, and undergo fragmentation, leading to the formation of hadronic jets initiated by bottom quarks: b jets. Therefore, the final state shown in Figure 2.20a contains four b jets, which are formed by colourless stable particles (hadrons and/or leptons).

2.3.4 Associated top-quark-pair and b-jets production

The $t\bar{t}b\bar{b}$ process is the dominant and almost irreducible background to $t\bar{t}H(b\bar{b})$ production, since it has exactly the same final state as the $t\bar{t}H(b\bar{b})$ process. The corresponding leading order Feynman diagram of the $t\bar{t}b\bar{b}$ production in the dileptonic decay channel of the $t\bar{t}$ system is shown in Figure 2.20c.

Therefore, it is especially important to properly estimate its contribution. Furthermore, since a limit on the $t\bar{t}H$ cross section is planned to be extracted from the dijet mass distribution in the region around the Higgs boson mass $100 < m_{b\bar{b}} < 140$ GeV, its shape directly affects the measurement. Hence, in order to achieve a reliable measurement, the shape of the $m_{b\bar{b}}$ spectrum has to be properly simulated and checked by measurement.

The possible effect from the $t\bar{t}b\bar{b}$ background modeling is illustrated in Figure 2.21. The observation of a $H \rightarrow b\bar{b}$ mass peak is the main reason behind a

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

(b) $t\bar{t}Z(b\bar{b})$ process
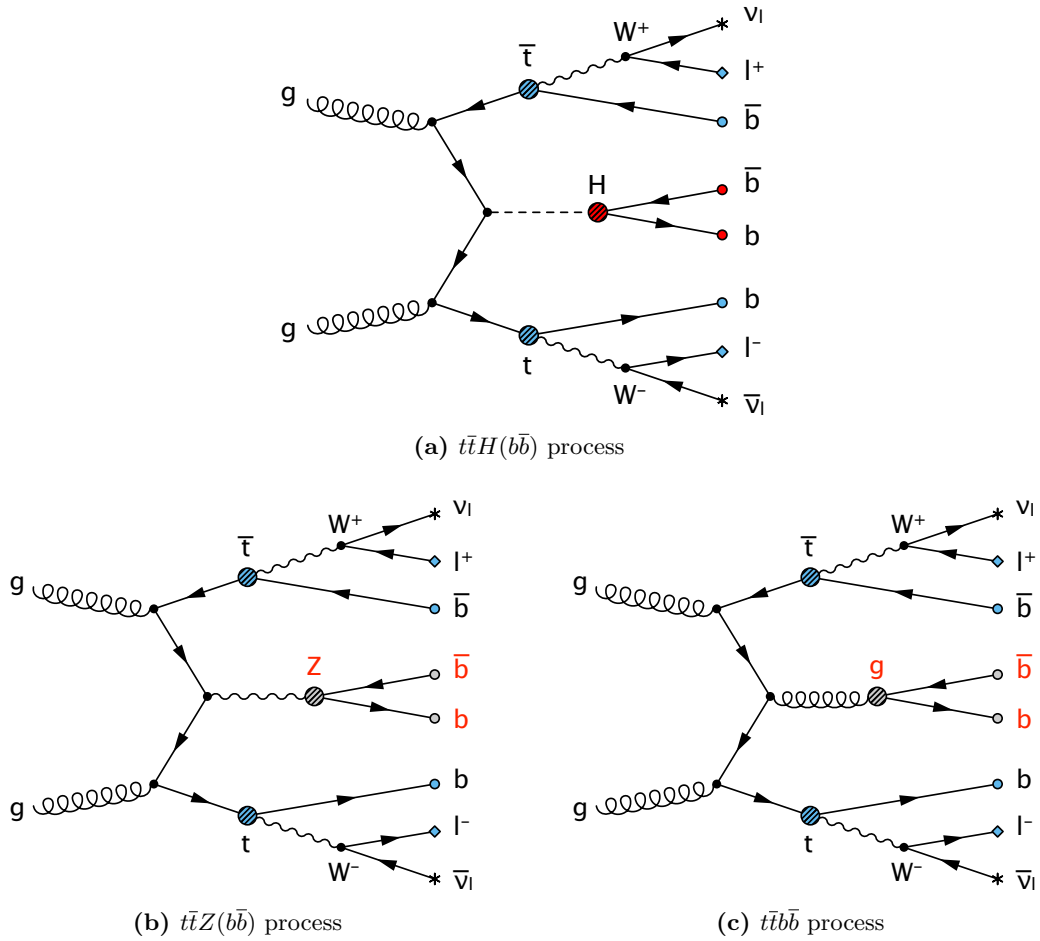
(c) $t\bar{t}b\bar{b}$ process

Figure 2.20: Feynman graphs of the $t\bar{t}H(b\bar{b})$, $t\bar{t}Z(b\bar{b})$ and $t\bar{t}b\bar{b}$ processes with the dileptonic final state of the $t\bar{t}$ system. Each final state particle is denoted by a marker. Bottom quarks, marked with empty circles, undergo fragmentation and are detected as b jets. Neutrinos, marked with asterisks, are not directly detectable. All the three processes produce identical final states at leading order.

dedicated measurement of the differential cross section of $t\bar{t}b\bar{b}$ production as a function of additional b jet properties, including $m_{b\bar{b}}$.
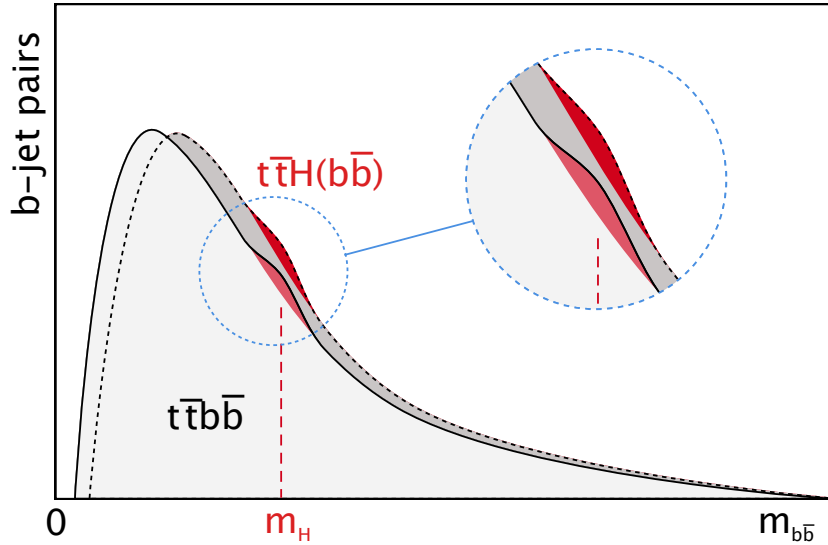
Figure 2.21: Illustration of shapes of the dijet-mass spectrum from $t\bar{t}H(b\bar{b})$ signal and $t\bar{t}b\bar{b}$ background. The two gray filled areas represent two potentially different modelings of the $t\bar{t}b\bar{b}$ shape, which leads to the change in observed excess from $t\bar{t}H$, represented by the red filled area.

The position of the peak in the $t\bar{t}b\bar{b}$ background distribution directly depends on the lower p_T threshold for the additional b jets. Therefore, the threshold has to be chosen as low as possible in order to move the peak of the background shape to lower mass values, and to reduce its contribution in the region around the Higgs-boson mass (m_H).

2.3.5   Associated top-pair and $Z \rightarrow b\bar{b}$ production

Another very important background to the $t\bar{t}H(b\bar{b})$ production is the $t\bar{t}Z$ process, where the Z boson decays in to a $b\bar{b}$ pair. This process has exactly the same final state as $t\bar{t}H(b\bar{b})$, as shown in Figure 2.20b. Its cross section combined with the $Z \rightarrow b\bar{b}$ branch fraction leads to an about 2 to 3 times smaller cross section than that of $t\bar{t}H(b\bar{b})$ production.

Although it is a smaller background contribution than the $t\bar{t}b\bar{b}$ process, it imposes a problem of a different kind. While the $m_{b\bar{b}}$ spectrum of the $t\bar{t}b\bar{b}$ process is exponentially falling, the $t\bar{t}Z(b\bar{b})$ process would be visible as a resonant peak around the Z-boson mass of 91.2 GeV, which is very close to the Higgs-boson mass. Unless the measured-mass resolution becomes better than the difference in masses of the Higgs and Z bosons, the $t\bar{t}Z(b\bar{b})$ process will enter the $t\bar{t}H(b\bar{b})$ signal region. Thus, the rate of the $t\bar{t}Z(b\bar{b})$ process, as well as the shape of its $m_{b\bar{b}}$ spectrum, have to be known precisely.

Chapter 3

Existing measurements

Several measurements of both $t\bar{t}H$ and $t\bar{t}b\bar{b}$ production have been performed using data recorded at the LHC by the ATLAS and CMS experiments. The results of these measurements serve as a cross-check for the analyses described in this thesis and have to be consistent within uncertainties, taking into account potential differences between the analyses. An overview of the main results of the existing measurements is given in the following sections.

3.1 Associated top-pair and Higgs-boson production

After the discovery of the Higgs boson, the measurement of its properties has become one of the primary analysis topics among the two experiments capable of such measurements, ATLAS and CMS. $t\bar{t}H$ production is not an exception, and despite the small number of such events expected in the data recorded at the LHC, several analyses on this topic have been performed by the two collaborations.

There are two measurements [115, 116] from the ATLAS collaboration in which searches for $t\bar{t}H$ production have been performed in two different decay channels of the Higgs boson:

- **$H \rightarrow \gamma\gamma$:** [115] all data recorded at the LHC has been analysed ($\sqrt{s} = 7/8$ TeV), searching for $t\bar{t}H$ with any final state of the $t\bar{t}$ system: dileptonic, semileptonic and full hadronic, with optimised selection for each channel. In order to separate $H \rightarrow \gamma\gamma$ decays from the background process, a localised excess of events is searched for in the $m_{\gamma\gamma}$ spectrum within the range $120 < m_{\gamma\gamma} < 130$ GeV. In total 5 candidate events have been found with the SM expectation of $4.6^{+1.3}_{-0.9}$ background events and 1.3 signal events. The measured events are consistent with the SM expectation of $t\bar{t}H$ signal, although no significant excess over the predicted background-only hypothesis has been observed, leading to an upper limit on the $t\bar{t}H$ production cross section of 6.5 times the SM cross section times BR($H \rightarrow \gamma\gamma$) at 95 % confidence level (CL). The measured limit is shown for the leptonic and full-hadronic channels individually and combined in Figure 3.1a.

- **$H \rightarrow b\bar{b}$:** [116] all data recorded at $\sqrt{s} = 8$ TeV has been analysed, searching for $t\bar{t}H$ with dileptonic or semileptonic final state of the $t\bar{t}$ system. To

achieve high sensitivity of the measurement, events are categorised in terms of jet and b jet multiplicities and a multivariate technique (MVA) has been used in each individual category by combining multiple quantities that discriminate between the $t\bar{t}H$ signal and background processes. By combining all categories and decay channels of the $t\bar{t}$ system, a combined fit of the MVA output distributions had been performed, and no significant excess of events above the background expectation was found. An observed upper limit on the $t\bar{t}H(b\bar{b})$ cross section of 4.1 times the Standard Model cross section has been obtained at 95 % CL, which is a better constraint than in the $H \to \gamma\gamma$ measurement. The measured limit is shown for dileptonic and semileptonic decay channels of the $t\bar{t}$ system individually and combined in Figure 3.1b.
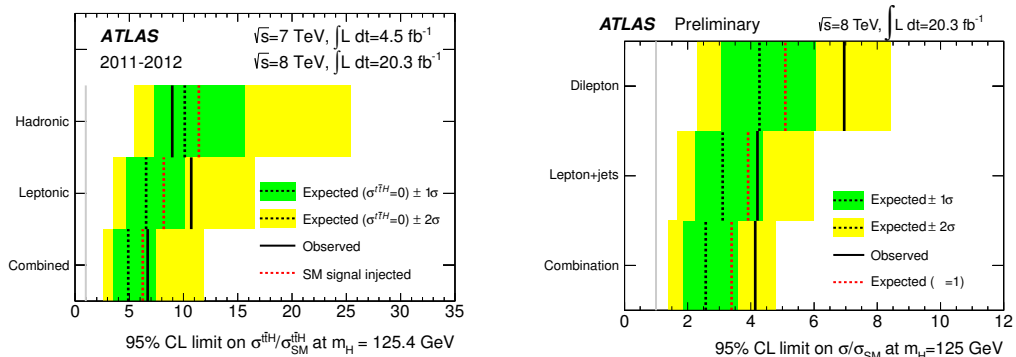
(a) Observed and expected 95 % CL upper limits on the $t\bar{t}H$ production cross section in the $H \to \gamma\gamma$ decay channel. The limits are given relative to the SM expectations at $m_H = 125.4\,\mathrm{GeV}$ [115].

(b) 95 % CL limit on $t\bar{t}H$ production cross section in the $H \to b\bar{b}$ decay channel relative to the SM prediction, σ/σ_{SM} for the individual channels as well as their combination, assuming $m_H = 125\,\mathrm{GeV}$ [116].

Figure 3.1: Results of searches for $t\bar{t}H$ production by the ATLAS collaboration in the $H \to \gamma\gamma$ (a) and $H \to b\bar{b}$ (b) decay channels.

A similar set of analyses [117] has been performed by the CMS collaboration, searching for $t\bar{t}H$ production with a wider set of possible decay channels of the Higgs boson and using the same amount of data recorded at $\sqrt{s} = 7/8\,\mathrm{TeV}$. All the channels have been combined into a single measurement and can be divided into three basic categories:

- **$H \to hadrons$:** includes searches for $t\bar{t}H$ production with $H \to b\bar{b}$ and $H \to \tau\tau$ final states, where in the $H \to \tau\tau$ channel both τ leptons decay hadronically and can stem either from the direct Higgs boson decays or $H \to WW$ or $H \to ZZ$ followed by the decays of the W and Z bosons into τ leptons. The analysed events are required to have at least one isolated lepton from W boson decays from the top quarks, which means that semileptonic and dileptonic final states of the $t\bar{t}$ system are searched for. Multivariate analysis (MVA) techniques are employed to tag the jets coming from b quark or τ lepton decays and to separate $t\bar{t}H$ events from the large $t\bar{t} + jets$ backgrounds.

- **$H \to photons$:** the search is focused exclusively on the $H \to \gamma\gamma$ decays and all possible final states of the $t\bar{t}$ system are considered. Similarly to the

ATLAS measurement [115], the invariant mass of the two photons is used to separate the $t\bar{t}H$ signal from background processes.

- **$H \rightarrow leptons$:** the leptons arise as secondary decay products from $H \rightarrow WW$, $H \rightarrow ZZ$ and $H \rightarrow \tau\tau$ decays. To optimise the signal-to-background ratio, events are required to have either a pair of same-sign charged leptons, or three or more charged leptons. MVA techniques are used to separate signal leptons (arising from W-boson, Z-boson and τ-lepton decays) from background leptons (coming from b-quark or c-quark decays, or jets misidentified as leptons) as well as to separate $t\bar{t}H$ events from background processes.

An upper limit on the $t\bar{t}H$ cross section of 4.5 times the Standard Model prediction at 95 % CL has been observed from a combination of all the different Higgs boson decay channels, as shown in Figure 3.2b. In addition, the dependence of the limit on the Higgs boson mass has been verified in the range $110 < m_H < 140$ GeV and appears not be significant according to Figure 3.2a [117]. In the $H \rightarrow b\bar{b}$ decay channel, a similar limit of 4.1 times the SM has been observed as in the measurement from the ATLAS experiment.
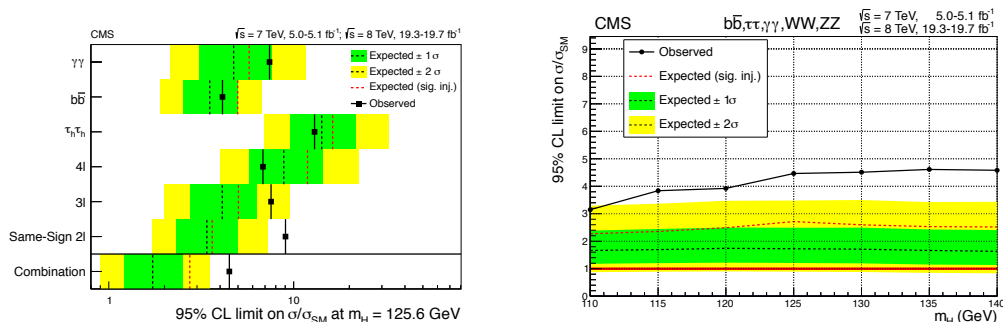
(a) Observed and expected 95 % CL upper limits on the $t\bar{t}H$ production cross section with respect to the Standard Model cross section in all the analysed Higgs boson decay channels individually and combined. The limits are given relative to the SM expectations at $m_H = 125.6$ GeV [117].

(b) The observed and expected limits as a function of the Higgs boson mass for all the channels combined [117].

Figure 3.2: Results of searches for $t\bar{t}H$ production by the ATLAS collaboration in $H \rightarrow \gamma\gamma$ (a) and $H \rightarrow b\bar{b}$ (b) decay channels.

The main feature of the existing $t\bar{t}H(b\bar{b})$ searches is that an output distribution from a complicated MVA tool is used for the discrimination between signal and background processes, which is a non-trivial combination of multiple quantities. This is in contrast to the approach harnessed in the analysis presented in Chapter 10 of this thesis, which aims for a separation between $t\bar{t}H$ and background processes based on a better-understood and more physically meaningful quantity, the invariant mass of the two b jets not stemming from top-quark decays. A comparison of the results from the existing measurement with results obtained by the current analysis is given in Section 10.5.

3.2 Associated top-pair and b jets production

A rough estimate of the heavy flavour content of jets produced in addition to a $t\bar{t}$ pair has been performed by the ATLAS collaboration [118] by measuring the $\sigma_{t\bar{t}+HF}/\sigma_{t\bar{t}+jets}$ ratio, where $t\bar{t}+HF$ denotes events with a $t\bar{t}$ pair accompanied by at least one heavy-flavour jet (bottom or charm) and $t\bar{t}+jets$ represents events with a $t\bar{t}$ pair accompanied by any jet. For such cross sections to have a well-defined value, a lower boundary on the allowed transverse momentum (p_{T}) of the jets is applied: $p_{\mathrm{T}} > 25\,\mathrm{GeV}$. Only data recorded at $\sqrt{s} = 7\,\mathrm{TeV}$ had been used for the measurement and the fraction of $t\bar{t}+HF$ cross section was estimated by a template fit of the vertex mass distribution of the additional jet, resulting in a value of $\sigma_{t\bar{t}+HF}/\sigma_{t\bar{t}+jets} = 6.2\pm2.1\%$. The obtained result was found to be consistent with LO prediction by both ALPGEN +HERWIG and POWHEG+HERWIG simulations [118].

A more specific analysis has been performed by the CMS collaboration, measuring the $\sigma_{t\bar{t}b\bar{b}}/\sigma_{t\bar{t}jj}$ cross section ratio using a larger set of data recorded at $\sqrt{s} = 8\,\mathrm{TeV}$ [119]. In the ratio, $t\bar{t}b\bar{b}$ denotes a $t\bar{t}$ pair produced in association with at least two b jets, while $t\bar{t}jj$ denotes a $t\bar{t}$ pair accompanied by at least two jets, regardless of their flavour. The ratio has been measured for two thresholds on the additional jet p_{T}: $p_{\mathrm{T}} > 40\,\mathrm{GeV}$ and $p_{\mathrm{T}} > 20\,\mathrm{GeV}$. Unlike the measurement by ATLAS, events with additional c jets are in the denominator of the ratio. The fraction of $t\bar{t}b\bar{b}$ events has been estimated from a template fit of the b-tag-discriminant distribution of the third and fourth additional jets, which is used to experimentally distinguish between b jets and jets of other flavours. The used distributions are shown in Figure 3.3 and allowed to measure the ratio with a precision of $\mathcal{O}(27\%)$: $\sigma_{t\bar{t}b\bar{b}}/\sigma_{t\bar{t}jj} = 2.2 \pm 0.6\%$. The ratio does not change with different values of the jet p_{T}-threshold, while the total cross sections of $t\bar{t}b\bar{b}$ and $t\bar{t}jj$ production individually do change significantly. The measured cross section for $p_{\mathrm{T}} > 40\,\mathrm{GeV}$ was compared to an NLO calculation, which yielded a lower cross section, but was compatible within 1.6 standard deviations [119].

The analysis of this thesis, presented in Chapter 9, goes one step further, measuring $\sigma_{t\bar{t}b\bar{b}}$ differentially as a function of additional b jet properties, using the jet-$p_{\mathrm{T}} > 20\,\mathrm{GeV}$ threshold. Since the integral of the differential cross section over the corresponding phase space should be equal to the inclusive cross section, a comparison of the results presented in this thesis to the total cross section from the existing measurement by CMS is an important cross-check. A detailed comparison and interpretation of differences between the two analyses is given in the corresponding Section 9.10.
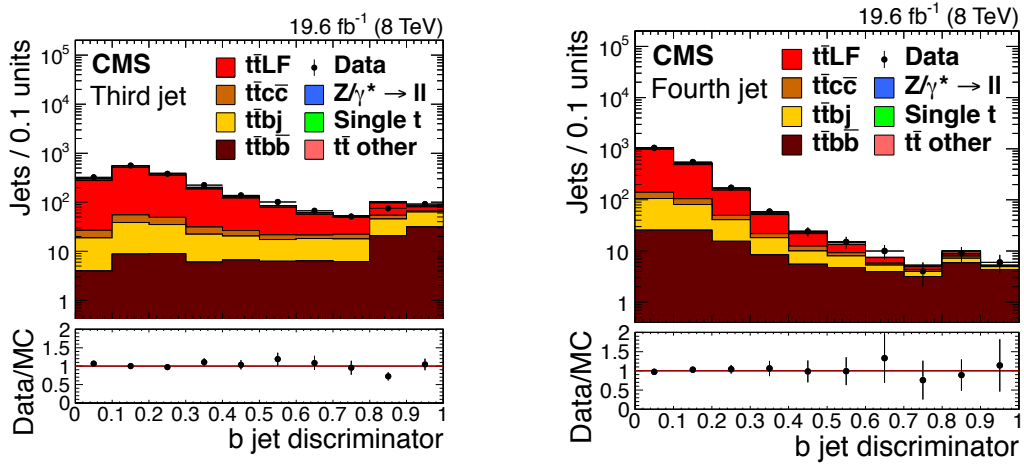
Figure 3.3: Distributions of *b* jet discriminator for the third (left) and fourth (right) jet in decreasing order of b-tagging discriminator value, after the full event selection. Points represent recorded data, while stacked histograms represent simulation of signal and background processes. The ratio of the number of data events to the total number of events in simulation after the template fit is shown in the lower panels [119].

Chapter 4

The CMS experiment at the LHC

As already mentioned in the introduction, the Standard Model has been extremely successful in describing numerous phenomena that have been experimentally observed during the last century. The majority of such observations were made at particle accelerators, with their collision energy gradually increasing over time, allowing to test the theory at higher and higher energy scales. Before the so-called "LHC era", the previous most powerful particle accelerator was the Tevatron, located at the Fermi National Laboratory in the USA. Its major achievement is the discovery of the top quark in 1995 by the CDF [97] and DØ [96] collaborations, which made the Higgs boson the last non-observed particle of the Standard Model.

The absence of positive results from searches at the Tevatron proton-antiproton collider and at the LEP electron-positron collider implied that if it were to exist, its mass should be higher than 114.4 GeV [120]. Therefore, a more powerful accelerator, capable of producing collisions with energies higher than that provided by the Tevatron, had to be built. Such a high-intensity accelerator has been constructed, and is called the Large Hadron Collider (LHC) [10].

Both the technical and physical contributions described in this thesis were carried out with the detector of the Compact Muon Solenoid (CMS) experiment [11], one of the experiments built to analyse hadron collisions taking place at the LHC. In the following sections a basic overview of the LHC is given, as well as a brief description of the main features of the CMS detector and of its main components. The last section is devoted specifically to the CMS tracking detector, which is the subject of the technical contribution of this thesis.

4.1 The LHC accelerator

The LHC is the world's largest and most powerful man-made particle accelerator, which was built between 1998–2008. It is located at the border between Switzerland and France, and is operated by the European Organisation for Nuclear Research (CERN, from French *"Conseil Européenne pour la Recherche Nucléaire"*).

Being a proton-proton collider, it accelerates two counter-rotating beams of protons with separate magnetic fields and vacuum chambers, with common sections

only at the insertion regions where the experimental detectors are located. The accelerating rings are placed underground inside the 26.7 km long tunnel from LEP, the world's highest energy electron-positron collider, which was closed in 2000. The tunnel was excavated at a depth varying from 175 m (under the Jura mountain) to 50 m (towards Lake Geneva), and has a slight slope of 1.4 %, in order to minimise the length of the vertical shaft tunnel that lies under the Jura. The geographical map of the LHC location together with the seven main experiments served by it are shown in Figure 4.1.

Figure 4.1: Hydro-geologic map of the LHC area with four active interaction points, where the particle beams cross, marked with names of the experiments using it. Some interaction points are shared by more than one experiment. Based on the image by Serge Dailler [121].

Each ring can accommodate up to 2808 bunches with 1.15×10^{11} protons in each, and with a time spacing of 24.95 ns. The proton beams cross in four interaction points, around which the detectors of the corresponding experiments are located. The four large-scale experiments are CMS, ATLAS, LHCb and ALICE, while LHCf, TOTEM and MoEDAL are much smaller experiments with a narrower range of research topics. The design centre-of-mass energy of the proton-proton collisions at the LHC is $\sqrt{s} = 14$ TeV. Besides protons, LHC was designed to also be capable of accelerating fully stripped lead ions ($^{286}\mathrm{Pb}^{82+}$) to an energy of 2.76 TeV/nucleon, leading to ion-ion collisions at the nucleon-nucleon centre-of-mass energy $\sqrt{s} = 5.52$ TeV.

The primary subject of measurements at particle colliders is a cross section (σ) of a particular process, which is deduced from the number of observed events (N_{ev}) of this process and the integral of the instantaneous luminosity ($\mathcal{L}$) over the time of event recording:

$$N_{ev} = \sigma \times \int \mathcal{L}(t)dt \,. \tag{4.1}$$

At the two largest experiments, ATLAS and CMS, the designed peak instantaneous luminosity is $\mathcal{L} = 10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ [10].

History of operation

The LHC started its operation on September 10-th, 2008, when the first proton beam has been successfully accelerated. Unfortunately, nine days later, on September 19-th, a faulty electrical connection between two magnets caused a number of magnet quenches (phase transition from the super-conducting to a normal-conducting state) that led to damaging the integrity of the liquid helium's enclosure followed by its explosion. As a result, 53 superconducting magnets were damaged and had to be repaired, delaying the operation of the LHC by more than a year. On November 20-th, 2009 proton beams were successfully circulated again at the energy of $450\,\mathrm{GeV}$ per beam.

Since March of 2010 till February of 2013 the LHC has been running without any problems. During 2010–2011 proton beams collided at $\sqrt{s} = 7\,\mathrm{TeV}$ with a peak instantaneous luminosity of $3.6 \times 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ delivering to the ATLAS and CMS experiments an integrated luminosity of about $6.2\,\mathrm{fb}^{-1}$.

During 2012 the energy of proton the beams was increased to $4\,\mathrm{GeV}$ leading to a centre-of-mass collision energy of $\sqrt{s} = 8\,\mathrm{TeV}$. The peak instantaneous luminosity during this run has increased to $7.7 \times 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-1}$ and the delivered integrated luminosity over 2012 was $23.3\,\mathrm{fb}^{-1}$.

Several weeks in 2010 and 2011 were dedicated to special heavy-ion runs, during which collisions of two beams of lead ions were taking place. At the beginning of 2013 LHC operated asymmetric collisions constituted by a proton beam with energy of $4\,\mathrm{TeV}$ and a lead-ion beam with energy of $1.58\,\mathrm{TeV}$ per nucleon, corresponding to the nucleon-nucleon $\sqrt{s} = 5.02\,\mathrm{TeV}$ [122].

In March 2013 the LHC was shut down for almost two years for the upgrade of the machine to the increased beam energy of $6.5\,\mathrm{TeV}$, which would allow to achieve a centre-of-mass energy of $\sqrt{s} = 13\,\mathrm{TeV}$ and later the designed $\sqrt{s} = 14\,\mathrm{TeV}$.

Technical layout

The protons at the LHC do not achieve their final energy in the main ring alone. There is a series of smaller accelerators, which accelerate the proton bunches to increasingly higher energies in several stages and are used by standalone smaller experiments. The basic layout of the complete CERN accelerator complex is sketched in Figure 4.2.

The complicated process of proton beam acceleration starts from a bottle of hydrogen gas, which is exposed to an electric field for stripping the atoms of their electrons and yielding bare protons. These protons are directed into the first accelerator in the chain, LINAC 2, which accelerates them to an energy of $50\,\mathrm{MeV}$. The beam is then injected into the Proton Synchrotron Booster (PSB), which accelerates the protons to $1.4\,\mathrm{GeV}$, followed by the Proton Synchrotron (PS), which pushes the beam to $25\,\mathrm{GeV}$. Protons are then sent to the Super Proton Synchrotron (SPS) where they are accelerated to $450\,\mathrm{GeV}$.
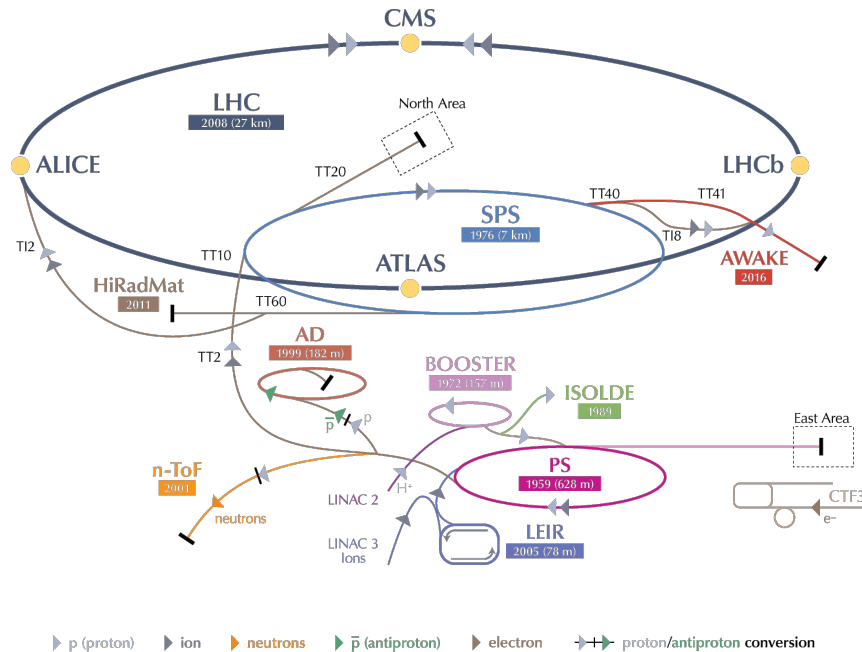
Figure 4.2: Sketch of the CERN accelerator complex layout [123]. The first number under each accelerator's name represents the proton beam energy at the extraction point from the accelerator. The number in brackets represents the year in which the accelerator started to operate.

The protons are finally transferred to the two beam pipes of the LHC. The beam in one pipe circulates clockwise while the beam in the other pipe circulates anticlockwise. Eight radiofrequency (RF) superconducting cavities are installed in long straight sections for each beam to accelerate the protons at a frequency of 400 MHz and keep the proton bunches tightly bunched to maximize the number of collisions. In order to retain the proton bunches in the circular orbits of the beam pipes, strong superconducting dipole magnets are installed in the arcs of the tunnel operating at 1.9 K, producing a magnetic field of 8.33 T [124]. It takes 4 minutes and 20 seconds to fill each LHC ring, and 20 minutes for the protons to reach their maximum energy of 4 TeV.

The accelerator complex also includes the Antiproton Decelerator and the Online Isotope Mass Separator (ISOLDE) facility, and feeds the CERN Neutrinos to Gran Sasso (CNGS) project and the Compact Linear Collider test area, as well as the neutron time-of-flight facility (nTOF) [125].

Lead-ion collisions also take place at the LHC. They start from a source of vaporised lead and enter Linac 3 before being collected and accelerated in the Low Energy Ion Ring (LEIR). They then follow the same route to maximum energy as the protons [125].

A really peculiar fact about the LHC accelerator is that due to its large size and an exceptional beam positioning precision, the accelerator is sensitive to the Moon orbiting around the Earth. Like water tides, which everyone is familiar with, analogous "ground tides" are caused by the gravitational field of the Moon raising the ground crust by around 25 cm in the Geneva area. These movements lead to a variation of the 26.7 km LHC circumference of the order of 1 mm [124]. Since the

orbits of accelerated proton beams have to be perfectly aligned at the interaction points and inside the focusing quadrupole magnets, the trajectories of the proton beams are regularly readjusted to optimise the delivered instantaneous luminosity. The predicted variations in beam energy due to the tides is shown in Figure 4.3 and are used by the LHC operators to compensate the accelerator displacement [126].
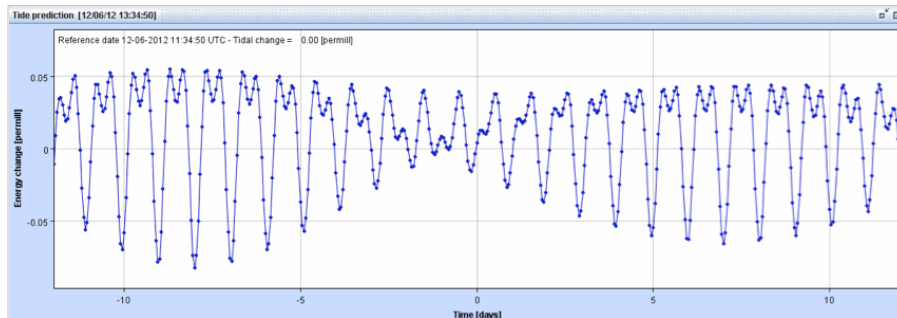

Figure 4.3: Prediction of beam energy variations due to the Moon tides as used by the LHC operators for the beam orbit correction. Each double-peak represents a single day-night, while the external modulation is caused by the interference between the Moon and the Sun gravitation, due to the relative movement of the Sun with respect to the Moon during a month [126].

Furthermore, the LHC is also affected by the hydrostatic pressure created by the water level in Lake Geneva, which deforms the tunnel shape as well. And even more surprisingly, it took months of observations and a train company strike to discover that some particular periodic, every-day, and fixed-time perturbations were caused by the passing of the fast TGV train linking Geneva to Paris, since it releases a lot of electrical energy into the ground [126].

4.2   The CMS experiment

The Compact Muon Solenoid (CMS) experiment is one of the two multi-purpose experiments operating at the LHC. The two experiments have very similar physics programs and complement the measurements of each other. For a discovery made by one experiment to be trustworthy, it has to be confirmed by a corresponding measurement from the other experiment. In particular, the synchronous observation of the Higgs boson resonance by the ATLAS and CMS experiments left no doubt about the existence of the observed particle.

The CMS detector is installed underground at the interaction point near the French village of Cessy at a depth of about 100 m. Its location with respect to the accelerator complex and other experiments can be seen in Figures 4.1 and 4.2. The beam energy and the design luminosity of the LHC accelerator have been chosen to allow studies of physics at the TeV energy scale. Therefore, these conditions require a very careful detector design that would satisfy a number of crucial criteria, which, together with building cost considerations, shaped the final design of the detector and its components. The basic criteria that had to be satisfied by the CMS detector are the following:

- **High granularity with good time resolution** to provide low occupancy under conditions of multiple proton-proton collisions (pile-up) taking place

during a single bunch crossing and to minimise confusion of particles from different pile-up interactions;

- **Hard-radiation sustainable detectors and front-end electronics** to withstand high radiation level caused by a large flux of particles coming from the interaction region;

- **Good charged-particle momentum resolution** to provide efficient identification of b jets and τ leptons;

- **Good muon identification and momentum resolution** for possible discoveries of processes that involve muons in final states;

- **High electromagnetic energy resolution** to be capable of good diphoton and dielectron mass resolution ($\approx 1\,\%$ at $m_{ee/\gamma\gamma} = 100\,\mathrm{GeV}$) as well as efficient photon and lepton isolation at high luminosities;

- **Fine lateral segmentation of hadronic calorimeters** in order to reach high missing-transverse-energy ($\not{E}_{\mathrm{T}}$) and dijet-mass (m_{jj}) resolution.

The CMS detector weighs about $14\,000$ tonnes and has diameter of $15\,\mathrm{m}$. This enormous size is needed to fit inside a sophisticated measurement apparatus with complicated internal structure that is organised into several basic components, which are schematically shown in Figure 4.4. Nevertheless, the name of the experiment says that the detector is "compact" and, indeed this makes sense, when comparing it to the other general-purpose detector, the ATLAS experiment, which has a diameter of $25\,\mathrm{m}$ [12].

Like most of its predecessors, the CMS detector has an onion-like structure, in which different subdetectors are arranged in layers around the interaction point, where the proton beams intersect. As typical for this kind of detectors, the layer closest to the interaction point is the all-silicon tracking detector, followed by the homogeneous crystal-scintillator electromagnetic calorimeter (ECAL), and the brass-scintillator hadronic calorimeter (HCAL). These subdetectors are arranged inside a superconducting solenoid, which generates a very strong magnetic field of $3.8\,\mathrm{T}$ and is a distinctive feature of the CMS detector that allows to reach the designed charged particle momentum resolution within such a compact size. The magnet is surrounded by a steel return yoke that is used to absorb the energy of muons, which is measured by gas-ionising chambers installed between the iron plates.

An overview of each of the components as well as a description of the CMS coordinate system are given in the following sections.

4.2.1 Coordinate system

In the right-handed coordinate system used across CMS, the origin is located at the interaction point in the center of the detector. The X axis points towards the center of the LHC accelerator ring, the Y axis points up perpendicularly to the LHC plane, and the z axis points along the counterclockwise-beam direction. Considering this arrangement of axes, the azimuthal angle ϕ is defined as the angle in the $X-Y$ plane measured from the X axis, and the radial coordinate in this plane is denoted
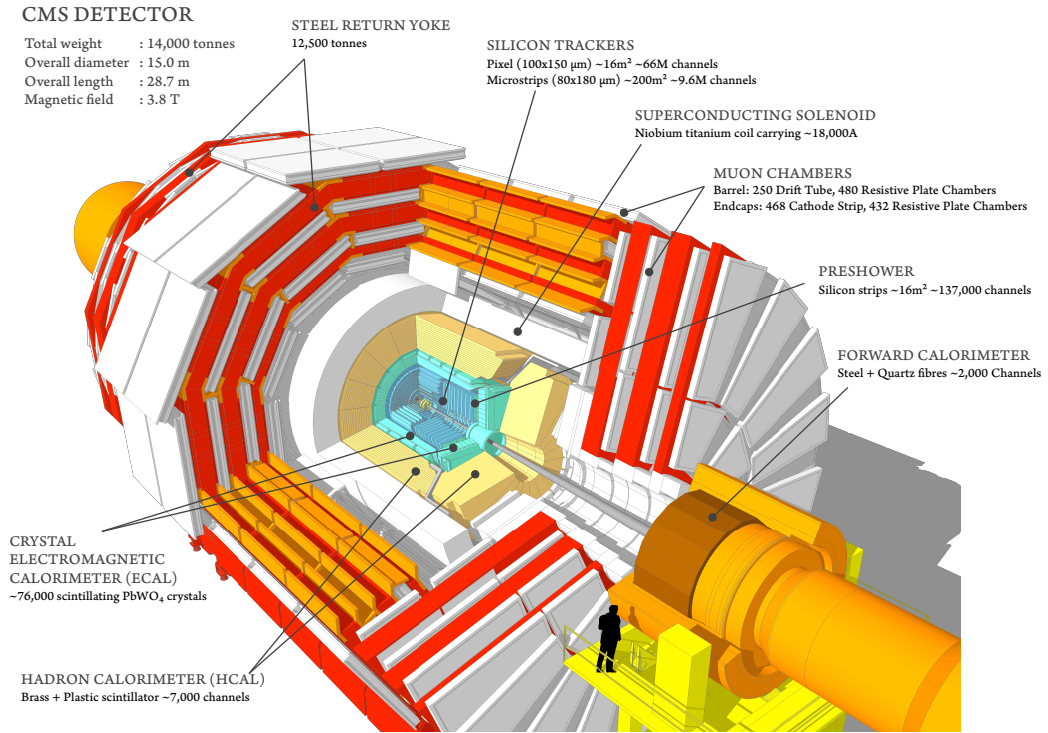
Figure 4.4: Schematic view of the CMS detector. Main components of the detector are marked with basic information about them. Image by Tai Sakuma [127].

by r. Therefore, the $X - Y$ plane is usually referred to as $r\phi$ plane. The polar angle θ is measured in the $Y - Z$ plane from the z axis.

In experimental high energy physics, particles are often characterised by their *rapidity* (y), which is defined as:

$$y \equiv \frac{1}{2} \ln \frac{E + p_{\mathrm{z}}}{E - p_{\mathrm{z}}} \,, \tag{4.2}$$

and in the ultrarelativistic limit ($m \ll p$) is transformed to a spatial coordinate *pseudorapidity* (η):

$$\eta \equiv \frac{1}{2} \ln \frac{|\mathbf{p}| + p_{\mathrm{z}}}{|\mathbf{p}| - p_{\mathrm{z}}} \equiv - \ln \left[\tan \frac{\theta}{2} \right]. \tag{4.3}$$

An important quantity used in nearly any analysis is the distance between two objects in $\eta - \phi$ space, which is referred to as ΔR and is defined as follows:

$$\Delta R \equiv \sqrt{(\Delta \eta)^2 + (\Delta \phi)^2} \,, \tag{4.4}$$

where $\Delta \eta$ and $\Delta \phi$ represent the difference between two objects in pseudorapidity and azimuthal angle respectively. This is the primary quantity employed by jet clustering algorithms, which are described in detail in Section 7.4.

The fact that the hard scattering processes studied at CMS occur as a result of head-on proton-proton collisions from antiparallel proton beams leads to the initial momentum of the proton-proton system in the plane perpendicular to the beam

direction being negligible. Thus, the momentum conservation law implies that the vector sum of momenta in the $r\phi$ plane of all final state particles must be equal to zero:

$$\sum_{all} \vec{p}_{\mathrm{T}} = 0 \,. \tag{4.5}$$

Therefore, *transverse momentum* (p_{T}) is an important property of a particle and is defined as:

$$p_{\mathrm{T}} \equiv \sqrt{p_{\mathrm{x}}^2 + p_{\mathrm{y}}^2} \,. \tag{4.6}$$

Since some particles, like neutrinos, almost do not interact with matter, they escape the CMS detector without leaving any energy deposit in the calorimeter. Nevertheless, the magnitude of the vector sum of their transverse momenta can be derived from the imbalance of the detected transverse momenta and defined as *missing transverse energy*:

$$\not{E}_{\mathrm{T}} \equiv - \left| \sum_{measured} \vec{p}_{\mathrm{T}} \right| \,. \tag{4.7}$$

Missing transverse energy is especially important for detecting the dileptonic $t\bar{t}$ events, since they are characterised by the presence of two neutrinos from W boson decays, as described in Section 2.3.1, which implies significant $\not{E}_{\mathrm{T}}$ in the detector.

4.2.2   Superconducting solenoid

The superconducting solenoid is a vital part of the CMS detector. It was designed to provide a magnetic field with a flux density of 4 T in a free bore with a diameter of 6 m and a length of 12.5 m. The flux is returned through a 10 000 t yoke comprising 5 wheels and 2 endcaps, composed of three disks each.

The coil of the magnet constitutes of the 4-layer winding made from a stabilised reinforced NbTi conductor and operates at a temperature of 4.6 K provided by the liquid helium cryostat. The high ratio between the stored energy (2.6 GJ) and the cold mass (220 t) causes a large mechanical deformation (0.15 %) during the energising of the coil, which exceeds the values of previous solenoidal detector magnets.

Three main features differentiate the superconducting magnet of the CMS experiment from its predecessors:

- **Large number of winding layers** due to the required number of ampere-turns to generate a 4 T magnetic field (4 in CMS compared to usual 1 in ALEPH [128] or DELPHI [129] and 2 in ZEUS [130] and BaBar [131]);

- **Conductor mechanically reinforced with an aluminium alloy** in order to achieve the required energy-over-mass ratio;

- **Large dimensions** to fit the complete tracking detector and calorimeters.

The current, applied to the solenoid during its operation, was lowered with respect to the designed one, resulting in magnetic field strength of $3.8\,T$, which was done to cover possible aging effects of the coil. Exact values of the magnetic field flux throughout the detector are important for proper simulation of particle detection. Inside the tracker volume, the magnetic field was measured by Hall probes [132]. In the steel return yoke, the magnetic field was mapped using events with cosmic muons [133].

4.2.3   Tracking detector

The inner tracking system of CMS was designed to provide a precise and efficient measurement of the charged particle trajectories, which would also allow to precisely reconstruct secondary vertices emerging from hadron decays. Being enclosed inside the superconducting solenoid, the tracker makes use of the strong magnetic field, which, via the Lorentz force, bends trajectories of charged particles that travel not in the plane of the magnetic field direction. Through the measurement of the complete trajectory of a particle, a number of its important properties can be deduced:

- **point of origin**, by finding several trajectories starting from a single point (vertex);

- **sign of charge**, from the direction of bending with respect to the magnetic field direction;

- **transverse momentum**, from the radius of the particle's helical trajectory.

Since the tracker is located closest to the interaction point, it undergoes severe radiation damage. In addition, the need of the high precision of trajectory reconstruction combined with the high pile-up from multiple collisions require fine segmentation of the detector with high power density of read-out electronics. Therefore, an efficient cooling is needed, while keeping to the minimum the amount of material, in order to limit multiple scattering, bremsstrahlung, photon conversion and nuclear interactions.

These considerations led to a design entirely based on silicon detector technology with the arrangement of the silicon modules into a barrel part around the interaction point and endcap parts at the positive and negative z-axis sides, as shown in Figure 4.5.

This is one of the most ambitious parts of the CMS detector, which required the development of production methods and quality control procedures that were new to the field of particle physics detectors and took a period of 12 to 15 years of collaboration of 51 institutes [11].

Since the tracking detector is the subject of the technical contribution to this thesis, it is described in greater detail in Section 4.3.
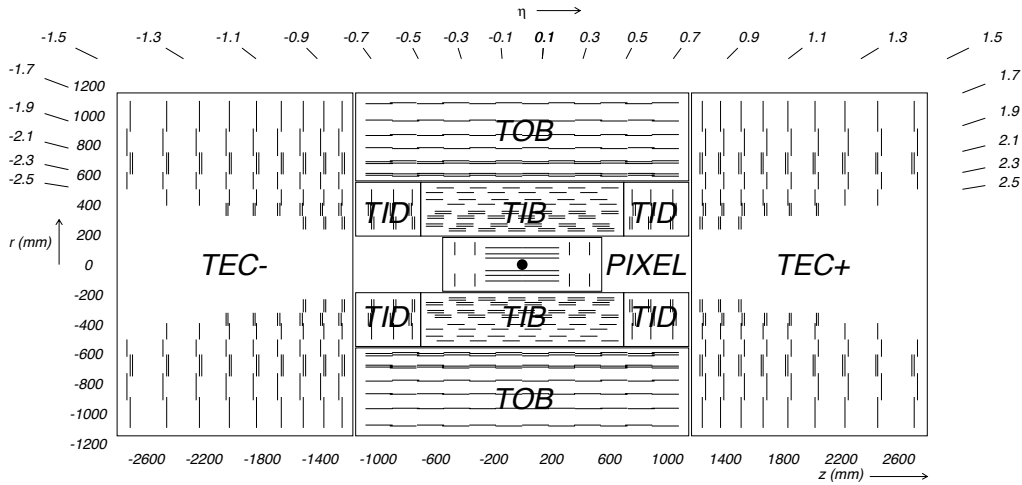
Figure 4.5: Schematic view of the CMS tracking detector. Shown are different subdetectors with a detailed explanation in Section 4.3 [11].

4.2.4   Calorimeters

While the tracker measures the trajectories of charged particles, calorimeters measure their energy by stopping them and converting energy into some detectable property. A particle passing through a calorimeter interacts with it, initiating a particle shower, which is collected and measured by a segmented set of individual calorimeter modules.

Traditionally for multipurpoose detectors, the CMS calorimeter consists of two distinct subdetectors: electromagnetic calorimeter (ECAL) and hadronic calorimeter (HCAL). The ECAL is used to measure energy of particles interacting electromagnetically, mainly leptons and photons, and covers the pseudorapidity region $|\eta| < 3.0$. The HCAL detects energy depositions from hadrons and due to hadronic showers being broader and deeper compared to electromagnetic ones, it has larger volume and surrounds the ECAL. The HCAL covers the same pseudorapidity region, while an additional forward calorimeter (HF) extends it to $|\eta| < 5.0$. In order to provide a reliable measurement of missing tranverse energy, the CMS calorimeter has large hermetic geometrical coveregare.

A distinctive feature of the CMS calorimeter system is that it is enclosed within the superconducting solenoid. While having the advantage of low energy loss by particles before entering the calorimeter, this design places stringent constraints on its size. As a result of the compact size, high-energetic showers can not be completely contained withing the space inside the solenoid. Therefore, to minimise leakage of such showers, an additional layer of tail-catching calorimeter is installed around the solenoid, and is referred to as the outer hadronic calorimeter (HO).

In addition to the calorimeters located within the CMS detector, two very forward calorimeters are also installed: the hadronic calorimeter CASTOR, installed in the $-Z$ side of CMS ($-6.6 < \eta < -5.2$) and the Zero Degree calorimeter (ZDC) with both electromagnetic and hadronic sections installed at both sides of CMS ($|\eta| > 8.3$).

Besides the energy of a particle, its type can also be identified based on the shape of the developed shower (shower shape). This shape can be characterised by three basic parameters of the detecting material:

- **Radiation length (X_0):** mean distance traveled by an electron before its energy is reduced by a factor $\frac{1}{e}$ (characterises depth of electromagnetic showers);

- **Nuclear interaction length (λ_I):** mean distance traveled by a hadron before undergoing an inelastic nuclear interaction (characterises depth of hadronic showers);

- **Molière radius:** radius of a cylinder that on average contains 90 % of the electromagnetic shower's energy deposition (characterises the width of the shower).

Electromagnetic calorimeter

The ECAL is a homogeneous calorimeter, meaning that the material used for producing a shower is also used for the deposited energy measurement. The detector consists of 75 848 PbWO$_4$ crystals, which are arranged in a barrel (EB) covering the pseudorapidity range $|\eta| < 1.479$ and two endcaps (EE) covering the forward region $1.479 < |\eta| < 3.0$. A graphical overview of the ECAL layout is given in Figure 4.6.
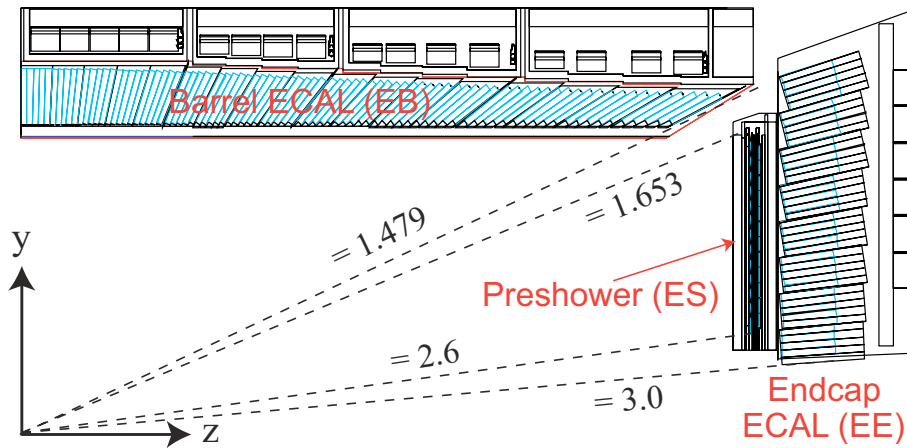

Figure 4.6: Schematic view of a longitudinal cross section of a quarter of the electromagnetic calorimeter (ECAL). Shown are the barrel (EB), the endcap (EE) and the preshower detector (ES). Dashed lines represent the pseudorapidity coverage of the subdetectors [134].

The PbWO$_4$ crystals have high density ($8.28\,\mathrm{g\,cm^{-1}}$), short radiation length ($0.89\,\mathrm{cm}$) and small Molière radius ($2.2\,\mathrm{cm}$), which allows a high resolution of shower shapes. The crystals act as scintillators, while the emitted light is measured with photodiodes. The decay time of the scintillation light is close to the time between LHC bunch crossing with $\sim 80\,\%$ of the light emitted in $50\,\mathrm{ns}$. In order to minimise the fluctuations of the scintillation light output, a water cooling system stabilises the temperature of the ECAL at $18\,^\circ\mathrm{C}$, at which 4.5 photoelectrons per MeV are collected by photodiodes.

To avoid cracks aligned with particle trajectories, the crystals are oriented towards the collision point with a small tilt. The EB crystals are mounted with a 3°

tilt in the η and ϕ projections with respect to the interaction point direction. The EE crystals point to a focus, which is $1.3\,\text{m}$ beyond the interaction point.

In front of the EE, the preshower detectors (ES) are mounted, which are sampling calorimeters made of two layers of lead radiator/silicon strip sensors. These sensors improve the position resolution of electromagnetic particles as well as the neutral pion identification.

The ECAL energy resolution was measured under a test beam of electrons and is given by:

$$\left(\frac{\sigma_E}{E}\right)^2 = \left(\frac{2.8\%}{\sqrt{E(\,\text{GeV})}}\right)^2 + \left(\frac{12\%}{E\,\text{GeV}}\right)^2 + (0.3\%)^2\,, \tag{4.8}$$

where the first term is stochastic, the second due to noise from electronics and the third is a constant term due to nonuniformity in the light collection, energy leakages and calibration errors [135]. The energy resolution of the ECAL for electrons is shown in Figure 4.7.
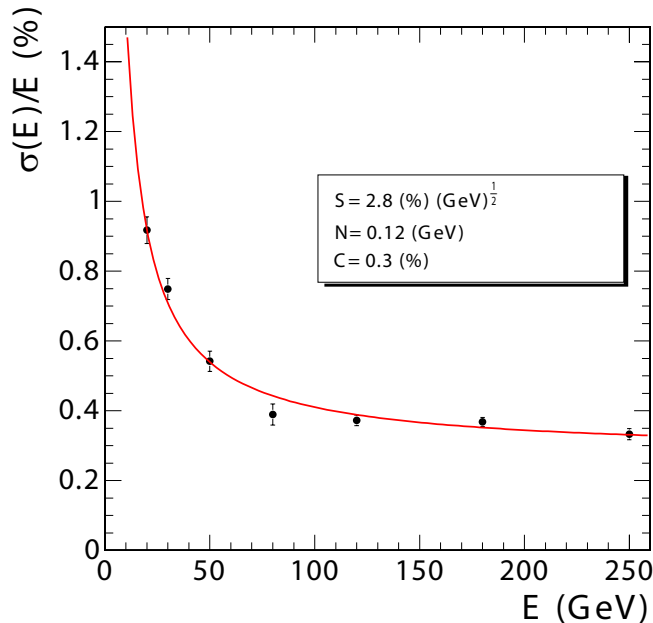

Figure 4.7: Energy resolution of the ECAL as a function of electron energy, as measured from a beam test. The points correspond to events taken restricting the incident beam to a narrow $(4\,\text{mm} \times 4\,\text{mm})$ region [11]. The stochastic (S), noise (N), and constant (C) terms are given as in (4.8) and are determined from the fit (red line) to the points.

The energy resolution for electrons from $Z \rightarrow ee$ decays is better than $2\,\%$ in the central region of the EB ($|\eta| < 0.8$) and better than $5\,\%$ in other ECAL regions. The final energy resolution of photons from $H \rightarrow \gamma\gamma$ decays is $1.2\,\%$–$2.6\,\%$ in the barrel region and $2.2\,\%$–$5\,\%$ in the endcaps. At the end of the 2011 data taking period, the percentage of active ECAL channels was $99.1\,\%$ in EB, $98.6\,\%$ in EE and $96.1\,\%$ in ES [136].

Hadronic calorimeter

A vital part of the CMS detector for the measurement of hadronic jets and missing tranverse energy ($\not{E}_T$) is the HCAL. The HCAL is a sampling calorimeter and consists of four subdectors: barrel (HB), endcap (HE), outer (HO) and forward (HF) calorimeters. The schematic view of the arrangement of the HCAL subdetectors with the tower segmentation is shown in Figure 4.8.
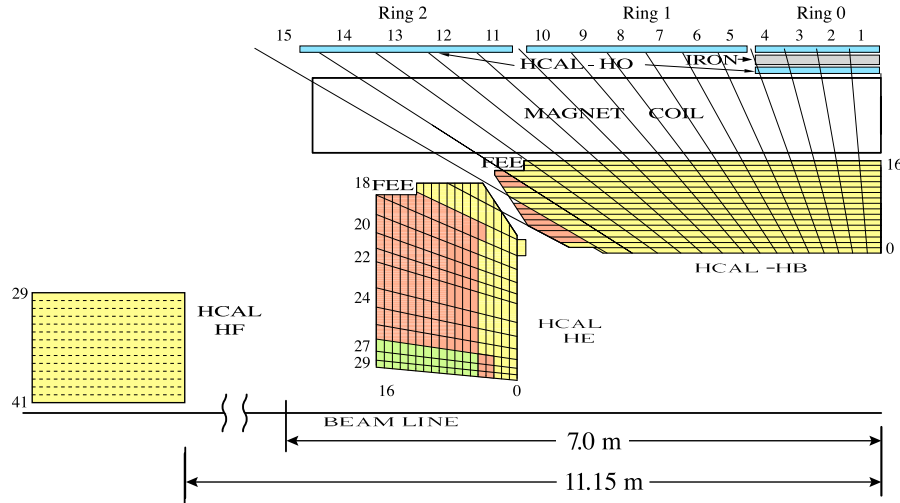

Figure 4.8: Schematic view of a longitudinal cross section of a quarter of the hadronic calorimeter (HCAL) in the rZ plane. Shown are the barrel (HB), the endcap (HE), the outer (HO) and the forward (HF) calorimeters. Numbered segments represent individual towers. The signals of the tower segments with the same colour are added optically for "longitutidinal" segmentation. HB, HE and HF are built of 36 identical azimuthal wedges with $\Delta\phi = 20°$ [137].

In the HB two double layers of steel plates followed by brass plates are used as an absorber, while in the HE only brass is used to not be influenced by the magnetic field of the solenoid. Plastic scintillators are used to produce light that is collected by wavelength-shifting fibers connected to photodetectors. The HB consists of 16 layers of absorber plates and 17 layers of scintillator tiles, covering the pseudorapidity region $|\eta| < 1.3$, while the HE comprises 17 layers of absorbing layers and 18 scintillating layers, covering $1.3 < |\eta| < 3.0$ range.

The scintillator layers are arranged in towers with individual readouts of size $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$ in HB, which match 5×5 arrays of ECAL crystals, while in HE readouts have dimension of $\Delta\eta \times \Delta\phi = 0.17 \times 0.17$. The thickness of HB and HE is $5.8 - 10.6\lambda_I$ depending on pseudorapidity.

The HO, being installed outside the solenoid, uses its coil as an absorber. The outer calorimeter consists of two layers of scintillator with an iron absorber in the middle, extending the total depth of the CMS calorimeter to at least $11.8\lambda_I$.

The forward calorimeter is a radiation hard component that extends the pseudorapidity coverage to $|\eta| = 5.0$ and uses Cherenkov-based technology. It consists of iron absorber and quartz fiber scintillators interfaced with photomultiplier tubes (PMT) to convert light into electric signals [134].

The energy resolution of the HCAL for jets, when combined with ECAL, is [138]:

$$\left(\frac{\sigma_E}{E}\right)^2 = \left(\frac{100\%}{\sqrt{E(\,\mathrm{GeV})}}\right)^2 + (5\%)^2 \,. \tag{4.9}$$

The jet transverse-momentum resolution as a function of jet p_T is shown in Figure 4.9 for three $|\eta|$ regions. It is typically $10\,\%$ at high p_T, and can reach $50\,\%$ for jets with $p_\mathrm{T} < 20\,\mathrm{GeV}$.
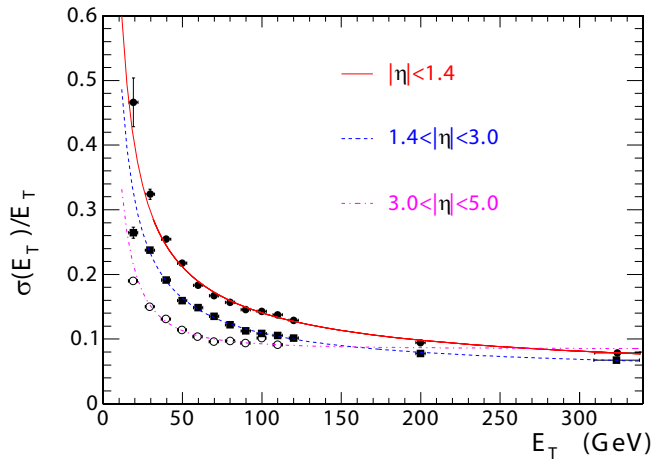

Figure 4.9: Jet transverse-momentum resolution as a function of the jet E_T for jets in the barrel region (red), in the endcap region (blue) and forward region (violet). The jets are reconstructed with the iterative cone $R = 0.5$ [11].

4.2.5   Muon system

As the name CMS says, muon identification has been a high priority in the design of the detector. A possible observation of the Higgs boson with four muons in the final state $(H \to ZZ \to 4\mu)$ was one of the motivations for the emphasis on muon measurements due to low radiative losses compared to electrons, providing the best 4-particle mass resolution.

The muon system has three main functions: muon identification, momentum measurement, and triggering. Good muon momentum resolution and trigger capability are enabled by the strong field solenoidal magnet and its flux-return yoke, which also serves as a hadron absorber for the identification of muons. The material thickness traversed by muons through the whole detector complex in different η directions is shown in Figure 4.10 [11].

The gaseous particle detectors (muon chambers) used in CMS are of three types: drift tube chambers (DT), cathod strip chambers (CSC), and resistive plate chambers (RPC). Constituting $25\,000\,\mathrm{m}^2$ of detection planes, they were designed to be inexpensive. A layout of the muon system is shown in Figure 4.11.

The DT chambers (stations) are installed in the barrel region, covering $|\eta| < 1.2$, where the neutron-induced background is small, the muon rate is low, and the 3.8-T magnetic field is uniform and mostly contained in the steel yoke. The drift tube chambers are arranged into four stations installed between the layers of the return
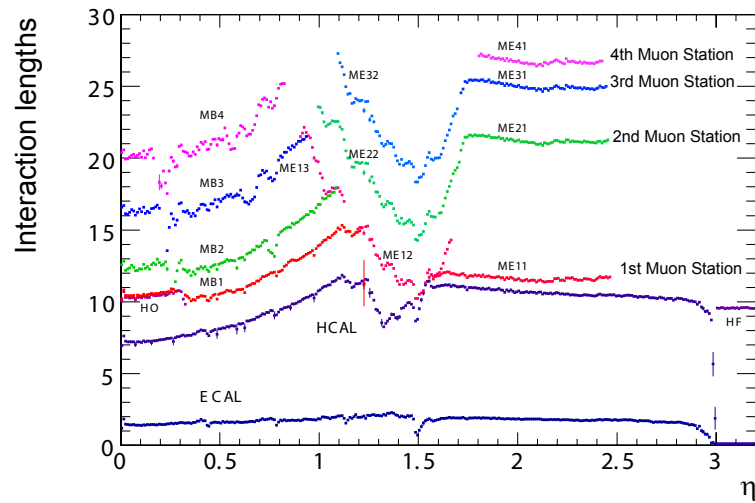
Figure 4.10: Material thickness of the CMS detector as a function of pseudorapidity, expressed in interaction lengths after the ECAL, HCAL, and at the depth of each muon station [11]. The dips around $\eta = 1.5$ are caused by the reduced material budget at the transition point between the barrel and endcap subdetectors.
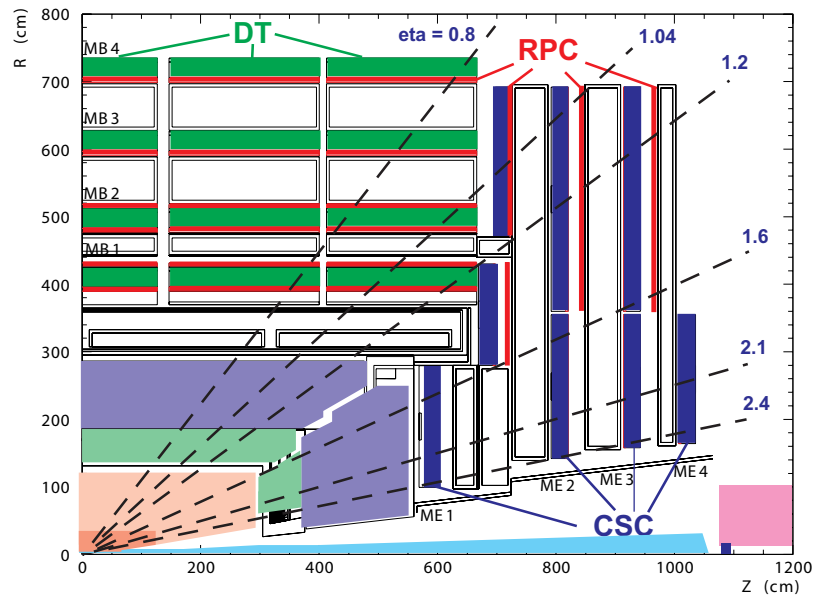

Figure 4.11: A schematic view of the muon system layout. Dashed lines represent different values of pseudorapidity. Muon chambers are arranged such that muons at any pseudorapidity almost always pass through four layers [11].

yoke. The first three stations provide measurements in the $r\phi$ plane and in the z direction, while the fourth station has no modules for measurements along the z axis.

The CSC chambers are installed in the endcaps, covering the pseudorapidity region $0.9 < |\eta| < 2.4$, due to high muon rates and background levels, as well as the strong and non-uniform magnetic field in the z direction. To work in such an environment, the CSC chambers have fast response time, fine segmentation and high radiation resistance. Four CSC stations are installed between flux return yokes in each muon-detector endcap, perpendicularly to the beam direction. The cathode strips of each chamber run radially outward, provide a measurement in the $r\phi$ bending plane, while the anode wires run approximately perpendicular to the strips, providing measurements of η.
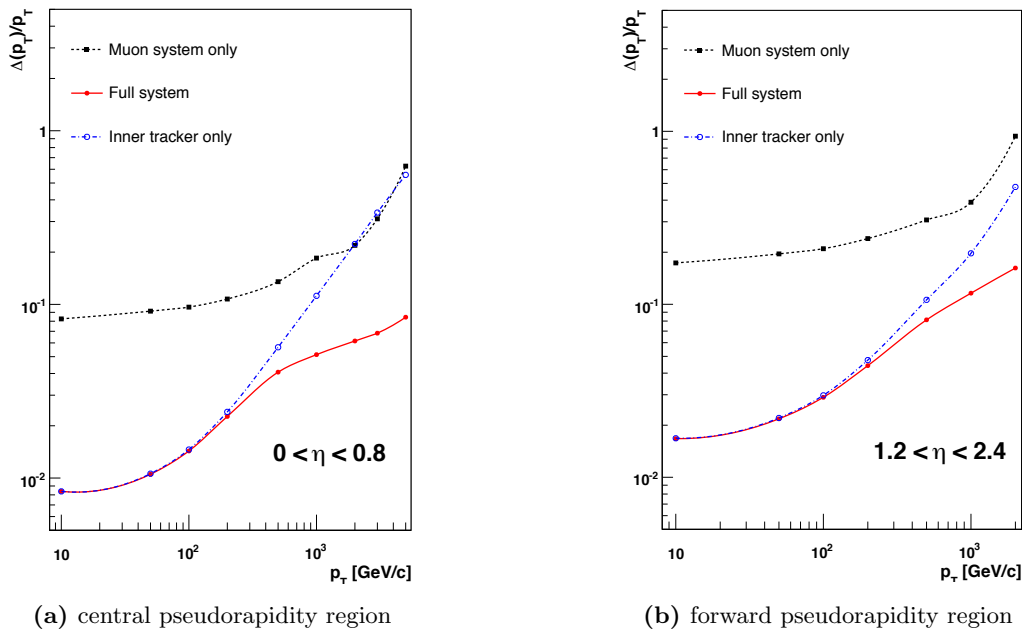

(a) central pseudorapidity region   **(b)** forward pseudorapidity region

Figure 4.12: The muon transverse-momentum resolution as a function of the transverse-momentum (p_T) using the muon system only, the inner tracking only, and both [11].

In addition to the DT and CSC stations, the PRC chambers are installed in both the barrel and endcap parts of the muon system, covering the pseudorapidity range $|\eta| < 1.6$. They have coarser spatial segmentation and, as a result, lower spatial resolution but their fast response with good time resolution makes them very efficient for triggering. Furthermore, the PRC chambers help to resolve ambiguities in attempts of making tracks from multiple hits in a chamber.

Both the inner tracker and the muon system provide independent measurements, but their combination yields the highest resolution, as can bee seen in Figure 4.12. The combination of the two in a global fit is especially important for muons with very high momentum allowing to reach a momentum resolution of $\mathcal{O}(5\%)$ for muons with $p_T = 1\,\text{TeV}$ [11].

4.2.6 Particle identification and triggering

All the complicated and diverse internal structured of the CMS detector, which combines numerous subdetectors of different types, is essential for the distinction between different kinds of particles and their properties. This distinction is especially important for the efficient use of the available computing and information-storage resources.

As mentioned earlier, the proton bunches cross in the centre of the CMS detector every 50 ns, while the average number of proton-proton collisions in each bunch crossing is equal to 21. From this enormous number of collisions, only a tiny fraction has hard scattering of the protons, which leads to the final states that are worth to be recorded and used for the wide range of physics analyses. Therefore, it is essential to have a good distinction between the interesting events and the non-interesting ones, which can be performed on the event-by-event basis.

This task is addressed by the CMS Trigger system [139], whose main purpose is to quickly identify the presence of possible high-energetic particles, which would be a signature of the hard scattering process in the event. It allows to reduce the bunch-crossing rate of 20 MHz to the recorded-event rate of 100 Hz to 300 Hz.

Particle identification

For both the trigger system and for the later offline analysis, the identification of particles and their properties is crucial. Various types of particles interact with matter differently, leaving a typical or no signal in specific subdetectors correspondingly, as illustrated in Figure 4.13.
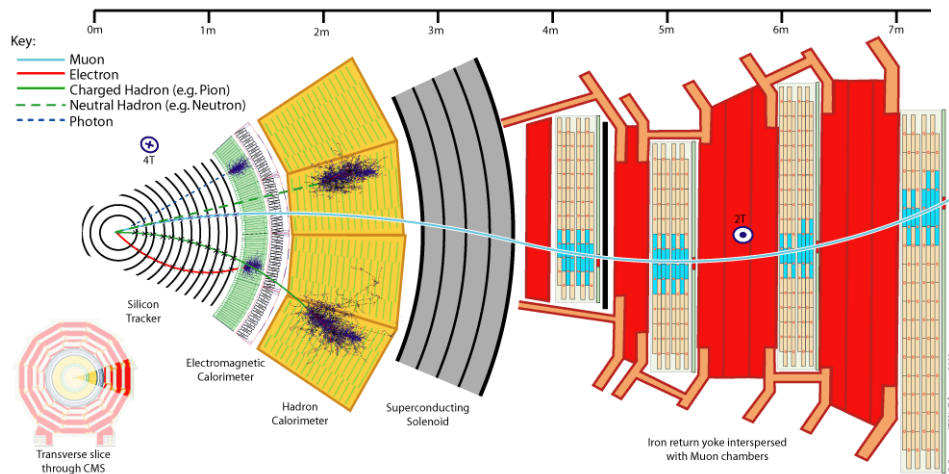
Figure 4.13: Schematic view of the principle behind the identification of different particle types. It is based on the fact that different types of particles leave signals in different specific structures of the CMS detector, as shown on the figure. Taken from [140].

In particular, charged particles, i.e. electrons, muons or charged hadrons, leave a signal in the silicon tracker. The direction of the magnetic field is constant, and particles always propagate through the detector from the centre to the outer part.

Therefore, the charge of such a particle can be identified based on the side to which its trajectory is bent.

Signals from the electromagnetic and hadronic calorimeters allow to distinguish between charged leptons and hadrons. The distinction between charged and neutral hadrons can be made by looking for energy deposits in the ECAL that coincide with the clusters in the HCAL.

The fact that photons penetrate the silicon modules of the tracker without interaction, allows to distinguish them from leptons, which have tracks that point to the corresponding clusters in the electromagnetic calorimeter.

Muons are identified by combining the tracks in the silicon tracker with the clusters in the muon-system chambers. The fact that the muon system is located outside the superconducting solenoid leads to the muon trajectory consisting of two arcs bent in opposite directions.

Level-1 trigger

The first step in the event identification is performed by a set of *level-1 triggers* (L1), which are implemented in customizable hardware logic boards. These triggers use the quickly-processable information from the calorimeters and the muon system with reduced granularity in order to achieve the processing time of $3.2\,\mu s$ [134]. Primitive particle candidates and characteristic properties of the event are constructed using simplified and fast algorithms. The simple reconstruction procedure is first performed locally, then regionally on the level of each individual subdetector, and then the information from all the subsystems is combined in a global reconstruction. The properties of hypothetical electrons, photons, jets and muons are analysed, as well as event variables, like the missing transverse energy. If a set of requirements specific for the activated trigger are fulfilled, the event is kept for further checks by the trigger system.

High level trigger

If an event fulfilled the requirements of the active L1 trigger, it is processed by a *high level trigger* (HLT) [134]. An HLT is software-based and runs on a dedicated computer farm. This step involves a more detailed reconstruction of the event using information from all the detector components, which allows to build higher-level objects, like muons or jets. In order to keep the processing time below $50\,ms$ per event, the reconstruction of the complex objects is performed only in the regions of the detector that were marked as interesting by the L1 triggers.

Typically, HLT requirements include certain thresholds for p_T, η and isolation of some minimum number of objects, i.e. jets, leptons, photons. In addition, some minimum amount of $\not{E}_T$ or $H_T = \sum_{jets} p_T^{jet}$ can be required. A number of different HLT paths exist, which are optimised for signatures of specific processes of interest. If the event fulfils the requirements of at least one HLT path, it is stored. The set of HLT paths that were fired by the event is used as a first step of the event selection during the later offline analysis.

In order to reduce the amount of information to be stored, some trigger paths with high event rates are prescaled, meaning that only a fraction of the selected events is recorded. The prescale factor is then used to obtain the total event rate of the HLT path.

Triggers used for the physics analyses

Since the dileptonic final state of the $t\bar{t}$ system is one of the key signatures of the studied $t\bar{t}b\bar{b}$ and $t\bar{t}H(b\bar{b})$ processes, the following set of triggers is used to select such events:

- **Double-muon HLT:** at least two muons with $p_T^{\mu 1} > 17\,\text{GeV}$ and $p_T^{\mu 2} > 8\,\text{GeV}$. The muons are reconstructed from the measurements of the tracker and muon systems. A second HLT path is used for the double-muon event selection, which requires that the second leading muon is reconstructed only in the inner tracker.

- **Double-electron HLT:** at least two electrons with $p_T^{e1} > 17\,\text{GeV}$ and $p_T^{e2} > 8\,\text{GeV}$. The electrons are reconstructed by matching tracks in the inner tracking detector to clusters from the calorimeter. In addition, minimal identification and isolation requirements are applied to triggering electrons.

- **Electron-muon HLT:** at least a pair of an electron and a muon with $p_T^{\mu/e} > 17\,\text{GeV}$ and $p_T^{\mu/e} > 8\,\text{GeV}$. Electrons and muons are reconstructed as in the double-electron and double-muon triggers respectively.

The actual HLT paths used for the analyses are listed in Appendix B.1.

4.2.7 Data quality certification

The high quality of data used for the physics analyses is ensured by the dedicated Data Quality Monitoring process (DQM) [134, 141]. The purpose of the DQM is to evaluate the detector conditions during the data taking, as well as in the simulated samples.

The DQM process in CMS is split into two steps:

- **online DQM:** the status of each detector subsystem is monitored in real time during the data-taking process, to discard events with low quality of the measured physics objects;

- **offline DQM:** reconstruction and calibration is certified in the recorded data using deeper checks, for a selection of events that can be used in physics analyses.

The offline-DQM certification is applied at the level of luminosity sections, which correspond to the period of 2^{18} revolutions of the proton beams ($23.31\,\text{s}$). The list of luminosity sections that are certified for physics analyses is provided centrally by CMS [142].
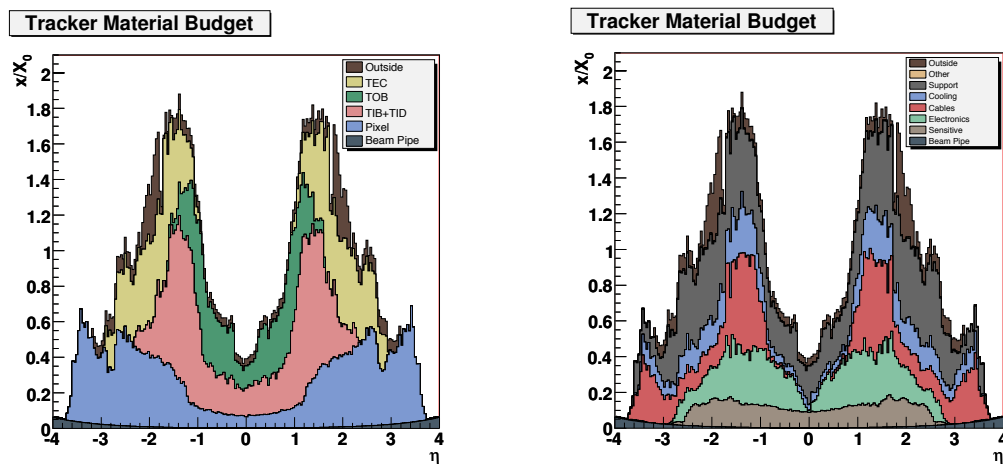
4.3   The CMS tracker

As mentioned in Section 4.2.3, the purpose of the tracker is to measure trajectories of charged particles. Since the particle flux is high at the interaction point, occupancy of the silicon modules must be kept low. Therefore, the CMS tracker has fine granularity that decreases at larger distances from the centre of the detector, where the particle flux is smaller.

Another requirement for the tracking detector has been to achieve good transverse momentum resolution: around $10\,\%$ for electrons with $p_{\mathrm{T}} = 1\,\mathrm{TeV}$. The transverse momentum resolution of a single muon can be related to the parameters of the tracking detector via a simplified Glückstern formula [143] (neglecting material interactions and assuming constant resolution across all sensors):

$$\frac{\sigma(p_{\mathrm{T}})}{p_{\mathrm{T}}} = \frac{\sigma_{hit}}{0.3BL^2}\sqrt{\frac{720}{N_{hit}+4}}\,, \tag{4.10}$$

where B is the magnetic field strength (in T) and L represents the the radial extension of the tracker (in cm), σ_{hit} is the spatial resolution of a single sensor and N_{hit} corresponds to the number of measured hits in the track under question [144]. According to the formula, better track-momentum resolution can be achieved by a larger diameter of the tracker, stronger magnetic field or finer granularity to increase the N_{hit}.

While material interactions are not included in (4.10), it is clear that they degrade the momentum resolution and must be kept to a minimum, since large amounts of material lead to significant energy losses, multiple scattering of electrons and to photon conversions. The material budget of the CMS tracking detector varies from $0.4X_0$ (see Section 4.2.4) in the very central region to $1.8X_0$ at the pseudorapidity regions that cross both the barrel and the endcap detectors, as shown in Figure 4.14.

(a) break-down by different subdetector parts (explained in Section 4.3.1)

(b) break-down by functional contributions

Figure 4.14: Material budget of the CMS tracker, including the beam pipe, as a function of pseudorapidity (η), expressed in radiation lengths (X_0) [11].

4.3.1 Tracker layout

The CMS tracker has a length of 5.8 m and a diameter of 2.5 m and consists of 25 684 silicon sensors, which measure the hits from the tracks passing through them. Two conceptually different types of sensors are used in the tracker:

- **Pixel sensors:** consist of a two-dimensional (2D) array of pixel cells providing 2D hit-position measurements on the surface of the sensor;

- **Microstrip sensors:** consist of a set of parallel or quasi-parallel strips providing only 1D hit-position measurements (in a single direction).

Depending on the type of the sensors, the tracker is divided into two distinct parts: the pixel tracker and the strip tracker. Furthermore, based on the arrangement and location of the sensors, six large structures are distinguished within the tracker, referred to as "subdetectors".

The general layout of the tracker with all the subdetectors is shown in Figure 4.5, while a more detailed view on a part of the tracker is given in Figure 4.15. An overview of the pixel and microstrip subdetectors as well as corresponding numbers of sensors are given in Table 4.1.

Table 4.1: Number of individual sensors in each subdetector of the pixel and microstrip detectors. The colour coding matches the one in Figure 4.15.

STRIP (1D)				PIXEL (2D)	
TEC	TOB	TID	TIB	FPIX	BPIX
6400	5208	816	2724	672	768
24 244				1440	

Total: 25 684

In the barrel part of the tracker, sensors are arranged in concentric cylindrical layers, consisting of sensors located at fixed radius r, while in the forward part layers consist of modules located at variable radial distance and at fixed positions along the z axis. Each layer is further segmented into rings: along the z axis in the barrel subdetectors and along the radius r in the forward parts of the tracker. The barrel and the endcaps of the tracker together provide a pseudorapidity coverage of the $|\eta| < 2.5$ region.

Pixel tracker

The pixel tracker is a vital part of the tracking system as it is located directly next to the beam pipe and must provide very high spatial resolution to resolve the large amount of tracks coming from the collision point. Using the 2D pixel sensors arranged in layers allows to achieve similar track resolution in both the $r\phi$ and z directions. Thus, a 3D vertex reconstruction in space is possible, which is especially important for the identification of secondary vertices with a low number of tracks associated to them.
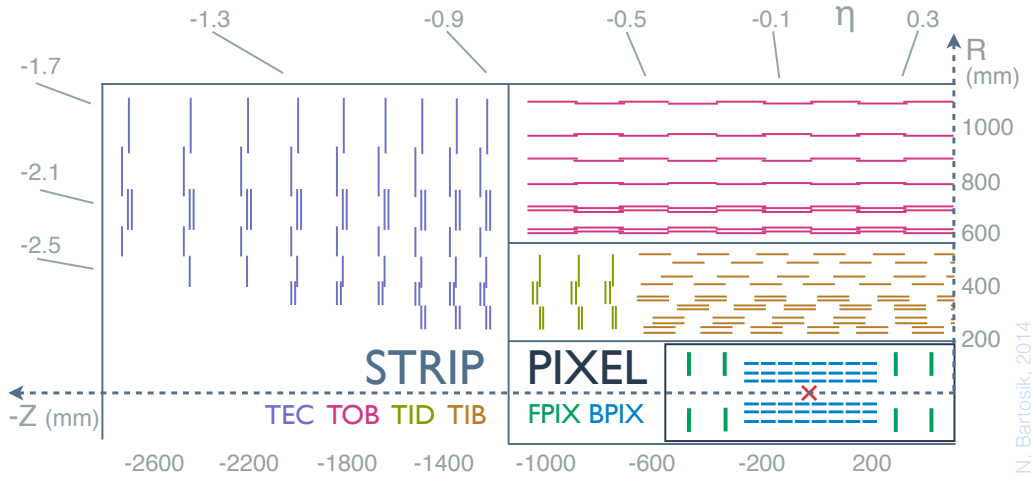
Figure 4.15: Detailed layout of a part of the CMS tracker. Sensors of each subdetector are marked by a separate colour. Double lines in some parts of the microstrip (strip) detector represent double-sided modules, which consist of a pair of microstrip sensors attached back-to-back. Based on the layout image from [11].

Mechanically, the pixel detector consists of two independent subdetectors: the barrel pixel detector (BPIX) and the forward pixel detector (FPIX). Together they cover the pseudorapidity range of $|\eta| < 2.5$, matching the coverage of the whole tracking detector. A schematic view of the pixel-detector sensor layout is shown in Figure 4.16.
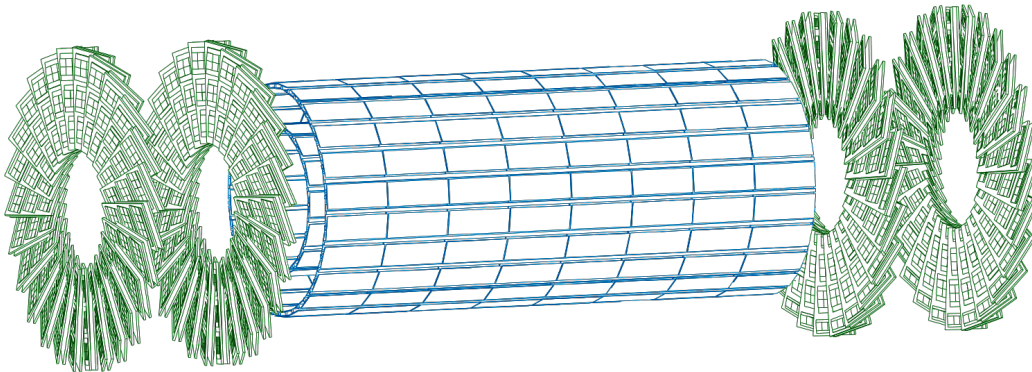

Figure 4.16: Detailed layout of the pixel tracker. The barrel part (BPIX) is coloured in blue while the forward disks (FPIX) are marked with green colour, matching the colour coding in Figure 4.15. The sensors in BPIX are aligned along the global z axis, while in FPIX sensors are tilted at 20° from the plane perpendicular to the z axis. Based on a layout image from [11].

The BPIX is 57 cm long and comprises 3 layers of pixel sensors located at mean radii of 4.4, 7.3 and 10.2 cm. It is mechanically composed by the upper $(+y)$ and lower $(-y)$ half-barrels, while the two halves at $+z$ and $-z$ have independent electric power supplies. The drift of the electrons to the collecting pixel implant is perpendicular to the 3.8 T magnetic field of the superconducting solenoid. Therefore, the resulting Lorentz drift (Lorentz drift angle $\theta_{LA} = 32°$) leads to a spread of the collected signal charge over more than one pixel, which allows to achieve a

spatial resolution of 15–20 µm using charge interpolation [11]. Within a single layer, modules located at the same ϕ have a common orientation of their sensitive layer, while the next rows of modules in ϕ has an opposite orientation. Thus, a drift direction of charge-carriers alternates in adjacent ladders of BPIX modules along global ϕ.

The FPIX extends from 6 to 15 cm in radius and consists of 2 layers located at $|z| = \pm 34.5$ cm and $|z| = \pm 46.5$ cm. The sensors in the FPIX disks are arranged into blades, which are tilted at 20° in a turbine-like geometry to induce charge-sharing. The charge-sharing is mainly due to the geometric effect of particles entering the detector at an average angle of 20° away from normal incidence [145]. The tilt of the FPIX sensors leads to a slight Lorentz drift, which further enhances the charge sharing between neighboring pixels, leading to a nominal hit position resolution of approximately 15 µm. This arrangement of the 3 barrel layers and 2 forward disks at each z side ensures 3 track hits in the pixel tracker over almost the full pseudorapidity range.

All the pixel sensors have a width of 285 µm with an equivalent structure of 100×150 µm pixel cells (with minor differences in processing between barrel and disk sensors) and operate at a bias voltage of 600 V. This structure provides a resolution of about 15 µm in the global $r\phi$ direction and 20 µm in the orthogonal direction. In total, 9 different geometries of the sensors are employed by the pixel detector with a varying number of read-out chips (ROC):

- **2 in BPIX:** full sensors (2×8 ROCs) across the whole barrel except for the edges of each layer constructed of half-sensors (1×8 ROCs);

- **7 in FPIX:** ranging from 1×2 ROCs to 2×5 ROCs with 2–3 sensors installed on a single wafer of the disk blade.

Considering the blades in FPIX as single structure units, it is more appropriate to use the term *module* instead of *sensor*, especially for the alignment purposes described in Chapter 5. Therefore, modules are equivalent to sensors in BPIX and represent groups of sensors sharing the same blade in FPIX.

Strip tracker

The strip tracker surrounds the pixel detector and consists of microstrip sensors, which provide 1D measurements in a single direction. The sensors are arranged into 4 distinct subdetectors: the tracker inner (TIB) and outer (TOB) barrels covering the central region around the interaction point accompanied by the two tracker inner disks (TID) and two tracker outer endcaps (TEC), one from each side along the z axis. The relative layout of the subdetectors can be seen in Figure 4.15.

Similarly to the pixel tracker, in the strip detector the silicon sensors are arranged into cylindrical layers along r in the TIB and TOB, while in the TID and TEC layers are made of distinct disks along z. TIB and TOB modules that are located next to each other along z, have alternating orientations, which leads to charge carriers of adjacent modules drifting in opposite directions. Unlike FPIX, which has the disks segmented only along ϕ, the microstrip endcaps cover a significantly larger area and, therefore, have additional segmentation along r into rings.

The variety of sensor geometries is significantly wider in the strip detector compared to the one in the pixel tracker, and is needed to efficiently equip all the regions in the CMS tracker. Firstly, the sensors differ by their thickness. The thin sensors with thickness of d = 320 ± 20 μm are installed in TIB, TID and the inner 4 rings of the TECs. The thick sensors (d = 500 ± 20 μm) are installed in the remaining 3 outer TEC rings and in the TOB. Secondly, 15 different sensor shapes are employed by the strip tracker: 2 rectangular each for TIB/TOB and 11 wedge-shaped for TID/TEC. A schematic overview of the sensor geometries used in the strip tracker is shown in Figure 4.17. The sensors have either 512 or 768 strips, reflecting the read-out modularity of 256 channels and operate under the bias voltage of 450 V.
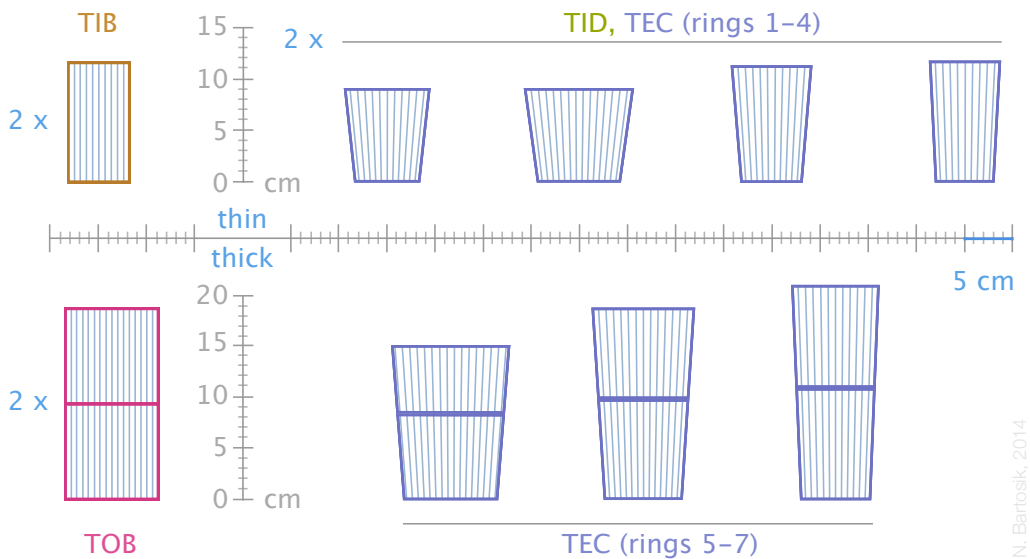

Figure 4.17: Geometries of microstrip sensors utilized in the CMS tracker. In the outer barrel and outer TEC rings the sensors are paired to form a single module. The TIB and TOB sensors exist in two types, of same area but with different pitch (distance between neighboring strips). The sensors in the TID and inner TEC rings exist in two different versions (only the TEC version is shown). Strips are parallel and equidistant in the rectangular sensors. In the wedge-shaped sensors, strips are quasi-parallel, originating from a single imaginary focus point located at the beam line, with equidistant angular spacing. Based on a similar image from [11].

A significant drawback of the microstrip sensors compared to the pixel sensors is that a hit position is measured only in a single direction perpendicular to the strips and is derived from a serial number of the signal strip. The hit position along the strip is therefore unknown. For instance, in the inner or outer barrel, strips in the modules extend along the global z axis and provide a precise measurement of the ϕ angle of the hit as well as its radius, which is constrained by the alignment of the module along r. The position along the z axis is only limited by the length of the module, which can reach 12 cm. Hence, such modules are referred to as $r\phi$ *modules*.

In order to have measurements of the hit position along the strips, some layers of the barrel and endcaps consist of double-sided modules, which comprise a pair of modules mounted back-to-back. One of the modules is an $r\phi$ module, which has the same orientation as the regular single-sided modules, while the other one is tilted around the module's normal by a stereo angle of 100 mrad and is called a *stereo*

module. Thus the hit position along the strip can be estimated from the effective area covered by the two skew strips. Such double-sided modules are marked by double lines in Figure 4.15 and constitute the two inner layers of the TIB and TOB as well as the two inner TID/TEC rings and the TEC ring 5. An overview of the microstrip module properties sorted by their position in different subdetectors is listed in Table 4.2.

Table 4.2: Characteristics of silicon microstrip modules, sorted by their location in the subdetectors. The modules in TIB/TOB are separated by layers (L), in TID/TEC by rings (R). Given are the number of modules, their thickness, the pitch of the strips and the number of readout channels per module [146].

Substructure	Modules	Thickness [µm]	Pitch [µm]	Channels
TIB L1–2	1536	320	80	768
TIB L3–4	1188	320	120	512
TOB L1–4	3528	500	183	512
TOB L5–6	1680	500	122	768
TID R1	288	320	97	768
TID R2	288	320	128	768
TID R3	240	320	143	512
TEC R1	288	320	97	768
TEC R2	576	320	128	768
TEC R3	640	320	143	512
TEC R4	1008	320	126	512
TEC R5	1440	500	143	768
TEC R6	1008	500	184	512
TEC R7	1440	500	158	512
Total:	**16 588**		$\approx 9.3 \times 10^6$ **strips**	

The strip pitch of the sensors increases at larger distances from the beam line, and ranges from 80 µm to 184 µm, which translates into a resolution along the sensitive coordinate (perpendicular to the strip direction) varying from 23.1 µm to 59.2 µm. The stereo modules in combination with the corresponding $r\phi$ modules provide a resolution around $230 - 530$ µm along the strip direction depending on the position of the module.

4.3.2 Operation of silicon modules

The silicon sensors of the CMS tracker are composed of an n-type doped wafer with the p-type doped pixels/strips applied on it, forming a pn-junction. In the vicinity of the border between p-type and n-type material, the loosely bounded electrons of the donors diffuse into the p-type material and can combine with the holes of acceptors, while holes diffuse from the p-type to the n-type. Therefore, a certain region is depleted of free charge carriers and an electric field builds up between the p-type and n-type material. This electric field is amplified by a bias voltage of the

same polarity as the initial potential. The required voltage (V_{depl}) to deplete the whole silicon sensor of thickness d is given by:

$$V_{\text{depl}} = \frac{ed^2}{2\epsilon_{\text{Si}}\epsilon_0}|N_{eff}|, \tag{4.11}$$

where $|N_{eff}|$ is the effective doping concentration, ϵ_{Si} and ϵ_0 represent the dielectric constant of silicon and vacuum respectively, and e is the electron charge [144].

When a charged particle penetrates the depleted region it causes ionisation that produces electron-hole pairs. Under the applied bias voltage, the charge carriers drift through the material of the sensor leading to a small measurable electric current pulse that is detected by the corresponding readout chip.

Irradiation damage

A high energy particle traversing the silicon sensor can displace an atom in its lattice. A vacancy and an interstitial are produced (Frenkel pair), which both can combine with impurities of the silicon bulk, thus forming defects. The displaced atom can have high energy, which is lost along the path by ionization and displacements of other atoms, leading to the formation of dense defect agglomerates. Hence, the silicon damage due to irradiation is the dominant source of the tracker performance degradation at the LHC. The macroscopic effects from the irradiation damage are a change of the effective doping concentration, an increase of the leakage current and a reduction of the charge collection efficiency.

The high particle flux from hadron collisions and the corresponding amount of hadronic interactions places strict requirements on the radiation hardness of the silicon sensors. The performance of the tracker modules under radiation was tested under electron/positron test beams at the DESY II synchrotron [144]. Since the tests showed that low temperatures reduce the irradiation damage of the sensors, the tracker modules operate at a temperature of about $-10\,°\text{C}$. Therefore, a special pipeline system for cooling liquid is connected to all modules in order to compensate the heating from readout electronics and power supplies. Furthermore, a thermal shield is placed around the outer tracker in order to isolate it from the electromagnetic calorimeter, which operates close to room temperature ($\approx 18\,°\text{C}$).

Lorentz angle effect

The silicon sensors in the tracker are aligned such that the magnetic field from the solenoid is perpendicular (barrel region) or nearly parallel (endcap region) to the electric field inside the sensors. The combination of the two non-parallel fields induces a Lorentz force on the charge carriers drifting towards the readout channels. Therefore, the trajectory of the drifting charge carrier is deflected by a certain Lorentz angle (θ_{LA}), which affects the charge sharing between multiple channels, as shown in Figure 4.18, and is referred to as *Lorentz angle effect*. This effect has to be corrected in the reconstruction step and is one of the main subjects of the technical contribution to this thesis. It is described in detail in Section 5.4.
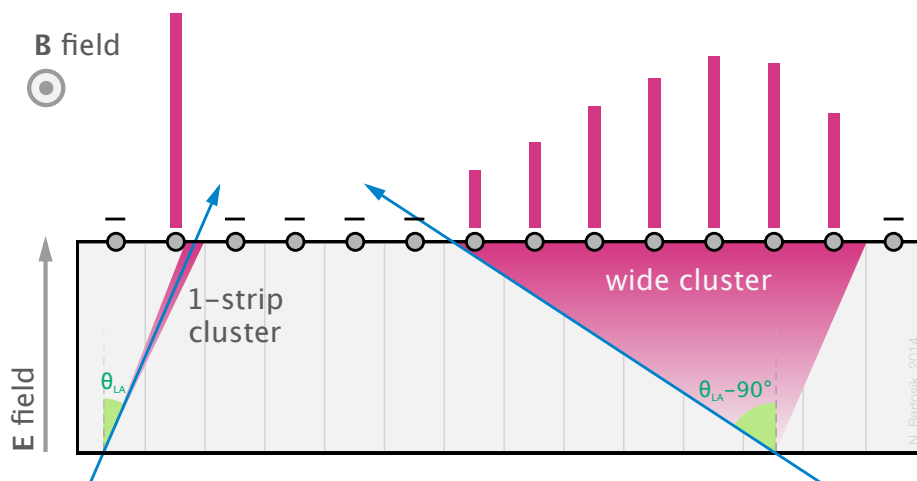
Figure 4.18: Schematic view of a drift of charge carriers in a silicon sensor under the Lorentz force. Shown is a barrel strip module with the electric field in the sensor perpendicular to the magnetic field. The black dots mark the strips, the magenta bars above represent the induced signal on the channel. The charge carrier drift direction is deflected from normal by a Lorentz angle (θ_{LA}). The direction of the left track coincides with θ_{LA} leading to a minimal signal dispersion. The right track traverses the module by an angle perpendicular to θ_{LA} causing the maximum distribution of the read out signal across multiple strips.

Backplane effect

Although the CMS detector was primarily designed to measure particles originating from pp collisions, it can detect any particles traversing it. For example, high-energetic cosmic muons can penetrate the ground and reach the CMS detector. While such particles are treated as background during hadron collisions and are rejected, they are measured and recorded for alignment purposes when there is no beam in the accelerator, as described in Section 5.3.2.

Two different signal readout schemes are applied to the tracker modules in order to optimise its performance under different conditions:

- **peak mode:** long readout time that allows to measure the full signal shape from the module and is primarily used during cosmic data taking;

- **deconvolution mode:** short readout time limited by the bunch crossing rate. Only the turn-on curve of the signal is measured at three points, while the shape of the signal is extrapolated.

The 50 ns time spacing between bunch crossings make the deconvolution (deco) mode the only one suitable for the normal data taking during hadron-collisions. Since not all charge carriers reach the channels during the short readout time, the effective depletion area becomes thinner, leading to a change of the cluster shape, as shown in Figure 4.19. This effect leads to a shift of the measured hit position depending on the track incident angle, and is corrected by a dedicated *backplane correction*. The calibration of this correction is the second part of the technical contribution to this thesis, and is described in Section 5.5.
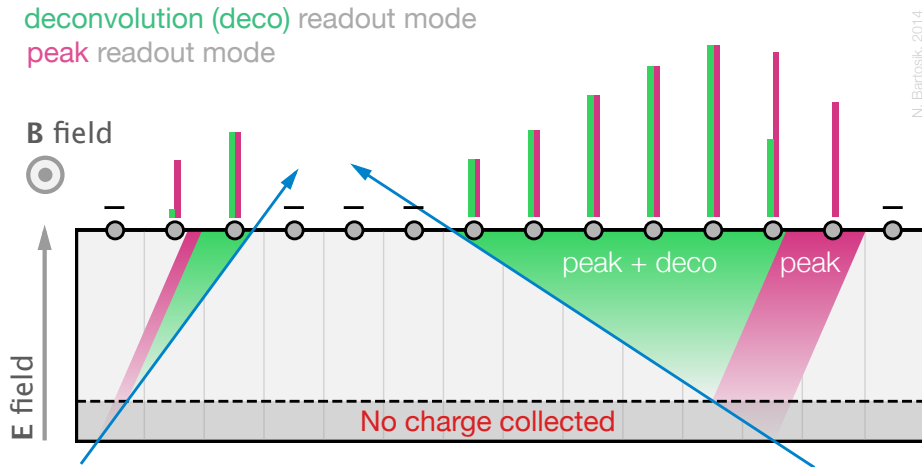
Figure 4.19: Schematic view of the backplane effect in a silicon sensor. In the *peak* readout mode, all charge from the sensor back plane is collected. In the *deconvolution* mode, the readout time is too short to collect all charge, leading to a reduced cluster size from one side, which depends on the track incident angle.

The analogue signal in the strip modules is read out by custom integrated circuits called Analogue Pipeline Voltage (APV). Each APV spans 128 neighboring strips. In the pixel modules, read-out chips (ROC) are used, each covering a sector of 52×80 pixels. The channels show a significant increase of the noise when going to the edges of the modules. This effect needs to be taken into account in the reconstruction step. The measured signals are transmitted via optical fibres to the radiation free zone for further processing [146].

4.3.3   Track reconstruction

The reconstruction of charged particle trajectories (tracks) in the CMS tracker is not an easy task due to the high particle multiplicity ($\approx 1000/event$) as expected at the design luminosity of the LHC. The reconstruction is based on two consecutive steps:

- **local reconstruction:** calculation of positions of individual measured hits using the digitised detector readout of the raw signal from the silicon sensors;

- **global reconstruction:** search for possible combinations of hits that make up continuous tracks corresponding to potential trajectories from charged particles that are bent by the magnetic field.

To successfully reconstruct a track, additional event-independent information is required, like calibration of the detector, positions of individual modules, etc. All this information has to be available at the beginning of the global reconstruction.

Local Reconstruction

At first, raw data from the readout electronics (electric signals) is digitised and stored, suppressing the noise. The local reconstruction takes place in two subsequent

steps: clustering and hit conversion. The procedure differs between the pixel and strip detector.

During the clustering step, signals from neighbouring channels and fulfilling certain threshold criteria are grouped into clusters taking into account the calibration of the readout electronics. The threshold criteria include a specific signal-to-noise ratio for the individual channels as well as for the whole cluster.

In the hit conversion step, the clusters are translated into hypothetical hit measurements using a cluster parameter estimator (CPE). To each hit a position and a corresponding uncertainty are assigned in the local coordinate frame of the silicon module. The calculated hit coordinates with uncertainties, as well as the charge and the profile of the clusters, form the final output of the local reconstruction.

Global Reconstruction

An input for the global reconstruction includes the output of the local reconstruction, as well as important information about the modules. On one hand it is the geometrical arrangement of the modules, as well as their shape and calibration parameters, which are all determined by the tracker-alignment procedure described in Section 5.3. On the other hand, records about the quality of the individual modules are taken into account. These records contain information about the status of the module itself, as well as of the readout electronics and the cabling. The readout from channels or even whole modules, which do not provide signals, or which suffer from the excessive noise is ignored.

The global reconstruction is based on three successive steps: the seed finding, the pattern recognition and the final fit. The default tracking algorithm used in the reconstruction and described here, is the combinatorial track finder (CTF). A detailed description of the CTF, together with studies of its performance, can be found in [147]. The tracking algorithm is based on a Kalman filter [148], which is used for both track finding and fitting. The tracking reconstruction is an iterative approach, during which CTF is run multiple times. Hits used in previous iterations are not considered anymore, and CTF is run again with progressively loosened requirements, which mainly affects only the seed finding step. The requirements for different consecutive iterations are listed in Table 4.3.

The seed finding searches for starting points, which are called seeds, consisting of a first hit combination with coarse estimates of the track parameters. Seeds are generated from hit triplets, or hit pairs for cosmic rays and in combination with the beam line for particles originating from hadron-collisions. Mainly neighbouring layers are used for seed finding, especially the inner layers. The starting parameters of the trajectory are obtained from a helix fit through the three points. Different seeding layers are requested at the different iteration steps [146].

The pattern recognition is the most computationally expensive part of track reconstruction. The way the Kalman Filter based CTF pattern recognition works is described in Figure 4.20. From each seed a propagation to the next surface is attempted. Hits are looked for in a window whose width is related to the precision of the track parameters. If a hit is found in the expected position it is added to the candidate trajectory and the track parameters are updated. As hits are added to the candidate trajectory the knowledge of the track parameters improves, thus allowing smaller windows to be opened when propagating to the next surface. If

Table 4.3: Iterative tracking steps [149]. Given are the seeding layers and track require-
ments for different steps (min. p_T, max. d_0, max. d_z, min. number of hits).
The * indicates the impact parameter with respect to a pixel vertex.

Step	Seeding structures	p_T [GeV]	d_0 [cm]	d_z [cm]	# hits
0	pixel triplets	0.8	0.2	3.0σ	3
1	pixel pairs	0.6	0.05	0.2*	3
2	pixel triplets	0.075	0.2	3.3σ	3
3	triplets: pixel, TIB L1–2, TID/TEC R1–2	0.25–0.35	2.0	10.0	3
4	pairs: TIB L1–2, TID/TEC R1–2	0.5	2.0	12.0	6
5	pairs: TOB L1–2, TEC R5	0.6	6.0	30.0	6

more than one consecutive hit is not found in the predicted position, the trajectory
is rejected as a fake and is not propagated anymore [147].

In addition, one candidate is always added without a hit in the specific layer,
since sometimes real hits cannot be reconstructed. Therefore, one, and sometimes
no layer without hit, are allowed in different iteration steps. For each new candidate
the propagation to the next layer is probed until the last layer is reached. Based on
the χ^2 of the fit and the number of hits at each layer, only the five best candidates
are kept for further propagation [146].

For the remaining candidates after the pattern recognition, the final fit is per-
formed using the Kalman filter, a so-called "dynamic" Least Squares Method [148].
The trajectory in a homogeneous magnetic field can be assumed to be a helix, de-
fined by five parameters. Due to the interaction with material of the detector, this
assumption is not completely correct. But effects like multiple scattering and energy
loss are taken into account. On each individual crossed sensor however, the assump-
tion of a helix is sufficiently correct. The trajectory state vector $\vec{T}$ on a module's
surface is then defined as

$$\vec{T} = \left(\frac{q}{p}, \frac{p_x}{p_z}, \frac{p_y}{p_z}, x, y \right) . \tag{4.12}$$

The components are: the inverse signed momentum, as well as the incident angle and
interception point in the two orthogonal directions in a local coordinate system of
the module. The Kalman filter passes through all hits of the candidate trajectories
and iteratively estimates the track parameters. The first hit of a candidate is the
starting point, $\vec{T}$ is initialised with an estimate from the pattern recognition, and the
uncertainties are set to high values. Then, three steps are carried out sequentially
for all hits of the track:

1. The state vector and the covariance matrix are propagated to the next mod-
   ule's surface to get a track prediction on it.
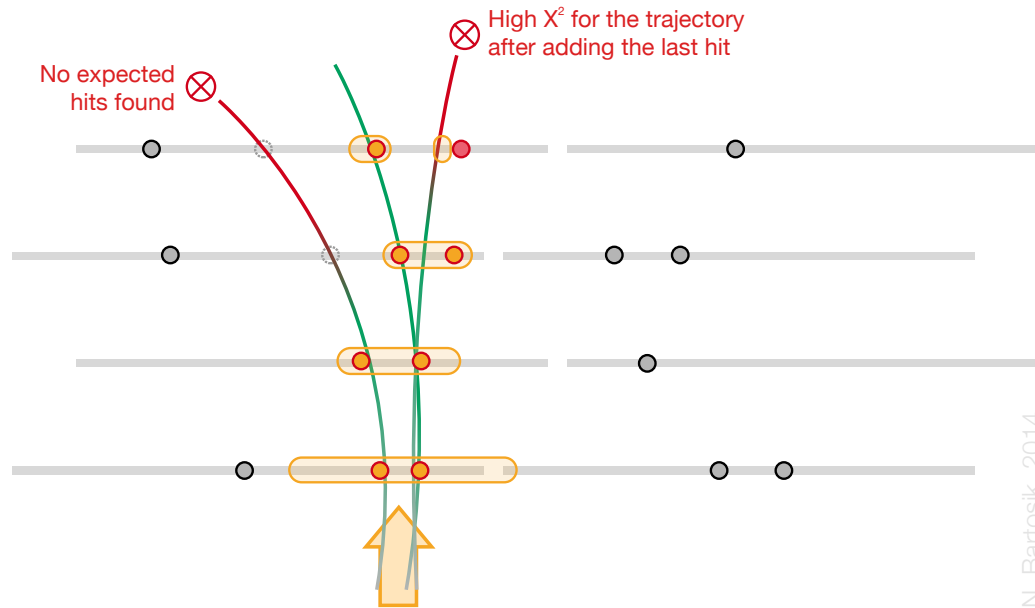
Figure 4.20: Schematic view of the track pattern recognition based on the Kalman filter. The arrow at the bottom shows the direction of the seed-finding propagation. Rounded rectangles denote the area in which hits consistent with the assumed track should lie. Dashed circles represent imaginary hits that should be present if the track was real. The reevaluation of candidate trajectories including the top layer makes the two trajectories, marked by a red cross, inconsistent with the hits. Based on a figure from [147].

2. The obtained state gets combined to a new state, including the information of the hit on this module.

3. The χ^2 value of the fit is increased, reflecting the compatibility of the track prediction and the hit.

The drawback of this method is, that the final precision is obtained only for the last hit. Thus, a second fit is applied in the opposite direction and order. The two obtained states are combined with a Kalman smoother to retrieve the final estimate of track parameters and their uncertainties. A schematic view of this process is shown in Figure 4.21. During the global reconstruction, the hit parameters are estimated again with a more sophisticated CPE using information from the track, mainly the incident angle on the sensor's surface.

Finally, a cleaning of the track collection is applied to reduce ambiguities, and quality flags are assigned. The cleaning is based on the fact, that only the best track should be kept when tracks share the majority of hits [146].

4.3.4   Track refit

The software for track reconstruction in CMS allows to repeat the final fit, which is referred to as *track refit*. Since the input, which is needed for the refit is stored independently of the applied alignment geometry and some calibrations, these conditions can be changed for the refit, and their influence can be studied. On the
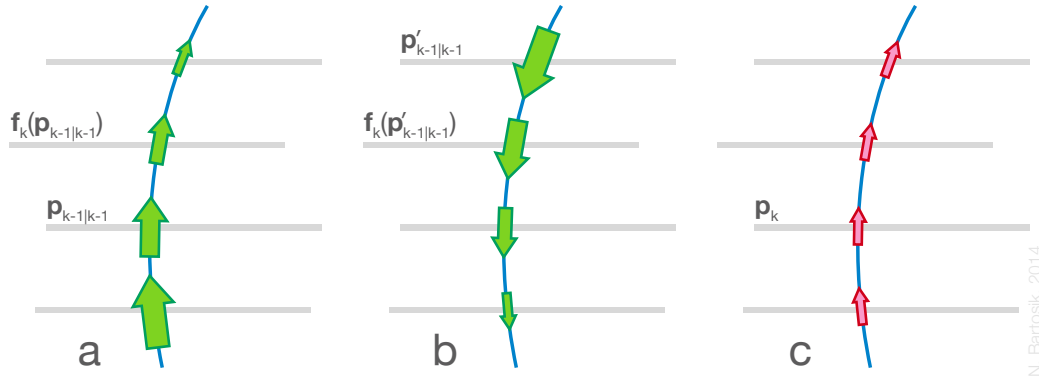
Figure 4.21: Track fitting. The three sketches represent: forward fitting (a), backward fitting (b), trajectory smoothing (c). $\mathbf{p}_{k-1|k-1}$ is the state vector on the surface $k-1$ calculated using the first $k-1$ hits. $\mathbf{f}_k(\mathbf{p}_{k-1|k-1})$ is the predicted state on the surface k; same for primed symbols. The forward and backward fit information is combined to give the best estimate of the trajectory state on each surface. Based on image from [147].

calibration side, important parameters of the reconstruction like the magnetic field or the Lorentz angle can be adjusted. On the alignment side, all alignment parameters and corresponding errors can be changed. The refit itself is configurable as well, allowing to e.g. use another CPE.

During the refit, each track that is stored as the output of the global reconstruction is fitted again. The clusters corresponding to the hits of the track and the original track parameters define the starting point. Only tracks that were found in the global reconstruction step can be refitted. But even if the conditions are identical to the ones used in the reconstruction, the refit may not converge in some cases, e.g. due to linearisation effects. Sometimes the tracks get shorter after the refit, since the refit converges only for a part of the track. Furthermore, an outlier rejection is usually applied, which discards hits that are far off from the predicted trajectory. This leads to the fact, that the number and the length of the tracks from the refit are altered by different conditions, especially from the alignment side [146].

Chapter 5

Combined alignment and calibration of the CMS tracker

The CMS tracker is a very complex detector, which is used in nearly every physics analysis performed at CMS. Its primary task is to reconstruct trajectories of charged particles traversing the tracker volume. The point of interactions of such a particle with a tracker sensor is referred to as a *hit*, which has a position determined from the cluster of charge carriers collected by the readout of the sensor. A single trajectory of a particle is constructed from multiple hits left by it in multiple tracker sensors standing on its way.

The internal structure of the CMS tracker has been described in detail in Section 4.3. Basically, it consists of 25 684 silicon sensors, which can be divided into the two categories: pixel sensors, which provide 2D measurements and strip sensors, which provide measurements only along a single direction. Some sensors are combined into single tracker modules, which can consist either of a single sensor or a pair of sensors. The spatial resolution of a single silicon sensor ranges from 15 µm in BPIX to 53 µm in TEC, although it can be slightly improved by including in the calculation a weighted charge collected in the adjacent channels.

Nevertheless, this is the precision of the hit-position measurement within the surface of a single tracker module. But the complete trajectory of a charged particle is reconstructed from hit positions in the global CMS coordinate system, which depend on actual positions of every single silicon sensor that was penetrated by the track. Therefore, in order to use the single-sensor resolution to its maximum potential, their positions have to be known with a precision of few µm.

The remaining misalignment has to be reflected by uncertainties on the positions of the modules, which are determined with a dedicated method. In fact, a too large uncertainty on positions of certain microstrip modules in TEC allowed to discover an inconsistency between the design geometry and the actual mechanical composition for these modules [146].

Furthermore, certain effects have a significant impact on the measured hit position even if positions of the modules are known very precisely. These can be the Lorentz-angle effect or the backplane effect, which were briefly introduced in Section 4.3.2. These calibrations were integrated into the complex procedure of determination of the precise tracker geometry. An overview of the alignment and calibration procedure, as well as a summary of the obtained results, are given in the following sections.

5.1   Coordinate system of the CMS tracker

An own global coordinate system $(r|r\phi|z)$ is defined for the CMS tracker, which has specific location and orientation in the global CMS coordinate system. The global tracker coordinate system is defined as the centre of gravity of all modules of the TOB, since it is expected to be more stable than other subdetectors. Thus, positions of other modules are fixed in a way that TOB as a whole has the position and orientation as in the design geometry.

For each silicon module of the tracker, a local Cartesian coordinate system $(u|v|w)$ is defined. Its origin is placed in the geometric centre of the active area of the module and axes are oriented as shown in Figure 5.1.
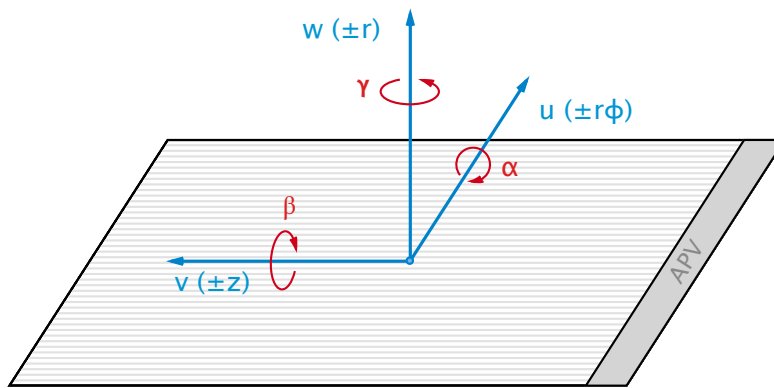
Figure 5.1: Schematic view of a rectangular silicon microstrip module with the axes of its local coordinate system: u, v and w. Rotation angles (α, β, γ) around the local axes are shown in red. The corresponding global orientations $(r\phi, z, r)$ shown in brackets are true for TIB and TOB modules.

The w-axis is always oriented in the direction of the electric field in the sensor and is normal to its surface. Thus, the strips and pixels are located on the positive w-side, while the backplane of the module is placed on the negative w-side.

The u-v-plane of the local coordinate system is defined by the surface of the silicon sensors. Along the w-axis, it is positioned in the middle of the active material. The origin of the local coordinate system is located in the middle of the module along all directions, i.e. between the two innermost pixels or strips and at the half length of the strip.

In the strip modules, the v-axis is oriented along the central strips, while the u-axis is orthogonal to it. The v-axis points away from the direction of the readout cabling (APV). In rectangular barrel modules, only the u-coordinate has a sensitive measurement across multiple strips. In the wedge-shaped sensors in TID and TEC, the measurements along the u-axis become correlated with the v-coordinate when going closer to the edges of the modules, where strips are not parallel to the central ones.

For pixel modules, the u-axis is defined as the one that is more sensitive to Lorentz angle effects. In the barrel part this is the direction along the smaller pitch in the $r\phi$ direction, while in the forward pixel it is along the wider pitch in the r direction [146].

5.2 Hardware-based alignment

Several independent strategies are used by CMS for the alignment of the tracking
detector using hardware systems, which utilise special reference markers and optical
systems.

Tracker Survey

The first step, tacker survey, is responsible for the monitoring of the assembly pre-
cision. It employs a number of different surveying measurements to obtain the
sensor-mounting accuracy of the order of $10\,\mu m$. The mounting precision of the
mechanical-support structures was estimated during the construction, and was found
to be typically within $100\,\mu m$ from the design values and maximum deviations of
about $200\,\mu m$. To some components markers were attached to perform additional
measurements after the assembly [150].

Laser Alignment System

The Laser Alignment System (LAS) [151] provides the permanent surveillance of the
mechanical structure with a precision better than $100\,\mu m$ for translation and better
than $100\,\mu rad$ for rotations. The LAS uses laser beams in the near infrared range
(1070 nm), for which the silicon is partially transparent. Special silicon sensors are
mounted in TIB, TOB and TEC for monitoring movements of these composite struc-
tures. Each TEC side is equipped with 8 beams, which are uniformly distributed in
global ϕ. Additional 8 beams are used to align TIB and TOB. The LAS layout is
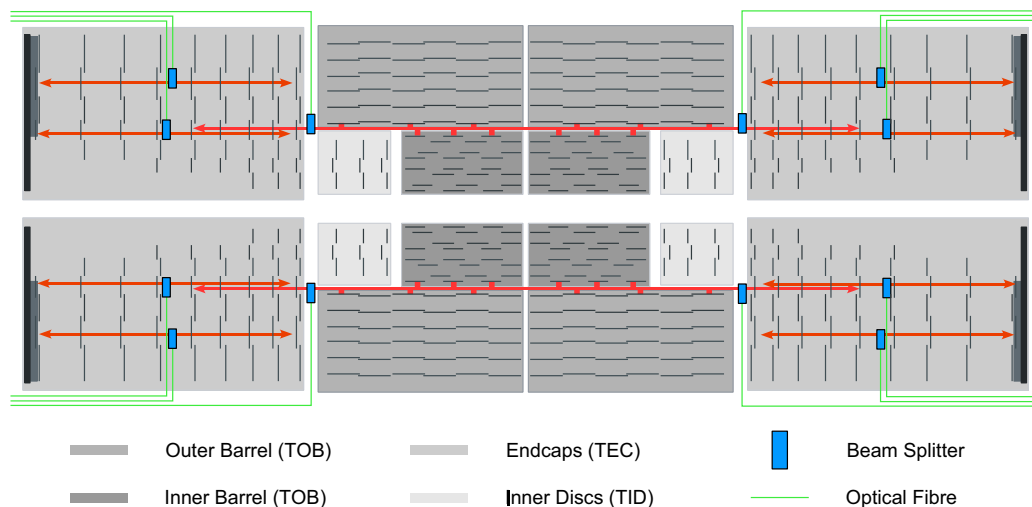schematically shown in Figure 5.2.

Figure 5.2: Schematic layout of the laser beams in the Laser Alignment System of the
CMS tracker [151].

5.3   Track-based alignment

The level of precision provided by the hardware alignment systems is not sufficient for the anticipated trajectory reconstruction. Thus, information from the hardware-based alignment is used only as starting and boundary conditions for the more sophisticated track-based alignment.

5.3.1   Track-hit residuals

A track-hit residual is the vital quantity that makes the track reconstruction and module alignment possible. To every hit i, measured on the surface of a module, a corresponding *measured hit position*, m_i, is assigned. The pattern recognition combines multiple hits that are fitted to obtain a track j, to which the corresponding *track parameters*, $\mathbf{q}_j$, are assigned. The fitted track also depends on parameters of the detector, which are referred to as *global parameters*, $\mathbf{p}$. These global parameters can be e.g. positions and orientations of the tracker modules. A track-hit residual is defined as the difference between the measured hit position m_{ij} and its projection on the corresponding track prediction f_{ij}:

$$r_{ij}(\mathbf{p}, \mathbf{q}_j) = m_{ij} - f_{ij}(\mathbf{p}, \mathbf{q}_j).\tag{5.1}$$

The effect of misaligned tracker modules on the fitted track prediction and on the track-hit residuals is schematically shown in Figure 5.3. Wrong assumption of the tracker geometry leads to the wrong fitted-track parameters, causing an increase of the track-hit residuals. Besides the pure geometry-related characteristics of the tracker modules, other global parameters related to the operation of tracker modules can be wrong, leading to a similar increase of the residuals.
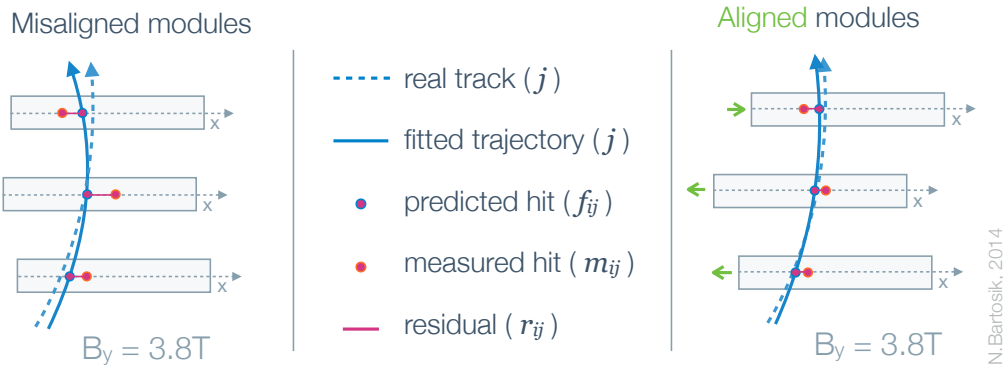

Figure 5.3: Schematic view of the fitted track prediction depending on the alignment of tracker modules. Local measured hit positions are constant, but their global positions depend on the alignment of the corresponding modules. Misaligned modules pull the measured hit positions away from the real trajectory, leading to an increase of the track-hit residuals and to a distortion of the trajectory.

Misalignment of the tracker leads to a reduced efficiency of the pattern recognition, since some hits of the track can migrate outside the search window. Also hits from distant tracks can migrate inside the search window, introducing additional ambiguities. These effects are directly translated to higher-level objects that

involve tracking information for their identification or reconstruction, e.g. charged
leptons, primary and secondary vertices, photons. In particular, b-tagging of jets
heavily relies on secondary vertices, for which the impact parameter is evaluated
from multiple intersecting tracks [152].

Validation of an alignment with residuals

Similarly to how residuals are used in a track-based alignment, they can also be
used to validate a determined alignment. In particular, one of the most sensitive
ways to detect misalignment or miscalibration of a certain parameter, is by using
a *distribution of the median of the residuals* (DMR). It is calculated for each single
module as a median of the residuals from all tracks that have crossed the module.
A distribution of such medians from a set of modules is referred to as DMR.

The width of such distribution is a measure of the hit-reconstruction resolution,
and its mean value should be 0. Any significant deviation of the mean value from 0
is a result of a systematic bias of some parameter that affects the reconstructed hit
positions.

5.3.2   Alignment methodology

The increase of track-hit residuals due to a misalignment of the tracker is the key fea-
ture that allows to determine a geometry that is as close as possible to the true one.
Considering that residuals are functions of global parameters $\mathbf{p}$ of the tracker, they
are determined by minimising the sum of squares of normalised track-hit residuals
from a large amount of tracks:

$$\chi^2(\mathbf{p}, \mathbf{q}) = \sum_{j}^{\text{tracks}} \sum_{i}^{\text{hits}} \left(\frac{m_{ij} - f_{ij}(\mathbf{p}, \mathbf{q}_j)}{\sigma_{ij}} \right)^2 , \qquad (5.2)$$

where σ_{ij} is the uncertainty of the measured hit position m_{ij}.

Several algorithms were developed at CMS to minimise the χ^2 function (5.2),
which differ by their complexity and capabilities. Their brief overview can be found
in Appendix A.1. All the improvements to the tracker-alignment procedure, which
are described in this thesis, were implemented in the framework of Millepede II
alignment algorithm. Thus, the main concept of Millepede II is briefly discussed
in this section.

Global-parameter corrections

Based on the design geometry, hardware-based alignment measurements and results
of previous track-based alignments, an approximate description of the geometry $\mathbf{p}_0$
is available. This starting geometry is also used to determine approximate track
parameters $\mathbf{q}_{0j}$. Corrections to track parameters due to difference between the real
and starting geometries are assumed to be small, such that f_{ij} can be linearised
around the initial values. If corrections are not small, several iterations of the
alignment procedure have to be performed. Thus, in the Millepede II algorithm,

the χ^2-function (5.2) is expressed in terms of global-parameter corrections $\delta\mathbf{p}$ with respect to the starting parameters $\mathbf{p}_0$:

$$\chi^2(\mathbf{p},\mathbf{q}) \simeq \sum_j^{\text{tracks}} \sum_i^{\text{hits}} \frac{1}{\sigma_{ij}^2} \left(m_{ij} - \left[f_{ij}(\mathbf{p}_0,\mathbf{q}_{0j}) + \frac{\partial f_{ij}}{\partial \mathbf{p}}\boldsymbol{\Delta}\mathbf{p} + \frac{\partial f_{ij}}{\partial \mathbf{q}_j}\boldsymbol{\Delta}\mathbf{q}_j \right] \right)^2 . \qquad (5.3)$$

The minimisation of the χ^2-function leads to a system of linear equations of least squares:

$$\mathbf{Ca} = \mathbf{b}, \text{ where } \mathbf{a}^{\mathrm{T}} = (\boldsymbol{\Delta}\mathbf{p},\boldsymbol{\Delta}\mathbf{q}) . \qquad (5.4)$$

Using the special structure of the matrix C, shown in Figure 5.4, the large system of equations (5.4) is reduced to a smaller system only for global parameters:

$$\mathbf{C}'\boldsymbol{\Delta}\mathbf{p} = \mathbf{b}' , \qquad (5.5)$$

where $\mathbf{C}'$ and $\mathbf{b}'$ sum contributions from all tracks used for alignment. Individual track-fit equations $\mathbf{C}_j\boldsymbol{\Delta}\mathbf{q}_j = \mathbf{b}_j$ have to be solved in order to derive $\mathbf{b}'$. A dramatic reduction of the matrix size from $\mathbf{C}$ to $\mathbf{C}'$ is one of the key features of the MILLEPEDE II approach [153].
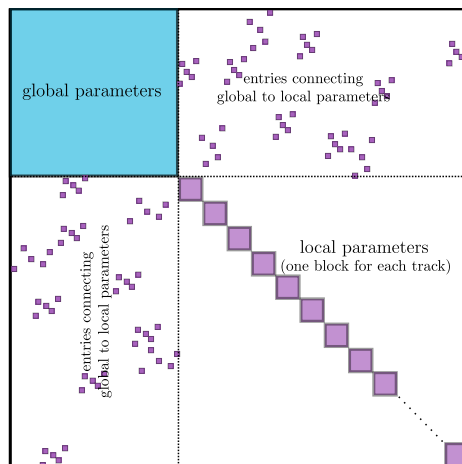

Figure 5.4: Structure of the matrix of the equations (5.5), as used in MILLEPEDE II alignment. Local fits are solved individually, and the global-parameter matrix is updated accordingly. The matrix equation for the global-parameter matrix is solved by the iterative MINRES algorithm. Taken from [154].

Global-parameter constraints

The CMS tracker is built in a hierarchical way from smaller mechanical substructures, e.g. three BPIX half-layers form each of the two BPIX half-shells. To treat translations and rotations of these substructures as a whole, six alignment parameters $\boldsymbol{\Delta}\mathbf{p}_l$ for each of the considered substructures can be introduced. These parameters allow to align complete substructures instead of individual modules if the track sample is too small, or to align them simultaneously with individual modules to take into account coherent movements of large structures. This hierarchical approach introduces redundant degrees of freedom, since movements of the large

structures can be expressed either by their alignment parameters or by combinations of the parameters of their components. In order to eliminate these degrees of freedom, alignment parameters $\mathbf{\Delta p}$ must fulfil certain linear equality constraints. These constraints ensure that translations or rotations of a large structure are equal to the sum of translations or rotations of its components [153].

Weak modes

Certain coherent changes of alignment parameters $\mathbf{\Delta p}$ can be compensated by changes of tracks parameters $\mathbf{\Delta q}$. Such changes are referred to as *weak modes*, since they do not change the track-hit residuals and can not be avoided without additional information included in equation (5.3).

In particular, $Z^0 \to \mu^+\mu^-$ decays are used to replace the two independent muon-track parameterizations by a two-body-decay parameterization with 9 parameters: position of the decay vertex, the momentum of the resonance candidate, two angles defining the direction of the decay products in the rest-frame of the resonance, and the mass of the resonance. Furthermore, this implies a vertex constraint as well, since the coordinates of the decay vertex are parameters of the combined fit object [153]. Such virtual measurements allow to avoid the so-called *twist deformation*, which represents the systematic distortion of the geometry by rotating modules at fixed global z coordinate around the beam line proportionally to their longitudinal position: $\Delta\phi \propto z$, as shown in Figure 5.5.
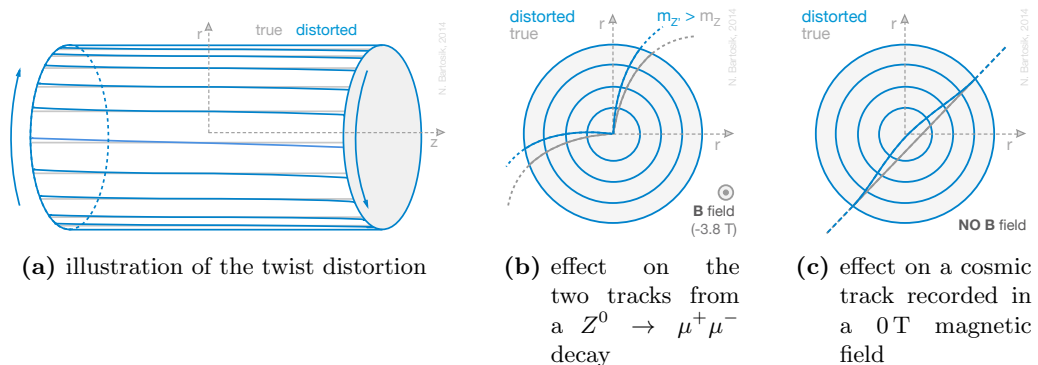

(a) illustration of the twist distortion

(b) effect on the two tracks from a $Z^0 \to \mu^+\mu^-$ decay

(c) effect on a cosmic track recorded in a $0\,\mathrm{T}$ magnetic field

Figure 5.5: Effect of the twist distortion of the tracker geometry on different types of tracks. Track-hit residuals of $0\,\mathrm{T}$ cosmic tracks and of muon tracks from $Z^0 \to \mu^+\mu^-$ decays increase due to such a distortion, while other track topologies are not sensitive to the twist distortion.

An additional type of tracks that allow to correlate relative movements of distant modules are tracks from cosmic muons. Unlike tracks from the collision point, which originate from the centre of the detector, cosmic tracks penetrate the detector from outside and can leave hits in two opposite parts of the tracker. Such tracks are especially important to avoid the so-called *telescope weak mode*, which is characterised by the shift of modules along the z axis proportionally to their distance from the beam line: $\Delta z \propto r$, as shown in Figure 5.6. Parameters of collision tracks weakly depend on such deformations, since they only cross modules in a single half of the tracker. In contrast, cosmic tracks can penetrate both parts, and the hits at opposite radii will not fit to the predicted trajectory if such a distortion is present.
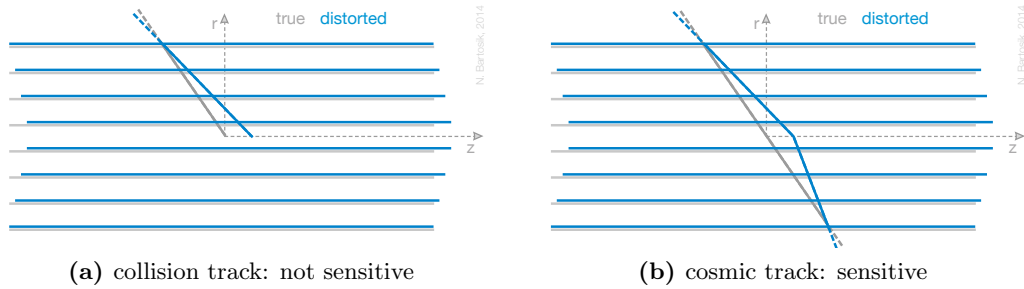
(a) collision track: not sensitive   **(b)** cosmic track: sensitive

Figure 5.6: Effect of the telescope distortion of the tracker geometry on collision and
cosmic tracks. Track-hit residuals of collision tracks do not increase due
to such a distortion. Residuals of cosmic tracks do increase, because they
have to be almost straight in the rz plane, while the telescope distortion
introduces a kink between the two parts of the track.

Cosmic tracks recorded without magnetic field are almost straight, with minor
deviations from a straight line due to multiple scattering. Track-hit residuals of such
tracks would also increase in case of the twist deformation, as shown in Figure 5.5c.

5.3.3   Technical implementation

In order to linearise the alignment problem, the dependence of the track prediction
$f_j(\mathbf{q}_j, \mathbf{p})$ on $\mathbf{q}_j$ and $\mathbf{p}$ has to be known.

Track parameterization

In general, a trajectory of a charged particle in a homogeneous magnetic field can
be parameterized by a helix, which is defined by five parameters, as listed in equa-
tion (4.12). But in reality, particles interact with the material in the tracker (see
Figure 4.14), undergoing multiple scatterings.

These scatterings lead to incorrect residuals if a helix parameterization is used,
and have to be taken into account during the track fitting process. Thus, multiple-
scattering treatment is achieved by adding two thin scatterers for each scattering,
which increases the number of track parameters: $n_{\mathrm{par}} = 5 + 2 \cdot n_{\mathrm{scat}}$, where n_{scat} is the
number of virtual thin scatterers penetrated by the particle, as shown in Figure 5.7.
This is implemented in MILLEPEDE II by the General Broken Lines track refit [155].
More details about the track parameterization can be found in [153]. In the CMS
tracker, the material is concentrated at the sensor planes, while cabling, cooling
pipes, electronic circuitry is considered as part of the sensor.

Alignment parameterization

In principle six rigid-body parameters are defined for a module, i.e. three shifts
(u, v, w) and three rotations (α, β, γ), as shown in Figure 5.1. But $r\phi$ strip modules
have only five degrees of freedom, because they are not sensitive to shifts along the
v axis, which is directed parallel to the strips.

The level of precision achieved by the MILLEPEDE II alignment is so high that
even deviations of sensor-surface shapes from a flat surface, as assumed by the
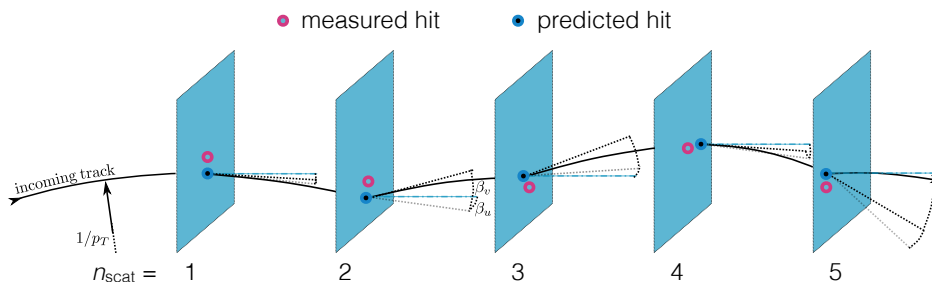
Figure 5.7: Schematic view of the treatment of virtual scatterers in the Broken Lines
trajectory model. Extended version of the sketch from [154].

CMS track reconstruction, become relevant. Therefore, instead of six rigid-body
parameters per module, nine degrees of freedom per sensor are used.

To relate track-hit residuals with corrections to global parameters ($\mathbf{\Delta p}$) and with
predicted hit positions (f_{ij}), 9 relevant derivatives have to be defined for the Δu
and Δv corrections of the predicted hit position. These derivatives depend on the
measured hit positions in the local coordinate frame of the module (u, v) and the
track angles (ψ, θ), which are illustrated in Figure 5.8. The exact 18 derivatives (9
parameters $\times$ 2 coordinates) can be found in [153].
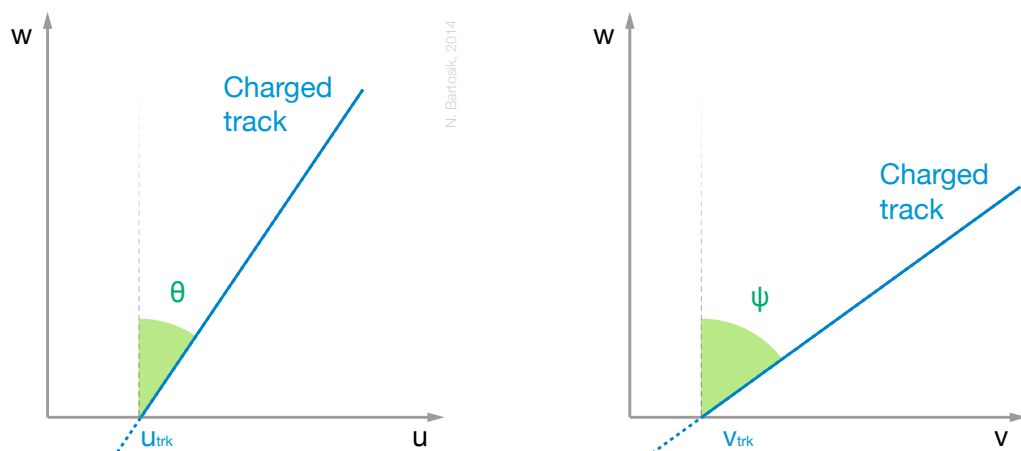
Figure 5.8: Track angles measured from the normal to the sensor surface in the local
coordinate system. Names of the two angles that are used internally for the
alignment are swapped with respect to the ones defined in this figure. Based
on sketches from [153].

Differential alignment

The tracker can undergo significant changes of conditions that can lead to the change
of its geometry. Such changes include temperature variations, which can happen
during the opening of the detector for repairs, or magnetic-field variations when
reenergising the superconducting solenoid. The mechanical structure of the CMS
tracker makes it very unlikely that positions of individual modules change relative
to each other, but large structures, i.e. half-barrels and endcaps of the strip tracker,
BPIX layers and FPIX half-disks in principle can shift or rotate independently.

Thus, all the module-level alignables are considered to be fixed in time, so that a single value is determined for each module-level global parameter. In contrast, several parameters corresponding to different periods of time are assigned to the larger structures that can move. Such periods are referred to as *intervals of validity* (IOV), and are defined by starting and ending run numbers, such that a sufficient amount of tracks of different topologies is present in each IOV. The initial fine IOV splitting is analysed to determine run ranges during which no significant movements can be seen. The adjacent IOVs without significant movements are then merged to maximise the amount of tracks per IOV.

Input tracks for alignment

Tracks are the key component of the track-based alignment, since they are used as a reference for the residual calculation. In order to have reliable alignment, only high-quality tracks and high quality hits are used for the fit. To overcome the weak-mode problem, tracks of different topologies are used from the four dedicated datasets: isolated muons, minimum bias events, $Z^0 \to \mu^+\mu^-$ decays, cosmic rays. Some quality criteria are common, while some are specific to tracks from particular datasets. An overview of the selection criteria can be found in Appendix A.2.

The actual sequence in which the alignment is performed, consists of two separate steps with individual technical features, which are described in Appendix A.4.

5.4　Lorentz-angle calibration

The Lorentz force is a crucial physical phenomenon that allows to estimate the momentum of charged particles traversing the tracker volume. According to the laws of physics, a single elementary charge q that travels in a magnetic field $\mathbf{B}$ with the velocity $\mathbf{v}$ will undergo the Lorentz force ($\mathbf{F}_\mathrm{L}$), which is orthogonal to both the directions of movement and of the magnetic field:

$$\mathbf{F}_\mathrm{L} = q\left(\mathbf{v} \times \mathbf{B}\right) . \tag{5.6}$$

This force leads to a bending of tracks in the plane perpendicular to the direction of the magnetic field of the superconducting solenoid, which allows to use the curvature of the track for the momentum estimate of the corresponding charged particle.

5.4.1　Lorentz-angle effect

The Lorentz force has also a side effect on charge carriers in the silicon sensors of the tracker, which is also known as *Hall effect*. According to it, charge carriers that travel through the silicon sensor perpendicularly to the direction of the magnetic field are deflected by some angle with respect to the electric field, which is referred to as *Lorentz angle* (θ_LA). Electrons and holes are not deflected by the same amount, because their drift velocities are different, as illustrated in Figure 5.9. In the silicon sensors of the CMS tracker, holes are the charge carriers that are detected by pixels or strips, therefore, later in the text, *Lorentz-angle effect* refers explicitly to the effect on holes.

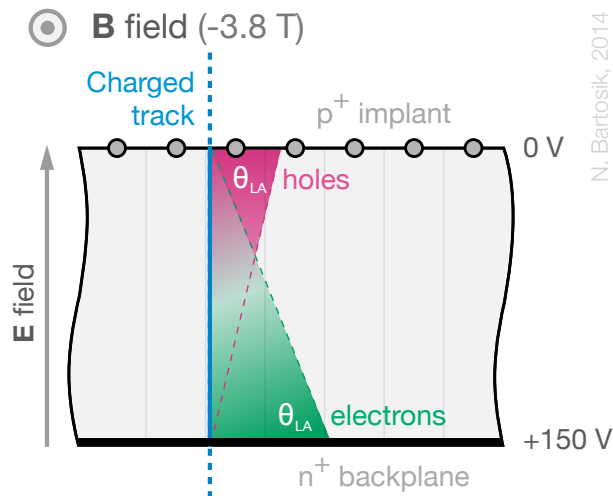Considering the drift of charge carriers, the Lorentz-angle effect has two main consequences:

Figure 5.9: Illustration of different deflection angles of electrons and holes due to the
Lorentz force in the material of a silicon sensor. The charge-carrier velocity
is constant across the whole sensor leading to a linear drift. In the CMS
tracker, holes are collected by the readout electronics in both the pixel and
microstrip detectors.

- degraded hit-position resolution due to the increased cluster size;

- systematic shift of the hit positions in the direction of the Lorentz angle due
  to the extended cluster width only in one direction.

This effect is relevant only for sensors where the electric-field direction is not par-
allel to the magnetic-field direction, since otherwise the drift direction lies in the
magnetic-field plane. Thus, it has to be measured most precisely for modules in the
barrel region (BPIX, TIB, TOB). In the FPIX, the effect is much smaller, although
not negligible, since its modules are arranged in a manner similar to the blades in
a turbine, with a small tilt around the v axis. In TID and TEC, no Lorentz-angle
effect is present, since microstrip modules in endcap regions are orthogonal to the
direction of the z axis.

In general, the hit position in a single module is obtained from a charge-weighted
average position of the cluster. During the track reconstruction, the Lorentz angle
has to be taken into account to properly estimate the hit position from the measured
cluster.

5.4.2   Standalone measurements of Lorentz angle

There are different ways to measure the Lorentz angle, which can be either hardware
based or track based, similarly to the tracker alignment.

Hardware-based measurement

A dedicated measurement of the Lorentz angle was performed in the test beam for
irradiated pixel sensors under different bias voltages [156]. The measurement shows
that θ_{LA} varies in the 4 T magnetic from 26° for an unirradiated sensor to 8° for

a highly irradiated sensor due to a required increase of bias voltage from $150\,\mathrm{V}$ to $600\,\mathrm{V}$ For a $285\,\mathrm{\mu m}$ thick pixel module this value corresponds to a shift of the charge center by up to $70\,\mathrm{\mu m}$.

Such measurements provide a good estimate of the starting conditions and allow to test simulations of the Lorentz-angle dependence on different factors. But it can not reflect all the varying conditions inside the CMS tracker during its lifetime. In particular, the Lorentz angle of a single module depends on the accumulated irradiation dose, its temperature, the applied bias voltage, etc. All these conditions are not uniform across the tracker, and their development with time is not known precisely enough to have optimal predictions of the Lorentz angle for every module during its whole operation time.

Track-based calibration

Similar to the track-based alignment of the tracker, a much better precision can be obtained by using reconstructed tracks to determine the Lorentz angle. Furthermore, it can be determined for subsequent periods of time to monitor the evolution of θ_{LA} in different parts of the detector. Such results are achieved by the so-called *grazing angle* method [157, 158], which was used to calibrate the Lorentz angle in BPIX as a function of integrated luminosity, covering the whole 2012 data-taking period [159].

In general, the idea of the grazing-angle method is to obtain a distribution of the charge-drift distance d as a function of the depth z, at which the charge drifting was initiated by the passing track. These quantities can be calculated as described in [158]:

$$
\begin{aligned}
d &= \Delta x - \Delta y \cdot \tan\gamma\,, \\
z &= \Delta y - \tan\beta\,.
\end{aligned}
\tag{5.7}
$$

The actual angles used in (5.7) are schematically shown in Figure 5.10a. θ_{LA} is equal to the slope of this distribution, and is obtained from the linear fit of the spectrum, as shown in Figure 5.10b. The obtained distributions of the Lorentz-angle evolution in each of the three layers (at different r) in BPIX, as well as a comparison of θ_{LA} in different rings (at different z) of each layer, are shown in Figure 5.11.

A clear time dependence is visible from the calibration, which is different in the three different layers of BPIX. The determined θ_{LA} increases with time due an increase of irradiation dose, and changes most rapidly in layer, which is closest to the interaction point. Within a single layer, an offset is visible between the two halves of BPIX along the z axis, which could not be explained up to now.

In microstrip modules, a slightly different method is used to measure Lorentz angle [160]. It is based on the fact that tracks, crossing a module at the incident angle that is equal to the Lorentz angle, produce clusters of the smallest width, as illustrated in Figure 4.18. Thus, using a large amount of cosmic tracks, cluster width is scanned across a range of track angles, and θ_{LA} is determined as the track angle leading to the smallest cluster width, as shown in Figure 5.12.
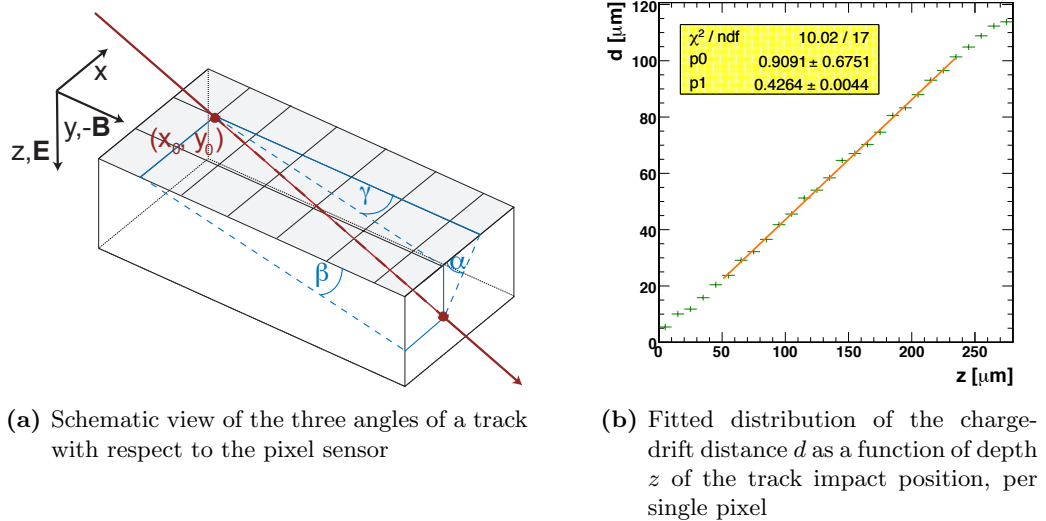
(a) Schematic view of the three angles of a track with respect to the pixel sensor

(b) Fitted distribution of the charge-drift distance d as a function of depth z of the track impact position, per single pixel

Figure 5.10: Ingredients of the grazing-angle method for track-based calibration of Lorentz angle in BPIX [158].

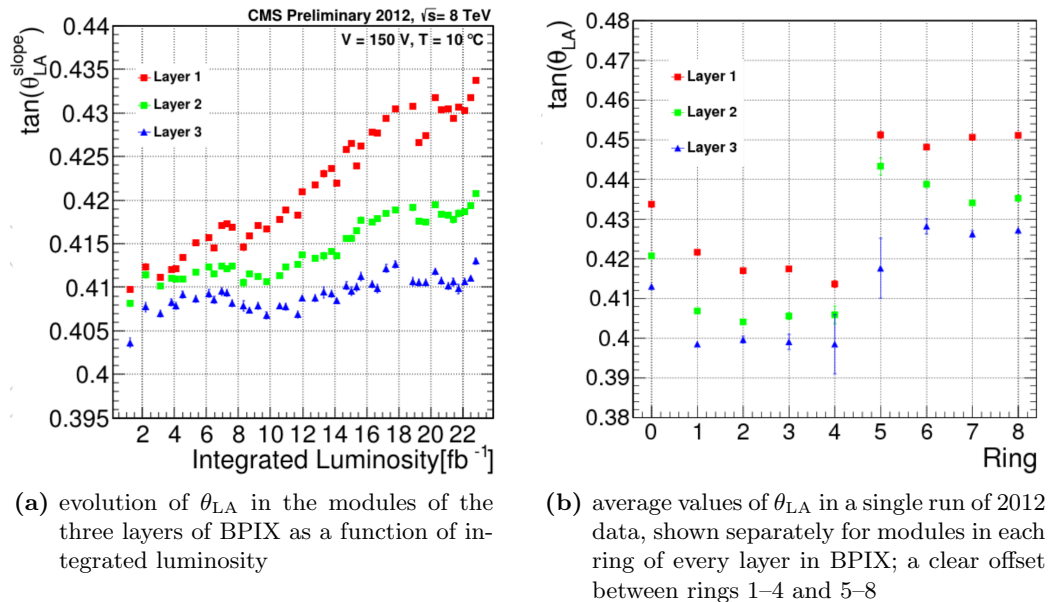**(a)** evolution of θ_{LA} in the modules of the three layers of BPIX as a function of integrated luminosity

(b) average values of θ_{LA} in a single run of 2012 data, shown separately for modules in each ring of every layer in BPIX; a clear offset between rings 1–4 and 5–8

Figure 5.11: Lorentz-angle values for BPIX modules, calibrated with the grazing-angle method, using tracks recorded during the 2012 data-taking period [159]. The measured $\tan(\theta_{LA})$ varies from 0.405 to 0.435 during the full 2012 data-taking period.

(a) $B = 0\,\mathrm{T}$ **(b)** $B = 3.8\,\mathrm{T}$
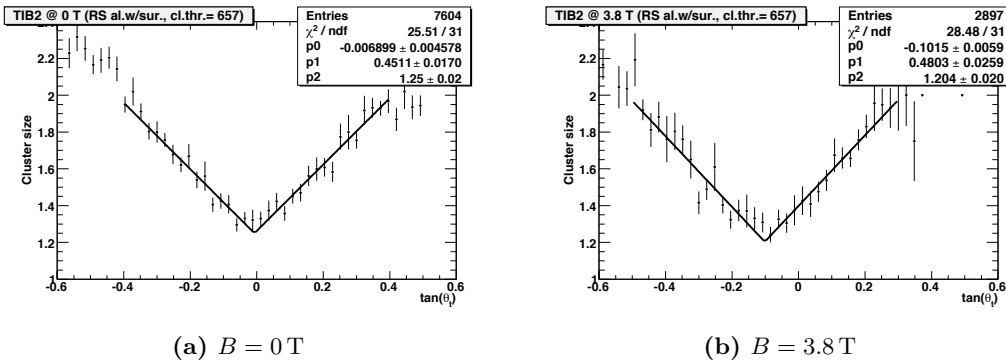
Figure 5.12: Cluster size versus $\tan(\theta_{\mathrm{track}})$ in TIB modules of the second layer. The dependence determined from cosmic tracks recorded at different strengths of magnetic field (B) [160]. At $B = 0\,\mathrm{T}$, the smallest cluster width is observed when tracks cross modules at normal angle, which reflects the expected behaviour without Lorentz-angle effect. The measured tangent of Lorentz angle is $\tan(\theta_{\mathrm{LA}}) = -0.102$.

5.4.3   Simultaneous alignment and Lorentz-angle calibration

While the grazing-angle method allows to reach high precision of the Lorentz-angle calibration with possibility to determine it for separate groups of modules and monitor changes with time, it still has a weak point. Tracks that are used to obtain the $d \propto z$ distribution, are reconstructed based on fixed assumptions about the tracker geometry from track-based alignment. Thus, possible misalignment can lead to a loss of resolution or a bias of θ_{LA} values calibrated with this method.

Similarly, fixed assumptions of the Lorentz angle from standalone measurements are used for track fits during the MILLEPEDE II alignment, which can also affect the quality of the aligned geometry. Therefore, this correlation between the Lorentz angle and aligned module positions makes it very attractive to combine the two, namely integrate the Lorentz-angle calibration into the track-based MILLEPEDE II alignment.

Constraining module positions

The drift of charge carriers under the Lorentz force leads to a shift of the measured hit positions in a module, as mentioned in Section 5.4.1. Since the track-based alignment tries to determine a geometry such that track-hit residuals are minimised, it can correct the module positions by an average shift due to the wrong θ_{LA}, compensating the possible miscalibration. Nevertheless, such a compensation would not be optimal, since it is known that θ_{LA} changes with accumulated irradiation dose, due to the reduced resistance of the silicon, and increased bias voltage that is required to fully deplete the sensor. But the determined positions of the tracker modules are stable in time, as mentioned in Section 5.3.3, making it impossible for the module-level alignment to compensate a changing Lorentz angle during a long time period.

In order to disentangle the Lorentz-angle miscalibration and module alignment, an orthogonal source of information is needed. The nature of the Lorentz-angle effect

directly suggests that tracks recorded in different magnetic-fields could provide such
an information, since the positions of the modules do not depend on the magnetic
field, while the Lorentz-angle effect does. The effect of the magnetic field strength
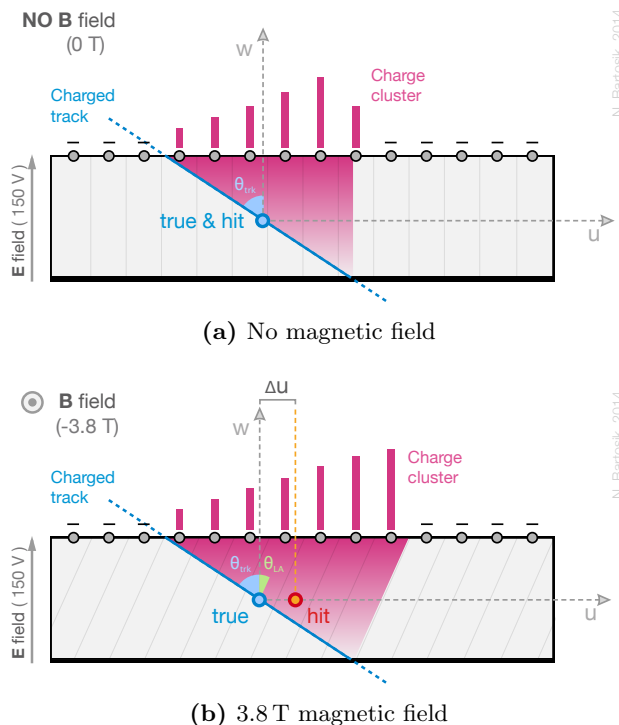on hit-position measurements is illustrated in Figure 5.13.

(a) No magnetic field

(b) 3.8 T magnetic field

Figure 5.13: Measured and true hit positions in magnetic fields of different strength.
The position of the hit is determined as a position of the barycentre of the
charge cluster.

During some periods of time, the superconducting solenoid is not energised,
leading to no magnetic field in the tracker. This usually happens during dedicated
cosmic runs, when no *pp* collisions take place, and only cosmic tracks are recorded.
Thus, cosmic data recorded with the magnet switched on is referred to as *3.8T
Cosmics*, and with the magnet switched off is called *0T Cosmics*.

The effect of compensation of Lorentz angle miscalibration by shifts of the mod-
ules is illustrated by splits between DMRs (see Section 5.3.1) of modules with op-
posite orientations of their *w* axis, as shown in Figure 5.14. The idea behind this
validation is that in modules of opposite orientations, measured hit positions are
shifted by the Lorentz force in opposite directions, which leads to the corresponding
shift of DMRs. Thus, the difference between mean values of the DMRs from modules
with different orientations is a good measure of Lorentz-angle miscalibration or the
Lorentz-angle-related bias of the tracker geometry. It can be seen from Figure 5.14,
that the alignment compensates for Lorentz-angle miscalibration, since the DMR
difference due to a wrong Lorentz angle in the correct geometry is the same as due
to a biased geometry with no Lorentz-angle effect.

It was mentioned earlier in Section 5.3.2 that 0T Cosmic tracks can help to
avoid the twist weak mode. But as demonstrated in Figure 5.14, tracks recorded
in different magnetic fields shift modules in opposite directions. Thus, without the
proper treatment of the Lorentz-angle effect in alignment, simultaneous usage of the
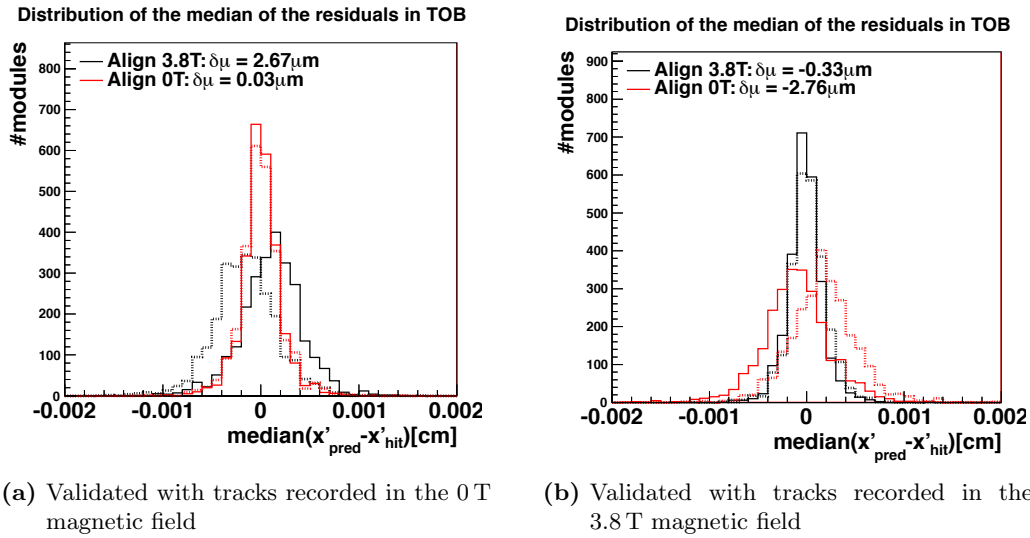
(a) Validated with tracks recorded in the 0 T magnetic field

(b) Validated with tracks recorded in the 3.8 T magnetic field

Figure 5.14: Distributions of the median of the residuals per BPIX module, obtained using tracks recorded in different magnetic fields during 2010. The DMRs were determined separately for modules in even (solid) and in odd (dashed) rods, which have opposite orientations of the modules. The two compared alignments are determined using tracks only from 3.8 T Cosmics (black) or 0 T Cosmics (red) datasets. In both alignments, no Lorentz-angle calibration was performed. The splits between DMRs of oppositely-oriented modules are caused by the module-level alignment systematically shifting modules, depending on their global orientation, to compensate for the remaining Lorentz-angle miscalibration (a) or due to the miscalibrated Lorentz-angle values applied during track reconstruction (b). Statistically independent sets of tracks were used for the alignment and for the validation. $\delta\mu$ denotes the difference between the mean values of the two DMRs that represent modules with opposite orientations.

both track topologies would create a tension between them, making it impossible to
fully compensate the effect by shifts of the modules.

Lorentz-angle parameterization

A shift of the measured hit position in the u direction (Δu) due to the Lorentz-angle
effect can be expressed in terms of the effective sensor thickness (d') and Lorentz
angle (θ_{LA}) as:

$$\Delta u = \frac{d'}{2} \tan(\theta_{\mathrm{LA}}) \,. \tag{5.8}$$

The effective thickness corresponds to the thickness (d) of the sensor's active zone,
when operating in peak readout mode:

$$d'_{\mathrm{peak}} = d \,. \tag{5.9}$$

In other words, it is the largest distance that a charge carrier can travel in the
normal direction to the sensor's surface. But in a sensor operating in deconvolution
mode, a partial loss of signal from the backplane region of the sensor reduces the
effective length that is traveled by measured charge carriers (see Figure 4.19). Thus,
in deconvolution mode, the effective sensor thickness is smaller by the width of the
backplane zone (d_{bp}):

$$d'_{\mathrm{deco}} = d - d_{bp} \,. \tag{5.10}$$

More details about the width of the backplane zone will be given in Section 5.5.

Since Lorentz angle depends on the strength of magnetic field $|\mathbf{B}|$, the quantity
that is usually calibrated is the *Hall mobility* (μ_{H}), which relates the two as [157]:

$$\tan(\theta_{\mathrm{LA}}) = \mu_{\mathrm{H}} \cdot |\mathbf{B}| \,. \tag{5.11}$$

Values of μ_{H} are stored into a database and are used during event reconstruction.
This is especially useful considering that the magnetic field is not ideally uniform
across the tracker volume. Hence, using the stored μ_{H} values and a map of the
magnetic-field strength, an exact value of θ_{LA} can be calculated for every single
tracker module.

Considering the explicit relation between the shift of the hit and μ_{H}, the cali-
bration of μ_{H} can be integrated into the MILLEPEDE II alignment procedure using
an additional set of global parameters. Similarly to other alignment parameters dis-
cussed in Section 5.3.3, derivatives have to be defined in the local coordinate system
of a module, and have to relate a variation of μ_{H} with the resulting change of the
track-hit residual, assuming that other global parameters do not change. Thus, the
derivative of a track-hit residual track for Lorentz angle is defined as follows:

$$\frac{\partial(\Delta u)}{\partial \mu_{\mathrm{H}}} = -B_v \cdot \frac{d - d_{bp}}{2} \,, \tag{5.12}$$

where $\mathbf{B}_v$ is the projection of magnetic field on the v axis, d represents the thickness
of the sensor's active volume, and d_{bp} corresponds to the width of the backplane zone
($d_{bp} = 0$ for pixel sensors and microstrip sensors in peak mode). In order to simplify
the dependence on d_{bp}, two separate sets of θ_{LA} values are determined for microstrip
modules: one for peak mode and one for deconvolution mode. The parameteriza-
tion (5.12) is correct only for modules in the barrel region, while for FPIX modules,

which have non-negligible B_w component, more complex derivatives should be used, as described in Appendix A.5. Nevertheless, only the derivative (5.12) was used in MILLEPEDE II for all tracker modules, since the mentioned second-order effects are not very significant. Furthermore, additional work required to implement these correction could not be completed within the time schedule of this thesis.

Conceptual feature of the combined calibration

It is worth mentioning that the θ_{LA} determined with this combined approach is not necessarily the same quantity as determined by the grazing-angle method or by hardware-based measurements. The purpose of standalone measurements is to determine the physical parameter, which should be used in a model that predicts the resulting hit position. In contrast, the combined alignment and calibration aim at direct determination of the effect on measured hit positions, which may include some other unknown effects that have similar behaviour. Therefore, results of the combined approach may differ slightly from the results of standalone measurement. But in this particular case, the aim is to achieve the best performance of the tracker-hit reconstruction.

Technical implementation

The code for Lorentz-angle calibration was implemented by Gero Flucke as a dedicated `SiPixelLorentzAngleCalibration` class of the `CommonAlignmentAlgorithm` package in the official CMS software framework. In the alignment procedure, input values of Hall mobility are taken from a database and used for track fit. The calibrated mobility values obtained from MILLEPEDE II alignment are stored in the same format as the ones stored in a database. This allows to use the determined μ_{H} values for further alignment iterations or for validation of the determined alignment in the same way as the official calibrations.

This approach introduced some complications for hit reconstruction in the pixel detector. There are two hit-reconstruction methods that can be used in the pixel tracker [161]:

- **generic:** the hit positions are reconstructed directly from the digitised output of the pixel modules. Values of μ_{H} are explicitly used to reconstruct hit positions based on the corresponding Lorentz drift;

- **template:** a set of templates of charge patterns is generated for various conditions, including irradiation effects, different track incident angles, etc. The hit position is obtained from the template that matches best the charge-cluster profile. Information about μ_{H} is implicitly included in the generated templates, but not explicitly accessible during hit reconstruction by default.

The template hit reconstruction is used by default during data taking, because it provides a much better position resolution, allowing to determine it with the precision of up to 1/8-th of a single-pixel width. In order to obtain a calibrated Lorentz-angle value, the input value has to be clearly defined and constant in time, which is not the case in pixel templates. Thus, a method was required for subtracting the assumed μ_{H} value in the template from the constant input value from a database, and

to correct the template-reconstructed hit position by the remaining Lorentz-angle effect. This task was addressed by a special patch to the hit-reconstruction code by its developers. It allowed to have a constant reference input value, with respect to which a correction is determined, even with the template hit reconstruction.

Since μ_{H} is not expected to have significant position-dependent variations, there is no sense in determining it separately for each single module. Furthermore, such approach would require a very large amount of tracks to obtain sufficient statistical precision. Therefore, modules are organised in groups of modules that are expected to have very similar Lorentz-angle values, as well as the same time evolution. An arrangement of groups of modules, for which a single value of a global parameter is determined, is referred to as *spatial granularity*. A sequence of time intervals (IOVs), during which a particular global parameter is assumed to be constant is referred to as its *time granularity*. Thanks to the development of the `TkModuleGroupSelector` class by Jörg Behr, spatial and time granularities for any global parameters can be defined independently. This feature was heavily used for numerous runs of combined Lorentz-angle calibration and tracker alignment with different configurations to find the optimal one.

5.5   Backplane correction

In addition to the Lorentz-angle effect described in Section 5.4, the backplane effect also significantly affects measured hit positions in the microstrip detector, as was briefly introduced in Section 4.3.2. More details about the effect itself and about its calibration method are given in the following sections.

5.5.1   Backplane effect

The very short time between consecutive proton-bunch crossings of 50 ns places an intrinsic upper limit on the distance (L_{max}) that a hole can travel through the silicon volume such that it is registered by the readout electronics. This limit for a single hole translates into a reduction of the thickness (d) of the effective active zone by d_{bp}:

$$d_{bp} = d - L_{\mathrm{max}} \,. \tag{5.13}$$

Furthermore, when the magnetic field is non-zero, the width of the backplane zone is larger due to the projection of L_{max} on the w axis reduced by the Lorentz-angle effect:

$$d_{bp} = d - L_{\mathrm{max}} \cdot \cos(\theta_{\mathrm{LA}}) \,. \tag{5.14}$$

The dependence of the measured hit position on the magnetic field and readout mode of a microstrip sensor is illustrated in Figure 5.15.

The loss of signal from the backplane zone affects the measured hit position, since the barycenter of the measured cluster is shifted from the middle of the sensor. Thus, in order to compensate this effect, a *backplane correction* (Δw) is defined for the strip tracker, such that:

$$d_{bp} = \Delta w \cdot d \,, \tag{5.15}$$

which corresponds to the fraction of the sensor's thinkness d, from which signal is lost. This correction is used during the hit reconstruction to properly estimate its

(a) No magnetic field
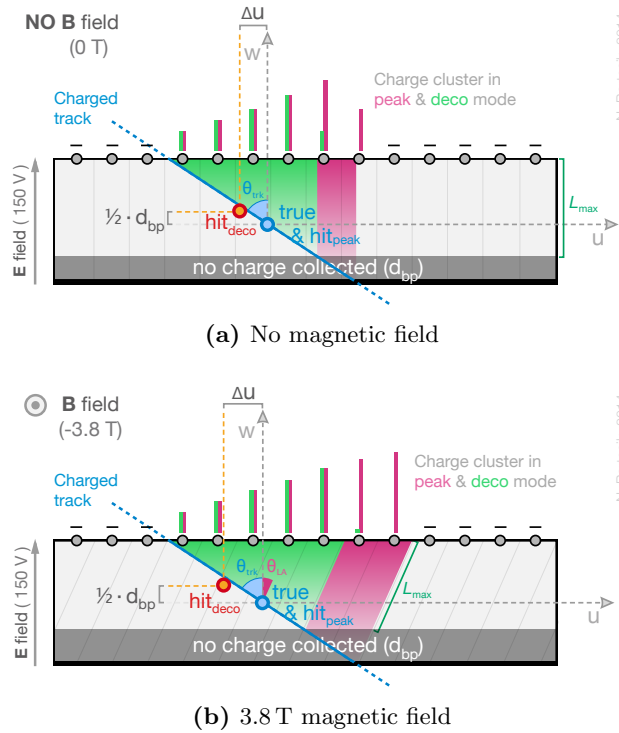

(b) 3.8 T magnetic field

Figure 5.15: Measured and true hit positions in a microstrip sensor in magnetic fields of different strength. Compared are drift direction of charge carriers in peak and deconvolution modes. The hit position is determined as a position of the barycenter of the charge cluster.

real position with only a fraction of charge being measured. Furthermore, its effect also has an impact on the Lorentz-induced shift, since the width of the backplane zone enters into its formula (5.12). The effect on μ_H values is addressed in the current implementation of MILLEPEDE II by calibrating the Lorentz angle separately for peak and deconvolution modes. Since the drift-velocity of charge carriers can change due to irradiation damage, the backplane correction can in principle vary with time. Therefore, its time dependence is also estimated.

5.5.2 Existing measurements of the backplane effect

The width of the backplane zone was measured using separate data samples from 2011, collected in deconvolution and in peak mode [162]. For ech of the two track samples, a dependence of the track-hit residual on the track angle θ_{trk} with respect to the Lorentz angle θ_{LA} was measured. The difference between the two results is used to determine the bias of the measured u coordinate of the hit position and of θ_{LA} correction due to the backplane effect. The difference between the track-hit residuals in the u direction, as obtained from tracks collected in the different readout modes, is shown in Figure 5.16 [163].

5.5.3 Combined alignment and backplane-shift calibration

The mentioned correlation between the reconstructed hit position, applied backplane correction and Lorentz angle makes it a natural decision to include all of them as

(a) track-hit residuals in u for peak and deconvolution readout modes

(b) fit of the difference between residuals from deconvolution and peak data
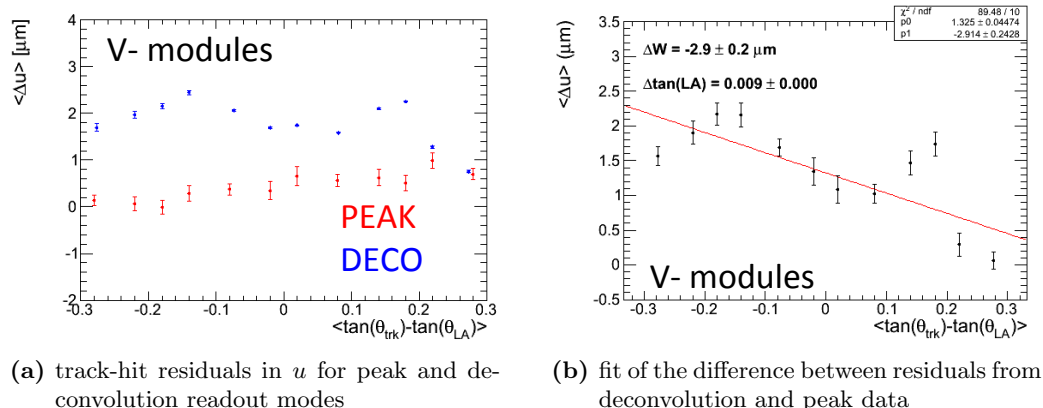
Figure 5.16: Results of standalone calibration of the backplane correction for modules of TIB as a function of track incident angle with respect to θ_{LA} [163]. "$V-$" corresponds to a subset of TIB modules of a certain orientation. The determined backplane correction is about $2.9\,\mu\mathrm{m}$, which is $1.8\,\%$ of the sensor's active-area thickness. Its effect on the Lorentz angle in deconvolution mode is $\Delta\tan(\theta_{\mathrm{LA}}) = 0.009$, which is about $8\,\%$ of the value in peak mode.

global parameters in the MILLEPEDE II alignment procedure. In order to differentiate the effect of a wrong backplane calibration from a wrong module alignment or a wrong Lorentz-angle calibration, tracks recorded in peak and deconvolution modes have to be used in the same alignment. This would ensure that module positions and μ_{H} values are constrained by tracks recorded in peak mode, where no backplane correction is needed, and are not biased by the backplane calibration.

Parameterisation of the backplane correction

In the peak readout mode, no backplane correction exists, therefore $\Delta w_{\mathrm{peak}} = 0$, and the corresponding derivative for the peak mode is:

$$\frac{\partial(\Delta u)}{\partial \Delta w_{\mathrm{peak}}} = 0\,. \tag{5.16}$$

In the deconvolution mode, the following derivative is defined in MILLEPEDE II:

$$\frac{\partial(\Delta u)}{\partial \Delta w_{\mathrm{deco}}} = d \cdot \frac{\mu_{\mathrm{H}} \cdot \mathbf{B}_v + \tan(\theta_{\mathrm{trk}})}{2}\,, \tag{5.17}$$

where $\tan(\theta_{\mathrm{trk}})$ corresponds to the angle between the track prediction and the module's surface, as shown in Figure 5.15 or 5.1. The increase of d_{bp} due to a non-zero θ_{LA} (see Equation 5.14) is not included in the derivative (5.17). Its effect on Δw is expected to be smaller than $1\,\%$, but for a perfect treatment this effect should be included in the derivative.

Technical implementation

The code for backplane calibration was implemented by Gero Flucke as a dedicated `SiStripBackplaneCalibration` class of the `CommonAlignmentAlgorithm` package

in the official CMS software framework (`CMSSW`). Similarly to the Lorentz-angle calibration, input values of the backplane correction are taken from a database, and calibrated values are stored in an `SQLite` file in the same format as the official corrections.

5.6   Combined alignment-calibration setup

Using the added classes for the Lorentz-angle and backplane calibrations, a typical MILLEPEDE II alignment setup was extended to introduce new global parameters for the calibrations.

5.6.1   Starting conditions

The starting tracker geometry used for the combined alignment and calibration from 2012 data is taken from the last IOV of the alignment on 2011 data. Latest Lorentz-angle and backplane calibrations from standalone measurements were used, which have a single IOV for the whole period of 2012. More details about the starting configuration can be found in Appendix A.6.

5.6.2   Global-parameter configuration

For the combined alignment and calibration, different sets of global parameters were configured with varying spatial and time granularity.

Alignment

Three shifts and three rotations of large structures were determined for multiple periods of times (IOVs):

- **BPIX layers, FPIX half-disks:** 15 IOVs;

- **TIB half-barrels:** 8 IOVs;

- **TID endcaps:** 10 IOVs;

- **TEC endcaps:** 8 IOVs.

The same number of IOVs does not mean that the boundaries of IOVs are the same, since individual parts of the detector can experience significant movements in different periods of time.

For modules in BPIX, FPIX, TID and TEC, 3 positions and 3 rotations are determined, which are fixed in time. For modules in TOB and TIB, 2 positions and 3 rotations are determined; alignment is not sensitive to shifts of these modules along the strip direction v. Furthermore, surface-shape parameters are determined for all tracker modules.

Lorentz-angle calibration parameters

Since BPIX is closest to the interaction point, it is assumed to have the largest
variations of the Lorentz angle due to the varying particle flux in different regions
of BPIX. Thus, an individual Lorentz-angle global parameter is assigned to a set of
modules belonging to a single ring within a single layer, as illustrated in Figure 5.17.
This leads to 24 spatial Lorentz-angle parameters for BPIX (3 layers × 8 rings). Due
to the non-optimal parameterization of the Lorentz-drift in FPIX, no calibration was
performed for this subdetector. To determine the time dependence of the Lorentz
angle in BPIX, 65 IOVs were defined for it. The IOV boundaries were selected such
that integrated luminosity of data recorded by CMS during different IOVs varies by
not more than 20 %.

(a) Division into rings along the z axis

(b) Division into layers at separate values of r
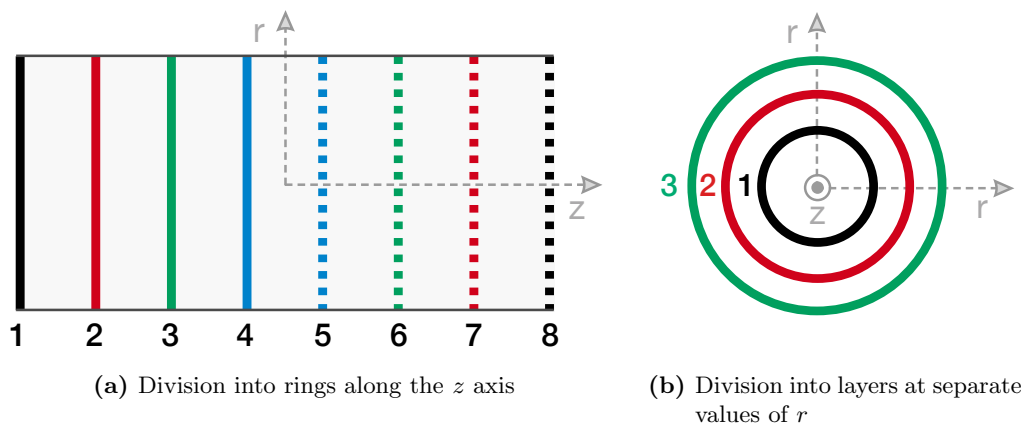
Figure 5.17: Sketch of spatial granularity of the Lorentz-angle calibration in BPIX. All
modules that belong to the same ring and the same layer share a common
Lorentz-angle global parameter.

Modules in TIB and TOB are arranged in 4 and 6 layers respectively, each
consisting of 12 groups of modules (12 rings) located at different z coordinates.
The first two layers in TIB and TOB consist of double-sided modules, which are
combinations of an $r\phi$ module and a stereo module, which is rotated along the w-
axis with respect to the $r\phi$ one. Due to a certain tension between the alignment and
calibrations, $r\phi$ and stereo modules are treated separately in calibrations. Thus,
an individual spatial Lorentz-angle parameter is assigned to modules belonging to
every ring of every layer of TIB and TOB, whereas the number of parameters is
doubled for the first two layers, in order to have separate parameters for $r\phi$ and
stereo modules. Studies with fine time granularity showed that variations of θ_{LA} in
microstrip modules with time are small. Therefore, the 2012 period is split only into
6 IOVs for the θ_{LA} in the deconvolution mode and into 2 IOVs for the peak mode.

Backplane calibration parameters

The backplane correction is determined only for microstrip modules. For modules
in TIB and TOB, the same spatial granularity is used as in the Lorentz-angle cal-
ibration. No backplane correction was determined for TID and TEC modules due
to technical difficulties in the proper application of the determined corrections in

software. No significant variations of the backplane correction with time were seen, thus 2 IOVs were defined for the whole 2012 period.

Summary of global calibration parameters

A summary of calibration parameters for different subdetectors, which are used in the full-scale alignment, is given in Table 5.1. In total, 3305 global calibration parameters are added to the standard set of about $200\,000$ alignment parameters. Although this is a small increase of the number of global parameters ($\sim 2\,\%$), the calibration parameters introduce a lot of correlations, which lead to a significant increase of computer resources that are required to determine them.

Table 5.1: Summary of calibration parameters used in the full-scale MILLEPEDE II alignment, separated by subdetectors to which they correspond.

Subdetector	Spatial granularity	Time granularity
Lorentz angle		
BPIX	24	65
TIB	72	6 (2)
TOB	96	6 (2)
Backplane correction		
TIB	72	2
TOB	96	2
Total		**3240 parameters**

Spatial granularity is defined as a number of groups of modules that share the same global parameter. Time granularity is defined as a number of periods of time during which all corresponding global parameters are configured as constant. Numbers for time granularity of Lorentz-angle calibration that are given in brackets correspond to the peak mode of microstrip sensors.

5.6.3   Input data

In order to sufficiently constrain all the variable global parameters a large set of tracks is required of different topologies and recorded in different conditions. The standard selection of datasets listed in Section A.2 already contained cosmic tracks, which were recorded with microstrip modules operating in peak mode. Hence, no specific additional data is required to differentiate the backplane corrections from the module-level alignment. But to separate the Lorentz-angle calibration from the module-level alignment and from the backplane calibration, tracks recorded without magnetic field ($0\,\mathrm{T}$ tracks) are required, which were not usually used in tracker-alignment.

0T tracks

Typically, only cosmic tracks are recorded without magnetic field. An important
aspect of such tracks is that they are almost perfectly straight, since there is no
Lorentz force to bend them. This implies that no momentum estimate can be made
for 0 T tracks, which is required for a proper treatment of multiple scattering in
the track fit. Thus, an assumption has to be made about the average p_T of such
tracks. Cosmic tracks are assumed to have $p = 10\,\text{GeV}$, which was estimated from
analogous tracks recorded with 3.8 T magnetic field long ago, since at the early stage
of tracker operation, only cosmic tracks were available for alignment. Unfortunately,
the amount of 0T cosmic tracks after the quality selections listed in Section A.2 is
rather low, which does not allow to constrain module alignment strong enough in
all regions of the tracker.

During a short period, the magnetic field was occasionally switched off, which
resulted in about $60\,\text{pb}^{-1}$ of pp collision data recorded in 0 T magnetic field. Since
such tracks were never used in alignment before, a dedicated study was performed,
which resulted in a special set of quality criteria, which allowed to select about 30
million of such tracks, and to use a fraction of them in alignment. The custom
parameters that were used for the selection criteria were implemented in the official
CMS software framework and were one of the motivations for a decision to have a
dedicated period during 2015 for collision data recorded in 0 T magnetic field. More
details about the selection criteria can be found in Appendix A.3.

Overview of input data

In order to fit the memory consumption of the full-scale alignment in the limits of the
available machines, a dedicated study was performed to optimise the composition
of tracks used in alignment. Based on these results, a more powerful machine for
MILLEPEDE II alignment was requested, which is already installed and allows to
perform alignments that require up 256 GB of memory. More details about the
data-composition study can be found in Appendix A.7.

The whole 2012 data-taking period is divided into 4 run eras, whose boundaries
can be defined in terms of progressing fractions of integrated luminosity: A (0–
20%), B (20–35%), C (35–70%) and D (70–100%). Different datasets with certain
types of tracks were recorded during some particular run periods. Cosmic data is
usually recorded at the beginning or in the end of a certain run period. The amount
of tracks from different datasets used in the combined alignment and calibration is
summarised in Table 5.2. In total, about 53 million tracks were used for the full-scale
alignment.

5.7   Calibration results

In this section, results of the full-scale simultaneous alignment and calibration setup
described in Section 5.6 are presented. For some results, a comparison is made with
calibrations obtained with other setups, which help to clarify possible effects of the
combined alignment and calibrations.

Table 5.2: Approximate amount of tracks of different types used for the combined alignment and Lorentz-angle and backplane calibrations.

Type of tracks	Magnetic field [T]	Run periods	# tracks [$\times 10^6$]
$Z^0 \to \mu^+\mu^-$ decays	3.8	A+B+C+D	3
Cosmic	3.8	A+B+C+D	2.6
Cosmic	($p = 10\,\mathrm{GeV}$) 0	A+C	0.4
Isolated muons	3.8	A+B+C+D	34
Minimum bias	($p = 3\,\mathrm{GeV}$) 0	C	8
Minimum bias	3.8	A+B+C+D	5
Total	**3.8 + 0**	**A+B+C+D**	**53**

The listed numbers of tracks correspond to the amount of tracks that pass the corresponding quality criteria and are used in the combined alignment and calibration. The different types of tracks are ordered by their priority for alignment. $Z^0 \to \mu^+\mu^-$ and cosmic tracks are most important due to their constraints of weak modes, $0\,\mathrm{T}$ tracks are important for the Lorentz-angle calibration and isolated muons provide clearly-measured tracks for the general module-level alignment. Cosmic tracks are recorded in peak mode of the microstrip detector. The momentum assumptions for tracks recorded without magnetic field is given in brackets.

5.7.1 Lorentz angle in the pixel detector

The evolution of the Lorentz angle in BPIX is presented by $\tan(\theta_{\mathrm{LA}}) = \mu_{\mathrm{H}} \cdot 3.8\,\mathrm{T}$ as a function of the integrated luminosity. The luminosity of each IOV was calculated based on its first and last run numbers, using the official CMS utility `lumiCalc2.py`. The calibrated Lorentz-angle values are shown separately for individual rings of every BPIX layer in Figure 5.18.

A clear time dependence of θ_{LA} is visible in every layer of BPIX, which is very similar in all rings within a single layer. Furthermore, in each layer, a clear offset between the negative ($z < 0$) and positive ($z > 0$) halves of BPIX is present. Finally, jumps at certain values of the integrated luminosity are observed consistently in all rings of each layer. These effects are discussed in more detail below.

Effect of pixel-template updates

The jumps in all the distributions in Figure 5.18 at the same IOVs are caused by the non-ideal performance of estimation of the μ_{H} value that is encoded in pixel templates. The inut pixel templates have 8 IOVs, each having an updated set of calibrations and thresholds for pixel-module readouts. As a result, the new calibrated μ_{H} values need a certain offset with respect to their true values, in order to provide the equivalent hit position when pixel templates are used. When the generic hit reconstruction is used, the evolution of θ_{LA} has a smooth shape, because the correction of μ_{H} is determined with respect to a constant input value. In contrast, when the template hit reconstruction is used, the input value differs from the one that is indirectly used in templates, and can not be reproduced with $100\,\%$ accuracy.

(a) BPIX layer 1

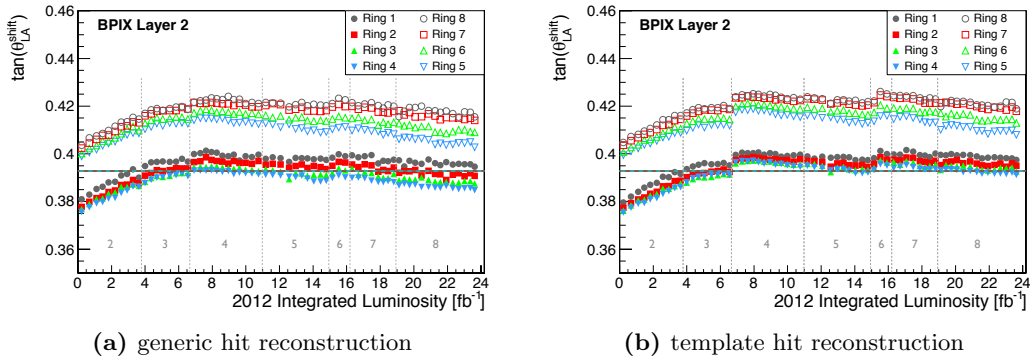**(b)** BPIX layer 2

(c) BPIX layer 3
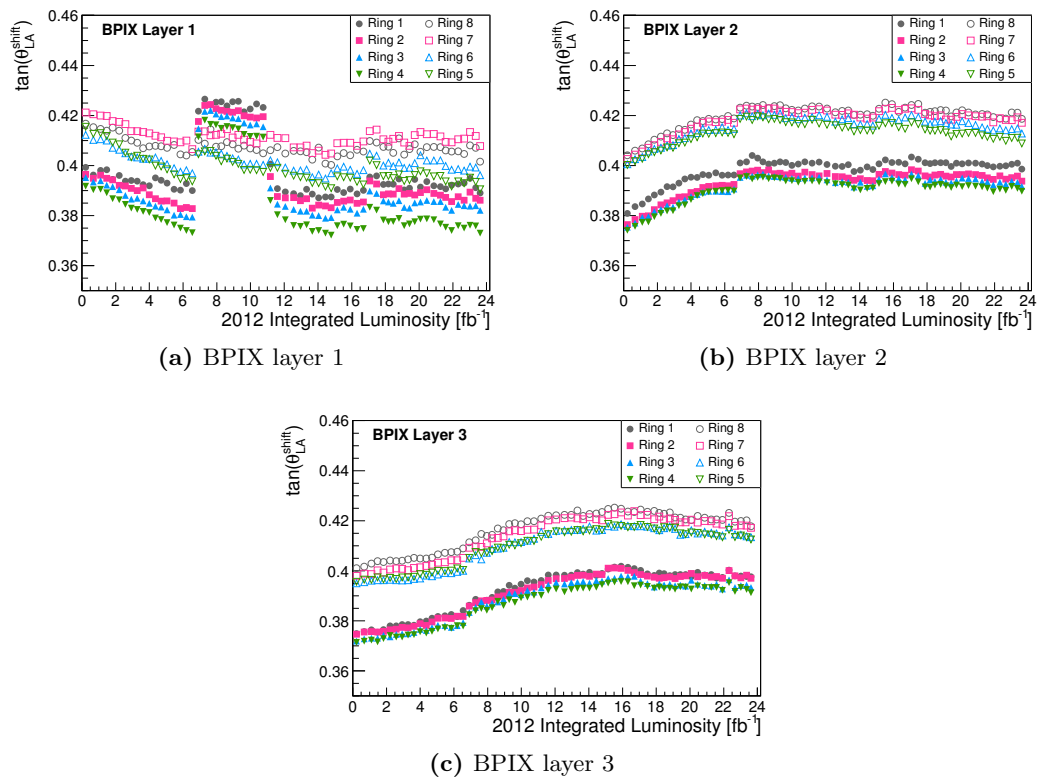
Figure 5.18: Distribution of $\tan(\theta_{LA})$ in BPIX modules as a function of integrated luminosity, separated by individual rings and layers. The ring numbering and colour coding correspond to the ones defined in Figure 5.17. The values are obtained from the combined alignment and calibration, using template hit reconstruction (see Section 5.4.3).

Therefore, offsets in calibrated values are such that they compensate the remaining differences in the treatment of the Lorentz angle between the two hit-reconstruction methods. The results obtained by the two methods are compared in Figure 5.19. For a perfect consistency, IOV boundaries of Lorentz-angle calibration should coincide with the ones of pixel templates. In the calibration setup presented here, IOV boundaries were selected only based on integrated luminosity. Nevertheless, the average width of the IOVs is so small that such boundary effects are negligible.

(a) generic hit reconstruction   **(b)** template hit reconstruction

Figure 5.19: Distribution of $\tan(\theta_{LA})$ in BPIX modules as a function of integrated luminosity, separated by individual rings of layer 2. Compared are values determined with either generic or template hit reconstruction in the pixel detector. The horizontal line represents the input $\tan(\theta_{LA}) = \mu_H \cdot 3.8\,\mathrm{T}$ value, with respect to which corrections were determined by MILLEPEDE II. The vertical dashed lines represent the boundaries of individual template IOVs with updated calibrations.

The large offset in Lorentz-angle values of the first layer for a half of the modules in the luminosity range of $7\,\mathrm{fb}^{-1}$ to $11\,\mathrm{fb}^{-1}$ is caused by a wrong Lorentz-angle value used in pixel templates for this period. But thanks to the combined alignment and calibration approach, it was corrected such that reconstructed hits have optimal positions.

Shape of Lorentz-angle evolution in BPIX

The calibrated values of $\tan(\theta_{LA})$ at the beginning of 2012 are in agreement with the results obtained by the grazing-angle method [159], which can be seen from a comparison of average values from individual plots in Figure 5.18 with individual curves in Figure 5.11a. In particular, $\tan(\theta_{LA})$ values for the modules of the first layer in the first IOVs of the both methods differ by not more than $1\,\%$. In outer layers, the absolute values become smaller in both results, but by a different amount. Thus, in layer 3 the difference reaches $5\,\%$

The evolution of $\tan(\theta_{LA})$ determined by the integrated approach has different slopes in individual rings of layer 1. In particular, the most rapid change of Lorentz angle happens in rings 4 and 5, which are closest to the interaction point, while in outermost rings 1 and 8 the changes are in the same direction but less rapid. This is an expected behaviour, considering a decrease of irradiation damage at larger distance from the collision point.

But the shapes of Lorentz-angle evolution in individual layers of BPIX is significantly different from the ones observed by the standalone calibration [159]. The values measured by the grazing-angle method are in agreement with expectations [157], according to which the μ_H (and $\tan(\theta_{LA})$) should increase with accumulated irradiation dose. In contrast, the values determined with the combined approach have opposite quasi-linear behaviour in layer 1 and non-linear behaviour in layers 2 and 3. Considering that the irradiation dose accumulated by layers at larger radius is smaller than by those that are close to the beam line, from purely empirical reasoning, it can be assumed that the shapes in the three individual layers represent different parts of a single dependence shown in Figure 5.20.

(a) evolution of $\tan(\theta_{LA})$ in modules located at similar z, but in different layers, determined with the generic hit reconstruction [164]

(b) assumed dependence of $\tan(\theta_{LA})$ on integrated luminosity
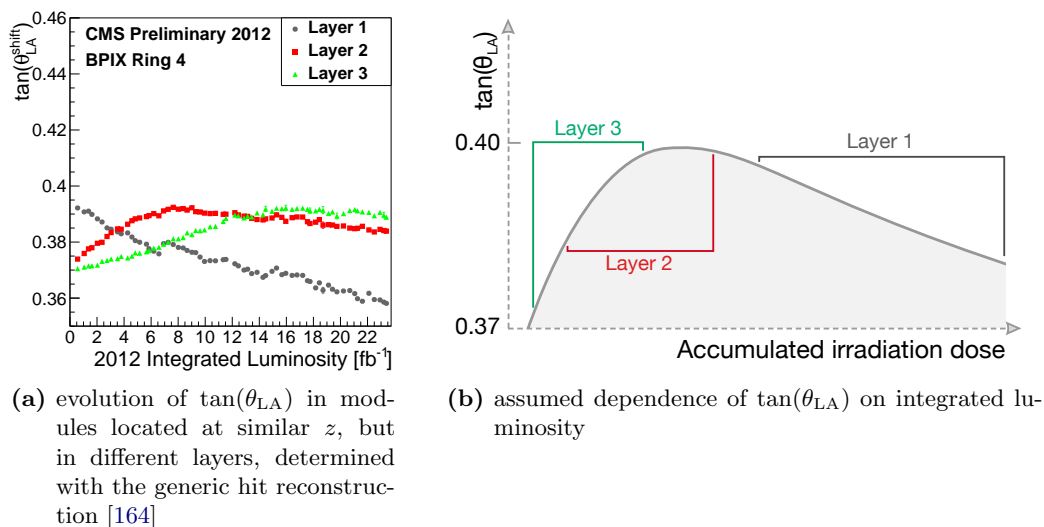
Figure 5.20: Assumed dependence of Lorentz angle on integrated luminosity, which could explain the observed difference in evolution of $\tan(\theta_{LA})$ between layers located at different distances from the beam line.

Another possible explanation of such a behaviour can be that the calibrated θ_{LA} is not the same quantity as determined by the grazing-angle method. Thus, some other effect may be present, with Lorentz-shift-like behaviour, but with an opposite time dependence. Since the individual modules are not allowed to move in the used setup, the θ_{LA} is the only quantity that could compensate such an effect. Hence, some time-dependent rotations of individual BPIX layers around the z axis could lead to such shifts of the hit positions that can be compensated by adjusting the Lorentz angle correspondingly.

Hit-position resolution

As mentioned in Section 5.4.3, compensations of other effects with Hall-like behaviour by the integrated Lorentz-angle calibration are expected, and the obtained hit-position resolution is the most relevant measure of performance. This was ensured by DMRs from modules in different layers of BPIX. Residuals were calculated using tracks from isolated muons recorded during the period corresponding to the range of integrated luminosity from $19.0\,\mathrm{fb}^{-1}$ to $19.7\,\mathrm{fb}^{-1}$ The comparison is shown in Figure 5.21. In all the three layers, the combined approach provides a smaller width of DMR, than can be achieved by alignment and Lorentz-angle calibration

performed separately. Although the largest difference in $\tan(\theta_{LA})$ between the two calibration approaches is in layer 1, DMRs show a similar width for both approaches. In contrast, in layer 3, the difference in $\tan(\theta_{LA})$ is smaller, but the DMR clearly shows two peaks corresponding to modules with opposite orientations.

A possible explanation of this observation is that $0\,T$ and $3.8\,T$ tracks pull modules in different directions if no integrated Lorentz-angle calibration is performed and standalone μ_H values are wrong. In this case positions and orientations of the modules are biased towards the tracks that have the largest statistical power. Considering the constant increase of instantaneous luminosity during 2012, there were more tracks at the end of 2012 than at the beginning. Thus, the large difference in $\tan(\theta_{LA})$ was compensated by shifts of the modules, which were biased towards the Lorentz-induced shift closer to the end of 2012. Therefore, the difference in DMRs of layer 1 is rather small for this period. In layer 3, the standalone $\tan(\theta_{LA})$ is almost constant, which means that the distance between the peaks in Figure 5.21c corresponds to the change of $\tan(\theta_{LA})$ that was obtained with integrated calibration from the beginning to the end of 2012. The change of the Lorentz angle during this period, $\Delta\tan(\theta_{LA}) \approx 0.02$ corresponds to the shift of the hit position by $\Delta\tan(\theta_{LA}) \cdot d/2 \approx 0.02 \cdot 285/2 \approx 2.8\,\mu m$, which is consistent with the distance between the peaks of the DMR in Figure 5.21c. The largest change of $\tan(\theta_{LA})$ in BPIX throughout 2012 is equivalent to a shift of the module by up to $4\,\mu m$.

The effect of the integrated Lorentz-angle calibration on the hit-position resolution was estimated in another way by Armin Burgmeier, using tracks with $p_T > 12\,GeV$ and with hits in each of the three layers of BPIX [166]. After removing the hit in layer 2, the track is refitted, and the distribution of the Δu residual of the removed hit is obtained for about 2 million tracks. The RMS of such a distribution divided by $\sqrt{3/2}$ is used as an intrinsic resolution of BPIX modules. Since adjacent ladders of BPIX modules in ϕ have opposite orientations, the spread of Δu along the global ϕ of the module position is sensitive to wrong Lorentz-angle values. The width of the spread of the ϕ of the modules as a function of Δu is used as a measure of a remaining Lorentz-angle miscalibration or of a consistent misalignment [166]. The comparison of the widths estimated in this way for different alignment setups is shown in Figure 5.22.

From the comparison in Figure 5.22a, it is clearly seen that the pixel-template hit reconstruction provides a much better resolution than the generic method, improving it by up to 15 %. The effect of the integrated Lorentz-angle calibration on the intrinsic resolution shows an improvement, which is larger when the generic hit reconstruction is used. When pixel templates are used, the effect is smaller, but in the direction of improvement throughout the whole analysed time period.

Furthermore, both generic and template hit-reconstruction methods can not overcome the problem of non-optimal Lorentz-angle calibration by standalone methods, which leads to variation of measured hit positions in modules at different ϕ by up to $1.5\,\mu m$. If pixel templates are used, these variations are reduced by about 10 % to 50 %, but they still vary with time. If combined alignment and Lorentz-angle calibration is performed, the variations are only about $0.5\,\mu m$, and are constant in time. Thus, the combined approach provides the highest intrinsic resolution, which is most homogeneous in space and most stable in time.

(c)
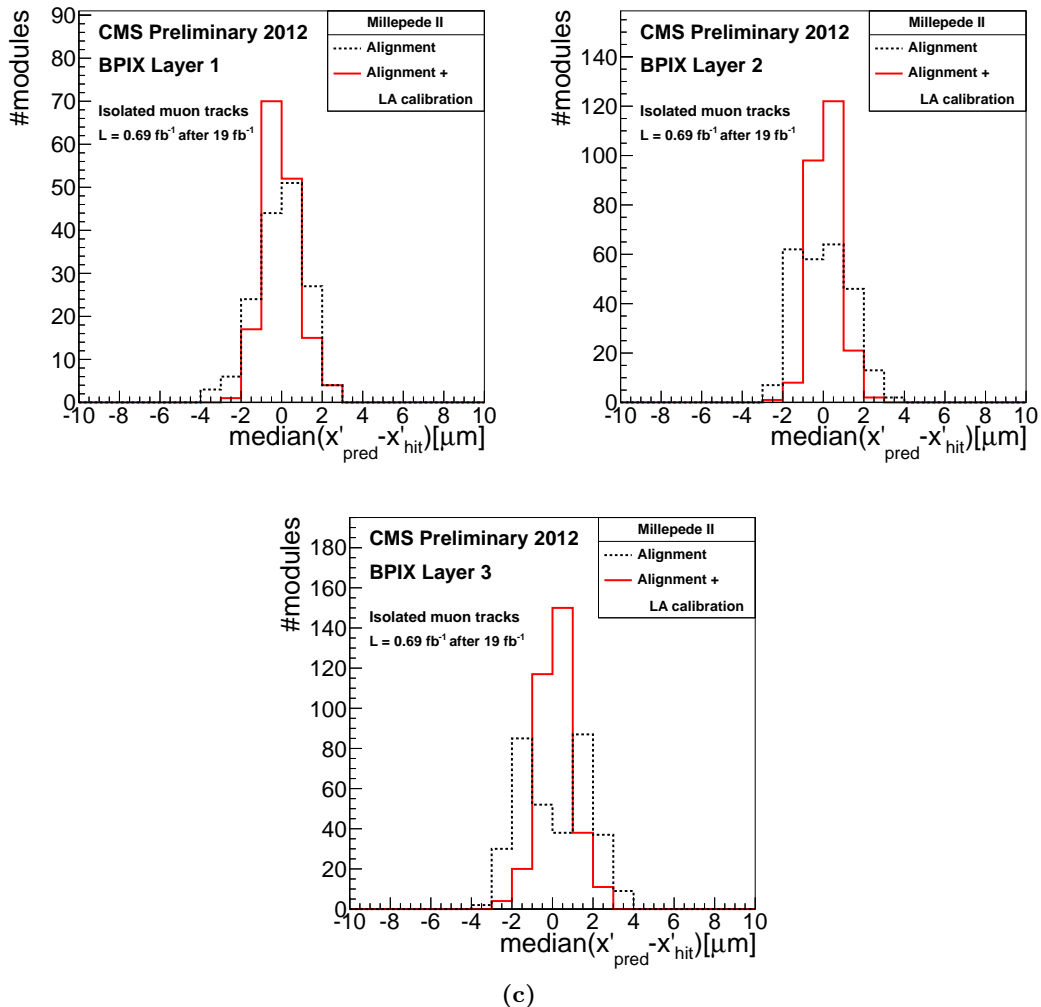
Figure 5.21: Distributions of the median of the residuals in different layers of BPIX compared between two geometries [165, 164]. Residuals are calculated from 2 million tracks of isolated muons recorded during the period corresponding to the integrated luminosity from $19.0\,\mathrm{fb}^{-1}$ to $19.7\,\mathrm{fb}^{-1}$. The *Alignment* setup was obtained by alignment without Lorentz-angle calibration, using standalone μ_H values from a database for alignment and validation. The *Alignment+LA calibration* setup was obtained by simultaneous alignment and Lorentz-angle calibration, using the calibrated μ_H values used for validation. Both alignments were performed using the generic hit reconstruction in the pixel detector.

(a) width of the Δu distribution

(b) width of the spread of Δu residual along the ϕ-coordinate of the module
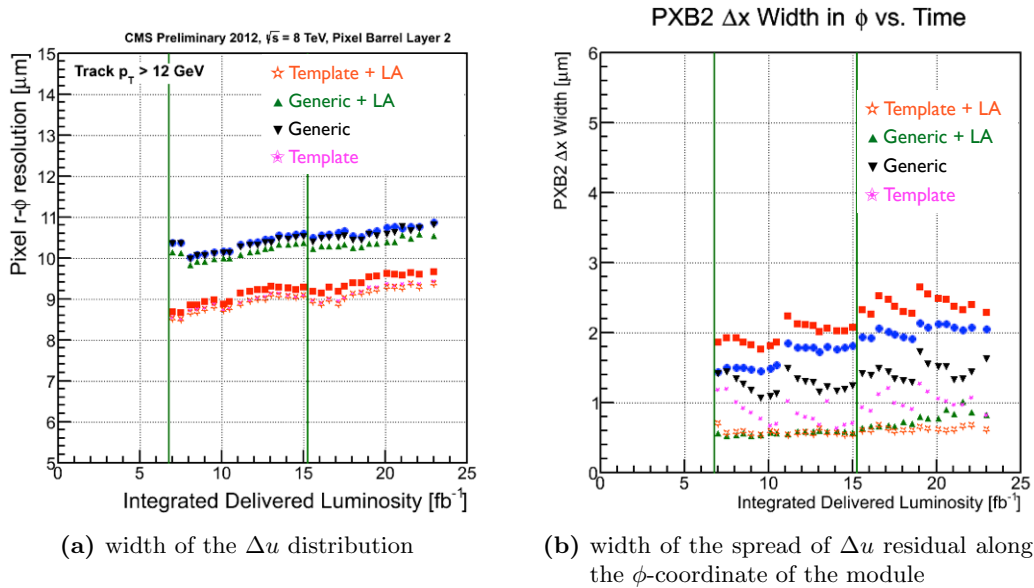
Figure 5.22: Widths of track-hit residuals for a hit in the second layer of BPIX removed from the track fit. The width $(RMS \cdot \sqrt{3/2})$ of the distribution of the residuals is defined as an intrinsic resolution of modules in layer 2. The spread of Δu in ϕ is an estimate of the effect of a remaining Lorentz-angle miscalibration on the measured u coordinate of the hits. Compared are different alignment setups; the most relevant ones are listed in the legend. The name of the alignment configuration denotes the hit-reconstruction method used for the alignment, while the presence of LA in the name means that the Lorentz-angle calibration was integrated into the corresponding alignment setup. The hit-reconstruction method used during alignment was consistently used for the resolution estimation [166].

Offset between $z < 0$ and $z > 0$ parts of BPIX

A prominent offset between $\tan(\theta_{\mathrm{LA}})$ values of the negative (rings 1–4) and positive (rings 5–8) halves of BPIX is clearly seen in Figure 5.18 in each layer of BPIX. A similar offset is observed by the standalone calibration, as shown in Figure 5.11.

No reason for this difference could be found up to now. A few assumptions were made by Morris Swartz and Danek Kotlinski [167] about the possible reason, but neither of them is confirmed:

- **Different bias voltage:** is the most obvious explanation, because the two parts are powered by individual power supplies. According to simulations by Morris Swartz, a difference in bias voltage of 20 V would be required to create such an offset. A direct measurement of the voltages from power supplies during the long shutdown showed the difference of only 1 V [167].

- **Missing grounding:** it was discovered during the shutdown, that one of the two parts of BPIX was not grounded. A new standalone measurement of $\tan(\theta_{\mathrm{LA}})$ using cosmic tracks recorded in 2015 and the upgraded CMS tracker with both parts being perfectly grounded, still shows a similar difference [167].

Possible improvements

An important step for improvement of the Lorentz angle calibration is a possible disentanglement of the Lorentz-angle calibration from other effects that must have been compensated by it in the combined approach. One of such effects could be module-level time-dependent geometry distortions. This would require to introduce several IOVs for individual-module alignment parameters and set only 1 IOV for the Lorentz angle. Such alignment should be performed with tracks recorded during a period of significant changes of the tracker environment, e.g. the temperature or the magnetic field, which are most likely to affect positions of the modules, but very unlikely to change the Lorentz angle. In particular, the period in 2012C, when the magnetic field was occasionally switched off, is a perfect candidate, since a single parameter has changed, which should not affect the Lorentz angle at all. Thus, tracks recorded before and after this period were recorded with potentially different module-level geometries, which can be easily detected.

Another improvement that does not require sophisticated studies is to take into account the position of the beamspot, which is not centered with respect to the BPIX in the $r\phi$ plane. This results in an uneven irradiation dose in the BPIX modules across the azimuthal angle, which is reflected by the leakage current in BPIX modules as a function of ϕ, as shown in Figure 5.23. Thus, it would make sense to segment the BPIX in azimuthal angle such that the leakage current does not change much within a single group of modules. Due to the difference in flux of particles, both absolute values and their time dependence can vary as a function of ϕ. This might be less relevant for outer layers, but in the innermost layer such a granularity should be checked at least.
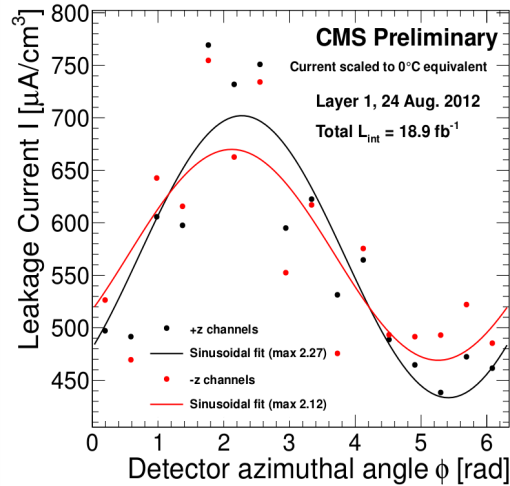
Figure 5.23: Leakage current as a fuction of azimuthal angle in BPIX modules of the innermost layer. It is not evenly distributed across modules located at different ϕ due to an offset in the LHC beam position in the transverse plane with respect to the center of the pixel detector. The data points taken in the positive and negative halves of the detector along z are fitted with sinusoidal curves [168].

5.7.2 Lorentz angle in the microstrip detector

In the microstrip modules, evolution of $\tan(\theta_{LA})$ is also presented as a function of the integrated luminosity throughout 2012. Since it was determined with a finer spatial granularity than in the pixel detector, it is only shown for the inner and outer layers of TIB and TOB. Results for modules in layer 2 are similar to those for layer 1, while other layers show a similar behaviour to the one in outermost layers.

Tracker Inner Barrel (TIB)

The layer 1 of TIB is split into 12 groups of modules along the z axis, which are further split into two subgroups of $r\phi$ modules and stereo modules. The two subgroups of modules have opposite orientations, and, consequently, Lorentz drifts in opposite directions. The evolution of $\tan(\theta_{LA})$ in the innermost layer of TIB in deconvolution readout mode is shown in Figure 5.24. The time dependence obtained with the fine granularity shows no significant time dependence, but a significant spread of absolute values. If the spatial granularity is reduced such that modules with opposite orientations have the same μ_H value, and that adjacent groups along the z axis are merged (6 groups along z), the spread of absolute values within a single layer becomes negligible. But modules in Ring 6 from Figure 5.24c, which correspond to modules in Rings 11 and 12 from Figures 5.24a and 5.24b, have a noticeable offset from other modules, which corresponds to only about 1 µm shift of the hit position due to the Lorentz-angle effect. The input value stored in a database differs by the same amount.

DMR validation of the two geometries showed no significant difference between them. Therefore, the differences in absolute values of $\tan(\theta_{LA})$ must be either compensated by corresponding shifts of the modules, which can reach 5 µm, or uncer-

(a) $r\phi$ modules

(b) stereo modules

(c) $r\phi$ and stereo modules combined, and a
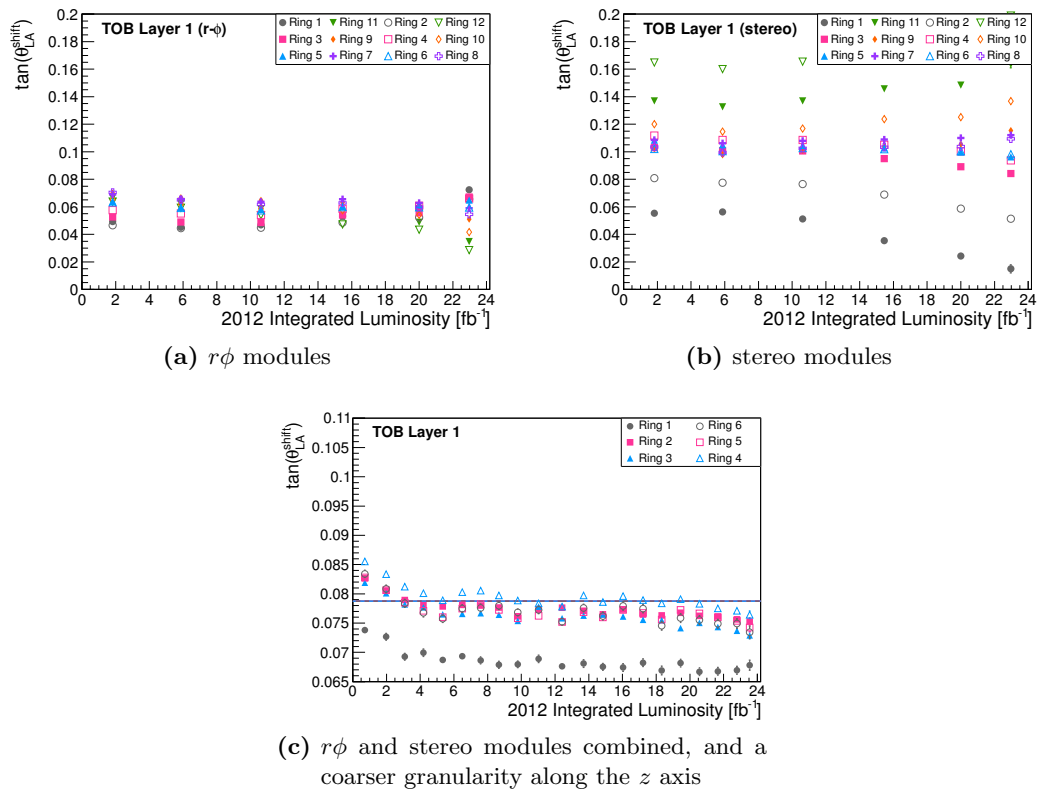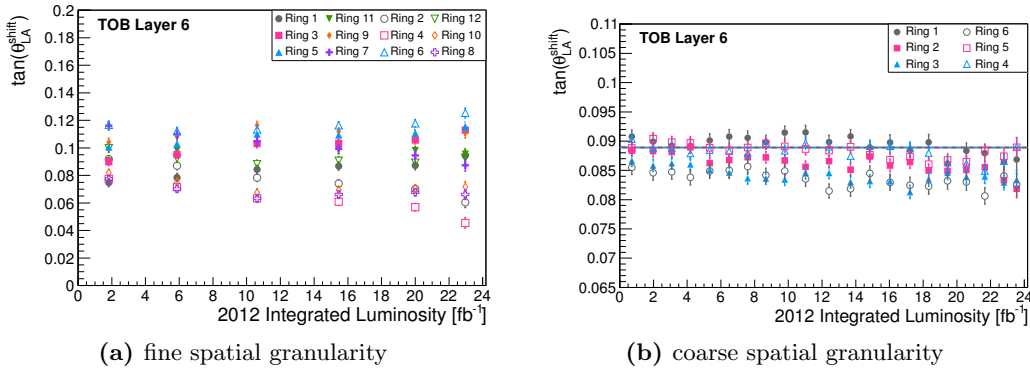coarser granularity along the z axis
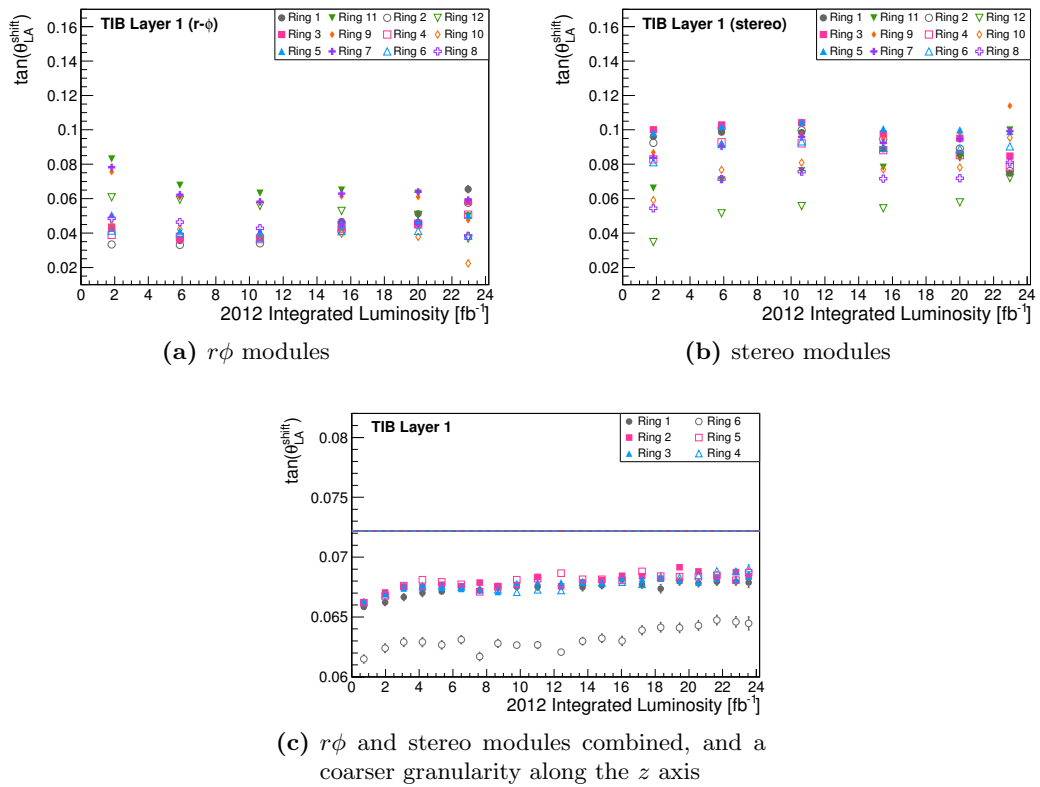
Figure 5.24: Evolution of $\tan(\theta_{\mathrm{LA}})$ in the modules of layer 1 in TIB, operating in deconvolution mode, separated by 12 rings along the z axis, and shown separately for $r\phi$ modules (a) and stereo modules (b). The evolution determined with a coarser granularity of 6 groups along the z axis, regardless of their orientation, is shown in (c). The horizontal line represents the input $\tan(\theta_{\mathrm{LA}})$ value from a database, to which corrections were determined by MILLEPEDE II. Uncertainties are estimated only for the calibration with the coarse spatial granularity.

tainties of the determined Δw values are larger than the spread observed for the fine granularity.

In the outermost layer of TIB (layer 4), which has no stereo modules, the same spread of absolute values is observed as in the first layer, but it is among $r\phi$ modules themselves. The calibrated values are in a good agreement with the values in a database, and no time dependence is observed. A separation is visible between even and odd ring numbers, as shown in Figure 5.25, which correspond to modules with opposite orientations of the local w axis. Potentially, such a dependence could reflect movements of oppositely oriented modules with respect to each other, which is compensated by a Lorentz shift, since individual modules are not allowed to change their positions with time. Comparison of DMRs for the two setups also showed no significant difference between the two granularities.

(a) fine spatial granularity

(b) coarse spatial granularity
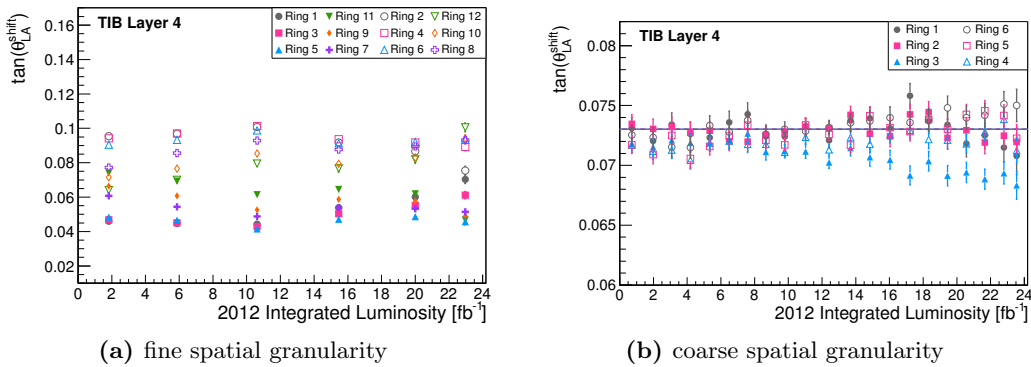
Figure 5.25: Evolution of $\tan(\theta_{LA})$ in the modules of layer 4 in TIB, operating in deconvolution mode, separated by 12 (a) or by 6 (b) rings along the z axis. The horizontal line represents the input $\tan(\theta_{LA})$ value from a database, to which corrections were determined by MILLEPEDE II. Uncertainties are estimated only for the calibration with the coarse spatial granularity.

Similar behaviour is observed in the Lorentz-angle values for the peak mode, which is shown in Figure 5.26.

Tracker Outer Barrel (TOB)

Time dependence of $\tan(\theta_{LA})$ in the innermost and outermost layers of TOB is shown in Figures 5.27 and 5.28 respectively, and shows a similar behaviour to the one observed in TIB. Calibrated values in all the layers are very close to the ones stored in a database. The time dependence of $\tan(\theta_{LA})$ in the innermost layer determined with the coarse spatial granularity is slightly stronger than in TIB and develops in the opposite direction. Its largest change throughout 2012 is equivalent to a shift of $2.5\,\mu m$. In the outermost layer almost no time dependence is observed.

Unification of Lorentz angle calibrations

Lorentz-angle values obtained for deco and peak modes have very similar absolute values. This allows to conclude that the readout mode does not make any significant influence on the Hall effect inside microstrip sensors. This is an expected behaviour, since the Lorentz angle should be only a property of the sensor material, but not of

(a) fine spatial granularity

(b) coarse spatial granularity
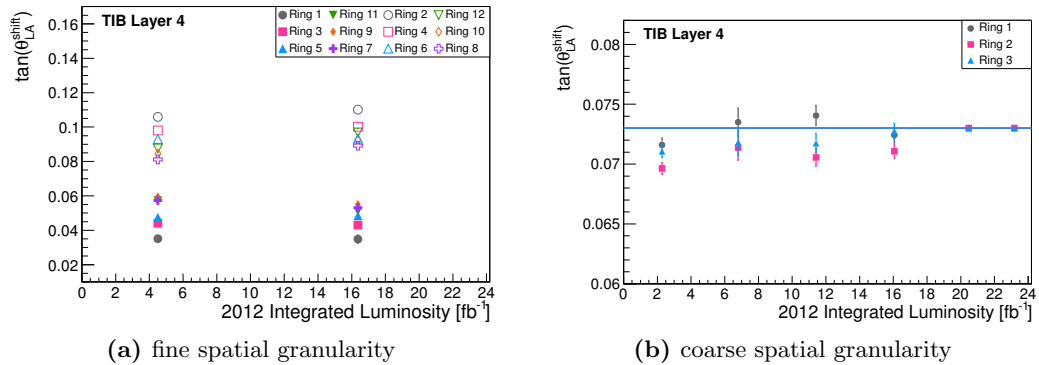
Figure 5.26: Evolution of $\tan(\theta_{\mathrm{LA}})$ in the modules of layer 4 in TIB, operating in peak mode, separated by 12 (a) or by 3 (b) groups along the z axis. The horizontal line represents the input $\tan(\theta_{\mathrm{LA}})$ value from a database, to which corrections were determined by MILLEPEDE II. The last two points in (b) have no correction, because no tracks recorded in peak mode were available for this period.

(a) $r\phi$ modules

(b) stereo modules

(c) $r\phi$ and stereo modules combined, and a coarser granularity along the z axis

Figure 5.27: Evolution of $\tan(\theta_{\mathrm{LA}})$ in the modules of layer 1 in TOB, operating in deconvolution mode, separated by 12 rings along the z axis, and shown separately for $r\phi$ modules (a) and stereo modules (b). The evolution determined with a coarser granularity of 6 groups along the z axis, regardless of their orientation, is shown in (c). The horizontal line represents the input $\tan(\theta_{\mathrm{LA}})$ value from a database, to which corrections were determined by MILLEPEDE II.

(a) fine spatial granularity   **(b)** coarse spatial granularity

Figure 5.28: Evolution of $\tan(\theta_{LA})$ in the modules of layer 6 in TOB, operating in deconvolution mode, separated by 12 (a) or by 6 (b) rings along the z axis. The horizontal line represents the input $\tan(\theta_{LA})$ value from a database, to which corrections were determined by MILLEPEDE II.

the time used to collect the signal, at least in first order. Therefore, it is possible to make a generic implementation of the Lorentz angle calibration, which does not only take into account the correlation between the Lorentz angle and backplane correction in peak mode, but also propagates it to deconvolution mode. In this way, a single set of μ_H values would be used, which are defined by a single derivative, while in deco mode only a backplane correction would be required in addition.

5.7.3   Backplane correction in the microstrip detector

The calibrated backplane correction is presented by a fraction of the thickness of the module Δw as a function of integrated luminosity. The evolution of Δw in the innermost layer of TIB is shown in Figure 5.29. Similarly to the Lorentz angle, when the fine granularity is used, a spread is observed around the values obtained with the coarse granularity. Furthermore, unphysical negative values are obtained for some groups of modules, which mean that the barycenter of the cluster should be moved in the opposite direction with respect to the one that can be caused by the backplane effect. The largest value of Δw is about 8 %, which for 320 μm thick TIB modules corresponds to the shift of the effective hit position by up to 13 μm along the w axis.

For TOB, only the coarse granularity is shown. In the innermost layer of TOB, the two outermost rings 1 and 6 have negative values, even when the coarse granularity is used, as can be seen in Figure 5.30. In the second layer, the two outermost rings also have a large offset from the rest of the modules, but in the opposite direction. Corrections determined for other modules can reach 25 %, which for 500 μm thick TOB modules corresponds to the effective shift of hit positions by up to 62 μm along the w axis.

A possible explanation of the large offsets in the outermost rings of the two innermost TOB layers is that they are biased by TEC modules, which are very close to them. Since no backplane correction is determined for TID and TEC modules due to some technical problems, it can be that a possibly wrong backplane correction in TEC modules introduces a tension between tracks recorded in different readout modes, which leads to a shift of TEC modules along the global z axis. Such shifts

(a) $r\phi$ modules

(b) stereo modules

(c) $r\phi$ and stereo modules combined, and a
coarser granularity along the z axis
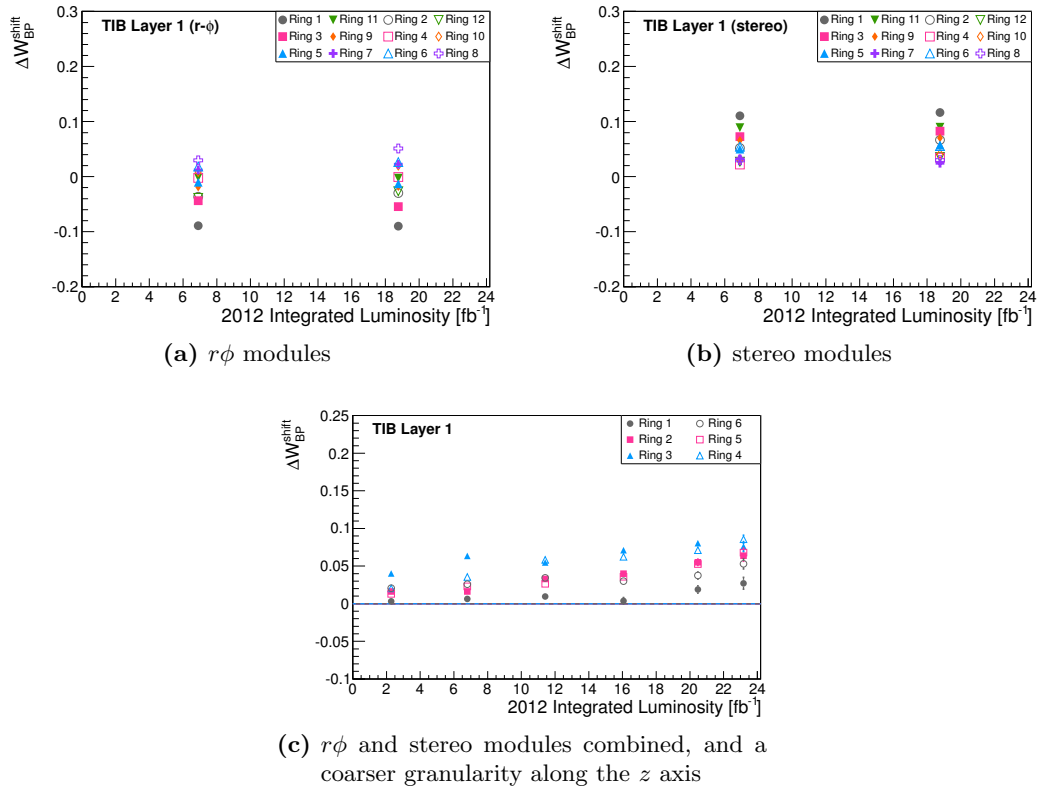
Figure 5.29: Evolution of Δw in the modules of layer 1 in TIB, separated by 12 rings
along the z axis, and shown separately for $r\phi$ modules (a) and stereo mod-
ules (b). The evolution determined with a coarser granularity of 6 groups
along the z axis, regardless of their orientation, is shown in (c). The
horizontal line represents the input Δw value from a database, to which
corrections were determined by MILLEPEDE II.
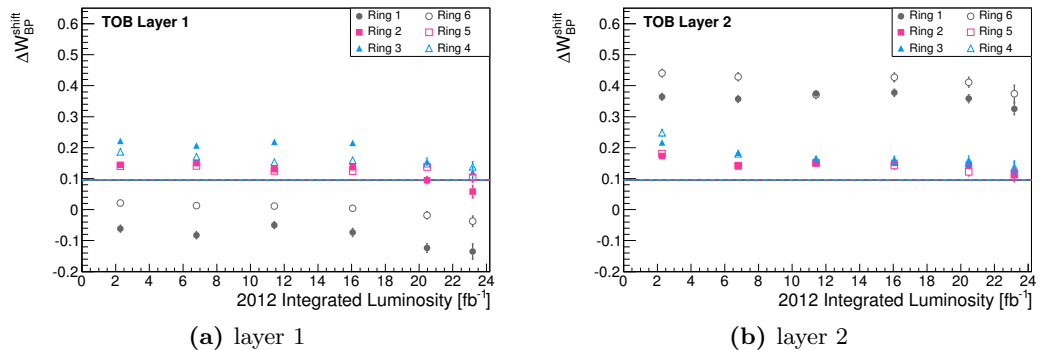
(a) layer 1

(b) layer 2

Figure 5.30: Evolution of Δw in the modules of layer 1 and layer 2 of TOB, separated
by 6 groups of modules along the z axis. The horizontal line represents the
input Δw value from a database, to which corrections were determined by
MILLEPEDE II.

would increase residuals in TOB modules that are adjacent to TEC, which are compensated by a rotation and an increased backplane correction for such modules.

This is also reflected by a validation distribution that shows the dependence of $(u_{\text{hit}} - u_{\text{track}})/\tan(\alpha)$ on the u coordinate of the hit, where α is the rotation of the module around the u axis in its local coordinate frame. This distribution is sensitive to surface deformations of the module, and is shown separately for outermost rings of layer 2, which have large offsets of backplane correction, and for all other modules of the layer. The distributions are shown in Figure 5.31, and the shape of the distribution for rings 1 and 6 looks like the modules are rotated around the v axis with respect to the optimal orientation. A similar effect but in the opposite direction was observed for the layer 1 of TOB. Thus, when the backplane calibration is integrated into alignment, the average value is almost perfectly corrected, but the slope of the distribution remains the same, which means that something is distorting the orientations of TOB modules that are close to TEC.

(a) rings from 2 to 5   (b) rings 1 and 6
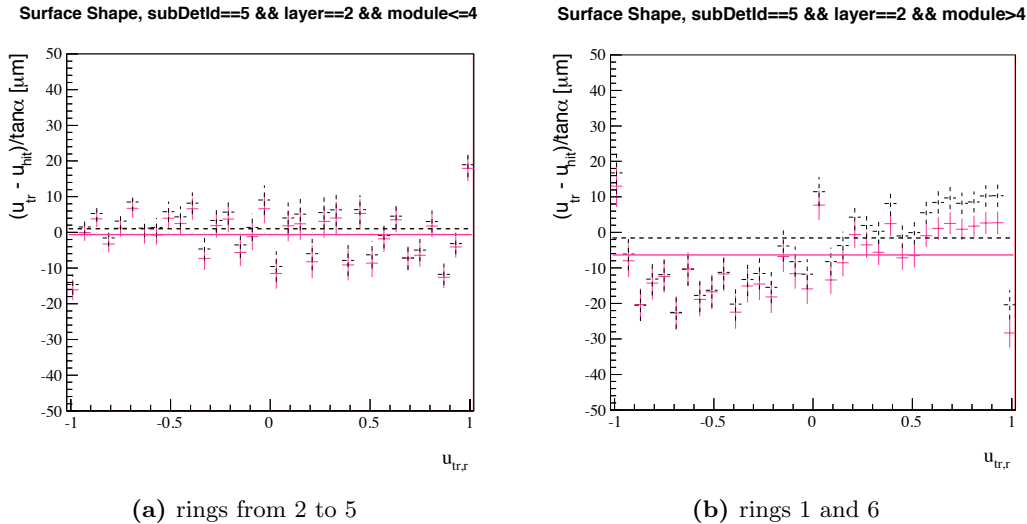
Figure 5.31: Surface shape distributions for modules of TOB in layer 2, shown separately for modules that have a Δw value close to the average and for modules that have a large offset in Δw (see Figure 5.30). The x axis represents the u coordinate normalised to the double of the width of the module along u. Compared are the geometry that was determined using standalone backplane corrections (solid magenta) and the one using backplane calibration integrated into alignment (dashed black). The shape of the distributions in (b) looks as if the modules were rotated around the v axis with respect to the real orientation. Horizontal lines represent an average of the corresponding distribution. The average residual in (b) is corrected by the backplane calibration from 6 μm to 2 μm.

5.8   Conclusions

The existing MILLEPEDE II alignment algorithm was extended by the calibrations of the Lorentz angle and of the backplane correction. The Lorentz-angle calibration is very important for a most precise estimation of measured hit positions in the magnetic field. The backplane correction has to be precisely determined in order

to properly compensate for the loss of charge during collision-data taking, i.e. deco
mode, and to have the possibility of a consistent use of cosmic tracks recorded in
peak mode in the alignment, which allow to significantly constrain certain weak
modes. Both calibrations were determined as functions of integrated luminosity in
order to account for irradiation damage and other possible time-dependent effects.
This is required to ensure that optimal values are used throughout the whole period
of data taking.

The calibrated Lorentz-angle values in the pixel detector show a significant time
dependence, which is opposite to the one determined by standalone methods and
contradicts expectations from simulations and hardware-based measurements. Nev-
ertheless, these values lead to a better hit-position resolution, which is constant in
time, in contrast to standalone results, which have significant variations with time.
An offset between Lorentz-angle values of the positive and negative halves of the
BPIX along z was observed, which is consistent with standalone measurements, but
no proven explanation of this offset exists. The possible explanation of this be-
haviour is that some other effect is absorbed by the Lorentz-angle calibration, which
can be some module-wise geometry distortions that are not constant in time.

The Lorentz-angle calibration in the strip detector showed a smaller time depen-
dence. When a very fine spatial granularity is defined for the calibration, unexpected
spreads of values are observed, which do not provide a better resolution than if a
coarser granularity is used. This might be an indication of a weak mode or of large
uncertainties of the determined values. Lorentz-angle values obtained for the peak
and deconvolution readout modes are consistent and allow to use a more generic
definition, which would correlate the Lorentz-angle calibration with the backplane
correction, and better constrain the two.

The backplane calibration showed almost no time dependence, and the same un-
expected spreads of absolute values if a very fine granularity is used. With a coarser
granularity, spatial variations are minimal. The only exception is the boundary
modules of the two innermost layers in TOB, which are located close to the positive
and negative sides of TEC. It can be explained by possible misalignment or back-
plane miscalibration in TEC modules, which bias the TOB modules that are close
to them.

In summary, the best possible full-scale alignment of the CMS tracker was per-
formed with the two new calibrations integrated into the MILLEPEDE II algorithm.
This is a big step in the direction of achieving the best possible performance of the
CMS tracker, since all major effects that can bias the alignment process are deter-
mined simultaneously, using the potential of correlations between different global
parameters. Changes of θ_{LA} are expected to be significantly larger during the future
runs of the LHC due to a higher irradiation dose, which makes the studied calibra-
tion Lorentz-angle calibration method especially relevant. This technique will be
used to align and calibrate the upgraded CMS tracker during the next run of the
LHC in 2015. The occasional dataset of collision tracks recorded without magnetic
field was proven to be suitable for the Lorentz-angle calibration and is an important
ingredient of the full-scale alignment. The developed technique for selection of high-
quality tracks allowed to use them in alignment together with commonly used track
topologies. The performed studies showed their positive impact on the alignment
performance, which was one of the motivations for a dedicated period of time in
2015, during which CMS will record such tracks on purpose.

Chapter 6

Event simulation

A real measurement of some process initiated by proton-proton collisions at the LHC involves the actual collision of protons, the measurement of final state particles, the reconstruction of the desired quantities and the conversion to quantities that can be theoretically calculated. In order to obtain the most correct and precise results, it is necessary to calibrate all the detector components and to properly correct for possible misreconstruction of measured quantities, as well as to optimise the selection criteria for a certain process that has to be measured. Furthermore, for searches for new physics or not-yet-discovered standard model processes, it is essential to estimate the possible sensitivity to particular processes under different conditions.

All this can be achieved by a computer simulation of the measured events, which is based on the Monte Carlo (MC) method and is widely used in the area of experimental particle physics. This is a stochastic technique, which provides the numerical computation through a repeated random sampling, which in the final result produces events in exactly the same format as the measured ones, but with the great advantage of having full access to the evolution of each event from the proton-proton collision to the reconstructed objects.

The process of simulation of any process is rather complicated, but can be divided into two independent parts: generation of the event and simulation of the detector response. The two stages are briefly described in the following sections.

6.1   Event generation

In accordance with the factorisation theorem (see Section 2.2.2), the generation of a collision event is divided into several steps, as schematically illustrated in Figure 6.1.

The event generation is in general divided into two subsequent steps: matrix element (ME) calculation of the hard scattering process and calculation of higher order QCD effects by a parton shower (PS) generator.

6.1.1   Matrix element generators

The first step in the event generation is the calculation of the hard process at a fixed order of α_s, using the momenta of the incoming partons, which are distributed
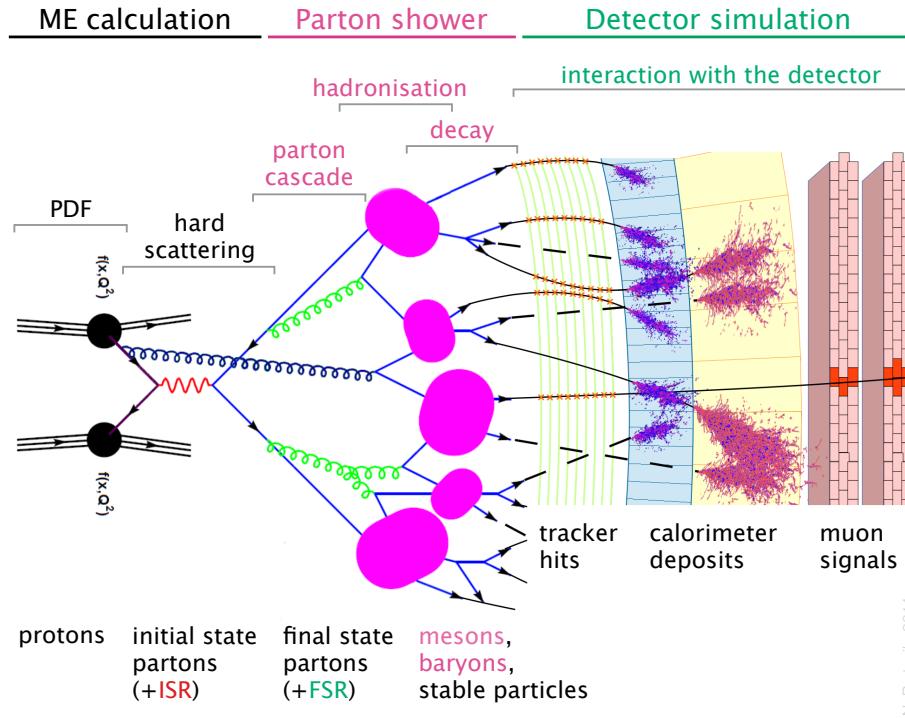
Figure 6.1: Schematic view of the typical event-generation process for a proton-proton collision. Based on the diagram from [169].

according to a random sampling of the input PDF. The final state depends on a set of interactions that are fed into the generator.

The ME calculation of the $t\bar{t}b\bar{b}$ process was performed by the MADGRAPH generator [170], which calculates the $t\bar{t}$ process at the leading-order (LO) of QCD with up to nine final state particles. In the case of dileptonic $t\bar{t}$ events, the decay of top quarks leads to six particles: two electrons/muons, two neutrinos and two b quarks. Thus, the remaining three are additional particles, which are calculated using only tree-level diagrams. The MADGRAPH generator is interfaced with the MADSPIN software package [171], which adds the spin-correlation effects to the $t\bar{t}$ pair and to its decay products. The TAUOLA package [172] is also used with the MADGRAPH generator to simulate the decays of τ leptons. Branching fractions for individual decay modes of W bosons in the MADGRAPH simulation are set to exact multiples of $\frac{1}{9}$, as described in Section 2.3.1, which is not fully correct. Thus, branching fractions of individual dileptonic and semileptonic channels, as well as of all full-hadronic channels combined, are corrected to the values listed in Table 2.6 using additional weights calculated in every $t\bar{t}$ event.

For systematic studies, alternative POWHEG [173] and MC@NLO [174] generators were used for the simulation of $t\bar{t}$ events at NLO accuracy, by including virtual and real-emission corrections. In these simulations, only additional gluon radiation is calculated by the ME generators, but not the splitting to a b$\bar{b}$ pair. Therefore, these simulations do not provide full LO calculations of the $t\bar{t}b\bar{b}$ process.

Other processes, including $t\bar{t}H$ production, are simulated with the PYTHIA [175] generator, which calculates a defined process at LO, without any additional radiation at ME level.

6.1.2   Parton-shower generators

After the matrix element calculation, higher order QCD effects, together with ISR and FSR, are added by parton-shower generators. Parton-shower models recursively split the radiated gluons into pairs of q$\bar{\text{q}}$ or gg pairs, reducing the energy scale of the event. This process starts at some highest energy scale Q_0, which is determined by the hard scattering, and continues until a cut-off scale Q_{min} is reached. At the scale Q_{min}, further splitting will not be simulated perturbatively, and the new particles might not be experimentally distinguishable, either due to their low energy or their collinear direction with respect to the parent parton. The values of the Q_0 and Q_{min} parameters are configurable in the simulation program, and define the amount of radiation to be produced by the showering process.

The two parton-showering models that will be discussed in this thesis are:

- PYTHIA: parton emission is ordered by decreasing momentum with $Q^2 = p_{\text{T}}^2$, where p_{T} represents the transverse momentum of the radiating parton;

- HERWIG: angular-ordered parton emission with $Q^2 = 2E_a^2(1 - \cos\theta)$, where E_a is the energy of the parent parton and θ represents the angle between the emitted and parent parton.

Matching parton shower to matrix element

When the full event generation is performed by the PYTHIA generator, the matrix-element and parton-shower calculation are completely compatible. When the ME calculation and parton showering are done by separate generators, the matching of the two can lead to double counting effects due to the same emission being produced simultaneously by both generators. Thus, a number of approaches exist for overcoming this effect.

The MLM approach [176] is used to remove double-counting terms when interfacing the MADGRAPH ME generator with the PYTHIA parton shower generator. This matching method completely excludes events that contain emissions from the parton shower that are already calculated by the matrix element. A *matching threshold* is introduced to separate events that should be produced by the ME generator from those that are taken from the PS generator, and is defined as p_{T} of additional radiation.

The POWHEG [177] matching applies special weights to the parton-shower model according to the matrix element term, such that it acts only on the phase-space configuration defined by the leading order expansion terms.

6.1.3   Hadronisation models

The evolution of the parton showers finishes at the energy scale of the order of a few GeV, at which the strong interaction leads to the formation of colourless hadrons. Several different hadronisation algorithms exist that model this process.

String model

The *string* (or Lund) hadronisation model [178], used in the PYTHIA [175] parton shower generator, is based on the long-distance colour-confinement approach. It assumes a uniform colour potential, which depends on the spatial separation between a q$\bar{\text{q}}$ pair [179]. According to it, the pair evolves in time by increasing its mutual distance and therefore the energy of the colour potential. When the potential energy is of the order of the hadron masses, it is energetically favourable for the string to break at some point along its length, creating a new quark-antiquark pair, q$'$$\bar{\text{q}}'$.

The new antiquark is created at the end of the string segment connected to the original quark q, and the new quark for the original antiquark $\bar{\text{q}}$. The string evolution stops when the string potential energy cannot produce more hadrons, i.e. when the separation between the quark and antiquark produces a potential energy smaller than the typical hadron masses. Figure 6.2 shows a representation of the colour potential and string fragmentation evolution [179].

(a) Graphical representation of the field lines between a u quark and a $\bar{\text{d}}$ quark.

(b) Sketch of the time-space evolution of the string hadronisation model.

Figure 6.2: Graphical illustration of the string hadronisation model. Modified versions of figures from [180].

Color-preconfinement model

An alternative hadronisation technique is based on the colour pre-confinement assumption [181], i.e. colour-connected partons are produced in a collinear direction. This process is graphically shown in Figure 6.3a. The HERWIG [182] program implements this hadronisation model. It produces colourless clusters of partons with a cluster-invariant-mass distribution independent of the process [180]. The clusters are considered as proto-hadrons, which are forced to decay into hadrons via a quasi-two-body decay. In this model, gluons are identified as colour-anticolour-pair lines and are forced to produce a q$\bar{\text{q}}$ pair before being assigned to a cluster [179]. The concept of the cluster-hadronisation model is illustrated in Figure 6.3b.

6.1.4   Underlying event

In addition to the hard scattering of the partons, secondary scattering or diffraction processes between the proton remnants can occur in a proton-proton collision. These secondary effects are collectively classified as *underlying event* (UE). The UE usually involves low energy processes and therefore the models are tuned to describe the experimentally measured data.

(a) Schematic representation of the colour preconfinement assumption. The gluons are represented by a colour-anticolour line pair connected at the vertices, and quarks are represented by single-colour lines.

(b) Representation of the cluster hadronisation model. Grey circles indicate the clusters. Gluons are forced to split into q$\overline{\text{q}}$ pairs before being assigned to a cluster, which later produces final state hadrons by a quasi-two-body decay.

Figure 6.3: Graphical illustration of the cluster hadronisation model. Figures (modified) obtained from [180].

The Z2* model was derived using data collected by the CMS experiment from proton-proton collisions at $\sqrt{s} = 0.9\,\text{TeV}$ and $7\,\text{TeV}$ [183]. This UE model is used in Monte Carlo samples produced with the PYTHIA parton shower. Simulation samples that use the HERWIG parton showering use two different UE parameterizations: the HERWIG model itself and the AUET2 model [184]. The AUET2 parameterization is derived by the ATLAS Collaboration using LHC collision data [179].

6.2   Detector response simulation

While the event generators model the physics of the interactions between particles produced from the pp collision, further decays and interactions with the detector material are simulated with the GEANT4 (GEometry ANd Tracking) [185][186] package, which is interfaced to the CMS software framework. All relevant interactions with matter consisting of different materials are included. Also the full CMS detector geometry, including support structures and active material with signal readout, noise and crosstalk are simulated, as well as the magnetic field effects on the detector response are considered. The detector response is built into the digitisation step of the simulated signals. Also a full emulation of the trigger is added. The output has the same format as real collision events, which allows to use it during analysis side by side with real measured data. Like for the event generators, there is a set of tuning parameters designed to optimise the agreement between data and simulation [146].

Chapter 7

Event reconstruction

The physics analyses described in the thesis involve processes with very complicated final states that include electrons, muons, neutrinos and jets. In order to have a convenient way to analyse such objects, it is not sufficient to use the raw output of the detector. It is much more appropriate to work with higher level physics objects, which represent the actual particles, which produced the corresponding signals in the detector read-out.

The main common part of all the analyses, described in this thesis, is the presence of the $t\bar{t}$ system with the dileptonic final state. Thus, the reconstruction of all the physics objects follows the recommendations of the CMS Top Quark Analysis Group (Top PAG) [187], which are based on the objects reconstructed with the Particle Flow (PF) algorithm [188], PF-objects. The PF-particle candidates are then used as input for the identification and reconstruction of more complex objects, like jets, which are made of clusters of PF-particles, or missing transverse energy.

In the next sections more details are given about the identification and reconstruction of specific types of objects that are relevant for the described measurements.

7.1 Primary vertex reconstruction

Considering the presence of multiple proton-proton collisions per single bunch crossing, it is very important to precisely identify the primary interaction point, which is referred to as the *primary vertex* (PV). Its position is essential for the reconstruction of any other physics object, since all the initially radiated particles are assumed to originate from the primary vertex. Therefore, the calculation of the propagation direction of each particle is defined with respect to the PV position.

In particular, this is crucial for the identification of secondary vertices that correspond to b-hadron decays, which are a key ingredient of the efficient identification of $t\bar{t}b\bar{b}$ and $t\bar{t}H(b\bar{b})$ processes. The process of identification of such decays is described in detail in Section 7.5.

The collection of reconstructed tracks measured by the silicon tracker is used to identify multiple vertices in the event. Each vertex is constructed from a set of tracks based on their distance from the vertex candidate along the z axis at the

point of closest approach to the z-axis. Then the position of each vertex is fitted based on the tracks associated to it [189].

Typical criteria for the selection of a good primary vertex are the minimum number of tracks associated to it and its distance from the beam spot in the transverse and longitudinal planes. The *beam spot* is defined as the average position of the beam crossing from a large number of tracks, in order to obtain a 3-dimensional profile of the beam. In case multiple vertices fulfil the requirements, the one with highest scalar sum of p_T^2 of the associated tracks is considered as the primary vertex.

Other vertices that are displaced with respect to the PV, but consistent with it according to momentum-directions of their associated tracks, are referred to as *secondary vertices* (SV). They correspond to decays of particles produced from the primary interaction, e.g. b or c hadrons.

Pile-up events

All the remaining vertices are considered as secondary proton-proton collisions, referred to as *pile-up* events (PU). The distribution of the number of PU interactions during 2012 is shown in Figure 7.1.

Figure 7.1: Distribution of the mean number of proton-proton collisions during a single bunch crossing during 2012. Taken from [190].

Tracks associated to such PU vertices are removed from the event before reconstruction of the physics objects, in order to reduce the effect on their properties. The contribution from neutral particles is corrected at the later step of jet energy calibration, described in Section 7.4.

7.2 Particle-flow algorithm

Single particles are reconstructed combining the reconstructed tracks and vertices inside the silicon tracker, energy deposits in the calorimeters and signals from the muon system. All physics objects used in this thesis are reconstructed using the *particle-flow* (PF) algorithm [191].

The PF technique attempts to use as much information from different parts of the detector as possible when reconstructing single objects, and treats the event as a whole. Thus, every stable particle, including muons, electrons, photons, charged and neutral hadrons, is reconstructed by the PF method, and is referred to as *PF candidate*. A schematic illustration of the PF concept is presented in Figure 7.2.

Figure 7.2: Schematic view of the particle-flow concept: combination of signals from different subdetectors to construct objects that represent individual particles with defined types. Based on [192].

The PF candidates are used to reconstruct higher level physics objects, like jets, missing transverse energy, etc. as discussed in the following sections.

7.3 Lepton reconstruction

One of the important signatures of the processes described in the thesis is the dileptonic final state of the $t\bar{t}$ system. Thus, electrons and muons have to be properly identified and reconstructed.

7.3.1 Muon reconstruction

The CMS detector is specifically optimised to measure muons with very high precision. These are the only particles that can penetrate the whole detector, reaching the muon chambers outside the superconducting solenoid. Thanks to the wide minimum ionisation region and long lifetime of the muons, they leave only small energy deposits in the calorimeters, and leave hits in both the inner tracker and the muon system. This signature, referred to as *minimum ionising particle* (MIP) is an essential part of the muon identification.

Several algorithms exist for the reconstruction of muons, which use different combinations of information from the tracker, calorimeters and muon system. A schematic overview of the different algorithms is shown in Figure 7.3, while detailed performance studies can be found in [193]. The algorithms relevant for the described analyses are the *tracker muons* and *global muons*.

Tracker muons are built using the inside-out approach, starting from the tracks in the tracker. Tracks that have total momentum $p > 2.5$ GeV and transverse momentum $p_\mathrm{T} > 0.5$ GeV are considered as possible muon candidates, and are extrapolated to the muon system. The magnetic field, average expected energy losses

Figure 7.3: Schematic view of different algorithms for the muon reconstruction. Shown are the reconstruction chains for the different algorithms together with the input used by each of them. Based on the diagram from [194].

and multiple Coulomb scattering in the detector material are taken into account for the extrapolation. A track in the silicon tracker is defined as a Tracker Muon if it matches at least one muon segment in the CSC or DT [193].

Global muons are reconstructed using the outside-in approach, starting from the tracks in the muon-system, referred to as *standalone muons*. The matching to the tracks in the silicon tracker is performed by a comparison of the two tracks extrapolated to a common surface. The matching of a tracker track to the standalone muon is performed by a fit of the global-muon-track candidate using the Kalman-filter technique [195]. This algorithm is more computing-intensive than the Tracker Muons, but provides an improved momentum resolution for the muons with large transverse momentum, $p_{\mathrm{T}} \gtrsim 200\,\mathrm{GeV}$ [193].

7.3.2   Electron reconstruction

Electrons are reconstructed by combining the information from the tracking detector and the electromagnetic calorimeter (ECAL). The electron-energy measurement in the ECAL is hampered by the material of the tracker, which is distributed in front of the ECAL, and by the presence of the strong magnetic field [196].

Furthermore, electrons, traversing the layers of silicon modules in the tracker, radiate bremsstrahlung photons, and the energy reaching the ECAL has a significant spread in the azimuthal direction ϕ. The spread is taken into account in the clustering of ECAL cells and in the building of superclusters (clusters of clusters), in order to collect the bremsstrahlung energy. Supercluster-driven pixel-seed finding is then used to initiate the building of trajectories in the inner tracker using a specific energy loss modeling.

The electron energy is estimated from the combination of the tracking and calorimetry measurements. The electron direction is obtained from the associated primary track.

7.3.3 Lepton isolation

In order to reject leptons from semi-leptonic hadron decays, an isolation criterion is imposed on the lepton candidates: $I_{rel} < 0.15$. I_{rel} is defined as the sum of transverse energy deposits from charged and neutral hadrons and photons inside a cone in $\eta - \phi$ space of $\Delta R < 0.3$ around the lepton direction, relative to the transverse momentum of the lepton:

$$I_{rel} = \frac{\sum_{\text{photons}} E + \sum_{\text{hadrons}}^{\text{charged}} E + \sum_{\text{hadrons}}^{\text{neutral}} E}{p_{\text{T, lepton}}} \tag{7.1}$$

7.4 Jet reconstruction

Confinement of the strong interaction leads to the hadronisation and fragmentation of a coloured parton, which produces a collimated set of colour-neutral particles, as mentioned in Section 2.1.2. Such sets of particles are reconstructed as single physics objects, referred to as *jets*.

In this thesis, the building blocks for the jet reconstruction are particle-flow candidates, which are clustered into jets by the anti-k_{T} algorithm [197], which is a special case of sequential recombination jet algorithms. The clustering proceeds by recursively grouping pairs of entities together to form new compound entities. Following this approach, entities with nearly parallel momenta are merged until no entities that are close enough are left. The final separated entities are then defined as jets.

Technically grouping conditions are defined by the two special distance parameters:

$$\begin{aligned} d_{ij} &= \min\left(k_{\text{T}_i}^{2p}, k_{\text{T}_j}^{2p}\right) \frac{\Delta R_{ij}^2}{R^2}, \\ d_i &= k_{\text{T}_i}^{2p}, \end{aligned} \tag{7.2}$$

where k_{T} is the transverse momentum and $\Delta R_{ij}^2 = (y_i - y_j)^2 + (\phi_i - \phi_j)^2$ is the angular distance between the two entities.

The parameter p effectively defines the order in which jets are formed, and for the anti-k_{T} algorithm is set to $p = -1$. According to the defined ordering, first jets are formed from high energetic entities, which makes them insensitive to the emission of soft particles. From a theoretical point of view this is important, since soft radiation can not be calculated perturbatively and might introduce unnecessary bias to the properties of jets with respect to the properties of the original partons. From an experimental point of view this property reduces the effect of soft-particle misreconstruction, as well as spatial-resolution effects. The radius parameter R effectively defines the size of the jet cone and is set to $R = 0.5$.

Since all PF objects are clustered to jets, leptons can appear in the event twice: as individual leptons and as jet constituents. This is corrected by the *top projection* algorithm [198], which excludes all isolated leptons from the list of input objects for the jet clustering. The lepton isolation criteria are the same as defined in Section 7.3.3.

7.4.1   Energy correction

When entities constituting a jet are combined by the jet-clustering algorithm, their four-momenta are summed. Thus, the four-momentum of the jet reflects the four-momentum of the initial parton. While this is true for jets at the generator level, at the reconstruction level significant deviations can appear. This is mainly caused by the non-uniform and non-linear response of the calorimeter, missing or excess energy due to some particles being outside the jet area, or undetected neutrinos or additional clustered particles, e.g. from the pile-up interactions. In order to correct for these effects, a calibration of the measured jet momenta is applied [199], which is referred to as *Jet Energy Scale* (JES).

The JES is centrally determined by the CMS *JetMET* group [200] using a factorised jet calibration method, which consists of several steps illustrated in Figure 7.4.

Figure 7.4: Schematic view of the jet-energy-calibration process. Some steps are applied only in real data, while some also in MC simulations. Based on [200].

At the first step, an excess of jet energy caused by neutral hadrons originating from the pile-up collisions (PU) is removed, and determined separately for data and MC. Additional corrections (MC) compensate for the non-linear and non-uniform response of the calorimeters, and are determined as function of jet p_T and η using MC predictions. The remaining differences between data and simulation are compensated by residual corrections applied in data, which are parameterized as functions of p_T and η. The angular-dependent corrections are derived in dijet events exploiting the p_T-balance of the two jets. The p_T-response corrections are determined in Drell-Yan events that have additional jets, and exploit the accurate measurement of the photon and Z-boson energy in the ECAL.

The final energy correction is provided for several jet p_T and η ranges [201]. In general, the corrections are smaller than 5 % for jets in the kinematic region considered for the analyses described in the thesis.

Besides generic jet energy corrections, specific corrections for b jets were developed by a group analysing $H \to b\bar{b}$ final states [202]. Such corrections aim at the recovery of neutrinos from semileptonic b-hadron decays, but are not centrally provided by CMS, and were not used in the described physics analyses.

Jet-energy-resolution scale factors

The differences in the jet energy resolution between data and MC simulations are corrected by $|\eta|$-dependent data-to-simulation factors applied to the simulated samples [201]. The correction factors are listed in Table 7.1.

Table 7.1: Jet energy resolution (JER) correction factors with the total uncertainty, in different ranges of reconstructed jet $|\eta|$ [201].

$\boldsymbol{|\eta|}$ **range**	$0-0.5$	$0.5-1.1$	$1.1-1.7$	$1.7-2.3$	$2.3-2.8$	$2.8-3.2$	$3.2-5.0$
JER factor	1.079	1.099	1.121	1.208	1.254	1.395	1.056
$\boldsymbol{1\sigma}$ **uncertainty**	± 0.026	± 0.028	± 0.029	± 0.046	± 0.062	± 0.063	± 0.191

7.5   Tagging of heavy-flavour jets

One of the distinct signatures of the $t\bar{t}b\bar{b}$ and $t\bar{t}H(b\bar{b})$ processes, which are studied in this thesis, is the presence of 4 jets, which are initiated by bottom quarks, referred to as *b jets*. Therefore, it is crucial for the measurement to identify or tag such b jets at the reconstruction level. The identification processes is referred to as *b-tagging*.

The performance of an arbitrary b-tagging algorithm can be defined by the two quantities:

- **b-tagging efficiency:** the probability of a real b jet to be tagged by the algorithm;

- **mistagging probability:** the probability of a non-b jet, i.e. a c jet or a light-flavour jet, to be b-tagged jet.

Obviously, the best algorithm should have 100 % b-tagging efficiency and 0 % mistagging probability. For the measurements that require at least 4 b jets, it is especially important to have low mistagging probability, since the large number of considered jets exponentially amplifies the probability of selection of a wrong-flavour jet.

All the existing tagging methods rely on the unique properties of b jets compared to other types of jets: the relatively large mass of b hadrons and their large lifetime $\tau_B \sim \mathcal{O}(1\,\mathrm{ps})$, which results in travelling a distance of $\mathcal{O}(1\,\mathrm{mm})$ before decaying. Such displaced decays of b hadrons are identified as secondary vertices, which are reconstructed from tracks using their distance of closest approach to the primary vertex. The layout of the primary vertex, tracks and the secondary vertex is schematically shown in Figure 7.5a.

A number of different b tagging methods were developed at CMS that use different sets of information. The algorithm used in this thesis is the *Combined Secondary Vertex* (CSV) [203], which combines the secondary vertices and track-based lifetime information into a multivariate CSV-discriminant, which has different shapes for jets of different flavour. It was chosen due to its best performance, i.e. highest b-tagging efficiency for the mistagging probability for light-flavour jets below 1 %, as shown in Figure 7.5b. The CSV discriminant values range from 0.0 to 1.0 for the majority of jets, while values smaller than 0.0 are possible if the discriminant could not be evaluated, e.g. due to the missing tracks that correspond to the jet. More details about the performance of the CSV algorithm can be found in [204].

Three working points are centrally defined for the CSV discriminant in terms of the mistagging probability of light-flavour jets, as listed in Table 7.2. For the described analyses the Medium working point (CSVM) is used to define a b jet at the reconstruction level, which provides the 1 % mistagging probability for light-flavour jets.

(a) Schematic view of the b jet signature compared to the light jet signature.

(b) Probability of tagging a light-flavour jet as a function of different b-tagging-efficiency values for the different b-tagging algorithms [203].

Figure 7.5: Schematic view of the distinctive features of b jets, which are exploited by b tagging algorithms and comparison of the performance of the b-tagging algorithms developed at CMS.

It should be kept in mind, that c hadrons also have a substantial lifetime, while on average it is smaller than that of b hadrons. Therefore, the mistagging probability for c jets is in general higher than for light-flavour jets.

Table 7.2: CSV-discriminant thresholds for the different working points of the algorithm [204] with the corresponding light-jet misidentification and b-jet identification efficiencies.

Working point	Threshold	Mistag probability	b-tagging efficiency
Loose (CSVL)	≥ 0.244	$10\,\%$	$80\,\%$
Medium (CSVM)	≥ 0.679	$1\,\%$	$65\,\%$
Tight (CSVT)	≥ 0.898	$0.1\,\%$	$50\,\%$

7.6   Missing transverse energy

Another distinctive feature of the dileptonic final state of the $t\bar{t}$ event is the presence of neutrinos from W-boson decays. Individual neutrinos, being weakly interacting particles, can not be measured directly by the detector. Nevertheless, their presence, as well as the vectorial sum of their transverse momenta, can be estimated using the p_{T} balance of all particles measured by the detector.

Experimentally, the missing transverse momentum $\vec{E}_{\mathrm{T}}$ is defined as the negative vectorial sum of the transverse momenta of all the leptons and jets present in the event. Since all the $t\bar{t}$ events with dileptonic final state are expected to have at least two neutrinos, the actual direction of the $\vec{E}_{\mathrm{T}}$ would reflect the vectorial sum of

directions of all the neutrinos in the event. Its magnitude, referred to as the missing transverse energy $\not{E}_T$, provides the estimate of the energy of the neutrinos, and is defined as:

$$\not{E}_T = \sqrt{\not{E}_{T,x}^2 + \not{E}_{T,y}^2} = \left| -\sum_{\text{leptons}} \vec{p}_T - \sum_{\text{jets}} \vec{p}_T \right| .\qquad(7.3)$$

Since the $\not{E}_T$ is calculated from the reconstructed physics objects, it is very sensitive to possible detector malfunctioning, mismeasurements or misidentification of the detected particles.

The resolution of the $\not{E}_T$ depends on the measured lepton- and jet-energy resolution, which makes it sensitive to contributions from pile-up interactions. Thus, the MVA, used in the thesis, is reconstructed using a multivariate algorithm, described in [205], which combines jet-, PU- and event-related information to achieve best resolution in different conditions.

Differences in the reconstructed $\not{E}_T$ between data and MC simulations are compensated by a *recoil correction*, proposed in [206]. It is applied only to the simulation of Drell-Yan events, for which the correction was developed, and which do not have intrinsic $\not{E}_T$. The correction provides a significant improvement of the agreement between data and Drell-Yan MC simulation in the region of small missing transverse energy, as shown in Figure 7.6.
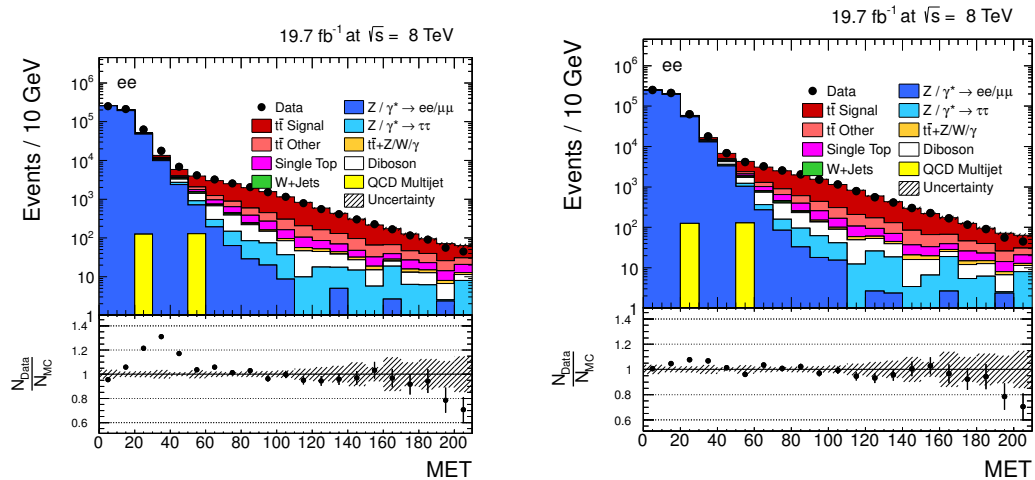
Figure 7.6: Missing transverse energy distribution in events with at least two opposite-sign isolated electrons. The distributions are shown before (left) and after (right) applying the recoil correction to Drell-Yan Monte Carlo events. Taken from [179].

Chapter 8

Heavy-flavour jet identification in simulation

With high energy collisions at the LHC, much attention is payed to jets that are produced due to the hadronisation of heavy flavour quarks, namely *heavy-flavour jets*. Such jets are initiated by bottom or charm quarks and are often an important part of signatures in many processes, e.g. top quark production, Higgs boson production etc. Since nearly every analysis at CMS uses Monte Carlo simulations of processes that contribute to the measurement, it is necessary to precisely identify such jets at generator level in simulated samples.

This is a quite straightforward task in case there is a single source of such heavy-flavour jets or the origin of each particular jet is not important. But there is a range of processes that have multiple heavy-flavour jets in the final state originating from different particles. An example of such process can be $t\bar{t}H(b\bar{b})$ or $t\bar{t}Z(b\bar{b})$ production with decays to at least 4 b quarks: 2 from the $t\bar{t}$ pair and 2 from the H or Z boson decays. Exactly the same final state can be a result of $t\bar{t}b\bar{b}$ production, where a second pair of b jets comes from additional gluon radiation, and would be an irreducible background to $t\bar{t}H(b\bar{b})$ and $t\bar{t}Z(b\bar{b})$ production. In such events, identification of a jet flavour alone is not enough because it is important also to know where each jet originates from.

Another aspect of such events is that due to limited spatial resolution of clustered jets they can overlap. Usually this happens if a pair of heavy flavour quarks is produced from collinear gluon splitting, leading to a very small spatial separation between the quarks, which can not be resolved at the jet level. Such processes are usually predicted with much lower precision than production of well separated heavy-flavour jets, therefore it is needed to deal with such jets with special care.

It is possible to find the origin of each heavy-flavour jet in MC simulations, which have complete evolution of initial particles down to final state particles that constitute a jet. Nevertheless, the methods that are used by the majority of analyses to distinguish jets from different processes are rather limited, and usually rely on finding a closest parton in $\eta - \phi$ space within some cone around the jet. Standard CMS tools [207] allow to identify a flavour of any jet or even a parton, associated to it, but their efficiency is not high enough in events with many heavy-flavour jets.

These issues are especially relevant for the $t\bar{t}b\bar{b}$ and $t\bar{t}H(b\bar{b})$ analyses described in this thesis. Therefore, a dedicated tool, $GenHFHadronMatcher$ [208], has been

developed as part of this thesis, which aims at the matching of each heavy-flavour jet
to the original b quark, and finding the origin of the b quark. This method consists
of the two main parts:

- matching hadrons to quarks;

- matching hadrons to jets.

The latter part is mainly performed by the standalone CMS jet-flavour-identification
tool, `JetFlavour` [207]. Therefore, the main task addressed by the tool developed in
the scope of this thesis is the matching of b hadrons to their corresponding partons.

Furthermore, in the context of $t\bar{t}b\bar{b}$ and $t\bar{t}H(b\bar{b})$ measurements, a study of meth-
ods for jet-charge identification at reconstructed level was performed by the DESY
ttH group [209]. The jet charge is meant to improve identification of b jets stemming
from the top-quark decays. Since leptons stemming from b-hadron decays provide a
more robust information on the charge of the original b hadron, identification of such
leptons was also implemented in the $GenHFHadronMatcher$. This information is
necessary for a proper calibration of the jet-charge identification method.

8.1   Jet flavour identification

If there is a heavy flavour quark produced, it fragments into a hadron, which later
decays, producing a shower of stable particles, which form a jet. It can also happen
that decay products of a single hadron are clustered to several different jets, making
it difficult to decide which of the jets should correspond to the heavy flavour hadron.
Usually the jet with highest p_T is taken as the jet representing the hadron, but this
might not always be the best solution.

During the development of this tool, another method was introduced, exploiting
the jet clustering algorithm that is used to produce the jets themselves. This is done
by adding so called *ghost hadrons* to the list of particles that are used to build jets.
The ghost hadrons have a negligibly small energy in order to not affect the energy of
the jet. With this approach, only the direction of the hadron is taken into account,
and the clustering algorithm decides into which jet the hadron should be clustered.
Such jets are then associated to the identified hadrons by finding them among the
jet constituents.

The hadron-jet matching method naturally allows to have several hadrons being
clustered to a single jet, meaning that the hadrons are so close to each other that
their decay products end up in the same jet, e.g. b hadrons from collinear gluon
splitting.

This approach was first implemented as a part of the $GenHFHadronMatcher$
tool to associate heavy flavour hadrons to jets. But later this method has been
adopted in the recent version of the official $JetFlavour$ tool [207] with a more
elegant technical implementation. It is based on the same ghost particle injection
technique to match partons, hadrons and leptons to jets.

Therefore, the $GenHFHadronMatcher$ is using the $JetFlavour$ tool for associ-
ation of heavy flavour hadrons to jets. Nevertheless, hadrons alone are not sufficient
for identification of the origin of each hadron associated to the jet. Thus, the origin
of each heavy flavour hadron must be identified.

8.2   Heavy-flavour hadron origin identification

The identification of the origin of a b hadron was developed to be as generic as possible, and works with the PYTHIA, HERWIG and SHERPA parton-shower generators. It is also independent from different particle-status schemes, which are different between versions 6 and 8 of the PYTHIA generator.

8.2.1   Particle chain scanning

In order to find a quark from which the b hadron has originated, the complete particle chain stored in a simulated sample is analysed for each generator-level jet. Generated particles are characterised by their *pdgId*, which are defined according to the Particle Data Group numbering scheme [210] employed by all major MC generators. According to this scheme, a particle is a b hadron if its *pdgId* is of type 5XX (b meson) or 5XXX (b baryon), where X is a digit from 0 to 9.

The scanning of the particles starts with a loop over constituents of a certain jet. For each jet constituent, the *pdgId* of its mother particle is checked. Then a mother of the mother is checked, and so on, until a b hadron is found. Only weakly decaying b hadrons are considered as final entities that are associated to jets. Such hadrons must have no sibling b hadrons of the same flavour. Otherwise, the b hadron is treated as intermediate, and is not associated to any jet. If the weakly decaying b hadron has been found, it is associated to the jet and the checking of mother particles continues until a particle is found that has no mothers any more. Apart from rare MC storage artefacts, this final particle is the incoming proton participating in a pp collision. In such a way, the history of each hadron is traced back through the particle chain in a MC simulation.

One of the major complications in this scanning method is related to the way particles produced by parton showers are stored in MC samples. A fragment of a typical particle chain as produced by PYTHIA parton-shower generator is shown in Figure 8.1. In such a chain, the scanning algorithm has to proceed from the b hadron to the initial top quark, which represents its origin. This path can be divided into two main parts, one of which is produced by a ME generator and another one by a parton-shower generator. The part produced by a ME generator provides unique relations between daughter and mother particles. In contrast to it, the part produced by a parton-shower, e.g. by PYTHIA, contains a *string* object, which has no unique relation to its ancestors. A string can have multiple mother particles, which, for example, can be multiple b quarks coming from different origins. In a HERWIG parton shower, an object with similar properties is present, which is called *cluster*, and also has multiple mother particles. This makes the scanning process much more complicated.

In order to address this feature, whenever a particle in a chain has multiple mothers, the scanning process splits into the corresponding number of branches, each of which proceeds further until the end of the particle chain is reached. This ensures that all particles participating in the process of creation of the b hadron are analysed.

After the whole chain is analysed, relations between all ancestor particles of each weakly decaying hadron are recorded into a separate list. This list is processed once more to find the origin of each b hadron, as shown in Figure 8.2. At first, the last b
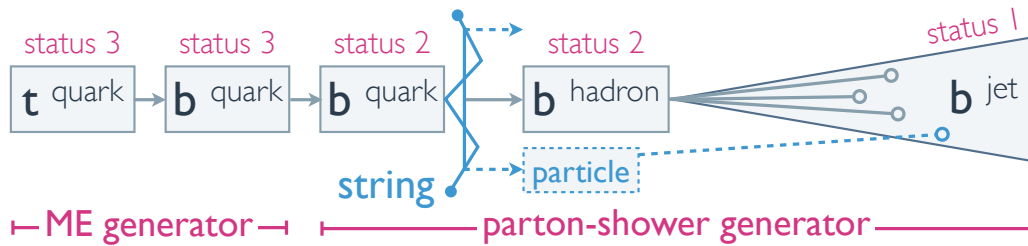
Figure 8.1: Schematic view of a part of a typical particle chain in a sample simulated
by PYTHIA 6 parton shower, which starts from a top quark and ends with
final state particles that enter a jet clustering algorithm. Particles produced
by the ME generator have status 3, while particles with status 2 or 1 are
produced by the parton-shower generator. Stable particles have status 1.

quark in the chain is found for each b hadron. Then the mother of the last b quark
is treated as the origin of the b hadron and is used in the analysis to differentiate
between b hadrons coming from different sources.
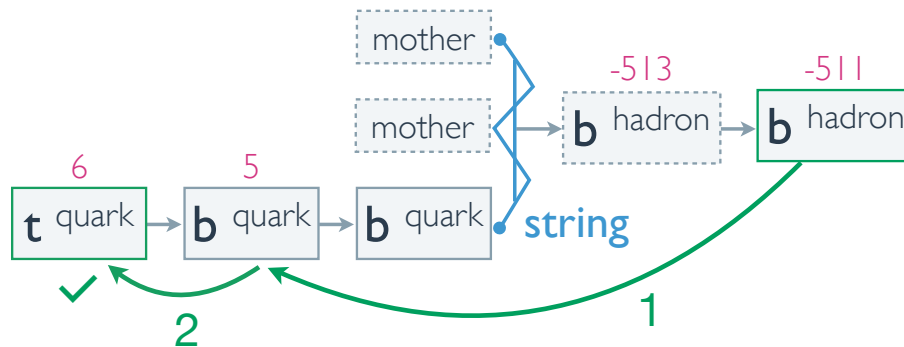
Figure 8.2: Schematic view of the origin-finding process for a b hadron in a sample with
PYTHIA 6 parton shower. The numbers above the boxes denote the *pdgId* of
corresponding particles. The step 1 denotes the search for a unique b quark
that fragmented into the b hadron. The step 2 denotes retrieving a mother
of the b quark, which is treated as the origin of the b hadron.

8.2.2 Ambiguity resolution

The fact that a string has multiple mothers, introduces an ambiguity to the process
of finding the ancestor b quark for each b hadron. In order to illustrate the process,
two MC simulations are compared: $t\bar{t} + jets$ and $t\bar{t}H(b\bar{b})$. The details of the two
samples are listed in Table 8.1, and allow to see how the complexity of the ambiguity-
resolution method depends on the number of b quarks in an event.

As demonstrated by the plots in Figure 8.3, a single b hadron can have more
than one candidate b quark matched to it, if only relations between particles are
checked as they are stored in an event record of a simulated sample. In particular,
about 11% of b hadrons in $t\bar{t} + jets$ events have multiple candidate b quarks, while
in $t\bar{t}H(b\bar{b})$ events with at least 4 b quarks per event, the number of ambiguous cases
increases to even 39%.

Table 8.1: MC samples used for the illustration of the b hadron origin identification process

Process	ME generator	Parton shower	# entries
$t\bar{t} + jets$	MADGRAPH	PYTHIA	62 131 965
$t\bar{t}H(b\bar{b})$	PYTHIA	PYTHIA	1 000 008
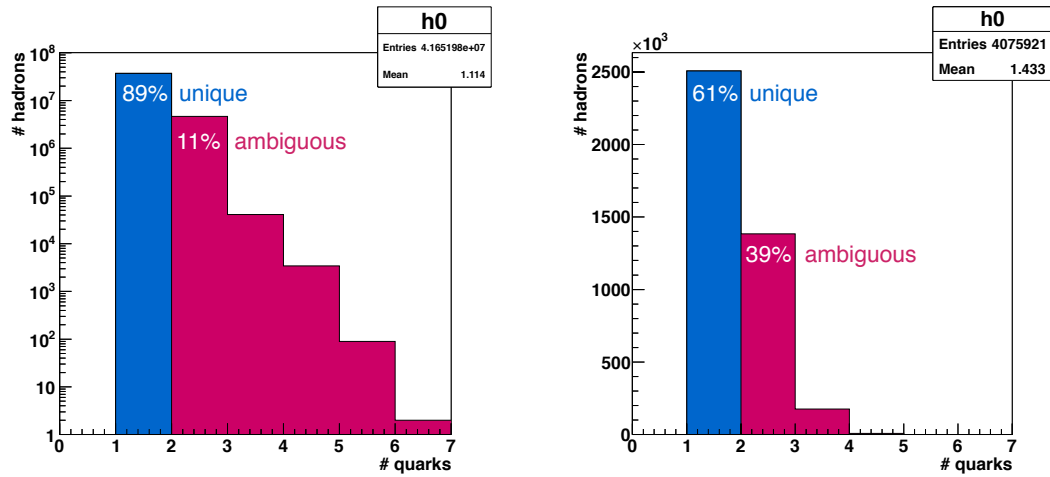
Figure 8.3: Multiplicity of distinct b quarks associated to a single b hadron after scanning the particle chain in the $t\bar{t} + jets$ (left) and $t\bar{t}H(b\bar{b})$ (right) samples using the PYTHIA parton shower.

The first step in resolving the ambiguous cases is to apply the requirement of hadron-quark flavour sign preservation, as shown in Figure 8.4a. Thus, in the case of multiple b quark candidates found as ancestors of a single b hadron, only b quarks that have the same flavour sign as the b hadron are considered. In terms of *pdgId* this means that a b hadron should have the same sign as the corresponding ancestor b quark if the hadron is a b baryon and the opposite sign if it is a b meson.

(a) requirement of the hadron-quark flavour sign preservation as applied in cases with multiple ancestor-quark candidates for a single hadron

(b) an example of a possible hadron-quark configuration with multiple ancestor-quark candidates of the same flavour sign as a hadron
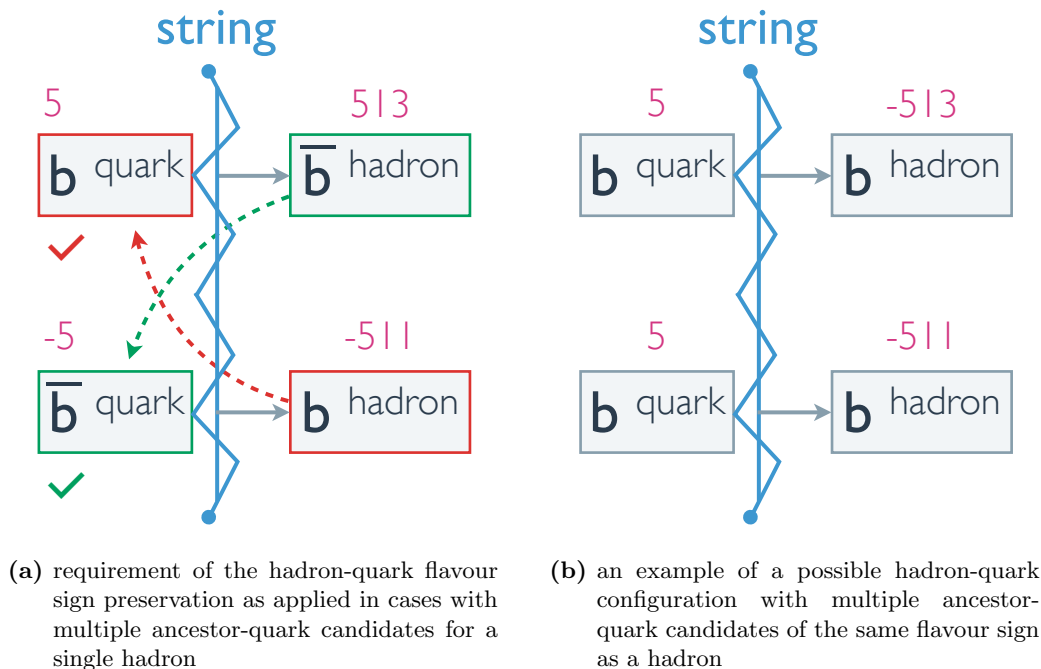
Figure 8.4: Diagrams of possible ambiguities in MC simulation that are resolved with unique hadron-quark assigment

After applying the requirement of the quark-hadron flavour sign preservation, most of the ambiguities are resolved, as shown in Figure 8.5. In the $t\bar{t} + jets$ sample 98% of b hadrons have a unique b quark associated to it that can be used to uniquely identify its origin. Nevertheless, in the $t\bar{t}H(b\bar{b})$ events 14% of b hadrons still have more than one b quark candidate with proper flavour.

Since the requirement of the quark-hadron flavour sign preservation is applied to all b quark candidates, a quark can be rejected even if it is the only candidate for the b hadron. This is illustrated in Figure 8.6, which shows the number of quarks with the proper flavour sign only for hadrons that have exactly one b quark candidate, which correspond to the first bin of the distributions in Figure 8.3.

In the remaining ambiguous cases, no decision can be made using the particle relations alone. Therefore, in order to assign a unique quark to the ambiguous hadron, the ΔR distance between the hadron and each candidate ancestor quark in $\eta - \phi$ space is calculated. The distances allow to find the quark that is closest to the hadron. This closest quark is then assigned to the hadron for identification of its origin. The performance of this approach is illustrated in Figure 8.7, which shows the 2D distribution of the distance of the hadron from its closest quark candidate versus the distance from the second closest quark candidate. For most of the ambiguous b hadrons, one b-quark candidate is much closer than the other one. Only in $\approx$
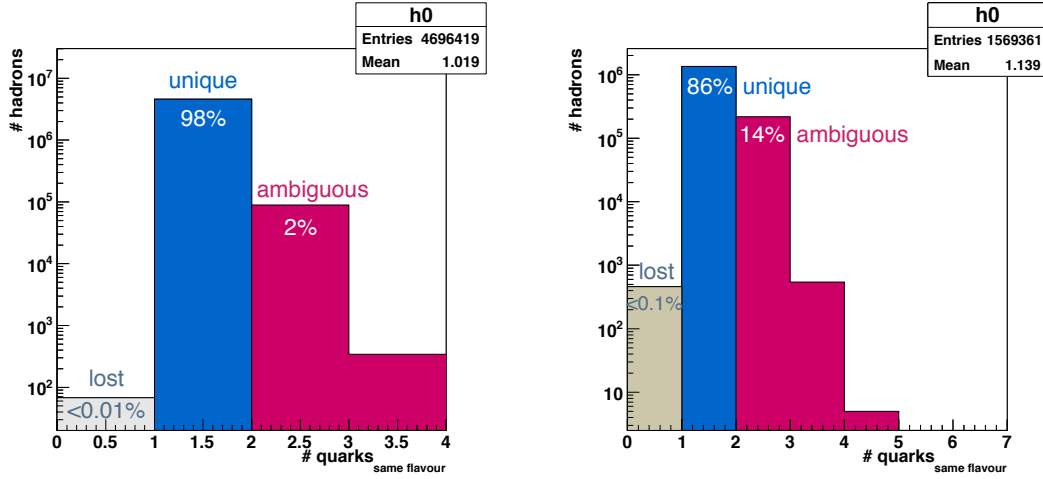
Figure 8.5: Multiplicity of distinct b quarks associated to a single b hadron after the requirement of the flavour-sign preservation in the $t\bar{t}+jets$ (left) and $t\bar{t}H(b\bar{b})$ (right) samples; only for hadrons that have more than one candidate b quark. A small fraction of the hadrons do not have a quark candidate with the proper flavour and are assigned with an undefined origin.
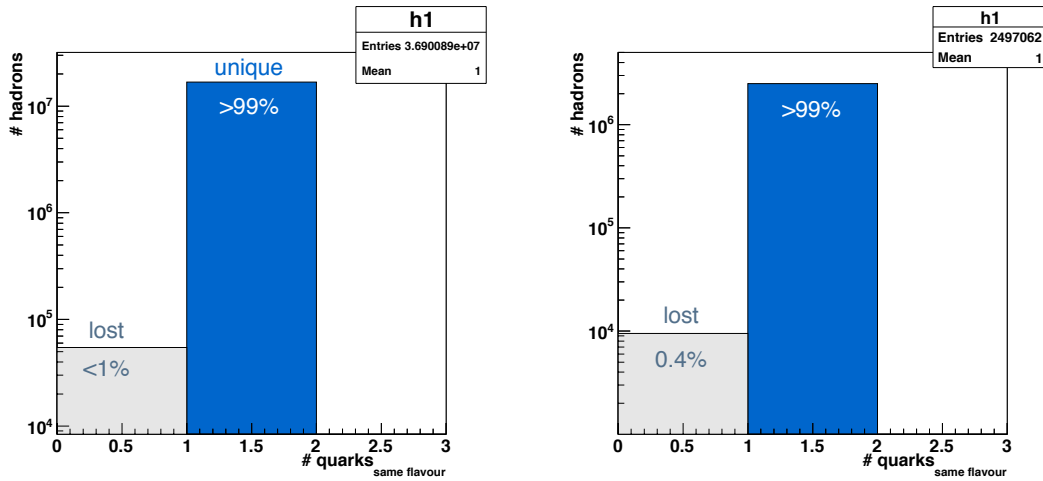
Figure 8.6: Multiplicity of distinct b quarks associated to a single b hadron after requirement of flavour sign preservation in $t\bar{t} + jets$ (left) and $t\bar{t}H(b\bar{b})$ (right) samples. A small fraction of hadrons do not have a quark candidate with the proper flavour and have undefined origin ($flavour = 0$).

4% of the ambiguous cases both quarks have similar distance from the b hadron:
$\Delta R_2/\Delta R_1 < 2$, leading to an almost random ancestor-quark assignment.

This 4% fraction of ambiguous cases can potentially be reduced even further by
giving preference to a quark with p_T closer to that of the hadron. Nevertheless, this
was not implemented in the $GenHFHadronMatcher$, since the absolute fraction of
such cases is below 0.6% in $t\bar{t}H(b\bar{b})$ events and completely negligible ($< 0.01\%$) in
$t\bar{t} + jets$ events.
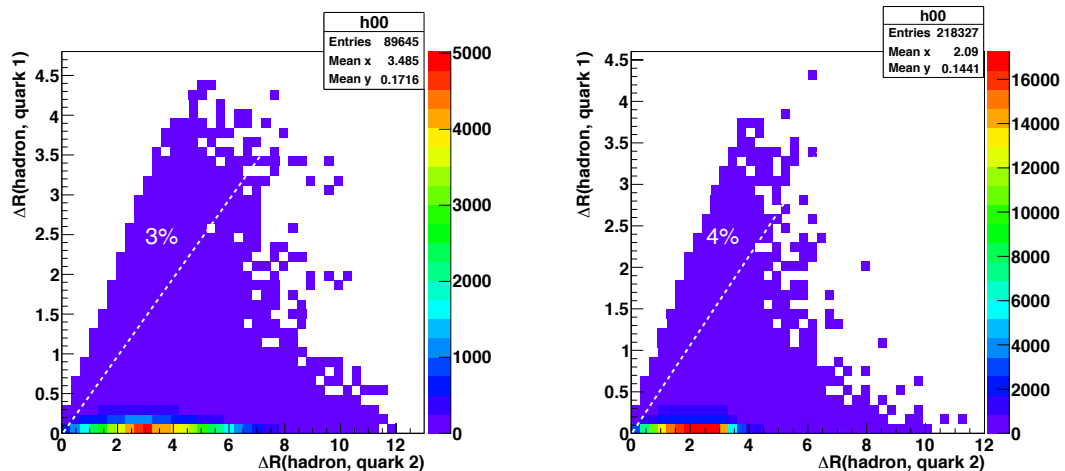
Figure 8.7: Correlation of ΔR distance between a b hadron and the first (Y axis) and
the second (X axis) closest b-quark candidate in $t\bar{t} + jets$ (left) and $t\bar{t}H(b\bar{b})$
(right) MC samples. The distributions represent only quarks that have the
same flavour sign as the hadron.

Thus, it can be concluded that this method finds a unique ancestor b quark
for more than 99% of the b hadrons, which should be sufficient for most analyses
working with b quarks.

8.2.3 Treatment of special cases

Although the resolution of all the ambiguities in the hadron $\rightarrow$ quark matching
process is covered by the step described before, there are some special cases that
can depend on a specific generator and need a dedicated treatment.

Hadron flavour oscillations

Since neutral b mesons can oscillate between their matter and antimatter forms,
this effect can be reflected in a particle chain of a simulated MC sample. In a
PYTHIA parton shower only the original b meson is stored, while its oscillated version
is skipped and the decay products of the oscillated meson are stored directly as
daughters of the original meson.

In contrast, HERWIG parton showers store both the original and the oscillated b
mesons in the particle chain with the oscillated version being a direct daughter of the
original hadron and having a $pdgId$ of the opposite sign. This leads to the fact that
in particle chains produced by the HERWIG parton shower the first b hadron found

during the scanning procedure can have flavour sign opposite to the corresponding b quark. In order to account for such flavour oscillations during the scanning process, the flavour of each b hadron is compared to the flavour of the ancestor b hadron, if it exists, and is inverted when the oscillation is detected. The inverted flavour is then used later for finding a quark with the proper flavour sign. This process is shown schematically in Figure 8.8a.
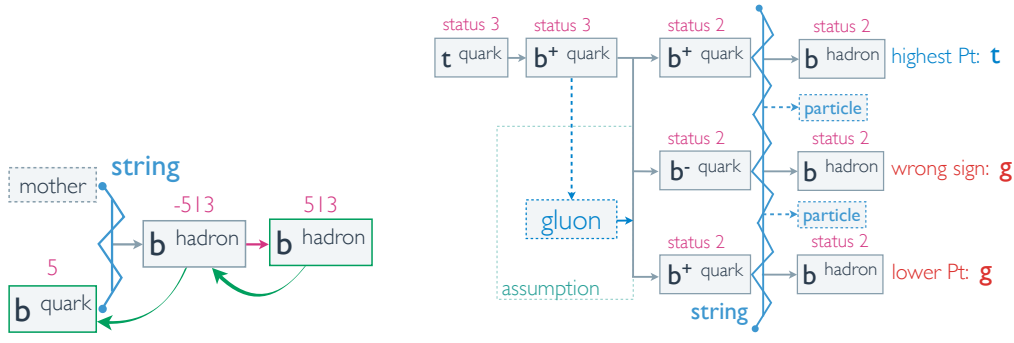
(a) identification of b-meson oscillation, which happens in particle chains produced by the HERWIG parton-shower generator

(b) an extra $b\bar{b}$ pair stored as a daughter of a single status 3 b quark; found in particle chains produced by PYTHIA 6 parton shower

Figure 8.8: Schematic views of special cases that require the dedicated treatment by the *GenHFHadronMatcher*.

Redundant $b\bar{b}$ pairs

Another strange occurrence, is the presence of multiple status 2 b quarks being daughters of a single status 3 b quark, which was found in particle chains produced by PYTHIA. Particles with status 3 are produced by the ME generator, while particles with status 2 are produced by the parton-shower generator. In about a dozen of such cases, which have been explicitly checked, the particle chain looks like in Figure 8.8b. In these cases there is a b quark that has status 3, which should normally have a single daughter that is a b quark with status 2. But in these problematic cases there is an additional pair of b and $\bar{b}$ quarks with status 2, which are also daughters of the same status 3 b quark. Since each of these three b quarks then hadronises into a distinct b hadron, they can not be ignored. In order to have a single b hadron originating from the status 3 b quark, an assumption is made that the two extra b quarks originate from a gluon radiated by the status 3 b quark, which is not stored in the chain for some reason. Among the two status 2 b quarks that have the same flavour sign as the status 3 b quark, the one that has higher p_T is treated as originating from the status 3 b quark. The other 2 b hadrons are treated as coming from a gluon splitting.

Infinite particles loops

The last issue with certain parton showers is an infinite particle loop, which is illustrated in Figure 8.9. Such loops occur when a particle has a mother that is also one of its siblings. Obviously this is unphysical. Nevertheless, such cases were

noticed in samples using the MC@NLO ME generator or a SHERPA parton shower. In such cases, the scanning algorithm would go back through the chain and ask each particle for its mother again. If no special treatment is done, the problematic particle will be scanned over and over again, until the memory is filled up and the program crashes.

This issue is solved by keeping track of all particles that have been analysed in each single branch of the particle chain. Thus, every mother is checked whether it has already been checked earlier in the same chain during the scanning process. If it has already been analysed, the scanning process for the given branch is interrupted and scanning of the next branch is initiated, if it exists.
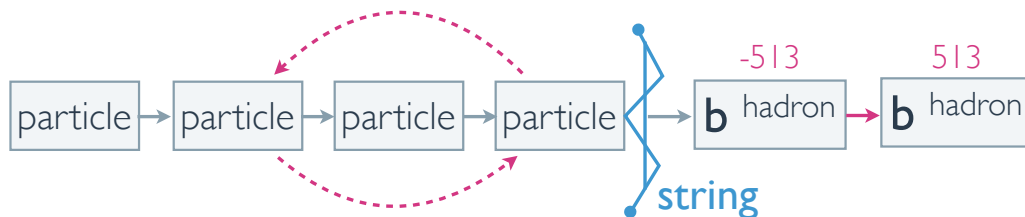
Figure 8.9: Schematic view of an infinite particle loop. The dashed arrows show the daughter-mother relations of two particles that create an infinite loop for the scanning process.

8.3 Test of performance

As stated earlier, for more than 99% of the b hadrons the $GenHFHadronMatcher$ tool finds a unique b quark that fragmented into the hadron and a mother of the b quark is defined as the origin of the b hadron. Having each b hadron matched to a generator level jet by the $JetFlavour$ tool, the flavour of each jet is defined as well as its origin. The jets are clustered from all final-state particles in the event, including neutrinos, using the anti-k_T algorithm [197] with the cone parameter $R = 0.5$. Neutrinos originating from W-boson decays are excluded from particles used for the jet clustering. The performance of the new method has been tested on the two samples listed in Table 8.1, which have a different minimal number of b quarks per event:

- $t\bar{t} + jets$: 2 b quarks from the $t\bar{t}$ pair decay

- $t\bar{t}H(b\bar{b})$: 4 b quarks (2 from the $t\bar{t}$ pair decay, 2 from the H decay)

8.3.1 Hadron decay identification

The easiest part of the $GenHFHadronMatcher$ is the matching of a b hadron to a jet, since it is mainly covered by the existing $JetFlavour$ tool. Since a b jet consists predominantly of decay products of the corresponding b hadron, the p_T of the jet should be close to that of the hadron. This comparison is demonstrated in Figure 8.10 and shows that the majority of b hadrons constitute 60-100% of the p_T of the jet associated to it. The tail in low values of the p_T ratio can be caused by

additional radiation contributing to the same jet. The cases with the b-hadron p_T being higher than that of the jet is a result of the possible wide decay angle of the hadron, which leads to only a fraction of its decay products being clustered to a single jet.
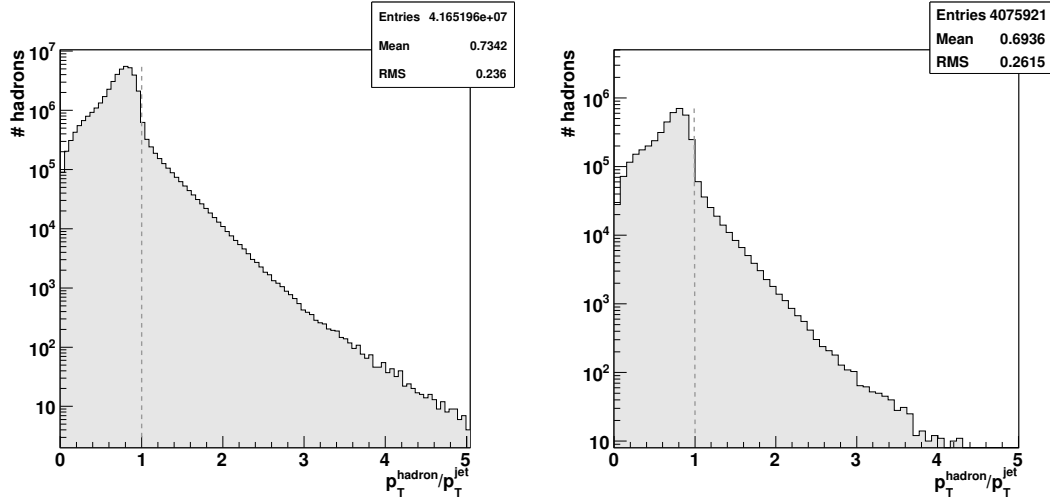
Figure 8.10: Ratio of the p_T of a b hadron over the p_T of the jet associated to it by the jet clustering algorithm in $t\bar{t} + jets$ (left) and $t\bar{t}H(b\bar{b})$ (right) events. Shown are all found b hadrons regardless of their origin.

The $GenHFHadronMatcher$ is also capable of identifying charged leptons stemming from b-hadron and c-hadron decays. Hence, the multiplicity of leptons among decay products of each b hadron is shown in Figure 8.11. The lepton can be either an electron or a muon, while a b hadron can also have a τ lepton among its decay products. Therefore, for each identified lepton a flag is stored, which tells whether it originates from a direct b hadron decay or via a τ lepton.
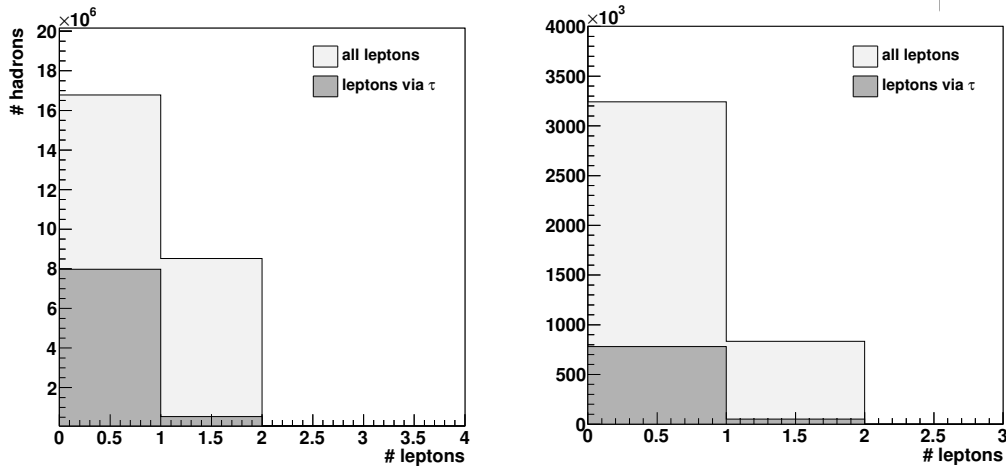
Figure 8.11: Multiplicity of leptons among decay products of each b hadron in $t\bar{t} + jets$ (left) and $t\bar{t}H(b\bar{b})$ (right) events. The multiplicity of leptons via τ is shown only for b hadrons that have at least one lepton among its decay products.

8.3.2 Comparison to ΔR matching

For testing purposes, a simple quark-hadron-jet ΔR matching method has been
compared to the approach based on the *GenHFHadronMatcher*. This method
relies on the spatial distance between quarks, hadrons and jets in $\eta - \phi$ space by
treating two objects that are closest in ΔR as matched. This usually implies some
maximum allowed distance between two objects for the distance matching to be
trustworthy, which can reduce matching efficiency.

Quark → hadron → jet matching

In the ΔR method quark-hadron-jet matching is done in two steps. First for each
b quark in the event the closest b hadron is found within a cone $\Delta R = 0.5$ around
the quark. If there is such a hadron, the closest jet must be found within a cone
$\Delta R = 0.5$ around the hadron. At each of these steps, ΔR cut can lead to a reduced
matching efficiency and possible mismatching.

To demonstrate the efficiency reduction, for each b quark in the event, a matched
b hadron is found by each of the two methods. Then the ΔR distance between a
quark and the matched hadron is compared between the methods. In case of the
ΔR matching, this is just the closest b hadron regardless of the distance between
the two, while in case of the *GenHFHadronMatcher* this is the b hadron matched
to the quark after scanning the whole particle chain. The comparison is performed
using the $t\bar{t} + jets$ and $t\bar{t}H(b\bar{b})$ MC samples, with a b quark originating either from
a top quark decay or from a Higgs boson decay.

The distribution of the distance is shown in Figure 8.12. It looks almost identical
for both methods, which means that for most of the b hadrons the same b quark is
assigned. Nevertheless, any matching method that is relying on the distance between
two objects requires some limit on the allowed distance. For our comparison we have
used the $\Delta R < 0.5$ cone, which means that all hadrons with the closest quark being
further than $\Delta R(hadron, quark) = 0.5$ will not have a b quark assigned. This
leads to the reduction of the matching efficiency by about 6% for the ΔR method
compared to the *GenHFHadronMatcher*.

Once a b hadron is identified for each b quark of interest, we can look at the
number of distinct b hadrons from top quark and Higgs boson decays. In $t\bar{t} + jets$
events, there are 2 b quarks from the $t\bar{t}$ system decay and each of them should have
a separate corresponding b hadron. In $t\bar{t}H(b\bar{b})$ events, two additional b quarks exist
from the Higgs boson decay leading to 4 distinct b hadrons in case of the ideal hadron
origin identification.

Thus, the multiplicity of unique b hadrons originating explicitly from the top
quark and the Higgs boson decays was calculated separately in $t\bar{t} + jets$ and $t\bar{t}H(b\bar{b})$
samples. The distributions are shown in Figure 8.13.

As can be seen from the distribution for the $t\bar{t} + jets$ events in Figure 8.13a, the
GenHFHadronMatcher finds 2 distinct b hadrons in 100% of events, which is 12%
higher than the ΔR matching with the $\Delta R(hadron, quark) < 0.5$ criteria applied.
Even without the ΔR cut, it still performs a few percent worse, which means that
sometimes a single b hadron is closest to the both b quarks from the $t\bar{t}$ decay.

The improvement from the *GenHFHadronMatcher* is even higher in $t\bar{t}H(b\bar{b})$
events with higher b quark multiplicity, where it still finds all 4 unique b hadrons in

(a) 2 b quarks from top-quark decays in $t\bar{t} +$ *jets* events

(b) 2 b quarks from Higgs-boson decays in $t\bar{t}H(b\bar{b})$ events
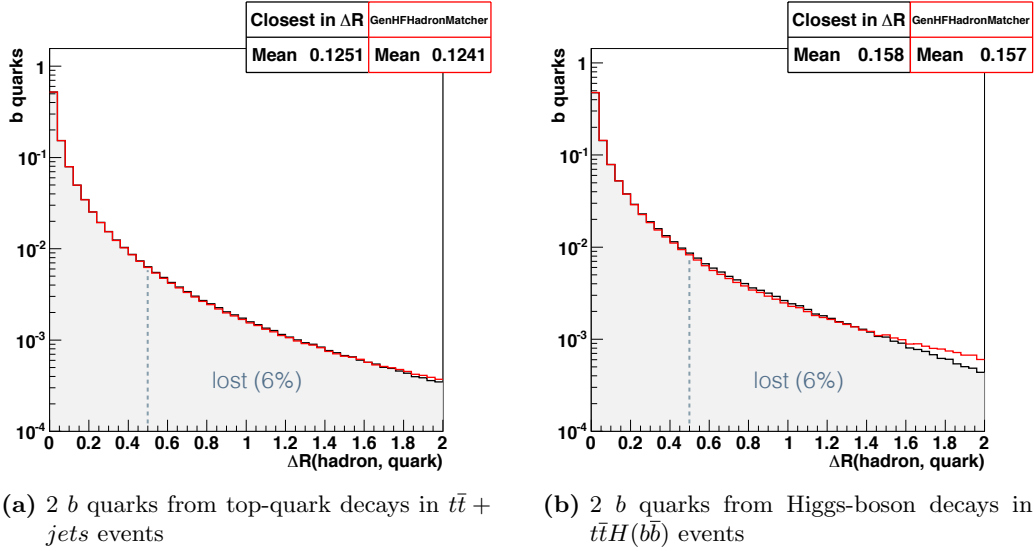
Figure 8.12: Distance between each b hadron and the b quark matched to it by the ΔR method (black) and *GenHFHadronMatcher* (red). Quarks at $\Delta R(hadron, quark) > 0.5$ are lost if the ΔR method is used. Mostly the same hadron is matched to b quarks by the both methods.

(a) 2 b quarks in $t\bar{t} + jets$ events

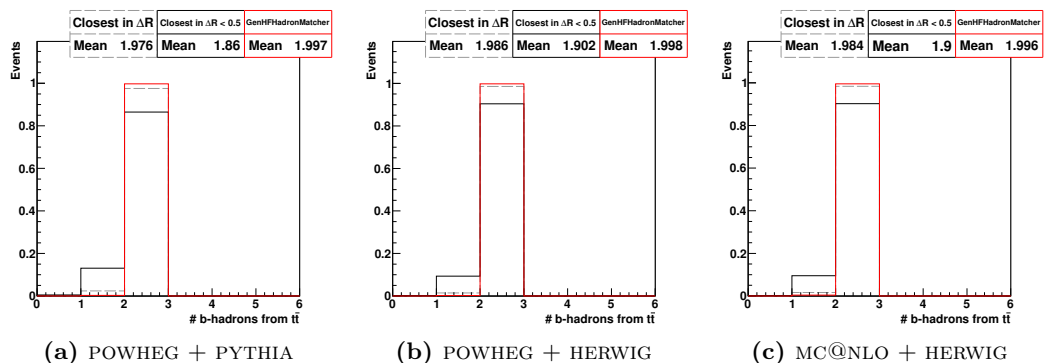**(b)** 4 b quarks in $t\bar{t}H(b\bar{b})$ events
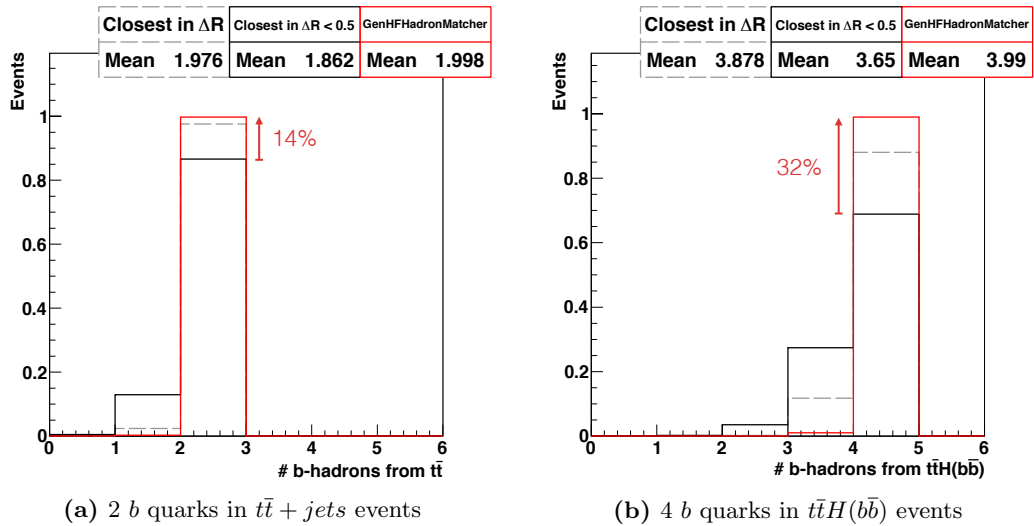
Figure 8.13: Number of distinct b hadrons from top quark and Higgs boson decays, as identified by the three methods: closest hadron to each b quark (dashed), closest hadron to each b quark within $\Delta R(hadron, quark) < 0.5$ cone (black), chain scanning by the *GenHFHadronMatcher* (red).

nearly 100% of events. The efficiency of the ΔR matching is 32% lower. Even if no ΔR criteria is applied, its efficiency is still about 10% lower compared to that of the $GenHFHadronMatcher$.

The next step is the identification of unique b jets at generator level by matching them to the identified distinct b hadrons. In case of the ΔR matching, a closest jet is found for each distinct hadron with or without application of the $\Delta R(hadron, jet) <$ 0.5 requirement. In case of the $GenHFHadronMatcher$, a jet is assigned to each b hadron by the jet clustering algorithm, as implemented in the $JetFlavour$ [207] tool.

The distributions of multiplicity of distinct jets associated to the previously identified b hadrons are shown in Figure 8.14. The inefficiency of the hadron identification shown in Figure 8.13 translates into a similar inefficiency at the jet level.

(a) 2 b quarks in $t\bar{t} + jets$ events   **(b)** 4 b quarks in $t\bar{t}H(b\bar{b})$ events
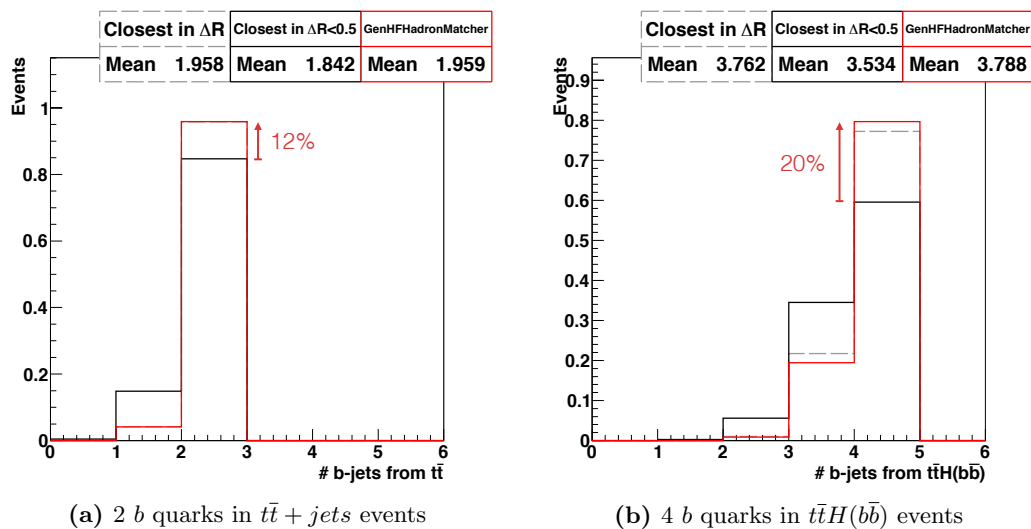
Figure 8.14: Number of distinct b jets associated to b hadrons from the top quark and the Higgs boson decays, as identified by the three methods: closest jet to each b hadron identified by the ΔR matching (dashed), closest jet to each b hadron identified by the $\Delta R(hadron, quark) < 0.5$ matching and is within $\Delta R(jet, hadron) < 0.5$ cone (black), jet matched to hadrons identified by the $GenHFHadronMatcher$ and associated by the jet clustering algorithm via the $JetFlavour$ tool (red).

The improvement of the new method is of the same order as it was for the b hadrons, but slightly smaller due to the lower spatial resolution for jets. Thus, two hadrons can be not far enough from each other, i.e. at $\Delta R(hadron_1, hadron_2) < 0.5$, leading to the high probability of the two hadrons being matched to the same jet, which was clustered with the $R = 0.5$ parameter [197]. This is the source of minor contributions in bins below 2 for the $t\bar{t} + jets$ sample and in bins below 4 for the $t\bar{t}H(b\bar{b})$ sample in Figure 8.14.

It can also be seen that if one would not use any constraint on the quark-hadron and hadron-jet distances, the same number of distinct b jets is found by the ΔR matching (dashed line) and the new method (red line) in the $t\bar{t} + jets$ sample. This means that the slightly more precise identification of b hadrons by the $GenHFHadronMatcher$ is absorbed by clustered jets and is not visible any more. Nevertheless, there is a small difference visible in $t\bar{t}H(b\bar{b})$ events, where it is more

probable to match such a wrong b hadron that it can not be absorbed by the jet clustering.

Quark $\rightarrow$ jet matching

An easier way to apply ΔR matching for a b jet identification can be finding directly a jet closest to the b quark, ignoring b hadrons completely. Similarly to the case of the quark $\rightarrow$ hadron $\rightarrow$ jet matching, the quark-jet distance is constrained to be within the $\Delta R(quark, jet) < 0.5$ cone. The multiplicity of distinct b jets identified by the quark $\rightarrow$ hadron $\rightarrow$ jet and the quark $\rightarrow$ jet matching is compared to the *GenHFHadronMatcher* in Figure 8.15.

(a) $t\bar{t} + jets$ events with 2 b quarks

(b) $t\bar{t}H(b\bar{b})$ events with 4 b quarks
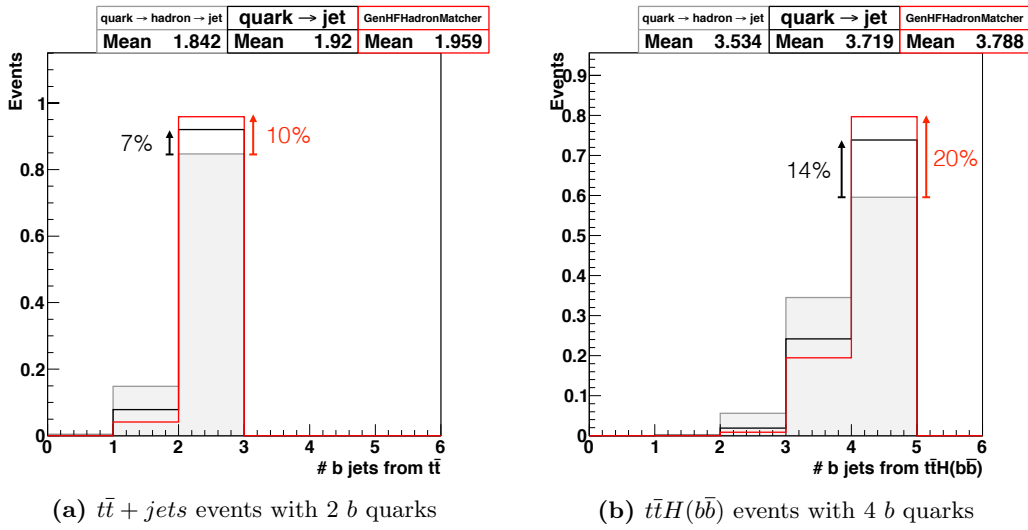
Figure 8.15: Number of distinct b jets associated to b quarks from the top-quark and the Higgs-boson decays, as identified by the three methods: quark $\rightarrow$ hadron $\rightarrow$ jet matching with $\Delta R < 0.5$ requirement (grey), quark $\rightarrow$ jet matching with the $\Delta R < 0.5$ requirement (black), the new approach of the *GenHFHadronMatcher + JetFlavour* tools (red).

It can be seen that eliminating the intermediate step of matching b quarks to b hadrons increases the b jet identification efficiency by 7% and 14% in the $t\bar{t} + jets$ and $t\bar{t}H(b\bar{b})$ events respectively. Nevertheless, the approach using the *GenHFHadronMatcher* still has better performance with 10% and 20% increase in the multiplicity of distinct b jets in the $t\bar{t} + jets$ and $t\bar{t}H(b\bar{b})$ events respectively.

Thus, it is worth to look at distances in $\eta - \phi$ space between the b quarks and jets matched to them by the ΔR matching with and without the intermediate step with b hadrons. This allows to see the differences in actual jets identified by the different methods and is shown in Figure 8.16. It can be seen that matching jets directly to the b quarks provides more distinct b jets with a cost of non-optimal jet assignment when the distance between the quark and the jet is $\Delta R(quark, jet) > 0.1$.

This mismatching can also be seen by explicitly checking whether a jet matched to a b quark by a ΔR method is the same as the jet matched to it by the method based on *GenHFHadronMatcher*. This comparison is given in Figure 8.17, and shows that the direct quark $\rightarrow$ jet matching picks a different jet more often than the ΔR matching via a b hadron.

(a) 2 b quarks from top quark decays in $t\bar{t} + jets$ events

(b) 4 b quarks from top-quark and Higgs-boson decays in $t\bar{t}H(b\bar{b})$ events
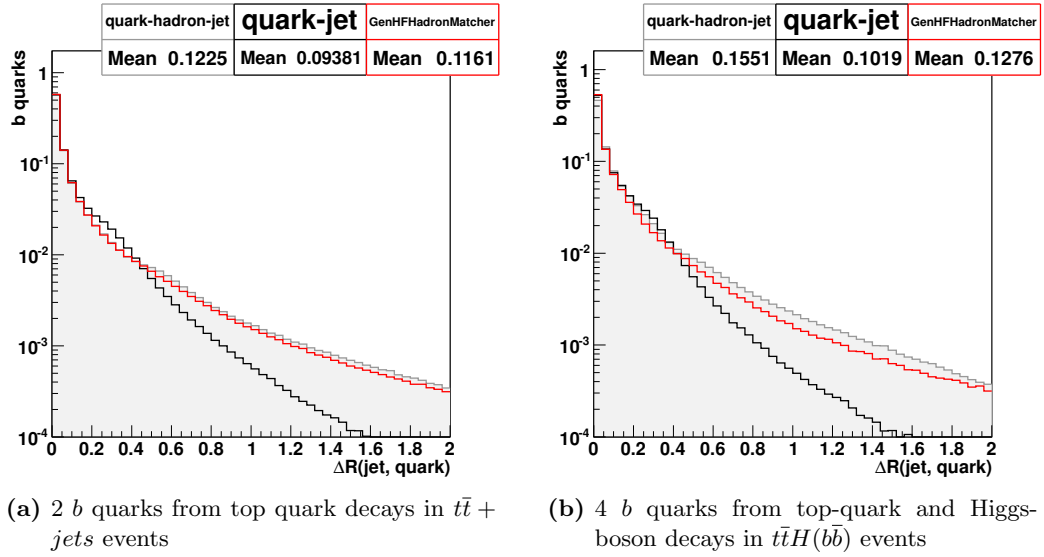
Figure 8.16: Distance between each b quark and a jet matched to it by the three methods without any ΔR cuts applied. The direct quark $\rightarrow$ jet matching prefers jets that are closer to the b quark than the matching based on the *GenHFHadronMatcher*.

(a) 2 b quarks from top quark decays in $t\bar{t} + jets$ events

(b) 4 b quarks from top-quark and Higgs-boson decays in $t\bar{t}H(b\bar{b})$ events
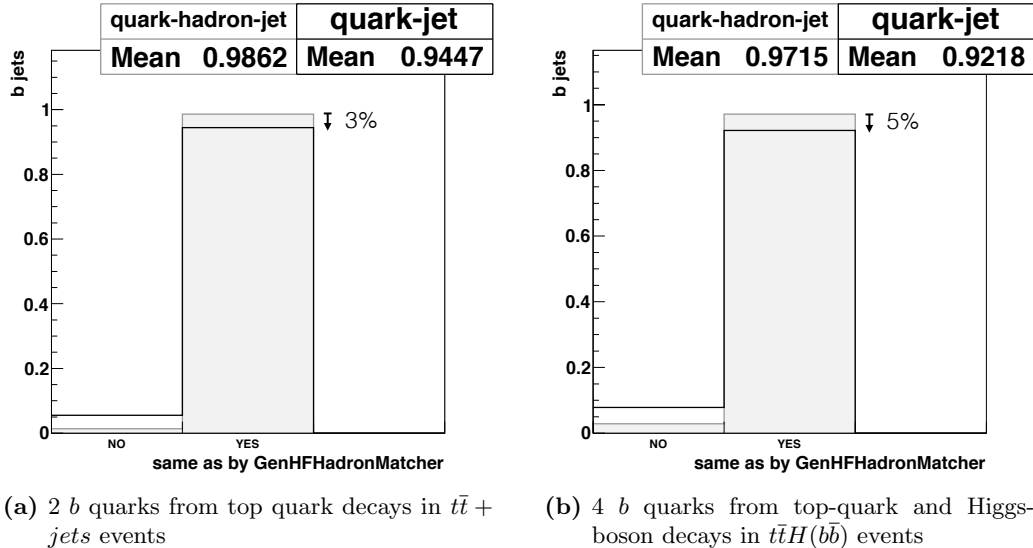
Figure 8.17: Fraction of jets matched to b quarks from the top-quark and the Higgs-boson decays by the two different ΔR methods that are the same as the ones matched by the *GenHFHadronMatcher*-based method. The plots are filled only for those quarks that are matched to jets by all the three compared methods.

Overlaps of heavy-flavour jets

Furthermore, the developed heavy-flavour jet matching provides the possibility to easily identify a heavy flavour content of each jet, even if it was formed by more than one hadron, which is not so well defined with the simple ΔR matching.

For instance, the new method allows to precisely identify how often b jets from the top quark decays overlap with b jets initiated by additional radiation. In order to enhance the effect, a $t\bar{t} + b\,jets$ subset of the $t\bar{t} + jets$ samples was selected, which has at least one additional b jet, which is initiated by a b hadron not originating from the $t\bar{t}$ decay. A difference in the probability of such b-jet overlaps, as modeled by MC simulations of the $t\bar{t} + b\,jets$ and the $t\bar{t}H(b\bar{b})$ processes, is demonstrated in Figure 8.18.
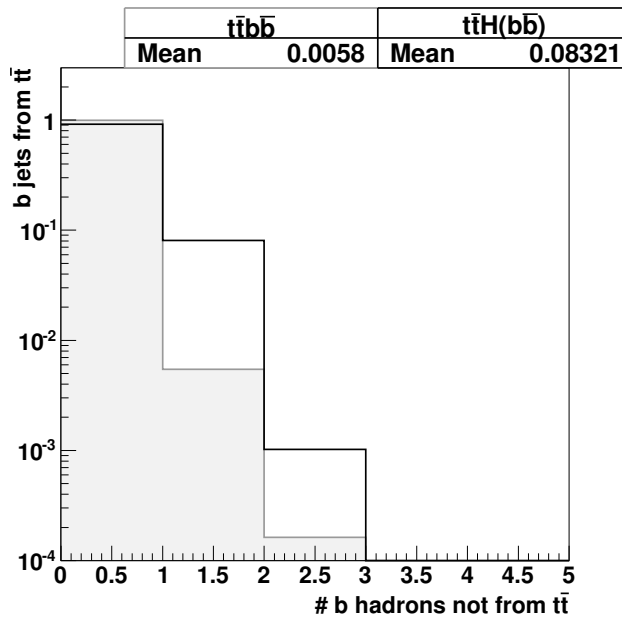
Figure 8.18: Multiplicity of b hadrons not originating from top quark decays in b jets with b hadrons from the top quark decays. The distribution is compared between the $t\bar{t} + b\,jets$ (filled grey) and $t\bar{t}H(b\bar{b})$ (hollow black) events.

The distributions indicate that b jets from the Higgs boson decays overlap with b jets from the $t\bar{t}$ system decay more often than additional b jets in the $t\bar{t} + b\,jets$ events. This can be explained by the large mass of the Higgs boson, which leads to a larger opening angle between the b quarks from its decay compared to the spatial separation of b quarks originating from gluon splitting.

Another effect of the overlapping b jets is that the presence of multiple secondary vertices from b hadrons in the vicinity of such a jet makes it more probable to be identified as a b jet at the reconstruction level. Thus, overlapping jets may need a special treatment in analyses that are sensitive to such contributions.

8.3.3   Comparison of different Parton Showers

All the tests shown before were performed with samples that use PYTHIA parton shower. The actual samples are listed in Table 8.1.

One of the features of the *GenHFHadronMatcher* is its generic approach for
the particle chain scanning, which relies only on *pdgId*'s of the particles and not
on their status codes. The status-code conventions are not unique across different
parton shower simulations and even across different versions of a single parton shower
simulation. But the convention of the *pdgId* assignments to different particles is
established by the Particle Data Group [210] and is the same in all major parton
shower simulators.

The second mostly used parton shower generator used in CMS is HERWIG. There-
fore, in addition to the samples listed in Table 8.1, the *GenHFHadronMatcher*
performance has been tested with several other $t\bar{t} + jets$ simulations that use dif-
ferent combinations of generator and parton shower, as listed in Table 8.2. Since
no $t\bar{t} + jets$ sample using SHERPA parton shower was available for tests, it is not
included in the list of compared samples.

Table 8.2: Additional MC samples of the $t\bar{t} + jets$ process simulated with different com-
binations of ME generators and parton showers.

Process	ME generator	Parton shower	# entries
$t\bar{t} + jets$	POWHEG	PYTHIA	21 675 970
$t\bar{t} + jets$	POWHEG	HERWIG	27 684 235
$t\bar{t} + jets$	MC@NLO	HERWIG	32 852 589

One of the best quantities that represent the *GenHFHadronMatcher* perfor-
mance is the multiplicity of distinct b hadrons from top quark decays, which has
already been shown in Figure 8.13 for the samples listed in Table 8.1. The same
result is obtained with the samples from Table 8.2, as shown in Figure 8.19.

(a) POWHEG + PYTHIA   (b) POWHEG + HERWIG   (c) MC@NLO + HERWIG

Figure 8.19: Number of distinct b hadrons from the top-quark decays, as identified by
the three methods: closest hadron to each b quark (dashed), closest hadron
to each b quark within $\Delta R(hadron, quark) < 0.5$ cone (black), chain scan-
ning by the *GenHFHadronMatcher* (red).

While the multiplicity of distinct b hadrons from $t\bar{t}$ decays looks the same in all
the simulations, additionally radiated b quarks can look different. In particular, this

can be illustrated by the ΔR distance between two closest b hadrons and the two closest b jets in events with at least three b jets that have $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ and $|\eta| < 2.4$ ($t\bar{t} + b\,jets$). Thus, in addition to the two b jets from the $t\bar{t}$ system, there is at least one b jet initiated by additional b hadrons. The distribution is shown in Figure 8.20 for all the different samples, including the $t\bar{t}H(b\bar{b})$ simulation, which should always have two b hadrons from the Higgs boson decay.
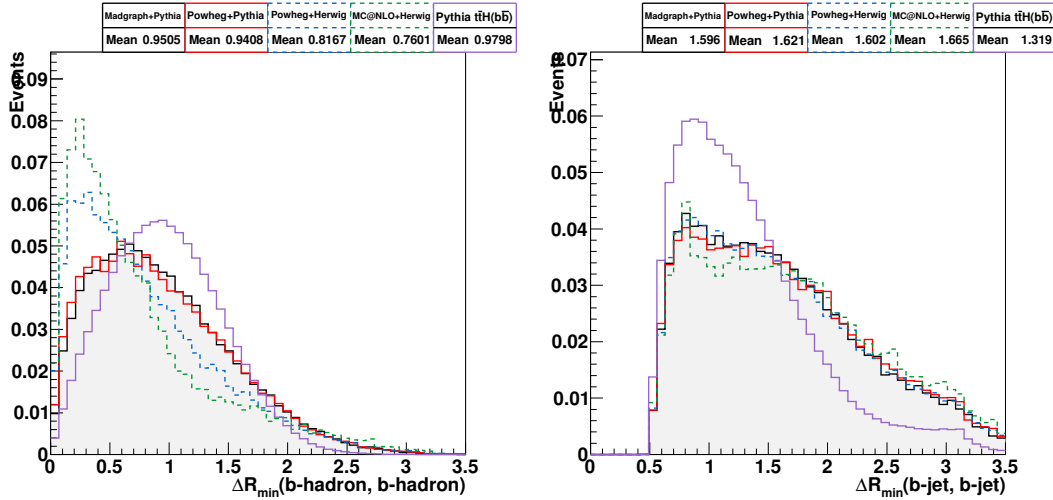
Figure 8.20: Distance in $\eta - \phi$ space between the two closest b hadrons (left) and the two closest b jets in all the different MC simulations. The dashed lines represent the samples simulated with POWHEG parton showers, while the solid line denotes the PYTHIA parton showers.

The first obvious difference is the one between the $t\bar{t} + b\,jets$ and $t\bar{t}H(b\bar{b})$ samples. In the $t\bar{t} + b\,jets$ sample, the two closest b hadrons usually originate from gluon splitting, while in $t\bar{t}H(b\bar{b})$ events those are either two b hadrons from the Higgs decay or one from the Higgs and one from the top quark decay. The large mass of the Higgs boson leads to a bigger separation between the b hadrons in $t\bar{t}H(b\bar{b})$ events compared to that in $t\bar{t} + b\,jets$ events, which is reflected by the larger distance between the closest b hadrons in $t\bar{t}H(b\bar{b})$ events.

The second difference is related to parton showers used in $t\bar{t} + jets$ process simulation. The two samples simulated with HERWIG have significantly more events with b hadrons very close to each other compared to the samples that use PYTHIA parton showers. When jets are clustered with the cone parameter $R = 0.5$, b hadrons separated by $\Delta R < 0.5$ are usually contained in the same jet, leading to the increased distance between two closest b jets in an event, as can be seen from the right plot in Figure 8.20. Compared to the $t\bar{t} + b\,jets$ events, this increase is much smaller in the $t\bar{t}H(b\bar{b})$ process, since the b hadrons themselves are well separated, each creating an own jet.

This merging of b hadrons into a single jet affects the total number of b jets in $t\bar{t} + b\,jets$ events, which looks different depending on a parton shower used in simulations, as shown in Figure 8.21. Samples simulated by HERWIG parton showers have smaller multiplicity of b jets due to the overlaps between them, while the $t\bar{t}H(b\bar{b})$ events have the highest number of b jets.
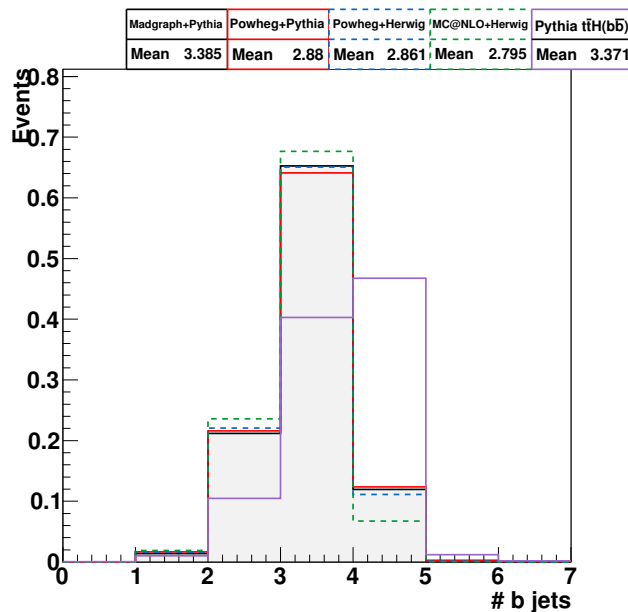
Figure 8.21: Multiplicity of b jets in the $t\bar{t} + b\,jets$ and $t\bar{t}H(b\bar{b})$ samples simulated by different combinations of generators and parton showers.

8.4   Conclusions

It can be concluded that the new method for b-jet identification is very reliable, and allows to not only identify b jets, but also to detect their origin using almost 100% unique b hadron $\rightarrow$ b quark matching. This approach is especially important in events with high multiplicity of b quarks, e.g. in the $t\bar{t}H(b\bar{b})$ processes. Furthermore the developed method allows to detect overlapping b or c jets as well as leptonic decays of the heavy flavour hadrons.

Being based only on daughter-mother relations of the particles in MC simulations and on their $pdgId$'s, the $GenHFHadronMatcher$ performs equally well with samples simulated by the two major parton-shower models: PYTHIA and HERWIG. This allows to use the tool without any special configuration and to easily compare effects of different generators and parton shower models on heavy-flavour-jet related quantities. The tool is integrated into the CMSSW framework and has technical instructions in the dedicated TWiki page [208].

Chapter 9

Measurement of associated top-quark-pair and b-jet production

The $t\bar{t}b\bar{b}$ production is an important QCD process that is addressed in a number of theoretical calculations at the NLO accuracy [211, 212, 213]. Furthermore, as mentioned in Section 2.3.4, the $t\bar{t}b\bar{b}$ process is the dominant and almost irreducible background to the $t\bar{t}H(b\bar{b})$ production.

Since experimentally only final state particles can be measured, it is sometimes hard to distinguish signal processes from background processes. For a precise measurement, the contribution of background processes has to be either minimised or known very precisely.

It was also shown in Figure 2.21, that an excess of $t\bar{t}H(b\bar{b})$ events in the dijet-mass distribution of two b jets not coming from the $t\bar{t}$ decay is very sensitive to the shape of the $t\bar{t}b\bar{b}$ cross section. Therefore, a measurement of the $t\bar{t}b\bar{b}$ cross section as a function of kinematic properties of additional b jets is an important ingredient of a search for $t\bar{t}H(b\bar{b})$ production.

9.1   Choice of the final state

The dileptonic final state of the $t\bar{t}$ system is considered in the $t\bar{t}b\bar{b}$ cross-section measurement. It requires a presence of two leptons, each being either an electron or a muon coming from either $W \to e/\mu$ or $W \to \tau \to e/\mu$ decay. Thus, the considered $t\bar{t}b\bar{b}$ process in the dileptonic decay channel of the $t\bar{t}$ system can have four basic final states depending on the presence of τ leptons, as shown in Figure 9.1.

A large number of non-perturbative processes, taking place during the b jet fragmentation, introduce distortions of experimentally measured jet properties (e.g. energy, spatial direction, electric charge, flavour) with respect to the characteristics of the original b quark. These distortions place an intrinsic limitation on the resolution of any quantities derived from the b jets, e.g. the invariant mass of two b jets in a $t\bar{t}b\bar{b}$ or $t\bar{t}H(b\bar{b})$ process. Furthermore, to measure the properties of the b jets, it is necessary to properly identify those that do not originate from top-quark decays. From the experimental point of view, there is no obvious distinction between b jets

(a) no decays to τ leptons

(b) W^+ boson decays to a τ lepton

(c) both W bosons decay to τ leptons

(d) W^- boson decays to a τ lepton
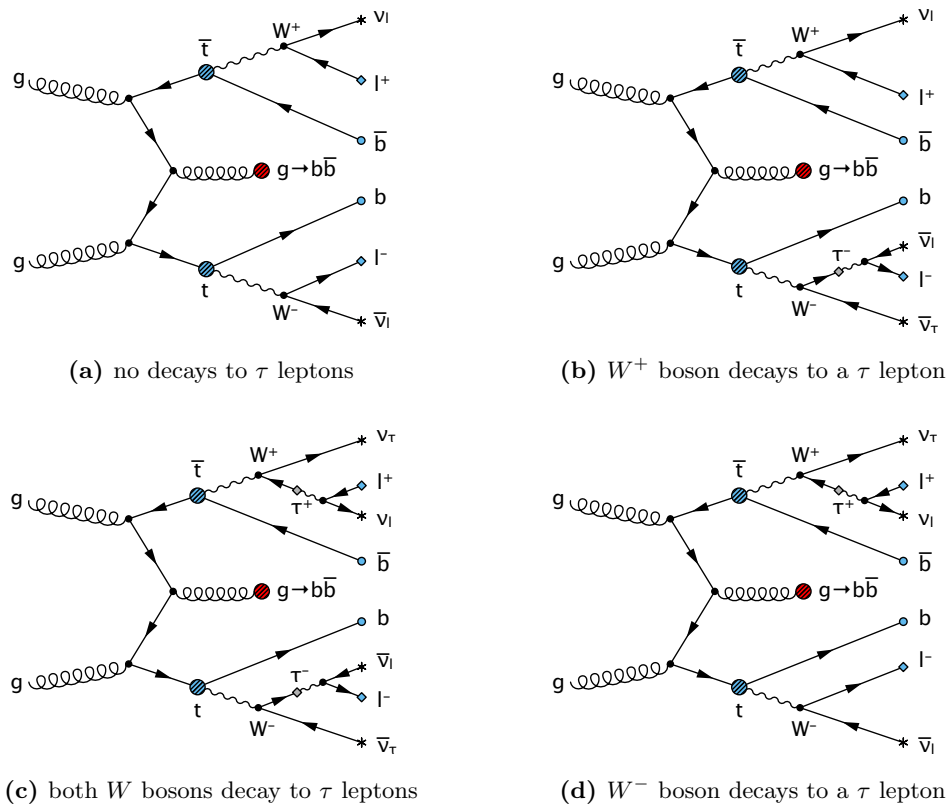
Figure 9.1: Feynman diagrams of possible dileptonic final states of the $t\bar{t}b\bar{b}$ process with respect to the presence of W boson decays into τ leptons. Additional b quarks from the gluon decay (red) are omitted to simplify the visual perception. On the right-hand side of each diagram l^+/l^- denotes an electron or a muon.

produced by b quarks from the $t\bar{t}$ system and b jets produced in addition. Thus, the probability of picking up the proper pair of b jets out of the four available ones is given by the inverse of the number of possible jet combinations C^2_{Njets}:

$$\mathrm{P}^{correct}_4 = \frac{1}{C^2_4} = \frac{2!(4-2)!}{4!} = \frac{1}{6} \,. \tag{9.1}$$

As discussed in Section 7.5, non-b jets can be misidentified as b jets, leading to an increased number of candidates for b jets from the Higgs-boson decay. Thus, in case of five b-jet candidates, the probability of selecting the proper combination is substantially reduced:

$$\mathrm{P}^{correct}_5 = \frac{1}{C^2_5} = \frac{2!(5-2)!}{5!} = \frac{1}{10} \,. \tag{9.2}$$

In events with the semileptonic or full-hadronic $t\bar{t}$ decays, more additional jets are present that can be mistagged, and can increase the number of potential fake b-jet-pair candidates.

This is one of the main reasons why the dileptonic channel of the $t\bar{t}$ system is considered in the analysis, despite its small branching fraction. Any additional jet from hadronic W boson decays would significantly reduce the fraction of properly reconstructed additional b-jet pairs.

9.2   Cross-section definition

As can be seen from the diagrams in Figure 9.1, the final state considered in the analysis is rather complicated. In order to properly interpret and compare a measurement of such a process to theory calculations, it is very important to precisely define the cross sections that are measured.

The $t\bar{t}b\bar{b}$ differential cross sections are measured as functions of kinematic properties of additional b jets in two phase spaces of the $t\bar{t}$ system:

- **visible phase space:** both additional b jets and the decay products of the $t\bar{t}$ system must be in the experimentally accessible region, which is defined later;

- **full phase space:** only the additional b jets are required to be in experimentally accessible region, while the $t\bar{t}$ system just has to be in a dileptonic final state, regardless of its experimental accessibility.

Since the full phase space is not directly accessible through experiment, it must rely on an extrapolation based on the theoretical prediction, which, in case of the current analysis, is based on a MC prediction of the $t\bar{t} + jets$ process. The advantage of a result in the full phase space is that it is easier to calculate theoretically, which allows to compare the measured cross section to a greater variety of theoretical predictions, including future ones.

The final state of the measured $t\bar{t}b\bar{b}$ process consists of leptons and jets, and the visible phase space definition is based on them. Therefore, the generator-level definitions of leptons and jets are given first, then the actual cross sections are defined based on the generator-level objects.

9.2.1   Object definitions

The only generator level objects directly involded in the measured $t\bar{t}b\bar{b}$ process at the visible phase space are leptons and b jets. For the full phase space only top quarks are relevant, which are always present in the corresponding $t\bar{t} + jets$ simulated samples.

Leptons

Leptons from decay of the $t\bar{t}$-system, which are used for the visible phase-space definition, correspond to the matrix-element level, before parton showering (see Section 6.1.1). This choice ensures that the momentum of a lepton from a $t\bar{t}$ decay is not affected by a collinear radiation of a photon by the lepton.

Technically the leptons are obtained using the standalone CMS package for reading information about a $t\bar{t}$ system from MC simulation: `TopQuarkAnalysis.TopEventProducers.sequences.ttGenEvent_cff`. The matrix-element level for leptons is configured via the parameter `decaySubset.fillMode = "kME"`.

Jets

Jets are obtained by recombining all final state particles in the event except of leptons originating from the $t\bar{t}$-system decay, as well as all their decay products or particles radiated from them, i.e. photons, leptons. The excluded leptons are electrons, muons, tau-leptons and their corresponding neutrinos. Other neutrinos, which do not originate from the $t\bar{t}$-pair decay are included in the jets. The clustering of neutrinos to jets is motivated by the jet-energy corrections applied to reconstruction-level jets, which use generated jets with neutrinos as a reference (see Section 7.4.1). Jets are built using the anti-k_{T} clustering algorithm [197] with the distance parameter $R = 0.5$, which are referred to as `ak5GenJets` within CMS.

Heavy-flavour jets

The heavy-flavour content of the jets is identified using the $GenHFHadronMatcher$, as described in Section 8.1. Basically three types of jets are distinguished:

- **b jet:** at least one b hadron associated to it;

- **c jet:** no b hadrons and at least one c hadron associated to it;

- **light jet:** no b hadrons and no c hadrons associated to it.

In addition, the origin of b/c hadrons and their multiplicity in each heavy-flavour jet is determined by the same tool. It is used to categorise simulated events as will be described later in Section 9.5.

Additional *b* jets

Since the heavy-flavour jet definitions are based on hadrons, additional *b* jets are defined with respect to the origins of the *b* hadrons associated to them. Three types of *b* hadrons are distinguished depending on their origin in a $t\bar{t}$ event, which are schematically shown in Figure 9.2:

- **from top:** contains a *b* quark directly from a top-quark decay;

- **additional:** contains a *b* quark radiated before top-quark decays;

- **pseudoadditional:** contains a *b* quark radiated after a top-quark decay;

(a) additional *b* quarks   **(b)** pseudoadditional *b* quarks
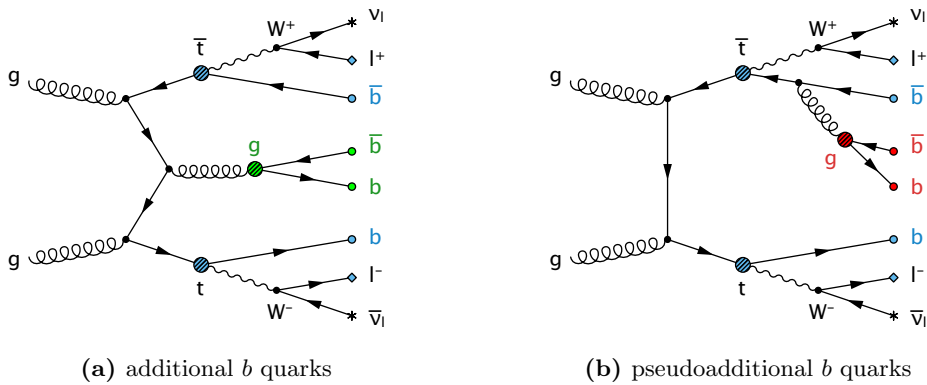
Figure 9.2: Schematic view of different *b*-hadron origins with respect to the $t\bar{t}$-system decay. Additional *b* hadrons contain *b* quarks radiated before the top-quark weak decays (marked by green circles). Pseudoadditional *b* hadrons contain *b* quarks radiated after the top-quark weak decays (marked by red circles). *b* hadrons from top contain *b* quarks stemming directly from top-quark weak decays (marked by blue circles).

Thus, additional *b* jets must be initiated by *b* hadrons that are not products of the $t\bar{t}$-pair decay. In terms of the *b*-hadron content, a *b* jet is defined as *additional* if it fulfils all the following requirements:

1. not associated to any *b* hadron *from top*;

2. associated to at least one *additional b* hadron;

The number of associated *pseudoaddional b* hadrons does not influence the definition of *additional b* jets.

9.2.2 Phase-space definition

Additional *b* jets are required to have $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ and pseudorapidity $|\eta| < 2.4$, which corresponds to the maximum kinematic region in which *b* jets can be reliably reconstructed and distinguished from light jets in real data. Additional *b* jets must satisfy these kinematic criteria in both the full and visible phase space of the $t\bar{t}$ system.

In the visible phase space, additional requirements are applied to the decay products of the $t\bar{t}$ system:

- **both leptons from top:** $p_T > 20\,\text{GeV}$, $|\eta| < 2.4$;

- **both b jets from top:** $p_T > 30\,\text{GeV}$, $|\eta| < 2.4$.

9.2.3   Cross sections

Depending on the multiplicity of additional b jets that satisfy the criteria mentioned before, two separate processes are distinguished, for which cross sections are measured:

- **$t\bar{t}b(\bar{b})$** : at least one additional b jet (the second b jet can be outside acceptance or overlap with the first additional b jet or with a b jet from the $t\bar{t}$ decay);

- **$t\bar{t}b\bar{b}$** : two or more additional b jets.

Only the origin of b hadrons matters in the definition of the process, as described in Section 9.2.1, while the number of b hadrons in each additional b jet does not matter. The processes with other sources of additional b jets, i.e. $t\bar{t}H$ or $t\bar{t}Z$, where the H or Z boson decay to a $b\bar{b}$ pair, are not considered as part of the $t\bar{t}b(\bar{b})$ or $t\bar{t}b\bar{b}$ processes.

For both the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ processes, differential cross sections are measured as functions of p_T and $|\eta|$ of the first or second additional b jet respectively. The first additional b jet is the one with the highest p_T among those that fulfil the acceptance requirements. The second additional b jet is the one with the second highest p_T among b jets in acceptance. In addition, the $t\bar{t}b\bar{b}$ cross section is measured as a function of the distance between the first and second additional b jets ($\Delta R_{b\bar{b}}$) and of the invariant mass of the two b jets ($m_{b\bar{b}}$).

9.2.4   Measurement strategy

As mentioned earlier, the final state considered in the analysis is rather unique, but it still can be produced by other processes at lower rates. Furthermore, the limited resolution of the detector used for the measurement can lead to other final states being misidentified as the process of interest. Therefore, the selected events from the recorded data will inevitably contain background processes, whose contribution has to be properly evaluated and subtracted from the measured data. For the estimation of different background contributions, Monte Carlo simulations of the corresponding processes are used, which are described in Section 9.3.

In order to minimise the influence of the background estimations on the measured cross sections, the data must be selected such that the easily-distinguishable or the most uncertain background contributions are significantly suppressed. The exact event selection used to fulfil these criteria is described in Section 9.4, while the process of background estimation using MC simulations is described in Section 9.5.

The fact that the measured cross sections are differential, leads to possible *bin-to-bin migrations* due to a misreconstruction or the limited resolution of the detector. A migration is the process of a measured value of the property appearing in a different bin than the true value, which leads to a distortion of the measured shape of the cross section. The process of correcting bin-to-bin migrations is referred to as *unfolding* and is described in Section 9.7.

Summarising the described steps, the differential cross section is calculated as:

$$\frac{d\sigma}{dX_i} = \frac{\sum_j A_{ij}^{-1} \left[N_j^{\text{data}} - N_j^{\text{bkg}} \right]}{\varepsilon \cdot \Delta x_i \cdot \mathcal{L}} \, , \tag{9.3}$$

where j represents the index of the measured bin, i is the index of the true bin, ε is the event-selection efficiency, Δx_i is the width of the bin and A_{ij} corresponds to the response matrix, which characterises the migration probability. In a measured bin j, N_j^{data} represents the total number of selected data events and N_j^{bkg} corresponds to the estimated number of background events. The $\mathcal{L}$ corresponds to the integrated luminosty of the data used for the measurement.

9.3   Data and simulated samples

The differential cross-sections for $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ production were measured using proton-proton collisions at $\sqrt{s} = 8\,\text{TeV}$ recorded by the CMS experiment during 2012. The data collection period was separated into four distinct sections according to the LHC running conditions: Run2012A, Run2012B, Run2012C and Run2012D. The complete dataset was certified for physics analysis by the DQM procedure as described in Section 4.2.7, and corresponds to the total integrated luminosity of $\mathcal{L} = 19.7 \pm 0.5$ pb.

Real data

The raw data recorded by CMS were re-reconstructed with the latest set of detector calibration constants and reconstructed-object corrections that existed on the 23-rd of January, 2012. The set of conditions used for the reprocessing of 2012 data is stored under the dedicated GlobalTag: `FT_R_53_V18`. Then the data were further analysed with the specific set of tools and a newer set of high-level corrections, which are stored in the GlobalTag: `FT_53_V21_AN6`.

The data used for the measurement were preselected by three different HLT triggers (see Appendix B.1) and organised into separate datasets with double-electron (ee), double-muon ($\mu\mu$) and electron-muon ($e\mu$) events. The list of actual datasets used for the measurement is given in Table 9.1.

MC simulation

The simulation of the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ processes is taken as a subset of the $t\bar{t} + jets$ sample, since no dedicated $t\bar{t}b\bar{b}$ simulation existed. In order to use MC simulations consistently with the data, all possible processes that can have a non-negligible contribution to the experimentally selected final state must be included. Therefore, numerous simulations of background processes are used as well, i.e. leptonically decaying W bosons accompanied by additional partons, Drell-Yan processes, single top quarks produced in a tW channel, multiple vector-boson production (diboson or triboson), multiple jet production (referred to as QCD multijet) and $t\bar{t}$ pairs with semileptonic or full-hadronic decays or $t\bar{t}$ produced in association with photons (γ) or W/Z bosons. The list of the simulated samples used for the measurement is summarised in Table 9.2.

Table 9.1: List of proton-proton collision datasets used in the analysis, separated by different dileptonic channels.

	Dataset name	Run range	Events
ee	/DoubleElectron/Run2012A-22Jan2013-v1/AOD	190456–193621	13M
	/DoubleElectron/Run2012B-22Jan2013-v1/AOD	193834–196531	23.5M
	/DoubleElectron/Run2012C-22Jan2013-v1/AOD	198022–203742	39M
	/DoubleElectron/Run2012D-22Jan2013-v1/AOD	203777–208686	34.5M
$\mu\mu$	/DoubleMu/Run2012A-22Jan2013-v1/AOD	190456–193621	5.6M
	/DoubleMuParked/Run2012B-22Jan2013-v1/AOD	193834–196531	29M
	/DoubleMuParked/Run2012C-22Jan2013-v1/AOD	198022–203742	37M
	/DoubleMuParked/Run2012D-22Jan2013-v1/AOD	203777–208686	38M
$e\mu$	/MuEG/Run2012A-22Jan2013-v1/AOD	190456–193621	2.5M
	/MuEG/Run2012B-22Jan2013-v1/AOD	193834–196531	15M
	/MuEG/Run2012C-22Jan2013-v1/AOD	198022–203742	21M
	/MuEG/Run2012D-22Jan2013-v1/AOD	203777–208686	22M

A,B,C and D in the dataset names correspond to the different data-taking periods during 2012, which were organised into separate datasets.

The simulated samples are normalised to the integrated luminosity of the recorded data using the experimentally measured or theoretically calculated cross sections of the corresponding processes. The cross-section values used for the normalisation of each MC sample are stated in Table 9.2.

All the listed samples will be used later in plots showing the comparison between the real data and MC simulations of the signal and background processes. In order to make the legend on the plots compact and readable, some processes are combined, while other are split into subprocesses that play an important role in the described analysis. Hence, the Drell-Yan, Diboson, Triboson and W+jets samples are combined into a single Electroweak (•) sample, and the $t\bar{t}W$ and $t\bar{t}\gamma$ samples are combined into a single $t\bar{t}W/\gamma$ (•) contribution.

The $t\bar{t}+jets$ sample is split into several subprocesses depending on the type and number of additional b jets in the full $t\bar{t}$ phase space, as defined in Sections 9.2.1 and 9.2.2:

- **$t\bar{t}b\bar{b}$:** dileptonic $t\bar{t}$ system with at least two additional b jets in acceptance;

- **$t\bar{t}2b$:** dileptonic $t\bar{t}$ system with exactly one additional b jet in acceptance, which is matched to more than one additional b hadron;

- **$t\bar{t}b$:** dileptonic $t\bar{t}$ system with exactly one additional b jet in acceptance, which is matched to exactly one additional b hadron;

- **$t\bar{t}Other$:** dileptonic $t\bar{t}$ system with no additional b jets in acceptance or semileptonic or full-hadronic $t\bar{t}$ system with any number of additional b jets;

The motivation for the splitting of the $t\bar{t}+jets$ sample will be explained in Section 9.5. The $t\bar{t}H$ sample is also split into the two components:

- **$t\bar{t}H(b\bar{b})$:** $t\bar{t}H$ production with the Higgs boson decaying into a $b\bar{b}$ pair, without any acceptance requirements for b jets either from the $t\bar{t}$ decay or from the $H \to b\bar{b}$ decay;

- **$t\bar{t}HOther$:** $t\bar{t}H$ production with all other decay channels of the Higgs boson, except of the $H \to b\bar{b}$.

The $t\bar{t}H(b\bar{b})$ process is plotted separately due to the same final state as the one of the $t\bar{t}b\bar{b}$ process, which makes it an important contribution, and because it is also the process of interest in the search described in Chapter 10. The $t\bar{t}HOther$ contribution is obtained from the simulation of the inclusive $t\bar{t}H$ process, in which all decays of the Higgs boson are present. Therefore, events with the $H \to b\bar{b}$ decays are excluded from the inclusive sample to avoid double counting.

Table 9.2: Summary of MC simulation samples used for the cross section measurement.

Process	Generator	Sample name	Events	σ [pb]	Source of σ
$t\bar{t}+jets$ ●●●○	MADGRAPH	/TTJets_MSDecays_central_TuneZ2star_8TeV-madgraph-tauola/	62.1M	241.5	measurement [214]
Single top ●	POWHEG	/Tbar_tW-channel-DR_TuneZ2star_8TeV-powheg-tauola/	0.5M	11.1	NLO+NNLL [215]
		/T_tW-channel-DR_TuneZ2star_8TeV-powheg-tauola	0.5M	11.1	
$t\bar{t}H$ ●●	PYTHIA	/TTH_HToBB_M-125_8TeV-pythia6/	1M	0.1293	NLO [114]
		/TTH_Inclusive_M-125_8TeV_pythia6/	1M	0.1293	
$t\bar{t}Z$ ●	MADGRAPH	/TTZJets_8TeV-madgraph_v2/	0.2M	0.2057	NLO [216]
$t\bar{t}W$ ●	MADGRAPH	/TTWJets_8TeV-madgraph/	0.2M	0.232	NLO [217]
$t\bar{t}\gamma$ ●	MADGRAPH	/TTGJets_8TeV-madgraph/	0.1M	1.8	LO [218]
W+jets ●	MADGRAPH	/WJetsToLNu_TuneZ2Star_8TeV-madgraph-tarball/	18.3M	36257.2	NNLO [219]
Drell-Yan ●	MADGRAPH	/DYJetsToLL_M-10To50filter_8TeV-madgraph/	7.1M	860.5	NNLO [219]
		/DYJetsToLL_M-50_TuneZ2Star_8TeV-madgraph-tarball/	30.5M	3503.71	
Diboson ●	PYTHIA	/WW_TuneZ2star_8TeV_pythia6_tauola/	10.0M	54.838	NLO [220]
		/WZ_TuneZ2star_8TeV_pythia6_tauola/	10.0M	33.21	
		/ZZ_TuneZ2star_8TeV_pythia6_tauola/	9.8M	17.654	
Triboson ●	MADGRAPH	/ZZZNoGstarJets_8TeV-madgraph/	0.23M	0.00553	NLO [221]
		/WWZNoGstarJets_8TeV-madgraph/	0.22M	0.05795	
		/WWWJets_8TeV-madgraph/	0.22M	0.08058	
QCD Multijet ●	PYTHIA	/QCD_Pt_20_MuEnrichedPt_15_TuneZ2star_8TeV_pythia6/	7.5M	134680.0	LO [222]
		/QCD_Pt_20_30_EMEnriched_TuneZ2star_8TeV_pythia6/	35.0M	2914860.0	
		/QCD_Pt_30_80_EMEnriched_TuneZ2star_8TeV_pythia6/	33.1M	4615893.0	
		/QCD_Pt_80_170_EMEnriched_TuneZ2star_8TeV_pythia6/	34.5M	183294.9	
		/QCD_Pt_20_30_BCtoE_TuneZ2star_8TeV_pythia6/	1.7M	167388.0	
		/QCD_Pt_30_80_BCtoE_TuneZ2star_8TeV_pythia6/	2.0M	167040.0	
		/QCD_Pt_80_170_BCtoE_TuneZ2star_8TeV_pythia6/	1.9M	12981.9	

All the samples correspond to the *Summer12_DR53X_PU_S10_START53_*/AODSIM production campaign, and are simulated with PYTHIA parton-shower.
Coloured bullets represent the colour of the corresponding samples, by which they are marked in control distributions later in text.

9.4   Event selection

The $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ cross sections are measured for the dileptonic final state of the $t\bar{t}$ system. The final-state topology of the $t\bar{t}$ pair is composed of two oppositely charged leptons, each of which can be either an electron or a muon, two or more neutrinos, which are detected by the missing transverse energy ($\not{E}_{\mathrm{T}}$), and two jets initiated by b quarks. In addition, at least one or two other b jets are present in the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ processes respectively.

From the list of considered processes that contribute to the chosen final state (Table 9.2) it can be seen that many processes have clearly larger cross section than the $t\bar{t} + jets$ process, which is a superset of $t\bar{t}b(\bar{b})$ or $t\bar{t}b\bar{b}$ production. In order to maximise the fraction of signal $t\bar{t}+b\,jets$ events in the selected data, specific features of the studied final state are used to reduce contributions from the background processes.

The current analysis is closely related to the measurement of differential $t\bar{t}$-production cross sections in the dileptonic channel [223]. Thus the measurement described in this thesis closely follows the basic dileptonic-$t\bar{t}$-event selection of the previous inclusive analysis [223].

9.4.1   Trigger

The first step in selection of dileptonic $t\bar{t}$ events is based on HLT paths that require the presence of at least two leptons, being either an electron or a muon with transverse momentum larger than 17 GeV or 8 GeV respectively. The detailed HLT paths used at this step are listed in Appendix B.1.

The dilepton trigger efficiencies in data and simulation are estimated using an ensemble of events selected with a set of orthogonal HLT paths based on $\not{E}_{\mathrm{T}}$ requirements. The fraction of these events that additionally fire the dilepton triggers define the trigger efficiency ϵ. Small differences in data and simulation efficiencies are corrected with a data-to-simulation scale factor (SF) [179]. The double-differential SF for a single trigger is derived as a function of pseudorapidity of the two triggered leptons, which is the ratio of the data efficiency over the simulation efficiency:

$$\mathrm{SF}_{trigger} = \frac{\epsilon_{data}}{\epsilon_{MC}} \; . \tag{9.4}$$

The SFs are found to be stable with respect to the vertex multiplicity, and are in general close to 1.0 across the full phase space for all the HLT paths, varying in the range $[0.97, 1.04]$ [179].

9.4.2   Lepton selection

The events considered in the analysis must have at least two oppositely-charged leptons with $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ and $|\eta| < 2.4$. Muon candidates must fulfil either the *tracker-muon* or *global muon* criteria, which were described in Section 7.3.1. Electron candidates must have a transverse impact parameter (IP) with respect to the primary vertex below 0.04 cm. In order to not select events with mis-reconstructed electron candidates passing the lepton selection, an additional requirement is applied to the multivariate discriminant (MVA) [224], which has to be greater than 0.5.

In the case that more than one oppositely-charged-lepton pair can be constructed from the leptons that fulfil all the mentioned requirements, the combination with the highest sum of the transverse momenta of the two leptons is chosen. Depending on the types of leptons in the selected pair, the event is categorised as an e^+e^-, $e^{\pm}\mu^{\mp}$ or $\mu^+\mu^-$.

After the lepton pair has been chosen, it must have the sufficient invariant mass $m_{\ell\ell} > 20\,\text{GeV}$. Otherwise the event is rejected, since the majority of events with lepton-pairs that have low invariant mass come from Drell-Yan and QCD multi-jet processes. The dilepton invariant mass distribution is well described by MC simulations, as shown in Figure 9.3.

(a) $ee + e\mu + \mu\mu$ channels combined   **(b)** $e\mu$ channel
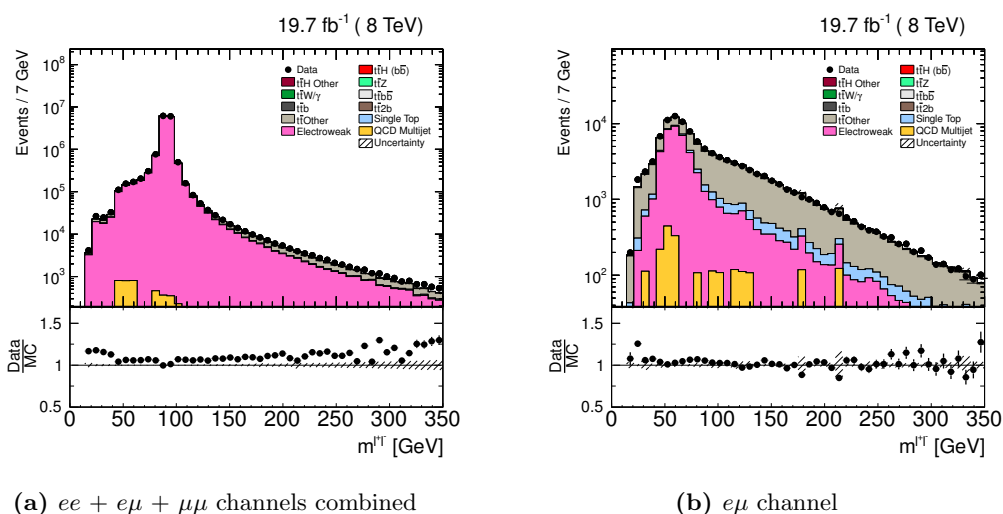
Figure 9.3: Distribution of the invariant mass of the oppositely-charged lepton pair after the trigger and lepton-pair selection. The peak in the combination of all dilepton channels represents the mass of the Z boson in Drell-Yan events.

While the $e\mu$ channel is dominated by $t\bar{t}$ events (see Figure 9.3b), in the ee and $\mu\mu$ channels the dominant contribution comes from Drell-Yan events that especially populate the dilepton mass region around the mass of the Z boson. In order to suppress the Drell-Yan contribution, events in the ee and $\mu\mu$ channels are rejected if the invariant mass of the selected lepton-pair is close to the mass of the Z boson: $76 < m_{\ell\ell} < 106\,\text{GeV}$, which is referred to as Z-veto. The dilepton mass distribution after applying the Z-veto in the ee and $\mu\mu$ channels is shown in Figure 9.4. The slightly underestimated normalisation of MC samples in the $\mu\mu$ channel, which is dominated by Drell-Yan events, can be seen in the Data/MC ratio in Figure 9.4b. This is not critical, since the Drell-Yan contribution will be greatly suppressed after further selection of dileptonic $t\bar{t}$ events.

9.4.3 Vertex selection

In order to reject events initiated by cosmic muons or interactions with a beam line, only events that have at least one good vertex are used in the analysis. The vertices are reconstructed from tracks of charged particles, and a good vertex must fulfil all of the following criteria:

(a) *ee* channel   **(b)** *μμ* channel
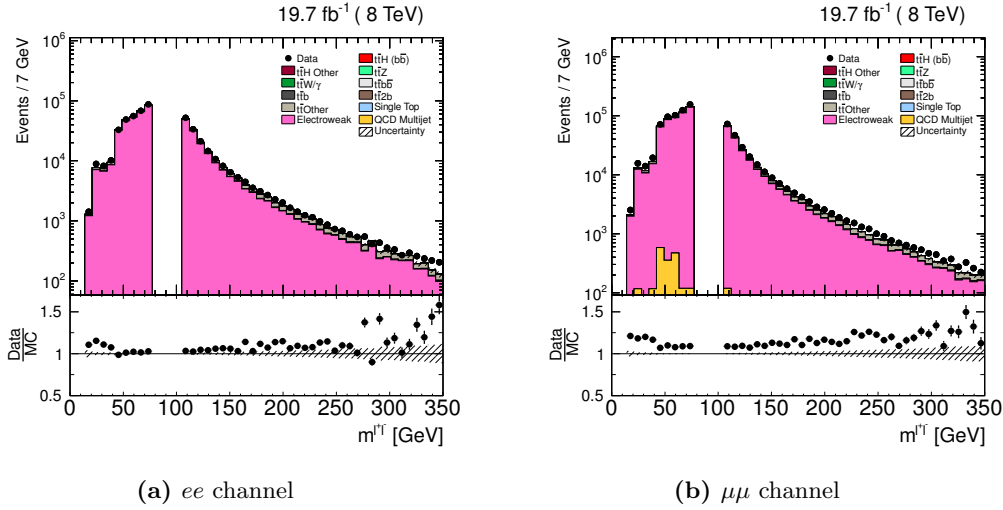
Figure 9.4: Distribution of the invariant mass of the oppositely-charged lepton pair after the trigger and lepton-pair selection and *Z*-veto cut. The $t\bar{t}$ contribution has increased significantly compared to the case without the *Z*-veto in Figure 9.3a.

- have sum of the associated-track weights larger than 4.0;

- be located within $|z| < 24$ cm from the beam spot;

- be located within $r < 2$ cm from the beam axis in the $r\phi$ plane.

Usually there are numerous vertices present in a single event due to multiple parton interactions taking place when the proton bunches cross. Thus, the vertex with the highest scalar sum of squared transverse momenta of the tracks associated to it is selected as the primary vertex, while all the remaining ones are treated as pile-up (PU) vertices. Leptons that are used to define the dileptonic channel have to be consistent with the primary vertex.

In real data, the average number of vertices depends on the total *pp* inelastic cross section and the instantaneous luminosity. The number of pile-up vertices in simulated events is generated based on a prior assumption about the LHC running conditions. Therefore, the distributions of the vertex multiplicity are not the same in data and MC simulation.

All the used MC samples were produced in a single production campaign with the same pile-up distribution in all samples, which corresponds to the `S10 Summer2012 MC` pile-up scenario [225]. The pile-up distribution in the data was obtained with the `pileupCalc.py` utility [226]. The difference in the vertex multiplicity was corrected by a bin-by-bin reweighting of the multiplicity distribution. The effect of the reweighting can be seen in Figure 9.5. Except of the low- and high-multiplicity tails, the shape in data is well described by the MC simulations after reweighting.

9.4.4 Jet selection

A distinctive feature of $t\bar{t}$ events compared to the dominant Drell-Yan background is the presence of energetic hadronic jets initiated by *b* quarks from top-quark decays and of softer jets from ISR and FSR. Since most background processes do
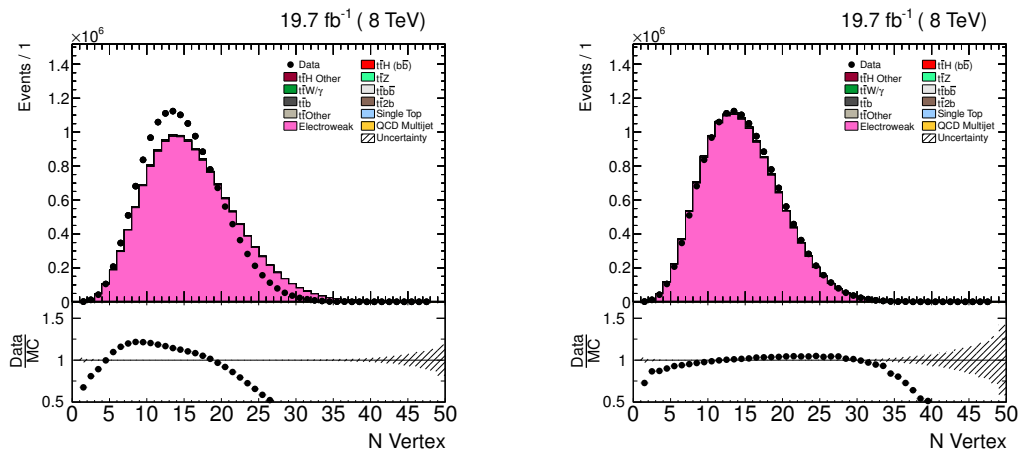
Figure 9.5: Vertex multiplicity before (left) and after (right) applying pile-up weights. The multiplicity is shown after requiring at least one good primary vertex, the trigger paths and after the lepton-pair selection described in the previous sections. The hatched area represents the statistical uncertainty of the MC simulations.

not have such hadronic activity, the requirement of jets significantly reduces non-$t\bar{t}$ contributions in the selected events.

Thus, only events that have at least two jets with $p_T > 20\,\text{GeV}$ and $|\eta| < 2.4$ are considered in the measurement. The jets are reconstructed as described in Section 9.4.6. Furthermore, the two leading jets in p_T are required to have $p_T > 30\,\text{GeV}$, while the remaining jets may have lower transverse momentum starting from 20 GeV. The effect of this selection is shown in Figure 9.6, which almost exclusively removes background events without affecting the $t\bar{t}$ process.
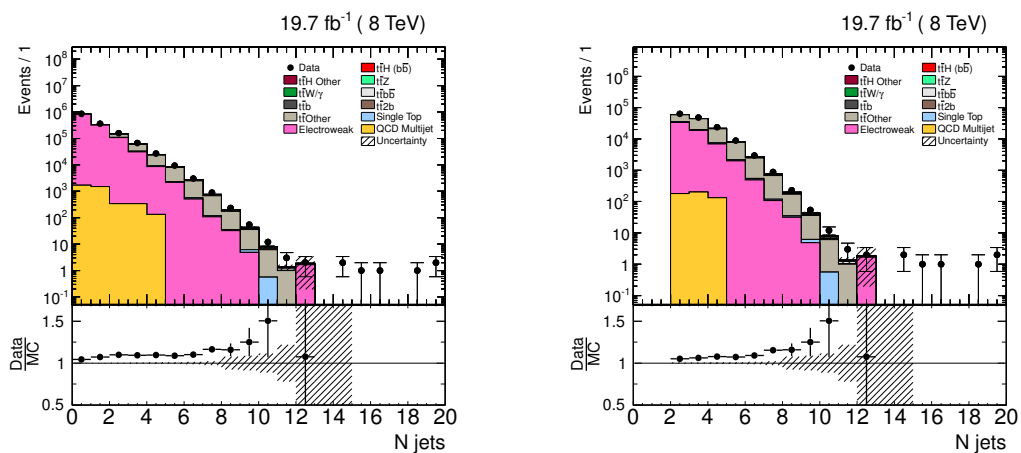
Figure 9.6: Multiplicity of jets with $p_T > 20\,\text{GeV}$ and $|\eta| < 2.4$ before (left) and after (right) the jet requirements described in Section 9.4.4. The distribution is shown for the ee, $e\mu$ and $\mu\mu$ channels combined. In addition to the rejection of events in the first two bins, the number of events with 3 or 4 jets is also partially reduced due to the requirement of the two leading jets to have $p_T > 30\,\text{GeV}$. The selection also affects the events in bins larger than 2 due to the special requirement of $p_T > 30\,\text{GeV}$ for the two leading jets.

The criterion about the two leading jets is applied in order to closely follow the $t\bar{t}$ event selection of the $t\bar{t}$ [223] and $t\bar{t} + jets$ [227] differential cross-section measurements, which serve as a starting point for the analysis described here and require jets with $p_{\text{T}} > 30\,\text{GeV}$. It was checked that lowering the p_{T} cut for the two leading jets to 20 GeV has negligible effect on the number of selected events after the complete set of selections is applied, including the ones described in later sections of this chapter. Thus, the standard $t\bar{t}$ selection is used for consistency.

9.4.5   Missing transverse energy

Apart from hadronic jets, another distinctive feature of $t\bar{t}$ events is the presence of large missing transverse energy ($\not{E}_{\text{T}}$) due to the neutrinos from W-boson and possible τ-lepton decays. In contrast, Drell-Yan events do not produce neutrinos, which would create a genuine transverse energy in the detector. The missing transverse energy is reconstructed as described in Section 7.6, and possible miscalibrations and wrong $\not{E}_{\text{T}}$ reconstruction can lead to non-zero missing tranverse energy detected even in Drell-Yan events.

Thus, the selected events in the ee and $\mu\mu$ channels are required to have substantial missing transverse energy, i.e. $\not{E}_{\text{T}} > 40\,\text{GeV}$. The specific cut value is motivated by the background composition in different regions of the $\not{E}_{\text{T}}$ distribution shown in Figure 9.7a. In the $e\mu$ channel, no requirement for the missing transverse energy is applied, since after the jet selection this channel is anyway dominated by $t\bar{t}$ events, as shown in Figure 9.7b.

(a) ee channel
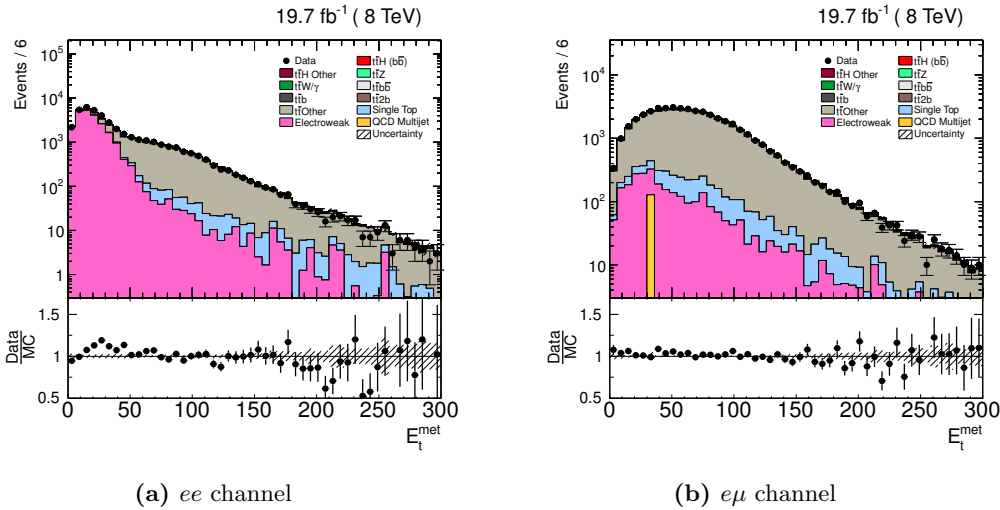
(b) $e\mu$ channel

Figure 9.7: Distribution of the missing transverse energy before the $\not{E}_{\text{T}}$ selection in the ee (left) and $e\mu$ (right) channels.

9.4.6   Selection of b jets

The final important signature of the $t\bar{t}$ events is the presence of at least two b jets from the top-quark decays. The reconstructed jets are identified as b jets with the Combined Secondary Vertex (CSV) algorithm, as described in Section 7.5.

Starting from this point, the event selection is different from the default event selection used in the $t\bar{t}$ cross-section measurement [223]. The standard $t\bar{t}$ selection requires the presence of at least one jet fulfilling the *loose* working point (CSVL), which corresponds to the CSV-discriminant value higher than 0.244. The obvious difference in the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ processes is the higher number of b jets that must be present in an event. The CSVL working point provides a mistagging probability of 10 %, which corresponds to the probability of misidentifing a light-flavour jet as a b-jet. Thus, in an event that has three or four b-tagged jets, the probability of having at least one mistagged jet raises to 30 % or 40 % respectively. This would lead to an inacceptable amount of selected $t\bar{t}$ events with additional c or light jets, which are represented in plots by the $t\bar{t}Other$ contribution.

Therefore, in order to reduce the contribution from $t\bar{t}Other$ processes and to maximise the fraction of the $t\bar{t} + b\,jets$ process in the selected events, the *medium* working point is used (CSVM). A CSVM-tagged jet must have the CSV-discriminant value higher than 0.679, which corresponds to the mistagging probability of 1 % and is shown in Figure 9.8a. A disagreement between the shapes of the CSV discriminant in data and MC simulations is clearly seen. This mismatch leads to a wrong description of the b-tagged-jet multiplicity, as demonstrated in Figure 9.8b, which must be corrected.
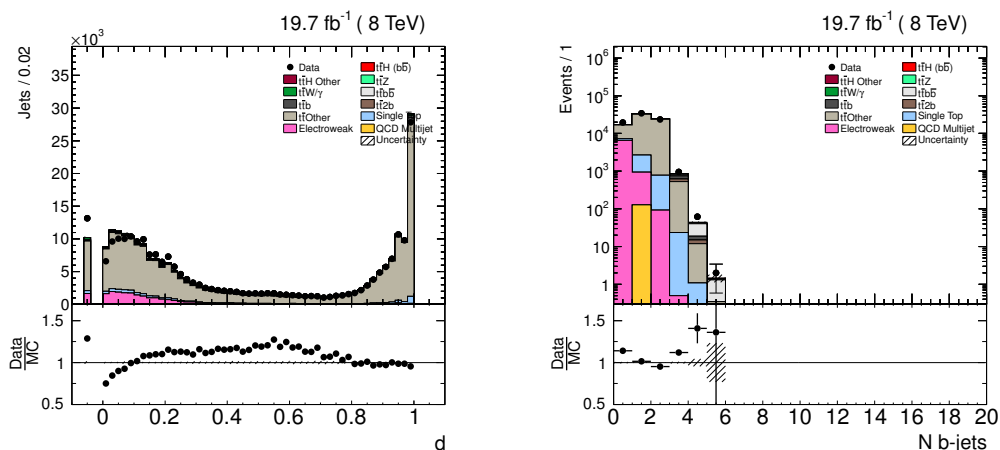
(a) CSV discriminant of every jet in acceptance ($p_T > 20\,\mathrm{GeV}$, $|\eta| < 2.4$)

(b) multiplicity of CSVM-tagged jets

Figure 9.8: Performance of the b-tagging without the CSV-shape reweighting. A jet is b-tagged if it fulfils the CSVM working point requirement. Jets with negative discriminant value do not have enough tracks associated to it for the proper value calculation. The wrong shape of the CSV distriminant in MC leads to the disacgreement in the number of b-tagged jets.

CSV reshaping

It is dangerous to perform any further selection based on b-tagged jets with these disagreements. Hence, the shape of the CSV discriminant in the MC is corrected to match the one in data using the dedicated *CSV-reshaping* technique developed by the CMS $t\bar{t}H$ group [228]. Two sets of scale factors are defined, one for real b jets and one for real light-flavour jets, which are parameterized as functions of the jet

CSV-discriminant value and transverse momentum. Scale factors for light-flavour jets are additionally parameterized as functions of the jet absolute pseudorapidity, since the dependence on $|\eta|$ is not negligible. For c-flavour jets, a flat scale factor of 1.0 is applied due to the absence of a data-driven calibration sample for charm jets, but it is assigned with a twice larger uncertainty than that of b jets. An example of the CSV scale factors as functions of the discriminant value is shown separately for light and b jets in Figure 9.9.

(a) light-flavour scale factor   (b) heavy-flavour scale factor
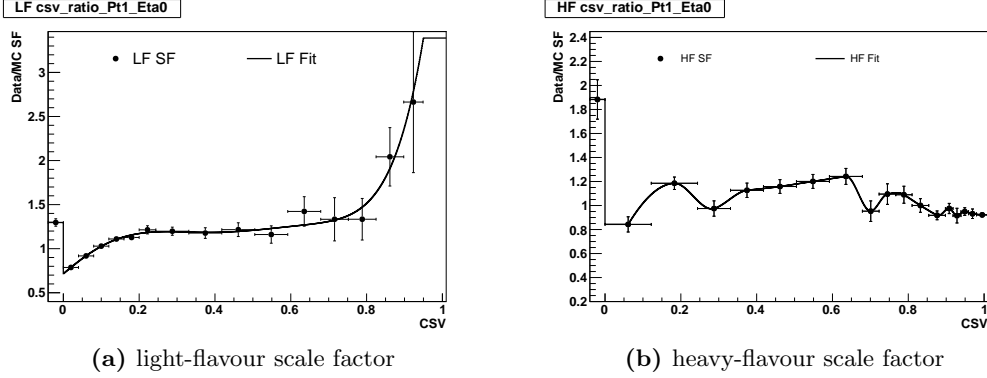
Figure 9.9: Example of the CSV scale factors as functions of the jet-discriminant values in a single p_{T} bin ($40.0 < p_{\mathrm{T}} < 60.0\,\mathrm{GeV}$) and for the light jets also in a single $|\eta|$ bin ($|\eta| < 0.8$). Taken from [228].

To correct the shape of the CSV-discriminant distribution, a single per-event scale factor is constructed from scale factors of each individual jet in an event [228]:

$$SF_{\mathrm{event}} = \prod_{i}^{N_{\mathrm{jets}}} SF_{\mathrm{jet}}^{i} = SF_{\mathrm{jet}}^{1} \cdot SF_{\mathrm{jet}}^{2} \cdot \ldots \qquad (9.5)$$

Weighting each event by the corresponding *CSV-weight*, the shape of the CSV-discriminant-value distribution in data gets described by Monte Carlo much better, as shown in Figure 9.10a. The improvement of the discriminant shape leads to the better description of the b-tagged-jet multiplicity, which is demonstrated in Figure 9.10b. The detailed description of the CSV-reshaping technique is given in the dedicated documentation [228].

A drawback of the described technique is the lack of data with light-flavour jets that have large CSV-discriminant value, which leads to the large statistical uncertainty on the scale-factor for mistagged jets, which is visible in the last bins of the distribution in Figure 9.9a. This is the reason behind the large statistical uncertainty in the last bin of the b-tagged-jet multiplicity in Figure 9.10b. Explicitly checking the MC events with 5 b-tagged jets, it was found that one of the events has 3 jets with extremely high CSV-discriminant values, above 0.99, and are scaled as light-flavour jets. A more detailed information about this event can be found in the dedicated presentation [229]. A possible explanation of such behaviour can be a wrongly reconstructed primary vertex leading to the wrong input for the CSV-discriminant calculation for all the jets.

The scale factors had originally been obtained for jets with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$, but later were reevaluated by adding one more bin: $20 < p_{\mathrm{T}} < 30\,\mathrm{GeV}$, to a large extent for the needs of the measurements described in this thesis.
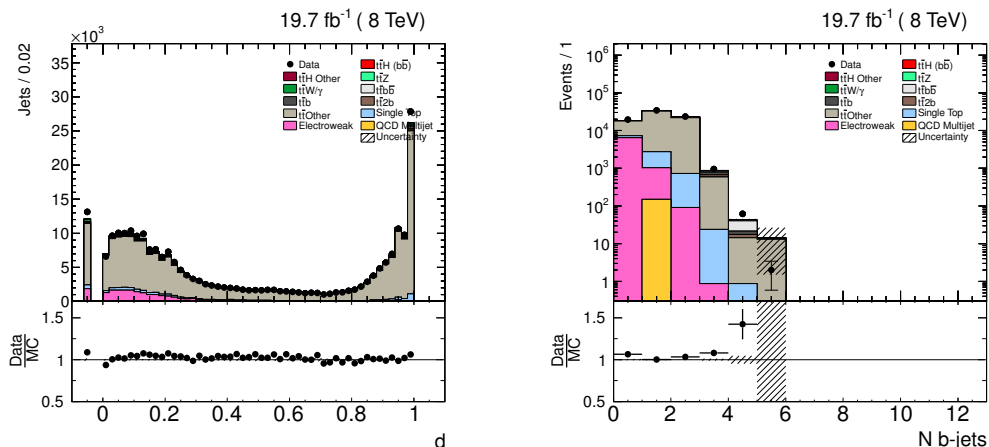
(a) CSV discriminant of every jet in acceptance ($p_T > 20\,\mathrm{GeV}$, $|\eta| < 2.4$)

(b) multiplicity of CSVM-tagged jets

Figure 9.10: Requirements of b-tagged jets for the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ processes. A jet is b-tagged if it fulfils the CSVM working point requirement. Jets with negative discriminant value do not have enough tracks associated to it for the proper value calculation. The number of b-tagged jets in an event must be 3 or 4 for the $t\bar{t}b(\bar{b})$ cross-section measurement and exactly 4 for the $t\bar{t}b\bar{b}$ measurement.

Event selection

Since a $t\bar{t}b(\bar{b})$ event has at least 3 b quarks in the final state, events with 3 or 4 b-jets are selected for the $t\bar{t}b(\bar{b})$ cross-section measurement. For the $t\bar{t}b\bar{b}$ cross section measurement, events with exactly 4 b jets are selected. There are only 2 events in data that have more than 4 b jets and their amount is inadequately higher in MC. These events are dominated by the $t\bar{t}Other$ process, which means that there are at least three mistagged jets, which are assigned with very large weights by the CSV-reshaping, which also have very large uncertainty, as demonstrated in Figure 9.9a. Thus, events with 5 or more b-tagged jets are not used for the measurement. The background composition in events with different numbers of b-tagged jets is shown in Figure 9.10b.

9.5   Estimation of the signal normalisation

After establishing the full event selection and corresponding object corrections, it is worth to look at the background composition of the selected events. This can be demonstrated by the distributions of the b-jet pseudorapidity shown in Figure 9.11.

The signal for the $t\bar{t}b(\bar{b})$ measurement is constituted by three contributions from the plot: $t\bar{t}b\bar{b}$, $t\bar{t}b$ and $t\bar{t}2b$. All other contributions are backgrounds and have to be subtracted from data. For the $t\bar{t}b\bar{b}$ cross-section measurement, only the $t\bar{t}b\bar{b}$ process is considered as signal.

It was mentioned in Section 9.4.6 that the CSV-reshaping improves the description of the b-tagged-jet multiplicity. Nevertheless, the description is still not optimal

(a) $t\bar{t}b$-event selection)

(b) $t\bar{t}b\bar{b}$-event selection
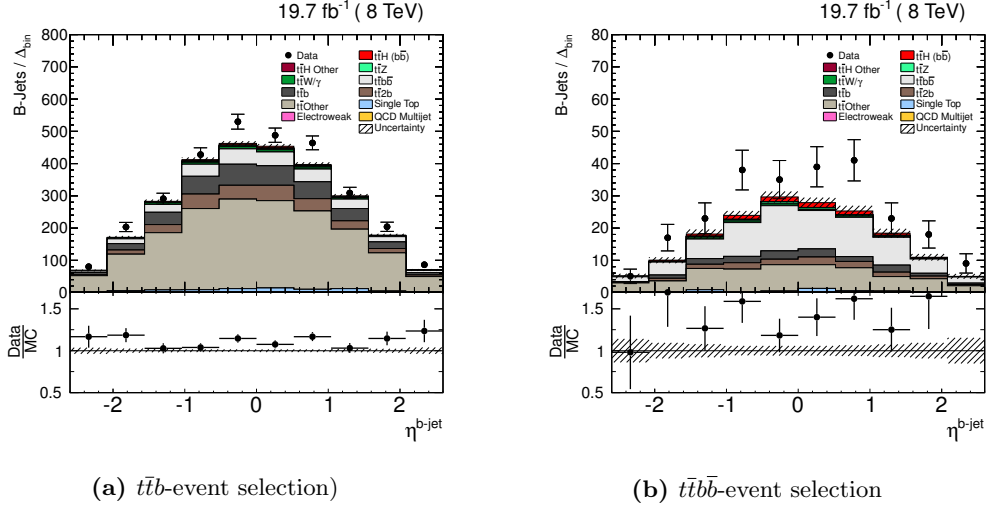
Figure 9.11: Pseudorapidity of b-tagged jets in acceptance after the full $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ event selection. The difference in normalisation of the data and MC simulation is caused by the remaining mismodelling of the b-tagged-jet multiplicity shown in Figure 9.10b.

and is the most obvious reason for the too low normalisation of the MC simulations, visible in Figure 9.11. Furthermore, a previous measurement of the inclusive $t\bar{t}b\bar{b}/t\bar{t}jj$ cross-section ratio, which used the same MC simulations of the $t\bar{t} + jets$ events, showed that the fraction of the $t\bar{t}b\bar{b}$ process is underestimated by about 40% [119].

The most obvious difference between the $t\bar{t}b\bar{b}$ process with respect to the $t\bar{t}jj$ process is in multiplicity of b jets. Therefore, the multiplicity of b-tagged jets is sensitive to the $t\bar{t}b\bar{b}$ fraction in the simulated sample.

9.5.1   Template fit configuration

It can be clearly seen from the Figure 9.10b that the low multiplicity region is dominated by the $t\bar{t}Other$ contribution, while at multiplicities above 2, the $t\bar{t}b\bar{b}$, $t\bar{t}b$ and $t\bar{t}2b$ contributions become dominant. The underestimation of the heavy-flavour-jets fraction leads to the increasing discrepancy between data and simulation at larger multiplicities of b-tagged jets. Events with 0 b-tagged jets have a significant contribution from Drell-Yan events, which add unnecessary dependence on its presence in the selected data. Thus, only events with at least 1 b-tagged jet are used, which are dominated by the $t\bar{t} + jets$ process, as shown in Figure 9.12a.

In order to correct the discrepancy, a template fit to the distribution of b-tagged-jet multiplicity has been employed. Three templates are used for the fit:

- **ttHF:** consists of the $t\bar{t}b\bar{b}$ and $t\bar{t}b$ processes (floating freely);

- **ttOther:** consists of the $t\bar{t}c\bar{c}$ and $t\bar{t} + light\,jets$ (floating freely);

- $t\bar{t}2b$**:** consists of the $t\bar{t}2b$ process (fixed);
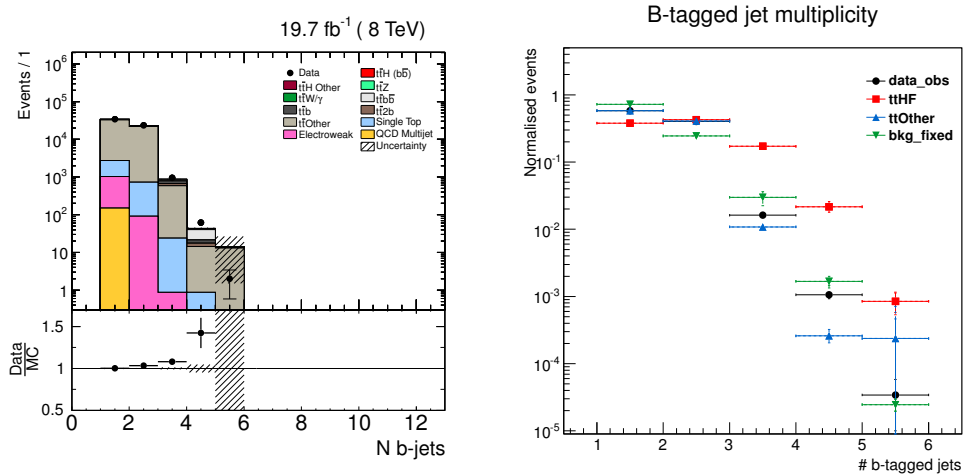
(a) Distribution of the number of *b*-tagged jets (CSVM) per event after the basic $t\bar{t}$ event selection (see Section 9.4) with the different requirement of jets: ≥ 2 jets, ≥ 1 *b*-tagged jet. At large b-tag multiplicities, the MC underestimates the number of events.

(b) Comparison of the shapes of the input templates as used in the fit. The multiplicity of *b*-tagged jets increases from the *background* to the *ttOther* to the *ttHF* template, providing sufficient separation power for the fit. Error bars represent the sum of statistical and systematic shape uncertainties included for the individual template.

Figure 9.12: Distributions of the *b*-tagged-jet multiplicity in the MADGRAPH+PYTHIA simulation, showing discrepancy with data (a) and the difference in shapes of the input templates (b).

- **background:** a combination of the remaining background processes, including all non-dileptonic $t\bar{t}$ events (fixed).

The $t\bar{t}c\bar{c}$ process is defined as a process that has no additional b jets and at least one c jet in acceptance ($p_{\mathrm{T}} > 20\,\mathrm{GeV}$, $|\eta| < 2.4$).

For the fit to be sensitive to the contribution of each template, their shapes have to be different, which is the case for the b-tag multiplicity, as can be seen in Figure 9.12b. The first two bins of the distribution, being dominated by the *ttOther* template, constrain its normalisation, while the next two bins can mainly define normalisation of the *ttHF* template. Therefore, the normalisations of the *ttHF* and *ttOther* templates are set to be the floating parameters of the fit. The *background* template is predominantly contained in the first bin of the distribution, and should be constrained by data even better, but its small contribution lowers the sensitivity of the fit to this contribution, which has anyway minor influence on the *ttHF* template in the 3-rd and 4-th bins of the b-tag multiplicity distribution.

The fit is performed by the Higgs `combine` tool [230], which is based on the collection of statistical tools, RooStats [231]. The used `combine` tool version is `v4.0.1-sl5`, and it was installed in the `CMSSW_6_1_2` release. The normalisations of the *ttHF* and *ttOther* templates were allowed to float freely in the range:

$$\frac{1}{100} \times N_{initial}^{t\bar{t}\mathrm{X}} < N_{final}^{t\bar{t}\mathrm{X}} < 100 \times N_{initial}^{t\bar{t}\mathrm{X}}\,, \tag{9.6}$$

with respect to its initial normalisation ($N_{initial}^{t\bar{t}\mathrm{X}}$) as provided by the MADGRAPH simulation. In the control card for the `combine` tool, the floating-normalisation parameters were configured as uniformly distributed (`lnU`).

The fit was performed in the background-only mode and the result of the fit is stored as new histograms of each template, which are corrected to fit the data distribution best. The template fit does not only change the normalisation of the input templates, but also their shape, according to the provided statistical and systematic variations of the input templates. For the flexibility of the analysis, the correction of the $t\bar{t} + b\,jets$ fraction is performed not in on the event-by-event basis, but to the whole sample. Thus, it was not possible to apply different scale factor to events with different number of b-tagged jets.

A single-number scale factor for each sample was obtained as a ratio of the fitted and input templates:

$$SF_{t\bar{t}\mathrm{X}} = \frac{\int t\bar{t}\mathrm{X}_{\mathrm{fitted}}}{\int t\bar{t}\mathrm{X}_{\mathrm{input}}}\,. \tag{9.7}$$

To every bin of the fitted template an uncertainty is assigned by the `combine` tool. A set of individual bin-uncertainties of a fitted template is translated into a single uncertainty on the scale factor by adding all the bin uncertainties linearly, as fully correlated:

$$\Delta_{SF} = \sum_{i}^{N_{bins}} \Delta_i\,. \tag{9.8}$$

9.5.2   Uncertainties on the shapes of the templates

Since the input templates are known to some finite precision, their shape can vary
within statistical uncertainty as well as due to systematic effects. Thus, statistical
uncertainties on the shapes of each template are included in the fit. A set of consid-
ered systematic uncertainties that change shapes of the *ttHF* and *ttOther* templates
is also included and consists of the following uncertainties:

- jet energy scale (JES);

- purity of the light-flavour samples used for the determination of the CSV
  discriminant scale factors for b jets (details in [228]);

- statistical uncertainty on the light-flavour sample used for the determination
  of the CSV discriminant scale factors (details in [228]).

For the *ttOther* template, additional sources of uncertainty are taken into account
since they change the shape significantly with respect to the statistical uncertainty
of the template distribution:

- purity of the heavy-flavour samples used for the determination of the CSV
  discriminant scale factors for light-flavour jets (details in [228]);

- statistical uncertainty on the heavy-flavour sample used for the determination
  of the CSV discriminant scale factors (details in [228]);

- cross section of the $t\bar{t}c\bar{c}$ process: varied up/down by 50%.

All of the systematic variations listed above were also included for the $t\bar{t}2b$ fixed
template, while its cross section was also varied as is discussed later.

It was ensured that other systematic effects, i.e. jet energy resolution, lepton-
isolation scale factors, trigger scale factors, pile-up correction, have negligible influ-
ence on the shapes of the templates compared to the statistical uncertainties and
were not included in the fit.

Normalisation of the $t\bar{t}2b$ process

The $t\bar{t}2b$ process, which corresponds to events with two additional b-hadrons that are
close enough to produce a single b-jet, is not included in the *ttHF* template. Instead,
its contribution is fixed, since the $t\bar{t}2b$ cross section is corrected externally by a factor
$1.74^{+0.69}_{-0.74}$, which is taken from the independent measurement of differential $Zb\bar{b}$ cross
section as a function of $\Delta R(b, \bar{b})$ [232]. The uncertainty on the $t\bar{t}2b$ cross section
is taken as an envelope of the uncertainties from the measurement and is extended
to include the value predicted by the MADGRAPH used for the current analysis,
meaning a correction factor of 1.0. The variation of the $t\bar{t}2b$ cross section is included
as a systematic uncertainty on the shape of the corresponding template.

Several methods were tried to include the $t\bar{t}2b$ as a floating parameter of the fit.
Using it directly as a floating parameter has certain sensitivity, because b jets initi-
ated by two close-by b hadrons have a higher probability to be b-tagged, since two
secondary vertices can be identified in such jets. Furthermore, the maximum mul-
tiplicity of secondary vertices per jet and sum of secondary-vertex invariant masses

were tried, which also have different spectra in $t\bar{t}b$ and $t\bar{t}2b$ processes. Nevertheless, the amount of available data is not large enough and statistical uncertainties are not small enough to constrain the $t\bar{t}2b$ contribution better than using the results from [232]. But with more data from the future runs of the LHC, the $t\bar{t}2b$ component could be treated as a floating parameter using information about secondary-vertex content of the b-tagged jets.

Technical configuration

Every effect of a single variation on the shape of the input template has to be stored as a separate histogram for the *combine* tool. Systematic uncertainties change consistently every bin of the distribution, which requires a single varied histogram to be stored for each up or down systematic variation.

The statistical uncertainty on the shape is not correlated across multiple bins of the template, and requires a separate histogram for every single-bin variation within its statistical uncertainty, keeping all other bins unchanged. Therefore, for the b-tagged-jet multiplicity with 5 bins and two variations within the statistical uncertainty (up and down), a total of $5 \times 2 = 10$ separate histograms of each template are provided to the fit.

9.5.3   Test of the fit performance on pseudodata

In order to check the performance of the template fit, it was tested on *pseudodata*, which is a combination of the MC samples used for the analyses, but treated as data in the fit. Since pseudodata is a direct combination of samples, the template fit should determine no additional correction, because the shape of the distribution in pseudodata exactly matches the one of the sum of the input MC templates. As shown in Figure 9.13a, the correction after the fit for pseudodata is equal to 1.0 ± 0.3.

In addition, the template fit was performed with the cross sections of the $t\bar{t}b\bar{b}$ and $t\bar{t}b$ processes in pseudodata increased by a factor 1.2, which distorts the shape of the distribution in pseudodata as shown in Figure 9.13b. After the fit, the proper correction factor 1.2 for the *ttHF* template is reproduced, making the pseudodata/MC ratio flat across all bins of the b-tag multiplicity distribution shown in Figure 9.13c. In case the $t\bar{t}b\bar{b}$ and $t\bar{t}b$ contributions are scaled by factor a 1.4 in pseudodata, it is also exactly reproduced by the template fit, as shown in Figure 9.13d.

The performed tests show that the template fit works as expected and can be safely applied on real data for the actual correction of the MC simulations.

9.5.4   Results of the fit using real data

When applied to real data, the template fit provides the following corrections:

- **ttHF:** 1.6 ± 0.4;

- **ttOther:** 1.01 ± 0.01.
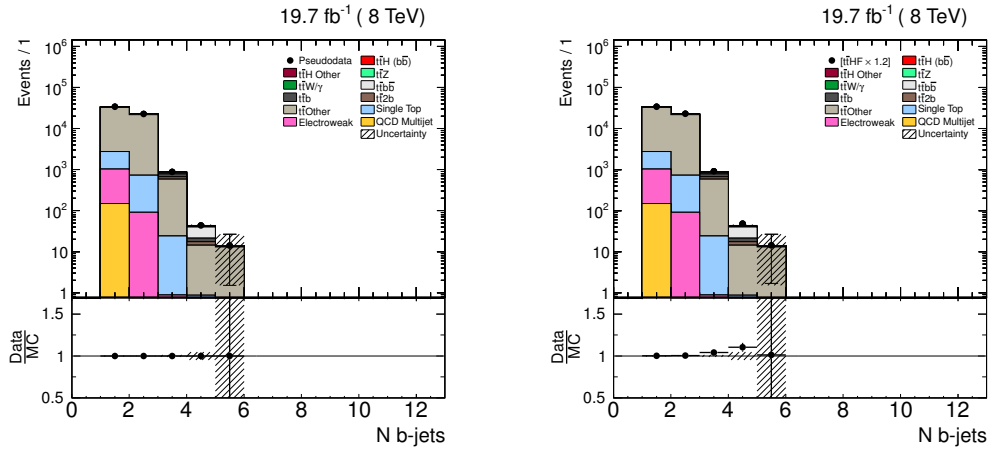
(a) direct sum of MC samples in pseudo-data (after the fit: $ttHF = 1.0 \pm 0.3$, $ttOther$: 1.0 ± 0.01)

(b) $\sigma_{t\bar{t}b\bar{b}}, \sigma_{t\bar{t}b} \times 1.2$ (before the fit)
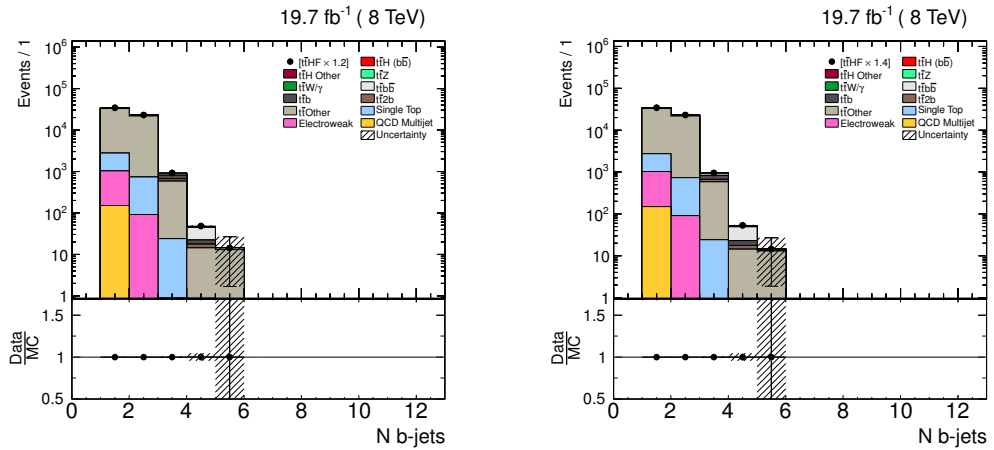
(c) $\sigma_{t\bar{t}b\bar{b}}, \sigma_{t\bar{t}b} \times 1.2$ (after the fit: $ttHF \times 1.2 \pm 0.43$, $ttOther \times 1.0 \pm 0.01$)

(d) $\sigma_{t\bar{t}b\bar{b}}, \sigma_{t\bar{t}b} \times 1.4$ (after the fit: $ttHF \times 1.4 \pm 0.39$, $ttOther \times 1.0 \pm 0.01$)

Figure 9.13: Distributions of the b-tagged jet multiplicity in MC simulation and pseudodata with modified normalisation of the $t\bar{t}b\bar{b}$ and $t\bar{t}b$ contributions in pseudodata. The template fit successfully reproduces the applied modifications.

The determined corrections eliminate the discrepancy shown in Figure 9.12a leading to the improved description of the b-tagged jet multiplicity in data by the MAD-GRAPH simulation, which is shown in Figure 9.14. The fact that the correction factor for the *ttOther* template is very close to 1.0 is an important cross check of the method, since the initial normalisation of the $t\bar{t} + jets$ sample is based on the latest inclusive $t\bar{t}$ cross section measurement [233]. Therefore, the *ttOther* component should be correct, except of the small $t\bar{t}c\bar{c}$ contribution, which is not known with high enough precision.
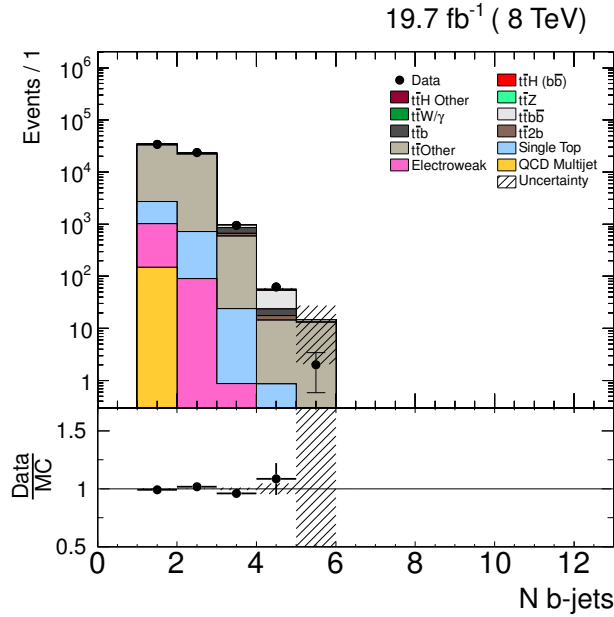
Figure 9.14: Distribution of the number of b-tagged jets (CSVM) per event with $t\bar{t}b\bar{b}$ and $t\bar{t}b$ cross sections corrected in accordance with the template fit results.

The application of the estimated correstions to the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ cross sections leads to the good agreement in normalisation between data and MC simulations, as shown in Figure 9.15.

Cross check with the tag-and-probe method

In order to ensure that the correction factors, determined by the template fit, can be safely applied in the analysis, a cross check has been performed using an independent quantity that is sensitive to the relative contribution from the $t\bar{t}b\bar{b}$, $t\bar{t}b$ and $t\bar{t}2b$ processes. For this purpose, a so-called *tag-and-probe* method was used in events with at least 4 jets, among which 3 or more are b-tagged. The jets are ordered in descending order by the value of the CSV discriminant. The medium working point (CSVM) used for counting the number of b-tagged jets for the multiplicity distribution corresponds to the value $d = 0.679$.

The third jet is always b-tagged and is referred to as a *tag jet*, while the fourth jet can have CSV discriminant value either above (tagged) or below (not tagged) 0.679 and is referred to as a *probe jet*. The distribution of the CSV discriminant value for the probe jet should be closer to 1.0 for real b jets that are present in the

(a) $t\bar{t}b(\bar{b})$-event selection

(b) $t\bar{t}b\bar{b}$-event selection
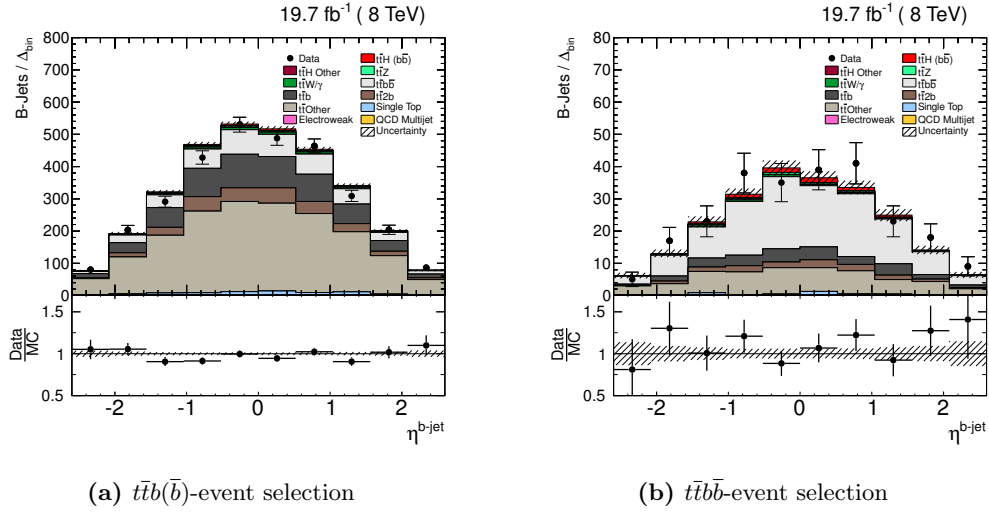
Figure 9.15: Pseudorapidity of b-tagged jets after full $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ events selection with the $t\bar{t} + HF$-fraction correction applied.

$t\bar{t}b\bar{b}$ process and should tend to 0 for non-b-jets, which dominate in the rest of the processes. Thus, the distribution of the CSV discriminant for the probe jet before and after applying the $t\bar{t} + HF$ fraction correction has been compared and is shown in Figure 9.16. It can be seen from the plots that the relative contribution of the $t\bar{t}b\bar{b}$ process increases with higher value of the b-tag discriminant (d), as does the disagreement between data and MC simulation. After applying the correction, the agreement improves significantly across the whole range of d except of the very first bin, which can be due to the remaining mismodelling of the CSV discriminant shape after the applied shape reweighting [228].

(a) without the $t\bar{t} + HF$ correction

(b) with the $t\bar{t} + HF$ correction
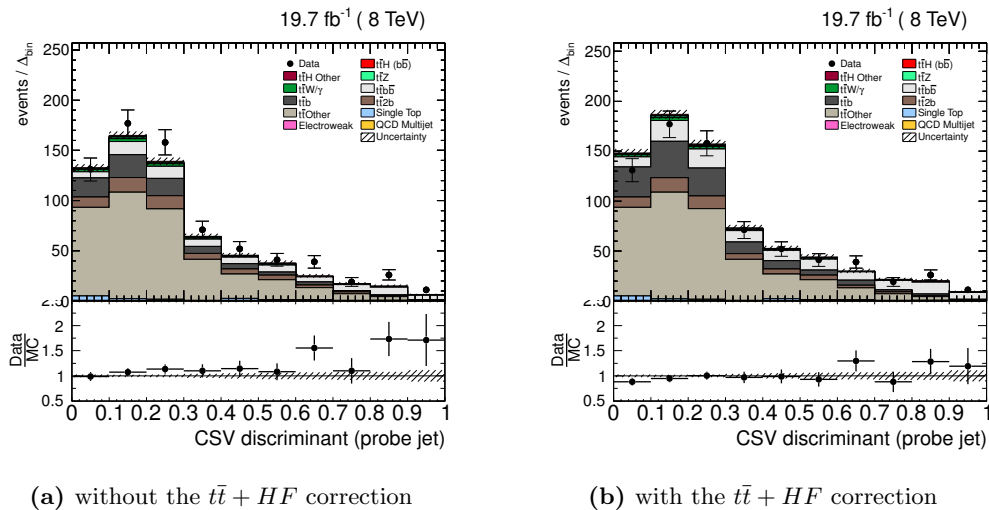
Figure 9.16: Distribution of the CSV discriminant value (d) for the fourth probe jet in data and in MC simulations with and without the correction of the $t\bar{t} + HF$ fraction.

Since after the $t\bar{t}b$ event selection, almost $50\,\%$ of the events correspond to the

$t\bar{t}Other$ background, it is important to ensure that its normalisation after the selection is correct. The same tag-and-probe method was used to check this. Events were selected that have at least 3 jets, among which 2 or 3 are b-tagged with the CSVM working point. The probe jet in this case is the third jet, according to the descending ordering by the CSV discriminant value (d). The distribution of d for the probe jet is shown in Figure 9.17. The events with values below 0.679 correspond to 2 b-tagged jets, while the ones with values above 0.679 have 3 b-tagged jets.

(a) without the $t\bar{t} + HF$ correction   **(b)** with the $t\bar{t} + HF$ correction
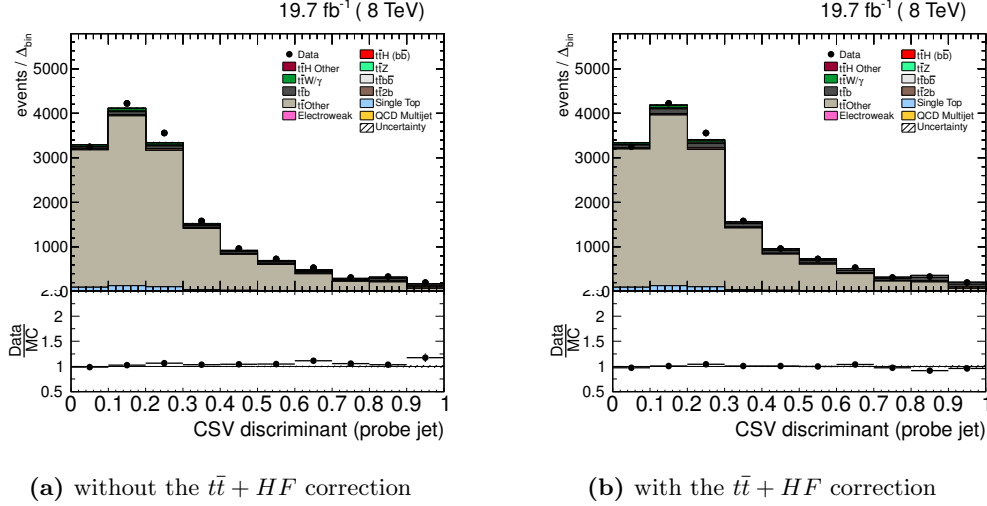
Figure 9.17: Distribution of the CSV discriminant value (d) for the third probe jet in data and in MC simulations with and without the correction of the $t\bar{t}+HF$ fraction.

The plots show that the shape of the Data/MC ratio is flat across the whole d range, which means that the normalisation of the $t\bar{t}Other$ component after the full $t\bar{t}b$-event selection is correct.

9.5.5   Event yields

Once the event selection and correction of MC simulations are established, it is worth to look at contributions of the individual signal and background processes to the final selected sample in data. An overview of the numbers is presented for the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ event selections in Table 9.3.

In total, 1007 and 62 events are selected in data for the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ measurements respectively. The total number of selected events in MC simulations is in agreement with data within statistical uncertainties.

After the full $t\bar{t}b(\bar{b})$ event selection, about 40 % of events correspond to the signal processes: $t\bar{t}b$, $t\bar{t}2b$ and $t\bar{t}b\bar{b}$. The number of events after the $t\bar{t}b\bar{b}$ event selection is much more limited by the available statistics. Thus, it is especially important to have high purity of the selected $t\bar{t}b\bar{b}$ sample, which is slightly higher than 50 %.

The dominant background to both the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ signal processes is $t\bar{t}$ production with additional light or charm jets. Another substantial background to the $t\bar{t}b\bar{b}$ process is $t\bar{t}$ production with a single additional b jet. Such backgrounds can survive the event selection either due to the higher p_T or lower $|\eta|$ of the reconstructed

Table 9.3: Number of selected events in data and in different simulated processes, after applying the full event selection criteria.

Sample	$t\bar{t}b(\bar{b})$	$t\bar{t}b\bar{b}$
Data	1007 ± 32	62 ± 8
Simulation (total)	1040 ± 5	57.1 ± 0.7

Simulated process	Fraction of events, %	
$t\bar{t}b\bar{b}$	**12.5**	**52.5**
$t\bar{t}b$	**18.6**	10.7
$t\bar{t}2b$	**8.1**	5.6
$t\bar{t}c\bar{c}$	4.0	4.0
$t\bar{t} + light\,jets$	50.2	17.6
$t\bar{t}$ non-dilepton	1.3	2.0
Single top	2.3	1.4
$t\bar{t}\gamma$	0.9	0.9
$t\bar{t}Z$	0.6	1.3
$t\bar{t}H(b\bar{b})$	0.9	3.2
Other backgrounds	0.6	0.8

The numbers are shown separately for the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ event selections. The events from simulation are normalised to the data luminosity $\mathcal{L} = 19.7$ pb. All the corrections described in Sections 9.4 and 9.5 are applied to the simulations. Fractions of the signal processes for each event selection are given in bold. In control distributions, the $t\bar{t}c\bar{c}$, $t\bar{t} + light\,jets$ and non-dileptonic $t\bar{t}$ decays are combined into a single $t\bar{t}Other$ contribution.

b jets or due to light or charm jets being mistagged with the CSV discriminant value higher than 0.679.

Single-top-quark events can pass the selection due to the presence of an additional W boson decaying leptonically, which fulfils the dilepton channel requirement and an additionally radiated hard jet can pass the jet requirements. The minor $t\bar{t}\gamma$ background process can, for example, be selected if one of the top quarks decays hadronically, producing a jet that passes the jet selection, while the photon is misidentified as an electron, which fulfils the dilepton channel requirement.

In processes like $t\bar{t}H$ and $t\bar{t}Z$, the additionally produced bosons can decay to a pair of b quarks, giving rise to a pair of genuine b jets. Other minor backgrounds mainly represent multiple-vector-boson production and $t\bar{t}H$ production with all other Higgs-boson decay channels except $H \to b\bar{b}$. Since no simulated QCD Multijet events have more than 1 b-tagged jets, it does not contribute to the selected sample. Therefore, this contribution is removed from the legend in the plots shown hereafter.

Control distributions

In order to ensure that the MC simulations describe data, a set of control distributions is shown separately for the $t\bar{t}b$ and $t\bar{t}b\bar{b}$ event selections. The distributions of the number of jets per event and their transverse momentum are shown in Figure 9.18. The transverse momentum and pseudorapidity of b-tagged jets are shown in Figure 9.19. The number of leptons in each event and the distance in ϕ between the two that were used to classify the event as dileptonic, are shown in Figure 9.20. The missing transverse energy and its direction in ϕ is shown in Figure 9.21.

From all the plots it can be concluded that MC simulations describe real data within the available statistical precision. The $t\bar{t}$ sample, which constitutes almost $100\,\%$ of the selected events, has 10-times larger statistics than the recorded data. Therefore, some bins of the control distributions are filled with simulated entries, but have no contributions from the real data.

9.6   Identification of additional b-jets

Even after suppressing events that have no additional b jets using the optimal event selection, another issue becomes relevant. Even in the clean $t\bar{t}b\bar{b}$ process, the cross section has to be measured as a function of properties of the additional b-jets; not any b jets.

In the recorded data, there is no information about the origin of each b jet. Thus, it is only possible to identify with a certain probability whether a jet is a b jet. But additional b jets do not have any particular properties that would allow to distinguish them from b jets originating from the $t\bar{t}$-system decay.

Therefore, it is easier to invert the problem and first identify b jets from the $t\bar{t}$ decay. This approach is more realistic, because the $t\bar{t}$ system as a whole and individual top quarks imply certain correlations between properties of the b jets, leptons and $\not{E}_{\mathrm{T}}$ originating from the $t\bar{t}$ decay.

(a) multiplicity of jets in acceptance

(b) transverse momentum of jets in acceptance
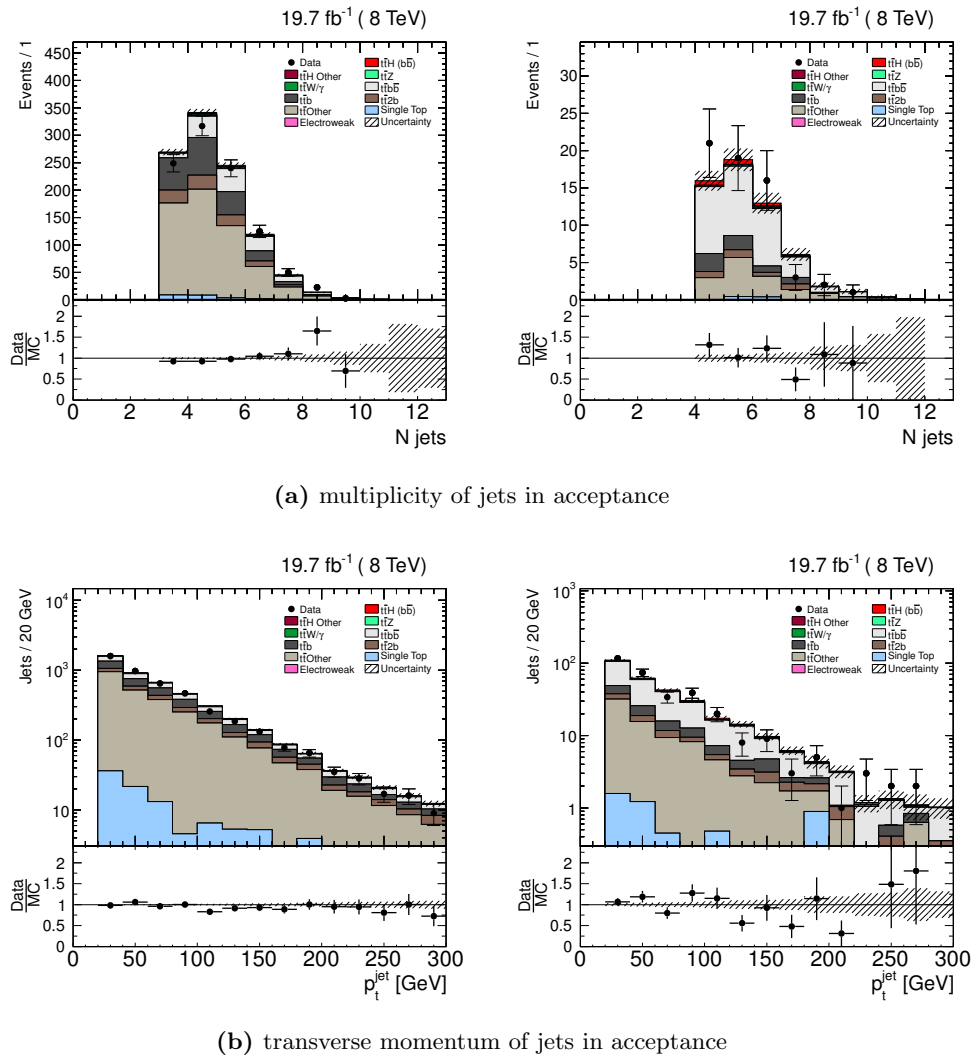
Figure 9.18: Properties of the jets in acceptance after the full event selection for the $t\bar{t}b$ (left) and $t\bar{t}b\bar{b}$ (right) cross section measurement.
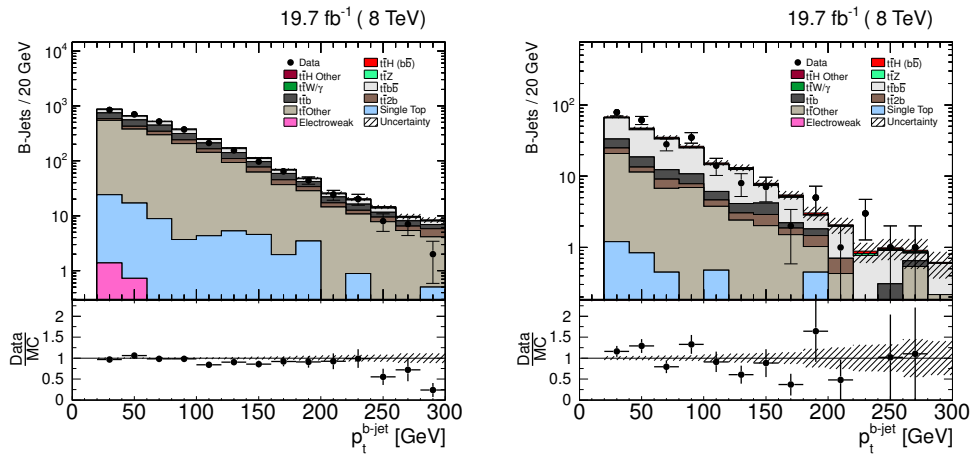
(a) transverse momentum of *b*-tagged jets in acceptance
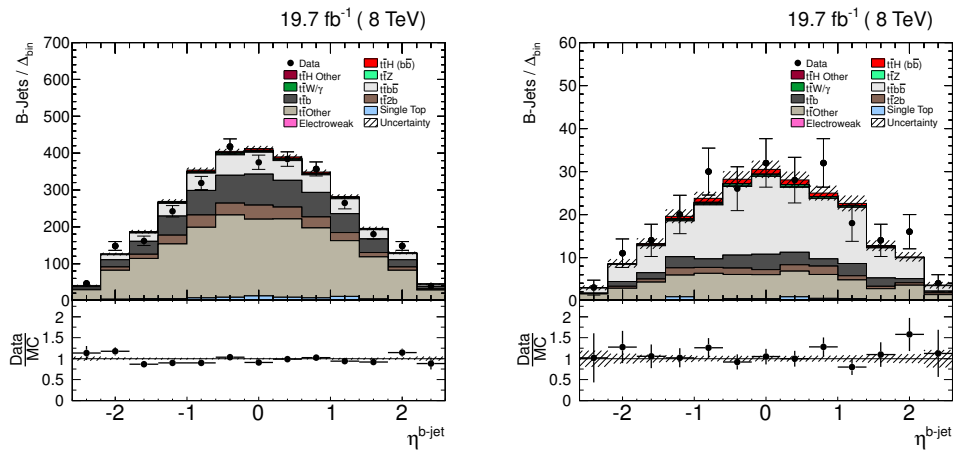
(b) pseudorapidity of *b*-tagged jets in acceptance

Figure 9.19: Properties of the *b*-tagged (CSVM) jets in acceptance after the full event selection for the $t\bar{t}b$ (left) and $t\bar{t}b\bar{b}$ (right) cross section measurement.
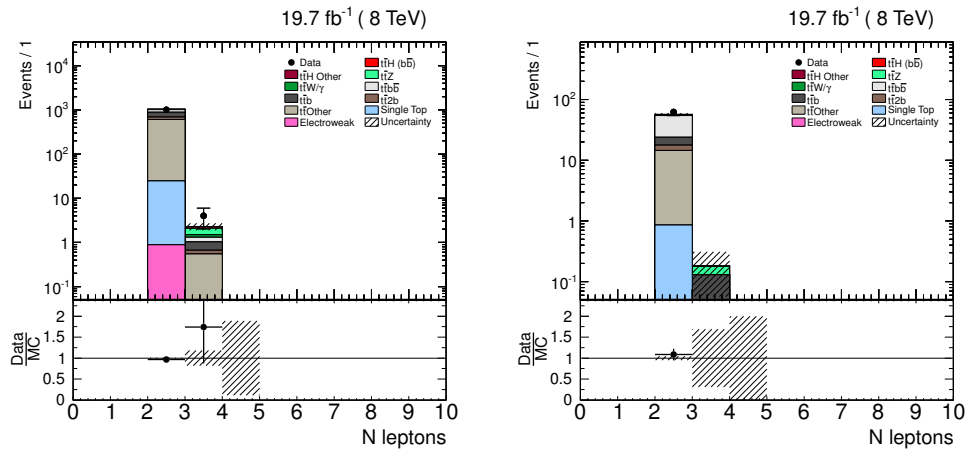
(a) multiplicity of leptons in acceptance
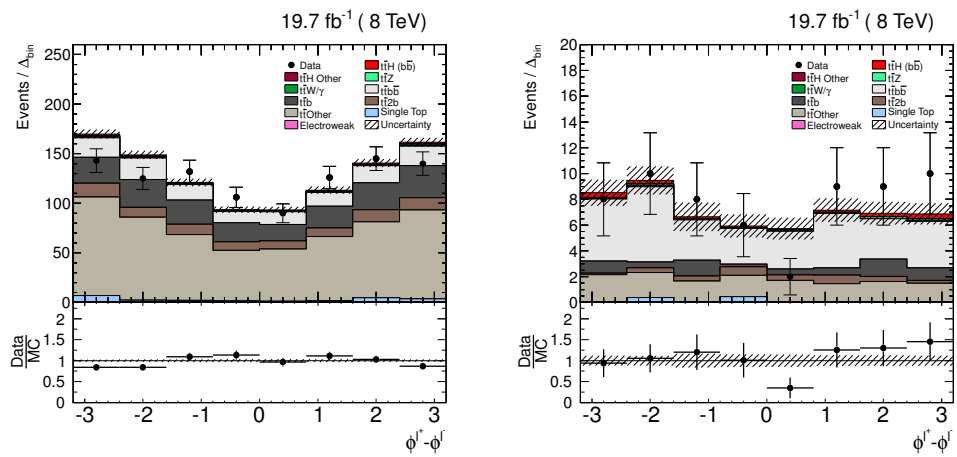
(b) difference in ϕ of the two oppositely charged leptons with highest p_T

Figure 9.20: Properties of leptons in acceptance after the full event selection for the $t\bar{t}b$ (left) and $t\bar{t}b\bar{b}$ (right) cross section measurement.

(a) missing transverse energy

(b) direction of the missing transverse energy in ϕ
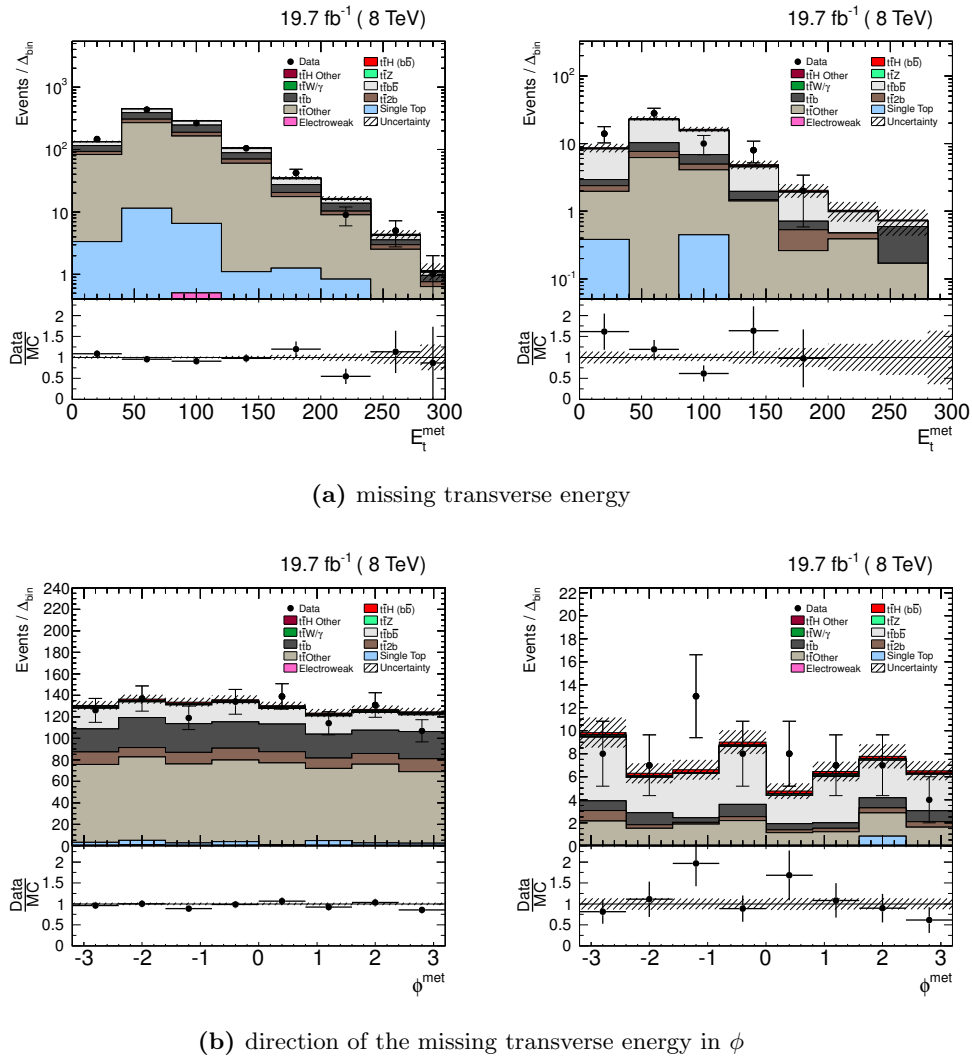
Figure 9.21: Properties of the $\not{E}_T$ after the full event selection for the $t\bar{t}b$ (left) and $t\bar{t}b\bar{b}$ (right) cross section measurement.

Additional b jets at the reconstruction level

Additional b jets are defined as all the b-tagged jets in an event, which are not assigned to the $t\bar{t}$ system. All the remaining jets are ordered by their p_T in descending order. Thus, the definition of additional b jets at the reconstruction level is as follows:

- **1-st additional b jet:** the first highest-p_T b-tagged jet not assigned to the $t\bar{t}$ system;

- **2-nd additional b jet:** the second highest-p_T b-tagged jet not assigned to the $t\bar{t}$ system.

While additional b-jets are required to be b-tagged, the jets assigned to the $t\bar{t}$ system can have any CSV-discriminant value.

Identification of b jets from the $t\bar{t}$ system

A standard approach for the identification of $t\bar{t}$ decay products is based on its kinematic reconstruction from the 4-momenta of leptons, jets and $\not{E}_T$. It was used in the differential $t\bar{t} + jets$ cross-section measurement [223] and showed high identification efficiency above $80\,\%$. Nevertheless, the problematic feature of the kinematic reconstruction is that it gives preference to b-tagged jets when assigning a jet to the $t\bar{t}$ system. This approach performs well for the $t\bar{t} + jets$ process, where most of events do not have additional b jets. On the contrary, in the $t\bar{t}b\bar{b}$ process with at least 4 b jets per event, such a decision does not make much sense and leads to a reduced efficiency of proper b-jet-pair identification.

Therefore, another approach was used for identification of the b jets from the $t\bar{t}$ system decay, based on the Multivariate Analysis (MVA) technique. Its main purpose is not the reconstruction of the whole $t\bar{t}$ system, but only the identification of b jets originating from the $t\bar{t}$ decay. This method is optimised specifically for events with high b-jet multiplicities. The method was developed by the DESY $t\bar{t}H$ group and is documented in [227].

In the following sections, just a brief overview of the MVA method is given, while the test of its performance is described in more detail, because it was a part of the analysis described in this thesis.

9.6.1   MVA strategy

In an event with 4 jets, 6 different dijet combinations can be constructed. Thus, the probability of a random assignment of a pair of jets to the $t\bar{t}$ system being correct is $\dfrac{1}{6} \approx 17\,\%$. The task of the MVA is to use the combination of different properties of dijet pairs to decide which one is most probable to originate from the $t\bar{t}$-system decay with higher success rate than just random assignment.

The first step in setting up the MVA is a definition of a set of dijet properties that are distributed differently for dijet pairs from $t\bar{t}$ decays and for all other jet pairs. Such properties should provide a separation power between different types of jet pairs allowing the MVA to distinguish between the two.

Then the MVA is trained on MC events in which the type of each jet pair is known. With a proper set of variables and enough statistics the MVA can produce a complicated function that converts the combination of dijet-pair properties into a probablity of the pair to originate from the $t\bar{t}$ decay.

Input variables

The MVA was developed to perform optimally for $t\bar{t}b\bar{b}$ and $t\bar{t}H(b\bar{b})$ measurements and not to bias one measurement towards the other. A set of variables for the MVA training was chosen carefully to avoid dependence on the kinematic properties of additional jets, especially the dijet mass, which is particularly sensitive to the $t\bar{t}H(b\bar{b})$ contribution. The variables combine information about the two final-state leptons, the $\not{E}_\mathrm{T}$ and the jets themselves.

All the variables need a distinction between b and $\bar{b}$ jets at the reconstruction level. This distinction is achieved on a statistical basis using the relative-track-momentum-weighted jet-charge definition, as described in detail in a dedicated note [209]. In the particular MVA used for the analysis, the inclusive jet-charge definition was used, with the squeezing parameter $x = 0.8$. A more advanced definition with special treatment of electrons and muons inside a jet was not yet available at the time of this analysis.

Based on the jet charge, in each jet pair, one jet is hypothetically treated as a b jet, and the other one as a $\bar{b}$ jet. A candidate jet pair, which is assigned by the MVA to the $t\bar{t}$ system, can be classified as one of the three types:

- **correct:** both jets originate from the $t\bar{t}$ decay, and are assigned with the proper charge;

- **swapped:** both jets originate from the $t\bar{t}$ decay, but each jet is assigned with the opposite charge;

- **wrong:** at least one of the jets does not originate from the $t\bar{t}$ decay.

This classification is schematically shown in Figure 9.22.
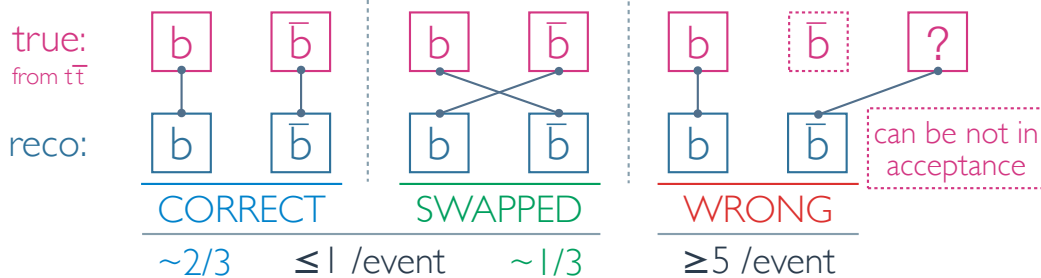
Figure 9.22: Types of jet pairs depending on the jet charge assignment in a $t\bar{t}b\bar{b}$ events with at least 4 b jets corresponding to at least 6 b-jet combinations.

An example of separation power of several variables used in the MVA is shown in Figure 9.23. It can be seen that the distributions for correct pairs differ in shape from the wrong combinations, while some variables for swapped pairs are more similar to the wrong ones. Nevertheless, the reconstruction-level jet charge makes it possible to have the proper assignment of b and $\bar{b}$ jets within a jet pair in about 67 %. The complete list of variables with corresponding explanations can be found in [227].

(a) average distance in ϕ between each b-jet candidate and the missing transverse energy

(b) distance in $\eta-\phi$ between a b jet and a lepton from hypothetically different t-quark decay

(c) difference in invariant masses of the jet pair and the hypothetical $b\bar{b}\ell^+\ell^-$ system

(d) difference in invariant masses of the jet pair and of all the other jets in the event
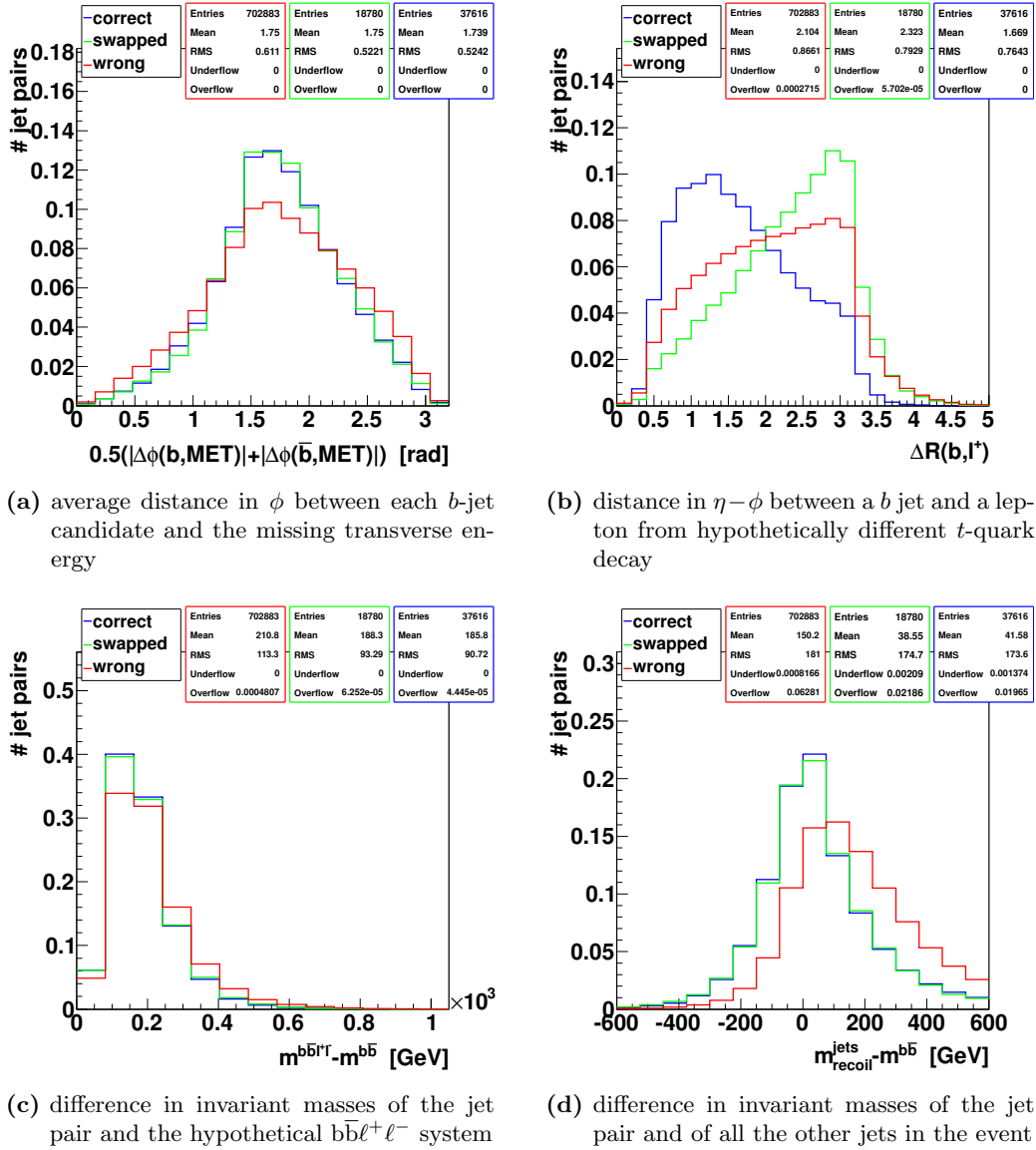
Figure 9.23: Example distributions of several input variables of the MVA, separated by three jet-pair types, as shown in Figure 9.22. The distributions are normalised to 1.0 for the shape comparison. On average there are 10 wrong pairs for each correct or swapped jet combination due to the multiple number of jets: ≥ 4.

Training setup

While the MVA was developed and set up by other members of the DESY ttH group, the actual training and testing of the performance was a part of the described analysis.

The MVA was designed to distinguish between b jets from the $t\bar{t}$ decay and additional b jets, which can originate from gluon splitting in $t\bar{t}b\bar{b}$ process or from the Higgs-boson decay in the $t\bar{t}H(b\bar{b})$ process. The subset of $t\bar{t}b\bar{b}$ events in the $t\bar{t} + jets$ sample used for the measurement was not suitable for the MVA training due to its limited statistics. For $t\bar{t}H$-related analyses a set of $t\bar{t}H(b\bar{b})$ samples were produced by the CMS collaboration, with different Higgs-boson-mass assumptions. Such samples perfectly suit the needs of the MVA training, because every event has 4 genuine b quarks, two of which originate from top-quark decays. Thus, 6 different $t\bar{t}H(b\bar{b})$ samples with m_H ranging from 110 GeV to 140 GeV in 5 GeV steps were combined for the training.

On the one hand, the MVA should be trained on events selected with the same criteria that are used in the measurement, i.e. 3 or 4 b-tagged jets. On the other hand, it is necessary to maximise the available statistics for the training in order to achieve stable performance of the MVA. Therefore, the event selection described in Section 9.4 was loosened for the training with the following changes:

- ≥ 4 jets in acceptance ($p_T > 20\,\mathrm{GeV}, |\eta| < 2.4$)

- ≥ 1 b-tagged jet (CSVM)

Since the swapped combinations behave differently depending on the particular variable, they are omitted from the training completely. Thus, in every $t\bar{t}H$ event that fulfils the selection, at most 1 correct pair and at least 5 wrong pairs were used in the training. After the combination of all the samples and full event selection, 56 176 correct and 858 147 wrong jet pairs were available for the training. The actual number of jet pairs is twice higher, but half of them are used for testing purposes. The testing is required to estimate possible bias of the MVA and needs a statistically independent sample. Further details on the training setup can be found in the documentation [227].

9.6.2   Performance of the *b*-jets identification

The output of the MVA is obtained after providing the values of all the variables used for its training. The value, return by the MVA for a single jet combination can be interpreted as the probability of a jet pair to represent the two b jets from the $t\bar{t}$ decay. The actual range of the value is arbitrary, which is a purely technical feature of the MVA.

The distribution of the MVA output for different types of jet pairs is shown in Figure 9.24. Since the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ cross sections are measured as functions of additional b jets, both *correct* and *swapped* combinations are equally good for the analysis. From the MVA-output distributions it can be seen that the *correct* pairs tend to have higher values, while the peak for *wrong* combinations is shifted to the smaller-values region. The output for the *swapped* combinations is almost uniformly distributed.
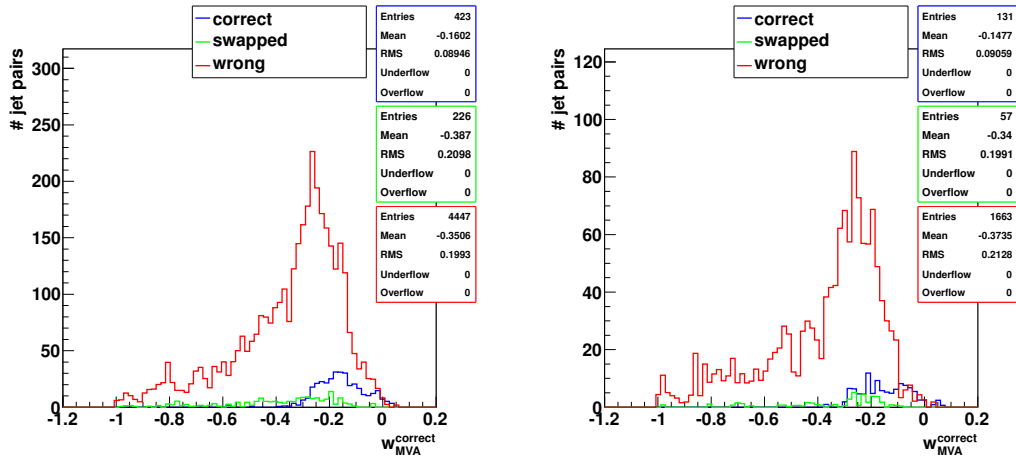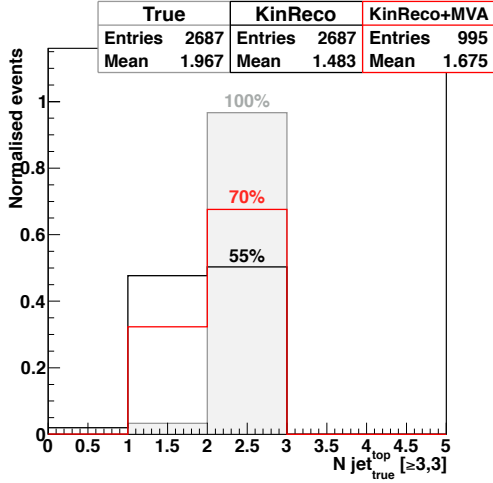
(a) $t\bar{t}b$ and $t\bar{t}2b$ events with exactly 3 b-tagged jets in acceptance

(b) $t\bar{t}b\bar{b}$ events with exactly 4 b-tagged jets in acceptance
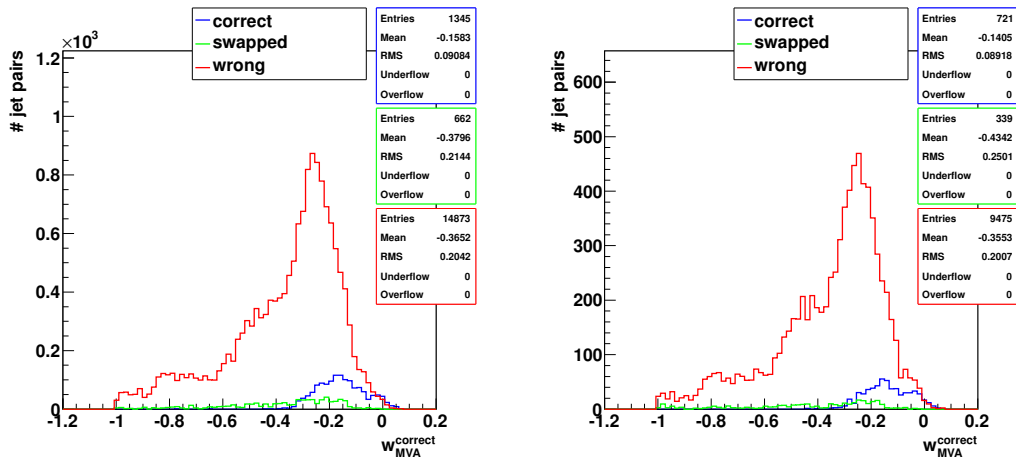
(c) $t\bar{t}Other$ events with exactly 3 b-tagged jets in acceptance

(d) $t\bar{t}H(b\bar{b})$ events with exactly 4 b-tagged jets in acceptance

Figure 9.24: Distributions of the MVA output for the three different types of jet pairs. The pair with the largest value of the MVA output is assigned to the $t\bar{t}$ system. There is clear separation between the *correct* and *wrong* jet combinations.

Thus, in order to identify a jet pair that most-likely originates from the $t\bar{t}$ decay, the pair with the highest MVA-output value is chosen. Therefore, the probability of the proper jet pair identification can be expressed as a ratio of the number of events with proper jet-pair identification to the number of events that have two distinct b jets from the $t\bar{t}$ decay in acceptance.

The proper jet-pair identification decay in MC simulation is established by the True method, which consists of the following steps:

- find the two b hadrons from the $t\bar{t}$ decay using the $GenHFHadronMatcher$, as described in Section 8.2;

- to each of the b hadrons, associate a generator-level b jet, as described in Section 8.1;

- for each of the generator-level b jets, find a closest reconstruction-level jet in $\eta - \phi$ space that fulfils both of the following requirements:

$$\Delta R(\mathrm{jet_{reco.}}, \mathrm{jet_{gen.}}) < 0.4$$

$$-0.5 < \frac{p_{\mathrm{T}}^{\mathrm{gen.}} - p_{\mathrm{T}}^{\mathrm{reco.}}}{p_{\mathrm{T}}^{\mathrm{gen.}}} < 0.6. \tag{9.9}$$

In order to better understand the level of performance provided by the MVA, it is also compared to the kinematic reconstruction method [223]. To estimate the probability of proper jet-pair assignment to the $t\bar{t}$ system, in every event three different jet multiplicities are checked:

- **True:** reconstruction-level jets assigned to the $t\bar{t}$ system using generator-level information (see Chapter 8);

- **KinReco:** jets assigned to the $t\bar{t}$ system by the kinematic reconstruction, but are also among the True jets;

- **MVA:** jets assigned to the $t\bar{t}$ system by the MVA, but are also among the True jets.

Thus, the number of events that have 2 *True* jets in acceptance is the highest possible success rate that can be achieved by the KinReco or MVA methods. The corresponding jet multiplicities are shown in Figure 9.25.

The comparison of the MVA and kinematic reconstruction does not show a clear advantage of the MVA for all the processes. Thus, a compromise must be made and the more important process has to be chosen. Considering the very limited statistics of events with 4 b-tagged jets in the data, the 20 % improvement for the $t\bar{t}b\bar{b}$ process compared to the KinReco method is more substantial than the 23 % reduction for the $t\bar{t}b$ and $t\bar{t}2b$ process.

In principle, the MVA could be applied in a more complicated way to improve the fraction of proper jet-pair identifications. For example, one might completely reject events in which the highest MVA-output value is lower than some threshold, e.g. $output_{\mathrm{MVA}} < 0.15$. This would eliminate events in which *wrong* combinations have values in the region of *correct* pairs. It is also possible to completely reject

(a) $t\bar{t}b$ and $t\bar{t}2b$ events with exactly 3 b-tagged jets in acceptance

(b) $t\bar{t}b\bar{b}$ events with exactly 4 b-tagged jets in acceptance

(c) $t\bar{t}Other$ events with exactly 3 b-tagged jets in acceptance

(d) $t\bar{t}H(b\bar{b})$ events with exactly 4 b-tagged jets in acceptance
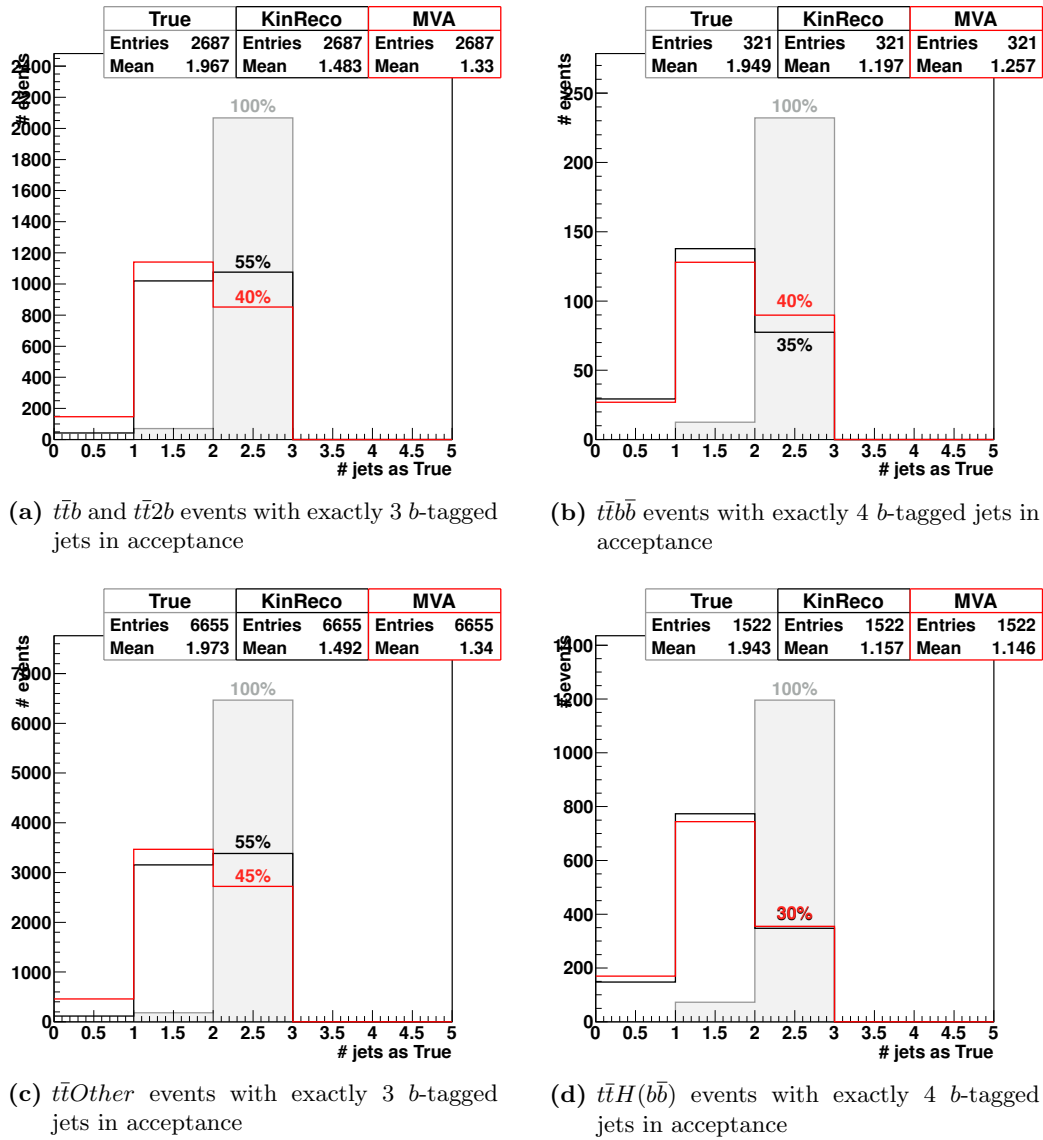
Figure 9.25: Multiplicity of properly identified b jets from the $t\bar{t}$ system by the kinematic reconstruction [223] and by the MVA. The identification method based on the generator-level information is marked as True. The distributions are shown for different processes with corresponding event selections.

events in which the difference between MVA-output values of the two best jet pairs is too small. In the remaining events the best pair would be more separated from others, hopefully leading to the increased probability of the pair being *correct*.

Nevertheless, these options are not applied in the described analysis, because the available statistics is very small with the existing event selection. Reducing it further by additional MVA-dependent event selection is not realistic in such situation.

Potential of the combined KinReco+MVA approach

Besides the pure multiplicity of properly identified jets, which was compared between the two methods in the previous Section, it is worth to look at the individual jets themselves. Thus, in the list of p_T-ordered jets, an index is assigned to each jet, ranging from 0 to $N_{jets} - 1$. The index corresponds to the position of the jet in the list. Thus, smaller indices correspond to jets with higher p_T and vice versa.

Using the described notation, each jet that is assigned to the $t\bar{t}$ system by the KinReco or MVA method, is compared to a potentially the same jet identified with the True method. The correlation between the indices of the jets identified by the True and the KinReco or MVA methods is shown in Figure 9.26a.

When looking at the properly identified jets, which are filled in the diagonal bins of the distribution, it can be seen that the MVA outperforms the KinReco in the region of small indices, while the KinReco performs better at large indices. Similarly, the wrongly identified jets tend to have smaller indices in the case of the MVA and larger indices in the case of the KinReco. Since smaller indices correspond to jets with higher p_T, it can be concluded that both methods have some bias in different directions.

This behaviour proves that the two methods are largely independent, which is a result of the completely different approaches used by the two. Thus, a combination of the KinReco and MVA can be another promising approach for optimisation of the jet-pair-identification performance.

The bias in different directions could be used to enhance the fraction of properly identified jet pairs by rejecting events in which the two methods have different results. The correlation between the KinReco and the MVA is shown in Figure 9.26b. The correlation plot shows that only in 50 % of the $t\bar{t}b\bar{b}$ events the two methods assign at least one common jet to the $t\bar{t}$ system. For the optimal definition of additional b jets, both jets from the $t\bar{t}$ system have to be properly identified. Thus, both methods assign identical jet pairs to the $t\bar{t}$ system only in 25 % of $t\bar{t}b\bar{b}$ events.

Nevertheless, if the same jet pair is identified by the both methods, there is a higher chance that the pair is correct. This is reflected in a multiplicity of properly identified jets by individual KinReco or MVA methods and by their combination. The result of the comparison is shown in Figure 9.27.

From the plots it can be seen that the combination of the two methods increases the fraction of properly identified jet pairs by 50 %, which is a very significant improvement. Unfortunately, this technique can not be used in the current analysis due to statistical limitations of the recorded data. Nevertheless, in the future, with a larger amount of data, the combination of the two independent methods would allow to select b-jet pairs originating from the $t\bar{t}$-system decay with a remarkable purity, especially in events that have multiple additional b jets.
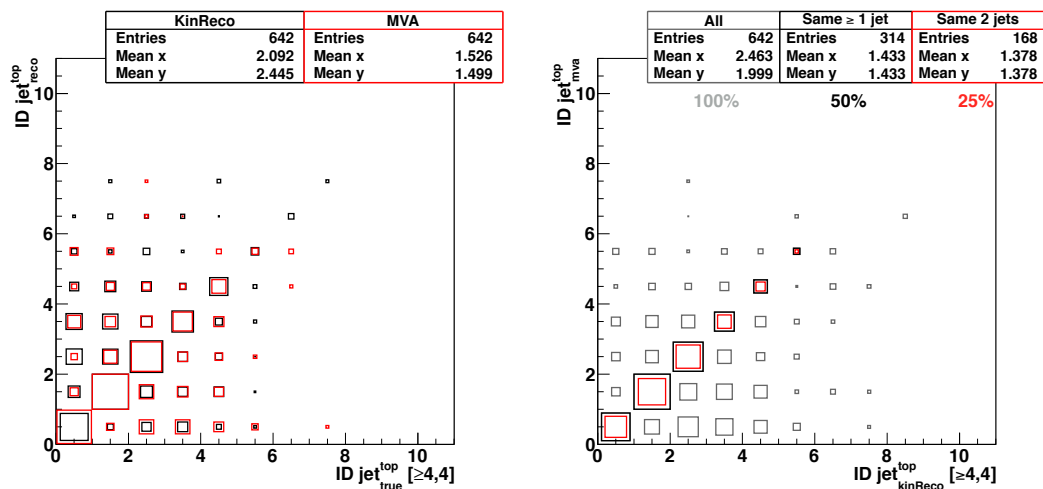
(a) if a jet in a pair identified by the KinReco or MVA is a true b-jet from the $t\bar{t}$ system, it enters the diagonal bins. Otherwise the jet enters an off-diagonal bin. The MVA tends to assign jets with higher p_T to the $t\bar{t}$ system, compared to the KinReco.

(b) if the same jet is idendified by both the KinReco and MVA, it is in the black diagonal bins (50 % of jets). If the other jet in a pair is also identified by the both methods, it enters the red diagonal (25 % of jets).

Figure 9.26: Indices of jets assigned to the $t\bar{t}$ system by the True, KinReco and MVA methods in $t\bar{t}b\bar{b}$ events after the full events selection with 4 b-tagged jets.

9.6.3 Effect on the properties of additional b-jets

While the plots, shown in Figure 9.25, represent the cleanest test of the MVA and KinReco approaches themselves, it is more relevant for the actual measurement how the difference in performance affects the properties of additional b jets.

For this purpose, the additional b jets are defined at generator level as described in Section 9.2.1 and at reconstruction level as described in Section 9.6. For the optimal measurement, the properties of the jets at the reconstruction level should be as close as possible to generator-level values.

The correlation between the two is shown in Figure 9.28 for the combination of $t\bar{t}b$, $t\bar{t}2b$ and $t\bar{t}b\bar{b}$ processes after the basic events selection with requirement of exactly 3 b-tagged jets. Since the binning of the distributions was optimised to minimise migrations of the reconstructed values away from the diagonal bins, the difference between the KinReco and MVA approaches is negligible for $|\eta|$ of the 1-st additional b-tagged jet. For the transverse-momentum correlation, minor differences are visible, but they are not in the same direction in all bins.

Thus, it can be concluded that the reduction in the proper-identification probability of the MVA, shown in Figure 9.28, can be neglected for the differential $t\bar{t}b$ cross-section measurement.

In contrast to the $t\bar{t}b$ case, the improvement of the MVA in $t\bar{t}b\bar{b}$ events does make a noticeable difference. The effect can be seen from the correlation plots in Figure 9.29. For the p_T and $|\eta|$ of the second additional b-tagged jet, the difference between the KinReco and MVA is negligible, as it was for the first additional b-tagged jet in Figure 9.28. But for the distance between the two leading additional b-tagged jets, the difference is clearly visible between Figures 9.29f and 9.29e.

(a) $t\bar{t}b$ and $t\bar{t}2b$ events with exactly 3 b-tagged jets
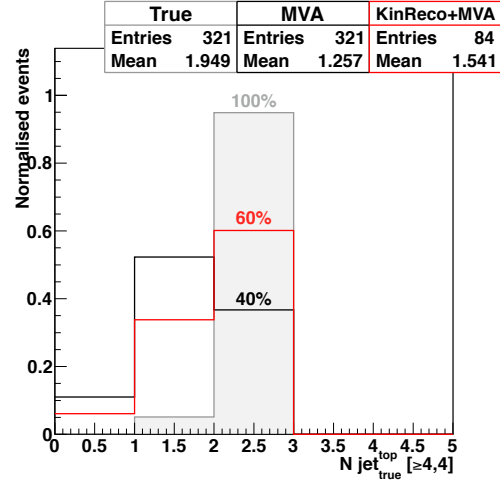
(b) $t\bar{t}b\bar{b}$ events with exactly 4 b-tagged jets
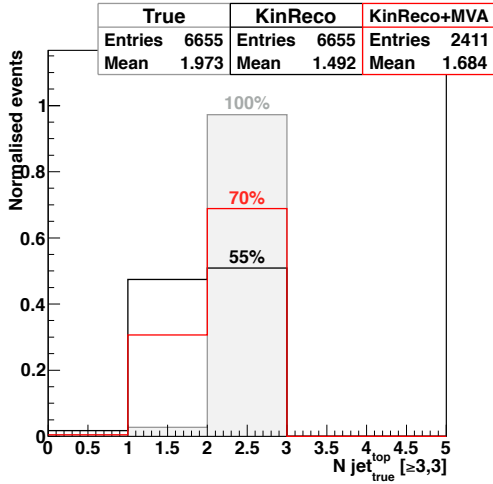
(c) $t\bar{t}Other$ events with exactly 3 b-tagged jets
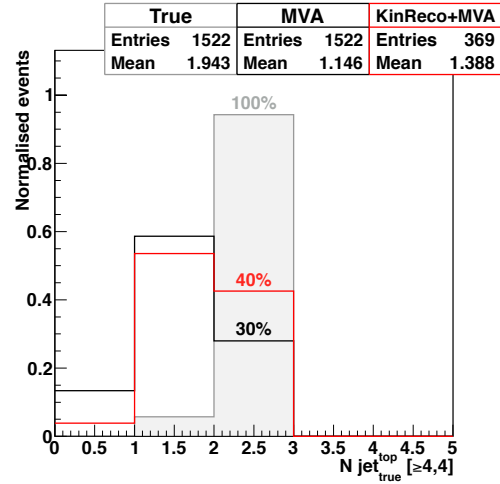
(d) $t\bar{t}H(b\bar{b})$ events with exactly 4 b-tagged jets

Figure 9.27: Multiplicity of properly identified b jets from the $t\bar{t}$ decay in all pairs selected by the MVA and only in pairs that were selected by both the MVA and KinReco methods. The distributions are normalised to 1.0 for comparison of the fraction of properly identified jets, regardless of the event-selection efficiency. The combined method is compared to either the KinReco or to the MVA approach, depending on which one performed best for the corresponding process in Figure 9.25.

(a) KinReco

(b) MVA
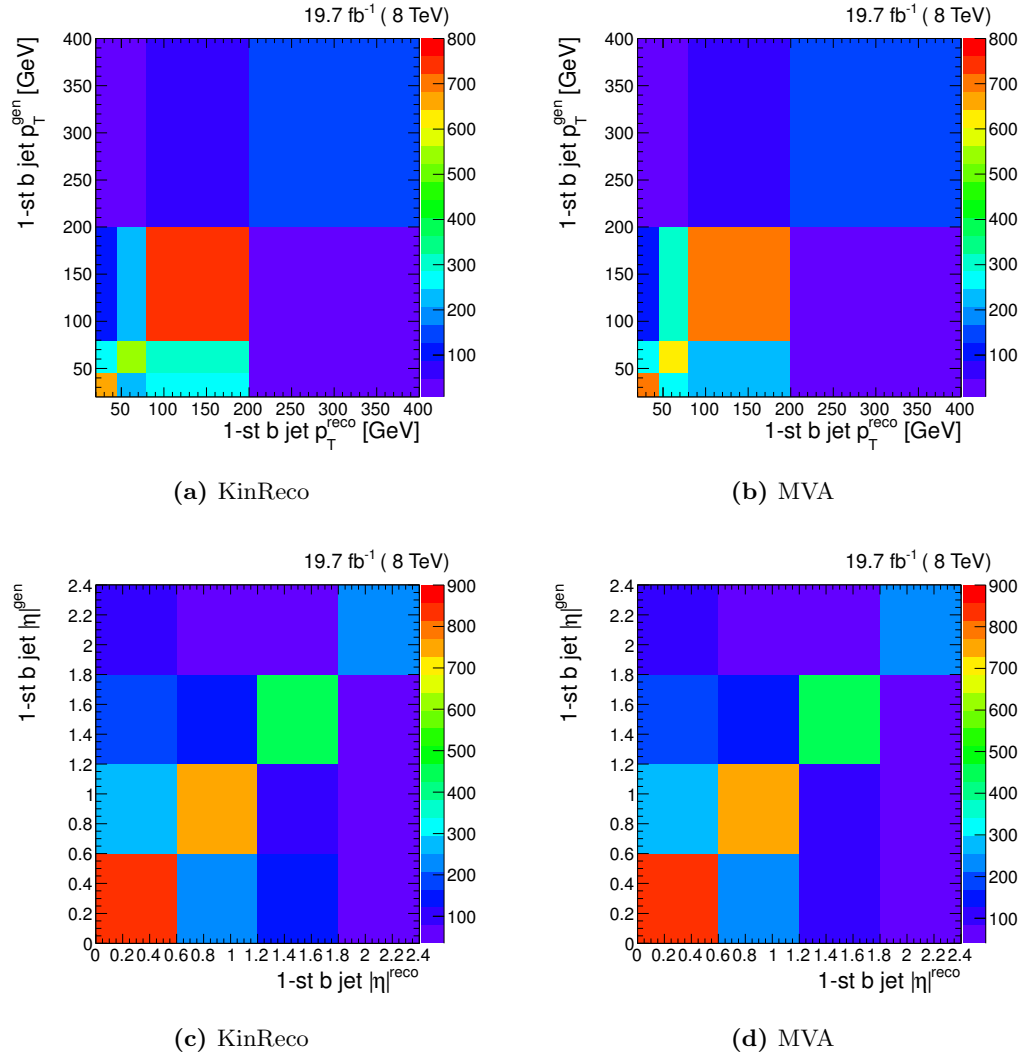
(c) KinReco

(d) MVA

Figure 9.28: Correlation between the generated and reconstructed properties of the first
additional b jet; compared between the KinReco and MVA methods for the
identification of b jets from the $t\bar{t}$ system. The distributions represent the
combination of $t\bar{t}b$, $t\bar{t}2b$ and $t\bar{t}b\bar{b}$ processes after the requirement of exactly
3 b-tagged jets.

In the first two bins of the $\Delta R_{b\bar{b}}$ distribution, the MVA provides about $50\,\%$ more events with the reconstructed value being in the same bin as the generated one. Only in the last bin there is about $20\,\%$ decrease.

Misreconstruction of the $\Delta R_{b\bar{b}}$

The remaining contribution in the non-diagonal bins of the $\Delta R_{b\bar{b}}$ correlation in Figure 9.29f is dominated by events where the real additional b jets are closer to each other than the reconstructed ones. Checking explicitly the events contributing to those bins, two different misreconstruction scenarios were found, which are shown in Figure 9.30.

It was concluded that the effect is mainly caused by the wrong identification of b jets from the $t\bar{t}$ system, as depicted in Figure 9.30b. In less than $10\,\%$ of the events, the effect was caused by an additional mistagged jet in the acceptance, while one of the real additional b jets was not b-tagged. This minor contribution is shown in Figure 9.30c.

Obviously, the dominating effect from the wrong identification of b-jets from the $t\bar{t}$ decay is enhanced when using the kinematic reconstruction, which has worse performance in $t\bar{t}b\bar{b}$ events. This is reflected by the higher contribution to the off-diagonal bins in Figure 9.29e.

Since the MVA provides the highest rate of the proper b-jet association to the $t\bar{t}$ system, the remaining contributions in the off-diagonal bins must be included in the measurement. The misreconstruction is taken into account by the *unfolding* procedure, which is described later in Section 9.7.

9.6.4 Control distributions

The misreconstruction mentioned above changes the reconstructed values with respect to the true ones. This effect should be the same in both MC simulations and in real data, since the response of the detector is precisely simulated and remaining differences are corrected. In order to check this, the reconstruction level distributions of each measured quantity are presented in Figure 9.31.

The distributions show in general good agreement between the simulations and data. Significant statistical fluctuations are present in the data, especially in the distributions that involve second additional b-tagged jet. Thus, the cross sections have to be measured with a coarser binning to reduce statistical fluctuations. Furthermore, the exponentially falling variables, i.e. p_{T}, $m_{b\bar{b}}$, have to be measured with bins of variable width in order to have sufficient statistics in low-populated regions. More information on the exact binning of each variable, taking into consideration not only statistical precision but also bin-to-bin migrations, will be given in the corresponding Section 9.7.1.

9.7 Unfolding procedure

Possible misreconstruction and limited detector resolution can smear or bias the reconstructed value of the quantity x^{reco} with respect to its true value x^{gen}.

(a) KinReco

(b) MVA

(c) KinReco

(d) MVA

(e) KinReco

(f) MVA
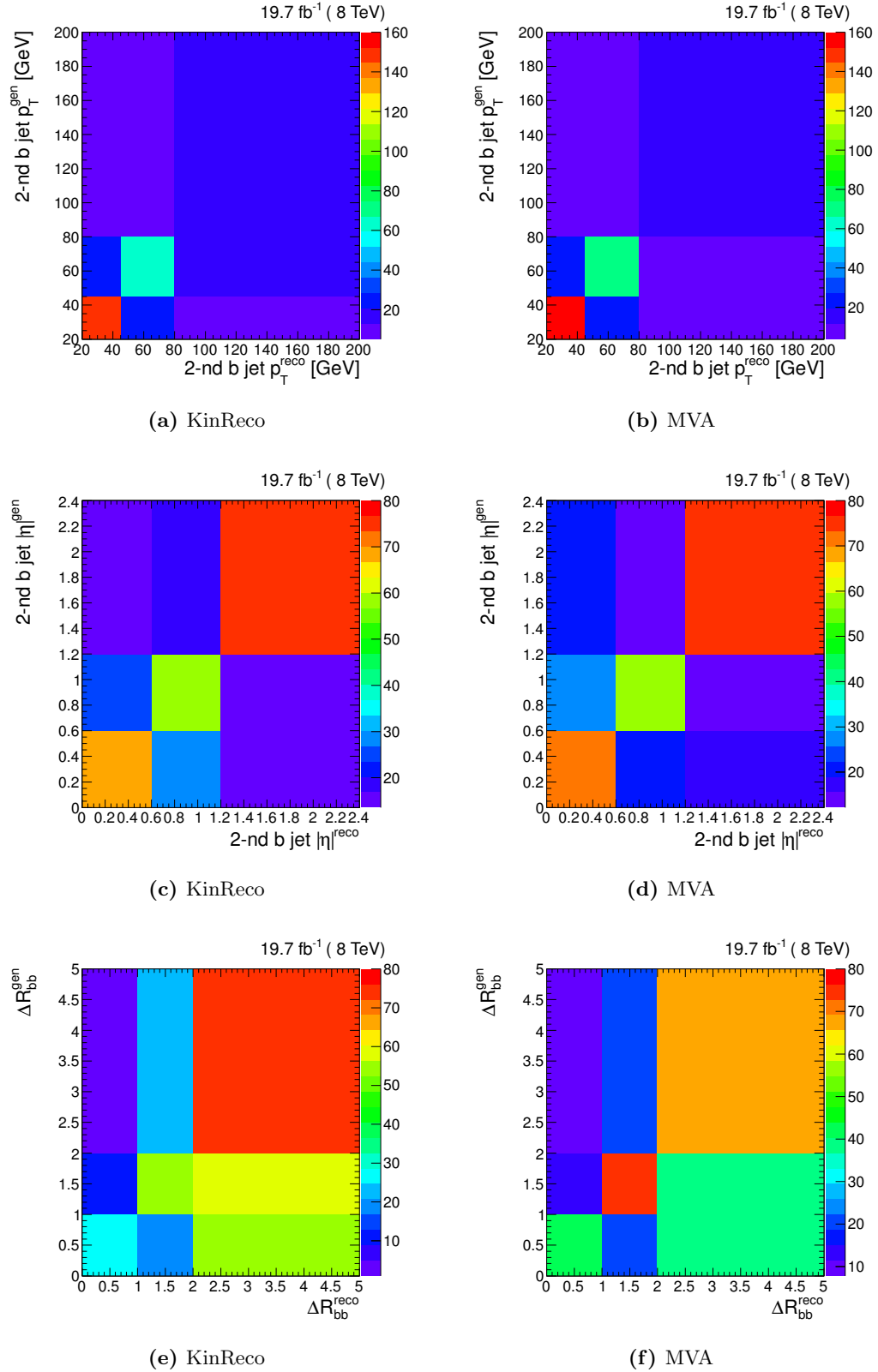
Figure 9.29: Correlation between the generated and reconstructed properties that involve the second additional b jet; compared between the KinReco and MVA methods for the identification of b jets from the $t\bar{t}$ system. The distributions represent the $t\bar{t}b\bar{b}$ process after the requirement of exactly 4 b-tagged jets.

(a) ideal case

(b) wrong identification of jets from the $t\bar{t}$ system ($\geq 90\,\%$)

(c) wrong b-tagging of additional jets ($\leq 10\,\%$)
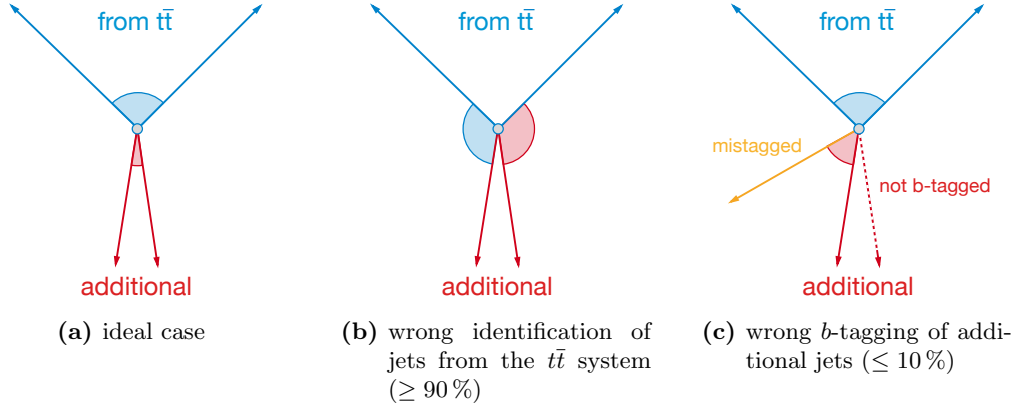
Figure 9.30: Schematic view of possible scenarios of additional-b-jet misreconstruction. These scenarios were found in events with small generated and large reconstructed $\Delta R_{b\bar{b}}$ contributing to the non-diagonal bins in Figure 9.29f.

Limited statistics of the recorded data requires a limited number of bins in the measured distribution $f(x^{\text{reco}})$ in order to have sufficient statistical precision for each bin. Thus, the mentioned effects can lead to the reconstructed value x^{reco} being assigned to a different bin than the true value x^{gen}, causing a change in the shape of the measured distribution $f(x^{\text{reco}})$ with respect to the true distribution $f(x^{\text{gen}})$.

9.7.1 Optimisation of the binning

Such migrations can be visualised as contributions to non-diagonal bins of the plots showing correlations between the true and observed values, as demonstrated for all the measured quantities in Figure 9.29e. In order to reduce the dependence on the mentioned effects, it is important to optimise the number of bins and their width such that the off-diagonal contributions are kept to a minimum.

Such optimisations of the binning are based on the two parameters for each individual bin i:

- **purity (p_i):** fraction of signal events reconstructed in bin i (N_i^{reco}) that are also generated in the same bin;

- **stability (s_i):** fraction of signal events generated in bin i (N_i^{gen}) that are also reconstructed in the same bin.

Thus, the purity for bin i can be expressed as a ratio:

$$p_i = \frac{N_i^{\text{reco \& gen}}}{N_i^{\text{gen}}} , \qquad (9.10)$$

and characterizes the probability of the true value in the bin i to migrate outside the bin. The stability, on the other hand, reflects the probability of a different true value to migrate into the bin i, and can be expressed by the following ratio:

$$s_i = \frac{N_i^{\text{reco \& gen}}}{N_i^{\text{reco}}} . \qquad (9.11)$$

(a) transverse momentum of the first additional b-tagged jet

(b) absolute pseudorapidity of the first additional b-tagged jet

(c) transverse momentum of the second additional b-tagged jet

(d) absolute pseudorapidity of the second additional b-tagged jet

(e) distance in $\eta - \phi$ space between the first and second additional b-tagged jets

(f) invariant mass of the pair of the first and second additional b-tagged jets

Figure 9.31: Comparison of distributions of the measured quantities of the first and second additional b-tagged jets between data and the MC simulations. As requested by the CMS Higgs group, the $100 < m_{b\bar{b}} < 140$ GeV region is blinded in the data due to the enhanced sensitivity to the not-yet-discovered $t\bar{t}H(b\bar{b})$ contribution.
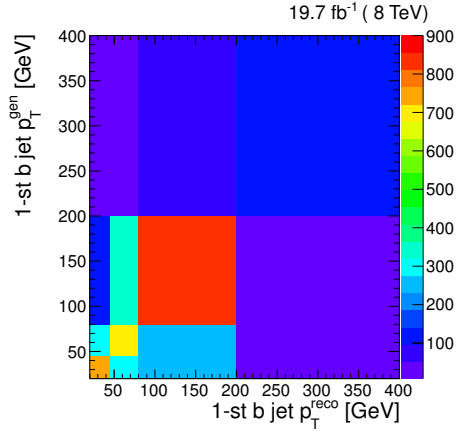
(a) transverse momentum of the first additional b-tagged jet
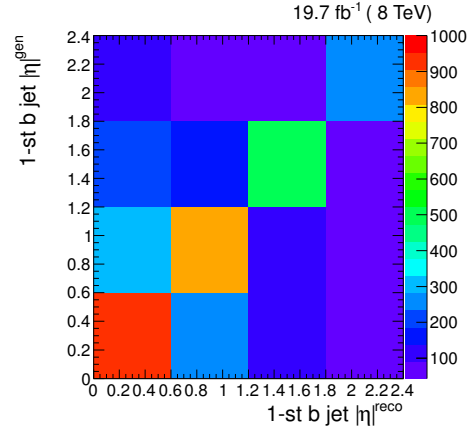
(b) absolute pseudorapidity of the first additional b-tagged jet
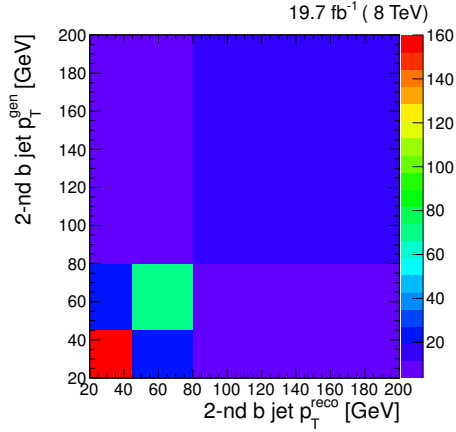
(c) transverse momentum of the second additional b-tagged jet
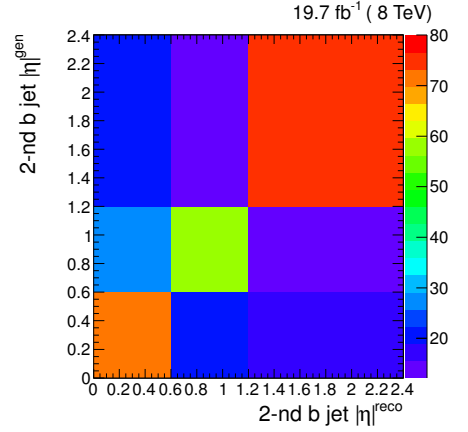
(d) absolute pseudorapidity of the second additional b-tagged jet
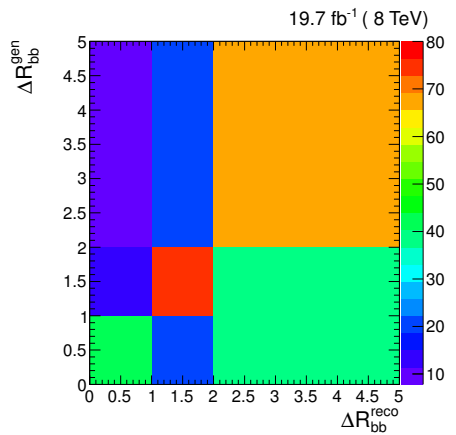
(e) distance in $\eta - \phi$ space between the first and second additional b-tagged jets
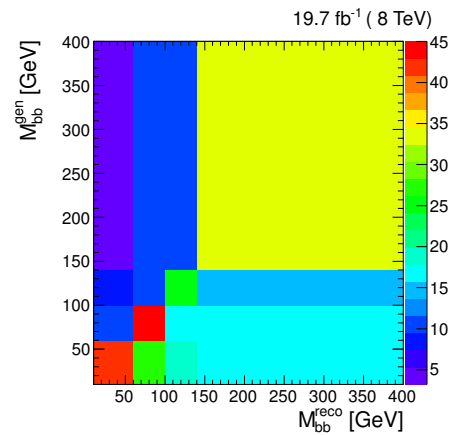
(f) invariant mass of the pair of the first and second additional b-tagged jets

Figure 9.32: Correlations between the measured and generated $t\bar{t}b$ and $t\bar{t}b\bar{b}$ quantities, as determined from the corresponding simulated signal events. The used $t\bar{t}b$ and $t\bar{t}b\bar{b}$ signal definitions correspond to the full phase space of the $t\bar{t}$ system. The correlations for the visible phase space can be found in Appendix B.2.

The shapes of distribution for each measured quantity differ, as well as statistics available for each event selection. Therefore, optimisation of the binning was performed for every single quantity. The corresponding plots showing purity and stability in every bin are presented for each measured variable in Figure 9.33. The main criteria for the applied binning were the following:

- purity and stability not lower than 40 % in every measured bin;

- at least three bins for each distributions, in order to have a shape estimate.

As can be seen from the plots, purity and stability are higher than 40 % in every bin of each distribution, except of the $m_{b\bar{b}}$. The bin $100 < m_{b\bar{b}} < 140$ GeV has purity of about 35 %, which is still acceptable, because this bin is blinded and is not measured. In some bins purity is higher than the probability of proper identification of b jets from the $t\bar{t}$ decay, which is explained by different types of jets that are considered. Thus, a non-b-tagged jet can be assigned to the $t\bar{t}$ system, while the proper additional b-jets still can be chosen, if they are b-tagged and have higher p_T than the missed b-jet from the $t\bar{t}$ decay. Furthermore, a wrongly identified additional b jet can have properties that are close enough to the ones of the correct jet, such that it is reconstructed in the proper bin of the measured distribution.

The pseudorapidity distributions have rather low level of migrations, corresponding to purity and stability on average above 60 %. The transverse-momentum distributions are well correlated with migration-probabilities above 50 % in every bin.

The most significant migrations are present in the $\Delta R_{b\bar{b}}$ and $m_{b\bar{b}}$ distributions, which are caused by the misidentification of additional b-jets at reconstruction level, as described earlier in Section 9.6.3. In the outer bins of the distributions either purity or stability reach 40 %. Thus, the high purity in one outer bin is compensated by low stability in the same bin. Due to the nature of the misidentifications, the opposite behaviour is present in the other outer bin.

9.7.2   Regularised unfolding

The aim of the differential cross-section measurement is to obtain the true distribution, which requires a correction of the migration effects present in the observed spectrum. The number of events in data after subtraction of background (N_i^{data}) in the bin i can be related to the number of true signal events (N^{true}) in all bins through a response matrix $\mathcal{A}$:

$$N_i^{\text{data}} = \sum_j \mathcal{A}_{ij} N_j^{\text{true}} . \tag{9.12}$$

The response matrix is obtained from the signal MC simulation after all the relevant corrections described in the previous sections. Every element $\mathcal{A}_{ij}$ of the matrix $\mathcal{A}$ is defined as the fraction of events generated in bin j and reconstructed in bin i among all events generated in bin j:

$$\mathcal{A}_{ij} = \frac{N_{\text{gen},j}^{\text{reco},i}}{N_{\text{gen},j}} . \tag{9.13}$$
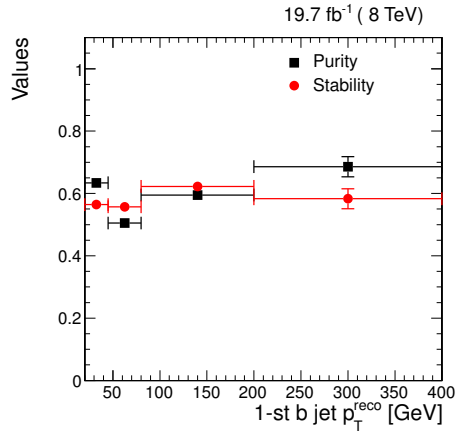
(a) transverse momentum of the first additional b-tagged jet
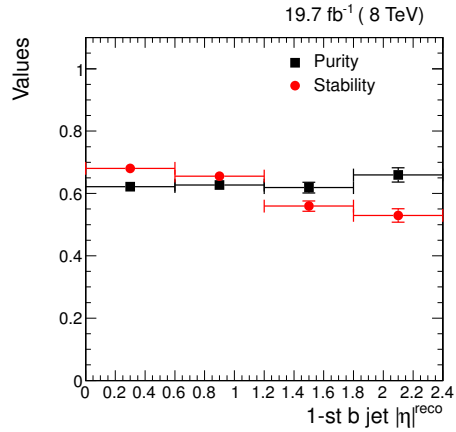
(b) absolute pseudorapidity of the first additional b-tagged jet
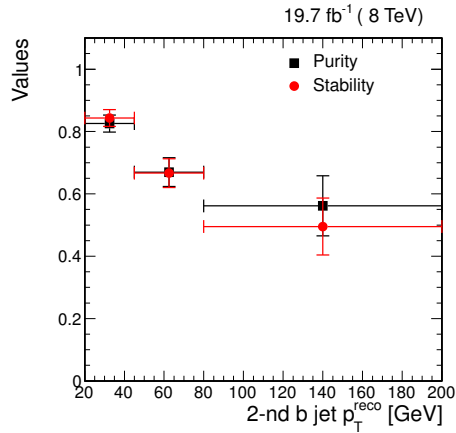
(c) transverse momentum of the second additional b-tagged jet
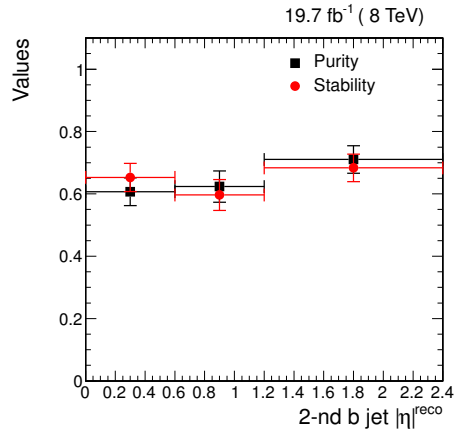
(d) absolute pseudorapidity of the second additional b-tagged jet
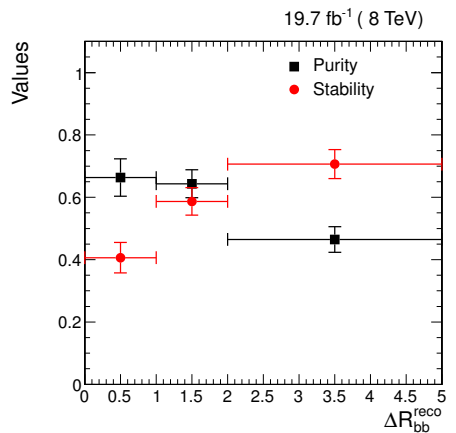
(e) distance in $\eta - \phi$ space between the first and second additional b-tagged jets
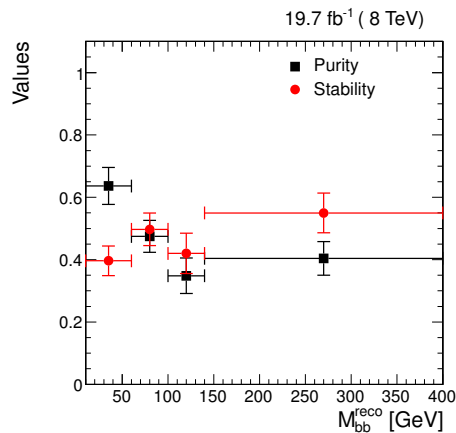
(f) invariant mass of the pair of the first and second additional b-tagged jets

Figure 9.33: Purity and stability of the measured $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ quantities as determined from the corresponding simulated signal events. The used $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ signal definitions correspond to the full phase space of the $t\bar{t}$ system. The vertical error bars represent the statistical uncertainty. The purity and stability in the visible phase space can be found in Appendix B.2.

The efficiency of the events with true values in bin j to be reconstructed can be represented by the sum over all matrix elements corresponding to this bin:

$$\epsilon_j = \sum_i \mathcal{A}_{ij} \,. \tag{9.14}$$

Thus, the true distribution can be obtained from the observed one by solving the Equation 9.12, which is referred to as an *unfolding*. In principle this can be done by simple inversion of the response matrix:

$$N_i^{\text{true}} = \sum_j \left(\mathcal{A}_{ij} \right)^{-1} N_j^{\text{data}} \,. \tag{9.15}$$

Nevertheless, such a simple approach leads to unstable results, since the direct inversion of the matrix $\mathcal{A}$ leads to the amplification of statistical fluctuations in the observed distribution [234].

This problem is addressed by imposing a measure of smoothness on the unfolded result in the process called *regularisation* [235, 236]. Its purpose is to suppress statistically insignificant contributions, when solving the Equation 9.12. The regularisation method used in this thesis is the Singular Value Decomposition (SVD) approach, as described in [237], and is technically implemented in the `TSVDUnfold` [238] package.

Choice of the regularisation strength

An important parameter of the unfolding procedure is the regularisation strength τ, which determines the amount by which the unfolded result will be smoothed. If the τ parameter is too small, the effect of the unfolding will be similar to the result of the direct-matrix-inversion method. If it is too large, the unfolded distribution can be overcorrected and artificially biased towards the simulated prediction used for the regularisation.

An optimal value of the regularisation-strength parameter is obtained individually for every measured quantity based on minimisation of global-correlation between different bins, as described in [239]. In order to determine the optimal value of τ, the root-mean-square (RMS) global correlation coefficient $\bar{\rho}$ is calculated for a series of τ-values in the range 0.01–100. The value leading to the lowest $\bar{\rho}$ is used for the final unfolded result.

Examples of the dependence of the RMS global correlation on the scanned regularisation strength is shown for the p_{T} of the first additional b-jet and for the $\Delta R_{b\bar{b}}$ in the full $t\bar{t}$ phase space in Figure 9.34. Distributions for other variables, as well as for the visible phase space, are presented in Appendix B.3.

9.7.3   Test with pseudodata

In order to check the behaviour of the unfolding, it is performed with the real collision data replaced by the reconstructed MC simulations, which is referred to as *pseudodata*. The pseudodata is normalised to the integrated luminosity of the real data and is corrected in the same way as normal MC samples, such that it exactly coincides with the sum of all signal and background MC simulations.
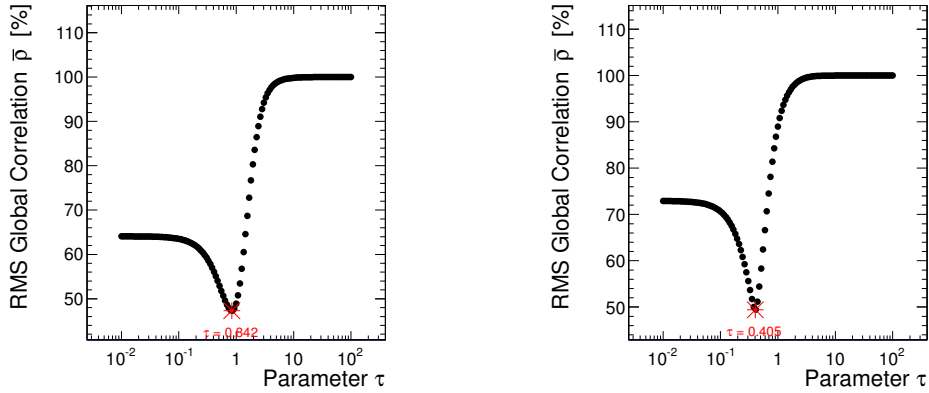
(a) transverse momentum, p_T, of the first additional b-jet

(b) distance in $\eta - \phi$ space, $\Delta R_{b\bar{b}}$, between the first and second additional b-jets

Figure 9.34: An example of the dependence of the RMS global correlation $\bar{\rho}$ on the regularisation-strength parameter τ for several measured quantities. The value of τ corresponding to the smallest global correlation is marked in red and is used for the final unfolded result. The scans were performed with the response matrices obtained for the full $t\bar{t}$ phase space.

The analysis is performed with exactly the same procedure as with the real data, and unfolding is done with the same response matrices and regularisation parameters as in the real measurement. For the illustration, checks are performed for only the two variables: p_T of the first additional b jet and for the $\Delta R_{b\bar{b}}$ between the first and second additional b jets. Results for the other variables lead to the same conclusions, and can be found in Appendix B.3.

The first test is performed with the pseudodata matching exactly the sum of MC simulations, as shown on the control distributions in Figure 9.35. In this case, the unfolded distribution matches exactly the true distribution from the MC simulation that was used for the unfolding. This means, that the unfolding procedure does not introduce any bias distorting the measured shape.

Pseudodata with distorted signal shapes

The next step is to check whether the applied unfolding procedure leads to a bias of the measured spectrum towards the true distribution, which was used for the unfolding. In particular, it has to be confirmed that the regularisation strength parameter used for the SVD unfolding is optimal. In order to check this, the MC samples corresponding to the signal process, are reweighted such that the shape of the corresponding variable becomes distorted. The reweighted samples are used only in pseudodata, while the original ones are used in the MC part.

Thus, when calculating the cross section, the distorted observed shape in pseudodata is unfolded using the original response matrix of the MC simulation. In the ideal case, the unfolded cross section should be identical to the distorted true distribution and not to the original true distribution, which was used to obtain the response matrix.
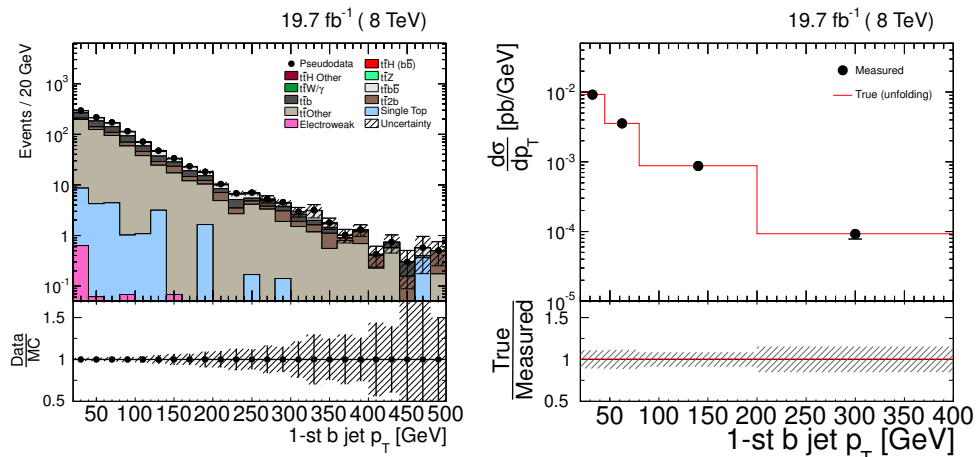
(a) p_T of the first additional b-jet
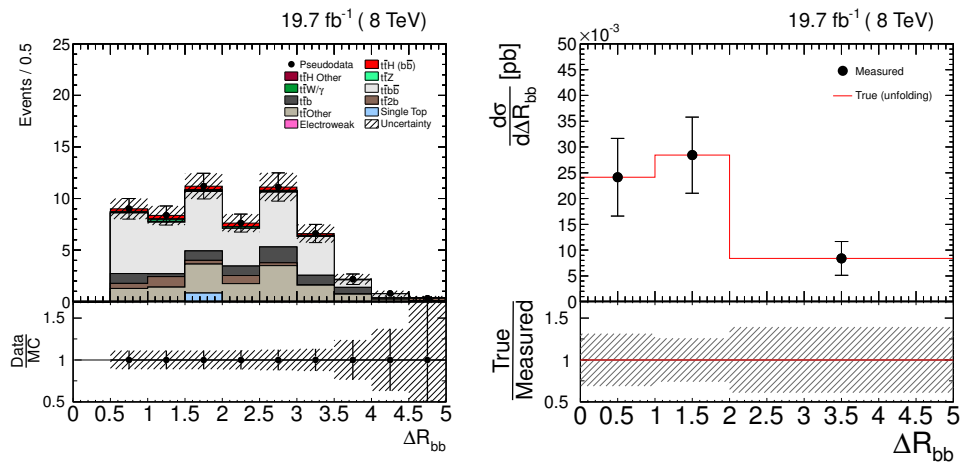
(b) p_T of the second additional b-jet

Figure 9.35: Control distribution of a measured quantity (left) and unfolded cross sec-
tion as a function of the quantity (right) with the real data replaced by
pseudodata. Pseudodata is constructed by a raw sum of all MC simulations
with the same event selection and corrections as in MC. The *Measured*
markers represent the unfolded cross section obtained from pseudodata.
The *True (unfolding)* line represents the true distribution, which was used
for the unfolding.

The tested shape distortions are simple slopes of the corresponding shapes. The effect of the distortion of the first additional b-jet p_{T} spectrum is shown for the control distribution and for the unfolded cross section in Figure 9.36. For the more reliable check, two reweightings are performed with the same slope in the opposite directions.

(a) $w = 1 + 0.3\dfrac{p_{\mathrm{T}} - 100}{100}$

(b) $w = 1 - 0.3\dfrac{p_{\mathrm{T}} - 100}{100}$
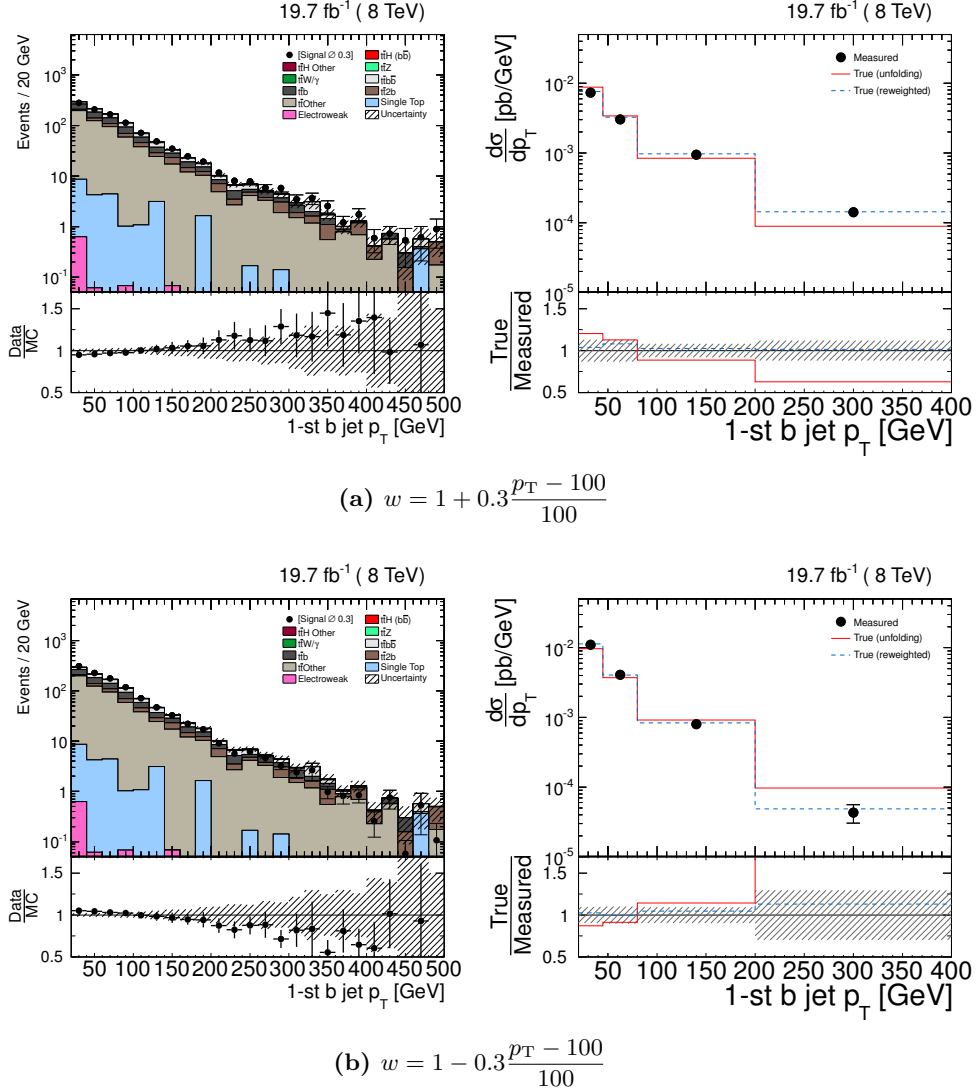
Figure 9.36: Control distribution (left) and unfolded cross section (right) as a function of the p_{T} of the first additional b jet in the full $t\bar{t}$ phase space. Signal samples are reweighted with an event-weight w, in order to introduce a slope in the p_{T} spectrum in pseudodata. The weight w is a function of the p_{T} of the first additional b jet, and is applied to every event that has at least one additional b jet, as defined in Section 9.2.1. The reweighted signal samples correspond to the $t\bar{t}b$, $t\bar{t}2b$ and $t\bar{t}b\bar{b}$ processes. The *Measured* markers represent the unfolded cross section obtained from reweighted pseudodata. The *True (unfolding)* solid line represents the undistorted true distribution, which was used for the unfolding. The *True (reweighted)* dashed line represents the reweighted true distribution, which should match the *Measured* distribution if the unfolding does not introduce a bias towards the *True (unfolding)* spectrum.

The same plots for the $\Delta R_{b\bar{b}}$ between the first and second additional b jets are shown in Figure 9.37.

(a) $w = 1 + 0.3\dfrac{\Delta R_{b\bar{b}} - 1.5}{1.5}$

(b) $w = 1 - 0.3\dfrac{\Delta R_{b\bar{b}} - 1.5}{1.5}$
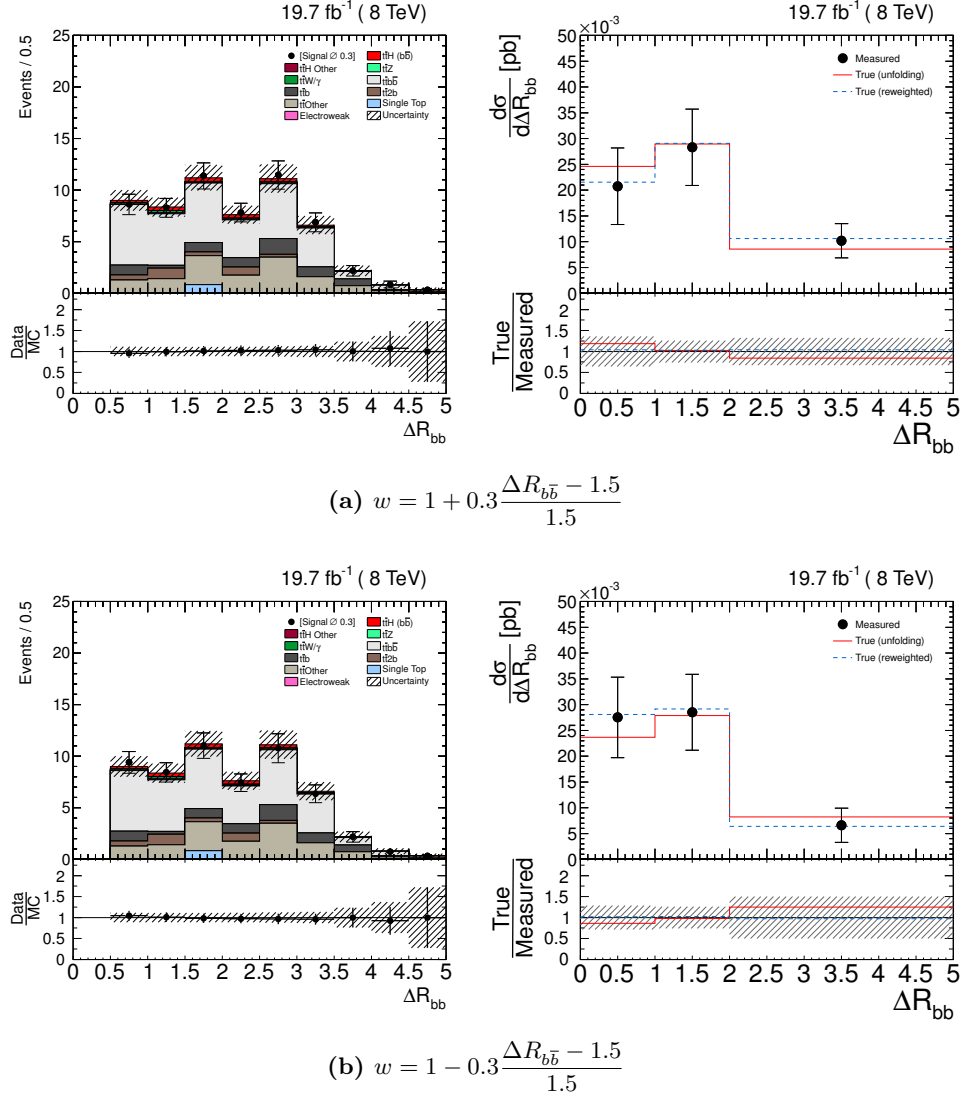
Figure 9.37: Control distribution (left) and unfolded cross section (right) as a function of the $\Delta R_{b\bar{b}}$ between the first and second additional b jets in the full $t\bar{t}$ phase space. The weight w is a function of the $\Delta R_{b\bar{b}}$, and is applied to every event that has at least two additional b jets, as defined in Section 9.2.1. The only reweighted signal sample corresponds to the $t\bar{t}b\bar{b}$ process. Meaning of the legend is the same as in Figure 9.36.

From the illustrated plots, it can be concluded that the unfolded cross section clearly reflects the spectrum present in data and not the true distribution used for the unfolding. In general the bias towards the true distribution is negligible. A slight bias of the unfolded distribution towards the one used for the unfolding can be seen for the p_{T} of the first additional b jet in Figure 9.36b. Nevertheless, the strength of the bias for the applied distortion is small enough to be covered by the statistical uncertainty. In real data the difference in shape with respect to the true distribution is expected to be even smaller, since the description of the real data is better than in the case of reweighted pseudodata.

The results of the same tests for other measured variables can be found in Appendix B.3 and support the mentioned conclusions. Therefore, the determined regularisation strength parameters are close to optimal and the adopted unfolding procedure can be safely used for the measurement with the real data.

9.8 Systematic uncertainties

From the control distributions, shown in Section 9.5.5, it can be seen that the statistical uncertainties in the data are quite large, especially for the variables involving the second additional b jet.

Nevertheless, systematic effects biasing the MC simulations used for the measurement need to be considered. Such effects can occur either due to the limited understanding of the detector performance and its simulation, or due to wrong assumptions in the theoretical predictions of the MC samples used in the measurement. Since the background shapes, subtracted from the distributions in data, are obtained from the MC predictions, corresponding systematic effects can change the measured cross sections. In order to account for possible changes of the result, an additional uncertainty originating from each effect is assigned to the measured cross section, which should account for a possible bias of both the normalisation and shape of the result.

Thus, systematic uncertainties can be split into the two categories: experimental and modelling uncertainties. The experimental uncertainties, considered in the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ measurements, are discussed in Section 9.8.1, while the influence of modelling variations on the final results is covered in Section 9.8.2.

Procedure of systematic-uncertainty determination

For a set of reconstructed quantities a correction factor is applied to the MC simulations in order to properly describe real measured quantities in data. The correction factors have certain uncertainties, most of which are determined by the dedicated groups, and are provided centrally for all analyses that apply appropriate object selection criteria and corrections.

In order to calculate the effect of each systematic effect on the final result, every correction factor for MC simulations is varied up or down by its corresponding uncertainty, and the complete analysis is repeated, leading to a varied cross-section result. The results obtained with the default correction factors, without any variation, are referred to as *Nominal*, while the ones obtained with a corresponding variation of a certain correction factor are referred to as *varied* results. The difference between the Nominal and a single varied result in each bin of the measured cross section is treated as the individual systematic uncertainty caused by the corresponding variation.

Uncertainties on the MC-to-data correction factors can be non-symmetric, as well as their effect on the measured cross section. Therefore, variations up and down are analysed separately. The effect of each individual up or down variation in every bin is added to effects of other variations in quadrature to obtain the total systematic uncertainty. In case both variations change the cross section in the same direction, the larger effect is taken and symmetrized. The total systematic uncertainty is added in quadrature to the statistical uncertainty to obtain the total uncertainty on the measured cross section in each bin of a distribution.

9.8.1   Experimental uncertainties

Differences in object-selection efficiency in data and MC simulations are corrected
by data-to-simulation scale factors (SF) and are determined with finite precision.
Their variation by the corresponding uncertainty leads to a change of the both
normalisation and shape of the measured cross sections.

Luminosity

The luminosity is measured centrally by CMS using the forward hadronic calorimeter
and the silicon pixel detector. The actual number is obtained by a cluster counting
method, and has a total uncertainty of 2.6%: 2.5% (syst.) $\oplus$ 0.5% (stat.) [240].
Since all the MC samples are normalised to the measured luminosity in the data,
its variation by the corresponding uncertainty will lead to a change of the normal-
isation of all the MC samples with respect to the real data. Nevertheless, its effect
on the measured cross section is negligible, since the normalisations of the main
backgrounds are constrained by the template fit of the $t\bar{t} + HF$ contribution. Its
typical effect on the measured cross sections is below 1%.

Pile-up

According to the pile-up (PU) model, the estimated average number of secondary
interactions per event in the data used for the measurement is about 20. The
estimation of the number of PU events is based on the total inelastic pp cross section.
The cross section is centrally determined by CMS, and equals to $64.9\,\mathrm{pb} \pm 5\%$.

The varied cross section is used to obtain different estimates of the PU dis-
tribution. The variation conservatively accounts for the possible variations of the
luminosity, pp cross section, as well as the modelling of pile-up used to reweight the
MC samples. The average effect of these variations on the measured cross sections
is below 3%.

Trigger and lepton-selection efficiency

The data-to-simulation scale factors for the lepton and trigger selection are deter-
mined as described in [241]. The trigger scale factors are determined as functions of
the lepton pseudorapidity and have typical uncertainty of 1%.

The lepton identification and isolation scale factors are determined separately
for electrons and muons as functions of their p_T and η. The average precision of
the scale factors is of the order of 0.3%, with additional 1% uncertainty to cover
possible differences between the $t\bar{t}$ event topology and the Drell-Yan events, which
were used for the scale-factor estimation. The effect of these variations on the $t\bar{t}b\bar{b}$
cross sections is below 3%, while the effect on $t\bar{t}b(\bar{b})$ cross sections is completely
negligible.

Jet energy scale

The calibration of the energy of jets is provided centrally by CMS as a function of η and p_T of the jets [242]. The uncertainty on the jet-energy scale (JES) is largest for the jets with low transverse momentum. In the kinematic region considered for the analysis, the JES uncertainty is not greater than 5 %.

Since the JES variation changes the p_T distribution of the jet, it influences the number of b-tagged jets in the acceptance, which is used in the template fit of the $t\bar{t} + HF$ fraction. Therefore, this uncertainty is included in the fit and only its effect on the shapes of the measured cross sections is evaluated separately, while the influence on the normalisation is included in the uncertainty of the fit. This is one of the dominant systematic uncertainties. Its effect on the measured cross sections is 1 % to 15 %, and an average uncertainty due to the JES variations is about 10 %.

Jet energy resolution

The jet-energy resolution in the simulation is corrected to describe data by the data-to-simulation scale factors, which depend on the pseudorapidity of a jet. The scale factors are measured with a precision of 3 % to 6 %, as described in [243]. Like the JES, this uncertainty is included in the template fit and only its effect on the shapes of the cross sections is evaluated specifically for the JER variation. The corresponding uncertainty is 2 % to 10 %.

Missing transverse energy

Since the calculation of the missing transverse energy is based on the transverse momenta of the leptons and jets in the event, any variation of their properties leads to a corresponding change in the estimated $\not{E}_T$. The $\not{E}_T$ is reevaluated after any change of kinematic properties of the particles in each event. Thus, no specific uncertainty is assigned to the $\not{E}_T$ variation, in accordance with the recommendations of the CMS JetMET group [244].

Reshaping of the b-tagging discriminant

The selection of the b-jets at the reconstruction level in data and in MC simulations is matched with the CSV reshaping technique [228]. The scale factors are parameterized by the p_T and η of a jet and are separate for b jets and light-flavour jets. A set of uncertainties is associated with the determined scale factors, categorised by their nature: statistical or systematic.

Scale factors for b jets are determined from the control sample of $t\bar{t}$ events. The main source of systematic uncertainty for the b-jet scale factors is the rate of light-flavour jets in the selected sample, which are reconstructed instead of a real b jet from the $t\bar{t}$ decay. The systematic effect on the scale factors for light-flavour jets originates from the contamination of $Zb\bar{b}$ events in the Z-boson-enriched control sample. Such a contamination leads to a non-negligible contribution of b jets in the assumed sample of light-flavour jets, which would lead to a bias of the light-flavour scale factors. The c-jet scale factors are constant and equal to 1.0, and are assigned

by a relative uncertainty that is twice as large as the total uncertainty on the b-jet
scale factors, which includes both the statistical and systematic uncertainties [228].

Since the variation of the jet-energy scale can change the p_T of a jet, on which the
b-tagging scale factor depends, the two uncertainties are correlated. For this reason,
a set of different b-tagging scale factors is provided by the developers of the reshaping
method, that were determined with with consistently varied jet energies. Thus,
when calculating a cross section with a certain JES variation, the corresponding set
of b-tagging scale factors is used, as recommended by the developers of the CSV-
reshaping technique [245].

The effect of the b-tagging-related uncertainties on the normalisation of the sam-
ples is included in the uncertainty on the results of the template fit. Therefore, only
the effect on the shapes of the measured cross section is evaluated.

Variations of the b-tagging scale factors is another dominant source of systematic uncertainty. Uncertainties on the scale factors for b and light-flavour jets affect
the shapes of $t\bar{t}b(\bar{b})$ cross sections by up to $13\,\%$ and $10\,\%$ respectively. The corre-
sponding uncertainties on the $t\bar{t}b\bar{b}$ cross sections is up to $20\,\%$ and $14\,\%$ respectively.
Uncertainty due to variation of the scale factors for c-flavour jets is $1\,\%$ to $5\,\%$.

Background modelling

A cross section of each simulated process is used to determine its normalisation with
respect to the recorded data. The main background for both the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$
measurements is the $t\bar{t}Other$ component, which includes $t\bar{t} + light\,jets$, $t\bar{t}c\bar{c}$, and
$t\bar{t}$ events with full-hadronic or semi-leptonic final states. From the event-yields in
Table 9.3 it can be seen that the non-dileptonic contributions are negligible and their
variation will not have noticeable effect on the measured cross sections. The cross
sections of the $t\bar{t}c\bar{c}$ process is conservatively varied by $50\,\%$.

The normalisation of the $t\bar{t} + light\,jets$ process is constrained by the template
fit, as well as the normalisations of the $t\bar{t}b$ and $t\bar{t}b\bar{b}$ processes. Thus, the uncertainty
on their contributions is addressed by the variation of the correction factors deter-
mined by the template fit by their corresponding uncertainties, which are stated in
Section 9.5.4.

The normalisation of the $t\bar{t}2b$ process is not determined by the fit. Thus, its
uncertainty is taken from the results of the differential $Zb\bar{b}$ cross-section measure-
ment [232], as described in Section 9.5.2, which equals to $\sigma(t\bar{t}2b)^{+40\%}_{-43\%}$. The variation
of the $t\bar{t}2b$ cross section is relevant only for the $t\bar{t}b\bar{b}$ measurement, since for the $t\bar{t}b$
cross section it is a part of the signal and is not subtracted from data.

Considering the large statistical uncertainties of the measured distribution and
small contributions of other background processes, illustrated in Table 9.3, uncer-
tainties related to other processes, e.g. $t\bar{t}\gamma$, single top quark, Drell-Yan production,
are assumed to be negligible and are not explicitly evaluated.

Not varied parts of the analysis

Certain parts of the analysis are set up once, and are not reevaluated for each
variation of MC samples.

MVA for the identification of b jets from the $t\bar{t}$-system decay. It is trained only once on the set of different $t\bar{t}H(b\bar{b})$ samples using the default MC-to-data correction values. In a systematically-varied version of the analysis, the choice of the reconstruction-level jets that should be assigned to the $t\bar{t}$ system is based on the output of the same MVA. On the other hand, the input to the MVA, used for the evaluation of the output for each jet pair, does change in accordance with the corresponding systematic variation. All the variables used for the training of the MVA, and then for the evaluation of its output, are well described by the MC simulations [227]. Since no event selection is performed based on the MVA output, no MC-to-data efficiency correction is needed. Therefore, there are no specific variations concerning the MVA that would affect the final result.

Normalisation of the $t\bar{t}b$ and $t\bar{t}b\bar{b}$ processes. The template fit, described in Section 9.5.1, is performed only once with the MC samples processed with the default correction factors. The fit results are obtained as histograms respresenting the corrected distributions of the b-tagged-jet multiplicity. The correction factor to the normalisation of the $t\bar{t}b$ or $t\bar{t}b\bar{b}$ process is obtained as a ratio of the integral of the fitted histogram to the integral of the original histogram in MC. Thus, the fitted histogram does not change, while the original histogram in the MC changes according to the applied systematic variation. This leads to a change of the correction factors applied to the $t\bar{t}Other$, $t\bar{t}b$ and $t\bar{t}b\bar{b}$ processes, since their normalisation after the correction is fixed. The normalisation of the different processes is effectively constrained by the result of the template fit, significantly reducing the effect of systematic variations on the measured cross sections.

Since all systematic uncertainties that noticeably change the shapes of the b-tagged-jet multiplicity are included in the template fit, their effect on the absolute values of the measured cross sections is covered by the uncertainty on the correction from the fit. Therefore, for the variations that are included in the template fit, only their effect on the shapes of the measured cross sections is considered. Other variations that do not change the shape of the b-tagged-jet multiplicity are corrected by a factor that is very close to the one determined for the Nominal case.

Regularisation-strength parameter of the unfolding. The strength of regularisation is determined by the τ-parameter scan, as described in Section 9.7.2, using the Nominal configuration of the MC samples. Each systematic variation changes the actual response matrix used for the unfolding, but the value of the τ parameter is fixed for each measured quantity and is not reevaluated for each systematic variation. The values of τ are below 1.0 for all the quantitites, and it was shown in the differential $t\bar{t}$ cross-section measurement that only increase of τ by a factor larger than 50 leads to a visible bias of the unfolded cross sections [179]. The optimal τ values are not expected to change so drastically due to the studied systematic variations.

9.8.2   Modelling uncertainties

After the full event selection for the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ cross-section measurements, the dominant backgrounds are different subsets of the $t\bar{t}+jets$ process, i.e. $t\bar{t}+light\,jets$, $t\bar{t}c\bar{c}$ and for the $t\bar{t}b\bar{b}$ cross section also $t\bar{t}b$ and $t\bar{t}2b$ processes. Besides the possible misreconstruction of leptons, jets or $\not{E}_{\mathrm{T}}$ at the detector level, it is also possible that

the theoretical assumptions made for the $t\bar{t}$ event simulation were wrong. Such assumptions include the parton distribution function of the proton, hard scattering factorisation and renormalisation scales, the matching threshold between the parton shower and matrix element calculation and the top-quark mass. Furthermore, different hard scattering models as well as different parton showering and hadronisation models can lead to different kinematic properties of the final state particles, changing both the normalisation and shapes of the simulated cross sections of background processes.

In order to evaluate the effect of a certain variation on the final result, the size of the varied MC sample has to be comparable to or larger than the size of the nominal sample. Otherwise the difference between the two results will not only represent the systematic effect, but also the statistical fluctuations in the varied sample.

The experimental variations are obtained from the nominal sample by changing the corresponding parameters of the object reconstruction. In contrast, most of the modelling variations require the complete MC sample to be regenerated. Unfortunately, the varied samples have too small amount of generated signal events, making it impossible to properly evaluate the effect of different modelling parameters on the final cross sections using the existing samples. Any difference or no difference between the nominal and varied simulated cross sections in this case could be just a result of a statistical fluctuation.

Rough estimate of modelling uncertainties

Considering large statistical uncertainties of the measured cross sections, it is sufficient to make only rough estimates of the modelling uncertainties. Thus, systematic uncertainties due to modelling variations were estimated from differences between the existing nominal and varied samples at the generator-level, before any event selection at the reconstructed level.

The effect of modelling variations on the measured inclusive cross section is assumed to be negligible, since the template fit constrains the normalisation of the background processes, and its possible variations should be covered by the uncertainties of the fit parameters (see Section 9.5). Therefore, only their effect on the shape of the measured cross sections is considered. The only exception is the PDF uncertainty, since no dedicated samples are needed for its evaluation.

Matching threshold between matrix element and parton shower

The matching of a matrix-element calculation to a parton shower model is characterised by a *matching threshold* (see Section 6.1.2), which defines the phase space for events that should be taken from ME calculation or from PS generator. For the MadGraph+pythia simulations the matching threshold is 20 GeV. In order to evaluate the effect from another threshold choice, the two samples produced with 40 GeV and 10 GeV thresholds were used. The shape differences between the measured-quantity distributions at generator level in the Nominal and varied samples were used as an estimate of systematic effect from the matching threshold variation. The p_{T} spectra of the first and second additional b jets with the matching scale variations are shown in Figure 9.38. The effect on $|\eta|$ is smaller, while the effect on $m_{b\bar{b}}$ and $\Delta R_{b\bar{b}}$ is of similar size. The variations in signal shapes are within the statistical uncertainties of the Nominal distribution.

(a) p_T of the first additional b jet

(b) p_T of the second additional b jet
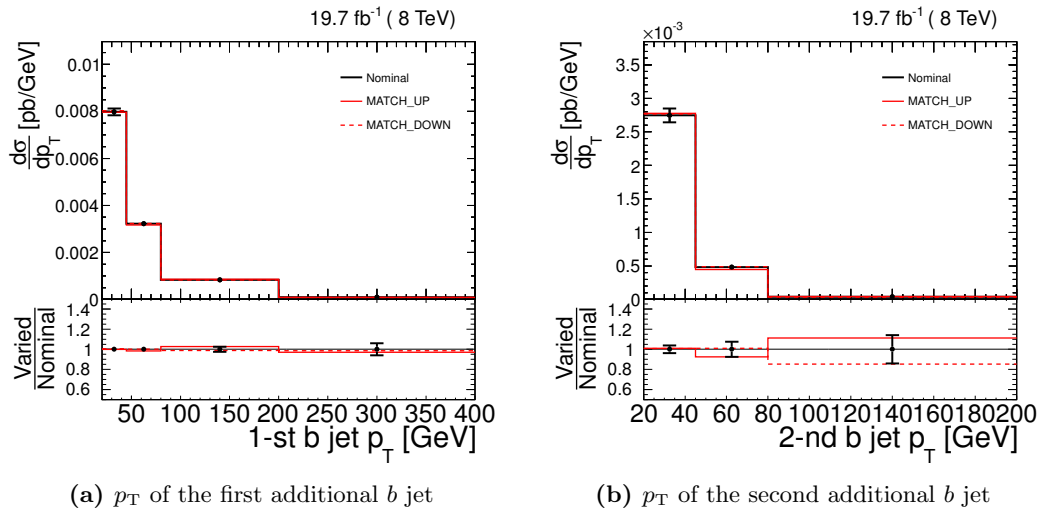
Figure 9.38: Generator-level cross sections produced by the MADGRAPH+PYTHIA simulation in the Nominal configuration and with the ME-PS matching threshold varied up and down by factor 2.

Hard scattering renormalisation and factorisation scale

A dependence of the ME calculation on the factorisation (μ_F) and renormalisation (μ_R) scales (see Sections 2.2.2 and 2.1.4) is addressed by comparison to special MC simulations with varied scales. In the nominal sample simulated with the MADGRAPH ME generator, both μ_F and μ_R are set to the Q^2 scale of the event, which is defined as:

$$Q^2 = m_t^2 + \sum p_\mathrm{T}^2 \,, \tag{9.16}$$

where the sum runs over all the additional jets in the event. Thus, the samples with varied scales are generated using $\mu_\mathrm{F} = \mu_\mathrm{R} = 4Q^2$ and $Q^2/4$ [246].

Simultaneously with the variations of μ_F and μ_R in the MADGRAPH generator, starting and evolution scales of parton showers in the PYTHIA generator are varied by the same factors (see Section 6.1.2). This allows to also account for changes in the initial- and final-state radiation that is added by the PS generator [246].

The applied variations of a factor 4 and 1/4 turned out to be too large, since they lead to deviations that are not compatible with data. This is supported by the results of the jet-multiplicity measurement in dileptonic $t\bar{t}$ events at $\sqrt{s} = 8\,\mathrm{TeV}$ [247]. The corresponding distribution of jet multiplicity, regardless of their flavour, is shown in Figure 9.39. Thus, in order to assign an uncertainty to this source, the value was conservatively taken from the inclusive $t\bar{t}b\bar{b}$ cross-section measurement [119], which is equal to 8 %, and was assigned to each bin of all the measured differential cross sections.

Hard-scattering model

The central values for cross sections are obtained with the MADGRAPH+PYTHIA simulation of signal and dominant backgrounds. In order to evaluate how a different choice of hard-scattering model would affect the measurement, a sample simulated with POWHEG+PYTHIA generator is used. The modelling of parton showers by

(a) $p_\mathrm{T}^\mathrm{jet} > 30\,\mathrm{GeV}$ **(b)** $p_\mathrm{T}^\mathrm{jet} > 60\,\mathrm{GeV}$
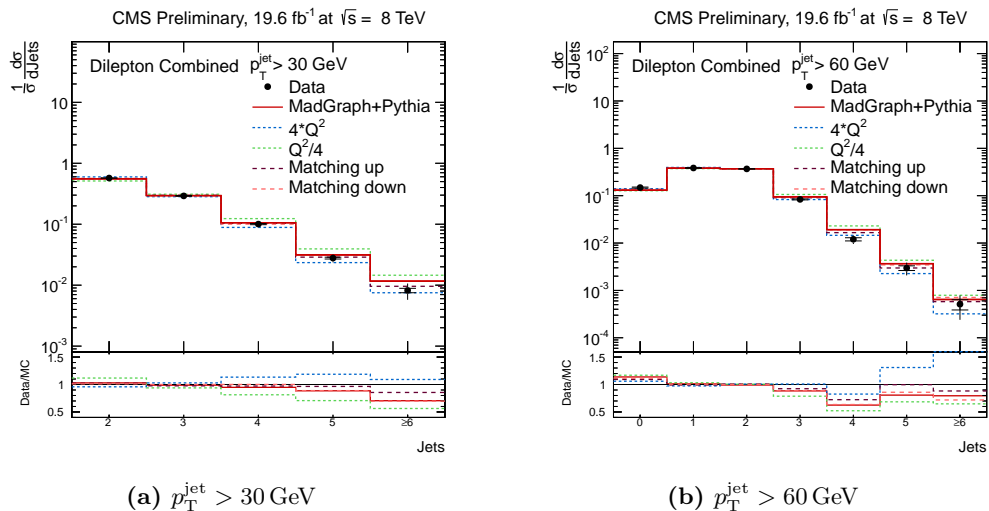
Figure 9.39: Normalised differential cross section as a function of jet multiplicity for jets
with different p_T thresholds. Compared are the measured cross section and
predictions with different variations of μ_F and μ_R, and ME-PS matching
threshold in the MADGRAPH+PYTHIA simulation, which are identical to
the ones used in the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ measurements described in this thesis.
Bins with 3 or 4 jets correspond to 1 or 2 additional jets, since 2 jets
correspond to b jets from the $t\bar{t}$ decay. The predictions obtained from
simulations with varied μ_F and μ_R scales are too large, since they are
incompatible with the measured cross sections. Taken from [247].

PYTHIA in both samples is assumed to be identical. The difference in the shapes of
the generator-level distributions between the two samples is used as an estimate of
the corresponding systematic uncertainty, and is shown in Figure 9.40. Variations
up to 40 % are observed in the hight-p_T region of the first additional b jets. The
effect is stronger for the $t\bar{t}2b$ process than for $t\bar{t}b$, as demonstrated in Figure 9.41,
which presumably is due to a different treatment of quasi-collinear gluon splitting
in the two ME generators. For the $t\bar{t}b\bar{b}$ cross section, the effect is not larger than
the statistical uncertainties.

Top-quark mass

The Nominal $t\bar{t} + jets$ simulation is generated with a top-quark mass of $m_t^\mathrm{MC} = 172.5\,\mathrm{GeV}$, while the experimentally measured value is $m_t^\mathrm{meas.} = 173.3\pm0.8\,\mathrm{GeV}$ [248].
In order to check the effect from a different assumption about the top-quark mass,
two samples generated with $m_t^\mathrm{MC} = 171.5\,\mathrm{GeV}$ and $m_t^\mathrm{MC} = 173.5\,\mathrm{GeV}$ are used. The
difference in the shapes of generator-level distributions between the two samples is
used as an estimate of the systematic uncertainty due to the top-quark-mass vari-
ation, and is shown in Figure 9.42. Its effect on the shape of the signal is below
5 %.

Parton distribution function

The Nominal $t\bar{t} + jets$ MC sample is simulated with the MADGRAPH generator,
for which the CTEQ6L1 [249] PDF is used (see 2.2.1). Since no uncertainties are

(a) p_T of the first additional b jet

(b) p_T of the second additional b jet
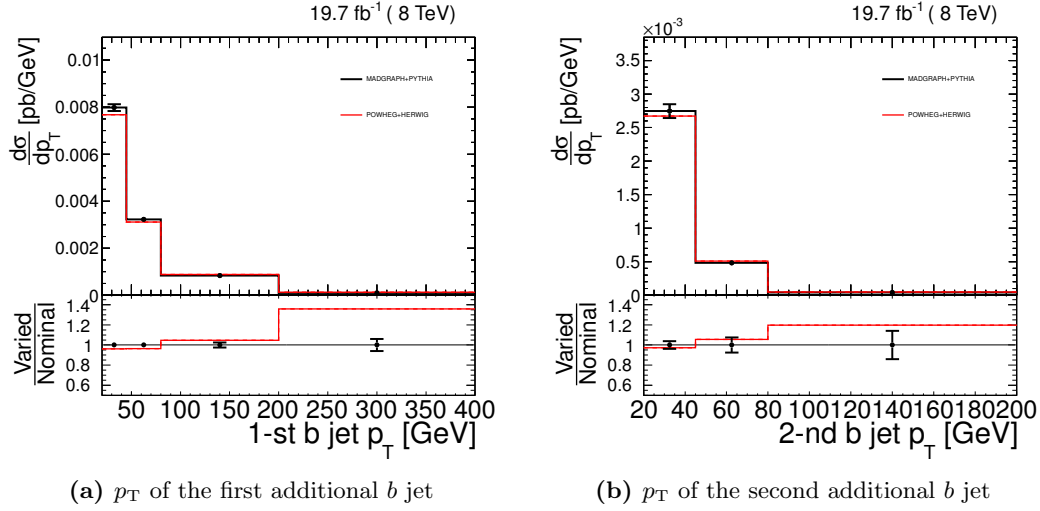
Figure 9.40: Generator-level cross sections produced by the MadGraph+pythia and powheg+pythia simulations.

(a) $t\bar{t}b$ process
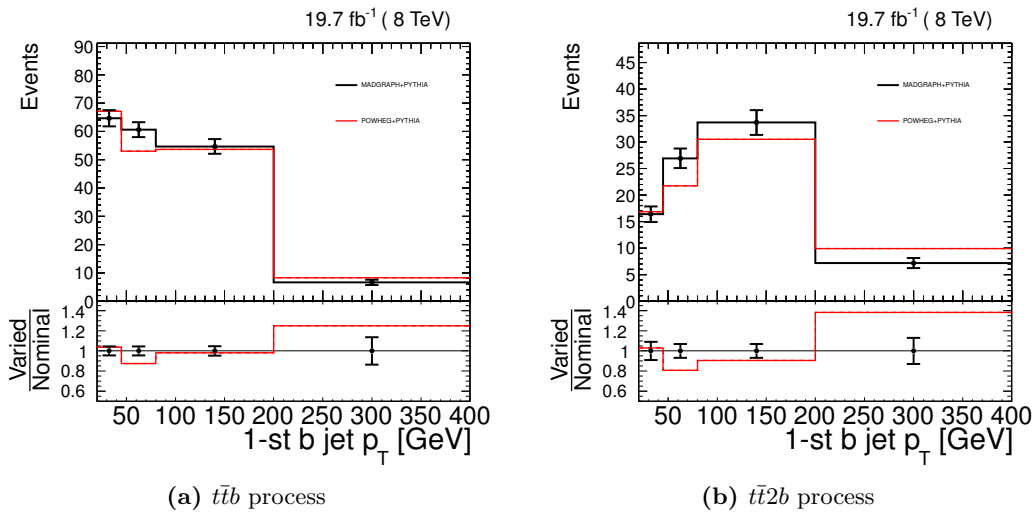
(b) $t\bar{t}2b$ process

Figure 9.41: Generator-level spectra of p_T of the first additional b jet produced by the MadGraph+pythia and powheg+pythia simulations. The $t\bar{t}b$ and $t\bar{t}2b$ processes are shown separately to demonstrate the larger effect of hard-scattering model on the production of two b jets that can not be resolved by the jet clustering algorithm.

(a) p_{T} of the first additional b jet

(b) p_{T} of the second additional b jet
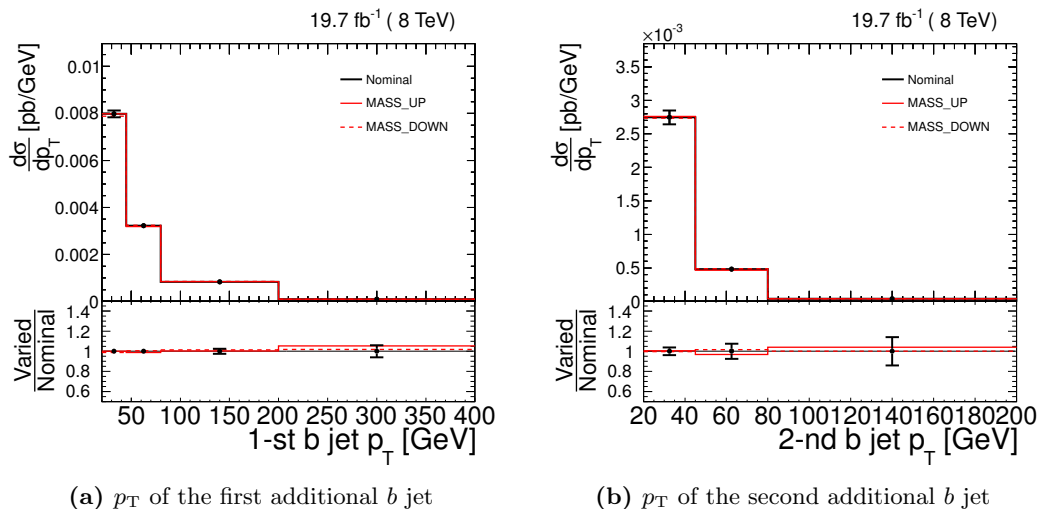
Figure 9.42: Generator-level cross sections produced by the MADGRAPH+PYTHIA simulation with different assumptions about the top-quark mass.

provided for the LO CTEQ6L1 PDF, the CT10 [250] PDF set is used to estimate the PDF uncertainty, following the recommendations of the Top PAG [251] and PDF4LHC group [252]. For the CT10 PDF set, uncertainties are provided by a total of 53 individual sets of weights, which contain 1 central PDF and 26 up/down variations. The PDF uncertainty on the measured cross section is evaluated by comparing the 52 reweighted samples to the sample reweighted by the central CT10 PDF set. The two variations that provide the largest up or down deviation of the inclusive cross section are identified for each individual measured quantity. These two variations are used to evaluate their effect in each bin of the measured distribution. The difference between the varied and central PDF sets is taken as a systematic uncertainty due to PDF variations, and is added in quadrature to other systematic uncertainties. The largest variation of the measured cross sections in a single bin is 5 %. Its effect on the $t\bar{t}b(\bar{b})$ cross section is even smaller, because it mainly affects only the normalisation of the $t\bar{t}b\bar{b}$ process, which is measured from data.

Possible uncertainty evaluation using the Nominal sample

In principle, the nominal sample could be reweighted by applying the difference between the nominal and varied distributions at the generator level. In this way, the propagation of the changes at generator level to the reconstructed quantities would reflect the effect of the corresponding variation on the shapes of background distributions. Furthermore, such variations can also change the measured cross section in the full $t\bar{t}$ phase space, since the shape of the simulated signal distribution is used to extrapolate from the measured visible phase space to the full phase space. This approach is technically rather complicated, and it could not be implemented within the scope of this thesis.

9.8.3 Summary of uncertainties

The summary of typical statistical and systematic uncertainties for the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ cross sections is listed in Table 9.4. The typical uncertainty from a certain source for a single process is caluated as an average of median uncertainties across all the cross-sections measured for the process. A median of a single differential cross section with N bins is defined as the $[\frac{N}{2}]$-th or $[\frac{N}{2}+1]$-th largest uncertainty of a single bin among the odd or even number of bins respectively.

The dominant systematic uncertainty is related to b-tagging. Nevertheless, all the systematic uncertainties are much smaller than the statistical uncertainty. Therefore, in the future, when more data will become available, special care should be taken to improve the precision of b-tagging or to develop such measurement methods that would make the measured results less sensitive to its performance.

Table 9.4: Summary of typical uncertainties on the absolute differential cross sections in the visible $t\bar{t}$ phase space.

Uncertainty source	Uncertainty [%]	
	$t\bar{t}b(\bar{b})$	$t\bar{t}b\bar{b}$
Statistical	28	50
Experimental		
Trigger and lepton id. efficiency	1	1
Background normalisation	4	6
Jet Energy Scale	6	9
Jet Energy Resolution	4	5
Pileup	2	3
b-tagging	13	13
Theory modelling		
Hard scattering model	4	8
Top-quark mass	2	2
Q^2 scale	8	8
ME-PS matching threshold	3	6
PDF	2	1
Total	**31**	**54**

All uncertainties are treated as uncorrelated, and are added in quadrature.

9.9　Results

All the cross sections, defined in Section 9.2, have been measured separately in the full and visible $t\bar{t}$ phase space, using all the procedures described in the previous sections. The results are presented as binned distributions, with each plot having 5 entries:

- **Data:** the measured unfolded cross section (represented by black markers);

- **Madgraph+Pythia:** the cross section predicted by the MADGRAPH generator interfaced with PYTHIA parton shower, as used in the $t\bar{t} + jets$ simulated sample (represented by a red solid line).

- **Powheg+Pythia:** the cross section predicted by the POWHEG generator interfaced with PYTHIA parton shower (represented by a blue solid line).

- **Powheg+Herwig:** the cross section predicted by the POWHEG generator interfaced with HERWIG parton shower (represented by a green dashed line).

- **MC@NLO+Herwig:** the cross section predicted by the MC@NLO generator interfaced with HERWIG parton shower (represented by an orange dashed line).

The MADGRAPH+PYTHIA prediction provides ME calculations of the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ processes using only tree-level diagrams that contribute to the cross section at up to NLO (see Section 6.1.1). Other MC predictions, i.e. POWHEG+PYTHIA, POWHEG+HERWIG and MC@NLO+HERWIG, provide calculations of the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ processes only by parton-shower generators.

Calculations of differential $t\bar{t}b\bar{b}$ cross sections at the NLO accuracy matched by parton showers are approached by theorists, and some results have been already achieved [213, 212, 211]. But the process definitions used by the calculations do not exactly match the ones used in the described analysis. Furthermore, usually results are not available publicly for a comparison. Thus, calculations of NLO predictions were approached by M.Garzelli [213] for the 4 cross sections of the $t\bar{t}b\bar{b}$ process with the exact definitions described in Section 9.2. Due to time constraints, the predictions could not be obtained in time to be included in this thesis, but are in preparation.

Inclusive cross sections

The inclusive $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ cross section are defined only by the phase space definition, thus the values of p_{T}, $|\eta|$, $\Delta R_{b\bar{b}}$ or $m_{b\bar{b}}$ can exceed the range in which the differential cross sections are measured. The values of the measured and predicted inclusive cross-sections are presented in Table 9.5.

The measured $t\bar{t}b\bar{b}$ inclusive cross sections in the full $t\bar{t}$ phase space are about 90 % higher than the most of LO MC predictions, while the $t\bar{t}b(\bar{b})$ cross section is 30 % higher than predicted by the MC. The only exception is the MC@NLO+HERWIG simulations, which underestimates the inclusive cross sections by up to factor of 5. The underestimated MC predictions can be explained by missing higher-order corrections from perturbative calculations, as explained in Section 2.1.4. The difference between cross sections calculated at LO and NLO was estimated in [212] for slightly higher b-jet p_{T} threshold of 25 GeV and slightly wider pseudorapidity range of $|\eta| < 2.5$. The increase of the NLO cross section of $t\bar{t}$ productions with only 1 additional b jet was estimated to be 25 %, which is not much different from the measured increase of 30 %. On the other hand, the theoretically estimated increase from LO value for the $t\bar{t}b\bar{b}$ process is only 20 % compared to the measured 90 %. The difference presumably is caused by different jet-p_{T} thresholds, which affects the cross sections more strongly if two such b jets are required.

The same result was obtained by the more precise measurement of the inclusive $t\bar{t}b\bar{b}$ cross section [119] by CMS. A more thorough comparison between the two analyses is made in Section 9.10.

Table 9.5: Measured and predicted inclusive cross sections of the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ production.

Phase space	$\sigma_{t\bar{t}b\bar{b}}$ [pb]	$\sigma_{t\bar{t}b}$ [pb]
Full		
• Measured	0.090 ± 0.022	0.435 ± 0.039
• MADGRAPH+ PYTHIA	0.047 ± 0.002	0.326 ± 0.011
• POWHEG+ PYTHIA	0.052 ± 0.002	0.341 ± 0.012
• POWHEG+ HERWIG	0.044 ± 0.002	0.362 ± 0.013
• MC@NLO+ HERWIG	0.017 ± 0.001	0.240 ± 0.008
Visible		
• Measured	0.033 ± 0.007	0.159 ± 0.014
• MADGRAPH+ PYTHIA	0.017 ± 0.001	0.120 ± 0.004
• POWHEG+ PYTHIA	0.019 ± 0.001	0.123 ± 0.004
• POWHEG+ HERWIG	0.016 ± 0.001	0.129 ± 0.005
• MC@NLO+ HERWIG	0.006 ± 0.001	0.084 ± 0.003

Only statistical uncertainties are quoted for the MC prediction. For the measured cross-sections, statistical and total systematic uncertainties are included. The total systematic uncertainties include all the systematic uncertainties listed in Table 9.4 added in quadrature.

Absolute differential cross sections

Every predicted cross section is normalised to the corresponding inclusive measured cross section, to better see the differences in shapes. Both the measured and predicted differential cross sections in each bin are divided by the width of the bin. The differential cross sections of the $t\bar{t}b(\bar{b})$ process are shown for the full and visible $t\bar{t}$ phase space in Figure 9.43. The differential cross section of the $t\bar{t}b\bar{b}$ process are shown for the full phase space in Figure 9.44, and for the visible phase space in Figure 9.45.

The points of the measured cross sections have outer vertical bars, representing the total uncertainty, while the inner bars represent the statistical uncertainty. In the bottom section of each plot, a ratio of the predicted cross section by the measured one is presented.

The differential cross sections of the $t\bar{t}b(\bar{b})$ process have an average total uncertainty of about $30\,\%$ in every bin, which is dominated by the statistical uncertainty. Compared to it, systematic uncertainties are almost negligible and visible only in the cross section as a function of the first-additional-b-jet p_{T} in the range $45\,\mathrm{GeV}$ to $80\,\mathrm{GeV}$. It also should be kept in mind that different bins of the measured distributions have become correlated after the applied unfolding (see Section 9.7), according to the correlation matrices shown in Figure 9.32. Thus, the statistical uncertainties in different bins should not be interpreted as fully independent.

Besides the differences in the absolute cross sections shown in Table 9.5, the shapes of the p_{T} and $|\eta|$ are slightly different. In particular, all the MC predictions tend to have a slightly harder p_{T} spectrum of the first additional b jet than the measured one, although in the region of $p_{\mathrm{T}} < 80\,\mathrm{GeV}$ the predicted and measured

spectra agree within the statistical uncertainty. With this precision, data does not yet provide sufficient sensitivity to discriminate between different models.

(a) full $t\bar{t}$ phase space

(b) visible $t\bar{t}$ phase space
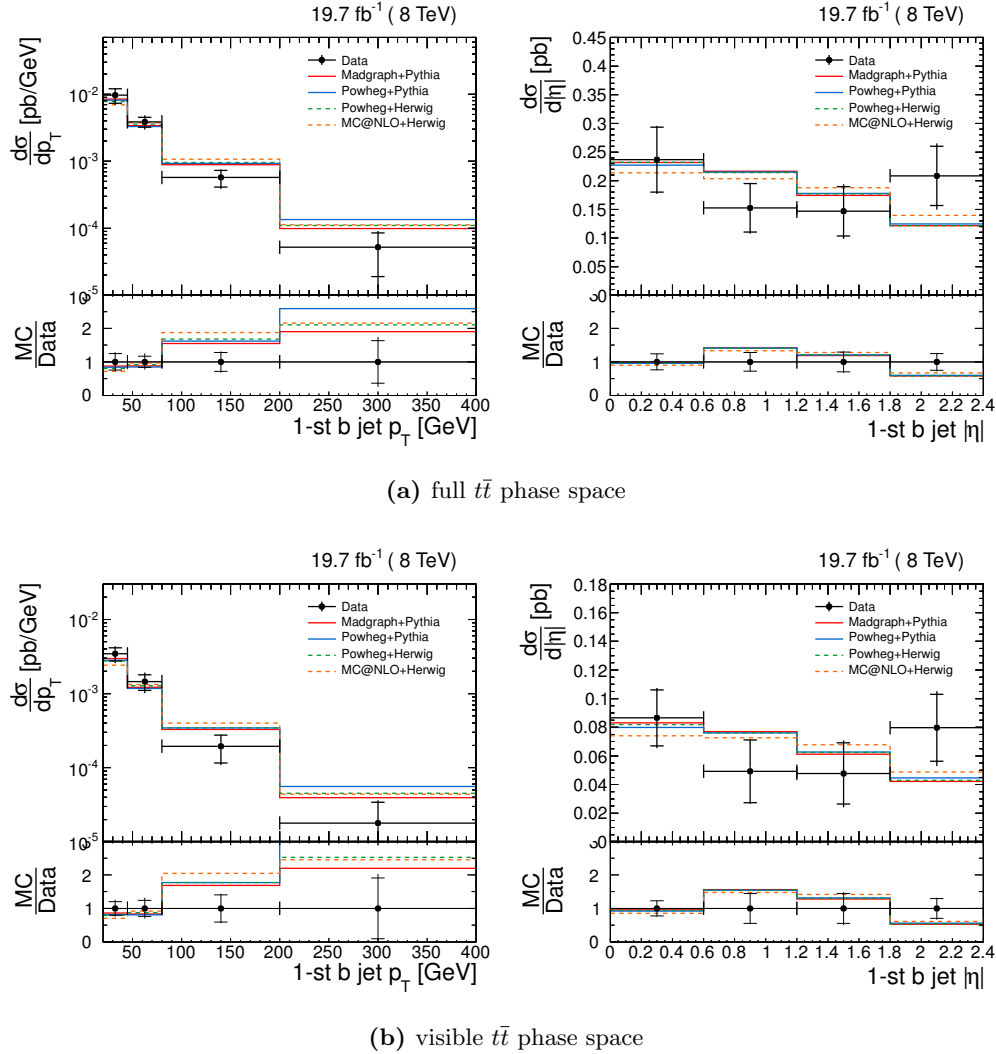
Figure 9.43: Differential cross sections of the $t\bar{t}b(\bar{b})$ process in the full and visible $t\bar{t}$ phase space as functions of the quantities mentioned in the captions.

The properties of the second additional b jet and of the pair of the first two b jets are measured with worse precision due to the very limited statistics of the selected sample. The shapes of the measured and predicted spectra of the p_{T} and $|\eta|$ of the second additional b jet are in agreement, considering the uncertainties of 20 % to 100 % per bin.

The measured $\Delta R_{b\bar{b}}$ spectrum is in agreement with the predicted one within uncertainties, and is compatible at 68 % confidence level. The MC@NLO+HERWIG simulation has a significant deviation from the measured spectrum and does not describe the data. The measured $m_{b\bar{b}}$ spectrum is well described by the MC predictions. The range $100 < m_{b\bar{b}} < 140\,\mathrm{GeV}$, which is most sensitive to the not-yet-discovered $t\bar{t}H(b\bar{b})$ process, is not shown in data. In this region, all the MC simulations provide similar results, except of MC@NLO+HERWIG, which does not describe the data, as concluded from the $\Delta R_{b\bar{b}}$ spectrum in Figure 9.45b. The remaining bins constitute an important control region for the $t\bar{t}H(b\bar{b})$ search. The visible agreement between

the simulations and data proves that the studied predictions of the $t\bar{t}b\bar{b}$ process can be used for the estimation of the shape of the main background to the $t\bar{t}H(b\bar{b})$ process, at least in the kinematic range of the b jets that is considered in the analysis. The normalisation of the $t\bar{t}b\bar{b}$ process is well constrained by the template fit to data, as described in Section 9.5.

(a) properties of the 2-nd additional b jet

(b) properties of the pair of the first and second additional b jets
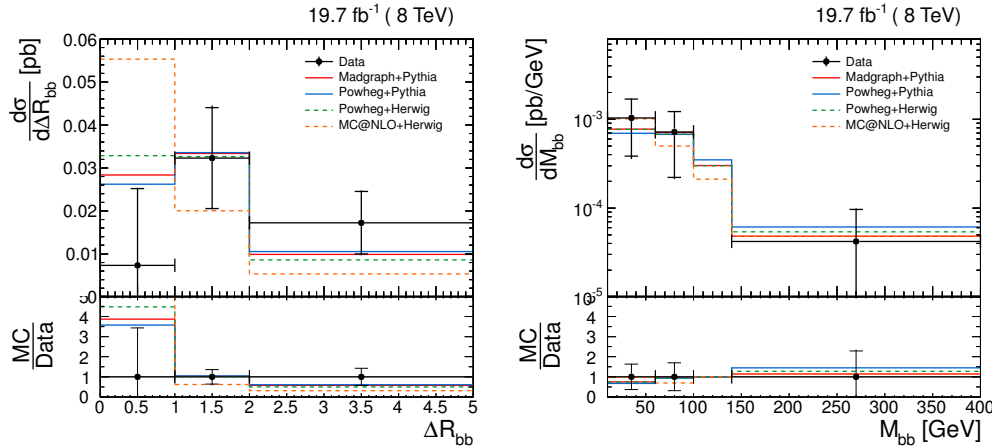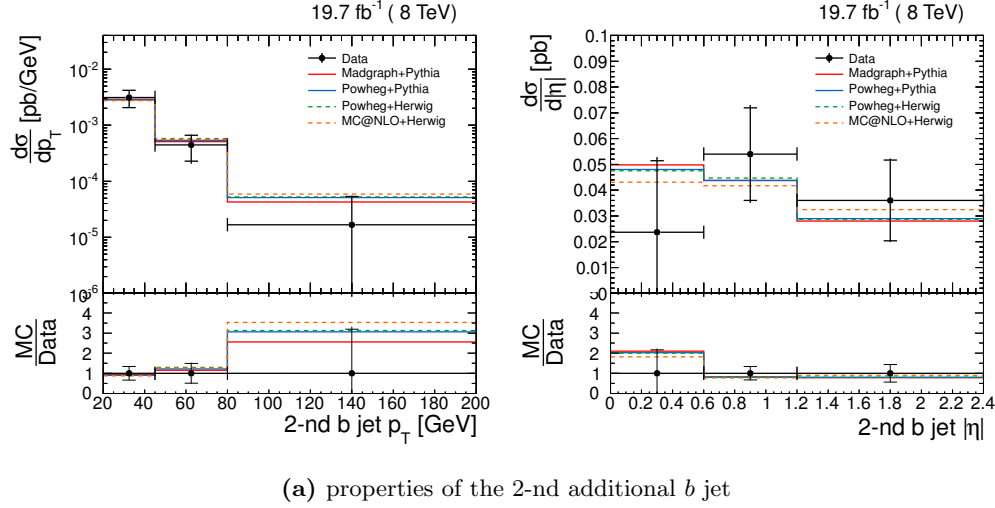
Figure 9.44: Differential cross sections of the $t\bar{t}b\bar{b}$ process in the full $t\bar{t}$ phase space as functions of the quantities mentioned one the X axes.

9.10   Comparison to the existing measurements

As shown in Table 9.5, the measured inclusive $t\bar{t}b\bar{b}$ and $t\bar{t}b$ cross section are higher than predicted by the MADGRAPH generator interfaced with PYTHIA parton shower.

A similar disagreement was found by the dedicated measurement of the inclusive $t\bar{t}b\bar{b}$ cross section [119], which is an ingredient of the measurement of the ratio:

$$R = \frac{\sigma_{t\bar{t}b\bar{b}}}{\sigma_{t\bar{t}jj}}, \tag{9.17}$$
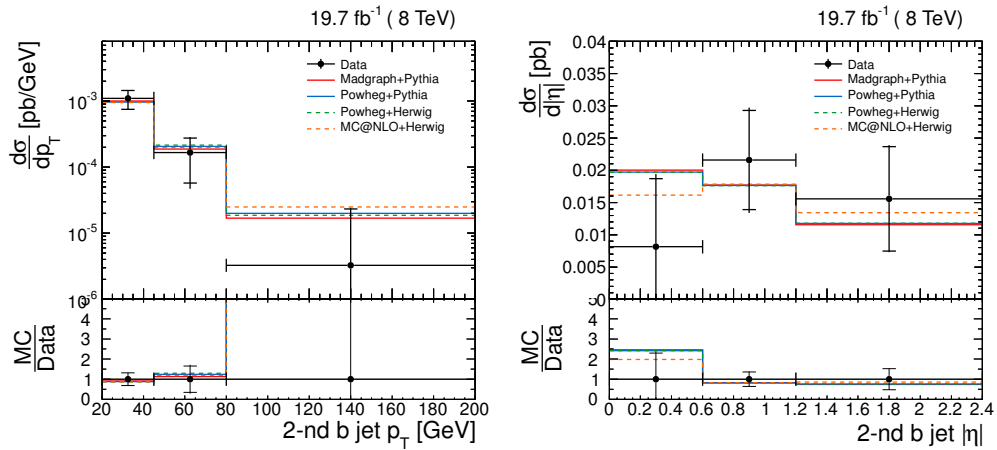
(a) properties of the 2-nd additional b jet
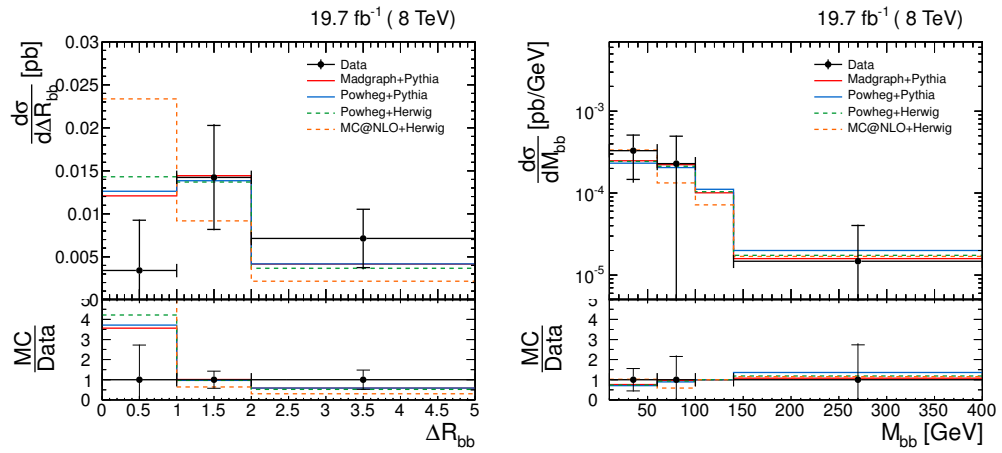
(b) properties of the pair of the first and second additional b jets

Figure 9.45: Differential cross sections of the $t\bar{t}b\bar{b}$ process in the visible $t\bar{t}$ phase space as functions of the quantities mentioned one the X axes.

where the $t\bar{t}jj$ represents the $t\bar{t}$ produced in association with at least two jets of any flavour. Before the direct comparison of the cross sections of the two measurements it is important to thoroughly check all the differences in the two cross-section definitions.

9.10.1   Differences in the cross-section definitions

What is relevant for the comparison are the definitions of the $t\bar{t}b\bar{b}$ process in the current analysis and in the previous inclusive measurement [119]. The major part of the definitions used in the inclusive measurement is documented in the corresponding paper [119], while some details were obtained through a private communication with the main author of the paper, Tae Jeong Kim [253].

In the inclusive cross-section measurement, several values are quoted: for the full and visible phase. In addition, the cross section for the full phase space is quoted with two different p_{T} thresholds for the additional b jets, for comparison with theoretical calculations. Nevertheless, only the value for the visible phase space can be compared, since the signal definition at the full phase space is done at parton level, which can not be easily implemented in the current analysis.

Thoroughly checking all the details, a set of differences in the cross-section definitions in the visible phase space were found, which are summarised in Table 9.6. Other definitions, including lepton selection, dileptonic-channel definition and τ-decay treatment was ensured to be identical. Thus, in order to make a direct comparison between the results of the two analyses, the definitions from the previous measurement were adopted were possible.

Table 9.6: Differences between the cross-section definitions of the $t\bar{t}b\bar{b}$ process

Current analysis	Inclusive analysis [119]
$|\eta|$ **of the additional b-jets**	
2.4	2.5
$|\eta|$ **of the b-jets from the $t\bar{t}$ decay**	
2.4	2.5
p_{T} **of the b jets from the $t\bar{t}$ decay**	
30 GeV	20 GeV
required combination of b jets	
2 from $t\bar{t}$ and ≥ 2 additional	any 4 b jets (regardless of origin)
exclusion of the $t\bar{t}$-decay leptons from jets	
leptons excluded before jet clustering	jets with $\Delta R(jet, l^{\pm}) < 0.5$ removed

Adapting the signal definition in several steps

In order to better understand the effect of the listed differences on the measured inclusive cross section, the change of the signal definition was performed in several steps. Thus, the following steps are defined:

Step 1 looser selection of b jets for the visible phase space;

- **additional b jets:** $p_{\mathrm{T}} > 20\,\mathrm{GeV}, |\eta| < 2.5$
- **b jets from the $t\bar{t}$ decay:** $p_{\mathrm{T}} > 20\,\mathrm{GeV}, |\eta| < 2.5$

Step 2 no requirement of specific origin of each b jet

- **$t\bar{t}b\bar{b}$ signal:** any 4 b jets in acceptance

Step 3 additional rejection of jets with $t\bar{t}$-decay leptons $(l^{\pm})$

- **use only jets sparated from the leptons:** $\Delta R(jet, l^{\pm}) > 0.5$

The inclusive cross section was calculated for each individual step to see how the changes in the signal definition are reflected in the measured cross section.

After implementing all the three steps, the only remaining difference from the definition in the inclusive measurement [119] is that jets in the current analysis are clustered without the leptons from the $t\bar{t}$ decay. Nevertheless, the effect of this difference is expected to be negligible, since the leptons that are at larger distance from a jet than $\Delta R = 0.5$, are very unlikely to be clustered into the jet. This is motivated by the fact that the generator level jets are clustered with the distance parameter $R = 0.5$, and such strongly-displaced particles must have p_{T} much smaller than that of a jet in order to be clustered into it.

9.10.2   Results of the comparison

The inclusive cross sections calculated after subsequent change of the $t\bar{t}b\bar{b}$ process definition in accordance to the three steps listed in Section 9.10.1, are given in Table 9.7.

Table 9.7: Differences between the cross-section definitions of the $t\bar{t}b\bar{b}$ process

Step	$\sigma_{t\bar{t}b\bar{b}}$ [pb]
1	0.038 ± 0.009
2	0.038 ± 0.009
3	0.030 ± 0.007
Inclusive analysis	$0.029 \pm 0.003 \pm 0.008$

Several conclusions can be made from the obtained comparison results. First of all, the $t\bar{t}b\bar{b}$ cross section measured in the current analysis is completely compatible with the more precise value obtained by the previous measurement. This confirms that the methods used for the identification of additional b jets, and for the correction of the normalisation of different subsets of $t\bar{t}+jets$ events do not introduce a bias for the $t\bar{t}b\bar{b}$ measurement, which would make the result incompatible with the existing result.

Secondly, the absence of any difference between the step 1 and step 2 means that if there are at least 4 b jets in the acceptance, the 2 from the $t\bar{t}$ decay are also in the acceptance. Furthermore, it shows that the pseudoadditional b jets radiated after the $t\bar{t}$ decay do not make any noticeable contribution to the cross section, since in

this case there is no explicit requirement that the additional b jets must be radiated before the top decay. Nevertheless, this is only true if both the b jets from top and additional b jets have the same phase space requirement. In the default definition of the differential measurement, the b jets from the $t\bar{t}$ decay have a higher p_{T} threshold, which would make it more probable that a third additional or pseudoadditional b jet fulfils the acceptance requirements, while one of the b jets from top would not be in the acceptance.

Finally, the striking reduction of the cross section after step 3 suggests, that the requirement $\Delta R(jet, l^{\pm})$ is too strong for the $t\bar{t}b\bar{b}$ process. This might not make such a dramatic effect in the $t\bar{t}$ measurement, where the number of jets is much smaller and the two leptons from the $t\bar{t}$ decay are well separated. But in a $t\bar{t}b\bar{b}$ event, there are at least 4 b jets with large transverse momentum, which occupy a significant part of the available phase space. Therefore, it is much more probable that at least one of the leptons will be close enough to one of the b jets, causing it to be removed from the event.

9.11   Conclusions

Differential cross sections of $t\bar{t}$ production in association with at least one ($t\bar{t}b(\bar{b})$) or two ($t\bar{t}b\bar{b}$) additional b jets were measured in the dileptonic decay channel of the top-quark pair, using all data recorded by CMS during 2012. The cross sections were measured as functions of kinematic properties of the first or second additional b jets. Furthermore, for the production of two additional b jets, the cross section was measured as a function of angular distance between the two b jets ($\Delta R_{b\bar{b}}$), and as a function of their invariant mass ($m_{b\bar{b}}$). This measurement provides an important test of QCD predictions of additional heavy-flavour radiation.

Using the developed method for estimation of the fraction of events with additional b jets, it was found that the amount of additional heavy-flavour radiation is underestimated by MC simulations by about 60 %, which is consistent with observations from a previous inclusive measurement of the same process. The statistical precision that could be achieved with the available data is rather limited, which significantly reduces the relevance of systematic effects for this measurement. The typical uncertainty in a single bin of the measured cross sections is about 30 % for the $t\bar{t}b(\bar{b})$ process, and about 55 % for the $t\bar{t}b\bar{b}$ measurement.

If not taking into account the underestimated production rate of the studied processes, most of the measured spectra are in a good agreement with the measurement. The measured shape of the p_{T} spectrum of the first additional b jet is slightly softer than predicted by the studied MC generators, but is within uncertainties. A slight tension between MC predictions and data is observed in the $\Delta R_{b\bar{b}}$ distribution, which demonstrates that most of the compared MC simulations overestimate the cross section of additional b jets produced with a small angular separation. Neither of the studied MC simulations provide full NLO calculations of the $t\bar{t}b\bar{b}$ process at ME level; additional radiation is added only by tree-level diagrams in the MADGRAPH+PYTHIA simulation, or by parton-shower generators in the other simulations. Therefore, the discrepancy in $\Delta R_{b\bar{b}}$ spectrum can reflect an underestimation of the production rate of two pairs of additional quasi-collinear b-quark pairs that can not be resolved at the jet level, which would have a larger angular distance. In this respect, a comparison to a full next-to-leading order calculation of such cross

sections would be very interesting. Such calculations are in progress, but could not
be finalised in the scope of this thesis.

The measurement of the $t\bar{t}b\bar{b}$ cross section as a function of $m_{b\bar{b}}$ is especially
important in the view of a search for $t\bar{t}H$ production, where the Higgs boson decays
to a b$\bar{\text{b}}$ pair. The $t\bar{t}b\bar{b}$ process is its dominant and almost irreducible background.
Thus, in order to see a peak from the Higgs boson in the invariant-mass distribution
of the two additional b jets, not only the rate of $t\bar{t}b\bar{b}$ process has to be known,
but also the shape of its $m_{b\bar{b}}$ spectrum. With the available statistical precision,
the measurement shows that all the tested leading order MC predictions properly
describe the shape of the $m_{b\bar{b}}$ distribution, and that differences between various
generators are rather small. The only exception is the MC@NLO generator, which
differs from other prediction by up to 30 % in the potential signal region, close to the
Higgs-boson mass. Thus, it can be concluded that the existing MC predictions of the
$t\bar{t}b\bar{b}$ process can be used for estimation of the dominant background to the $t\bar{t}H(b\bar{b})$
production, without the need for sophisticated data-driven background-estimation
methods.

Chapter 10

Sensitivity to t$\bar{\text{t}}$H production

As explained in Section 2.3.4, and shown in Figure 9.31f, the $t\bar{t}b\bar{b}$ process is the most important and irreducible background to $t\bar{t}H(b\bar{b})$ production, due to exactly the same final state and a much larger cross section. Contributions from other $t\bar{t} + jets$ processes can be suppressed by a tighter selection of b-tagged jets that provides a smaller mistagging rate. This, of course, would inevitably reduce the available statistics.

The correction of different $t\bar{t} + jets$ subprocesses, described in Section 9.5, has defined the proper normalisation of the $t\bar{t}b\bar{b}$ process, while the shape of the invariant mass of additional b jets, $m_{b\bar{b}}$, is well described by the MADGRAPH prediction, as shown in Figure 9.44. Thus, this distribution is an excellent candidate for a search for the $t\bar{t}H(b\bar{b})$ process, for which it should be distributed around the Higgs-boson mass.

Considering the very low amount of data that is available after the full event selection, only an estimate of sensitivity is done using MC predictions. The sensitivity is estimated for the available amount of data, as well as for potential data that will be collected during the future runs of the LHC at $\sqrt{s} = 14\,\text{TeV}$. This is only a first evaluation, which would need to be extended by a full treatment of systematic effects when a sufficient amount of data will have been collected to see a $t\bar{t}H$ signal.

10.1   Signal definition

The $t\bar{t}H(b\bar{b})$ signal is defined as a $t\bar{t}$ pair produced in association with the SM Higgs boson. All the dileptonic decay channels of the $t\bar{t}$ system are considered and no kinematic requirements are applied to its decay products, which corresponds to the *full $t\bar{t}$ phase space*, as defined in Section 9.2.2. The Higgs boson has a mass of $125\,\text{GeV}$, and must decay to a pair of bottom quarks at the matrix-element level: $H \to b\bar{b}$.

10.2   Analysis strategy

The event selection for the rejection of background processes, as well as all object and event corrections exactly follow the ones used in the $t\bar{t}b\bar{b}$ measurement, as described

in Sections 9.4 and 9.5. The definitions of different $t\bar{t} + jets$ subprocesses ($t\bar{t}b\bar{b}$, $t\bar{t}b$, $t\bar{t}2b$, $t\bar{t}Other$) remain the same as in Section 9.3.

The distribution of the invariant mass of the two additional b jets ($m_{b\bar{b}}$) in the $t\bar{t}H(b\bar{b})$ signal should represent the mass of the Higgs boson, and is used to increase its significance by restricting the expected signal region to the Higgs-boson mass peak. Thus, the two additional b jets, which do not originate from the $t\bar{t}$ decay, have to be identified. Otherwise, one or more wrong jets in the combination will lead to an $m_{b\bar{b}}$ value that does not represent the Higgs-boson mass.

Considering the complicated final state of the $t\bar{t}H(b\bar{b})$ event, which has at least 4 b jets, the misidentification issue is especially important. In an event with 4 b jets it is possible to construct 6 dijet combinations, while only one of them would be correct. Thus, in addition to the background contributions from other processes, the *combinatorial smearing* can be substantial in a properly identified $t\bar{t}H(b\bar{b})$ event. This and other effects on the resolution of the Higgs-boson mass, reconstructed from a pair of jets, are shown in Figure 10.1.

(a) $t\bar{t}H(b\bar{b})$ sample ($m_H = 125\,\text{GeV}$)   **(b)** $t\bar{t}b\bar{b}$ sample
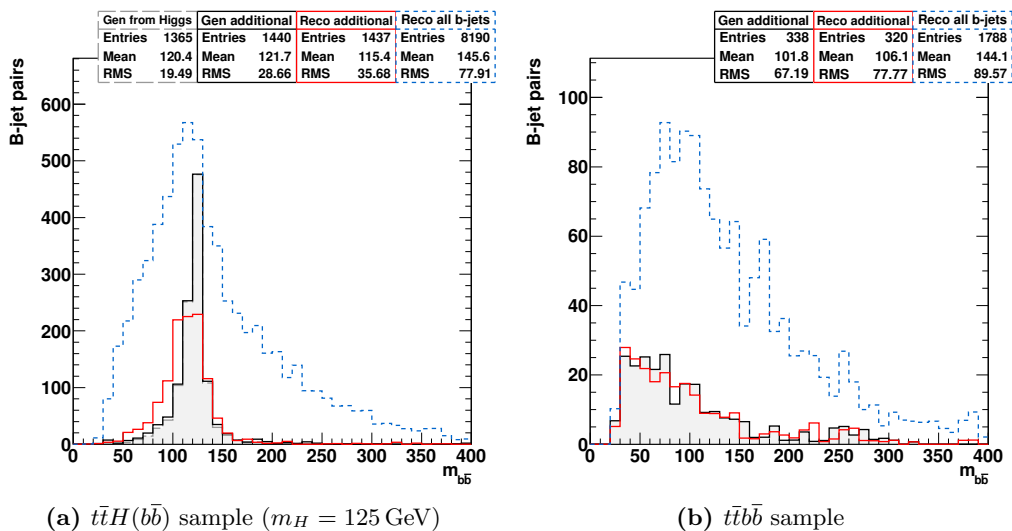
Figure 10.1: Resolution of the dijet mass estimated with different levels of generator-level information involved, and compared between the $t\bar{t}H(b\bar{b})$ and $t\bar{t}b\bar{b}$ processes. Distributions are obtained from events after full $t\bar{t}b\bar{b}$ event selection (see Section 9.4), which have exactly 4 b-tagged jets. At the generator level, b jets and their origin, e.g. stemming from $t\bar{t}$ pair or the Higgs boson, are identified by the *GenHFHadronMatcher* (see Chapter 8). At the reconstruction level, b jets are defined as b-tagged jets using the CSVM working point. Identification of jets from the $t\bar{t}$ system at the reconstruction level is done by matching them to the generator-level b jets from the $t\bar{t}$ decay. *Gen from Higgs* denotes only pairs of generated b jets originating from the $H \rightarrow b\bar{b}$ decay, and is relevant only for the $t\bar{t}H(b\bar{b})$ simulation (1 entry/event). *Gen additional* represents all pairs of generated b jets that do not originate from the $t\bar{t}$ decay ($\geq$ 1 entry/event). *Reco additional* represents all pairs of reconstructed b-tagged jets that are not matched to generated jets from the $t\bar{t}$ decay ($\geq$ 1 entry/event). *Reco all* denotes all pairs of reconstructed b-tagged jets without matching to generated jets from the $t\bar{t}$ decay (6 entries/event).

In order to numerically evaluate how well the Higgs boson mass is reproduced

by the dijet mass, each distribution in Figure 10.1a is fitted by a gaussian function:

$$f(x) = a \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) . \tag{10.1}$$

The parameter μ represents the $m_{b\bar{b}}$ value corresponding to the peak of the fitted distribution and should be as close as possible to the $m_H = 125\,\text{GeV}$, which was used in the generated MC simulation. The parameter σ corresponds to the width of the gaussian shape and is a measure of the mass resolution provided by the fitted distribution. The two parameters were evaluated for each distribution by fitting it in the range $80 < m_{b\bar{b}} < 160\,\text{GeV}$, in order to cover a peak, while reducing the effect of possible tails of the distribution. The obtained values of the peak position μ and the mass resolution σ are listed in Table 10.1.

Table 10.1: Mean values of the dijet mass and its resolution, as obtained from the fit of each distribution in Figure 10.1a by a gaussian.

Jets used in combinations	Mean value [GeV]	Resolution [GeV]
● Gen from Higgs	122	10
● Gen additional	122	11
● Reco additional	115	18
● Reco all	113	37

Fit performed in the range $80 < m_{b\bar{b}} < 160\,\text{GeV}$. Mean and resolution represent the μ and σ parameters of the gaussian. Names of jet-combinations correspond to the ones defined in Figure 10.1

.

As is clearly seen from the numbers in Table 10.1, the Higgs-boson mass is best reproduced by the generator-level b-jets matched to the b-quarks originating from the Higgs-boson decay. The invariant mass of such jet pairs is very close to the generated m_H value, and its resolution is 10 GeV, which is the best resolution that can theoretically be achieved with jets. When using all pairs of b jets that do not originate from the $t\bar{t}$ decay, the resolution degrades only very slightly, by 1 GeV. This is in contrast to the falling spectrum observed in the $t\bar{t}b\bar{b}$ process. A peak is present in the $t\bar{t}b\bar{b}$ process at $m_{b\bar{b}}$ values near the jet-p_T threshold, but it is far enough from the Higgs-boson-mass region.

A much larger effect comes from the transition to the reconstruction level, when using all pairs of b-tagged jets that do not originate from the $t\bar{t}$ decay. Jet energy miscalibration, finite acceptance, mistagging of non-b jets and low b-tagging efficiency lead to the 80 % worse resolution, $\sigma = 18\,\text{GeV}$.

Finally, the strongest degradation of the invariant-mass resolution is caused by the combinatorial smearing, when all possible pairs of b-tagged jets are considered for the $m_{b\bar{b}}$ distribution. The presence of wrong combinations makes the resolution worse ($\sigma = 37\,\text{GeV}$). Furthermore, the shapes of the $m_{b\bar{b}}$ spectra in both the $t\bar{t}b\bar{b}$ and $t\bar{t}H(b\bar{b})$ processes become very similar. This is mainly caused by the large top-quark mass, which leads to the b jets originating from $t\bar{t}$ decays having similar kinematic properties to those originating from $H \to b\bar{b}$ decays. Thus, the number of wrong jet combinations has to be minimised, as will be described in Section 10.3, to optimise the resolution of the dijet-mass distribution.

The numbers of entries in the obtained dijet-mass spectrum corresponding to the $t\bar{t}H(b\bar{b})$ signal process ($N_{t\bar{t}H}$) and to the background processes (N_{bkg}) are then used to quantify the potential sensitivity with the two quantities:

- **signal significance:** good measure of sensitivity for a small amount of data;

$$\mathcal{S}_{ign} = \frac{N_{t\bar{t}H}}{\sqrt{N_{t\bar{t}H} + N_{\mathrm{bkg}}}} \qquad (10.2)$$

- **signal-to-background ratio (S/B):** measure of the maximum sensitivity of an analysis strategy in case of a big amount of data;

$$S/B = \frac{N_{t\bar{t}H}}{N_{\mathrm{bkg}}} . \qquad (10.3)$$

Both the signal significance and S/B are evaluated in the fixed region of the expected $t\bar{t}H$ signal contribution, which was chosen as $[100, 140]$ GeV. The different contributions and the dijet-mass range are shown schematically in Figure 10.2. Thus, the highest sensitivity to the signal contribution can be achieved if its ratio to the background is high and if the overall amount of data together with a good control of systematic effects lead to statistical errors smaller than the signal contribution.
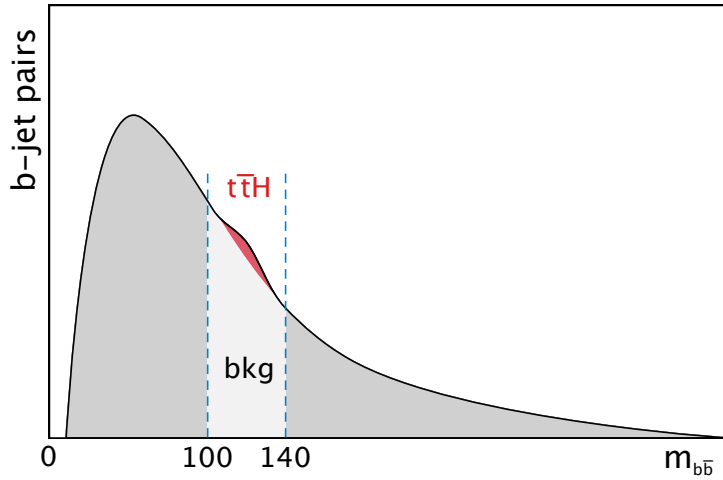
Figure 10.2: Schematic view of the dijet-mass distribution with contributions from the signal $t\bar{t}H(b\bar{b})$ process and from background processes (bkg). A bump is produced by $H \to b\bar{b}$ decays, which have invariant mass smeared around the Higgs-boson mass. A contribution from the $t\bar{t}Z(b\bar{b})$ process, which would produce a similar bump in a smaller-mass region, is neglected, since its cross section is significantly smaller than that of $t\bar{t}H(b\bar{b})$ production. The range used for the calculation of the signal significance and the signal-to-background ratio is shown by the vertical dashed lines.

By convention, in experimental high energy physics, evidence of some process can be concluded if its significance is larger than 3σ. A discovery can be claimed only if the significance is larger than 5σ. Thus, considering the $m_{b\bar{b}}$ quantity and the developed background suppression methods, an estimate is done of the luminosity needed for evidence using data collected by the LHC running at $\sqrt{s} = 8$ TeV and $\sqrt{s} = 14$ TeV. The estimated numbers will be presented in Section 10.4.

10.3   Suppression of combinatorial smearing

It was shown in the previous Section 10.2 that the presence of multiple possible jet combinations dramatically degrades the Higgs-mass resolution and increases background contributions in the signal region, as shown in Figure 10.1. Therefore, several methods were tried to reduce the number of jet combinations that do not represent the $H \to b\bar{b}$ pair, while keeping most of the correct jet pairs.

10.3.1   A pair of leading additional *b*-tagged jets

The first and most simple method studied is the one used in the $t\bar{t}b\bar{b}$ measurement, as described in Section 9.6. The MVA is used to identify the b jets from the $t\bar{t}$ system and the invariant mass of the remaining two b-tagged jets with highest p_{T} is plotted.

In this case, in every event that fulfils the event selection, a single pair of b-tagged jets is used. Thus, background contributions can be rejected only via event selection. The obtained dijet-mass distributions are shown in Figure 10.3.

(a) $t\bar{t}H(b\bar{b})$ overlaid and normalised to the sum of backgrounds

(b) $t\bar{t}H(b\bar{b})$ stacked
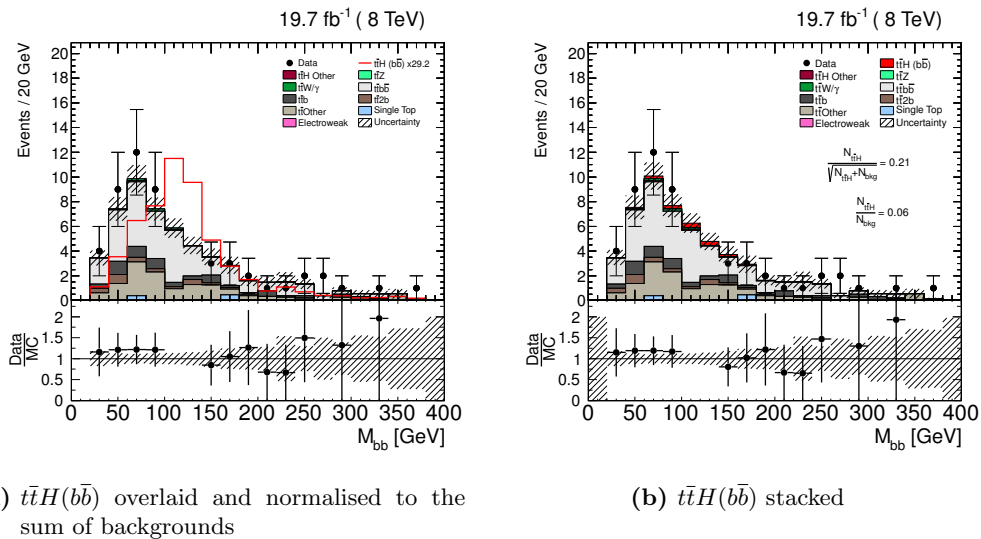
Figure 10.3: Dijet mass distribution for a pair of the two highest-p_{T} b-tagged jets that are not assumed to originate from the $t\bar{t}$ decay according to the MVA, as described in Section 9.6.1. The evaluated signal significance and the S/B are shown for the stacked plot. The signal region is blinded in data according to the request of the CMS Higgs group.

Considering the uncertainty on the background shape, which is caused by the limited statistics of the $t\bar{t} + jets$ sample, the number of background events, N_{bkg}, which is used for the calculation of signal significance and S/B, is increased by its statistical uncertainty. Thus, the evaluated signal significance and S/B ratio represent the potential lowest values, in order to avoid possible overestimation of the performance due to statistical fluctuations in MC simulations.

From the distributions it can be clearly seen that the shapes of the $t\bar{t}H(b\bar{b})$ signal and of its dominant backgrounds are different. While the $t\bar{t} + jets$ processes have a falling spectrum with the smeared kinematic edge closer to the jet-p_{T} threshold,

the $t\bar{t}H$ distribution has a clear peak around the Higgs-boson mass. The $t\bar{t}H$ contribution is about 30 times smaller than that of the background processes, leading to the signal significance of 0.21.

10.3.2   Cut on MVA output

A different approach for the suppression of combinatorial smearing was tried, which is based on the same MVA. But instead of having a single jet combination in every selected event, multiple or no pairs are allowed depending on whether they fulfil the selection requirement, which is schematically shown in Figure 10.4. Another difference from the previous method is that only pairs of b-tagged jets are considered. Thus, only b-tagged jets can be assigned to the $t\bar{t}$ system and to the $H \to b\bar{b}$ decay.
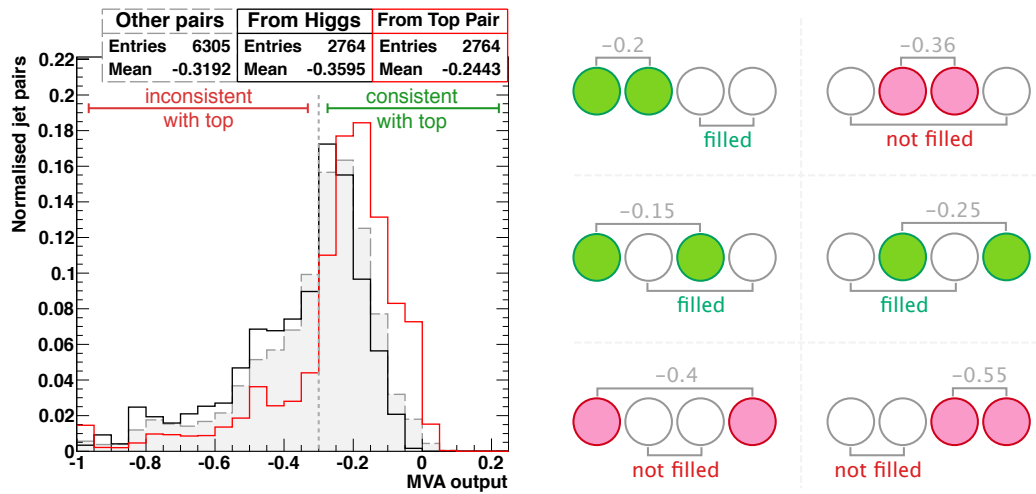
Figure 10.4: Schematic view of the jet-pair-selection principle based on the MVA output of the other jet pair in the event with 4 b-tagged jets. A threshold on the MVA output defines the region for jet pairs that are assumed to be consistent with the $t\bar{t}$ system. A pair of jets enters the invariant-mass distribution only if the pair of the other two jets is consistent with the $t\bar{t}$ system. Moving the threshold to larger values allows to suppress more pairs that do not represent the $H \to b\bar{b}$ decay. *From Top Pair* denotes the pair of b-tagged jets matched to the generated b-jets originating from the $t\bar{t}$ decay. *From Higgs* represents the pair of b-tagged jets matched to the generated b-jets originating from the $H \to b\bar{b}$ decay. *Other pairs* stands for all other pairs of b-tagged jets.

Three different thresholds on the MVA output of jet pairs consistent with the $t\bar{t}$ system were tested: -0.3, -0.2 and -0.1. The corresponding thresholds are shown for the $t\bar{t}H$ and $t\bar{t}b\bar{b}$ processes in Figure 10.5. The invariant-mass control distributions obtained with different MVA-output thresholds are shown in Figure 10.6.

Besides the dominant $t\bar{t} + jets$ background, the $t\bar{t}Z$ process becomes visible, which is mainly represented by $Z \to b\bar{b}$ decays. Considering the mass of the Z boson, which is close to the Higgs-boson mass, $m_Z = 91.19\,\text{GeV}$ [27], it is also an important background, which produces a peak close to the Higgs-boson signal region.

The control distributions show that the MVA-based jet-pair selection provides a better result in terms of sensitivity to the $t\bar{t}H$ signal, compared to the previ-

Other pairs		From Higgs		From Top Pair	
Entries	6305	Entries	2764	Entries	2764
Mean	-0.3192	Mean	-0.3595	Mean	-0.2443

Other pairs		From Top Pair	
Entries	5327	Entries	568
Mean	-0.3402	Mean	-0.2312
RMS	0.191	RMS	0.1438

(a) $t\bar{t}H(b\bar{b})$ events   **(b)** $t\bar{t}b\bar{b}$ events
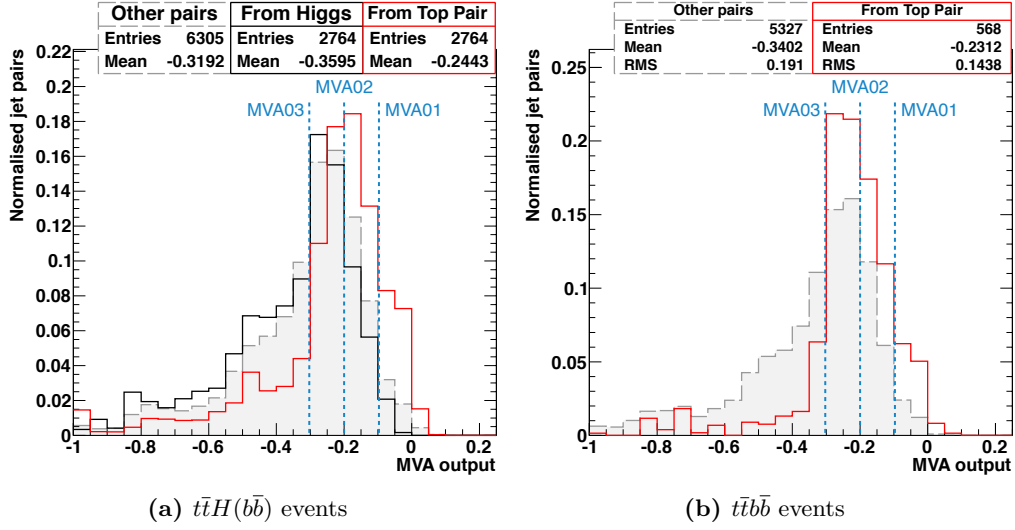
Figure 10.5: MVA-output distributions for different pairs of b jets in $t\bar{t}H(b\bar{b})$ and $t\bar{t}b\bar{b}$ processes, labelled as in Figure 10.4. Three different thresholds are shown by vertical dashed lines for jet-pairs that are assumed to be consistent with the $t\bar{t}$ system. Differences in shapes of the jet pairs from $t\bar{t}$ decays between the $t\bar{t}H(b\bar{b})$ and $t\bar{t}b\bar{b}$ processes allow to reduce the $t\bar{t}b\bar{b}$ background more than the $t\bar{t}H(b\bar{b})$ signal, in spite of the identical final states.

ously described selection of a single pair of leading b-tagged jets. By applying the $\mathrm{MVA}_{t\bar{t}} \geq -0.3$ threshold, almost the same S/B ratio is obtained as for the leading additional b-jet pair, while the signal significance is $50\,\%$ larger. Hence, the presence of multiple jet pairs per event effectively increases the statistical precision, and a larger fraction of b-jet pairs representing $H \to b\bar{b}$ decays are selected, which allows to better constrain the $t\bar{t}H$ contribution.

Reducing the number of used jet pairs by a tighter threshold, $\mathrm{MVA}_{t\bar{t}} \geq -0.1$, the S/B ratio is improved by about factor 2 with respect to the other studied approaches, and corresponds to about $1/9$. Thus, varying the MVA-output-threshold for $t\bar{t}$-consistent jet pairs, the sensitivity of the $t\bar{t}H$ search can be adjusted to the available amount of data.

10.4 Results

For the final sensitivity estimation, three different approaches are compared for the invariant-mass distribution:

- **MVA_leading:** the pair of additional b-tagged jets with highest p_T, as described in Section 10.3.1;

- **MVA03:** all pairs of b-tagged jets that satisfy the criteria $\mathrm{MVA}_{t\bar{t}} > -0.3$, as explained in Figure 10.4;

- **MVA01:** all pairs of b-tagged jets that satisfy the criteria $\mathrm{MVA}_{t\bar{t}} > -0.1$, as explained in Figure 10.4.

(a) $\text{MVA}_{t\bar{t}} \geq -0.3$

(b) $\text{MVA}_{t\bar{t}} \geq -0.2$

(c) $\text{MVA}_{t\bar{t}} \geq -0.1$
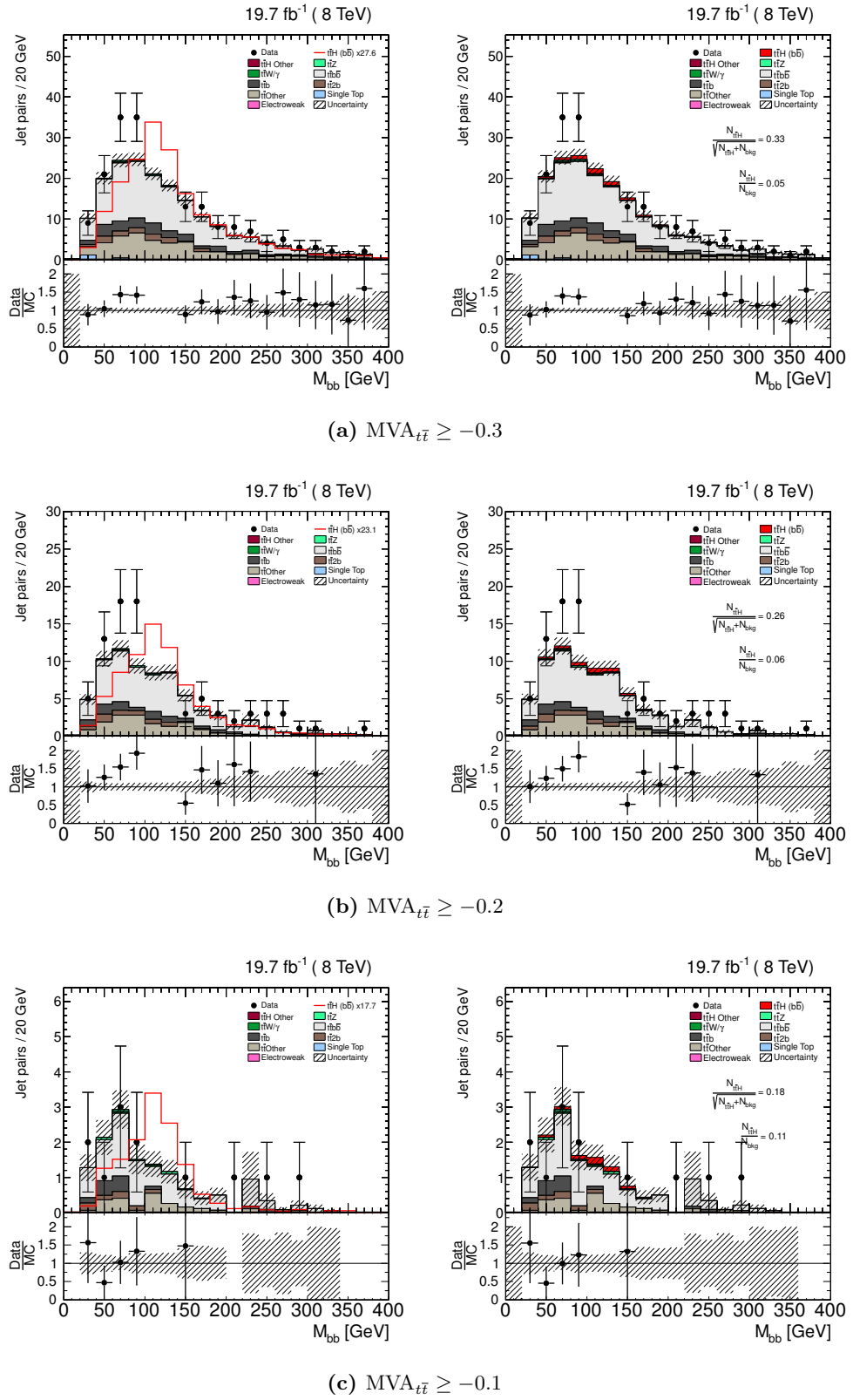
Figure 10.6: Invariant-mass distributions for jet pairs that fulfil each of the three different MVA-output thresholds, as schematically shown in Figure 10.4. On the left, the $t\bar{t}H(b\bar{b})$ contribution is drawn overlaid and normalised to the sum of the background processes. On the right, the $t\bar{t}H(b\bar{b})$ contribution is stacked, and shows also the calculated signal significance and S/B ratio.

Thus, the *MVA_leading* represents the simple approach with a single jet pair per event, while the *MVA03* and *MVA01* are optimised for higher signal significance and S/B ratio respectively.

10.4.1   Invariant-mass resolution

The first parameter that is relevant for the $t\bar{t}H$ search is the resolution of the Higgs-boson mass, which is reconstructed from the invariant mass of a jet pair. In addition to the three approaches mentioned above, the *reference resolution* is estimated, which is calculated from the invariant mass distribution of the two reconstructed jets that are matched to the two generated b jets originating from the $H \rightarrow b\bar{b}$ decay. This *reference resolution* represents the highest possible value that can be achieved with the CMS detector and all the applied corrections, if the two b jets from $H \rightarrow b\bar{b}$ decay were identified with 100 % efficiency.

The reference resolution is obtained from events that have two distinct jets at the reconstruction level that originate from the $H \rightarrow b\bar{b}$ decay. Such jets are first identified at the generator level using the $GenHFHadronMatcher$, and are matched to reconstruction-level jets by finding the closest in ΔR, using the requirements from (9.9).

The invariant-mass distribution for jet pairs obtained by three different approaches are compared for the $t\bar{t}H(b\bar{b})$ and $t\bar{t}b\bar{b}$ processes in Figure 10.7. Also the reference resolution is shown for the $t\bar{t}H$ process, which has real b jets from the $H \rightarrow b\bar{b}$ decay.

(a) $t\bar{t}H(b\bar{b})$ events   **(b)** $t\bar{t}b\bar{b}$ events
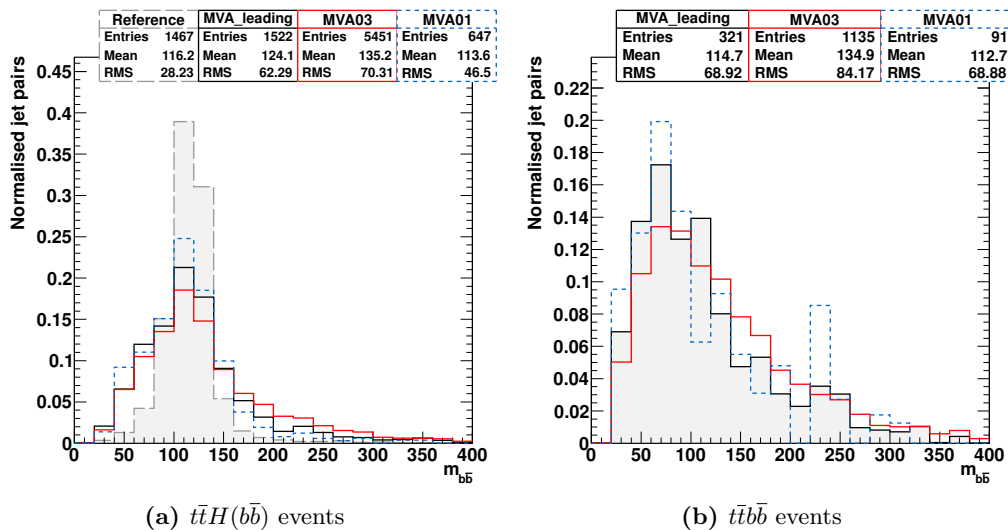
Figure 10.7: Distributions of the jet-pair invariant mass obtained by different pair-selection approaches. *Reference* is shown only for the $t\bar{t}H$ process, which has genuine a genuine pair of b jets representing the $H \rightarrow b\bar{b}$ decay.

It can be clearly seen from the plots that the most narrow peak is obtained by the reference pair selection, which represents the best possible mass resolution. Naturally, other methods, which do not use any generator-level information about the jet origin, provide a worse mass resolution. The *MVA_leading* approach has a smaller tail in the large-mass region than the *MVA03* approach, while the narrowest

peak is obtained by the *MVA01* method. In the $t\bar{t}b\bar{b}$ process, the *MVA01* and *MVA_leading* methods reflect the kinematic edge best, while the *MVA03* has a visible effect from combinatorial smearing, which leads to a significant contribution in the Higgs-mass region.

The position of the Higgs-mass peak and its resolution are evaluated by a fit in the range $80 < m_{b\bar{b}} < 160\,$GeV for each of the distributions in Figure 10.7a. The obtained values are summarised in Table 10.2, and show that the *MVA_leading* and *MVA01* methods provide the best and almost identical invariant-mass resolutions of about $26\,$GeV. Also both methods reproduce the same invariant mass, which is very close to the reference one.

Table 10.2: Mean values of the dijet mass and its resolution, as obtained from the fit of each distribution in Figure 10.7a by a gaussian.

Dijet selection method	Mean value [GeV]	Resolution [GeV]
• Reference	115	17
• MVA_leading	113	27
• MVA03	112	34
• MVA01	114	26

Fit performed in the range $80 < m_{b\bar{b}} < 160\,$GeV.

The difference in the shapes of the reconstructed dijet invariant mass spectra in the $t\bar{t}H(b\bar{b})$ and $t\bar{t}b\bar{b}$ processes proves the main idea of the MVA approach, which aims at the identification of b jets from the $t\bar{t}$ decay without significant bias towards the Higgs-boson mass. The difference in shapes of the $m_{b\bar{b}}$ spectra between the $t\bar{t}b\bar{b}$ and $t\bar{t}H(b\bar{b})$ process shown in Figures 10.6 and 10.3 supports this idea. This is an important strength of the developed method, which allows to achieve good separation between signal and background spectra.

Jet-energy-correction optimisation

While the *reference resolution* represents the best resolution that could be achieved with the current jet energy corrections, its mean value is $10\,$GeV lower than the Higgs-boson mass used in the simulation. This is a consequence of jet-energy corrections that are not optimal for b jets. Such corrections can potentially be improved, as was observed in dedicated studies [202].

The energy of all reconstruction-level jets that are used in the analysis is corrected as described in Section 7.4.1. The purpose of this correction is to make the energy of a reconstructed jet as close as possible to the energy of the corresponding generator-level jet. This is a generic correction, which does not have any flavour-specific treatment. Since, most of jets recorded by CMS are light-flavour, this correction is mainly designed for such jets.

On the other hand, b hadrons have about a $10\,\%$ probability to decay semileptonically: $B \rightarrow l\nu_l X$. Also b hadrons usually decay to c hadrons, which can also decay to a lepton and a neutrino: $B \rightarrow DX \rightarrow l\nu_l X$. This leads to about $35\,\%$ of b jets having non-optimal energy corrections due to the undetected neutrinos. This is

demonstrated by the ratio of reconstructed transverse momentum to the generated one (p_T^{reco}/p_T^{gen}) for b jets and other jets in Figure 10.8a. While non-b jets have a gaussian distribution with center at 1, for b jets it is 0.95, which means that on average the reconstructed p_T of a b jet is about 5% smaller than the true value.
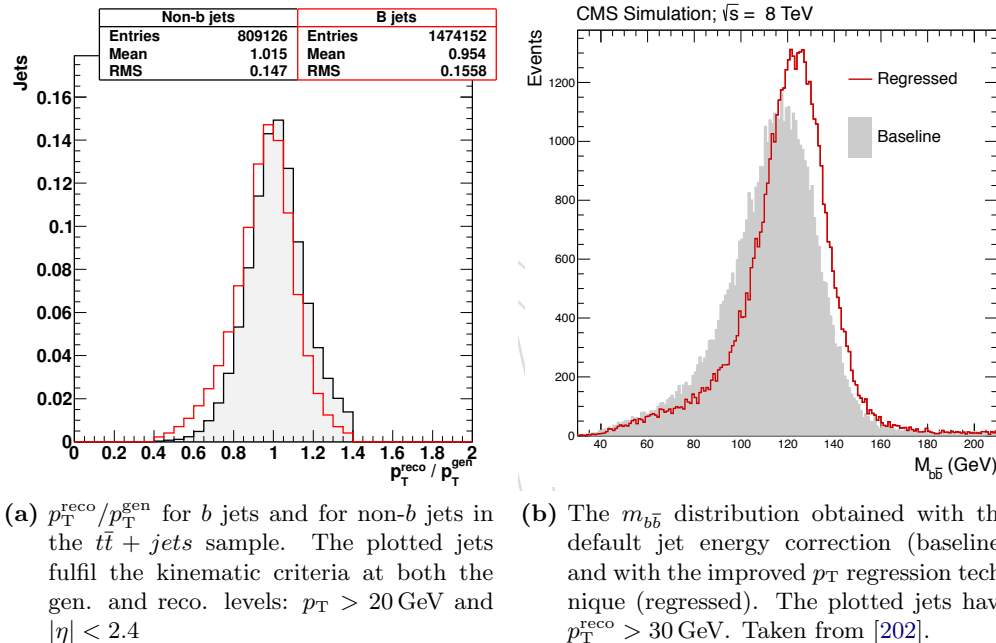
(a) p_T^{reco}/p_T^{gen} for b jets and for non-b jets in the $t\bar{t} + jets$ sample. The plotted jets fulfil the kinematic criteria at both the gen. and reco. levels: $p_T > 20\,\text{GeV}$ and $|\eta| < 2.4$

(b) The $m_{b\bar{b}}$ distribution obtained with the default jet energy correction (baseline) and with the improved p_T regression technique (regressed). The plotted jets have $p_T^{reco} > 30\,\text{GeV}$. Taken from [202].

Figure 10.8: Potential effect of the optimised correction of jet p_T as proposed in [202].

A special *jet-p_T regression* technique was developed at CMS specifically to improve the Higgs-boson-mass resolution in $H \to b\bar{b}$ decays [202]. This method uses a multivariate approach to fine-tune the jet-p_T calibration in many relevant jet properties, which automatically addresses the problem of semi-leptonic b-hadron decays. A deeper explanation of the technique can be found in [202, 254, 255].

The potential effect of the improved jet-energy correction on the reconstructed $m_{b\bar{b}}$ spectrum is demonstrated in Figure 10.8b. The shift of the $m_{b\bar{b}}$ peak to the higher values would correct the mean values listed in Table 10.2. It can also be seen that the width of the regressed spectrum is reduced, which leads to a better invariant-mass resolution. This effect can improve the sensitivity by allowing to use a smaller range of $m_{b\bar{b}}$ for the $t\bar{t}H$ search, which would effectively reduce the amount of background contributions.

10.4.2   Expected sensitivity

The obtained signal significance and S/B ratio for each of the compared jet-pair-selection methods allow to make assumptions about the amount of data required to reach a certain significance, e.g. 2σ for a visible $m_{b\bar{b}}$ contribution from the $t\bar{t}H(b\bar{b})$ production in the dileptonic $t\bar{t}$ decay channel.

Obviously, such assumptions allow only approximations at this stage, since they do not include possible systematic variations of the background and signal shapes. Also the statistical precision of the available $t\bar{t} + jets$ sample is not high enough to

make firm conclusions about the background contribution, especially for the *MVA01* approach, which implies very strict requirements for the potential *b*-tagged-jet pairs. Nevertheless, it allows to make a rough estimate of the potential of the developed methods and of their relevance for future analyses.

Summarising the numbers from Figures 10.3 and 10.6, the signal significance and S/B ratio for each of the compared jet-pair-selection methods are presented in Table 10.3. The *MVA_leading* approach is most straightforward, but provides neither the highest signal significance nor best S/B ratio. Among the other approaches with a variable threshold on jet-pair selection, applying a tighter selection reduces the signal significance and improves the S/B ratio. Thus, the *MVA01* provides the best sensitivity for an infinite amount of data, and the *MVA03* allows to reach the highest signal significance with the available amount of data recorded during 2012.

Table 10.3: Signal significance ($\mathcal{S}_{ign}$) and the signal-to-background ratio (S/B) for the compared jet-pair-selection methods.

Dijet-selection method	Signal significance	S/B ratio
• MVA_leading	0.21σ	1/17
• MVA03	0.33σ	1/20
• MVA01	0.18σ	1/9

The numbers are evaluated in the range $100 < m_{b\bar{b}} < 140\,\text{GeV}$, using the equations (10.2) and (10.3).

Expected evidence at $\sqrt{s} = 8\,\text{TeV}$

Since the numbers listed in Table 10.3 were obtained with the data collected at $\sqrt{s} = 8\,\text{TeV}$, the cross sections for the $t\bar{t}H(b\bar{b})$ and background processes remain unchanged. The signal significance is a measure of the size of the signal contribution compared to the statistical uncertainty of the measured data. Thus, with a constant S/B ratio, but with a factor K larger amount of data, the expected signal significance will improve, and can be expressed as:

$$\mathcal{S}_{ign}^* = \frac{K}{\sqrt{K}} \cdot \frac{N_{t\bar{t}H}}{N_{t\bar{t}H} + N_{\text{bkg}}} = \sqrt{K} \cdot \mathcal{S}_{ign} \,. \tag{10.4}$$

Consequently, the desired factor K required to obtain an arbitrary signal significance $\mathcal{S}_{ign}^*$ can be calculated as:

$$K = \left(\frac{\mathcal{S}_{ign}^*}{\mathcal{S}_{ign}} \right)^2 \,. \tag{10.5}$$

The estimated luminosity required for a 2σ contribution of the $t\bar{t}H(b\bar{b})$ production in the dileptonic channel is listed for the described methods in Table 10.4. It can be seen from the numbers, that about 40–120 times more data would be required at $\sqrt{s} = 8\,\text{TeV}$ for an evidence of the $t\bar{t}H(b\bar{b})$ process depending on the MVA-output threshold, if systematic uncertainties were neglected.

Table 10.4: Expected luminosity required for a contribution of the $t\bar{t}H(b\bar{b})$ process in the dileptonic $t\bar{t}$ decay channel at $\sqrt{s} = 8\,\text{TeV}$ with the significance of 2σ, if more data would be recorded.

Dijet-selection method	Expected luminosity [fb^{-1}]
• MVA_leading	1800
• MVA03	750
• MVA01	2450

Expected evidence at $\sqrt{s} = 14\,\text{TeV}$

When the LHC will operate at its design energy, with proton beams colliding at $\sqrt{s} = 14\,\text{TeV}$, the cross sections of both the $t\bar{t}H(b\bar{b})$ and $t\bar{t}b\bar{b}$ processes will increase. According to the theoretical predictions of $t\bar{t}$ cross section calculated at NNLO+NNLL accuracy [256] for $\sqrt{s} = 13\,\text{TeV}$ and of the $t\bar{t}b\bar{b}$ cross section at NLO accuracy for $\sqrt{s} = 14\,\text{TeV}$ [257], the $t\bar{t}b\bar{b}$ cross section will increase by a factor 5.5 at $\sqrt{s} = 14\,\text{TeV}$ compared to the one at $\sqrt{s} = 8\,\text{TeV}$. This factor was evaluated for the additional-b-jet p_T threshold of 20 GeV and top-quark mass $m_t = 173.5\,\text{GeV}$.

The inclusive $t\bar{t}H$ cross section will increase by a slightly smaller amount, by a factor 4.7, from 0.13 pb to 0.61 pb, according to the NLO calculations at $\sqrt{s} = 8\,\text{TeV}$ [114] and at $\sqrt{s} = 14\,\text{TeV}$ [258] for the $m_H = 125.0\,\text{GeV}$. But the $p_\text{T} > 20\,\text{GeV}$ threshold on b jets from the $H \to b\bar{b}$ decays would make this ratio larger, which is assumed to be very close to the one for the $t\bar{t}b\bar{b}$ cross section. Therefore, both the $t\bar{t}b\bar{b}$ and $t\bar{t}H(b\bar{b})$ cross sections are taken to increase by the same factor 5.5 at $\sqrt{s} = 14\,\text{TeV}$.

The increase of cross sections of other background processes is assumed to be the same, although it will be smaller for most of the processes. This simplification is acceptable considering their small contribution, and the obtained result will be more conservative. Thus, the integrated luminosity $\mathcal{L}^*$ required to observe the $t\bar{t}H(b\bar{b})$ contribution with a certain significance $\mathcal{S}^*_{ign}$ at $\sqrt{s} = 14\,\text{TeV}$ can be calculated from the significance $\mathcal{S}_{ign}$ estimated with $19.7\,\text{fb}^{-1}$ of data recorded at $\sqrt{s} = 8\,\text{TeV}$:

$$\mathcal{L}^* = \frac{19.7\,\text{fb}^{-1}}{5.5} \cdot \left(\frac{\mathcal{S}^*_{ign}}{\mathcal{S}_{ign}} \right)^2 . \tag{10.6}$$

The required luminosity for a 2σ contribution of the $t\bar{t}H(b\bar{b})$ process is estimated for each of the methods using the equation (10.6), and is listed in Table 10.5.

Table 10.5: Expected luminosity required for a detection of the $t\bar{t}H(b\bar{b})$ process in the dileptonic $t\bar{t}$ decay channel at $\sqrt{s} = 14\,\text{TeV}$ with the significance of 2σ.

Dijet-selection method	Expected luminosity [fb^{-1}]
• MVA_leading	330
• MVA03	130
• MVA01	450

The projection of the sensitivity to the $t\bar{t}H(b\bar{b})$ signal for the future run of the LHC at $\sqrt{s} = 14\,\text{TeV}$ shows that using the loose jet-pair selection the observation of a 2σ contribution of the $t\bar{t}H$ production can be achieved with just about $130\,\text{fb}^{-1}$ of data. This amount of data might be available even before the Longs Shutdown 2 (LS2) at $\sqrt{s} = 13\,\text{TeV}$, according to the LHC Run2 schedule [259]. With a significantly larger amount of data, corresponding to the integrated luminosity of about $450\,\text{fb}^{-1}$, a mass peak from $H \to b\bar{b}$ decays can be observed with the very high S/B ratio of $1/9$. These numbers might still be improved with a potential increase of sensitivity, as described below.

Potential improvement of sensitivity

Besides the improved jet-energy calibration, discussed in Section 10.4.1, a conceptual difference in colour flow in the $t\bar{t}b\bar{b}$ and $t\bar{t}H(b\bar{b})$ processes might be used to further suppress the $t\bar{t}b\bar{b}$ background. The fact that the Higgs boson is a colour singlet leads at the leading order to a colour connection only between its $H \to b\bar{b}$ decay products. But no colour-connection is present between the b quarks from the $t\bar{t}$ decay and b quarks from the Higgs-boson decay. In contrast, in a $t\bar{t}b\bar{b}$ event, b quarks from the $t\bar{t}$ decay are colour-connected with each other and with initial state radiation. Experimentally this could be observed as an absence of hadronic activity in the corresponding parts of the phase space for $t\bar{t}H$. Oppositely, in the $t\bar{t}b\bar{b}$ process additional b quarks are colour-connected to b quarks from the $t\bar{t}$ decay, leading to a different pattern of hadronic activity in the event. However such observable would be distorted by ISR and FSR, pile-up events, etc., requiring a careful study of such an approach.

This feature could be used as an additional event-selection criterion that would differentiate between the $t\bar{t}b\bar{b}$ and $t\bar{t}H(b\bar{b})$ processes, although they have identical final states. Furthermore, it can be used within a $t\bar{t}H(b\bar{b})$ event to reduce the number of candidate jet pairs originating from the $t\bar{t}$ system. An example of an observable that reflects the colour flow in an event is discussed in the phenomenological study [260].

The estimated sensitivity at $\sqrt{s} = 14\,\text{TeV}$ might be better in reality even without any dedicated technical improvements, if the detector resolution does not degrade significantly during the future runs of the LHC at higher $\sqrt{s}$. Considering the higher average energy of particles produced in $t\bar{t}H$ events at the higher centre-of-mass energy, the event-selection efficiency might increase, leading to a larger amount of selected events and therefore to a better signal significance.

10.5 Comparison to existing measurements

A number of $t\bar{t}H$ searches in different final states were performed by the CMS [261, 262] and ATLAS [263, 264] collaborations using the data recorded at $\sqrt{s} = 7\,\text{TeV}$ and $\sqrt{s} = 8\,\text{TeV}$. Since the amount of recorded data is still too small for a measurement of the $t\bar{t}H$ cross section, only upper limits on the cross section are usually presented for a quoted value of the Higgs-boson mass.

Combined decay channels of the Higgs boson

A search targeting the widest range of final states was performed by the CMS Collaboration in the multivariate analysis [261], which includes decays of the Higgs boson to hadrons, leptons and photons. The $H \rightarrow b\bar{b}$ decay channel is a part of the hadronic category and is analysed separately with an optimised analysis approach. In this measurement not only dileptonic but also semi-leptonic decays of the $t\bar{t}$ system are used, which increases the available statistics considerably. Furthermore, events with multilepton and diphoton final states of the Higgs boson further increase the sensitivity to the $t\bar{t}H$ signal. For the $H \rightarrow \gamma\gamma$ and $H \rightarrow b\bar{b}$ channels, additional $5 \, \text{fb}^{-1}$ of data are used from the 2011 LHC run at $\sqrt{s} = 7 \, \text{TeV}$. The estimated expected 95% CL upper limit on the $t\bar{t}H(b\bar{b})$ cross section is 5.0 times the Standard Model prediction for the $m_H = 125.6 \, \text{GeV}$. The limit is significantly improved by adding all the other decay channels of the Higgs boson, leading to the upper limit on the $t\bar{t}H$ cross section of 2.7 times the Standard Model prediction.

$t\bar{t}H(b\bar{b})$ searches with $H \rightarrow b\bar{b}$ decays

Also dedicated searches for the $t\bar{t}H$ production with $H \rightarrow b\bar{b}$ decays were performed by the CMS [262] and ATLAS [263] experiments. Both analyses employ the so-called *matrix element method* [265], and use both semileptonic and dileptonic decay channels of the $t\bar{t}$ system. The amount of data used by the two analyses is very close, and ranges from $19.5 \, \text{fb}^{-1}$ to $20.3 \, \text{fb}^{-1}$. Some differences in the analyses lead to a significant difference between the achieved sensitivities. Thus, the CMS measurement [262] has set the upper limit of 7.8 times the SM for the dilepton channel alone and 4.1 times the SM for the semileptonic and dileptonic channels combined. The ATLAS measurement [263] has constrained the $t\bar{t}H(b\bar{b})$ cross section much better by measuring the upper limit of 4.7 times the SM for the dilepton channel alone and 3.1 times the SM for the combined semileptonic and dileptonic decay channels.

$t\bar{t}H$ search with $H \rightarrow \gamma\gamma$ decays

A search for $t\bar{t}H$ production with two photons in the final state was also performed by ATLAS [264]. This analysis uses $4.5 \, \text{fb}^{-1}$ of data recorded at $\sqrt{s} = 7 \, \text{TeV}$ and $20.3 \, \text{fb}^{-1}$ recorded at $\sqrt{s} = 8 \, \text{TeV}$. Due to the small branching fraction of the $H \rightarrow \gamma\gamma$ decay, all decay channels of the $t\bar{t}$ system were considered in the analysis to gain statistics. The estimated expected limit on the $t\bar{t}H(\gamma\gamma)$ cross section is 4.9 times the standard model.

Upper limits of the presented analysis

The results of the current analysis were presented in terms of expected signal significance. They can be easily converted to expected upper limits on the $t\bar{t}H(b\bar{b})$ cross section. The signal significance (10.2) represents the ratio between the expected size of the signal contribution and the statistical uncertainty on the total number of signal and background entries.

Thus, the inverse of this quantity would represent the largest amount of $t\bar{t}H$ signal that would be consistent with the expected uncertainty in the data. In order

to obtain the upper limit at the 95 % confidence level, a 1.96σ deviation from data should be allowed, which leads to the following simplified equation for the expected upper limit (UL):

$$\mathrm{UL} = 1.96 \cdot \frac{1}{\mathcal{S}_{ign}} \,. \tag{10.7}$$

This is a very simplified calculation, but it allows to make a rough estimate of the performance for comparison to other results. The values of expected upper limits on the Standard Model $t\bar{t}H(b\bar{b})$ cross section for different jet-pair-selection methods are listed in Table 10.6.

Table 10.6: Expected 95 % CL upper limits on the SM $t\bar{t}H(b\bar{b})$ cross section estimated with a simplified approach on data corresponding to the integrated luminosity of $19.7\,\mathrm{fb}^{-1}$ recorded at $\sqrt{s} = 8\,\mathrm{TeV}$.

Dijet-selection method	Expected upper limit
• MVA_leading	9.3
• MVA03	6.0
• MVA01	11.0

The numbers are propagated from signal-significance values listed in Table 10.3, which were calculated in the range $100 < m_{b\bar{b}} < 140\,\mathrm{GeV}$.

Thus, the best limit is expected to be obtained by the *MVA03* approach, which provides the highest signal significance. This performance is comparable to the one provided by the matrix-element method used in the $t\bar{t}H(b\bar{b})$ search [262] with the dileptonic channel of the $t\bar{t}$ system. Nevertheless, this analysis also uses semileptonic final states, which improves its expected upper limit from 7.8 to 4.1 times the SM.

Other analyses perform even better thanks to a wider range of considered final states and, consequently, better statistical precision. Especially the low expected limit achieved by the matrix-element method of the ATLAS measurement [263] in the $H \rightarrow b\bar{b}$ channel alone is largely caused by a better-performing b-tagging algorithm and an optimisation of the event-classification MVA technique for events with different jet-flavour compositions.

Nevertheless, with the restricted amount of data from dileptonic $t\bar{t}H(b\bar{b})$ events, the developed method can provide a competitive sensitivity.

10.6 Conclusions

The technique developed for the identification of b jets from the $t\bar{t}$ system using the MVA approach allows to optimise the selection of the proper jet pair among 6 possible combinations. The correction of additional b-jet activity in the $t\bar{t} + jets$ sample allowed to optimise and check the description of the dominant background to $t\bar{t}H(b\bar{b})$ production.

Besides the simple selection of a pair of highest-p_T additional b-tagged jets, a more sophisticated algorithm for the jet-pair selection was implemented and tested. Using a variable cut on the MVA output, it allows to adjust the fraction of b-jet pairs

representing $H \rightarrow b\bar{b}$ decays in order to obtain the best performance with available data. Furthermore, this method allows to suppress contributions from processes that have an identical final state, e.g. $t\bar{t}b\bar{b}$, thanks to different shapes of the MVA output in different processes.

Three different approaches for jet-pair selection were compared and the obtained invariant mass resolution was estimated. The best Higgs-mass resolution is 26 GeV, which is larger than the best that could be achieved with perfect jet-pair identification by about 10 GeV. Optimised jet energy corrections could improve the resolution further by recovering undetected neutrinos from semileptonic b-hadron decays.

The performance of the different jet-pair-selection methods was evaluated in terms of the expected signal significance and signal-to-background ratio neglecting certain effects. The obtained values allowed to estimate the amount of data required for a 2σ observation of the $t\bar{t}H(b\bar{b})$ process with the existing analysis strategy. Also an estimate for the future run of the LHC at $\sqrt{s} = 14$ TeV was made, which showed that with the current performance and with the expected changes of the $t\bar{t}H$ and $t\bar{t}b\bar{b}$ cross sections, about $130\,\mathrm{fb}^{-1}$ to $450\,\mathrm{fb}^{-1}$ of data would be required for a 2σ observation of the $H \rightarrow b\bar{b}$ mass peak. The data required for the observation with the loose jet-pair-selection might be available even before Long Shutdown 2.

With the tightest selection of b-jet pairs, the $t\bar{t}Z(b\bar{b})$ contribution has become visible. This has a lower cross section than that of the $t\bar{t}H(b\bar{b})$ process, since the rate of Higgs-boson production in association with a pair of top quarks is enhanced by the large coupling of the Higgs boson to the very massive top quark. Nevertheless, the mass of the Z boson is only 35 GeV lower than that of the Higgs boson, which leads to a large $t\bar{t}Z$ contribution in the $t\bar{t}H$ signal region. Therefore, in future analyses with higher statistical precision, this will be a relevant background contribution, and the shape of its $m_{b\bar{b}}$ spectrum has to be known.

In comparison to other $t\bar{t}H$ searches performed by the CMS and ATLAS collaborations, the developed method provides a comparable performance. Considering its simplicity and flexibility, the developed method can be integrated into a larger-scale analysis as a specific part for dileptonic final states of the $t\bar{t}$ system. It can also be preferable for early analyses aiming at first data from the LHC at $\sqrt{s} = 13$ TeV and $\sqrt{s} = 14$ TeV, since its implementation is more straightforward compared to other approaches, which would allow to obtain results faster.

The main strength of the developed strategy is that it is the only one that uses a pure physical quantity for the $t\bar{t}H$ signal extraction, the invariant mass of two b jets. Thus, the developed approach allows to see the $t\bar{t}H$ signal by eye, as a peak in the $m_{b\bar{b}}$ distribution, much in the same way as the Higgs boson was discovered. It is not aiming at highest sensitivity as other measurements do, which use multiple different final states or very clean signatures. Since the Higgs boson is already discovered, after an evidence of $t\bar{t}H$ process by other measurements, the developed method would provide a very important complementary result, which would have a much more straightforward interpretation.

Chapter 11

Summary and conclusions

All the four distinct contributions to the CMS physics programme that were addressed in this thesis, have been successfully completed using data recorded by CMS at $\sqrt{s} = 8\,\text{TeV}$. Two of them are technical, while the other two are physics analyses.

The new calibration method of Lorentz-angle and backplane correction, simultaneously with the track-based alignment of the CMS tracker, was studied in the framework of MILLEPEDE II method. As a result of this study, a new configuration of the full-scale tracker alignment was prepared, which was used as a basis to obtain the best tracker geometry using tracks recorded during 2012. Furthermore, the integrated-calibration methods will be used during the future runs of the LHC at $\sqrt{s} = 13\,\text{TeV}$ and $\sqrt{s} = 14\,\text{TeV}$. In particular, already in 2012, the official backplane corrections were obtained with this method instead of a standalone calibration. The calibrated values of Lorentz angle showed a clear improvement of the hit-position resolution in the pixel detector, which is also more stable in time compared to the results of standalone calibrations. Significant differences in time-dependence of the Lorentz angle between the integrated and standalone methods, as well as dependence on spatial granularity in the microstrip detector suggest that the developed configuration can still be improved to achieve even better performance. During the future runs of the LHC, the radiation damage of tracker modules will significantly increase due a higher instantaneous luminosity. This makes the integrated alignment and calibration approach a valuable contribution to the tracker performance, which is relevant for nearly any physics analysis.

The method to identify the origin of any heavy-flavour jet in MC simulations was developed from scratch and showed excellent performance for b jets in simulations produced using different parton-shower generators, which could not be achieved by tools existing in CMS before. This was a crucial component of the physics analyses described in this thesis, which are characterised by the presence of a large number of b jets and need a clear distinction between them. Due to the high performance and versatility of the developed method, and due to its importance for other analyses, it was integrated into the CMS software framework and used as a basis of b-jet definitions for future analyses.

The measured differential cross sections of the $t\bar{t}b(\bar{b})$ and $t\bar{t}b\bar{b}$ processes in the dileptonic final state of the $t\bar{t}$ pair showed that the amount of additional b-jet radiation is underestimated by leading-order simulations. The obtained results are in a good agreement with the previous measurement of inclusive cross section, not

only qualitatively but also quantitatively. The shapes of the measured spectra are in general well described by LO theoretical predictions. The only exception is the angular distance between the two additional b jets ($\Delta R_{b\bar{b}}$), which shows that in LO predictions, additional b jets tend to be produced more closely to each other than they appear in data. This might be explained by an underestimation of double quasi-collinear gluon splitting to $b\bar{b}$ pairs, which can not be resolved at jet level, producing two largely separated b jets. Theoretical calculations at NLO accuracy show that the contribution from such events can be substantial. The $m_{b\bar{b}}$ spectrum of additional b jets is well described by all studied MC simulations, which makes them a good option for the background estimation in $t\bar{t}H$ measurements. With the available amount of data, the precision of the measurements is limited by statistical uncertainties, which are about 30% for the $t\bar{t}b(\bar{b})$ cross sections and about 50% for the $t\bar{t}b\bar{b}$ cross sections. Thus, a better statistical precision obtained with future data from the LHC is very much anticipated for such a measurement. Theoretical calculations of the studied processes at least at NLO accuracy will be especially interesting for a comparison to the $\Delta R_{b\bar{b}}$ and p_{T} spectra in the data.

Finally, the good description of the $m_{b\bar{b}}$ shape by MC predictions allowed to make a simulation-based estimate of the sensitivity to the $t\bar{t}H(b\bar{b})$ process, neglecting systematic effects. The key idea of the studied approach is to see a peak around the Higgs-boson mass in the $m_{b\bar{b}}$ distribution. The multivariate approach for identification of b jets originating from the $t\bar{t}$ decay, which was also used in the presented $t\bar{t}b\bar{b}$ cross-section measurement, allowed to obtain clearly distinctive spectra in $t\bar{t}b\bar{b}$ and $t\bar{t}H(b\bar{b})$ processes. The difference between signal and background shapes of a clear physical quantity is the main strength of the studied method compared to other approaches that use complicated multivariate quantities for such a distinction. The best invariant-mass resolution that was achieved for the Higgs boson was $26\,\mathrm{GeV}$. Based on the available data, it was estimated that about $130\,\mathrm{fb}^{-1}$ of data recorded at $\sqrt{s} = 14\,\mathrm{TeV}$ should be sufficient to see a mass peak from the $H \to b\bar{b}$ decays with a significance of 2σ. This amount of data is expected to be available before the Long Shutdown 2. At this level of precision, contributions from the $t\bar{t}Z(b\bar{b})$ process will also become relevant, because they will produce a Higgs-like peak around $m_Z = 91.2\,\mathrm{GeV}$, which requires a good control of their shape.

Of course, each of the results could be further improved if time did not pass so fast or if the knowledge obtained during the years of work was there from the very beginning. Most of the studies are complemented with ideas about possible ways of their improvement, and provide a good starting point for new, more advanced and precise analyses, taking advantage of significantly larger amount of data that will be available in the future.

Appendix A

Details of the combined tracker alignment and calibration

A.1 Alignment algorithms at CMS

The three alignment methods that are implemented in CMS software framework are:

- **Hit and Impact Points algorithm (HIP)** [266] separately computes global parameters for each individual alignable structure in an iterative approach. Therefore, all correlations between alignment parameters are ignored in this method. In each consecutive iteration, track parameters are updated based on global parameters from the last iteration. Due to the neglection of correlations between global parameters, it is not possible to properly estimate uncertainties of the global parameters. This method is well suited for early stages of data taking after significant changes in the tracker geometry, when the amount of available tracks for alignment is limited and the starting geometry is significantly different from the true one. In particular, after the restart of the LHC at $\sqrt{s} = 13\,\mathrm{TeV}$, the HIP algorithm might be used to obtain the first alignment of the upgraded tracker geometry in 2015, when the amount of tracks recorded with the upgraded detector will be limited.

- **Kalman Filter algorithm** [267] is based on the Kalman-filter track fit [268]. In this approach, all global parameters are updated after adding each track. This method includes correlations between alignment parameters, which allows to update parameters of structures that were not even crossed by the added track. Nevertheless, technical limitations of available memory and CPU power do not allow to update all parameters. Therefore, only parameters that have significant correlations with the crossed alignable structures are updated.

- **Millepede II algorithm** [269] minimises the sum of the normalised squared track-hit residuals for all tracks at once. Solving the system of equations (5.4) for all tracks requires a large amount of memory and time. Therefore, the computation of improved track parameters is skipped and special methods are used to reduce the system of equations, as briefly described in Section 5.3.2. Nevertheless, the updated track parameters can be taken into account in the next iteration of the MILLEPEDE II alignment. Usually, up to 3-5 iterations

247

can be required to reach the optimal global parameters, which do not change significantly after further iterations. Furthermore, a special MINRES method is used to solve matrix equations in an iterative approach, without the need of ivnersion of the whole matrix. In order to optimise the computing-resources requirements, it takes advantage of a usually sparse structure of the matrix $\mathbf{C}$ and stores only non-zero elements. For matrix elements that sum contributions from only a few tracks, single precision storage is used, which allows to further reduce the memory consumption.

A.2   Quality criteria of tracks used in alignment

Tracks and their constituent hits have to fulfil the following quality criteria, which may differ between the pixel and strip regions of the tracker:

- the signal-over-noise ratio of the strip hits must be higher than 12 (18) in the deconvolution (peak) mode (see Section 4.3.2);

- the probability of the pixel hit to match the template shape [270] must be higher than 0.001 (0.01) in the u (v) direction;

- the angle between the track and the module surface must be larger than 10° (20°) for tracks from pp collisions (from cosmis rays);

- a track must contain at least two hits on pixel or strip stereo modules, to ensure a reliable determination of the polar track angle θ; the hits must fulfil the requirements mentioned above;

- tracks from pp collisions must fulfil the *high purity* quality criteria [161] of the CMS track-reconstruction code;

- hits that have a normalised track-hit residual above certain threshold are identified as *outlier hits*, and are removed from the fit;

- in the final track fit within MILLEPEDE II, tracks are rejected if their χ^2 value is larger than the 3σ deviation of the χ^2 distribution for the number of degrees of freedom of the track.

In order to optimally select tracks of different topologies, more specific quality criteria are applied in addition to the ones mentioned above:

- **Isolated muons:** *global muons* (see Section 7.3.1), with hits in both the tracker and the muon system, that fulfil all of the following quality criteria:

  1. at least 10 tracker hits, out of which at least 1 is in the pixel detector;
  2. transverse momentum $p_\mathrm{T} > 5\,\mathrm{GeV}/c$;
  3. distance from the axis of the closest jet $\Delta R > 0.1$ (see (4.4)).

- **Minimum bias:** events are selected with a combination of triggers based on signals indicating the crossing of two filled proton bunches, beam scintillator counters or moderate requirements on hit and track multiplicity in the pixel detector. Such tracks must fulfil the following requirements:

1. at least 8 tracker hits;

2. momentum $p > 8\,\text{GeV}/c$;

- **$Z^0 \to \mu^+ \mu^-$ decays:** events are requested to have at least two *global* muons of opposite charge. The two corresponding tracks must fulfil the following criteria:

  1. at least 10 tracker hits, out of which at least 1 is in the pixel detector;

  2. transverse momentum $p_{\text{T}} > 15\,\text{GeV}/c$;

  3. the invariant mass of the reconstructed dimuon system $85.8 < M_{\mu^+\mu^-} < 95.8\,\text{GeV}/c^2$;

  These track pairs are reparameterized with 9 parameters (and those for multiple scattering). A virtual mass measurement of $M_{\mu^+\mu^-} = 90.86 \pm 1.86\,\text{GeV}/c^2$ is added for each such track pair.

- **Cosmic rays:** events are recorded with the strip tracker operating in the peak or deconvolution readout mode. Data in peak mode were recorded during dedicated periods for cosmic data taking, before the restart of the LHC and when no beams are in the accelerator. Cosmic tracks must fulfil the following criteria:

  1. at least 8 tracker hits;

  2. momentum $p > 4\,\text{GeV}/c$;

A.3   Selection of high-quality $0\,\text{T}$ collision tracks

Due to some problems related to the superconducting solenoid, during a short period of time magnetic field in the tracker was reset, with the $|\mathbf{B}|$ decreasing from $3.8\,\text{T}$ to $0\,\text{T}$, and then ramping up back to the designed $3.8\,\text{T}$. Since this happened during pp-collision data taking, a rather unique set of about 200×10^6 tracks was recorded, originating from the collision point but without a magnetic field. This type of tracks was never used in tracker alignment before, and a dedicated study was required to select tracks of a good quality from this amount.

Events from this period were recorded with minimum-bias requirements, and stored in 4 datasets: `/ZeroBias[1-4]/Run2012C-TkAlMinBias-v2/ALCARECO` [271]. In contrast to cosmic muons, $0\,\text{T}$ collision tracks were recorded under unintended and non-optimised conditions, which required a close look at their properties and comparison to minimum-bias tracks recorded at $3.8\,\text{T}$, in order to find the optimal average-momentum estimation for such tracks. Thus, a series of alignments was performed with different momentum estimations for the $0\,\text{T}$ collision tracks ranging from $0.1\,\text{GeV}$ to $5\,\text{GeV}$, to compare the fractions of tracks rejected after refit and average χ^2 of selected tracks with the values of from $3.8\,\text{T}$ minimum-bias tracks. The momentum estimate of 3 GeV was found to provide the closest performance to the one of $3.8\,\text{T}$ tracks, and was chosen as a default value for these tracks.

Furthermore, track-selection criteria had to be optimised for such tracks. The basic idea of track-selection optimisation was to have many hits in the tracker, which would all lie on the same line, ensuring that the track is fitted well. The

largest gain from this dataset is expected to be in FPIX, TID and TEC, since very few cosmic tracks cross the detector in such forward directions. Thus, setting the threshold too high would suppress forward tracks due to the lack of enough layers in the forward direction (see Figure 4.15). Furthermore, the tracks were required to have their last hit either at large r or at large z, which, in combination with the requirement of a pixel hit, ensures that the track penetrates the whole detector to its outer edge. The selection of tracks based on the last-hit position was added to the `AlignmentTrackSelector` class of the `CommonAlignmentProducer` package for this purpose. Finally, to reject possible fake tracks, they are selected only if no other tracks are close to them and if they are not displaced far from primary vertices with many degrees of freedom. Thus, the 0 T collision tracks have to fulfil the following special requirements in addition to the ones listed for minimum-bias tracks in Section 5.3.3:

- at least 10 tracker hits, out of which at least 2 in the pixel detector;

- at least 3 hits should provide 2D measurements, either in pixel modules or stereo modules of the microstrip detector;

- distance of closest approach from a good primary vertex in transverse (longitudinal) direction smaller than 7 cm (18 cm);

- position of the last hit should be either at $r > 90$ cm or $|z| > 220$ cm;

Furthermore, tracks are selected only from events that have at least one good primary vertex, which should be constructed from at least 4 tracks and be displaced form the beamspot no more than by 2 cm (15 cm) in r (z) direction.

The adding of 0 T collision tracks to the alignment procedure was found to better constrain absolute values of μ_{H}. As can be seen from Figures A.1 and A.2, with added 3.8 T collision tracks, shapes of Lorentz-angle evolution in BPIX and FPIX do not change, but absolute values for neighbouring groups of modules are closer, which is more intuitive considering similar operating conditions of these modules.

A.4   Millepede II alignment sequence

The name MILLEPEDE II consists of the two distinct parts: *mille* and *pede*, which correspond to independent stages of the alignment procedure. During the first (*mille*) step, tracks from input datasets are fitted, selected according to the quality criteria listed in Appendix A.2, and used to accumulate the derivatives, residuals, uncertainties and certain bookkeeping information, which are stored in binary files. In order to fit and select good tracks, the track and hit selection is defined at this stage, as well as starting conditions of the tracker, which are used to determine track parameters. During this process, tens of millions of tracks have to be analysed by the CMS software, which is efficiently performed by analysing hundreds of data subsets in parallel at different nodes of the CMS batch farm.

The next (*pede*) step determines the alignment parameters in a single run of the standalone `pede` program. This step is parallelised within a single process, which is run on multiple processor cores of the same computer, since the whole matrix for more than 200 000 alignable parameters has to be accessible from memory to solve

(a) Without 0 T collision tracks

(b) With 0 T collision tracks
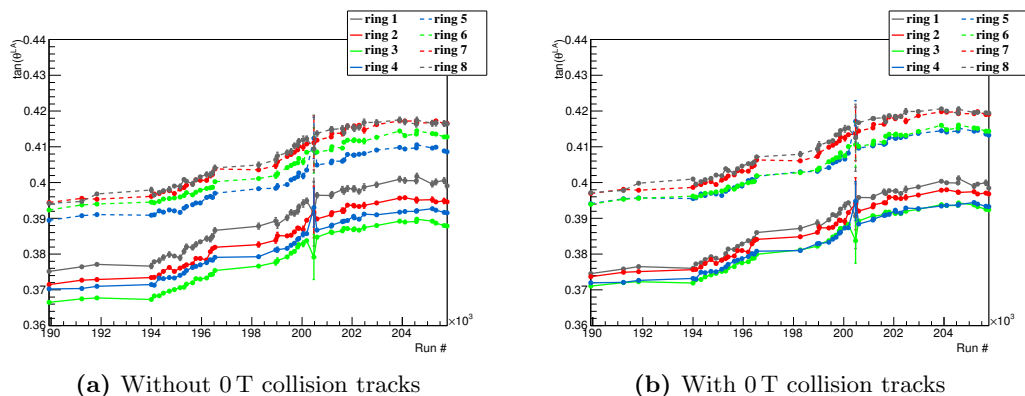
Figure A.1: Distribution of $\tan(\theta_{\mathrm{LA}}) = \mu_{\mathrm{H}} \cdot \mathbf{B}_v$, as a function of run number in BPIX
modules of layer 3, as determined by MILLEPEDE II. The 45 points in
each curve correspond to 45 IOVs corresponding to approximately the same
integrated luminosity. Gaps without points correspond to run ranges during
which no data were recorded. Colour coding corresponds to the one defined
in Figure 5.17. The evolution of $\tan(\theta_{\mathrm{LA}})$ is shown for calibration results
obtained with and without adding 0 T collision data to the alignment input.
The time dependence is identical in the both cases, but absolute values for
modules in neighbouring BPIX rings are significantly closer if 0 T collision
tracks are used in alignment.

(a) Without 0 T collision tracks

(b) With 0 T collision tracks
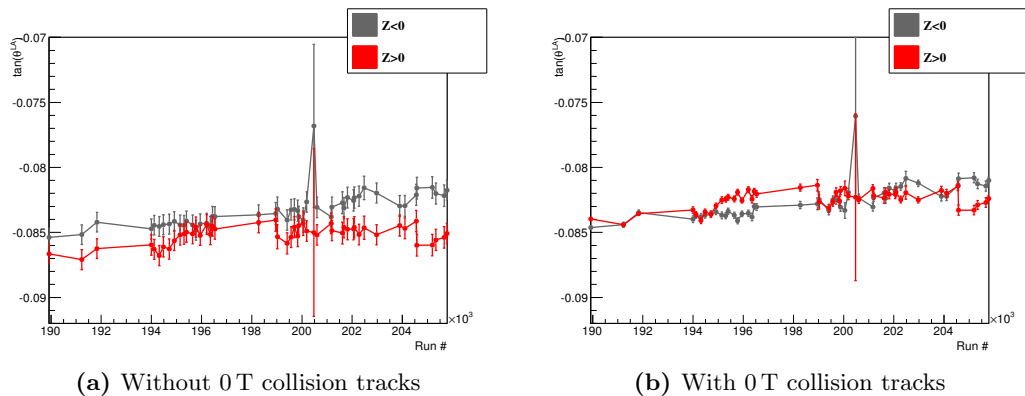
Figure A.2: Distribution of $\tan(\theta_{\mathrm{LA}}) = \mu_{\mathrm{H}} \cdot \mathbf{B}_v$, as a function of run number in FPIX
modules. The 45 points in each curve correspond to 45 IOVs corresponding
to approximately the same integrated luminosity. Gaps without points
correspond to run ranges during which no data were recorded. Modules
of the two sides of FPIX located at negative and positive z have separate
Lorentz-angle values. The evolution of $\tan(\theta_{\mathrm{LA}})$ is shown for calibration
results obtained with and without 0 T collision data to the alignment input.
The time dependence is identical in the both cases, but absolute values for
modules in the two sides of FPIX are significantly closer if 0 T collision
tracks are used in alignment. Also uncertainties are significantly reduced
due to the increased amount of tracks in forward direction.

the system of linear equations (5.5). Its output is the list of determined alignment parameters in a text format, which is then used by the CMS software.

The memory consumption of the `pede` grogram during the full-scale alignment of the tracker is very high, thus two special machines with 48 GB of Random Access Memory (RAM) are available at the batch farm. These machines serve a dedicated `cmscafexclusive` queue for the pede jobs of Millepede II alignment. This limit of available memory constrains the amount of data that can be used for the full-scale alignment, leading to a fraction of data not being included in alignment in order to keep the memory consumption below 48 GB. Specifically for tracker-alignment-related tasks, a dedicated `desy-cms011` machine is available at DESY, which has 64 GB of memory. Nevertheless, running pede on this machine requires a lot of manual actions, i.e. transferring of mille binaries, job submission, conversion of alignment results to the CMS format, etc. Thus, it was used only a few times for an estimate of memory consumption with more data and more global parameters to be determined, as described in Appendix A.7.

The global parameters obtained from a single run of the Millepede II alignment can be used as starting conditions for the next iteration of the alignment sequence. Thus, the input for the mille step is reconfigured and the alignment is performed on the same set of tracks.

A.5 Proper derivatives for Lorentz-angle in FPIX

The parameterization (5.12) is correct only for modules in the barrel region of the tracker, i.e. BPIX, TIB and TOB, in which B_u and B_w components of the magnetic field are 0. In contrast, FPIX modules are tilted in turbine-blades-like manner, leading to non-negligible $\mathbf{B}_v$ and $\mathbf{B}_w$ components. For example, in FPIX modules, which are tilted by 20° around the v axis, the components are: $B_v = -|\mathbf{B}| \cdot \sin(20°) \approx -0.34 \cdot |\mathbf{B}|$ and $B_w = \pm|\mathbf{B}| \cdot \cos(20°) \approx \pm 0.94 \cdot |\mathbf{B}|$ [272]. The sign of B_w alternates according to the FPIX-module orientation. The large component of the magnetic field in the w direction leads to non-zero second-order terms in the derivative (5.12)), as well as to shifts in the v direction becoming relevant [273]:

$$
\begin{aligned}
\frac{\partial(\Delta u)}{\partial \mu_\mathrm{H}} &= -\frac{d'}{2} \cdot \frac{\mathbf{B}_v + 2\mu_\mathrm{H}\mathbf{B}_u\mathbf{B}_w - \mu_\mathrm{H}^2\mathbf{B}_v\mathbf{B}_w^2}{(1 + \mu_\mathrm{H}^2\mathbf{B}_w^2)^2}, \\[2mm]
\frac{\partial(\Delta v)}{\partial \mu_\mathrm{H}} &= -\frac{d'}{2} \cdot \frac{\mathbf{B}_u + 2\mu_\mathrm{H}\mathbf{B}_v\mathbf{B}_w - \mu_\mathrm{H}^2\mathbf{B}_u\mathbf{B}_w^2}{(1 + \mu_\mathrm{H}^2\mathbf{B}_w^2)^2}.
\end{aligned}
\tag{A.1}
$$

A.6 Starting conditions for the combined alignment and calibration setup

The starting geometry of the tracker is obtained from the last IOV #23 of the `mp1098` alignment. This includes positions and orientations of individual modules and large structures, as well as surface deformations of individual sensors. Latest input values of Lorentz angle (μ_H) and backplane correction (Δw) were taken from a database using the following tags:

- **Pixel Lorentz angle:**

  - `SiPixelLorentzAngle_v03_offline` [274];
  - input μ_{H} values: BPIX: $0.1031\,\mathrm{T}^{-1}$, FPIX: $0.064\,04\,\mathrm{T}^{-1}$;

- **Strip Lorentz angle:**

  - `SiStripLorentzAngleDeco_GR10_v1_offline_BPCorrected`

- **Strip backplane correction:**

  - `SiStripBackPlaneCorrection_deco_GR10_v3_offline`

All other conditions were taken from the global tag `FT_R_53_V21` [275], which is a large collection of database objects containing a large number of corrections, calibrations and other quantities, which are not explicitly relevant for the discussed calibrations.

A.7 Optimisation of data composition

In the ideal case, all data that is available should be used for the MILLEPEDE II alignment. In the real world this is not possible, since larger amount of data corresponds to a larger amount of memory required for the *pede* step. In particular, a larger amount of tracks will increase the number of elements of the global-parameter matrix that have enough entries to not be removed as insignificant, or to increase the storage precision. Furthermore, tracks that introduce significant correlations between different global parameters will make the global-parameter matrix less sparse.

Determination of surface deformations of tracker modules has a significant impact on the required memory for the *pede* step. Therefore, during the development of the calibration approach, modules were assumed to be flat, which resulted in about 90 000 alignment parameters. This allowed to use about 65 million of tracks of different configurations in a single run of alignment. But some certain validations showed that the surface deformations in some modules are quite significant, e.g. in TOB layer 2 hits were shifted by up to 6 μm due to shape deformations. These shifts were then wrongly compensated by backplane corrections and rotations of the modules.

Thus, for a reliable alignment and calibration, surface-shape parameters had to be determined as well. This increased memory requirements for the *pede* step to more than 64 GB, which made it impossible to obtain any results, since using swap memory from a hard drive would reduce the speed of calculations by several orders of magnitude.

In order to still run the full-scale alignment and calibration with non-flat sensors, amount of data used in alignment had to be reduced. It was found that tracks from $Z^0 \to \mu^+\mu^-$ decays make the biggest impact on the memory consumption, because they introduce correlations between a large amount of global parameters via the mass constraint, and therefore increase the number of non-zero elements of the correlation matrix that is stored in memory. Thus, the number of tracks from $Z^0 \to \mu^+\mu^-$ events was reduced, as well as the number of minimum-bias tracks recorded in 0 T and 3.8 T magnetic fields.

Based on the results of the memory-consumption study, a machine with an increased computational power was requested specifically for the MILLEPEDE II alignment. One of the main criteria were the large amount of Rapid Access Memory ($\geq 96\,\text{GB}$) and preferably Solid State Drive (SSD). In particular, the SSD is important because a large fraction of time during the *pede* step is spent on reading the binary output from *mille* step into the memory, which performs much faster on SSD.

Two machines that partially fulfil the requirements were installed in 2014. They have $256\,\text{GB}$ of RAM, which should allow to use all $Z^0 \rightarrow \mu^+\mu^-$ events recorded in 2012, as well as more $0\,\text{T}$ collision tracks and isolated muons. This is especially important for time-dependent module alignment, since a large amount of tracks could be used from a short period of time, allowing to check possible movements of individual modules. The relevance of such approach is discussed in Section 5.7.1.

Appendix B

Details of the differential $t\bar{t}b\bar{b}$ cross-section measurement

B.1 HLT trigger paths

As stated in Section 9.4.1, the first step in selection of dileptonic $t\bar{t}$ events is based on the high level double-lepton trigger. A number of different trigger paths are combined with the logical `OR` operation to select events with either an electron-pair, or a muon-pair or an electron-muon-pair. The actual HLT paths for the different final states are given in Table B.1 and are separated between the collected data and MC simulations.

Table B.1: Summary of the HLT-trigger paths used in data and MC simulation for the three dilepton channels: ee, $e\mu$ and $\mu\mu$.

	Type	HLT Path
ee	Data	`HLT_Ele17_CaloIdT_CaloIsoVL_TrkIdVL_TrkIsoVL_TrkIsoVL_v*`
	MC	`HLT_Ele17_CaloIdT_CaloIsoVL_TrkIdVL_TrkIsoVL_TrkIsoVL_v17`
$\mu\mu$	Data	`HLT_Mu17_Mu8_v*` `HLT_Mu17_TkMu8_v*`
	MC	`HLT_Mu17_Mu8_v17` `HLT_Mu17_TkMu8_v10`
$e\mu$	Data	`HLT_Mu17_Ele8_CaloIdT_CaloIsoVL_TrkIdVL_TrkIsoVL_v*` `HLT_Mu8_Ele17_CaloIdT_CaloIsoVL_TrkIdVL_TrkIsoVL_v*`
	MC	`HLT_Mu17_Ele8_CaloIdT_CaloIsoVL_TrkIdVL_TrkIsoVL_v7` `HLT_Mu8_Ele17_CaloIdT_CaloIsoVL_TrkIdVL_TrkIsoVL_v7`

B.2   Binning of the $t\overline{t}b\overline{b}$ measurement

In Section 9.7.1, correlation matrices, purity and stability for each bin of the measured $t\overline{t}b$ and $t\overline{t}b\overline{b}$ quantities were shown for the full $t\overline{t}$ phase space. Below, the same distributions are shown for the visible phase space. In this case the number of entries in the correlation matrices is lower, since a part of the signal events is removed due to the stricter requirements to the $t\overline{t}$ final state.

B.3   Unfolding for the $t\overline{t}b\overline{b}$ measurement

In Section 9.7.2 an example of the dependence of the RMS global correlation on the scanned regularisation-strength-parameter value was shown for a few variables in the full $t\overline{t}$ phase space. A set of the same plots for other variables is shown in Figure B.3. The dependence for all the measured variables in the visible phase space is shown in Figure B.4.

(a) transverse momentum of the first additional b-tagged jet

(b) absolute pseudorapidity of the first additional b-tagged jet

(c) transverse momentum of the second additional b-tagged jet

(d) absolute pseudorapidity of the second additional b-tagged jet

(e) distance in $\eta - \phi$ space between the first and second additional b-tagged jets

(f) invariant mass of the pair of the first and second additional b-tagged jets

Figure B.1: Correlations between the measured and generated $t\bar{t}b$ and $t\bar{t}b\bar{b}$ quantities, as determined from the corresponding simulated signal events. The used $t\bar{t}b$ and $t\bar{t}b\bar{b}$ signal definitions correspond to the visible phase space of the $t\bar{t}$ system.

(a) transverse momentum of the first additional b-tagged jet
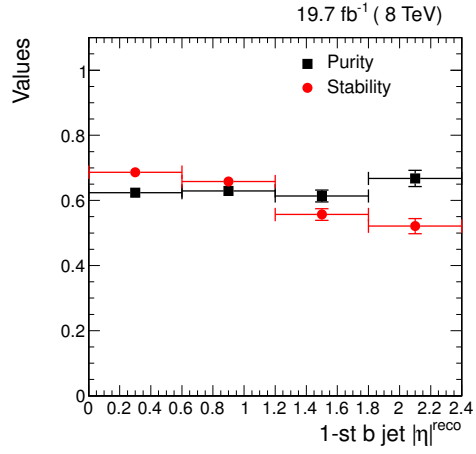
(b) absolute pseudorapidity of the first additional b-tagged jet
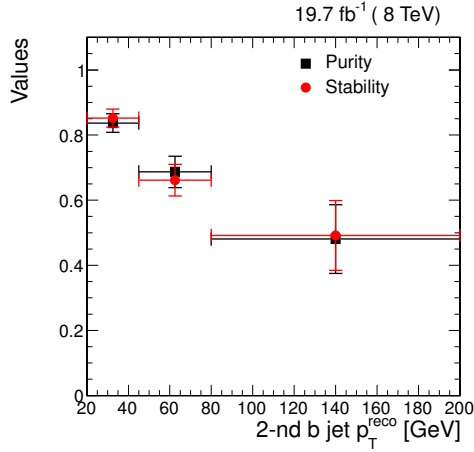
(c) transverse momentum of the second additional b-tagged jet
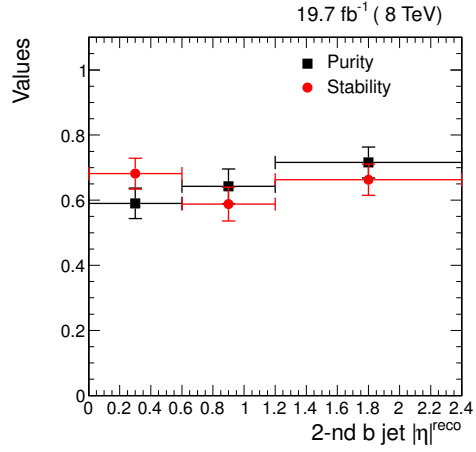
(d) absolute pseudorapidity of the second additional b-tagged jet
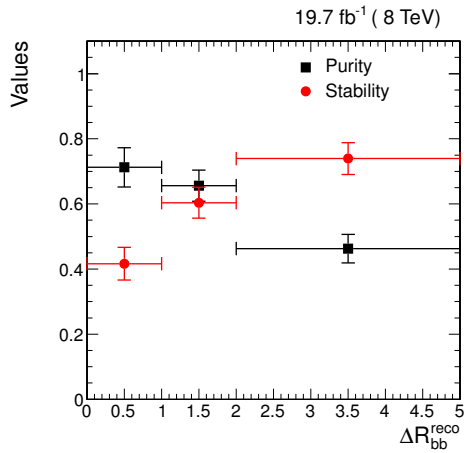
(e) distance in $\eta - \phi$ space between the first and second additional b-tagged jets
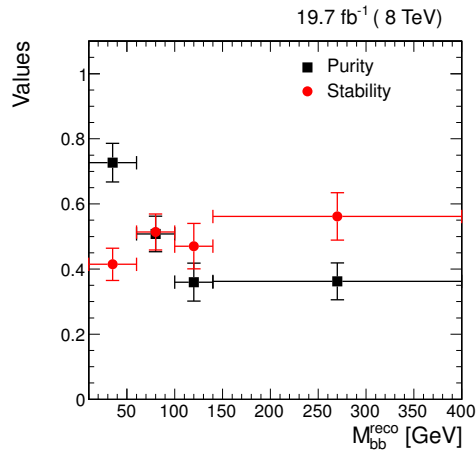
(f) invariant mass of the pair of the first and second additional b-tagged jets

Figure B.2: Purity and stability of the measured $t\bar{t}b$ and $t\bar{t}b\bar{b}$ quantities as determined from the corresponding simulated signal events. The used $t\bar{t}b$ and $t\bar{t}b\bar{b}$ signal definitions correspond to the visible phase space of the $t\bar{t}$ system. Color represents a number of MC events entering each bin of the distribution according to the color axis.
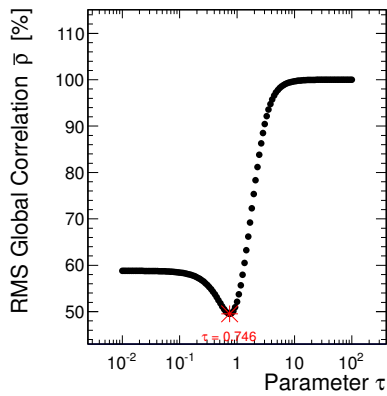
(a) $|\eta|$ of the first additional b-jet
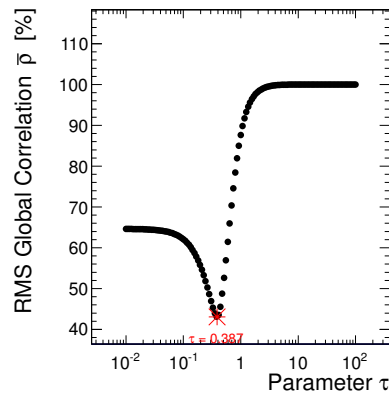
(b) $|\eta|$ of the second additional b-jet
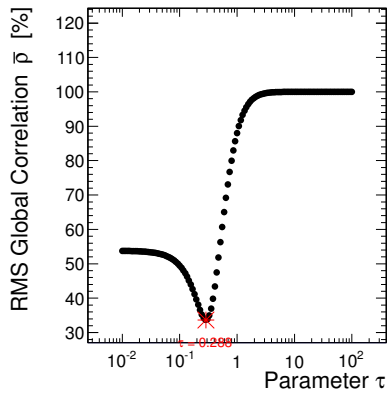
(c) p_{T} of the second additional b-jet
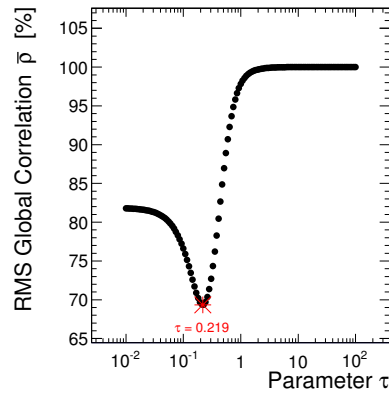
(d) $m_{b\bar{b}}$ of the pair of the first and second additional b-jets

Figure B.3: The dependence of the RMS global correlation $\bar{\rho}$ on the regularisation-strength parameter τ for several measured quantities in the full $t\bar{t}$ phase space. The value of τ corresponding to the smallest global correlation is marked in red and is used for the final unfolded result.
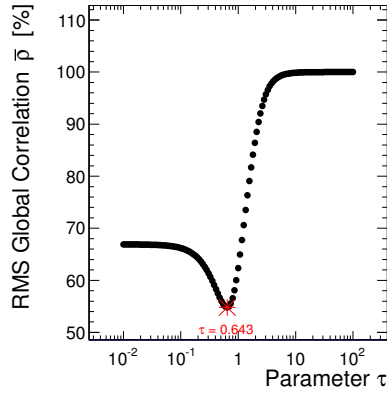
(a) p_{T} of the first additional b-jet
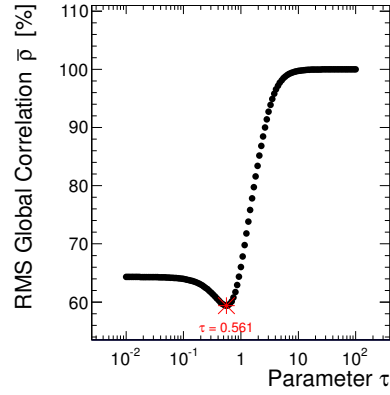
(b) $|\eta|$ of the first additional b-jet
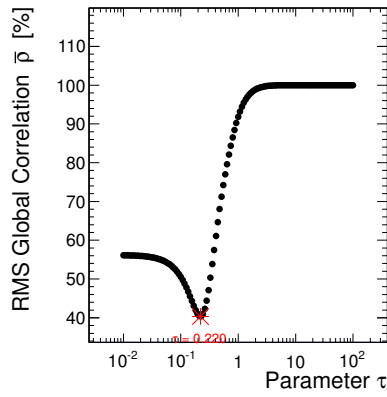
(c) p_{T} of the second additional b-jet
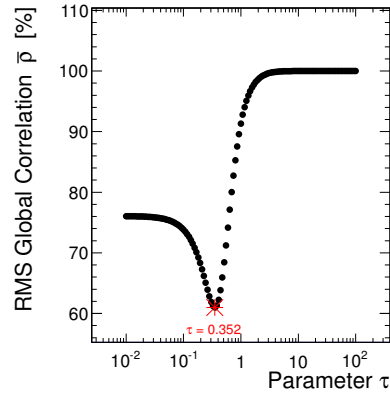
(d) $|\eta|$ of the second additional b-jet
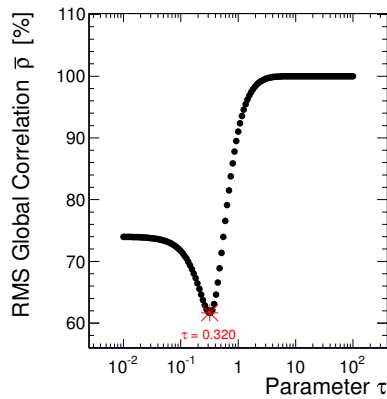
(e) $\Delta R_{b\overline{b}}$ between the first and second additional b-jets
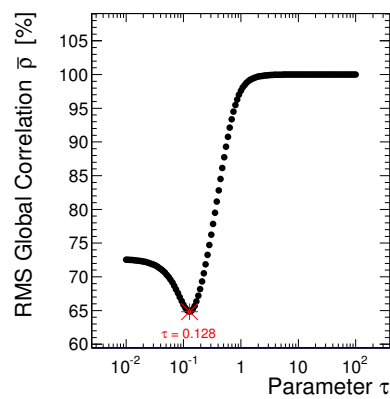
(f) $m_{b\overline{b}}$ of the pair of the first and second additional b-jets

Figure B.4: The dependence of the RMS global correlation $\bar{\rho}$ on the regularisation-strength parameter τ for all the measured quantities in the visible $t\overline{t}$ phase space. The value of τ corresponding to the smallest global correlation is marked in red and is used for the final unfolded result.

(a) p_{T} of the first additional b jet

(b) $|\eta|$ of the first additional b jet

(c) p_{T} of the second additional b jet

(d) $|\eta|$ of the second additional b jet

(e) $\Delta R_{b\bar{b}}$ between the first and second additional b jets

(f) $m_{b\bar{b}}$ of the pair of the first and second additional b jets
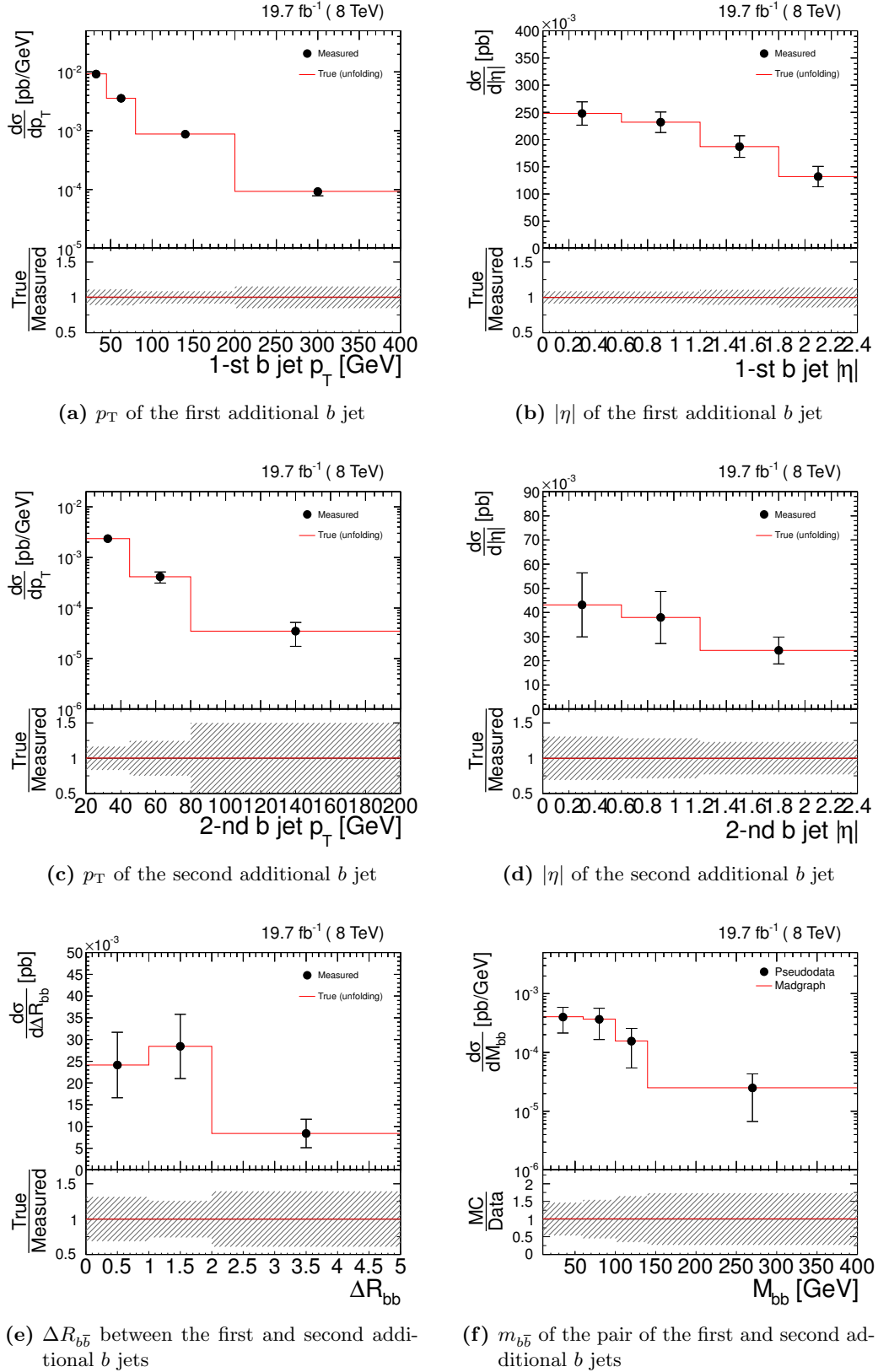
Figure B.5: Unfolded cross sections as functions of all the measured variables in the full $t\bar{t}$ phase space, obtained with undistorted pseudodata, which exactly matches the sum of MC simulations.

(a) $w = 1 + 0.3\dfrac{m_{b\overline{b}} - 100}{100}$

(b) $w = 1 - 0.3\dfrac{m_{b\overline{b}} - 100}{100}$
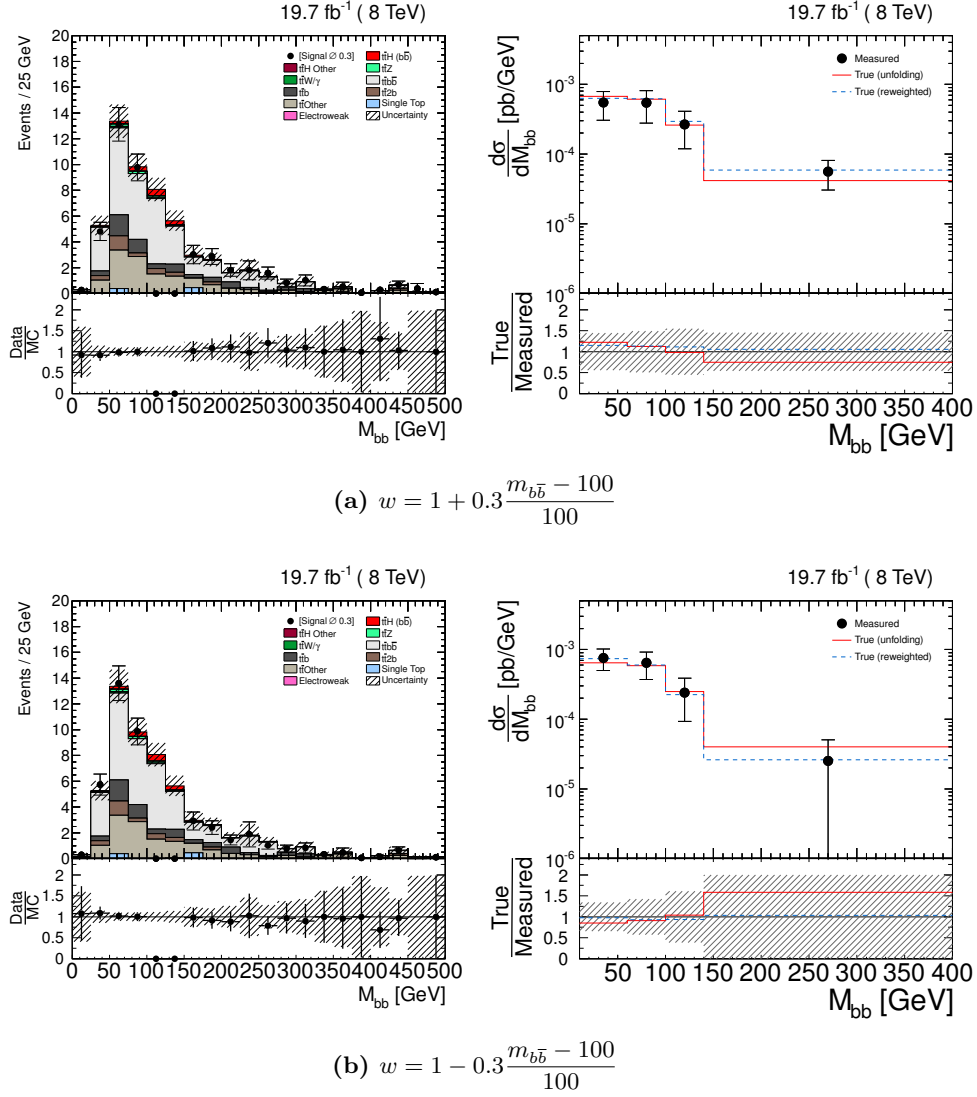
Figure B.6: Control distribution (left) and unfolded cross section (right) as a function of the $m_{b\overline{b}}$ of the pair of the first and second additional b jets in the full $t\overline{t}$ phase space. The weight w is a function of the $m_{b\overline{b}}$, and is applied to every event that has at least two additional b jets, as defined in Section 9.2.1. The only reweighted signal sample corresponds to the $t\overline{t}b\overline{b}$ process.

(a) $w = 1 + 0.3\dfrac{\Delta R_{b\bar{b}} - 60}{60}$

(b) $w = 1 - 0.3\dfrac{\Delta R_{b\bar{b}} - 60}{60}$
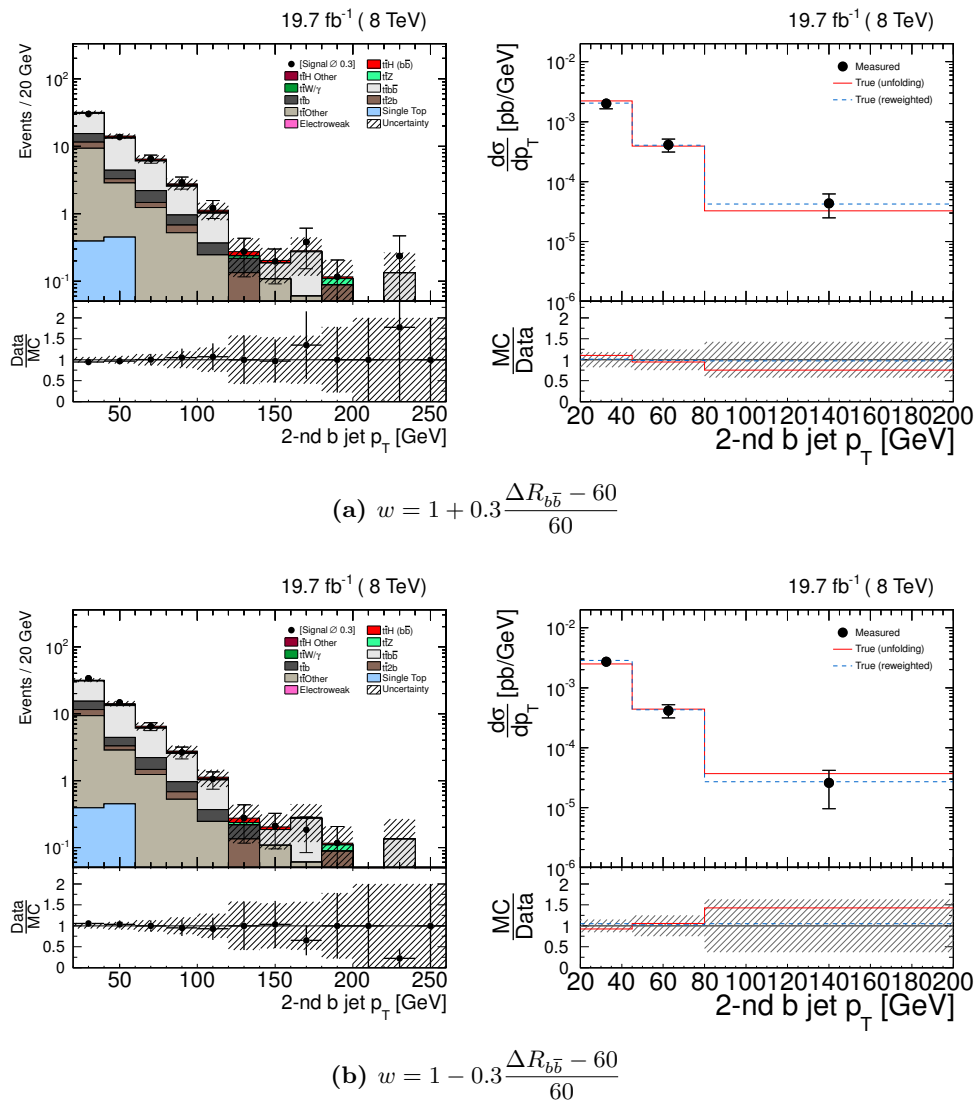
Figure B.7: Control distribution (left) and unfolded cross section (right) as a function of the p_{T} of the second additional b jet in the full $t\bar{t}$ phase space. The weight w is a function of the p_{T} of the second additional b jet, and is applied to every event that has at least two additional b jets, as defined in Section 9.2.1. The only reweighted signal sample corresponds to the $t\bar{t}b\bar{b}$ process.

(a) $w = 1 + 0.3(|\eta| - 1)$

(b) $w = 1 - 0.3(|\eta| - 1)$
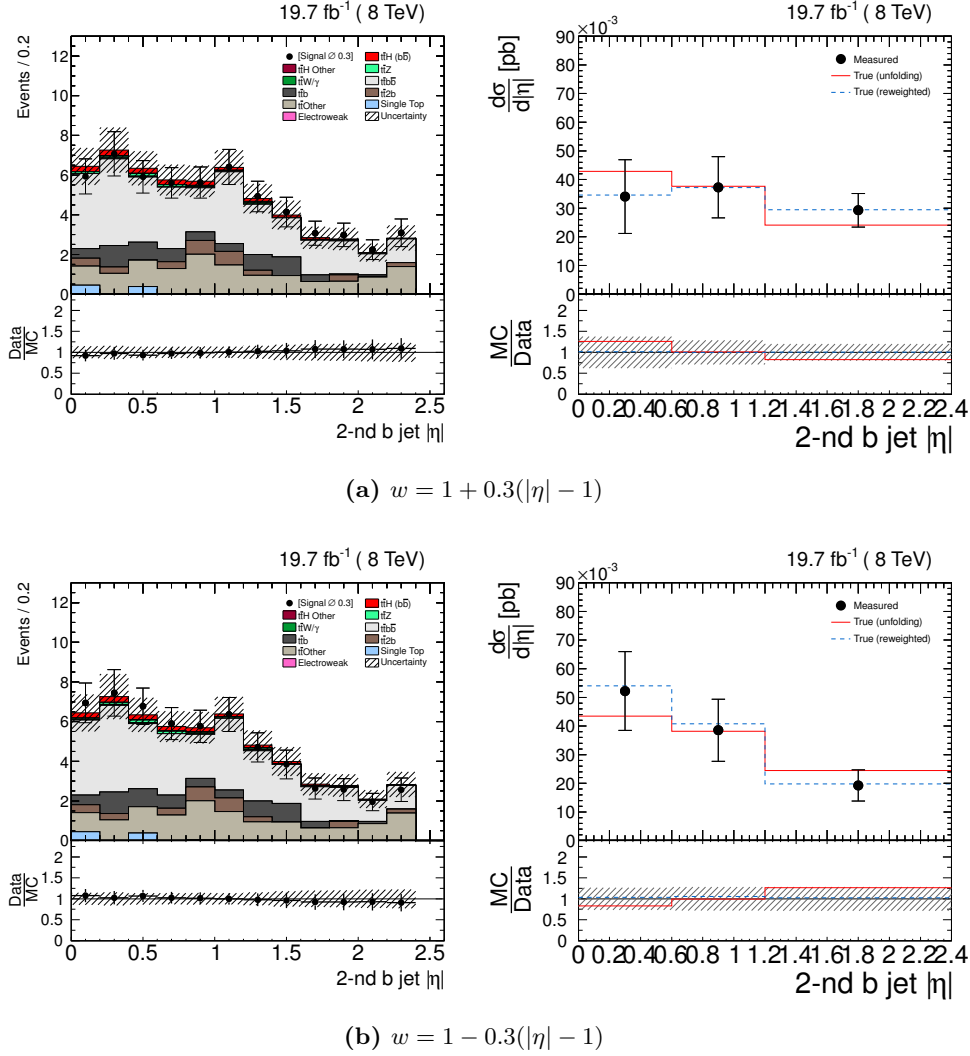
Figure B.8: Control distribution (left) and unfolded cross section (right) as a function of the $|\eta|$ of the second additional b jet in the full $t\bar{t}$ phase space. The weight w is a function of the $|\eta|$ of the second additional b jet, and is applied to every event that has at least two additional b jets, as defined in Section 9.2.1. The only reweighted signal sample corresponds to the $t\bar{t}b\bar{b}$ process.

(a) $w = 1 + 0.3(|\eta| - 1)$

(b) $w = 1 - 0.3(|\eta| - 1)$
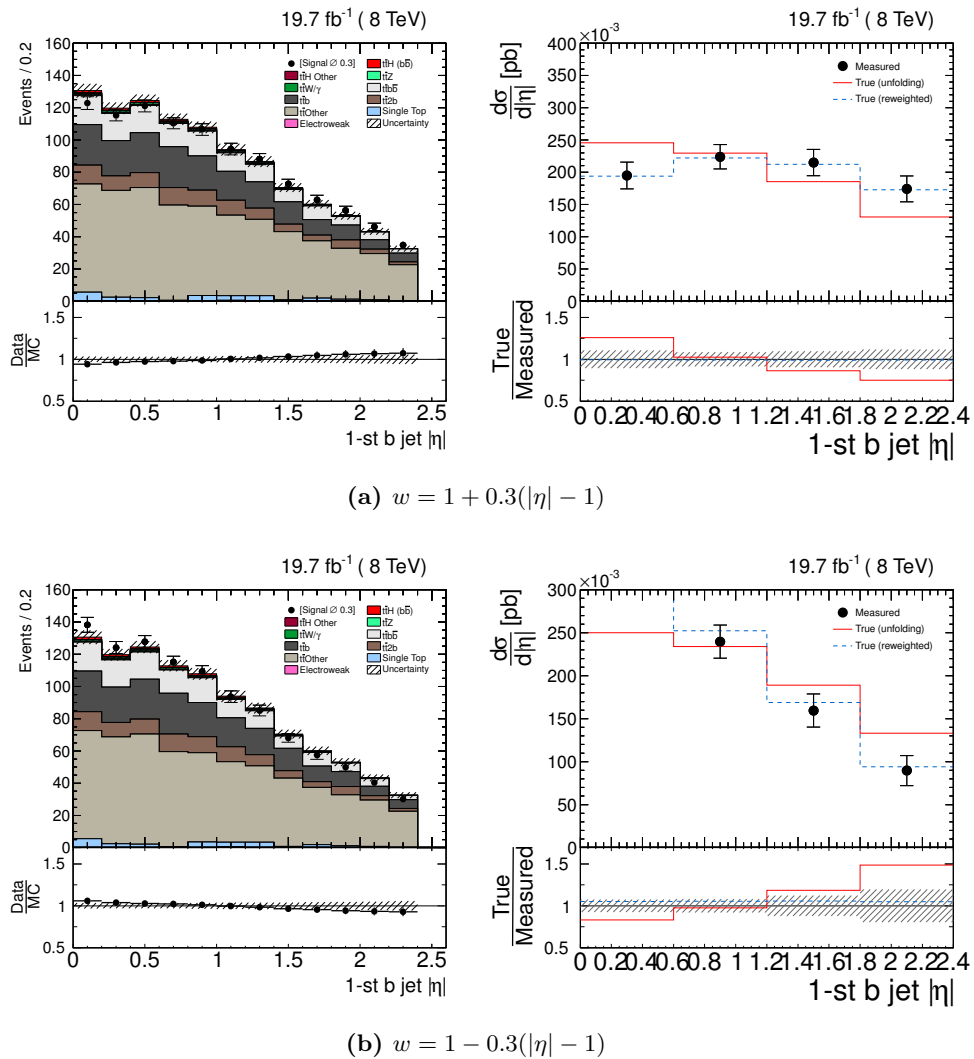
Figure B.9: Control distribution (left) and unfolded cross section (right) as a function of the $|\eta|$ of the first additional b jet in the full $t\bar{t}$ phase space. The weight w is a function of the $|\eta|$ of the first additional b jet, and is applied to every event that has at least one additional b jet, as defined in Section 9.2.1. The reweighted signal samples corresponds to the $t\bar{t}b$, $t\bar{t}2b$ and $t\bar{t}b\bar{b}$ processes.

List of Tables

List of Figures

269

Bibliography

[1] Sylvia Berryman. Ancient Atomism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter 2011 edition, 2011.

[2] Philip P. Wiener. *Dictionary of the History of Ideas*. Charles Scribner's Sons, 1973. `http://xtf.lib.virginia.edu/xtf/view?docId=DicHist/uvaBook/tei/DicHist1.xml`.

[3] Mrinalkanti Gangopadhyaya. *Indian Atomism: History and Sources*. Bagchi, K.P., 1980.

[4] Literary and Philosophical Society of Manchester. *Memoirs of the Literary and Philosophical Society of Manchester*. The Society, 1805. URL: `http://books.google.de/books?id=LJNIAAAAYAAJ`.

[5] J. Dalton. *A New System of Chemical Philosophy*. S. Russell, 1808. URL: `http://books.google.de/books?id=Wp7QAAAAMAAJ`.

[6] H. Geiger. On the Scattering of the α-Particles by Matter. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 81(546):174–177, 1908. URL: `http://rspa.royalsocietypublishing.org/content/81/546/174`, `arXiv:http://rspa.royalsocietypublishing.org/content/81/546/174`, `doi:10.1098/rspa.1908.0067`.

[7] H. Geiger and E. Marsden. The laws of deflexion of α particles through large angles. *Philosophical Magazine Series 6*, 25(148):604–623, 1913. URL: `http://dx.doi.org/10.1080/14786440408634197`, `arXiv:http://dx.doi.org/10.1080/14786440408634197`, `doi:10.1080/14786440408634197`.

[8] E. Rutherford. The scattering of α and β particles by matter and the structure of the atom. *Philosophical Magazine Series 6*, 21(125):669–688, 1911. URL: `http://dx.doi.org/10.1080/14786440508637080`, `arXiv:http://dx.doi.org/10.1080/14786440508637080`, `doi:10.1080/14786440508637080`.

[9] J.J. Thomson. On the structure of the atom: an investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure. *Philosophical Magazine Series 6*, 7(39):237–265, 1904. URL: `http://dx.doi.org/10.1080/14786440409463107`, `doi:10.1080/14786440409463107`.

[10] L. Evans and P. Bryant. LHC Machine. *JINST*, 3(S08001), 2008. `http://iopscience.iop.org/1748-0221/3/08/S08001`.

[11] CMS Collaboration. The CMS experiment at the CERN LHC. *JINST*, 3(S08004), 2008. http://iopscience.iop.org/1748-0221/3/08/S08004.

[12] ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3(S08003), 2008. http://iopscience.iop.org/1748-0221/3/08/S08003.

[13] ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716(1):1–29, 2012. URL: http://www.sciencedirect.com/science/article/pii/S037026931200857X, doi:10.1016/j.physletb.2012.08.020.

[14] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716(1):30–61, 2012. URL: http://www.sciencedirect.com/science/article/pii/S0370269312008581, doi:10.1016/j.physletb.2012.08.021.

[15] F. Englert and R. Brout. Broken Symmetry and the Mass of Gauge Vector Mesons. *Phys. Rev. Lett.*, 13:321–323, Aug 1964. URL: http://link.aps.org/doi/10.1103/PhysRevLett.13.321, doi:10.1103/PhysRevLett.13.321.

[16] Peter W. Higgs. Broken Symmetries and the Masses of Gauge Bosons. *Phys. Rev. Lett.*, 13:508–509, Oct 1964. URL: http://link.aps.org/doi/10.1103/PhysRevLett.13.508, doi:10.1103/PhysRevLett.13.508.

[17] The Royal Swedish Academy of Sciences. The Nobel Prize in Physics 2013. http://www.nobelprize.org/nobel_prizes/physics/laureates/2013/press.pdf, October 2013.

[18] Enrico Fermi. On the Quantization of the Monoatomic Ideal Gas. arXiv:cond-mat/9912229, 1999. english translation by Alberto Zannoni.

[19] P. A. M. Dirac. On the Theory of Quantum Mechanics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 112(762):661–677, 1926. URL: http://dx.doi.org/10.1098/rspa.1926.0133, doi:10.1098/rspa.1926.0133.

[20] Wolfgang Pauli. Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren. *Zeitschrift für Physik*, 31(1):765–783, 1925. from german: On the Connexion between the Completion of Electron Groups in an Atom with the Complex Structure of Spectra. URL: http://dx.doi.org/10.1007/BF02980631, doi:10.1007/BF02980631.

[21] Satyendra Nath Bose. Plancks Gesetz und Lichtquantenhypothese. *Zeitschrift für Physik*, 26(1):178–181, 1924. translation into german by Albert Einstein. URL: http://dx.doi.org/10.1007/BF01327326, doi:10.1007/BF01327326.

[22] Albert Einstein. Quantentheorie des einatomigen idealen Gases. *Sitzungsberichte der Preussischen Akademie der Wissenschaften, Physikalisch-Mathematische Klasse*, pages 261–267, 1924. from german: Quantum Theory of the Monatomic Ideal Gas. URL: http://books.google.de/books?id=oRFFHAAACAAJ.

[23] Donald H. Perkins. *Introduction to High Energy Physics.* Cambridge University Press, 4th edition, 2000. Cambridge Books Online. URL: `http://dx.doi.org/10.1017/CBO9780511809040`, `doi:10.1017/CBO9780511809040`.

[24] Mark Thomson. *Modern Particle Physics.* Cambridge University Press, 4th edition, 2013. URL: `http://www.cambridge.org/de/academic/subjects/physics/particle-physics-and-nuclear-physics/modern-particle-physics`.

[25] Markus Fierz. Über die relativistische Theorie kräftefreier Teilchen mit beliebigem Spin. *Helvetica Physica Acta*, 12(1):3–37, 1939. URL: `http://retro.seals.ch/digbib/view?pid=hpa-001:1939:12::13`, `doi:10.5169/seals-110930`.

[26] Wolfgang Pauli. The Connection Between Spin and Statistics. *Physical Review*, 58(8):716–722, October 1940. URL: `http://link.aps.org/doi/10.1103/PhysRev.58.716`, `doi:10.1103/PhysRev.58.716`.

[27] K.A. Olive et al. (Particle Data Group). PDG Particle Listings. *Chin. Phys. C*, 38, 2014. URL: `http://pdg.lbl.gov/2014/listings/contents_listings.html`.

[28] G. Guralnik, C. Hagen, and T. Kibble. Global Conservation Laws and Massless Particles. *Physical Review Letters*, 13:585–587, November 1964. URL: `http://link.aps.org/doi/10.1103/PhysRevLett.13.585`, `doi:10.1103/PhysRevLett.13.585`.

[29] Emmy Noether. Invariant variation problems. *Transport Theory and Statistical Physics*, 1(3):186–207, 1971. `doi:10.1080/00411457108231446`.

[30] M. Weitz, A. Huber, F. Schmidt-Kaler, D. Leibfried, and T. W. Hänsch. Precision measurement of the hydrogen and deuterium 1 S ground state Lamb shift. *Phys.Rev.Lett.*, 72(3):328–331, 1994. `http://prl.aps.org/abstract/PRL/v72/i3/p328_1`.

[31] D. Berkeland, E. Hinds, and M. Boshier. Precise Optical Measurement of Lamb Shifts in Atomic Hydrogen. *Phys.Rev.Lett.*, 75(13):2470–2473, 1995. `http://prl.aps.org/abstract/PRL/v75/i13/p2470_1`.

[32] Richard P. Feynman. *QED: The Strange Theory of Light and Matter.* Princeton University Press, 1985.

[33] Richard P. Feynman. Space-Time Approach to Quantum Electrodynamics. *Phys. Rev.*, 76:769–789, Sep 1949. URL: `http://link.aps.org/doi/10.1103/PhysRev.76.769`, `doi:10.1103/PhysRev.76.769`.

[34] David Kaiser. Physics and Feynman's Diagrams. *American Scientist*, 93(2):156–165, 2005. `http://web.mit.edu/dikaiser/www/FdsAmSci.pdf`. `doi:10.1511/2005.2.156`.

[35] H. J. Bhabha. The Scattering of Positrons by Electrons with Exchange on Dirac's Theory of the Positron. *Proc.R.Soc.*, A154:195–206, 1936. `http://rspa.royalsocietypublishing.org/content/154/881/195`.

[36] F. Jegerlehner. The running fine structure constant via the Adler function. *Nuclear Physics B - Proceedings Supplements*, 181–182:135–140, 2008.

Proceedings of the International Workshop on e+e- Collisions. URL: `http://www.sciencedirect.com/science/article/pii/S0920563208001424`, `doi:10.1016/j.nuclphysbps.2008.09.010`.

[37] L3 Collaboration. Measurement of the running of the fine-structure constant. *Physics Letters B*, 476(1–2):40–48, 2000. URL: `http://www.sciencedirect.com/science/article/pii/S0370269300001222`, `doi:10.1016/S0370-2693(00)00122-2`.

[38] L3 Collaboration. Measurement of the running of the electromagnetic coupling at large momentum-transfer at LEP. *Physics Letters B*, 623(1–2):26–36, 2005. URL: `http://www.sciencedirect.com/science/article/pii/S0370269305010439`, `doi:10.1016/j.physletb.2005.07.052`.

[39] R. Devenish and A. Cooper-Sarkar. *Deep Inelastic Scattering*. Oxford University Press, 2004. URL: `https://books.google.de/books?id=LsSumroE7nsC`.

[40] Belle Collaboration. Observation of a Resonancelike Structure in the $\pi^{+-}\psi^{'}$ Mass Distribution in Exclusive $B \to K\pi^{+-}\psi^{'}$ Decays. *Physical Review Letters*, 100:142001, Apr 2008. URL: `http://link.aps.org/doi/10.1103/PhysRevLett.100.142001`, `doi:10.1103/PhysRevLett.100.142001`.

[41] LHCb Collaboration. Observation of the Resonant Character of the $Z(4430)^-$ State. *Physical Review Letters*, 112:222002, Jun 2014. URL: `http://link.aps.org/doi/10.1103/PhysRevLett.112.222002`, `doi:10.1103/PhysRevLett.112.222002`.

[42] James Cronin. CP Symmetry Violation – The Search for Its Origin. *Nobel Lectures*, 1980. `http://www.nobelprize.org/nobel_prizes/physics/laureates/1980/cronin-lecture.html`.

[43] Val Fitch. The Discovery of Charge – Conjugation Parity Asymmetry. *Nobel Lectures*, 1980. `http://www.nobelprize.org/nobel_prizes/physics/laureates/1980/fitch-lecture.html`.

[44] M. Kobayashi and T. Maskawa. CP Violation in the Renormalizable Theory of Weak Interaction. *Prog.Theor.Phys.*, 49:652–657, 1973. `http://ptp.ipap.jp/link?PTP/49/652/`.

[45] CMS Collaboration. Measurement of the t-channel single-top-quark production cross section and of the $|\ V_{tb}\ |$ CKM matrix element in pp collisions at $\sqrt{s}=$ 8 TeV. *JHEP*, 1406:090, 2014. `arXiv:1403.7366`, `doi:10.1007/JHEP06(2014)090`.

[46] S. L. Glashow. Partial-symmetries of weak interactions. *Nucl.Phys.*, 22(4):579–588, 1961. URL: `www.slac.stanford.edu/spires/find/hep/www?j=NUPHA,22,579`.

[47] Abdus Salam and J.C. Ward. Weak and electromagnetic interactions. *Il Nuovo Cimento Series 10*, 11(4):568–577, 1959. URL: `http://dx.doi.org/10.1007/BF02726525`, `doi:10.1007/BF02726525`.

[48] Steven Weinberg. A Model of Leptons. *Phys. Rev. Lett.*, 19:1264–1266, Nov 1967. URL: `http://link.aps.org/doi/10.1103/PhysRevLett.19.1264`, `doi:10.1103/PhysRevLett.19.1264`.

[49] P.D.B. Collins, A.D. Martin, and E.J. Squires. *Particle Physics and Cosmology*. A Wiley-Interscience publication. Wiley, 1989. URL: `http://books.google.de/books?id=MBR8F8oxrzEC`.

[50] P.L. Anthony et al. Precision measurement of the weak mixing angle in Moller scattering. *Phys.Rev.Lett.*, 95:081601, 2005. `arXiv:hep-ex/0504049`, `doi:10.1103/PhysRevLett.95.081601`.

[51] Valerio Ippolito. *Measurement of properties of the new particle observed within the search for the Standard Model Higgs Boson in the $H \to ZZ^{(*)} \to 4l$ decay channel at ATLAS*. PhD thesis, Spienza – Università di Roma, Oct 2014.

[52] Goldstone, Jeffrey and Salam, Abdus and Weinberg, Steven. Broken symmetries. *Phys. Rev.*, 127:965–970, Aug 1962. URL: `http://link.aps.org/doi/10.1103/PhysRev.127.965`, `doi:10.1103/PhysRev.127.965`.

[53] CMS Collaboration. Precise determination of the mass of the Higgs boson and studies of the compatibility of its couplings with the standard model. Technical Report CMS-PAS-HIG-14-009, CERN, Geneva, 2014. URL: `http://cds.cern.ch/record/1728249`.

[54] G. 't Hooft. Dimensional regularization and the renormalization group. *Nuclear Physics B*, 61(0):455–468, 1973. URL: `http://www.sciencedirect.com/science/article/pii/0550321373903763`, `doi:http://dx.doi.org/10.1016/0550-3213(73)90376-3`.

[55] Steven Weinberg. New Approach to the Renormalization Group. *Phys. Rev. D*, 8:3497–3509, Nov 1973. URL: `http://link.aps.org/doi/10.1103/PhysRevD.8.3497`, `doi:10.1103/PhysRevD.8.3497`.

[56] J. Einasto. Dark Matter, 2009. `http://arxiv.org/abs/0901.0632`.

[57] G. Paál, I. Horváth, and B. Lukács. Inflation and compactification from galaxy redshifts? *Astrophysics and Space Science*, 191(1):107–124, 1992. URL: `http://dx.doi.org/10.1007/BF00644200`, `doi:10.1007/BF00644200`.

[58] Adam G. Riess et al. Observational evidence from supernovae for an accelerating universe and a cosmological constant. *Astron.J.*, 116:1009–1038, 1998. `arXiv:astro-ph/9805201`, `doi:10.1086/300499`.

[59] Th. M. Nieuwenhuizen. Do non-relativistic neutrinos constitute the dark matter? *EPL (Europhysics Letters)*, 86(5):59001, 2009. URL: `http://stacks.iop.org/0295-5075/86/i=5/a=59001`.

[60] Richard A. Battye and Adam Moss. Evidence for Massive Neutrinos from Cosmic Microwave Background and Lensing Observations. *Phys. Rev. Lett.*, 112:051303, Feb 2014. URL: `http://link.aps.org/doi/10.1103/PhysRevLett.112.051303`, `doi:10.1103/PhysRevLett.112.051303`.

[61] J. Wess and B. Zumino. Supergauge transformations in four dimensions. *Nuclear Physics B*, 70(1):39 – 50, 1974. URL: `http://www.sciencedirect.com/science/article/pii/0550321374903551`, `doi:10.1016/0550-3213(74)90355-1`.

[62] Ian J. R. Aitchison. Supersymmetry and the MSSM: An Elementary Introduction, 2005. http://arxiv.org/abs/hep-ph/0505105.

[63] P. Binetruy. *Supersymmetry:Theory, Experiment, and Cosmology: Theory, Experiment, and Cosmology.* Oxford Graduate Texts. OUP Oxford, 2006. URL: http://books.google.de/books?id=StDjdRnT72AC.

[64] M. Shifman. Large Extra Dimensions: Becoming Acquainted With An Alternative Paradigm. *International Journal of Modern Physics A*, 25(2&3):199–225, Jan 2010. URL: http://dx.doi.org/10.1142/S0217751X10048548, doi:10.1142/S0217751X10048548.

[65] N. Arkani-Hamed, S. Dimopoulos, and G. Dvali. The hierarchy problem and new dimensions at a millimeter. *Phys.Lett.*, B429(3–4):263–272, 1998. http://arxiv.org/abs/hep-ph/9803315.

[66] D. Kapner, T. Cook, E. Adelberger, J. Gundlach, B. Heckel, C. Hoyle, and H. Swanson. Tests of the Gravitational Inverse-Square Law below the Dark-Energy Length Scale. *Phys.Rev.Lett.*, 98:021101, 2007. URL: http://arxiv.org/abs/hep-ph/0611184.

[67] Paul Wesson. *Space-Time-Matter: Modern Kaluza-Klein Theory.* 1999.

[68] V. Barger and R. Phillips. *Collider Physics.* Frontiers in physics. Addison-Wesley Publishing Company, 1997. URL: http://books.google.de/books?id=k-iBmqBk6y4C.

[69] Paul H. Frampton, P.Q. Hung, and Marc Sher. Quarks and leptons beyond the third generation. *Physics Reports*, 330(5–6):263–348, 2000. URL: http://www.sciencedirect.com/science/article/pii/S0370157399000952, doi:10.1016/S0370-1573(99)00095-2.

[70] John M. Campbell and R.K. Ellis. MCFM for the Tevatron and the LHC. *Nuclear Physics B - Proceedings Supplements*, 205–206(0):10–15, 2010. Loops and Legs in Quantum Field Theory. Proceedings of the 10th DESY Workshop on Elementary Particle Theory. URL: http://www.sciencedirect.com/science/article/pii/S0920563210001945, doi:10.1016/j.nuclphysbps.2010.08.011.

[71] CDF Collaboration. Studying the Underlying Event in Drell-Yan and High Transverse Momentum Jet Production at the Tevatron. *Phys.Rev.*, D82:034001, 2010. URL: http://arxiv.org/abs/1003.3146, doi:10.1103/PhysRevD.82.034001.

[72] Ringaile Placakyte. Parton Distribution Functions. 2011. URL: http://arxiv.org/abs/1111.5452.

[73] T. Gleisberg, S. Hche, F. Krauss, M. Schnherr, S. Schumann, F. Siegert, and J. Winter. Event generation with SHERPA 1.1. *Journal of High Energy Physics*, 2009(2):7, 2009. URL: http://stacks.iop.org/1126-6708/2009/i=02/a=007, doi:10.1088/1126-6708/2009/02/007.

[74] Richard P. Feynman. Very High-Energy Collisions of Hadrons. *Phys. Rev. Lett.*, 23:1415–1417, Dec 1969. URL: http://link.aps.org/doi/10.1103/PhysRevLett.23.1415, doi:10.1103/PhysRevLett.23.1415.

[75] J. D. Bjorken and E. A. Paschos. Inelastic Electron-Proton and γ-Proton Scattering and the Structure of the Nucleon. *Phys. Rev.*, 185:1975–1982, Sep 1969. `doi:10.1103/PhysRev.185.1975`.

[76] Y. L. Dokshitzer. Calculation of the Structure Functions for Deep Inelastic Scattering and e+ e- Annihilation by Perturbation Theory in Quantum Chromodynamics. *Sov.Phys.JETP*, 46:641–653, 1977.

[77] V. N. Gribov and L. N. Lipatov. Deep inelastic e p scattering in perturbation theory. *Sov.J.Nucl.Phys.*, 15:438–450, 1972.

[78] L. N. Lipatov. The parton model and perturbation theory. *Sov.J.Nucl.Phys.*, 20:94–102, 1975.

[79] G. Altarelli and G. Parisi. Asymptotic freedom in parton language. *Nucl.-Phys.*, B(2):298–318, 1977. URL: `http://adsabs.harvard.edu/abs/1977NuPhB.126..298A`.

[80] H1 and ZEUS Collaboration. Combined measurement and QCD analysis of the inclusive e$^{\pm}$p scattering cross sections at HERA. *Journal of High Energy Physics*, 2010:1–63, 2010. 10.1007/JHEP01(2010)109. URL: `http://dx.doi.org/10.1007/JHEP01(2010)109`.

[81] CTEQ Collaboration. New Generation of Parton Distributions with Uncertainties from Global QCD Analysis. *Journal of High Energy Physics*, 2002(07):012, 2002. URL: `http://stacks.iop.org/1126-6708/2002/i=07/a=012`.

[82] A.D. Martin, W.J. Stirling, R.S. Thorne, and G. Watt. Heavy-quark mass dependence in global PDF analyses and 3- and 4-flavour parton distributions. *Eur.Phys.J.C*, 70:51–72, 2010. `arXiv:1007.2624`, `doi:10.1140/epjc/s10052-010-1462-8`.

[83] Richard D. Ball et al. Parton distributions for the LHC Run II. *JHEP*, 04:040, 2015. `arXiv:1410.8849`, `doi:10.1007/JHEP04(2015)040`.

[84] ZEUS Collaboration: S. Chekanov et al. Measurement of the neutral current cross section and $\mathbf{F_2}$ structure function for deep inelastic $\mathbf{e^+p}$ scattering at HERA. *The European Physical Journal C - Particles and Fields*, 21(3):443–471, 2001. `doi:10.1007/s100520100749`.

[85] H1 Collaboration: C. Adloff et al. Deep-inelastic inclusive ep scattering at low x and a determination of α_s. *The European Physical Journal C - Particles and Fields*, 21(1):33–61, 2001. `doi:10.1007/s100520100720`.

[86] DØ Collaboration: B. Abbott et al. High-p_T jets in $\overline{p}p$ collisions at $\sqrt{s} = 630$ and 1800 GeV. *Phys. Rev. D*, 64:032003, Jul 2001. `doi:10.1103/PhysRevD.64.032003`.

[87] T. Affolder et al. Measurement of the inclusive jet cross section in $p\overline{p}$ collisions at $\sqrt{s} = 1.8$ TeV. *Phys. Rev. D*, 64:032001, Jun 2001. `doi:10.1103/PhysRevD.64.032001`.

[88] A.C. et al. Benvenuti. A High Statistics Measurement of the Proton Structure Functions F(2) (x Q^2) and R from Deep Inelastic Muon Scattering at High Q^2. *Phys. Lett. B*, 223:485, 1989. `doi:10.1016/0370-2693(89)91637-7`.

[89] M. Arneodo et al. Accurate measurement of F2(d) / F2(p) and R^d - R^p. *Nucl.Phys. B*, 487:3–26, 1997. `arXiv:9611022, doi:10.1016/S0550-3213(96)00673-6`.

[90] U. K. Yang et al. Measurements of F_2 and $xF_3^\nu - xF_3^{\overline{\nu}}$ from CCFR ν_μ-Fe and $\overline{\nu}_\mu$-Fe Data in a Physics Model-Independent Way. *Phys. Rev. Lett.*, 86:2742–2745, Mar 2001. `doi:10.1103/PhysRevLett.86.2742`.

[91] Maltoni, Fabio and McElmurry, Thomas and Putman, Robert and Willenbrock, Scott. Choosing the Factorization Scale in Perturbative QCD. 2007. `arXiv:0703156`.

[92] CTEQ Collaboration. Handbook of perturbative qcd. *Rev. Mod. Phys.*, 67:157–248, Jan 1995. `doi:10.1103/RevModPhys.67.157`.

[93] Rick Field. Min-Bias and the Underlying Event at the LHC. 2012. `arXiv:1202.0901`.

[94] P. Bartalini et al. Multi-Parton Interactions at the LHC. 2011. `arXiv:1111.0469`.

[95] Makoto Kobayashi and Toshihide Maskawa. CP-Violation in the Renormalizable Theory of Weak Interaction. *Progress of Theoretical Physics*, 49(2):652–657, 1973. `doi:10.1143/PTP.49.652`.

[96] DØ Collaboration. Observation of the Top Quark. *Phys.Rev.Lett.*, 74(14):2632–2637, 1995. `arXiv:9503003`.

[97] CDF Collaboration. Observation of Top Quark Production in $p\bar{p}$ Collisions with the Collider Detector at Fermilab. *Phys.Rev.Lett.*, 74:2626–2631, 1995. `arXiv:hep-ex/9503002, doi:10.1103/PhysRevLett.74.2626`.

[98] W. Bernreuther. Top-quark physics at the LHC. *Journal of Physics G: Nuclear and Particle Physics*, 35(8):083001, 2008. `doi:10.1088/0954-3899/35/8/083001`.

[99] Joseph R. Incandela, Arnulf Quadt, Wolfgang Wagner, and Daniel Wicke. Status and prospects of top-quark physics. *Progress in Particle and Nuclear Physics*, 63(2):239–292, 2009. `doi:10.1016/j.ppnp.2009.08.001`.

[100] Frank-Peter Schilling. Top quark physics at the LHC: a review of the first two years. *International Journal of Modern Physics A*, 27(17):1230016, July 2012. `doi:10.1142/S0217751X12300165`.

[101] Nikolaos Kidonakis. NNLL threshold resummation for top-pair and single-top production. *Phys.Part.Nucl.*, 45(4):714–722, 2014. `arXiv:1210.7813, doi:10.1134/S1063779614040091`.

[102] Micha Czakon, Paul Fiedler, and Alexander Mitov. Total Top-Quark Pair-Production Cross Section at Hadron Colliders Through $O(\alpha_s^4)$. *Phys.Rev.Lett.*, 110:252004, 2013. `arXiv:1303.6254, doi:10.1103/PhysRevLett.110.252004`.

[103] M. Beneke, P. Falgari, and C. Schwinn. Soft radiation in heavy-particle pair production: All-order colour structure and two-loop anomalous dimension. *Nucl.Phys.B*, 828:69–101, 2010. `arXiv:0907.1443`, `doi:10.1016/j.nuclphysb.2009.11.004`.

[104] Michal Czakon, Alexander Mitov, and George F. Sterman. Threshold Resummation for Top-Pair Hadroproduction to Next-to-Next-to-Leading Log. *Phys.Rev.D*, 80:074017, 2009. `arXiv:0907.1790`, `doi:10.1103/PhysRevD.80.074017`.

[105] Matteo Cacciari, Michal Czakon, Michelangelo Mangano, Alexander Mitov, and Paolo Nason. Top-pair production at hadron colliders with next-to-next-to-leading logarithmic soft-gluon resummation. *Phys.Lett.B*, 710:612–622, 2012. `arXiv:1111.5869`, `doi:10.1016/j.physletb.2012.03.013`.

[106] Tevatron New Physics Higgs Working Group, CDF Collaboration, DØ Collaboration. Updated Combination of CDF and D0 Searches for Standard Model Higgs Boson Production with up to 10.0 fb^{-1} of Data. 2012. `arXiv:1207.0449`.

[107] LHC Higgs Cross Section WG Picture Gallery. `https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWGCrossSectionsFigures`.

[108] Nima Arkani-Hamed, Andrew G. Cohen, and Howard Georgi. Electroweak symmetry breaking from dimensional deconstruction. *Phys.Lett.B*, 513:232–240, 2001. `arXiv:hep-ph/0105239`, `doi:10.1016/S0370-2693(01)00741-9`.

[109] N. Arkani-Hamed, A.G. Cohen, E. Katz, and A.E. Nelson. The Littlest Higgs. *JHEP*, 0207:034, 2002. `arXiv:hep-ph/0206021`, `doi:10.1088/1126-6708/2002/07/034`.

[110] Marcela S. Carena, Eduardo Ponton, Jose Santiago, and Carlos E.M. Wagner. Light Kaluza Klein States in Randall-Sundrum Models with Custodial SU(2). *Nucl.Phys.B*, 759:202–227, 2006. `arXiv:hep-ph/0607106`, `doi:10.1016/j.nuclphysb.2006.10.012`.

[111] Roberto Contino, Leandro Da Rold, and Alex Pomarol. Light custodians in natural composite Higgs models. *Phys.Rev.*, D75:055014, 2007. `arXiv:hep-ph/0612048`, `doi:10.1103/PhysRevD.75.055014`.

[112] CMS Collaboration. Inclusive search for a vector-like T quark by CMS. Technical Report CMS-PAS-B2G-12-015, CERN, Geneva, 2013.

[113] ATLAS Collaboration. Search for heavy top-like quarks decaying to a Higgs boson and a top quark in the lepton plus jets final state in *pp* collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. Technical Report ATLAS-CONF-2013-018, CERN, Geneva, Mar 2013.

[114] S Heinemeyer et al. Handbook of LHC Higgs Cross Sections: 3. Higgs Properties. 2013. `arXiv:1307.1347`, `doi:10.5170/CERN-2013-004`.

[115] ATLAS Collaboration. Search for $H \to \gamma\gamma$ produced in association with top quarks and constraints on the Yukawa coupling between the top quark and the Higgs boson using data taken at 7 TeV and 8 TeV with the ATLAS detector. *Phys.Lett.B*, 740:222–242, 2014. `arXiv:1409.3122`, `doi:10.1016/j.physletb.2014.11.049`.

[116] ATLAS Collaboration. Search for the Standard Model Higgs boson produced in association with top quarks and decaying to $b\bar{b}$ in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector at the LHC. Technical Report ATLAS-CONF-2014-011, CERN, Geneva, Mar 2014.

[117] CMS Collaboration. Search for the associated production of the Higgs boson with a top-quark pair. *JHEP*, 1409:087, 2014. `arXiv:1408.1682, doi:10.1007/JHEP09(2014)087,10.1007/JHEP10(2014)106`.

[118] ATLAS collaboration. Study of heavy-flavor quarks produced in association with top-quark pairs at $\sqrt{s} = 7$TeV using the ATLAS detector. *Phys.Rev.*, D89(7):072012, 2014. `arXiv:1304.6386, doi:10.1103/PhysRevD.89.072012`.

[119] CMS Collaboration. Measurement of the cross section ratio $\sigma_{t\bar{t}b\bar{b}}/\sigma_{t\bar{t}jj}$ in pp collisions at $\sqrt{s} = 8$TeV. 2014. `arXiv:1411.5621`.

[120] W-M Yao etal. Review of particle physics. *Journal of Physics G: Nuclear and Particle Physics*, 33(1):1, 2006. URL: `http://stacks.iop.org/0954-3899/33/i=1/a=001`.

[121] Serge Dailler. Hydro-geologic map of the LHC area. Carte hydrogeologique de la zone des ouvrages du CERN. LHC-PHO-1997-173. AC Collection. Legacy of AC. Pictures from 1992 to 2002., Jul 1997. URL: `http://cds.cern.ch/record/842406/`.

[122] Public CMS Luminosity Information. URL: `https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults`.

[123] Fabienne Marcastel. CERN's Accelerator Complex. La chaîne des accélérateurs du CERN. General Photo. OPEN-PHO-CHART-2013-001, Oct 2013. URL: `https://cds.cern.ch/record/1621583`.

[124] C Lefevre. LHC: the guide. Jan 2008. URL: `http://cds.cern.ch/record/1092437`.

[125] The accelerator complex. URL: `http://home.web.cern.ch/about/accelerators`.

[126] Pauline Gagnon. Is the moon full? Just ask the LHC operators. Quantum Diaries, June 2012. URL: `http://www.quantumdiaries.org/2012/06/07/is-the-moon-full-just-ask-the-lhc-operators/`.

[127] 3D SketchUp images of the CMS detector. URL: `https://cms-docdb.cern.ch/cgi-bin/PublicDocDB//ShowDocument?docid=11514`.

[128] ALEPH Collaboration. ALEPH: A detector for electron-positron annnihilations at LEP. *Nucl.Instrum.Meth.A*, 294:121–178, 1990. `doi:10.1016/0168-9002(90)91831-U`.

[129] DELPHI Collaboration. The DELPHI detector at LEP. *Nucl.Instrum.Meth.A*, 303:233–276, 1991. `doi:10.1016/0168-9002(91)90793-P`.

[130] ZEUS Collaboration. The ZEUS detector, 1993. Full, scanned version of the Status Report. URL: `http://www-zeus.desy.de/bluebook/bluebook.html`.

[131] Bernard Aubert et al. The BaBar detector. *Nucl.Instrum.Meth.A*, 479:1–116, 2002. `arXiv:hep-ex/0105044`, `doi:10.1016/S0168-9002(01)02012-5`.

[132] V.I. Klyukhin et al. Measurement of the CMS Magnetic Field. *Applied Superconductivity, IEEE Transactions on*, 18(2):395–398, June 2008. `doi:10.1109/TASC.2008.921242`.

[133] CMS Collaboration. Precise mapping of the magnetic field in the CMS barrel yoke using cosmic rays. *Journal of Instrumentation*, 5(03):T03021, 2010. URL: `http://stacks.iop.org/1748-0221/5/i=03/a=T03021`.

[134] CMS Collaboration. *CMS Physics: Technical Design Report. Volume 1: Detector Performance and Software*. Technical Design Report CMS. CERN, Geneva, 2006. URL: `http://cds.cern.ch/record/922757`.

[135] P Adzic. Energy resolution of the barrel of the CMS Electromagnetic Calorimeter. *Journal of Instrumentation*, 2(04):P04004, 2007. URL: `http://stacks.iop.org/1748-0221/2/i=04/a=P04004`.

[136] CMS Collaboration. Energy calibration and resolution of the CMS electromagnetic calorimeter in pp collisions at *sqrts* = 7 TeV. *Journal of Instrumentation*, 8(09):P09009, 2013. URL: `http://stacks.iop.org/1748-0221/8/i=09/a=P09009`.

[137] CMS Collaboration. Performance of the CMS hadron calorimeter with cosmic ray muons and LHC beam data. *Journal of Instrumentation*, 5(03):T03012, 2010. URL: `http://stacks.iop.org/1748-0221/5/i=03/a=T03012`.

[138] CMS Collaboration. CMS Physics Technical Design Report, Volume II: Physics Performance. *Journal of Physics G: Nuclear and Particle Physics*, 34(6):995, 2007. URL: `http://stacks.iop.org/0954-3899/34/i=6/a=S01`.

[139] CMS Collaboration. *CMS TriDAS project: Technical Design Report, Volume 1: The Trigger Systems*. Technical Design Report CMS. 2000. URL: `http://cds.cern.ch/record/706847`.

[140] CMS Collaboration. Interactive Slice of the CMS detector. [Unpublished. Online, accessed 26 Mar. 2015]. URL: `https://cms-docdb.cern.ch/cgi-bin/DocDB/ShowDocument?docid=4172`.

[141] CMS Physics Performance and Dataset Group. CMS DQM graphical user interface. [Unpublished. Online, accessed 27 Mar. 2015]. URL: `https://cmsweb.cern.ch/dqm/offline`.

[142] CMS Collaboration. Certified collision data from 2012. [Lists of run and luminosity-section ranges in the `JSON` format]. URL: `https://cms-service-dqm.web.cern.ch/cms-service-dqm/CAF/certification/Collisions12/8TeV/`.

[143] R.L. Glückstern. Uncertainties in track momentum and direction, due to multiple scattering and measurement errors. *Nuclear Instruments and Methods*, 24(0):381–389, 1963. `doi:10.1016/0029-554X(63)90347-1`.

[144] Markus Stoye. *Calibration and Alignment of the CMS Silicon Tracking Detector*. PhD thesis, Deutsches Elektronen-Synchrotron, (DESY), July 2007. DESY-THESIS-2007-026. URL: http://www.physnet.uni-hamburg.de/services/fachinfo/dissfb12_2007.html.

[145] M. Atac et al. Beam test results of the us-cms forward pixel detector. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 488(1–2):271–281, 2002. doi:10.1016/S0168-9002(02)00472-2.

[146] Johannes Hauk. *Measurement of Associated Z^0-Boson and b-Jet Production in Proton-Proton Collisions with the CMS Experiment*. PhD thesis, Deutsches Elektronen-Synchrotron, (DESY), 2012. DESY-THESIS-2012-030. URL: https://cms.desy.de/e128524/e205118/Johannes___Hauk.pdf.

[147] P. Lenzi, C. Genta, and B. Mangano. Track reconstruction of real cosmic muon events with CMS tracker detector. *Journal of Physics: Conference Series*, 119(3):032030, 2008. URL: http://stacks.iop.org/1742-6596/119/i=3/a=032030.

[148] R. Frühwirth, M. Regler, R. K. Bock, H. Grote, and D. Notz. *Data Analysis Techniques for High-Energy Physics*. Cambridge Monographs on Particle Physics, Nuclear Physics and Cosmology. Cambridge University Press, 2000. URL: http://www.cambridge.org/ar/academic/subjects/physics/particle-physics-and-nuclear-physics/data-analysis-techniques-high-energy-physics-2nd-edition.

[149] Kevin Stenson. Iterative Tracking Steps. CMS Twiki, Nov. 2012. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideIterativeTracking.

[150] Edmund Erich Widl. *Global Alignment of the CMS Tracker*. PhD thesis, Vienna University of Technology, 2008. CERN-THESIS-2008-123, CMS-TS-2008-026.

[151] Bruno Wittmer, Roman Adolphi, Michael Henke, Andrei Ostaptchouk, Maarten Thomas, et al. The laser alignment system for the CMS silicon microstrip tracker. *Nucl.Instrum.Meth.A*, 581:351–355, 2007. doi:10.1016/j.nima.2007.08.002.

[152] CMS Collaboration. Impact of Tracker Misalignment on the CMS b-Tagging Performance. *PAS*, BTV-07-003, Aug 2009. http://cdsweb.cern.ch/record/1198691/.

[153] CMS Collaboration. Alignment of the CMS tracker with LHC and cosmic ray data. *JINST*, 9:P06009, 2014. arXiv:1403.2286, doi:10.1088/1748-0221/9/06/P06009.

[154] V. Blobel, C. Kleinwort, and F. Meier. Fast alignment of a complex tracking detector using advanced track models. Trondheim, 2010. Conference on Computational Physics. URL: http://cms.desy.de/sites/site_cms/content/e53612/e234315/e241962/e241763/posterCCP2010_brokenlines.pdf.

[155] Claus Kleinwort. General broken lines as advanced track fitting method. *Nuclear Instruments and Methods in Physics Research*, A673:107–110, May 2012. `arXiv:1201.4320`, `doi:10.1016/j.nima.2012.01.024`.

[156] A. Dorokhov, Claude Amsler, D. Bortoletto, V. Chiochia, L. Cremaldi, et al. Tests of silicon sensors for the CMS pixel detector. *Nucl.Instrum.Meth.*, A530:71–76, 2004. `arXiv:physics/0311050`, `doi:10.1016/j.nima.2004.05.050`.

[157] B. Henrich and R. Kaufmann. Lorentz-angle in irradiated silicon. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 477(1–3):304–307, 2002. 5th Int. Conf. on Position-Sensitive Detectors. URL: `http://www.sciencedirect.com/science/article/pii/S0168900201018654`, `doi:http://dx.doi.org/10.1016/S0168-9002(01)01865-4`.

[158] L. Wilke, V. Chiochia, and T. Speer. Lorentz angle calibration for the barrel pixel detector. 2008.

[159] Jelena Luetić and Darko Mekterović. Lorentz angle measurement in CMS Pixel Barrel Detector. CMS Note 2013/371, 2013.

[160] Vitaliano Ciulli, Raffaello D'Alessandro, Simone Frosali, and Chiara Genta. Determination of the Lorentz Angle in Microstrip Silicon Detectors with Cosmic Muons. Technical Report CMS-NOTE-2008-006, CERN, Geneva, Sep 2007. URL: `http://cds.cern.ch/record/1095045`.

[161] CMS Collaboration. Description and performance of track and primary-vertex reconstruction with the CMS tracker. *JINST*, 9(10):P10009, 2014. `arXiv:1405.6569`, `doi:10.1088/1748-0221/9/10/P10009`.

[162] Andrea Venturi. SiStrip Back Plane Correction Analysis, 2012. [Online, accessed on 10 April, 2015]. URL: `https://twiki.cern.ch/twiki/bin/view/CMS/BackPlaneCorrections`.

[163] Ben Hooberman. Investigation of the backplane correction, Jan. 2012. [Online, accessed on 10 April, 2015]. URL: `https://indico.cern.ch/event/167355/contribution/6`.

[164] Nazar Bartosik. Simultaneous alignment and Lorentz angle calibration in the CMS silicon tracker using Millepede II. *PoS*, EPS-HEP2013:074, 2013.

[165] CMS Collaboration. CMS Tracker Detector Performance Results: Alignment. [Online, accessed on 15 April, 2015]. URL: `https://twiki.cern.ch/twiki/bin/viewauth/CMSPublic/TkAlignmentPerformance2012`.

[166] Armin Burgmeier. Explanation and results of the hit-position-resolution estimatation in BPIX, Jun 2013. [Private communication].

[167] Morris Swartz and Danek Kotlinski. Discussion about Lorentz angle difference in forward and backward regions of BPIX, 2013–2015. [Private communication].

[168] V. Veszpremi. Operation and performance of the CMS tracker. *JINST*, 9:C03005, 2014. `arXiv:1402.0675`, `doi:10.1088/1748-0221/9/03/C03005`.

[169] Martin Görner and Peter Schleper. *Differential Cross Sections for Top-Quark-Pair Production in the e/μ+Jets Final State at* $\sqrt{s} = 8$ *TeV in CMS*. PhD thesis, Hamburg U., 2014. URL: http://cds.cern.ch/record/1754332.

[170] J. Alwall, M. Herquet, F. Maltoni, O. Mattelaer, and T. Stelzer. MadGraph 5: Going beyond. *JHEP*, 1106:128, 2011. arXiv:1106.0522, doi:10.1007/JHEP06(2011)128.

[171] P. Artoisenet, R. Frederix, O. Mattelaer, and R. Rietkerk. Automatic spin-entangled decays of heavy resonances in Monte Carlo simulations. *JHEP*, 1303:015, 2013. arXiv:1212.3460, doi:10.1007/JHEP03(2013)015.

[172] Zbigniew Was. TAUOLA for simulation of tau decay and production: perspectives for precision low energy and LHC applications. *IFJPAN*, IV-2011-1, 2011. http://arxiv.org/abs/1101.1652.

[173] S. Alioli, S.-O. Moch, and P. Uwer. Hadronic top-quark pair-production with one jet and parton showering. *JHEP*, 1201:137, 2012. arXiv:1110.5251, doi:10.1007/JHEP01(2012)137.

[174] S. Frixione and B. R. Webber. Matching NLO QCD computations and parton shower simulations. *JHEP*, 0206:029, 2002. arXiv:hep-ph/0204244, doi:10.1088/1126-6708/2002/06/029.

[175] T. Sjostrand, S. Mrenna, and P. Z. Skands. PYTHIA 6.4 physics and manual. *JHEP*, 0605:026, 2006. arXiv:hep-ph/0603175, doi:10.1088/1126-6708/2006/05/026.

[176] M. L. Mangano, M. Moretti, F. Piccinini, and M. Treccani. Matching matrix elements and shower evolution for top-quark production in hadronic collisions. *JHEP*, 0701:013, 2007. arXiv:hep-ph/0611129, doi:10.1088/1126-6708/2007/01/013.

[177] S. Frixione, P. Nason, and C. Oleari. Matching NLO QCD computations with Parton Shower simulations: the POWHEG method. *JHEP*, 0711:070, 2007. arXiv:0709.2092, doi:10.1088/1126-6708/2007/11/070.

[178] B. Andersson. *The Lund model*. Cambridge University Press, 1998. Cambridge Books Online. URL: http://dx.doi.org/10.1017/CBO9780511524363.

[179] Ivan Asin Cruz. *Measurement of Top-Quark-Pair Differential Cross Sections in Proton-Proton Collisions at* $\sqrt{s} = 8$ *TeV with the CMS Experiment*. PhD thesis, Deutsches Elektronen-Synchrotron, (DESY), 2014. DESY-THESIS-2014-045. URL: http://www-library.desy.de/cgi-bin/showprep.pl?desy-thesis-14-045.

[180] A. Buckley, J. Butterworth, S. Gieseke, D. Grellscheid, S. Hoche, et al. General-purpose event generators for LHC physics. *Phys. Rept.*, 504:145–233, 2011. arXiv:1101.2599, doi:10.1016/j.physrep.2011.03.005.

[181] D. Amati and G. Veneziano. Preconfinement as a property of perturbative QCD. *Phys. Lett.B*, 83:87, 1979. doi:10.1016/0370-2693(79)90896-7.

[182] G. Corcella et al. HERWIG 6.5 Release Note. *arXiv*, hep-ph/0210213v2, 2005. URL: http://arxiv.org/abs/hep-ph/0210213.

[183] CMS Collaboration. Measurement of the underlying Event Activity at the LHC with $\sqrt{s} = 7\,\text{TeV}$ and comparison with $\sqrt{s} = 0.9$ TeV. *JHEP*, 1109:109, 2011. `arXiv:1107.0330`, `doi:10.1007/JHEP09(2011)109`.

[184] ATLAS Collaboration. ATLAS tunes of PYTHIA 6 and Pythia 8 for MC11. 2011. ATL-PHYS-PUB-2011-009.

[185] S. Agostinelli et al. Geant4—a simulation toolkit. *Nucl.Inst.Meth.*, A506(3):250–303, 2003. URL: `http://www.sciencedirect.com/science/article/B6TJM-48TJFY8-5/2/23ea98096ce11c1be446850c04cfa498`.

[186] J. Allison et al. Geant4 developments and applications. *IEEE Trans. Nucl. Sci.*, 53:270, 2006. URL: `http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1610988`.

[187] Top Physics Analysis Group. Top PAG reference selections. [Unpublished. Online, accessed 26 Mar. 2015]. URL: `https://twiki.cern.ch/twiki/bin/view/CMS/TopEventSelectionRun1`.

[188] CMS Collaboration. Particle-flow event reconstruction in CMS and performance for jets, taus, and MET. 2009. CMS-PAS-PFT-09-001.

[189] The CMS Collaboration. Description and performance of track and primary-vertex reconstruction with the CMS tracker. *Journal of Instrumentation*, 9(10):P10009, 2014. URL: `http://stacks.iop.org/1748-0221/9/i=10/a=P10009`.

[190] CMS Collaboration. Public CMS Luminosity Information. [Unpublished. Online, accessed 27 Mar. 2015]. URL: `https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults`.

[191] The Top Quark Physics Object Group. CMS Public Software Guide - Particle Flow Implementation in the Physics Analysis Toolkit. [Unpublished. Online, accessed 29 May, 2015]. URL: `https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuidePF2PAT`.

[192] Florian Beaudette. Performance of the Prticle Flow Algorithm in CMS. [Unpublished. Online, accessed 29 May, 2015]. URL: `http://indico.cern.ch/event/73513/session/45/contribution/825`.

[193] Serguei Chatrchyan et al. Performance of CMS muon reconstruction in *pp* collision events at $\sqrt{s} = 7\,\text{TeV}$. *JINST*, 7:P10002, 2012. `arXiv:1206.4071`, `doi:10.1088/1748-0221/7/10/P10002`.

[194] CMS Collaboration. Muon Analysis. [Unpublished. Online, accessed 27 Mar. 2015]. URL: `https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookMuonAnalysis`.

[195] R. Fruhwirth. Application of Kalman filtering to track and vertex fitting. *Nucl.Instrum.Meth.A*, 262:444–450, 1987. `doi:10.1016/0168-9002(87)90887-4`.

[196] S. Baffioni, C. Charlot, F. Ferri, D. Futyan, P. Meridiani, et al. Electron reconstruction in CMS. *Eur.Phys.J.C*, 49:1099–1116, 2007. `doi:10.1140/epjc/s10052-006-0175-5`.

[197] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-k_t jet clustering algorithm. *JHEP*, 04:063, 2008. `arXiv:0802.1189`, `doi:10.1088/1126-6708/2008/04/063`.

[198] The Top Quark Physics Object Group. Top projections, or avoiding double-counting. [Unpublished. Online, accessed 29 May 2015]. URL: `https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuidePF2PAT#Top_projection_or_avoiding_doubl`.

[199] The CMS collaboration. Determination of jet energy calibration and transverse momentum resolution in CMS. *Journal of Instrumentation*, 6(11):P11002, 2011. URL: `http://stacks.iop.org/1748-0221/6/i=11/a=P11002`.

[200] CMS Jet Energy Resolution and Corrections Subgroup. Jet Energy Resolution and Corrections. [Unpublished. Online, accessed 29 May 2015]. URL: `https://twiki.cern.ch/twiki/bin/viewauth/CMS/JetEnergyScale`.

[201] K. Goebel, J. Haller, J. Ott, and H. Stadie. Jet transverse momentum resolution measurement using dijet events at $\sqrt{s} = 8\,\text{TeV}$. 2014. [Unpublished].

[202] P. Azzurri, J. Bernardini, D. Caiulo, S. Donato, G. Rauco, P. Spagnolo, C. Vernieri S. Alderweireldt, S. Bansal, X. Janssen, T. Cornelis, J. Lauwers, N. van Remortel S. de Visscher, and K. Kousouris. Updated search for the standard model Higgs boson produced in vector boson fusion and decaying to bottom quarks using the full $\sqrt{s} = 8\,\text{TeV}$ 2012 data sample. 2014. CMS-AN-2013/358.

[203] Serguei Chatrchyan et al. Identification of b-quark jets with the CMS experiment. *JINST*, 8:P04013, 2013. `arXiv:1211.4462`, `doi:10.1088/1748-0221/8/04/P04013`.

[204] CMS Collaboration. Performance of b tagging at sqrt(s)=8 TeV in multijet, ttbar and boosted topology events. 2013. CMS-PAS-BTV-13-001.

[205] CMS Collaboration. Performance of missing transverse momentum reconstruction algorithms in proton-proton collisions at $\sqrt{s} = 8\,\text{TeV}$ with the CMS detector. 2012. CMS-PAS-JME-12-002.

[206] J. M. H. Salfeld-Nebgen. *Search for the Higgs boson decaying into τ-leptons in the di-electron channel*. PhD thesis, University of Hamburg, 2014. DESY-THESIS-2014-019.

[207] Jet Flavour Identification (MC Truth). `https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideBTagMCTools`.

[208] Nazar Bartosik and Johannes Hauk. Hadron based origin identification of heavy flavour jets at generator level. `https://twiki.cern.ch/twiki/bin/view/CMSPublic/GenHFHadronMatcher`.

[209] J. Garay Garcia. Jet charge studies in the context of $t\bar{t}H(b\bar{b})$ and $t\bar{t}b\bar{b}$ analyses. 2015. CMS AN-2015/049.

[210] K.A. Olive et al. (Particle Data Group). PDG Particle Listings. *Chin. Phys. C*, 38, 2014. URL: `http://pdg.lbl.gov/2014/mcdata/mc_particle_id_contents.html`.

[211] Axel Bredenstein, Ansgar Denner, Stefan Dittmaier, and Stefano Pozzorini. Production of $t\bar{t}b\bar{b}$ at the LHC at NLO QCD. *Nucl.Phys.Proc.Suppl.*, 205–206:80–85, 2010. `arXiv:1006.2653`, `doi:10.1016/j.nuclphysbps.2010.08.023`.

[212] Fabio Cascioli, Philipp Maierhfer, Niccolo Moretti, Stefano Pozzorini, and Frank Siegert. NLO matching for $t\bar{t}b\bar{b}$ production with massive b-quarks. *Phys.Lett.B*, 734:210–214, 2014. `arXiv:1309.5912`, `doi:10.1016/j.physletb.2014.05.040`.

[213] M.V. Garzelli, A. Kardos, and Z. Trcsnyi. Hadroproduction of $t\bar{t}b\bar{b}$ final states at LHC: predictions at NLO accuracy matched with Parton Shower. *JHEP*, 1503:083, 2015. `arXiv:1408.0266`, `doi:10.1007/JHEP03(2015)083`.

[214] Combination of ATLAS and CMS ttbar charge asymmetry measurements using LHC proton-proton collisions at 7 TeV. Technical Report CMS-PAS-TOP-14-006, CERN, Geneva, 2014. URL: `https://cds.cern.ch/record/1670517`.

[215] N. Kidonakis. NNLL threshold resummation for top-pair and single-top production. *Phys. Part. Nucl.*, 45(4):714–722, 2014. `arXiv:1210.7813`, `doi:10.1134/S1063779614040091`.

[216] M.V. Garzelli, A. Kardos, C.G. Papadopoulos, and Z. Trocsanyi. $t\bar{t}$ $W^{\pm}$ and $t\bar{t}$ Z hadroproduction at NLO accuracy in QCD with parton shower and hadronization effects. *JHEP*, 1211:056, 2012. `arXiv:1208.2665`, `doi:10.1007/JHEP11(2012)056`.

[217] J. M. Campbell and R. K. Ellis. $t\bar{t}$ $W^{\pm}$ production and decay at NLO. *JHEP*, 1207:052, 2012. `arXiv:1204.5678`, `doi:10.1007/JHEP07(2012)052`.

[218] K. Melnikov, M. Schulze, and A. Scharf. QCD corrections to top quark pair production in association with a photon at hadron colliders. *Phys. Rev.*, D83:074013, 2011. `arXiv:1102.1967`, `doi:10.1103/PhysRevD.83.074013`.

[219] The CMS Collaboration. Measurement of the $t\bar{t}$ production cross section in the dilepton channel in pp collisions at $\sqrt{s} = 8$ TeV. *JHEP*, 1402:024, 2014. `arXiv:1312.7582`, `doi:10.1007/JHEP02(2014)024`.

[220] J. M. Campbell, R. K. Ellis, and C. Williams. Vector boson pair production at the LHC. *JHEP*, 1107:018, 2011. `arXiv:1105.0020`, `doi:10.1007/JHEP07(2011)018`.

[221] Piergiulio Lenzi, Sanjay Padhi, Guillelmo Gomez Ceballos Retuerto, Frank Wuerthwein, Tomislav Seva, and Henry Yee-Shian Tong. Standard Model Cross Sections for CMS at 8 TeV. `https://twiki.cern.ch/twiki/bin/viewauth/CMS/StandardModelCrossSectionsat8TeV`.

[222] T. Sjostrand, S. Mrenna, and P. Z. Skands. PYTHIA 6.4 physics and manual. *JHEP*, 0605:026, 2006. `arXiv:hep-ph/0603175`, `doi:10.1088/1126-6708/2006/05/026`.

[223] Ivan Asin, Nazar Bartosik, Olaf Behnke, Carmen Pardos Diez, Ganna Dolinska, Tyler Dorland, Johannes Hauk, Jan Kieseler, Ievgen Korol, and Andreas B. Meyer. Measurement of differential top-quark pair production cross sections in the dilepton final state at $\sqrt{s} = 8$ TeV. CMS Note 2013/266, 2013.

[224] Volker Adler. Multivariate Electron Identification. [Unpublished. Online, accessed 8 April, 2015]. URL: https://twiki.cern.ch/twiki/bin/view/CMS/MultivariateElectronIdentification.

[225] Michael Hildreth. Pileup Scenarios. [Unpublished. Online, accessed on 10 March 2015]. URL: https://twiki.cern.ch/twiki/bin/viewauth/CMS/Pileup_MC_Gen_Scenarios.

[226] Michael Hildreth. Utilities for Accessing Pileup Information for Data. [Unpublished. Online, accessed on 10 March 2015]. URL: https://twiki.cern.ch/twiki/bin/viewauth/CMS/PileupJSONFileforData#Calculating_Your_Pileup_Distribu.

[227] Maria Aldaya, Ivan Asin, Nazar Bartosik, Carmen Pardos Diez, Ganna Dolinska, Jasone Garay Garcia, Johannes Hauk, Jan Kieseler, Ievgen Korol, and Eleni Ntomari. Measurement of $t\bar{t}$ production with additional jet activity and properties of the additional jets, including $t\bar{t}b\bar{b}$, in the dilepton channel at $\sqrt{s} = 8\,\text{TeV}$. CMS Note 2014/274, 2014.

[228] N. Bartosik, S. Boutle, A. Brinkerhoff, A. Descroix, T. Dorland, J. Garay Garcia, J. Hauk, M. Hildreth, R. Hughes, U. Husemann, C. Jessop, J. Kieseler, J. Kolb, K. Lannon, W. Luo, R. Mankel, A. Meyer, H. Mildner, C. Neu, C. Diez Pardos, P. Lobelle Pardo, D. Puigh, A. Raspereza, J. Slaunwhite, J. Smith, N. Valls, J. Wagner-Kuhr, S. Williamson, B. Winer, M. Wolf, J. Wood, and A. Woodard. Calibration of the Combined Secondary Vertex b-Tagging discriminant using dileptonic ttbar and Drell-Yan events. CMS Note 2013/130, 2013.

[229] Nazar Bartosik. CSV reweighting of light jets, Aug. 2014. [ttH bb+tau Working Meeting. Online, accessed 10 March 2015]. URL: https://indico.cern.ch/event/333571/session/10/contribution/11.

[230] Giovanni Petrucciani. Documentation of the roostats-based statistics tools for higgs pag. https://twiki.cern.ch/twiki/bin/viewauth/CMS/SWGuideHiggsAnalysisCombinedLimit.

[231] Lorenzo Moneta, Kevin Belasco, Kyle S. Cranmer, S. Kreiss, Alfio Lazzaro, et al. The RooStats Project. *PoS*, ACAT2010:057, 2010. arXiv:1009.1003.

[232] Serguei Chatrchyan et al. Measurement of the cross section and angular correlations for associated production of a Z boson with b hadrons in pp collisions at $\sqrt{s} = 7\,\text{TeV}$. *JHEP*, 1312:039, 2013. arXiv:1310.1349, doi:10.1007/JHEP12(2013)039.

[233] CMS Collaboration. Combination of ATLAS and CMS top quark pair cross sectionmeasurements in the emu final state using proton-proton collisions at 8 TeV. Technical Report CMS-PAS-TOP-14-016, CERN, Geneva, 2014. URL: https://cds.cern.ch/record/1950834.

[234] V. Blobel. A survey of unfolding methods for particle physics, 2010. Terascale Statistics Tools School Spring 2010. [Online, accessed 23 Mar. 2015]. URL: https://www.wiki.terascale.de/images/a/a8/Blobel_Unfold_Note_20100524.pdf.

[235] David L. Phillips. A Technique for the Numerical Solution of Certain Integral Equations of the First Kind. *J. ACM*, 9(1):84–97, Jan 1962. `doi:10.1145/321105.321114`.

[236] A. N. Tikhonov. On the solution of improperly posed problems and the method of regularization. *Soviet Math.*, 5:1035, 1963.

[237] A. Hocker and V. Kartvelishvili. SVD approach to data unfolding. *Nucl. Instrum. Meth. A*, 372:469–481, 1996. `arXiv:hep-ph/9509307`, `doi:10.1016/0168-9002(95)01478-0`.

[238] K. Tackmann, A. Hoecker, and H. Lacker. TSVDUnfold software package. [Online, accessed 23 Mar. 2015]. URL: `http://root.cern.ch/root/html/TSVDUnfold.html`.

[239] D.-J. Fischer. *Inclusive neutral current ep cross sections with HERA II and two-dimensional unfolding*. PhD thesis, University of Hamburg, 2011. DESY-THESIS-2011-020. URL: `http://www-library.desy.de/preparch/desy/thesis/desy-thesis-11-020.pdf`.

[240] CMS Collaboration. CMS luminosity based on pixel cluster counting - Summer 2013 update. 2013. CMS-PAS-LUM-13-001.

[241] C. Diez Pardos and J. Kieseler. Dilepton trigger and lepton identification efficiencies for the top quark pair production cross section measurements at 8 TeV in the dilepton decay channel. 2013. CMS AN-2012/389.

[242] CMS Collaboration. Determination of jet energy calibration and transverse momentum resolution in CMS. *JINST*, 6:P11002, 2011. `arXiv:1107.4277`, `doi:10.1088/1748-0221/6/11/P11002`.

[243] K. Goebel, J. Haller, J. Ott, and H. Stadie. Jet transverse momentum resolution measurement using dijet events at $\sqrt{s} = 8\,\text{TeV}$. 2014. CMS-AN-2013/416.

[244] CMS Collaboration. Performance of missing transverse momentum reconstruction algorithms in proton-proton collisions at $\sqrt{s} = 8\,\text{TeV}$ with the CMS detector. 2012. CMS-PAS-JME-12-002.

[245] Darren Puigh. Event reweighting of the CSV distriminant shape using scale factors calculated with a tag and probe method. [Unpublished. Online, accessed on 10 March 2015]. URL: `https://twiki.cern.ch/twiki/bin/viewauth/CMS/BTagShapeCalibration`.

[246] Martijn Gosselink. Signal modelling uncertainties in CMS. Winchester, Sep. 2012. TOP 2012 - 5th International Workshop on Top Quark Physics. [Online, accessed on 10 April, 2015]. URL: `http://indico.cern.ch/event/180665/contribution/17`.

[247] CMS Collaboration. Measurement of the Jet Multiplicity in dileptonic Top Quark Pair Events at 8 TeV. 2013.

[248] CMS D0 ATLAS, CDF. First combination of Tevatron and LHC measurements of the top-quark mass. 2014. `arXiv:1403.4427`.

[249] J. Pumplin, D.R. Stump, J. Huston, H.L. Lai, Pavel M. Nadolsky, et al. New generation of parton distributions with uncertainties from global QCD analysis. *JHEP*, 0207:012, 2002. `arXiv:hep-ph/0201195`, `doi:10.1088/1126-6708/2002/07/012`.

[250] Hung-Liang Lai, Marco Guzzi, Joey Huston, Zhao Li, Pavel M. Nadolsky, et al. New parton distributions for collider physics. *Phys.Rev.*, D82:074024, 2010. `arXiv:1007.2241`, `doi:10.1103/PhysRevD.82.074024`.

[251] Top Physics Analysis Group. TOP Systematic Uncertainties (Run1). [Unpublished. Online, accessed 24 Mar. 2015]. URL: `https://twiki.cern.ch/twiki/bin/view/CMS/TopSystematicsRun1`.

[252] Michiel Botje, Jon Butterworth, Amanda Cooper-Sarkar, Albert de Roeck, Joel Feltesse, et al. The PDF4LHC Working Group Interim Recommendations. 2011. `arXiv:1101.0538`.

[253] Tae Jeong Kim. Details of the $t\bar{t}b\bar{b}$ cross-section definition. [Private communication].

[254] P. Azzurri, J. Bernardini, D. Caiulo, P. Spagnolo, C. Vernieri S. Alderweireldt, S. Bansal, X. Janssen, J. Lauwers, T. Cornelis, N. van Remortel S. de Visscher, and K. Kousouris. Search for the Standard Model Higgs Boson Produced in Vector Boson Fusion and Decaying to Bottom Quarks. 2013. CMS-AN-2012/395.

[255] Caterina Vernieri. Regression in H(bb) searches. [Unpublished. Online, accessed 8 April, 2015]. URL: `https://twiki.cern.ch/twiki/bin/view/CMS/HiggsToBBjetEnergyRegression`.

[256] CMS Collaboration. ATLAS-CMS recommended predictions for top-quark-pair cross sections using the Top++v2.0 program. [Unpublished. Online, accessed 8 April, 2015]. URL: `https://twiki.cern.ch/twiki/bin/view/LHCPhysics/TtbarNNLO`.

[257] G. Bevilacqua and M. Worek. On the ratio of $t\bar{t}b\bar{b}$ and $t\bar{t}jj$ cross sections at the CERN Large Hadron Collider. *JHEP*, 1407:135, 2014. `arXiv:1403.2046`, `doi:10.1007/JHEP07(2014)135`.

[258] Rei Tanaka. SM Higgs production cross sections at $\sqrt{s} = 13$-14 TeV. [Unpublished. Online, accessed 8 April, 2015]. URL: `https://twiki.cern.ch/twiki/bin/view/LHCPhysics/CERNYellowReportPageAt1314TeV`.

[259] Jorgen Beck Hansen. Physics at the LHC in Run2, 2013. [Unpublished. Online, accessed 8 April, 2015]. URL: `http://indico.cern.ch/event/251191/session/0/material/slides/0?contribId=1`.

[260] Jason Gallicchio and Matthew D. Schwartz. Seeing in Color: Jet Superstructure. *Phys.Rev.Lett.*, 105:022001, 2010. `arXiv:1001.5027`, `doi:10.1103/PhysRevLett.105.022001`.

[261] CMS Collaboration. Search for the associated production of the Higgs boson with a top-quark pair. *JHEP*, 1409:087, 2014. `arXiv:1408.1682`, `doi:10.1007/JHEP09(2014)087,10.1007/JHEP10(2014)106`.

[262] CMS Collaboration. Search for a standard model Higgs boson produced in association with a top-quark pair and decaying to bottom quarks using a matrix element method. 2015. `arXiv:1502.02485`.

[263] ATLAS Collaboration. Search for the Standard Model Higgs boson produced in association with top quarks and decaying into $b\bar{b}$ in pp collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector. 2015. `arXiv:1503.05066`.

[264] ATLAS Collaboration. Search for $H \to \gamma\gamma$ produced in association with top quarks and constraints on the Yukawa coupling between the top quark and the Higgs boson using data taken at 7 TeV and 8 TeV with the ATLAS detector. *Phys.Lett.B*, 740:222–242, 2015. `arXiv:1409.3122`, `doi:10.1016/j.physletb.2014.11.049`.

[265] K. Kondo. Dynamical Likelihood Method for Reconstruction of Events With Missing Momentum. 1: Method and Toy Models. *J.Phys.Soc.Jap.*, 57:4126–4140, 1988. `doi:10.1143/JPSJ.57.4126`.

[266] F.-P. Schilling V. Karimäki, T. Lampén. The HIP Algorithm for Track Based Alignment and its Application to the CMS Pixel Detector. CMS Note 2006/018, 2006.

[267] W. Adam E. Widl, R. Frühwirth. A Kalman Filter for Track-based Alignment. CMS Note 2006/022, 2006.

[268] R. Frühwirth, T. Todorov, and M. Winkler. Estimation of detector alignment parameters using the Kalman filter with annealing. *Journal of Physics G: Nuclear and Particle Physics*, 29(3):561, 2003. URL: `http://stacks.iop.org/0954-3899/29/i=3/a=309`.

[269] V. Blobel. Software alignment for tracking detectors. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 566(1):5 – 13, 2006. Proceedings of the 1st Workshop on Tracking in High Multiplicity Environments. URL: `http://www.sciencedirect.com/science/article/pii/S0168900206007984`, `doi:10.1016/j.nima.2006.05.157`.

[270] M. Swartz, D. Fehling, G. Giurgiu, P. Maksimovic, and V. Chiochia. A new technique for the reconstruction, validation, and simulation of hits in the CMS pixel detector. *PoS*, VERTEX2007:035, 2007.

[271] CMS Collaboration. CMS Tracker Alignment 2012. [Online, accessed on 10 April, 2015]. URL: `https://twiki.cern.ch/twiki/bin/view/CMS/TrackerAlignment2012`.

[272] Morris Swartz. Notes on second order Lorentz-angle effect relevant in FPIX, 2012. [Unpublished. Online, accessed on 18 April, 2015.]. URL: `https://twiki.cern.ch/twiki/pub/CMS/TkAlignmentLegacyThoughts/electron_drift1.pdf`.

[273] Gero Flucke. Thoughts on future of CMS tracker alignment, 2013. [Unpublished. Online, accessed on 18 April, 2015.]. URL: `https://twiki.cern.ch/twiki/bin/view/CMS/TkAlignmentLegacyThoughts`.

[274] CMS Collaboration. CMS Tracker Silicon Pixel Lorentz Angle Tags. [Online, accessed on 10 April, 2015]. URL: `https://twiki.cern.ch/twiki/bin/view/CMS/SiPixelLorentzAngleHistory`.

[275] CMS Collaboration. Global tags for conditions data. [Online, accessed on 10 April, 2015]. URL: `https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideFrontierConditions`.

Statement of authorship

I hereby declare, on oath, that I have written the present dissertation by my own and have not used other than the acknowledged resources and aids.

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Hamburg, 04 Mai 2015

Acknowledgements

Apart from my own work on this thesis, a series of people influenced it in a direct or indirect way. I would like to acknowledge those people who have made the largest impact on this work.

First of all I would like to thank to my supervisor, Achim Geiser, in particular for suggesting such an interesting subject for my physics analysis long before it has become a really hot topic in the outside world. Your guidance, support, long and interesting discussions (usually late in the evening), and care about personal issues are very valuable, and are not something one can expect from every supervisor. Thank you for this great PhD experience!

A large fraction of the described results were achieved thanks to a young, but very devoted physicist, Johannes Hauk. Your continuous motivation, optimism and the hard endeavour to work as a team have helped me a lot. Especially the *GenHFHadronMatcher* would not be even close to its success without your promotion and technical guidance. Thank you for moving this train forward no matter what, and for being more personal than the work requires.

Furthermore, I want to thank Carmen Diez Pardos for the great opportunity to transform a simple estimation of $t\bar{t}b\bar{b}$ background into a standalone differential cross-section measurement, which will potentially have a much greater value than a chapter in a thesis. I would like to thank Alexei Raspereza for the genuine interest in the described physics analyses and for the help with the technical setup of one of the key parts of the $t\bar{t}b\bar{b}$ measurement. I also appreciate inputs from Ivan Asin Cruz, Maria Martin Aldaya, Jasone Garay Garcia, Eleni Ntomari, Andreas Meyer, Tyler Dorland, Ievgen Korol and Ganna Dolinska, which were important at different stages of my work.

A significant part of my PhD was devoted to the tracker-alignment activities, which gave me the pleasure to work with such people as Gero Flucke, Jörg Behr, Rainer Mankel, Justyna Tomaszewska, Gregor Hellwig, and Claus Kleinwort. I would like to especially thank Gero Flucke for the great ideas of studies within the tracker-alignment project, continuous assistance and promotion of the obtained results, and for the involvement in these studies even after leaving CMS. You have made my work in the CMS tracker-alignment project much more than just a technical task.

I would also like to thank Achim, Johannes, Jasone, Gero and Carmen for voluntarily reviewing different parts of my thesis. In addition, I am very grateful to Johannes Haller, Elisabetta Galo, Maria Vittoria Garzelli and Dieter Horns for agreeing to join the Dissertation Committee, and for being very active and supportive in the most crucial parts of the formal procedure. Finally, the defence of