Karlsruhe Institute of Technology

IEKP-KA/2015-24

Analysis of Standard Model Higgs Boson Decays to Tau Pairs with the CMS Detector at the LHC

Thomas Müller

Zur Erlangung des akademischen Grades eines
DOKTORS DER NATURWISSENSCHAFTEN
von der Fakultät für Physik des
Karlsruher Instituts für Technologie

genehmigte

DISSERTATION

von

Dipl.-Phys. Thomas Müller
aus Ehringshausen

Tag der mündlichen Prüfung: 20.11.2015

Referent: Prof. Dr. Günter Quast
Korreferent: Prof. Dr. Wim de Boer

Abstract

The search for the Standard Model Higgs boson decaying into pairs of τ leptons performed at the complete CMS run I data set comprising integrated luminosities of 4.9 fb^{-1} at centre-of-mass energies of $\sqrt{s} = 7$ TeV and 19.7 fb^{-1} at 8 TeV, respectively, is presented with a particular focus on the analysis of the di-muon final state. Multivariate discrimination techniques are used to handle the two most dominant backgrounds of $\mathrm{Z} \to \mu\mu$ and $\mathrm{Z} \to \tau\tau$ events. Exclusion limits are set as no excess over the background-only hypothesis is found. The published CMS $\mathrm{H} \to \tau\tau$ analysis is summarised and an outlook to future measurements in this channel is given by extrapolating the $\mathrm{H} \to \tau\tau$ analysis to the expected luminosity of the present $\sqrt{s} = 13$ TeV data taking period. Here, only the most sensitive channels $\mu\tau_{\mathrm{h}}$, $\mathrm{e}\tau_{\mathrm{h}}$, $\mathrm{e}\mu$ were used.

Declaration

This dissertation is the result of my own work, except where explicit reference is made to the work of others, and has not been submitted for another qualification to this or any other university.

Thomas Müller

Contents

Introduction

The current best knowledge about elementary particles and their interactions is described by a quantum theory called the Standard Model of particle physics [1–3]. It explains the electromagnetic, weak and strong interactions between the fermions, the smallest building blocks of matter, with the exchange of force carriers, the bosons. The masses of elementary particles differ by orders of magnitude, which has a strong impact on their kinematics and the properties of the interactions. Fifty years ago, a theory called the electroweak symmetry breaking has been predicted, postulating the Higgs boson as being responsible for generating the masses of elementary particles [4–7]. The Higgs mechanism contains one free parameter to be determined experimentally: the mass of the Higgs boson. An overview of the relevant parts of the Standard Model is given in chapter 1.

Before the LHC era it was not possible to find evidence for the existence of the Higgs boson. The LHC collides protons at centre-of-mass energies which are larger by almost one order of magnitude compared to the Tevatron, which was the most powerful collider until the LHC started its operation. One of the major goals of the two general-purpose detectors ATLAS and CMS at the LHC was and still is the investigation of the origin of the electroweak symmetry breaking. An overview of the CMS experiment is given in chapter 2.

July 4^{th}, 2012 marked an enormous progress in this field: a new resonance of a boson with a mass of 125 GeV decaying into pairs of photons or into four leptons was observed by both the ATLAS and the CMS collaborations [8, 9]. The discovery raised a central question: Is the discovered particle the Higgs boson as it is predicted by the Standard Model of particle physics? Since the observation, all ATLAS and CMS analyses focussing on the Higgs boson addressed this question.

The Higgs boson is predicted to couple to the mass of elementary particles: the heavier the particle, the stronger the coupling to the Higgs boson. However, this mechanism is fundamentally different for couplings of the Higgs boson to bosons and to fermions. In the low Higgs boson mass range below $m_{\mathrm{H}} = 150$ GeV, direct couplings of the Higgs boson to fermions can best be probed in decays to pairs of τ leptons, the heaviest siblings of electrons. This is caused by the large branching ratio $\mathcal{BR}\left(\mathrm{H} \to \tau\tau\right)$ for low Higgs boson mass hypotheses below 150 GeV and by the well controllable and comparably

small background contamination in this channel. A fraction of 6.32 % of Higgs bosons with a mass hypothesis of 125 GeV is predicted to decay into pairs of τ leptons. However, the analysis of this channel is complicated by the two to four neutrinos in the final state, depending on the di-τ decay mode.

In chapter 3, the search for Higgs bosons decaying into pairs of τ leptons in the di-muon final state is presented, before the combined results of all $H \to \tau\tau$ sub-analyses are summarised in the first part of chapter 4. The analysis is based on the complete CMS data from the first running period of the LHC in the years 2011 and 2012 comprising integrated luminosities of 4.9 fb^{-1} at $\sqrt{s} = 7$ TeV and 19.7 fb^{-1} at 8 TeV, respectively. The analysis of the di-muon channel shares analysis techniques with all other $H \to \tau\tau$ channels, such as the mass reconstruction of the invariant di-τ mass or the embedding method for the modelling of the $Z \to \tau\tau$ background. Moreover, the analysis is complicated by the additional and huge background from $Z \to \mu\mu$ events which represents more than 99 % of all backgrounds. Multivariate analysis techniques performing the simultaneous suppression of the two main backgrounds are developed. It is shown that the sensitivity of this channel can be brought to a level that is comparable to the $e\mu$ final state. The combination of all $H \to \tau\tau$ sub-analyses provides evidence for Higgs boson decays into fermions [10, 11].

In the second running period of the LHC (since 2015), the centre-of-mass energy is raised to $\sqrt{s} = 13$ TeV and the instantaneous luminosity is expected to increase. Consequently, the higher signal production rates allow for measurements of the Higgs boson and its couplings with higher precision. The second part of chapter 4 provides an outlook on the analysis of the upcoming data from the second running period of the LHC in the $H \to \tau\tau$ channel. The appendix completes some details of the descriptions of the analyses and presents a general software framework developed for data analysis applications in high energy physics, which has been introduced during the upgrade phase between run I and run II.

1 The Standard Model of Particle Physics

Since ancient times, mankind searches for the fundamental building blocks of matter in the microcosm as well as for a description of the structure and the evolution of our universe. Particle physics focuses on the first part, whereas the investigation of the structure of the matter cannot be fully separated from the study of cosmological problems.

Two prominent pioneers should be named: the Greek philosopher Democritus postulated indivisible constituents of all matter which he called atoms [12]. This, which is possibly the oldest model of particle physics, has been stated more than two thousands years ago. The era of modern particle physics was mainly pioneered by Ernest Rutherford and his scattering experiment at the beginning of the last century [13]. By radiating beams of α and β particles at a gold foil and analysing the scattering angles of the elastically scattered particles, he discovered the atomic nuclei, a concentration of matter in only a small fraction of the macroscopic volume.

Since this experiment, the methods of modern particle physics have changed only slightly. While still scattering experiments are performed, the energy of the colliding particles has increased by orders of magnitude and the focus moved to inelastic scattering processes. More and more small nuclear structures of composite particles can be resolved and new unstable particles can be created by transforming energy into mass. At present, the collider flagship LHC at CERN collides protons at a centre-of-mass energy of 13 TeV, which can be considered as the current high energy frontier.

The current best knowledge about fundamental particles and their interactions is comprised by the standard model of particle physics (SM). Matter is built from spin-$\frac{1}{2}$[1] particles, the fermions, and the interactions between them are mediated by spin-1[1] force carriers, the bosons. Additionally, the scalar Higgs boson takes the responsibility of generating the invariant masses of the elementary particles.

Four fundamental forces are known. All processes involving interactions between two systems can be traced back to either one of these forces or a combination of them. Firstly, there is the gravitational force that describes the interaction between massive particles. Secondly, an electromagnetic force between electrically charged particles is known. Thirdly, the weak interaction is responsible for nuclear processes such as beta decays. At last, there is the strong interaction describing the forces between

[1] The spin of elementary particles is measured in units of $\hbar$. The unit is often omitted.

colour-charged particles such as quarks and gluons. This force holds together the nuclei of the atomic nucleus.

Quantum field theories exist that provide a microscopic description for all these interactions apart from the gravitation[2]. The interactions between particles are mediated by gauge bosons and the relevant charge of the particle is a measure of the coupling strength. Feynman graphs provide an intuitive illustration of the microscopic processes. Figure 1.1 exemplary shows a Feynman graph for the elastic scattering between two electrons which is mediated by the exchange of a virtual photon.

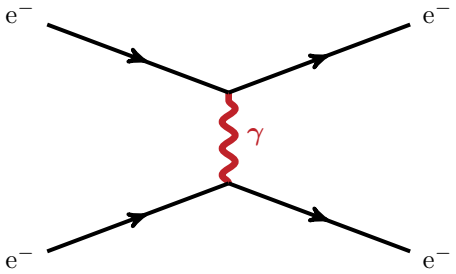


Figure 1.1.: Feynman graph illustrating the elastic scattering of two electrons mediated by the exchange of a virtual photon between them. The abscissa shows the time and the ordinate is a measure for the space variable.

Table 1.1 gives an overview of the gauge bosons. The electromagnetic interaction couples to the electric charge. Quarks and down-type leptons, namely electrons, muons and taus, take part in this interaction. The vanishing mass of the photons lead to the infinite range of this interaction. $W^{\pm}$ and Z bosons coupling to the weak isospin mediate the weak interaction. Only particles with left-handed chirality[3] take part in this interaction. The high mass of the vector bosons causes this interaction to be weak and short-ranged. The gluons responsible for the strong interaction couple to the colour charge of quarks. Gluons itself are colour-charged. The confinement leads to the short range in the order of the proton radius.

Table 1.1.: Gauge bosons and their interactions

Interaction	**Gauge Boson**	**Mass** / GeV	**Range** / m
Electr. magn.	Photon γ	0	∞
Weak	Z^0	91.18	10^{-15}
	$W^{\pm}$	80.40	
Strong	8 Gluons g	0	10^{-18}

All elementary fermions, the leptons and the quarks, appear in three generations or flavours, as shown in table 1.2. These generations are approximately distinguished by the mass of the particles. Stable matter particles such as protons and neutrons are constructed from elementary particles belonging

[2]Presently, the gravitation can only be described classically by the theory of general relativity. A quantum field theory that fits into the scheme of the Standard Model is required to describe gravitational phenomena at high energies or small distances. For the actual research at the energy scale of colliders, gravitation does not play any role, since its force is negligible in comparison with the other three fundamental forces.

[3]The chirality of a particle refers to its helicity, the projection of the spin onto the direction of momentum. Right-handed particles (and left-handed antiparticles) have a positive helicity whereas left-handed particles (and right-handed antiparticles) are characterised by a negative helicity.

to the first generation as they have the lightest masses. The charges are a measure of the couplings between the particles and the force carriers and therefore determine the interaction strength. Each fermion has a sibling with opposite charges which is referred to as its antiparticle.

Table 1.2.: Left-handed elementary fermions. Right-handed elementary fermions do not carry any weak isospin.

Fermions	**Generation**			**Charge**		
	1	2	3	El. Charge	Weak Isospin	Colour
Leptons	ν_{e}	ν_{μ}	ν_{τ}	0	$+\frac{1}{2}$	0
	e	μ	τ	$-e$	$-\frac{1}{2}$	
Quarks	u	c	t	$+\frac{2}{3}\,e$	$+\frac{1}{2}$	r, g, b
	d	s	b	$-\frac{1}{3}\,e$	$-\frac{1}{2}$	

In the following, a short introduction into the theoretical framework describing quantum electrodynamics, the weak interaction and its unification with the electromagnetic interaction as well as the Higgs mechanism is presented. The reader is pointed to textbooks for more details, such as [14, 15], or to the summarising article [16]. The chapter concludes with a phenomenological overview of Higgs boson production and decay in the Standard Model and a summary of the latest results of Higgs boson analyses.

1.1. Quantum Field Theory

In classical mechanics the dynamics of a point-like system are described by a Lagrangian $L\left(q, \dot{q}\right) = E_{\mathrm{kin}} - E_{\mathrm{pot}}$. The action $S = \int L \,\mathrm{d}t$ is minimised by every state transition, from which an equation of motion, the Euler-Lagrange equation, can be derived.

In quantum field theory particles are referred to as wave functions $\psi(x)$, where x denotes the space-time. These continuous systems are described by a Lagrange density function $\mathcal{L}\left(\psi, \partial_{\mu}\,\psi\right)$ where $\partial_{\mu}\,\psi$ denotes the derivative $\frac{\partial \psi}{\partial x^{\mu}}$. The action

$$S = \int \mathcal{L}\left(\psi, \partial_{\mu}\,\psi\right)\,\mathrm{d}^4 x_{\mu}$$

is minimised leading to the Euler-Lagrange equation

$$\partial_{\mu}\left(\frac{\partial \mathcal{L}}{\partial(\partial_{\mu}\,\psi)}\right) - \frac{\partial \mathcal{L}}{\partial \psi} = 0$$

that expresses the equation of motion for the wave function $\psi(x)$. For instance, the Lagrangian for a free fermion with mass m and the corresponding Euler-Lagrange equation are

$$\mathcal{L} = \mathrm{i}\,\psi^{\dagger}\,\gamma^{\mu}\,\partial_{\mu}\,\psi - m\,\psi^{\dagger}\,\psi \quad \Rightarrow \quad \left(\mathrm{i}\,\gamma^{\mu}\,\partial_{\mu} - m\right)\psi = 0 \tag{1.1}$$

The fermion fields ψ are four-component Dirac spinors and $\psi^\dagger$ denotes the adjoint wave function. The equation of motion is called the Dirac equation. The gamma matrices are denoted by γ^μ.

1.1.1. Quantum Electrodynamics

Quantum Electrodynamics describes the interaction between electrically charged particles by exchanging photons as force carriers between them. For the introduction of interactions between particles the principle of local gauge invariance has proved to be adequate. A Physical system lets the Lagrangian remain unchanged under symmetry transformations.

In this theory, global gauge transformations are phase rotations of the wave functions coming from the underlying $\mathrm{U}(1)$ symmetry group. They are expected to have no influence on any measurement because only the probability density $|\psi|^2$ can be measured. Local $\mathrm{U}(1)$ symmetry gauge transformations for both the wave function and its derivative can be expressed as

$$\psi \;\to\; \psi' = \mathrm{e}^{\mathrm{i}\alpha(x)}\,\psi \quad \text{and} \quad \partial_\mu\,\psi \;\to\; \partial_\mu\,\psi' = \mathrm{e}^{\mathrm{i}\alpha(x)}\left(\mathrm{i}\,\psi\,\underbrace{\partial_\mu\,\alpha(x)}_{\neq\,0} + \partial_\mu\,\psi\right)$$

depending on the local parameter $\alpha(x)$. The derivative of this parameter yields an additional term in the derivative of the wave function that destroys the invariance of the Lagrangian. A compensating field A can be included to conserve its local gauge invariance.

$$\partial_\mu \;\to\; \mathrm{D}_\mu = \partial_\mu + \mathrm{i}\,e\,A_\mu(x) \quad \text{with} \quad A_\mu \;\to\; A'_\mu = A_\mu - \frac{1}{e}\,\partial_\mu\alpha(x)$$

The Lagrangian of a free fermion (1.1) has to be completed with terms resulting from the new field A.

$$\begin{aligned}&\mathcal{L}_{\mathrm{QED}} = \psi^\dagger\left(\mathrm{i}\,\gamma^\mu\,\partial_\mu - m\right)\psi - j^\mu\,A^\mu - \frac{1}{4}\,F_{\mu\nu}\,F^{\mu\nu} \\ &\text{with} \quad j^\mu = q\,\psi^\dagger\,\gamma^\mu\,\psi \quad \text{and} \quad F^{\mu\nu} = \partial^\mu\,A^\nu - \partial^\nu\,A^\mu\end{aligned} \tag{1.2}$$

The four-current is denoted by $j_\mu = (\rho, \vec{j})$ and $A^\mu = (\Phi, \vec{A})$ describes the electromagnetic four-potential as known from classical electrodynamics. The second term expresses the coupling of the fermion with charge q to the field A which can be identified as the photon field. Thus the photon is introduced as the gauge boson mediating the electromagnetic interaction between charged particles. The last term, a kinematic term for the photon field, is added for completeness. A mass term $\sim m_\gamma^2 A^\mu A_\mu$ would spoil the local gauge invariance: the photon is therefore massless. The Dirac equation as well as the Maxwell equations follow from the QED Lagrangian.

$$\left(\mathrm{i}\,\gamma^\mu\,\partial_\mu - m\right)\psi = 0 \quad \text{and} \quad \partial_\mu\,F^{\mu\nu} = j^\nu \tag{1.3}$$

1.1.2. The Weak Interaction

The weak interaction describes transitions between fermions, either leptons or quarks, by the exchange of W or Z bosons. These transitions are elegantly expressed in terms of a spin formalism. Left-handed

fermions are arranged in weak isospin doublets and transitions are described by rotations in the isospin space. For instance, a muon with a third component of its weak isospin of $T_3 = -\frac{1}{2}$ is converted into a muon neutrino with $T_3 = +\frac{1}{2}$ by emitting a W^- boson with $T_3 = -1$. Since in the Standard Model right-handed fermions do not couple to W bosons, their weak isospin is zero. They are therefore regarded to as isospin singlets.

1.1.3. Glashow-Weinberg-Salam Theory – The Electroweak Unification

In analogy to the theory of the electromagnetic interaction, the weak interaction has been tried to be described by a similar theory, a Yang-Mills theory. Sheldon Glashow [1], Abdus Salam [3] and Steven Weinberg [2] formulated a unified theory that is capable to describe both the electromagnetic and the weak interaction. The formalism of the electroweak interaction is based on local $\mathrm{SU}(2) \otimes \mathrm{U}(1)$ gauge symmetry transformations. The transformations of the wave functions are expressed depending on four local parameters $\vec{\alpha}(x)$ and $\beta(x)$.

$$\psi_\mathrm{L} \;\to\; \psi'_\mathrm{L} = \mathrm{e}^{\,\mathrm{i}\,\vec{\alpha}(x)\,\frac{\vec{\sigma}}{2} + \mathrm{i}\,\beta(x)\,Y}\,\psi_\mathrm{L} \quad \text{and} \quad \psi_\mathrm{R} \;\to\; \psi'_\mathrm{R} = \mathrm{e}^{\,\mathrm{i}\,\beta(x)\,Y}\,\psi_\mathrm{R}$$

Here, the hypercharge $Y = 2\left(\frac{q}{e} - T_3\right)$ denotes the generator of the $\mathrm{U}(1)$ symmetry group describing phase rotations whereas the Pauli matrices $\vec{\sigma}$ are the generators of the $\mathrm{SU}(2)$ symmetry group initiating rotations in the isospin space. The left-handed fermion fields are referred to as isospin doublets of Dirac spinors. For example, the leptonic doublet containing a charged lepton $\ell = \mathrm{e}, \mu, \tau$ and its neutrino ν_l can be written as the following doublet.

$$\psi_\mathrm{L} = \begin{pmatrix} \psi_{\nu_l} \\ \psi_l \end{pmatrix}_L$$

Consequently, the Lagrangian (1.1) for free fermions needs to be extended in order to preserve its local gauge invariance under these $\mathrm{SU}(2) \otimes \mathrm{U}(1)$ symmetry transformations. Four new gauge fields $\vec{W}_\mu$ and B_μ need to be introduced that transform as follows.

$$\begin{aligned}
\partial_\mu \;&\to\; \mathrm{D}_\mu = \partial_\mu - \mathrm{i}\,g_2\,\frac{\vec{\sigma}}{2}\,\vec{W}_\mu - \mathrm{i}\,g_1\,\frac{Y}{2}\,B_\mu \\
\vec{W}_\mu \;&\to\; \vec{W}'_\mu = \vec{W}_\mu - \frac{1}{g_2}\,\partial_\mu\,\vec{\alpha}(x) - \vec{\alpha}(x) \times \vec{W}_\mu \\
B_\mu \;&\to\; B'_\mu = B_\mu - \frac{1}{g_1}\,\partial_\mu\,\beta(x)
\end{aligned}$$

Then the Lagrangian describing the electroweak interaction reads as the following.

$$\begin{aligned}
&\mathcal{L}_\mathrm{EWK} = \psi^\dagger\,\mathrm{i}\,\mathrm{D}_\mu\,\gamma^\mu\,\psi - \frac{1}{4}\,\vec{W}_{\mu\nu}\,\vec{W}^{\mu\nu} - \frac{1}{4}\,B_{\mu\nu}\,B^{\mu\nu} \qquad (1.4)\\
&\text{with} \quad \vec{W}^{\mu\nu} = \partial^\mu\,\vec{W}^\nu - \partial^\nu\,\vec{W}^\mu \quad \text{and} \quad B^{\mu\nu} = \partial^\mu\,B^\nu - \partial^\nu\,B^\mu
\end{aligned}$$

Couplings between the fermions and the new gauge bosons are given by the mixed terms with fermion and boson fields originating from the covariant derivative. The coupling constants are denoted by g_1 for the couplings to the field B_μ and by g_2 for the couplings to the three fields $\vec{W}_\mu$. Again, the Lagrangian contains no mass term since such a term would destroy the gauge invariance.

The physical fields of the corresponding weak and electromagnetic force carriers are linear combinations of the newly introduced fields parametrised by the mixing angle θ_W, the Weinberg angle $\sin^2\theta_\mathrm{W} \approx 0.23$ [2].

$$W_\mu^\pm = \frac{1}{\sqrt{2}}\left(W_\mu^1 \mp W_\mu^2\right) \quad \text{and} \quad \begin{pmatrix} Z_\mu \\ A_\mu \end{pmatrix} = \begin{pmatrix} \cos\theta_\mathrm{W} & -\sin\theta_\mathrm{W} \\ \sin\theta_\mathrm{W} & \cos\theta_\mathrm{W} \end{pmatrix} \begin{pmatrix} W_\mu^3 \\ B_\mu \end{pmatrix}$$

$$\text{with} \quad \sin^2\theta_\mathrm{W} = \frac{g_2^2}{g_1^2 + g_2^2}$$

The charged $\mathrm{W}^\pm$ bosons mediate charged weak couplings whereas the neutral Z boson is connected to neutral currents. The field A represents the photon γ responsible for the electromagnetic interaction.

1.1.4. The Higgs Mechanism – Spontaneous Symmetry Breaking

The quantum field theory derived from the fundamental principle of local gauge invariance describes the (three) fundamental forces very well, but it is not capable of explaining the masses of the weak gauge bosons $\mathrm{W}^\pm$ and Z that can not be neglected. Peter W. Higgs[4] [4,5], Robert Brout[4] and François Englert[4] [6] as well as Gerald S. Guralnik, Carl R. Hagen and Tom W. B. Kibble [7] suggested a mechanism of spontaneous electroweak symmetry breaking by which the weak gauge bosons acquire mass. The mechanism is shortly known as the Higgs mechanism.

A new complex scalar field Φ is introduced based on the following Lagrangian term which is invariant under $\mathrm{SU}(2) \otimes \mathrm{U}(1)$ symmetry gauge transformations. The two complex components correspond to four real fields. This is the simplest way of introducing at least three degrees of freedom that are needed to describe the vector boson masses.

$$\mathcal{L}_\mathrm{Higgs} = \left(\mathrm{D}^\mu \Phi\right)^\dagger \left(\mathrm{D}_\mu \Phi\right) \underbrace{-\mu^2 \Phi^\dagger \Phi - \lambda\left(\Phi^\dagger \Phi\right)^2}_{=-V(\Phi)} \quad \text{with} \quad \Phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} \tag{1.5}$$

The potential, $V(\Phi)$, which is known as a Mexican hat potential because of its shape, contains a mass term with the mass-type constant μ^2 and self-couplings with the positive dimensionless constant λ. For negative values of μ^2 the potential has not only one minimum at $\langle\Phi\rangle_0 = 0$, as it is the case for $\mu^2 > 0$, but there are multiple ground states $\langle\Phi\rangle_0$ for the scalar field fulfilling the following condition, which

[4] In 2013, after the discovery of the Higgs boson, François Englert and Peter W. Higgs were awarded the Nobel Prize in Physics for the theoretical discovery of the Higgs mechanism. (Robert Brout had died in 2011.)

can be expressed based on a vacuum expectation value v.

$$\left|\langle\Phi\rangle_0\right|^2 = -\frac{\mu^2}{2\,\lambda} \equiv \frac{v^2}{2}$$

Since there is no distinctive ground state, the system chooses one at random. Without loss of generality because of the $\mathrm{SU}(2)$ invariance, the chosen ground state can be written as follows.

$$\langle\Phi\rangle_0 = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix}$$

After the system has fallen into such a ground state, the $\mathrm{SU}(2)$ rotation symmetry is no more apparent. It is said that the symmetry is spontaneously broken. After transforming the scalar field Φ and expressing it in terms of the Higgs field H which is one of the four real fields,

$$\Phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix}$$

the Lagrangian (1.5) can be expanded around the ground state $\langle\Phi\rangle_0$. By doing so, the kinetic term $\left|\mathrm{D}_\mu\,\Phi\right|^2$ results in mass terms for the vector bosons $\mathrm{W}^\pm$ and Z as well as terms describing the couplings between the weak gauge bosons and the Higgs boson and kinematic terms for the Higgs field.

$$\begin{aligned} \left|\mathrm{D}_\mu\,\Phi\right|^2 &= \frac{1}{8}\,v^2\,g_2^2\left|W_\mu^1 + \mathrm{i}\,W_\mu^2\right|^2 + \frac{1}{8}\,v^2\left|g_2\,W_\mu^3 - g_1\,B_\mu\right|^2 + \ldots \\ &= \frac{1}{2}\,m_W^2\,W_\mu^+\,W^{-\mu} + \frac{1}{2}\,m_Z^2\,Z_\mu\,Z^\mu + \ldots \end{aligned}$$

Three massless Goldstone bosons are absorbed by the breaking of the symmetry and the scalar Higgs boson field remains, with the Higgs boson as a physical particle. The three degrees of freedom of the Goldstone bosons are absorbed by the mass parameters of the vector bosons that can be expressed in terms of the vacuum expectation value v. The relationship between the W mass and the Z mass is given by the Weinberg angle.

$$m_W = \frac{v}{2}\,g_2 \quad \text{and} \quad m_Z = \frac{v}{2}\,\sqrt{g_1^2 + g_2^2} \quad \text{with} \quad \frac{m_W}{m_Z} = \frac{g_2}{\sqrt{g_1^2 + g_2^2}} = \cos\theta_\mathrm{W}$$

In turn, the expansion of the potential terms yields a mass term for the Higgs boson, the particle connected with the Higgs field.

$$V = \frac{1}{2}\,\mu^2\,(v+H)^2 + \frac{1}{4}\,\lambda\,(v+H)^4 \quad \Rightarrow \quad m_H = -\sqrt{2}\,\mu = v\,\sqrt{2\,\lambda}$$

The vacuum expectation value can be expressed in terms of the measured vector boson masses and the Weinberg angle, but the value of one parameter is not constrained by the theory and therefore has to be measured. This parameter is usually chosen as the Higgs boson mass m_H. From theoretical considerations merely an approximate upper boundary of $m_H < 650$ GeV can be derived [16].

By this mechanism, only the $\mathrm{SU}(2)$ symmetry is spontaneously broken and the weak force carriers acquire mass. Since the $\mathrm{U}(1)$ symmetry remains exact, the photon is still massless. The same is true for the massless gluons as the gauge fields of the QCD.

1.1.5. Yukawa Interaction

The fermion masses are neither described by the electroweak Lagrangian (1.4), since this would mix up left-handed and right-handed terms, nor are they introduced by the Lagrangian describing the scalar field (1.5). However, these masses can be described in a similar manner as the vector boson masses.

The coupling between the fermion fields ψ and the scalar field Φ is expressed by a Yukawa interaction Lagrangian. For leptons it is expressed as the following.

$$\mathcal{L}_{\mathrm{Yukawa}} = -\lambda_l \, \psi_{\mathrm{L}}^{\dagger} \, \Phi \, \psi_{\mathrm{R}} = -\frac{1}{\sqrt{2}} \, \lambda_f \, \psi^{\dagger} \, (v + H) \, \psi \tag{1.6}$$

After spontaneous symmetry breaking it is expressed in terms of the remaining Higgs field H. This Lagrangian introduces both the fermion masses m_f and their couplings to the Higgs field g_{Hff}.

$$m_f = \frac{v}{2} \, \lambda_f \quad \text{and} \quad g_{Hff} = \frac{m_f}{v}$$

Since the parameters λ_f are not fixed, the fermion masses appear as additional parameters of the theory. The Yukawa coupling of the fermions to the Higgs boson is proportional to their mass.

1.1.6. The Standard Model Lagrangian

In summary, the full Lagrangian of the Standard Model comprises the following four terms, that have been introduced previously[5]. The summations over all fermions and the complex conjugated terms are not shown explicitly.

$$\mathcal{L}_{\mathrm{SM}} = \mathcal{L}_{\mathrm{EWK}} + \mathcal{L}_{\mathrm{Higgs}} + \mathcal{L}_{\mathrm{Yukawa}} + \mathcal{L}_{\mathrm{QCD}}$$

Feynman rules describe the procedure of predicting physical observables such as cross section or angular distributions of flight directions of decay products. This thesis focuses on the experimental verification of such predictions or expectations. Therefore a more phenomenological approach is chosen for the introduction of Higgs physics at the LHC in the next section.

[5] The theory of the strong interaction, quantum chromodynamics (QCD) is established in analogy to the electroweak theory. $\mathrm{SU}(3)$ symmetry operations describe rotations in the three-dimensional colour space. This symmetry is preserved by introducing eight massless gluon fields mediating interactions between colour charged objects based on a Lagrangian term $\mathcal{L}_{\mathrm{QCD}}$.

1.2. Higgs Boson Searches at the LHC

As an impressively successful theory the Standard Model has been verified by many particle physics experiments and its parameters are precisely measured. exemplary the electroweak precision measurements at LEP [17] and the latest results of the CMS Standard Model Physics group [18] should be mentioned here. In July 2012, the discovery of a new boson at both the ATLAS and the CMS experiment [8,9] highlighted an important step in probing the mechanism of the electroweak symmetry breaking as the last building block of the SM remaining to be established.

The section first outlines the environment of the LHC as a hadron collider and then presents the Higgs boson production and decay modes, focussing on the $H \to \tau\tau$ channel. After this, an overview of the latest CMS results is given. For an experimental introduction to the LHC and the CMS detector the reader is referenced to the next chapter 2.

1.2.1. Hadron Colliders and Parton Distribution Functions

The LHC is operated at the high energy frontier providing particle collisions at the highest centre-of-mass energies that can artificially be produced today. The protons that are brought to collisions, are composite particles and its constituents, three valence quarks defining the protons quantum numbers and a see of quarks and gluons, are called partons. A collision of two protons can in general lead to a hard interaction of two partons with large momentum transfer, accompanied by an underlying event consisting of low energy interactions of the other proton constituents. The longitudinal momentum of the partons taking part in the hard interaction before the collision is unknown as it is only a fraction of the proton momentum. Parton distribution functions (PDFs) describe the distribution $x\, f_i(x, Q^2)$ of the fraction x of the momentum a parton of a given flavour i is carrying with respect to the proton momentum at the scale Q^2, see figure 1.3. The momentum carried by the valence quarks is on average significantly higher than the one of the sea quarks.

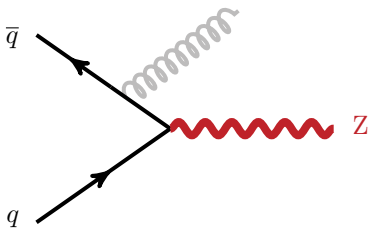


Figure 1.2.: Production of Z bosons at hadron colliders.

This has consequences for the production of massive resonances at the LHC. Z bosons are produced via the annihilation of a $q\overline{q}$ pair as illustrated in figure 1.2. At leading order, without initial state radiation, the Z boson is not produced in rest but carries momentum because of the asymmetric initial state characterised by the lower momentum of the anti-quark with respect to the valence quark. In contrast, Higgs boson events without jets in the event are predominantly produced more centrally via gluon fusion as illustrated in figure 1.5a, because of the symmetric initial state of the two gluons.

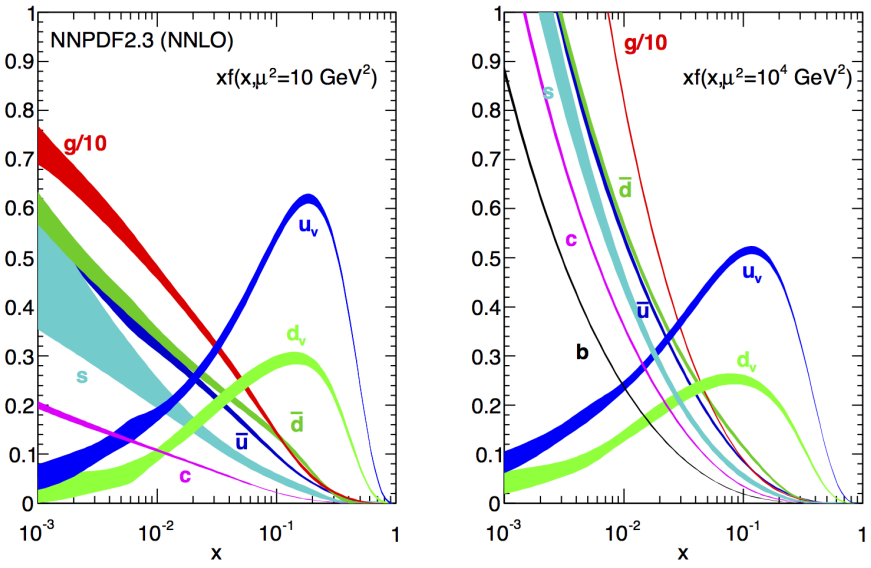


Figure 1.3.: Parton distribution functions describing the distribution of the momentum fraction of the proton carried by a parton of a given flavour, for lower scales (left) and higher ones (right) [19]. Valence quarks have on average a higher momentum than sea quarks or gluons. With increasing momentum of the proton the fraction of sea quarks and gluons with high momenta increases with respect to the one of valence quarks.

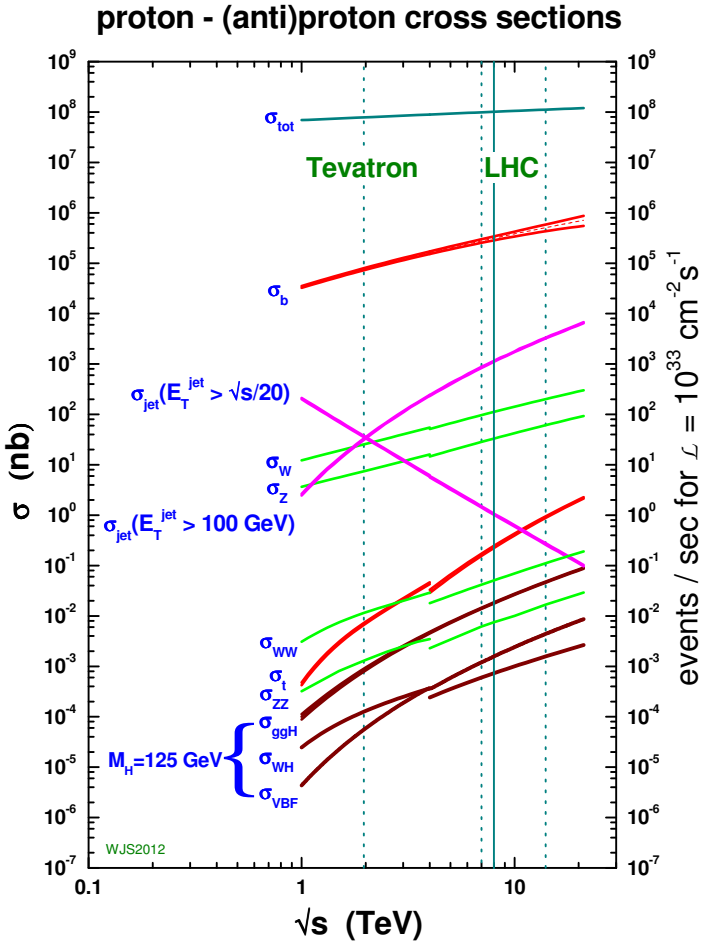


Figure 1.4.: Standard Model cross sections as a function of the centre-of-mass energy [20].

Beside the fact that analyses at hadron colliders cannot exploit any longitudinal momentum conservation they have to cope with another difficulty with respect to lepton colliders. Due to the coloured initial state of partons, the production cross sections for jets are very large compared to the ones for the production of heavy bosons. The cross sections for important SM processes are depicted in figure 1.4 as a function of the colliders centre-of-mass energy. The total cross section is dominated by jet production.

1.2.2. Standard Model Higgs Boson Production and Decay

At the LHC there are four main Higgs boson production modes at leading order. Their Feynman graphs are shown in figure 1.5. Table 1.3 gives the production cross sections at centre-of-mass energies of $\sqrt{s} = 7$, 8 and 13 TeV for a Higgs boson mass hypothesis of $m_{\mathrm{H}} = 125$ GeV. Data from pp collisions at these centre-of-mass energies are analysed in the scope of this thesis. The cross sections at $\sqrt{s} = 13$ TeV are enhanced by a factor of approximately two with respect to the ones at $\sqrt{s} = 8$ TeV. Figure 1.6 shows the Higgs boson production cross sections at centre-of-mass energies of 8 and 14 TeV, respectively, as a function of the Higgs boson mass hypothesis.

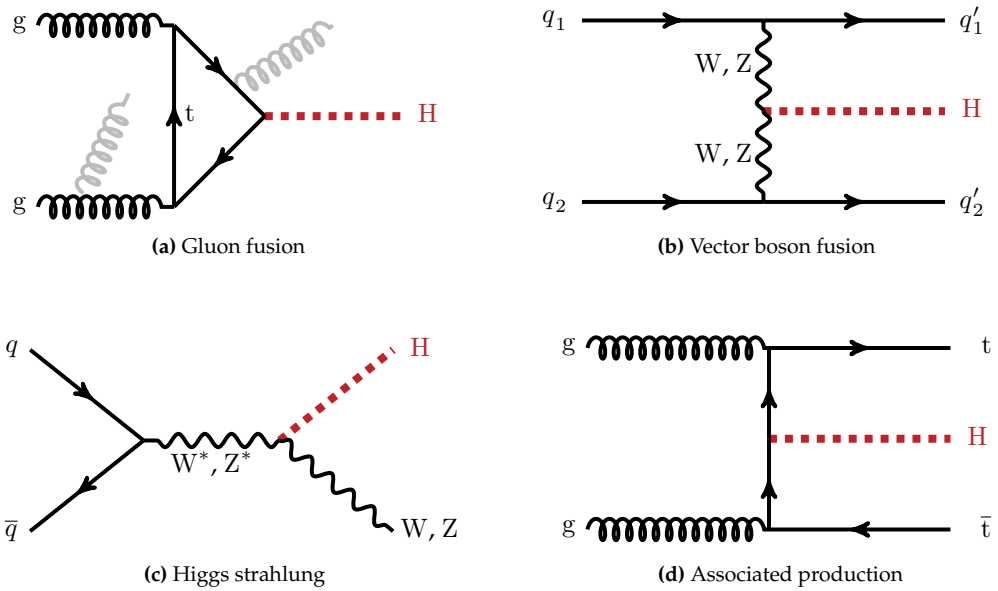
The main production mode is the gluon fusion (figure 1.5a). Since the Higgs boson only couples to the mass of particles, the production is mediated by a loop of heavy quarks or W bosons. The theoretical description of this loop is a source of substantial systematic uncertainties.

The cross section for the second important production mode, vector boson fusion (figure 1.5b), is about one order of magnitude smaller. The importance of this production mode is given by the two forward jets that radiate the vector bosons forming the Higgs boson. It is also characterised by only low hadronic activity in the central part of the detector between the two quark jets. Other SM processes are rarely produced with this signature. Therefore this production mode provides a good signal to background ratio in the search for Higgs bosons.

Especially for light Higgs masses two other processes contribute: the Higgs strahlung (figure 1.5c) together with W or Z boson and the associated production (figure 1.5d) together with two heavy quarks. Leptonic vector boson decays and b-tagged jets from the top quark decays help exploiting these processes although their cross sections are about two orders of magnitude smaller than the one of the gluon fusion process. The Higgs strahlung process has been the most important Higgs boson production mode at the proton-antiproton collider Tevatron. At the LHC this production mode is suppressed, since at a proton-proton collider the antiquark has to be a sea quark.

Since the Higgs boson couples to the mass of elementary particles, it preferably decays into the heaviest particle-antiparticle pairs that are kinematically allowed. Figure 1.7 depicts the branching ratios of the different decay modes as a function of the Higgs boson mass.

Light Higgs bosons primarily decay into b quark or τ^- lepton pairs. The decay into pairs of gluons, which is mediated via a heavy-quark loop, is experimentally not accessible due to a large irreducible QCD background, although the branching ratio is comparably large. Higgs events with b jets in the final state require a sophisticated b-tagging strategy due to the large hadronic background. This augments the importance of studying Higgs bosons decaying into pairs of τ leptons as presented in this thesis.



(a) Gluon fusion **(b)** Vector boson fusion

(c) Higgs strahlung **(d)** Associated production

Figure 1.5.: Leading order Feynman diagrams of the four main Higgs production modes at hadron colliders.

Table 1.3.: Predicted Higgs boson production cross sections at the LHC and their uncertainties from scale variations and PDF and α_S uncertainties at three centre-of-mass energies [21–23]. The values are given for a Higgs boson mass hypothesis of $m_H = 125$ GeV.

Production Mode	**Cross Section / pb $\pm$ QCD Scale Uncertainty $\pm$ PDF and α_S Uncertainty**		
	$\sqrt{s} = 7$ **TeV**	$\sqrt{s} = 8$ **TeV**	$\sqrt{s} = 13$ **TeV**
ggH	$15.13\,^{+7.1}_{-7.8}\,\%\,^{+7.6}_{-7.1}\,\%$	$19.27\,^{+7.2}_{-7.8}\,\%\,^{+7.5}_{-6.9}\,\%$	$43.62\,^{+7.4}_{-7.9}\,\%\,^{+7.1}_{-6.0}\,\%$
$pp \to qqH$	$1.222 \pm 0.3\,\% \,^{+2.5}_{-2.1}\,\%$	$1.578 \pm 0.2\,\% \,^{+2.6}_{-2.8}\,\%$	$3.748 \pm 0.7\,\% \pm 3.2\,\%$
$pp \to WH$	$0.5785 \pm 0.9\,\% \pm 2.6\,\%$	$0.7046 \pm 1.0\,\% \pm 2.3\,\%$	$1.380\,^{+0.7}_{-1.5}\,\% \pm 2.2\,\%$
$pp \to ZH$	$0.3351 \pm 2.9\,\% \pm 2.7\,\%$	$0.4153 \pm 3.1\,\% \pm 2.5\,\%$	$0.8696 \pm 3.8\,\% \pm 2.2\,\%$
$pp \to t\bar{t}H$	$0.08632\,^{+3.2}_{-9.3}\,\% \pm 8.4\,\%$	$0.1293\,^{+3.8}_{-9.3}\,\% \pm 8.1\,\%$	$0.5085\,^{+5.7}_{-9.3}\,\% \pm 8.8\,\%$

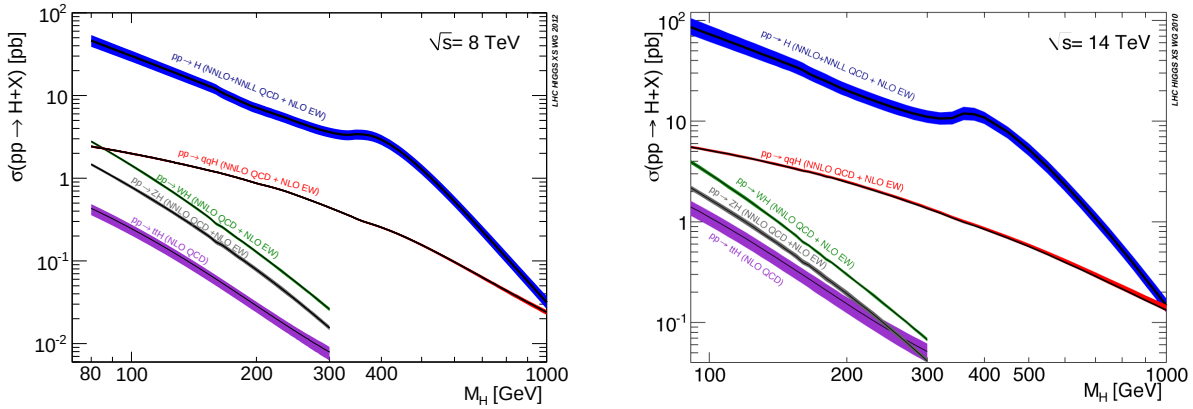


Figure 1.6.: Predicted Higgs boson production cross sections at the LHC and their total uncertainties as a function of the Higgs boson mass hypothesis for centre-of-mass energies of $\sqrt{s} = 8$ TeV (left) and 14 TeV (right) [21–23].

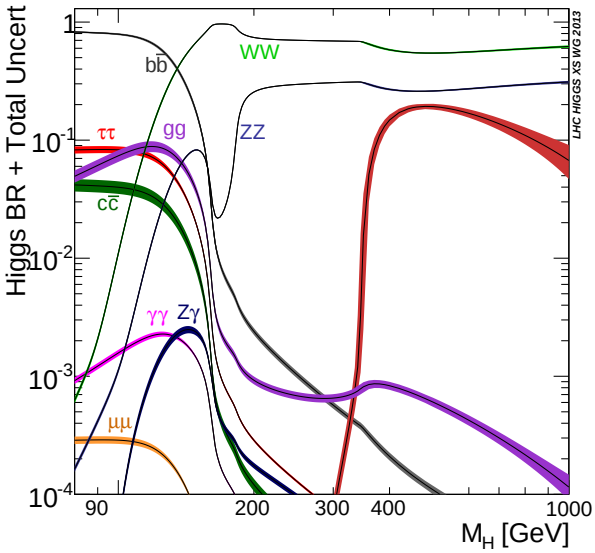


Figure 1.7.: Standard Model Higgs boson decay branching ratios [21–23].

For low masses the investigation of Higgs bosons decaying into pairs of photons (via a quark loop) is very powerful. Although the branching ratio is relatively small, this decay mode gains from the fact that it is almost only overlayed by the well known Drell-Yan background, which can be modelled easily. Above the WW and ZZ mass thresholds the decay of heavy Higgs is dominated by these pairs of vector bosons. From about 350 GeV the decay into $\mathrm{t\bar{t}}$ pairs gets relevant.

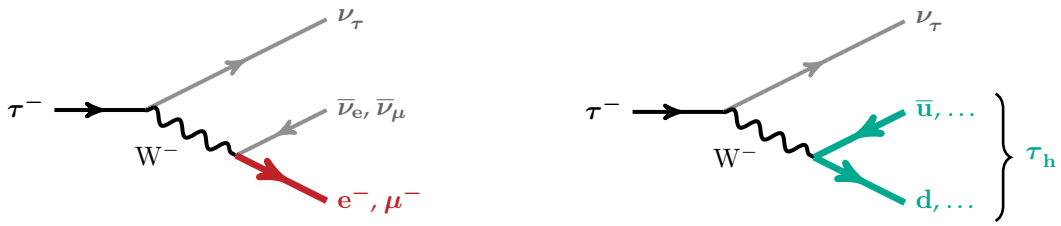
Decay into Pairs of τ leptons

In the search for light Higgs bosons the decay into pairs of τ leptons is very important. One reason is the comparably high branching ratio (see figure 1.7). About almost 4 % to 9 % of all Higgs bosons with masses below 140 GeV decay into pairs of τ leptons. Another reason is the fact that the irreducible background from Z boson decays, is well known and less dominant background processes can be suppressed with little effort. Together with the $\mathrm{b\bar{b}}$ channel the $\tau\tau$ channel shares the opportunity to intuitively probe the Yukawa-type fermionic couplings of the Higgs boson[6].

Figure 1.8 shows the Feynman diagrams for leptonic and hadronic τ decays at leading order. Two neutrinos in the leptonic and one in the hadronic mode carry away energy and momentum that cannot be directly measured with a detector such as CMS. This makes a full reconstruction of the four-momentum of the original τ lepton impossible. Hadronic τ decays may lead to resonances in the spectrum of the visible decay products.

Hadronic tau decays are classified in one, three or five prong decays. The number of prongs refers to the number of charged hadrons (mostly pions) within the τ jet. They are accompanied by neutral

[6]The production via gluon fusion is expected to be driven by the fermionic coupling to top quarks. However, W bosons can also contribute in the loop. This makes the direct probing of fermionic couplings in the production mode difficult.



(a) Leptonic decay **(b)** Hadronic decay

Figure 1.8.: Feynman diagrams for the leptonic and hadronic decay of the τ lepton.

Table 1.4.: Leptonic and hadronic τ decay modes together with their branching ratios [24, 25]. Although the pions represent the majority of the hadronic decays, the suppressed decay into kaons is possible too, which is not included in the table.

Decay Mode	Resonance	Branching Ratio / %
$\tau^- \to e^- \overline{\nu}_e \nu_\tau$		17.8
$\tau^- \to \mu^- \overline{\nu}_\mu \nu_\tau$		17.4
$\tau^- \to \pi^- \nu_\tau$	$\pi(140)$	11.6
$\tau^- \to \pi^- \pi^0 \nu_\tau$	$\rho(770)$	26.0
$\tau^- \to \pi^- \pi^0 \pi^0 \nu_\tau$	$a_1(1260)$	10.8
$\tau^- \to \pi^- \pi^+ \pi^- \nu_\tau$	$a_1(1260)$	9.8
$\tau^- \to \pi^- \pi^+ \pi^- \pi^0 \nu_\tau$		4.8
Other hadronic modes		1.7
All hadronic modes		64.8

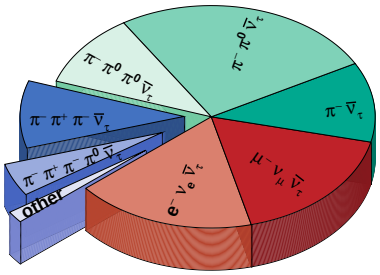
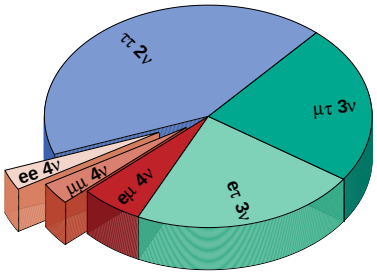


Table 1.5.: Branching for decays of $\tau\tau$ pairs. There are three fully leptonic, two semi-leptonic/hadronic and one fully hadronic decay mode.

Decay Mode	Branching Ratio / %
$\tau\tau \to \mu + \mu + \nu\nu\nu\nu$	3.0
$\tau\tau \to e + e + \nu\nu\nu\nu$	3.2
$\tau\tau \to e + \mu + \nu\nu\nu\nu$	6.2
$\tau\tau \to \mu + \tau_h + \nu\nu\nu$	22.5
$\tau\tau \to e + \tau_h + \nu\nu\nu$	23.1
$\tau\tau \to \tau_h + \tau_h + \nu\nu$	42.0



hadrons (again mostly pions). Table 1.4 shows the most prominent τ decay modes together with their branching fractions.

Tauons having a mass of $m\tau = 1777$ MeV are the heavy siblings of the lighter leptons, electrons and muons. Since they decay only weakly, high energetic τ leptons can fly mean distances of $c\,\Delta t = 87$ µm before they decay. This enables the possibility to resolve impact parameters of the tracks of the decay products and secondary vertices distinct from the primary production vertex for specific decay modes. The sufficiently high mass is also the reason for large branching fraction of about 65 $\%$ of hadronic τ decays.

Further details on τ decays and their reconstruction are given in section 2.4.4.

The branching ratios listed above lead to the branching ratios for the decays of $\tau\tau$ systems as listed in Table 1.5. There are three fully leptonic, two semi leptonic/hadronic and one fully hadronic decay modes.

1.2.3. Experimental Verification

In experimental Higgs analyses direct and indirect searches for the Higgs boson are distinguished. Indirect measurements allow for probing the theory for much higher Higgs masses as they are directly accessible.

The masses of the vector bosons, W and Z, have been precisely measured at the electron-positron collider LEP and also at the Tevatron. They depend on the Higgs boson mass via quantum loop corrections and therefore it is possible to constrain the Higgs mass from fits to the vector boson masses. Such fits to electroweak precision measurements [17] yield a most probable Higgs mass of $m_H = 94^{+29}_{-24}$ GeV at 68 $\%$ confidence level. At 95 $\%$ confidence level this analysis could exclude Higgs masses above 152 GeV. By direct searches for the Higgs boson at LEP and the Tevatron, masses below 114.4 GeV and between 156 and 177 GeV could be excluded at 95 $\%$ confidence level.

Due to the higher centre-of-mass energy of proton collisions at the LHC it is possible to probe Higgs boson masses that could not be excluded at predecessor colliders. As already mentioned, at both multi-purpose detectors ATLAS and CMS the observation of a new boson with a mass around 125 GeV at the 5 σ was published in July 2012 [8, 26]. This result was based on a combination of all analyses available at that time, where mainly the golden channels $\mathrm{H} \to \gamma\gamma$[7] and $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ contributed. Both channels profit from a very precise Higgs boson mass resolution and from low and well controlled background contributions. Figure 1.9 shows mass distribution for the complete data from the run-I data set taken in 2011 and 2012. A significant excess over the background-only hypothesis is visible and compatible in both channels.

Based on the same data set analyses in all accessible decay channels have been performed. In addition to the analyses in the $\mathrm{H} \to \gamma\gamma$ [27] and the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ [28] channels the $\mathrm{H} \to \mathrm{WW}$ [29], $\mathrm{H} \to \tau\tau$ [10] and $\mathrm{H} \to \mathrm{bb}$ [30] channel analyses have to be mentioned. In all of these channels apart from

[7] Couplings between the Higgs boson and pairs of photons are mediated via a loop of heavy quarks. Direct couplings between Higgs bosons and the massless photons are forbidden since the Higgs boson only couples to massive particles.

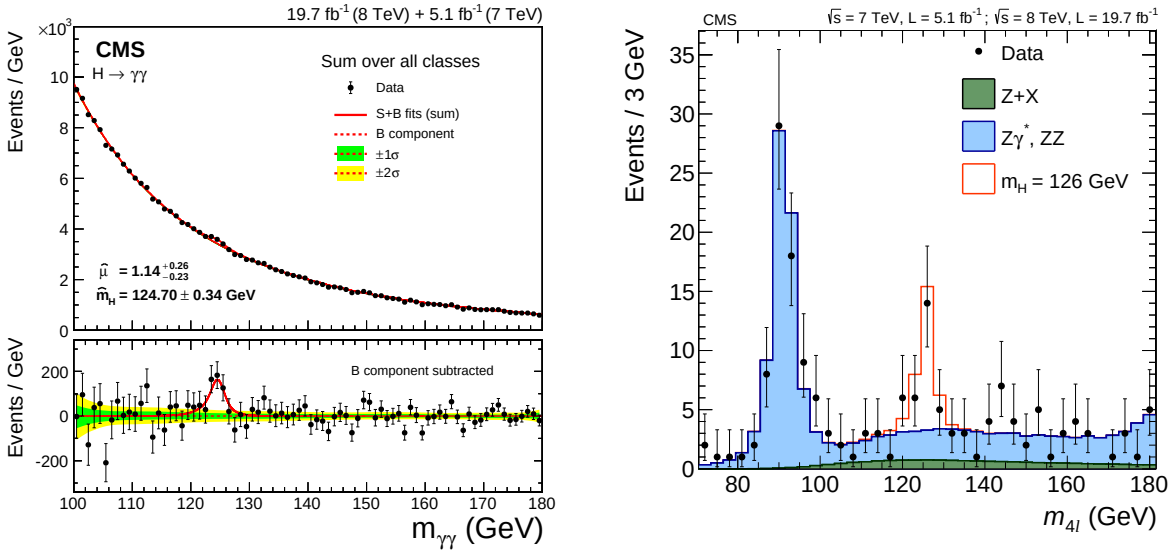


Figure 1.9.: Higgs boson mass resonances found in the $\mathrm{H} \to \gamma\gamma$ channel (left) and the $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ channel (right) [27, 28].

$\mathrm{H} \to \mathrm{bb}$ an excess is seen with a significance of at least 3 σ that is compatible with the Higgs boson observation at 125 GeV. The $\mathrm{H} \to \mathrm{bb}$ channel suffers from the large QCD multi-jet background. A larger data set is needed to reach the precision needed for a 3 σ evidence.

The results of all channels have been combined and various measurements of the Higgs boson properties have been performed [26]. The Higgs boson mass is measured to yield the following value.

$$m_{\mathrm{H}} = 125.02\ ^{+0.26}_{-0.27}\,(\mathrm{stat})\ ^{+0.14}_{-0.15}\,(\mathrm{syst})\ \ \mathrm{GeV}$$

As already pointed out all other properties if the SM Higgs boson can be predicted as a function of one parameter, the Higgs boson mass. This provides a tool for comparing the observed resonance with the SM prediction.

Measurements of the signal strengths in various sub-channels can be transformed into coupling parameters describing the couplings of Higgs bosons to other particles or classes of them. Two main results are shown in figure 1.10.

The individual channels couple differently to vector bosons and fermions, depending on the Higgs boson production and decay modes. Thus it is possible to derive constraints on the coupling strength modifiers for vector bosons and fermions, κ_{V} and κ_f. The 1 σ contours are shown for the individual channels as well as for their combination in figure 1.10 (left). The results agree with the SM expectation of $\kappa_{\mathrm{V}}^{\mathrm{SM}} = \kappa_f^{\mathrm{SM}} = 1$ within 1 σ. It is clearly visible that analyses of vector boson final states provide stronger constraints on the vector boson coupling, κ_{V}, and that analyses in fermionic final states, mainly the $\mathrm{H} \to \tau\tau$ analysis, yield a higher precision for the fermionic coupling, κ_f. However, interferences between heavy quarks and vector bosons in the loops mediating the couplings to massless particles such as gluons in the production and photons in the decay make it impossible to fully disentangle couplings to fermions and vector bosons.

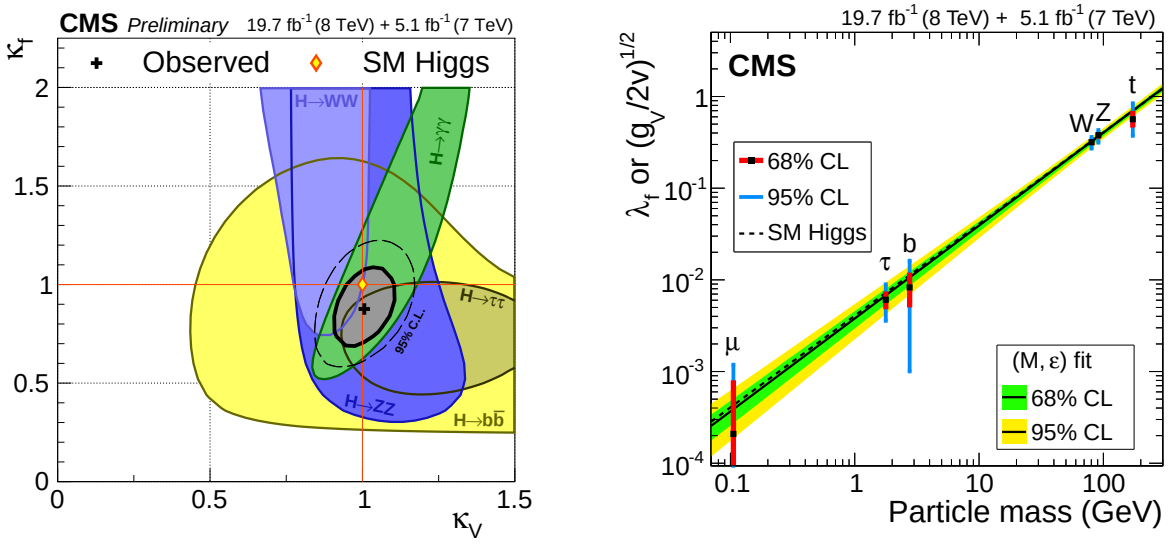


Figure 1.10.: Measurements of the Higgs boson couplings [26].

In a similar way the coupling to the individual particles are determined in a combined analysis of all channels. Figure 1.10 (right) shows the corresponding coupling constants as a function of the particle mass. The coupling scales are different for fermions and bosons are different since the coupling structures are fundamentally different: the coupling between massive vector bosons and the Higgs boson is proportional to the square of the vector boson mass and Yukawa couplings are proportional to the fermion masses. The fit of these couplings nicely follows the SM prediction of couplings proportional to the mass of the particles. This is a strong indication for the Higgs boson being responsible for the generation of the mass of elementary fermions and bosons.

The next chapter 2 gives an overview of the CMS detector at the LHC, before the search for $H \to \tau\tau$ events in the di-muon channel is presented in detail in chapter 3 and an overview of the $H \to \tau\tau$ analysis in general is given in chapter 4 in the form of an outlook to the future analysis of the 13 TeV data from CMS before the results of the 7+8 TeV analysis have been summarised.

2 The CMS Detector at the LHC

2.1. The Large Hadron Collider

The large hadron collider (LHC) is a proton-proton collider hosted by CERN[1] situated near Geneva (Switzerland) at the French border. The hadron collider delivers particle collisions at the highest centre-of-mass energies that are reached up to now. With a designed centre-of-mass energy of 14 TeV it supersedes the Tevatron collider near Chicago (USA) with centre-of-mass energies of up to 1.96 TeV. Probing the electroweak symmetry breaking was mainly driving the design planes for the accelerator as well as for the experiments. A detailed technical introduction is given in the LHC Design Reports [31–33].

Before proton bunches are injected into the 27 km long LHC ring at energies of 450 GeV, they pass several pre-accelerators as shown in figure 2.1. In the ring accelerators the bunches gain energy from microwave radiation in cavities. The cavities also ensure a longitudinal stability of the bunches. The bending is performed by superconducting dipole magnets cooled by superfluid helium which are designed for magnetic fields of up to 8.3 T in the LHC ring. The radial focussing of the beam particles is performed by quadrupole and higher order magnets.

The LHC provided proton collisions in the first data-taking period from 2010 to 2012 and from 2015 on in the second data-taking period. The start-up phase until the end of 2010 was mainly focussed on machine commissioning. In 2011, collisions at a centre-of-mass energy of 7 TeV have been achieved. This measure was increased to 8 TeV in 2012. This period is regarded to as LHC run I. After a long shut-down in the years 2013 to 2015, the recommissioning of the machine started with first collisions at $\sqrt{=}13$ TeV centre-of-mass energy in 2015. This marks the beginning of the LHC run II phase. The following section refer to the first running period of the LHC and CMS. An overview of the differences between run I and II is given in section 2.3.2.

The LHC provides four major experiments at four interaction points with collision events. Both ATLAS[2] [35] and CMS[3] [36] are general-purpose particle detectors. The as well suited for the study of a

[1] Conseil Européen pour la Recherche Nucléaire (engl.: European Organisation for Nuclear Research)
[2] A Toroidal LHC Apparatus
[3] Compact Muon Solenoid

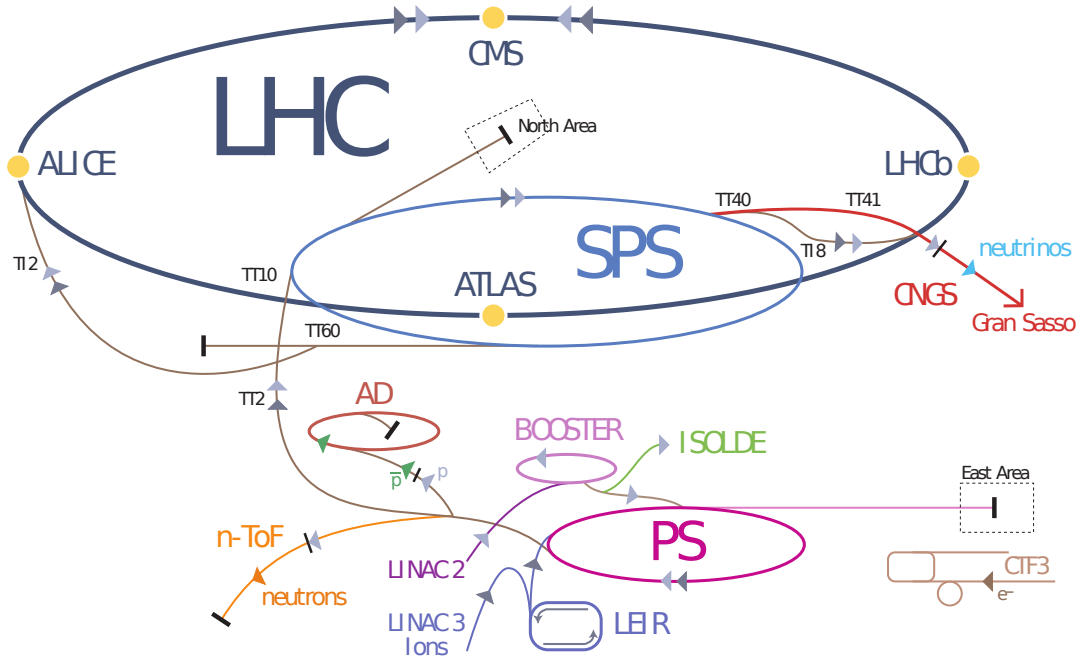


Figure 2.1.: The CERN accelerators complex [34] showing the linear accelerator (LINAC 2), the circular pre-accelerators, BOOSTER, Proton Synchrotron (PS) and Super Proton Synchrotron (SPS), and the LHC ring with its four interaction points, where the detectors ATLAS, CMS, ALICE and LHCb are placed.

variety of fields, starting at QCD processes over top and Higgs physics leading to the exploration of possible physics beyond the Standard Model. However, they have been optimised for the discovery of Higgs bosons. The two experiments provide the opportunity to independently confirm each other. Additionally, ALICE[4] [37] focuses on studying heavy ion collisions and the resulting quark-gluon-plasma whereas the asymmetric detector LHCb[5] [38] concentrates on the CP violation in B hadrons and other rare decays.

2.2. The Compact Muon Solenoid Detector

The compact muon solenoid (CMS) experiment is one of the two general-purpose detectors designed to investigate particle physics up to the TeV scale. One main task is to study the electroweak symmetry breaking by directly searching for the predicted Higgs boson (see section 1.1.4). Additionally, the Standard Model has to be probed at the TeV energy scale. As one of the largest detectors ever built, CMS has been designed to achieve an excellent muon detection, a good di-photon mass resolution and a spatial tracking resolution that enables the reconstruction of secondary vertices of τ leptons and b-quarks. The high demands from physicists and the technological feasibility require an elaborate construction. The cooling, the mechanical stability and the radiation hardness are big challenges.

The basic onion-type structure is depicted in figure 2.2. In the barrel region sub-detectors are arranged in layers around the collision point. End-caps with the same sub-detectors cover these tubes.

[4] A Large Ion Collider Experiment
[5] Large Hadron Collider beauty

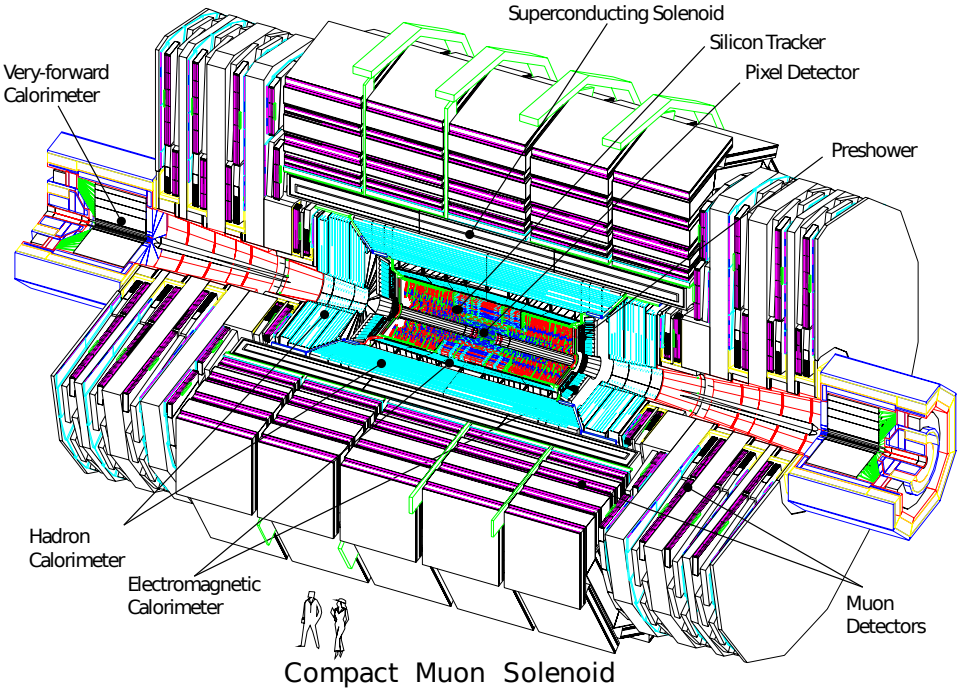


Figure 2.2.: Three-dimensional schematic view of CMS [39].

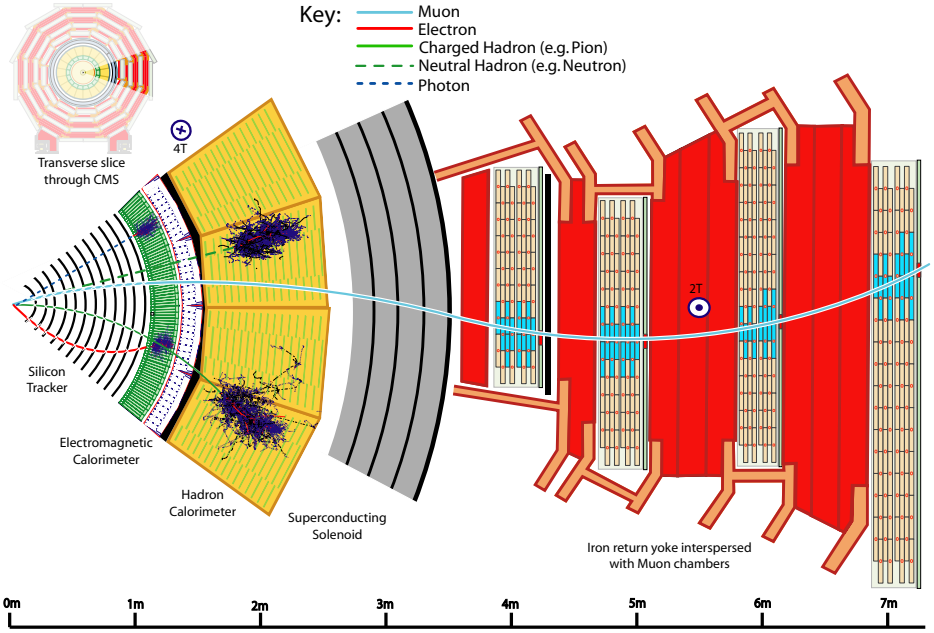


Figure 2.3.: Slice trough the CMS detector showing signatures of different types of particles traversing the various sub-detectors [39]

A solenoid magnet incorporates the tracking detector and the calorimeters. This requires a compact design. Outside the coil is the muon system. A brief overview of these components is given in the following sub-sections. More detailed information can be taken from the CMS Physics Technical Design Reports [40, 41] and from a more recent official documentation [36]. In the following all numbers are taken from these sources unless they are indicated differently.

The most important measurements comprise the three-momentum and the energy of high energetic particles traversing the detector material and their tracks close to the interaction point. The basic principles will be described in the following. Figure 2.3 shows the different kinds of particle interactions with the detector components.

In total, CMS measures 21.6 m in length, 14.6 m in diameter and about 12500 t in weight. About 10^8 data channels can be read out every 25 ns.

2.2.1. Coordinate System

The CMS coordinate system has its origin in the nominal collision point. In Cartesian coordinates, the x axis points radially inward with respect to the LHC ring, the y axis points vertically upward and the z axis points along the beam axis. Together, the three axes form a right-handed coordinate system.

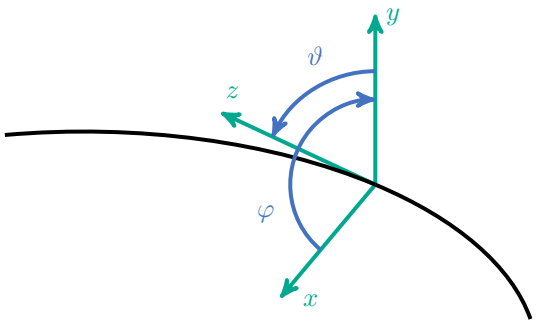


Figure 2.4.: The coordinate system of the CMS detector with its origin at the nominal collision point.

In spherical coordinates r denotes the distance from the origin of the coordinate system whereas φ is the azimuthal angle in the x-y plane measured from the x axis and ϑ is the polar angle in the y-z plane measured from the beam line. The angle φ ranges from $-\pi$ to $+\pi$. Usually, the polar angle ϑ is translated into the non-dimensional pseudorapidity η defined as

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

The pseudorapidity yields the value 0 for points in the x-y plane with $\vartheta = \frac{\pi}{2}$ and increases while ϑ degreases to zero. In comparison to the rapidity y, the pseudorapidity only takes into account the flight direction of a particle given by its three-momentum and is therefore independent of the particle energy and mass. Angular distances, ΔR, between to direction vectors can then be quoted according to the following formula.

$$\Delta R = \sqrt{\Delta\varphi^2 + \Delta\eta^2}$$

The azimuthal difference is between the two vectors is denoted by $\Delta\varphi$ and $\Delta\eta$ is the difference in the pseudorapidity.

Longitudinal directions (index L) always refer to the beam line. The x-y plane is referred to as the transverse plane (index T). The transverse momentum $p_{\mathrm{T}} = \sqrt{p_x^2 + p_y^2}$ and the missing transverse energy $\not{E}_{\mathrm{T}}$ coming from the transverse energy imbalance are examples for variables defined in the transverse plane.

2.2.2. Silicon Tracking Detector

Tracking detectors reconstruct hits of charged particles, the points where the particles traversed the tracking material. These hits are the reconstructed as tracks of individual particles. Inside the almost homogeneous magnetic field of the magnet the tracks of charged particles get bent. The bending radius of the tracks is a measure for the transverse component of the momentum, given the measured magnetic field and hypothesis of the particle mass. The full three-momentum is inferred from the orientation of the track within the detector.

The inner silicon tracking tracking system reconstructs hits of tracks from charged particles as close as possible to the interaction point. The hits are detected in silicon semiconductor cells. A p-n junction in reverse direction has a region with almost no free charges between the p-type and the n-type semiconductor. A ionising particle passing through this zone causes a measurable current that is reconstructed as hit at the position of the cell. That is the functional principle of each silicon tracking cell [42].

The beam pipe is surrounded by three cylindrical layers of pixel detectors at radii between 4.4 and 10.2 cm that are covered by two layers of disc pixel modules at each side. Each pixel gives a three-dimensional information per hit. Its high spatial resolution ensures the capability to precisely measure particle momenta and to reconstruct secondary decay vertices. For example, the transverse impact parameter resolution of high p_T tracks with a value of 10 µm is dominated by the pixel size of 100×150 µm^2 in r-φ and z.

Adjacent to the pixel detector, multiple silicon strip detectors are installed up to an outer radius of $r = 116$ cm, a longitudinal distance of $|z| = 282$ cm and a maximum pseudorapidity of $|\eta| = 2.5$. These components only provide a two-dimensional information. The single layers are tilted towards each other in order to obtain three-dimensional information with multiple modules, as long as the particle flux is small enough to avoid ambiguities. In total, 66 million pixels and 9.3 million strips are read out by this sub-detector which covers an overall area of about 200 m^2. A schematic view of the sub-detector is shown in figure 2.5.

This innermost sub-detector is exposed the high particle flux near the collision point. Although all modules are designed for this environment, radiation damages reduce its lifetime to about one decade. Another problem is the high electrical power density. The necessary cooling system reduces the volume of active detector material and yields additional unwanted inference of the particles with matter.

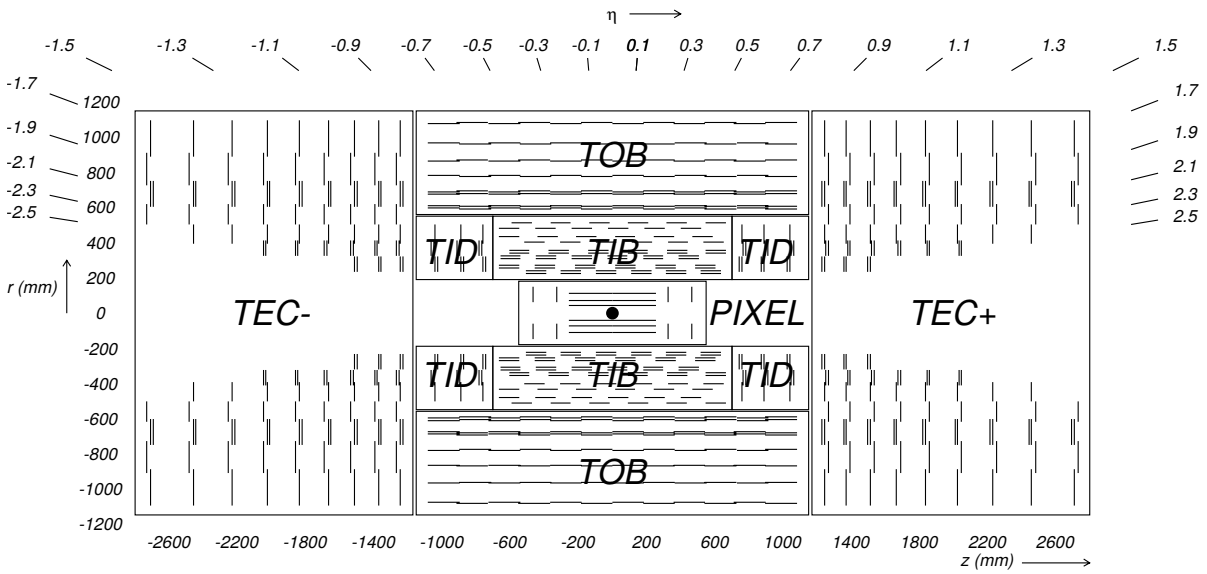


Figure 2.5.: The inner silicon tracking system of CMS [36]. The pixel detector (PIXEL) is surrounded by the strip tracker. As parts of this component the Tracker Inner Barrel and Disks (TIB/TID) are followed by Tracker Outer Barrel (TOB) and the Tracker EndCaps (TEC).

2.2.3. Electromagnetic Calorimeter

In the electromagnetic calorimeter basically photons and electrons, are stopped. They deposit their energy in the dense absorber material by producing electromagnetic showers. Incoming electrons radiate photons via bremsstrahlung, whereas photons themselves produce pairs of electrons and positrons until their energy lies below the pair production threshold. Therefore a shower of secondary particles spreads out over the calorimeter. Scintillators are then used to detect these particles. Its material is excited by high energetic photons and radiates scintillation light that can be detected by photo diodes. The energy of the incoming particle is proportional to the number of generated photons. Therefore the statistically induced energy uncertainty is $\sqrt{E}$ [42].

CMS uses homogeneous lead tungstate ($\mathrm{PbWO_4}$) crystals both as absorber with a high density and a short radiation length and as scintillator material. This allows a compact calorimeter. The small Molière radius results in a high granularity of the sub-detector and the scintillation decay time is short enough to detect approximately 80 % of the light within the time of 25 ns between the proton bunches. 61200 crystals in the central barrel part of the detector and 7324 ones in each of the end-caps surround the collision point hermetically up to $|\eta| = 3$. The scintillation light is read out by avalanche photo diodes (APDs) in the barrel and vacuum photo diodes (VPTs) in the end-caps. Additionally, a pre-shower sampling calorimeter is installed in front of the end-cap crystals to reject neutral pions. Figure 2.6 shows a schematic view of the sub-detector.

The main goal of the design of the electromagnetic calorimeter was the capability to resolve the to two photons coming from a possible Higgs boson with a good energy resolution. The energy resolution of the ECAL, σ, consists of the stochastic term due to the counting of photons in the scintillators, a noise

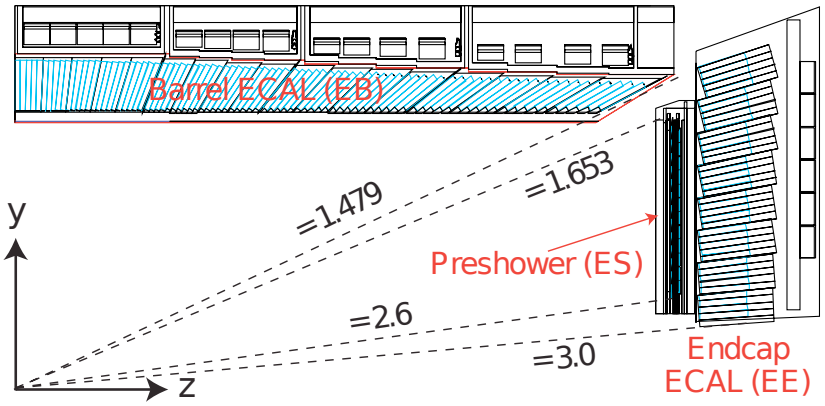


Figure 2.6.: Schematic view of the electromagnetic calorimeter of CMS [40]. The dashed lines signalise polar angles in terms of the pseudorapidity η.

term proportional to the energy and a constant offset.

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{2.8\,\%}{\sqrt{E/\mathrm{GeV}}}\right)^2 + \left(\frac{12\,\%}{E/\mathrm{GeV}}\right)^2 + (0.3\,\%)^2$$

2.2.4. Hadronic Calorimeter

Hadronic calorimeters are used to measure the energy of hadrons. Their functional principle is the same as the one of electromagnetic calorimeters. Hadronic particles, basically protons, neutrons, pions and kaons within jets, produce showers in dense absorber materials depositing their energy before they are stopped. The deposit of their energy can be measured with scintillators and photo diodes.

CMS used a hadronic sampling calorimeter where brass absorber plates alternate with plastic scintillators. Only a small fraction of the energy of an incoming particle is deposited in the scintillator and gets measured. In order to account for this effect, the energy has to be corrected. It also causes a worse energy resolution compared with the homogeneous electromagnetic calorimeter. The absorber thickness in the barrel region is between 5.82 and 10.6 interaction lengths depending on the polar angle while the electromagnetic calorimeter adds about 1.1 interaction lengths. Since not all high energetic hadrons can be stopped within the volume limited by the surrounding magnet coil, an additional outer calorimeter detects tails of these showers. The CMS calorimeters are completed by a forward calorimeter placed 11.2 m away from the beam crossing which is used to measure the instantaneous luminosity. A schematic view of the sub-detector is shown in figure 2.7.

The energy resolution of the HCAL, σ, consists of the stochastic term due to the counting of photons in the scintillators and a constant offset.

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{100\,\%}{\sqrt{E/\mathrm{GeV}}}\right)^2 + (5\,\%)^2$$

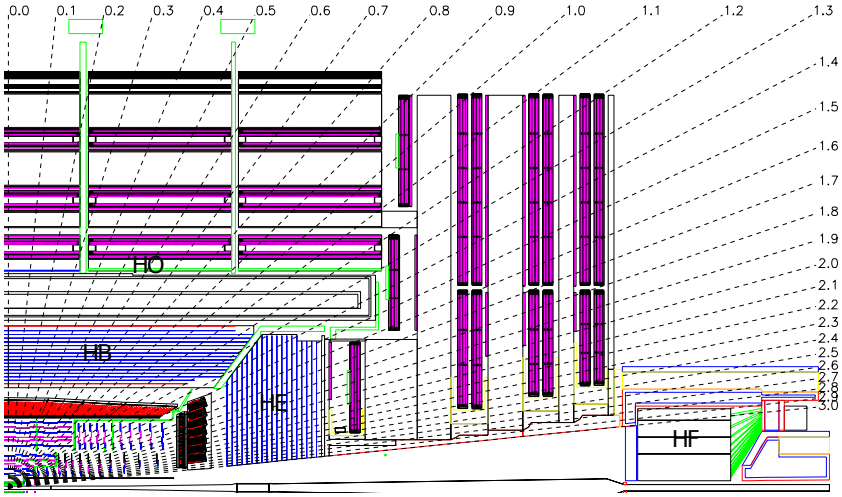


Figure 2.7.: The CMS detector with respect to the hadronic calorimeter [36]. The hadron barrel (HB) and end-cap (HE) calorimeters are located inside the magnet coil, whereas the outer (HO) calorimeter measures the tails of high energetic showers outside the coil. The forward calorimeter (HF) is placed 11.2 m away from the beam crossing. The dashed lines signalise polar angles in terms of the pseudorapidity η.

On account of the large spatial coverage of the calorimeter the total energy deposit in the detector can be well measured which helps drawing conclusions about the missing transverse energy originating from undetected neutrinos or exotic particles.

2.2.5. Superconducting Solenoid

A superconducting solenoid magnet comprising the tracking detector and the electromagnetic and inner hadronic calorimeters provides a longitudinal magnet field with a maximum field strength of 4 T. The magnetic field is needed to bend the tracks of charged particles. The curvature of a track of a charged particle, which is bended by the magnetic field, indicates the momentum of the particle. A high momentum resolution is achieved by a good spatial resolution of the tracking detectors and a high magnetic field strength. The precision of the momentum measurements decreases with increasing momentum as the tracks get less bent.

Due to the compact design of the inner sub-detectors, these could be placed inside the magnetic coil with an inner radius of about 3 m and a length of 12.5 m. The magnetic flux is returned by a 10000 t steel yoke.

2.2.6. Muon System

Most of the particles traversing the detector are shielded by the inner detector components and the solenoid magnet. Of the known particles, only muons and the undetectable neutrinos reach beyond the calorimeters. Thus, a tracking detector specialised on muons is installed outside the magnet coil between the yoke elements. Low noise levels result from this shielding effect. The muon system is built from gaseous particles detectors. The working principle is the same as for the silicon tracking detectors

near to the collision point. Traversing charged particles (respectively muons) ionise gas atoms. Then the charge carriers are collected at wires and measured as a current. The spatial information of hits and then tracks is taken from the position of the wires that showed a signal.

CMS uses three different kinds of gaseous particle detectors covering an angular interval of $|\eta| < 2.4$ in total. The barrel region is equipped with drift tubes whereas the end-caps use cathode strip chambers that are capable of working in inhomogeneous magnetic fields. Additional fast resistive plate chambers are taken for triggering. A schematic view of the sub-detector is shown in figure 2.8.

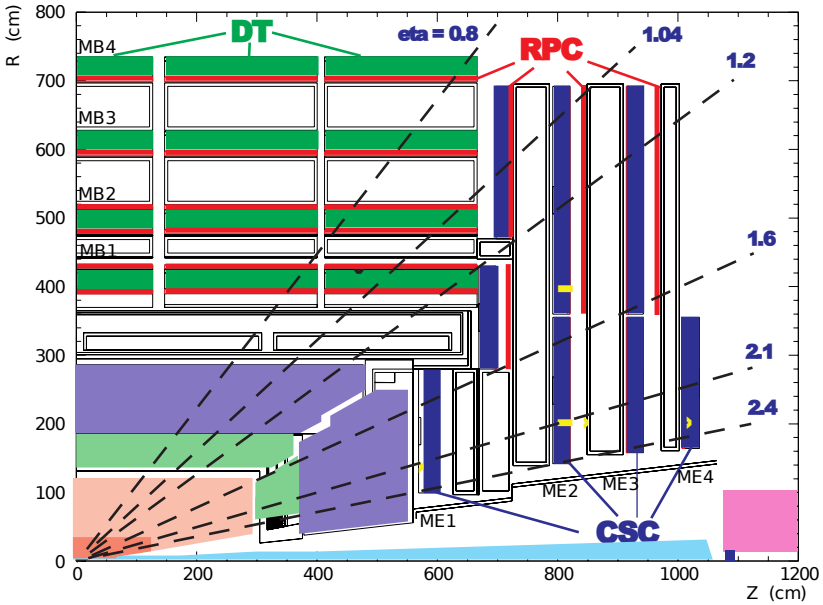


Figure 2.8.: The CMS detector with respect to the muon system [40]. The barrel drift tubes (DT) and the end-cap cathode strip chambers (CSC) are completed by resistive plate chambers (RPC). Iron yoke components intersect the muon system to return the magnetic flux of the solenoid.

The momentum resolution of the muon system on its own (about 9 % for muons with $p_T \leqslant 200$ GeV) can be improved by one order of magnitude in combination with the inner tracking system. Muons are objects of high interest, since they appear in many final states and can be detected with high efficiencies.

2.2.7. Data Acquisition

The LHC is designed for a bunch crossing rate of 40 MHz and an instantaneous luminosity of $L = 34\ \mathrm{cm}^{-2}\,\mathrm{s}^{-1}$), which results in about 20 proton-proton interactions per bunch crossing. The CMS detector has to cope with these high collision rates. One bunch crossing produces about 1-2 MB uncompressed data. Therefore it is obviously impossible to read out the full detector at this high rate and store the complete data in real time. The data rate has to be dramatically reduced. Events with low momentum transfers that are not of interest for physics analyses have to be discarded The ones with objects of high transverse momentum have to be kept.

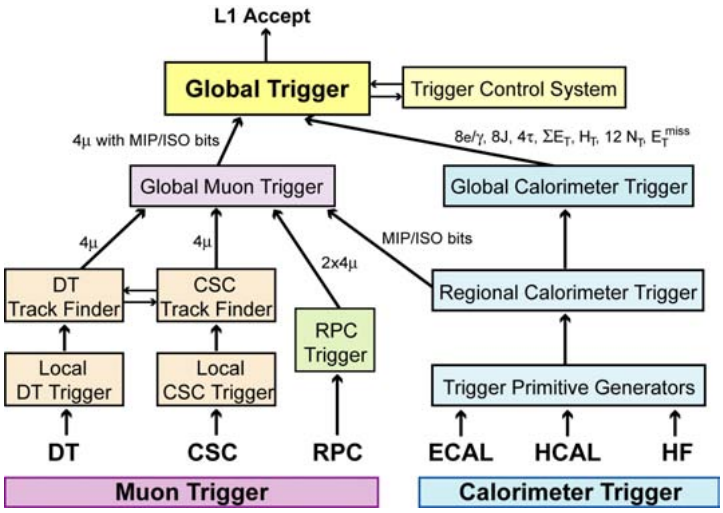


Figure 2.9.: Architecture of the Level-1 Trigger [36].

CMS employs a two-staged triggering system to decide, which events to discards and which ones to store. At the first stage, the Level-1 trigger is implemented in programmable hardware like FPGAs, ASICs or LUTs. The high-resolution detector data is stored in pipelined memories in the front-end electronics while the trigger system evaluates corser data within up to 3.2 µs. First, energy deposits in the calorimeters and track candidates in the muon system are evaluated and ranked locally. The highest ranked triggering objects are then passed to a global stage which ranks all objects from the complete detector. The Level-1 trigger reduces the event rate from 40 MHz to 100 kHz. Figure 2.9 illustrates the architecture of the Level-1 trigger.

Events passing the Level-1 trigger are computed by a second-stage trigger, the so-called High Level Trigger (HLT). Since this trigger is implemented in software, it is able to evaluate the full detector information. Large computer farms are employed to process the events in parallel. The HLT algorithms reconstruct and evaluate the data very similar to the offline reconstruction software. The algorithms are constantly updated and adjusted to the increasing instantaneous luminosities in order to maintain the rate of events to be stored on tape. The High Level Trigger reduces the event rate by three orders of magnitude to reduce the amount of data that has to be stored to technically processable size. In the first running period of the LHC, about 300 events per second were promptly reconstructed and stored on tape. Further 700 events per second were stored in the raw format and reconstructed during the LHC shut-down. Figure 2.8 illustrates the data acquisition system of CMS.

It is obvious that such huge amounts of data requires for a sophisticated offline storage and processing of the data. High energy physicists chose a distributed computing infrastructure called the worldwide LHC computing grid (WLCG) [43,44]. Computing centres, referred to as sites, are located all over the world and share both data and computing power. The system is organised in a tiered structure. More details about the CMS computing model can be found in reference [45,46].

At CERN the sole Tier-0 site is located. Raw data coming from the data acquisition systems is stored on tape archives together with first reconstructed samples. This site also distributes the data to several Tier-1 sites [47] which are large computing centres responsible for the storage of copies of the data as

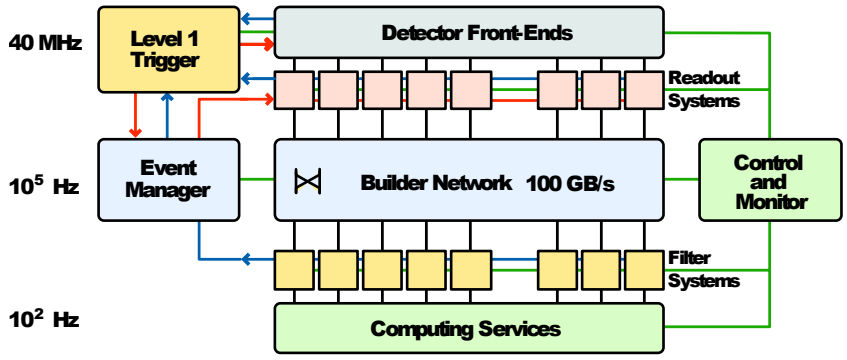


Figure 2.10.: The CMS data acquisition system [36].

well as for the large-scale re-reprocessing and skimming of the data. Fast network connections enable them to exchange data with the Tier-2 sites. Local resources for user analyses are combined by the multiple Tier-2 sites. Beside a grid-based analysis they are supposed to provide resources for event simulation. At last, Tier-3 sites form the final stage. These are small clusters that do not have to fulfil any demands towards the WLCG and are mainly used for user analyses. The access to grid resources is mediated by so-called middleware, a software layer that allows a standardised interaction between user software and the inhomogeneous grid hardware.

2.3. LHC and CMS Performance

In a first running period, the LHC provided collisions for physics in the years 2010 to 2013, before the operation was interrupted by a planned technical stop in the years 2013 to 2015. In 2015 the machine operation was resumed with the second running period. The next section 2.3.1 gives an overview of the run I and after that, section 2.3.2 presents the updates in the run II. The following list summarises the dates of important milestones.

- **September 10th, 2008:** first beams circulated in the LHC accelerator.
- **September 19th, 2008:** incident during powering test of the dipole magnets. A finite resistance between two superconducting magnets lead to a sudden release of helium and therefore to a mechanical damage of several dipole magnets in one sector of the LHC [48].
- **November 20th, 2009:** first beams after shut-down. Supplementary to the repair of the dipole magnets, additional fuses were installed in the helium systems to prevent mechanical damages after quenching of magnets.
- **March 30th, 2010:** first collisions at 7 TeV mark the begin of LHC run I
- **April 5th, 2012:** first collisions at 8 TeV
- **February 2nd, 2013:** end of LHC run I and begin of first long shut-down
- **June 6th, 2015:** first collisions at 13 TeV mark the begin of LHC run II

2.3.1. Performance in the First Data-taking Period

An important performance quantity of a collider is the instantaneous luminosity L. It is a measure for the number of particles colliding at a specific area per time interval and is based on machine parameters such as the geometry of the beam, the number of bunches and the number of particles per bunch. The product of the instantaneous luminosity and the cross section σ for a specific process yields the event rate $\frac{\mathrm{d}}{\mathrm{d}t}N$ for this process.

$$\frac{\mathrm{d}}{\mathrm{d}t}N = \sigma\, L \quad \text{with} \quad L = N_b\, N_{\mathrm{p}}^2\, \frac{f}{A}$$

Here, N_b denotes the number of bunches per beam, N_{p} the number of protons per bunch, f the revolution frequency of the beams and the area A is a measure for the geometrical cross section of the beams[6].

The instantaneous luminosity for the collisions in the CMS detector is determined in two ways. The number of primary vertices reconstructed based on the tracking detector or the activity measured in the forward calorimeter are relative measures for the instantaneous luminosity. The absolute scale is determined in special LHC runs, the Van Der Meer scans [49]. The achieved instantaneous luminosities[7] in 2011 reached up to values of $L = 4.02$ Hz/nb and in 2012 up to 7.67 Hz/nb [50].

The integrated luminosity $\mathcal{L}$ is a quantity measuring the amount of data accumulated by a colliding experiment, in case σ denotes the total cross section.

$$N = \sigma\, \mathcal{L} \quad \text{with} \quad \mathcal{L} = \int L\ \mathrm{d}t \tag{2.1}$$

With the cross section for a certain process is is also used to determine the number of expected events for this process.

Figure 2.11 depicts the cumulative distribution of the total integrated luminosity for the first running period (left and centre). The integrated luminosity delivered by the LHC is compared with the one of the data set recorded by the CMS experiment. The efficiency that is reduced by dead time of the detector, is above 90 % and therefore signalises a good operation of the detector. The data that has been certified for the physics analysis corresponds to integrated luminosities of 4.9 fb^{-1} at $\sqrt{s} = 7$ TeV and 19.7 fb^{-1} at 8 TeV, respectively. This means that 80 to 85 % of the delivered collisions is used for physics analyses.

The integrated luminosities of the three data sets of the first running period are compared in figure 2.12. In 2010, 45 pb^{-1} were collected. Due to ongoing machine development this is much less data compared to the 5.55 fb^{-1} taken in 2011 and the 21.79 fb^{-1} taken in 2012. Therefore, the data set from 2010 is neglected in most physics analysis because no significant improvement in the precision of the results based on the data sets from 2011 and 2012 is expected that would justify the additional

[6] The area, A, is usually expressed in terms the machine parameters γ, the Lorentz factor of the protons, ϵ_n, the normalized transverse emittance, β^*, the betatron function at the interaction point, and F, the reduction factor due to the crossing angle: $A = 4\pi\, \epsilon_n\, \beta^* \frac{1}{\gamma F}$

[7] Hz/nb $= 10^{33}/\mathrm{cm}^2/\mathrm{s}$

studies needed. The actual size of the data sets in the analysis is smaller than the one of the total data set. Only selected triggers are evaluated in the analysis.

The instantaneous luminosity is directly connected to the amount of pile-up interactions that analyses have to cope with. In most of the cases, a single bunch crossing leads to zero or one interaction between two protons that triggers the HLT. Several other protons of the two bunches can lead to secondary interactions with on average lower momentum transfer. They are called in-time pile-up interactions. On top of that, small bunch spacings can lead to the so-called out-of-time pile-up. This is activity in the detector that remained from the previous collisions and gets read out together with the current event. The number of reconstructed primary vertices is a measure for the amount of in-time pile-up interactions. Figure 2.13 compares its distribution for the different data sets. It is clearly visible that the amount of pile-up significantly increased in 2012 with respect to 2011 because of the increased instantaneous luminosity.

2.3.2. LHC/CMS Phase 0 Upgrade and Outlook to the Second Data-taking Period

The LHC and especially its experiments have well performed during the more than three years of the run I period. Most of the machine components have already been used many years before during the testing and commissioning phases. As expected, some components suffer from damages especially cause by radiation. Also some of the components deserve an update with state-of-the art techniques. On top of that, the LHC has not yet been run at its design energy. After the incident in 2008, only a quarter of the LHC ring underwent some renovations due to the needed repairs. Therefore, the main goals of the long shut-down of the LHC and its experiments between the run I and run II periods were the repair and consolidation of the machine and the preparations for larger upgrades foreseen for the phase 1 upgrade.

As the complete LHC ring was warmed up and therefore accessible, all splices between the superconducting magnets have been repaired and additional protection against quenching of magnets has been installed. At the same time, other parts of the accelerator complex have been consolidated and updated. After the longs shut-down 1 it is therefore possible to achieve collision energies of 13 TeV. The aim is to perform collisions at the design energy of 14 TeV starting from 2016. Additionally, it is aimed for bunch spacings of 25 ns instead of the 50 ns of run 1 in order to further increase the instantaneous luminosity. However, first runs at 13 TeV have still been performed with 50 ns bunch spacing.

The period of the shut-down has already been used by all experiments to perform repairs and consolidation works. In the CMS detector, damaged cells in all sub-detectors have been replaced. Especially for the muon chambers, new channels have been added. Additionally, the read-out electronics and the data acquisition software have been updated. The triggers are updated to cope with the higher instantaneous luminosities.

Figure 2.11 (right) shows the cumulative distribution of the integrated luminosity of the 2015 data set recorded until the end of October 2015. Until then, data corresponding to roughly 1.5 fb^{-1} have been certified for the use in physics analyses, whereas more than 3 fb^{-1} of collisions have been delivered. The reason are problems with the cooling of the superconducting magnet of CMS. The safe run of the

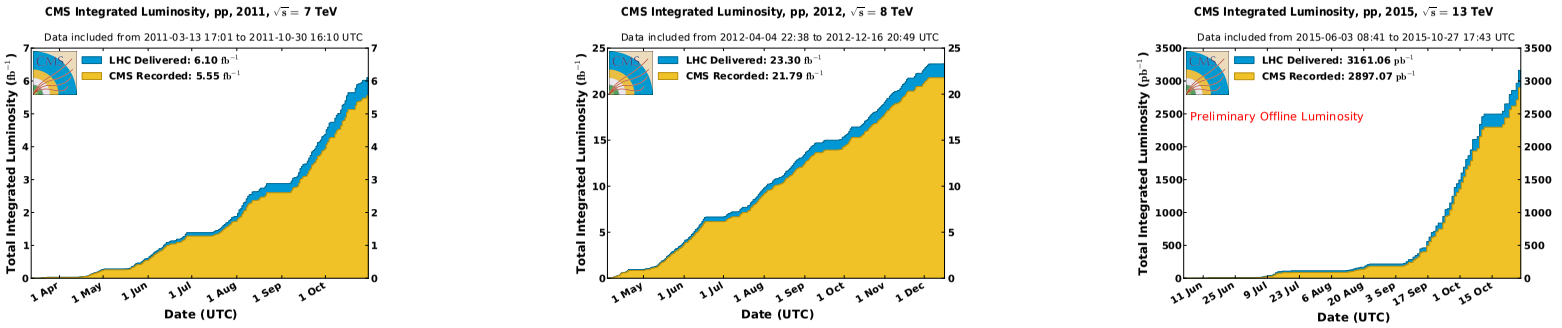


Figure 2.11.: Cumulative distribution of the total integrated luminosities delivered by the LHC and recorded by the CMS detector in the years 2011 (left), 2012 (centre) and 2015 (right) as a function of the date [50]. The recording efficiencies from the first data-taking period are with values above 90 % an indication for well understood and operated systems. The data-taking of 2015 is still ongoing.

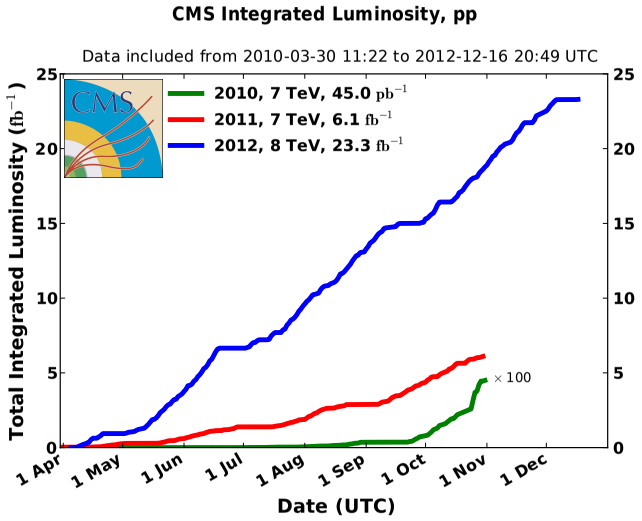


Figure 2.12.: Cumulative distribution of the total integrated luminosities delivered at the CMS detector in the years 2010 to 2012 as a function of the date [50]. The data taken in 2010 is not analysed in the scope of this thesis because the small size of the data set does not justify the additional studies needed. The impact on the precision of the results is negligible.

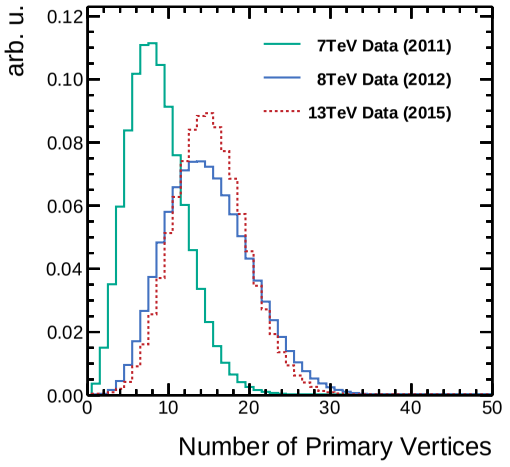


Figure 2.13.: The distributions of the number of reconstructed primary vertices in data taken in the years 2011, 2012 and 2015 are a measure for the amount of pile-up interactions.

magnet could not be guaranteed during the entire running time of the LHC and collisions without magnet and therefore without momentum measurements have been performed that cannot be used for many physics analyses. This issue is currently addressed by experts. The number of pile-up interactions is on average smaller than it was in the 8 TeV data, as shown in figure 2.13. The bunch spacing was reduced to 25 ns, but the bunches are not filled with as many protons as in the 8 TeV runs. It is planned, to fill more protons into the bunches from 2016 on, which then leads to the expectation of much higher number of pile-up interactions that future analyses have to cope with.

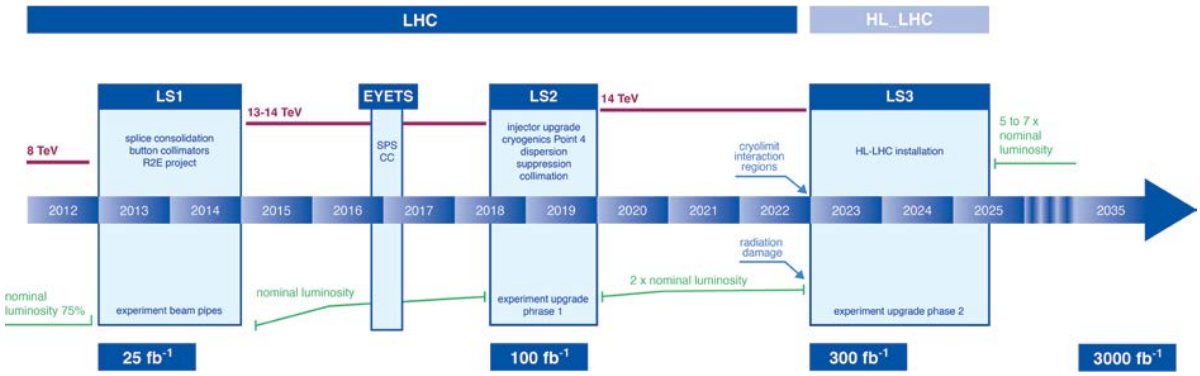


Figure 2.14.: Current plan for the schedule of the LHC [51].

Figure 2.14 provides a brief overview of the currently planned LHC schedule. In the second running period it is aimed for providing roughly 100 fb^{-1} of collision data at centre-of-mass energies of 13 to 14 TeV, before a second long shut-down is planned in order to again increase the instantaneous luminosity by a factor of roughly two and to collect about 300 fb^{-1} of data in the third running period after 2019.

2.4. Event Reconstruction and Particle Identification

The reconstruction of collision events in the CMS detector exploits the Particle Flow (PF) algorithm. A detailed documentation of the algorithm is given in reference [52]. Five different classes of particles are reconstructed and identified based on information from all parts of the detector.

Reconstructed tracks from the inner tracking detector are linked to measurements from the muon chambers to form muons. In case energy deposits in the electromagnetic or hadronic calorimeters are found to be consistent with extrapolated tracks these information is associated with electrons or charged hadrons, respectively. Energy deposits without associated tracks are assigned to photons or neutral hadrons, respectively. For all candidates of particles of these five classes the four-momentum and further properties like their electric charge are determined.

The PF candidates are then taken as input in further event reconstruction algorithms such as the jet clustering and the reconstruction of the missing transverse energy. Th PF algorithm stands for a efficient and optimal event reconstruction since it exploits and combines the information available from all sub-detectors.

2.4.1. Tracks and Vertices

The reconstruction of tracks of charged particles emerging the collisions is one of the most important steps for the analysis. The momenta of charged particles are determined based on the track parameters. The track reconstruction in CMS is performed in several steps [53]. The algorithm is initiated by seeds containing two or three hits of the innermost tracking layers. They provide a first estimate of the track parameters. The tracks are then iteratively extrapolated layer by layer in the tracking detector base on the Kalman filter technique [54]. In the appropriate uniform magnetic field within the coil an helical shape of the tracks is assumed. Ambiguities between the reconstructed trajectories are resolved by applying quality criteria based on the number of hits shared by multiple trajectories.

The resolution of the transverse momenta of the reconstructed tracks depends on the transverse momentum itself. The resolution for low p_{T} tracks of $\mathcal{O}(1-10)$ GeV is between 0.6 and 2 $\%$ depending on the pseudorapidity η, whereas for tracks with p_{T} of $\mathcal{O}(100)$ GeV the resolution increases to 1-2 $\%$ for $|\eta| < 1.6$ and in the barrel, the resolution worsens significantly up to 10 $\%$.

The tracks are then used to reconstruct vertices which is of importance for the identification of decays of long-lived particles as well as for the mitigation of pile-up interactions. Vertices are then reconstructed based on the adaptive vertex fitting technique [55] after all tracks have been clustered and assigned to one possible vertex they originate from. The primary vertex is defined as the vertex with the largest sum of squares of p_{T} of the tracks assigned to this vertex. This vertex is required to lie within the beamspot region to suppress beam-induced backgrounds. All other vertices are assumed to originate from pile-up interactions. Secondary vertices of the decays of long-lived particles such as B hadrons from b decays are fitted separately based on the tracks assigned to these decays.

2.4.2. Electrons

The detector signature of electrons consists of a track in the tracking detector and energy deposited in the ECAL. However, this simple picture has to be expanded because of the fact of usual bremsstrahlung radiation of the electrons. The radiation of bremsstrahlung is supported by the material budget in front of the ECAL as well as by the bending of the electron tracks in the magnetic field. In a second step these photos can again convert into pairs of electrons and positrons.

In total, the energy deposit caused by electrons spreads in the φ-direction of the ECAL. Clusters from single electrons or photons have to be combined to so-called superclusters. Also the track finding and fitting is affected by the energy loss of the electrons. A Gaussian sum filter (GSF) is exploited to handle these peculiarities of electron tracks [56]. The energy of electrons is reconstructed with a resolution below 2 $\%$.

Multivariate discriminators are used to identify electrons and discriminate them from misidentified hadronic showers. These discriminators employ variables describing the compatibility between the track and the ECAL clusters as well as variables characterising the cluster itself.

Details about the electron reconstruction techniques in CMS are documented in reference [57]. Measurements of the performance of the reconstruction in 7 and 8 TeV can be found in the references [58, 59], respectively.

2.4.3. Muons

Muons used for physics analyses as the ones presented in this thesis are reconstructed as so-called global muons. Information from the inner tracking system as well as from the muon chambers outside the magnet coil is taken into account. First the tracks in both systems are fit independently. Then matching tracks are combined in a global track fit which also takes into account the energy loss and multiple Coulomb scattering in the traversed detector material.

The momentum resolution of muons with low transverse momenta is driven by the inner tracking system. The higher the transverse momentum is, the more the muon chambers contribute to the precision of the muon measurement.

The identification and rejection from the small background of charged hadrons can be improved based on requirements on the track fit quality, the number of hits in the two tracking systems and on the longitudinal distance from the beam spot. Details of the reconstruction, identification and the performance can be found in reference [60].

2.4.4. Hadronically Decaying Tau Leptons

The detector signature of hadronically decaying τ leptons is very similar to the one of jets from quarks or gluons (see next section 2.4.5) with the main difference that the multiplicity of the jet constituents is much lower than in typical QCD jets. These hadronically decaying τ leptons are usually referred to as taus, τ_{h}. They are reconstructed based on the hadrons plus strips (HPS) algorithm [61, 62].

The HPS algorithm is seeded by reconstructed jets ($\mathrm{anti}{-}k_t$ with $\Delta R = 0.5$, see next section 2.4.5). Profiting from the reconstruction of single charged and neutral hadrons by the PF algorithm, four different τ decay hypotheses according to table 1.4 can be tested. The HPS algorithm aims to cluster neutral pions in strips in the η-φ plane, where the strips are expanded in the φ-direction. The reason are $\pi^0 \to \gamma\gamma$ decays followed by photon conversions into $\mathrm{e}^+\mathrm{e}^-$ pairs before they reach the ECAL. According to the number of charged hadrons and strips reconstructed within a signal cone the candidate is classified as one of the four decays: three charged hadrons, one charged hadron plus one or two strips and one charged hadrons. In the case that more than one hypothesis matches, the one with the highest p_{T} of the candidates is selected.

QCD jets but also electrons or muons can be misidentified as hadronically decaying τ leptons. The main handle against the misidentified jets is an isolation requirement. Together with other variables the momentum sums for additional charged and neutral hadrons around the τ candidate multivariate discriminators are exploited to suppress the misidentification of QCD jets. Similar discriminators are available for the suppression of misidentified electrons and muons.

2.4.5. Jets

Because of its colour charge quarks and gluons cannot be observed as free particles, but they fragment and hadronise. This leads to a collimated spray of energetic hadrons that is usually called a jet. From the reconstruction of these jets it is tried to deduced parameters like the four-momentum of the initial partons. Jet reconstruction algorithms group particles into jets and defined the calculation of the parton parameters [63].

Most of the CMS analyses exploit a sequential recombination algorithm called the $\mathrm{anti}{-}k_t$ algorithm [64]. The algorithm is known to be collinear- and infra-red-safe, meaning the number of jets is not affected by soft collinear gluon emission or parton splitting. It is implemented within the FastJet package [65] interfaced to the CMS reconstruction software. The analyses presented in this thesis use jets reconstructed by the $\mathrm{anti}{-}k_t$ algorithm with a distance parameter of $\Delta R = 0.5$.

In general, the measured four-momentum of a reconstructed jet does not agree with the momentum on parton or hadron level. Corrections are applied in a multiplicative way to the four-momentum vectors to account for different experimental effects [66]. First, an offset term is supposed to remove energy resulting from processes apart from the hard scattering such as from detector noise or from pile-up. Then a second term only for simulated events is assigned to correct the energy of the reconstructed jets on average to the generated ones. Finally, residual corrections are applied to account for differences between the simulation and the measured data. The performance of the jet energy calibration and the related uncertainties can be found in reference [67].

Jets originating from b quarks can be distinguished from jets initiated by a charm or a light quark or a gluon by exploiting lifetime variables of the resulting B mesons. Secondary vertices (CVs) can be reconstructed based on fits of well reconstructed tracks of jet constituents. The impact parameter with respect to the primary vertex (PV) is combined together with other variables into a combined secondary vertex discriminator (CSV). Working points are defined yielding specific selection efficiencies and fake rates. Details and performance information can be found in reference [68].

2.4.6. Missing Transverse Energy

Missing transverse energy (MET) as the negative vectorial sum of all measured momenta in the transverse plane quantifies the momentum carried away by undetected particles such as neutrinos. This variable is very sensitive to experimental effects and mismeasurements such as contributions from pile-up and underlying event collisions, jet energy resolution, detector noise and the finite detector acceptance.

The response and the resolution of the simple PF-based definition of the MET show a quite strong dependence on the number of primary vertices in the event and therefore to the amount of pile-up interactions. Multivariate techniques help to mitigate this effect. The method is documented in [69,70].

The MET is calibrated in $Z \to \mu\mu$ events. The transverse momentum of the Z boson, usually denoted as $\vec{q}_T$, is well reconstructed and balanced against a hadronic recoil, $\vec{u}_T$.

$$\vec{q}_T + \vec{u}_T + \vec{\not{E}}_T = 0$$

Corrections to the hadronic recoil are applied in terms of two BDT regressions: the first for its azimuthal angle and the second for its magnitude [70].

2.5. Simulation and Software

2.5.1. Monte Carlo Event Generators

One important task of experimentalists is to compare predictions described by a theory or model with the outcome of their measurements. Theorists provide Monte Carlo event generators to simulate the outcome of given experiments.

PYTHIA is a general-purpose event generator used for hadron and lepton collisions [71]. PYTHIA simulates the whole scattering of particles with a partonic substructure. Therefore, the results a comparable with the ones of a hypothetical ideal detector.

Starting with particles having a partonic substructure described by parton distribution functions, PYTHIA describes the gluon-radiation of colour-charged objects in the initial state by a parton shower model. After that, two partons participate in the hard interaction process. The calculations base on a library containing descriptions for various $2 \to 1/2/3$ processes at next-to-leading order (NLO). Final state radiation is considered as well as the hadronisation of colour charged objects in the final state. This is done based on the so-called Lund string model [72]. In addition to that, remnants of the initial hadrons not participating in the hard process are described by the underlying event model of PYTHIA. Even multiple partonic interactions are possible. Since details aspects of certain processes may be treated better by other packages, PYTHIA allows to replace some calculations with plug-ins.

MADGRAPH is a general-purpose leading order (LO) matrix-element based event generator [73]. In contrast to PYTHIA, only the hard process is calculated. Interfaces to event generators such as PYTHIA are available. MADGRAPH has been extended to the aMC@NLO package [74], which contains also one-loop matrix elements for Standard Model processes.

For given sets initial and final state particles, MADGRAPH computes all possible feynman diagrams at leading order and generates code to calculate the matrix elements. MADGRAPH is able to deal with every possible renormalisable or effective $2 \to n$ theory that is based on a Lagrangian. The included tool MADEVENT can then be used to generate events.

POWHEG is, similar to MADGRAPH and event generator that can be interfaced to parton showers such as the one integrated in PYTHIA [75]. The events are calculated at next-to-leading order.

TAUOLA is a package for the simulation of τ decays [76]. It is specialised on handling spin and polarisation effects of the decay products of the τ leptons and can be interfaced to the other event generators.

2.5.2. Detector Simulation

After generating pure events originating from proton-proton-collisions, the detector response has to be simulated in order to yield outputs that are compatible with real obtained data. Most physics analyses in CMS use the full simulation method[8].

The detector simulation uses as input the so-called HepMC [77] files that are produced by event generators. The whole detector geometry is modelled in detail by the GEANT 4 simulation tool-kit [78]. This enables the simulation of traversing particles through the detector material and its interactions. Besides this physical simulation of the detector behaviour, the tool-kit also covers the digitising step, where the read-out electronics is simulated. Its output is then passed through the same reconstruction software as real data.

2.5.3. Software Frameworks

Large amounts of data require sophisticated tools to perform the data analysis. The analysis software development is an important issue for experimenters. This section describes the common frameworks in general. A detailed overview of the actual work-flow for the analyses presented in this thesis is given in the appendix A.

ROOT is understood as the most fundamental framework for all event-based data analyses in the high energy physics environment [79]. This object oriented framework provides a great variety of classes that cover almost every step of the analysis procedure from event generation and detector simulation over event reconstruction and data acquisition to the data analysis. Data structures such as n-tuples and trees which allow access to the data event by event are used for final analyses. Also histograms with various functionalities including fitting routines as well as visualisation capabilities are part of the most frequently used classes.

TMVA is a tool-kit for multivariate data analysis with ROOT [80]. It contains a large variety of machine learning algorithms for the classification of events and the estimation of quantities. It implements the training, testing, performance evaluation and application of all these methods.

For the analyses presented in this thesis the commonly used boosted decision trees (BDTs) are of special importance, although this method has also been compared to other algorithms like simple likelihood ratios or neural networks. Strictly-speaking, a boosted decision tree is build from a forest of simple decision trees. Each node of the tree makes a decision based on one input variable whether the event belongs to one or the other class of events the BDT is trained to discriminate

[8] There is also a fast simulation method which is used for very large scale productions of simulated samples because it is faster by about three orders of magnitude compared to the full simulation method. It is based on a simplified detector model and observables are further smeared according to resolutions measured with the real detector.

between. The depth of the tree, i.e. the number of layers of nodes, is a parameter of the decision tree. Usually, several hundreds of small trees are combined with three to five layers of nodes. The algorithm optimises the input variables the single trees is based on with a boosting strategy. Events that are mis-classified by the existing trees, are reweighted to gain importance in the training of the next tree. Finally, the decision of the entire BDT is taken as the average of all decisions of the single trees. Therefore the output is usually a continuous quantity. Smaller values signalise that events belong to the first and larger values that they belong to the second class of events.

CMSSW is a software framework provided by the CMS collaboration [46, 81]. This software based on the ROOT framework covers all parts of the analysis from data taking or generation to final studies on small sets of selected events.

CMSSW defines data formats for all CMS data and simulation by a so-called event data model (EDM). There exist different formats for different levels of details. The RAW or FEVT data contains the full event content from the detector output (or of the detector simulation in case of simulated events). The format is reduced by the reconstruction step in CMSSW to the RECO or the analysis object data (AOD) format. The latter one is usually used for physics analysis. In 2015, the levels of details here have further been reduced by the introduction of a so-called miniAOD format in order to speed up the processing time. In a subsequent skimming step the AOD format is usually reduced to a analysis-specific n-tuple format containing only the very needed information for the special analysis. This data allow then for fast turn-around-cycles of few hours for running the entire analysis. This skimming step uses the official analysis modules provided by CMSSW to process the AOD format, whereas the output of this stage is often independent of CMSSW.

Combine is a tool developed by the ATLAS and CMS collaborations in the context of the LHC Higgs Combination Group [82]. Various statistical methods are available to compare and quantity the compatibility of data with different models for signal and background. A detailed description of the methods used in the scope of this thesis are given in section 3.8.

3

The Analysis of Standard Model Higgs Bosons Decaying into τ Leptons in the Di-muon Final State

The search for the Standard Model Higgs boson in $\tau\tau$ final states is subdivided into several channels. Six main channels are distinguished. They cover fully hadronic ($\tau_{\mathrm{h}}\tau_{\mathrm{h}}$), semi-leptonic ($\mu\tau_{\mathrm{h}}$ and $\mathrm{e}\tau_{\mathrm{h}}$) and fully leptonic ($\mathrm{e}\mu$, $\mu\mu$, ee) decays of τ pairs. In this chapter, the analysis of the di-muon channel is presented in detail, before the next chapter 4 gives and overview of the complete $\mathrm{H} \to \tau\tau$ analysis in CMS as published in reference [10] and puts the $\mathrm{H} \to \tau\tau \to \mu\mu$ analysis in its context.

The di-muon analysis is restricted to events were both τ leptons decay leptonically into a muon and two neutrinos. In total, four neutrinos in the final state lead to substantial amount of missing transverse energy and consequently to a degraded mass resolution. The overwhelming background from $\mathrm{Z} \to \mu\mu$ events and the low branching ratio $\mathcal{BR}(\tau\tau \to \mu\mu)$ of approximately 3 $\%$ complicate this analysis and distinguish it from the analyses in more significant channels like the semi-leptonic one (because of the better mass resolution and the higher branching ratio) and the $\mathrm{e}\mu$ channel (because of the strongly suppressed background from Drell Yan events). Because of similar challenges, the analyses of both the di-muon and the di-electron channels share a common analysis strategy.

The analysis presented in this thesis is based on the full data set taken at the CMS experiment in the years 2011 and 2012 corresponding to 4.9 fb^{-1} and 19.7 fb^{-1}, respectively, at centre-of-mass energies of 7 TeV and 8 TeV, respectively. It extends and improves an earlier version based on a smaller data set presented in reference [83]. This chapter focusses on the improvements introduced in the published analysis with respect to the preliminary analysis based on the full data set as documented in reference [84]. For the details of the analysis in the di-electron analysis it is referred to reference [85].

3.1. Signal Signature and Background Processes

The detector signature of the $H \to \tau\tau$ signal events in the di-muon channel are characterised by two oppositely charged muons. The accompanying four neutrinos escape direct detection.

$$pp \to H + X \to \tau^+\tau^- + X \to \mu^+\mu^- \nu\nu\nu\nu + X$$

The signal signature in the detector is characterised by two muons with large transverse momenta. The four neutrinos in the final state lead to substantial amount of missing transverse energy. The analysis is optimised for the signal production via gluon fusion and vector boson fusion (see section 1.2.2). However, the signal produced in association with vector bosons or top quarks is also taken into account.

Standard model processes with di-muon final states have to be considered as sources for background events. The by far most dominant source of background events are leptonic Z decays that can be distinguished into $Z \to \mu\mu$ and $Z \to \tau\tau \to \mu\mu$ events. Minor contaminations arise from $t\bar{t}$ events, di-boson decays (WW, WZ, ZZ) as well as QCD and W+jets processes.

3.1.1. Data Sets and Simulation

The entire data set recorded at the CMS experiment in the years 2011 and 2012 has been analysed. Depending on the available triggers for the sub-analyses this corresponds to an integrated luminosity of 4.9 fb^{-1} at the centre-of-mass energy of 7 TeV and 19.7 fb^{-1} at 8 TeV, respectively, in the main channels including the di-muon analysis. The events are selected based on dedicated trigger algorithms for pairs of two leptons, depending on the analysis channel.

The signal samples for the gluon fusion and the vector boson fusion processes are simulated using the POWHEG event generator in version 1.0 [75] whereas the production in association with vector bosons and top quarks is simulated by PYTHIA 6.4 [71]. The POWHEG samples are generated at next-to-leading order (NLO) precision and the PYTHIA sample at leading order (LO) precision. Correction factors from NNLO calculations [86] are applied to the Higgs boson p_T spectrum of the gluon fusion sample.

The Z+jets, W+jets, $t\bar{t}$+jets and di-boson background samples are generated with MADGRAPH 5.1 [73] at LO precision. In order to increase the statistical precision in signal-sensitive selections, the Z+jets and W+jets samples are complemented by simulations of samples with fixed jet multiplicities from one to four. They are weighted such to preserve the total cross sections.

The POWHEG and MADGRAPH generators are interfaced with PYTHIA for modelling the parton shower and fragmentation. PYTHIA is also used for the simulation of additional proton-proton collisions that are then added to the simulation of the hard interaction in order to account for pile-up interactions. All τ lepton decays are modelled with TAUOLA [76]. Finally, the CMS detector is modelled with GEANT 4 [78]. For an overview of the simulation software it is referred to section 2.5.

Table 3.1 summarises all simulated samples used in the analysis presented here and lists the corresponding cross sections. The simulated events are normalised to NNLO cross sections. For the signal, Higgs boson mass hypotheses from 90 to 145 GeV have been studied.

Table 3.1.: Summary of the simulated samples and their cross sections [21–23, 87, 88] used in the SM $H \to \tau\tau$ analysis. The signal cross sections are given for a Higgs boson mass hypothesis of $m_H = 125$ GeV.

Process	**Generator**	**Cross Section / pb**	
		$\sqrt{s} = 7$ TeV	8 TeV
$(W \to \ell\nu) + \text{jets}$	MADGRAPH	31314	36257
$(Z/\gamma^* \to \ell\ell) + \text{jets}$	MADGRAPH	3048	3504
$t\bar{t}$ +jets	MADGRAPH	158	225
$(WW \to \ell\nu\ell\nu) + \text{jets}$	MADGRAPH	4.78	5.818
$(WZ \to qq'\ell\ell) + \text{jets}$	MADGRAPH	1.79	2.268
$(WZ \to \ell\nu\ell\ell) + \text{jets}$	MADGRAPH	0.857	1.093
$(ZZ \to \ell\ell q\bar{q}) + \text{jets}$	MADGRAPH	0.777	2.493
$(ZZ \to \ell\ell\nu\nu) + \text{jets}$	MADGRAPH	0.251	0.713
$(ZZ \to \ell\ell\ell\ell) + \text{jets}$	MADGRAPH	0.0642	0.18
SM $gg(H \to \tau\tau)$	POWHEG	0.96	1.22
SM $qq(H \to \tau\tau)$	POWHEG	0.077	0.010
SM $Z(H \to \tau\tau)$+$W(H \to \tau\tau)$+$t\bar{t}(H \to \tau\tau)$	PYTHIA	0.063	0.079
SM $gg(H \to WW)$	POWHEG	0.34	0.43
SM $qq(H \to WW)$	POWHEG	0.028	0.036

3.2. Event Selection and Categorisation

Events that are considered for this analysis must have been selected by a double-muon high level trigger. The reconstructed muons have to match with the objects that fired the trigger. The trigger thresholds on the transverse momentum are 13 (7) GeV for the leading (trailing) muon. The leading muon is the one with larger transverse momentum. The trigger efficiencies have been measured for the preliminary analysis using the same data set via tag-and-probe methods [83]. Correction factors are applied to the simulation in order to match the trigger efficiency as a function of the transverse momentum to the one of the data.

The two muons have to be reconstructed by the Particle Flow algorithm [52] as global muons passing the tight muon identification definitions as summarised in section 2.4.3. Furthermore, they are required to be oppositely charged and pass the following kinematic requirements

$$
\begin{aligned}
&p_T(\mu_1) > 20 \text{ GeV} \quad \text{and} \quad |\eta(\mu_1)| < 2.1 \\
&p_T(\mu_2) > 10 \text{ GeV} \quad \text{and} \quad |\eta(\mu_2)| < 2.4 \\
&m(\mu\mu) > 35 \text{ GeV}
\end{aligned}
$$

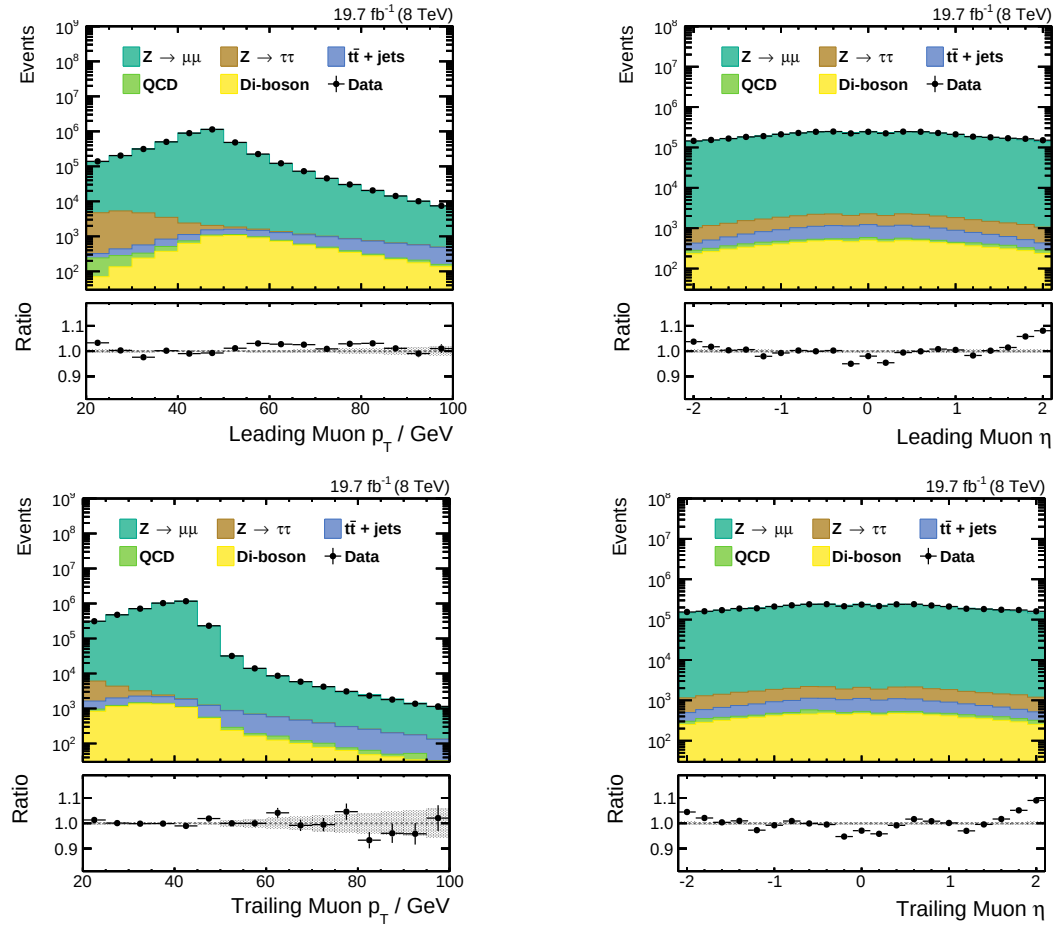


Figure 3.1.: Transverse momentum (left) and pseudorapidity (right) of the leading (top) and the trailing (bottom) muon for the 8 TeV analysis. The ratio compares the observation in data to the expectation given by the sum of all backgrounds. The statistical errors are shown in the error band. The discrepancies are due to detector effects (see figure 3.2). The effect is well covered by the corrections applied to the Z $\to \mu\mu$ background and its systematic uncertainties (see section 3.6.1).

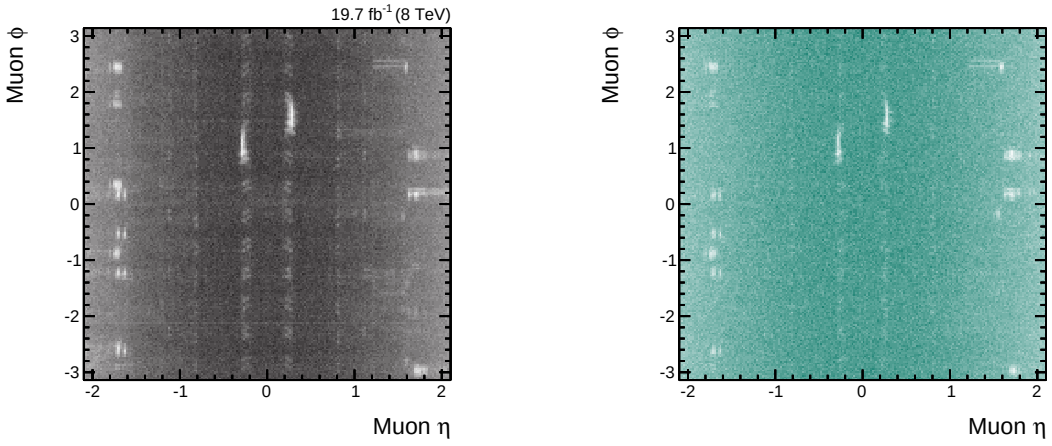


Figure 3.2.: Event distribution of data (left) and in the simulation of Z $\to \mu\mu$ events (right) in the η-φ plane of the reconstructed muons (based on a very loose selection) in 8 TeV analysis. Clear differences in the modelling of the chimneys for the magnet cryogenic lines in the central region and of the reconstruction efficiencies of the muon modules in the end-caps lead to the discrepancies observed in figure 3.1

where μ_1 denotes the leading muon and μ_2 the trailing muon. The corresponding control plots are shown in figure 3.1 for the leading the trailing muon of the 8 TeV analysis. The ratio comparing the observation in data to the expectation given by the sum of all backgrounds reveals significant discrepancies with respect to the statistical uncertainties. They can be explained by detector effects that are illustrated in figure 3.2, where the modelling of the muon reconstruction in the simulation of $\mathrm{Z} \to \mu\mu$ events (right) is compared to the observation measured in the detector (left). It is visible that both the chimneys for the magnet cryogenic lines in the central parts around $0.2 < |\eta| 0.25$ and some muon chambers in the end-caps are not well modelled by the simulation[1]. The residual effect on the final test statistic is reduced by a dedicated background estimation method for the $\mathrm{Z} \to \mu\mu$ background and systematic uncertainties are assigned that cover these discrepancies (see section 3.6.1).

Additionally, the two muons are required to be isolated from other activity in the detector in order to reduce the background of from QCD multi-jet production. The isolation variable, I, is defined as the activity within a cone in the η-φ-space around the muon flight direction originating from the same primary vertex excluding the muon itself. The cone size is defined by a upper cut on ΔR which is chosen to be 0.4.

$$I = \sum_{\substack{\text{charged,}\\ \text{non-pile-up}}} p_{\mathrm{T}} + \max\left(0, \sum_{\text{neutral}} p_{\mathrm{T}} - \beta \sum_{\substack{\text{charged,}\\ \text{pile-up}}} p_{\mathrm{T}}\right) \tag{3.1}$$

Non-pile-up activity from charged particles is selected by requiring that their tracks originate from the primary vertex. For neutral particles there is no such possibility because of the absence of track measurements. Therefore the contribution from all neutral particles is corrected by subtracting a term proportional to the charged contribution from pile-up vertices. The proportionality factor $\beta = \frac{1}{2}$ is evaluated based on the simulation and roughly agrees with the overall fraction of neutral to charged activity in the calorimeters. The relative isolation $I_{\mathrm{rel}} = I/p_{\mathrm{T}}$ of a muon with the transverse momentum, p_{T}, is required to undermatch values of 0.1.

Jets are reconstructed according to the documentation in section 2.4.5 and are required to pass the following kinematic cuts and have to be separated from the muons in the η-φ plane.

$$p_{\mathrm{T}}(\text{jet}) > 30\ \text{GeV} \quad \text{and} \quad \left|\eta(\text{jet})\right| < 4.7 \quad \text{and} \quad \Delta R(\text{jet}, \mu) > 0.5$$

The corresponding control plots are shown in figure 3.3 for the leading and the sub-leading jet as well as for the number of jets per event in figure 3.4 for the 8 TeV analysis.

Events containing jets tagged as originating from b quarks are vetoed in order to suppress the background contributions arising from $\mathrm{t\bar{t}}$+jets events. For b-tagged jets a medium identification is applied.

$$p_{\mathrm{T}}(\mathrm{b\ jet}) > 20\ \text{GeV} \quad \text{and} \quad \left|\eta(\mathrm{b\ jet})\right| < 2.4 \quad \text{and} \quad \mathrm{CSVM(b\ jet)} > 0.679$$

According to the summary in section 2.4.5, CSVM is defined as the b-tagging discriminator that yields a medium misidentification probability for light parton jets of 1 % [68].

[1] More recent simulations are available that correct for these effects. In order to keep the consistency with all $\mathrm{H} \to \tau\tau$ analysis, these samples have not been used here.

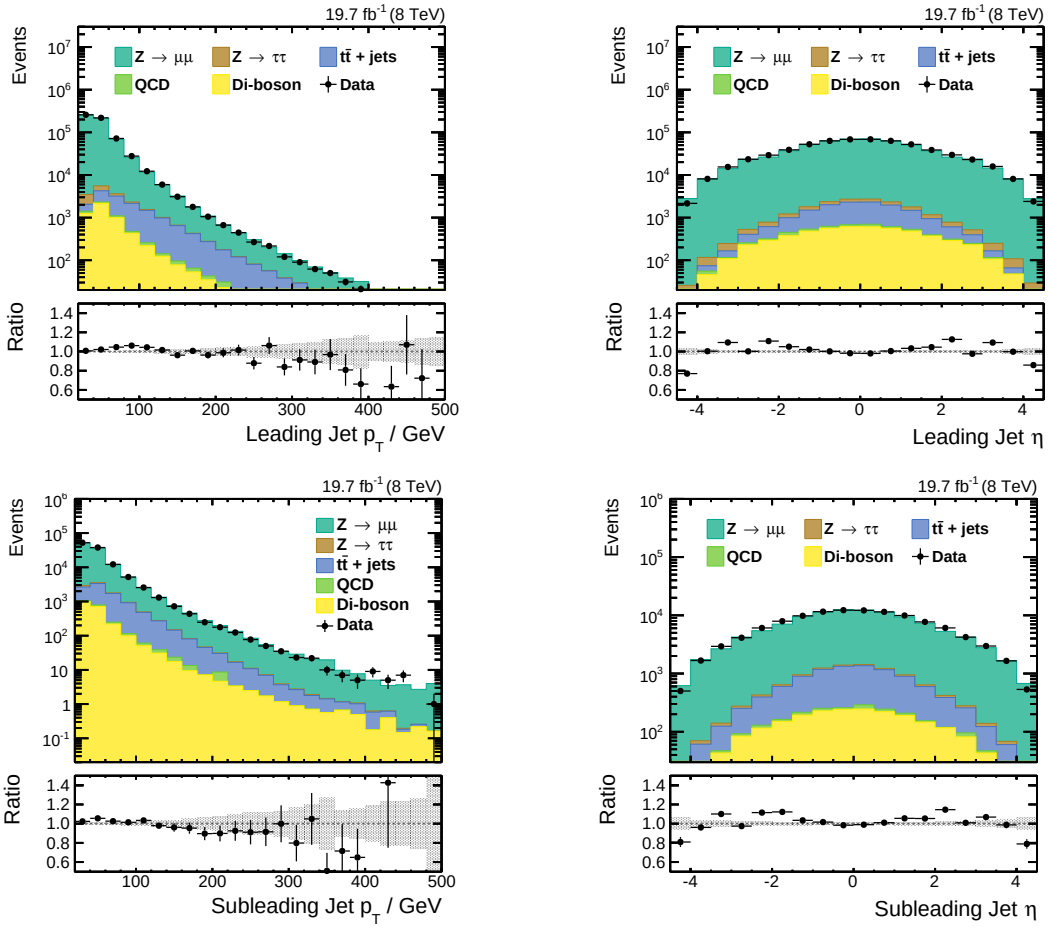


Figure 3.3.: Transverse momentum (left) and pseudorapidity (right) of the leading (top) and the sub-leading (bottom) jet for the 8 TeV analysis The ratio compares the observation in data to the expectation given by the sum of all backgrounds. The statistical errors are shown in the error band. The agreement between data and the sum of all backgrounds is reasonably good, at least in the distributions of the transverse momenta of the jets. The remaining discrepancies are well covered by systematic uncertainties.

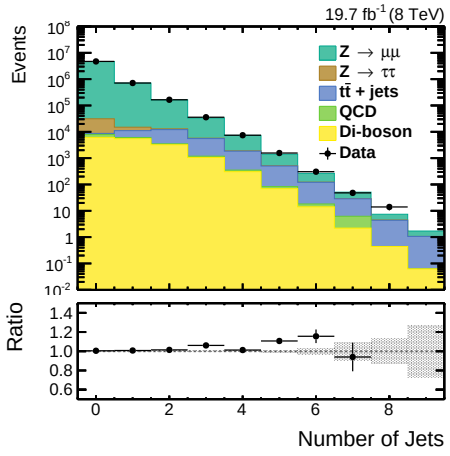


Figure 3.4.: Number of jets per event for the 8 TeV analysis. The ratio compares the observation in data to the expectation given by the sum of all backgrounds. The statistical errors are shown in the error band. There is a good agreement between data and the sum of all backgrounds up to the two jets that are considered in this analysis.

The missing transverse energy (MET) is reconstructed in a multivariate fashion (see section 2.4.6). As in the other H $\to \tau\tau$ channels the simulated MET needs to be corrected to match the distribution of the data. This is an important step since the MET is one of the most crucial variables in the analysis. In contrast to the other channels, the same-flavour di-lepton channels do not apply correction factors as weights but perform an event-by-event mapping of the simulated events based on the cumulative distribution functions of the two components of the MET. This calibration follows the same technique as the calibration of the DCA variable as documented in section 3.4. Details are found in reference [83]. This procedure has the advantage that correlations between the MET and other variables are preserved. Figure 3.5 displays an 8 TeV control plot for the calibrated MET.

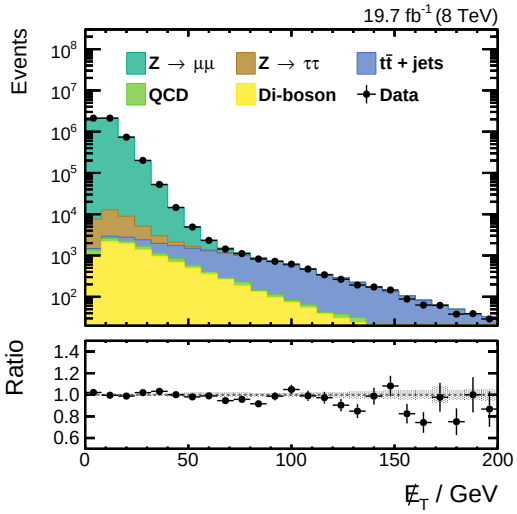


Figure 3.5.: Calibrated missing transverse momentum for the 8 TeV analysis. The ratio compares the observation in data to the expectation given by the sum of all backgrounds. The statistical errors are shown in the error band. The agreement in the low MET region which is dominated by Z and also H is acceptably good. Systematic uncertainties are assigned to cover the remaining discrepancies.

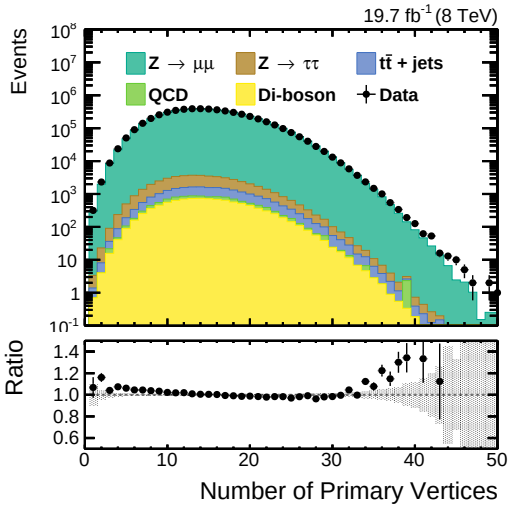


Figure 3.6.: Number of primary vertices after pile-up reweighting for the 8 TeV analysis. The ratio compares the observation in data to the expectation given by the sum of all backgrounds. The statistical errors are shown in the error band. The simulation is well modelled in the central region of the distribution.

Reconstructed events are required to have at least one well reconstructed primary vertex (PV). The distribution of the number of PVs is shown in figure 3.6 after the simulation has been reweighted to match the one in data.

The pre-selected events are divided into five different event categories in order to exploit different signal production mechanisms and to enhance the signal to background ratio in dedicated signal categories. Firstly, events are categorised by their jet multiplicity as in all other sub-analyses. The details of the event categorisation are introduced in the following. Table 3.2 lists the event yields for data, the backgrounds and the signal in all five categories for the 8 TeV analysis. Respectively for the 7 TeV analysis the values are found in table B.1.

The 2-jet category is optimised to explore the VBF signal production mechanism. The two jets with the highest transverse momentum are assumed to originate from the two quarks characterising the VBF production mechanism. There must not be any further high-p_T jet in the gap in the pseudorapidity, $\Delta\eta_{jj}$, between these two jets. Vetoing significant hadronic activity in the central part of the detector reduces the SM background contributions considerably, making this category the most sensitive one. There is no additional cut on quantities defining the di-jet system like in the other $H \to \tau\tau$ channels or the preliminary analysis of this channel because these variables are exploited in the subsequent multivariate analysis. For this reason a significant fraction of signal events produced via the gluon fusion process is selected in this category. The notable theory uncertainties related to the modelling of the two additional jets are mentioned in section 3.7.

Events which do not fall into the 2-jet category and that contain at least one jet are considered for the 1-jet categories. The remaining events are assigned to the 0-jet categories. The lower the jet multiplicity in the event is the larger is the contamination with background events. Especially the 0-jet categories are well suited for constraining the uncertainties related to detector effects, the events reconstruction and the modelling of the backgrounds that are correlated among all categories.

The 1-jet and 0-jet categories are then further subdivided by a cut on the transverse momentum of the leading muon. A cut of $p_T(\mu_1) = 35$ GeV is taken in analogy to the other leptonic channels, ee and eμ. Due to the higher Higgs boson mass hypothesis compared to the Z boson mass the spectrum of the transverse momentum of signal events is expected the be harder than the one of the background. This is the reason for the more pronounced sensitivity to the signal of the high-pt categories.

Table 3.2.: Event yields after the pre-selection and the categorisation for data, the background processes and the signal for the 8 TeV analysis. The number of signal events are given for a Higgs boson mass hypothesis of 125 GeV. It is clearly visible that the $Z \to \mu\mu$ background deserves the most crucial treatment since it is orders of magnitude larger than the other backgrounds. It is also noticeable that the signal events are to more enriched with qqH events the higher the jet multiplicity is. The corresponding numbers for the 7 TeV analysis are documented in table B.1

Process	**0-jet**		**1-jet**		**2-jet**
	low-pt	**high-pt**	**low-pt**	**high-pt**	
Data	**873709**	**3776365**	**40606**	**646549**	**164469**
$Z \to \mu\mu$	850731	3763980	38522	634395	157571
$Z \to \tau\tau$	17797	5443	1887	1893	860
$t\bar{t}$+jets	16	323	144	2002	2428
WW + WZ + ZZ	400	4033	238	5140	3511
QCD + W+jets	616	698	130	502	254
Sum Background	**869559**	**3774476**	**40922**	**643931**	**164623**
ggH	25.30	30.53	6.98	15.07	6.29
qqH	0.34	0.43	1.00	2.22	3.92
VH + $t\bar{t}$H	0.64	3.52	0.57	4.36	2.60
Sum Signal	**26.27**	**34.47**	**8.55**	**21.65**	**12.82**
$S/\sqrt{S+B}$	0.028	0.018	0.042	0.027	0.032

In general, the categories are not defined as tight as in the channels involving hadronically decaying τ leptons. One reason is the comparatively small branching ratio $\mathcal{BR}(\tau\tau \to \mu\mu)$ of approximately 3 %. The small number of signal events remaining after the pre-selection complicates tight selections. Secondly, the search for the signal in the same-flavour di-lepton channels is performed based on a multivariate discriminator (see section 3.5). Here the separation power of several variables is combined into one single discriminator, whereas the other channels follow the strategy of cutting on certain variables and using one remaining variable (the invariant mass of the di-τ system) as a final discriminator. The robustness of multivariate discriminators increases with the number of training events. This is another reason, why the categories are defined less exclusively.

The table 3.2 shows the background composition in the five event categories. It is obvious that resonant Z decays deserve the most thorough treatment. $Z \to \mu\mu$ decays constitute by far the most dominant background. The numbers of $Z \to \mu\mu$ events are larger then the ones sum of all other backgrounds by one to three orders of magnitude. Minor contributions arise from non-resonant processes like $t\bar{t}$+jets events and resonant di-boson decays. The QCD multi-jet background is already strongly reduced by requiring isolated muons.

The next section 3.3 concentrates on the mass reconstruction in the $H \to \tau\tau \to \mu\mu$ channel and section 3.4 introduces the distance of closest approach of the two muon tracks as one of the most crucial variables in this analysis before the actual signal extraction via multivariate methods is described in section 3.5. The focus of the explanations lies exemplary on the 1-jet high-pt category. All other categories are analysed similarly. The 1-jet high-pt category is best suited for illustration purposes due to its fairly high sensitivity paired with sufficiently large numbers of events in all plots. The main differences in the 2-jet category are mentioned as well.

3.3. Di-τ Pair Mass Definitions

The reconstruction of the invariant di-τ pair mass is complicated by the two to four neutrinos in the final state that escape direct measurement in the detector. The analyses of the same-flavour di-lepton channels exploit three different mass definitions. These mass definitions are important variables to discriminate between signal and background events. However, the special composition of the backgrounds in the analysis leads to the need of more than one definition for the mass to efficiently perform this separation.

3.3.1. Visible Mass

The mass of the visible τ decay products, here the di-muon mass, $m_{\mu\mu}$, neglects the four neutrinos. Therefore it underestimates the true mass of the resonance and its diluted resolution decreases the separation power for the discrimination between signal and Z+jets events. Its distribution in the 1-jet high-pt category for the 8 TeV analysis is shown in figure 3.7 (left). The corresponding shapes of the distributions for the two most dominant background processes, $Z \to \mu\mu$ and $Z \to \tau\tau$, are compared with the one for the signal with a Higgs boson mass hypothesis of 125 GeV in figure 3.8 (left). Events originating from $Z \to \mu\mu$ events result in a sharp Z mass, whereas $Z \to \tau\tau$ events yield smaller di-muon

mass values due to the neglected momentum from the neutrinos. H $\to \tau\tau$ events yield slightly higher di-muon mass values on average, depending on the Higgs boson mass hypothesis. The shape of the di-muon mass of Z $\to \mu\mu$ events suggests a mass window cut around the Z mass. Taking into account the absolute number of events it is comprehensible that such a cut still leaves a large fraction of Z $\to \mu\mu$ events in the tails of the Z mass peak.

3.3.2. Reconstructed Invariant Mass of the Di-τ Pair

The missing transverse energy gives information about the energy carried away by the sum of all neutrinos in the transversal plane. The likelihood-based method summarised in general below estimates the invariant τ pair mass as a function of the measured four-momenta of the visible decay products of the two τ leptons, p_1 and p_2, and the two components of the MET, $\vec{\not{E}}_{\mathrm{T}}$ [10].

The full τ lepton four-momentum can be parametrised as a function of the measured four-momenta of the visible decay products, p_ℓ for leptonic and $p_{\tau_{\mathrm{h}}}$ for hadronic decays, the fraction of the τ lepton energy in the laboratory frame carried by the visible decay products, x, the azimuthal angle of the τ lepton direction in the laboratory frame, φ, and the mass of the two-neutrino system, $m_{\nu\nu}$ in case of leptonic τ decays. The set of unknown parameters is denoted by $\vec{a}$.

$$p_\tau^\ell = p_\ell + p_{\nu\nu} = p_\tau^\ell\left(p_\ell, x, \varphi, m_{\nu\nu}\right) \equiv p_\tau^\ell\left(p_\ell, \vec{a}_l\right)$$
$$p_\tau^{\mathrm{h}} = p_{\tau_{\mathrm{h}}} + p_\nu = p_\tau^{\mathrm{h}}\left(p_{\tau_{\mathrm{h}}}, x, \varphi\right) \equiv p_\tau^{\mathrm{h}}\left(p_{\tau_{\mathrm{h}}}, \vec{a}_h\right)$$

For leptonic decays, the physically meaningful values for x range from zero to one and $m_{\nu\nu}$ can have values up to $m_\tau\sqrt{1-x}$. For hadronic decays, x is further bound from below by the value $m_{\mathrm{vis}}^2/m_\tau^2$, where m_{vis} denotes the invariant mass of the reconstructed τ jet.

Then the mass of the τ lepton pair can be determined as a function of the two visible momenta and the two sets of unknown parameters.

$$m_{\tau\tau} = m_{\tau\tau}\left(p_1, p_2, \vec{a}_1, \vec{a}_2\right)$$

A likelihood model, $\mathcal{L}$, is constructed in order to estimated the most probable values of the parameters $\vec{a}_1$ and $\vec{a}_2$. The likelihood function is a product of three terms.

$$\mathcal{L}\left(p_1, p_2, \vec{\not{E}}_{\mathrm{T}}, \vec{a}_1, \vec{a}_2\right) = \mathcal{L}_{\tau_{\mathrm{l}}}\left(p_1, \vec{a}_1\right) \cdot \mathcal{L}_{\tau_{\mathrm{l}}}\left(p_2, \vec{a}_2\right) \cdot \mathcal{L}_{\mathrm{miss}}\left(\vec{\not{E}}_{\mathrm{T}}, \vec{a}_1, \vec{a}_2\right)$$

The first two, $\mathcal{L}_{\tau_{\mathrm{l}}}$ and $\mathcal{L}_{\tau_{\mathrm{l}}}$, model the kinematics to the τ leptons based on matrix elements for unpolarised leptonic ($\mathcal{L}_{\tau_{\mathrm{l}}}$) and hadronic ($\mathcal{L}_{\tau_{\mathrm{h}}}$) decays.

$$\mathcal{L}_{\tau_{\mathrm{l}}}\left(p_\ell, x, \varphi, m_{\nu\nu}\right) = \frac{\mathrm{d}\Gamma}{\mathrm{d}x\,\mathrm{d}\varphi\,\mathrm{d}m_{\nu\nu}} \propto \frac{m_{\nu\nu}}{4\,m_\tau^2}\left[\left(m_\tau^2 + 2\,m_{\nu\nu}^2\right)\left(m_\tau^2 - m_{\nu\nu}^2\right)\right]$$
$$\mathcal{L}_{\tau_{\mathrm{h}}}\left(p_{\tau_{\mathrm{h}}}, x, \varphi\right) = \frac{\mathrm{d}\Gamma}{\mathrm{d}x\,\mathrm{d}\varphi} \propto \frac{1}{1 - m_{\mathrm{vis}}^2/m_\tau^2}$$

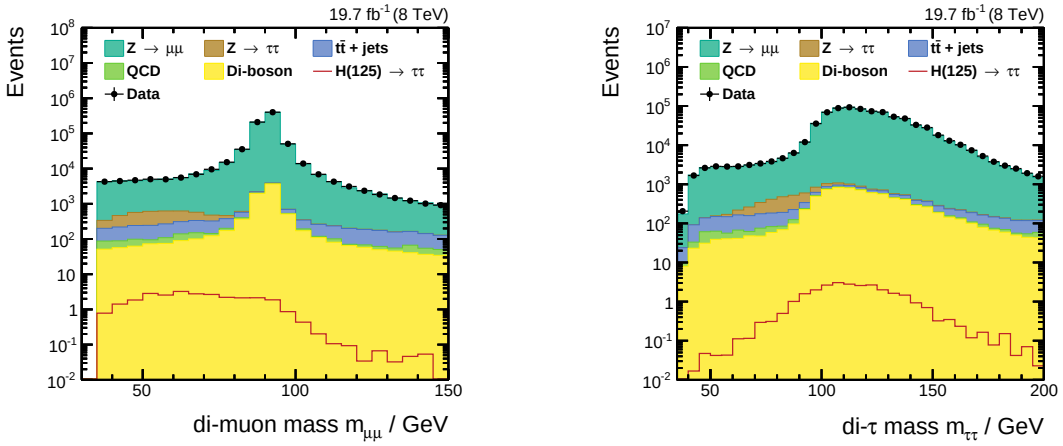


Figure 3.7.: Mass distributions in the 1-jet high-pt category.

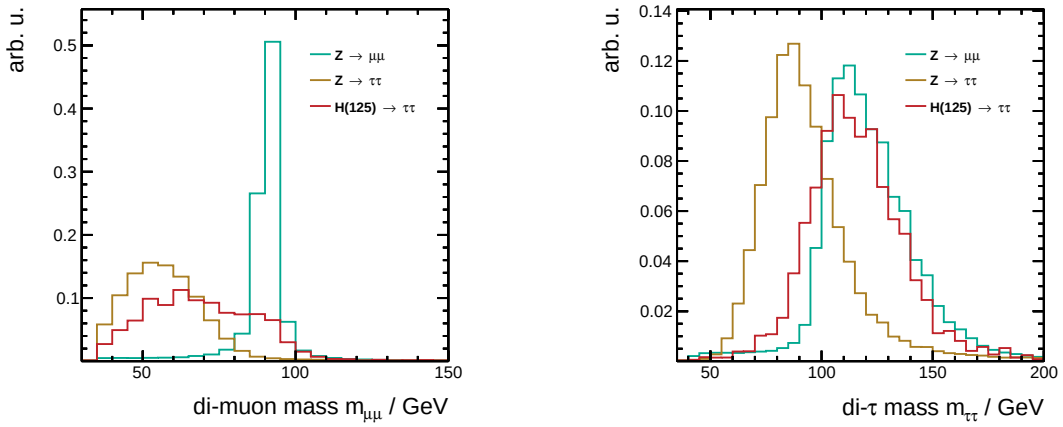


Figure 3.8.: Shapes (normalised to unity) of the mass distributions in the 1-jet high-pt category for the main sources of background events, $Z \to \mu\mu$ and $Z \to \tau\tau$ and for the $H \to \tau\tau$ signal.

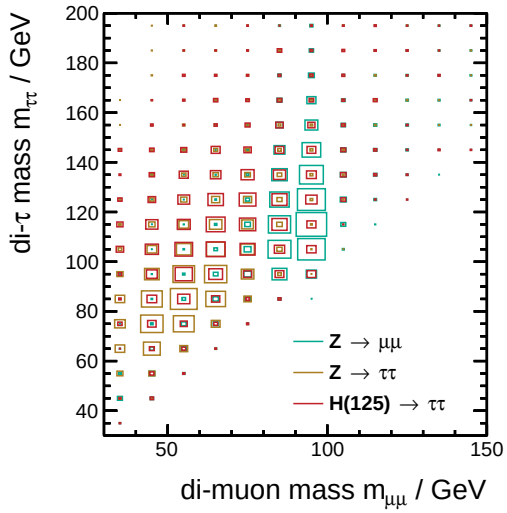


Figure 3.9.: Correlations of the mass variables in the 1-jet high-pt category for the main sources of background events, $Z \to \mu\mu$ and $Z \to \tau\tau$ and for the $H \to \tau\tau$ signal.

The third component, $\mathcal{L}_{\text{miss}}$, models the compatibility of the measured MET, $\vec{\not{E}}_{\text{T}}$, including its uncertainty given by the covariance matrix C, with the di-τ momentum hypothesis.

$$\mathcal{L}_{\text{miss}}\left(\vec{\not{E}}_{\text{T}}, \vec{a}_1, \vec{a}_2\right) = \frac{1}{2\pi\sqrt{\det C}} \exp\left(-\frac{1}{2}\vec{\not{E}}_{\text{T}}^{\text{T}} \cdot C^{-1} \cdot \vec{\not{E}}_{\text{T}}\right)$$

Finally, the probability, P, for a given mass hypothesis, $m_{\tau\tau}^{\text{hyp}}$, is given as the integral of the likelihood function over all values of the unknown parameters $\vec{a}_1$ and $\vec{a}_2$ that are physically allowed. The dependency on x and $m_{\nu\nu}$ is introduced via these integration boundaries whereas $\mathcal{L}_{\text{miss}}$ introduces the dependency on φ

$$P\left(m_{\tau\tau}^{\text{hyp}}\right) = \int\!\!\cdots\!\!\int \mathrm{d}\vec{a}_1\, \mathrm{d}\vec{a}_2\, \delta\left[m_{\tau\tau}^{\text{hyp}} - m_{\tau\tau}\left(p_1, p_2, \vec{a}_1, \vec{a}_2\right)\right] \mathcal{L}\left(p_1, p_2, \vec{\not{E}}_{\text{T}}, \vec{a}_1, \vec{a}_2\right)$$

The reconstructed invariant di-τ mass is identified as the value $\hat{m}_{\tau\tau}^{\text{hyp}}$ that maximises the probability P. In the di-muon channel it provides additional information compared to the di-muon mass. However, the mass reconstruction suffers from the four neutrinos and therefore the mass resolution is worse compared to the final states involving hadronically decaying τ leptons. This is of course related to the worse MET resolution. The distribution of the reconstructed di-τ mass in the 1-jet high-pt category is shown in figure 3.7 (right) for the 8 TeV analysis. Correspondingly, the shapes of the main backgrounds and the signal are illustrated in figure 3.8 (right). The shape comparison shows the benefit of the reconstruction in separation $\text{Z} \to \tau\tau$ from $\text{H} \to \tau\tau$ events with respect to the one of the visible mass. Events without genuine MET get shifted towards higher values due to a misinterpretation of the measured MET. This is the reason why the reconstructed mass of the di-τ candidate almost completely looses its power to separate $\text{Z} \to \mu\mu$ from signal events.

Figure 3.8 supports the assumption that the separation between the three processes, $\text{Z} \to \mu\mu$, $\text{Z} \to \tau\tau$ and $\text{H} \to \tau\tau$, is improved based on the correlation of the two mass definitions as shown in figure 3.9 which is reflected in the fact that the three samples roughly concentrate in different regions of the masses plane. On the other hand, it is obvious that other discriminating information is needed in order increase the sensitivity of this channel to a reasonable level.

3.3.3. Collinear Approximation

Thirdly, a collinear approximation for the mass reconstruction is considered. Caused by the difference in mass between the heavy boson and the light τ lepton the latter ones are usually boosted significantly. Consequently, the τ decay products escape in a narrow cone around the τ flight direction. The collinear approximation now assumes that the flight directions of the visible and the invisible decay products are exactly the same.

$$p(\tau_{1/2}) = \left(1 + x_{\nu\nu,1/2}\right) p(\mu_{1/2}) \quad \text{with} \quad x_\mu, x_{\nu\nu,1/2} \in \mathbb{R} \tag{3.2}$$

Furthermore, the sum of the invisible momenta is assumed to match the measured MET in the transversal plane.

$$\vec{\not{E}}_T = x_{\nu\nu,1}\,\vec{p}_T(\mu_1) + x_{\nu\nu,2}\,\vec{p}_T(\mu_2) \tag{3.3}$$

Figure 3.10 illustrates these two assumptions. The two equations (3.3) are then solved for the parameters $x_{\nu\nu,1/2}$ and the parametrisation (3.2) is then used the calculate the invariant di-τ mass. However, in some cases one or both parameters yield negative results indicating neutrinos flying in the opposite direction of the muons. For these non-physical solutions no mass reconstruction is possible using the collinear approximation. Because of this artificial loss of events this mass definition itself is not used in the analysis. Instead, the information whether a physical solution is available is used. This binary variable provides information about the correlation of the MET and the flight directions of the muons in the transversal plane.

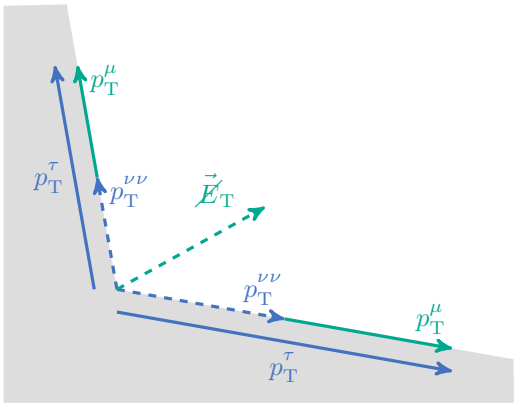


Figure 3.10.: Illustration of the collinear approximation for the reconstruction the invariant di-τ mass. If the missing momentum vector, $\vec{\not{E}}_T$, points into the grey shaded region, the approximation yields non-physical results.

3.4. Distance of Closet Approach of the Muon Tracks

$Z \to \mu\mu$ events are characterised by prompt muons originating from the Z decay at the position of the primary vertex. Muons from $Z/H \to \tau\tau \to \mu\mu$ events originate from the secondary vertices of the τ decays. This difference between $Z \to \mu\mu$ events on the one hand and $Z \to \tau\tau$ and $H \to \tau\tau$ events on the other hand is exploited in the distance of closest approach (DCA) between axes aligned with the muon flight directions (which is usually abbreviated by the DCA of the two muon tracks) as illustrated in figure 3.11.

On reconstruction level, the distributions are smeared due to the limited track resolution and the vertex reconstruction (see section 2.2.2). Especially for prompt muons this leads to non-vanishing but small values of the DCA. The resolution of the DCA depends on the opening angle of the two muon flight directions, $\cos(\mu^+, \mu^-)$, as well as on the transverse momenta of the two muons. The more the opening angle deviates from zero, the better the resolution of the DCA is. The same is true for higher transverse muon momenta. To account for these differences in the resolution, the significance of the

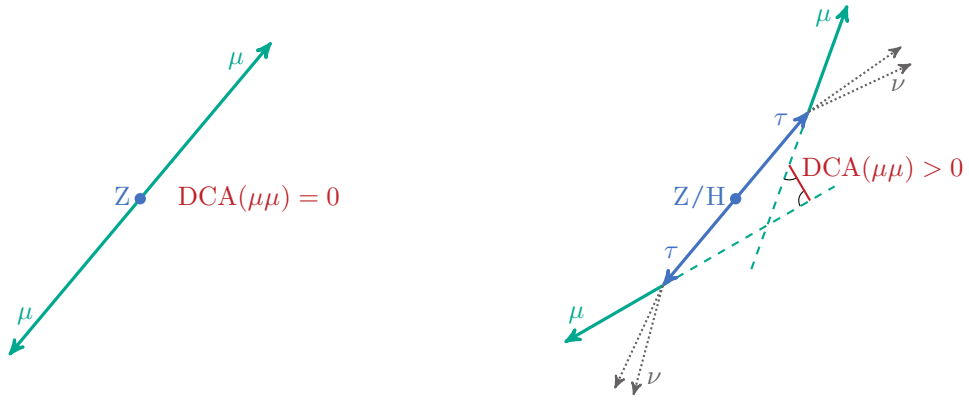


Figure 3.11.: Illustration of the distance of closest approach (DCA) between axes aligned with the muon flight directions for $Z \to \mu\mu$ and $Z/H \to \tau\tau \to \mu\mu$ events on generator level. This variable distinguishes between Z boson decays into prompt muons (left) and $\tau\tau$ final states (right).

DCA is ecploited:

$$\text{DCA} = \sqrt{\vec{d}_{\text{PCA}}^{\,\text{T}} \cdot \vec{d}_{\text{PCA}}} \quad \text{and} \quad \sigma_{\text{DCA}} = \sqrt{\frac{\vec{d}_{\text{PCA}}^{\,\text{T}} \cdot \text{C} \cdot \vec{d}_{\text{PCA}}}{\vec{d}_{\text{PCA}}^{\,\text{T}} \cdot \vec{d}_{\text{PCA}}}}$$

$$\text{DCA significance} = \frac{\text{DCA}}{\sigma_{\text{DCA}}}$$

where $\vec{d}_{\text{PCA}}$ denotes the distance vector between the points of closest approach and and the reconstruction uncertainties are described by the covariance matrix C. It has been found that the longitudinal component of the DCA does not add any information to the transverse DCA and therefore the two dimensional DCA is used in the following, expressed as $\log_{10}\text{DCA}(\mu\mu)$. For simplicity reasons, the dimensionless DCA significance is from now on consistently abbreviated by DCA.

The DCA is one of the most crucial variables used in this analysis as it provides strong separation power for the discrimination of prompt muons and muons originating from τ decays and as it is the basis for the data-driven estimation of the $Z \to \mu\mu$ background. Therefore a good agreement of the measurement and the simulation of this variable is essential. A calibration is applied in order to correct for possible disagreements. The calibration is done by shifting the simulated DCA values on event level rather than reweighting the whole distribution of simulated events in order to preserve the correlation to all other event quantities.

The calibration is performed in the following steps. First, the distributions of the DCA variable are parametrised by a fit function both for simulated $Z \to \mu\mu$ events and for data events. A double asymmetric Gaussian is chosen for the central part and two exponential functions for the lower and upper tail of the fit function. The fit is performed in the following di-muon phase space regions that differ in the precision of the DCA measurement.

- $\cos(\mu^+, \mu^-)$: [-1.0, -0.5, 0, 0.5, 1.0]

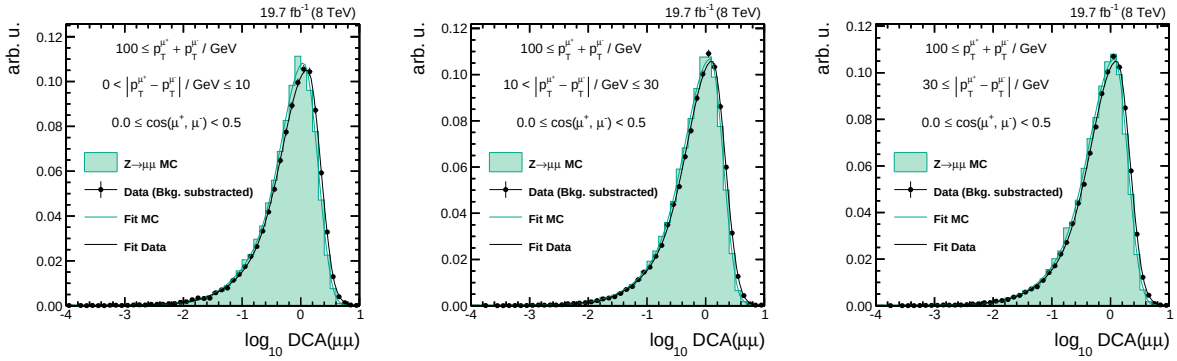


Figure 3.12.: Distributions of the DCA variable in simulated $Z \to \mu\mu$ and data events are parametrised by a fit function. The fits are performed in bins of the angle between the flight directions of the two muons, $\cos(\mu^+, \mu^-)$, and the transverse momenta of the two muons, $p_T^{\mu^+} + p_T^{\mu^-}$ and $|p_T^{\mu^+} - p_T^{\mu^-}|$. The examples shown here vary in the difference of the transverse muon momenta.

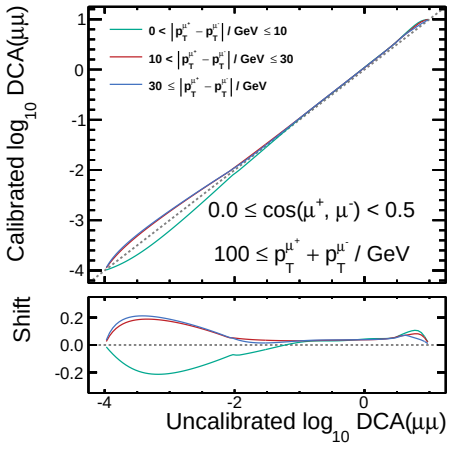


Figure 3.13.: Calibration of the DCA variable for simulated $Z \to \mu\mu$ events exemplary shown for the three phase space regions shown above.

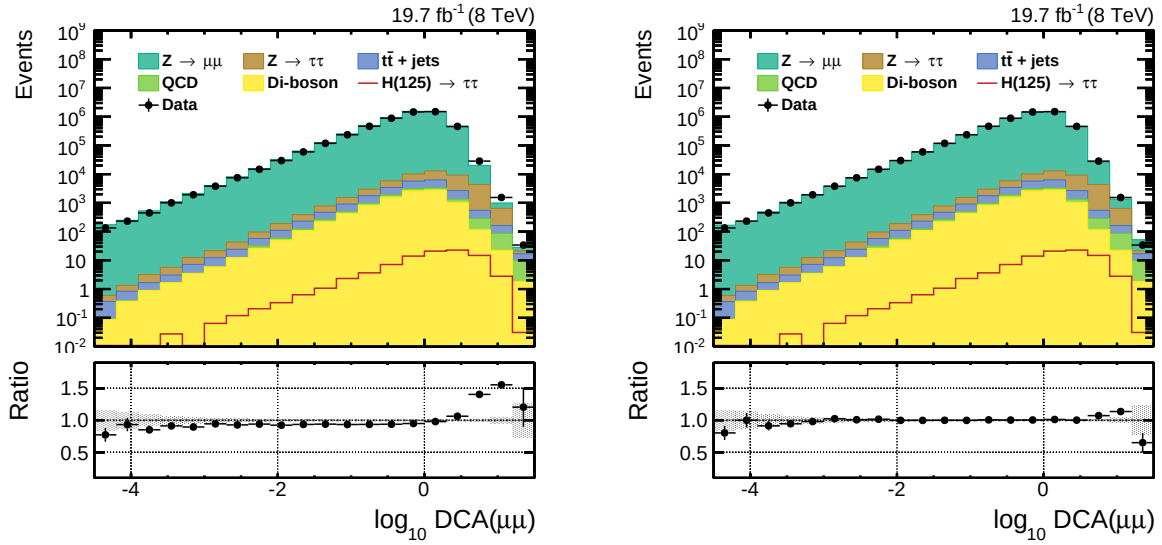


Figure 3.14.: Distribution of the DCA variable before (left) and after (right) the calibration of the simulated $Z \to \mu\mu$ events. The ratio compares the observation in data to the expectation given by the sum of all backgrounds. The statistical errors are shown in the error band.

- $\left(p_{\mathrm{T}}^{\mu^+} + p_{\mathrm{T}}^{\mu^-}\right)$ / GeV: [0, 30, 50, 100, ∞]
- $\left|p_{\mathrm{T}}^{\mu^+} - p_{\mathrm{T}}^{\mu^-}\right|$ / GeV: [0, 10, 30, ∞]

Before fitting the distributions in data, the expected contribution from all backgrounds apart from the $\mathrm{Z} \to \mu\mu$ simulation are subtracted by taking into account the final background estimation as described in section 3.6. Especially the $\mathrm{Z} \to \tau\tau$ contribution is taken from the embedded data sample and therefore does not need to be calibrated. Figure 3.12 shows three examples for these fits where $\left|p_{\mathrm{T}}^{\mu^+} - p_{\mathrm{T}}^{\mu^-}\right|$ is varied for constant opening angles and $p_{\mathrm{T}}^{\mu^+} + p_{\mathrm{T}}^{\mu^-}$.

The next step is based on the normalised parametrisations of the fits from the step before. These are taken as probability densities to measure a given value of the DCA in the $\mathrm{Z} \to \mu\mu$ simulation or in data, respectively. Based on the corresponding cumulative distributions each DCA value is mapped to a quantile which specifies the probability to measure a DCA value smaller than or equal as the given one. Each DCA value in the $\mathrm{Z} \to \mu\mu$ simulation is then shifted to the value which gives the same quantile in the parametrisation for data. Thereby, the probability density of the DCA values in the simulation is matched to the one in data. Figure 3.13 illustrates the calibration step for the three example fits shown in figure 3.12.

Figure 3.14 shows the improved agreement of data and the sum of all backgrounds after the DCA calibration for the 8 TeV analysis after putting together the results from the fits in all phase space regions. Although the distributions of data and simulation in the individual phase space regions, that have been used for the calibration, are expected to match per construction, there are several reasons that explain the remaining discrepancies in the tails of the full calibrated DCA distribution: The most important reason is the difficulty to parametrise the tails of the DCA distributions in phase space regions with low numbers of events. The fit parameters yield substantial uncertainties in these cases and fluctuations of the real events result in differences between the actual distributions and the fits. Secondly, the full distribution is a combination of calibrations in multiple categories, by which discrepancies can be enhanced in some cases, which is especially true for the tail regions. The calibration is performed similarly for the 7 TeV analysis.

3.5. Multivariate Signal Extraction

It has been pointed out in the previous sections that the sensitivity of the search in the di-muon channel based on mass variables is quite low after the pre-selection and subsequent event categorisation. Additional selections of signal-like events or a final discriminator with more separation power is needed. The preliminary analysis [83, 84] followed the first approach. The pre-selection and event categorisation is followed by a multivariate selection of signal-like events and concluded by a fit of two-dimensional mass distributions ($m_{\mu\mu}$ vs. $m_{\tau\tau}$) as the final discriminator.

For the MVA step boosted decision trees have been trained on a set of discriminating variables. Four different trainings have been performed for the 0/1-jet categories and the 2-jet category and

separately for the 7 TeV and the 8 TeV analysis. Cut thresholds on the BDT discriminators then have been optimised based on the $S/\sqrt{S+B}$ ratio separately for all five event categories.

An improvement of this method could be the exploitation of more significant variables, tighter cuts on the MVA discriminator or a completely new MVA approach. Since for physics reasons the mass variables are already among the most significant ones, further variables are not expected to improve the sensitivity significantly. Tighter cuts on the MVA discriminators result in distributions containing less events. This complicates the modelling of the background processes and increases therefore both the statistical and systematic uncertainties. In the scope of this thesis, a new multivariate approach has been implemented and studied.

3.5.1. Two-staged Approach

The signal extraction presented here follows a new idea but is based on the same variables. As already pointed out in section 3.3, two dominant irreducible backgrounds, $Z \to \mu\mu$ and $Z \to \tau\tau$, have to be considered. Consequently two discriminators are defined for the suppression of these backgrounds.

The first discriminator is aimed to distinguish between $\tau\tau$ final states and all other backgrounds. The main purpose of this discriminator is a suppression of the $Z \to \mu\mu$ background. The task of the second discriminator is then to distinguish between $Z \to \tau\tau$ and $H \to \tau\tau$ events. As the largest background is already suppressed by the first stage, this discriminator can be fully optimised on the separation of these very similar signatures.

Finally, the two discriminators are combined into a single discriminator rather than cutting on them. More important than avoiding to discard too many signal events is the advantage that also all the background events after the pre-selection remain and can be used to constrain the background-related uncertainties in the background-dominated regions of the discriminator.

3.5.2. Trainings of the Boosted Decision Trees

For the two discriminators boosted decision trees (BDTs) from the TMVA package of ROOT (see section 2.5.3) are trained. Each discriminator is trained separately in the 0/1-jet categories and in the 2-jet category and separately for the 7 TeV and the 8 TeV analysis, the same way as it has been done previously. The result are four trainings. Independent training samples have been used to avoid learning and reproduction of statistical fluctuations. The samples used for the trainings follow closely the estimation methods as described in section 3.6. Especially for $Z \to \tau\tau$ events the embedded data set and for QCD events the same-sign muon charge selection in data have been used. Background events are weighted according to their cross section and the integrated luminosity of the analysed data set. For the $H \to \tau\tau$ signal a superposition of all samples with Higgs boson masses from 110 to 145 GeV in 5 GeV steps is taken in order to avoid a bias towards a single mass hypothesis. For all trainings all important Higgs boson production modes are considered according to their cross sections: gluon fusion, vector boson fusion and the production in association with vector bosons or quarks. The complete signal sample is reweighted such that the ratio S/B for the training is one.

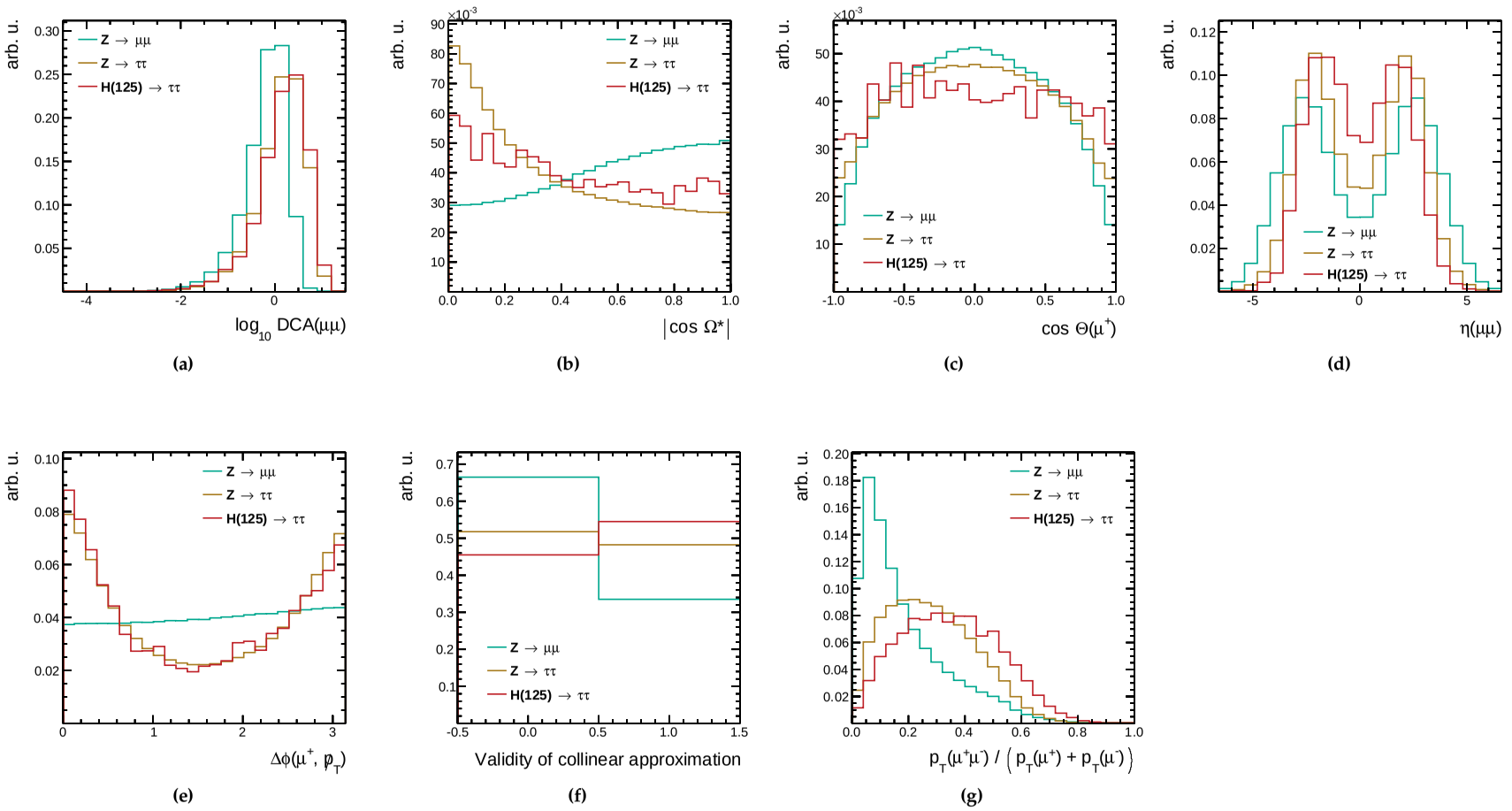


Figure 3.15.: Discriminating variables used in the BDT trainings for the 0/1-jet categories. The distributions normalised to unity of the most dominant background events, $Z \to \mu\mu$ and $Z \to \tau\tau$, are compared to the one of the $H \to \tau\tau$ signal.

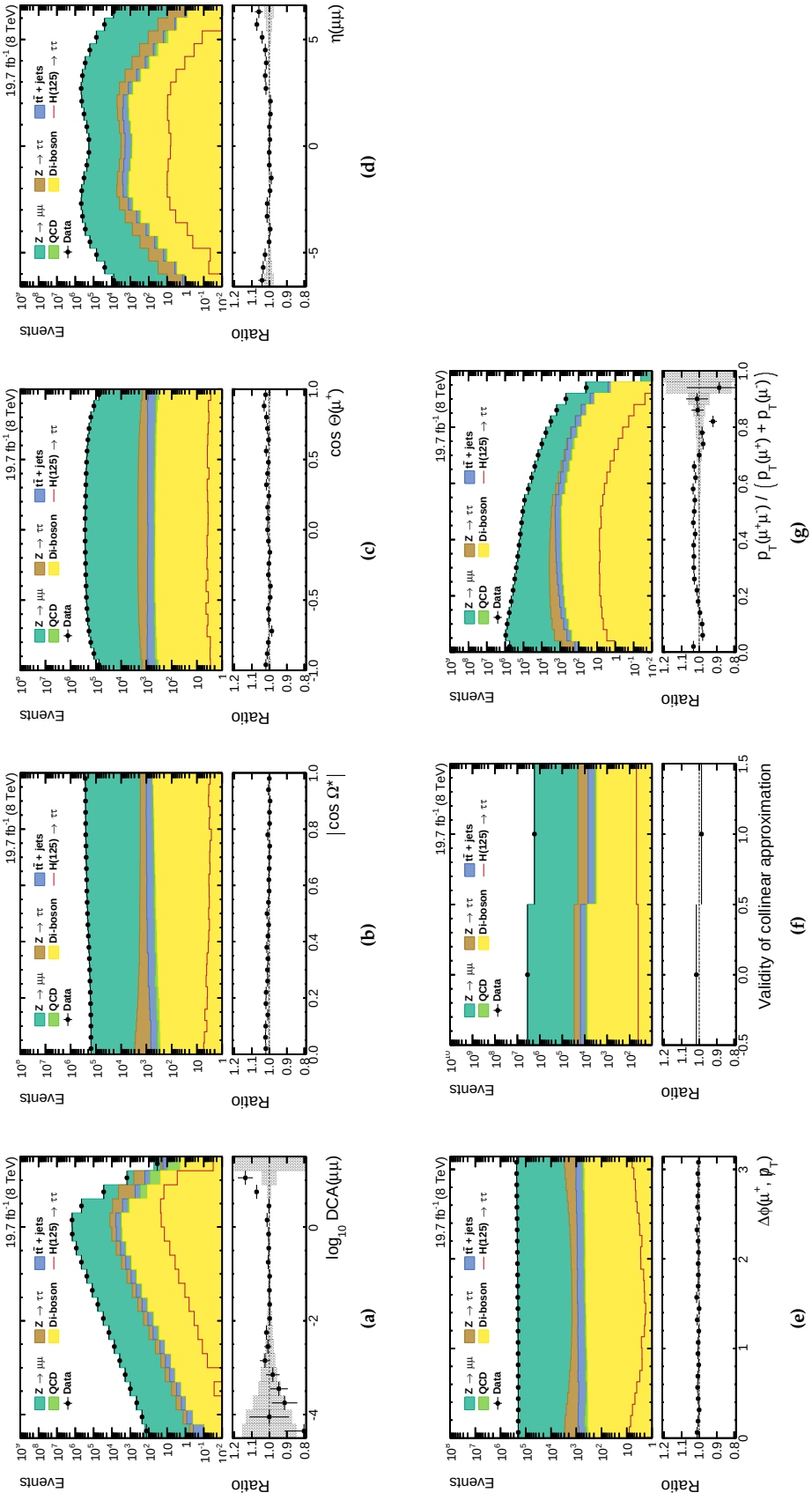


Figure 3.16.: Distributions of the discriminating variables used in the BDT trainings for the 0/1-jet categories. The ratio compares the observation in data to the expectation given by the sum of all backgrounds. The statistical errors are shown in the error band. The discrepancies between the observation and the background-only expectation are below 5 % in almost all distributions. These are addressed by the background estimation studies based on the resulting discriminator.

The baseline selection of input variables is the same as in the preliminary analysis. The following variables are exploited in the 0/1jet trainings for both BDTs. Their separation power is illustrated in figure 3.15 and their full distributions for all samples are compared with the one of data in figure 3.16.

- Distance of closest approach of the two muon tracks, $\log_{10} \mathrm{DCA}(\mu\mu)$, see figures 3.15a and 3.16a.
- Angle between the μ^+ and the normal of the di-μ production plane spanned by the di-μ momentum direction and the beam direction, $|\cos\Omega^*|$, see figures 3.15b and 3.16b.
- Polar angle $\cos\Theta(\mu^+)$ of the μ^+ in the di-μ rest system, see figures 3.15c and 3.16c.
- Pseudorapidity of the di-μ system, $\eta(\mu\mu)$, see figures 3.15d and 3.16d.
- Azimuthal angle between the μ^+ and the missing momentum in the transverse plane, $\Delta\phi(\mu^+, \not{E}_\mathrm{T})$, see figures 3.15e and 3.16e.
- Validity of the collinear approximation, see figures 3.15f and 3.16f.
- Transverse momentum ratio, $p_\mathrm{T}(\mu\mu) / \left(p_\mathrm{T}(\mu^+) + p_\mathrm{T}(\mu^-)\right)$, see figures 3.15g and 3.16g.

These variables are complemented by the mass variables that previously have been used for final discriminator. As pointed out in section 3.3 the di-muon mass, $m_{\mu\mu}$, provides a good separation between $\mathrm{Z} \to \mu\mu$ events and Z/H decays with neutrinos in the final state and is therefore used as an input variable in the first stage training for the suppression of the $\mathrm{Z} \to \mu\mu$ background. On the other hand, the di-τ mass is used in the second stage training as this reconstruction provides a better separation between $\mathrm{Z} \to \tau\tau$ and $\mathrm{H} \to \tau\tau$ events because of the different invariant mass of the intermediate boson.

Together with the mass variables the DCA variable is among the most powerful input variables, especially in the first stage training. The reason is its good separation power and the small correlation to the mass variables.

Additional information is incorporated based on variables exploiting the MET. It has already been pointed out that the resolution of the MET increases with the boost of the hadronic recoil in the event. The combined 0/1-jet trainings contain a large fraction of events without jets, where the H or Z boson is produced almost at rest in the laboratory frame. This results in a worse MET resolution compared to 2-jet events and large systematic uncertainties. Therefore, the MET, $\not{E}_\mathrm{T}$, is not used itself as input variable. Instead, the correlation of the MET with the muon momenta is exploited. Events with genuine MET caused by neutrinos show a strong correlation between the muon flight directions and the missing momentum in the transversal plane, whereas in other events like $\mathrm{Z} \to \mu\mu$ events there is no physically motivated connection between the measured MET and the momenta of the muons.

The azimuthal correlation between the momentum of the positively charged muon and the missing momentum, $\Delta\phi(\mu^+, \not{E}_\mathrm{T})$ indicates that the neutrinos in $\mathrm{Z/H} \to \tau\tau$ events prefer to fly in the same direction as the muons, whereas to $\mathrm{Z} \to \mu\mu$ almost no correlation is seen. Similarly the calculation of the di-τ mass following the collinear approximation fails in a larger fraction for $\mathrm{Z} \to \mu\mu$ events compared to $\mathrm{Z/H} \to \tau\tau$ events because of the missing correlation between the muons and the MET.

The pseudorapidity of the di-muon system, $\eta(\mu\mu)$ exploits the differences in the production of Z and H bosons at the LHC as introduced in section 1.2.1. Due to the asymmetric initial state cause by the $q\bar{q} \to Z$ production and the resulting longitudinal boost of the Z boson, the system of the visible decay products tends to be less central than the ones of H bosons preferably produced in rest via the gluon fusion. For similar reasons, the ratio of the transverse momenta, $p_T(\mu\mu)/\left(p_T(\mu^+) + p_T(\mu^-)\right)$ shows lower values for the $Z \to \mu\mu$ values compared to $H \to \tau\tau$ signal. Both variables together with the angular variables are important for the separation of $H \to \tau\tau$ and $Z \to \tau\tau$ events.

The list of input variables for the trainings of both BDTs in the 2-jet category is slightly adjusted to the specialities of the VBF production mechanism. The baseline variables in these trainings are the following ones. Again, these variables are complemented by the two mass definition in the same way as for the trainings in the 0/1-jet categories. Their separation power is illustrated in figure B.1 and their full distributions for all samples are compared with the one of data in figure B.2.

- Distance of closest approach of the two muon tracks, $\log_{10} \mathrm{DCA}(\mu\mu)$, see figures B.1a and B.2a.
- Angle between the μ^+ and the normal of the di-μ production plane spanned by the di-μ momentum direction and the beam direction, $|\cos\Omega^*|$, see figures B.1b and B.2b.
- Polar angle $\cos\Theta(\mu^+)$ of the μ^+ in the di-μ rest system, see figures B.1c and B.2c.
- Validity of the collinear approximation, see figures B.1e and B.2e.
- Azimuthal angle between the μ^+ and the missing momentum in the transverse plane, $\Delta\phi(\mu^+, \not{E}_T)$, see figures B.1f and B.2f.
- Magnitude of the missing transverse energy, $\not{E}_T$, see figures B.1g and B.2g.
- Di-jet mass, m_{jj}, see figures B.1d and B.2d.
- Difference in the pseudorapidity between the two jets, $\Delta\eta_{jj}$, see figures B.1h and B.2h.

The most important change with respect to the 0/1-jet trainings lies in the exploitation of di-jet variables quantifying the VBF-likeliness. High values of both the di-jet mass, m_{jj} and the gap in the pseudorapidity between the two jets, $\Delta\eta_{jj}$ discriminate the $H \to \tau\tau$ signal from the background of Z decays.

Because of the higher MET resolution in this case where the H boson recoils against to di-jet system, the MET, $\not{E}_T$, is used as an input in these trainings. Related systematic uncertainties are discussed below. The better MET resolution manifests itself also in the better discrimination power of the variable quantifying the validity of the collinear approximation.

Figure 3.17 shows the outputs of the BDT trainings for the first and the second stage, respectively, evaluated on an independent sample. The left plots compare the shapes of the distributions for the most important samples normalised to unity. It is clearly visible that the first stage BDT, B_1, discriminates between $\tau\tau$ final states and the $Z \to \mu\mu$ background whereas the second stage BDT, B_2, is able to separate between $H \to \tau\tau$ and $Z \to \tau\tau$ events. However, the plots also reveal that not all of the three samples can be distinguished from each other based on one single discriminator. In the first stage,

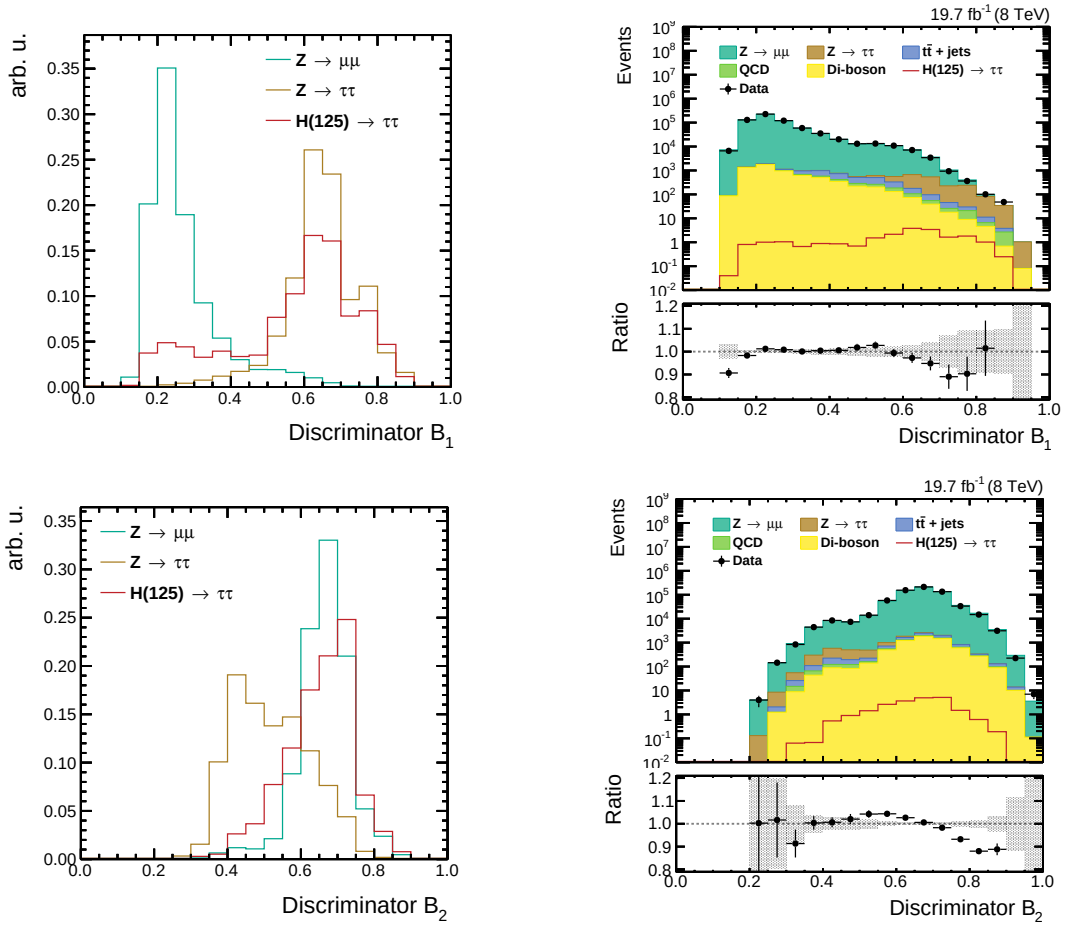


Figure 3.17.: Output of the first (top) and the second (bottom) stage BDT in the 1-jet high-pt category of the 8 TeV analysis. The first stage BDT discriminates between $\tau\tau$ final states and mainly the $Z \to \mu\mu$ background, whereas the second stage BDT is optimised on the discrimination between $H \to \tau\tau$ signal and $Z \to \tau\tau$ background (left). The complete distribution of the backgrounds is compared with data (right). The ratio compares the observation in data to the expectation given by the sum of all backgrounds. The statistical errors are shown in the error band. The visible discrepancies between data and the background-only expectation are subject of the background estimation methods described in section 3.6. A complete set of plots for all categories is shown in figures B.3 and B.4.

$H \to \tau\tau$ events with di-muon masses close to the Z mass, see figure 3.8 (left), result in BDT outputs in the range that is dominated by $Z \to \mu\mu$ events. These events are hardly distinguisable if one only looks at the first stage BDT. The second stage BDT provides nearly no separation between the $H \to \tau\tau$ signal and the $Z \to \mu\mu$ background because to the similarity in the di-τ mass, see figure 3.8 (right). This is a more technical justification for the need of two discriminators to handle the suppression of the two most dominant backgrounds.

3.5.3. Two-staged final discriminator

The full information of the two discriminators is given by their correlations in the form of two-dimensional distributions. However, a limited number of simulated events complicates the exploitation

of the full two-dimensional distributions. The binning of the histograms would have to be sufficiently coarse to ensure a reasonable number of events in every bin. This limits the sensitivity of the discriminator.

Another option would be the construction of a third discriminator. One can for example imagine a third BDT or a likelihood ratio constructed from the two individual discriminators. This option has been studied previously [89,90]. It has been shown that the discrimination is not better than the one of a single BDT trained to separate the H $\to \tau\tau$ signal from all other backgrounds, because the combined discriminator is also optimised on separating the signal against a composition of different backgrounds, where the Z $\to \mu\mu$ background is the most dominant one. Therefore the combined discriminator focusses on the separation against this huge background. An option would have been to reweight the composition of backgrounds in order to let the training concentrate more on events of the less dominant backgrounds. This option has been rejected because of a lack of a good motivation for the applied weights. Their choice would need an optimisation only under technical considerations.

The solution is an analytical formula for the combination of the two discriminators, which does not depend on the composition of the background samples nor on the signal to background ratio.

$$D_{\mathrm{sig/bkg}}(B_1, B_2) = \int_{-\infty}^{B_1} \mathrm{d}B_1' \int_{-\infty}^{B_2} \mathrm{d}B_2' \; P_{\mathrm{sig/bkg}}(B_1', B_2') \tag{3.4}$$

This procedure referred to as PDF integration method, is illustrated in figure 3.18. The full two-dimensional probability density functions (PDFs) of the signal and backgrounds as functions of the two BDT discriminators, B_1 and B_2, are denoted by $P_{\mathrm{sig/bkg}}(B_1, B_2)$. For each event yielding the BDT outputs B_1 and B_2 the combined discriminator is then defined as the integral of these PDFs in the rectangular two-dimensional range from $(-\infty, -\infty)$ up to the point (B_1, B_2)[2]. Thus, the method is also independent of the binning of the two-dimensional distributions and there is no problem of limited statistical precision because of the integration over large numbers of events in the signal-dominated regions.

Two choices for the PDFs are possible: $P_{\mathrm{bkg}}(B_1, B_2)$ and $P_{\mathrm{sig}}(B_1, B_2)$. The background PDF is based on the sum of all backgrounds as they are expected in data (excluding the signal expectation). The estimation of the backgrounds is performed similarly to the final estimation as described in section 3.6. Especially the Z $\to \tau\tau$ background is taken from the embedded data set and the QCD background is also determined in a data-driven way. The signal PDF is constructed from the superposition of the samples for all Higgs boson mass hypotheses from 110 to 145 GeV in steps of 5 GeV the same as it has been done for the trainings of the BDTs. Again, all important Higgs boson production modes are considered according to their cross sections: gluon fusion, vector boson fusion and the production in association with vector bosons or quarks. The reason is to avoid any bias towards a given Higgs boson mass.

[2] In practice, the integral is not calculated for each event but pre-calculated look-up tables with a very fine binning in B_1 and B_2 are used instead in order to accelerate the process of retrieving the discriminator values for all events. These tables correspond to the cumulative distributions shown in figure 3.18

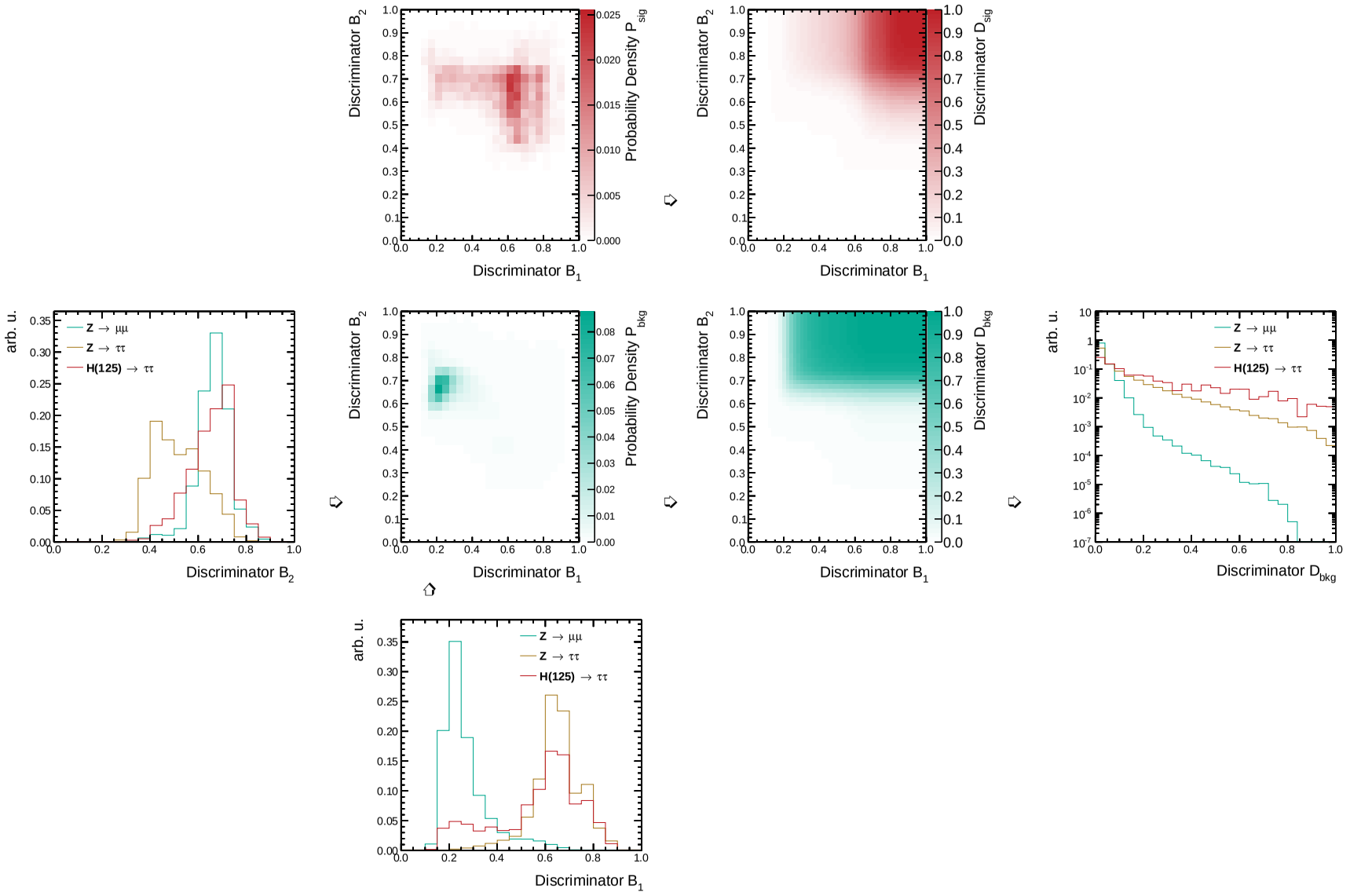


Figure 3.18.: Illustration of the PDF integration method for combining the two discriminators B_1 (bottom) and B_2 (left) into the final discriminator $D = D_{\mathrm{bkg}}$ (right). The second column shows the probability densities for the sum of all signal events with Higgs boson masses from 110 to 145 GeV (top) and the sum of all background events (centre). The third column contains the cumulative distributions as determined by formula 3.4.

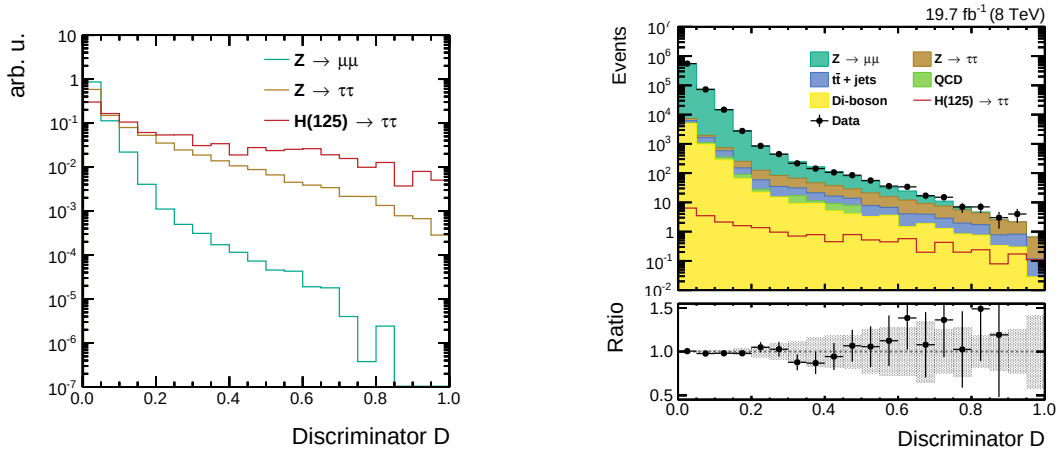


Figure 3.19.: Output of the final discriminator, D, in the 1-jet high-pt category of the 8 TeV analysis which is an analytical combination of the two BDT discriminators, B_1 and B_2. It is clearly visible that the discriminator yields significantly different shapes for the main processes considered, $Z \to \mu\mu$, $Z \to \tau\tau$ and $H \to \tau\tau$ events (left). The complete distribution of the backgrounds is compared with data (right). The ratio compares the observation in data to the expectation given by the sum of all backgrounds. The statistical errors are shown in the error band. Additional systematic uncertainties are assigned to cover the remaining discrepancies. A complete set of plots for all categories is shown in figure B.5.

The resulting discriminators, $D_{\text{bkg}}(B_1, B_2)$ and $D_{\text{sig}}(B_1, B_2)$, are found to provide a similar sensitivity. As final discriminator the background version is chosen.

$$D = D(B_1, B_2) \equiv D_{\text{bkg}}(B_1, B_2)$$

A given value of this discriminator D yields the probability that the values of the two BDT discriminators, B_1 and B_2, are both smaller in the background sample. The complete distribution of the final discriminator in the 1-jet high-pt category is displayed in figure 3.19. The agreement between the data and the background-only expectation is acceptable, but small trends in the ratio are visible. They are covered by the statistical and systematic uncertainties, where the latter ones are discussed in the following.

3.6. Background Modelling

The production cross sections for the background processes are orders of magnitude larger than the ones for the signal. The signal excess is searched in phase space regions that are still dominated by the backgrounds. This requires a thorough prediction of the background yields and of the shapes of the discriminators, which is described in this section. Systematic uncertainties arising from these methods are summarised together with the complete set of uncertainties in the next section 3.7.

The dominant backgrounds from Z boson decays are modelled in a data-driven way in order to reduce the systematic uncertainties. The method based on DCA template fits for the $Z \to \mu\mu$ background and the embedding technique for the $Z \to \tau\tau$ background are explained in the following. The other sources of background processes ($t\bar{t}$, di-boson, W+jets and QCD multi-jet events) only contribute

marginally to the overall yield measured in data. These backgrounds are estimated based on the simulation.

3.6.1. Data-driven Estimation of the Z $\to \mu\mu$ Background – DCA Template Fits

The Z $\to \mu\mu$ process is the most dominant source of background events for the search in the H $\to \tau\tau \to \mu\mu$ channel and therefore requires a particularly thorough estimation technique. The final discriminator, D, is a combination of multiple variables that all show some smaller or larger discrepancies in the agreement between data and simulation on their own. Some of these differences cancel each other, some of them get enhanced and reveal significant trends in the ratios of data over the sum of the expected backgrounds. There is no analytical or even a physical description of the shape of the discriminators. Therefore the following data-driven method is chosen for the modelling of the Z $\to \mu\mu$ background.

Scale factors for the Z $\to \mu\mu$ simulation are derived from DCA template fits to data in bins of the di-muon mass, $m_{\mu\mu}$, and the reduced discriminator, D_{red}, which results in both a normalisation and a shape correction of the simulation of the final discriminator D. The DCA is well suited for these template fits because of two reasons that have already been exploited advantageously for the BDT trainings: firstly the simulated DCA values of Z $\to \mu\mu$ events are calibrated to the data and secondly the DCA is characterised by only weak correlations with the other event quantities, especially the mass variables and the other MVA input variables. Because of the latter advantage, the template fits are performed in bins of the di-muon mass, $m_{\mu\mu}$, and of the reduced discriminator, D_{red}, per event category. The reduced Discriminator differs from the final discriminator only in the set of input variables. The masses and and the DCA are excluded from these dedicated BDT trainings.

Figure 3.20 shows three examples for these fits in the 1-jet high-pt category. The shapes of the Z $\to \mu\mu$ and Z $\to \tau\tau$ distributions are fit to data, where the contributions from all other backgrounds have been subtracted. The resulting fit parameters and its statistical errors are shown below the distributions. The Z $\to \tau\tau$ background is not subtracted from data beforehand. Instead, the Z $\to \tau\tau$ sample is fit simultaneously with the Z $\to \mu\mu$ sample in order to account for possible H $\to \tau\tau$ signal. Because of the similar DCA shapes of Z $\to \tau\tau$ and H $\to \tau\tau$ events (see figure 3.16a), the fit parameter for the Z $\to \tau\tau$ normalisation is allowed to freely float in the range from 0.9 to 1.1. Its fit parameter is not used in the following.

In figure 3.21 the scale factors (left) and statistical errors resulting from the fits (right) are shown for all fits in the 1-jet high-pt category. The bins of the examples in figure 3.20 are highlighted. The highest mass bin ranges up to 1000 GeV. The Z $\to \mu\mu$ contributions are scaled with these factors and therefore mapped to the measurement from data. Several effects are identified: firstly, the correction in this category leads to scaling up of the simulation for low values of the discriminator, D_{red}, and to a scaling down of the simulation of high discriminator values. Therefore the shape is altered. Secondly, as expected the corrections are smaller in the region around the di-muon mass and larger in its tails. Also the statistical precision is directly connected to the number of events in the different fits. The precision increases for di-muon masses in the region of the Z mass peak and for low discriminator values, D_{red} that are dominated by the Z $\to \mu\mu$ background.

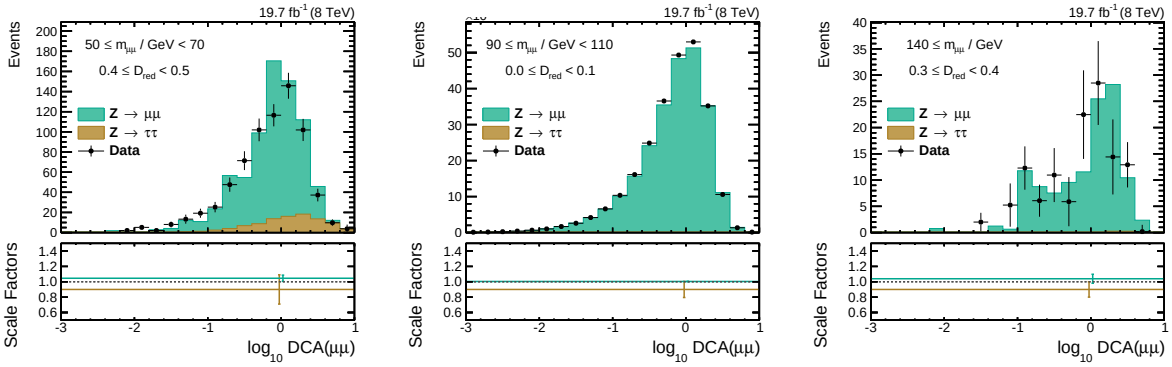


Figure 3.20.: Examples for the DCA template fits. The normalisations of $Z \to \mu\mu$ and $Z \to \tau\tau$ events are fit to the one of data, where all other background contributions are subtracted, in bins of the di-muon mass $m_{\mu\mu}$ and the reduced discriminator D_{red}. The sub-plots show the resulting fit parameters. The $Z \to \tau\tau$ scale factor is fit to avoid biases from signal, which has a similar shape as $Z \to \tau\tau$ events. This scale factor is not further used.

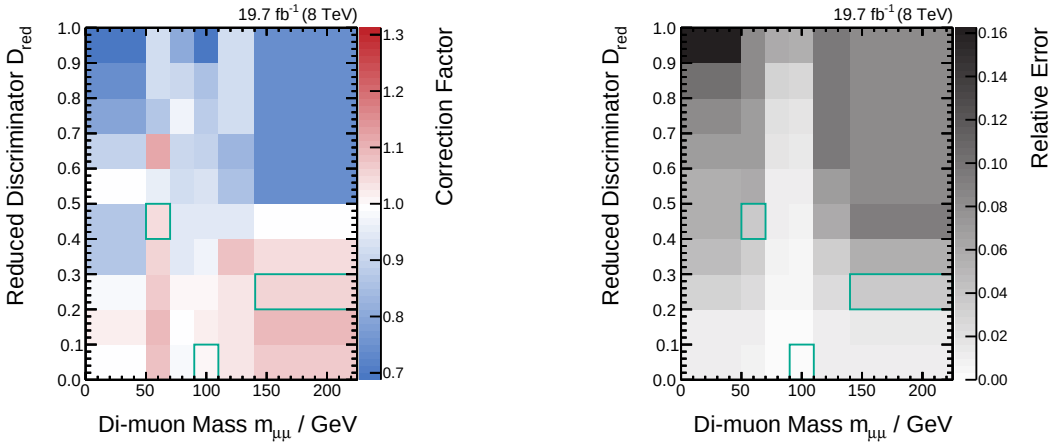


Figure 3.21.: Correction factors that are applied on the $Z \to \mu\mu$ background in bins of the di-muon mass $m_{\mu\mu}$ and the reduced discriminator D_{red}. The three marked bins correspond to the illustration in figure 3.20. The factors shown here are applied in the 1jet high pt category. A complete set of the correction factors can be found in the appendix in figure B.7

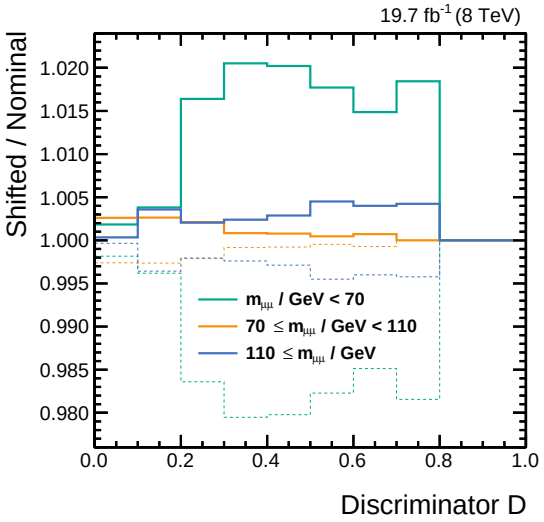


Figure 3.22.: Shape-altering uncertainties for assigned for the estimation of the $Z \to \mu\mu$ background in the 1-jet high-pt category. Ratios between the up and down variation and the nominal distribution are shown. The uncertainty is split into three different di-muon mass regions accounting for different precision of the fit results. A complete set of these shape uncertainties for all event categories is shown in figure B.8 for the 8 TeV analysis.

Systematic uncertainties are assigned both for the yields of the $Z \to \mu\mu$ background as well as for the shape of the discriminators. Figure 3.22 shows the shape-altering uncertainties compared to the nominal distribution. From the errors of the fit parameters as shown in figure 3.20 three different shape uncertainties are constructed accounting for variations in different di-muon mass regions, one for the lower tail of the Z mass peak, one for the region around the mass peak and one for the upper tail. The largest uncertainties are assigned to the signal-dominated regions where also the statistical precision of the $Z \to \ell\ell$ sample is limited. A complete set of these shape uncertainties for all event categories is shown in figure B.8 for the 8 TeV analysis.

It has to be pointed out that there is no simple mapping between the values of the two discriminators, D_{red} and D. Because of the fact that the reduced discriminator lacks important discriminating variables, its discrimination power is reduced compared to the final discriminator. Therefore, more $Z \to \mu\mu$ events results in values close to one, whereas the final discriminator does not contain any $Z \to \mu\mu$ events in the region above 0.8. This is the reason, why correction factors and uncertainties exist for values $D_{\text{red}} > 0.8$, but no final uncertainty can be assigned to events for values $D > 0.8$. Similarly, the estimation method is optimised for the shapes of the PDF integration discriminators. The agreement between data and the background-only expectation is acceptable. But this does not necessarily mean that this has to be true for all other distributions as well. As already pointed out, there are remaining disagreements in some of the BDT input variables.

3.6.2. The Embedding Technique

The second most dominant background of $Z \to \tau\tau$ events is estimated in a data-driven way, the embedding technique [91], which is applied in the same way in all $H \to \tau\tau$ channels. A sample of $Z \to \mu\mu$ events is selected in data. The two muons originating from the Z boson decays are removed from the event content measured in data. Then taus are generated to match the four-momenta of the muons. Their decays are modelled using TAUOLA [76]. Subsequently, the detector simulation and PF reconstruction is performed for the simulated taus. The resulting objects are inserted into the event content from the data and replace the reconstructed muons.

The result is a hybrid event. Apart from the simulated taus the rest of the reconstructed event remains from data, which is especially true for activity measured from the underlying event and from pile-up interactions. Thus, the reconstruction of the MET and jet energies is less affected by systematic uncertainties originating from the simulation. Scale factors are derived for correcting the inclusive yield of the embedded sample to the one from simulation. The final discriminator shapes as well as the categorisation efficiencies are taken from the embedded sample.

3.7. Systematic Uncertainties

Various uncertainties are assigned to the yield and shapes of the discriminators to account for imprecisely known quantities used in the analysis. The sources of systematic uncertainties range from theoretical uncertainties over the reconstruction and identification of physics objects to uncertainties

related to the event categorisation and background modelling methods. Uncertainties are taken as either fully correlated among multiple channels and categories or fully uncorrelated. In case an uncertainty accounts for an effect in several channels or categories, it is taken as a correlated uncertainty.

This section outlines the systematic uncertainties that are considered in the analysis of the di-muon channel. Small adaptations with respect to the preliminary analysis [83, 84] are needed in order to account for the new discriminator structure and the new modelling of the $Z \to \mu\mu$ background. Table 3.3 summarises all uncertainties considered in the di-muon channel.

Table 3.3.: Systematic uncertainties, affected samples, and change in acceptance resulting from a variation of the nuisance parameter equivalent to one standard deviation. Several systematic uncertainties are treated as (partially) correlated for different decay channels and/or categories.

Uncertainty	**Affected Processes**	**Change in Acceptance**
Muon ID & trigger	signal & sim. backgrounds	4 %
Jet energy scale	signal & sim. backgrounds	shape unc.
Jet b-tagging efficiency	$t\bar{t}$+jets	up to 5 %
MET scale (2-jet category)	signal & sim. backgrounds	shape unc.
Norm. $Z \to \mu\mu$	$Z \to \mu\mu$	0.1 - 2 %
Shape $Z \to \mu\mu$ (3 mass bins)	$Z \to \mu\mu$	shape unc.
Norm. $Z \to \tau\tau$	$Z \to \tau\tau$	3 %
$Z \to \tau\tau$ categories	$Z \to \tau\tau$	6 - 9 %
Norm. $t\bar{t}$+jets	$t\bar{t}$+jets	8 - 10 %
Norm. diboson	di-boson	3 %
Norm. QCD multi-jet	QCD multi-jet	9 - 100 %
Luminosity	signal & sim. backgrounds	2.6 %
PDF (qqH)	qqH signal	3.6 %
PDF (qqH, 2-jet category)	qqH signal	shape unc.
PDF (ggH)	ggH signal	9.7 %
PDF (ggH, 2-jet category)	ggH signal	shape unc.
PDF (VH)	VH signal	1 - 4 %
Scale variation (2-jet category)	signal	shape unc.
Scale variation	signal	0.8 - 18.2 %
Underlying event & parton shower	signal	0.4 - 8.9 %
Limited number of events	all	shape unc.

The uncertainty on the muon trigger, identification and isolation requirement is determined via a tag-and-probe method [83] and amounts to 2 % per muon.

The uncertainty on the jet energy as documented in the references [66, 67] has an impact both on the event yield als also on the shape of the final discriminator, as the jet multiplicity is one of the main ingredients for the event categorisation and the some di-jet variables are exploited in the discriminator of the 2-jet category. In this analysis the four-momentum is scaled up and down by 3 % and 7 % in the

barrel and the end-cap regions, respectively. The effect with respect to the nominal discriminator is shown in figure B.9 for all categories in the 8 TeV analysis.

Events containing b-tagged jets are vetoed in order to suppress the $\mathrm{t\bar{t}}+$jets background. Normalisation uncertainties of up to 5 $\%$ cover the varying b-tag efficiencies in the $\mathrm{t\bar{t}}+$jets background sample.

The MET scale uncertainty mainly affects the 2-jet category where the MET is directly taken as an input for the multivariate discriminator. The MET scale and its uncertainties for the different affected samples have been determined previously [83, 84]. The effect of MET scale shifts with respect to the nominal discriminator is shown in figure B.10 for the 2-jet category in the 8 TeV analysis. The effect of the MET scale uncertainty in the other categories is only due to the reconstructed invariant di-τ mass which is therefore covered by the uncertainties related to the background estimation methods.

The estimation of each background is covered by a dedicated normalisation uncertainty which is correlated among all event categories. The imperfections of the embedding technique are covered by additional uncertainties for the $\mathrm{Z} \to \tau\tau$ sample that are assumed to be partially correlated and partially uncorrelated among the event categories. The shape uncertainties for the $\mathrm{Z} \to \mu\mu$ sample are described in the previous section 3.6.1. Additionally, all simulated samples are allowed to vary within the correlated uncertainty on the integrated luminosity.

Theory predictions like cross section but also the simulation of events dependd on uncertainties related to the PDF, the renormalisation and factorisation scales as well as on effects due to finite-order calculations. Additionally, imprecisions result from the simulation of the parton showering and the underlying event. The normalisation uncertainties mentioned above also account for the cross section predictions. Dedicated studies for the theory-related uncertainties in the signal samples have been performed previously [92]. The variations of the PDFs and of the factorisation scale are propagated through the analysis and the resulting shape uncertainties as shown in figure B.11 are considered in the 2-jet category. Substantial contributions from the gluon fusion production in this category lead to sizeable uncertainties from the theory prediction. In all other categories normalisation uncertainties are considered. The same is true for the uncertainty related to the modelling of the underlying event and the parton shower.

The limited statistical precision of the templates for each sample are accounted for by bin-by-bin uncertainties that are uncorrelated among all analysis bins.

3.8. Statistical Inference and Results

Given are the distributions of the final discriminator (invariant di-τ mass in most of the channels) for the observation in the data as well as for the backgrounds and the predicted signal. The compatibility of the observation with the signal-plus-background and the background-only hypotheses are compared. In the absence of signal, exclusion limits on the signal production cross section are given. If there is an excess over the background-only hypothesis its significance is quantified.

3.8.1. Exclusion Limits and Signal Significances

This section describes in general the determination of exclusion limits and signal significances based on the CL_s [93,94] method that is commonly used for the Higgs boson searches in ATLAS and CMS [82].

The signal model predicts $s = \{s_i\}$ signal events in the phase space regions or bins i, whereas $b = \{b_i\}$ background events are expected. A global signal strength modifier μ is introduced scaling the signal expectation s across all bins. A value of $\mu = 0$ corresponds to the background-only hypothesis. The number of observed events in data is denoted by $n = \{n_i\}$. Systematic uncertainties are introduced via a set of nuisance parameters $\theta = \{\theta_j\}$. The following profile likelihood ratio q_μ is chosen to set upper exclusion limits on the signal strength μ.

$$q_\mu = -2 \ln \frac{\mathcal{L}(n \,|\, \mu, \hat{\theta}_\mu)}{\mathcal{L}(n \,|\, \hat{\mu}, \hat{\theta}_{\hat{\mu}})} \quad \text{with} \quad 0 \leqslant \hat{\mu} \leqslant \mu$$

where $\hat{\mu}$ and $\hat{\theta}_{\hat{\mu}}$ are the parameters that globally maximise the likelihood function and $\hat{\theta}_\mu$ denotes the values for the nuisance parameters that maximise the likelihood function for a fixed value of μ. The values of the signal strength parameter μ have to be non-negative since negative cross sections would be non-physical. The signal strength, μ, is constrained to be larger than $\hat{\mu}$ in order to obtain one-sided upper limits. The smaller the values of q_μ are, the more the observation favours the model μ over $\hat{\mu}$.

The event numbers in each bin are assumed to be Poisson-distributed. The likelihood function can be written in the following form.

$$\mathcal{L}(n \,|\, \mu, \theta) = \prod_i \frac{\left[\mu\, s_i(\theta) + b_i(\theta)\right]^{n_i}}{n_i!} \exp\left[-\mu\, s_i(\theta) - b_i(\theta)\right] \cdot \prod_j p(\theta_j \,|\, \tilde{\theta}_j) \tag{3.5}$$

Here $p(\theta_j \,|\, \tilde{\theta}_j)$ denotes the probability density function (pdf) for the nuisance parameter θ_j, whereas $\tilde{\theta}_j$ is the predicted value. Both the predicted value as well as the pdf are usually taken from auxiliary measurements or based on physics models or assumptions.

The CL_s value is defined as the ratio of two p-values, the one for the signal-plus-background hypothesis and the one for the background-only hypothesis.

$$\mathrm{CL}_s(\mu) = \frac{p_\mu}{1 - p_0} \quad \text{with} \quad p_\mu = \int_{q_\mu^{\mathrm{obs}}}^{\infty} \mathrm{d}q_\mu\, p(q_\mu \,|\, \mu, \hat{\theta}_\mu) \quad \text{and} \quad 1 - p_0 = \int_{q_0^{\mathrm{obs}}}^{\infty} \mathrm{d}q_0\, p(q_0 \,|\, 0, \hat{\theta}_0)$$

Again, $p(q_\mu \,|\, \mu, \hat{\theta}_\mu)$ denotes the pdf for the test statistic values for a given value of μ. Signal strengths larger than the tested value of μ are excluded with a CL_s confidence level of $1 - \mathrm{CL}_s(\mu)$. The relation $\mathrm{CL}_s(\mu)$ is usually inverted in order to state exclusion limits for μ at a chosen confidence level. A common choice are 95 % confidence levels for exclusion limits.

In case the signal is not excluded but the analysis is expected to have sensitivity for the model in question, the significance of a possible excess is quantified. For this purpose, the compatibility of the observation with the background-only hypothesis ($\mu = 0$) is assessed. The signal significance can be understood as a measure for the level of fluctuations of the background that would reproduce the

observed excess. The test statistic q_0 distributed according to the pdf $p(q_0 \mid 0, \hat{\theta}_0)$ is used.

$$q_0 = -2 \ln \frac{\mathcal{L}(n \mid 0, \hat{\theta}_0)}{\mathcal{L}(n \mid \hat{\mu}, \hat{\theta}_{\hat{\mu}})} \quad \text{with} \quad 0 \leqslant \hat{\mu}$$

The larger the values of the test statistic, the more the background-only hypothesis is disfavoured by the observation.

Then the p-value p_0 is determined giving the probability that larger values of the test statistic than the observed one are measured.

$$p_0 = P(q_0^{\text{obs}} \leqslant q_0) = \int_{q_0^{\text{obs}}}^{\infty} \mathrm{d}q_0 \, p(q_0 \mid 0, \hat{\theta}_0)$$

This p-value can be transformed into a significance in units of the width σ of a one-sided Gaussian distribution.

The exact pdf $p(q_\mu \mid \mu, \hat{\theta}_\mu)$ is not known analytically and its determination based on generated toy data sets is a computationally expensive procedure. A very good approximation is given by an asymptotic formula [95] which is used for all results presented in this thesis.

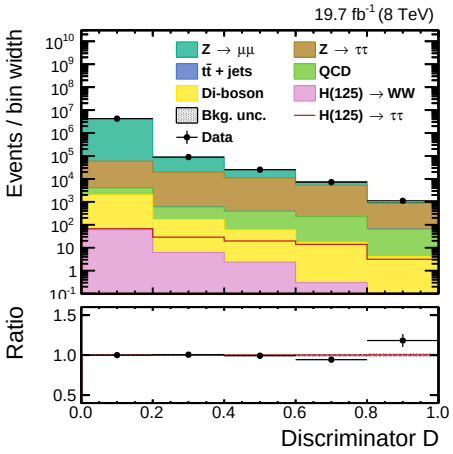
Exclusion limits and p-values are not only determined based on the actual observation but they are also compared with expectations based on different hypotheses for the signal strength, μ. In order to measure the expected exclusion limit or p-value, the observation, n, in the formulas above is replaced by toys that follow the expectation for the hypothesis in question. For each generated toy, the limit or p-value is calculated. The median of multiple of these results is quoted together with the one and two σ uncertainties. The number of toys needed is chosen such that the statistical errors on the median and the uncertainties are negligible. In practice, this toy-based approach is again replaced by using the asymptotic formula.

The expected exclusion limits are based on toys that follow the background-only ($\mu = 0$) hypothesis and therefore quantify the sensitivity of the analysis. For a given confidence level, the expected limit on μ gives the upper limit on the signal strength the analysis could exclude in case there is no signal. The lower the expected limit is, the higher is the sensitivity of the analysis.

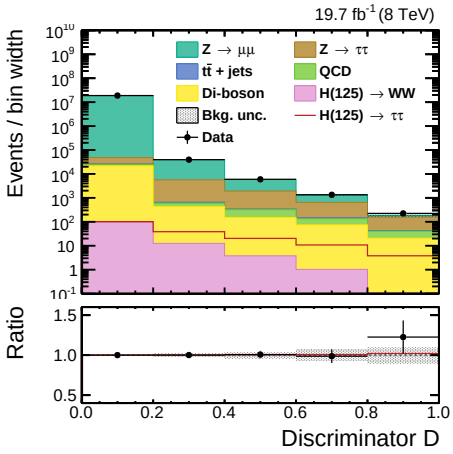
Expected p-values are based on toys that follow the signal-plus-background hypothesis ($\mu = 1$) and therefore help digesting the compatibility of an observed excess with the expected one.

3.9. Results

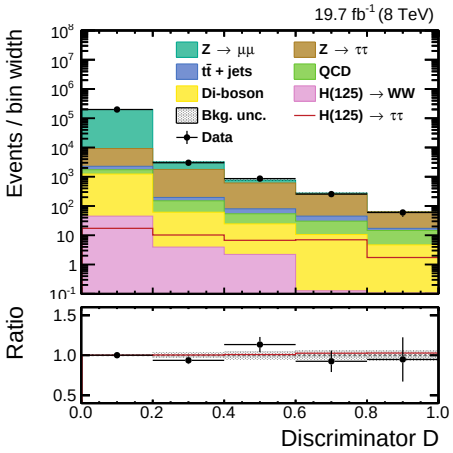
The global fit, as described before, is performed based on the final discriminator, D. Figure 3.23 shows the post-fit distributions of the discriminator for the 8 TeV analysis for all five event categories. The respective plots for the 7 TeV analysis are shown in the appendix in figure B.12. For these plots the values of the nuisance parameters are set to the values resulting from the likelihood fit. The error band contains the total post-fit uncertainty. The SM H $\to \tau\tau$ signal is shown for a mass hypothesis of



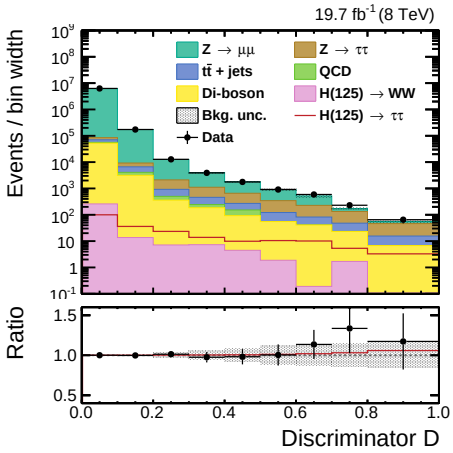
(a) 0-jet low-pt category



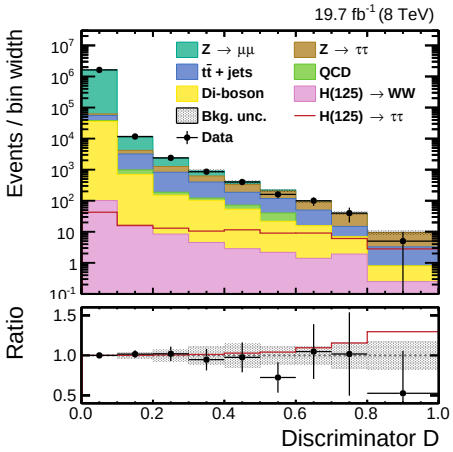
(b) 0-jet high-pt category



(c) 1-jet low-pt category



(d) 1-jet high-pt category



(e) 2-jet category

Figure 3.23.: Post-fit distributions of the final discriminator D in the five event categories for 8 TeV data. The ratio compares the observation in data to the expectation given by both the background-only and the signal-plus-background for a Higgs boson mass of 125 GeV hypotheses. The total post-fit errors are shown in the error band. The observation is compatible with both hypotheses in most of the bins.

125 GeV. The observation is compatible with both the background-only and the signal-plus-background hypothesis in most of the bins. This already shows that the analysis is not sensitive enough to find or exclude the SM Higgs boson signal. Therefore exclusion limits on the signal strength are quoted below.

The binning of the distributions that went into the global fit is choosen coarser than then one of the pre-fit plots shown in figure 3.19, where the focus was on a qualitative depiction of the discriminator shapes. The coarser binning here accounts for the level of agreement between data and the background-only expectation and for the population of bins in the low-statistics samples, which is especially important for the $W+$jets and QCD multi-jet samples. The chosen binning ensures that the result is robust against statistical fluctuations in the templates of all samples.

The $H \to \tau\tau$ analysis is also sensitive to Higgs boson decays into W bosons where the W bosons pairs dominantly decay into pairs of τ leptons and two additional neutrinos. However, the contributions from the $H \to WW$ signal are small in the di-muon channel. The $H \to WW$ process with a Higgs boson mass hypothesis of 125 GeV as indicated by recent measurements [29] is considered as a background in order to be able to retrieve the signal strength for $H \to \tau\tau$ decays independently of the presence of these $H \to WW$ events. The sample is included in the postfit plots and shows a yield that is smaller than the one of the $H \to \tau\tau$ signal at a Higgs boson mass of 125 GeV, proving that the analysis is more sensitive to the selection of $H \to \tau\tau$ events compared to $H \to WW$ events.

Expected and observed exclusion limits are shown in figure 3.24 separately for the 8 TeV analysis (left) and for the combination of the 7 and 8 TeV analyses (right). The sensitivity of the analysis quantified by the expected limit based on the background-only hypothesis reaches down to signal strengths of 2.96 times the SM cross section times branching ratio at $m_H = 125$ GeV. The sensitivity decreases both towards lower and higher Higgs boson mass hypotheses and has its minimum between 120 and 130 GeV. For lower masses the overlap with the Z peak deteriorates the separation between signal and background events and for higher masses the branching ratio for $H \to \tau\tau$ decays decreases rapidly since new decay channels open up. The observation is compatible with the background-only hypothesis within the 1 σ band. No sensitivity to the SM Higgs boson is expected, also no excess is seen in data. Nevertheless, the result is compatible with the combined CMS result from all $H \to \tau\tau$ channels as it will be shown in section 4.2 and especially in figure 4.6.

Figure 3.25 (left) shows the sensitivity of the different event categories. The expected limits are shown for the zero, one and two jet categories separately for the 8 TeV analysis. The effect already visible in the discriminator shapes manifests itself here quantitatively: The higher the jet multiplicity in the event is, the more increased is the sensitivity of the corresponding category. Background-related nuisance parameters are constrained in the 0-jet category. 1-jet events with their boosted topology profit from a better MET and mass resolution. The background contamination is comparatively low in the 2-jet category that is optimised for the analysis of the VBF Higgs boson production mechanism.

The right plot of figure 3.25 shows the effect of combining the 8 TeV data with the 7 TeV one. Due to limited statistical precision in the 7 TeV samples, fluctuations of the limit as a function of the Higgs boson mass are larger than in the 8 TeV analysis. Overall the 7 TeV data amount up to about 10 % to the combined sensitivity. The 8 TeV data are significantly more sensitive to the signal due to the larger signal cross sections and because of the data set which is by a factor of four larger than the 7 TeV one.

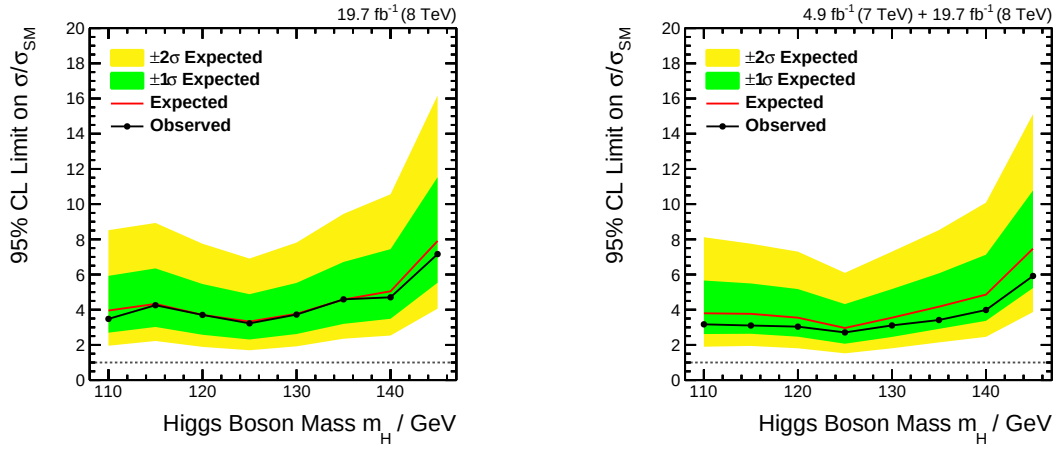


Figure 3.24.: Expected and observed limits for the 8 TeV analysis (left) and the combination of the 7 and 8 TeV analyses (right). The analysis is sensitive to signal strengths of values down to 2.96 times the SM cross section times branching ratio at $m_{\mathrm{H}} = 125$ GeV. No excess is seen, the observation follows the background-only expectation within the 1 σ band in the entire mass range.

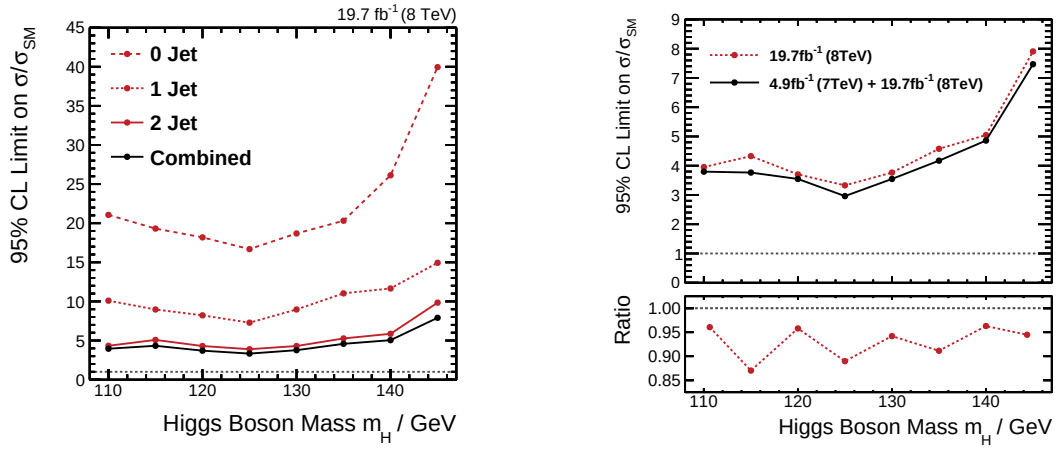


Figure 3.25.: Expected limits split into categories according to the jet multiplicity for the 8 TeV analysis (left) and the combination of the 7 and 8 TeV analyses (right). The higher the jet multiplicity is, the higher is the sensitivity of the analysis. 2-jet events contribute most to the combined sensitivity. Events without jets help constraining the background-related systematic uncertainties. The 7 TeV analysis amounts up to about 10 % to the combined sensitivity.

Based on the size of the data set by a factor of roughly 1.25, the exclusion limit is expected to scale by $\sqrt{(1.25)} = 1.12$. Therefore, it can be claimed that the increase in sensitivity is compatible with the expectation from the two points mentioned.

The improvement of the new signal extraction method based on the discriminator, D, compared the preliminary method is quantified in figure 3.26. The expected limit is improved by a factor of about 20 $\%$ in the Higgs boson mass range between 125 and 135 GeV. The CL_s limit is approximately proportional to the reciprocal square root of the integrated luminosity of the data set. Therefore an improvement in the limit of 20 $\%$ is compatible with the increased sensitivity based on a hypothetical data set with an integrated luminosity that is increased by 56 $\%$. Towards higher Higgs boson mass hypothesis the improvement decreases. The reason is the increased separation power of the mass variables on its own.

Finally, the performance of the analysis presented in this chapter is put in the context of the complete $H \to \tau\tau$ analysis. The $\mu\mu$ channel is compared with the other di-lepton channels, ee and $\mathrm{e}\mu$. The sum of the branching ratios of the same-flavour di-lepton channels, $\mu\mu$ and ee, is equal to the branching ratio of the $\mathrm{e}\mu$ channel because of lepton universality of the weak interaction. A comparison of the di-lepton limits is shown in figure 3.27. The sensitivity of the $\mu\mu$ and ee channels separately is significantly worse than the one of the $\mathrm{e}\mu$ channel because of the additional $Z \to \ell\ell$ background (with $\ell\ell = \mu\mu, \mathrm{ee}$) and the factor of 2 in the branching ratio. The latter is eliminated by combining the same-flavour di-lepton channels as shown in the dashed line. By comparing the combination of all di-lepton channels with the $\mathrm{e}\mu$ channel, it can be inferred that the $\mu\mu$ and ee channels add approximately 20 $\%$ sensitivity which again corresponds to a gain in sensitivity that could be expected from a hypothetical data set with an integrated luminosity increased by 56 $\%$ compared to the analysed one. The ratio of the combined limit over the $\mathrm{e}\mu$ limit and the combination of the $\mu\mu$ and ee channels is shown below to illustrate the contribution of the same-flavour and different-flavour events to the combined performance.

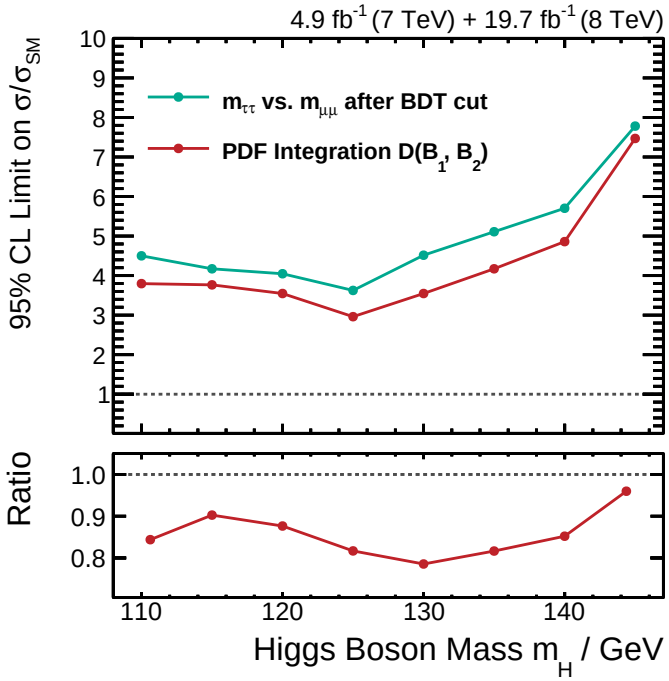


Figure 3.26.: Comparison of the sensitivity of the preliminary method, where the final 2D mass discriminator was fit after a cut on a BDT discriminator, with the published one, where the PDF integration method has been used to construct the final discriminator, based on the expected limits for the background only hypothesis. At the Higgs boson mass hypothesis of 125 GeV the improvement is almost 18 %.

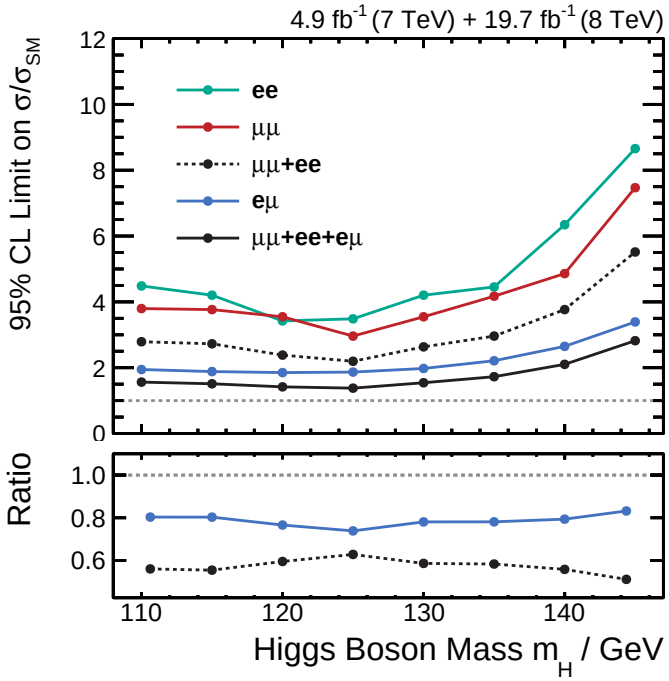


Figure 3.27.: Comparison of the expected limits for the di-lepton channels (ee, $\mu\mu$ and eμ) and combinations of them. The ratio of the same-flavour di-lepton combination and the eμ limit over the combined limit ($\mu\mu$ + ee + eμ) is shown at the bottom. By the help of the same-flavour di-lepton the full di-lepton combination yields a limit that is better by about 20 % then the single eμ channel result.

3.10. Summary

In the first part of this thesis the search for Higgs bosons decaying into pairs of τ leptons has been presented. The focus was on the analysis in the di-muon channel. Despite all the peculiarities and difficulties in the channel it was possible to contribute to the main result of finding evidence for $H \to \tau\tau$ decays with a significance of more than 3 σ.

The performance of the analysis of the di-muon channel has been driven by the new approach of combining two BDT discriminators into an one-dimensional final discriminator. This method has then also been applied in the analysis of the di-electron channel. The sensitivity in this channel alone required the determination of upper limits on the signal cross section times branching ratio. No excess in data has been seen. The results are in agreement with both the background-only and the SM expectation including the Higgs boson with a mass of 125 GeV.

The performance of this channel has been compared with the one of the other di-lepton channels that are characterised by similarly small branching fractions. It has been shown that the combination of the same-flavour channels contributes significantly performance in the combination of all di-lepton channels. This is a remarkable outcome considering the composition of the backgrounds. Whereas the same-flavour channels suffer from the additional and most dominant $Z \to \ell\ell$ background, these events are completely excluded from the search in the $e\mu$ channel.

4

Prospects of the $\mathrm{H} \to \tau\tau$ Analysis for the CMS Run II

The Higgs boson has first been observed in the bosonic decay channels $\mathrm{H} \to \mathrm{ZZ} \to 4\ell$ and $\mathrm{H} \to \gamma\gamma$, which are characterised by a good mass resolution, by the ATLAS [8] and the CMS experiment [9]. Together with the measurement of the signal strengths in these channels, this already provides evidence for Higgs boson couplings to vector bosons (via the $\mathrm{H} \to \mathrm{ZZ}$ decay) and top quarks (in the loop of the production via gluon fusion).

A major step for the identification of the observed particle being the Standard Model Higgs boson, responsible for the generation of the masses of all elementary particles, is the direct probing of Yukawa couplings to leptons. Here, the $\mathrm{H} \to \tau\tau$ channel is the most promising one. In contrast to $\mathrm{H} \to \mathrm{bb}$ decays, the di-τ final states are characterised by a cleaner signature in the detector. Furthermore, the high branching ratio of $\mathrm{H} \to \tau\tau$ decays in the low mass range (see figure 1.7) is an advantage of the search in this channel. Decays into top quarks do not play any role in the low mass range. At the same time, the di-τ channel is best suited to directly probe fermionic couplings of the Higgs boson. The branching ratio for decays into the next lighter leptons, muons, is suppressed by a factor $(m_\mu/m_\tau)^2 \approx 0.0035$, although this channel profits from a much better mass resolution.

After the search for the Higgs boson decaying into τ lepton pairs in the di-muon final state has been described in the previous chapter 3, this chapter puts it into the context of the complete $\mathrm{H} \to \tau\tau$ analysis which is published in reference [10]. Quantitative details are given in the analysis notes [96–103]. This chapter first summarises the combined results of the various different sub-channels of the $\mathrm{H} \to \tau\tau$ analysis based on the complete CMS run I data set collected at centre-of-mass energies of 7 and 8 TeV. The second part of this chapter gives an outlook of the $\mathrm{H} \to \tau\tau$ analysis in the main sub-channels as it is prepared for the data that are being collected at the time of finishing this thesis in the second CMS running period at centre-of-mass energies of 13 TeV. Basic principles of the analysis channels are explained and extrapolations of expected uncertainties for the upcoming measurements are presented.

4.1. Combination of all Analysis Sub-channels

The H $\rightarrow \tau\tau$ is split in multiple sub-analyses. There are six analyses for the six different decay modes of τ pairs: fully hadronic ($\tau_h\tau_h$), semi-leptonic ($\mu\tau_h$, $e\tau_h$) and fully leptonic ($e\mu$, $\mu\mu$, ee). Furthermore there are dedicated VH analyses that cover the Higgs boson production in association with vector bosons, where the W or Z boson decays leptonically (involving electrons or muons) and the Higgs boson is required to decay into pairs of τ leptons. These higher lepton multiplicities provide a handle to suppress the backgrounds arising from SM processes. The descriptions in this chapter focus on the six main channels.

According to the channel, each sub-analysis selects events, where two oppositely charged leptons of the specific flavour have been identified and reconstructed. Vetoes on events containing additional leptons, that are selected based on looser criteria, remove possible overlap with other analyses, mainly the one of the VH channels. The selection criteria define the composition of the background for each sub-analysis as depicted in figure 4.1. The overall number of events selected differs from channel to

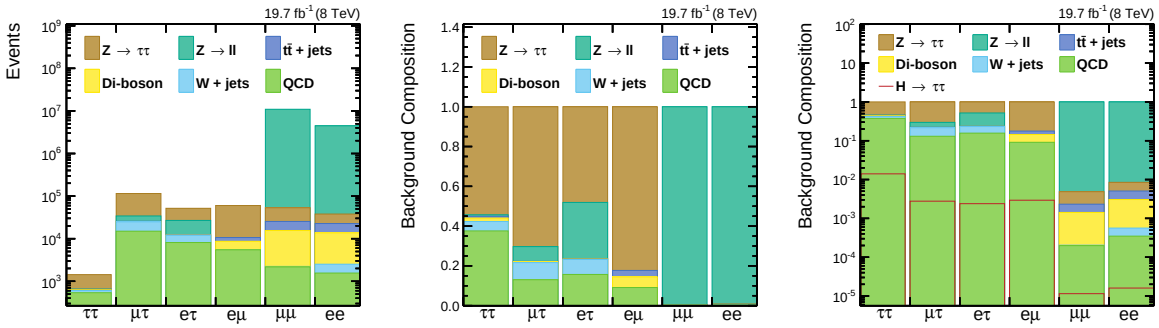


Figure 4.1.: Composition of the background in the six main channels of the 8 TeV H $\rightarrow \tau\tau$ analysis. The absolute numbers of events per channel in the inclusive selection (left) are compared with the version, where these numbers are normalised to the sum of all backgrounds per channel (centre and right). The logarithmic representation (right) also includes the fraction of H $\rightarrow \tau\tau$ signal with a mass hypothesis of 125 GeV compared to the sum of all backgrounds per channel. The characteristics of the background composition for each channel (as described in the text) are clearly visible.

channel. The $\tau_h\tau_h$ channel needs to apply tighter kinematic cuts on the two τ leptons in order to reduce the background from jets being misidentified as taus. Therefore this channel yields the smallest inclusive number of events. The same-flavour di-muon channels suffer from the additional Z $\rightarrow \ell\ell$ background ($\ell = e, \mu$) as described in chapter 3 which constitutes at least 99 % of the sum of all backgrounds. In all other channels, the Z $\rightarrow \tau\tau$ background is the most dominant one. Because of the very similar final state compared to the H $\rightarrow \tau\tau$ signal which only differs in the mass of the resonance, this background is the most difficult one to suppress. The more hadronically decaying τ leptons a sub-analysis uses, the larger the background from QCD multi-jet events is, which amounts to 38 % in the $\tau_h\tau_h$ channel. The W+jets background deserves a special treatment in the semi-leptonic decay modes as well as the $t\bar{t}$+jets background in the $e\mu$ channel. Here, these backgrounds are suppresses by additional requirements that are described in the second part of this chapter. The signal to background ratio also depends on the channel. Not only the composition of the backgrounds has an effect here, but also the branching ratios for the decays of τ leptons (see section 1.2.2). Here, the $\tau_h\tau_h$ channel profits from the large branching

ratio of 64.8 % for hadronic τ decays. The semi-leptonic decay modes provide a good compromise between the signal rate and the possibilities to suppress the backgrounds.

A second important characteristic differing between the six main channels is the resolution of mass variables. The primary choice for the mass definition is the reconstructed invariant mass of the di-τ pair as described in section 3.3.2 and documented in reference [10]. This method tries to reconstruct the most probable configuration for the four-momentum of the system of the two to four neutrinos in the final state depending on the channel as a function of the measured missing transverse energy and the momenta of the visible τ decay products. Figure 4.2 compares the resolution of the mass of the visible decay products with the reconstructed invariant di-τ mass of a simulated H $\rightarrow \tau\tau$ sample with a mass hypothesis of 125 GeV for three decay channels that differ in the number of neutrinos in the final state. Two effects are visible. Firstly, the mass resolution is improved significantly with respect to the mass of

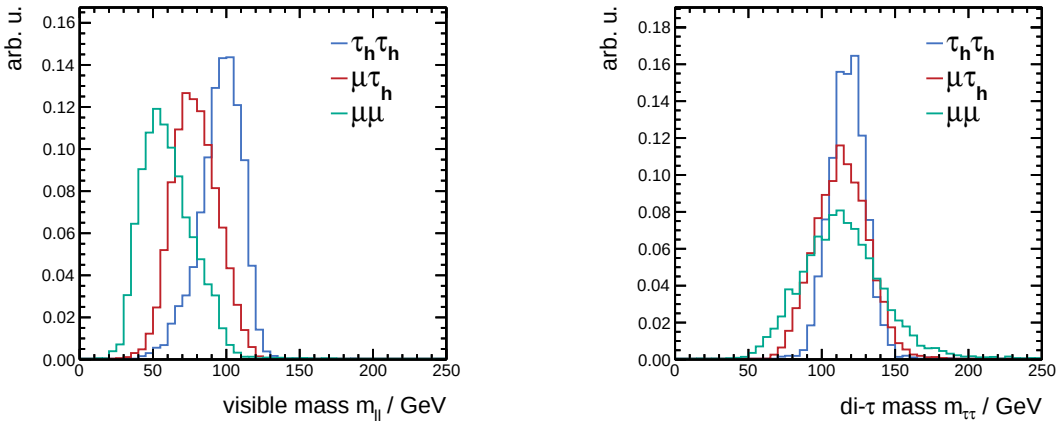


Figure 4.2.: Comparison of the mass of the visible decay products (left) and the reconstructed invariant di-τ mass (right) of a simulated H $\rightarrow \tau\tau$ sample with a mass hypothesis of 125 GeV. The mass reconstruction improves both the mean and the width of the mass peak. The resolution depends on the number of neutrinos in the final state.

the visible decay product that neglects the neutrino momenta. In addition, the mean is shifted towards the nominal mass of the resonance. Secondly, it is also noticeable that the resolution improves with decreasing numbers of neutrinos in the final state. The $\tau_h\tau_h$ channel provides the best resolution of the reconstructed invariant di-τ pair mass. The effect of the missing momentum due to the number of neutrinos is also seen in the distribution of the visible mass. The underestimation of the true mass is more pronounced in the $\mu\mu$ channel than in the $\tau_h\tau_h$ channel.

Neural network regression techniques have been studied for the mass reconstruction [89] and can serve as cross checks for the method above which is used for nearly all channels of the H $\rightarrow \tau\tau$ analysis. Neural networks have been trained to estimated the true mass of the resonance and show a similar performance in terms of resolution compared to the likelihood based method. The disadvantage of a missing physics motivation or understanding of this method is contrasted by the fact that the computation time is decreased by orders of magnitude compared to the cost of the multidimensional integration of the likelihood function.

Each analysis improves its sensitivity by defining an event categorisation. Some categories show an improved signal-to-background ratio and others are dominated by background events and therefore help constraining the background-related systematic uncertainties. The definition of categories depends on the individual channels, their event topology, the background composition and the size of the selected data set. In common, they share categories based on the jet multiplicity and the transverse momentum of one reconstructed lepton, as it has already been described for the $\mu\mu$ channel in chapter 3. On top of that, some sub-analyses apply further selection criteria and define even tighter categories in order to further improve the sensitivity of the analysis.

4.2. The CMS Result of the Run I Data Set

The published result combines all analysis channels and event categories in a single global fit. The parameter of interest is the signal strength μ which is a measure of the Higgs boson production cross section time decay branching fraction in units of the SM expectation. Usually it is abbreviated to $\mu = \sigma/\sigma_{\mathrm{SM}}$. This factor defines a common scaling for all considered Higgs boson production mechanisms. The fit is performed separately for each Higgs boson mass hypothesis. Contributions from the $\mathrm{H} \rightarrow \mathrm{WW}$ production with a Higgs boson mass of 125 GeV as seen in the $\mathrm{H} \rightarrow \mathrm{WW}$ analysis [29] are considered as background process.

Figure 4.3 shows the exclusion limit for the observation in data and the expectation derived from generated pseudo-data. Although the analysis has the sensitivity to exclude possible Higgs bosons with production cross sections smaller than the one predicted by the standard model, σ_{SM}, in the full studied mass range from 90 to 145 GeV as the expected limit shows, the standard model Higgs boson cannot be excluded above 100 GeV[1]. An excess over the background-only expectation is clearly visible

Therefore, the significance of the excess is quantified as shown in figure 4.4. An excess with significances larger than 3 σ is seen for Higgs boson masses between 115 and 130 GeV which is compatible with the SM expectation. The reason for this broad peak is the limited mass resolution of the di-τ mass caused by the neutrinos in the final state. Figure 4.5 shows the di-τ mass peak for the four most significant decay channels ($\mu\tau_{\mathrm{h}}$, $\mathrm{e}\tau_{\mathrm{h}}$, $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ and $\mathrm{e}\mu$). After weighting the contributions from the individual event categories according to the $S/(S+B)$ ratio in the given category, the signal excess can already be seen[2]. The inlay plot shows compares the signal expectation with the data, where the predicted contributions from all backgrounds have been subtracted.

Finally, figure 4.6 shows the best fit signal strength μ for the combination of all channels as well as for the individual channels separately. In the latter case, the data in these channels is also fit separately for single channels. The combined result of $\mu = 0.78 \pm 0.27$ is compatible with the SM expectation of $\mu = 1$. The split into individual channels indicates that the signal is seen consistently in all channels within their uncertainties. The error bars are a measure for the sensitivity of the channels. The $\mu\tau_{\mathrm{h}}$ and $\mathrm{e}\tau_{\mathrm{h}}$ provide the strongest constraints on μ.

[1] The lower sensitivity below 100 GeV is caused by the strong overlap ot the mass peaks of the Z and the H boson at this mass hypotheses

[2] The significance of the excess cannot be extracted from this display option.

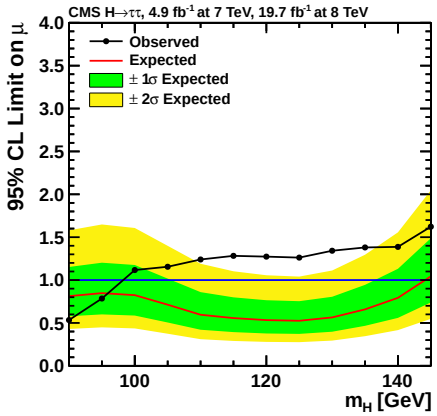


Figure 4.3.: Exclusion limits from the combination of all analysis channels and event categories [10]. The expected limit based on the background-only hypothesis shows, that the analysis is sensitive to the Standard Model Higgs boson in the entire studied mass range. The observation shows an excess over the background-only expectation with a significance of more than 2 σ from 135 to 135 GeV.

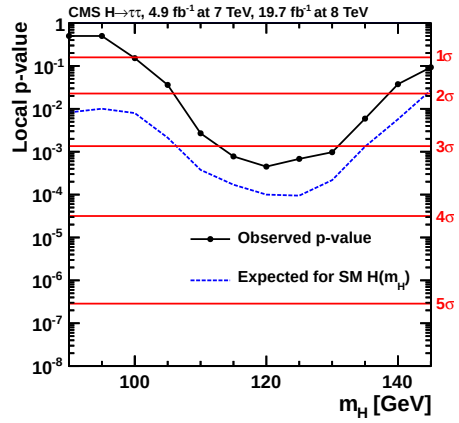


Figure 4.4.: Significances and p-values from the combination of all analysis channels and event categories [10]. The observed excess at 125 GeV has a significance of 3.2 σ which is compared to the 3.7 σ of the expected excess.

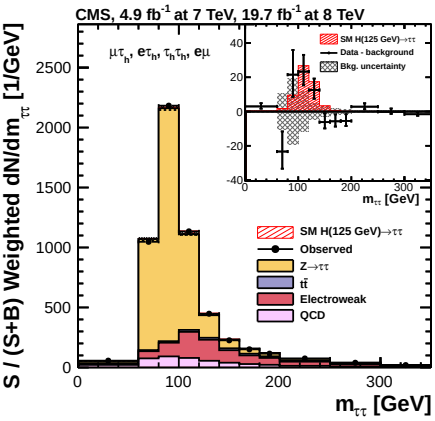


Figure 4.5.: The di-τ mass distribution in the four main channels after weighting the contributions from the single categories according to their ratio $S/(S+B)$ shows the signal excess. [10].

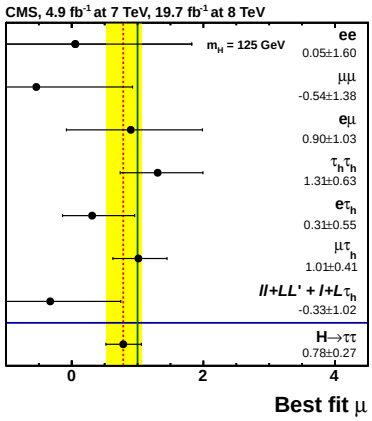


Figure 4.6.: Best fit values for the signal strength parameter, $\mu = \sigma/\sigma_{\mathrm{SM}}$ for all individual sub-analysis and the combined $\mathrm{H} \to \tau\tau$ analysis [10]. All channels agree within their uncertainties with the combined result.

Further measurements have been performed and can be found documented in reference [10]. The are all compatible with the Standard Model expectation within 2 σ standard deviation. Their relevance will increase in the CMS run II data-taking in the following part of this chapter focussing on the future analysis.

4.3. Data Sets and Simulation at 13 TeV

During the upgrade phase between the first and the second data-taking periods in the years between 2013 and 2015, both the accelerator and the detectors have been modified (see section 2.3.2). As a consequence, the LHC is now able to provide proton-proton collisions with increased instantaneous luminosity, which is mainly achieved by performing collisions with 25 ns bunch spacing, at an increased centre-of-mass energy of 13 TeV. However, in the start-up phase of the run II the LHC was operated at 50 ns bunch spacing as in the 8 TeV run. The first 71.52 pb^{-1} of data have been recorded and certified for this setting. The outlook to the run II data presented in this chapter concentrates on this first data set taken at $\sqrt{s} = 13$ TeV.

The increased centre-of-mass energy leads to larger parton luminosities as shown in figure 4.7, caused the larger momentum the partons of the protons carry on average. This is the main reason for increased production cross sections at 13 TeV compared to 8 TeV. The two effects, the higher parton luminosity and the increased instantaneous luminosity lead to larger production rates for SM Higgs bosons. Moreover, the reach in sensitivity for physics beyond the Standard Model is raised as hypothetical new particles with higher masses can be produced. This sets the scope for future H $\rightarrow \tau\tau$ analyses. In the field of the SM Higgs boson analysis, measurements of couplings and other properties of the newly found boson can be performed with higher accuracies. As many natural extensions to the Standard Model, such as Two-Higgs-Doublet models, predict enhanced couplings of additional Higgs bosons to down-type fermions, the di-τ final state is a very important channel for probing new physics in the Higgs sector. The latter is not discussed in the scope of this thesis. However, the appendix A outlines technical developments as a preparation for covering multiple follow-up analyses.

Table 4.1 lists the simulated samples, its generators, the numbers of generated events and the cross sections used in the new SM H $\rightarrow \tau\tau$ analysis for the run II data-taking period. Again, as for the 7 and 8 TeV analyses, the parton shower and underlying event is generated by PYTHIA and the event simulation is completed by the full detector simulation based on GEANT 4. It has to be noted that these samples are generated for the 25 ns bunch crossing scenario, as this is the plan for the largest parts of the collisions provided by the LHC in 2015. Discrepancies between the data and the simulation presented below might at least partially be caused by this inconsistency. However, as already more data with 25 ns bunch spacing are recorded than for the 50 ns setting, future analyses won't be based on these very first 13 TeV collisions and the agreement between data and the simulation is investigated in detail in the upcoming studies.

At the time of writing this thesis, the simulation of events was still ongoing and an even larger production campaign that could serve for the analysis of the larger expected data sets was in planning.

Table 4.1.: Summary of the simulated samples and their cross sections [23, 104] used in the 13 TeV SM H $\to \tau\tau$ analysis. The signal cross sections are given for a Higgs boson mass hypothesis of $m_{\mathrm{H}} = 125$ GeV.

Process	Generator	Number of Events	Cross Section / pb
$(\mathrm{W} \to \ell\nu) + \mathrm{jets}$	aMC@NLO	24 151 270	61526.7
$(\mathrm{Z}/\gamma^* \to \ell\ell) + \mathrm{jets}$ $(10 < m_{\ell\ell}/\mathrm{GeV} < 50)$	aMC@NLO	30 535 559	18610.0
$(\mathrm{Z}/\gamma^* \to \ell\ell) + \mathrm{jets}$ $(m_{\ell\ell} > 50$ GeV$)$	aMC@NLO	28 825 132	6025.0
$\mathrm{t\bar{t}}$+jets	POWHEG	19 899 500	831.76
$(\mathrm{WW} \to qq'qq') + \mathrm{jets}$	POWHEG	1 995 200	45.20
$(\mathrm{WW} \to \ell\nu qq') + \mathrm{jets}$	POWHEG	1 995 200	43.53
$(\mathrm{WW} \to \ell\nu\ell\nu) + \mathrm{jets}$	POWHEG	1 930 000	10.481
$(\mathrm{WZ} \to \ell\nu q\bar{q}) + \mathrm{jets}$	aMC@NLO	24 711 046	10.96
$(\mathrm{WZ} \to qq'\ell\ell) + \mathrm{jets}$	aMC@NLO	31 054 519	6.415
$(\mathrm{WZ} \to \ell\nu\ell\ell) + \mathrm{jets}$	POWHEG	1 925 000	4.42965
$(\mathrm{ZZ} \to q\bar{q}q\bar{q}) + \mathrm{jets}$	aMC@NLO	35 817 626	6.956
$(\mathrm{ZZ} \to q\bar{q}\nu\nu) + \mathrm{jets}$	aMC@NLO	36 559 813	4.301
$(\mathrm{ZZ} \to \ell\ell q\bar{q}) + \mathrm{jets}$	aMC@NLO	18 898 680	3.135
$(\mathrm{ZZ} \to \ell\ell\ell\ell) + \mathrm{jets}$	aMC@NLO	10 333 043	1.256
$(\mathrm{ZZ} \to \ell\ell\nu\nu) + \mathrm{jets}$	POWHEG	6 652 512	0.5644
SM $\mathrm{gg}(\mathrm{H} \to \tau\tau)$	POWHEG	1 497 400	2.775744
SM $qq(\mathrm{H} \to \tau\tau)$	POWHEG	1 478 412	0.2368736
SM $\mathrm{W}^+(\mathrm{H} \to \tau\tau)$	POWHEG	440 704	0.043608
SM $\mathrm{W}^-(\mathrm{H} \to \tau\tau)$	POWHEG	443 200	0.043608
SM $\mathrm{Z}(\mathrm{H} \to \tau\tau)$	POWHEG	585 353	0.05495872

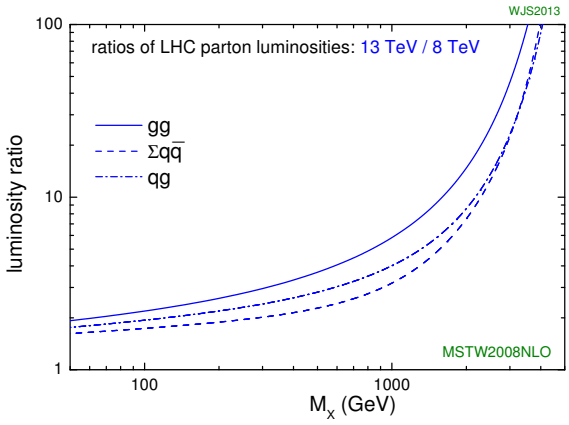


Figure 4.7.: For a resonance of the mass of 125 GeV, the parton luminosities at 13 TeV are increased by a factor of approximately two with respect to 8 TeV [105].

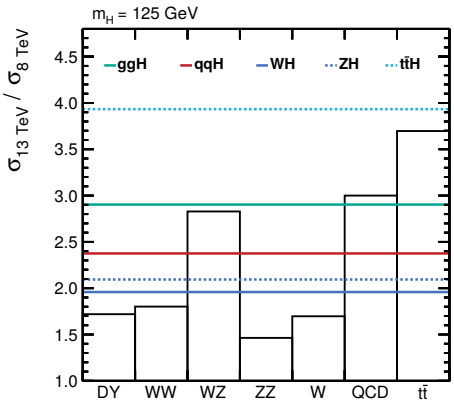


Figure 4.8.: Comparison of cross sections for background and signal for 8 TeV and 13 TeV [21–23, 88, 104]. The signal production rate increases more than the one for the most dominant backgrounds.

These preliminary samples do not have the necessary size to guarantee a small statistical uncertainty related to the simulation. This will be pointed out in the studies presented below.

In figure 4.8 ratios of cross sections at $\sqrt{s} = 13$ TeV over ones at 8 TeV are shown for the backgrounds and compared with the ones for the $\mathrm{H} \rightarrow \tau\tau$ signal. It is noticeable that the increase in production ratio for the signal is larger than the one for most of the background processes, especially the most dominant ones like Drell-Yan events. The signal cross sections of dominant processes increase by factors of two to three. On the other hand, the increase in luminosity leads to an increase in pile-up activity, which has to be studied in detail as soon as enough data are available.

4.4. Event Selection and Modelling of the Backgrounds

The analysis is performed in the four most sensitive channels: $\tau_\mathrm{h}\tau_\mathrm{h}$, $\mu\tau_\mathrm{h}$, $\mathrm{e}\tau_\mathrm{h}$ and $\mathrm{e}\mu$. For each channel a separate sub-analysis is performed although all channels share common parts. Events considered in the analyses have to be selected by high level triggers based on one or two leptons, according to the channel. They are listed in table 4.2. Apart from the $\tau_\mathrm{h}\tau_\mathrm{h}$ channel, either of two trigger decisions is considered. The reconstructed leptons are required to match with the trigger objects within $\Delta R < 0.5$.

Table 4.2.: High level triggers required for the events selection in the four main channels. The numbers after the lepton legs indicate the trigger thresholds in GeV. In the $\mathrm{e}\tau_\mathrm{h}$ channel, the names differ for data and simulation (in parentheses).

Channel	High Level Trigger
$\boldsymbol{\tau_\mathrm{h}\tau_\mathrm{h}}$	HLT_DoubleMediumIsoPFTau40_Trk1_eta2p1_Reg
$\boldsymbol{\mu\tau_\mathrm{h}}$	HLT_IsoMu24_eta2p1 HLT_IsoMu17_eta2p1_LooseIsoPFTau20
$\boldsymbol{\mathrm{e}\tau_\mathrm{h}}$	HLT_Ele32_eta2p1_WPTight_Gsf (HLT_Ele32_eta2p1_WP75_Gsf) HLT_Ele22_eta2p1_WPLoose_Gsf_LooseIsoPFTau20 (HLT_Ele22_eta2p1_WP75_Gsf_LooseIsoPFTau20)
$\boldsymbol{\mathrm{e}\mu}$	HLT_Mu8_TrkIsoVVL_Ele23_CaloIdL_TrackIdL_IsoVL HLT_Mu23_TrkIsoVVL_Ele12_CaloIdL_TrackIdL_IsoVL

The leptons and other physics objects are reconstructed and identified with the CMS algorithms based on Particle Flow as described in section 2.4. Electrons are required to pass the tight MVA-based identification yielding an efficiency of 80 %. Muons are selected based on a medium working point. For taus it is required that a valid decay mode is reconstructed by the HPS algorithm. In the $\mu\tau_\mathrm{h}$ channel the a tight MVA-based veto against muons being misidentified as τ jets and a medium MVA-based veto against such electrons is applied. In the $\mathrm{e}\tau_\mathrm{h}$ channels the tight and medium vetos are applied vice-versa. Additionally, events with third leptons (electrons or muons) are vetoed based on looser identification criteria.

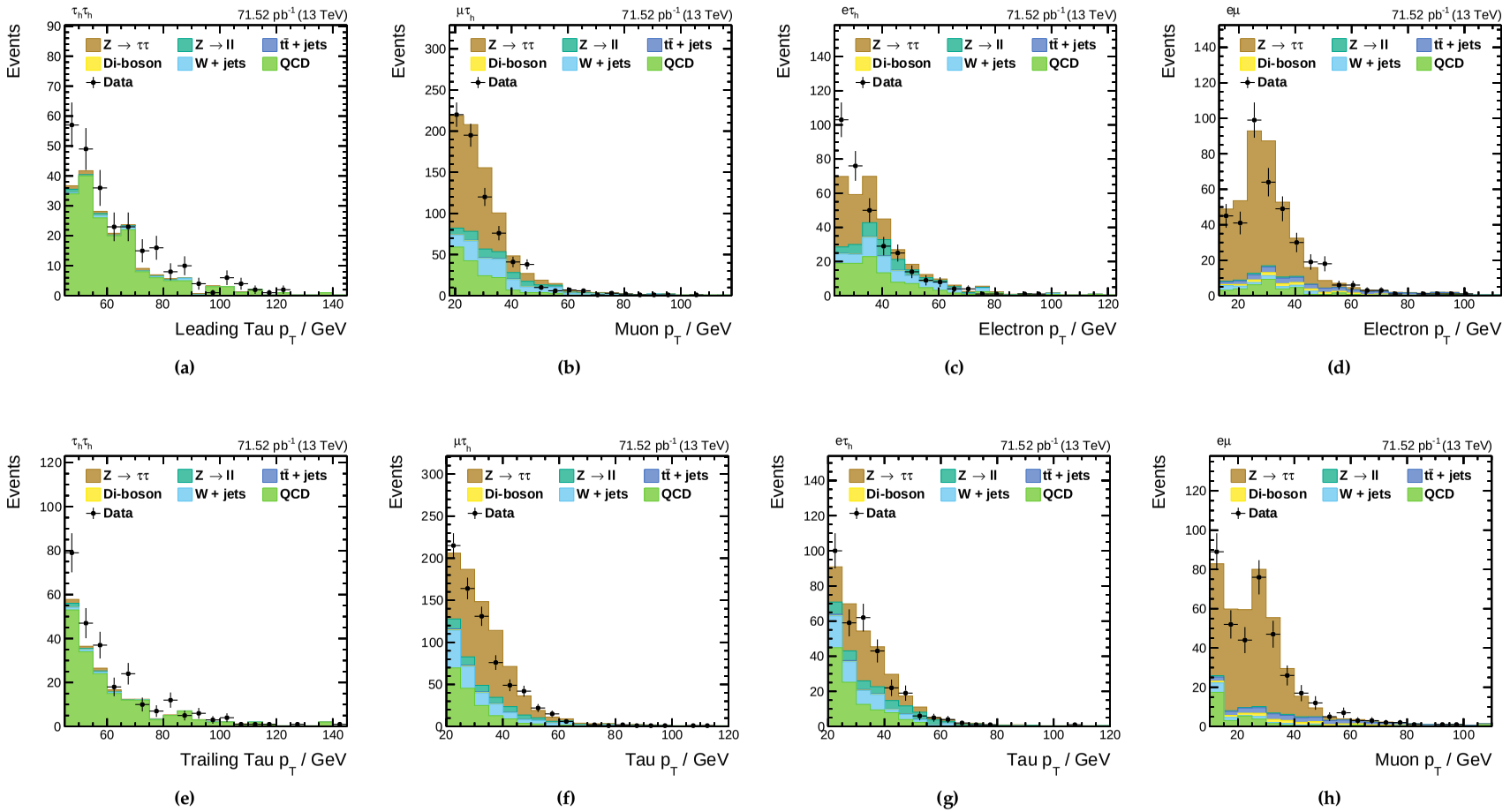


Figure 4.9.: Transverse momenta of the two leptons in the four channels $\tau_h\tau_h$ (1st column), $\mu\tau_h$ (2nd column), $e\tau_h$ (3rd column) and $e\mu$ (4th column). The observation in data is compared to the expectation given by the sum of all backgrounds. The agreement is acceptable given the preliminary simulation and the small data set. Larger discrepancies are visible in the $\tau_h\tau_h$ channel, that is mostly dominated by the QCD multi-jet background, and for lower electron momenta in the $e\tau_h$ channel, where the simulation of the $Z \to \tau\tau$ background seems to underestimate the data.

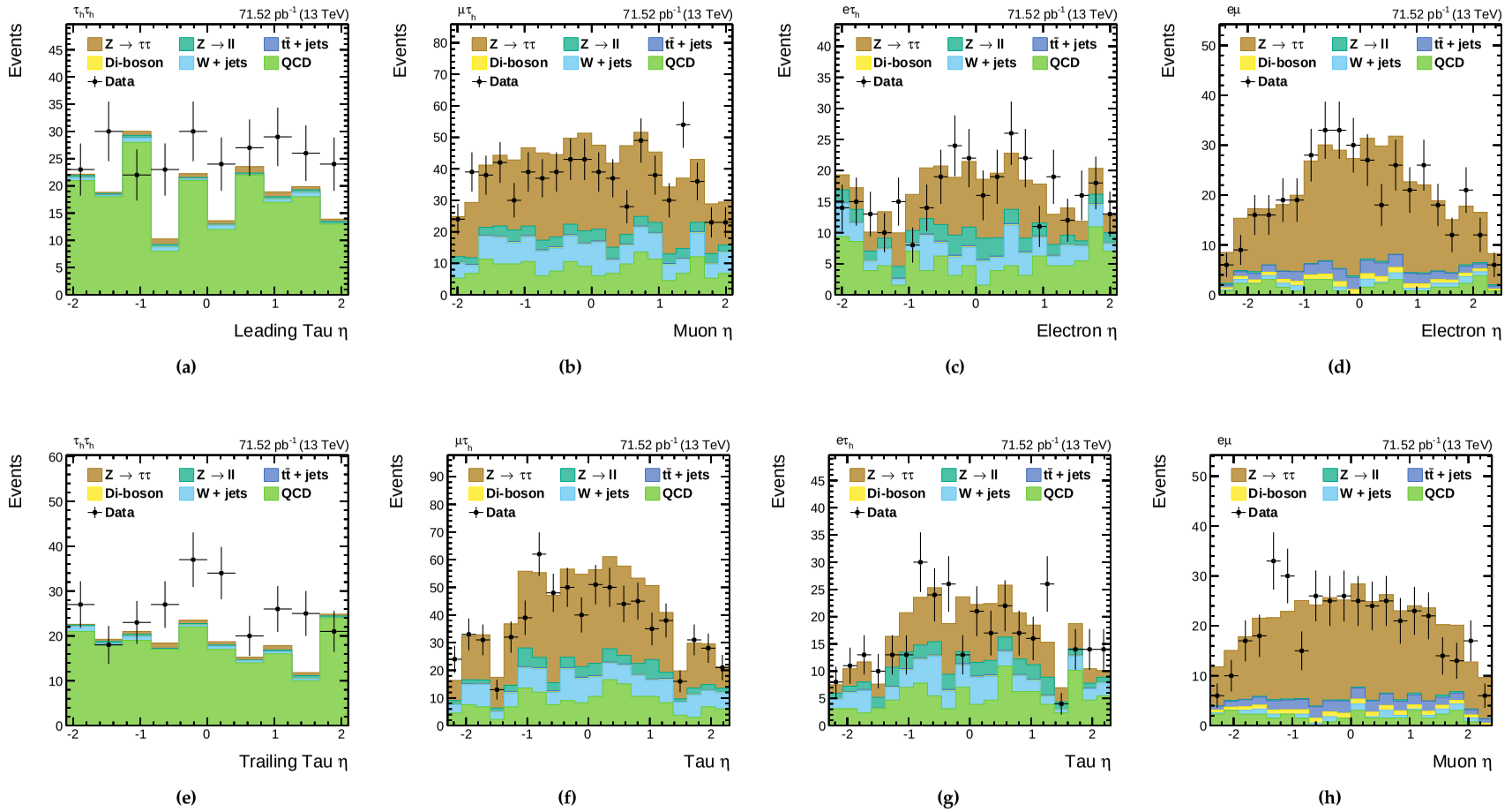


Figure 4.10.: Pseudorapidities of the two leptons in the four channels $\tau_h\tau_h$ (1^{st} column), $\mu\tau_h$ (2^{nd} column), $e\tau_h$ (3^{rd} column) and $e\mu$ (4^{th} column). The observation in data is compared to the expectation given by the sum of all backgrounds. The agreement is acceptable within the statistical uncertainties given the preliminary simulation and the small data set. Larger discrepancies are visible in overall normalisation of the background expectation in the $\tau_h\tau_h$ channel, that is mostly dominated by the QCD multi-jet background.

Table 4.3.: Kinematic cuts for the lepton selection in the four main channels.

$\boldsymbol{\tau_h\tau_h}$	$\boldsymbol{\mu\tau_h}$	$\boldsymbol{e\tau_h}$	$\boldsymbol{e\mu}$
$p_T(\tau) > 45$ GeV	$p_T(\mu) > 18$ GeV	$p_T(e) > 23$ GeV	$p_T(e) > 13$ GeV
	$p_T(\tau) > 20$ GeV	$p_T(\tau) > 20$ GeV	$p_T(\mu) > 10$ GeV
$\lvert\eta\rvert(\tau) < 2.1$	$\lvert\eta(\mu)\rvert < 2.1$	$\lvert\eta(e)\rvert < 2.1$	$\lvert\eta(e)\rvert < 2.5$
	$\lvert\eta(\tau)\rvert < 2.3$	$\lvert\eta(\tau)\rvert < 2.3$	$\lvert\eta(\mu)\rvert < 2.4$
$I(\tau) < 1.5$ GeV	$I_{rel}(\mu) < 0.1$	$I_{rel}(e) < 0.1$	$I_{rel}(e) < 0.15$
	$I(\tau) < 1.5$ GeV	$I(\tau) < 1.5$ GeV	$I_{rel}(\mu) < 0.15$

The kinematic cuts on the lepton four-momenta are listed in table 4.3. As the Higgs boson decays do not involve any strong interactions with exchange of colour charges, the leptons are required to be isolated from other activity in the detector. This reduces the QCD multi-jet background in the selection. The isolation I uses the $\Delta\beta$ corrections for the contribution of neutral particles in cones around the lepton flight directions are introduced in equation (3.1). For the taus a MVA-bases isolation variable is used.

Tracks from electrons or muons are required to originate from the primary vertex in order to reduce contributions from pile-up interactions. Distances of the lepton tracks to the primary vertex have to undermatch the following requirements in both the transverse plane and the longitudinal direction.

$$d_{xy}(\ell, \mathrm{PV}) < 0.45 \text{ mm} \quad \text{and} \quad d_z(\ell, \mathrm{PV}) < 2 \text{ mm}$$

The lepton selection lead to more than two leptons considered in single events. The resulting ambiguity for the definition of the lepton pair is resolved by constructing the pair from the two most isolated leptons. It has been validated on the H $\to \tau\tau$ signal sample that the efficiency for selecting the right pair is sufficiently large.

The figures 4.9 and 4.10 show the transverse momentum and the pseudorapidities of the leptons, respectively, for the four decay channels. The backgrounds in these plots are predicted based on the methods described below. The agreement between data and the sum of all backgrounds differs from channel to channel because of the different background compositions. Extensive studies of the QCD multi-jet background have not been possible based on the available simulated samples and the small data set. Discrepancies are therefore expected and seen in the $\tau_h\tau_h$ channel which is dominated by the background from QCD multi-jet events. The data in the other channels are well described by the predicted backgrounds within the statistical uncertainties. However, small systematic discrepancies remain, for example for lower electron momenta in the $e\tau_h$ channel, were differences are related to the simulation of the Drell-Yan background and the understanding of the data in this phase space region has to be improved.

As in the run I analyses, a multivariate regression technique is employed for the reconstruction of the missing transverse energy [106]. Figure 4.11 shows the distributions of this MET variable in the

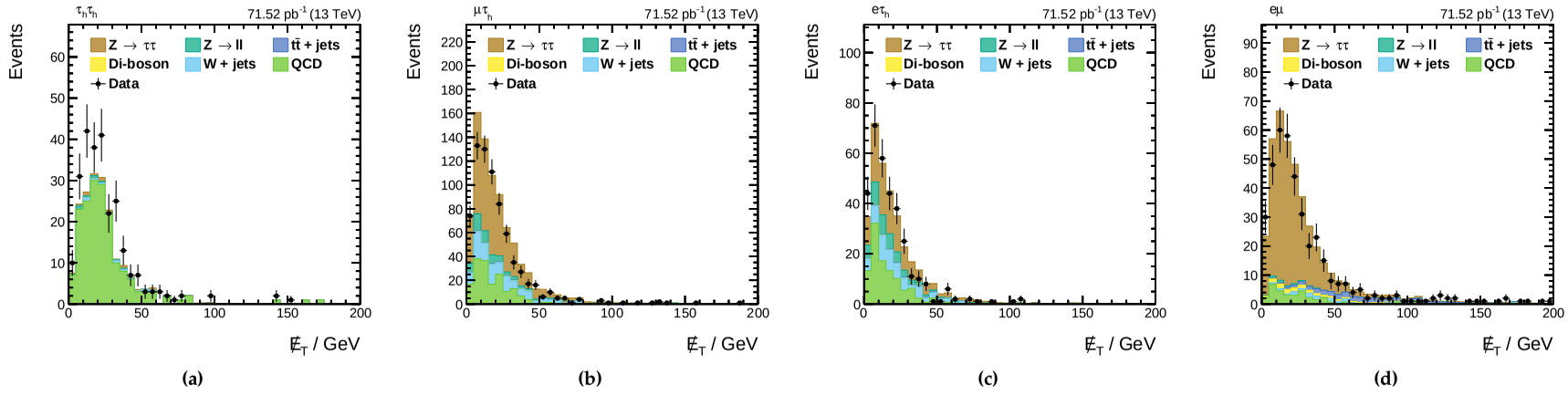


Figure 4.11.: Missing transverse energy in the four channels $\tau_h\tau_h$ (1st column), $\mu\tau_h$ (2nd column), $e\tau_h$ (3rd column) and $e\mu$ (4th column). The observation in data is compared to the expectation given by the sum of all backgrounds. The agreement is good within the statistical uncertainties given the preliminary simulation and the small data set. Only the $\tau_h\tau_h$ channel shows normalisation discrepancies. It is clearly visible that the resolution in the semi-leptonic final states is better then in the $e\mu$ channel, due to the smaller number of neutrinos.

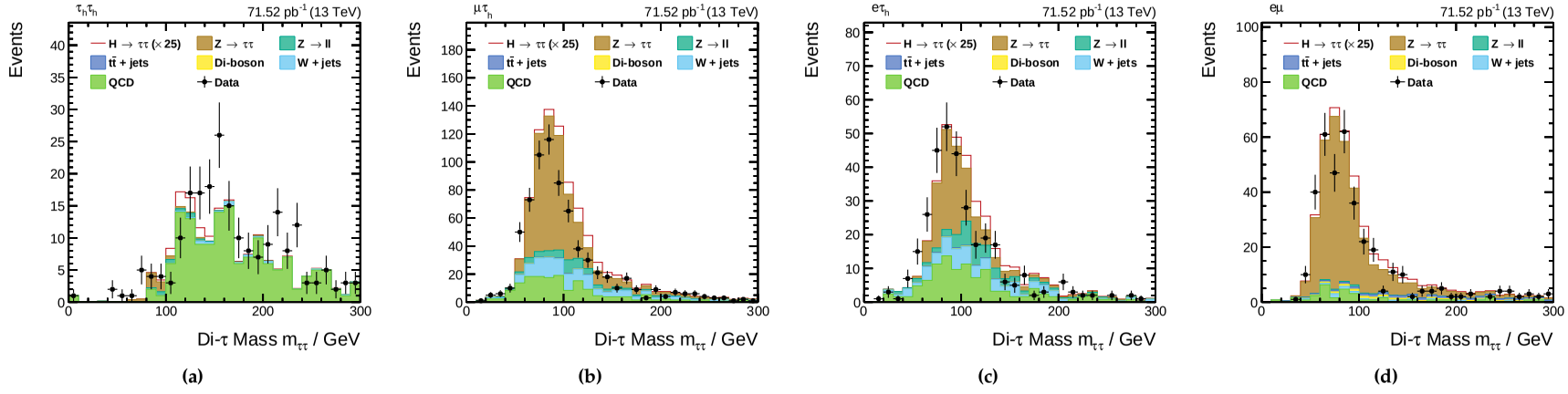


Figure 4.12.: Reconstructed invariant di-τ mass in the four channels $\tau_h\tau_h$ (1st column), $\mu\tau_h$ (2nd column), $e\tau_h$ (3rd column) and $e\mu$ (4th column). The observation in data is compared to the expectation given by the sum of all backgrounds. The agreement is good within the statistical uncertainties given the preliminary simulation and the small data set. Only the $\tau_h\tau_h$ channel shows normalisation discrepancies. The mass peak is sharper in the $\mu\tau_h$ and $e\tau_h$ channels compared to the $e\mu$ channel because of the better MET resolution. All distributions show tails to higher mass values, which potentially reduce the discrimination power when the mass is compared to the one of the $H \to \tau\tau$ signal. The 125 GeV signal is added scaled by a factor of 25. However, the amount of data is not yet sufficient to see sensitivity to the expected signal.

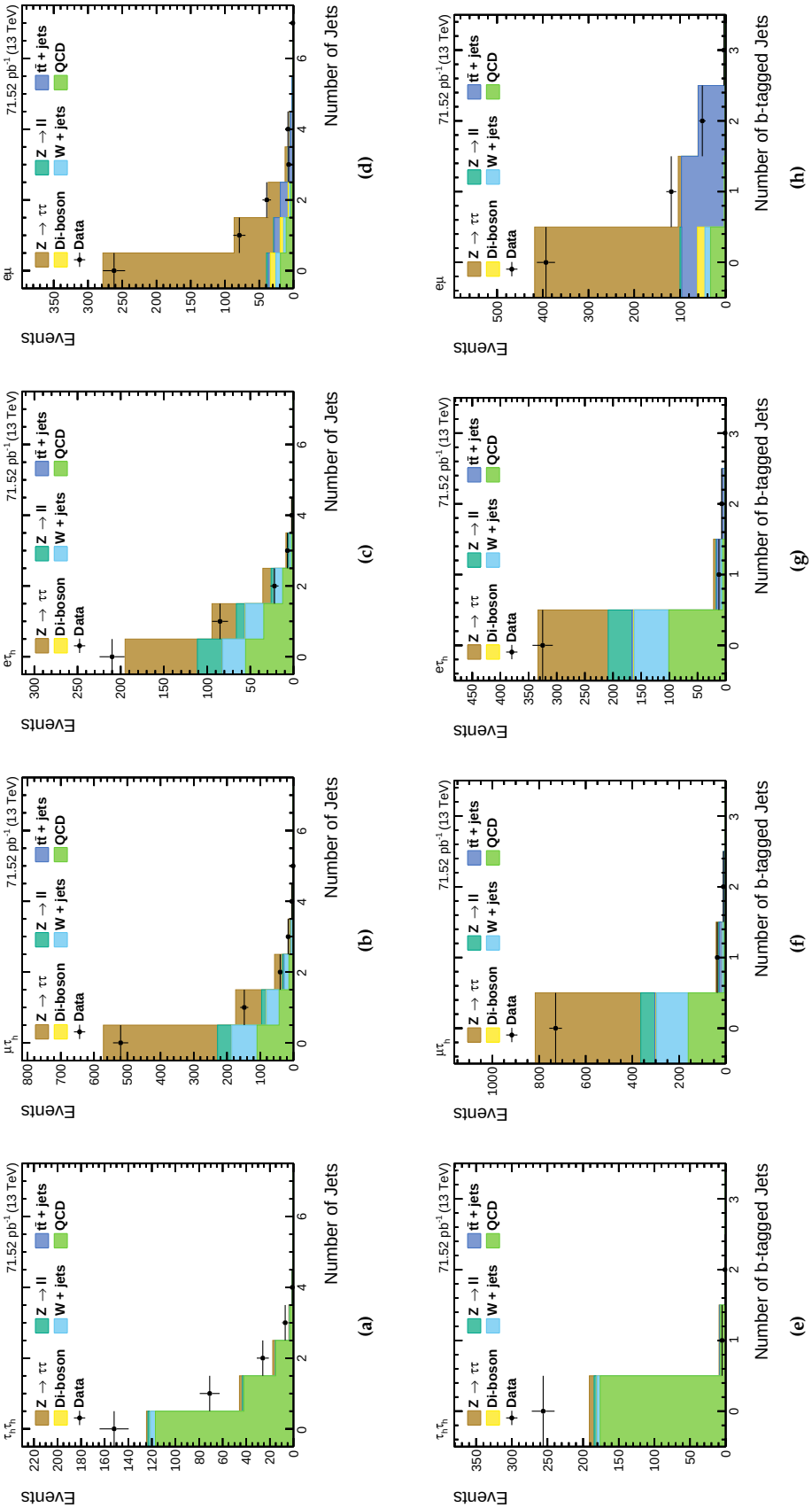


Figure 4.13.: Numbers of jets (top) and b-tagged jets (bottom) in the four channels $\tau_h\tau_h$ (1st column), $\mu\tau_h$ (2nd column), $e\tau_h$ (3rd column) and $e\mu$ (4th column). The observation in data is compared to the expectation given by the sum of all backgrounds. The agreement is acceptable considering the statistical uncertainties given by the preliminary simulation and the small data set. Only the $\tau_h\tau_h$ channel shows normalisation discrepancies. However, the jets deserve dedicated studies as soon as the complete simulated samples are available and the predictions of different generators can be compared to each other. It is visible that the requirement of zero b-tagged jets reduces the background from $t\bar{t}$+jets events in the $e\mu$ channel.

four channels. The dependency of the resolution on the number of neutrinos in the final state is clearly visible.

Again, as in the run I analyses, the reconstruction of the invariant di-τ mass is determined using the maximum likelihood method summarised in section 3.3.2. The resulting distributions are shown in figure 4.12. Small discrepancies between data and the sum of all backgrounds are mostly covered by the statistical uncertainties. The resolution of the missing transverse energy depending on the number of neutrinos in the final state directly translates to the resolution of the reconstructed τ pair mass. However, the tails towards higher masses seem to be more pronounced than in the run I analyses. This can be related to known issues of the covariance matrix of the MET that describes the reconstruction uncertainty on the MET, and has to be studied carefully in the future.

Jets are identified based on a loose identification working point after their four-momenta have been corrected explained in section 2.4.5. These corrections have been newly derived on the run II data set. The following acceptance cuts are applied.

$$p_{\mathrm{T}}(\text{jet}) > 30 \text{ GeV} \quad \text{and} \quad \left|\eta(\text{jet})\right| < 4.7$$

Possible overlap with the reconstructed leptons is removed by requiring that each jet has to be separated in the η-φ space from the one of the leptons.

$$\Delta R(\text{jet}, \ell) > 0.5$$

Jets originating from b decays are identified based on a medium working point of a multivariate CSV discriminator. Events containing these b-tagged jets are vetoed in the analysis in order to reduce the background from $\mathrm{t\bar{t}}$+jets events. The distributions of the numbers of jets and b-tagged jets are shown in figure 4.13 for the four channels. Distributions of the transverse momenta of the leading and sub-leading jets as well as their pseudorapidities are shown in figures C.1 and C.3. As these distributions require the existence of at least one or two jets, respectively, the number of events is significantly smaller than in the inclusive selection. The uncertainties are completely dominated by the statistical errors of data and the simulation.

All plots for the 13 TeV analysis shown so far include the methods for the background modelling described in the following sections. In summary, it has to be noted that the agreement between the early 13 TeV data and the sum of all backgrounds is good within the statistical uncertainties related to the data and also to the simulation. The following caveats have to be noted:

- The simulation is based on a 25 ns bunch spacing scenario whereas the early 13 TeV data comprising 71.52 pb^{-1} has been produced with 50 ns bunch spacing. Additionally, the preliminary simulated samples, that are available, are characterised by large statistical uncertainties due to their small number of generated events.

- Identification and reconstruction efficiencies are not yet studied. Discrepancies between data and the simulation would have to be corrected in the simulated samples. As an example, the disagreement in the low electron momenta in the $\mathrm{e}\tau_{\mathrm{h}}$ channel seem to be caused by different identification efficiencies in data and in the simulation.

- Compared to the 8 TeV analysis, looser selection criteria are chosen in order to select a sufficient number of events. This selection needs to improved in order to optimise the selection of signal-like events. Events remaining in the selection that are less signal-like or that are even misidentified are expected to deteriorate the agreement between data and the simulation.
- The reconstruction of physics objects and higher level calculations are not yet fully checked and optimised to the new data. This is especially true for the multivariate regression for the missing transverse energy and the reconstruction of the invariant di-τ mass. In the distribution of the mass large tails towards high values are visible that decrease the separation power of this quantity as a discriminator between signal and background.

4.4.1. Modelling of the Drell-Yan Background

In the run I analyses, the most dominant background from $\mathrm{Z} \rightarrow \tau\tau$ events has been modelled with the embedding technique in all channels (see section 3.6.2). In the run II data-taking period the embedding technique needs to be revised due to the increased amount of pile-up interactions. Some studies are already available [91,107], while the production of embedded data samples is still in planning. Therefore, the simulated Drell-Yan samples are used for the estimation of the $\mathrm{Z} \rightarrow \tau\tau$ and $\mathrm{Z} \rightarrow \ell\ell$ backgrounds in the scope of this thesis.

4.4.2. Modelling of the $\mathrm{t\bar{t}}$+jets Background

Leptonic decays of $\mathrm{t\bar{t}}$+jets events constitute an important background in the $\mathrm{e}\mu$ channel (see figure 4.1), as the number of $\mathrm{Z} \rightarrow \tau\tau$ events is reduced compared to the other channels due to the small branching ratio $\mathcal{BR}(\tau\tau \rightarrow \mathrm{e}\mu)$ of 6.2 % (see table 1.5) and the $\mathrm{Z} \rightarrow \ell\ell$ background is strongly suppressed by the selection of two leptons of different flavour. The 7 and 8 TeV analysis in the $\mathrm{e}\mu$ channel exploited a multivariate suppression of the $\mathrm{t\bar{t}}$+jets background, which was mainly based on the variables p_ζ^{miss} and p_ζ^{vis} and the number of b-tagged jets. Because of the small data set, the studies presented in this thesis follow a cut-based approach based on these variables.

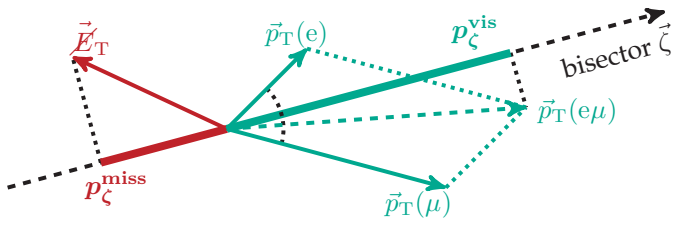


Figure 4.14.: Illustration of the determination of the variables p_ζ^{miss} and p_ζ^{vis} that are used to discriminate against the background from $\mathrm{t\bar{t}}$+jets events.

Given the bisector, $\vec{\zeta}$, between the flight directions of the two leptons in the transversal plane

$$\vec{\zeta} = \frac{\vec{p}_{\mathrm{T}}(\mathrm{e})}{p_{\mathrm{T}}(\mathrm{e})} + \frac{\vec{p}_{\mathrm{T}}(\mu)}{p_{\mathrm{T}}(\mu)}$$

the variables p_ζ^{miss} and p_ζ^{vis} are defined as the projection of the missing transverse energy or the vectorial sum of the visible lepton momenta, respectively [108], as illustrated in figure 4.14.

$$p_\zeta^{\text{miss}} = \vec{\not{E}}_{\text{T}} \cdot \frac{\vec{\zeta}}{|\vec{\zeta}|} \quad \text{and} \quad p_\zeta^{\text{vis}} = \left(\vec{p}_{\text{T}}(\text{e}) + \vec{p}_{\text{T}}(\mu)\right) \cdot \frac{\vec{\zeta}}{|\vec{\zeta}|}$$

The variable p_ζ^{miss} exploits the fact that the spin and flight directions of the taus and the tau neutrinos are correlated. Because of their vanishing mass and the fact that they are always left-handed, the spin of the neutrino has to point in the opposite direction of its momentum. The tau neutrinos from τ decays preferably fly into the same direction. Therefore, $\text{Z/H} \to \tau\tau$ decays tend to result in positive values of p_ζ^{miss}. However, also the neutrinos from the leptonic W decays contributed to the missing transverse energy, leading to a smearing of the p_ζ^{miss} variable. In $\text{t}\bar{\text{t}}$+jets events, the flight directions of the two leptons and the missing transverse energy are much less correlated and a tail towards negative values appears. The combination $p_\zeta = p_\zeta^{\text{miss}} - 0.85\, p_\zeta^{\text{vis}}$ is found to show a good separation between $\text{t}\bar{\text{t}}$+jets and signal events, as shown in figure 4.16.

The $\text{t}\bar{\text{t}}$+jets background is mainly reduced by vetoing events containing b-tagged jets. Additional suppression is given based on the p_ζ variable. Singal-like events are selected based on the cut $p_\zeta > -20$ GeV. The region below $p_\zeta = -20$ GeV is identified as a control region for the $\text{t}\bar{\text{t}}$+jets background. In this region, the yield of the $\text{t}\bar{\text{t}}$+jets contribution is corrected to the one measured in data, where small contributions from other backgrounds have been subtracted based on the prediction by the simulation. The scale factor for $\text{t}\bar{\text{t}}$+jets normalisation in the signal region is found to be 1.19 ± 0.35. The statistical uncertainty here is considered as a systematic uncertainty accounting for the modelling of this background.

In all other channels, the $\text{t}\bar{\text{t}}$+jets contribution is predicted based on the simulation.

4.4.3. Modelling of the W+jets Background

In the semi-leptonic channels, $\mu\tau_{\text{h}}$ and $\text{e}\tau_{\text{h}}$, the background from W+jets events remains large after the selection of two oppositely charged leptons. The hadronically decaying τ lepton can be a misidentified jet and the lighter lepton can originate from the decay of a W boson. This background can be suppressed based on the transverse mass of the light lepton and the missing transverse energy, $m_{\text{T}}(\ell, \vec{\not{E}}_{\text{T}})$.

$$m_{\text{T}}\left(\ell, \vec{\not{E}}_{\text{T}}\right) = \sqrt{2\, p_{\text{T}}(\ell) \not{E}_{\text{T}} \left(1 - \cos\Delta\varphi\right)} \quad \text{with} \quad \Delta\varphi = \varphi(\ell) - \varphi(\not{E}_{\text{T}})$$

Figure 4.15 shows the distributions for this variable in the two channels. In contrast to the τ leptons, the much heavier W bosons are produced with less transverse momentum which leads to a larger angular separation between the flight directions of the lepton and the neutrino in the W decays. In τ decays, the decay products are more collimated due to the higher boost of the τ lepton originating from the decay of a heavy resonance.

The baseline selection defines an upper cut on the transverse mass, $m_{\text{T}}(\mu, \vec{\not{E}}_{\text{T}})$ or $m_{\text{T}}(\text{e}, \vec{\not{E}}_{\text{T}})$ of 30 GeV. The region above $m_{\text{T}} = 70$ GeV is dominated by W+jets events and is therefore used to control

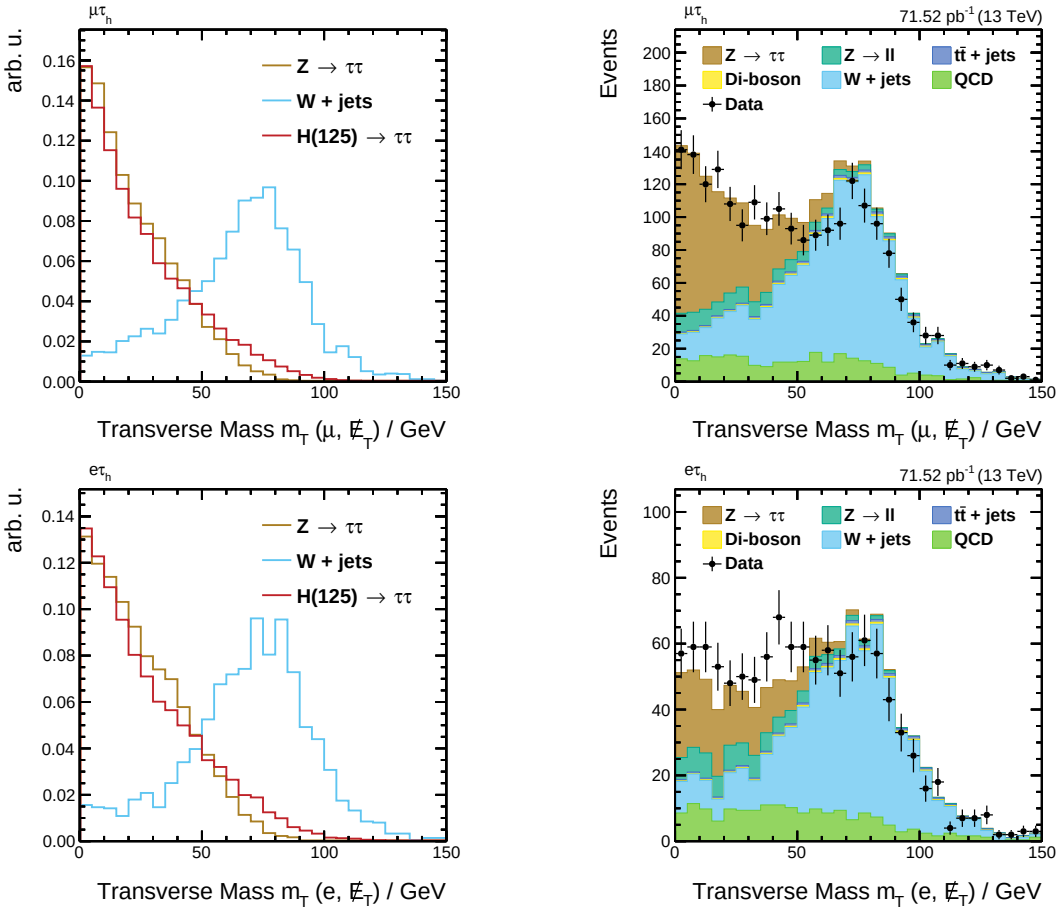


Figure 4.15.: The background from W+jets events in the semi-leptonic channels ($\mu\tau_h$ top and $e\mu$ bottom) is suppressed by requiring transverse masses below 30 GeV. The normalisation of the W+jets background is corrected to the one in data measured in the control region above 70 GeV, which is dominated by this background. The right plots show the result of this data-driven estimation. Discrepancies for low transverse mass values are due to other backgrounds, mainly from $Z \rightarrow \tau\tau$ events.

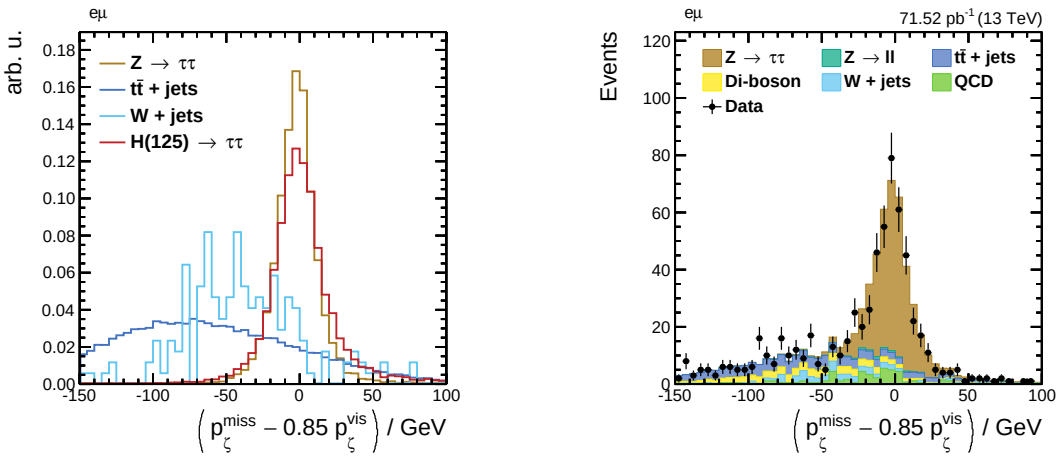


Figure 4.16.: The background from $t\bar{t}$+jets events in the $e\mu$ channel is suppressed by requiring $p_\zeta > -20$ GeV. The normalisation of the $t\bar{t}$+jets background is corrected to the one in data measured in the control region defined by $p_\zeta < -20$ GeV, which is dominated by this background. The right plot shows the result of this data-driven estimation. The agreement between the data and the sum of all background is good.

this background with data. The yield of the W+jets simulation is corrected to the one measured in data, where the small contributions from other backgrounds have been subtracted based on the prediction from the simulation. The scale factor the for W+jets yield in the signal region is found to be 1.04 ± 0.15 in the $\mu\mu$ channel and 1.10 ± 0.24 in the $\mathrm{e}\tau_{\mathrm{h}}$ channel. The statistical uncertainty here is considered as a systematic uncertainty accounting for the modelling of this background.

In the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ and $\mathrm{e}\mu$ channels, this background is estimated fully based on the simulation.

4.4.4. Modelling of the Di-boson Background

The background prediction of di-boson events, WW, WZ, and ZZ, is based on the simulation, as it only plays a subordinate role. The event kinematics look very similar to the ones of Z decays as dominantly only one boson is produced on the mass shell. This makes it difficult to define a control region that is only dominated by this background.

4.4.5. Modelling of the QCD Multi-jet Background

As hadronically decaying τ leptons have a very similar signature in the detector as quark or gluon jets, the probability to select QCD multi-jet events, where jets have been misidentified, increases with the number of hadronically decaying τ leptons in the final state. These non-resonant events are estimated in a data-driven way. As there is no correlation between the charges of leptons or jets in QCD multi-jet events, the modelling is based on the data sample from which leptons are selected with same-sign charges. For this selection, the small contributions of the all other backgrounds are subtracted. Especially the W+jets background in the same-sign charge selection is estimated following the same principle as described in section 4.4.5 for the semi-leptonic channels. The yields of the same-sign charge selection are scaled by 6 %, a factor for the transition from the same-sign to the opposite-sign charge selection that has been measured in the 8 TeV analysis.

4.5. Analysis Strategy and Event Categorisation

The aim of the analysis is to measure the properties of H $\rightarrow \tau\tau$ couplings. These measurements are based on a quantity discriminating between background and signal events. For basic measurements such as ones of coupling strength parameters, the reconstructed invariant di-τ mass provides this information, it yields values around the Higgs boson mass for the signal and the dominant background from Z $\rightarrow \tau\tau$ events peaks near the Z boson mass. A signal-like excess is searched at the upper tail of the Z boson mass peak.

On the one hand, the sensitivity of the analysis depends on the overall contamination of the signal region with background events and the difference in the shapes of the mass variable for the various processes. Basic background suppression mechanisms have already been discussed. On the other hand, the sensitivity is improved by the event categorisation. More signal-like and more background-like

event categories are defined differing in the signal-to-background ratio compared to the one of the inclusive selection. The background-like event categories are used to constrain the background-related systematic uncertainties, whereas the signal-like event categories are used to establish the signal. At the same time, the event categorisation exploits different Higgs boson production mechanism and their characteristic kinematics.

The categorisation of this prototype analysis follows as much as possible the strategy of the run I analysis. However, the preliminary simulated samples and the much smaller available data set requires are less tight categorisation. The statistical precision related to the size of the simulated samples has a strong impact on the uncertainties of the result as will be shown below. The amount of the available data is relevant for the modelling of the data-driven background predictions, as they are used for the backgrounds from $\mathrm{t\bar{t}}$+jets, W+jets and the QCD multi-jet events. The estimations are completely done in the inclusive selection and the categorisation is applied on top of this modelling step. This means that the efficiencies of the cuts which define the event categories are taken from the simulation in the $\mathrm{t\bar{t}}$+jets ar

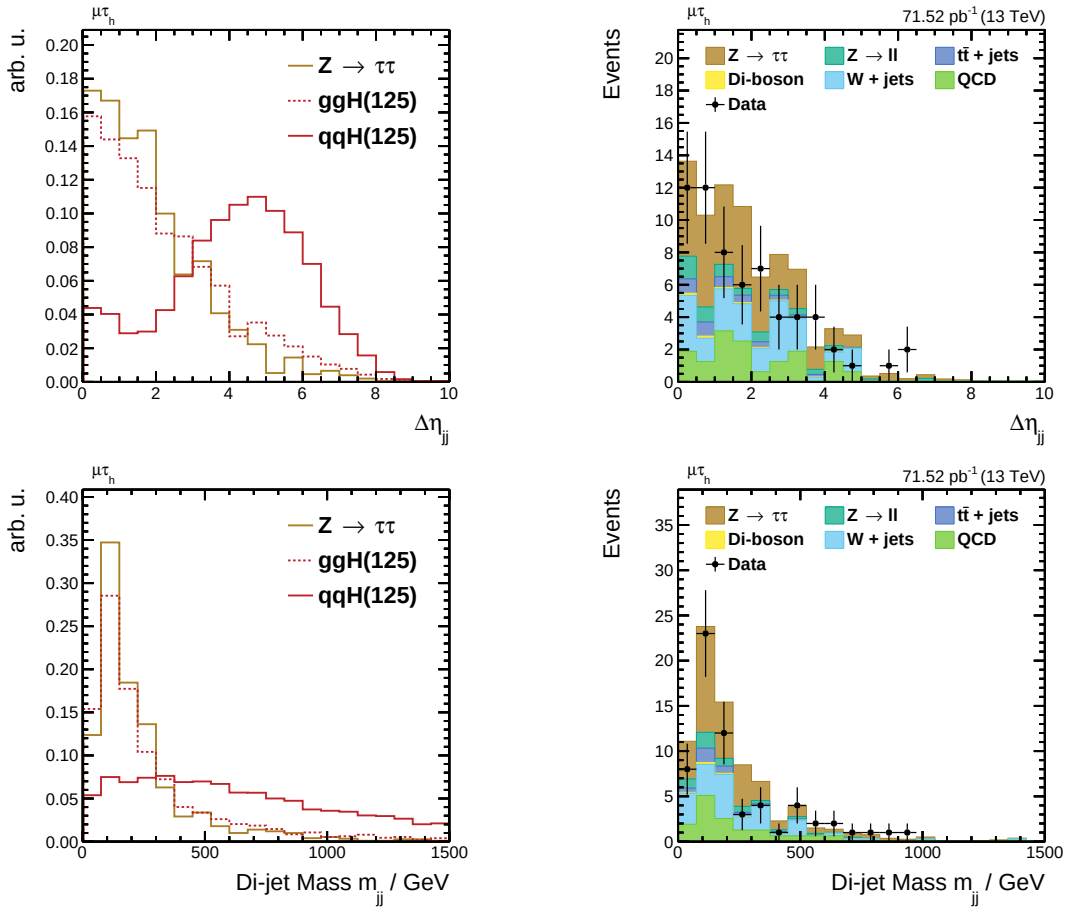


Figure 4.17.: Signal events produced via vector boson fusion (qqH) are distinguished from gluon fusion signal (ggH) and SM background events based on the di-jet system. Large differences in the pseudorapidity and the large mass of the two jets characterise the VBF production mechanism. The distributions of both quantities are well modelled by the simulation (right).

The event categorisation mainly exploits the number of jets and the transverse momentum of one lepton. The jet multiplicity is a measure for the production mode. Three categories are distinguished:

VBF: Events are required to contain at least two jets with $p_{\mathrm{T}}(\text{jet}) > 30$ GeV and a di-jet mass of $m_{jj} > 200$ GeV. These jets should be separated in the pseudorapidity by $\Delta\eta_{jj} > 2$.

1-jet: Contains events not passing the VBF selection criteria and containing at least one jet with $p_{\mathrm{T}}(\text{jet}) > 30$ GeV.

0-jet: Contains events that do not pass the VBF and 1-jet selection criteria.

VBF signal events ($qq\mathrm{H}$) are characterised by the two forward jets and low activity in the central part of the detector apart from the signal signature. The chosen requirements on the longitudinal separation of the two jets and their mass distinguish these events from other signal productions, mainly gluon fusion (ggH), and the SM backgrounds, as exemplarily shown in figure 4.17 for the $\mu\tau_{\mathrm{h}}$ channel.

Cuts on the transverse momenta of the leptons exploit the fact that signal events are characterised by a harder spectrum than $\mathrm{Z} \rightarrow \tau\tau$ events because of the larger mass of the intermediate boson. The 1-jet and 0-jet categories are therefore split into two sub-categories, respectively: A low-pt and a high-pt category, where the first is more dominated by backgrounds. The threshold is chosen at $p_{\mathrm{T}}(\ell) = 35$ GeV, where ℓ denotes the tau in the semi-leptonic channels and the muon in the $\mathrm{e}\mu$ channel. The tau is taken, because it provides a better discrimination between signal and background. The transverse momenta of misidentified light leptons tend to yield higher values.

In total, five event categories are constructed for the semi-leptonic and fully-leptonic channels. Because of the low number of selected events the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel is not concluded in the extrapolation studies presented in the following. Due to similar statistical caveats, the data-driven modelling of the shapes of the $\mathrm{W}+$jets background in the $\mu\tau_{\mathrm{h}}$ and $\mathrm{e}\tau_{\mathrm{h}}$ channels and of the QCD multi-jet background in the $\mu\tau_{\mathrm{h}}$, $\mathrm{e}\tau_{\mathrm{h}}$ and $\mathrm{e}\mu$ channels is modified: the requirements on the lepton isolation are relaxed for the definition of the discriminator shapes, while the inclusive yields are maintained for the original isolation requirements. This procedure is justified, since the shapes do not change significantly. For light leptons, electrons and muons, the isolation requirement is fully omitted while taus are required to pass a loose requirement of $I(\tau) < 10$ GeV. A larger number selected events provides smoother shapes that better resemble the physical process.

The di-τ mass distributions in all event categories are shown in the appendix in figures C.3, C.4 and C.5.

4.6. Extrapolations of Experimental Precisions

Preparations for the $\mathrm{H} \rightarrow \tau\tau$ analysis of CMS run II data have been shown so far. They have been presented based on the data set taken with 50 ns bunch spacing corresponding to an integrated luminosity of 71.52 pb^{-1}. In the CMS run I data set roughly 25 fb^{-1} of data and a very sophisticated analysis was needed to find evidence for $\mathrm{H} \rightarrow \tau\tau$ couplings. Since the signal cross sections at a centre of mass energy of $\sqrt{s} = 13$ TeV increase only slightly more than the ones of the backgrounds, no Higgs boson signal is expected to be observed, rather then studies are possible, in the first few fb^{-1} of the CMS run II data.

In general, analyses are done in a blind fashion, meaning that analysts agree to not looking at the data in signal regions before the details of the complete analysis are defined and the data is optimally understood in background-dominated phase spaces. Once the analysis strategy is fixed and all optimisations based on the simulation are finalised, data in the signal region are studied. This procedure guarantees results that are not biased towards expectations or hopes of the analysts.

In the scope of this thesis there is no need of blinding the signal region, although the data taking is under way and many aspects of the data have still to be thoroughly studied and optimisations for the sensitivity of the analysis have to be carved out. The reason is the low signal expectation in the first 71.52 pb^{-1} of data: In the inclusive selection, 2.5 signal events are expected in the $\mu\tau_{\mathrm{h}}$ channel, one event in the $\mathrm{e}\tau_{\mathrm{h}}$ channel and 1.5 events in the $\mathrm{e}\mu$ channel. The expectation refers to the SM Higgs boson of a mass of 125 GeV. Compared to the background expectation of 300 to 800 events in these channels, the sensitivity of the analysis to the signal is extremely low.

Therefore, the outlook of this thesis is dedicated to preliminary extrapolations of the precision that can be reached for measures of the $\mathrm{H} \to \tau\tau$ signal as a function of the size of the data set. These predictions have to be seen as approximate upper thresholds on the precision that future analyses are going to reach as the starting point is the status of the analysis as presented in the sections before and as the extrapolations cover a time span that will allow analysts to significantly improve the analysis.

All extrapolations presented in the following are based on maximum likelihood fits as described in section 3.8.1. These fits are based on so-called Asimov data sets. This means that the observation in the likelihood function of equation (3.5) is defined by a signal-plus-background hypothesis. By doing so, the sensitivity of the analysis is examined. The signal prediction is taken as the $\mathrm{H} \to \tau\tau$ signal with a mass hypothesis of $m_{\mathrm{H}} = 125$ GeV. The parameters of interest for the fits depend on the model to be tested. It is resigned to present exclusion limits or p-values, as based on the combination of the run I results from ATLAS and CMS an excess of more than 5 σ is already observed, which finally proved the existence of $\mathrm{H} \to \tau\tau$ decays [109]. Therefore, the primary goal of the run I measurements in the Standard Model field focus on measurements of these couplings.

4.6.1. Systematic Uncertainties

The treatment of systematic uncertainties constitutes a crucial part for the study of the precision of future measurements. In the scope of this thesis, the following conservative assumptions are made:

- The performance of the identification and reconstruction of physics objects by the detector and the reconstruction software remains the same.
- Theoretical uncertainties, that mainly affect the $\mathrm{H} \to \tau\tau$ signal, stay unchanged.
- Uncertainties covering the imprecisions resulting from the modelling of the background samples are usually estimated based on the statistical uncertainties introduced by the modelling procedure. Future improvements in this area are not considered.

These assumptions describe a worst-case scenario. Experimentalists will try to continuously improve the identification and reconstruction of physics objects as well as to reduce the systematic uncertainties introduced by their measurements. Similarly, theorists will provide calculations and simulation with higher precision as time evolves. Therefore, the resulting measurement uncertainties have to be taken as upper boundaries of what will be achieved by future analyses. Other scenarios are studied by both the ATLAS and the CMS collaboration as documented in the references [110, 111].

The actual systematic uncertainties assigned to the analysis are motivated by the run I analysis. Many sources of uncertainties have not yet been studies and the results of the 8 TeV analysis have been taken. Exceptions are described below. A list of all systematic uncertainties is provided in table 4.4.

Table 4.4.: Systematic uncertainties, affected samples, and change in acceptance resulting from a variation of the nuisance parameter equivalent to one standard deviation as applied to the extrapolation studies. Several systematic uncertainties are treated as correlated for different decay channels and/or categories.

Uncertainty	**Affected Processes**	**Change in Acceptance**
Tau energy scale	signal & $Z \rightarrow \tau\tau$	shape unc.
Tau ID & trigger	signal & sim. backgrounds	8-19 %
Tau misidentification	$Z \rightarrow \ell\ell$	30 %
Muon ID & trigger	signal & sim. backgrounds	2 %
Electron ID & trigger	signal & sim. backgrounds	2 %
Jet energy scale	signal & sim. backgrounds	shape unc.
Jet b-tagging efficiency	$t\bar{t}$+jets	4-7 %
MET scale	signal & sim. backgrounds	2-3 %
Norm. Z+jets	$Z \rightarrow \tau\tau$ & $Z \rightarrow \ell\ell$	3 %
Norm. $t\bar{t}$+jets	$t\bar{t}$+jets	10-120 %
Norm. diboson	di-boson	15 %
Norm. W+jets	W+jets	20-52 %
Norm. QCD multi-jet	QCD multi-jet	6-100 %
Luminosity	signal & sim. backgrounds	2.6 %
PDF	signal	2.2-7.1 %
Scale variation	signal	0.7-7.9 %
Underlying event & parton shower	signal	up to 10.6 %
Limited number of events	all	shape unc.

The resolution of the τ energy scale was measured as 3 % in the 8 TeV analysis [10]: A fit assessing the compatibility of the data with the sum of all backgrounds is performed. The fit optimises a correction factor for the τ energy scale. In this analysis, τ energy shifts of 3 % are propagated through the entire analysis chain and the resulting effect on the di-τ mass shapes is considered as a shape uncertainty. The τ energy scale affects the four momentum of the τ leptons which is an input for the di-τ mass reconstruction and which is used for the categorisation. Similarly, the uncertainty provided with the determination of the jet energy corrections is covered by shape uncertainties based on the propagation of the resulting shifts through the entire analysis chain. This uncertainty covers the imprecisions introduced by the categorisation based on the jet multiplicity.

The effect of the data-driven estimation of yields of background contributions is covered by normalisation uncertainties that are determined based on the statistical error on the scale factor for the transitions from the control to the signal regions and on the statistical precision of the yield obtained from the simulation that is scaled. This uncertainty has the largest effect in the VBF and in the high-pt categories that contain only few entries.

The available simulated samples are characterised by a low statistical precision due to the low numbers of generated events. So-called bin-by-bin uncertainties are assigned to cover these effects. In the run I analysis the precision of the results was not significantly limited by the statistical precision of the simulated samples. The size of the simulated samples is expected to grow with the size of the recorded data set. Therefore, these uncertainties are excluded from the studies presented below.

4.6.2. Signal Strength and Coupling Modifiers

A leading-order motivated framework, the κ-framework, is used to interpret data measuring Higgs boson couplings [112]. This framework defines signal strength and coupling modifiers that scale the Standard Model predictions and are therefore used to quantify possible deviations from the SM. In general, the production of the Higgs boson via the process i and its decay via the process f can be distinguished. In the narrow width approximation the signal strength modifier μ is defined relative to the SM Higgs boson production cross section, σ_{SM}, times the branching ratio for its decay, $\mathcal{BR}_{\mathrm{SM}}$.

$$\mu = \mu_i \cdot \mu_f = \frac{\sigma(i)}{\sigma_{\mathrm{SM}}(i)} \cdot \frac{\mathcal{BR}(f)}{\mathcal{BR}_{\mathrm{SM}}(f)}$$

Coupling modifiers for Higgs boson decays are commonly expressed in terms of the partial decay widths, $\Gamma(f)$.

$$\kappa_i^2 = \frac{\sigma(i)}{\sigma_{\mathrm{SM}}(i)} \quad \text{and} \quad \kappa_f^2 = \frac{\Gamma(f)}{\Gamma_{\mathrm{SM}}(f)}$$

The scaling of any $i \rightarrow f$ process can be expressed based on these κ factors.

$$\sigma(i) \cdot \mathcal{BR}(f) = \kappa_i^2 \, \sigma_{\mathrm{SM}}(i) \cdot \frac{\kappa_f^2 \, \Gamma_{\mathrm{SM}}(f)}{\kappa_{\mathrm{H}}^2 \, \Gamma_{\mathrm{H}}^{\mathrm{SM}}} \tag{4.1}$$

The scaling of the couplings of the individual processes can have an influence on the total decay width of the Higgs boson, Γ_{H}. The corresponding modifier is denoted by κ_{H}. In the Standard Model, all values μ and κ are positive and equal to unity.

In the scope of the H $\rightarrow \tau\tau$ analysis, the final state is limited to H $\rightarrow \tau\tau$ decays. Via the Higgs boson production mechanisms this channel is sensitive to couplings to other particles. The gluon fusion production is based on couplings to mainly top quarks, but also bottom quarks contribute as well as interferences appearing in the loop. The vector boson fusion production and the production in association with vector bosons are based on couplings to W and Z bosons. Table 4.5 lists the coupling modifiers for the relevant processes.

Table 4.5.: Coupling modifiers for the processes relevant in the H $\to \tau\tau$ analysis [109]. The factors modifying the couplings to the elementary particles are derived from SM calculations and involve interferences in case of loops that are indicated by the negative terms.

Process	**Multiplicative Coupling Modifier**
Gluon fusion: $\sigma\left(\mathrm{pp} \to \mathrm{gg} \to H\right)$	$1.06\,\kappa_\mathrm{t}^2 + 0.01\,\kappa_\mathrm{b}^2 - 0.07\,\kappa_\mathrm{t}\kappa_\mathrm{b}$
Vector boson fusion: $\sigma\left(\mathrm{pp} \to qq \to H\right)$	$0.74\,\kappa_\mathrm{W}^2 + 0.26\,\kappa_\mathrm{Z}^2$
Associated production: $\sigma\left(\mathrm{pp} \to \mathrm{WH}\right)$	κ_W^2
Associated production: $\sigma\left(\mathrm{pp} \to qq/qg \to \mathrm{ZH}\right)$	κ_Z^2
Associated production: $\sigma\left(\mathrm{pp} \to \mathrm{gg} \to \mathrm{ZH}\right)$	$2.27\,\kappa_\mathrm{Z}^2 + 0.37\,\kappa_\mathrm{t}^2 - 1.64\,\kappa_\mathrm{Z}\kappa_\mathrm{t}$
Partial decay width: $\Gamma\left(\mathrm{H} \to \tau\tau\right)$	κ_τ^2
Total decay width: Γ_H	$0.57\,\kappa_\mathrm{b}^2 + 0.22\,\kappa_\mathrm{W}^2 + 0.09\,\kappa_\mathrm{g}^2 + 0.06\,\kappa_\tau^2$
	$+0.03\,\kappa_\mathrm{Z}^2 + 0.03\,\kappa_\mathrm{c}^2 + 0.0023\,\kappa_\gamma^2 + 0.0016\,\kappa_{\mathrm{Z}\gamma}^2$
	$+0.0001\,\kappa_\mathrm{s}^2 + 0.00022\,\kappa_\mu^2$

4.6.3. Measurement of the Signal Strength

The simplest assumption for the signal strength modifiers is made for the measurement of signal strengths: all signal strength modifiers μ_i and μ_f are assumed to have the same value. Then the combined signal strength is expressed as follows[3].

$$\mu = \frac{\sigma\left(\mathrm{pp} \to \mathrm{H}\right) \cdot \mathcal{BR}\left(\mathrm{H} \to \tau\tau\right)}{\sigma_\mathrm{SM}\left(\mathrm{pp} \to \mathrm{H}\right) \cdot \mathcal{BR}_\mathrm{SM}\left(\mathrm{H} \to \tau\tau\right)}$$

The extrapolation of the parameter μ as a function of the integrated luminosity is shown in figure 4.18 (left). The best fit value always yields the expectation for the SM Higgs boson with a mass hypothesis of 125 GeV, because the fit has been performed on an Asimov data set following the signal-plus-background hypothesis. The plot distinguishes three components of 1 σ uncertainties on this parameter. The statistical uncertainty of the data set roughly evolves proportional to $1/\sqrt{N}$, where the number of events, N, is again proportional to the integrated luminosity. Secondly, the systematic uncertainties as described in section 4.6.1 result in an almost constant contribution of 11 $\%$ to the total uncertainty[4]. Thirdly, the statistical uncertainty caused by the limited number of simulated events is shown. It is clearly visible that the latter uncertainty dominates the total uncertainty. However, this uncertainty is based on the simulated samples that are currently available. This clearly shows that larger samples for the simulation are required in order to fully exhaust the statistical precision which will be reached with the expected data set of 100 fb^{-1} and even more. In the following, this uncertainty is omitted.

In figure 4.18 (right) statistical and systematic uncertainties on the best fit signal strength are shown. At 20 fb^{-1} the total uncertainty amounts to 54 $\%$. With a five times larger data set of 100 fb^{-1}, that is expected to be collected in the second running period of the LHC until 2018, the total uncertainty can be reduced by a factor of approximately 52 $\%$.

[3]The parameter is usually abbreviated by $\mu = \sigma/\sigma_\mathrm{SM}$ omitting the branching ratios.
[4]The systematic contribution is determined in a fit of $\sqrt{(a\,\mathcal{L})^{-1} + b^2}$ to the combined uncertainty.

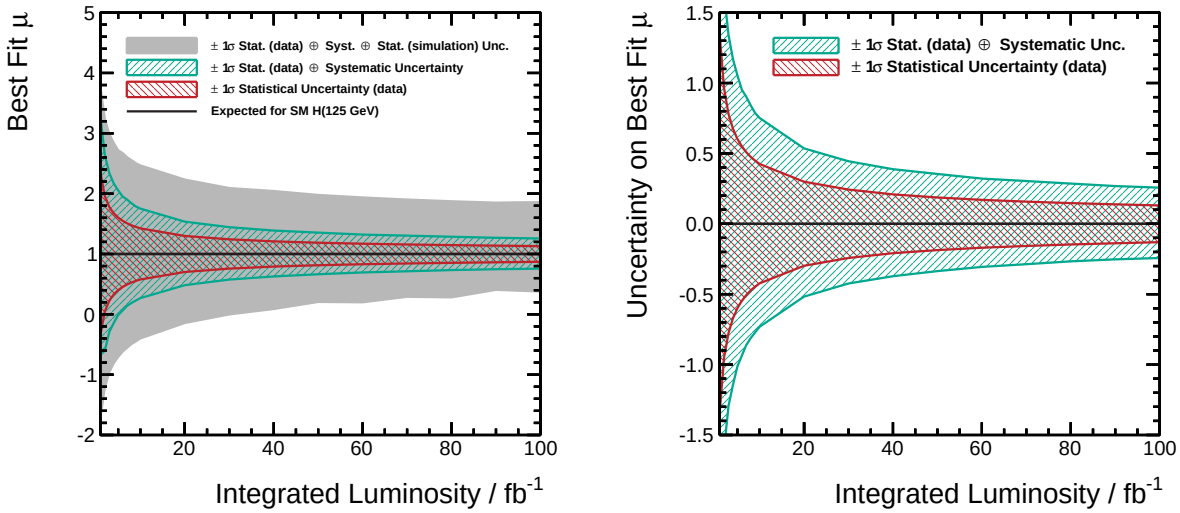


Figure 4.18.: Extrapolation of the signal strength as a function of the integrated luminosity. The fit combining the three channels $\mu\tau_{\mathrm{h}}$, $\mathrm{e}\tau_{\mathrm{h}}$ and $\mathrm{e}\mu$ is performed on an Asimov data set which follows the signal+background hypothesis. Therefore the fit always yields best values of one. Three kinds of uncertainties are distinguished: the statistical uncertainty of the data set, the systematic uncertainty and the statistical uncertainty of the simulated samples. The latter dominates the precision as shown left. As more data is recorded, more simulated events will be generated in order to reduce this uncertainty. Right shows the statistical uncertainty of the data set and the systematic uncertainty on the signal strength.

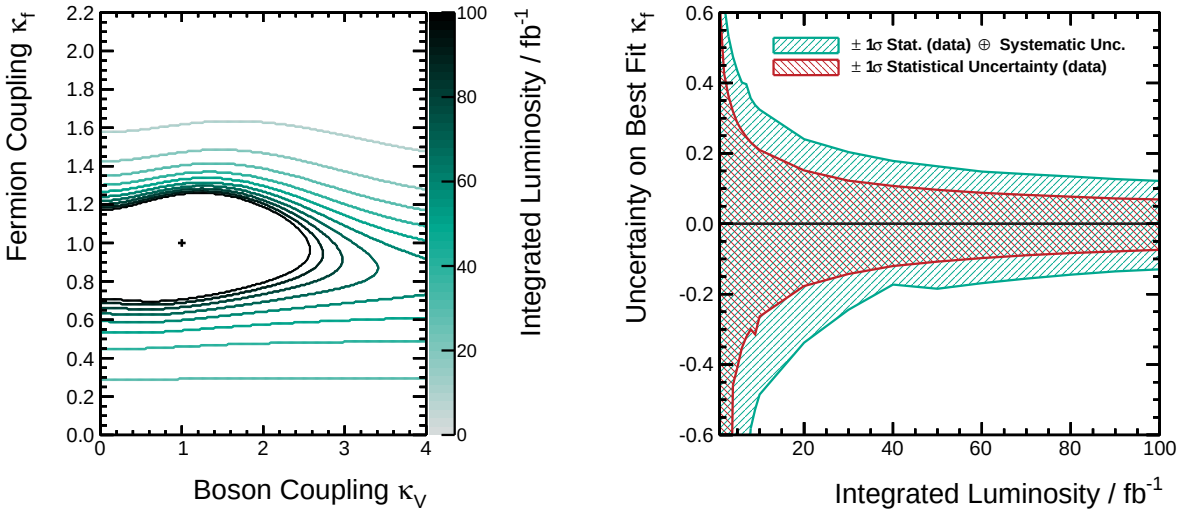


Figure 4.19.: Extrapolation of the fermionic and bosonic coupling modifiers as a function of the integrated luminosity. The 1 σ contour lines of a two-dimensional scan in the parameter space κ_{V}-κ_f is shown left. The precision for the measurement of the fermion coupling modifier, κ_f is shown right, which is determined in a separate fit where κ_{V} is allowed to freely vary.

Published results quote the uncertainties for integrated luminosities of 300 fb^{-1} at a centre-of-mass energy of 14 TeV, which covers the next decade of data-taking (see figure 2.14). They expect a precision of 14 % in the combined H $\rightarrow \tau\tau$ analysis [110] or 22 % in the VBF analysis [111] for similar conservative scenarios of the systematic uncertainties. The studies presented here yield a precision of 18 %, when they are extrapolated to 300 fb^{-1} at a centre-of-mass energy of 13 TeV. This is in good agreement with the published results.

As a cross check, the analysis presented here has been performed on the 8 TeV samples and a fit of the corresponding Asimov data set has been run. The resulting best fit signal strength has a precision of 48 %, which has to be compared with the 54 % for the analysis presented in this chapter. It would have been expected to observe a better performance based on the 13 TeV analysis compared to 8 TeV because of the signal production cross section, which is enhanced slightly more than the one for the main backgrounds. This indicates that the expected improvement is counteracted by the fact that this analysis is not yet sufficiently optimised. The same caveats already mentioned in section 4.4 have to be considered as the main reasons.

- The selection of events, especially the identification criteria for the leptons, is very loose resulting in a signal-to-background ratio that can further be improved by tighter selections of signal like objects.
- The distributions of the reconstructed di-τ mass reveal large tails towards higher values that deteriorate the discrimination power of this variable.

Finally, a comparison with the published result for the complete run I data set (25 fb^{-1} at 7 and 8 TeV) [10], which yielded a precision of 35 %, shows that an optimisation of the analysis inspired by the run I version can easily give a further relative improvement of at least 30 %. The most significant improvement in terms of sensitivity is expected from an optimisation of the VBF category. The analysis presented in this thesis employs a very loose definition of this category, whereas the full run I analysis is based on multiple VBF categories that are defined as tight as possible in order to maximise the sensitivity to the signal produced in the VBF production mode. Secondly, the addition of more channels, especially the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel which is characterised by the largest branching fraction, will add more sensitivity.

4.6.4. Measurement of Higgs Boson Couplings to Fermions and Vector Bosons

It has been pointed out that Higgs boson couplings to fermions, namely Yukawa couplings, are fundamentally different from Higgs boson couplings to vector bosons that are gauge couplings (see section 1.1.4). Therefore it is important to measure these couplings separately. The most straightforward way to perform this measurement is based on the assumptions that the scaling of Higgs boson couplings all fermions is the same and similarly for the vector bosons.

$$\kappa_f = \kappa_\tau = \kappa_{\mathrm{t}} = \dots$$
$$\kappa_{\mathrm{V}} = \kappa_{\mathrm{W}} = \kappa_{\mathrm{Z}}$$

Based on this parametrisation and the scaling defined in equation (4.1), the Asimov data set is fit to the signal-plus-background hypothesis with these two parameters of interests: κ_f and κ_V. The 1 σ contours of the negative log-likelihood are shown in figure 4.19 (left) for integrated luminosities from 10 to 100 fb^{-1} in steps of 10 fb^{-1}. A one-dimensional scan for the κ_f parameter is shown in figure 4.19 (right). In this fit, the parameter κ_V was considered as a free parameter. The systematic uncertainties contribute 8 % to the total uncertainty[4], whereas the statistical component evolves as expected proportional to $1/\sqrt{\mathcal{L}}$.

The 2D scan shows that the $H \to \tau\tau$ analysis is very important for constraining the fermionic Higgs boson couplings, whereas the sensitivity to couplings to vector bosons is limited. The latter sensitivity is mainly given by the VBF category. The 1D projection emphasises the expected improvement in precision for the measurement of the fermionic couplings. The relative improvement for the run II dataset of 100 fb^{-1} with respect to 20 fb^{-1} is approximately a factor of two and reaches a level around 12 % at 100 fb^{-1}.

Similar measurements can be performed looking at the Higgs boson production only. The results for the parameters μ_{ggH} (fermionic production) and $\mu_{qqH,VH}$ (production with vector bosons) are shown in the appendix in figure C.6.

4.7. Summary

In this chapter an outlook to the analysis of the upcoming 13 TeV data in the $H \to \tau\tau$ channel has been presented. The preliminary studies presented here have been based on the first 13 TeV data set of 71.52 pb^{-1}. The event selection for the four most sensitive channels ($\mu\tau_h$, $e\tau_h$, $e\mu$ and $\tau_h\tau_h$) has been demonstrated as well as the application of technique developed in the run I analysis like the reconstruction of the invariant di-τ mass. Extrapolations to higher integrated luminosities concluded this chapter.

For the analysis a new software framework has been developed (see appendix A), which is well prepared for the various studies that are planned in the field of di-τ final states and beyond. Its technical abilities have been demonstrated by the physics application in this chapter.

Conclusions

The search for Standard Model Higgs boson decays into pairs of τ leptons in the di-muon channel has been presented and put into the context of the complete $\mathrm{H} \to \tau\tau$ analysis, which combines all di-τ final states. The complete CMS run I dataset comprising integrated luminosities of 4.9 fb^{-1} at centre-of-mass energies of $\sqrt{s} = 7$ TeV and 19.7 fb^{-1} at 8 TeV, respectively, has been analysed.

In the scope of all di-τ decay modes of the $\mathrm{H} \to \tau\tau$ analysis, the same-flavour fully-leptonic channels $\mu\mu$ and ee are particularly challenging. Compared to other channels, the $\mathrm{H} \to \tau\tau \to \mu\mu$ channel suffers from the huge additional background from $\mathrm{Z} \to \mu\mu$ events, which accounts for more than 99 $\%$ of all background events after the inclusive event selection. In addition, the four neutrinos in the final state lead to a reduced di-τ mass resolution. Multivariate methods are necessary to sufficiently suppress the backgrounds and to bring the sensitivity of this channel to a level that is comparable to the $\mathrm{e}\mu$ channel, which gains from an intrinsic suppression of the $\mathrm{Z} \to \ell\ell$ background. In the combination of all fully leptonic di-τ states ($\mathrm{e}\mu$, $\mu\mu$, ee), the two same-flavour channels contribute 20 $\%$ to the sensitivity expressed in terms of expected upper limits on the Higgs boson production cross section. Both same-flavour channels follow the signal extraction method that has been developed in the $\mu\mu$ channel in the scope of this thesis. The sensitivity of the search was improved by 18 $\%$ at the Higgs boson mass hypothesis of 125 GeV with respect to the preliminary analysis.

No excess over the background-only hypothesis was seen in the $\mathrm{H} \to \tau\tau \to \mu\mu$ channel. Therefore, upper limits have been set. In the full $\mathrm{H} \to \tau\tau$ analysis, which is a combination of the analyses in all sub-channels, evidence for the $\mathrm{H} \to \tau\tau$ decays was found with a significance of more than 3 σ. All measurements are compatible with the Standard Model expectation within 2 σ. Finally Higgs boson couplings to τ leptons have been directly observed with a significance of more than 5 σ in the combination of the $\mathrm{H} \to \tau\tau$ searches of the ATLAS and the CMS collaborations [109].

Finding evidence for Higgs bosons coupling to fermions (and especially to leptons) reported on July 4$^{\mathrm{th}}$, 2012 marked a major step in probing the nature of the resonance observed at 125 GeV. Figure 5.1 shows the predicted mass-dependence of the Higgs boson couplings to elementary particles. At the lower boundary of the mass scale, τ leptons provide strong constraints on the proportionality of Higgs

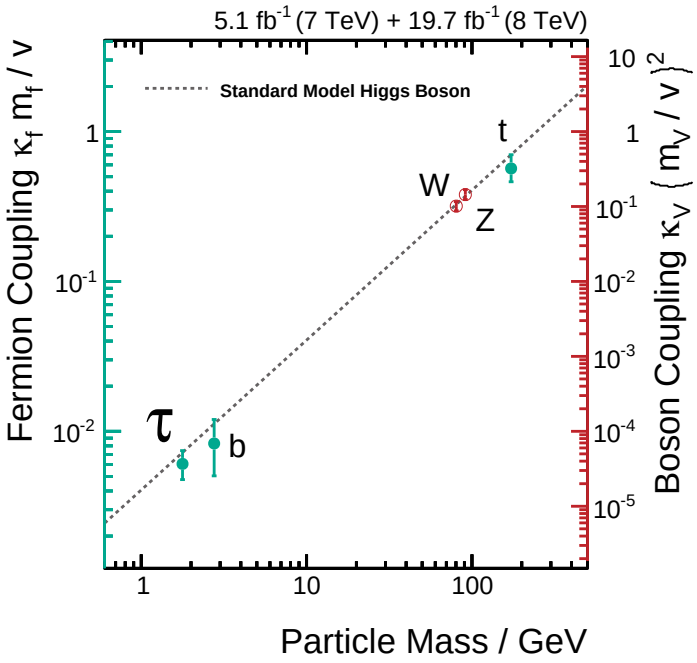


Figure 5.1.: Illustration of the Higgs mechanism based on the CMS run I analyses. The Higgs boson couples proportional to the mass of elementary fermions and proportional to the square of the mass of vector bosons. The values are taken from reference [26].

boson couplings to leptons. The coupling of the observed particle to leptons is proportional to their mass as it is predicted for the Standard Model Higgs boson. The coupling proportional to the square of the mass of vector bosons agrees fits in this picture. This provides strong evidence for the Higgs-nature of the 125 GeV resonance.

After the coupling of the Higgs boson to τ leptons has been established based on the data of the first LHC running period, the plans for the second running period turn towards precise measurements of these couplings. So far, all measurements are in good agreement with the Standard Model predictions. As both the instantaneous luminosity and the centre-of-mass energy of the run II collisions are increased with respect to run I, these measurements are expected to be performed with higher precision in order to either confirm the Standard Model hypothesis or to find deviations from it indicating physics beyond the Standard Model.

An introduction into a prototype $\mathrm{H} \to \tau\tau$ analysis for the CMS run II data and prospects for the expected measurement uncertainties have been presented by analysing the four main channels $\tau_\mathrm{h}\tau_\mathrm{h}$, $\mu\tau_\mathrm{h}$, $\mathrm{e}\tau_\mathrm{h}$ and $\mathrm{e}\mu$ final states. In the scope of this analysis, a new software framework has been developed (see appendix A). It is characterised by a modular design which allows experimentalists to use it for the multiple analysis needs in the future. Currently, this framework is already successfully employed by analysis groups in different fields of CMS physics, for example in the scope of the $\mathrm{H} \to \tau\tau$ analysis, as presented in this thesis, for the measurement of parton distribution functions and for jet energy calibrations.

The precision on the signal strength in the $H \to \tau\tau$ channel as well as on the coupling strength for fermionic couplings can be increased by a factor of two when the data set is increased from 20 fb^{-1} to 100 fb^{-1}. The upper boundary on the precision of the signal strength at 100 fb^{-1} in a conservative scenario, in which all systematic uncertainties are not reduced with respect to the 8 TeV analyses, is 26 %. Based on the same data set, fermionic couplings can be measured up to a precision of 12 %.

The $H \to \tau\tau$ channel offering unique chances for the measurement of properties of the Higgs boson is going to profit from the precision that is expected based on the complete CMS run II data set. Further analysis improvements will emphasise the relevance of this channel. Additionally, the di-τ final state is best suited for the search for additional Higgs bosons as they are predicted by two-Higgs-doublet models.

A

Development of Software Tools for Event-based Data Analyses

A.1. Common Workflow of High Energy Physics Analyses

Between the data taking at the detector or the generation of Monte Carlo events with the subsequent detector simulation and the showcasing of the physical results, the following analysis steps have to be performed.

The event reconstruction of the raw detector data is performed on large computing centres in the WLCG (see section 2.2.7). The reconstructed data is stored in different formats at multiple sites in the WLCG. Usually, the AOD or the miniAOD format is used for physics analyses. These formats contain as much information about the events as needed to serve the general needs of the analysis groups.

This step is followed by the skimming of data sets that are relevant for a specific analysis. Both the number of events as well as the event content is reduced in order to speed up the following processing of the data. By requiring data sets for certain high level triggers and applying cuts on event quantities, the number of events is reduced. Only the event quantities, that are relevant for a given analysis, are written out. The skimming step is usually performed on grid sites where the data is located. Sometimes, the data is also transferred to sites, where free computing resources are available. The output is an analysis-dependent n-tuple format that is stored at local file servers. The size of files needed for a certain analysis ranges from $\mathcal{O}(10)$ GB to few TB.

As an example, the $H \to \tau\tau \to \mu\mu$ analysis as presented in chapter 3, used a data set containing only events that were accepted by double-muon triggers. Additionally, there was no need to store reconstructed τ leptons (that decay hadronically) as there is no veto on taus like on electrons and third muons.

The skimmed n-tuples are finally analysed by user code on local computing resources. Events are selected and categorised, high-level calculations based on the physics objects are performed and the output is either again written to smaller n-tuples or histograms. The events are usually processed with different configurations depending on the analysis needs. The output is then used for post-processing

steps including the plotting of results. The post-processing does not necessarily require an event-based processing and therefore also needs on average the smallest computing resources.

A.2. The Artus Framework

The framework introduced in this section handles the last step, the user analysis. In many cases the code is developed without a strategy or structure according to the growing needs of the analysis. This leads to analysis code that is convoluted, tangled and unstructured. Such code is difficult to understand and to maintain and also difficult to hand over to new people joining the analysis effort. Secondly, code is rarely shared among different analysis groups, although basic parts of the analysis like the processing n-tuple-structured data or the selection and identification of physics objects follow the same principles. This usually leads to code duplication slowing down the analysis effort.

Following the structure of the CMSSW framework [46, 81], the Artus framework[1] [113] provides a front-to-end solution for the analysis of event-based data. The basic concept of modules processing the events within an event loop is complemented by tools for the configuration of every step. The complete structure is explained in the next section A.2.1.

This general framework is applicable to any kind of event processing analysis. The well-tested core is shared among various analyses, providing a reliable fundament for any new analysis. The highly modular structure guides the user to develop analysis software that is easily maintainable and constructively extensible. Thus, thes Artus framework avoids the aforementioned problems of separate user code for individual analyses.

The Artus framework is written in C++. The core of the framework only depends on the boost library[2]. Further dependencies may be introduced by the actual analysis code. C++ is chosen for performance reasons and in order to be able to integrate existing frameworks and libraries that are commonly used in high energy physics.

A.2.1. Structure of the Framework

The basic concept of the framework is taken from the code used for the $Z + \text{jet}$ analysis documented in reference [92]. It is generalised and extended to satisfy the needs of various analysis groups. The structure is illustrated in figure A.1.

At the beginning a configuration file in the JSON format is read in. It contains information on the input data and settings for the pipelines and processors described in the following. Each of the modules described in the following has access to the settings read in from the configuration file.

[1] https://github.com/artus-analysis/Artus
[2] http://www.boost.org/

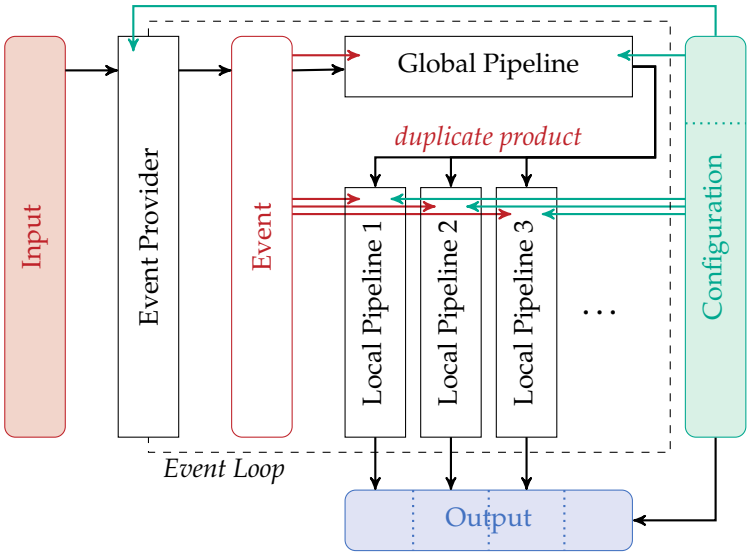


Figure A.1.: Structure of an Artus analysis. The input is read by an event provider. Within the pipelines the event content is analysed by the processors. Consumers in local pipelines write results to a common output. All parts of the analysis are configurable.

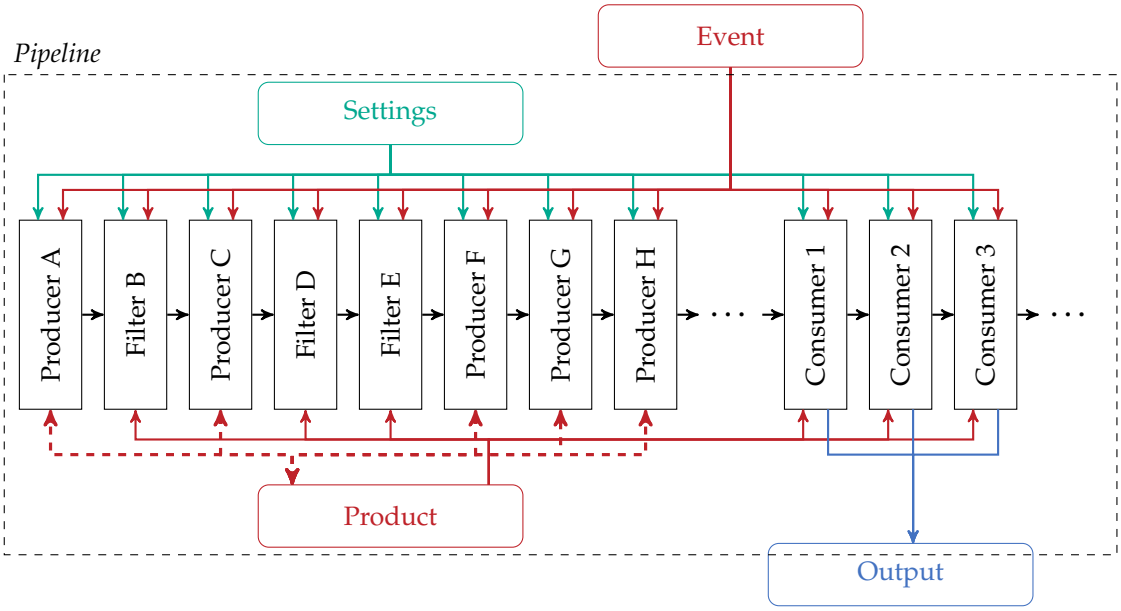


Figure A.2.: Organisation of processors in a pipeline. Producers and filters are run before the consumers. All processors can access the pipeline settings, the event content and the product. Only producers have write access to the product. Consumers can write results to the output file.

The input is handled by an event provider. Its purpose is to manage reading events contents from any kind of input into an event object and to provide information about the total number of events to be processed.

The events are then sequentially processed by processors organised in pipelines. First, a so-called global pipeline is run. Then the execution is split up into multiple so-called local pipelines that are also processed sequentially. The composition of a pipeline is illustrated in figure A.2.

Each pipeline has its own settings. The local pipelines can access both their own settings and the ones of the global pipeline. Thus, a mechanism exists that enables processing the same event with different configurations, whereas the costly reading of the input is only needed to be done once. Similarly, a product object is managed by each pipeline. The product is meant to store new quantities produced in the analysis. The product for the global pipeline is empty at the beginning of the processing of a new event. Local pipelines start with a copy of the global product.

A pipeline consists of an arbitrary number of processor modules that are run in a sequence. Three kinds of processors are distinguished:

Producers are meant to calculate new quantities based on the information in the event and the product and following the settings of the pipeline.

Filters apply cuts. In case an event does not survive a requirement defined by a filter, the processing of the subsequent producers and filters is skipped.

Consumers can be added to local pipelines. They are executed after the producers and filters. The task of consumers is to take quantities from the event and the product and store them in the output according to the pipeline settings.

Producers and filters can be configured in an arbitrary order. Consumers are executed after the chain of producers and filters. The output together with the complete configuration is written to a ROOT file. These processors are created and managed by a dedicated factory according to the configuration.

This structure extends the established concepts of existing frameworks like FWLite [114] from CMSSW to the following four main advantages:

Modular structure Analysis steps are implemented in processors, breaking down the complete analysis in well-defined and easily understandable and writeable parts.

Re-utilisation of code Processors can be shared among different analysis groups. Every user contributes to the testing of existing code and improves the confidence in this code.

Configurability Each module uses settings that are read from the configuration. Each pipeline is mapped to a separate sub-set of the configuration.

High performance The concept of pipelines allows to analyse the data with different configurations at the same time, while the data is only read in only once and common processing steps are shared in the global pipeline. Therefore the cost-intensive input operations are reduced to a minimum.

A.2.2. Building an End-user Analysis

In the Artus framework each concrete analysis is defined by a set of object types. It needs an implementation of the classes for settings, the event and the product. The settings class defines the tags that should be able to be read from the configuration file and the corresponding types of the values. The event class defines the event content being read from the input and the product class contains members for new quantities being calculated by the analysis. These three classes need to be derived from the base classes in the core of the framework.

An implementation of an event provider must be available. The event provider must know how to read event content from the input files and how to store them in the event object. Furthermore, a factory managing the creation of analysis-specific processors needs to exists. These classes also have to be derived from the Artus base classes.

The most important part of an analysis are the processors. The abstract processors of the Artus core need to be implemented to realise the physics needs of the analysis in the form of code. New processors have to be registered in the factory. After the basic structure of an analysis is set up, most of changes and additions are only related to the processors.

Analyses which share the input file format can easily also share processors that implement basic analysis steps in a general fashion. Analysis-specific parts can either derive from existing processor classer or be modelled in new processors.

A.3. HarryPlotter – A Python Post-processing Framework

The computationally expensive event processing steps are followed by the post-processing of the results and their presentation. From a technical point of view the focus here is rather on flexibility than on performance. The inputs for this step are either histograms filled in the previous step or n-tuples whose entries can be filled into histograms at low cost. For this reason, Python has been chosen as language for the HarryPlotter framework[3]. As the name implies, the main focus is on plotting of results. The modular structure allows to integrate multiple analysis steps and several output formats.

The structure of the framework is illustrated in figure A.3. It follows the idea of processors in the Artus framework. Three kinds of processor modules are executed subsequently:

1. One or, in special cases, more input modules are run to read in data from an input source. The input format is defined by the input module. The most commonly used input module is the one that reads histograms or other objects from ROOT files.

2. An arbitrary number of analysis modules is appended. They can perform calculations based on the inputs or the results of previously run analysis modules.

[3] https://github.com/artus-analysis/Artus/tree/master/HarryPlotter

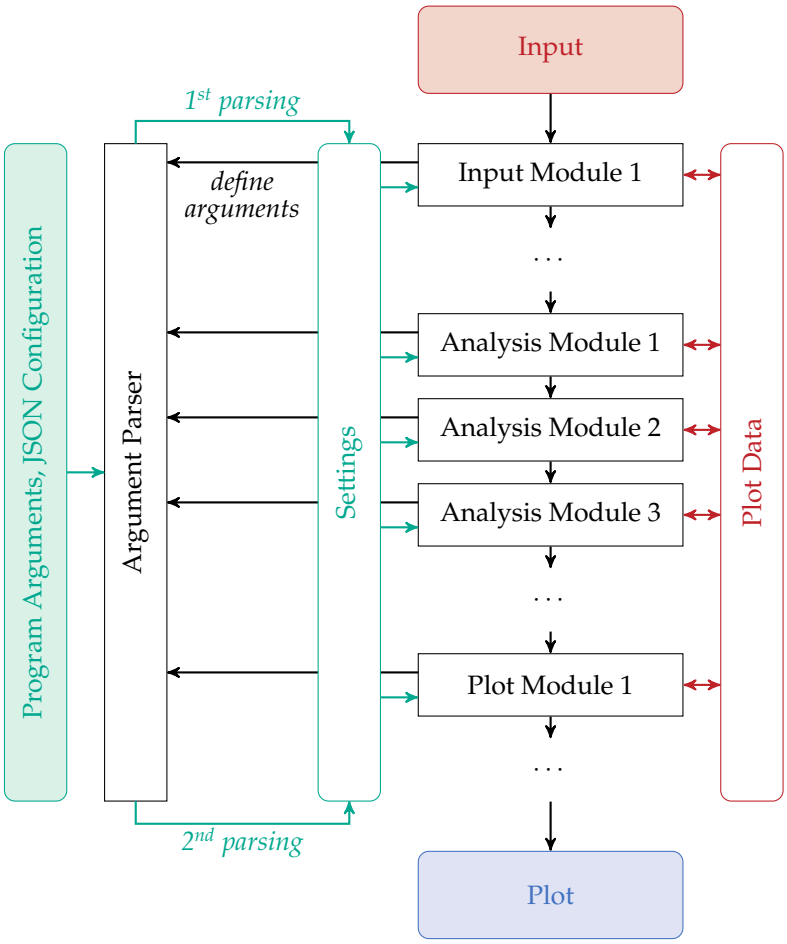


Figure A.3.: Structure of the HarryPlotter framework. The processing of input modules is followed by the run of analysis modules and concluded by the execution of plot modules. The configuration via a command line interface or via JSON files is parsed by an argument parser whose arguments are defined by the individual modules.

3. One or, in special cases, multiple plot modules are executed in the last step. Their function is to output the processed data in the specified format. In most of the cases, the output is a plot but it is also possible to store the objects in new ROOT files.

All processors share common plot data objects and possible meta-data where they can add new information and read the existing one. For example, the input modules add histograms read in from input files and the plot modules access these histograms.

The configuration is provided by an argument parser that can handle both arguments from a command line interface and content from JSON configuration files or Python dictionaries. This makes it possible to use either the HarryPlotter executable configured via the command line or call the main function within another Python script. The executable can perform one single plot per run. For the execution within a script functions exist to create multiple plots in parallel.

The parsing of the arguments is performed in two steps. The first time it determines the sequence of modules to be executed. These modules are then initialised and can define additional arguments

needed for their configuration, such that every module can manage its own settings but can also access the settings of entire program. After all arguments are defined the configuration is parsed a second time providing the full settings for the HarryPlotter run. Then all processors are executed as described above.

A.4. A Real World Application – the $H \to \tau\tau$ Analysis

Currently, the Artus framework is used for various user analysis of CMS data ranging from jet energy calibration and QCD studies to analyses in the scope of Higgs boson measurements and searches for physics beyond the Standard Model. As an example, the technical aspects of the $H \to \tau\tau$ analysis presented in this thesis are explained in the following.

The $H \to \tau\tau$ analysis developed in the scope of this thesis makes use of skimmed data sets in the Kappa format. The Kappa framework[4] is designed for skimming CMS data into small n-tuples that can be analysed independently of the CMSSW software. The only dependence is the ROOT framework. The n-tuples store high-level information such as four-momentum vectors, collections of leptons and vertices. Similar as the Artus framework, Kappa aims to satisfy general analysis needs.

A large part of the Artus software is specialised on the analysis of Kappa n-tuples[5]. An event provider is available for reading Kappa n-tuples. The event class contains members for all important physics objects used in analyses. The branches of the n-tuples, that should be read in from the input files, are configurable. Reducing the amount of information transferred from the input files accelerates the analysis significantly.

The structure of the pipelines is chosen such that for each decay channel a separate pipeline exists. Each of the pipelines is duplicated and modified in order to implement the different settings for shifts of quantities affected by systematic uncertainties. For example, the τ energy scale is shifted up and down with respect to the nominal value. The change of the τ four-momentum is then propagated through the entire analysis and the effect on the final result is studied. For this, each pipeline for decay channels involving hadronically decaying τ leptons is duplicated three times and the setting for the τ energy scale shift is modified accordingly. The global pipeline is used to perform actions that are common for all channels. This is for example the identification of jets and their correction as well as the filtering of valid data events.

The producers and filters are implemented to perform small tasks of the analysis. Main producers are the ones for the identification and the selection of valid leptons and jets. Further producers perform the matching with trigger or generator objects and do the splitting into decay channels or the categorisation. Also the mass reconstruction is done in a separate producer. The selection of events according to high level trigger decisions as well as the selection of decay channels are examples for filters. The filters are executed as early as possible in the chain of processors in order to avoid running subsequent processors without real need.

[4] `https://github.com/KappaAnalysis/Kappa`
[5] `https://github.com/artus-analysis/Artus/tree/master/KappaAnalysis`

Table A.1 shows a runtime measurement for the baseline part of the $H \to \tau\tau$ in the four main channels ($\mu\tau_h$, $e\tau_h$, $e\mu$ and $\tau\tau$). It illustrates the main advantage of the Artus structure: sub-analyses with different sets of configurations can be performed in the same run avoiding the multiple IO overhead when the configurations would be performed sequentially in different runs.

Table A.1.: Measurement of the runtime of the baseline $H \to \tau\tau$ analysis. The analysis is performed on 10 files with a total size of 6.1 GB containing 452229 simulated Z+jets events. The measurement is performed with different numbers of pipelines, where each pipeline implements the analysis of one sub-channel of the analysis. The values are compared with the execution without any pipelines and without global processors. It is clearly visible that the time needed for reading the inputs amounts to a large part of the overall runtime and that the simultaneous execution of multiple pipelines can reduce the runtime significantly compared to the case where different configuration are performed sequentially including separate I/O operations.

Pipelines	**Runtime / min**	**Event Rate / s^{-1}**
0	5:12	1470
1 ($e\mu$)	6:31	1174
1 ($e\tau_h$)	7:26	1028
1 ($\mu\tau_h$)	7:14	1052
1 ($\tau\tau$)	6:07	1247
2 ($\mu\tau_h$, $e\tau_h$)	8:32	896
3 ($\mu\tau_h$, $e\tau_h$, $e\mu$)	9:54	767
4 ($\mu\tau_h$, $e\tau_h$, $e\mu$, $\tau\tau$)	10:01	758

The outputs are flat n-tuples suitable for performing the background modelling step done in HarryPlotter. Histograms are read while cuts for the final event selection are applied. As an example, the cut on the transverse mass is applied on this level because events from control regions needed to be preserved for the background modelling step. For each background process, an analysis module has been developed performing the estimation of the samples. Most of them perform calculations on multiple histograms, defining the yield and the shape of the final quantity in signal and control regions. They add a single histogram to the plot data object containing the final estimation for a given sample. Similarly, ratios comparing the compatibility of data and the simulation for the sub-plots are determined and added this way.

The post-processing step either outputs the plots shown in this thesis or writes the histograms to ROOT files and passes them over to the limit calculation tools.

B

Supporting Material for the $\mathrm{H} \to \tau\tau \to \mu\mu$ Analysis

Table B.1.: Event yields after the pre-selection and the categorisation for data, the background processes and the signal for the 7 TeV analysis. The number of signal events are given for a Higgs boson mass hypothesis of 125 GeV. It is clearly visible that the $\mathrm{Z} \to \mu\mu$ background deserves the most crucial treatment since it is orders of magnitude larger than the other backgrounds. It is also noticeable that the signal events are to more enriched with $qq\mathrm{H}$ events the higher the jet multiplicity is.

Process	**0-jet**		**1-jet**		**2-jet**
	low-pt	**high-pt**	**low-pt**	**high-pt**	
Data	**266365**	**982442**	**18757**	**234390**	**33186**
$\mathrm{Z} \to \mu\mu$	260827	979322	17805	230852	31981
$\mathrm{Z} \to \tau\tau$	4923	1313	704	586	216
$\mathrm{t\bar{t}}$+jets	2	32	45	592	265
$\mathrm{WW} + \mathrm{WZ} + \mathrm{ZZ}$	124	858	91	1682	766
QCD + W+jets	140	143	48	188	48
Sum Background	**266017**	**981668**	**18694**	**233899**	**33275**
$\mathrm{gg}\mathrm{H}$	8.03	5.50	2.53	3.71	1.44
$qq\mathrm{H}$	0.06	0.06	0.41	1.38	0.22
$\mathrm{VH} + \mathrm{t\bar{t}H}$	0.086	0.33	0.32	1.93	0.71
Sum Signal	**8.17**	**5.89**	**3.26**	**7.01**	**2.37**
$S/\sqrt{S+B}$	0.016	0.006	0.024	0.015	0.013

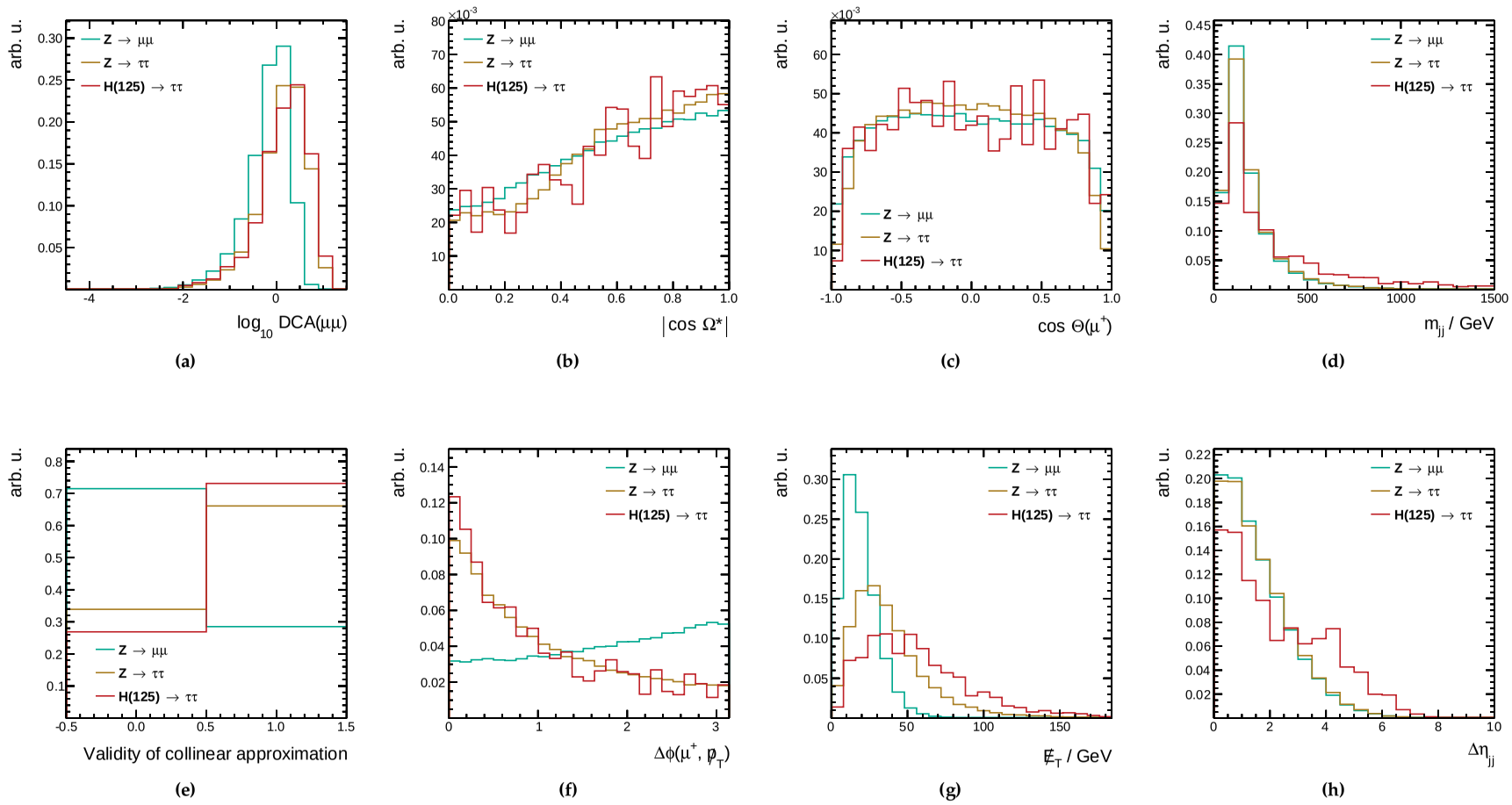


Figure B.1.: Discriminating variables used in the BDT trainings for the two jet category. The distributions normalised to unity of the most dominant background events, $Z \to \mu\mu$ and $Z \to \tau\tau$, are compared to the one of the $H \to \tau\tau$ signal.

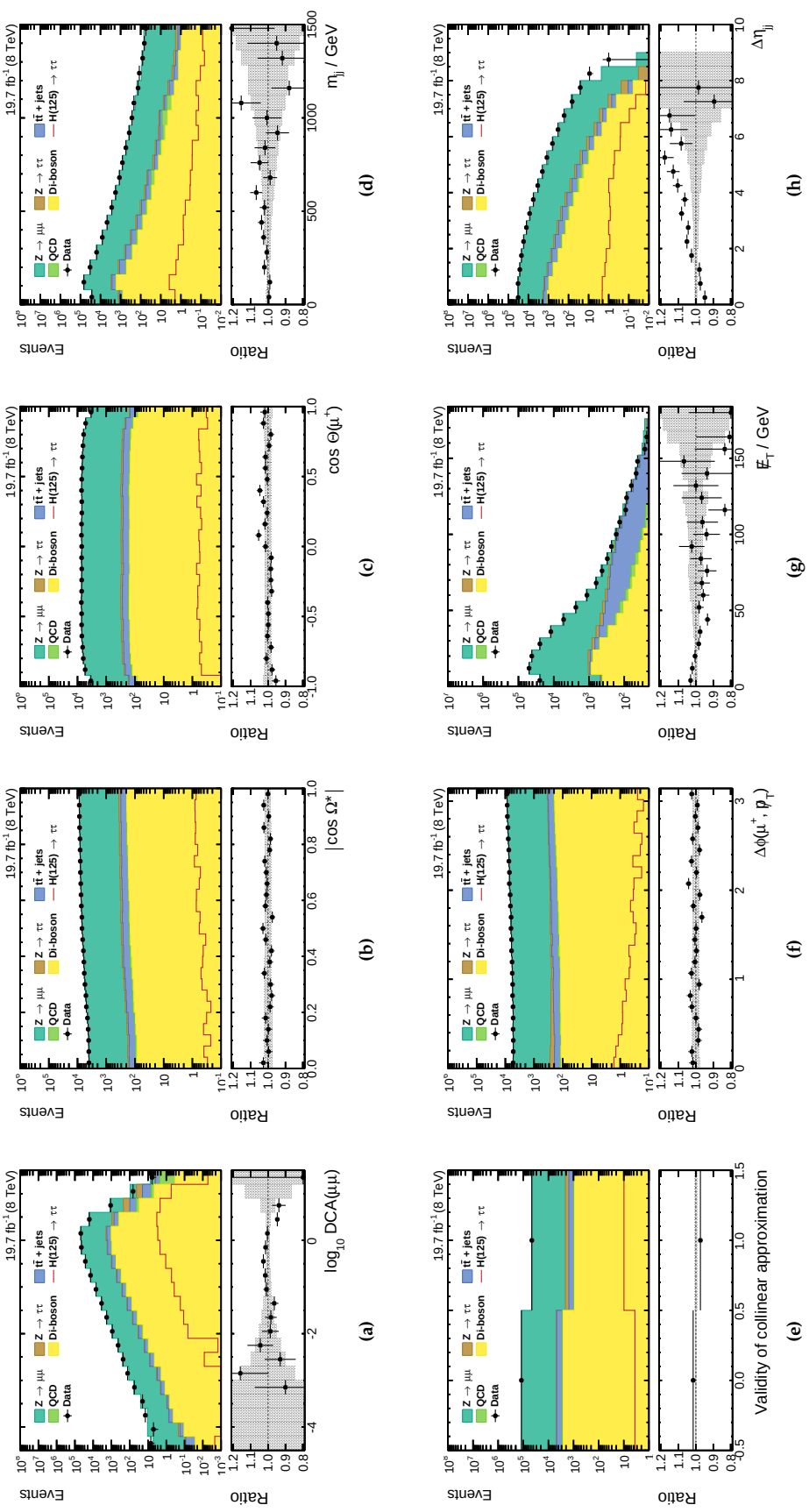
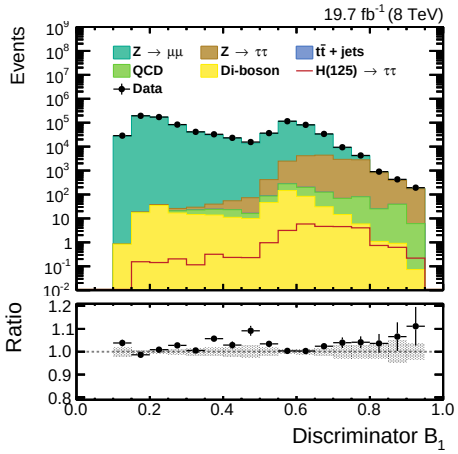
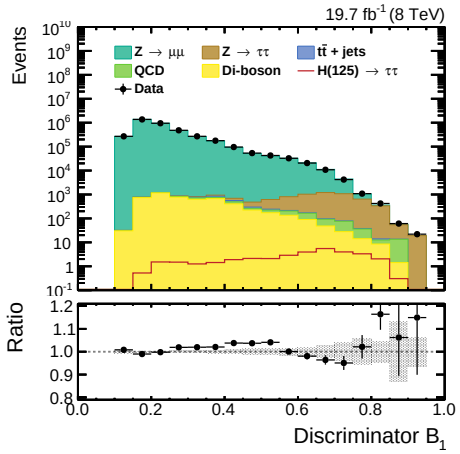


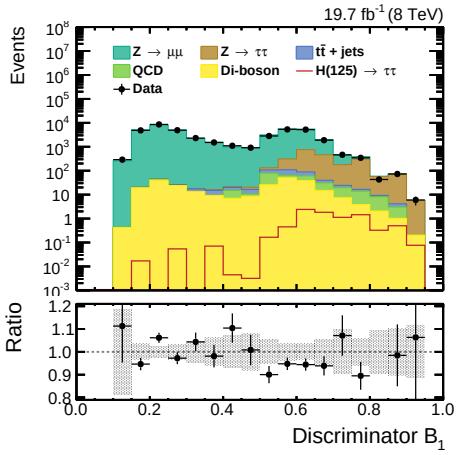
Figure B.2.: Distributions of the discriminating variables used in the BDT trainings for the two jet category. The ratio compares the observation in data to the expectation given by the sum of all backgrounds. The statistical errors are shown in the error band. The discrepancies between the observation and the background-only expectation are below 5 % in almost all distributions. These are addressed by the background estimation studies based on the resulting discriminator.



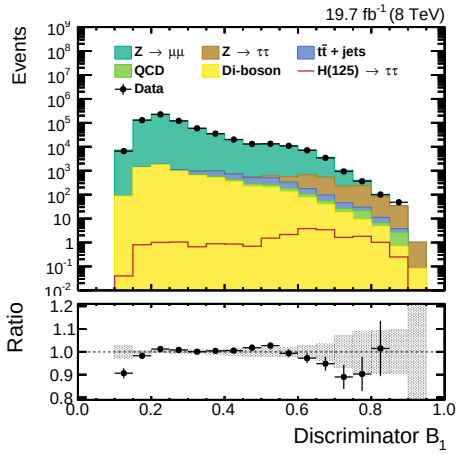
(a) 0-jet low-pt category



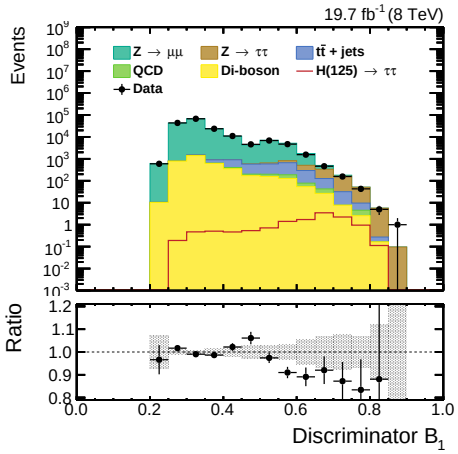
(b) 0-jet high-pt category



(c) 1-jet low-pt category

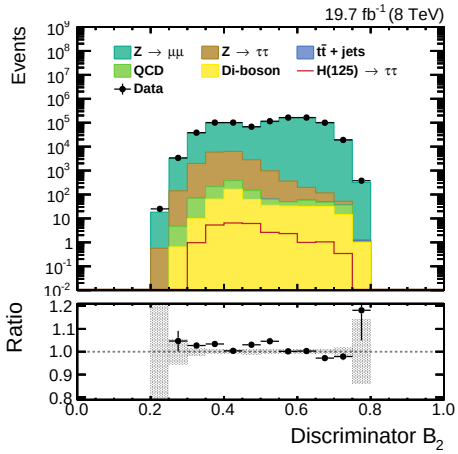


(d) 1-jet high-pt category

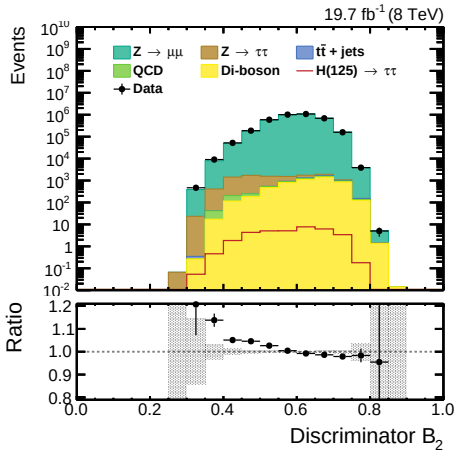


(e) 2-jet category

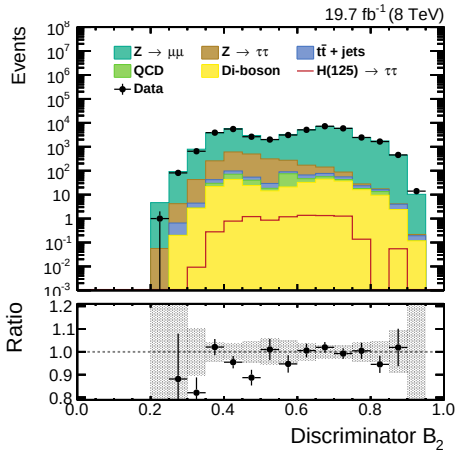
Figure B.3.: Distributions of the first stage BDT in the 8 TeV analysis, discriminating between $\tau\tau$ final states and mainly the $Z \to \mu\mu$ background. The ratio compares the observation in data to the expectation given by the sum of all backgrounds. The statistical errors are shown in the error band.



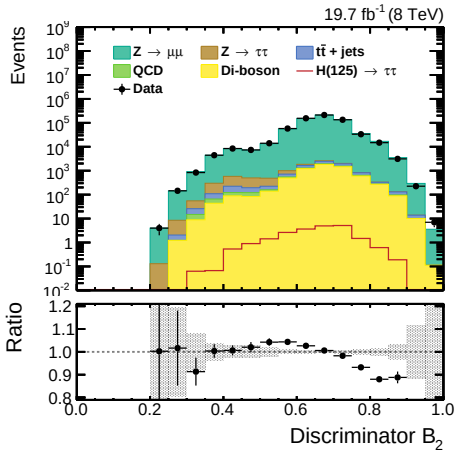
(a) 0-jet low-pt category



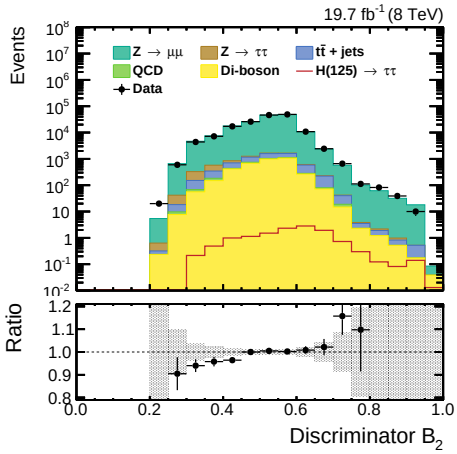
(b) 0-jet high-pt category



(c) 1-jet low-pt category

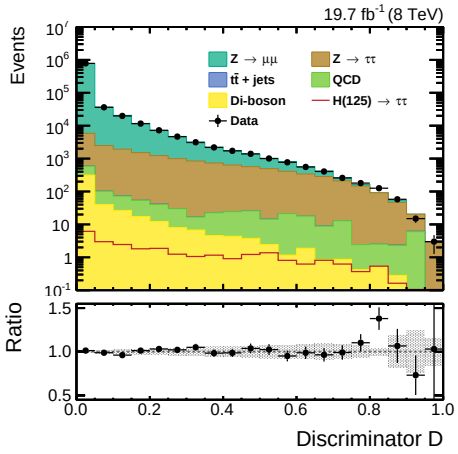


(d) 1-jet high-pt category

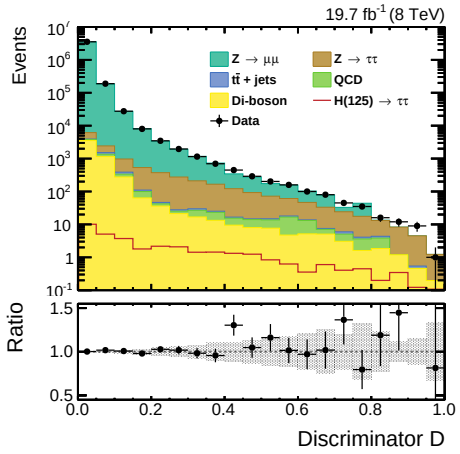


(e) 2-jet category

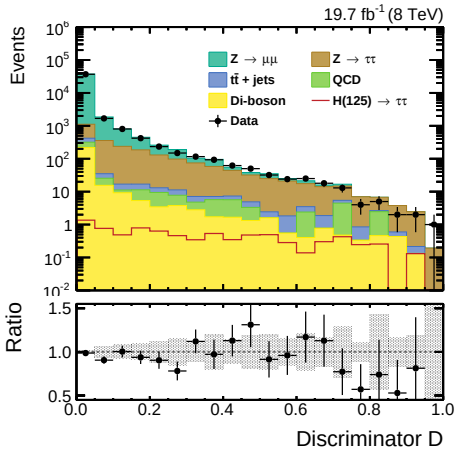
Figure B.4.: Distributions of the second stage BDT in the 8 TeV analysis, which is optimised on the discrimination between H $\to \tau\tau$ signal and Z $\to \tau\tau$ background. The ratio compares the observation in data to the expectation given by the sum of all backgrounds. The statistical errors are shown in the error band.



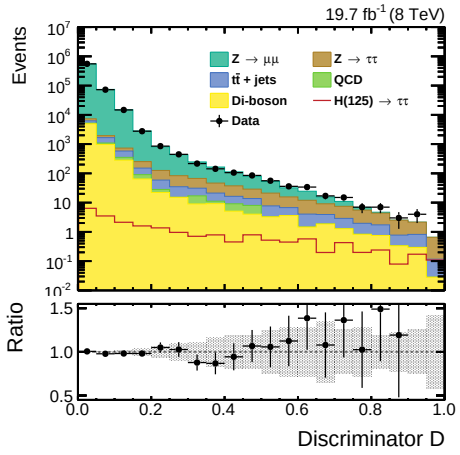
(a) 0-jet low-pt category



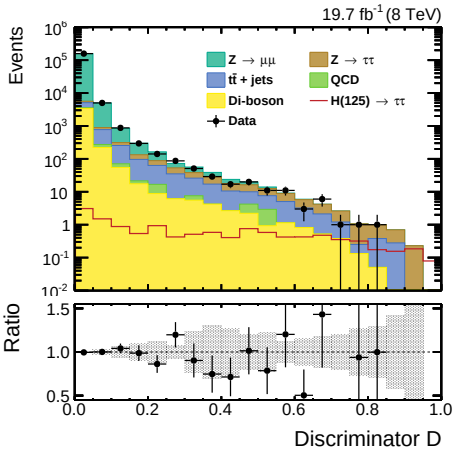
(b) 0-jet high-pt category



(c) 1-jet low-pt category

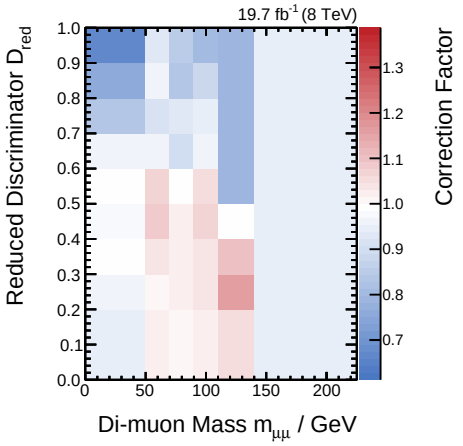


(d) 1-jet high-pt category

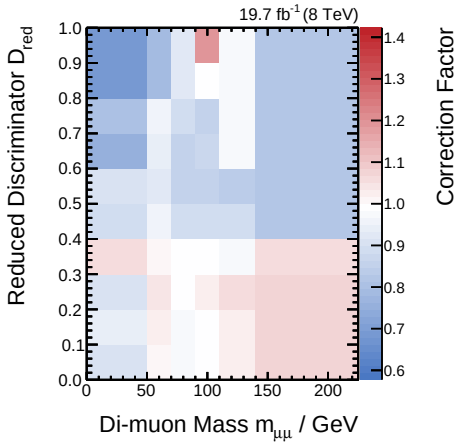


(e) 2-jet category

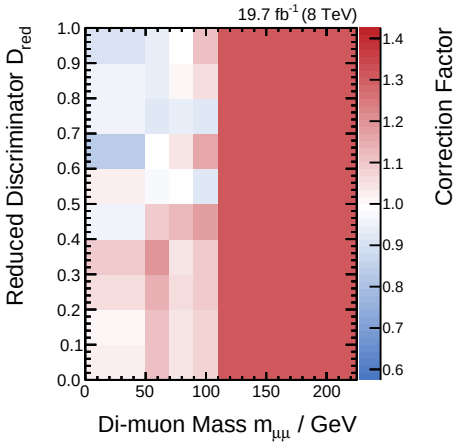
Figure B.5.: Distributions of the final discriminator, D, in the 8 TeV analysis which is an analytical combination of the two BDT discriminators, B_1 and B_2. The ratio compares the observation in data to the expectation given by the sum of all backgrounds. The statistical errors are shown in the error band.



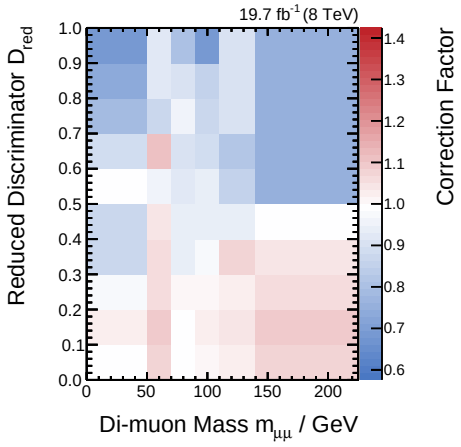
(a) 0-jet low-pt category



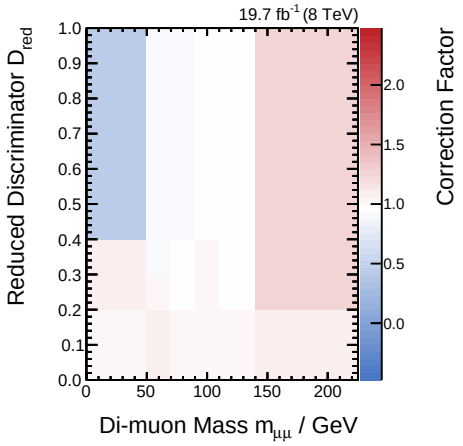
(b) 0-jet high-pt category



(c) 1-jet low-pt category

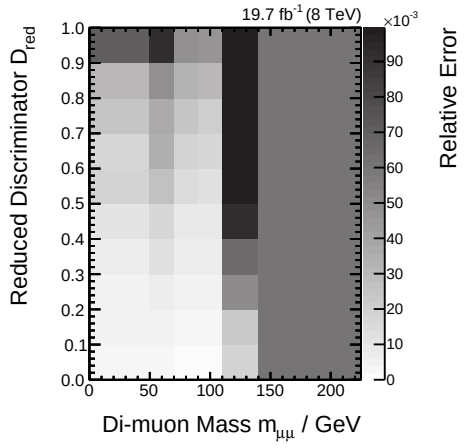


(d) 1-jet high-pt category

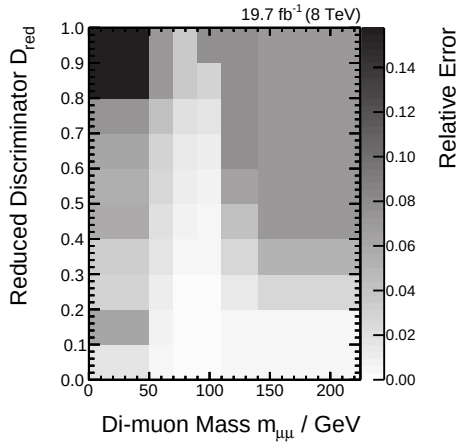


(e) 2-jet category

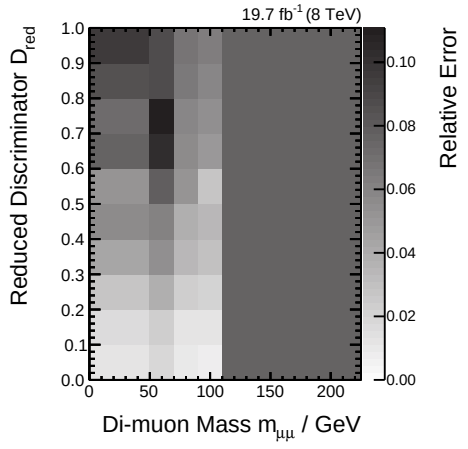
Figure B.6.: Complete set of $Z \to \mu\mu$ correction factors for the analysis of 8 TeV data, that are derived from the DCA template fits.



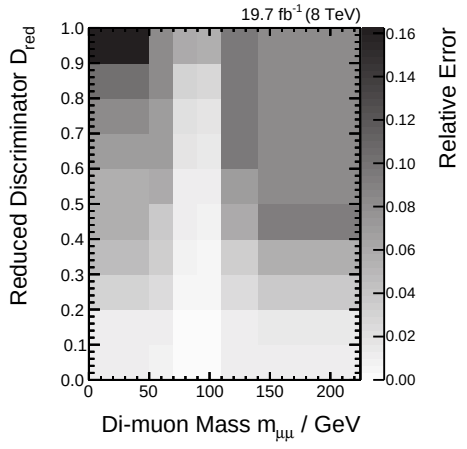
(a) 0-jet low-pt category



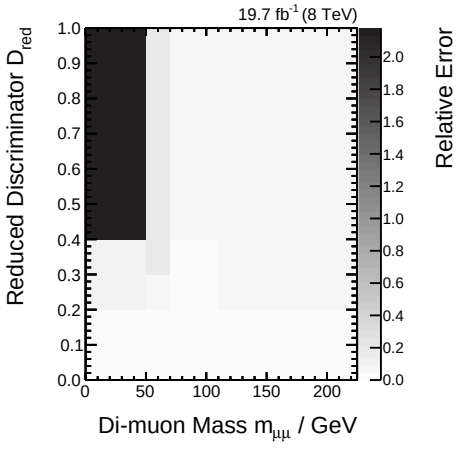
(b) 0-jet high-pt category



(c) 1-jet low-pt category

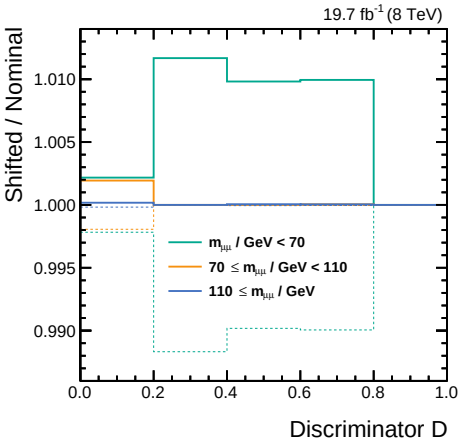


(d) 1-jet high-pt category

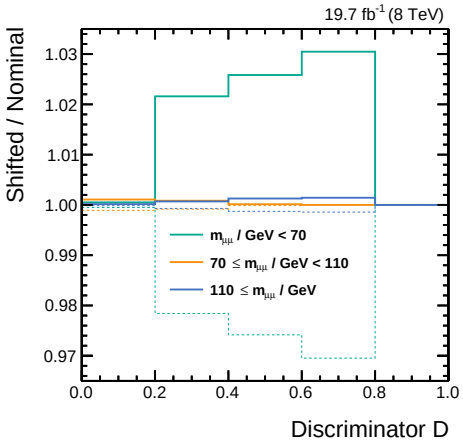


(e) 2-jet category

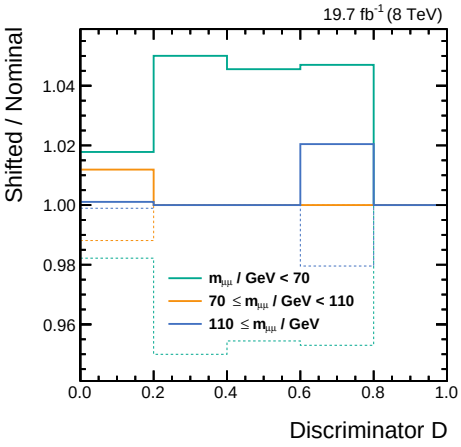
Figure B.7.: Complete set of statistical uncertainties on the $Z \to \mu\mu$ correction factors for the analysis of 8 TeV data, that are derived from the DCA template fits.



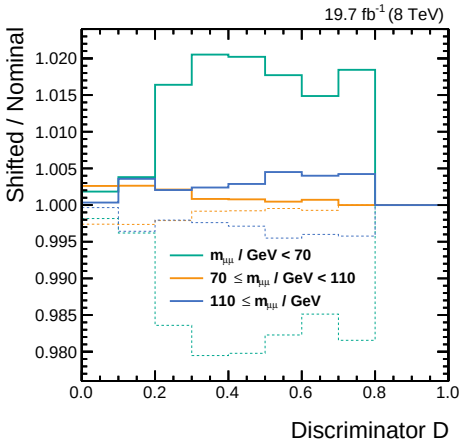
(a) 0-jet low-pt category



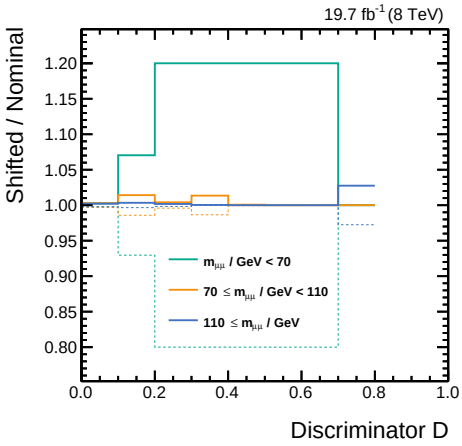
(b) 0-jet high-pt category



(c) 1-jet low-pt category

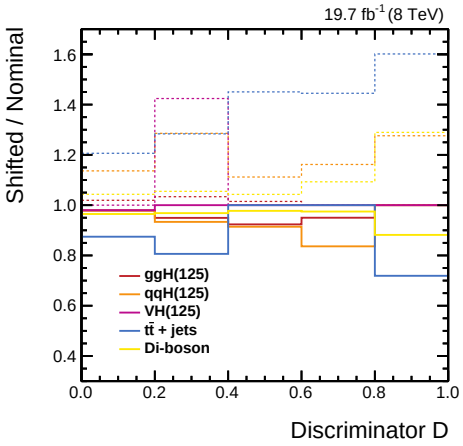


(d) 1-jet high-pt category

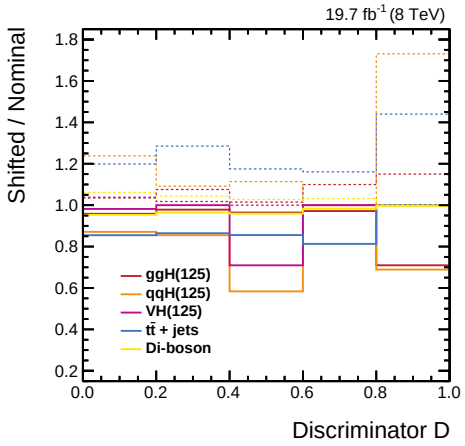


(e) 2-jet category

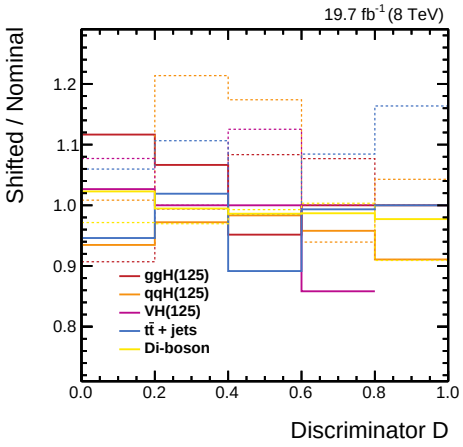
Figure B.8.: Effect of the Z $\to \mu\mu$ shape variation by 1 σ up (solid) and down (dashed) in the 8 TeV analysis, that is assigned in bins of the di-muon mass in order to allow for an altering of the discriminator shape in regions of phase space with different signal-to-background ratios The uncertainties grow with the the value of the reduced discriminator because of the steeply falling distribution for the Z $\to \mu\mu$ background.



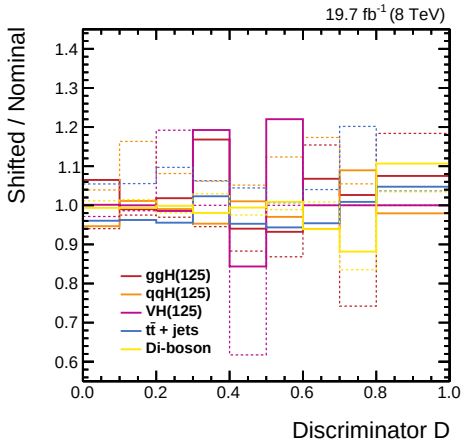
(a) 0-jet low-pt category



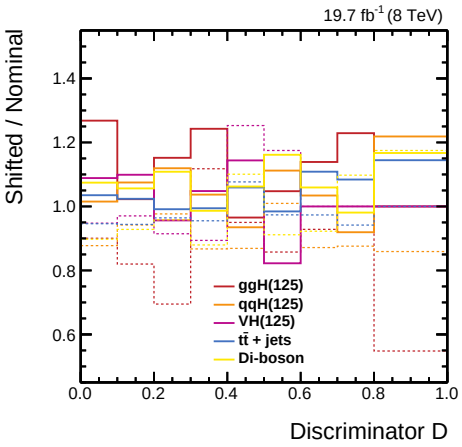
(b) 0-jet high-pt category



(c) 1-jet low-pt category



(d) 1-jet high-pt category



(e) 2-jet category

Figure B.9.: Effect of the jet energy scale variation by 2 σ up (solid) and down (dashed) in the 8 TeV analysis. The most inportant influence on the categorisation is visible in the fact, that the upward fluctuation leads to a smaller yield in the 0-jet categories and to a higher yield in the 2-jet category.

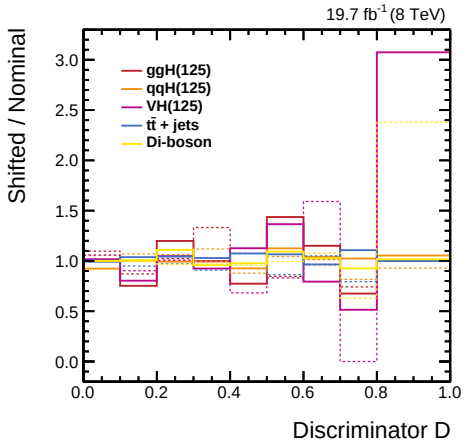
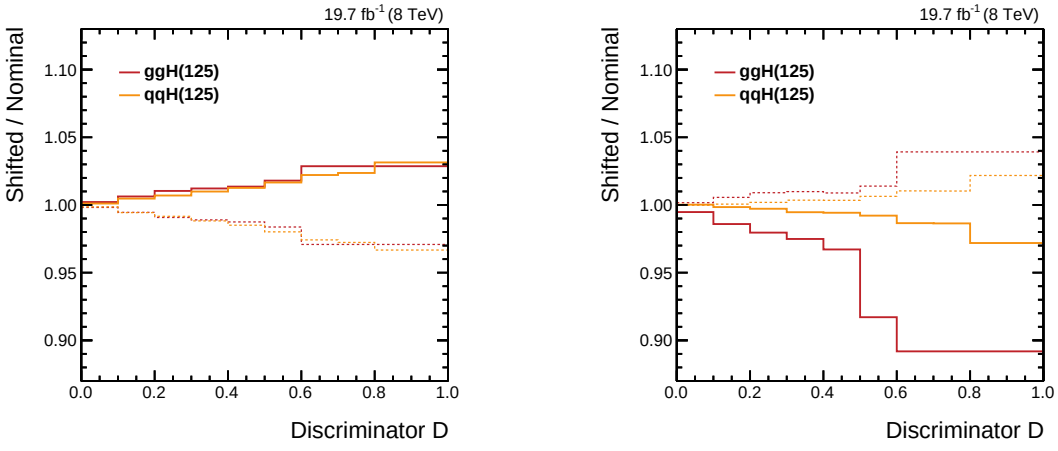
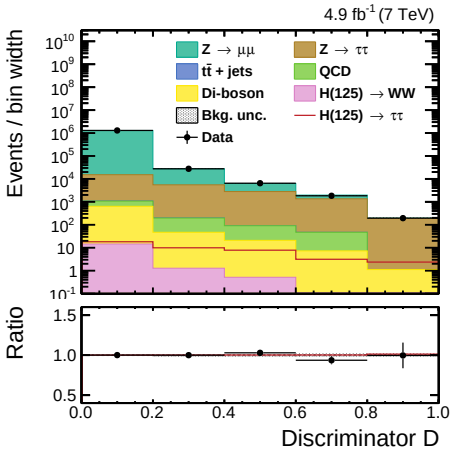


Figure B.10.: Effect of the MET scale variation by 2 σ up (solid) and down (dashed) for the considered processes in the 2-jet category in the 8 TeV analysis.

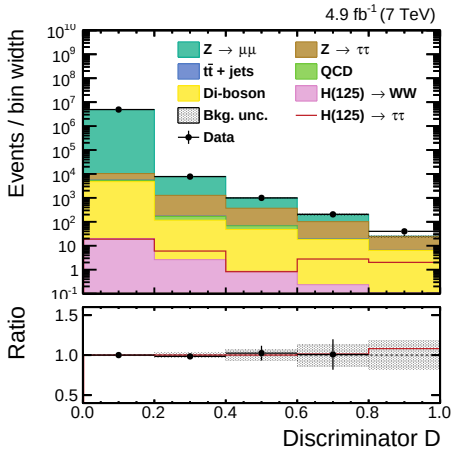


(a) PDF **(b)** Scale

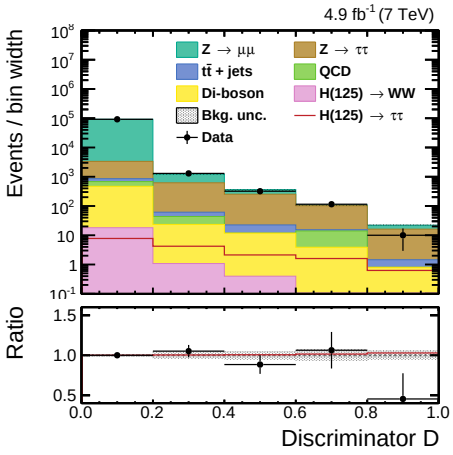
Figure B.11.: Effect of variation of the PDF scale (left) and the QCD scale (right) by 1 σ up (solid) and down (dashed) for the signal processes in the 2-jet category in the 8 TeV analysis.



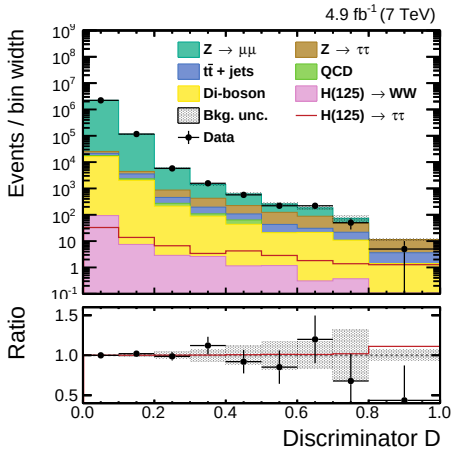
(a) 0-jet low-pt category



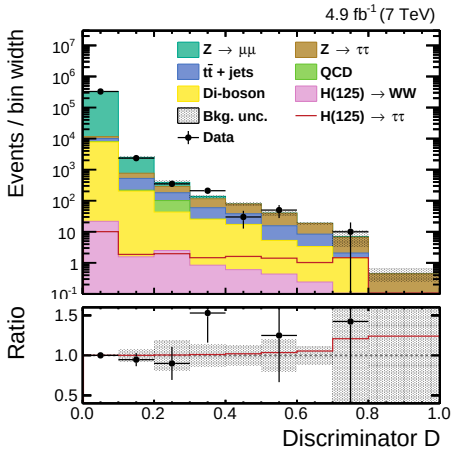
(b) 0-jet high-pt category



(c) 1-jet low-pt category

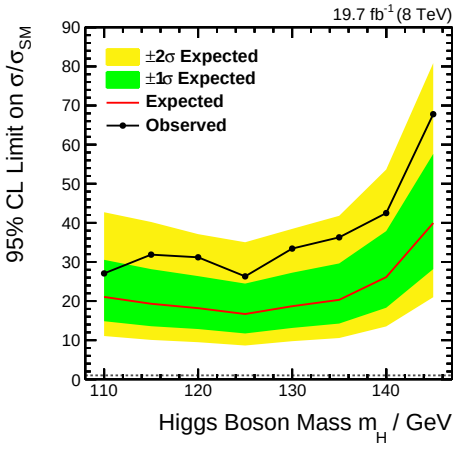


(d) 1-jet high-pt category

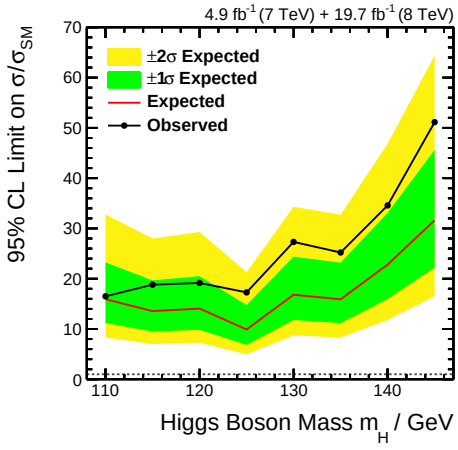


(e) 2-jet category

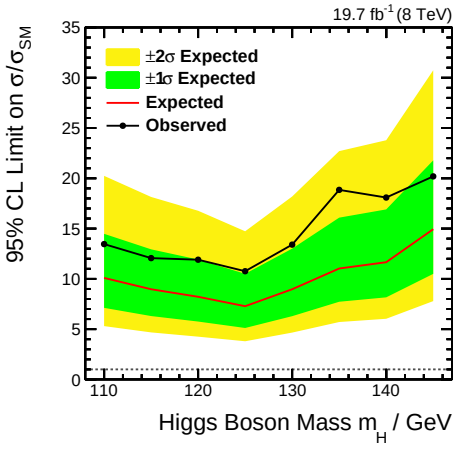
Figure B.12.: Post-fit distributions of the final discriminator D in the five event categories for 7 TeV data. The ratio compares the observation in data to the expectation given by both the background-only and the signal-plus-background for a Higgs boson mass of 125 GeV hypotheses. The total post-fit errors are shown in the error band. The observation is compatible with both hypotheses in most of the bins.



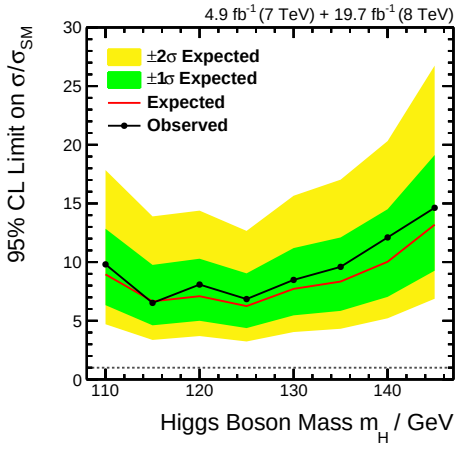
(a) 0-jet, 8TeV



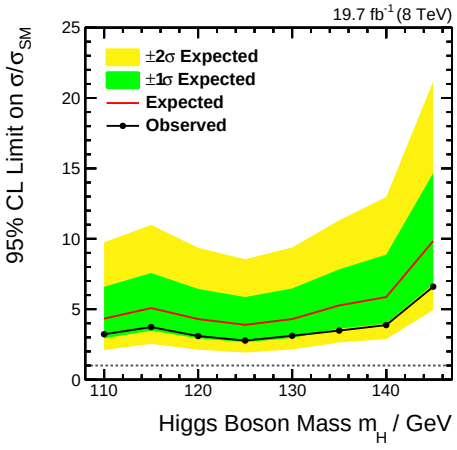
(b) 0-jet, 7+8TeV



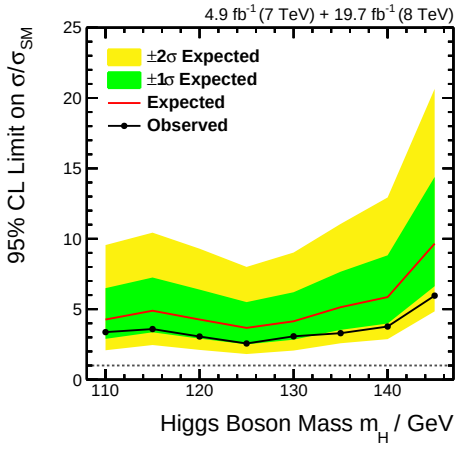
(c) 1-jet, 8TeV



(d) 1-jet, 7+8TeV



(e) 2-jet, 8TeV



(f) 2-jet, 7+8TeV

Figure B.13.: Expected and observed upper limits on the H $\to \tau\tau$ signal production cross section times branching ratio as a function of the Higgs boson mass hypothesis. The limits are shown for the combination of the 0-jet category (top), the 1-jet categories (centre) and the 2-jet category (bottom). The result for the 8 TeV analysis are shown left, whereas the combination of 7 and 8 TeV is shown right. It is clearly visible, that the small deficit seen in data is caused by the 2-jet category, which is the most sensitive category.

C

Supporting Material for the Run II $\mathrm{H} \to \tau\tau$ Analysis

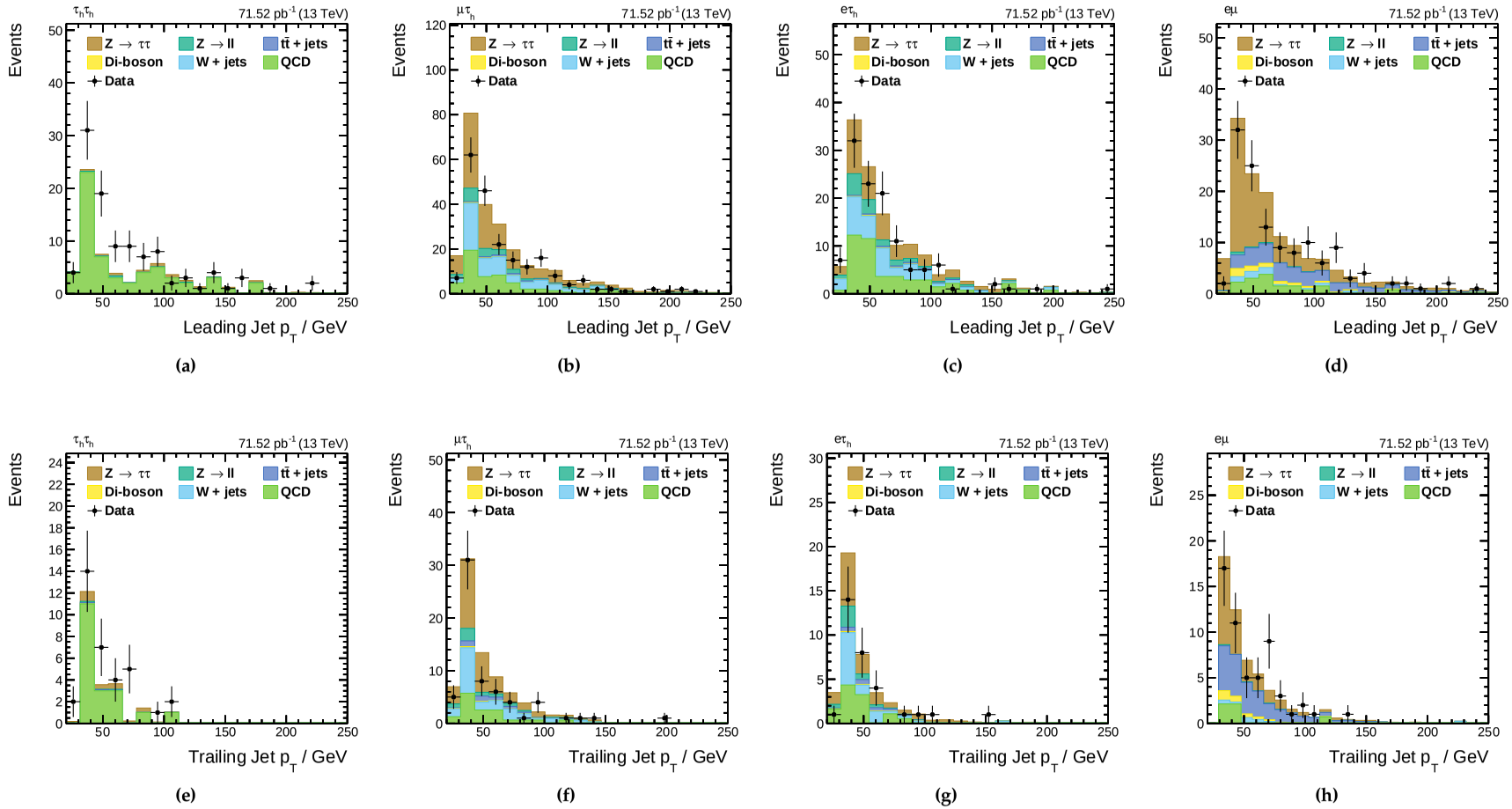


Figure C.1.: Transverse momenta of the leading (top) and sub-leading (bottom) jets in the four channels $\tau_h\tau_h$ (1st column), $\mu\tau_h$ (2nd column), $e\tau_h$ (3rd column) and $e\mu$ (4th column). The observation in data is compared to the expectation given by the sum of all backgrounds. The agreement is acceptable given the preliminary simulation and the small data set.

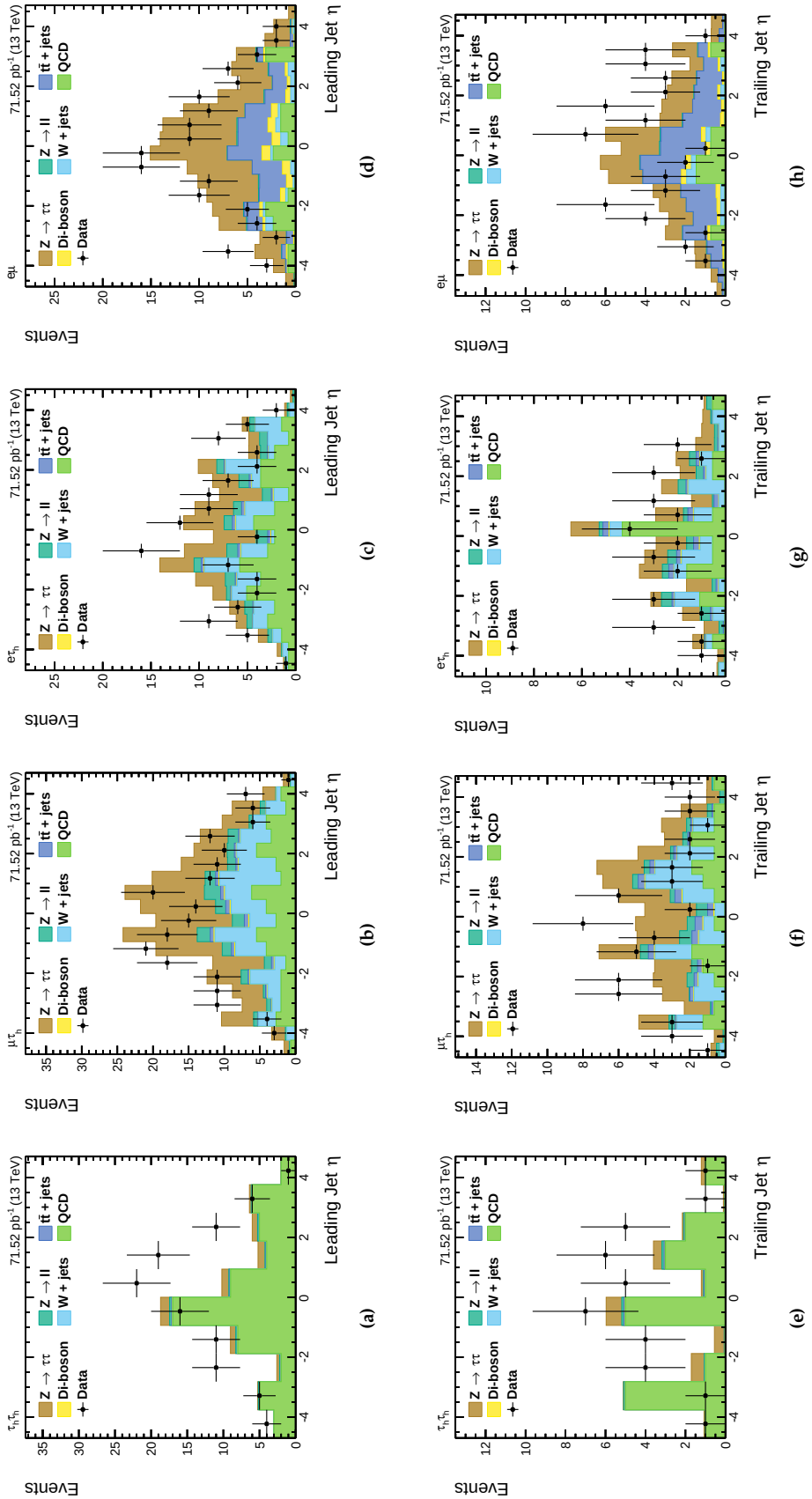
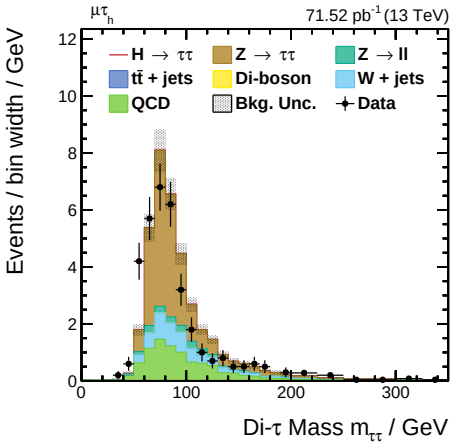
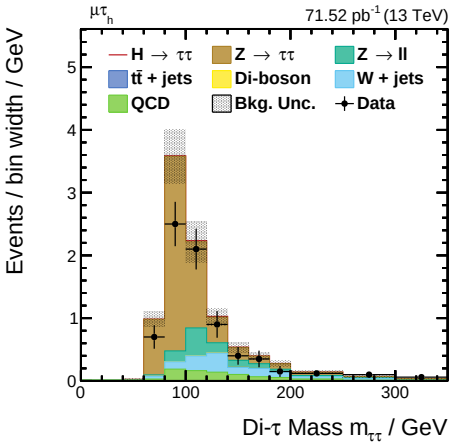


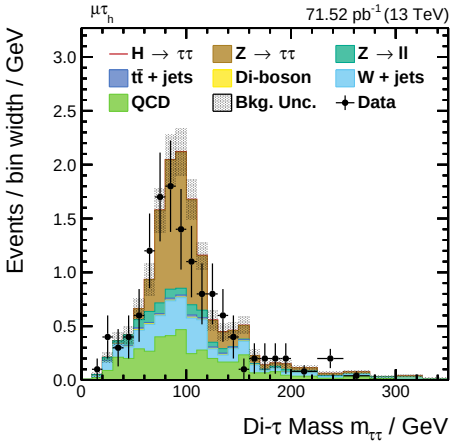
Figure C.2.: Pseudorapidities of the leading (top) and sub-leading (bottom) jets in the four channels $\tau_h\tau_h$ (1^{st} column), $\mu\tau_h$ (2^{nd} column), $e\tau_h$ (3^{rd} column) and $e\mu$ (4^{th} column). The observation in data is compared to the expectation given by the sum of all backgrounds. The agreement is acceptable given the preliminary simulation and the small data set.



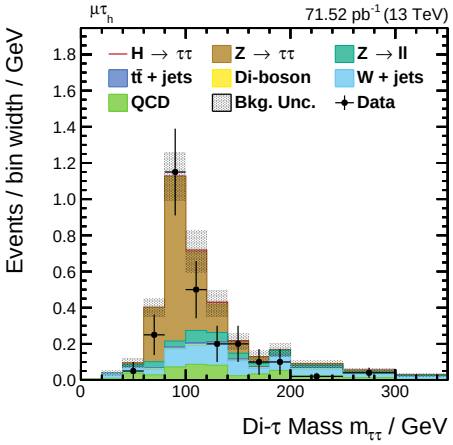
(a) 0-jet low-pt category



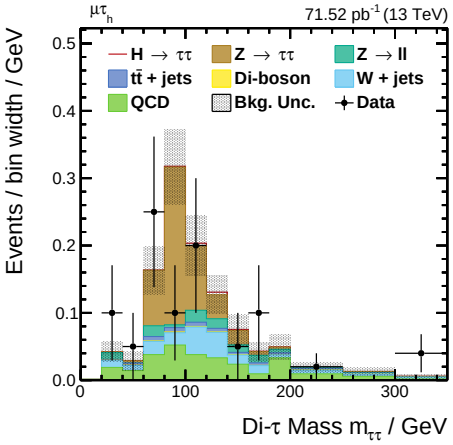
(b) 0-jet high-pt category



(c) 1-jet low-pt category

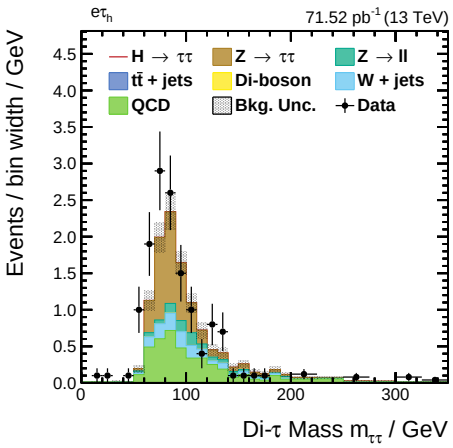


(d) 1-jet high-pt category

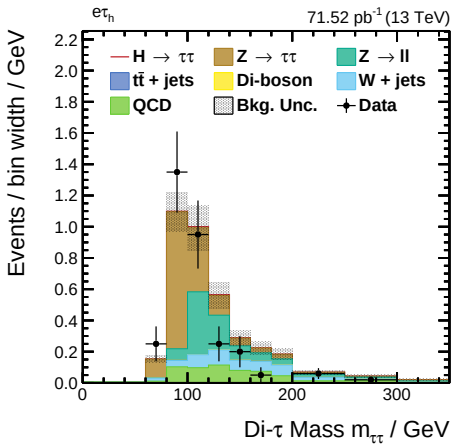


(e) 2-jet category

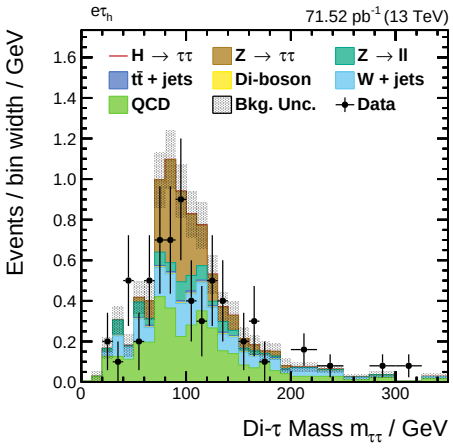
Figure C.3.: Prefit di-τ mass distributions in the five categories, that are exploited in the $\mu\tau_h$ decay channel. The observation in data is compared to the expectation given by the sum of all backgrounds. The observation is compatible with the sum of all backgrounds within the statistical uncertainties. For the extrapolation studies the observation shown here is not taken into account but is replace by an Asimov data set following the signal-plus-background hypothesis.



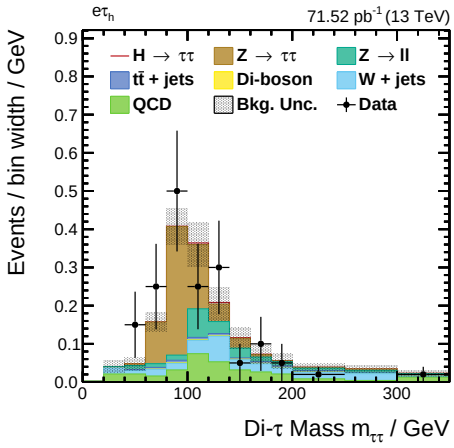
(a) 0-jet low-pt category



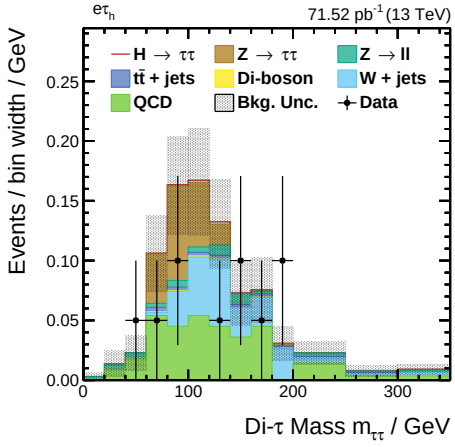
(b) 0-jet high-pt category



(c) 1-jet low-pt category

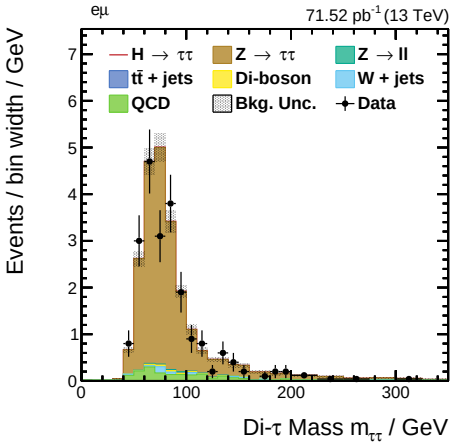


(d) 1-jet high-pt category

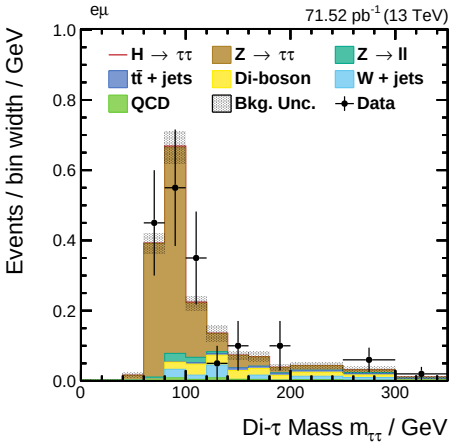


(e) 2-jet category

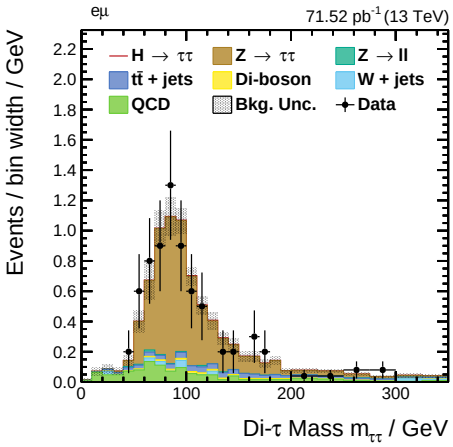
Figure C.4.: Prefit di-τ mass distributions in the five categories, that are exploited in the $e\tau_h$ decay channel. The observation in data is compared to the expectation given by the sum of all backgrounds. The observation is compatible with the sum of all backgrounds within the statistical uncertainties. For the extrapolation studies the observation shown here is not taken into account but is replace by an Asimov data set following the signal-plus-background hypothesis.



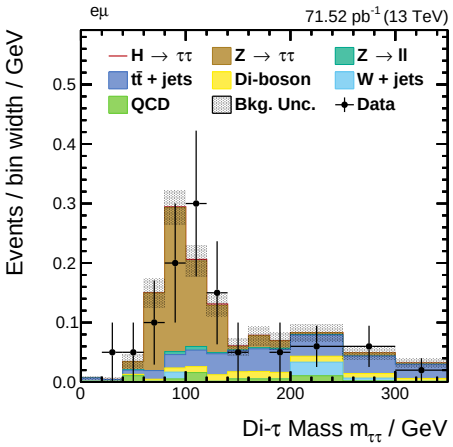
(a) 0-jet low-pt category



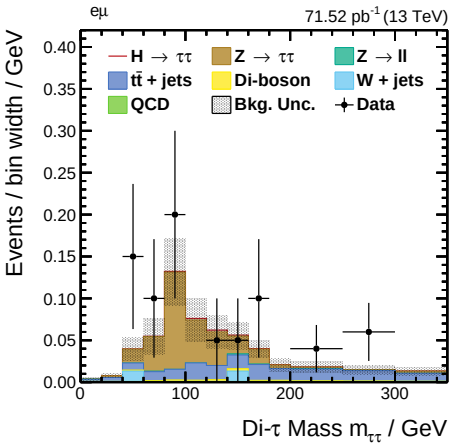
(b) 0-jet high-pt category



(c) 1-jet low-pt category



(d) 1-jet high-pt category



(e) 2-jet category

Figure C.5.: Prefit di-τ mass distributions in the five categories, that are exploited in the eμ decay channel. The observation in data is compared to the expectation given by the sum of all backgrounds. The observation is compatible with the sum of all backgrounds within the statistical uncertainties. For the extrapolation studies the observation shown here is not taken into account but is replace by an Asimov data set following the signal-plus-background hypothesis.

The measurement of Higgs boson couplings to fermions and bosons can be restricted to a scaling of the different production mechanisms. The gluon fusion production is mediated by the loop of heavy quarks (mainly top quarks) and therefore is sensitive to the Higgs boson coupling to fermions. On the other hand, the vector boson fusion process as well as the production in association with vector bosons provides sensitivity to vector boson couplings to the Higgs boson.

Figure C.6 (left) shows the 1 σ contour lines for a fit of the signal strength modifier for the femionic production, μ_{ggH} and for the bosonic production, $\mu_{qq\mathrm{H,VH}}$ as a function of the integrated luminosity. Figure C.6 (right) shows the fit of the fermion production scaling where the boson scaling is allowed to freely vary. The systematic uncertainties contribute 16 % to the total uncertainty, whereas the statistical component evolves as expected proportional to $1/\sqrt{\mathcal{L}}$.

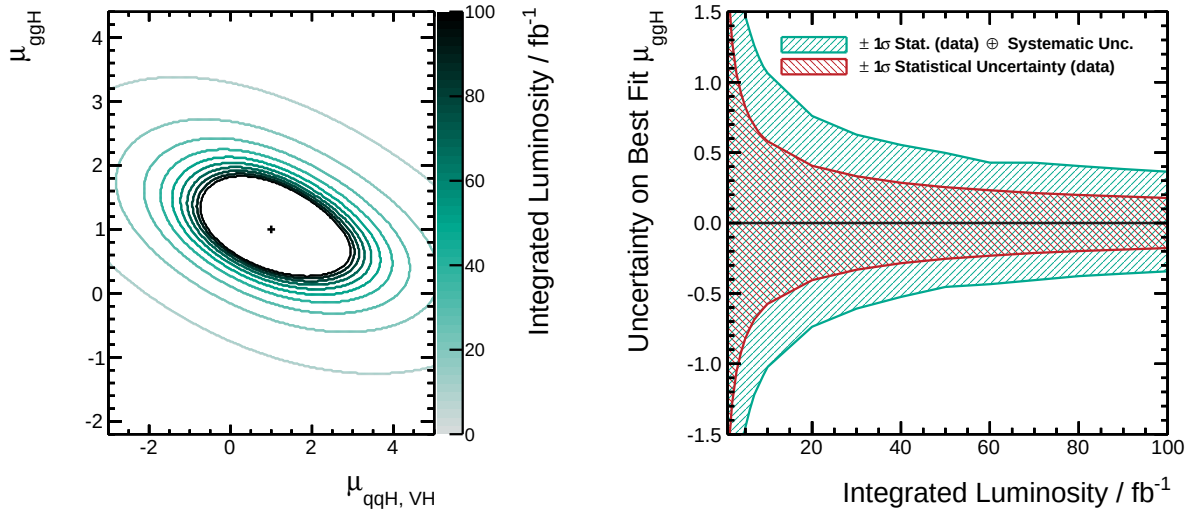


Figure C.6.: Extrapolation of the fermionic and production strength modifiers as a function of the integrated luminosity. The 1 σ contour lines of a two-dimensional scan in the parameter space $\mu_{qq\mathrm{H,VH}}$-μ_{ggH} is shown left. The precision for the measurement of the fermion coupling modifier, μ_{ggH} is shown right, which is determined in a separate fit where $\mu_{qq\mathrm{H,VH}}$ is allowed to freely vary.

Bibliography

[1] S. L. Glashow, "Partial-symmetries of weak interactions", *Nuclear Physics* **22** (1961), no. 4, 579 – 588, doi:10.1016/0029-5582(61)90469-2.

[2] S. Weinberg, "A Model of Leptons", *Phys. Rev. Lett.* **19** (11, 1967) 1264–1266, doi:10.1103/PhysRevLett.19.1264.

[3] A. Salam, "Weak and Electromagnetic Interactions", *Conf.Proc.* **C680519** (1968) 367–377.

[4] P. Higgs, "Broken symmetries, massless particles and gauge fields", *Physics Letters* **12** (1964), no. 2, 132–133, doi:10.1016/0031-9163(64)91136-9.

[5] P. W. Higgs, "Broken Symmetries and the Masses of Gauge Bosons", *Phys. Rev. Lett.* **13** (10, 1964) 508–509, doi:10.1103/PhysRevLett.13.508.

[6] F. Englert and R. Brout, "Broken Symmetry and the Mass of Gauge Vector Mesons", *Phys. Rev. Lett.* **13** (08, 1964) 321–323, doi:10.1103/PhysRevLett.13.321.

[7] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, "Global Conservation Laws and Massless Particles", *Phys. Rev. Lett.* **13** (11, 1964) 585–587, doi:10.1103/PhysRevLett.13.585.

[8] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Phys.Lett.* **B716** (2012) 1–29, doi:10.1016/j.physletb.2012.08.020, arXiv:1207.7214.

[9] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", *Phys. Lett. B* **716** (Jul, 2012) 30–61. 59 p.

[10] CMS Collaboration, "Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons", *JHEP* **1405** (2014) 104, doi:10.1007/JHEP05(2014)104, arXiv:1401.5041.

[11] CMS Collaboration, "Evidence for the direct decay of the 125 GeV Higgs boson to fermions", *Nature Phys.* **10** (Jan, 2014) 557–560. 25 p.

[12] R. Löbl, "Demokrits Atomphysik". Erträge der Forschung; 252. Wiss. Buchges., Darmstadt, 1987.

[13] E. Rutherford, "The Scattering of the Alpha and Beta Rays and the Structure of the Atom", *Proceedings of the Manchester Literary and Philosophical Society IV* (1911) 18–20.

[14] H. Hilscher, "Elementare Teilchenphysik". Facetten. Vieweg, Wiesbaden, 1996.

[15] M. E. Peskin and D. V. Schroeder, "An introduction to quantum field theory". The advanced book program. Westview Press, 1995.

[16] A. Djouadi, "The Anatomy of electro-weak symmetry breaking. I: The Higgs boson in the standard model", *Phys. Rept.* **457** (2008) 1–216, doi:10.1016/j.physrep.2007.10.004, arXiv:hep-ph/0503172.

[17] "The LEP Electroweak Working Group", October, 2015. http://lepewwg.web.cern.ch.

[18] "CMS Standard Model Physics Group", October, 2015. https://twiki.cern.ch/twiki/bin/view/CMSPublic/PhysicsResultsSMP.

[19] R. D. Ball et al., "Parton distributions with LHC data", *Nucl. Phys.* **B867** (2013) 244–289, doi:10.1016/j.nuclphysb.2012.10.003, arXiv:1207.1303.

[20] "Martin-Stirling-Thorne-Watt Parton Distribution Functions", August, 2015. http://projects.hepforge.org/mstwpdf/.

[21] LHC Higgs Cross Section Working Group et al., "Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables", *CERN-2011-002* (CERN, Geneva, 2011) arXiv:1101.0593.

[22] LHC Higgs Cross Section Working Group et al., "Handbook of LHC Higgs Cross Sections: 2. Differential Distributions", *CERN-2012-002* (CERN, Geneva, 2012) arXiv:1201.3084.

[23] LHC Higgs Cross Section Working Group et al., "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties", *CERN-2013-004* (CERN, Geneva, 2013) arXiv:1307.1347.

[24] Particle Data Group Collaboration, "Review of particle physics", *J. Phys.* **G37** (2010).

[25] M. S. Bachtis et al., "Performance of tau reconstruction algorithms with 2010 data in CMS",. CMS AN-2011/045.

[26] CMS Collaboration, "Precise determination of the mass of the Higgs boson and tests of compatibility of its couplings with the standard model predictions using proton collisions at 7 and 8 TeV", *Eur. Phys. J.* **C75** (2015), no. 5, 212, doi:10.1140/epjc/s10052-015-3351-7, arXiv:1412.8662.

[27] CMS Collaboration, "Observation of the diphoton decay of the Higgs boson and measurement of its properties", *The European Physical Journal C* **74** (2014), no. 10, doi:10.1140/epjc/s10052-014-3076-z.

[28] CMS Collaboration, "Measurement of the properties of a Higgs boson in the four-lepton final state", *Phys. Rev. D* **89** (May, 2014) 092007, doi:10.1103/PhysRevD.89.092007.

[29] CMS Collaboration, "Measurement of Higgs boson production and properties in the WW decay channel with leptonic final states", *JHEP* **1401** (2014) 096, doi:10.1007/JHEP01(2014)096, arXiv:1312.1129.

[30] CMS Collaboration, "Search for the standard model Higgs boson produced in association with a W or a Z boson and decaying to bottom quarks", *Phys. Rev. D* **89** (Oct, 2013) 012003. 49 p.

[31] O. S. Brüning et al., "LHC Design Report, Volume I: the LHC Main Ring". CERN, Geneva, 2004.

[32] O. S. Brüning et al., "LHC Design Report, Volume II: the LHC Infrastructure and General Services". CERN, Geneva, 2004.

[33] M. Benedikt et al., "LHC Design Report, Volume III: the LHC Injector Chain". CERN, Geneva, 2004.

[34] C. Lefevre, "LHC: the guide (English version). Guide du LHC (version anglaise)",.

[35] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider", *Journal of Instrumentation* **3** (2008), no. 08, doi:10.1088/1748-0221/3/08/S08003.

[36] CMS Collaboration, "The CMS experiment at the CERN LHC", *Journal of Instrumentation* **3** (2008), no. 08, doi:10.1088/1748-0221/3/08/S08004.

[37] ALICE Collaboration, "The ALICE experiment at the CERN LHC", *Journal of Instrumentation* **3** (2008), no. 08, doi:10.1088/1748-0221/3/08/S08002.

[38] LHCb Collaboration, "The LHCb Detector at the LHC", *Journal of Instrumentation* **3** (2008), no. 08, doi:10.1088/1748-0221/3/08/S08005.

[39] "CMS Media", August, 2015. http://cmsinfo.web.cern.ch/cmsinfo/Media/.

[40] A. de Roeck et al., "CMS Physics Technical Design Report Volume I: Detector Performance and Software". Technical Design Report CMS. CERN, Geneva, 2006.

[41] CMS CollaborationM. Grunwald, et al., "CMS physics Technical Design Report, Volume II: Physics Performance", volume 34. 2007.

[42] C. Berger, "Elementarteilchenphysik – von den Grundlagen zu den modernen Experimenten". Springer-Lehrbuch. Springer, Berlin, 2., aktualisierte und überarb. aufl. edition, 2006.

[43] C. Eck et al., "LHC computing Grid: Technical Design Report. Version 1.06 (20 Jun 2005)". Technical Design Report LCG. CERN, Geneva, 2005.

[44] "Worldwide LHC Computing Grid", August, 2015. http://wlcg.web.cern.ch/documents-reference.

[45] G. L. Bayatyan et al., "CMS computing: Technical Design Report". Technical Design Report CMS. CERN, Geneva, 2005. Submitted on 31 May 2005.

[46] "The CMS Offline WorkBook", August, 2015. https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBook.

[47] A. Scheurer et al., "German Contributions to the CMS Computing Infrastructure", *Journal of Physics: Conference Series* **219** (2010), no. 6, doi:10.1088/1742-6596/219/6/062064. ID:CHEP299.

[48] M. Bajko et al., "Report of the Task Force on the Incident of 19th September 2008 at the LHC",.

[49] S. van der Meer, "Calibration of the effective beam height in the ISR",.

[50] "Public CMS Luminosity Information", October, 2015. https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults.

[51] "HL-LHC Preliminary Design Report: Deliverable: D1.5",.

[52] CMS Collaboration, "Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET",.

[53] W. Adam, B. Mangano, T. Speer, and T. Todorov, "Track Reconstruction in the CMS tracker",.

[54] R. Frühwirth, "Application of Kalman filtering to track and vertex fitting", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **262** (1987), no. 2 - 3, 444 – 450, doi:http://dx.doi.org/10.1016/0168-9002(87)90887-4.

[55] W. Waltenberger, R. Frühwirth, and P. Vanlaer, "Adaptive vertex fitting", *Journal of Physics G: Nuclear and Particle Physics* **34** (2007), no. 12, N343.

[56] W. Adam, R. Frühwirth, A. Strandlie, and T. Todor, "Reconstruction of Electrons with the Gaussian-Sum Filter in the CMS Tracker at the LHC",.

[57] S. Baffioni et al., "Electron reconstruction in CMS", *The European Physical Journal C* **49** (2007), no. 4, 1099–1116, doi:10.1140/epjc/s10052-006-0175-5.

[58] CMS Collaboration, "Electron Reconstruction and Identification at $\sqrt{s} = 7$ TeV",.

[59] CMS Collaboration, "Electron performance with 19.6 fb^{-1} of data collected at $\sqrt{s} = 8$ TeV with the CMS detector.",.

[60] CMS, "Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV", *Journal of Instrumentation* **7** (2012), no. 10, P10002.

[61] CMS Collaboration, "Performance of τ-lepton reconstruction and identification in CMS", *J. Instrum.* **7** (Sep, 2011) P01001. 33 p.

[62] CMS Collaboration, "Reconstruction and identification of tau lepton decays to hadrons and tau neutrino at CMS", arXiv:1510.07488.

[63] G. P. Salam, "Towards jetography", *The European Physical Journal C* **67** (2010), no. 3-4, 637–686, doi:10.1140/epjc/s10052-010-1314-6.

[64] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k_t jet clustering algorithm", *Journal of High Energy Physics* **2008** (2008), no. 04, 063.

[65] M. Cacciari, G. P. Salam, and G. Soyez, "FastJet user manual", *The European Physical Journal C* **72** (2012), no. 3, doi:10.1140/epjc/s10052-012-1896-2.

[66] CMS, "Determination of jet energy calibration and transverse momentum resolution in CMS", *Journal of Instrumentation* **6** (November, 2011) 11002, doi:10.1088/1748-0221/6/11/P11002, arXiv:1107.4277.

[67] CMS Collaboration, "8 TeV Jet Energy Corrections and Uncertainties based on 19.8 fb^{-1} of data in CMS",.

[68] CMS Collaboration, "Performance of b tagging at sqrt(s)=8 TeV in multijet, ttbar and boosted topology events",.

[69] CMS Collaboration, "MET performance in 8 TeV data",.

[70] CMS Collaboration, "Performance of the CMS missing transverse momentum reconstruction in pp data at $\sqrt{s}$Âă=Âă8ÂăTeV", *Journal of Instrumentation* **10** (2015), no. 02, P02006.

[71] T. Sjöstrand, S. Mrenna, and P. Z. Skands, "PYTHIA 6.4 Physics and Manual", *Journal of High Energy Physics* **5** (2006) doi:10.1088/1126-6708/2006/05/026.

[72] B. Andersson, G. Gustafson, G. Ingelman, and T. Sjöstrand, "Parton fragmentation and string dynamics", *Physics Reports* **97** (1983), no. 2-3, 31–145, doi:10.1016/0370-1573(83)90080-7.

[73] J. Alwall et al., "MadGraph 5 : Going Beyond", *Journal of High Energy Physics* **6** (2011) doi:10.1007/JHEP06(2011)128.

[74] J. Alwall et al., "The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations", *JHEP* **07** (2014) 079, doi:10.1007/JHEP07(2014)079, arXiv:1405.0301.

[75] S. Alioli, P. Nason, C. Oleari, and E. Re, "A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX", *JHEP* **1006** (2010) 043, doi:10.1007/JHEP06(2010)043, arXiv:1002.2581.

[76] Z. Was, "TAUOLA the library for tau lepton decay, and KKMC/KORALB/KORALZ/... status report", *Nucl. Phys. Proc. Suppl.* **98** (2001) doi:10.1016/S0920-5632(01)01200-2.

[77] "HepMC event record", August, 2015. http://lcgapp.cern.ch/project/simu/HepMC/.

[78] S. A. et al., "Geant4 – a simulation toolkit", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **506** (2003), no. 3, 250–303, doi:10.1016/S0168-9002(03)01368-8.

[79] R. Brun and F. Rademakers, "ROOT – An object oriented data analysis framework", *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* **389** (1997), no. 1-2, 81–86, doi:10.1016/S0168-9002(97)00048-X. New Computing Techniques in Physics Research V.

[80] A. Hoecker et al., "TMVA: Toolkit for Multivariate Data Analysis", *PoS* **ACAT** (2007) 040, arXiv:physics/0703039.

[81] "CMS Software", August, 2015. http://cms-sw.github.io/index.html.

[82] ATLAS, CMS, LHC Higgs Combination Group Collaboration, "Procedure for the LHC Higgs boson search combination in Summer 2011",.

[83] A. Bethani, "Neutral Higgs boson searches in the $H \to \tau\tau \to \mu\mu$ decay channel",. Dissertation, KIT Karlsruhe.

[84] CMS Collaboration, "Search for the Standard-Model Higgs boson decaying to tau pairs in proton-proton collisions at sqrt(s) = 7 and 8 TeV",.

[85] J. Salfeld-Nebgen, "Search For The Higgs Boson Decaying Into τ-Leptons In The Di-Electron Channel",. Dissertation, Hamburg University.

[86] D. de Florian, G. Ferrera, M. Grazzini, and D. Tommasini, "Higgs boson production at the LHC: transverse momentum resummation effects in the H->2gamma, H->WW->lnu lnu and H->ZZ->4l decay modes", *JHEP* **1206** (2012) 132, doi:10.1007/JHEP06(2012)132, arXiv:1203.6321.

[87] "Standard Model Cross Sections for CMS at 7 TeV", October, 2015. https://twiki.cern.ch/twiki/bin/viewauth/CMS/StandardModelCrossSections.

[88] "Standard Model Cross Sections for CMS at 8 TeV", October, 2015. https://twiki.cern.ch/twiki/bin/viewauth/CMS/StandardModelCrossSectionsat8TeV.

[89] T. Müller, "Unterdrückung des Z-Untergrundes zur Higgs-Suche im Kanal $H \to \tau\tau$ durch Multivariate Analyse von $\tau\tau$-Endzuständen in pp-Kollisionen am LHC",. KIT, Diplomarbeit, 2012.

[90] R. Friese, "Multivariate Methoden zur Identifikation von $H \to \tau\tau \to \mu\mu$ Zerfällen",. KIT, Diplomarbeit, 2013.

[91] A. Burgmeier, "Position Resolution and Upgrade of the CMS Pixel Detector and Search for the Higgs Boson in the $\tau^+\tau^-$ Final State",. Dissertation, KIT Karlsruhe.

[92] J. Berger, "Search for the Higgs Boson Produced via Vector-Boson Fusion in the Decay Channel $H \to \tau\tau$",. Dissertation, KIT Karlsruhe.

[93] A. L. Read, "Presentation of search results: the CL s technique", *Journal of Physics G: Nuclear and Particle Physics* **28** (2002), no. 10, 2693.

[94] B. Mistlberger and F. Dulat, "Limit setting procedures and theoretical uncertainties in Higgs boson searches", arXiv:1204.3851.

[95] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics", *European Physical Journal C* **71** (February, 2011) 1554, doi:10.1140/epjc/s10052-011-1554-0, arXiv:1007.1727.

[96] CMS Higgs to Tau Tau Working Group Collaboration, "Physics Objects in the Higgs to Tau Tau Analysis",.

[97] CMS Higgs to Tau Tau Working Group Collaboration, "Search for Higgs to Tau Tau in the Electron-Muon Channel",.

[98] CMS Higgs to Tau Tau Working Group Collaboration, "Search for Higgs to Tau Tau in the Muon-Tau and Electron-Tau Channels",.

[99] CMS Higgs to Tau Tau Working Group Collaboration, "Search for the Higgs boson decaying into TauTau in the full hadronic channel",.

[100] CMS Higgs to Tau Tau Working Group Collaboration, "Search for Higgs boson decays to tau pairs in the di-muon and di-electron channels",.

[101] CMS Higgs to Tau Tau Working Group Collaboration, "Search for a Standard Model Higgs boson decaying to tau pairs produced in association with a W or Z boson",.

[102] CMS Higgs to Tau Tau Working Group Collaboration, "Theoretical uncertainty for the Higgs production via VBF and Gluon Fusion process",.

[103] CMS Higgs to Tau Tau Working Group Collaboration, "Search for the Standard-Model Higgs boson decaying to tau pairs in proton-proton collisions at $\sqrt{s}$ = 7 and 8 TeV",.

[104] "Standard Model Cross Sections for CMS at 13 TeV", October, 2015. `https://twiki.cern.ch/twiki/bin/viewauth/CMS/StandardModelCrossSectionsat13TeV`.

[105] J. Stirling, "Parton luminosity and cross section plots", October, 2015. `http://www.hep.ph.ic.ac.uk/~wstirlin/plots/plots.html`.

[106] CMS Collaboration, "Multivariate Determination of the Missing Energy in the Transverse Plane (E_T^{miss}) at $\sqrt{s} = 13$TeV",.

[107] B. Treiber, "Estimation of the Background from Z $\to \tau\tau$ in H $\to \tau\tau$ Analyses",. Karlsruher Institut für Technologie (KIT), Masterarbeit, 2015.

[108] C. Cuenca Almenar, "Search for the neutral MSSM Higgs bosons in the ditau decay channels at CDF Run II", `doi:10.2172/953708`.

[109] "Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC pp collision data at $\sqrt{s}$ = 7 and 8 TeV",.

[110] CMS Collaboration, "Projected Performance of an Upgraded CMS Detector at the LHC and HL-LHC: Contribution to the Snowmass Process",.

[111] ATLAS Collaboration, "Projections for measurements of Higgs boson cross sections, branching ratios and coupling parameters with the ATLAS detector at a HL-LHC",.

[112] LHC Higgs Cross Section Working Group Collaboration, "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties: Report of the LHC Higgs Cross Section Working Group",.

[113] J. Berger et al., "ARTUS - A Framework for Event-based Data Analysis in High Energy Physics", arXiv:1511.00852.

[114] "Framework-light", August, 2015. https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookFWLite.

List of figures

List of tables

Acknowledgements

First of all, I am grateful to my supervisor Prof. Dr. Günter Quast for giving me the opportunity to pursue this research in the Institut für experimentelle Kernphysik at the Karlsruhe Institute of Technology and for offering the almost two years at CERN, which were a valuable experience for me. Thanks for all his support, guidance and help on physical topics. I would also like to thank Prof. Dr. Wim de Boer for acting as a co-referee and for all discussions about this thesis.

Secondly, I thank my advisor Dr. Alexei Raspereza for introducing me into the physics of the $H \to \tau\tau \to \mu\mu$ analysis and for integrating me into his group. I appreciated very much all the physics discussions and his support in the $H \to \tau\tau$ group at CERN. Similarly, I would like to thank Dr. Roger Wolf for all his guidance, motivation and support in the second part of my work and for the proofreading of this thesis.

Many thanks go to Fabio Colombo for the perfect collaboration during the last three years. I profited very much from the mutually complementing work and the innumerable discussions on analysis details. I also want to thank Dr. Manuel Zeise for introducing me into the group and especially into data analysis at the CMS experiment in general.

I thank all my colleagues for the friendly atmosphere, the help and feedback in various points and for the proofreading of this thesis. Thank you very much for accepting the challenges of a common software framework.