Universität Bonn

Physikalisches Institut

Measurement of the Higgs Boson Decay to τ-Pairs in the Lepton plus Hadrons Final State in Proton-Proton Collisions with the ATLAS Detector

Jessica Kristina Liebal

In this thesis, the decay of the Standard Model Higgs boson into one hadronically and one leptonically decaying τ lepton is investigated using multivariate methods. Proton-proton collisions recorded by the ATLAS experiment at the LHC in 2011 and 2012 at a center-of-mass energy of $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV are analysed. The data correspond to an integrated luminosity of 4.5 fb^{-1} and 20.3 fb^{-1}, respectively. Within the analysis two event categories are defined, which are sensitive to the two main Higgs production mechanisms, gluon-gluon fusion and vector-boson fusion. Further improvements of the data-driven technique used to model the irreducible $Z \rightarrow \tau\tau$ background in $\tau\tau$ final states were performed. Under the assumption of a Higgs boson with mass m_H=125 GeV, an excess in data over the expected background is found with an observed (expected) significance of 2.5 (2.4) σ. The measured signal strength of $\mu = 1.1^{+0.5}_{-0.5}$ is in agreement with the Standard Model prediction. The presented analysis is an important part of the combined analysis of the $H \rightarrow \tau\tau$ decay channel, which includes all possible τ decay modes. The combined analysis yields an excess of events over the predicted background with an observed (expected) significance of 4.5 (3.4) σ and a measured signal strength of $\mu = 1.4^{+0.4}_{-0.4}$. The combined result is evidence of the Higgs boson Yukawa coupling to τ leptons. Furthermore, studies are presented which address potential advancements of the analysis strategy towards future operation times of the LHC.

Physikalisches Institut der
Universität Bonn
Nussallee 12
D-53115 Bonn

BONN-IR-2016-05
October 2016
ISSN-0172-8741

Measurement of the Higgs Boson Decay to τ-Pairs in the Lepton plus Hadrons Final State in Proton-Proton Collisions with the ATLAS Detector

Dissertation
zur
Erlangung des Doktorgrades (Dr. rer. nat.)
der
Mathematisch-Naturwissenschaftlichen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

von
Jessica Kristina Liebal
aus
Leverkusen

Bonn, 2016

Dieser Forschungsbericht wurde als Dissertation von der Mathematisch-Naturwissenschaftlichen Fakultät der Universität Bonn angenommen und ist auf dem Hochschulschriftenserver der ULB Bonn `http://hss.ulb.uni-bonn.de/diss_online` elektronisch publiziert.

1. Gutachter: Prof. Dr. Norbert Wermes
2. Gutachter: Prof. Dr. Jochen Dingfelder

Tag der Promotion: 22.09.2016
Erscheinungsjahr: 2016

Contents

CHAPTER 1

Introduction

The question of the origin of matter in the universe and the nature of the fundamental physical interactions has driven science for centuries. The development of classical physics replaced earlier rudimentary theories and forms the basis of our understanding of nature. With the discovery of various chemical elements a scientific ordering scheme was developed at the end of the 19th century which arranged the assumed elementary particles in a periodic table. Further experimental observations showed that the chemical elements consist of electrons in the atomic shell and an atomic nucleus which is built from protons and neutrons. At the beginning of the 20th century physics experienced a paradigm shift introduced by quantum mechanics and the theory of relativity. In the following years it was experimentally confirmed that protons and neutrons are not elementary particles but have a substructure. Not only the fundamental particles which build protons and neutrons were identified but also further composite and elementary particles were discovered. Those discoveries led to the development of the Standard Model of particle physics (SM) in the 1960's. So far, the SM is the most fundamental theory which exists for the description of elementary particles and their interactions. It turned out to be an outstanding theory of modern physics, since it did not only describe physical processes already known, but also correctly predicted the existence of new particles, like e.g. the charm quark or the τ lepton, which were later discovered.

The SM incorporates three fundamental forces: strong interaction, weak interaction and electromagnetism. The theory of electromagnetic and weak interactions were unified in the theory of electroweak interaction. In addition, the SM includes twelve fermions, which are grouped according to their interactions into six quarks, three charged leptons and three neutral neutrinos. These fermions interact with each other via the exchange of bosons, the mediators of the three interactions. The mediators of strong and electromagnetic interactions are massless, while the bosons of the weak interaction are massive.

The SM not only describes the interactions between elementary particles but also gives an explanation for the origin of the masses of quarks and leptons as well as the massive bosons. In this theory all masses are generated by the so-called Brout–Englert–Higgs mechanism. A consequence of this mechanism is another massive scalar particle, the Higgs boson. The Brout–Englert–Higgs mechanism is an essential part of the SM: If electrons had been massless no atoms would have been formed, because massless electrons would have escaped from the nuclei with the speed of light according to the theory of relativity. In addition, the weak interaction would be much stronger than the electromagnetic force. As a consequence, the universe would look completely different. For that reason, the experimental evidence of the Brout–Englert–Higgs mechanism and thereby the origin of particle masses is crucial for the validation of the SM and our understanding of nature.

In order to confirm the existence of the mass generating mechanism, the Higgs boson has to be observed experimentally. The Higgs boson is the last missing particle predicted by the SM. Its experimental evidence is one of the major goals of the searches for physical processes at the Large Hadron Collider (LHC) at CERN. In 2012, the ATLAS and CMS experiments at the LHC discovered a new particle with a mass of around 125 GeV in the bosonic Higgs decay channels [1, 2]. The latest mass measurements by the ATLAS and CMS experiment yield a mass of $m_H = 125.36 \pm 0.41$ GeV [3] and $m_H = 125.03 \pm 0.30$ GeV [4], respectively. Also other so far investigated properties of this particle are consistent with the predicted SM Higgs boson [5–7].

In order to confirm that this new particle is indeed the SM Higgs boson, an experimental evidence of its decay into fermions must be found. The search for the Higgs boson decay into two τ leptons with the ATLAS detector is the subject of this thesis. This Higgs decay channel is the first fermionic decay channel for which an experimental evidence was achieved. The two τ leptons originating from the Higgs decay cannot be observed directly in the detector due to the short lifetime of the τ lepton. Instead, they are indirectly detected by their decay products. The τ lepton can either decay into the lighter leptons, i.e. electrons or muons, or into hadrons, i.e. bound states of quarks. Depending on the decay products, three $\tau\tau$ decay channels are distinguished in the Higgs search: (i) both τ leptons decay hadronically, (ii) both τ leptons decay leptonically and (iii) one τ lepton decays hadronically and the other one leptonically.

The largest branching ratio with ca. 46% is revealed by channel (iii). The corresponding decay is identified in the ATLAS detector via the τ decay products, i.e. one electron or muon and hadronic activity associated with a hadronically decaying τ lepton. In addition, a significant amount of so-called missing transverse energy is expected as a consequence of the undetected neutrinos which are produced by the decay of the τ leptons. This experimental signature in the detector is, however, not unique to the decay of the Higgs boson. The decay of the Z boson, one of the mediators of the weak interaction, into two τ leptons reveals the same signature and is therefore one of the most probable physical processes which can be mistaken for a Higgs decay into two τ leptons.

In this thesis, the search for the Higgs boson in channel (iii) is presented based on data taken with the ATLAS experiment in 2011 and 2012. The thesis is part of the analysis of the combined channels (i)-(iii), which is based on the complete 2011 and 2012 dataset and was published in Reference [8]. In order to achieve a high sensitivity with respect to the Higgs boson decay the analysis strategy of earlier studies of ATLAS data has been further developed and extended to a larger dataset. The studies performed during the work on this thesis significantly contributed to this ATLAS publication and point to future improvements of the analysis strategy. One important aspect of this thesis concerns the improvement and validation of the modelling technique used for the dominant background, which originates from decays of the Z boson into two τ leptons. This modelling technique is of high relevance for the Higgs search in $\tau\tau$ final states at the LHC. A comprehensive overview of the technique was therefore published in Reference [9].

The thesis is structured as follows: Chapter 2 starts with a short introduction to the theoretical concept of the SM and the Higgs searches at the LHC. In Chapter 3, the LHC and the components of the ATLAS detector are summarised. Chapter 4 briefly discusses how different physical processes are simulated with the ATLAS software. Further, it is explained how the various particles are reconstructed within the ATLAS detector. After a short overview of the experimental signatures of the Higgs boson decay into τ leptons and the background suppression in Chapter 5, Chapter 6 is dedicated to the modelling of the dominant background from the decay of the Z boson into two τ leptons. In Chapter 7, the multivariate analysis of channel (iii) is presented. In Chapter 8 the published ATLAS analysis of the Higgs boson decay considering all possible τ decay modes is shortly summarised. An optimised multivariate analysis, which suggests further improvements of the analysis concept, is given in Chapter 9. Finally, in Chapter 10 conclusions are drawn on the obtained results.

CHAPTER 2

The Standard Model and the phenomenology of the Higgs boson

The Standard Model of particle physics (SM) is a theoretical framework describing the interaction of fundamental particles. Developed in the 1960's through an interplay of experimental observations and theoretical concepts, it led to a deeper understanding of matter. In the following decades, many predictions made by the SM could be confirmed experimentally. A recent milestone was reached in 2012 when the ATLAS and CMS experiments at the Large Hadron Collider (LHC) found a new particle [1, 2] consistent with the Higgs boson, which was postulated by the Brout–Englert–Higgs mechanism[1] [10–12].

In this chapter, the theoretical concepts of the SM are briefly explained with emphasis on the Brout–Englert–Higgs mechanism as the essential part of this theory. The discussion closely follows [13–15]. In Section 2.2 important theoretical concepts which are necessary to describe proton-proton collisions are explained. Further, the searches for the Higgs boson at the LHC are introduced according to [16–19].

2.1 The Standard Model

In the early 20th century protons, neutrons and electrons were the only known particles. The discovery of more and more particles in the following years, however, exposed that the structure of matter is far more complex. For example, various hard scattering experiments revealed that neutrons and protons must have a substructure and are built up from quarks q and gluons g, also referred to as partons, with interactions on sub-atomic scales. Therefore, protons and neutrons are only two particles in the spectrum of many composite particles.

The SM was developed to explain the fundamental interactions between the elementary particles within a highly predictive scheme. An overview of the particles and the interactions of the SM is given in Figure 2.1. According to the SM the universe is made of elementary particles with spin $1/2$ - called fermions - , which can be further subdivided into quarks, leptons and neutrinos. The SM contains six quarks with charges $-1/3e$ or $+2/3e$ and six anti-quarks. They are arranged in three generations according to their mass and quantum numbers. Bound states of quarks ($q\bar{q}$ or qqq, $\bar{q}\bar{q}\bar{q}$) are also referred to as hadrons. Matter, observed in everyday life, contains only the two lightest quarks of the first generation - namely the up (u) and down (d) quarks. Protons and neutrons, which form atomic nuclei, are made up

[1] In the following, this mechanism is called the Higgs mechanism.

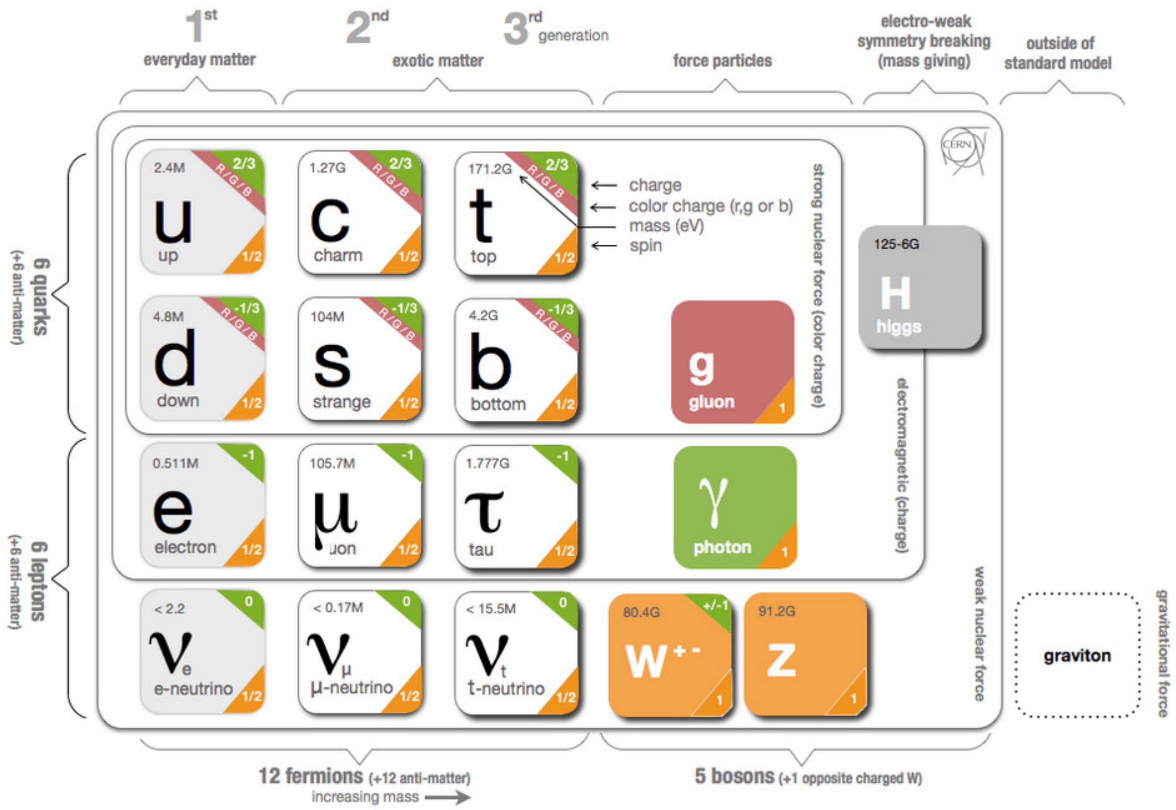


Figure 2.1: Particles and interactions described by the SM. The figure is taken from [20].

of three light quarks. While protons consist of two u quarks and one d quark, neutrons contain two d quarks and one u quark. Other composite particles, i.e. mesons[2] ($q\bar{q}$) or baryons (qqq, $\bar{q}\bar{q}\bar{q}$), containing the heavier quarks of the higher generations are referred to as exotic matter. They can be created, e.g. in high-energy experiments.

The SM further contains three charged (anti-) leptons e, μ, τ and three electrically neutral (anti-) neutrinos ν_e, ν_μ, ν_τ, which are arranged in three generations. Each generation consists of one charged lepton and its corresponding neutrino. An additional quantum number L_i ($i = e, \mu, \tau$) is introduced for all generations, which is conserved in interactions. While the masses of the charged leptons significantly increase in the higher generations, the SM treats neutrinos as massless particles.

The interaction between these fundamental particles is described in the SM via three forces: the electromagnetic, the strong and the weak interaction. Gravity is not part of the SM, since so far no suitable quantum theory exists. However, the effect of the gravitational force is negligible when considering interactions between particles at nuclear or at even smaller scales.

The fundamental interactions are mediated by gauge bosons. While the photon γ and the gluons g, which mediate the electromagnetic and the strong interaction respectively, are massless, the mediators $W^\pm$ and Z of the weak interaction are massive. To explain the interactions between particles in detail different specific theories have been developed using the formalism of Lagrangian densities[3]. Quantum electrodynamics (QED) is the theoretical basis of the electromagnetic interaction (see Section 2.1.1), while

[2] The three lightest mesons are the un-/charged pions $\pi^\pm$ and π^0. They are bound states of u and d quarks and can be produced, e.g. in τ decays (see Section 4.3.7).

[3] An introduction into the Lagrangian formalism can be found in several textbooks, e.g. [13].

the strong interaction is described by quantum chromodynamics (QCD) [21–23] (see Section 2.1.2). In order to bind quarks e.g. into protons, the strong force has to outbalance the repulsive Coulomb field between their constituents. The strong interaction couples to the *colour* charge, the counterpart of the electric charge in electromagnetic interactions. Only quarks and gluons carry colour. Each quark carries one colour, *red* (r), *green* (g) or *blue* (b), which allows to distinguish quarks of the same flavour. Gluons are bicoloured objects carrying one colour and one anti-colour. The various combinations of colour quantum numbers lead to eight distinguishable gluons g_i, which, in contrast to photons, can also couple to each other.

In nature, only colourless bound states of quarks are observed. The confinement of quarks and gluons into hadrons results from the dependence of the strong coupling constant on the energy scale. For low energy scales, corresponding to large distances, the coupling strength increases, which makes the observation of free quarks impossible. In contrast, for large energy scales, corresponding to small distances, i.e. within the confined hadrons, they act basically as free particles. This effect is referred to as *asymptotic freedom*.

The lifetime of the neutron and some other particles is much longer than expected from strong interaction and therefore hinted at the existence of another interaction, known as weak interaction. It is mediated by the heavy Z and $W^{\pm}$ gauge bosons. One example for weak interactions is the β-decay, in which a proton (neutron) decays into a neutron (proton) via the emission of a positron (electron) and a (anti-) neutrino. Both leptons and quarks participate in weak interactions. Since neutrinos are electrically neutral, they can only be observed in weak interactions.

The electromagnetic and weak interaction can be combined in the electroweak theory. The unification of both theories, completed by Glashow, Salam and Weinberg [24–26] in the 1960's, includes a spontaneous symmetry breaking mechanism to account for a massless electromagnetic mediator, the photon, and the three heavy mediators, Z and $W^{\pm}$, of the weak interaction. This mechanism is referred to as the Higgs mechanism and introduces a new scalar field, the Higgs field. The mechanism was first suggested by Brout and Englert [27], Higgs [10, 11] and Guralnik, Hagen and Kibble [12]. In combination with the Higgs mechanism an additional scalar particle, the Higgs boson, was postulated. Its mass is one of the free parameters of the SM.

In total, the SM is described by 19 free parameters. Their values are not predicted by theory but have to be determined experimentally.

Up to now, the SM is a very successful model, since it does not only postulated the existence of so far unknown fundamental particles like the top quark or the τ lepton but also precisely predicted interaction rates. One missing part essential for the validation of the SM was the experimental evidence of the Higgs boson and thus the Higgs mechanism. This was finally achieved by the ATLAS and CMS experiments at the LHC in 2012 [1, 2].

Although the SM is a success story, which guided physicists for decades to a deeper understanding of nature, there are still questions which remain unanswered. For example, the experimental evidence of neutrino oscillation is an evidence for massive neutrinos. This discovery might hint to physics beyond the current SM. Further, since the masses of fermions are free parameters of the SM, the SM cannot explain the large mass difference between fermions. Further questions arise from the observation of the dominance of matter over antimatter in the universe, which cannot be completely explained by the CP violation in weak interaction. Also, the origin of dark matter is an open question. These are only a few of the open questions, which go beyond the SM and are currently under investigation. Their discussion goes beyond the scope of this thesis. A summary of how such questions can be addressed by the experiments situated at the LHC is given in [28].

2.1.1 The Lagrangian density of quantum electrodynamics

The interaction of electrons with electromagnetic fields is described within the theory of quantum electrodynamics (QED). In this section, the main concepts of field theory are introduced on the basis of QED. The same concepts are also used later for the formulation of quantum chromodynamics and of the electroweak theory. The free Dirac Lagrangian for a spin 1/2 particle is given by

$$\mathcal{L} = i\bar{\psi}\gamma^{\mu}\partial_{\mu}\psi - m\bar{\psi}\psi \tag{2.1}$$

with the particle's four-component wavefunction ψ and the Dirac matrices γ^{μ} ($\mu = 0, 1, 2, 3$) defined in [13]. It is invariant under global phase transformation $\psi(x) \to e^{-ie\alpha}\psi(x)$, but not under local phase transformation $\psi(x) \to e^{-ie\alpha(x)}\psi(x)$. In the latter case, an additional dependence of the phase $e\alpha(x)$ on space and time leads to a non vanishing extra term $e\bar{\psi}\gamma^{\mu}\psi\partial_{\mu}\alpha$, so that the Lagrangian after transformation is given by

$$\mathcal{L}' = \bar{\psi}\gamma^{\mu}\partial_{\mu}\psi - m\bar{\psi}\psi + e\bar{\psi}\gamma^{\mu}\psi\partial_{\mu}\alpha\ . \tag{2.2}$$

Since the Lagrangian in Equation 2.1 is required to be locally phase invariant, an additional term, which implies the coupling between a field A_{μ} to a Dirac particle, is added:

$$\mathcal{L}_{\text{inv}} = i\bar{\psi}\gamma^{\mu}\partial_{\mu}\psi - m\bar{\psi}\psi - (e\bar{\psi}\gamma^{\mu}\psi)A_{\mu}\ . \tag{2.3}$$

This new field transforms according to $A_{\mu} \to A_{\mu} + \partial_{\mu}\alpha$, which is analogous to the transformation of the electromagnetic field in the Maxwell equations, and compensates the extra term in $\mathcal{L}_{\text{inv}}$ in order to ensure local phase invariance. The final Lagrangian of QED includes a further term to account for the free field dynamics of the massless[4] field A_{μ}

$$\mathcal{L}_{\text{QED}} = i\bar{\psi}\gamma^{\mu}\partial_{\mu}\psi - m\bar{\psi}\psi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} - (e\bar{\psi}\gamma^{\mu}\psi)A_{\mu} \tag{2.4}$$

with $F^{\mu\nu} = \partial^{\mu}A^{\nu} - \partial^{\nu}A^{\mu}$ being the field strength tensor. The equations of motion are obtained from the Euler-Lagrange equations and directly lead to the Maxwell equation of electrodynamics, which are known to be invariant under gauge transformations of the form $A_{\mu} \to A_{\mu} + \partial_{\mu}\alpha$. In the final step, the equations of motion are solved in perturbation theory. This approach can be used, since an expansion in the coupling strength $\alpha = e^2/4\pi \approx 1/137$ converges quickly due to the small size of α.
QED is a good example to illustrate the concept of symmetry groups, which is also highly relevant to the formulation of quantum chromodynamics and weak interaction: The interaction structure is set by defining a symmetry group U with $U^{\dagger}U = 1$, under which the Lagrangian is required to be invariant, i.e transforms like

$$\psi \to U\psi\ .$$

For the Lagrangian density of QED such a unitary symmetry group is of the degree 1 and named U(1). In this case, the electric charge e is the generator of the U(1) group. Further terms are added to the Lagrangian in order to preserve gauge invariance.
The same concept is used for the formulation of the Lagrangian densities of strong and weak interaction, which are based on the SU(3) and SU(2) symmetry group, respectively. Also, more fermions can be

[4] The dynamics are obtained from the Lagrangian for spin 1 vector fields with mass m [14]. To ensure gauge invariance the field A_{μ} is required to be massless.

incorporated by adding kinetic and interaction terms for each new fermion, leading to the Lagrangian of the SM.

2.1.2 The Lagrangian density of quantum chromodynamics

Deep inelastic scattering experiments performed around 1970 provided the first evidence that protons have a substructure. This was in line with the quark model postulated by Gell-Mann and Zweig in 1964. Today, the quark model with six colour-charged quarks of different flavours is experimentally established. The dynamics of strong interaction between quarks are explained by QCD. Due to the colour charge each quark has three additional degrees of freedom and hence the gauge symmetry is defined by the SU(3) group. This results in eight gauge bosons, called gluons, as mediators of the strong force, in contrast to one gauge boson as in QED. However, the QCD formalism can be derived equivalently to QED by following the same strategies.
Again, the free Lagrangian is given by

$$\mathcal{L} = i\bar{q}\gamma^{\mu}\partial_{\mu}q - m\bar{q}q \tag{2.5}$$

where q now denotes the quark field

$$q(x) = \begin{pmatrix} \psi_{\text{red}} \\ \psi_{\text{blue}} \\ \psi_{\text{green}} \end{pmatrix}(x) \tag{2.6}$$

and where ψ_{red}, ψ_{blue} and ψ_{green} are the four-component Dirac spinors that carry the colour charge *red*, *blue* and *green*, respectively. The quark field transforms locally under SU(3) gauge symmetry according to

$$q(x) \to e^{i\alpha_a(x)\cdot\lambda_a/2} \ . \tag{2.7}$$

λ_a refers to the 3x3 Gell-Mann matrices [13] and the index a takes values from one to eight. For the Gell-Mann matrices, it holds

$$[\lambda_a, \lambda_b] = 2if_{abc}\lambda_c \ . \tag{2.8}$$

f_{abc} refers to the structure constants of QCD[5]. Applying the local transformation in Equation 2.7 to the free Lagrangian (Equation 2.5), results in an additional term, which destroys the local gauge invariance as in QED. The interaction term is absorbed by introducing the covariant derivative

$$D_{\mu} = \partial_{\mu} + ig_s\frac{\lambda_a}{2}G_{\mu}^{a} \tag{2.9}$$

resulting in

$$\mathcal{L}_{\text{inv}} = i\bar{q}\gamma^{\mu}\partial_{\mu}q - m\bar{q}q - \frac{g_s}{2}(\bar{q}\gamma^{\mu}\lambda_a q)G_{\mu}^{a} \ . \tag{2.10}$$

The parameter g_s refers to the coupling strength between quarks and gluons in strong interaction. The covariant derivative also introduces additional gauge fields G_{μ}^{a} representing the eight gluons. As a last

[5] The structure constants are fully antisymmetric and describe the commutation relation between generators of the symmetry group. A definition of all structure constants f_{abc} is given in [13].

step, the free gluon Lagrangian density $\mathcal{L}_{\text{gluon}}$ has to be added:

$$\begin{aligned}\mathcal{L}_{\text{gluon}} &= -\frac{1}{2}\text{Tr}[G_{\mu\nu}G^{\mu\nu}] \\ &= -\frac{1}{4}G^{a}_{\mu\nu}G^{a,\mu\nu} \,. \end{aligned} \tag{2.11}$$

$G^a_{\mu\nu}$ denotes the gluon field strength tensor

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu - g_s f^{abc} G^b_\mu G^c_\nu \,. \tag{2.12}$$

In contrast to the field tensor $F_{\mu\nu}$ of QED, $G^a_{\mu\nu}$ provides an extra term due to the non-vanishing structure constants f_{abc}, so that $\mathcal{L}_{\text{gluon}}$ not only provides kinetic energy terms, but also features self-interaction between gluons. The final Lagrangian density of QCD reads:

$$\mathcal{L}_{\text{QCD}} = i\bar{q}\gamma^\mu\partial_\mu q - m\bar{q}q - \frac{1}{4}G^a_{\mu\nu}G^{a,\mu\nu} - \frac{g_s}{2}(\bar{q}\gamma^\mu\lambda_a q)G^a_\mu \,. \tag{2.13}$$

It describes the kinetics and the interaction of gluons, which are required to be massless in order to preserve gauge invariance, with quarks. Additional terms depending on g_s and g_s^2 are associated with three-gluon and four-gluon vertices. Those gluon self-couplings cause an increase in the coupling strength with decreasing energy scale Q as pointed out earlier. The dependence of the strong coupling strength α_s on Q^2 can be written as [13]

$$\alpha_s(Q^2) = \frac{g_s^2(Q^2)}{4\pi} = \frac{12\pi}{33 - 2n_f \log(Q^2/\Lambda^2)} \tag{2.14}$$

where n_f refers to the number of light quark flavours, which - depending on their mass - contribute at the energy scale Q^2. The QCD scale Λ is a free parameter and must be determined from experiments. It can be interpreted as a threshold marking the transition between asymptotic freedom and the confinement of partons in hadrons. The method used to solve the equation of motion depends on the exact value of $\alpha_s(Q^2)$: For low values of $\alpha_s(Q^2)$, a perturbative expansion of α_s is possible. This regime corresponds to high energy scales Q^2, in which the distances between partons are small. Due to their small coupling, they can be regarded as quasi-free particles. In contrast, for large α_s, i.e. low energy scales, the coupling between partons increases. The consequence of the behaviour of α_s is that quarks are confined inside hadrons. Attempts to dig partons out of hadrons cause the scattered parton to hadronise. These hadronisations result in parton showers (see Section 2.2), which are reconstructed as jets (see Section 4.3.6) in the detector.

2.1.3 The electroweak unification and the Higgs mechanism

The theory of electroweak interaction describes the coupling of massive leptons to the massless electromagnetic field and to the massive Z and $W^\pm$ fields. Further, it reproduces the experimental observations, which demonstrated that the weak interaction is maximal parity violating [29].
Glashow, Salam and Weinberg suggested such a unified theory based on a $\text{SU(2)}_\text{L} \times \text{U(1)}_\text{Y}$ gauge group, which describes the coupling of fermions to three vector fields W^μ with the coupling strength g and the coupling to one single vector field B^μ with strength $g'/2$. The corresponding Lagrangian $\mathcal{L}_{\text{Weinberg-Salam}}$ is written as

$$\mathcal{L}_{\text{Weinberg-Salam}} = \mathcal{L}_{\text{weak bosons}} + \mathcal{L}_{\text{interaction}} \,. \tag{2.15}$$

The first term describes the self-interaction and kinetic energies of the four gauge bosons

$$\mathcal{L}_{\text{weak bosons}} = -\frac{1}{4}W^i_{\mu\nu}W^{i,\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} \tag{2.16}$$

with the field tensors of SU(2) and U(1):

$$W^i_{\mu\nu} = \partial_\mu W^i_\nu - \partial_\nu W^i_\mu - g\epsilon^{ijk}W^j_\mu W^k_\nu \tag{2.17}$$
$$B_{\mu\nu} = \partial_\mu B_\nu - \partial_\nu B_\mu \; . \tag{2.18}$$

In analogy to QCD, ϵ^{ijk} refers to the SU(2) structure constants[6]. The interaction of these fields with fermions and the fermions' kinetic energy are expressed in the second term of Equation 2.15:

$$\mathcal{L}_{\text{interaction}} = \bar{\chi}_L\gamma^\mu\left(i\partial_\mu - g\frac{1}{2}\tau\cdot W_\mu - g'\frac{Y}{2}B_\mu\right)\chi_L + \bar{\psi}_R\gamma^\mu\left(i\partial_\mu - g'\frac{Y}{2}B_\mu\right)\psi_R \tag{2.19}$$

where τ_i $(i = 1, 2, 3)$ are 2×2 matrices defines as

$$\tau_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tau_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \tau_1 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} . \tag{2.20}$$

The weak isospin $I_i = \tau_i/2$ and the hypercharge Y denote the generators of $SU(2)_L$ and $U(1)_Y$, respectively. The hypercharge connects the third component of the isospin I_3 with the electric charge Q according to

$$Y = 2Q - I_3 \; . \tag{2.21}$$

The interaction term in Equation 2.19 distinguishes between left-handed (χ_L) and right-handed (ψ_R) particles. Left-handed particles are arranged in weak isospin doublets. They interact with the fields W_μ of $SU(2)_L$ and with the field B_μ of $U(1)_Y$. In contrast, right-handed particles form isospin singlets and only interact with the field B_μ. Since neutrinos are electrically neutral ($Q = 0$), right-handed neutrinos have assigned an hypercharge of $Y = 0$ and hence do not interact with any field. This reflects the fact that no right-handed neutrinos are observed in nature. All isospin doublets and singlets are listed in Table 2.1.

The isospin doublet χ_L locally transform under $SU(2)_L xU(1)_Y$ as

$$\chi_L \to e^{i\alpha(x)I + i\beta(x)Y}\chi_L \; , \tag{2.22}$$

while the isospin singlet ψ_R transforms according to

$$\psi_R \to e^{i\beta(x)Y}\psi_R \; . \tag{2.23}$$

By adding the covariant derivative to Equation 2.19, local gauge invariance is preserved.

[6] ϵ^{ijk} is equal to +1 (-1) in the case of (anti-) cyclic permutations of 123. In all other cases ϵ^{ijk} is equal to zero.

Table 2.1: Weak isospin I and hypercharge Y of multiplets formed by left-(L) and right-(R) handed quarks and leptons. The singlets of right-handed massless neutrinos are not displayed, since they do not participate in any interaction. Each multiplet is characterised by the same hypercharge Y and same isospin I.

Quark	I	I_3	Y	Lepton	I	I_3	Y
u_L	1/2	+1/2	+1/3	ν_L^e	1/2	+1/2	-1
d_L	1/2	-1/2	+1/3	e^-	1/2	-1/2	-1
c_L	1/2	+1/2	+1/3	ν_L^μ	1/2	+1/2	-1
s_L	1/2	-1/2	+1/3	μ^-	1/2	-1/2	-1
t_L	1/2	+1/2	+1/3	ν_L^τ	1/2	+1/2	-1
b_L	1/2	-1/2	+1/3	τ^-	1/2	-1/2	-1
u_R	0	0	+4/3				
d_R	0	0	$-2/3$	e_R^-	0	+0	-2
c_R	0	0	+4/3				
s_R	0	0	$-2/3$	μ_R^-	0	+0	-2
t_R	0	0	+4/3				
b_R	0	0	$-2/3$	τ_R^-	0	+0	-2

The fields of the electroweak theory can be associated with the observed bosons of weak and electromagnetic interaction. While W^1_μ and W^2_μ correspond to the charged gauge bosons

$$W^{\pm} = (W^1 \mp iW^2)/\sqrt{2}\,, \tag{2.24}$$

the other two fields B_μ and W^3_μ are neutral. They are connected to the physical fields Z_μ and A_μ via the relation

$$\begin{pmatrix} Z_\mu \\ A_\mu \end{pmatrix} = \begin{pmatrix} \cos\Theta_W & -\sin\Theta_W \\ \sin\Theta_W & \cos\Theta_W \end{pmatrix} = \begin{pmatrix} W^3_\mu \\ B_\mu \end{pmatrix} \tag{2.25}$$

where $\Theta_W = \text{atan}(g'/g)$ denotes the weak mixing angle. Further, the elementary charge e can be expressed by

$$e = \frac{gg'}{\sqrt{g^2 + g'^2}} = g' \cos\Theta_W\,. \tag{2.26}$$

So far, the derivation of the electroweak Lagrangian density followed the same theoretical concepts, which were also used earlier in the context of QED and QCD. At this point, it describes the interaction of massless particles with massless gauge fields. In the case of weak interactions, however, the mediators Z and $W^{\pm}$ are bosons with large masses. For this reason, the electroweak gauge theory also has to provide massive gauge fields.
Because the simple attempt of adding mass terms to the Lagrangian density violates the local gauge invariance, the concept of spontaneous symmetry breaking is used. As a consequences of this symmetry breaking mechanism three massive gauge fields and a massive Higgs field are obtained. This Higgs mechanism assigns masses to the gauge bosons of weak interactions and enables to unify weak and electromagnetic interaction in the electroweak theory of Weinberg and Salam.
In the following, the concepts of spontaneous symmetry breaking and the Higgs mechanism are discussed. Furthermore, the Lagrangian density for a spontaneous broken local $SU(2)_L xU(1)_Y$ symmetry is derived, which is necessary to take heavy gauge bosons into account and thus complement the electroweak unification in Equation 2.15.

The concept of spontaneous symmetry breaking

Before considering how the local SU(2) symmetry for weak interaction is broken, the mathematical concept of spontaneous symmetry breaking and the resulting mechanism of mass generation is discussed. The Lagrangian density for a complex scalar field $\Phi = (\Phi_1 + i\Phi_2)/\sqrt{2}$ with global gauge symmetry is given by

$$\mathcal{L} \;=\; (\partial_\mu\Phi)^*(\partial^\mu\Phi) - \mu^2\Phi^*\Phi - \lambda(\Phi^*\Phi)^2 \tag{2.27}$$

with $\lambda > 0$. The last two terms correspond to the potential $V(\Phi) = \mu^2\Phi^*\Phi + \lambda(\Phi^*\Phi)^2$ with a shape depending on the sign of μ^2. For $\mu^2 > 0$, $V(\Phi)$ describes a complex field with mass μ. Its ground state is given by $\Phi_1 = \Phi_2 = 0$ and reflects the symmetry of the Lagrangian. In the case of $\mu^2 < 0$, the minima of the potential are given by

$$v^2 = -\frac{\mu^2}{\lambda} = \Phi_1^2 + \Phi_2^2\,. \tag{2.28}$$

There is an infinite number of possible vacuum states. All solutions are found on a circle in the plane spanned by Φ_1 and Φ_2 as depicted in Figure 2.2. In order to solve the equation of motion one particular

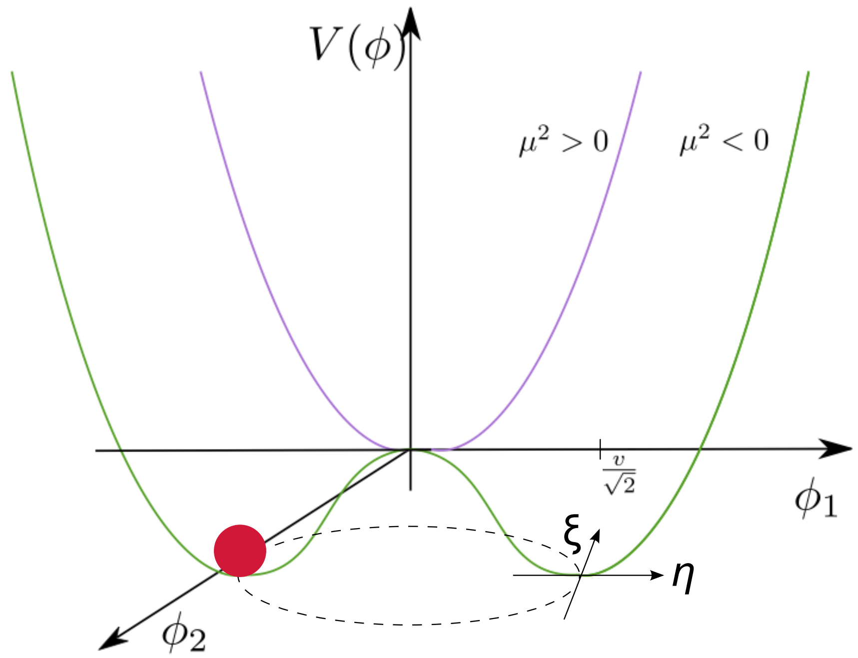


Figure 2.2: The Higgs potential $V(\Phi)$ for a complex scalar field with $\lambda > 0$ and (a) $\mu^2 > 0$ in violet and (b) for $\mu^2 < 0$ in green. For the latter case the minima lie on a circle with radius v and $\Phi_1^2 + \Phi_2^2 = v$. Choosing one particular minimum breaks the local symmetry. The direction of the fields η and ξ are indicated by black arrows.

vacuum state has to be chosen. In doing so, the symmetry is spontaneously broken. The fluctuation around the minimum energy is calculated by arbitrarily choosing the ground state where $\Phi_1 = v$ and $\Phi_2 = 0$. Then, the Lagrangian in Equation 2.32 is expanded around this ground state with respect to the fields η and ξ, i.e. substituting

$$\Phi = \frac{1}{\sqrt{2}}[v + \eta(x) + i\xi(x)] \,. \tag{2.29}$$

The resulting Lagrangian then reads

$$\mathcal{L}_{\text{exp}} = \frac{1}{2}(\partial_\mu \xi)^2 + \frac{1}{2}(\partial_\mu \eta)^2 + \mu^2\eta^2 + \text{higher orders} \,. \tag{2.30}$$

The higher order terms describe interactions and higher order corrections. The second and third term of $\mathcal{L}_{\text{exp}}$ correspond to the kinetic energy of the η-field and its mass. Indeed, the third term can be written as

$$\mu^2\eta^2 = -\frac{1}{2}m_\eta^2\eta^2 \tag{2.31}$$

and hence has the form of a mass term with mass $m_\eta = \sqrt{-2\mu^2}$. The first term of Equation 2.30, $\frac{1}{2}(\partial_\mu\xi)^2$, corresponds to the kinetic energy of the ξ-field and has no associated mass term. Hence, due to the spontaneous symmetry breaking, a massless scalar, known as *Goldstone boson* is introduced.

Spontaneous symmetry breaking is a phenomenon which is found in many physical systems. It is observed when the lowest energy state is not unique. For example, the Lagrangian describing ferromagnets is invariant under rotation in space. In the ground state, the spins are aligned in a particular direction and thus break the rotational symmetry.

Spontaneous symmetry breaking in the local SU$(2)_\mathrm{L}$xU$(1)_\mathrm{Y}$ gauge symmetry

The example above demonstrates how the concept of spontaneous symmetry breaking is used to generate mass terms. In the following, it is extended to the more complex SU$(2)_\mathrm{L}$xU$(1)_\mathrm{Y}$ symmetry group and the masses of the three mediators of the weak interaction, Z and $W^\pm$, are derived.
Starting point is again a Lagrangian density with global gauge symmetry given by

$$\mathcal{L}_\mathrm{Higgs} = (\partial_\mu\Phi)^\dagger(\partial^\mu\Phi) - \mu^2\Phi^\dagger\Phi - \lambda(\Phi^\dagger\Phi)^2 \, . \tag{2.32}$$

Φ denotes an isospin doublet (I=1/2) of complex scalar fields and weak hypercharge Y=1:

$$\Phi = \frac{1}{\sqrt{2}}\begin{pmatrix}\Phi_1 + i\Phi_2 \\ \Phi_3 + i\Phi_4\end{pmatrix} \, . \tag{2.33}$$

The global phase transformation of Φ is given by $\Phi \rightarrow e^{-i\alpha_a(x)\tau_a/2+i\beta(x)Y}\Phi$. Here $\alpha_a(x)$ and $\beta(x)$ are real numbers, while τ_a and Y denote the three generators of the SU$(2)_\mathrm{L}$ group and U$(1)_\mathrm{Y}$ group, respectively. To ensure local phase invariance the covariant derivative

$$D_\mu = \partial_\mu + ig\frac{\tau_a}{2}W^a_\mu + ig'\frac{Y}{2}B_\mu \tag{2.34}$$

is introduced, so that the Lagrangian reads as

$$\mathcal{L}_\mathrm{Higgs} = \left(i\partial^\mu\Phi - g\frac{1}{2}\tau W^\mu\Phi - g'\frac{Y}{2}B^\mu\Phi\right)^\dagger \left(i\partial_\mu\Phi - g\frac{1}{2}\tau W_\mu\Phi - g'\frac{Y}{2}B_\mu\Phi\right) - \mu^2\Phi^\dagger\Phi - \lambda(\Phi^\dagger\Phi)^2 \, . \tag{2.35}$$

The Lagrangian $\mathcal{L}_\mathrm{Higgs}$ describes the Higgs potential $V(\Phi)$ and the interactions between the Higgs field and the electroweak gauge fields. It is added to $\mathcal{L}_\text{Weinberg-Salam}$ in Equation 2.15, so that the full Lagrangian of electroweak theory yields:

$$\begin{aligned}
\mathcal{L}_\text{Weinberg-Salam} &= \mathcal{L}_\text{weak bosons} + \mathcal{L}_\text{interaction} + \mathcal{L}_\text{Higgs} \\
&= -\frac{1}{4}W^i_{\mu\nu}W^{i,\mu\nu} - \frac{1}{4}B_{\mu\nu}B^{\mu\nu} \\
&\quad + \bar{\chi}_L\gamma^\mu\left(i\partial_\mu - g\frac{1}{2}\tau\cdot W_\mu - g'\frac{Y}{2}B_\mu\right)\chi_L + \bar{\psi}_R\gamma^\mu\left(i\partial_\mu - g'\frac{Y}{2}B_\mu\right)\psi_R \\
&\quad + \left(i\partial^\mu\Phi - g\frac{1}{2}\tau W^\mu\Phi - g'\frac{Y}{2}B^\mu\Phi\right)^\dagger\left(i\partial_\mu\Phi - g\frac{1}{2}\tau W_\mu\Phi - g'\frac{Y}{2}B_\mu\Phi\right) \\
&\quad - \mu^2\Phi^\dagger\Phi - \lambda(\Phi^\dagger\Phi)^2 \, .
\end{aligned} \tag{2.36}$$

For $\lambda > 0$ and $\mu^2 > 0$ Equation 2.35 describes the interaction of three massless gauge bosons W_μ and the massless gauge boson B_μ, whereas the symmetry is broken if $\lambda > 0$ and $\mu^2 < 0$. The minimum of

$V(\Phi)$ is given by

$$\Phi^\dagger\Phi = -\frac{\mu^2}{2\lambda} = \frac{1}{2}v\,. \tag{2.37}$$

Choosing the ground state

$$\Phi_0 = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\ v\end{pmatrix} \tag{2.38}$$

breaks the symmetry of $\mathrm{SU(2)_L}$ and $\mathrm{U(1)_Y}$, but not of $\mathrm{U(1)_{em}}$. This is due to the fact that applying the generator $Q = I_3 + Y/2$ of $\mathrm{U(1)_{em}}$ on the vacuum expectation value yields

$$Q\Phi_0 = 0\,. \tag{2.39}$$

Hence, Φ_0 is invariant under $\mathrm{U(1)_{em}}$ transformation. As a consequence, the photon which is the gauge field of $\mathrm{U(1)_{em}}$ remains massless. The massive gauge bosons are produced by the three symmetry breaking generators of $\mathrm{SU(2)_L x U(1)_Y}$. The fluctuation of Φ around its ground state Φ_0 is evaluated and can be written as

$$\Phi(x) = \frac{1}{\sqrt{2}}\begin{pmatrix}\Theta_2(x) + i\Theta_1(x)\\ v + h(x) - i\Theta_3(x)\end{pmatrix} \approx \frac{1}{\sqrt{2}}e^{i\tau\cdot\Theta(x)/v}\begin{pmatrix}0\\ v + h(x)\end{pmatrix}\,. \tag{2.40}$$

It is fully described by four independent real fields $\Theta_1(x)$, $\Theta_2(x)$, $\Theta_3(x)$ and $h(x)$. The key point is that due to the choice of the gauge transformation terms in Equation 2.35 with dependences on the three fields $\Theta_i(x)$ vanish and mass terms of the three W_μ gauge bosons are introduced. In order to calculate the masses of the three W_μ gauge bosons the vacuum expectation value is substituted in the relevant term of Equation 2.35. The relevant term is given by

$$\begin{aligned}\mathcal{L}_{\mathrm{mass}} &= \left(\left(-ig\frac{1}{2}\tau W_\mu - i\frac{g'}{2}B_\mu\right)\Phi\right)^\dagger\left(\left(-ig\frac{1}{2}\tau W^\mu - i\frac{g'}{2}B^\mu\right)\Phi\right)\\ &= \frac{1}{8}v^2g^2\left[\left(W_\mu^1\right)^2 + \left(W_\mu^2\right)^2\right] + \frac{1}{8}v^2\left(W_\mu^3, B_\mu\right)\begin{pmatrix}g^2 & -gg'\\ -gg' & g'^2\end{pmatrix}\begin{pmatrix}W^{3\mu}\\ B^\mu\end{pmatrix}\,.\end{aligned} \tag{2.41}$$

Each of the first two terms is associated to the mass term of one boson, $\frac{1}{2}MW_\mu^2$, with mass $M = \frac{1}{2}vg$. Since the fields W_μ^1 and W_μ^2 are related to the charged bosons $W^\pm$ according to Equation 2.24, these terms are responsible for the masses of the $W^\pm$ bosons. The masses of the neutral gauge bosons are derived from the mixing term of W_μ^3 and B_μ in $\mathcal{L}_{\mathrm{mass}}$:

$$\mathcal{L}_{\mathrm{mass,neutral}} = \frac{1}{8}v^2\left(W_\mu^3, B_\mu\right)\begin{pmatrix}g^2 & -gg'\\ -gg' & g'^2\end{pmatrix}\begin{pmatrix}W^{3\mu}\\ B^\mu\end{pmatrix}\,. \tag{2.42}$$

Expressing $\mathcal{L}_{\mathrm{mass,neutral}}$ in the eigenbasis of the 2×2 matrix yields

$$\mathcal{L}_{\mathrm{mass,neutral}} = \frac{1}{8}v^2\left[gW_\mu^3 - g'B_\mu\right]^2 + 0\left[g'W_\mu^3 + gB_\mu\right]^2\,. \tag{2.43}$$

These terms refer to the mass terms of the physical fields Z and A which are given by:

$$\frac{1}{2}M_Z^2Z^2 + \frac{1}{2}M_A^2A^2\,. \tag{2.44}$$

In agreement with experiments, one obtains three massive gauge bosons with masses

$$\begin{aligned} m_{W^\pm} &= \frac{1}{2}vg, \\ m_Z &= \frac{m_W}{\cos\Theta_W} \end{aligned} \tag{2.45}$$

and one mass eigenstate equal to zero

$$m_A = 0\,, \tag{2.46}$$

which is associated with the massless photon. The vacuum expectation value v depends on the Fermi coupling constant G_F, which is derived from measurements of the muon lifetime [30]. Based on the relation $v = (\sqrt{2}G_F)^{-1/2}$ the vacuum expectation value v is determined to be $v \approx 246\,\text{GeV}$ [31]. Since the massive Z boson is a mixed state of the fields W^3_μ and B_μ, its mass is larger than the mass of the other weak gauge bosons $W^\pm$.
As seen in the example before, the spontaneous symmetry breaking of $SU(2)_L$ introduces four real fields and gives masses to the gauge fields W_μ. The fields $\Theta_i(x)$ refer to the massless Goldstone bosons, which are eliminated by the choice of the gauge transformation. The Lagrangian $\mathcal{L}_{\text{Higgs}}$ also provides an additional particle - the Higgs boson - which is identified with the Higgs field $h(x)$. The requirement of electric charge conservation forces the Higgs boson to be a neutral scalar. Its mass is given by

$$m_H = \sqrt{2\mu^2}\,. \tag{2.47}$$

The coupling of the Higgs boson to the gauge bosons $W^\pm$ and Z is proportional to the squared masses m_W^2 and m_Z^2. In addition, the chosen Higgs potential in Equation 2.35 results in terms, describing the Higgs self-coupling of the order h^3 and h^4.
The experimental evidence of the Higgs boson is essential for the verification of the Higgs mechanism. Within the SM, however, the Higgs mass is a free parameter, which depends on the exact shape of the Higgs potential. The fact that the Higgs mass is not predicted by the SM renders the experimental search difficult such that different mass hypotheses (see Section 2.3) have to be assumed.

Fermion masses

The fermion masses and the Yukawa coupling of fermions to the Higgs field are described by the Lagrangian:

$$\mathcal{L}_{\text{Yukawa}} = (G_1\bar{L}\Phi R + G_2\bar{L}\Phi_c R + \text{ hermitian conjugate}) \tag{2.48}$$

where Φ_c denotes the charge conjugated Higgs field. The same Higgs field Φ, which assigns masses to gauge bosons, also assigns masses to fermions. $\mathcal{L}_{\text{Yukawa}}$ can be rewritten in terms of two Lagrangian densities $\mathcal{L}_{\text{lepton}}$ and $\mathcal{L}_{\text{quark}}$ describing the coupling of the Higgs field to leptons and quarks, respectively. The Lagrangian

$$\mathcal{L}_{\text{lepton}} = -G_\ell\left[\left(\bar{\nu}_\ell, \bar{\ell}\right)_L \begin{pmatrix}\Phi^+ \\ \Phi^0\end{pmatrix} \ell_R + \bar{\ell}_R\left(\Phi^-, \bar{\Phi}^0\right)\begin{pmatrix}\nu_\ell \\ \ell\end{pmatrix}_L\right] \tag{2.49}$$

couples left- and right-handed leptons $\ell(=e,\mu,\tau)$. Choosing the same ground state as before and fixing the gauge yields

$$\mathcal{L}_{\text{lepton}} = -\frac{G_\ell}{\sqrt{2}} v \left(\bar{\ell}_L \ell_R + \bar{\ell}_R \ell_L\right) - \frac{G_\ell}{\sqrt{2}} \left(\bar{\ell}_L \ell_R + \bar{\ell}_R \ell_L\right) h \,. \tag{2.50}$$

The desired lepton mass is proportional to the coupling constant G_ℓ

$$m_\ell = \frac{G_\ell v}{\sqrt{2}} \,. \tag{2.51}$$

The last term of Equation 2.50 contains direct couplings of the Higgs field to the leptons. Since the Higgs boson couples more strongly to heavier leptons, the experimental proof of the proportionality between Higgs-to-fermion coupling is essential to confirm that the Higgs mechanism is indeed the origin of symmetry breaking in electroweak interaction.
The masses of the six quarks are generated similarly. But in contrast to the lepton doublets, masses must also be assigned for the upper elements of the quark doublets. Since the charge conjugated Higgs doublet

$$\Phi_c = \begin{pmatrix} -\bar{\Phi}^0 \\ \Phi^- \end{pmatrix} \tag{2.52}$$

transforms in the same way as Φ, it is used to generate masses for the upper elements, i.e. u, c and t quark.
In weak interactions the mixing of quarks between generations is possible. The mixing in the quark sector is accounted for by expressing the lower member of the quark doublet as a linear combination of flavour eigenstates. Hence the Lagrangian density, revealing masses for the quarks, reads

$$\mathcal{L}_{\text{quark}} = -G_d^{ij} \left(\bar{u}_i, \bar{d'}_i\right)_L \begin{pmatrix} \Phi^+ \\ \Phi^0 \end{pmatrix} d_{jR} - G_u^{ij} \left(\bar{u}_i, \bar{d'}_i\right)_L \begin{pmatrix} -\bar{\Phi}^0 \\ \Phi^- \end{pmatrix} u_{jR} + \text{h.c.} \tag{2.53}$$

with G_u^{ij} and G_d^{ij} being 3x3 matrices. The usual procedure results in

$$\mathcal{L}_{\text{quark}} = -m_d^i \bar{d}_i d_i \left(1 + \frac{h}{v}\right) - m_u^i \bar{u}_i u_i + \left(1 + \frac{h}{v}\right) \,. \tag{2.54}$$

Again, the coupling strength is proportional to the fermion mass.
The above discussion stresses the importance of the experimental evidence of the Higgs boson for the Higgs mechanism incorporated in the SM. The Higgs mechanism is minimal in a sense, that it allows to generate masses of the weak gauge bosons and the fermions consistent with experimental observations. However, their masses remain as free parameters of the SM. As pointed out earlier, the mass of the Higgs boson is a free parameter of the SM and has to be experimentally measured. As soon as the Higgs mass is determined, its production cross-section and decay rate can be calculated. For that reason, searches for the Higgs boson are performed for various mass hypotheses (Section 2.3). However, theoretical considerations are used to set bounds on the possible Higgs mass and hence confine searches at the LHC to certain mass ranges. Previous experimental and theoretical constrains suggested that the Higgs mass is between 114-250 GeV [28]. Indeed, the found Higgs boson with a mass of ≈ 125 GeV is in the favoured mass range.

2.2 Phenomenology of proton-proton collisions

In scattering experiments two particles collide and interact with each other. In this interaction, they can e.g. annihilate and produce new particles. While the initial state, i.e. the colliding particles, and the final state, i.e. the particles after the interaction, are experimentally observable, the transition between the initial and final state is unknown. In fact, different possible transitions refer to the same observation of the initial and final state. Each mathematically described transition can be represented by Feynman graphs. An introduction into the concept of Feynman rules and graphs is found in [14]. In Figure 2.3 two Feynman graphs belonging to the same physical process are illustrated. They visualise different transitions between identical initial and final states. On the left-hand side an electron-positron pair in

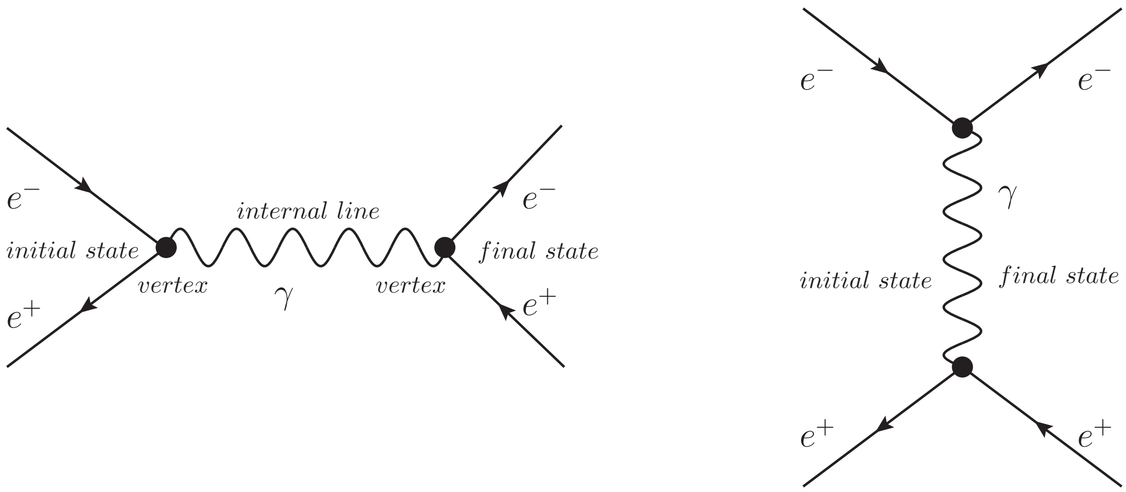


Figure 2.3: Two Feynman diagrams describing the scattering of an electron-positron pair.

the initial and an electron-positron pair in the final state are illustrated. The point where the lines meet is called interaction vertex. Two interaction vertices are connected with each other via internal lines. In this example, the internal line describes a virtual photon which originates from the annihilation of the electron and positron and decays into an electron-positron pair. Another possible transition is shown on the right-hand side. Here, a virtual photon is exchanged between the incoming electron and positron. For a particular process the total transition amplitude, also called matrix element $\mathcal{M}$, is defined by the sum of all possible transitions with amplitude $\mathcal{M}_i$. The probability for a particle to transit from the initial state to the final state is quantified by the cross-section σ which is given in units of an effective area. The differential cross-section $d\sigma$ is proportional to the square of the total transition amplitude $\mathcal{M}^2 = |\mathcal{M}_1 + \mathcal{M}_2 + \mathcal{M}_3 + ...|^2$:

$$d\sigma \propto |\mathcal{M}|^2 d\rho \,, \tag{2.55}$$

where $d\rho$ refers to all possible kinematic states. If the cross-section of a process is known the corresponding event rate N_{events} can be predicted according to

$$N_{\text{events}} = \sigma \cdot \int \mathcal{L}\, dt \,, \tag{2.56}$$

where $\int \mathcal{L}\, dt$ denotes the time integrated luminosity, which depends on the experimental setup. For the LHC, it depends on the number of bunches n, the number of protons per bunch in each beam

$N_{\text{proton beam 1,2}}$, the frequency of collisions f such as the spread of both beams in the x and y direction $\sigma_{x/y}$. It can be written as

$$\mathcal{L} = \frac{n f N_{\text{proton beam 1}} N_{\text{proton beam 2}}}{2\pi\sigma_x\sigma y} . \qquad (2.57)$$

The luminosity is measured as discussed in Section 3.3.
As in most high energy physics analyses, in the presented analysis the expected rates of various physical processes are determined in a particular phase space region and the theoretical expectations are compared to the observed data. In order to obtain the expected event rates, the cross-section has to be calculated. Its calculation is mainly based on perturbation theory. However, some important aspects of the model for proton-proton collisions sketched in Figure 2.4 have to be considered: First of all, since the proton is a compound of partons, the energy fraction carried by the interacting partons is of importance. The energy fraction is described by the probability density function as discussed below. Secondly, interactions observed in the same event take place at different energy scales. While the hard scattering, like the production of a Z or a Higgs boson, is calculated with high precision within perturbation theory, the soft scattering denotes processes at small energy scales at which α_s becomes too large to use perturbative calculations. Pileup and underlying events (see Section 2.2.5) cause significant contributions of soft scattering processes to the total cross-section. Both scattering processes are based on the theoretical framework of QCD, but, due to their nature, they have to be treated separately based on the factorisation theorem [32]. Partons in the initial and final state are accompanied by initial and final state radiation caused by the emission of additional gluons. The resulting parton showers and their hadronisation have to be modelled on a phenomenological basis.
In the following, the concept of perturbative QCD calculations and features of the different parts of the proton-proton collision model mentioned above are briefly introduced.

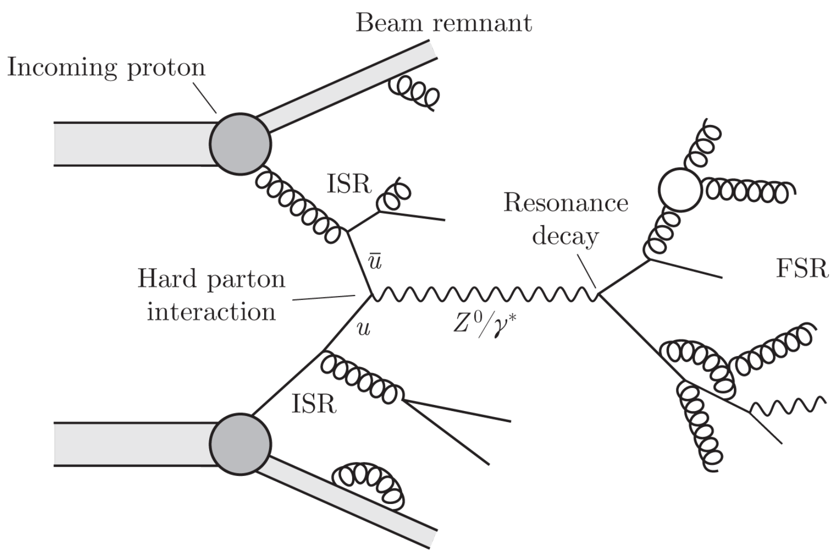


Figure 2.4: Sketch of a proton-proton interaction including soft and hard interactions at parton level: Two incoming protons illustrated as thick grey lines collide. An u quark and an $\bar{u}$ quark interact with each other. The two quarks annihilate and produce a boson which decays into two quarks. In the final state, two quarks, which cause a parton shower, and the remnant of the two protons (thin grey lines) are found. *ISR* and *FSR* denote the initial and final state radiation. The figure is taken from [33].

2.2.1 Concepts of perturbative QCD calculations

The calculation of the cross-section is based on the evaluation of the integral over $|\mathcal{M}|^2$. Since this integral can neither be solved numerically nor analytically, it has to be calculated using perturbation theory. In general, perturbation series arrange the different terms according to the number of interaction vertices as illustrated in Figure 2.5. The leading-order (LO) term describes transitions involving two interaction vertices. It is proportional to the coupling $\alpha_s(Q^2)$. The next higher term is called next-to-leading-order

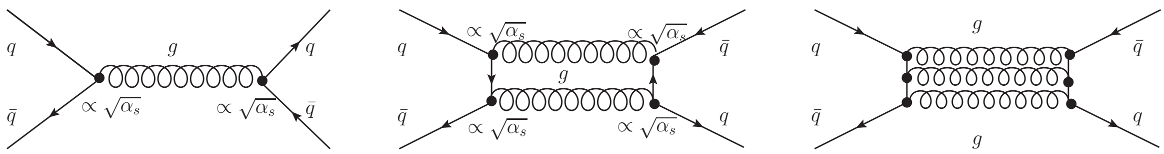


Figure 2.5: Feynman diagrams for a quark and anti-quark pair interacting via strong interaction. They show examples for transitions of different orders: leading-order (left), next-to-leading-order (middle) and next-to-next-to-leading-order (right).

(NLO) and is proportional to $\alpha_s^2(Q^2)$. If $\alpha_s(Q^2) < 1$, the perturbation series converges such that the perturbative expansion is often only evaluated up to next-to-next-to-leading-order (NNLO). In QCD, however, a perturbative expansion is only possible for large energy scales due to the dependence of the strong coupling $\alpha_s(Q^2)$ on the energy scale (see Section 2.1.2).
In the perturbative expansion divergent terms due to virtual loops can occur in higher orders. In order to ensure the renormalisability of the SM these divergences have to be removed [34]. Therefore, the experimentally measurable quantities like the coupling constants (g_{exp}) or the boson masses (m_{exp}) are assumed to be composed of a bare quantity (g_0 or m_0) and an extra factor (δ_g or δ_m). In the renormalisation step the bare quantities which refer to the divergent terms are subtracted from the result. The remaining renormalised quantities are associated to the experimentally observed quantities. The energy scale Q^2 at which the divergent terms are subtracted is called renormalisation scale μ_R. For strong interaction the coupling strength $\alpha_s(Q^2)$ depends on μ_R.

2.2.2 Parton density function

In proton-proton collisions, a parton q_1 belonging to proton h_1 interacts with a parton q_2 of proton h_2. Each parton carries an a-priori unknown momentum fraction $x_i P_i$ of the proton momentum P_i. However, the probability of a parton carrying a certain momentum fraction can be expressed by the parton density function (PDF) $f_{q_i/h_i}(x_i)$, so that the hadronic cross-section reads

$$\sigma_{h_1 h_2 \to X} = \sum_{q_1 q_2} \int_0^1 dx_1 \int_0^1 dx_1 f_{q_1/h_1}(x_1) f_{q_2/h_2}(x_2) \int d\hat{\sigma}_{\text{partonic}}(x_1 P_1, x_2 P_2) \,. \tag{2.58}$$

The partonic cross-section $d\hat{\sigma}_{\text{partonic}}(x_1 P_1, x_2 P_2)$ is weighted by the PDFs of the two corresponding partons and integrated over all possible momentum fractions x_i. Finally, the sum over all initial parton pairs is calculated.
This equation is justified by the factorisation theorem, which allows the calculation of the hard scattering process $\hat{\sigma}_{\text{partonic}}$ independent of soft scattering phenomena. Both, the partonic cross-section and the PDFs, depend on the so-called *factorisation scale* μ_F, which divides the energy scale Q^2 into the non-perturbative soft regime ($Q^2 < \mu_F$) and the region of hard scattering ($Q^2 > \mu_F$). The PDFs are evaluated at the factorisation scale. Divergences in the perturbative expansion of processes at low energy scales

($Q^2 < \mu_F$) are absorbed by the PDFs. The cross-section also depends on the renormalisation scale μ_R introduced in Section 2.2.1. The exact values of μ_F and μ_R are arbitrary. They are in general chosen to be identical to match the energy scale which corresponds e.g. to the invariant mass M of the final state leptons [35].
The PDF only depends on the energy scale Q^2 and can therefore be exploited for every kind of hard scattering process. Since it cannot be calculated using perturbative expansions, it is obtained from data fits based on measurements of earlier experiments. For example, the PDFs for the LHC are derived from other experiments, e.g. at HERA [36], at lower energy scales and extrapolated to higher energy scales [37]. Several sets of PDFs including uncertainties are provided by e.g. CTEQ [38] [39] or MSTW [40] and are shown in Figure 2.6.

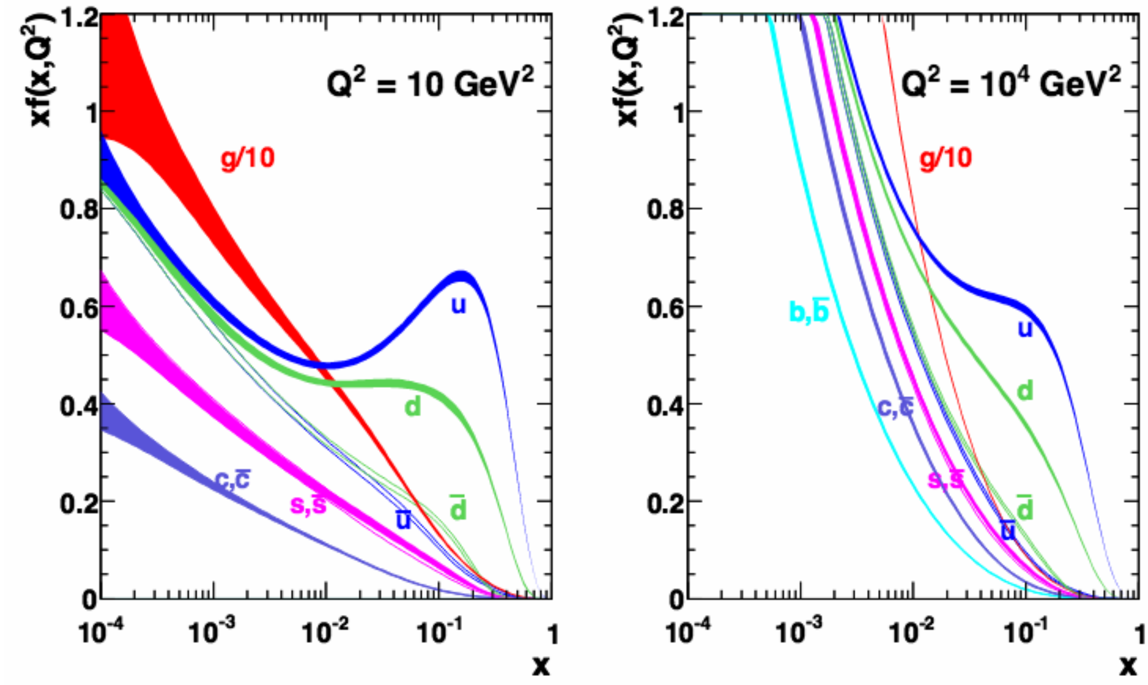


Figure 2.6: The parton density as a function of the energy x for quarks and gluons provided by MSTW 2008 NLO at the energy scales $Q^2 = 10\,\mathrm{GeV}^2$ (left) and $Q^2 = 10\,\mathrm{TeV}^2$ (right). The bands illustrate the 1σ uncertainty. The figures are taken from [40].

2.2.3 Parton shower

Scattering processes at high energies involving partons can be accompanied by an additional emission of gluons, quarks or photons. Although the calculations of many important physical processes at NLO consider the emission of one such particle, the description is not sufficient to explain the occurrence of a large number of partons in the final state. However, event generators like ALPGEN factorise the interaction into the hard scattering process as well as further final state radiation. While the hard process is determined by matrix element calculations, the showering of the additional partons must be approximated. In order to model parton showers, the parton at the energy scale of hard processes Q has to be subsequently split into further partons until the energy scale of hadronisation Q_0 is reached. The probability of the parton A to be split into two partons B and C at a given energy scale Q^2 is given by

$$dP_A = \frac{\alpha_s}{2\pi}\frac{dQ^2}{Q^2}P_{A\to BC}(z)dz \tag{2.59}$$

with z being the fraction of energy which parton A transfers to parton B. Considering all possible splittings of A into parton combinations and energy transfers z a set of differential equations is obtained. These equations still suffer from divergences in the soft (emission of partons with small momentum) and collinear (small opening angle between emitted and original parton) limit, that have to be eliminated by the so-called *Sudakov form factor* [41], which describes the probability of no parton splitting between two scales ΔQ^2. Based on these probabilities the parton splitting is repeated at different scales q until the hadronisation scale Q_0 is reached.
The several event generators use slightly different implementations, e.g. different starting scales Q. As a consequence, the parton showering differs slightly between generators. Additional matching algorithms [42] are implemented in order to resolve the conflicts that are caused if the emission of further partons is already included in calculated higher order corrections for the hard cross-section. A detailed overview of the implementation of models for parton showering is given in [43].

2.2.4 Hadronisation

After parton showering, partons in the final state are found at low energies where they form bound states, called hadrons, due to confinement. This transition from individual partons to composite hadrons in the final state is referred to as hadronisation. Since the hadronisation involves small momentum transfers, it cannot be simulated using perturbation theory. Instead phenomenological methods, which operate at energy scales well separated from the perturbative QCD region, are used for simulation.
Different models are implemented in the generators. While PYTHIA [44] employs the *Lund-String model*, which immediately transforms partons into hadrons, HERWIG [45] uses the *cluster hadronisation model*, which constructs massive cluster objects before decaying them into hadrons. In both models unstable hadrons decay into hadrons with long enough lifetimes. The parton shower and its hadronisation is sketched in Figure 2.7 for the *Lund-String model*. A detailed description of the hadronisation methods can be found in [46].

2.2.5 Pileup and underlying event

During the collision of two protons only two partons are involved in the hard scattering process. The remaining partons undergo soft scattering. This further activity is also referred to as underlying event. Several models within MC generators attempt to describe the underlying event structure based on observations rather than precise theoretical predictions.
Pileup activity denotes soft processes originating from other proton-proton interactions either in the same (*in-time*) or neighbouring (*out-of-time*) bunch crossings. Due to the huge number of protons in each bunch, the correct modelling of pileup activity is challenging. Pileup and underlying event can have a significant effect on the performance of momentum measurements and can also impact the training of multivariate classifiers (see Section 5.2) in physics analyses. For example, the $H \to \tau\tau$ analysis strongly relies on the modelling of VBF topologies, which are characterised by two jets associated to the VBF production mechanism (see Section 2.3.1). The energy measurement of these VBF jets (and all other jets) can be influenced by energy deposits in the calorimeter due to pileup and hence has to be corrected [47]. Although MC generators contain models for pileup and underlying events, which are tuned with respect to these activities, data-based-background estimates are preferred (see Chapter 6).

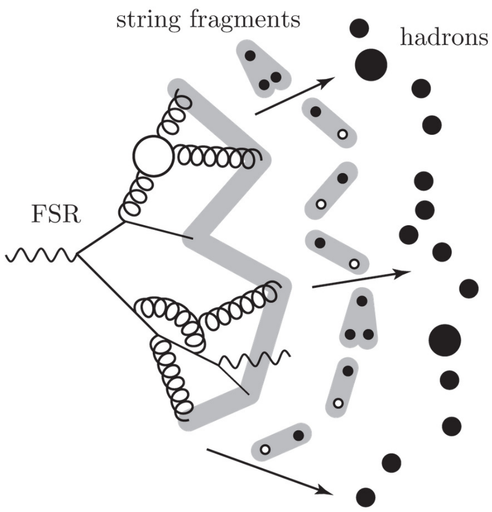


Figure 2.7: Sketch of the parton shower from Figure 2.4 and its hadronisation: All partons produced in the parton shower are connected by a string (long grey line). This string splits into string fragments (small grey bands) from which hadrons (black balls) are formed. The figure is taken from [33].

2.3 Higgs production and decay mechanisms at the LHC

From earlier experiments performed at colliders such as LEP at CERN and the Tevatron at Fermilab, some mass ranges could already be excluded from the search for the Higgs boson. Hence, in the first years of data taking the searches for the Higgs boson at the LHC focused on a mass range of 100 GeV$<m_H<$600 GeV and resulted in the discovery of a new particle, whose properties are consistent with the SM predictions. In this section, the predictions that can be made on the production and the decay of the Higgs boson once a fixed Higgs mass is chosen are presented.
Further details on the Higgs boson phenomenology at the LHC including theory calculations are provided by the Higgs cross-section working group. Their results are published in three reports [16–18] to which this section mainly refers to.

2.3.1 Production of the SM Higgs boson

The SM Higgs boson at hadron colliders is produced via one of four possible production mechanisms illustrated in Figure 2.8: gluon-gluon fusion (ggF), vector boson fusion (VBF), Higgs-Strahlung (VH) and Higgs production in association with one top pair ($t\bar{t}H$). The production cross-section of the Higgs boson strongly depends on the interacting partons, the Higgs mass and the center-of-mass-energy ($\sqrt{s}$) of the two colliding beams as indicated in Figure 2.9. Increasing the center-of-mass energy from 7 TeV to 14 TeV enhances the cross-section by 300-400% in the low Higgs mass region. The dominant production mechanism is the ggF process (Figure 2.8a), because gluons significantly contribute to the density function of protons at small momentum fractions. Since this process is mainly induced by top-quark loops, the corresponding QCD corrections are large. For a Higgs boson with mass m_H=125 GeV the ggF cross-section is 19.3 pb at $\sqrt{s}$ =8 TeV. The ggF cross-section contains contribution from the finite top and bottom mass in leading-order. QCD radiative correction at NLO are of the order of 80-100%.

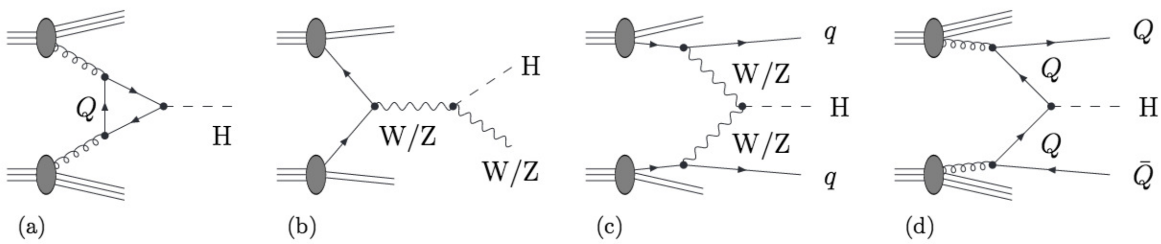


Figure 2.8: The main SM Higgs production processes at hadron colliders at leading-order, with q (Q) denoting light (heavy) quarks: (a) gluon-gluon fusion, (b) Higgs-Strahlung, (c) vector boson fusion, (d) Higgs production in association with heavy quarks. The figures are taken from [19].

Additionally, QCD corrections in the m_{top}-limit in the NNLO result in corrections of about 25%. The precision in NNLO is improved by the resummation of the contributions from the soft gluon emission in all orders at next-to-next-to-leading log (NNLL). Further, electroweak corrections strongly depend on the Higgs boson mass and are about +5% for m_H =125 GeV. Uncertainties on the theoretical predictions are discussed in the context of the analysis in Section 7.7.3.

The VBF production mechanism (Figure 2.8c) has the second largest cross-section. In this process, two quarks radiate a vector boson (W or Z). These two vector bosons then annihilate to a Higgs boson. In contrast to the ggF production mechanism in which partons annihilate, in the VBF production the two interacting quarks undergo a scattering process. Due the two scattered quarks being well separated in pseudorapidity the VBF process has a distinct signature in experimental searches. For that reason, it is a very suitable channel for Higgs searches despite of its small cross-section of 1.58 pb (for m_H=125 GeV at $\sqrt{s}$ =8 TeV). The cross-section for VBF production is calculated in strong and electroweak couplings at NLO. In contrast to ggF corrections, the NLO QCD corrections are small and only of the order of 5-10%. Higher NNLO QCD corrections are approximated by a structure function approach [48]. The size of electroweak corrections is conditioned by the renormalisation scheme and similar to the size of QCD corrections.

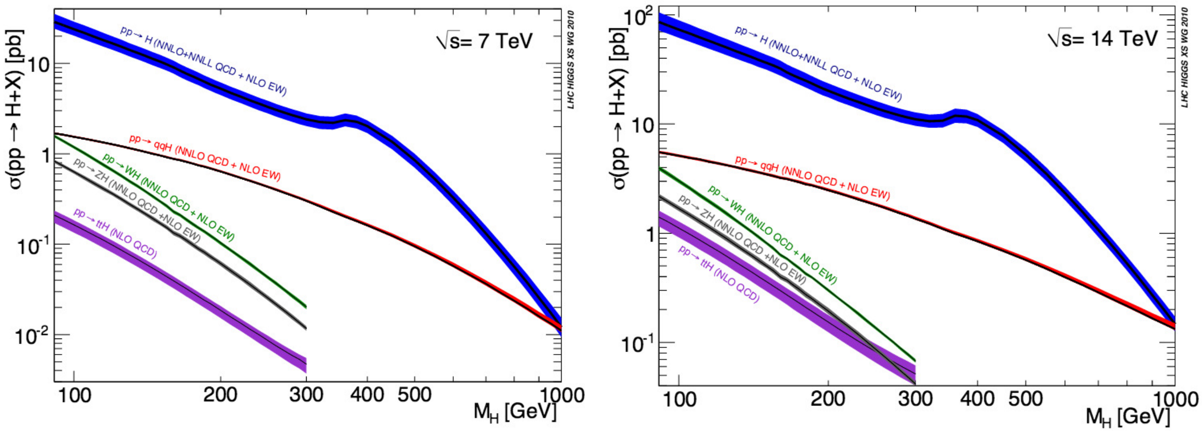
The process of Higgs-Strahlung (Figure 2.8b) is the third most likely production mechanism. Here, two quarks annihilate into a weak vector boson (W or Z), which radiates off a Higgs boson (H). The cross-section for WH (ZH) is 0.705 pb (0.415 pb) for m_H=125 GeV at $\sqrt{s}$ =8 TeV. While QCD corrections are calculated up to NNLO, electroweak corrections are determined only up to NLO.

The $t\bar{t}H$ process (Figure 2.8d) has the lowest cross-section, since the two top quarks in the final state limit the available phase space. Due to its small cross-section it can be neglected in the context of this thesis.

More details on the different generators used to simulate the Higgs signal such as the corrections from theory are discussed in Section 7.1.

2.3.2 Decay channels of the SM Higgs boson

The Higgs boson is an unstable particle decaying most likely into the heaviest particle-antiparticle pair which is kinematically allowed. In such a decay, the strength of the interaction is proportional to the masses of the fermions and vector bosons produced (m_f/v and m_V^2/v). Hence, for a given Higgs mass m_H the branching ratios for the various possible decay channels strongly depend on the mass of the decay products. Higgs boson searches were performed over a large range for m_H. Due to the mass dependences of the branching ratios, the accessible final states differ and the search strategies have to



(a) (b)

Figure 2.9: The SM Higgs production cross-section for different production mechanisms at the LHC as a function of the Higgs boson mass at (a) $\sqrt{s}$ =7 TeV and (b) $\sqrt{s}$ =14 TeV. The comparison of both figures shows the dependence of the cross-section on the center-of-mass energy. The figures are taken from [16].

be adjusted for the different mass hypotheses. The leading-order decay modes relevant for the low-mass SM Higgs search at the LHC are illustrated in Figure 2.10. While the Higgs directly couples to vector bosons and τ leptons or bottom quarks, the decay into photons or gluons is indirectly induced via heavy quark loops (or vector bosons in the case of $H \rightarrow \gamma\gamma$).
The total Higgs boson decay width considering all possible decays at the given mass covers a wide range

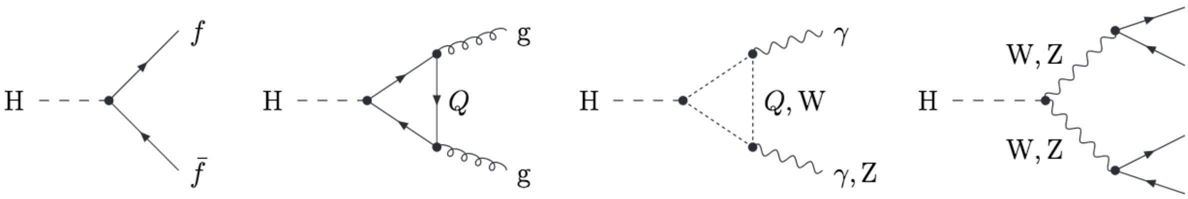


Figure 2.10: Leading-order diagrams for different decay channels of the SM Higgs boson. While f ($\bar{f}$) denote the decay into (anti-) fermions like b quarks or τ leptons, Q describes any heavy quark. The figures are taken from [19].

from 100 keV for Higgs masses below 10 GeV and 10^3 GeV at about 1 TeV. The mass dependence of the total decay width for the region accessible by the LHC is depicted in Figure 2.11. In the case of a Higgs boson with mass 125 GeV, the total width Γ_H is about 4 MeV [49]. Even in Higgs decay channels with very good mass resolution, this width is still about three orders of magnitude smaller than the experimental resolution and hence too small to be directly measured.
The branching ratios as well as the Higgs decay width are calculated using the software packages HDECAY [50] and PROPHECY4F [51], which generates the four particle decay $H \rightarrow WW/ZZ^{*} \rightarrow 4f$. The software determines all partial widths with high precision including corrections from QCD and electroweak processes, and derives the particular branching ratios from the complete set of partial widths. Corresponding theoretical uncertainties result from missing higher order corrections. In addition, parametric uncertainties are considered. Those arise from uncertainties on the SM input parameters like the strong coupling strength α_s.

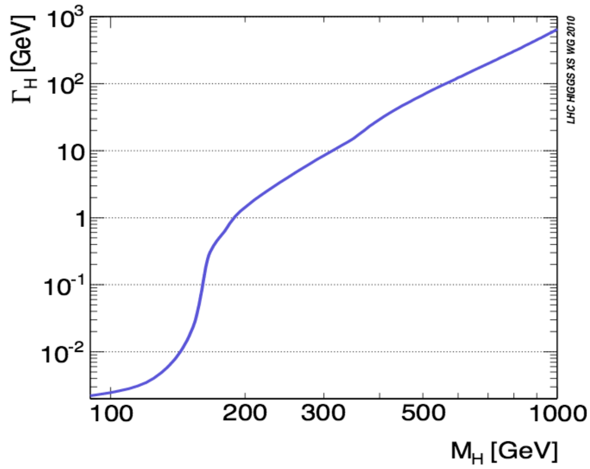
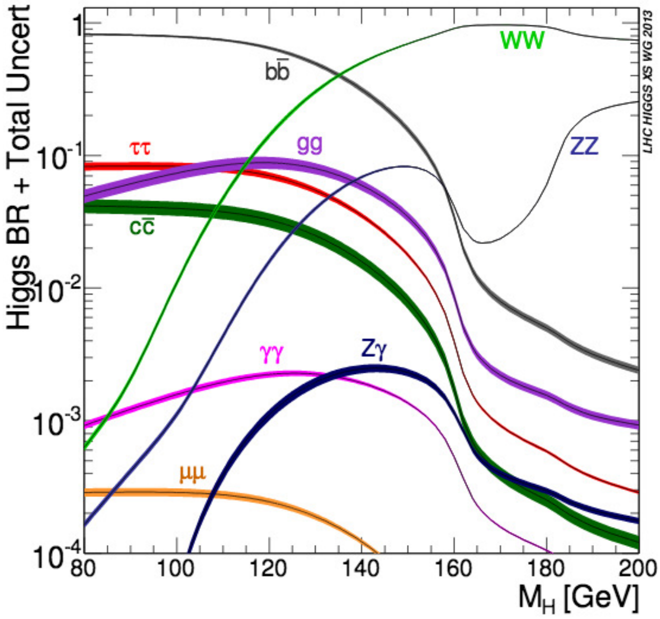
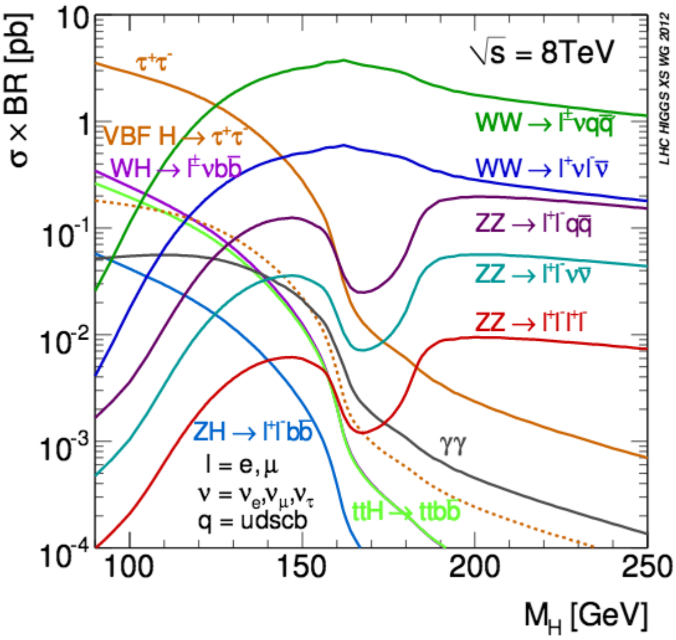


Figure 2.11: The total width of the SM Higgs boson as a function of the Higgs boson mass in the mass range accessible at the LHC. The figure is taken from [16].

As mentioned before, the branching ratio for the various decay modes strongly depends on the Higgs boson mass so that only certain combinations of Higgs production mechanisms and decays can be studied at the LHC. The mass dependence is shown in Figure 2.12a visualising the accessible decay modes as a function of the Higgs masses. The width of the bands illustrate the overall uncertainties originating from missing higher order effects. While at high masses the decay into the heavy vector bosons dominates ($H \rightarrow WW^*$ and $H \rightarrow ZZ^*$), for low masses the largest contribution arises from the decay into fermions, especially into b quarks. The range at around m_H =125 GeV is from physics perspective a very interesting region as it allows to analyse the decay of the Higgs boson in diverse channels. In particular, the predictions of the SM Higgs boson decays into fermions as well as bosons can be directly probed. All accessible decay channels at m_H =125 GeV are briefly introduced below. Details on the reconstruction and signature of particular physics objects are discussed in Chapter 4 and 5.
For the experimental evidence of the SM Higgs boson, the $H \rightarrow \gamma\gamma$ [1, 5, 52] decay is of high importance. Compared to other decay processes the branching ratio of the $H \rightarrow \gamma\gamma$ decay is rather small. However, it reveals a clear signature. The two photons are reconstructed with high accuracy and low misidentification probability. Further, prompt photons are distinguishable from photons originating from π^0 decays. This clear signature allows to identify the decay mode. The Higgs boson mass can be fully reconstructed by the four momenta of the two photons. This channel, however, is only accessible in a small mass range between 110 GeV< m_H <140 GeV.
Besides the di-photon decay, the $H \rightarrow ZZ^* \rightarrow \ell\ell\ell'\ell'$ ($\ell = e, /\mu$) decay channel [1, 5, 52] also allows to fully reconstruct the event with an excellent mass resolution. The clear signature counterbalances the low branching ratio at low Higgs boson masses. The analysis of the $H \rightarrow ZZ^*$ channel focuses on the Z boson decays into light leptons. These processes are almost background free because hadronic processes hardly contribute to the background. In contrast, hadronic decays indicate high misidentification probabilities. Further, the full reconstruction of invariant mass is impossible, because neutrinos cannot be detected. The decays of the Z bosons also allow studies of spin and CP quantum numbers.
The analysis of the $H \rightarrow WW^*$ [1, 5, 52] channel, which is the dominant decay channel in a wide Higgs boson mass range, concentrates on decays of the W bosons into muons or electrons in order to suppress large contributions from the W+jets or multijets background. Due to the neutrinos originating from the W decay the Higgs boson mass cannot be completely reconstructed. By investigating the angular



(a)



(b)

Figure 2.12: Top: Branching ratios for SM Higgs decays as a function of the Higgs boson mass m_H for the most important decay channels in low mass SM Higgs searches. The line width corresponds to the theoretical uncertainties [16]. Bottom: Production cross-section times branching ratio as a function of m_H. The $H \to \tau\tau$ channel produced in VBF (ggF) is illustrated as solid (dashed) orange line. The figures are taken from [16].

distributions of the decay products, conclusions on spin quantum numbers of the Higgs boson can be drawn.
For a Higgs with mass 125 GeV two fermionic decay modes are accessible at the LHC: The $H \to b\bar{b}$ channel [53] has the largest branching ratio at low Higgs boson masses. Due to the large background contributions from multijet events, the search for this channel is very challenging. The search focuses on the VH production processes, in which the vector boson decays leptonically. The focus on this particular production mechanism enables an effective lepton triggering.
The only leptonic decay mode observed with Run 1 data is the $H \to \tau\tau$ channel. This channel enables to directly study the Yukawa coupling of the Higgs boson to τ leptons. Hence, it is essential for probing the mechanism for the fermion mass generation. At m_H =125 GeV, the branching ratio for the $H \to \tau\tau$ channel is 6.3% with an uncertainty of about ± 6% [17]. The $H \to \tau\tau$ channel has one of the largest sensitivities with respect to the VBF production mechanism (see Figure 2.12b). By exploiting the VBF topology, large parts of the background processes can be rejected, which renders the analysis of the $H \to \tau\tau$ most sensitive towards VBF production.

CHAPTER 3

The LHC and the ATLAS experiment

The analysis presented in this thesis is based on data taken with the ATLAS detector [54], which is one of the experiments at the Large Hadron Collider (LHC) [55]. It is situated at the European Center For Nuclear Research (CERN) in Geneva, Switzerland (see Figure 3.1).
The main part of the LHC is a synchrotron ring with a circumference of 26.7 km. Two proton beams are injected into the ring in opposite directions and are then accelerated. The protons collide at four interaction points. At these points the main experiments are installed in order to study the collision products. With a design luminosity of $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-2}$ and a design center-of-mass energy of $\sqrt{s} = 14\,\mathrm{TeV}$, the LHC operates at the high-energy frontier of particle physics. Not only does it provide high precision measurements of SM parameters, but it also allows for searches for new physics phenomena.
The experiments are designed for several purposes: The ATLAS and CMS experiment are multi-purpose detectors. The LHCb experiment [56] is designed to investigate decays of hadrons containing c or b

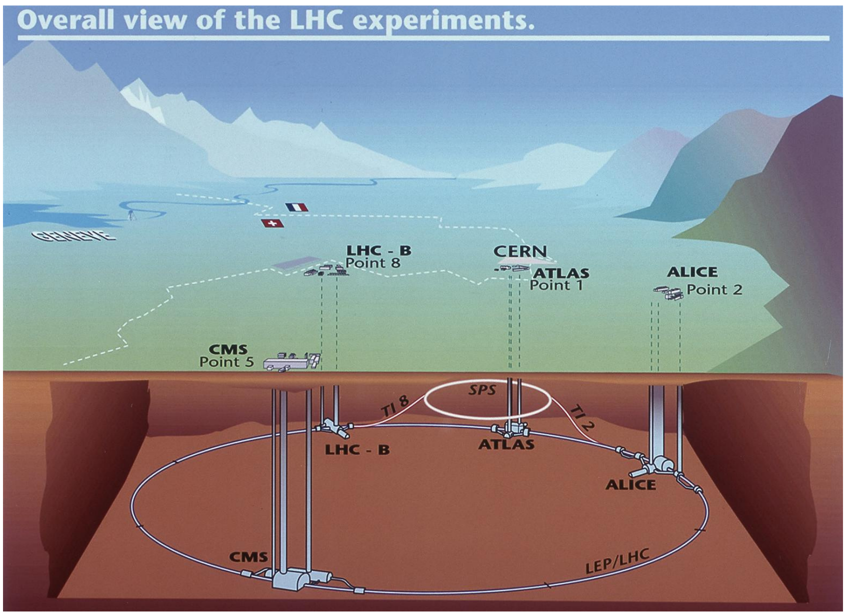


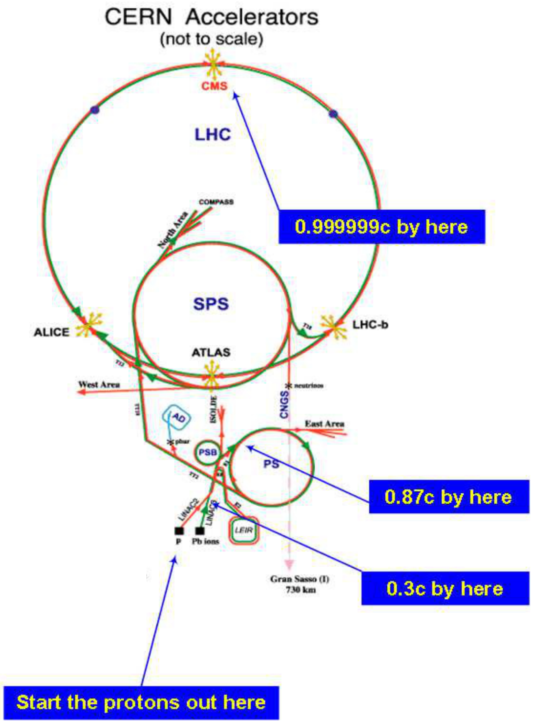
Figure 3.1: The LHC and the location of the ATLAS, CMS, LHCb and ALICE experiment. The figure is taken from [57].

quarks while the ALICE detector [58] collects data particularly from heavy ion collisions. Since the last two experiments investigate processes with higher cross-sections, they are supplied with less luminosity. Besides those large experiments some other experiments are situated at the LHC as well: The TOTEM experiment [59] studies elastic and diffractive scattering, whereas the LHCf [60] experiment measures neutral particles, which are emitted in forward region in proton-proton collisions, and is used for studies of cosmic rays. Finally, the MoEDAL experiment [61] is built to search for magnetic monopoles and other exotic particles.
After a short introduction to the proton acceleration at the LHC in Section 3.1 a brief explanation of the coordinate system used by ATLAS follows in Section 3.2. In Section 3.3 the different subdetectors of the ATLAS experiment are summarised. Finally, Section 3.4 gives an overview of the running conditions during the first data taking period between 2011-2012. The data collected in this period is the basis for this thesis.

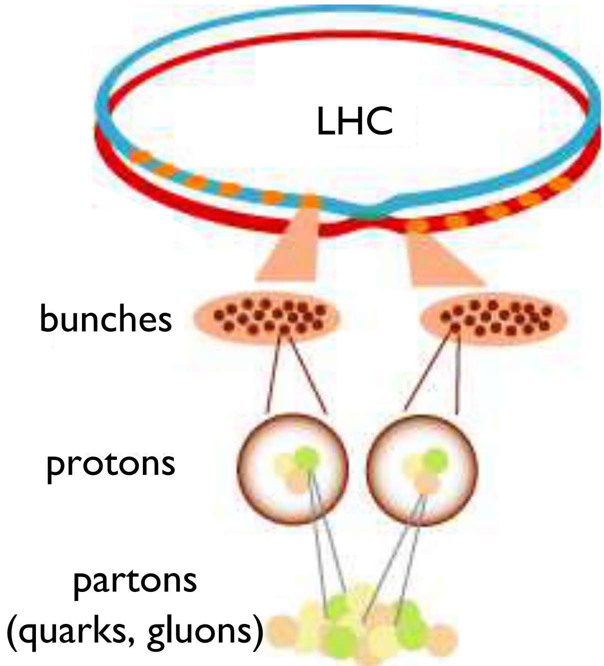
3.1 The LHC

The LHC [55] is situated in the tunnel of the former LEP (Large Electron Positron Collider) machine roughly 100 m below the surface. It is designed to collide proton-proton, proton-Pb and Pb-Pb beams. While in proton-proton collisions the fundamental interaction between partons is investigated, heavy ion collisions are an appropriate method to study QCD dynamics in hot and dense environments [58]. Compared to e^+e^- colliders, hadron colliders do not suffer from high energy losses due to synchrotron radiation, since protons are about a 1000 times heavier than electrons. Hence, collisions at very high center-of-mass energies become possible. Disadvantages of such colliders are additional activity in the detector originating from further interactions of the same or neighbouring proton-proton collision (see Section 2.2) and the unknown initial states.
Before the proton beams in the LHC can collide at the design center-of-mass energy of 14 TeV, several pre-accelerators illustrated in Figure 3.2a are used[1]: Protons obtained from ionising hydrogen gas are accelerated with linear accelerators up to 50 MeV and with the help of quadrupoles grouped in sets of bunches. Afterwards, they are injected into the *Proton Synchroton Booster* accelerating them to 1.4 GeV. The bunch train pattern which consists of 72 neighbouring bunches with a spacing of 25 ns is delivered by the *Proton Synchrotron*. This accelerator further accelerates the protons to 25 GeV. The last pre-accelerator is the *Super Proton Synchrotron* which accelerates the protons to 450 GeV and then injects two beams into the beam pipe of the LHC. There, they are bent on a circular trajectory by the magnetic field of 1232 superconducting dipole magnets. At 7 TeV the magnetic field strength has to be 8.33 T in order to deflect the beams onto the ring. Hence, the magnetic field is a limiting factor for high beam energies. The proton beams are focused by 392 quadrupole magnets. In addition, 8 superconductive cavities operating at 400 MHz generate an electric field which is necessary to further accelerate these protons and to compensate energy losses due to synchroton radiation. Per turn the particle bunches gain an energy of 485 keV. In total, the LHC can be filled with 39 bunch trains, each consisting of 2808 bunches. Each of the bunches contains about 10^{11} protons as sketched in Figure 3.2b. The proton beams with energies of 7 TeV are brought to collisions with a crossing angle of 285 μrad in four interaction points at which detectors are installed. The beam lifetime depends on the luminosity loss within the collisions and is about 14.9 h. In the first data-taking period during 2010-2012 the LHC was operated at center-of-mass energies of $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV and a bunch spacing of 50 ns was used.

[1] The following values refer to the design values of the LHC.



(a)



(b)

Figure 3.2: Left: Schematic figure of the CERN accelerators used to accelerate protons injected into the LHC ring [62]. Right: Sketch of collisions at the LHC [63]. Protons are arranged in bunches within the beam. These bunches are brought to collisions so that the constituents of the protons are interacting with each other.

3.2 The coordinate system and common variables

In the description of collisions events recorded by the ATLAS detector a right-handed coordinate system is used as shown in Figure 3.3: Its origin is given by the nominal proton-proton interaction point situated in the detector's center. The z-axis runs along the beam pipe. While the x-axis points from the interaction point to the center of the LHC ring, the y-axis points upwards such that the transverse plane spanned by the x and y axis is perpendicular to the beam pipe. Due to the cylindrical geometry of the detector, trajectories in the transverse plane are described in polar coordinates (r, ϕ, θ), where $r = \sqrt{x^2 + y^2}$ denotes the radial distance from the interaction point and $\phi = \arctan y/x$ ($\theta = \arctan(r/z)$) the azimuthal (polar) angle around the beam pipe. The pseudorapidity η is derived from the rapidity for massless particles and is given in terms of the polar angle θ as $\eta = -\ln\tan(\theta/2)$. The angular distance between two objects is measured in units of $\Delta R \equiv \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$ [54].
The partons carry a fraction x of the proton momentum which, for an ensemble of events, is predictable via parton distributions functions (see Section 2.2.2). With respect to a particular proton-proton interaction, however, the longitudinal and the transverse momentum of the initial parton cannot be determined. But in contrast to the longitudinal component the transverse momentum is approximately zero and is therefore neglected. For that reason at the LHC interactions are often considered in the transverse plane in which due to momentum conservation the sum over the transverse momenta of all scattered particles $\vec{p}_T^{\,i}$ is zero:

$$\sum \vec{p}_T^{\,i} = 0 \,. \tag{3.1}$$

The transverse momentum $\vec{p}_T^{\,\mathrm{miss}}$ of particles like neutrinos, which do not interact with the detector material, are estimated from momentum conservation

$$\vec{p}_T^{\,\mathrm{miss}} = \sum \vec{p}_T^{\,i} \tag{3.2}$$

taking all detected particles i into account. Its magnitude is called missing transverse energy E_T^{miss} (see Section 4.3.8). A large measured E_T^{miss} indicates neutrinos within the detected event.

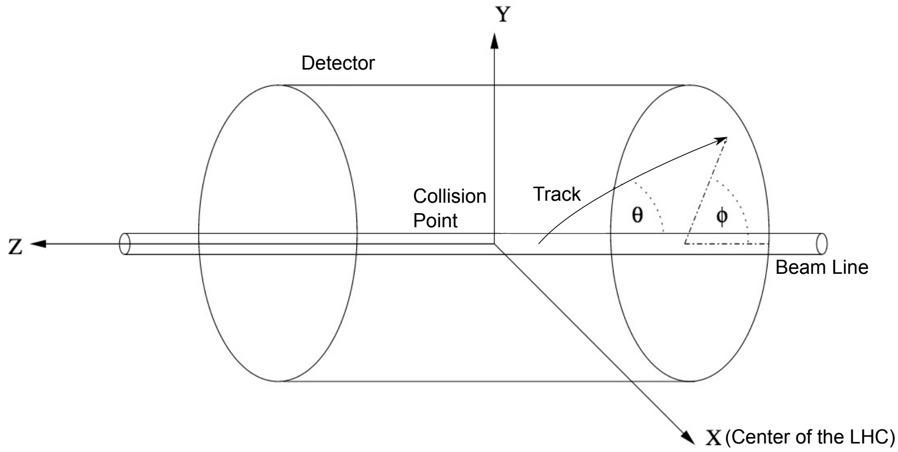


Figure 3.3: The coordinate system of the ATLAS detector. The figure is adopted from [64].

3.3 The ATLAS detector

In this thesis, data recorded with the multi-purpose ATLAS detector in 2011 and 2012 is analysed. The detector, which weighs 7 kt and has a diameter of 25 m, cylindricallly encloses the beam pipe (diameter of 36 mm) on a length of 44 m along the z-direction. Hence, it covers a solid angle of almost 4π around the point of interaction. To ensure a backward and forward symmetry the interaction point is located in the middle of the detector.
The ATLAS detector depicted in Figure 3.4 consists of three different subsystems: the *inner detector*, the *calorimeter* and the *muon spectrometer*. These subdetectors are designed to provide a good resolution of all detected objects in order to be able to precisely reconstruct hard scattered events. For the identification and reconstruction of b hadrons[2] and τ leptons a high spatial resolution of the reconstructed production vertex is necessary. In order to accurately reconstruct the tracks and the momenta of particles, the detector has to measure a high number of space points over large distances. At the same time, the energy loss due to bremsstrahlung, which occurs when a particle is slowed down while traversing through the detector material, has to be reduced as much as possible. In order to measure the momenta of particles their trajectories are bent by a superconducting solenoid magnet, which encloses the inner detector and creates a field strength of 2 T. In addition, the magnetic field of three toroid magnets is required to bend muon trajectories to allow for measurements of the muon momenta in the muon spectrometer over a wide momentum range. Moreover, the detector provides a precise measurement of energies and an accurate reconstruction of E_T^{miss}. This is achieved by a combination of electromagnetic and hadronic calorimeters with high energy resolutions.

[2] Bound states of b quarks and lighter quarks are referred to as b hadrons.

Due to the high luminosity multitudes of collisions take place. An efficient trigger system selects events of interest within this large number of collision events. The luminosity itself is determined with the help of the LUCID and ALFA detectors [54], which measure the number of charged particles in a distance of ±17 m and ±240 m, respectively, from the interaction point.
Starting from the inside to the outside, the following subsections introduce the different subsystems according to [54, 65].

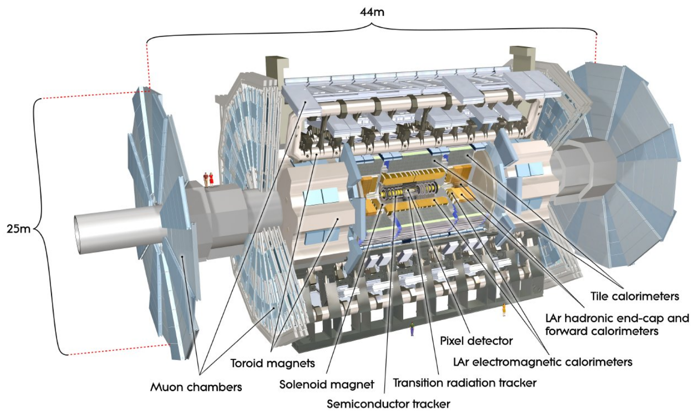


Figure 3.4: The ATLAS detector and its different components. The figure is taken from [54].

3.3.1 The inner detector

The *inner detector* (ID) shown in Figure 3.5 is the smallest component with a diameter of 2.1 m. It consists of three sub-detectors used to measure the momentum of charged particle tracks within $|\eta| < 2.5$. The magnetic field of the solenoid magnet bends the tracks, so that their momentum can be calculated according to the momentum-curvature relation.
The *pixel detector*, the component closest to the interaction point, is built up from three cylindrical layers in the barrel region. The innermost layer is also called b-layer, because it is of high importance for the determination of the secondary vertices for b quark decays. On these layers about 80.4 million pixels with a nominal size of 50×400 µm are located, which makes precise track and vertex reconstruction possible. Additional pixels are situated on three disks on each of the two endcaps. They are ordered on staves consisting of 13 modules. When a particle crosses the pixel detector it leaves about three hits. Each of these hits has a resolution of 10 µm in the (r, ϕ) plane and 115 µm in z-direction, respectively.
The *Silicon Microstrip Tracker* (SCT) is a silicon strip detector which consists of four layers in the barrel region and nine disks in each endcap. With an active area of 63 m the SCT is able to measure four space points in a range of $|\eta| < 2.5$. Two 285 µm thick sensors are built from 768 active strops with a strip pitch of 80 µm. They are attached to modules which enclose a stereo angle of 40 mrad.
The small angle enables the measurement of three-dimensional space points with a precision of 17 µm in the (r, ϕ) plane and 580 µm in z-direction, respectively. The sensors are single sided p-in-n silicon detectors. They are supplied with a voltage of 150 V. The voltage can be adjusted to ensure the efficient collection of charged particles. Like the pixel detector, the SCT is cooled in order to suppress radiation effects on the detector performance.
The outermost component is the *Transition Radiation Tracker* (TRT), which consists of polymide drift tubes with a diameter of 4 mm. The 144 cm long tubes, installed parallel to beam pipe, are convoluted with polyproyplen and polyethylen fibres. They are filled with a gas mixture based on Xenon.

Their working concept is similar to proportional counters. Charged particles traversing the tubes create charges due to gas ionisation, which are then collected by a gold plated tungsten wire inserted in the tubes. The tubes provide a measurement with an accuracy of 130 µm. Due to their geometry, however, the particle's position in z-direction is not detectable. Despite of the worse hit precision compared to the SCT, with roughly 36 hits per particle it is of large importance for the track reconstruction. Besides the detection of low-threshold TRT hits due to ionisation, high-threshold hits originating from transition radiation are measured with the help of the embedded fibres. The transition radiation is emitted if a particle crosses the boundary between two materials with different dielectric constants. Since the intensity of the transition radiation depends on the Lorentz factor $\gamma = E/m$, the TRT enables the identification of particles. In particular, it provides a separation of electrons from muons and pions over a wide momentum range.

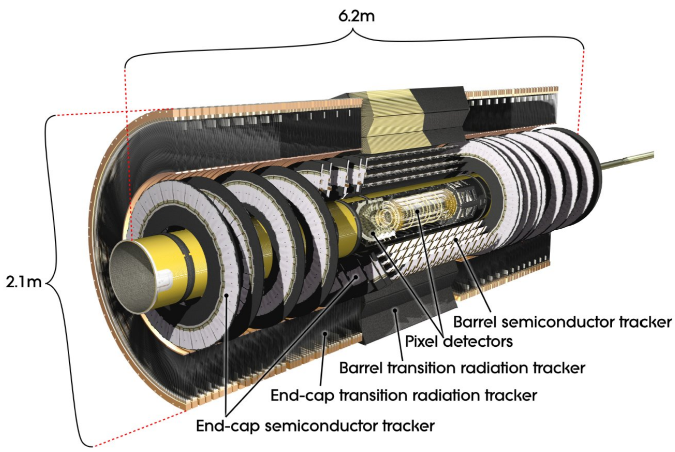
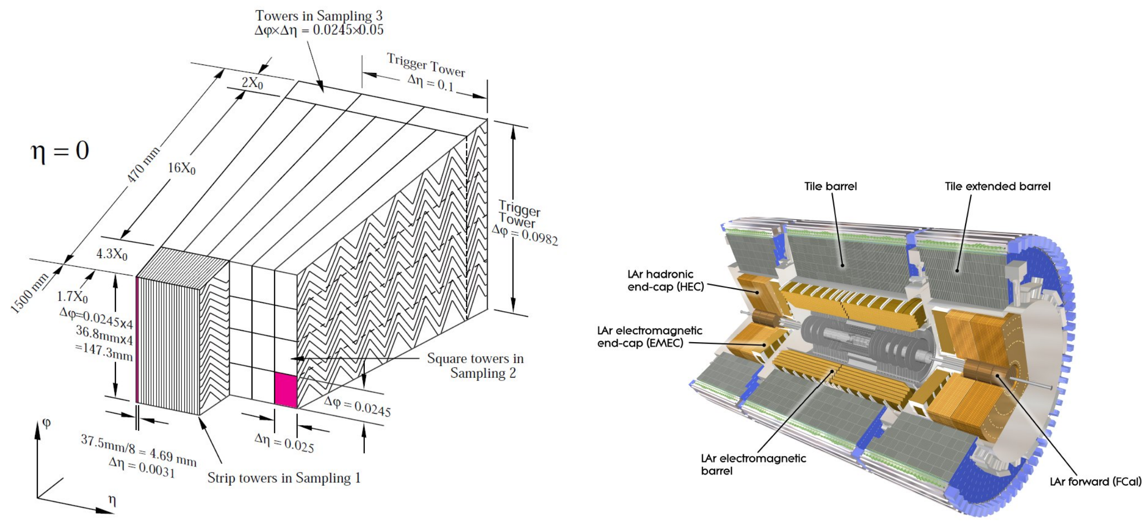


Figure 3.5: The inner detector and its three components: pixel detector, SCT and TRT. The figure is taken from [66].

3.3.2 The calorimeter

The *electromagnetic calorimeter* (ECAL) encloses the ID and is followed by the *hadronic calorimeter* (HCAL) (see Figure 3.6).
The ECAL is specialised to measure the energy of photons and electrons. Its center is built by the EM barrel calorimeter (diameter 2.8 m-4 m) with a coverage of $|\eta| < 1.4$ and a 4 mm gap at $|z| = 0$. In contrast, the EM endcap calorimeters have a coverage in forward region of $|\eta| <3.2$ and ranges from r =0.33 m to r =2.1 m. The ϕ-symmetric ECAL consists of three different layers in accordion shape as depicted in Figure 3.6. Each of these layers is built up from narrow strips with a granularity of $\Delta\eta \times \Delta\Phi = 0.003 \times 0.1$ resulting in an high resolution in η. It provides an overall good angular resolution for photon and electron reconstruction. Further, the ECAL also enables to identify $\pi^0 \to \gamma\gamma$ decays, which is of importance for the τ reconstruction. Due to its transverse segmentation $\pi^0 \to \gamma\gamma$ decays are distinguishable from primary photons. Since most of the electromagnetic showers are found in the first two layers (granularity of the second layer: $\Delta\eta \times \Delta\Phi = 0.025 \times 0.25$), the outer layer ($\Delta\eta \times \Delta\Phi = 0.05 \times 0.025$) is suitable to differentiate between electromagnetic and hadronic deposits.
In front of the EM calorimeter, a 11 mm thick liquid argon (LAr) layer is installed in order to correct for energy losses within the ID. Further, the EM calorimeter, which is also based on LAr, is penetrated by lead absorber plates with a thickness of 1.33 mm to 1.53 mm. Two wheels within a distance of 3 mm in

between form one of the EM endcaps. They have the same geometry as the barrel calorimeter.
For the τ identification, the subsequent HCAL, in which hadronic τ-decay products are measured, is essential. It consists of the hadronic tile, endcap and the LAr forward calorimeter. The cylindrical tile barrel calorimeter (radius 2.3 m- 4.3 m) employs scintillating tiles as active material intercepted by steel absorbers. Photomultiplier used for read out are installed at the tile edges with wavelength shifting fibres. As the ECAL, it is formed by three layers, but with a coarser granularity of $\Delta\eta \times \Delta\Phi = 0.1 \times 0.1$ (first two layers) and $\Delta\eta \times \Delta\Phi = 0.2 \times 0.2$ (third layer) respectively since hadronic showers are broader than electromagnetic ones in most cases. Two disks with a coverage of $1.5 < |\eta| < 3.2$ form the endcap calorimeter, which is designed as copper/LAr sampling calorimeter. Its granularity is $\Delta\eta\times\Delta\Phi = 0.1\times0.1$ ($\Delta\eta\times\Delta\Phi = 0.2\times0.2$) below (above) $|\eta| = 2.5$. The LAr calorimeter is a combination of electromagnetic and hadronic calorimeters and is built up from three layers. While in the first layer copper is used in order to measure electromagnetic showers, the last two layers provide the detection of hadronic showers using tungsten. The active material is LAr which is located in gaps between the absorber material.
Due to the hermetic geometry the calorimeter enables a precise reconstruction of E_T^{miss}, which is of high importance for the analysis of the $H \to \tau\tau$ decay channel. Further, the design of the calorimeter does not only allow the precise measurement of particle energies but also to differentiate between various objects. In addition, the thickness of the calorimeter prevents particles except for muons from reaching the muon spectrometer, which is necessary to measure the complete energy of all hadrons.



(a) (b)

Figure 3.6: Left: The ECAL with its three layers and cell granularity in the $\eta - \phi$ space. Right: The complete ATLAS calorimeter system. The figures are taken from [54].

3.3.3 The muon spectrometer

The outermost part of the ATLAS detector is the *muon spectrometer* (MS) which encloses the HCAL. Its purpose is the measurement of the position and the momentum of muons in a momentum range of 3 GeV-1 TeV with an acceptance of $|\eta| < 2.7$. Due to the calorimeter design muons are the only particles from primary interactions which are able to reach the MS. Since, in general, muons loose their energy via ionisation, they deposit only a small amount of their energy (roughly 3.5 GeV) in a detector. This requires a larger detector volume for track reconstruction compared to other particles. In order to bend

muon tracks three air-core toroid magnets, each consisting of eight coils, generate a magnetic field. The muons subsequently pass monitored drift tubes, cathode strip chambers and thin gap chambers, which have different response times. The monitored drift tubes, organised in the barrel in cylindrical layers in 5 m - 10 m distance from the interaction point as well as in the endcap region, are made of aluminium tubes filled with an Ar/CO2 gas mixture. They have a drift time of about 700 ns and a spatial resolution of 35 μm. Due to the high resolution and the resulting accurate hit positions the chambers are required to be perfectly aligned at a precision of 30 μm, which is realised by an optical alignment system. Further, tungsten-rhenium wires supplied with 3080 V are used to collect ionised electrons.
The cathode strip chambers operating in the region between $2.0 < |\eta| < 2.7$ are multi-wire proportional chambers. They allow to resolve the hits in two dimensions. Besides these chambers, *Resistive Plate Chambers* are installed on both sides (back side) of the middle (outer) drift tube layer in order to obtain a fast trigger system. In addition, they complement the information provided by the monitored drift tubes by a further two-dimensional position measurement. The Resistive Plate Chambers consist of two electrode plates which are paralleled in a distance of 2 mm creating a magnetic field of strength 4.9 kV mm^{-1}. The space between both plates is filled with a gas mixture.

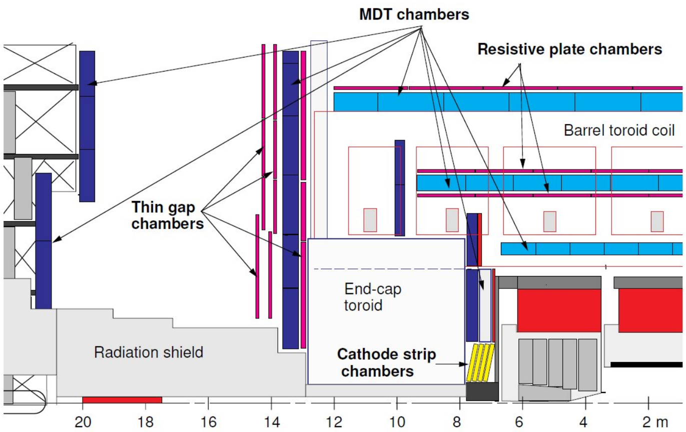


Figure 3.7: The muon spectrometer. The figure is taken from [66].

3.3.4 The ATLAS trigger system

Due to the high collision rate and the corresponding high luminosity a large number of inelastic interactions per bunch crossing is observed, which makes it technically impossible to store all events. A remedy is found by the installation of an appropriate trigger system filtering events with potentially interesting topologies and hence reducing the original event rate of 40 MHz to 200 Hz [67].
The implemented trigger system contains three levels as illustrated in Figure 3.8: the *level-1 hardware trigger* (L1), the *level-2 software trigger* (L2) and the *event filter trigger* (EF). Each of these trigger levels exploits the information obtained by the upstream trigger and includes additional conditions. Because of the short bunch spacing time data is written into a buffer. The L1 trigger makes a fast decision within 2.5 μs. Those decisions are based on information delivered by the different detector components. To obtain a decision, the L1 trigger scans the calorimeter for energy depositions in electromagnetic or hadronic trigger towers, which reveal a coarser granularity ($\Delta\eta \times \Delta\Phi = 0.1 \times 0.1$) than later used in the off-line reconstruction. Regions with large energy depositions are referred to as *regions of interest* (ROI) and are exploited to build different trigger objects. The L1 trigger is able to reduce the event rate from

40 MHz to 75 kHz. A further reduction is achieved by the high level software trigger processing the L1 triggered events and their ROIs. Based on those ROIs the L2 enables more sophisticated decisions by the use of the complete granularity of the ATLAS detector. Although its functionality is similar to the off-line reconstruction, the L2 does not provide a full event reconstruction. Instead, it only evaluates the event within the ROIs. Focusing exclusively on ROIs makes the L2 with 40 ms still rather fast. In the last step, the EF reconstructs the full event based on information of all detector subsystems. The event reconstruction is very similar to the off-line object reconstruction (discussed in Chapter 4) and thus can already reject large parts of possible background sources. Its processing time is about four seconds.
The reconstructed particle momenta and energies are checked against the required trigger conditions. The different selection requirements are provided by the so-called trigger-menu collecting different conditions suitable to identify certain topologies, e.g. isolation or kinematic criteria on τ leptons, jets, muons or electrons.

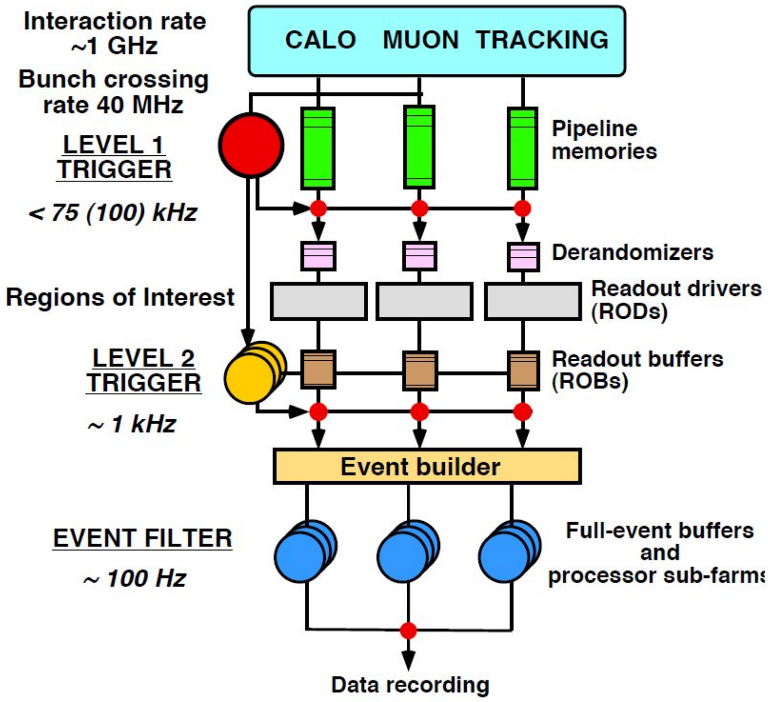


Figure 3.8: Scheme of the ATLAS trigger system. The figure is taken from [68]

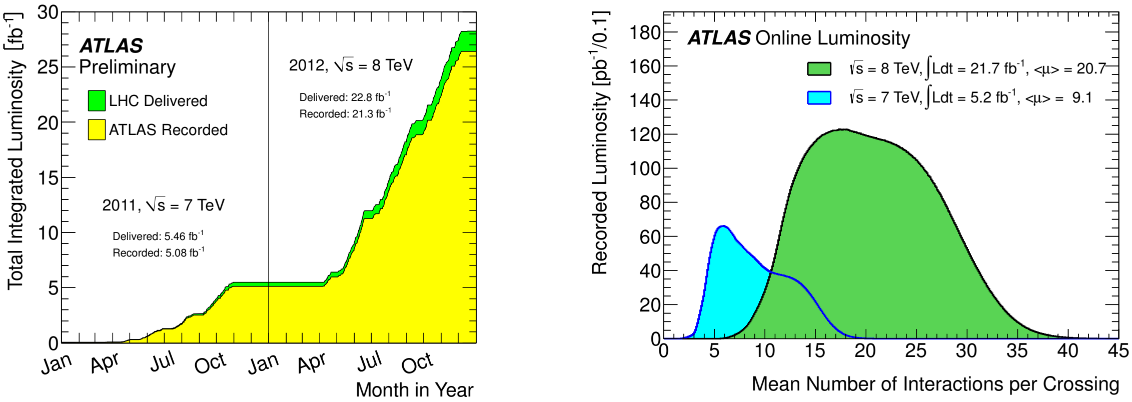
3.4 Run 1 period

The experiments at the LHC collected the first data between the end of 2009 and the beginning of 2013. In this data taking period, also referred to as Run 1, data from proton-proton as well as heavy ion collisions were successfully recorded. In the years of data taking, the beam energies and the delivered instantaneous luminosities have been increased. The operation periods of the LHC during Run 1, which are also summarised in Table 3.1, are briefly explained below:
The first proton-proton collisions took place at the end of 2009 at $\sqrt{s}$ =900 GeV. In the following weeks the center-of-mass energy was subsequently increased up to $\sqrt{s}$ =7 TeV. At this center-of-mass energy the LHC delivered a luminosity of 48.1 pb^{-1} in 2010. The 2010 dataset already permitted precision

Table 3.1: Delivered and recorded luminosities in proton-proton collisions in Run 1. This thesis is based on the datasets collected in 2011 and 2012. Further details on the luminosity measurements with ATLAS are found in [69].

Year	$\sqrt{s}$	Delivered luminosity	Recorded luminosity
2010	7 TeV	$48.1\,\mathrm{pb}^{-1}$	$45.0\,\mathrm{pb}^{-1}$
2011	7 TeV	$5.46\,\mathrm{fb}^{-1}$	$5.08\,\mathrm{fb}^{-1}$
2012	8 TeV	$22.8\,\mathrm{fb}^{-1}$	$21.3\,\mathrm{fb}^{-1}$

measurements of the SM. With a higher beam intensity but same beam energy in 2011, $5.25\,\mathrm{fb}^{-1}$ were recorded corresponding to an efficiency of about 90%. The largest Run 1 dataset with $21.7\,\mathrm{fb}^{-1}$ was obtained during the data taking period in 2012, in which the center-of-mass energy was enhanced to $\sqrt{s}$ =8 TeV. The cumulative luminosity as a function of time is shown in Figure 3.9a demonstrating a significant increase in the delivered luminosity in the middle of 2012.
In Run 1, the protons were grouped in bunches with a spacing of 50 ns, which is twice as the design bunch spacing (25 ns). Due to the large number of protons per bunch, however, a peak luminosity of $7.7 \times 10^{33}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-2}$ was achieved in 2012, which is slightly below the design luminosity of $10^{34}\,\mathrm{cm}^{-2}\,\mathrm{s}^{-2}$. One consequence of the increasing luminosity during these two years is a higher pileup



(a) (b)

Figure 3.9: Left: The cumulative luminosity delivered by the LHC (green) and recorded with the ATLAS detector (yellow) during proton-proton collisions taking place in 2011 and 2012 at centre-of-mass energy of $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV, respectively. Right: Mean number of interactions per bunch crossing in the 2011 (blue) and 2012 (green) dataset. The distributions are luminosity weighted and taken from [69].

activity in 2012 quantified by the number of interactions per bunch crossing. Figure 3.9b compares the number of interactions per bunch crossing in 2012 to the one in 2011. The mean number of interactions increased from about nine interactions in 2011 to 20 in 2012.
Breaks between proton-proton data taking periods were used for heavy ion collisions. These were carried out in the winters of 2010/2011, 2011/2012 and 2012/2013. Run 1 ended in 2013. From 2013 until May 2015 no data was collected in order to prepare the LHC for Run 2, which started in May 2015 with $\sqrt{s}$ =13 TeV. The data collected between May and November 2015 corresponds to a delivered luminosity of $4.2\,\mathrm{fb}^{-1}$ [70].
This thesis is based on the proton-proton collisions recorded with the ATLAS detector in 2011 and 2012.

For the analysis a *Good-Run-List* (GRL) is applied in order to ensure a high data quality. It contains a list of good-quality runs[3] and corresponding luminosity blocks[4]. A run is tagged as good-quality run if no defects in one of the detector's subsystems is significantly affecting the detector performance. In case of small defects the run is written into the GRL and affected blocks are masked. These blocks are not considered within the analysis. Based on blocks which are listed as good-quality data in the GRL, the overall luminosity has to be recalculated, which results in a slightly smaller luminosity than delivered by the ATLAS detector as visualised in Figure 3.9a.

[3] *Run* denotes a short period of data taking.
[4] Each run is subdivided into several luminosity blocks. Each of them contains about the same number of events.

CHAPTER 4

Monte Carlo simulation and event reconstruction

The ATLAS data is analysed with the ATHENA software framework [71] which provides algorithms for a wide range of data-processing application in ATLAS, e.g. the reconstruction of physics objects. In Run 1 the ATHENA off-line reconstruction provided several data formats differing in the amount of information available for each event and therefore suitable for different use cases. Since for physics analyses the experimental results have to be compared to theoretical predictions, the detector response of signal and background processes have to be simulated. Event generators are used to simulate event signatures at parton level. Afterwards the response of the ATLAS subdetector components is simulated. Finally, general reconstruction algorithms also used for collision data are employed. However, the reconstruction performance and efficiency of the object identification in Monte Carlo (MC) simulation slightly differs from collision data due to possible imperfections of the detector simulation. For that reason calibrations are required.

The reconstruction algorithm used for simulated events and collision data includes the reconstruction of physics objects starting from hits and tracks as well as the reconstruction of particle four-momenta and associated properties based on detector-level information.

This chapter focuses on the event simulation as well as the reconstruction and identification of physics objects relevant for the $H \to \tau\tau$ analysis. The description refers to algorithms used during the 2012 data-taking period.

4.1 Data formats in ATLAS

The ATLAS data model of Run 1 provides several data formats. They differ in the detail of information about hits, cells, clusters and physics objects. Higher level formats, which are used for analyses, are derived from lower level formats by applying several slimming and thinning algorithms.

The most fundamental data format is given by the RDO (Raw Data Objects) format. It contains information on hits originating from the various subsystems of the ID. The energy deposits in the cells of the electromagnetic and hadronic calorimeter are also stored. From this data format the higher level ESD (Event Summary Data) format is obtained. In ESDs, particle tracks are accessible. In addition, all information concerning the calorimeter are provided in a compactified way. The AOD (Analysis Object Data) format is the next higher data format, which is already suitable for user analyses. It only provides a selective access to lower level objects. For example, instead of cells, only energy clusters are stored. The DESD (Derived Event Summary Data) and DAOD (Derived Analysis Object Data) format are derived from the ESD and AOD format by applying filters. The filter requirements select physics objects with certain properties, e.g. $Z \to \mu\mu$ events, and hence reduce the number of stored events. The D3PD (Derived Physics Data) format is mainly used in physics analyses. It contains the reconstructed physics objects and their properties. The exact content can differ depending on the use case, e.g. D3PDs used in the $H \to \tau\tau$ analysis include observables related to the τ reconstruction.

4.2 Event generation and simulation

The search for the SM Higgs boson or other phenomena in high-energy physics strongly relies on an accurate simulation of physical processes within the ATLAS detector. This does not only imply the simulation of hard scattering processes, but also the simulation of pileup and underlying event contributions as well as the response of the different subdetector systems and triggers. These components are considered within the ATLAS simulation infrastructure [72]. The simulation consists mainly of three steps: First, events are generated using external event generators. The event generation contains the simulation of the hard scattering process including all immediate decays and also accounting for QCD effects and resulting parton showers. Pileup interactions from both, the same and neighbouring bunch crossings, are considered. An overview of the different event generators used in this thesis is given in Section 7.1.
In the second step, the transition of these particles through the ATLAS detector including interactions with the detector material, is simulated by the GEANT4 [73] toolkit. GEANT4 provides a full simulation of the detector geometry, hits and tracks for a large number of particles over a wide energy range. The different numerical models describing the interactions of particles are ordered in so-called physics-lists within GEANT4 [73]. In the third simulation step, the detector response is simulated by converting the created hits into *digits*. Digits result from an excess of voltage or current above a certain threshold in one of the various readout channels. Several subdetector digitisation packages are provided for channel specific differences. After the digitisation the simulated event is handed over to the common object reconstruction also used for collision data.
Compared to other steps of the simulation chain, the detector simulation requires large computing resources. For that reason, several attempts were made to simplify the models for the detector simulation and consequently safe computing resources. Such simplifications are possible, because for many use cases a detailed detector simulation is not compulsory. Comparisons of fast simulation programs FATRAS and FastCaloSim with GEANT4 [74] still show room for improvements. Nevertheless, regarding the increasing amount of data in Run 2, fast simulation will become of interest for many analyses.

4.3 Object reconstruction

The following section describes the reconstruction and identification of objects relevant for the $H \to \tau\tau$ analysis. Starting from fundamental objects, tracks and clusters, the reconstruction of physics objects which are necessary to find $H \to \tau\tau$ signatures are discussed. Finally, the reconstruction of E_T^{miss} which is an indicator for undetected neutrinos is introduced. Although for some objects several reconstruction algorithms exist, here, the focus lies on methods used for the analysis presented in this thesis.

4.3.1 Track reconstruction

Charged particles with sufficiently large lifetimes traverse the ID and leave hits in the different subdetectors. After converting the hit information into three-dimensional space points, tracks are reconstructed following the procedures described in [75]. One differentiates between two kinds of tracks: While *primary tracks* refer to tracks of particles originating from the primary vertex (see Section 4.3.2), *secondary tracks* summarise all other particle tracks, e.g. tracks from photon conversions. The first group of tracks is reconstructed by the INSIDE-OUT algorithm provided that the track transverse momentum exceeds 400 MeV. This algorithm starts searching for hits in the pixel and SCT detector and extrapolates them into the TRT taking into account material geometry as well as bending due to the magnetic field. The OUTSIDE-IN reconstruction is used for secondary tracks and extrapolates track segments in the TRT to the primary interaction point by matching SCT and pixel hits. Afterwards, both kinds of tracks are refitted using all matched space points.
A track is characterised by five parameters: d_0 and z_0 denote the distance between the track and the primary vertex in the transverse and longitudinal plane, respectively. The parameters are measured with respect to the point of closest approach of the track to the vertex. The angles ϕ_0 and Θ_0 describe the polar and azimuthal angle of the direction of the track in the transverse plane at the point of closest approach. A track is further characterised by the ratio of the particle's charge and the absolute momentum. Tracks are used for vertexing and further particle reconstruction. The latter implies the application of object-specific track quality requirements, like the number of hits in the different subdetectors used to reject badly reconstructed jets.

4.3.2 Vertex reconstruction

Vertices are reconstructed using an iterative finding algorithm [76]. They are required to have at least two associated tracks. For robustness more than two tracks are often required.
Vertex seeds are built from the z distribution of all reconstructed tracks at the beamline. Starting from

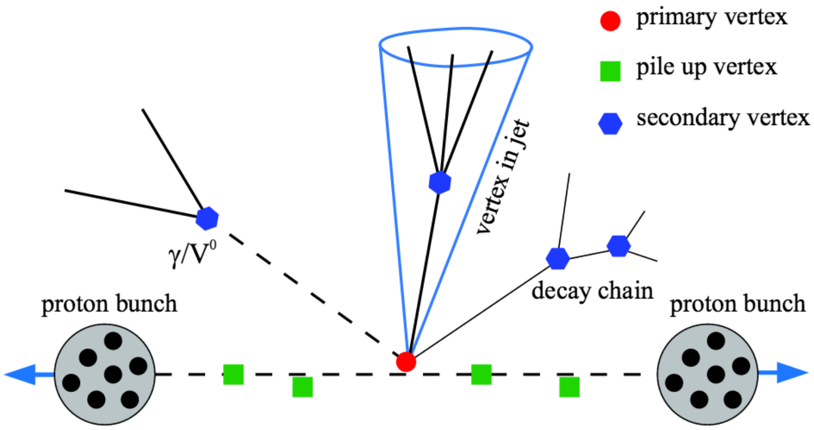


Figure 4.1: The three types of vertices relevant for physics analyses: primary (red), pileup (green) and secondary vertex (blue). The figure is taken from [77].

the maximum of the z distribution and nearby tracks a vertex is fitted using an iterative χ^2 fit. Each considered track obtains a weight which quantifies the compatibility of the track with the fitted vertex. If a track is displaced by more than 7σ it is not associated to the fitted vertex, but used as seed for a new vertex. This procedure is repeated as long as further vertices are found.

For each identified vertex, the sum of squared transverse momenta of all associated tracks is calculated. The vertex with the largest associated sum is marked as primary vertex. Other vertices reconstructed along the beam pipe are called pileup vertices. Another group of vertices, called secondary vertices, originates from the decay short-lived particles as shown in Figure 4.1.
The performance of the vertex reconstruction in collision data is summarised in [78].

4.3.3 Reconstruction of clusters

A particle traversing the calorimeter causes a particle shower with energy depositions in many calorimeter cells. ATLAS provides two algorithms [79], which arrange cells with energy deposits in clusters and calculate the total amount of energy. In parallel calorimeter noise is suppressed.
Since the measured energy differs from the true energy of the particle shower, the reconstructed cluster energies have to be corrected. Based on measurements performed with test beams all reconstructed clusters are calibrated to match the electromagnetic scale. In contrast to electromagnetic showers, the calorimeter features in general a smaller energy response with respect to energy deposits from electrons or photons in hadronic showers. This effect is also known as non-compensation and requires additional corrections. The calibration of the individual clusters depends on the shower type, i.e. electromagnetic or hadronic shower, and takes dead material, non-compensation effects and energy deposits outside the reconstructed cluster into account. Correction factors dedicated to hadronic showers are determined by comparing the reconstructed with the true energy of simulated pions.
The clustering algorithm starts with the search for a cell, which is appropriate to serve as cluster seed. Therefore, in each cell i the significance Γ_i of the energy deposit is calculated. It is defined as the cell energy divided by its noise level. The latter depends on the η region and the specific calorimeter module. A high significance ensures that the energy measurements remain unaffected by the calorimeter noise. A cell is considered as cluster seed if its significance exceeds a certain threshold $\Gamma > \Gamma_{\text{threshold}}$. All neighbouring cells with $\Gamma > \Gamma_{\text{neighbour}}$ are added. A neighbouring cell might be associated with two clusters. If this is the case, the potential clusters are merged. All other cells with $\Gamma < \Gamma_{\text{neighbour}}$ but above a certain threshold Γ_{cell} are matched to the most significant seed cluster in the adjacencies.
After building clusters considering all cells with energy deposits, they are split. In order to split a cluster, well isolated energy maxima have to be identified within the cluster. Such a local maximum is characterised by a cell with a deposited energy of 500 MeV. Further, the energy deposit in this cell has to exceed the energy deposits in all neighbouring cells. At least four adjacent cells must be associated with the same cluster as the cell seed. Starting from these maxima, new clusters are built in a similar way as described above based on already clustered cells but without merging clusters. For each cell, which is matched to two clusters, a weight is assigned considering the distance of the cell to the center of both clusters as well as the energy of the two clusters.
For each reconstructed cluster two copies are stored in two different cluster collections. While in the first collection clusters are calibrated according to the electromagnetic scale, in the second the energy of clusters is corrected to match the local cluster level [47].
Depending on properties of the reconstructed clusters like e.g. the shower depth the probability of the cluster to be caused by an electromagnetic or hadronic shower is determined. Based on the shower probability the cluster is weighted using corresponding calibration constants.
Topological clusters are used for the reconstruction of several physics objects like E_T^{miss} or jets. In the following, the cluster collection which is calibrated to match the local cluster level is used for the reconstruction of jets and hadronically decaying τ leptons, because jets reconstructed from these clusters reveal a better energy resolution than jets calibrated at the electromagnetic scale.

4.3.4 Muon reconstruction

For the reconstruction of muon candidates in Run 1 two algorithms [80], so-called *chains*, were developed. While the first chain is based on a statistical combination of track parameters considering the covariance matrix between them, the second chain refits the muon track starting from hits in the different subdetectors. For the embedding method discussed in Chapter 6 *chain 1* muons are used since the reconstruction of these muons does not require lower-level detector information. Those algorithms are replaced in Run 2 by the *chain 3* algorithm which combines features of the two old chains.
In order to reconstruct the muon trajectory in the muon spectrometer, each layer of the chamber is scanned for track segments. These track segments are later combined to one trajectory. In the ID, a second measurement of the trajectory takes place which is closer to the interaction point. The ID track has to fulfil several quality criteria concerning the number of hits in different ID components [80]. Several types of muons are provided. They differ in how or whether track segments in the muon spectrometer and ID are combined. The type with the highest muon purity is the *combined muon* which is based on the combination of two independent measurements of the muon trajectory in the muon spectrometer and ID. Except for some acceptance losses at $\eta = 0$ and $1.1 < \eta < 1.3$ muons in the range $\eta < 2.5$ are detected with 99% efficiency. Non-prompt muons, i.e. muons which originate from hadronic showers, are rejected using isolation requirements.

4.3.5 Electron reconstruction

Electrons in the central region (below $|\eta| < 2.47$) are reconstructed by matching an energy cluster in the electromagnetic calorimeter to a reconstructed ID track [81]. A *sliding-window* algorithm [79] is used to search for seed-clusters with an overall transverse energy larger than 2.5 GeV. Additionally, the seeds have to pass shower shape requirements. Clusters are reconstructed with high efficiencies between 95%-99.9% depending on the transverse energy and the detector region.
For the track reconstruction, the area in a cone with $\Delta R = 0.3$ around the EM cluster is scanned using pattern recognition algorithms [75, 82] based on charged pion and electron hypotheses. The electron hypothesis takes possible energy losses due to bremsstrahlung into account. This improves the performance of the electron track reconstruction. According to the chosen hypothesis, the track candidate is refitted using the ATLAS Global χ^2 Track Fitter [83]. Based on loose hit requirements and extrapolations of the track into the electromagnetic calorimeter, a geometrically sufficiently close enough track is associated with the cluster. After matching, the track is refitted using the optimised Gaussian Sum Filter algorithm [84], considering effects due to non-linear bremsstrahlung. Within the electron reconstruction, several tracks can be associated with the cluster. All of them are stored for later analyses, but the closest one to the cluster is marked as primary track from which the charge and kinematics of the electron candidate are measured. Electrons in the transition region between barrel and endcaps between $1.37 < |\eta| < 1.52$ are not considered in this thesis.
Objects like hadronic jets or electrons from photon conversions can be misidentified as prompt electrons, which e.g. originate from Z boson decays. In the different electron identification techniques discriminating variables are used to distinguish between true electron candidates and faked ones [81]. Different working points for the electron identification which depend on the transverse energy and η are introduced. These working points differ in their signal over background efficiency.

4.3.6 Jet reconstruction

Jets are an essential ingredient for many physics analyses, e.g. jet topologies are exploited in the search for the VBF Higgs boson production mechanism (see Section 5.6.3). They result from the hadronisation of partons (Section 2.2.4) and are characterised by collimated, neighbouring particles with high transverse momenta originating from the same vertex. In this thesis, jets are reconstructed based on topological clusters (Section 4.3.3) using the anti-k_t jet clustering algorithm [85, 86]. The anti-k_t algorithm clusters pair-wise objects starting with object i which reveals the largest transverse momentum. The distance d_{ij} between object i and object j as well the distance d_{iB} between object and beam is determined according to

$$d_{ij} = min\left(p_{T,i}^{-2}, p_{T,j}^{-2}\right) \frac{\Delta R_{ij}^2}{r^2} \tag{4.1}$$

$$d_{iB} = p_{T,i}^{-2} \tag{4.2}$$

with the resolution parameter $r = 0.4$. For all objects the distances d_{ij} and d_{iB} are calculated and their sizes are compared. As long as the smallest value is a distance d_{ij} between two objects, object i and j are combined and the list is updated. If instead d_{iB} is smaller than any d_{ij}, the jet is assumed to be complete and removed from the list of input objects. These calculations are subsequently repeated as long as unclustered input objects are found.

The reconstructed jet has to be corrected for effects which are not considered in the cluster calibration. This calibration accounts e.g. for energy deposits, not reconstructed in the calorimeter, or particles stopped in the ID. The calibration begins with the removal of energy depositions from multiple p-p interactions. In order to reduce the dependence on the number of vertices a technique is applied which determines the p_T density per area in the event and uses correction factors based on these jet areas [87]. The jet direction is recalculated assuming that all particles associated with the jet have their origin in the primary vertex. Finally, the jet is calibrated using p_T- and η-dependent correction factors which are derived from dijet simulation and data [47, 88, 89].

For the analysis only jets from the primary interaction are of interest, while jets from pile up vertices must be suppressed. To achieve their suppression, the jet vertex fraction (JVF) [90] is calculated for each reconstructed jet. The JVF of jet i with respect to the primary vertex (PV) j is given by

$$\mathrm{JVF}(\mathrm{jet}_i, \mathrm{PV}_j) = \frac{\sum_m p_T\left(\mathrm{track}_m^{\mathrm{jet}_i}, \mathrm{PV}_j\right)}{\sum_n \sum_l \left(\mathrm{track}_l^{\mathrm{jet}_i}, \mathrm{PV}_n\right)} \tag{4.3}$$

with m running over tracks which are associated with jet i and originating from the primary vertex j. The index n runs over all primary vertices, while l over tracks which are associated with jet i and belong to primary vertex n. Cuts on the JVF used in this thesis are documented in Section 7.2.3.

In particular, the rejection of b jets is of interest in $H \to \tau\tau$ searches. Since b hadrons have a long lifetime, jets containing b hadrons can be tagged. By identifying b jets, background contributions from $t\bar{t}$ and single top (Section 5.4) are efficiently reduced in the $H \to \tau\tau$ analysis. In other searches like $H \to b\bar{b}$, however, they are an essential part of the event signature [53]. Different algorithms are provided to identify b jets [91]. They make use of characteristic signatures of b-decays. In this thesis, the MV1 algorithm [92] is used for b jet identification. It combines several tagging algorithms in a neural network to improve the tagging efficiency. The working point used in this thesis corresponds to an efficiency of 70%.

4.3.7 Reconstruction of hadronically decaying τ leptons

With a mass of (1776.82 ± 0.16) MeV the τ lepton is the heaviest known lepton. Due to its short lifetime of $(290.6 \pm 0.1)\times10^{-15}$ s it can only be detected indirectly via its decay products. For that reason, the τ lepton is classified according to its decay products. Leptonically decaying τ leptons, which are also denoted as τ_ℓ in the following, decay into one of the lighter leptons ℓ and two neutrinos, i.e. into an electron and two neutrinos ($\tau_\ell \to e\nu_e\nu_\tau$) or a muon and two neutrinos ($\tau_\ell \to \mu\nu_\mu\nu_\tau$). Electrons and muons originating from τ_ℓ decays and prompt electrons and muons, e.g. originating from Z boson decays, are difficult to distinguish. The majority of τ leptons (≈65%), however, decays hadronically and can be used for τ identification. The hadronically decaying τ lepton is also referred to as τ_{h} lepton. In these hadronic decays, the τ lepton decays into a τ neutrino, mostly in combination with charged and uncharged pions and only rarely accompanied by kaons K. The main τ decay modes and their branching ratios are summarised in Table 4.1. The hadronic decay products of the τ decay are denoted as τ_{had} object so that the hadronic τ decay can also be written as $\tau_{\mathrm{h}} \to \tau_{\mathrm{had}}\nu_\tau$. These hadronic decay products are used for the reconstruction and identification of τ leptons.

Table 4.1: The most dominant τ decay modes including the corresponding branching ratios according to [31].

	Decay mode	Branching ratio [%]
Leptonically decays	$\tau^- \to e^-\bar{\nu}_e\nu_\tau$	17.83±0.04
	$\tau^- \to \mu^-\bar{\nu}_\mu\nu_\tau$	17.41±0.04
One prong τ_{h} decays	$\tau^- \to \pi^-\pi^0\nu_\tau$	25.52±0.09
	$\tau^- \to \pi^-\nu_\tau$	10.83±0.06
	$\tau^- \to \pi^-\pi^0\pi^0\nu_\tau$	9.30±0.11
	$\tau^- \to \pi^-\pi^0\pi^0\pi^0\nu_\tau$	1.05±0.07
	$\tau^- \to K^-\nu_\tau$	0.700±0.010
	$\tau^- \to K^-\pi^0\nu_\tau$	0.429±0.015
Three prong τ_{h} decays	$\tau^- \to \pi^-\pi^-\pi^+\nu_\tau$	8.99±0.06
	$\tau^- \to \pi^-\pi^-\pi^+\pi^0\nu_\tau$	2.70±0.08

The reconstruction of τ_{had} objects is based on the reconstruction of topological clusters [47]. These calibrated clusters serve as input to the anti-k_t jet finding algorithm with a radius parameter $r = 0.4$. To associate a track to a calorimeter cluster it has to fulfil several track quality criteria [93] concerning the momentum and distance to the cluster barycentre. After associating tracks and clusters, the momentum of the τ_{had} candidate is derived from energy depositions in the topological clusters and the charge is given by the sum of charges of the linked tracks. The signature of τ_{had} candidates within the detector is very similar compared to jets as shown in Figure 4.2. Although τ_{had} candidates usually have a more narrow shower shape and less associated tracks, these characteristics are not sufficient to separate both physics objects within the reconstruction. Instead, different identification algorithms are applied to the reconstructed candidates. These identification algorithms are based on variables which exploit characteristic properties of hadronically decaying τ leptons. They are derived from information provided by the calorimeter, ID or by combining information of both detector subsystems. In the following, the variables used in the default identification algorithm used for 2012 data are shortly introduced according to the description in [93]:

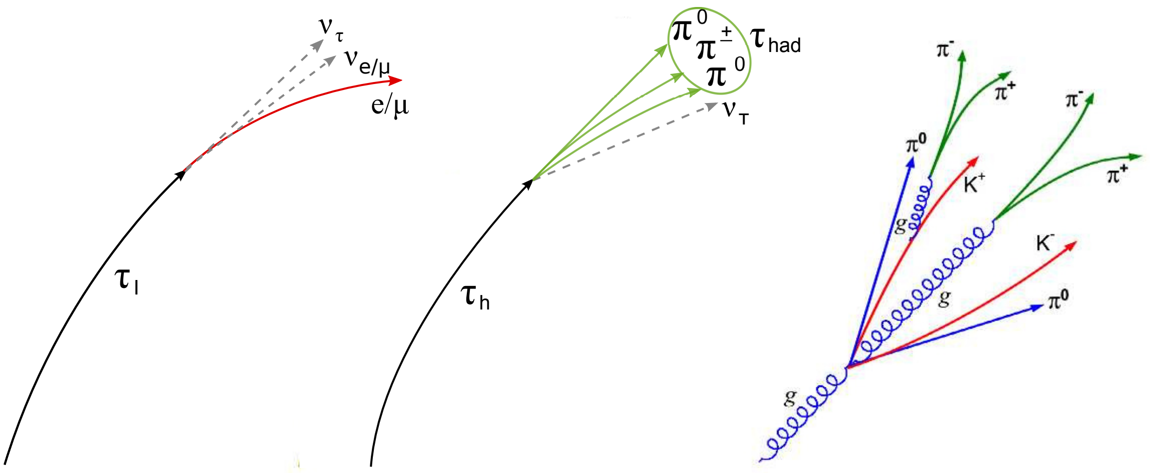


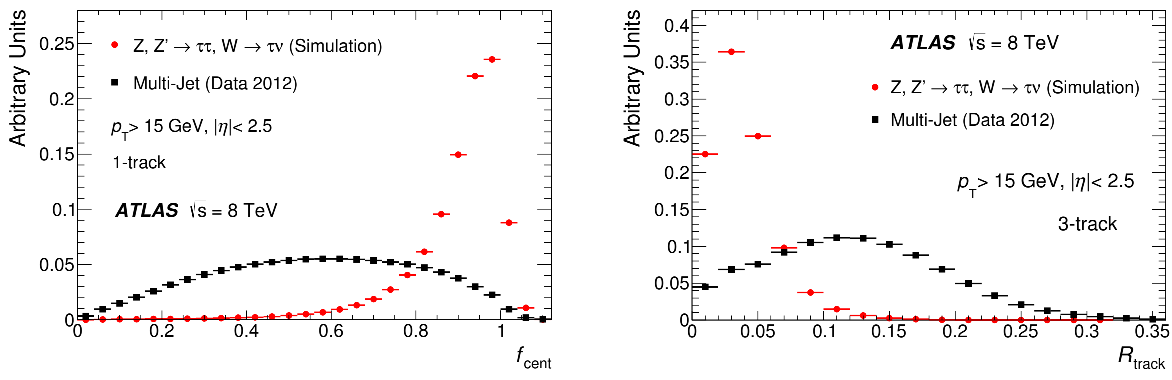
Figure 4.2: Left: Leptonically decaying τ lepton. Middle: One prong τ_{h} decay. Right: Gluon initiated hadronic shower. The right figure is taken from [94].

- The *central energy fraction (f_{cent})* quantifies the amount of energy deposited in a cone with $\Delta R < 0.1$ around the τ_{had} candidate compared to the deposited energy in a larger environment ($\Delta R < 0.2$). In contrast to jets, τ_{had} objects have a very narrow shower shape, for which the deposited energy is expected to be concentrated within a small area, i.e. they result in a high f_{cent}-value as illustrated in Figure 4.3a.
- Another important variable is the *leading track momentum fraction (f_{track})*. It is defined by the ratio of the transverse momentum of the leading associated τ_{had} track and the sum of all transverse energy.
- The *track radius (R_{track})* calculates the p_T-weighted distance between the directions of the τ_{had} track and the τ_{had} object. Figure 4.3b shows that the track radius is significantly smaller for true τ_{had} leptons than for misidentified τ_{had} candidates.
- All tracks which are matched to the τ_{had} candidate within $0.2 < \Delta R < 0.4$ are counted and referred to as *number of tracks in the isolation region (N^{iso}_{track})*.
- The *maximum ΔR (ΔR_{Max})* is defined as the maximum distance between a τ_{had} track and the direction of the τ_{had} candidate.
- The *leading track impact parameter significance ($S_{leadtrack}$)* is only calculated for one prong τ_{had} candidates. Therefore, the impact parameter of the leading τ track is determined and divided by its uncertainty. For three prong candidates the *transverse flight path significance (S^{flight}_{T})* is calculated which corresponds to the τ decay length divided by its uncertainty.
- The *track mass (m_{track})* is built from all tracks in the τ_{had}-relevant region under the assumption that each track belongs to a pion.

In addition, a π^0 cluster identification algorithm [93] was developed, which reconstructs the substructure of τ_{h} decays. Based on information from the calorimeter and strip layer neutral pions associated with the τ_{h} decay are reconstructed. In this thesis, only the identification with boosted decision trees [1], called

[1] The concept of boosted decision trees is discussed in Section 5.2.1.

BDT τ ID, is used in order to identify τ_{had} candidates and to reject jets and electrons. The BDT τ ID is evaluated for τ_{had} candidates with one associated track (one prong) or three matched tracks (three



(a) (b)

Figure 4.3: Distribution of the central energy fraction f_{cent} (left) and track radius R_{track} (right) for one prong and three prong τ_{had} candidates, respectively. The signal distribution (red) is based on simulation and contains reconstructed τ_{had} leptons which are matched to truth τ leptons on generator-level. The multi-jet distribution (black) is obtained from data. The figures are taken from [93].

prong). For each case a separate BDT is trained based on a short list of τ_{had} characteristic cell and track variables (see Table 4.2). Three different p_T-dependent working points *tight*, *medium* and *loose* [93] are defined with different efficiencies. The efficiencies for these working points in the case of one prong τ_{had} candidates with p_T >15 GeV are shown in Figure 4.4a. They are stable with respect to changing pileup conditions. In this thesis, only τ_{had} candidates satisfying the *medium* criteria are considered. This working point corresponds to a τ_{had} selection efficiency of about 55-60%. The probability to misidentify a jet as τ_{had} candidate is 1-2%.
Although electrons and muons[2] can also be misidentified as τ_{had} candidates, the misidentification prob-

Table 4.2: Variables used as input for the training of the BDT based τ ID for one prong and three prong τ_{had} candidates. The notation is adopted from [93].

Variable	One prong	Three prong
f_{track}	✓	✓
f_{cent}	✓	✓
R_{track}	✓	✓
$N^{\text{iso}}_{\text{track}}$	✓	
$S_{\text{leadtrack}}$	✓	
$S^{\text{flight}}_{\text{T}}$		✓
ΔR_{Max}		✓
m_{track}		✓

ability is much smaller compared to jets. While electrons have a very similar signature to one prong τ_{h} decays with one track and narrow shower, muons can be misidentified in cases where they interact with the material of the calorimeter and lose a large part of their energy. Since these objects have a high

[2] In the following, electrons and muons are also called light leptons or ℓ in order to distinguish between them and the τ lepton.

reconstruction efficiency, the majority of them can be removed by an overlap removal between different objects. This overlap removal is carried out on analysis level with a looser object selection than usually required for the final analysis objects. The overlap removal for the presented analysis is discussed in Section 7.2.3.
Besides the overlap removal, other means are provided to reject fakes caused by lighter leptons. For the rejection of electrons against τ_{had} candidates a BDT is trained based on information on hits, deposited energy and shower profile. Again, three different working points corresponding to different signal efficiencies are provided. The *medium* working point with an efficiency of about 85% is used for the presented analysis. The BDT is exclusively trained on one prong candidates, but also reveals a good performance for three prong candidates.
In the cases of misidentified τ_{had} candidates due to muons a muon veto built from requirements on the relation between the transverse momentum of the track and deposited energy is applied. The muon veto

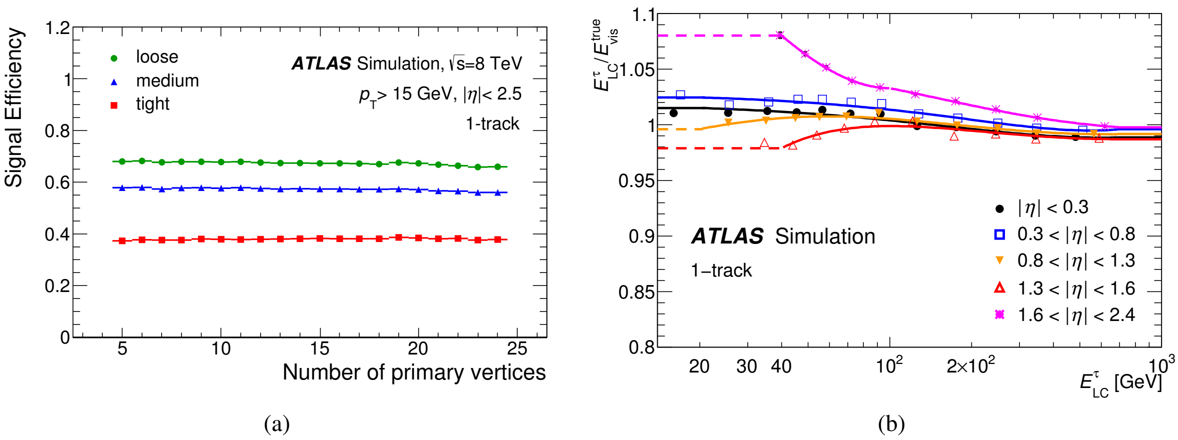


Figure 4.4: Left: The τ ID efficiency for one prong τ_{had} candidates for the three working points *tight* (red), *medium* (blue) and *loose* (green) as a function of the number of reconstructed interaction vertices. Right: Energy response curves for one prong τ_{had} candidates as a function of the reconstructed τ_{had} energy. The curves correspond to different pseudorapidity ranges. The figures are taken from [93].

removes roughly 40% of all misidentified muons. Due to the low misidentification probability, however, this veto is not used in the following.
As for jets, the clusters which are associated to the tracks of the τ_{had} candidate are calibrated. However, a pure cluster calibration at local cluster scale is not sufficient, since it is not optimised with respect to the cone size of $\Delta R = 0.2$ which is used to measure the momentum of the τ_{had} candidate. Further, it is not tuned for the specific mixture of hadrons found in τ_{h} decays. For that reason, additional corrections are applied to the reconstructed τ_{had} candidates. In the first calibration step, the τ energy scale is corrected in order to match the true energy scale at a few percent level. The corrections are based on energy dependent response curves, which are derived from simulated $Z \to \tau\tau$, $Z' \to \tau\tau$ and $W \to \tau\nu$ events. In these simulated events the reconstructed τ_{had} candidate, which has to pass several quality requirements, is truth-matched to τ leptons on generator level. The ratio of the reconstructed τ_{had} energy and the visible true energy is calculated and evaluated in intervals of the true visible energy and pseudorapidity. In each of these intervals a Gaussian fit is used to determine the mean value of the ratio distribution. Based on those mean values the response curves for different pseudorapidity ranges are fitted as a function of the average reconstructed τ_{had} energy. The response curves for one prong τ_{had} leptons is shown in Figure 4.4b. After the application of the response curves additional corrections are necessary to correct for

two further effects: First, the measured pseudorapidity is biased since the reconstructed cluster energies are underestimated in poorly instrumented regions of the detector. Second, pileup correction reduce the dependence of the reconstructed energy on the number of vertices. In the final calibration step, two complementary data driven methods are used to probe the calibration with real data. These techniques add some small corrections to the τ energy scale. They are also used to asses systematic uncertainties. Further details on these methods are documented in [93].

4.3.8 Reconstruction of missing transverse energy

The reconstruction of the missing transverse energy E_T^{miss} relies on all fully reconstructed objects discussed above, i.e. electrons, muons, τ_{had} leptons, photons and jets [95]. All cells associated with one of these objects are arranged together and the negative sum of the momenta of the matched objects is calculated. Afterwards, all terms are added:

$$E_T^{\text{miss}} = E_T^{\text{miss},e} + E_T^{\text{miss},\mu} + E_T^{\text{miss},\tau_{\text{had}}} + E_T^{\text{miss},\gamma} + E_T^{\text{miss,jet}} + E_T^{\text{miss,soft}} \tag{4.4}$$

The last term $E_T^{\text{miss,soft}}$ refers to cells with energy deposits which cannot be associated with one of the other physics objects and is scaled according to a soft-term vertex fraction [96]. This fraction is defined as the scalar sum of transverse momenta of unmatched tracks originating from the primary vertex divided by the sum of transverse momenta of all unmatched tracks in the event. Considering contributions from unmatched cells improves the reconstruction of E_T^{miss} in high pileup environments [97].
Each term is calibrated according to the corresponding physic objects, e.g. $E_T^{\text{miss},\tau_{\text{had}}}$ is calibrated to the τ energy scale [93] and electrons to the electron energy scale.
The missing transverse energy plays an important role for the analysis of the $H \to \tau\tau$ channel, since a large amount of the τ lepton energy is transferred to τ decay neutrinos which cannot be detected. The use of E_T^{miss} and its direction as a measure for the neutrinos within the analysis improves the rejection of background processes. Further, E_T^{miss} is an important ingredient in the reconstruction algorithm of the $\tau\tau$ invariant mass, discussed in Section 5.5.3.

CHAPTER 5

Properties of $H \to \tau_\ell \tau_h$ decays and background suppression

For the study of the $H \to \tau\tau$ decay channel three processes are distinguished: Either both τ leptons decay hadronically ($H \to \tau_{\mathrm{h}}\tau_{\mathrm{h}} \to \tau_{\mathrm{had}}\nu_\tau\tau_{\mathrm{had}}\nu_\tau$) or both leptonically ($H \to \tau_\ell\tau_\ell \to \ell\nu_\ell\nu_\tau\ell\nu_\ell\nu_\tau$) or one τ lepton decays leptonically and the other one hadronically ($H \to \tau_\ell\tau_{\mathrm{h}} \to \ell\nu_\ell\nu_\tau\tau_{\mathrm{had}}\nu_\tau$). Each of these decay channels is analysed separately by searching for channel-specific signatures as indicated in Figure 5.1.

Good knowledge of the characteristic event topologies of $H \to \tau_\ell\tau_h$ decays is of high importance for the

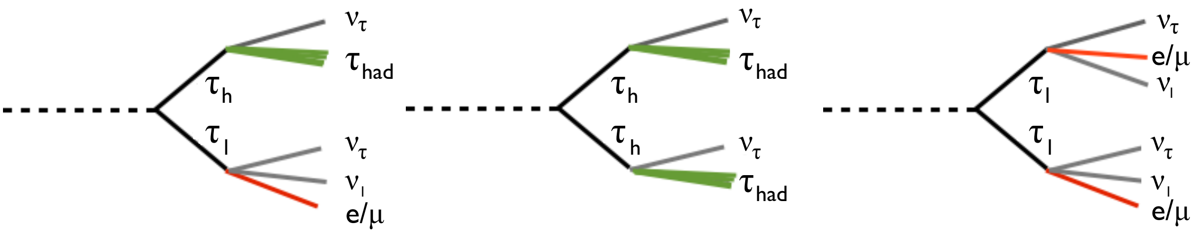


Figure 5.1: Sketch of the three analysis channels $H \to \tau_\ell\tau_h$ (left), $H \to \tau_h\tau_h$ (middle) and $H \to \tau_\ell\tau_\ell$ (right). The visible decay products of the hadronically (leptonically) decaying τ lepton are illustrated in green (red). Neutrinos emitted in the τ lepton decays are shown in grey. They remain undetected.

development of the analysis strategy. The $H \to \tau_\ell\tau_h$ signature is characterised by the τ decay mode as well as the Higgs production mechanism. It is identified by one τ_{had} candidate in the final state in combination with an electron or muon of opposite electric charge. Further, a significant amount of missing transverse energy is measured due to the undetected neutrinos originating from the τ decay. Additional jet activity points to different Higgs production mechanisms.

Several background processes reveal a very similar signature and can be misidentified as signal process. For example, the $Z \to \tau\tau$ process, which is the most dominant background process, exposes the same objects in the final state. The misidentification rate in combination with the large cross-section of these background processes is challenging for $H \to \tau_\ell\tau_h$ searches. Concepts have to be developed to efficiently suppress large parts of the background and hence increase the signal sensitivity.

In this chapter the analysis strategy is motivated. Based on the characteristics of $H \to \tau_\ell\tau_h$ signatures it is explained which physical processes can mimic signal processes and how differences in the event topologies can be used to suppress background processes. This chapter is organised as follows: After a short description of the $H \to \tau_\ell \tau_h$ analysis strategy in Section 5.1 and an introduction to multivariate classifiers in Section 5.2 the main features of the $H \to \tau\tau$ signature are summarised in Section 5.3. The most important background processes are discussed in Section 5.4. The reconstruction of the invariant $\tau\tau$ mass is explained in Section 5.5. In the last section, different variables which are suitable to discriminate between signal and background processes are described. The knowledge about the properties of the various processes will be used in the $H \to \tau_\ell \tau_h$ analysis in Chapter 7.

5.1 The analysis strategy

Although the signature of the three $H \to \tau\tau$ decay channels is distinct, the analyses of these channels are based on the same analysis strategy. In this thesis, the $H \to \tau_\ell \tau_h$ channel is studied (Chapter 7) with the intention to combine the results with the analyses of the two other channels (Chapter 8). The combined sensitivity of all channels is large enough to confirm the decay of the Higgs boson into two τ leptons.
Background processes can imitate the signal signature due to jets or light leptons misidentified as jets. Compared to these processes the cross-section for the Higgs production is several orders of magnitude smaller (see Figure 5.2). For these reasons, searches for $H \to \tau_\ell \tau_h$ signatures are strongly contaminated by background processes. Hence, a low signal-to-background ratio is expected. The analysis strategy

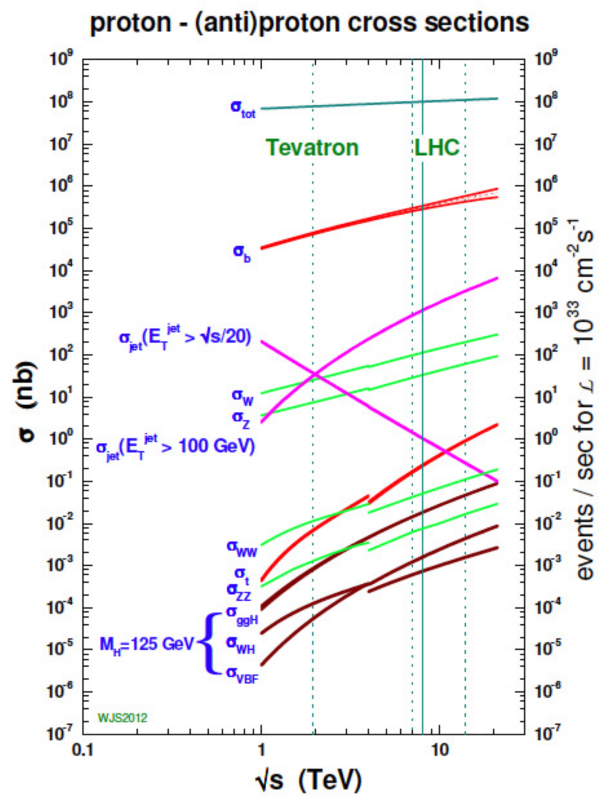
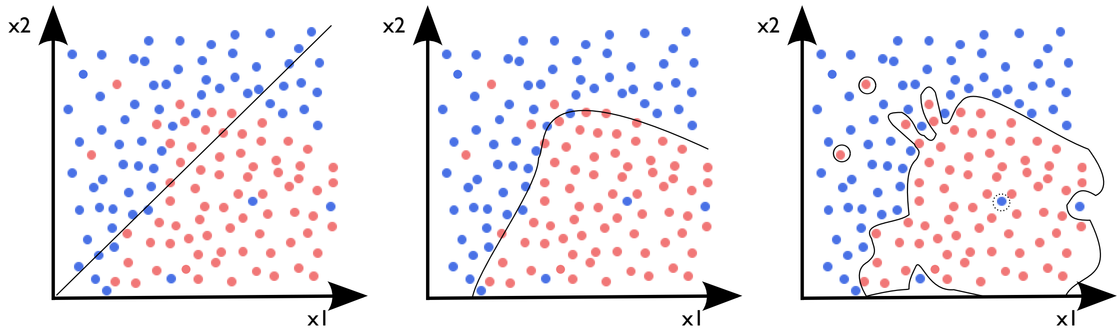


Figure 5.2: Calculated SM cross-section for selected processes at hadron-hadron colliders as a function of center-of-mass energy. The figure is taken from [98].

has to improve the signal-to-background ratio by an appropriate choice of signal-enriched phase spaces. Another challenge arises from the worse resolution of the invariant $\tau\tau$ mass due to the undetected neutrinos from τ decays. For the reconstruction of the $\tau\tau$ invariant mass an algorithm has to be chosen, which calculates the invariant mass as precise as possible. As discussed later, the resolution depends on the kinematics in the Higgs production mechanism. Hence, the topologies of the Higgs production mechanisms are exploited to improve the mass resolution.
An increased signal-to-background ratio is obtained in several steps. In the *preselection* events are selected which reveal exactly the objects expected in $H \to \tau_\ell \tau_h$ decays. Requirements on the quality of the physics objects (*object preselection*) already reduce some background contributions in particular due to misidentified τ_{had} candidates. Events passing the preselection are arranged in two *categories* provided that they pass the requirements for at least one of these categories. The definition of both categories, the *boosted* and *VBF* category, is driven by the kinematic properties of the two dominant Higgs production mechanisms (see Section 7.2.4). While in the VBF category events with at least two jets with high transverse momentum are selected, the boosted category contains events with a boosted Higgs candidate and which fail the VBF selection. The purpose of the categorisation is to enhance the number of signal events with respect to a certain Higgs production mechanism (VBF or ggF) rather than to reduce background contaminations. The background is mainly suppressed by the use of *multivariate classifiers* (Section 5.2), which split the phase space into signal-like and background-like phase space regions and declares the events either as signal or background. The construction of different signal-enriched regions and the resulting high sensitivity are the most important advantages of the application of multivariate classifiers. In each category, one classifier is trained based on discriminating variables which render a good separation power between the signal and background processes. Since the VBF and boosted categories are sensitive to the VBF and ggF production mechanism, respectively, a different set of variables is used for discrimination. The variables are chosen such that they have separation power with respect to the dominant Higgs production mechanism in the particular category.
All analysis steps are performed on data and on the *background-only (background-plus-signal) model*. The background-only (background-plus-signal) model is obtained from simulations and describes the expected number of the different background (and signal) processes in the signal regions. Inaccuracies in the modelling of the processes are origin for systematic uncertainties. In order to reduce the size of systematic uncertainties the dominant background sources, i.e. $Z \to \tau\tau$ and backgrounds due to misidentified τ_{had} signatures, are not derived from simulations but from data-driven methods. In order to extract the signal the expected event yields from the background-only (background-plus-signal) model are compared to the observed events in data. Therefore, a profile likelihood fit, which incorporates statistical and systematic uncertainties in the form of nuisance parameters, is performed based on the output distributions of the multivariate classifiers. The compatibility test of the observed data with the hypothesis of a $H \to \tau_\ell \tau_h$ signal determines an estimate for the signal strength and the observed significance.

5.2 Multivariate classifiers

In order to gain sensitivity in the signal categories with respect to a possible $H \to \tau_\ell \tau_h$ signal the discriminating variables introduced in Section 5.6 has to be employed. These variables can be used in different ways: In the cut-based analysis (see Section 8.4), signal-enriched sub-categories are defined by consecutively cutting on discriminating variables. Events which (do not) pass these cuts are classified as signal (background)-like. Instead of the cut-based classifier, more sophisticated methods, so-called multivariate methods, can be used to classify data events as signal-like or background-like. As the cut-based classifier, multivariate methods use the information gained from a set of discriminating variables in order to determine whether the event is signal or background. In general they have, however, a higher sensitivity compared to the cut-based classifier, since they form different signal-enriched regions, which are hypercubes of high dimensionality. Examples for such multivariate classifiers are likelihood estim-



(a) (b) (c)

Figure 5.3: Scheme showing the functionality of a classifier: The red (blue) points illustrate signal (background) events in a two dimensional phase space spanned by two parameters x_1 and x_2. The black line stands for the classifier separating the phase space into two regions. The left plot shows an example of a simple classifier. A more complex classifier is illustrated in the middle. The right plot is an example for a classifier suffering from overtraining. It is sensitive to fluctuations in the signal sample.

ation, neural networks or boosted decision trees.
Before the classifier is used in the analysis to determine the signal-likeness of events, it is trained on given (Monte Carlo) background and signal samples. For the training a small set of discriminating variables is used which reveal separation power between background and signal samples. These samples are known to be signal or background. Afterwards, the classifier is tested on other samples, the so-called testing samples, which are independent from the training samples.
Some classification examples are sketched in Figure 5.3: The red (blue) points illustrate signal (background) events in a two dimensional phase space. In Figure 5.3a all points below the black curve are classified as signal, while the rest is assumed to be background. Some events are misclassified as signal or background. In Figure 5.3b the classifier is optimised for the given two datasets and yields a smaller misclassification rate. In contrast, Figure 5.3c shows a worse classification. The classifier is sensitive to fluctuations in the datasets, which are indicated by the circles around the single points. The problem that the classifier exploits statistical fluctuations is referred to as overtraining.
In Chapter 7- 9 boosted decision trees (BDTs) are used as classifiers.

5.2.1 The BDT classifier

A decision tree is characterised by consecutive yes/no decisions (also called leaf nodes) based on discriminating variables[1] which split the phase space into several regions. The structure of a decision tree is illustrated in Figure 5.4. Starting from the root node several decisions are made. Since in the training it is known whether an event is signal or background, the true ratio of signal events to background events can be determined in the final regions. Depending on this ratio the regions are declared to be signal-like (red) or background-like (blue). Events which end up in these regions are regarded as signal-like or background-like corresponding to the name of the region. Such a simple decision tree corresponds to a

[1] Each decision is only based on one variable.

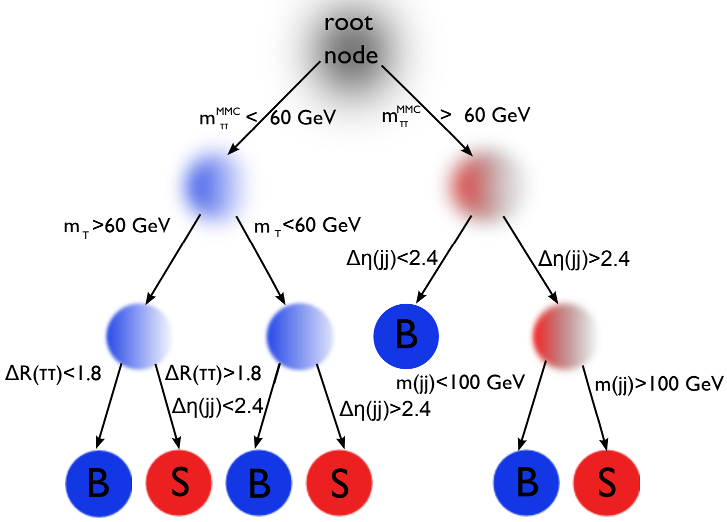


Figure 5.4: Schematic representation of a decision tree. Consecutive decisions based on a set of input variables lead to the separation of signal (S) and background (B) events. Each decision is based on events included in the corresponding node and on one variable.

cut-based analysis: The most signal-like leaf node, for which all decisions indicated signal, corresponds to the signal region of the cut-based analysis. Only the sensitivity of this single leaf node is exploited for the cut-based analysis. But it might be that the cut on one variable decided the event to be background-like, whereas the other decisions indicated signal. These events are not considered in the cut-based analysis. In contrast, analyses based on decision trees gain additional sensitivity, since they take further leaf nodes with high signal contributions into account. The consideration of additional signal-like leaf nodes is a huge advantage of decision trees[2].

The decision tree is optimised with respect to the used variables and their best cut-values with the aid of the so-called Gini-Index G. It is defined as $G = p \cdot (1 - p)$, where p stands for the signal purity in a given leaf node, and quantifies the separation between signal and background events. The Gini-Index is zero (has its maximal value), if the purity is 100% (50%). For the optimisation the Gini-Index of the parent node is compared to the weighted sum of the indices of both daughter nodes. The weighting is given by the relative fraction of events ending up in the nodes. The variables and corresponding cut values are chosen such that they have the highest increase between mother and daughter nodes. By repeating the optimisation procedure with the daughter nodes, the tree grows further until the maximal depth of the tree is reached or the number of events in the leaf node is getting to small. Both, the depth of the tree and the minimal fraction of events in the leaf nodes, are set by the user. For this thesis, a maximal depth of 5(4) is chosen for the VBF (boosted) category. In principle, it is possible to construct a tree based on so many decisions, such that each leaf node contains exactly one input event. However, such a classifier would be overtrained. In order to avoid overtraining, the depth of the tree and leaf size should be restricted, so that enough events end up in the leaf nodes. Further, all events used for the training are not used for the testing of the classifier's performance or within the analysis.

[2] The latest cut-based analysis also increased its sensitivity by introducing sub-categories with large signal contamination similar to the idea of the decision tree (see Section 8.4.).

Improvement of BDT performance via boosting

The boosted decision tree is characterised by the training of several trees based on the same training sample. The sum of all trained BDTs is referred to as *forest*. A forest enables to classify each event not only based on the decisions of a single tree but takes the classification of all trees into account. Events misclassified in the first tree obtain higher weights compared to the proper classified events. In the next tree to be trained, these events are of higher importance, so that the decisions and the resulting cuts might differ from the findings of the first one. This learning process, also called *boosting*, is iterated 600 (400) for the boosted (VBF) category resulting in a better separation of signal and background events. After boosting, for each event the BDT score value which is a number between -1 and 1 is defined. An increasing (decreasing) score value corresponds to a higher signal (background)-likeness.
Different boosting algorithms are provided by TMVA [99]. While the $H \to \tau_\ell \tau_h$ and $H \to \tau_\ell \tau_\ell$ analyses use the gradient boost [100], the $H \to \tau_h \tau_h$ analysis uses AdaBoost [101]. In the following, both algorithms are briefly introduced according to the description in [99].

- **AdaBoost:** The *Adaptive Boost* algorithm focuses on events which are misclassified in the previous tree and assigns a higher weight to them. For each tree the misclassification rate N_{miss} is determined, from which a boosting weight α is derived according to

$$\alpha = \frac{1 - N_{\mathrm{miss}}}{N_{\mathrm{miss}}} . \tag{5.1}$$

  In order to ensure that the sum over all weights remains unaffected by the reweighting, all obtained weights α are renormalised. Afterwards, they are applied to the samples used for the next tree iteration so that the final boosted classification $c_{\mathrm{boosted}}(x)$, which depends on the tuple of BDT input variables x, is given by

$$c_{\mathrm{boosted}}(x) \quad = \quad \frac{1}{N_{\mathrm{collection}}} \sum_{i}^{N_{\mathrm{collection}}} \ln(\alpha_i) \cdot f_i(x) , \tag{5.2}$$

  with $f_i(x)$ as classifier, which maps the value $+1$ to signal and -1 to background, and $N_{\mathrm{collection}}$ being the number of classifiers in the collection.

- **Gradient boost:** As the *Adaptive Boost* algorithm, the *Gradient Boost* minimises the deviation between the estimated and true class, i.e. background class or signal class, to which the events are associated. Therefore, a loss function is used. It slightly differs from the one used in the *Adaptive Boost* method and has the advantage of being more robust with respect to fluctuations. It is defined as

$$L(F, y) = \ln\left(1 + \exp^{2F(x)y}\right) , \tag{5.3}$$

  where $F(x, y)$ denotes the model response and y the true class, i.e whether the event is signal or background. In order to minimise the loss function its gradient is calculated. Afterwards, a classifier $f_i(x)$ is built in a way that its corrected leaf values correspond to the mean values of the gradient. By iterating this steepest-descent approach one obtains a forest which minimises

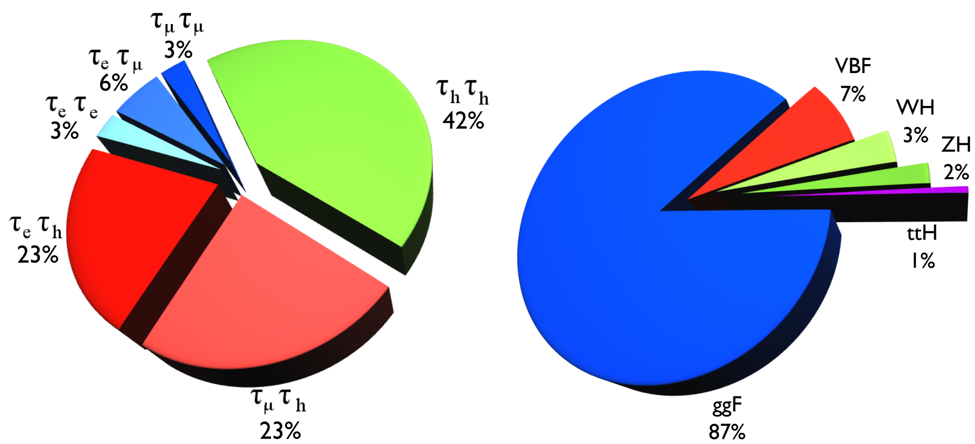
$F(x, y)$. The final classifier is given by the sum over all iterated classifiers $f_i(x)$:

$$f(x) = \sum_{i=0}^{N} w_i f_i(x) \ . \tag{5.4}$$

The classifier's output is finally transformed to a BDT score value between -1 and 1.
Calculating the gradient of the loss function is difficult, because the gradient associated to the trained classifier is only defined at certain points. These points x_p are given by the tuples x_i. Each tuple x_i corresponds to one training event and its variable values, used in the training. Only for these points x_p the true class is given. The gradient at all other points in the variable space are unknown. In order to apply the trained classifier also on events, which are not associated to the points x_p the gradient has to be estimated between the points x_p. Statistical fluctuations in the input samples might influence the estimate. A stabilisation is achieved by dividing the weights w_i by a factor a and hence reducing the learning rate artificially. This method is known as *shrinkage* method. It is particularly used in applications with complex correlations and increases the number of necessary iterations. By increasing the number of iterations the discriminating power of the classifier can be improved significantly.

5.3 Signature of $H \to \tau\tau$ processes

The signature of $H \to \tau\tau$ decays is characterised by the decay mode of the τ lepton and the Higgs production mechanism. The analysis strategy is optimised with respect to these aspects.
The branching ratios for the different $\tau\tau$ decays are depicted in Figure 5.5a. The $H \to \tau_\ell\tau_\mathrm{h}(\to \ell\nu_\ell\nu_\tau\tau_\mathrm{had}\nu_\tau)$ channel has the largest branching fraction with 45.6%, closely followed by the $H \to \tau_\mathrm{h}\tau_\mathrm{h}(\to \tau_\mathrm{had}\nu_\tau\tau_\mathrm{had}\nu_\tau)$ channel (42%). In contrast, the contribution of $H \to \tau_\ell\tau_\ell(\to \ell\nu_\ell\nu_\tau\ell\nu_\ell\nu_\tau)$ decays is relatively small (12.4%). The $H \to \tau_\ell\tau_h$ signature is characterised by one single light lepton (i.e. electron or muon) and one τ_had candidate. Due to charge conservation, both objects must have opposite electric charge. The τ decay products are in general emitted collinear to the τ lepton and are high-energy



(a) (b)

Figure 5.5: Left: Branching ratios of the different $\tau\tau$ decay modes. Decay modes summarised in the $H \to \tau_\ell\tau_h$ ($H \to \tau_\ell\tau_\ell$) analysis are shown in shades of red (blue). The $H \to \tau_h\tau_h$ decay is depicted in green. Right: Branching ratios for the different Higgs production mechanisms in the case of a Higgs boson with mass m_H=125 GeV.

because of the large mass difference between the τ lepton and the Higgs boson. This implies not only the visible decay products, the light lepton and τ_{had} candidate, but also the three undetected neutrinos. Due to those neutrinos $H \to \tau_\ell \tau_h$ events are expected to posses a significant amount of reconstructed missing transverse energy.
The reconstruction of the invariant $\tau\tau$ mass provides additional information about the Higgs boson. Its reconstruction, however, is difficult due to the undetected neutrinos, which carry a non-negligible amount of energy from the decaying τ lepton. Mass reconstruction algorithms were developed in order to cope with this challenge (Section 5.5).
The complete $H \to \tau\tau$ signature also depends on the Higgs production mechanism. All production modes exhibit different kinematic properties and result in additional jet activity in the event which effect the kinematics of the Higgs boson and its decay products. The ggF is the dominant production mode (see Figure 5.5b). In leading-order, no additional partons are contained in the final state. In this case, the transverse momentum of the Higgs boson is small and the two τ leptons are emitted in opposite directions as illustrated in the left sketch in Figure 5.6. The back-to-back emission of the τ leptons forces the neutrinos to fly in opposite directions. As a consequence, their contributions to the missing transverse energy almost cancel, which results in a small measured missing transverse energy. However, due to higher order QCD corrections additional partons in the final state are possible. The additional parton hadronises and causes a recoil of the Higgs boson against the jet. As a consequence, the Higgs boson is

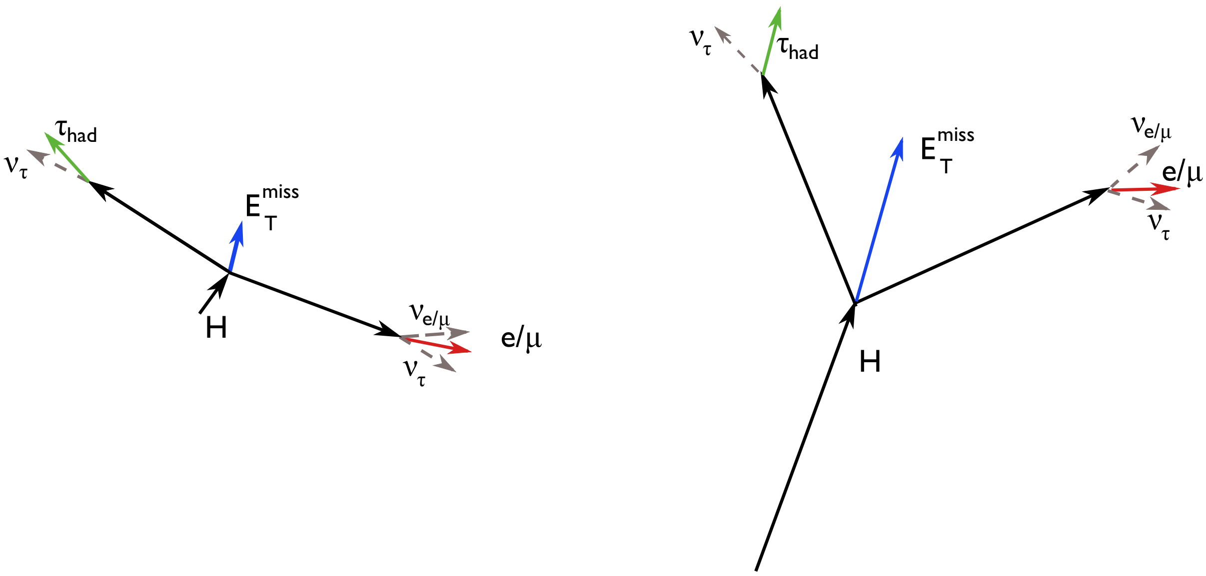


Figure 5.6: Sketch of the event kinematics in the case of a Higgs boson with low transverse momentum (left) and high transverse momentum (right).

boosted in the center-of-mass system and the transverse momenta of the τ decay products increase. In this case, the sum of the transverse neutrino momenta does not cancel and a larger missing transverse momentum is measured (see right sketch in Figure 5.6). Due to the increased momenta of the visible τ decay products and E_T^{miss} they are resolved which also improves the invariant $\tau\tau$ mass resolution.
The VBF mechanism illustrated in Figure 2.8c is characterised by two additional jets in the final state. The two additional jets are very suitable to tag this signal process, as they are well-separated in pseudorapidity. Its topology is clearly distinguishable from other processes, so that the $H \to \tau\tau$ analysis is most sensitive to this mechanism.
The Higgs production in association with a vector boson is also considered a signal process in this ana-

lysis. However, the signal categories are not optimised with respect to this particular Higgs production mechanism. Instead, a separate analysis is dedicated to search for the SM Higgs boson produced in this production mode and decaying into a τ pair [102]. Furthermore, the analysis is not sensitive to the Higgs production in association with $t\bar{t}$ due to its low cross-section[3].
The topologies of the dominant Higgs production processes, i.e. boosted Higgs boson and VBF jets, are exploited to define signal categories and to differentiate between signal and background processes.

5.4 Background processes in $H \to \tau_\ell \tau_\mathrm{h}$ searches

Processes which contribute to the background model in the $H \to \tau_\ell \tau_h$ search can be organised in three categories: events with a real light lepton and a real τ_h decay as well as events with a jet faking the τ_had candidate (jet→ τ_had fake) or a light lepton faking the τ_had candidate (electron/muon→ τ_had fake).
In general, background events with an identical signature are difficult to separate from signal processes. The reconstructed $\tau\tau$ mass is one of the best discriminators for these background processes. Event yields for background processes with fake τ_had signatures are difficult to obtain from simulations, since the actual τ_had misidentification rate in the experiment is not known beforehand. For that reason, a data-driven method described in Section 7.3.5 was developed to estimate the fake τ_had background in data.
In the following, the different background processes and their properties are briefly introduced.

- **$Z \to \tau\tau$+jets:** The Z boson decays into a fermion and its corresponding anti-fermion. Relevant for the analysis are the decays into two light leptons ($Z \to ee$ and $Z \to \mu\mu$) and into two τ leptons ($Z \to \tau\tau$). The $Z \to \tau\tau$ decay represents the largest irreducible background in $H \to \tau\tau$ searches, since it has the same final state and similar kinematics as the signal process. The exemplary Feynman graphs in Figure 5.7 show the production and decay of the Z boson. The Figures 5.7a- 5.7c show the production of the Z boson via the QCD interaction of two partons, while Figure 5.7d displays the electroweak production via the interaction of two vector bosons. The latter mechanism is accompanied by two jets, which makes the signature of electroweak Z boson production very similar to the VBF Higgs production signature.

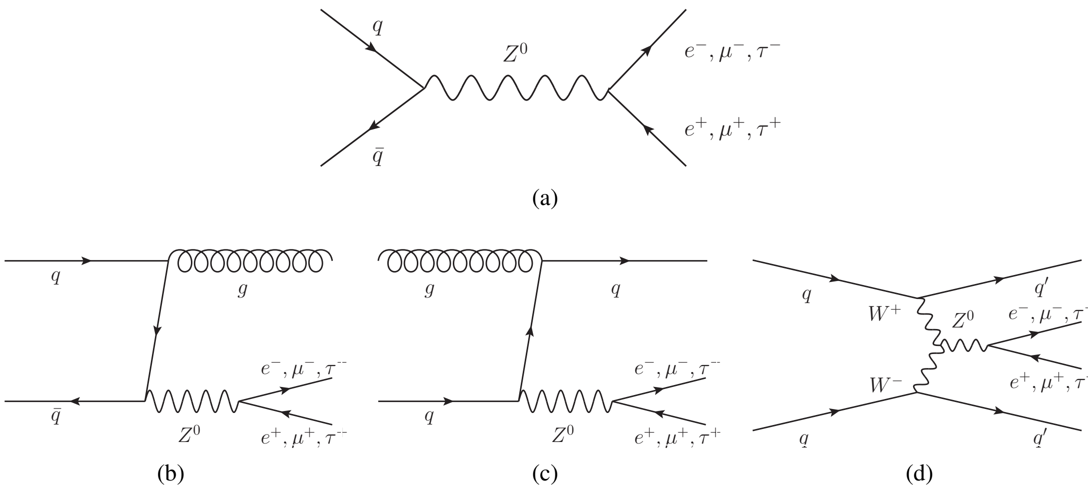


Figure 5.7: Exemplary Feynman graphs for the production and the decay of the Z boson into leptons in leading and higher orders.

[3] The ATLAS collaboration published another analysis which includes the search for this rare mode in Reference [103].

As the $H \rightarrow \tau_\ell \tau_h$ decay, $Z \rightarrow \tau\tau$ events are characterised by the signature of a real light lepton and a real τ_{h} decay accompanied by missing transverse energy. Due to the same signature as the signal process they are difficult to distinguish from $H \rightarrow \tau\tau$ decays. The kinematic properties of additional jets in the final state, e.g. the difference in pseudorapidity between both leading jets, have some discriminating power. However, the most important discriminator between both processes is the mass difference between the Z and the Higgs boson. In order to exploit the mass difference it is essential to reconstruct the $\tau\tau$ mass as precise as possible and with a small resolution. In this analysis, the MMC mass reconstruction is chosen because of its good performance (see Section 7.3.5). Since the high mass tail of the $Z \rightarrow \tau\tau$ background overlaps with the much smaller signal mass peak, also other variables are necessary to further separate both processes.

- **$Z \rightarrow ee$+jets and $Z \rightarrow \mu\mu$+jets:** $Z \rightarrow ee$ and $Z \rightarrow \mu\mu$[4] events can imitate the signal signature, if one of the two light leptons in the final state is not reconstructed as lepton. One distinguishes two cases sketched in Figure 5.8: The τ_{had} candidate is either faked by a light lepton (Figure 5.8a) or by a jet (Figure 5.8b). The background estimate differs depending on the particular object, which is misidentified as τ_{had} candidate. While for jet$\rightarrow \tau_{\mathrm{had}}$ fakes no charge asymmetry is expected, in the case of electron/muon$\rightarrow \tau_{\mathrm{had}}$ fakes the charges of the selected lepton and τ_{had} candidate are strongly correlated, i.e. more light leptons and τ_{had} candidate pairs with opposite charge than same charge are expected. Further, the invariant mass reconstructed from the true light lepton and the electron/muon$\rightarrow \tau_{\mathrm{had}}$ is close to the mass of the Higgs boson. These effects, charge correlation and similar invariant mass, makes the case of electron/muon$\rightarrow \tau_{\mathrm{had}}$ fakes more difficult to distinguish from signal. But, compared to jets, the τ_{had} misidentification rate for muons and electrons is rather low. As a consequence, only events with a jet faking the τ_{had} candidate significantly contribute to the $H \rightarrow \tau_\ell \tau_h$ background. These events can be suppressed by accepting only events with exactly one light lepton (see Figure 5.8b).

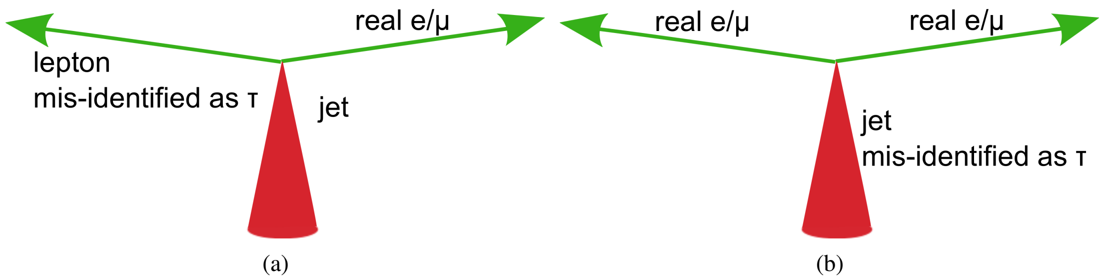


Figure 5.8: Sketch of possible $Z \rightarrow \ell\ell$ decays which might be selected when searching for $H \rightarrow \tau_\ell \tau_h$ signatures. While in Figure (a) one of the light leptons (electron or muon) from the Z boson decay is misidentified as τ_{had} candidate, in Figure (b) a jet is misidentified as τ_{had} candidate. In both cases, one of the light leptons originating from the Z boson decay is correctly identified. Background events of the second kind can be reduced if one requires one reconstructed light lepton at most in the event.

- **W+jets:** At the LHC, W bosons are mainly produced via quark anti-quark annihilation. Other production mechanisms involving quark-gluon interactions contribute in NLO. As illustrated in Figure 5.9, the production of the W boson in NLO is often accompanied by the emission of a quark or gluon which results in additional jet activity in the final state. These processes are also called W+jets events. The dominant decay process of the W boson with a branching ratio of

[4] In the following, the notation $Z \rightarrow \ell\ell$ is introduced for $Z \rightarrow ee$ and $Z \rightarrow \mu\mu$ decays.

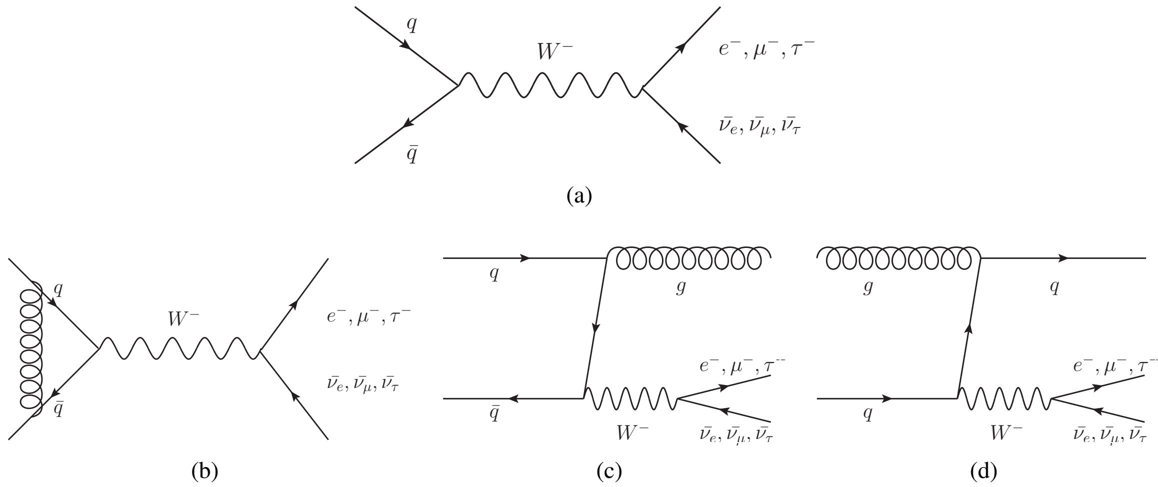
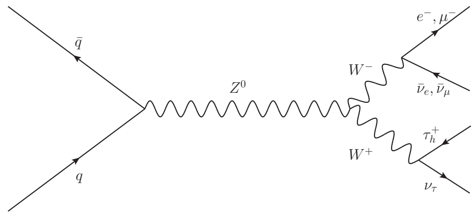


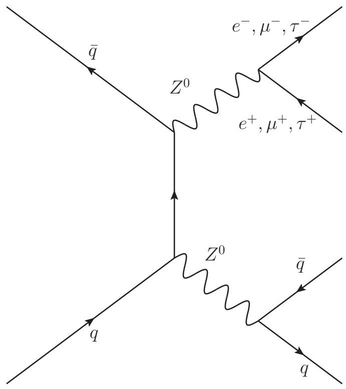
Figure 5.9: Exemplary Feynman graphs showing the production and the decay of the W boson in leading (top) and next-to-leading order (bottom).

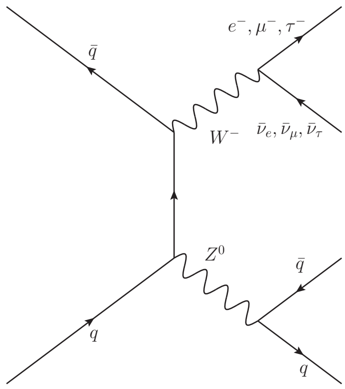
about 67.4% is the decay into a quark anti-quark pair. With a branching ratio of 10.9% it also decays into a charged lepton (e, μ, τ) and a neutrino $(\nu_e, \nu_\mu, \nu_\tau)$. Decays of the W boson can mimic the $H \to \tau_\ell\tau_h$ signature by a reconstructed light lepton originating from the W boson decay and a jet faking the τ_{had} candidate. Due to the large decay cross-section, which is one order of magnitude larger than for $Z \to \ell\ell$ processes, it significantly contributes to the background of $H \to \tau_\ell\tau_h$ searches. In particular, W+jets events are the dominant background source due to a misidentified τ_{had} object. Such events feature a significant amount of missing transverse energy because of the neutrino from the W boson decay. Since in most cases the jet originates from the hadronisation of an emitted quark, a charge anti-correlation between the fake τ_{had} candidate and the light lepton is found, so that slightly more fake τ_{had}/light leptons pairs with opposite charge than same charge are detected. One of the best ways to suppress W+jets events is a cut on the transverse mass (Section 5.6.2). For W+jets events, the transverse mass has its maximum slightly below the mass of the W boson, so that rejecting events with high transverse mass significantly reduces the contamination of this background.

- **Diboson ($ZZ, W^+W^-, W^\pm Z$:)** Processes in which a pair of electroweak bosons, ZZ, W^+W^- or $W^\pm Z$, is produced are referred to as diboson processes. Although a part of the diboson background can be rejected by requiring exactly one light lepton in the event, diboson events can still imitate $H \to \tau_\ell\tau_h$ signatures in several ways. Some possible decays are shown in Figure 5.10. For example, WW decays (Figure 5.10a) can have a real lepton and τ_{had} signature, if one W decays into a light lepton and the other one into a hadronically decaying τ lepton. The ZZ background (Figure 5.10b) mainly contributes if one Z boson decays leptonically. Signal and ZZ events can look alike due to the same reasons as described above for single Z boson decays. In WZ background processes (Figure 5.10c) a jet can be misidentified as τ_{had} candidate in similar ways as in single W or Z decays.
  Diboson events represent only a small background source for $H \to \tau_\ell\tau_h$ searches due to its small cross-section times branching ratio.



(a)

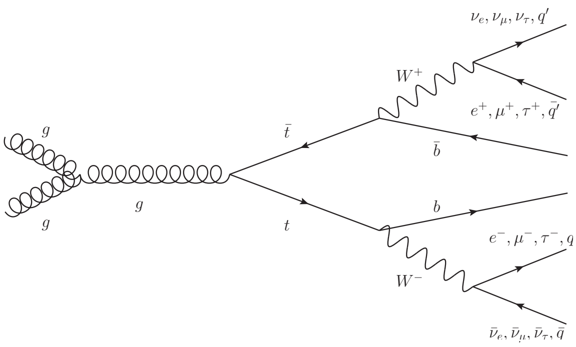




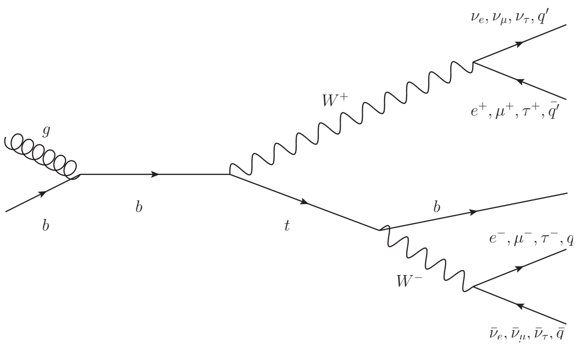
(b) (c)

Figure 5.10: Exemplary Feynman graphs for diboson processes which mimic $H \to \tau_\ell \tau_h$ signatures.

- ***$t\bar{t}$* and single top:** Another contribution to the background in $H \to \tau_\ell \tau_h$ searches arises from the production of two top quarks ($t\bar{t}$) or a single top. Examples for these processes are illustrated in Figure 5.11. Each top quark further decays into a b quark from which a b jet originates and a W boson which decays as described above. About 50% of $t\bar{t}$ and single top processes contribute



(a)



(b)

Figure 5.11: Exemplary Feynman graphs for $t\bar{t}$ (left) and single top (right) processes.

to the background due to a τ_h lepton from one W boson decay and a light lepton from the other W boson decay. The neutrinos from the W decay result in a non-negligible amount of missing transverse energy. The other half of these processes contains a light lepton originating from a W boson decay and a jet which is misidentified as τ_had candidate. Only few events contribute in which a lepton fakes the τ_had candidate.

In the case of $t\bar{t}$ events, two b jets are found in the final state, which can imitate the two jets from the VBF production mechanism of the Higgs boson. For single top events, one b jet is found in the final state. By vetoing b-tagged jets (Section 4.3.6) in the event, the contribution of $t\bar{t}$ and single top events to the background is reduced.

- **QCD events:** The QCD process is a generic term for all processes in which only quarks and gluons interact via strong interaction as e.g. displayed in Figure 5.12. Due to the hadronisation of these partons in the final state, these events are also referred to as multijet events. The QCD contamination results from two jets of which one is misidentified as τ_{had} candidate and the other as light lepton. Since the selected objects are jets they are not correlated with each other and as consequence no charge asymmetry is expected. The amount of missing transverse energy is ex-

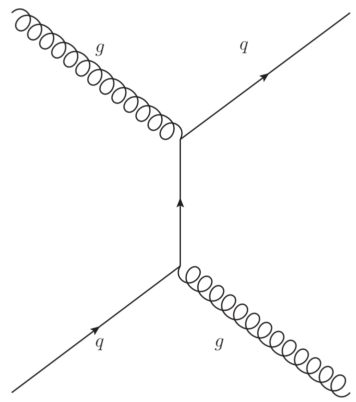


Figure 5.12: Exemplary Feynman graph for a QCD process.

pected to be small. This is due to the effect, that no neutrinos are produced in these processes and hence only misreconstruction and resolution effects contribute. The large cross-section of QCD processes is coped with by applying tight isolation and identification criteria within the object selection. Although such requirements significantly reduce the size of the QCD background, it remains a considerable background source in $H \to \tau_\ell \tau_h$ searches.

For the various processes different modelling methods are used as discussed in Section 7.3. The composition of these background processes in the $H \to \tau_\ell \tau_h$ channel is illustrated at analysis preselection level in Section 7.2.3. It differs from the composition in the other $H \to \tau\tau$ decay channels, where these background events contribute by imitating other signatures depending on the τ decay mode.
The preselection level, at which events with characteristic $H \to \tau_\ell \tau_h$ signature are selected, still show large contributions of background processes. As already indicated, different discriminating variables can be used to further suppress them. They employ differences in the kinematics and event topologies compared to the signal process and are introduced in detail below.

5.5 Reconstruction of the invariant $\tau\tau$ mass

For the study of the $H \to \tau\tau$ decay channel, the reconstruction of the invariant $\tau\tau$ mass is of high importance in order to separate the Higgs signal from the background. In particular for the reduction of the irreducible $Z \to \tau\tau$ background the invariant $\tau\tau$ mass is essential as it is the best variable to differentiate between both processes. The mass can be reconstructed in several ways. They differ in the treatment of the undetected neutrino momenta which have to be considered in order to fully reconstruct the invariant mass.

5.5.1 Visible and effective mass

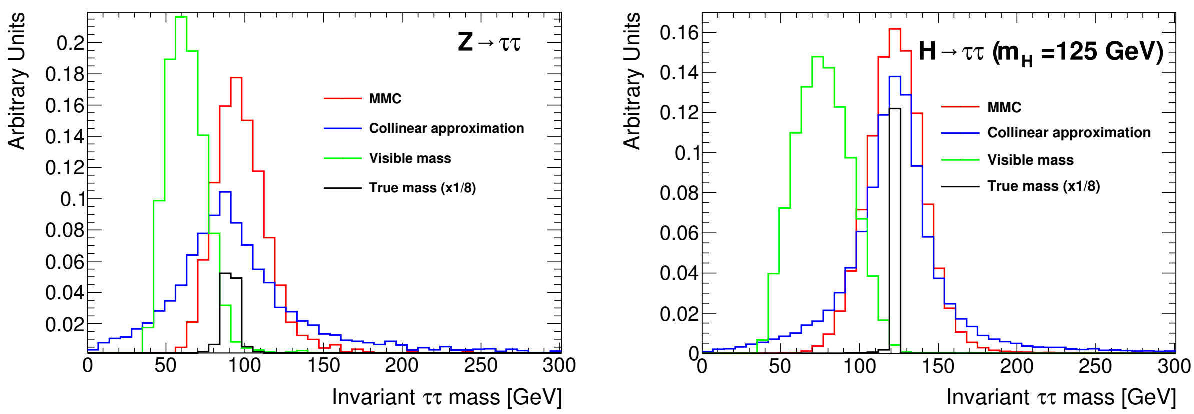
The visible mass m_{vis} is the simplest way to reconstruct the invariant $\tau\tau$ mass. It is exclusively based on the visible τ decay products:

$$m_{\text{vis}} = (\mathbf{p}_\ell + \mathbf{p}_{\tau_{\text{had}}})^2$$

with $\mathbf{p}_\ell$ and $\mathbf{p}_{\tau_{\text{had}}}$ denoting the four-vectors of the light lepton and the τ_{had} candidate. Since the visible mass does not consider the momenta of the invisible τ decay products, it is smaller than the true invariant $\tau\tau$ mass and has a worse mass resolution (see Figure 5.13).
Considering that the vectorial sum of all neutrinos from τ lepton decays is given by the measured missing transverse momentum yields the effective mass m_{eff}:

$$\begin{aligned} m_{\text{eff}} &= (\mathbf{p}_\ell + \mathbf{p}_{\tau_{\text{had}}} + \mathbf{p}_T^{\text{miss}})^2 \\ \text{with } \mathbf{p}_T^{\text{miss}} &= \begin{pmatrix} E_T^{\text{miss}} \\ \vec{p}_T^{\,\text{miss}} \end{pmatrix} . \end{aligned}$$

The effective mass is a better approach than m_{vis}, since it takes the missing transverse energy into account. However, it is still incomplete, since the longitudinal component of $\vec{p}_T^{\,\text{miss}}$ cannot be measured. To improve the mass reconstruction other techniques have to be used which are capable of estimating the longitudinal component.



(a) (b)

Figure 5.13: The reconstructed invariant $\tau\tau$ mass compared to the true $\tau\tau$ mass for the Z boson decay (left) and Higgs boson decay (right) with m_H=125 GeV. For the reconstruction different mass algorithms are used.

5.5.2 The collinear mass

Provided that the τ leptons have large momenta, i.e. are strongly boosted, the visible and invisible τ decay products fly in a very similar direction as the initial τ leptons. This behaviour is used in the calculation of the collinear mass [104]. In the collinear approximation, illustrated in Figure 5.14, the direction of the in-/visible decay products associated to the same τ lepton are assumed to be equal. Under this assumption, the missing transverse momentum can be projected onto the directions of the

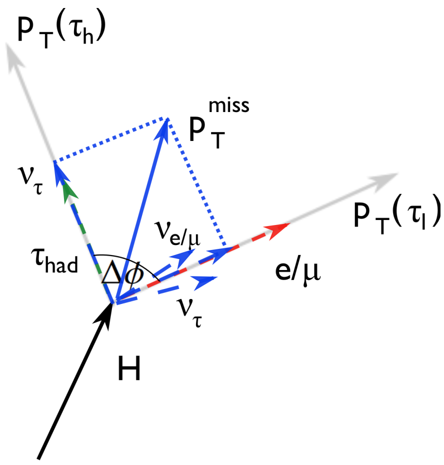


Figure 5.14: Sketch of the $H \to \tau_\ell\tau_h$ event kinematics in the transverse plane: In the collinear approximation, the projection of $\vec{p}_T^{\,\mathrm{miss}}$ on the direction of the visible τ lepton decay products, illustrated in red and green, is an estimate for the neutrino momenta. The projection is indicated by the blue arrows. The resulting transverse momenta of the τ leptons are shown in grey.

visible light lepton and the τ_{had} candidate. The projection is an estimate of the neutrino momentum[5]. Due to the high momenta of the τ leptons it holds that the invariant $\tau\tau$ mass is approximately given by

$$\mathrm{m}^2 \approx p_\tau p_\tau (1 - \cos\Delta\Phi(\ell, \tau_{\mathrm{had}})) \,.$$

Using geometrical relations the mass can be rewritten in terms of the visible momentum fractions

$$\begin{aligned} x_\ell &= \frac{p_T(\ell)}{p_T(\tau_l)} \\ \text{and } x_{\mathrm{h}} &= \frac{p_T(\tau_{\mathrm{had}})}{p_T(\tau_{\mathrm{h}})} \,. \end{aligned}$$

The resulting collinear mass m_{coll} is given by

$$m_{\mathrm{coll}} = \frac{m_{\mathrm{vis}}}{\sqrt{x_\ell x_{\mathrm{h}}}} \,.$$

Although this method provides an estimate for the longitudinal component of $\vec{p}_T^{\,\mathrm{miss}}$, it has limitations in experimental applications. For example, if both τ leptons fly in opposite directions the projections of the missing transverse momentum on the visible decay products are an invalid estimate of the neutrino momenta. This can lead to non-physical solutions for the energy fractions x_i and cause large mass tails as shown in Figure 5.13.

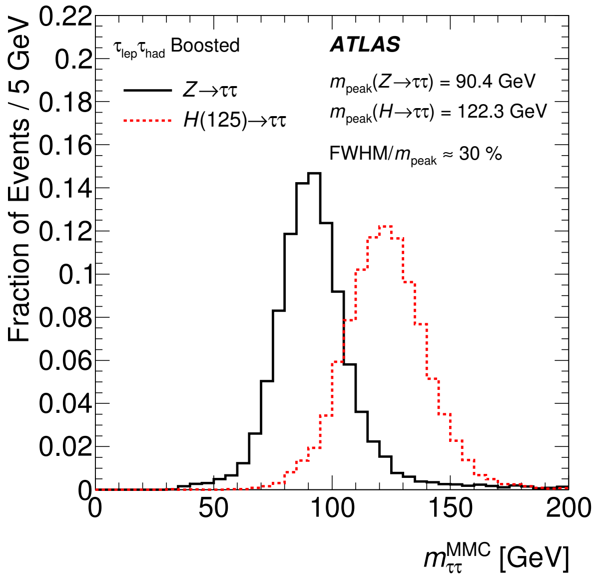
[5] In the case of the leptonically decaying τ lepton, two neutrinos are found in the final state. Hence the projection of the transverse momentum on the light lepton direction is an estimate for the vectorial sum of both neutrino momenta.

5.5.3 The $\tau\tau$ MMC mass

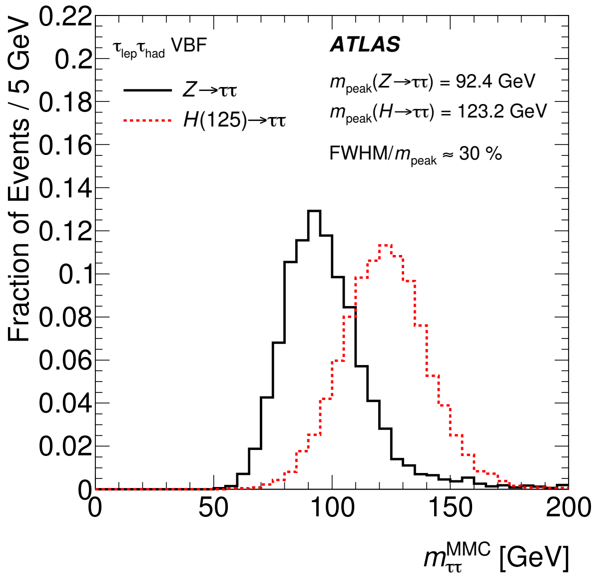
The missing mass calculator (MMC) [105] extends the concept of the collinear approximation and improves the modelling of the high mass tails (see Figure 5.13). Instead of requiring the τ decay products to be completely aligned, probability functions for the angle between the visible and invisible τ lepton decay products are derived and considered in the mass reconstruction. The MMC delivers an accurate reconstruction of the invariant $\tau\tau$ mass. The invariant mass calculated with the MMC is also referred to as MMC mass or $m_{\tau\tau}^{\mathrm{MMC}}$ in the following.
The exact reconstruction of the invariant $\tau\tau$ mass requires the knowledge of six unknowns, i.e. the three space components of the undetected momentum carried by the neutrinos from the decay of the two τ leptons. Depending on the τ decay mode up to two additional unknowns need to be considered. They result from τ_ℓ leptons, which reveal two neutrinos in the final state instead of one as for hadronically decaying τ leptons.
All unknowns are put in relation by the lower system of equations

$$\begin{aligned} E_{T,x}^{\mathrm{miss}} &= p_{\mathrm{mis},1} \sin\theta_{\mathrm{mis},1} \cos\phi_{\mathrm{mis},1} + p_{\mathrm{mis},2} \sin\theta_{\mathrm{mis},2} \cos\phi_{\mathrm{mis},2} \\ E_{T,y}^{\mathrm{miss}} &= p_{\mathrm{mis},1} \sin\theta_{\mathrm{mis},1} \sin\phi_{\mathrm{mis},1} + p_{\mathrm{mis},2} \sin\theta_{\mathrm{mis},2} \sin\phi_{\mathrm{mis},2} \\ M_{\tau,1}^2 &= m_{\mathrm{mis},1}^2 + m_{\mathrm{vis},1}^2 + 2\sqrt{p_{\mathrm{vis},1}^2 + m_{\mathrm{vis},1}^2}\sqrt{p_{\mathrm{mis},1}^2 + m_{\mathrm{mis},1}^2} - 2p_{\mathrm{vis},1}p_{\mathrm{mis},1}\cos\theta_{\nu,m1} \\ M_{\tau,2}^2 &= m_{\mathrm{mis},2}^2 + m_{\mathrm{vis},2}^2 + 2\sqrt{p_{\mathrm{vis},2}^2 + m_{\mathrm{vis},2}^2}\sqrt{p_{\mathrm{mis},2}^2 + m_{\mathrm{mis},2}^2} - 2p_{\mathrm{vis},2}p_{\mathrm{mis},2}\cos\theta_{\nu,m2} \end{aligned} \tag{5.5}$$

with $E_{T,x/y}^{\mathrm{miss}}$ being the space components of the E_T^{miss} vector and $p_{\mathrm{vis}1,2}$, $m_{\mathrm{vis}1,2}$, $\theta_{\mathrm{vis}1,2}$, $\phi_{\mathrm{vis}1,2}$ ($p_{\mathrm{mis}1,2}$, $m_{\mathrm{mis}1,2}$, $\theta_{\mathrm{mis}1,2}$, $\phi_{\mathrm{mis}1,2}$) the momenta, invariant masses, polar and azimuthal angles of the visible τ decay



(a)



(b)

Figure 5.15: Distributions of the invariant $\tau\tau$ mass calculated with the MMC for MC simulated $H \to \tau_\ell\tau_h$ (red) and embedded $Z \to \tau\tau$ (black) events. The events passed the boosted (left) or VBF category (right) of the $H \to \tau_\ell\tau_h$ analysis. The figures are taken from [8].

products (of the invisible part carried by the neutrinos). The angle between p_{vis} and p_{mis} is denoted as $\theta_{v,m1,2}$.
Although an exact solution of these equations is impossible because of its under-constraintment, one can distinguish solutions, which are more likely than others by using additional information provided by the distance $\Delta R(\text{vis}, \text{mis})$ between the visible and invisible τ lepton decay products: By scanning through different points in the parameter space different solutions for the unknown variables are found. Each scanned point obtains a weight which depends on the probability of the calculated $\Delta R(\text{vis}, \text{mis})$. The probability functions which depend on the τ momentum and the decay mode are derived from Monte Carlo simulations. Finally, the weighted invariant $\tau\tau$ mass is calculated for each scanned point and filled in a histogram. The maximum value of the obtained $m_{\tau\tau}$ distribution is the final estimator $m_{\tau\tau}^{\text{MMC}}$. Since the performance of the MMC algorithm is highly correlated with the resolution of E_T^{miss}, those resolution effects are also accounted for in the kinematic scan.
For the Higgs and Z boson decays the efficiency of the MMC algorithm is above 99%. For that reason, events failing the MMC are not further considered in the analysis of the $H \to \tau\tau$ decay channel. The small number of signal events, for which the MMC algorithm fails, are the result from huge fluctuations of E_T^{miss} or other scanned variables.
In Figure 5.15, the $m_{\tau\tau}^{\text{MMC}}$ distribution is depicted for $H \to \tau\tau$ and $Z \to \tau\tau$ events passing the $H \to \tau_\ell\tau_h$ analysis categories (see Section 7.2.4). The resolution which is given by the ratio between the maximum value and the full width at half maximum is about 30%. The high efficiency in combination with the good resolutions favours the MMC technique over the collinear approximation or other methods.

5.6 Discriminating event observables

Differences in the event topology between signal and background processes are exploited to determine signal-enriched phase space regions. In the following, event observables are introduced which describe characteristic properties of signal and background processes. They will be used in the analyses presented in Chapter 7- 9. Only few observables are employed to define signal categories which mainly aim for selecting different Higgs production mechanisms (see Section 7.2.4). The larger part of variables is used within classifiers to reject background processes and hence increases the signal-to-background ratio. The performance of these variables differs for the several background sources, since the two groups of background sources, processes with a fake and true τ_{had} signatures, have different kinematics. Some observables are mainly suitable to reject processes with a fake τ_{had} candidate, while others are necessary to reject background event with true τ_{had} signatures. A small selection of those observables is shown in Figure 5.16. In Figure 5.16a (Figure 5.16b) the shape of the ratio between the transverse momentum of the selected light lepton and the τ_{had} candidate (the difference in pseudorapidity) is shown for $H \to \tau_\ell\tau_h$ signal and $Z \to \tau\tau$ background events. Examples for shape differences between the $H \to \tau_\ell\tau_h$ signal and $W \to \mu\nu$ events are depicted in Figure 5.16c and Figure 5.16d. They illustrate the distance between the selected light lepton and τ_{had} candidate and the transverse mass, respectively.

5.6.1 Variables related to spin correlations

The kinematics of the light lepton and the τ_{had} candidate is affected by the spin of the decaying boson. While the Higgs boson is a particle with spin zero, the Z boson carries a spin of one. The different spin causes small differences in the momentum distributions of the visible decay products and hence provide additional discriminating power between the signal and background process. For example in the case of a Higgs boson decay, a larger momentum difference between the visible decay products $|p_T(\tau_{\text{had}}) - p_T(\ell)|$ than in Z boson decays is expected. The visible energy fraction x_ℓ and $x_{\tau_{\text{had}}}$ as well

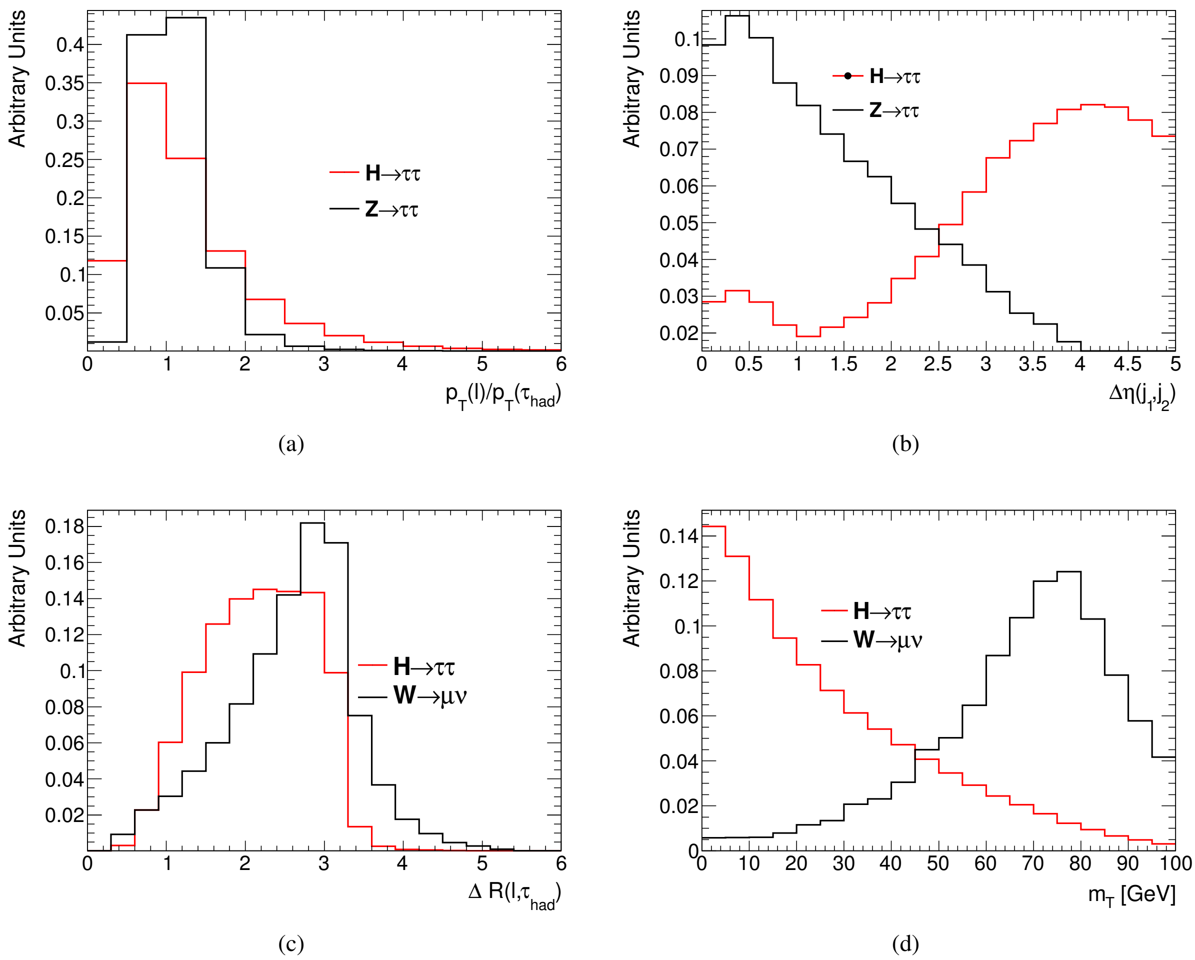


Figure 5.16: Shape comparisons for some event observables in $H \to \tau_\ell \tau_h$ decays (red) and background processes (black): (a) the ratio between the transverse momentum of the selected light lepton and the τ_{had} candidate, (b) difference in pseudorapidity between the two leading jets, (c) distance between the selected light lepton and the τ_{had} candidate, (d) the transverse mass. In the upper (lower) plots comparisons between signal and $Z \to \tau\tau$ ($W \to \mu\nu$) events are shown.

as the momentum ratio $p_T(\ell)/p_T(\tau)$ also provide some separation power due to their sensitivity to spin correlations.

5.6.2 E_T^{miss} related observables

$E_T^{\text{miss}}\phi$ centrality

Angular relations between the selected objects are used to reduce the contributions from W+jets decays, because in contrast to the signal process, the selected objects in W+jets events do not belong to the same mother particle. Figure 5.17 illustrates how the angular orientation in W+jets events between light lepton and (mis-) identified τ_{had} candidate differs from signal processes: In Figure 5.17a the orientations of the $H \to \tau_\ell \tau_h$ decay products are displayed. Due to the additional jet activity in the Higgs production the Higgs system is boosted against the jet. E_T^{miss} points in the direction of the neutrinos from both

τ decays and is therefore expected to lie between the direction of the light lepton and the true $\tau_{\rm had}$ candidate. In the case of W+jets events, illustrated in Figure 5.17b, the $\tau_{\rm had}$ candidate is faked by a jet. Only one neutrino from the leptonically τ decay is present in the event so that $E_T^{\rm miss}$ points into the direction of this neutrino. $E_T^{\rm miss}$ is not between the angle enclosed by the lepton and $\tau_{\rm had}$ candidate. The $E_T^{\rm miss}\phi$ centrality quantifies the relative angular position of $E_T^{\rm miss}$ with respect to the visible τ decay

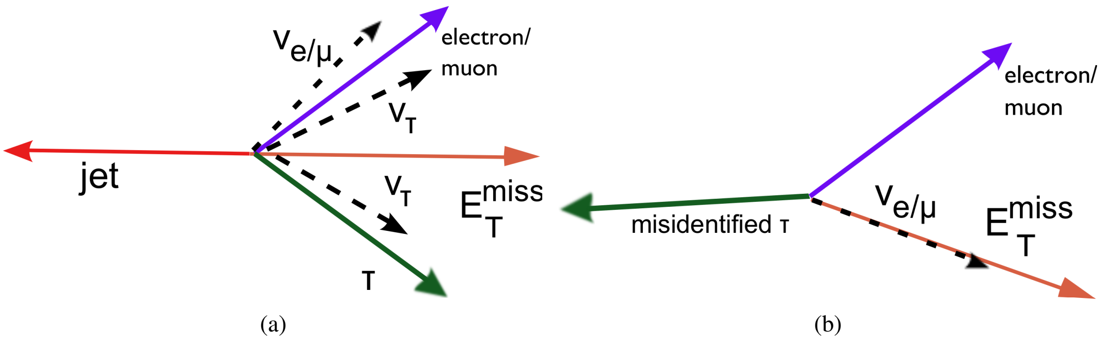


Figure 5.17: Scheme of the orientation of the light lepton, the (mis-) identified $\tau_{\rm had}$ candidate and $E_T^{\rm miss}$ in the transverse plane for (a) the $H \to \tau_\ell \tau_h$ decay and (b) the decay of a W boson.

products. It is defined as:

$$C(E_T^{\rm miss}\phi) = \frac{A(E_T^{\rm miss}) + B(E_T^{\rm miss})}{\sqrt{A(E_T^{\rm miss})^2 + B(E_T^{\rm miss})^2}} \tag{5.6}$$

with

$$A(E_T^{\rm miss}) = \frac{\sin(\phi(E_T^{\rm miss}) - \phi(\tau_{\rm had}))}{\sin(\phi(\ell) - \phi(\tau_{\rm had}))} \tag{5.7}$$

and

$$B(E_T^{\rm miss}) = \frac{\sin(\phi(\ell) - \phi(E_T^{\rm miss}))}{\sin(\phi(\ell) - \phi(\tau_{\rm had}))} \; . \tag{5.8}$$

According to its definition the variable has its maximum value of $\sqrt{2}$, if $E_T^{\rm miss}$ lies centrally between the visible τ decay products. A value of one corresponds to the case, in which $E_T^{\rm miss}$ is perfectly aligned with one of the visible decay products. For real $\tau\tau$ topologies $1 < C(E_T^{\rm miss}\phi) < \sqrt{2}$ is expected. If the $E_T^{\rm miss}\phi$ centrality is smaller than one, $E_T^{\rm miss}$ is not enclosed by the visible τ decay products. Hence, processes with jets faking the $\tau_{\rm had}$ candidate are expected to have $C(E_T^{\rm miss}\phi) < 1$.

The transverse mass

The transverse mass $m_T(\ell/\tau_{\rm had})$ also makes use of the correlation between the direction of $E_T^{\rm miss}$ and the visible τ decay products. It partially reconstructs the invariant $\tau\tau$ mass based on one of the visible τ decay products (either ℓ or $\tau_{\rm had}$) and the measured $E_T^{\rm miss}$:

$$m_T(\ell/\tau_{\rm had}) = \sqrt{2 \cdot p_T(\ell/\tau_{\rm had}) \cdot E_T^{\rm miss}\left(1 - \cos\Delta\phi(\ell/\tau_{\rm had}, E_T^{\rm miss})\right)} \; . \tag{5.9}$$

In W+jets events (Figure 5.17b), the angle between the lepton and E_T^{miss} is compatible large, which results in a large $m_T(\ell)$ component. Hence, $m_T(\ell)$ is already used at analysis preselection level to reject this background source (see Section 7.2.3). The equivalent $m_T(\tau_{\text{had}})$, based on the detected τ_{had} candidate, becomes larger if the energy of a jet is mismeasured. It is suitable to reject QCD background and is used as input for the BDT training in Section 7.6.

5.6.3 Boosted topologies

Summed transverse momentum

The sum of the absolute transverse momenta of the lepton, the τ_{had} candidate and all jets passing the jet quality criteria

$$\sum p_T = |p_T(\tau_{\text{had}})| + |p_T(\ell)| + \sum_i |p_T(jet_i)| \tag{5.10}$$

is a measure of the total activity. It is expected to be smaller in cases of weak boson decays than in $t\bar{t}$ or multijet processes.

Transverse momentum of the reconstructed Higgs boson

The transverse momentum of the reconstructed $\tau\tau$ system is referred to as $\vec{p}_T(H)$:

$$\vec{p}_T(H) = \vec{p}_T(\tau_{\text{had}}) + \vec{p}_T(\ell) + \vec{E}_T^{\text{miss}} \, . \tag{5.11}$$

In the case of additional partons, produced within ggF or Drell-Yan processes, the $\tau\tau$ system is recoiled, which results in a large $\vec{p}_T(H)$. Since the $\vec{p}_T(H)$ spectrum for signal events is harder, it has some discriminating power against $Z \to \tau\tau$ background events. However, the definition of the boosted event category[6] does not only benefit from the separation power of this observable, but also from its correlation to other observables. Due to the correlations an even better discrimination between signal and background events is obtained. With increasing $\vec{p}_T(H)$, $\sum E_T$ scales up as well and so the relative E_T^{miss} resolution is getting worse. Due to the dependence of $m_{\tau\tau}^{\text{MMC}}$ on the E_T^{miss} resolution, the relative $m_{\tau\tau}^{\text{MMC}}$ resolution improves with larger $\vec{p}_T(H)$, which further enhances the separation power of $m_{\tau\tau}^{\text{MMC}}$ for $H \to \tau\tau$ against $Z \to \tau\tau$. Another observable, which exploits correlations with $\vec{p}_T(H)$, is the distance between the two visible τ decay products $\Delta R(\ell, \tau_{\text{had}})$. For the lepton and τ_{had} candidate originating from τ leptons of a boosted Higgs boson decay, an angular relation is observed, whereas in processes with a fake τ_{had} candidate like multijet or $t\bar{t}$ events the kinematics of the visible products is not directly related with each other.

5.6.4 VBF topologies

The above discussed observables describe characteristics of boosted Higgs system independent of the Higgs production mechanism. Higgs bosons produced via VBF form a special case for boosted Higgs systems. Such VBF-like events are characterised by two well-separated jets with high momentum as illustrated in Figure 5.18. The following observables use properties of the VBF topology to discriminate between signal and background.

[6] Besides other requirements $p_T(H)$ has to exceed 100 GeV. For further details see Section 7.2.4.

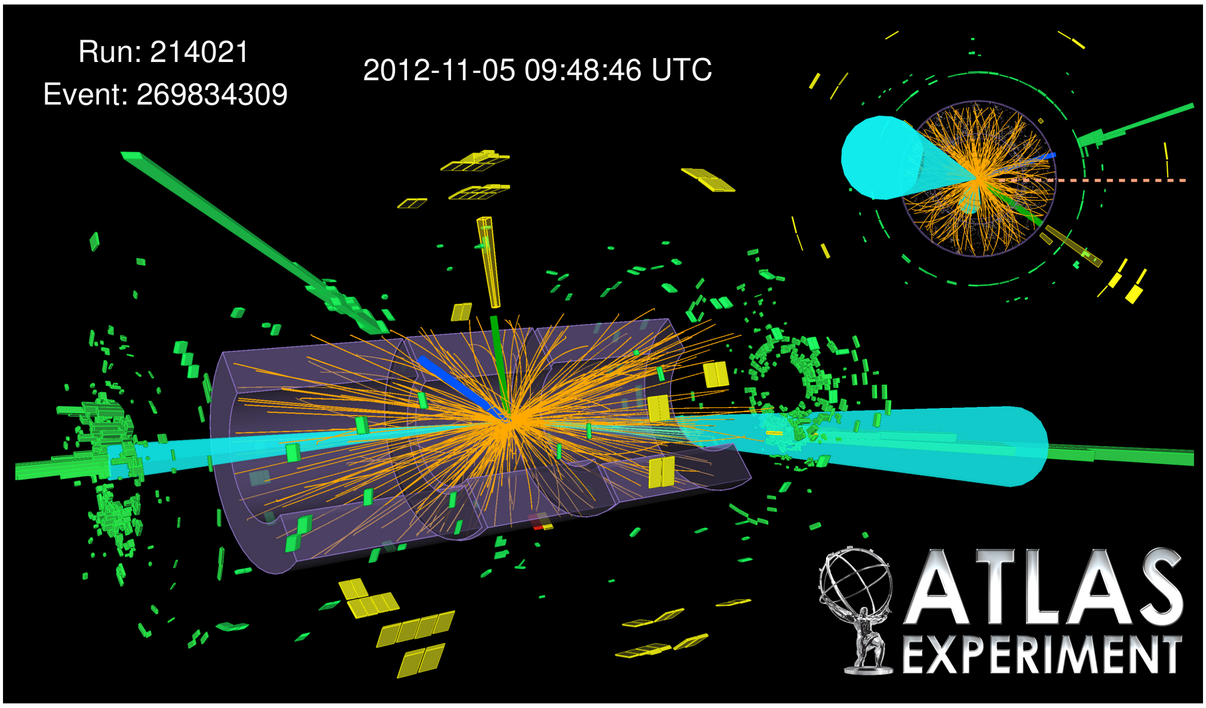


Figure 5.18: Event display of an VBF-like $H \to \tau_\ell \tau_h$ event: The electron originating from the τ decay is illustrated as blue track, while the visible decay product of the hadronically decaying τ lepton is represented by a green track and the yellow cluster. The direction of E_T^{miss} is indicated by the dashed line. The two turquoise cones visualise the two VBF jets. The figure is taken from [106].

Total transverse momentum

A suitable variable to measure additional activity in VBF-like events and to reject QCD background is p_T^{tot}. It is defined as the vectorial sum of the visible decay products, the two leading jets and $\vec{E}_T^{\text{miss}}$:

$$p_T^{\text{tot}} = |\vec{p}_T(\tau_{\text{had}}) + \vec{p}_T(\ell) + \vec{p}_T(\text{leading jet}) + \vec{p}_T(\text{subleading jet}) + \vec{E}_T^{\text{miss}}| \,. \tag{5.12}$$

In the case of electroweak quark scattering, p_T^{tot} vanishes, whereas for QCD events the momenta of these objects do not necessarily balance each other.

Difference in pseudorapidity

A cut on the pseudorapidity difference of the two leading jets $\Delta\eta(j_1, j_2) = \eta(j_1) - \eta(j_2)$ ensures well-separated jets. It is also used to define the VBF signal category in Section 7.2.4.

Product of pseudorapidities

The product of pseudorapidities $\eta(j_1) \times \eta(j_2)$ also accounts for the well-separated jets originating from VBF processes. The product becomes negative, if the two jets are detected in opposite detector hemispheres.

$\ell - \eta$ centrality

In VBF signal processes the direction of the Higgs boson typically points between the two jets so that the light lepton from the $H \to \tau_\ell\tau_h$ decay is found between these jets as well. The position of this lepton (η_ℓ) with respect to the directions of these jets (η_1 and η_2) is quantified by the $\ell - \eta$ centrality:

$$C_{\eta_1,\eta_2}(\eta_\ell) = \exp\left[-\frac{4}{(\eta_1-\eta_2)^2}\left(\eta_\ell - \frac{\eta_1+\eta_2}{2}\right)^2\right] . \tag{5.13}$$

It is defined such that it has its maximum value of 1 if η_ℓ is centrally between the two jets. If the lepton is perfectly aligned with one of the jets, the $\ell - \eta$ centrality drops to $1/e$. It lies below, if the lepton is not enclosed by both jets.

η centrality

A more general definition considers that not only the lepton but also the τ_{had} candidate is enclosed by both VBF jets. The η-centrality is defined as the product of $\ell - \eta$ centrality and the $\tau_{\text{had}} - \eta$ centrality:

$$C_{\eta_1,\eta_2} = C_{\eta_1,\eta_2}(\eta_\ell) \cdot C_{\eta_1,\eta_2}(\eta_{\tau_{\text{had}}}) . \tag{5.14}$$

Invariant dijet mass

The invariant dijet mass m_{jj} contains information about the momenta of the two jets and their angular correlation. Compared to multijet or $t\bar{t}$ or $Z \to \tau\tau$ processes, VBF jets are expected to have larger transverse momenta and thus a larger dijet mass. Also contributions from diboson events can be rejected using m_{jj}. For example, the signature of the ZZ decay in which one Z boson decays into two τ leptons and from the other two jets originate is very similar to the $H \to \tau\tau$ decay in which the Higgs is produced via VBF. But the invariant dijet mass is about the mass of the Z boson.

CHAPTER 6

Modelling of $Z \rightarrow \tau\tau$ processes using data-driven methods

The correct modelling of background processes is one of the essential ingredients of the $H \rightarrow \tau\tau$ analysis, because its experimental sensitivity relies on the study of fairly complex event signatures involving not only observables related to the Higgs candidate, but also global observables. For example, the missing transverse momentum serves as one of the inputs for the reconstruction of the MMC mass (see Section 5.5.3), which amongst others serves as a multivariate classifier for signal extraction, see Section 5.5.3.
Especially the modelling of the $Z \rightarrow \tau\tau$ process[1] - the largest irreducible background - is of huge importance for the analysis. Dedicated measurements in ATLAS [107–109] have illustrated, that current Monte Carlo simulations of Z+jets events do not correctly model the data. Hence, a data-based estimate is favoured to reduce potential effects due to mismodelling in MC and detector simulation. However, it is difficult to obtain sufficiently pure $Z \rightarrow \tau\tau$ samples from data, since such events cannot be selected without also including events from other background processes, in which e.g. jets are misidentified as τ_{had} leptons. More important, it is conceptually impossible to select a signal-free sample. As illustrated in Figure 6.1, the Higgs signal overlaps with the right tail of the Z mass peak, which makes it impossible to separate both processes. Nonetheless, this particular region must be well modelled.
The embedding method [9] offers the possibility to still model $Z \rightarrow \tau\tau$ events in a largely data-driven way, exploiting that the kinematics of $Z \rightarrow \mu\mu$ decays are - apart from the difference in mass between the muon and the τ lepton - the same as for $Z \rightarrow \tau\tau$ decays. In contrast to $Z \rightarrow \tau\tau$ events, $Z \rightarrow \mu\mu$ decays can be selected with high purity from data and due to the small muon mass the branching ratio of $H \rightarrow \mu\mu$ decays is several magnitudes smaller compared to $H \rightarrow \tau\tau$ decays. For that reason, the signal contamination is negligible.
Using the muon four momenta of selected $Z \rightarrow \mu\mu$ events, a $Z \rightarrow \tau\tau$ with identical kinematics is simulated. Afterwards the detector response of the original $Z \rightarrow \mu\mu$ decay is removed from the original event and replaced by the corresponding information of the simulated decay. The combination of both events takes place on cell and track level. In a final step, the hybrid event, which is referred to as embedded event, has to be re-reconstructed.
Embedded events have the advantage, that except for the simulated τ lepton decays, the kinematics of Z+jet events as well as underlying and pileup events are directly taken from data. The embedding

[1] In the following, the simplified notations $Z \rightarrow \tau\tau$ and $Z \rightarrow \mu\mu$ are used for $Z/\gamma^* \rightarrow \tau\tau$ and $Z/\gamma^* \rightarrow \mu\mu$ processes.

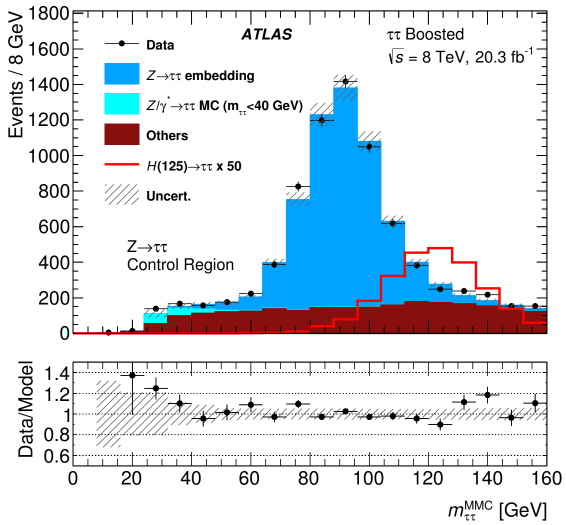


Figure 6.1: Invariant $\tau\tau$ MMC mass in the boosted $Z \to \tau\tau$ control region. In order to show the overlap with the signal (red) the cut on the MMC mass is removed. The $Z \to \tau\tau$ background is illustrated in blue. All other background processes are summarised in "others". The figure is taken from [9].

technique has been developed based on $Z \to \mu\mu$ MC simulations before the first data taking period of the LHC [110–112]. While embedded events served as cross-checks in early analyses, with time it has become an important ingredient for Higgs analyses in $\tau\tau$ final states [113–115] including the recent $H \to \tau\tau$ analysis (see Chapter 7 and 8). Its range of application has also been extended to single τ processes, e.g. searches for charged Higgs boson [116, 117] and analysis of $W \to \tau\nu$ [118, 119] decays.
During its application in various analyses a deeper understanding of the method was obtained which results in technical improvements as well as derivation of embedding specific systematic uncertainties. In particular, in the context of this thesis, further corrections and tools have been developed, which allow an almost standalone validation of the method. As later shown, these tools can also be used in the future to better estimate embedding-related uncertainties. Further, the effect of muon final state radiation on embedded events was quantified, demonstrating that these effects are covered by embedding-related uncertainties.
The performed studies were published in [9], to which this chapter mostly refers to. The analysis of the $H \to \tau\tau$ decay channel in Chapter 7 and 8 strongly benefit from the presented results discussed below and mainly drove the developments of the embedding technique within this thesis.
This chapter gives a detailed description of the embedding procedure and its validation: In Section 6.1 the embedding algorithm and its sub-steps are explained. Section 6.2 lists the various samples and selections for validation purposes. The discussion of embedding-specific properties distinguishing embedded samples from standard MC such as corrections applied to embedded events follows in Section 6.3 and 6.4. After the discussion of systematic uncertainties related to the embedding technique in Section 6.6, the validation of this technique is shown in Section 6.7. The performance of the embedding-specific corrections is discussed in Section 6.8. The last Section 6.9 suggests a further improvement of the method which might be useful for future applications.
If not otherwise stated this chapter focuses on the technical realisation and validation of embedded samples based on 2012 data recorded with $\sqrt{s}$ =8 TeV and used in the SM $H \to \tau\tau$ analysis.

6.1 The Embedding procedure

The embedding technique can be subdivided into five consecutive steps: After the selection of $Z \to \mu\mu$ input events, a $Z \to \tau\tau$ decay with identical decay kinematics as the corresponding $Z \to \mu\mu$ decay is generated and simulated. In the next step, the simulated event is merged with the input data event by replacing cells and tracks, which belong to the muons in the original event with the simulated τ decay products. Finally, the merged hybrid events have to be re-reconstructed, which implies the complete standard ATLAS reconstruction except for hits. Starting from the ESD format, all steps of the procedure are performed at the level of reconstructed tracks and calorimeter cells. In the following, the individual steps are described in detail according to the flowchart in Figure 6.2. Although the concept of the method corresponds the description below, the actual technical realisation is more complicated: The increasing amount of data taken, and therefore the larger number of embedded events has made it necessary to restructure the individual steps in order to exploit the available computer resources in a more efficient way during productions. The technical realisation within the ATLAS production setup is summarised in Appendix A.

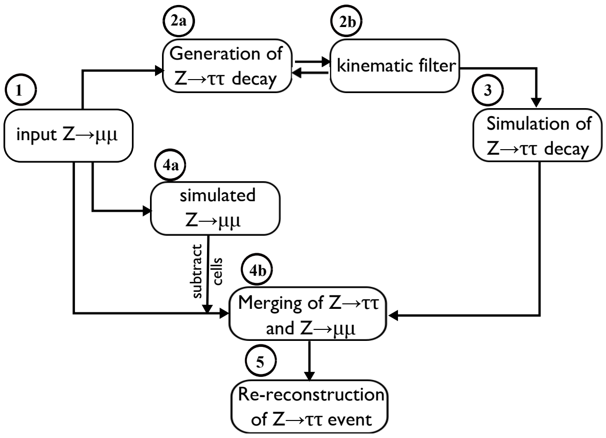


Figure 6.2: Flowchart of the embedding procedure.

- **1) Selection of $Z \to \mu\mu$ input events:**
  In order to obtain a rather pure $Z \to \mu\mu$ input sample with low background contamination, events have to pass several requirements listed in Table 6.1. The event has to be triggered by at least one of the two chosen muon triggers. The choice of the muon triggers is driven by the trigger selection and p_T thresholds used in the analyses in Chapter 7 and 8. The single muon trigger EF_MU24I_TIGHT triggers events with one isolated muon with transverse momentum of $p_T > 24$ GeV, whereas the di-muon trigger EF_MU18_TIGHT_MU8_EFFS requires two muons with $p_T >$18 GeV (8 GeV). The event has to contain at least two combined STACO muon candidates passing track quality criteria, which are provided by the muon performance group [80]. Muons also must have non-zero charge, $|\eta_\mu| < 2.5$ and $p_T >$15 GeV. For each event, all possible pairs of oppositely charged muon candidates with a common vertex are formed, where the transverse momentum of the leading candidate is required to exceed 20 GeV. Only events, that contain at least one such pair with an invariant mass $m_{\mu\mu}$ larger than 40 GeV, are selected and the muon pair with $m_{\mu\mu}$ closest to the Z boson mass is chosen as the $Z \to \mu\mu$ candidate. For the skimming of 7 TeV data no triggers are required and the momentum of the leading muon has only to exceed 15 GeV.

<table>
<tr><th></th><th>$\sqrt{s}$ =8 TeV</th><th>$\sqrt{s}$ =7 TeV</th></tr>
<tr><td>Trigger</td><td>EF_MU18_TIGHT_MU8_EFFS || EF_MU24I_TIGHT</td><td>not required</td></tr>
<tr><td>GRL</td><td>Not required</td><td>Required</td></tr>
<tr><td rowspan="5">Muons</td><td colspan="2">At least two combined STACO muons with $q = \pm 1e$ and $|\eta| < 2.5$</td></tr>
<tr><td colspan="2">Requirements on hits in the ID</td></tr>
<tr><td colspan="2">Following the recommendations of Muon performance group [80]</td></tr>
<tr><td>$p_T(\mu_1)$>20 GeV , $p_T(\mu_2)$>15 GeV</td><td>$p_T(\mu_1)$ $p_T(\mu_2)$>15 GeV</td></tr>
<tr><td colspan="2">$I(E_T, 0.2)/p_T$<0.2</td></tr>
<tr><td rowspan="3">Combination of muons</td><td colspan="2">Originating from same vertex</td></tr>
<tr><td colspan="2">$m_{\mu\mu}$>40 GeV</td></tr>
<tr><td colspan="2">$q_{\mu 1} * q_{\mu 2} < 0$</td></tr>
</table>

Table 6.1: Selection requirements for $Z \to \mu\mu$ data events used for the embedding technique.

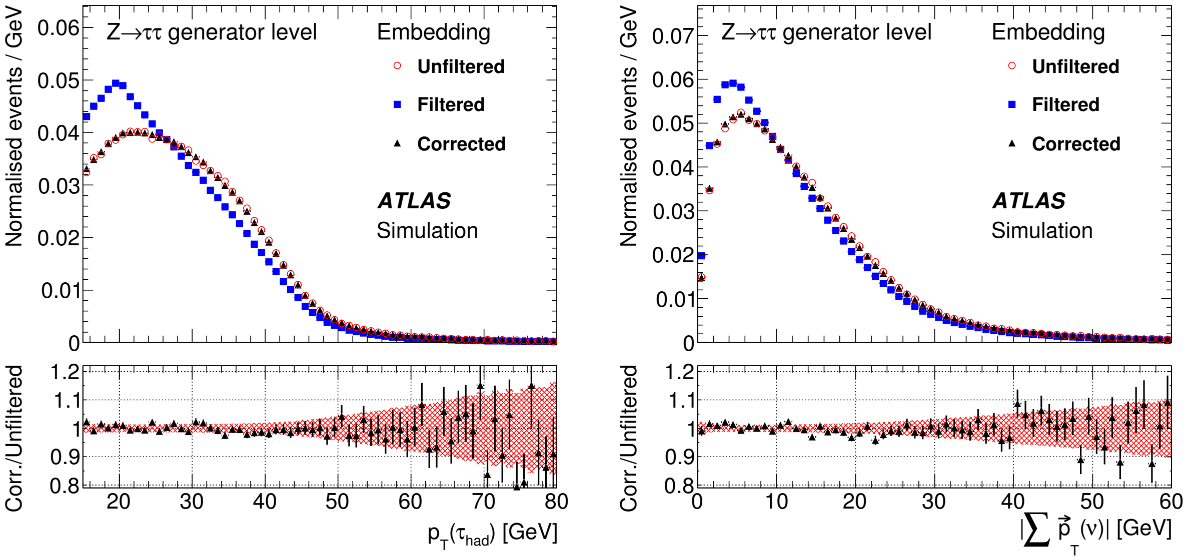
- **2a) Generation of $Z \to \tau\tau$ decays:**
In the next step, a $Z \to \tau\tau$ decay with identical kinematics as the original $Z \to \mu\mu$ decay is generated. The muons and the reconstructed Z boson are written in an ASCII file in HEPEVT format, which contains the particle's four momentum, production or decay vertex (in case of the Z boson), information about possible daughter or mother particles and its particle identification number. To declare the $Z \to \mu\mu$ decays as $Z \to \tau\tau$ decays the particle identification number of the muons is changed to that of the τ lepton and their four momenta is corrected by the factor

$$\frac{\sqrt{E_\mu^2 - m_\tau^2}}{|\vec{p}_\mu|} \tag{6.1}$$

in order to account for the difference in mass between both leptons. Next, the obtained $Z \to \tau\tau$ decay kinematics is handed over to TAUOLA [120] and PHOTOS [121] to generate the τ decay. Whereas the τ polarisation and spin correlation are correctly taken into account by TAUOLA the polarisation of the Z boson depends on the initial parton configuration which is not accessible in collision data. For that reason, TAUOLA assumes an average polarisation of zero and assigns a random helicity of ± 1 to each Z boson. To account for the non-zero average Z polarisation, weights are obtained by the TauSpinner program [122–124] as described in Section 6.5.

- **2b) Kinematic filter:**
$Z \to \tau\tau$ decays are generated based on physical probability distributions of the decay kinematics. As a consequence, many generated τ decay products, namely the visible part of hadronically decaying τ leptons and leptonically decaying τ leptons, cannot pass the selection criteria of physics objects in the $H \to \tau\tau$ analysis. The available background statistics is therefore significantly reduced in signal sensitive regions. To increase the statistics in these regions, a kinematic filter is implemented at generator level. For each $Z \to \mu\mu$ input data event, it generates 1000 $Z \to \tau\tau$ decays with different kinematic configurations of the τ decay products. These configurations correspond to the physical probability distributions. For each of those $Z \to \tau\tau$ decays, the filter tests whether the transverse momenta of the visible τ decay products exceed certain threshold values. These thresholds are driven by cut values, used for leptons in the $H \to \tau\tau$ analysis. The first event, passing the filter thresholds, is the only event handed over to the ATLAS simulation and finally embedded into the input data event. The results of the other iterations are used to calculate the

filter efficiency for each event. The particular choice of the filter thresholds and the resulting filter efficiencies are discussed in Section 6.3.1. Using the filter causes a kinematic bias as illustrated in Figure 6.3. This bias is corrected for by evaluating the calculated event-by-event filter efficiencies and applying these as weights to the event.



(a) (b)

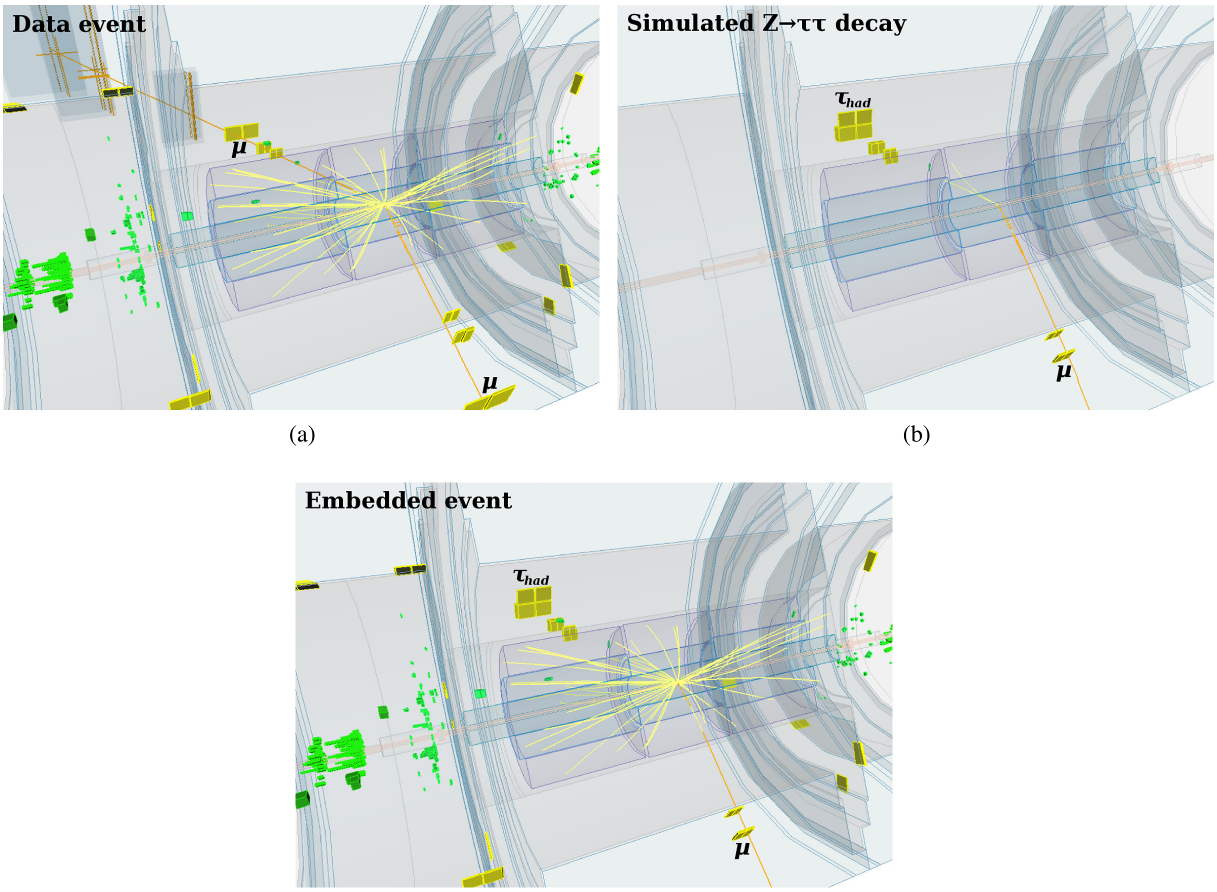
Figure 6.3: Comparisons of τ-embedded events on generator-level: (a) τ_{had} transverse momentum and (b) the summed transverse momenta of all neutrinos. (Un-) filtered events are shown in blue (red), events with applied filter weights in black. The lower ratio illustrates the relative deviation of the corrected distributions from the unfiltered ones. The error band (bars) correspond to the statistical uncertainty from the unfiltered (filtered) events. The figures are taken from [9].

- **3) Simulation of $Z \to \tau\tau$ decays**
  The resulting $Z \to \tau\tau$ decay, which corresponds to a standard event generator output like PYTHIA, is processed with the standard ATLAS detector simulation, digitisation and reconstruction. In order to prevent possible double counting in the merging step, the simulation of the calorimeter noise is disabled. The resulting simulated event contains only the pure $Z \to \tau\tau$ decay without underlying and pileup events and is therefore referred to as *mini event* in the following.

- **4) Merging of simulated $Z \to \tau\tau$ and $Z \to \mu\mu$ data event**
  The signature of the $Z \to \mu\mu$ decay in the input event needs to be removed and replaced by the corresponding tracks and cells of the simulated $Z \to \tau\tau$ decay. In order to remove the calorimeter cells associated with the muons of the original event, a $Z \to \mu\mu$ decay with the same kinematics as the input event but without underlying and pileup event is simulated. The simulated cell energies are subtracted from the corresponding cells in the data event and all calorimeter cells of the simulated $Z \to \tau\tau$ mini event are copied into the data event. The simulated tracks of the τ decay products are copied as well. These copied tracks and cells are the only part of the hybrid event, which is simulated and therefore the event properties are still very close to data conditions.

- **5) Re-reconstruction of the hybrid event**
  In the final step, the standard ATLAS reconstruction needs to be applied to the hybrid event, because the combination of the event information at the level of fully reconstructed physics objects is insufficient. For example, the identification of hadronic τ decays depends on the details of the calorimeter response. Also, the calculation of E_T^{miss} is potentially influenced by additional calorimeter energy depositions. By re-running the reconstruction algorithms, all physics objects are recreated except for tracks and cells, from which the reconstruction starts. In common applications, particle tracks are reconstructed based on detector hits. In the re-reconstruction this track reconstruction is impossible due to missing hit information. However, corresponding effects are negligible in the phase space of $H \to \tau\tau$ analysis. For that reason, starting the reconstruction from tracks and cells is sufficient for the embedding procedure.

The event displays in Figure 6.4 illustrate the three main steps of the procedure: A selected $Z \to \mu\mu$ input event, the corresponding mini event, in which one τ lepton decays hadronically and the other one into a muon, and the finally merged embedded $Z \to \tau\tau$ event.



(a) (b) (c)

Figure 6.4: Event displays of (a) a $Z \to \mu\mu$ input event selected from data, (b) the corresponding simulated mini event with one hadronically decaying τ lepton and the other decaying into a muon such as (c) the merged embedded $Z \to \tau\tau$ event. The figures are taken from [9].

6.2 Dedicated validation samples and selection

The simulated $Z \to \tau\tau$ MC reference sample as well as the $Z \to \mu\mu$ MC input sample are produced according to the description in Section 7.1. Embedded $Z \to \tau\tau$ or $Z \to \mu\mu$ events are generated as described in Section 6.1. The τ decays are simulated with Tauola [120], while the photon radiation from charged leptons is provided by Photos [121]. From simulated $Z \to \mu\mu$ MC or collision data events the following embedded samples are obtained:

- *τ-embedded* data: Muons from $Z \to \mu\mu$ collision data are replaced with τ leptons from simulated $Z \to \tau\tau$ decays. These are the standard event samples used in the $H \to \tau\tau$ analysis to model the $Z \to \tau\tau$ background. They are based on the full 2011 and 2012 ATLAS dataset and validated in dedicated control regions (see Section 6.7.3).
- *μ-embedded* data: The muons from $Z \to \mu\mu$ input data are replaced by muons from simulated $Z \to \mu\mu$ decays. Those events provide a tool to study systematic effects of the embedding procedure in comparatively simple final states as in Section 6.7.1. Since these samples are only used for validation purposes, they are based on a subset of ATLAS collision data which corresponds to an integrated luminosity of 1.0 fb^{-1}.
- *re-reconstructed* data: $Z \to \mu\mu$ collision events are re-reconstructed using the same configurations as for the re-reconstruction of embedded events (Section 6.7.1). These samples are used to validate the reconstruction based on cells and tracks instead of hits.
- μ- or *τ-embedded MC* or *re-reconstructed* MC: Instead of $Z \to \mu\mu$ collision data, $Z \to \mu\mu$ MC serves as input for the production of the samples mentioned above. They can be directly compared to standard $Z \to \mu\mu$ or $Z \to \tau\tau$ MC and help to disentangle embedding specific effects from imperfections of MC simulations.
- *generator-seeded embedding*: Effects originating from the reconstruction of the input muons and final state radiation are studied by μ- or τ-embedded MC samples, in which the kinematics of the embedded objects is obtained from the momenta of generator-level muons, instead of reconstructed muons. Embedded MC events, in which the decay kinematics is based on reconstructed momenta, are also called *detector-seeded* below.

For the studies in this chapter the listed samples have to pass one of the following sets of criteria. In all cases, standard quality criteria described in Section 7.2.3 are applied to ensure a fully operational detector and well-reconstructed events.

- *$Z \to \mu\mu$ selection:*
Collision events are selected using a combined dimuon trigger with p_T thresholds of 18 GeV for the leading muon and 8 GeV for the sub-leading muon or a single-muon trigger ($p_T(\mu) > 24$ GeV). Only events with at least two good-quality muons are accepted if the leading (sub-leading) muon fulfils $p_T(\mu) > 20$ (15) GeV. Both muons must be isolated in the ID, which is ensured by requiring the scalar sum of other track transverse momenta in an isolation cone of size $\Delta R = 0.4$ to be smaller than 20% of the muon transverse momentum ($I(p_T, 0.4)/p_T(\mu) < 0.2$). Only events containing at least one such opposite-charge muon pair with an invariant mass $m_{\mu\mu} > 40\,\text{GeV}$ are considered.
- *$Z \to \tau\tau$ selection:*
To enhance the available statistics for validation purposes, $Z \to \tau\tau$ events only have to pass the $H \to \tau\tau$ analysis preselection described in Section 7.2.3.

6.3 Special properties

With regard to most analysis aspects, embedded events can be treated as common ATLAS data. However, there are some special properties and limitations of embedded events which need to be considered.

- The selection of $Z \to \mu\mu$ input events introduces a bias due to trigger and reconstruction efficiencies. In particular, these efficiencies have impact on the selection of low p_T leptons, which play an important role in the $H \to \tau_\ell\tau_\ell$ channel. In order to account for that, correction factors are derived and applied, as documented in Section 6.4.

- Since the embedding procedure is based on tracks and cells, the trigger response of the hybrid event cannot be simulated. Such a simulation would require hit level information. The missing trigger responds with respect to the analysis trigger selection needs to be evaluated by parametrising the measured trigger efficiency in data. They enter the analysis as additional correction factors for the embedded events.

- Although $Z \to \mu\mu$ events are selected with high purity, the selected input samples are still contaminated by other processes like $t\bar{t}$ or diboson events. To estimate the contamination the number of $t\bar{t}$ and diboson in the VBF and boosted signal region containing two truth τ leptons with $p_T(\tau_1)$ >20 GeV, $p_T(\tau_2)$ >15 GeV, $|\eta|^{\tau_1,\tau_2} < 2.5$ and invariant mass $m_{\tau\tau}$>40 GeV is determined. While 4.1% (11.9%) of diboson events in the VBF (boosted) signal region survive the truth τ selection, 6.3% (9.2%) of $t\bar{t}$ events pass the selection as well. To avoid double counting in the final background model, these events are vetoed.

- The kinematics of the embedded $Z \to \tau\tau$ decay are derived from reconstructed muons implying that the truth kinematics of the Z decay are folded with the resolution of the muon reconstruction. In addition, the final state radiation of the selected input muons affect the kinematics of the embedded objects as well. These effects are conceptually unavoidable and cannot be separated when embedding leptons in $Z \to \mu\mu$ data events. However, by using $Z \to \mu\mu$ MC events as input, both effects can be studied separately, as documented in Section 6.7. Those studies confirm, that the effects are small.

- Although most of the event topology is taken from data, the embedded τ lepton is still taken from MC simulation and consequently systematic uncertainties related to the MC description of τ decays as well as the corresponding detector response have to be considered.

- Potential systematic uncertainties related to the procedure itself have been studied and are also addressed in the analysis. Embedding specific uncertainties, which are considered in Run 1 analyses are summarised in Section 6.6.

- The re-reconstruction based on tracks instead of hits might bias the track reconstruction of hadronic τ decays in the presence of pileup. The corresponding systematic effect is discussed in Section 6.6.1.

- The number of faked τ leptons in τ embedded events differs from standard $Z \to \tau\tau$ MC. In only 1% of the cases, the selected light lepton and hadronically decaying τ lepton are not matched to a true τ lepton, while this is the case for 2.79% of $Z \to \tau\tau$ MC events. The misidentification rate is lower in embedded events due to preselection of high momentum muons and the usage of the kinematic filter, which ensures that the τ decay products survive the kinematic analysis thresholds.

This difference is compensated in the $H \to \tau\tau$ analysis by normalising embedded events to data in a dedicated region and by the fake background model.

- Depending on the use case the normalisation of embedded samples to data, MC or an absolute normalisation is preferable. For the SM $H \to \tau\tau$ analyses, embedded samples are normalised to data (see Section 7.3.1), while for searches for Higgs bosons of the *Minimal Supersymmetric Standard Model* (MSSM) [125] an absolute normalisation is preferable, because obtaining a signal free Z control region is not possible due to low mass signal samples. The possibility of normalising embedded events absolutely has been tested within the MSSM analysis [126]. While the normalisation to data has been proven to be reliable, for an absolute normalisation, several efficiencies needed to be accounted for: Starting from lepton universality (i.e. assuming the branching ratios of $Z \to \tau\tau$ and $Z \to \mu\mu$ decays to be equal), the branching ratios for the different τ decay channels need to be considered as well as the efficiency and purity of the $Z \to \mu\mu$ input selection. Further, one needs to estimate the cut into the $\tau\tau$ phase space due to $\mu\mu$ acceptance cuts. For example one has to consider the different isolation requirements for muons and τ leptons.
- The size of embedded samples is naturally limited by the number of available $Z \to \mu\mu$ events in collected data. In particular, in the high BDT bins the number of embedded events is a statistical limiting factor. The available sample size of embedded events can be enhanced by the usage of the kinematic filter (see Section 6.3.1).

6.3.1 Efficiencies of the embedding technique

The size of τ-embedded samples is limited by the number of input $Z \to \mu\mu$ MC or data events. The number is further reduced due to the selection of $Z \to \mu\mu$ events. Testing the skimming efficiency on $Z \to \mu\mu$ MC results in a loss of about 54% of $Z \to \mu\mu$ events, which are no longer available for validation purposes. For the embedding of $Z \to \mu\mu$ collision data, however, the skimming is necessary in order to reduce background contaminations. Further, events are rejected due to the analysis thresholds on the kinematics of embedded objects. The percentage of simulated physics objects, which are below

Truth visible object	Analysis threshold	Fraction of objects passing threshold
τ_{had}	$p_T(\tau_{\text{had}})$ >20 GeV	66.9±0.9%
Muon	$p_T(\mu)$ > 26 GeV	24.0±0.6 %
Electron	$p_T(e)$ > 26 GeV	36.2±1.0%
$\mu\tau_{\text{had}}$		16.1±1.1%
$e\tau_{\text{had}}$		24.2±1.3%

Table 6.2: Fraction of visible simulated τ decay products and embedded events passing the analysis p_T thresholds.

the analysis thresholds is given in Table 6.2, showing significant losses of embedded events only due to kinematics of random TAUOLA $Z \to \tau\tau$ decays. The kinematic filter helps to enhance the statistics of events passing analysis thresholds by requiring the simulated visible τ decay products to exceed $p_T(e)$ >18 GeV ($p_T(\mu)$ >15 GeV) and $p_T(\tau_{\text{had}})$ >15 GeV. The values are driven by the thresholds of the lepton-τ triggers EF_tau20Ti_medium1_e18vh_medium1[2] and EF_tau20_medium1_mu15. Events below these thresholds are forced to be regenerated by the filter, which increases the number of embedded events used within the analysis. Since events selected by the lepton-τ trigger are, however, not considered in the analysis, the kinematic filter could have been further optimised by adjusting the thresholds

[2] The naming convention is introduced in Section 7.2.

to the requirements of the single lepton trigger EF_MU24I_TIGHT andEF_E24VHI_MEDIUM1, respectively. The distribution of the transverse momentum of the truth muon after $H \to \tau_\ell \tau_h$ objects preselection (Figure 6.5) shows, that 47% of the events are below the single lepton trigger threshold and hence not used for analysis. Prospective studies showed that if the thresholds of the kinematic filter are increased to 24 GeV a gain in statistics by a factor of about 2.8 is achieved. In this thesis, however, the lower thresholds ($p_T(e)$ >18 GeV, $p_T(\mu)$ >15 GeV, $p_T(\tau_{\text{had}})$ >15 GeV) are used[3]. For future applications the thresholds of the kinematic filter should be revisited in order to ensure that the maximal possible statistics for τ embedded events is obtained.

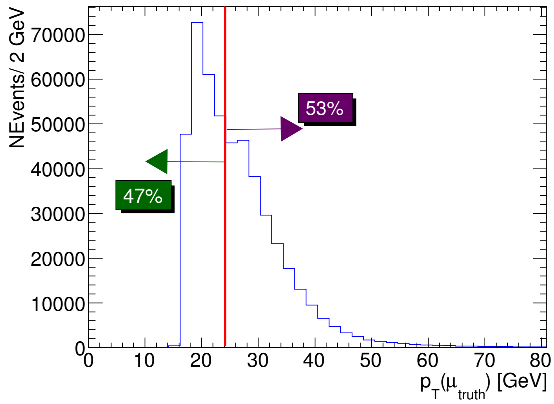


Figure 6.5: Transverse momentum of the truth muon after object preselection. The threshold of the muon trigger is indicated by the red line. All events below this threshold are not used for the analysis.

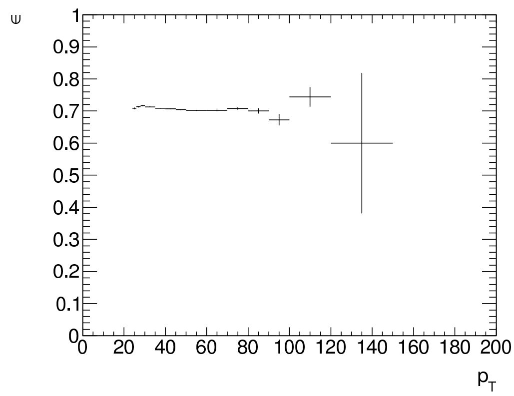
6.4 Embedding-specific corrections

Some of the above mentioned properties of embedded events require embedding-specific corrections. These corrections concern the bias due to the selection of $Z \to \mu\mu$ input events, effects due to the missing trigger information and polarisation effects. The following section explains how these correction factors are derived. Their impact on the shape of analysis relevant distributions is discussed in Section 6.8.

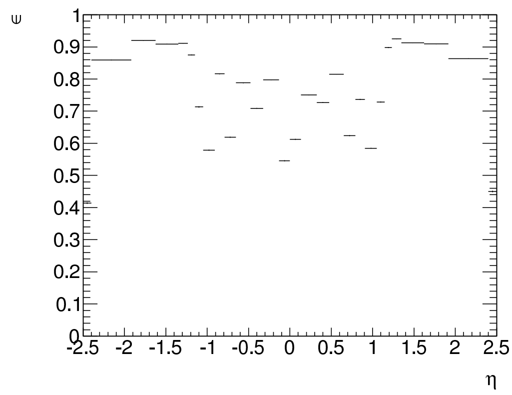
6.4.1 Muon selection efficiencies

The selection of $Z \to \mu\mu$ input events is subject to trigger and reconstruction efficiencies of input muons and hence biases the kinematics of the simulated $Z \to \tau\tau$ decays. To account for both effects, the embedded objects are weighted by the inverse product of the corresponding efficiencies. The trigger efficiencies for single and di-muon trigger are derived separately for different data taking periods in bins of η and muon p_T. An example for the single muon trigger efficiency is given in Figure 6.6. The muon reconstruction efficiency is a function of η and muon p_T. It is shown in Figure 6.7.

[3] By using lower filter thresholds more flexibility in the design of the $H \to \tau\tau$ analysis was obtained.

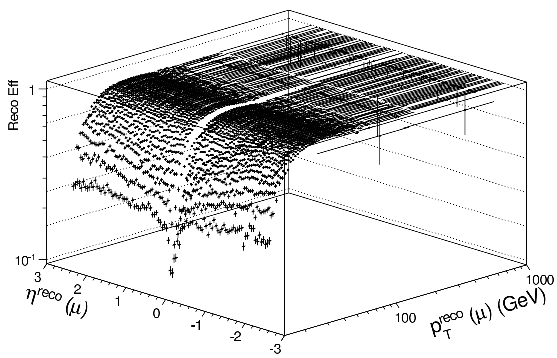


(a)



(b)

Figure 6.6: Efficiencies ϵ of the EF_MU24I_TIGHT trigger in period H-L as a function of transverse momentum p_T (left) and pseudorapidity η (right) of the muon. The figures are taken from [127].

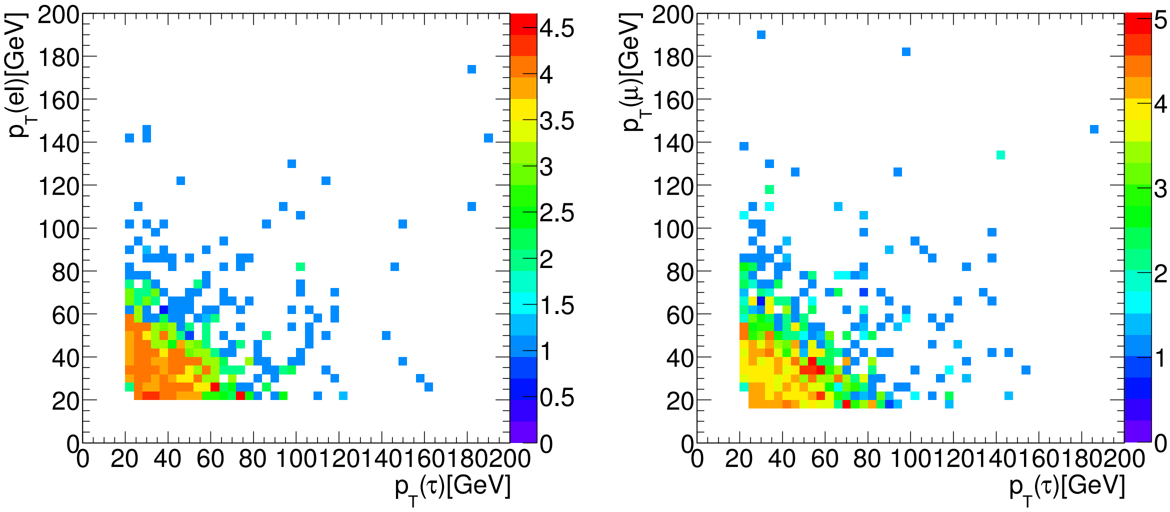


(a)

Figure 6.7: Reconstruction efficiencies of muons as a function of transverse momentum and pseudorapidity of the muon. The figure is taken from [127].

6.4.2 Trigger emulation correction

Since the trigger response of the hybrid event is not simulated, the trigger efficiencies are taken from respective trigger performance groups, which evaluate the efficiencies by parametrising the measured trigger efficiency in data. The remaining differences between emulation and regular offline trigger treatment are corrected for by another weight, which is obtained from standard ATLAS $Z \rightarrow \tau\tau$ MC. This weight is derived from shape comparisons between the transverse momentum of τ decay products (electron, muon and τ_{had}), using the offline trigger treatment and the trigger emulation. These weights, shown in Figure 6.8, are calculated for the $e\tau_{\text{had}}$ and $\mu\tau_{\text{had}}$ channel separately in bins of $p_T(\tau_{\text{had}})$ and $p_T(\ell)$. This additional correction accounts for imperfections of the trigger emulation like missing trigger-matching.

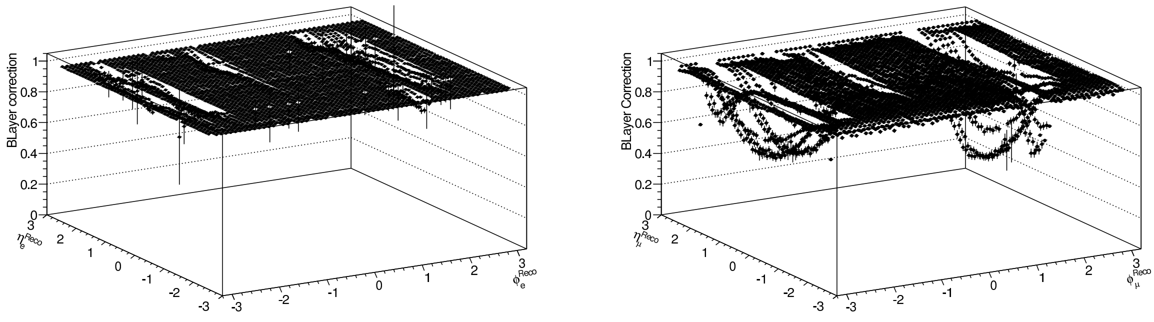


(a) (b)

Figure 6.8: Trigger emulation corrections as function of $p_T(\tau_{\text{had}})$ and $p_T(\ell)$ for $e\tau_{\text{had}}$ (left) and $\mu\tau_{\text{had}}$ (right) channel.

6.4.3 The B-layer correction

The B-layer correction is temporarily necessary due to incorrect information about the expected hit in the B-layer in the latest produced samples of $Z \to \tau\tau$ decays embedded in $Z \to \mu\mu$ collision data at $\sqrt{s}$ =8 TeV: Although the information, whether a B-layer hit is correctly simulated for the embedded muon and electron objects, the information is overwritten and always set to true during the re-reconstruction. Since this bug concerns ATLAS tools in a certain software release, this correction is not related to the embedding method itself and won't be necessary in future applications. For the $H \to \tau\tau$ analysis the incorrectly stored information results in inconsistencies to standard MC and collision data, in which certain B-layer requirements are only applied if a B-layer hit is indeed expected. For each embedded muon or electron object the corresponding B-layer correction weight is obtained from the map given in Figure. 6.9. It shows the reciprocal efficiency of requiring a B-layer hit, when none is expected, in bins of η and ϕ.



(a) (b)

Figure 6.9: B-layer correction for electrons (left) and muons (right) as a function of the η and ϕ of the electron and muon, respectively. The figures are taken from [127].

6.5 The TauSpinner tool

The polarisation of the τ leptons p_τ and Z boson influences the kinematics of τ decay products and hence needs to be correctly modelled. The TAUOLA program assigns the correct polarisation to τ leptons and spin correlations. The ratio between positive and negative polarised τ leptons, originating from Z/γ^* decays, depends on the probability

$$p_\tau^Z = \frac{\frac{\mathrm{d}\sigma}{\mathrm{d}\cos\theta}(s, \cos\theta, p_\tau = 1)}{\frac{\mathrm{d}\sigma}{\mathrm{d}\cos\theta}(s, \cos\theta, p_\tau = 1) + \frac{\mathrm{d}\sigma}{\mathrm{d}\cos\theta}(s, \cos\theta, p_\tau = -1)} \tag{6.2}$$

with

$$\frac{\mathrm{d}\sigma}{\mathrm{d}\cos\theta}(s, \cos\theta, p_\tau) = (1 + \cos^2\theta)F_0(s) + 2\cos\theta F_1(s) - p_\tau[(1 + \cos^2\theta)F_2(s) + 2\cos\theta F_3(s)] \tag{6.3}$$

where $\cos\theta$ is the scattering angle of τ leptons, s the squared center-of-mass energy of the hard process and $F_i(s)$ the form factors as function of the initial and final state couplings of fermions to the Z boson. To calculate the probability p_τ^Z, the flavour of the incoming partons must be known. Whereas common generators like PYTHIA provide this information, the configuration of incoming partons is unknown in embedded events. For that reason, TAUOLA assumes an average p_τ^Z, which is equal to $p_\tau^Z(\cos\theta = 0)$. However, the correct p_τ^Z can be stochastically restored by the TauSpinner tool [122–124] which determines the initial state configuration in four steps:

- Calculation of invariant Z/γ^* mass
- Calculation of scattering angle $\cos\theta$ in the $\tau\tau$ rest-frame
- Determination of fraction of momenta taken by partons
- Make probabilistic choice to assign flavors to the initial quarks and the sign of the scattering angle

Using this information, p_τ^Z is obtained as the weighted average over all possible initial quark configurations. The final output of the TauSpinner program is a spin weight, calculated for each event separately. The embedding is one of the main applications of the TauSpinner tool. But TauSpinner can also be used to remove tau spin effects from samples with generated spin effects and to change spin effects in samples, which are generated with a certain τ polarisation to other spin configurations, e.g. $Z/\gamma^* \to H$. To validate the TauSpinner tool two standard $Z \to \tau\tau$ MC samples are generated with and without spin effects and the shape of the truth visible $\tau\tau$ mass spectrum is compared with each other (Figure 6.10a). After applying TauSpinner (Figure 6.10b), both samples match each other. TauSpinner weights are then applied to the second sample and compared with the first one, see Figure 6.10.

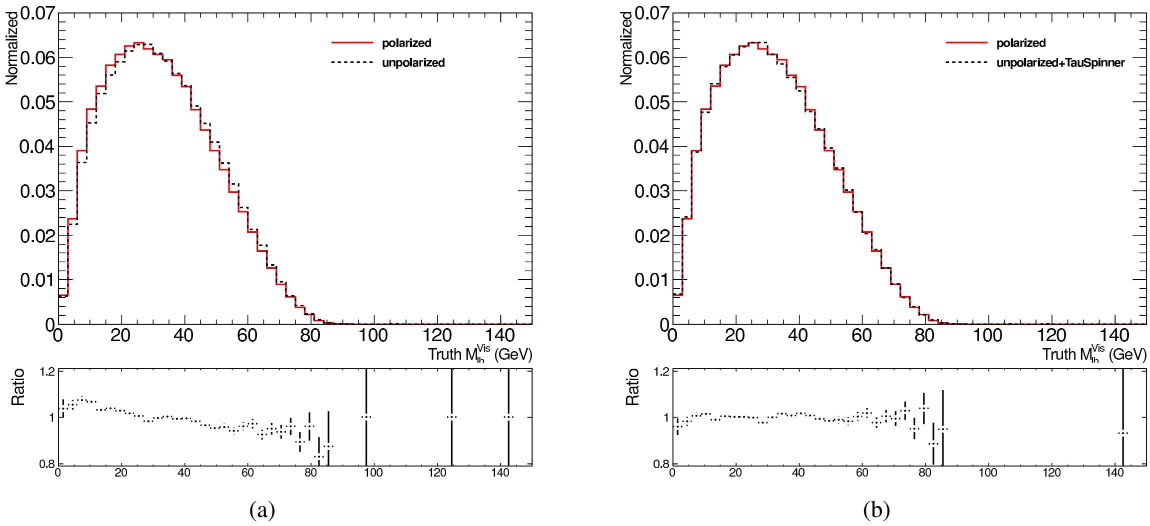
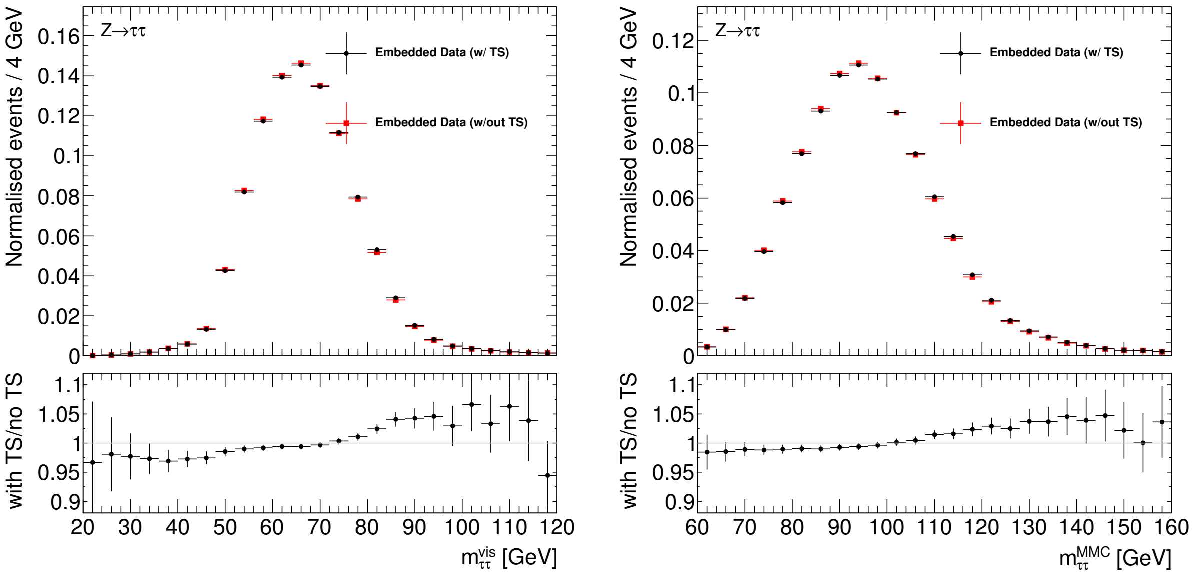


Figure 6.10: Distribution of the truth visible $\tau\tau$ mass in the $H \to \tau_\ell\tau_h$ channel for $\tau_h \to \pi\nu$. Left: Comparison of polarised (red) and unpolarised (black) $Z \to \tau\tau$ MC; Right: Comparison of polarised (red) and unpolarised $Z \to \tau\tau$ MC corrected with TauSpinner (black). The figures are taken from [127].

6.5.1 Influence of TauSpinner weights on $Z \to \tau\tau$ embedded events

The size and the direction of the TauSpinner correction in embedded events depend on the τ lepton decay mode and therefore on the three $H \to \tau\tau$ analysis channels. In all channels, TauSpinner corrections are small compared to other corrections. Comparing the impact of this correction between the channels, the largest effect is observed in the $H \to \tau_\ell\tau_\ell$ channel, in which the visible mass spectrum becomes harder after applying TauSpinner. The size of the correction increases for high masses and reaches values of about 10-15%. The direction of the correction changes in cases of the $H \to \tau_h\tau_h$ channel and reaches values of about 10%. The smallest polarisation effects are observed in the $H \to \tau_\ell\tau_h$ channel as shown in Figure 6.11a. In this case, effects due to hadronic and leptonic τ decays cancel. Since the visible mass spectrum in Figure 6.11a considers all hadronic τ_h decays and not only the decay of $\tau_h \to \pi\nu$ as in Figure 6.10a, the direction of the correction is switched. Also other quantities like the invariant $\tau\tau$ MMC mass in Figure 6.11b are only slightly influenced by spin weights, so that compared to other effects the polarisation is negligible in this particular channel.



(a) (b)

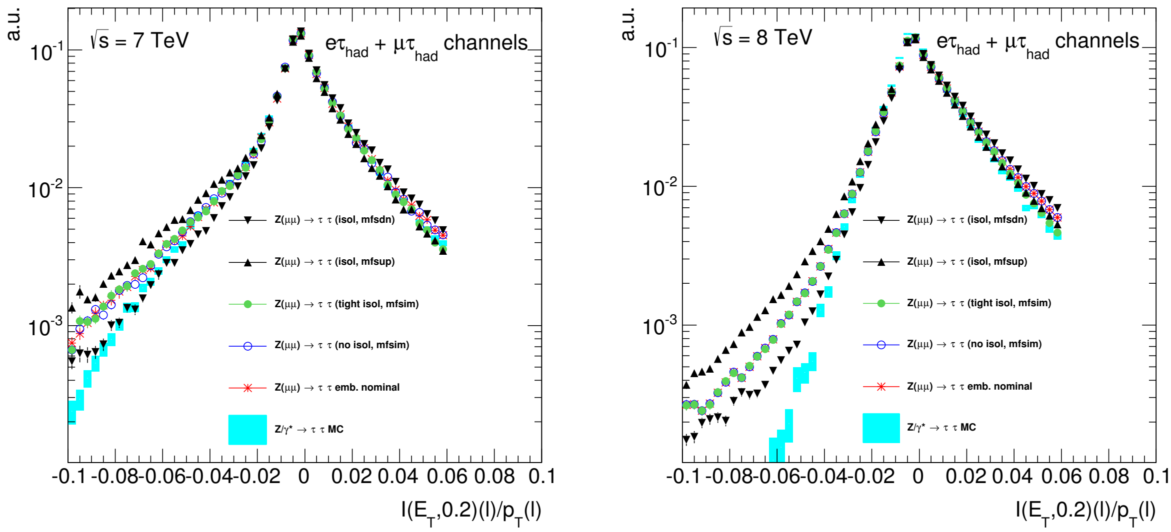
Figure 6.11: Comparison of τ-embedded events in the $H \to \tau_\ell \tau_h$ channel with (black) and without (red) TauSpinner weights applied. All possible (hadronic) τ-decay channels are considered: (a) visible $\tau\tau$ mass (b) invariant $\tau\tau$ MMC mass.

6.6 Systematic uncertainties

For the presented analysis in Section 7, two embedding-specific uncertainties have been identified:

- The isolation requirement, used to select $Z \to \mu\mu$ events, influences the environment of the embedded τ object. It has also impact on the background contamination from $\mu\mu$ final states with non-prompt muons. To estimate the effect the nominal isolation requirement $I(p_T, 0.2)/p_T(\mu) < 0.2$ is either tightened to $I(p_T, 0.4)/p_T(\mu) < 0.06$ and $I(E_T, 0.2)/p_T(\mu) < 0.04$ or removed completely.

- The cell energy, associated to the input muons, is removed by subtracting deposited energies in corresponding cells of simulated $Z \to \mu\mu$ decays with the same kinematics. This estimate results in potential large uncertainties, which are accounted for by varying the simulated energy deposition before subtraction by ±20% (±30%) for data recorded with $\sqrt{s}$ =8 TeV ($\sqrt{s}$ =7 TeV). The choice of the amount of the energy to be varied had to be made before the production of τ-embedded data events, used for the analysis, and was confirmed to be reasonable by opposing the lepton isolation in τ-embedded data events to standard $Z \to \tau\tau$ MC, e.g. as shown in Figure 6.12. The relative calorimeter isolation differs between τ-embedding and standard MC. However, varying the subtracted energy by ±30% as demonstrated in Figure 6.12a covers indeed remaining differences. Based on the good coverage due to the cell-related systematic uncertainty, the varied energy was reduced to ±20% for later data-taking periods at $\sqrt{s}$ =8 TeV (Figure 6.12b). The validity of the ad-hoc choice is also confirmed by comparisons based on τ-embedded $Z \to \mu\mu$ MC events. The latter comparison is more significant since it only involves pure MC events and hence potential effects due to differences between simulated cell energies and cells from collision data do not interfere.



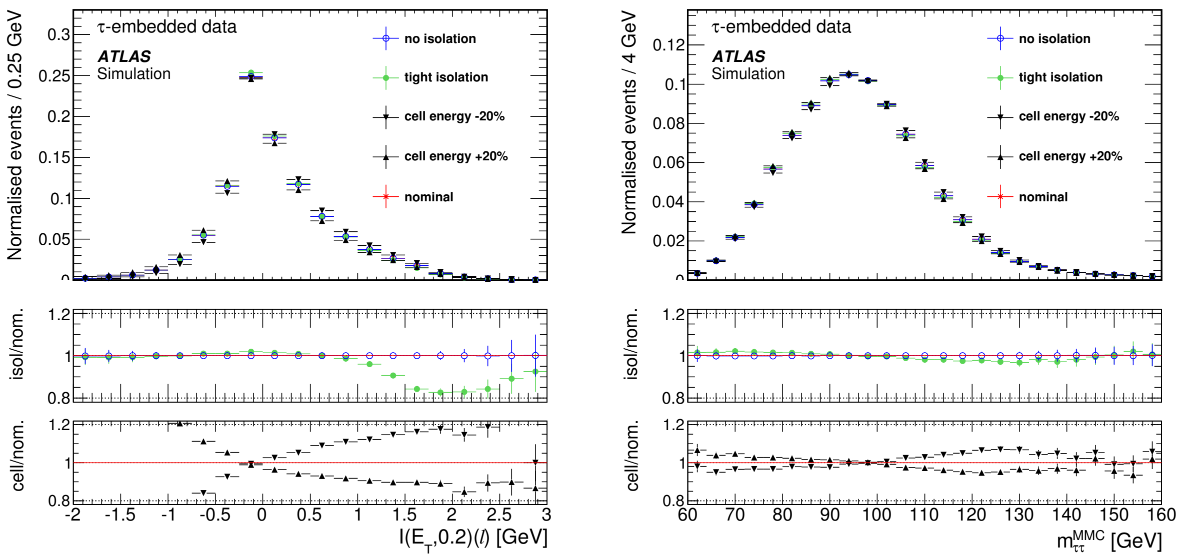
(a) (b)

Figure 6.12: Relative calorimeter isolation of standard $Z \to \tau\tau$ MC in light blue compared to nominal τ-embedded data events in red at $\sqrt{s}$ =7 TeV (left) and $\sqrt{s}$ =8 TeV (right). The triangles correspond to the upward/downward variation of the subtracted cell energy, the green (dark blue) circles to tight (no) isolation.

The described variations are produced in parallel for all embedded samples summarised in Section 6.2. The different datasets are used to validate the embedding procedure and the embedding-specific systematic uncertainties. Other potential sources of systematic uncertainties were studied and found to be small compared to the recent used systematics (see Section 6.6.1). In case of validation studies, the systematic variations are normalised to the default samples in order to absorb differences in selection efficiencies, e.g. due to tighter muon isolation requirements. The remaining shape uncertainties are symmetrised for each of the two systematic uncertainties to the larger variations, which particularly compensates for the non-symmetric isolation criteria. In Figure 6.13 the absolute calorimeter and $m_{\tau\tau}^{\text{MMC}}$ mass are shown, illustrating the influence of the embedding specific uncertainties on the shape of the nominal $Z \to \tau\tau$ embedded sample. In contrast to the cell subtraction systematic uncertainty the muon isolation has a small effect. In particular dropping the muon isolation requirement, which is done for the down variation, has no visible impact compared to the nominal. This is due to the rather strong isolation requirements used in the analysis ($I(p_T, 0.4)/p_T < 0.06$ and $I(E_T, 0.2)/p_T < 0.06$), which already reject most of the non isolated events.

6.6.1 Further sources of systematic uncertainties

Besides the two embedding-related uncertainties, described above, further possible sources have been investigated. These further uncertainties turned out to be negligable in the context of the $H \to \tau\tau$ analysis in Run 1. As discussed in Section 6.3, modifications of the input muon kinematics due to detector resolution or final-state radiation are unavoidable effects of the procedure and therefore could be considered as more fundamental sources of systematic effects. Although they do not directly enter, neither the muon isolation variation, nor the varied subtracted cell energy, their impact is expected to be correlated with these variations and indeed found to be small in comparison (see Section 6.7.3). Other potential sources of systematic uncertainties are the shower models used for the simulation of $Z \to \tau\tau$ decays or the composition of one and three prong τ decays as discussed below.

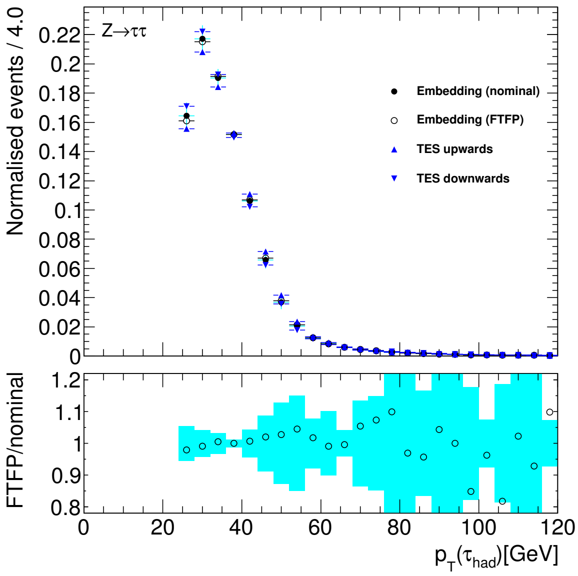


(a) (b)

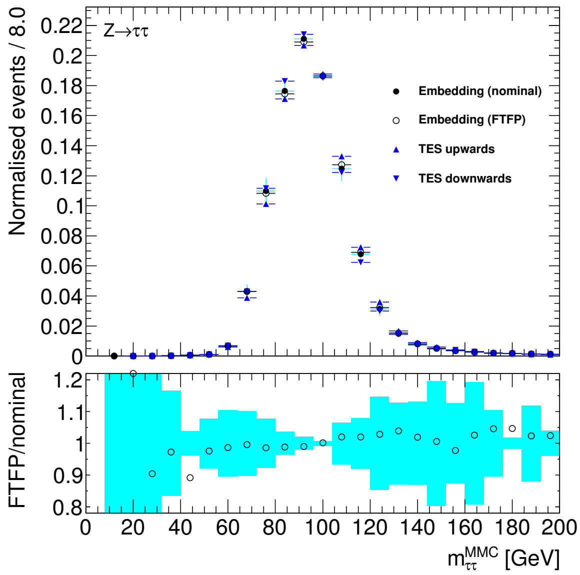
Figure 6.13: Effect of systematic variations on (a) the calorimeter isolation of the selected light lepton and on (b) the $\tau\tau$ invariant MMC mass. The lower ratio plots display the effect of no (tight) isolation in blue (green) as well as the effect of scaling the subtracted cell energy by +20% (−20%) presented as upward (downward) pointing triangles. The red lines illustrate the nominal embedded sample. The figures are taken from [9].

Shower model

For the simulation of physics processes in the ATLAS detector, GEANT4 offers a variety of different models, describing the interaction of particles with the detector material. GEANT4 summarises consistent physics models for all particles and different energy ranges in so-called physics lists. The user can choose between those lists and -if necessary- modify them, in order to balance the detail of modelled physics processes or the modelled description and necessary CPU time. A detailed description of various physics list is documented in [128]. Suitable models for operating at high energies are the *Quark-Gluon String (QGS)* and the *fragmentation model (FTF)* [129]. They are combined with the *Precompound model (P)* operating at low energies to de-excite nucleons. For primary protons, neutrons, pions and kaons below roughly 10 GeV the *Bertini cascade model (BERT)* is used.
The combined physics lists QGSP_BERT and FTFP_BERT show a good performance in the ATLAS and CMS experiment and are chosen for those as default. For the embedding QGSP_BERT is used. The choice of the physics list for the simulation of the $Z \to \tau\tau$ decay is another possible source of systematic uncertainties. In order to estimate its size, another τ-embedded sample produced with FTFP_BERT is compared to the default one. As examples, the distribution of the transverse momentum of the hadronically decaying τ lepton as well as the $\tau\tau$ invariant mass are shown in Figure 6.14. The chosen physics list has only a small impact on τ-embedded events. The tiny effect is clearly covered by the tau energy scale uncertainty, which already includes uncertainties on the modelling of true τ leptons and hence is correlated to the shower model.
Since the tau energy scale is considered as systematic uncertainty for τ embedded events and covers shape differences due to different shower models, no additional systematic uncertainty for the shower model is assigned to the embedding method.



(a)



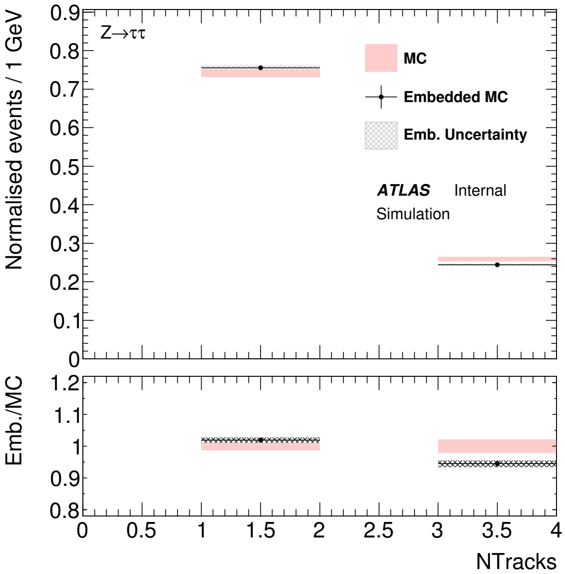
(b)

Figure 6.14: Comparison of nominal τ-embedded events and τ-embedded events simulated with the alternative physics list FTFP. The up and down variation of the tau energy scale is shown in dark blue. The blue error band in the ratio refers to the symmetrised up and downward variation of the nominal embedded sample with respect to the tau energy scale.

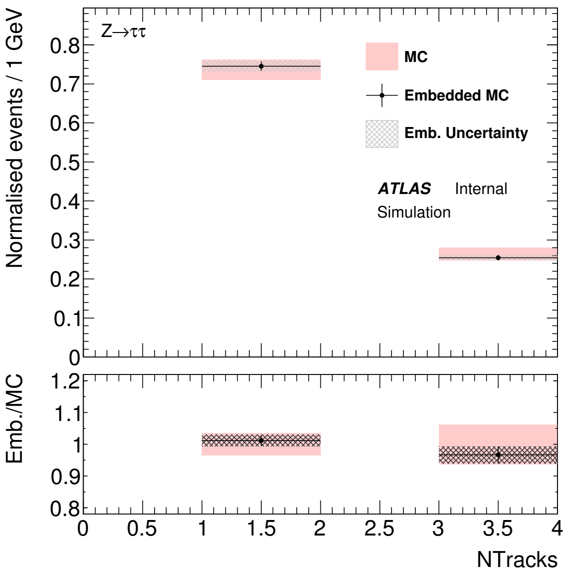
One prong/three prong composition

Another potential source of embedding-related uncertainties arises from the usage of reconstructed tracks from the simulated $Z \to \tau\tau$ mini event. These tracks are not re-reconstructed in the final embedding step opposed to the common ATLAS reconstruction, in which all objects of the full event are reconstructed from hits. In the presence of pileup the track reconstruction might be biased. Although it is assumed, that there is no impact on the larger fraction of events with one prong τ_{had} candidates, pileup might still influence the reconstruction of three prong τ_{had} candidates.
To estimate the size of the effect, in Figure 6.15 the distribution of the number of τ_{had} tracks in τ-embedded $Z \to \mu\mu$ MC events is compared to standard ATLAS $Z \to \tau\tau$ MC in the boosted and VBF signal region of the $H \to \tau_\ell\tau_h$ analysis. Small differences in the composition of one and three prong decays are observed, which are not quite covered by embedding-related systematics. In particular, less three prong τ decays are observed in τ embedding than expected from standard $Z \to \tau\tau$ MC. However, the corresponding control plots for the Z enriched control region in Figure 6.16 demonstrate that the composition is compatible with data within total uncertainties. The level of agreement to data is similar for τ embedded and standard $Z \to \tau\tau$ MC.
Despite of small visible effects it is very unlikely that they are relevant for the $H \to \tau_\ell\tau_h$ analysis and therefore would require to address an additional systematic uncertainty within the analysis. First of all, the analysis categories do not separate between one and three prong τ decays. Secondly, any overall mismodelling of efficiencies would be compensated by normalisation procedures[4]. Possible second order effects might arise, if the one/three prong composition biases other τ-related quantities, e.g. fake estimate or resolution: In order to test if the BDT shape used for signal extraction is influenced by the

[4] Embedded events are normalised to data in dedicated region, see Section 7.3.1.

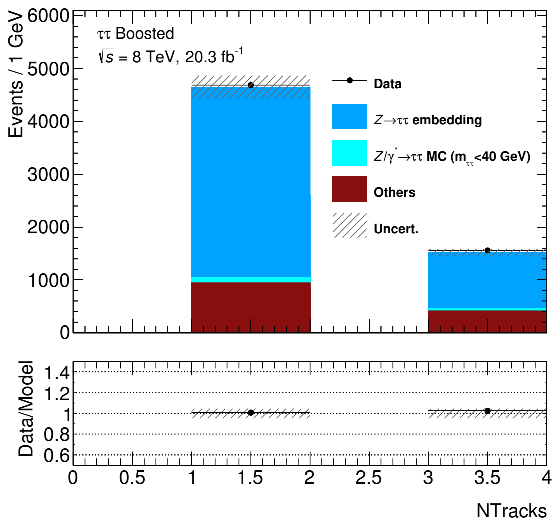


(a)

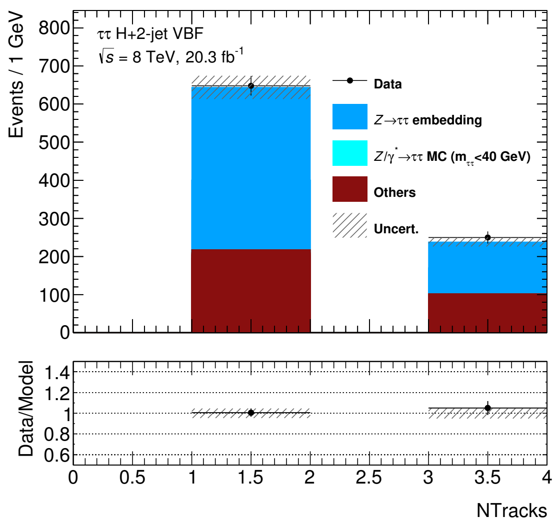


(b)

Figure 6.15: Comparison of τ-embedded $Z \to \mu\mu$ MC events (black) with $Z \to \tau\tau$ MC events (red) with respect to the ratio between one and three prong τ decays in the (a) boosted (b) VBF signal region of the $H \to \tau_\ell\tau_h$ analysis. The corresponding ratio plots below show the relative difference between the τ-embedded and $Z \to \tau\tau$ MC events. The embedding-related uncertainties are shown in grey.



(a)



(b)

Figure 6.16: Comparison of data with the combined background model for the one/three prong track composition in the (a) boosted and (b) VBF Z-enriched control region. The ratios show the relative difference of the data to the total background estimate. Except for the $Z \to \tau\tau$ background, all other background processes are summarised and are estimated as described in Section 7.3.

one/three prong ratio, the distributions of τ embedded events in Figure 6.15 are reweighted by 10% to match the prediction from standard ATLAS $Z \to \tau\tau$ MC. The reweighting is then applied to all BDT input variables and compared to the shapes of these distributions without reweighting. No significant impact is found.
To evaluate the effect further the complete analysis was repeated with reweighted embedded samples. With respect to the embedding normalisation no visible effect is observed. The same holds for the final signal strength, on which the impact of the one/three prong composition is less than <0.01% [130]. Similar checks for the $H \to \tau_h\tau_h$ analysis show, that the impact of the one-to-three prong ratio is below the pruning threshold for systematics [131]. Hence, indirect systematic effects due to the composition are negligible and no additional systematic is assigned.

6.7 Validation

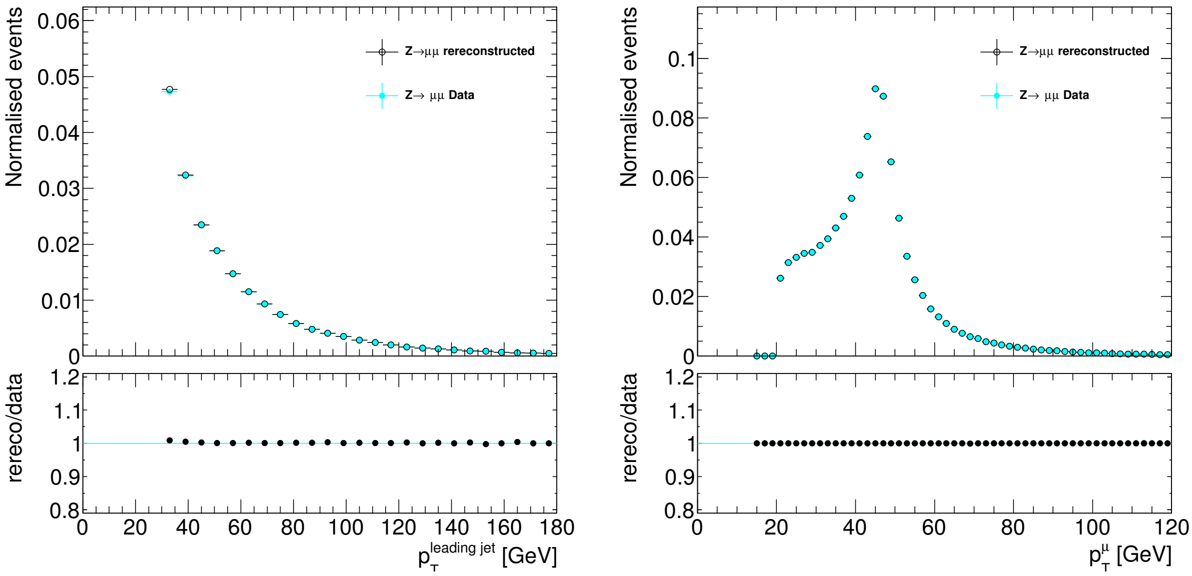
Since a precise $Z \to \tau\tau$ modelling is of high importance for the $H \to \tau\tau$ analysis, the embedding procedure has to be validated carefully. Various validation approaches have been developed in the context of this thesis, which are based on different combinations of the event samples, documented in Section 6.2. Each of them is suitable to validate particular aspects of the procedure and hence is an important part of the validation. In total their results give confidence that the method works properly. Nonetheless, some of them also show room for improvements for Run 2, e.g. tuning of the cell subtraction.
In this section the different validation methods are introduced. All distributions are normalised to unit area unless stated otherwise.

6.7.1 $Z \to \mu\mu$ based validation

Most parts of the embedding method can be validated without the embedding of more complex physics objects like electrons or hadronically decaying τ leptons. For example, testing for biases introduced by the configurations used to re-reconstruct events from tracks and cells or the removal and replacements of cells and tracks. Hence, comparisons of $Z \to \mu\mu$ events before and after running parts of the embedding algorithm provide a powerful validation. Since input and output distributions are obtained from the same events, they can be directly compared without considering caveats and external corrections listed in Section 6.3 and 6.4. Due to the correlation the statistical error is not displayed in the following ratio plots.
The validation described below is mainly based on $Z \to \mu\mu$ collision data. However, the same cross checks were also carried out using $Z \to \mu\mu$ MC as input for the re-reconstruction and μ-embedded events. The parallel investigation ensures that neither MC specific nor data specific effects biases the interpretation of the obtained results.

Pure re-reconstruction of $Z \to \mu\mu$ events

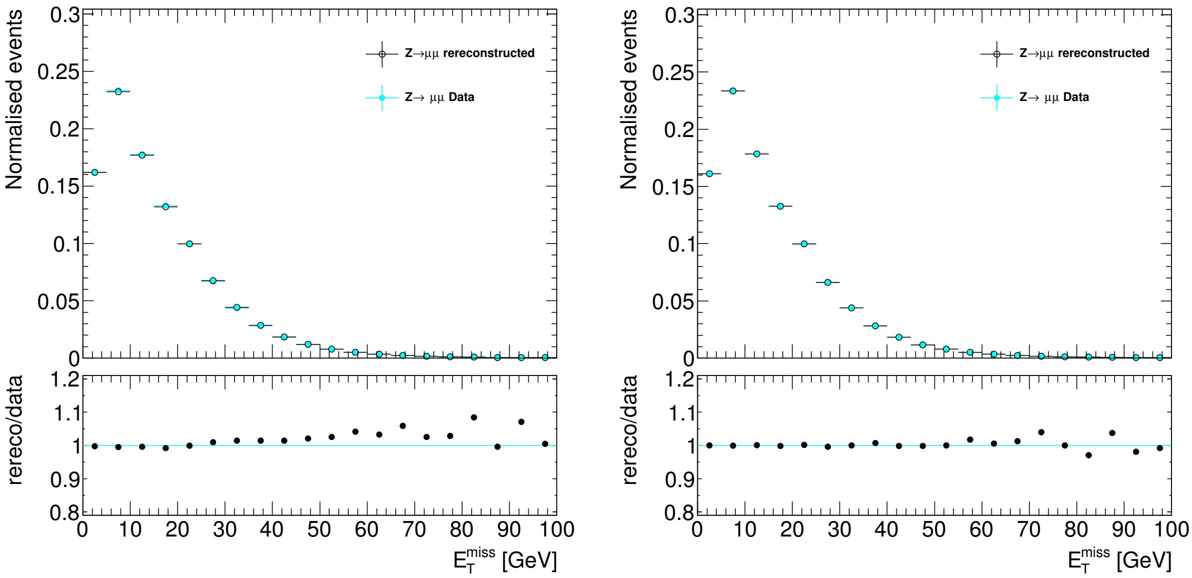
Comparisons before and after pure re-reconstructions of $Z \to \mu\mu$ input events might be considered as a trivial test. But the embedding is the only current use case for reconstruction based on cells and tracks instead of hits. Thus, reconstruction based on ESDs is neither officially supported, nor centrally validated within ATLAS. In the past, the ESD-based reconstruction was frequently broken by low-level changes in the reconstruction algorithms, underlying configurations or data formats. Hence, the reconstruction configurations in the embedding method have to be validated.
Distributions before and after re-reconstruction are not expected to be identical (even for a correctly configured reconstruction), because of the compactification of the calorimeter information within the



(a)

(b)

Figure 6.17: Comparison of $Z \rightarrow \mu\mu$ data events before (blue) and after (black) re-reconstruction in terms of transverse momentum of (a) the leading jet and (b) the leading muon. The corresponding ratio plots show the relative difference of the distributions after re-reconstruction indicated by the black points.

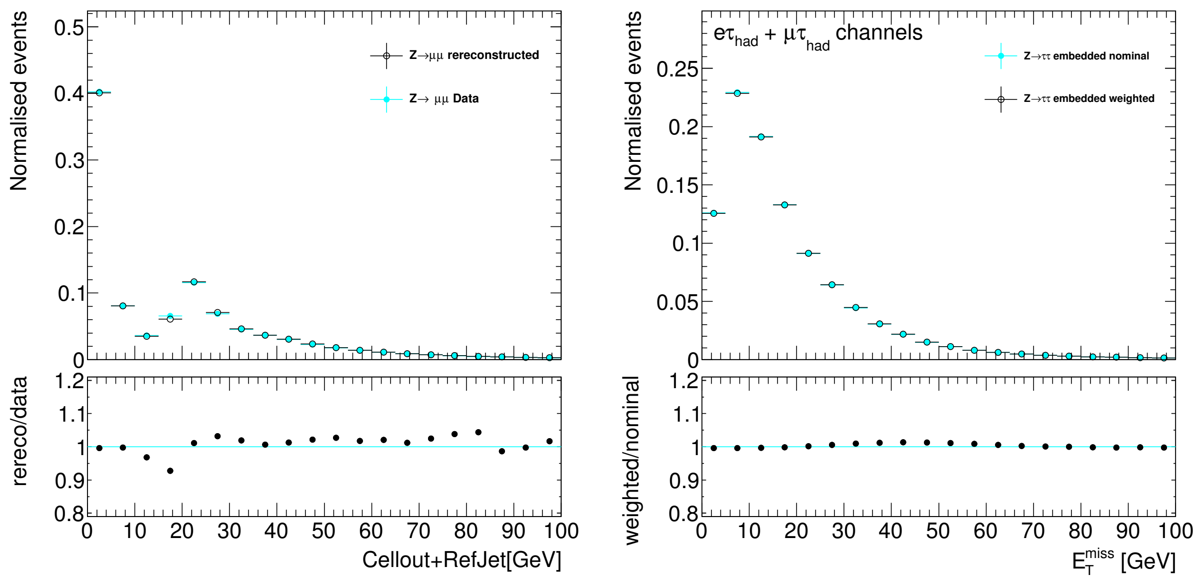


(a)

(b)

Figure 6.18: Distribution of the missing transverse energy in $Z \rightarrow \mu\mu$ events before (blue) and after (black) re-reconstruction: The re-reconstruction runs in the software release which is (a) dedicated for the production of embedded samples and (b) also used to reconstruct the input $Z \rightarrow \mu\mu$ events. The corresponding ratio plots show the relative difference of the distributions after re-reconstruction indicated by the black points.

ESD files [132], which is a very small effect. Indeed, when comparing distributions for a variety of quantities of $Z \to \mu\mu$ MC and ATLAS data before and after re-reconstruction, no significant differences are observed, as indicated by the examples in Figure 6.17. However, there is a hint of larger tails in the



(a) (b)

Figure 6.19: Left: Vectorial sum of jet and cell out component of E_T^{miss} before (blue) and after (black) re-reconstruction of $Z \to \mu\mu$ events. Right: E_T^{miss} in τ-embedded events with (black) and without (blue) E_T^{miss} reweighting.

distribution of missing transverse energy after re-reconstruction of both data and MC due to inconsistent reconstruction releases of $Z \to \mu\mu$ data.

The various ATLAS data and MC $Z \to \mu\mu$ events in ESD format used as input for the embedding technique are reconstructed in different releases. Ideally, the re-reconstruction in the embedding setup would run in the same software release as used for the different $Z \to \mu\mu$ input samples beforehand. However, due to technical reasons, the various embedding steps, i.e. the simulation and reconstruction of $Z \to \tau\tau$ decays as well as the re-reconstruction of hybrid events, have to run all in one software release dedicated for the $\sqrt{s}$ =8 TeV embedding production.

In order to confirm that the effect in the high-E_T^{miss} range originates from inconsistent production releases the $Z \to \mu\mu$ events are re-reconstructed based on tracks and cells in the software release, which was used before to reconstruct the input $Z \to \mu\mu$ data starting from hits (see Figure 6.18b). A good agreement of the E_T^{miss} distribution before and after re-reconstruction is achieved. A closer study of the E_T^{miss} components shows that the jet and cellout component are most affected by inconsistent releases. The distribution of the vectorial sum of jet and cellout term is plotted in Fig 6.19a before and after re-reconstruction. The distribution is plotted only for events, in which the jet component of E_T^{miss} does not vanish, because events, which have no jet component in E_T^{miss}, do not reveal release inconsistencies. The small effect of inconsistent releases in re-reconstruction cannot be avoided without reducing the efficiency of the production setup[5]. For that reason, it is of interest to study the impact of the differ-

[5] The current production setup is discussed in Appendix A, in which the consequences of different releases for different embedding steps are also discussed.

ence between re-reconstruction and standard ATLAS reconstruction (Figure 6.18a) on the τ embedded samples, used to model the $Z \to \tau\tau$ background in the $H \to \tau\tau$ analysis. Therefore weights are derived by a linear fit of the ratio plot of Figure 6.19a. The obtained weights $\mathrm{w}_{E_T^{\mathrm{miss}}}$ are:

- for RefJet+CellOut<20 GeV: $w_{E_T^{\mathrm{miss}}} = 1-0.04/10\,\mathrm{GeV}\cdot(\mathrm{RefJet+CellOut})$
- for 20 GeV:<RefJet+CellOut <80 GeV: $w_{E_T^{\mathrm{miss}}} = 1+0.04/60\,\mathrm{GeV}\cdot(\mathrm{RefJet+CellOut})$

These weights are then applied to the τ-embedded samples with a non-zero jet component in E_T^{miss} after analysis preselection. Finally, the reweighted missing transverse energy is compared to the nominal one. Figure 6.19b demonstrates that the reweighting has no significant impact and thus the reconstruction of missing transverse energy based on different re-reconstruction releases does not effect the analysis.

Embedding of $Z \to \mu\mu$ events in $Z \to \mu\mu$ data

Muon-embedded samples refer to events, in which the original muons are removed and replaced with muons from simulated $Z \to \mu\mu$ decays with identical kinematics. The muon embedding was already part of the embedding validation in earlier analysis [112]. However, further developments improved this validation approach significantly.

- The former muon embedding (Figure 6.20a) uses one simulated $Z \to \mu\mu$ decay not only to subtract the cell energy of the original muons but also to embed these simulated muons in the original event. This method is only conditionally suitable, because the same amount of energy is added to each cell as is subtracted from it beforehand. Hence, this former method can only be used to test if the cell subtraction algorithm is correctly implemented. An approach, which is closer to the actual τ-embedding procedure is preferable. The recent muon embedding (Figure 6.20b) makes use of two simulated $Z \to \mu\mu$ decays, which are both based on the same kinematics as the input $Z \to \mu\mu$ event. Whereas the first simulated decay is used to subtract the cell energies from the original event, the muons of the second decay are embedded. Thus, this method is sensitive to fluctuations in the estimation of the calorimeter energy of the input muons and implies cases, where the energy subtracted from input data is larger than the actually deposited energy in the input data cells.

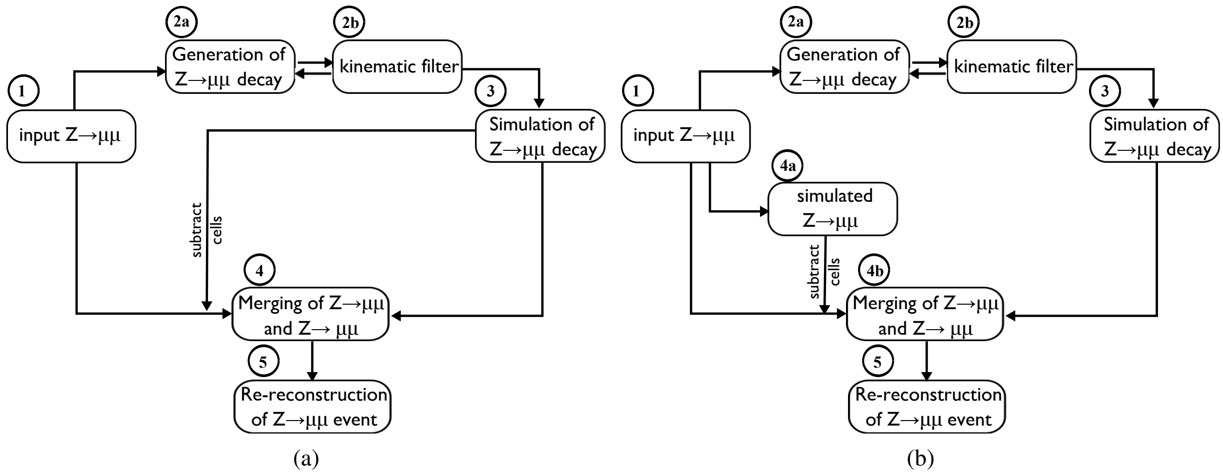


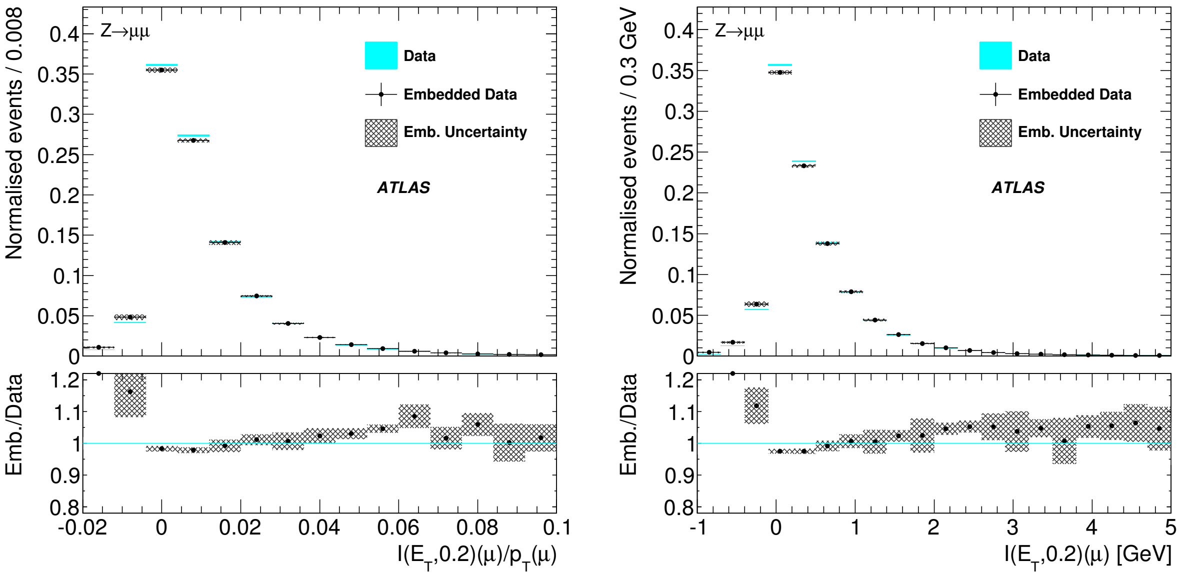
Figure 6.20: Flowchart of the old (left) and improved (right) muon embedding procedure.

- The embedding specific systematic uncertainties are taken into account within the validation of muon embedded samples.
- The muon embedding is repeated using $Z \to \mu\mu$ MC in order to validate the MC embedding, which serves as final closure test (see Section 6.7.3). Moreover, μ-embedded MC allows to study muon resolution and radiation effects. These effects are unavoidable in data embedding. In addition, comparing results of μ-embedded data with μ-embedded MC can help to clarify whether visible effects which might occur originate from merging data with MC instead of MC with MC.
- The direct comparison of those muon-embedded events with re-reconstructed instead of input $Z \to \mu\mu$ events exclude potential re-reconstruction effects, as discussed earlier.

Muon embedded events allow to investigate several aspects of the embedding technique:
The relative $(I(E_T, 0.2)/p_T)$ and absolute $(I(E_T, 0.2))$ muon calorimeter isolation are appropriate observables to study possible distortions of the detector response close to the input muons. They are displayed for μ-embedded data in Figure 6.21. The isolation systematic uncertainty is not shown in the error bands, because it is obtained by varying an explicit cut on the relative calorimeter isolation and thus is not well defined in these specific comparisons. For example in Figure 6.21a, the threshold for tight isolated muons is given by $I(E_T, 0.2)/p_T > 0.4$. Some of the embedded events with previous tight isolated muons are situated above this thresholds, because the embedding technique has a slight impact on the muon isolation. As consequence, input muons, which were just below this threshold, can result in embedded muons above it. In general, the isolation systematic becomes very large above this cut value. The observed changes in the distributions show fluctuations in the estimated calorimeter energy, which are not fully covered. This mainly concerns negative isolation values, which are far away from standard isolation requirements used for the $H \to \tau\tau$ analysis and the region above $I(E_T, 0.2)/p_T > 0.4$, where the isolation uncertainty drastically increases.
The cell energy $\mathrm{E}_{\mathrm{cell}}^{\mathrm{final}}$ after removing the energy deposited by the muons is given by

$$E_{\mathrm{cell}}^{\mathrm{final}} = E_{\mathrm{cell}}^{\mathrm{initial}} - k_\mu \cdot E_{\mathrm{cell}}^{\mathrm{simulated}} \tag{6.4}$$

with $E_{\mathrm{cell}}^{\mathrm{initial}}$ being the cell energy in the original event, $E_{\mathrm{cell}}^{\mathrm{simulated}}$ the energy in the simulated $Z \to \mu\mu$ event and k_μ the factor by which the subtracted energy is scaled. The scale factor k_μ is assumed to be equal to one and is varied by ±20% in the corresponding systematic variation. However, the distributions of relative and absolute calorimeter isolation suggest that the default value differs from 1. So, they can be used to optimise the default value k_μ in future applications.
Jets and their kinematics are not manipulated by the method and should remain unaffected, e.g. as illustrated for the transverse momentum of the leading jet and dijet mass in Figure 6.22.
For quantities directly related to the muon four-momenta most changes are found to be within the uncertainties: The p_T spectrum of the leading muon in Figure 6.23a shows some slight differences in its peak, but these differences are covered by isolation uncertainties. The same holds for the distribution of the transverse momentum of the Z boson (Figure 6.23b), in which small differences are observed at the low end of the spectrum. Larger effects are found in particular for the dimuon invariant mass shown in Figure 6.23c. Here, the largest deviations are observed at the Z mass, although some smaller differences are also visible in the region of low mass. The impression of large uncertainties in the tail originates from shape differences between tight isolation variation and nominal events. They are larger in the low mass region, because pileup can contribute an almost constant term to the cell energy deposits around the muons and the isolation efficiency depends e.g. on $p_T(\mu)$. The symmetrisation of the uncertainties further increases the impression of large uncertainties. Although not directly obvious,



(a) (b)

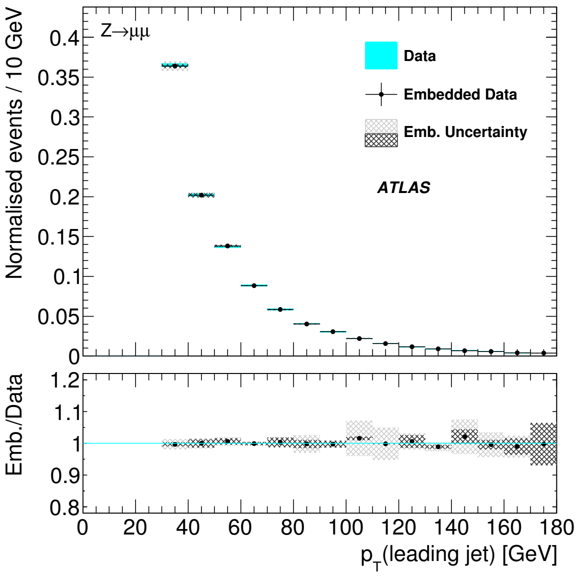
Figure 6.21: Comparison of $Z \to \mu\mu$ ATLAS data events before (blue) and after (black) μ-embedding in terms of (a) the relative and (b) the absolute calorimeter isolation in a cone of $\Delta R = 0.2$. The corresponding ratio plots below show the relative difference of the distributions after μ-embedding. The grey error band corresponds to the cell systematic uncertainties of the μ-embedded events. The figures are taken from [9].

global observables like E_T^{miss} also depend on the muon four-momentum, because muons also enter the E_T^{miss} calculation and hence cause small differences, see Figure 6.23d.

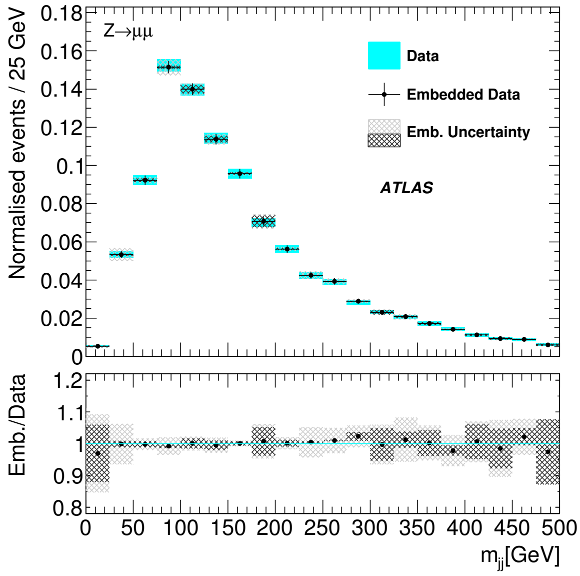
All these differences are expected, because the kinematics of the embedded $Z \to \mu\mu$ events are based on reconstructed input muons. As already mentioned in Section 6.3, the reconstructed muon momenta are used to seed the true momenta of the embedded muon objects. Afterwards the reconstruction is run again for the embedded hybrid event. Furthermore, it is impossible to access reconstructed muons before they radiate, which results in a softer p_T spectrum of the input muons. For that reason, they are certainly modified by the detector resolution as well as muon final state radiation.

The source of these effects can be confirmed by using generator seeded embedded samples, in which simulated $Z \to \mu\mu$ events serve as input and the kinematics of the embedded (muon-) objects are derived from generator level muon momenta, instead of the reconstructed information. This allows to remove muon reconstruction and final state radiation effects from input muons, which indeed improves the agreement in the muon-related distributions as shown in Figure 6.24: The small differences observed before in the distributions of $p_T(\mu)$, $p_T(Z)$ and E_T^{miss} in Figure 6.23 vanish. Also the agreement in the dimuon invariant mass is improved.

Although these generator-level based studies can confirm that the observed differences result from reconstruction and final state radiation effects, both effects unavoidably enter the embedding of ATLAS data events. For the eventual application of τ-embedding samples, those studies have to be repeated using generator-seeded τ-embedding events. As shown later in Section 6.7.3, in contrast to generator-seeded μ-embedding, these effects are much smaller and covered by already existing embedding-related uncertainties. Therefore, they can be neglected.



(a)



(b)

Figure 6.22: Comparison of $Z \to \mu\mu$ ATLAS data events before (blue) and after (black) μ-embedding in terms of (a) the transverse momentum of the leading jet p_T(leading jet) and (b) the dijet mass m_{jj}. The corresponding ratio plots below show the relative difference of the distributions after μ-embedding. The light (dark) grey error band corresponds to the sum in quadrature of cell and isolation (cell only) systematic uncertainties of the μ-embedded events. The figure 6.22a is taken from [9].

6.7.2 Embedding of $Z/\gamma^* \to ee$ decays in $Z \to \mu\mu$ events

The next more complex objects muons can be replaced with are electrons. The embedding of $Z/\gamma^* \to ee$ decays in $Z \to \mu\mu$ events allows stability checks of the missing transverse momentum, since it only involves prompt electrons in contrast to secondary electrons or muons from $Z \to \tau\tau$ decays. Further, in contrast to embedded $Z \to \mu\mu$ decays, it involves the embedding of calorimeter based objects, which can be used to study the manipulation of cell energies. Samples with embedded $Z/\gamma^* \to ee$ decays are also suitable to verify external embedding corrections related to the muon trigger and reconstruction efficiencies (see Section 6.4.1). The $Z/\gamma^* \to ee$ based approach has been recently developed [133] and probably will be part of future validations of embedded samples.

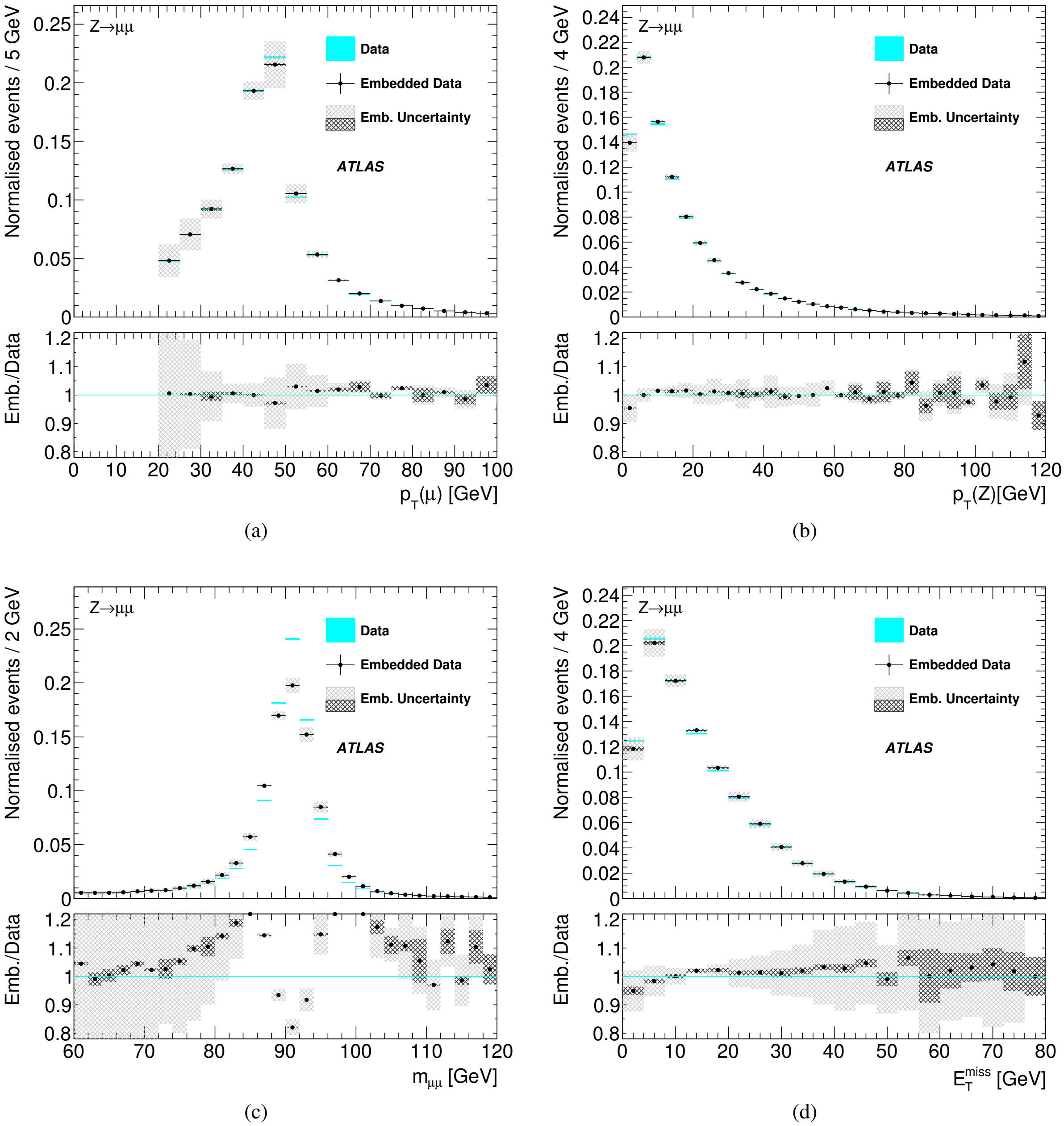
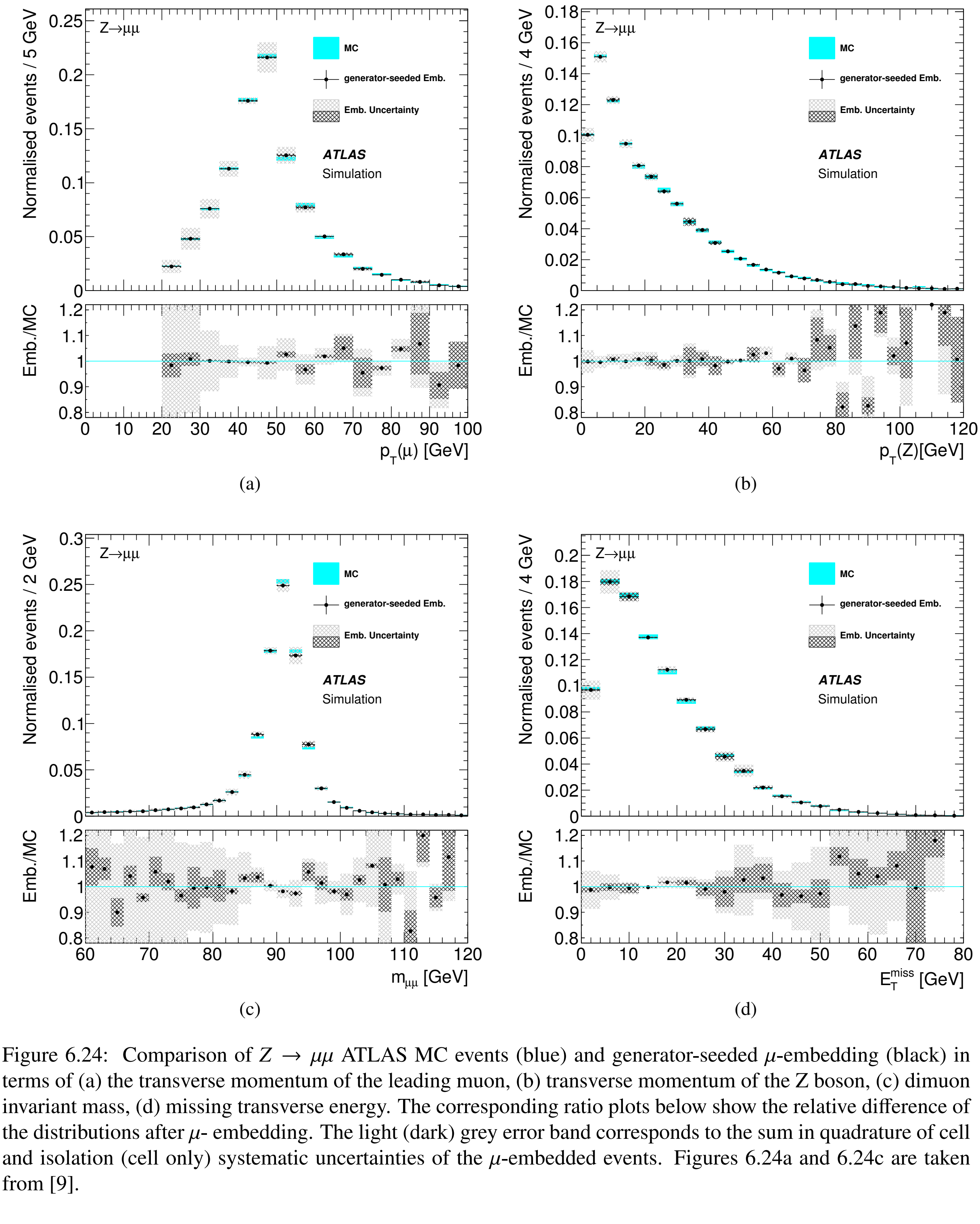


Figure 6.23: Comparison of $Z \to \mu\mu$ ATLAS data events before (blue) and after (black) μ-embedding in terms of (a) the transverse momentum of the leading muon, (b) transverse momentum of the Z boson, (c) dimuon invariant mass, (d) missing transverse energy. The corresponding ratio plots below show the relative difference of the distributions after μ- embedding. The light (dark) grey error band corresponds to the sum in quadrature of cell and isolation (cell only) systematic uncertainties of the μ-embedded events. The figures 6.23a and 6.23c are taken from [9].



(a) (b) (c) (d)

Figure 6.24: Comparison of $Z \to \mu\mu$ ATLAS MC events (blue) and generator-seeded μ-embedding (black) in terms of (a) the transverse momentum of the leading muon, (b) transverse momentum of the Z boson, (c) dimuon invariant mass, (d) missing transverse energy. The corresponding ratio plots below show the relative difference of the distributions after μ- embedding. The light (dark) grey error band corresponds to the sum in quadrature of cell and isolation (cell only) systematic uncertainties of the μ-embedded events. Figures 6.24a and 6.24c are taken from [9].

6.7.3 $Z \to \tau\tau$ based validation

The $Z \to \mu\mu$ based validation already addresses most aspects of the embedding technique and its results show no indication of any significant intrinsic biases introduced by the manipulation of the input event. More direct comparisons of $\tau\tau$ final states are desirable in order to complementary validate the modelling of the $Z \to \tau\tau$ background by τ-embedded samples. It would be preferable to compare τ-embedded samples directly to $Z \to \tau\tau$ samples from collision data. However, it is impossible to obtain such a reference sample signal-free and with sufficient purity. A comparison of the τ-embedded ATLAS data with standard $Z \to \tau\tau$ is not a conceptually clean approach either, because in that case embedding-specific effects due to event manipulation cannot be disentangled from imperfections in standard MC simulations. In earlier analyses [114, 134, 135], the latter was the only possibility (besides testing the combined background model including τ-embedded data in control regions, as e.g. in Section 6.7.3) to validate τ embedded events. A well-defined way to further study the embedding method at $\tau\tau$ level is provided by the τ-embedded $Z \to \mu\mu$ MC events and subsequent comparisons with standard $Z \to \tau\tau$ MC presented in the following. These τ-embedded $Z \to \mu\mu$ MC events serve as final closure tests of the embedding and are used to estimate effects related to the input muons.

Input muon radiation and reconstruction effects

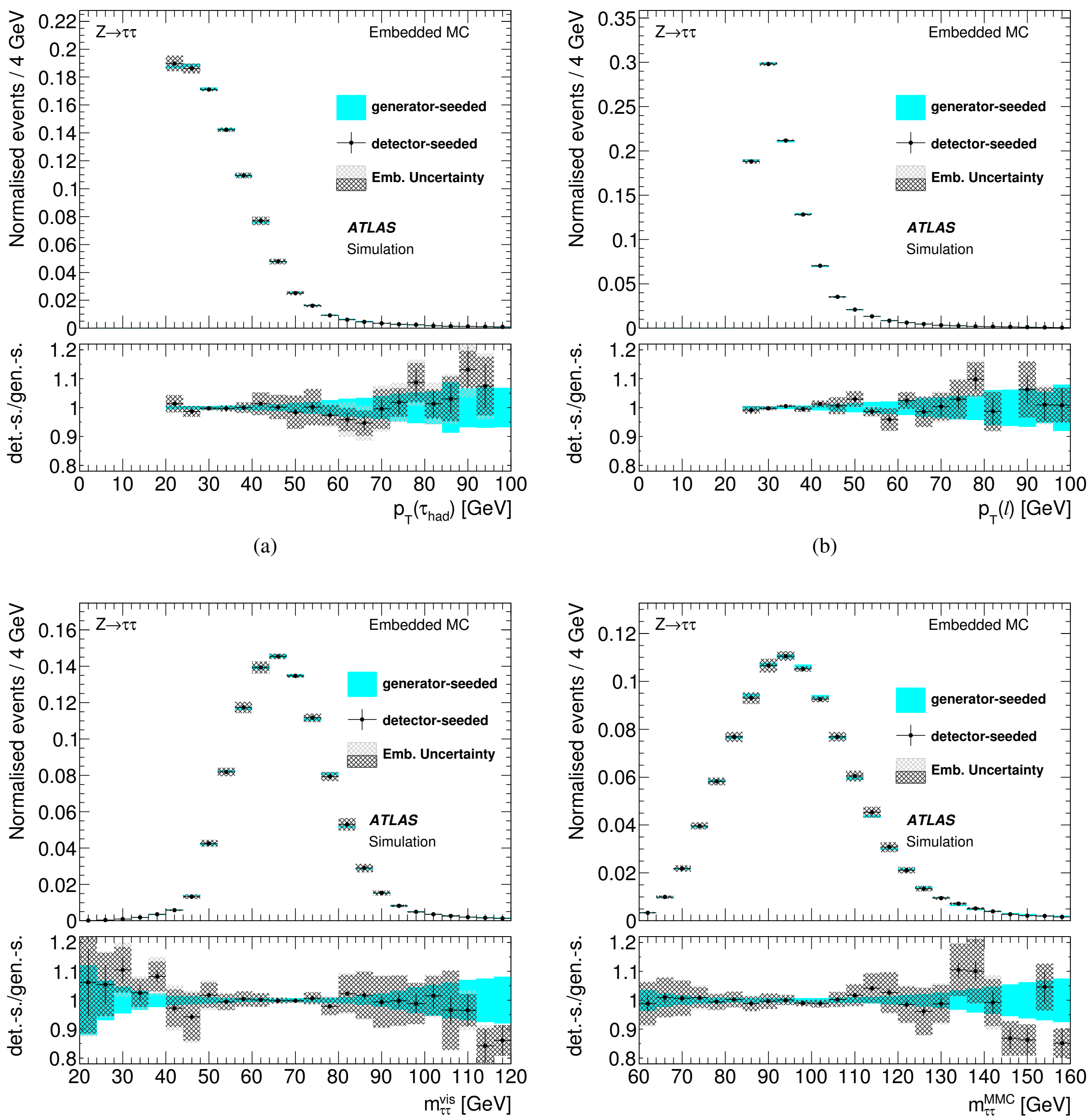
As discussed earlier, τ-embedded events include two effects which originate from the input muons and by construction are unavoidable: The derived kinematics of the $Z \to \tau\tau$ mini event is influenced by the resolution of the reconstructed momenta of the input muons as well as their final state radiation. Whereas these effects are clearly visible in the muon embedding (e.g. in Figure 6.23), the effect will be certainly less distinct in the τ-embedding, because in contrast to muons τ leptons have a worse energy resolution (about 10-20% for energies below 100 GeV [93]). Hence, muon resolution and radiation effects are partially covered by the resolution of reconstructed τ leptons.

In order to judge if these effects are indeed negligible for τ-embedding the distributions of $p_T(\ell)$, $p_T(\tau_{\text{had}})$, visible mass of the visible $\tau\tau$ decay products and invariant $\tau\tau$ MMC mass are compared in Figure 6.25. The blue distributions depict the generator-seeded τ-embedded $Z \to \mu\mu$ MC events. They represent the ideal case of τ embedding, in which the kinematics of the simulated $Z \to \tau\tau$ mini-event are derived from the truth four momenta of the input muons before they radiate. In reality, only detector-seeded τ embedding (illustrated in black) is accessible. An influence on the high-p_T tails is observed, which might be due to the higher accuracy of the τ momentum resolution[6]. In case of reconstructed $\tau\tau$ final states the mass resolution is dominated by the neutrinos produced in the τ decay, so that they are almost not affected.

All differences between both samples are covered by already existing embedding-specific systematic uncertainties. These comparisons confirm that the two effects related to the input muons are negligible and thus do not require to derive additional systematics. However, for future applications, it is conceivable that under consideration of correlations further embedding-specific uncertainties are deduced from these effects. This would especially be true, if the systematic related to the cell subtraction can be further reduced, so that it does not any longer cover remaining differences in Figure 6.25.

In primary order the combined effect of final state radiation and resolution on τ-embedding is of interest, because with respect to the embedding technique they have the same conceptual origin. They originate from the generation of $Z \to \tau\tau$ decays based on already reconstructed decay kinematics in $Z \to \mu\mu$ events. Nevertheless, the embedding toolbox also allows to study both effects separately: The effect

[6] The momentum resolution depends on the pseudorapidity as well as the momentum. The resolution decreases with increasing τ lepton momentum.



(a) (b) (c) (d)

Figure 6.25: Comparison of generator-seeded (gen.-s.) in blue and detector-seeded (det.-s.) in black τ-embedded $Z \to \mu\mu$ MC events: (a) transverse momentum of the visible decay products of the hadronically decaying τ lepton, (b) transverse momentum of the τ decay lepton, (c) visible mass of the visible $\tau\tau$ decay products, (d) $\tau\tau$ invariant mass calculated with the MMC. The corresponding ratio plots below show the relative difference of the distributions from detector-seeded embedding. The blue error band in the ratio plots corresponds to the statistical uncertainties of the generator-embedded events. The black error bars are the statistical uncertainties associated with the detector-seeded embedded events. The light (dark) grey error band corresponds to the sum in quadrature of cell and isolation (cell only) systematic uncertainties of the detector-seeded τ-embedded events. The figures 6.25b, 6.25c and 6.25d are taken from [9].

of final state radiation can be investigated by comparing τ-embedded $Z \to \mu\mu$ MC events, using truth muons before and after radiation as kinematic seeds. The impact of the muon reconstruction can be studied by opposing τ-embedded $Z \to \mu\mu$ MC events, for which truth muons after radiation serve as kinematic seeds and detector-seeded τ-embedded $Z \to \mu\mu$ MC event.

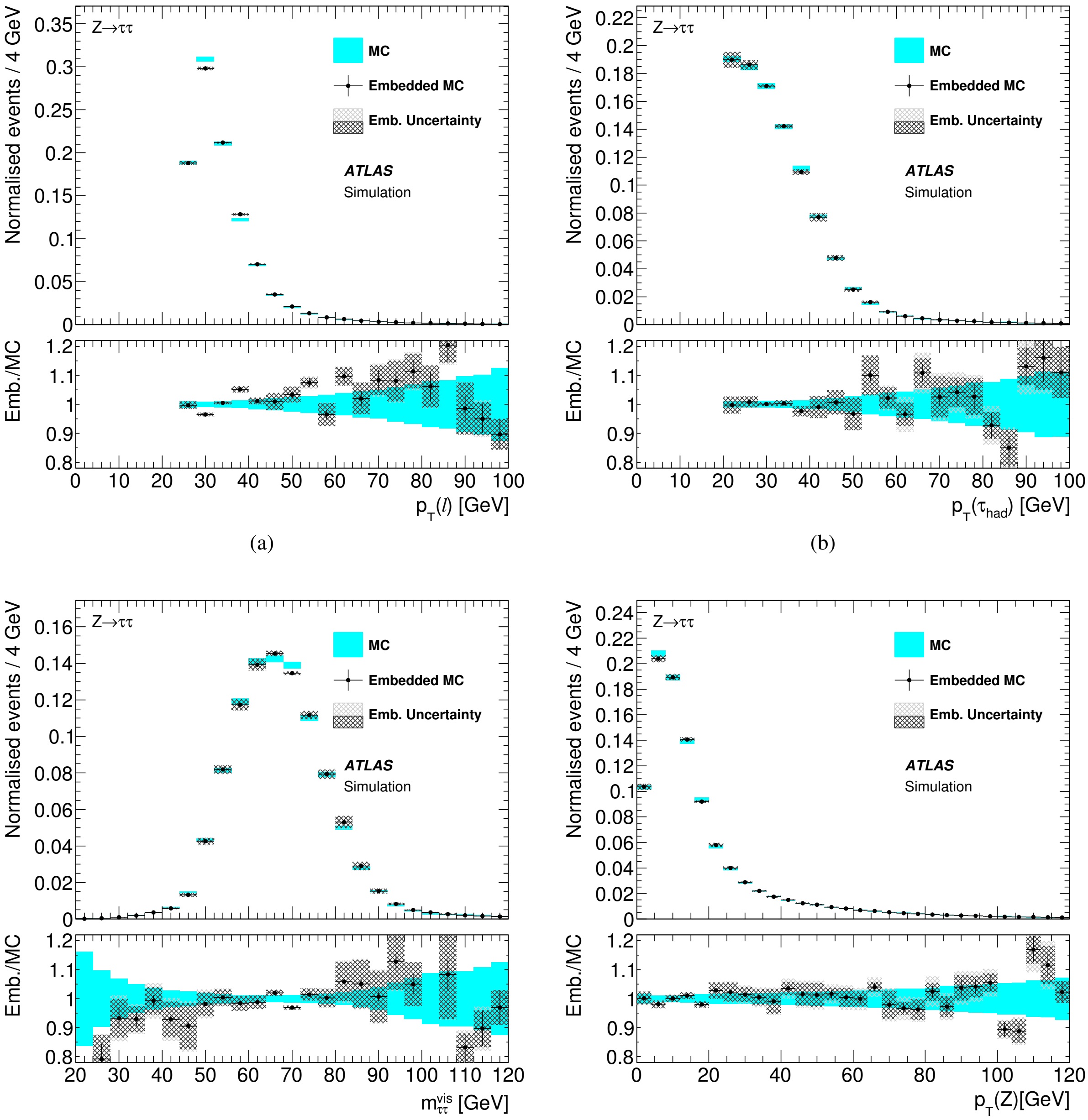
MC embedding

Although the τ-embedding technique was first developed based on input $Z \to \mu\mu$ MC events, $Z \to \mu\mu$ collision data are exclusively used as input since the start of Run 1. As a consequence, a powerful tool which provides a well defined way to further validate the embedding technique got lost: In contrast to comparisons between τ-embedded ATLAS data and $Z \to \tau\tau$ collision data (Section 6.7.3), MC embedding has the advantage, that contaminations from other background processes are excluded. Additionally, potential effects from imperfections of simulations are avoided, because the hybrid event is created by merging the $Z \to \tau\tau$ MC mini event with $Z \to \mu\mu$ MC. Hence, the kinematics of jets, pileup and underlying event are also taken from MC and generated the same way as the $Z \to \tau\tau$ MC reference sample. The latter argument implies, that for the generation of the input $Z \to \mu\mu$ MC sample the same models and consistent configurations as for the $Z \to \tau\tau$ MC reference are applied, so that any observed differences can be reliably interpreted as an effect of the embedding technique.
For the latest $H \to \tau\tau$ analysis, the τ-embedding based on $Z \to \mu\mu$ MC events has been re-developed and used as final closure test by comparing these to standard $Z \to \tau\tau$ MC. In contrast to the $Z \to \mu\mu$ based validation presented in Section 6.7.1, the opposed distributions refer to statistically independent events samples and their statistical error is now additionally displayed in the ratio plots. Also the corrections introduced in Section 6.3 and 6.4 must be applied. The influence of these corrections is separately discussed in Section 6.8.
Due to the missing trigger information, the embedded samples need to be separately normalised to the yields of standard $Z \to \tau\tau$ MC after preselection in the $\mu\tau_{\text{had}}$ and $e\tau_{\text{had}}$ channel. The percentage of MC embedded events after preselection, which enter the boosted and VBF category of the $H \to \tau\tau$ analysis, matches the expectations from standard MC (see Table 7.7).
Some examples used in the final closure test are collected in Figure 6.26 and Figure 6.27: Despite

Table 6.3: Percentage of preselected MC embedded and standard MC $Z \to \tau\tau$ events, that enter the boosted or VBF category of the $H \to \tau\tau$ analysis. The given uncertainties are statistical only.

Category	Embedding (stat)	MC (stat)
Boosted	4.35±0.06%	4.41 ± 0.06%
VBF	0.56±0.02%	0.57±0.02%

of small differences, the transverse momenta of the visible decay products $p_T(\ell)$ in Figure 6.26a and $p_T(\tau_{\text{had}})$ in Figure 6.26b are in agreement within uncertainties. Also other quantities, related to the kinematics of the Z boson like $m_{\tau\tau}^{\text{vis}}$ and $p_T(Z)$ in Figure 6.26c and Figure 6.26d, are in good agreement. Figures 6.27a and Figure 6.27b compare the distribution of the missing transverse energy originating from the simulated τ decay neutrinos and reconstruction effects and the $m_{\tau\tau}^{\text{MMC}}$. The $E_T^{\text{miss}}\phi$ centrality in Figure 6.27c is another example for one of the BDT input variables. For all of these distributions no significant differences are observed. The same conclusions are drawn for jet-related variables like the pseudorapidity difference between both jets in Figure 6.27d.



(a) (b) (c) (d)

Figure 6.26: Comparison of τ-embedded $Z \rightarrow \mu\mu$ MC events (black) with $Z \rightarrow \tau\tau$ MC events (blue): transverse momentum of the (a) light lepton and (b) τ_{had} candidate, (c) visible $\tau\tau$ mass, (d) visible transverse momentum of the Z boson. The corresponding ratio plots below show the relative difference of the τ-embedded distributions. The blue error band in the ratio plots corresponds to the statistical uncertainties of the MC events. The black error bars are the statistical uncertainties associated with the embedded MC events. The light (dark) grey error band corresponds to the sum in quadrature of cell and isolation (cell only) systematic uncertainties of the embedded MC events. Figures 6.26b and 6.26c are taken from [9].

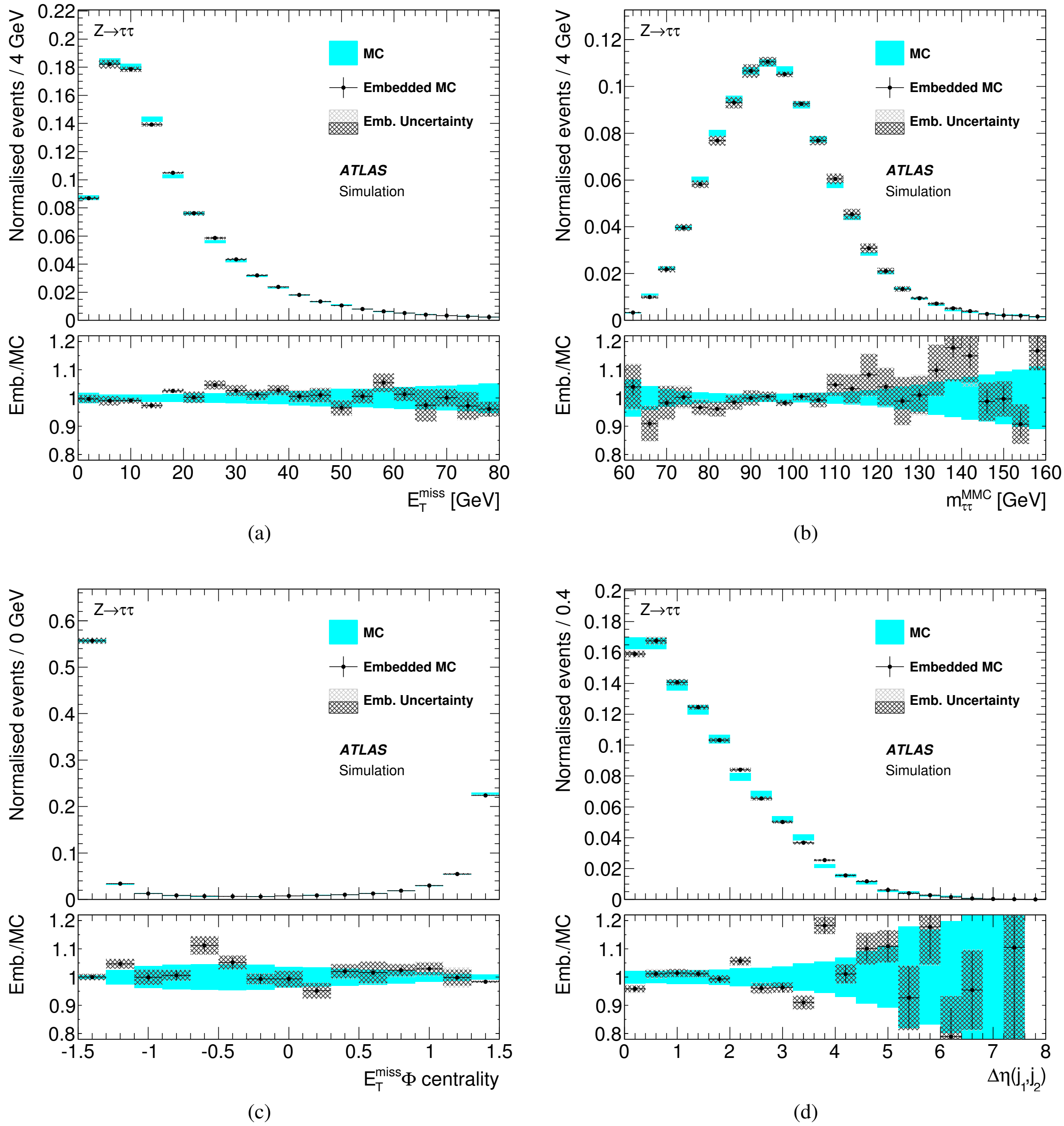


Figure 6.27: Comparison of τ-embedded $Z \to \mu\mu$ MC events (black) with $Z \to \tau\tau$ MC events (blue): (a) missing transverse momentum, (b) invariant $\tau\tau$ MMC mass, (c) $E_T^{\text{miss}}\phi$ centrality, (d) distance in pseudorapidity between the two leading jets. The corresponding ratio plots below show the relative difference of the τ-embedded distributions. The blue error band in the ratio plots corresponds to the statistical uncertainties of the MC events. The black error bars are the statistical uncertainties associated with the embedded MC events. The light (dark) grey error band corresponds to the sum in quadrature of cell and isolation (cell only) systematic uncertainties of the embedded MC events. The figures 6.27a and 6.27b are taken from [9].

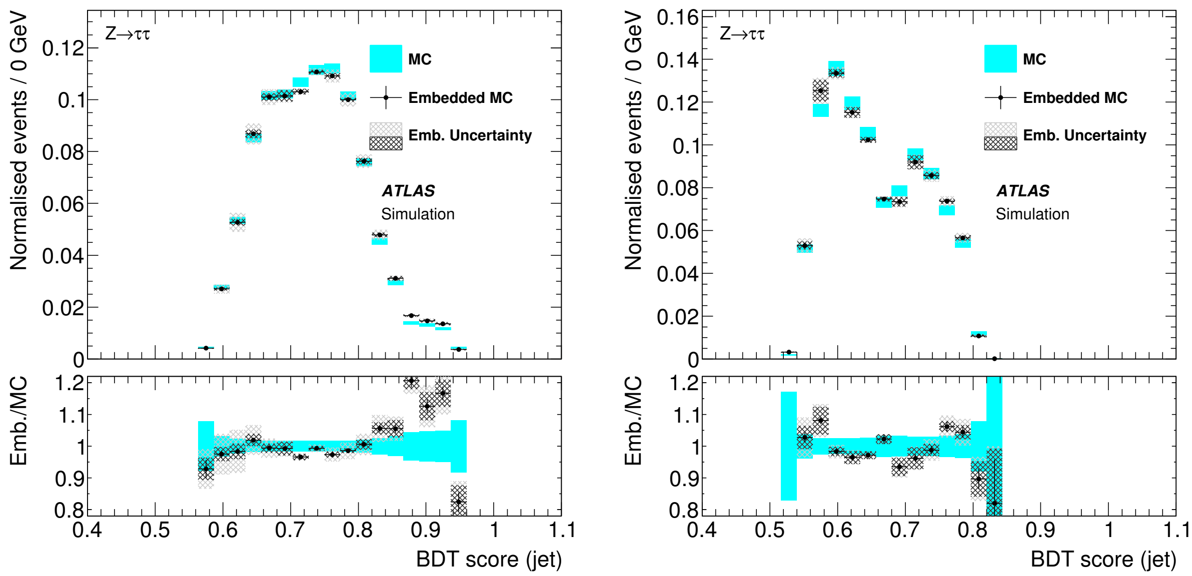
So far only high level observables related to the embedded τ object are discussed. Merging simulated τ objects with real data, however, might cause some modifications of lower level observables which are used to train BDTs in order to separate τ leptons from fake τ candidates. As explained in Section 4.3.7, for the identification of hadronic τ decays three p_T dependent working points loose, medium and tight,

Table 6.4: τ ID efficiencies in τ-embedding and standard ATLAS $Z \to \tau\tau$ MC for one prong decays.

One prong	Embedding (stat)(syst)	MC (stat)
Loose	86.2±0.7±3.7%	86.4 ± 0.1%
Medium	74.1±1.1±4.6%	73.9±0.1%
Tight	49.6±1.0±4.1%	49.4±0.1%

which correspond to different τ ID efficiencies, are provided by those BDTs. In order to study the impact of the embedding on the τ ID, for each working point the τ ID efficiencies in τ-embedded $Z \to \mu\mu$ MC are calculated by counting the number of events passing the $Z \to \tau\tau$ preselection, described in Section 6.2, but failing the τ ID.

The results are compared to the corresponding ones in standard ATLAS $Z \to \tau\tau$ MC in Table 6.4 for one and in Table 6.5 for three prong decays. Very good agreement is found with respect to the τ ID efficiencies of one prong decays. For three prong decays the obtained efficiencies match within uncertainties. Although the efficiencies for the different working points agree well, there are some slight differences visible in the underlying BDT scores shown in Figure 6.28. The BDT score distribution reveals devi-



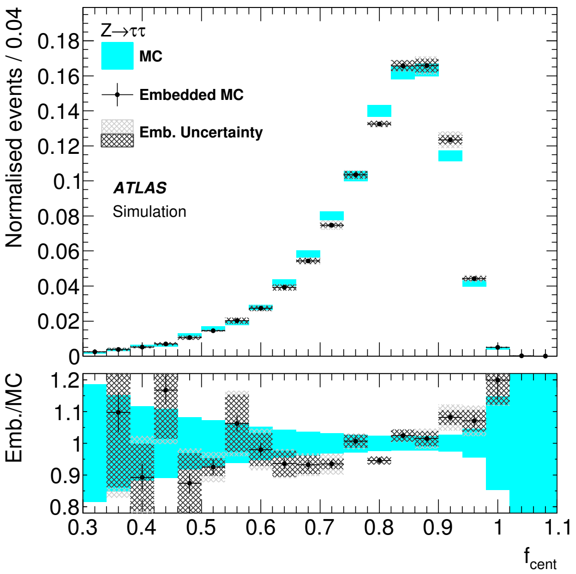
(a) (b)

Figure 6.28: Comparison of τ-embedded $Z \to \mu\mu$ MC events (black) with $Z \to \tau\tau$ MC events (blue) in terms of the τ BDT score for one prong (left) and three prong (right) decays. The corresponding ratio plots below show the relative difference of the τ-embedded distributions. The blue error band in the ratio plots corresponds to the statistical uncertainties of the MC events. The black error bars are the statistical uncertainties associated with the embedded MC events. The light (dark) grey error band corresponds to the sum in quadrature of cell and isolation (cell only) systematic uncertainties of the embedded MC events.

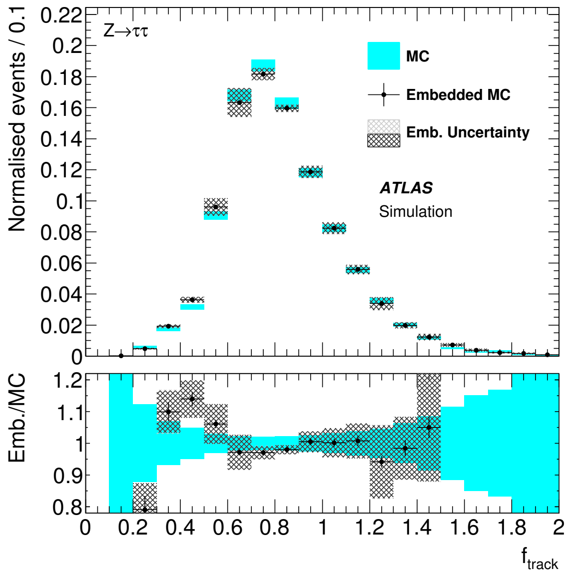
Table 6.5: τ ID efficiencies in τ-embedding and standard ATLAS $Z \to \tau\tau$ MC for three prong decays.

Three prong	Embedding (stat)(syst)	MC (stat)
Loose	71.0±0.8±4.7%	73.0 ± 0.2%
Medium	61.8±0.7±3.3%	63.0±0.2%
Tight	40.0±0.4±3.4%	42.3±0.2

ations in the high BDT score region above 0.8, which, however, lies above the tight working points and hence does not influence the resulting τ ID efficiencies. Those deviations arise from differences in the modelling of the BDT input variables (see Section 4.3.7) due to unavoidable software inconsistencies in simulation of the $Z \to \tau\tau$ mini event and the standard ATLAS $Z \to \tau\tau$ MC as similar explained for the re-reconstruction in Section 6.7.1. Two distributions of BDT input variables are shown in Figure 6.29: the central energy fraction and the leading-track momentum fraction of three prong τ decays. Both distributions agree within systematic uncertainties and hence indicate, that the detector response to embedded τ leptons does not significantly differ from τ leptons in standard MC.
In all relevant distributions τ-embedded $Z \to \mu\mu$ MC events agree with standard ATLAS $Z \to \tau\tau$ MC within embedding-related uncertainties. Additional effects like the modification of the input muon kinematics due to resolution and final state radiation as well as residual effects from the modelling of the Z boson polarisation on the τ decays are included in these comparisons. They are covered by current embedding-related uncertainties. Hence, no additional systematics need to be derived for these effects.

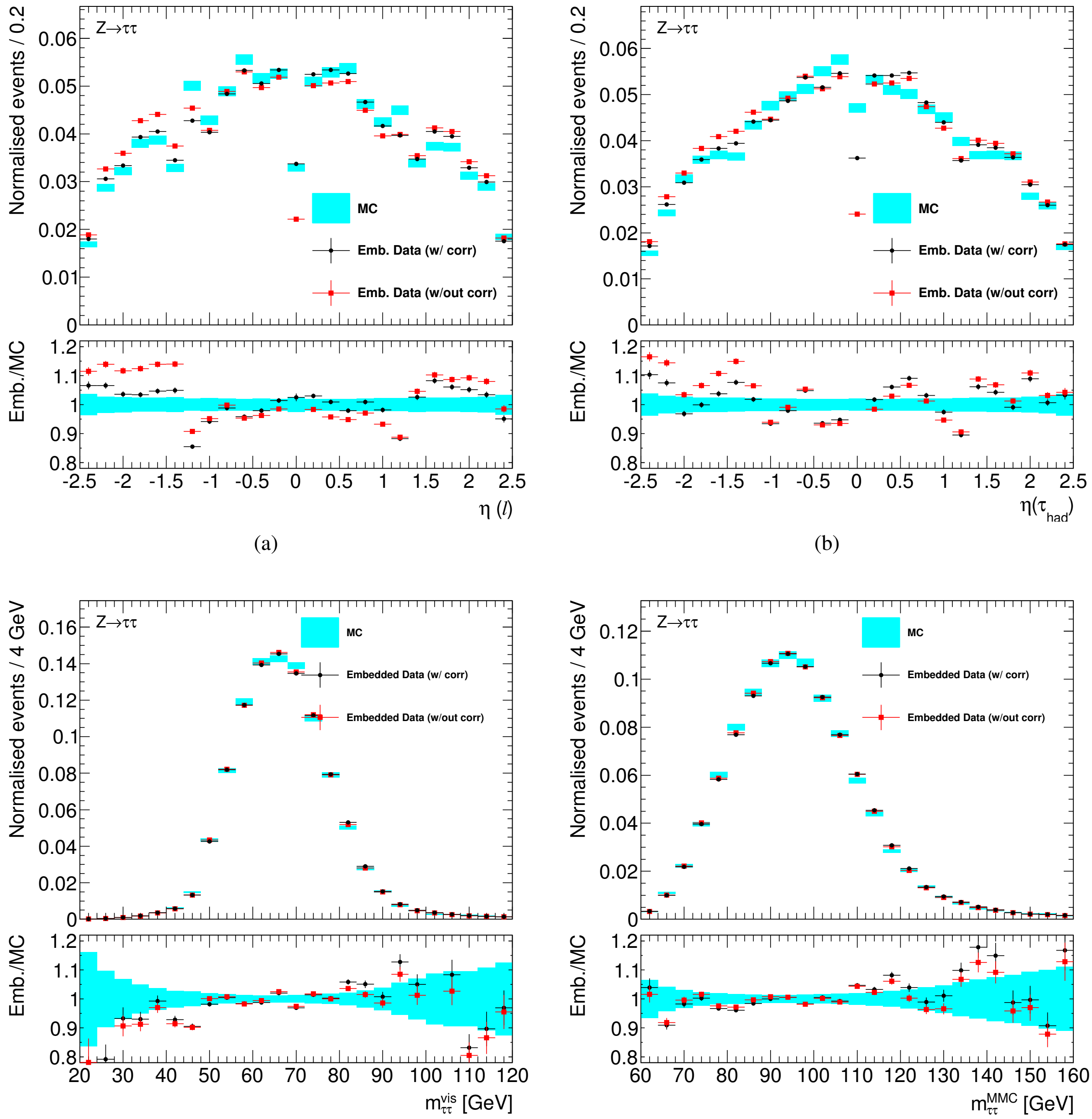


(a)



(b)

Figure 6.29: Comparison of τ-embedded $Z \to \mu\mu$ MC events (black) with $Z \to \tau\tau$ MC events (blue) for BDT input variables: (a) central energy fraction and (b) leading-track momentum fraction for three-prong hadronic τ decays. The corresponding ratio plots below show the relative difference of the τ-embedded distributions. The blue error band in the ratio plots corresponds to the statistical uncertainties of the MC events. The black error bars are the statistical uncertainties associated with the embedded MC events. The light (dark) grey error band corresponds to the sum in quadrature of cell and isolation (cell only) systematic uncertainties of the embedded MC events. The figures are taken from [9].



(a) (b) (c) (d)

Figure 6.30: Comparison of standard $Z \to \tau\tau$ MC (blue) with (un-) corrected τ-embedded events in black (red): the pseudorapidity of (a) the leptonically and (b) the hadronically decaying τ lepton, (c) the visible $\tau\tau$ mass and (d) the invariant $\tau\tau$ MMC mass.

Z control region

All validation methods, introduced so far, do not imply direct comparisons of embedded events with $Z \to \tau\tau$ data. They are carried out, before using embedded samples in the $H \to \tau\tau$ analysis. In the final step, the τ embedded $Z \to \mu\mu$ collision data events must be validated within the analysis as part of the combined background model in a dedicated region. This validation is performed in parallel for all three channels of the SM $H \to \tau\tau$ analysis. Examples for the $H \to \tau_\ell\tau_h$ use-case are summarised in Section 7.4.1.

6.8 Performance of embedding-specific corrections in embedded events

The combined effect of the external corrections discussed in Section 6.4 is studied using τ-embedded MC. In Figure 6.30, the effect is illustrated for four quantities: the pseudorapidity of the visible τ decay products, the visible $\tau\tau$ mass and the invariant $\tau\tau$ MMC mass. The pseudorapidity is strongly related to the source of corrections. Its mismodelling originates from differences in the detector acceptance between the original muons and the embedded objects. The corrections significantly improve the shape of the distributions. Especially in the case of the light lepton p_T spectrum a good agreement at $\eta = 0$ is found. After applying the corrections the pseudorapidity is still not perfectly modelled. However, the results in Chapter 7 indicate that the mismodelling of this particular variable has no impact on observables relevant for the $H \to \tau\tau$ analysis. While the corrections have a visible effect on η and ϕ their impact is less pronounced for other analysis-relevant quantities like the visible mass or the $\tau\tau$ invariant MMC mass (Figure 6.30c and Figure 6.30d). The actual size of the corrections differs between the three $H \to \tau\tau$ analysis channels.

6.9 Rotation around the axis of the Z boson

The embedding procedure can be further improved for future applications by inserting the simulated $Z \to \tau\tau$ decay into another region of the detector than the original $Z \to \mu\mu$ decay.
An example of such a transformation in η and ϕ is the rotation of the Z boson decay products around the Z boson axis: In the rest frame of the Z boson the decay leptons have opposite directions and enclose a fixed angle with the Z boson axis, while the other angle is arbitrary as illustrated in Figure 6.31a. For the transformation, the direction of the muons is determined in the Z boson rest-frame and afterwards rotated by a fixed angle α (Figure 6.31b). This angle is preferably chosen such that the new position significantly differs from the old one ($\alpha = \pi/2$). The rotated momenta are transformed back into the laboratory frame and are used to generate the $Z \to \tau\tau$ decay as described in Section 6.1. The original muon tracks and associated cells are removed and the simulated $Z \to \tau\tau$ decay products are embedded in a detector region which is distinct from the muon direction. Such a transformation has the advantage that the bias due to the selection of $Z \to \mu\mu$ events is reduced. Effects due to the muon cell subtraction and corresponding systematic uncertainties are also reduced.
Figure 6.32 shows the pseudorapidity distribution of the lepton and the τ_{had} candidate of rotated τ-embedded MC events in comparison with $Z \to \tau\tau$ MC events. All corrections described in Section 6.4 are applied except for corrections concerning the muon selection efficiencies. These η-dependent corrections are ignored in order to study the impact of the muon selection bias on simulated $Z \to \tau\tau$ events which are embedded into another detector region than the original $Z \to \mu\mu$ decay. Indeed, a good agreement between rotated embedded events and $Z \to \tau\tau$ MC events at around η =0 is achieved. In this region, the performance of rotated events is even better than the performance of $Z \to \tau\tau$ decays embedded according to the procedure described in Section 6.1 for which all corrections (including selection efficiency corrections) are applied[7]. The standard τ-embedded events are depicted in Figure 6.32 as red points. Although the agreement between standard embedded events and $Z \to \tau\tau$ MC improves after

[7] For simplification the latter class of embedded events is referred to as standard τ-embedded events.

applying corrections (see Figure 6.30), there are still some discrepancies visible in the pseudorapidity spectrum of the τ_{had} candidate. After rotation, however, the τ-embedded distribution at around $\eta = 0$ matches $Z \to \tau\tau$ MC within uncertainties. There are still some differences visible in the tails of the pseudorapidity distribution. The impact of embedding-specific corrections on other kinematic quantities like transverse momentum, missing transverse energy or invariant $\tau\tau$ mass are shown to be small [9]. For that reason, no significant influence due the rotation of the $Z \to \tau\tau$ decay is expected which is confirmed by distributions shown in Figure 6.33.
These quantities also benefit from inserting the $Z \to \tau\tau$ decay into a different detector region than the original muons, because the uncertainty originating from the cell subtraction is reduced. This reduction results from the fact that the cells in which the muons have deposited their energies are situated far away from the embedded objects. Hence, only cell energies far away from the embedded τ decay products are subtracted. The impact of the rotation on this uncertainty is depicted in Figure 6.34 for the invariant $\tau\tau$ MMC mass and the transverse momentum of the τ_{had} candidate. While for standard τ-embedded events (upper ratio plots) the size of the cell subtraction uncertainty is in the range of 5-7% in the relevant regions, the uncertainty is significantly reduced by rotating events.
The rotation of $Z \to \mu\mu$ decays slightly affect the $Z \to \mu\mu$ acceptance as illustrated in Figure 6.35: Muons in $Z \to \mu\mu$ decays are selected if they exceed 15 GeV. When rotating these muons the transverse momentum changes and can drop below the analysis thresholds. Some events can be recovered by the kinematic filter (see Section 6.1) of the embedding procedure. If the rotated transverse muon momentum - and consequently the true τ transverse momentum - is below the filter thresholds, the corresponding τ decay products can not exceed the filter threshold and are not available for the analysis. In contrast to standard embedded events, the efficiency of the analysis preselection is about 7.3% smaller for rotated events. This lost in acceptance should be corrected for by event weights.
Although this simple rotation results in slight modulations of the angle between the muon and the proton axis in Z boson rest-frame, the study of analysis relevant observables demonstrates that the modelling

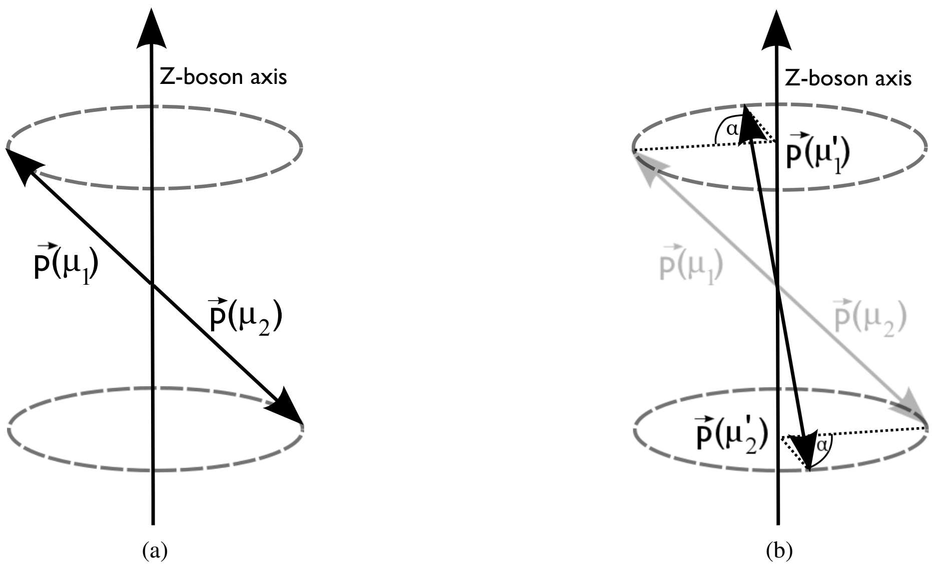
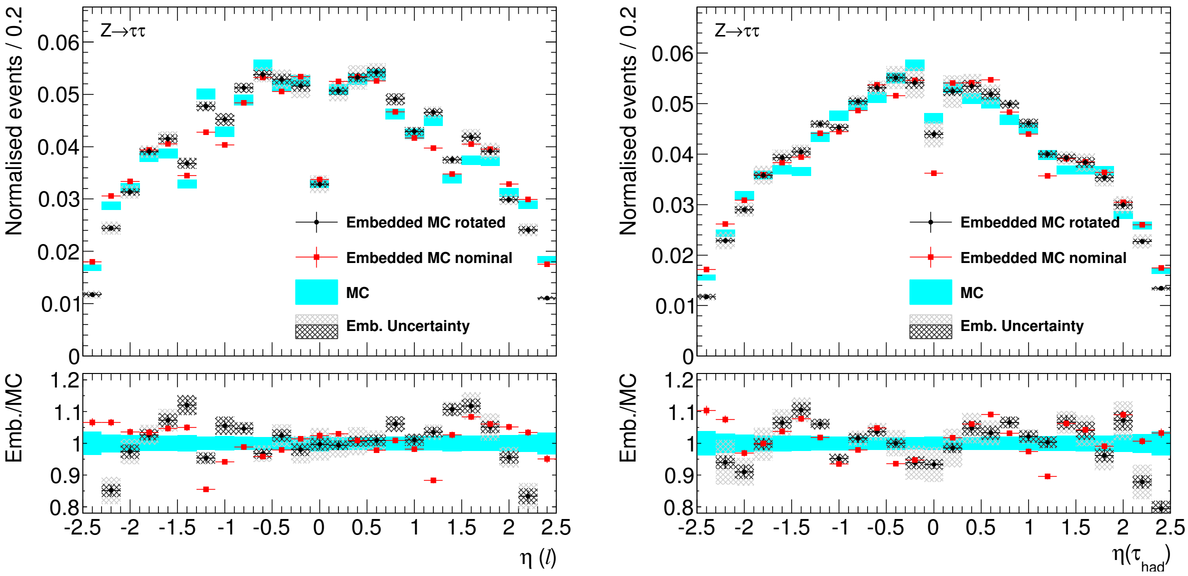


Figure 6.31: Scheme of a $Z \to \mu\mu$ decay in the Z boson rest-frame. The vertical line illustrates the Z boson axis, while the other two arrows indicate the direction of the muons. Additionally in (b) the direction of the muons is shown after rotating the muons around the angle α.



(a) (b)

Figure 6.32: Comparison of $Z \to \tau\tau$ MC events (blue) with τ-embedded $Z \to \mu\mu$ MC events after rotation (black) and standard τ-embedded $Z \to \mu\mu$ MC events (red): (a) pseudorapidity of the light lepton and (b) pseudorapidity of the τ_{had} candidate. The rotated events are not corrected for muon selection efficiencies. The corresponding ratio plots show the relative difference between the rotated (standard) τ-embedded and $Z \to \tau\tau$ MC distributions in black (red). The blue error bands in the ratio plots correspond to the statistical uncertainties of the $Z \to \tau\tau$ MC events. The black error bars are the statistical uncertainties associated with the rotated embedded events. The light (dark) grey error band corresponds to the sum in quadrature of cell and isolation (cell only) systematic uncertainties of the rotated τ-embedded events. The red error bars of the standard τ-embedded events illustrate the corresponding statistical error.

based on rotated τ-embedded events is compatible to standard τ-embedded events. The improvement of the η-modelling and the reduction of systematic uncertainties are good arguments to use a transformation of muon four-momenta in the future.

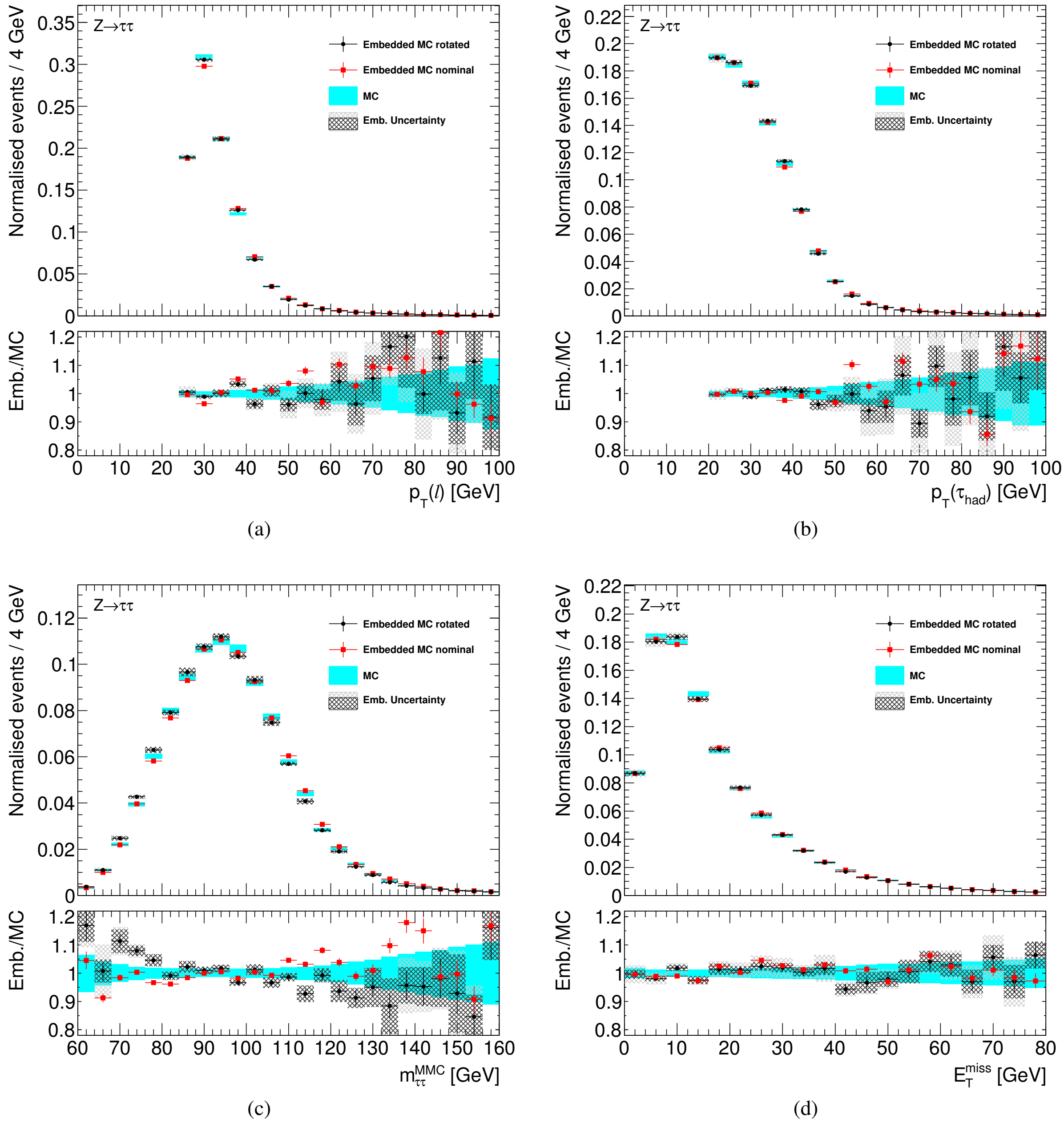
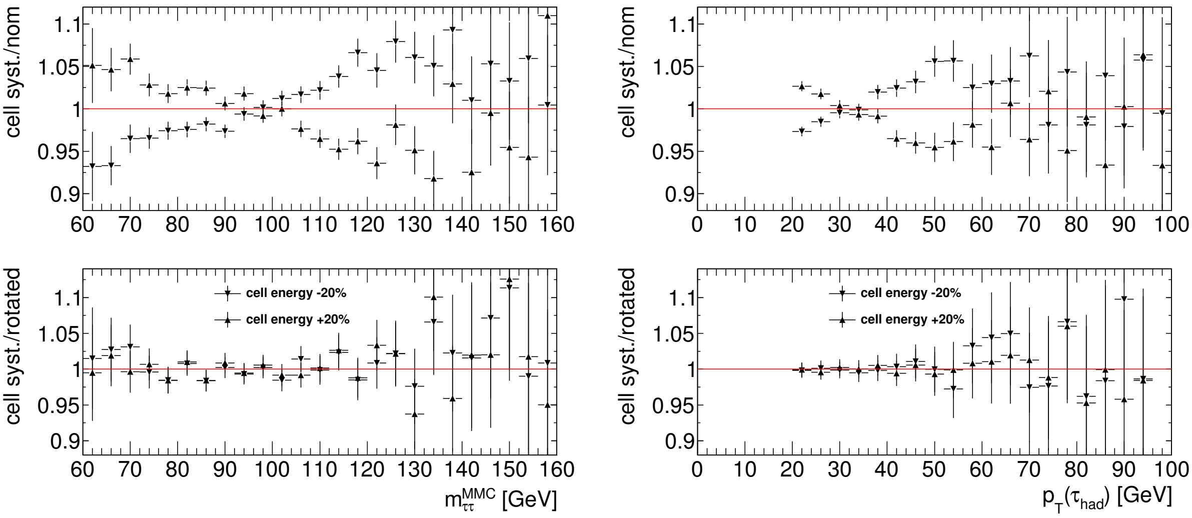


Figure 6.33: Comparison of $Z \to \tau\tau$ MC events (blue) with τ-embedded $Z \to \mu\mu$ MC events after rotation (black) and standard τ-embedded $Z \to \mu\mu$ MC events (red): (a) transverse momentum of the light lepton, (b) transverse momentum of the τ_{had} candidate, (c) invariant $\tau\tau$ mass calculated with MMC and (d) missing transverse energy. The rotated events are not corrected for muon selection efficiencies. The corresponding ratio plots below show the relative difference between the rotated (standard) τ-embedded and $Z \to \tau\tau$ MC distributions in black (red). The blue error bands in the ratio plots correspond to the statistical uncertainties of the $Z \to \tau\tau$ MC events. The black error bars are the statistical uncertainties associated with the rotated embedded events. The light (dark) grey error band corresponds to the sum in quadrature of cell and isolation (cell only) systematic uncertainties of the rotated τ-embedded events. The red error bars of the not-rotated τ-embedded events illustrate the corresponding statistical error.



(a) (b)

Figure 6.34: Upper ratio plots: Ratio between the cell varied τ-embedded events and standard τ-embedded events. Lower ratio plots: Ratio between the cell varied τ-embedded events and rotated τ-embedded events. The subtracted cell energy is varied by ±20%. Left: The invariant $\tau\tau$ mass. Right: The transverse momentum of the τ_{had} candidate. The error bars show the statistical uncertainty.

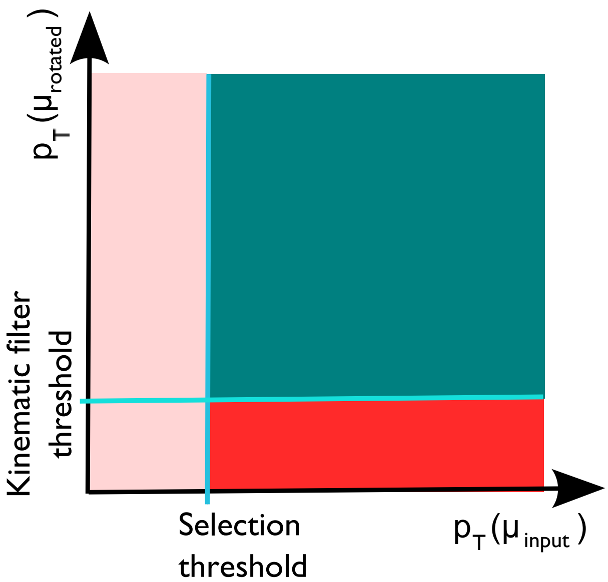


Figure 6.35: Scheme of muon transverse momenta before (x-axis) and after (y-axis) rotation. The vertical line illustrates the muon momentum threshold used to selected $Z \to \mu\mu$ input events while the horizontal line indicates the threshold of the kinematic filter applied in the embedding procedure on visible τ decay products. Events below this threshold are rejected. Events in the turquoise phase space are available for the analysis.

CHAPTER 7

Analysis of the $H \rightarrow \tau_\ell \tau_h$ decay channel with 4.5 fb^{-1}+ 20.3 fb^{-1}

The first preliminary evidence for a SM Higgs boson decaying into two τ leptons was achieved with the analysis of the 2012 dataset [135]. This analysis was based on the dataset taken at $\sqrt{s}$ =8 TeV, which corresponds to an integrated luminosity of 20.3 fb^{-1}. The analysis of this dataset results in an observed (expected) significance of 4.1 (3.2) standard deviations and a measured signal strength of $\mu = 1.4^{+0.5}_{-0.4}$. Whereas previous searches for SM $H \rightarrow \tau\tau$ decays [114, 134] were based on the fit of the invariant $\tau\tau$ mass in a signal-enriched phase space region[1], the claim of an evidence with the Run 1 dataset was possible by a drastic change of the analysis strategy using multivariate methods.
Multivariate methods are advantageous in cases of complex final states like $H \rightarrow \tau\tau$. They are used to improve the signal-to-background separation and hence the signal sensitivity. The good separation power, which allows to use rather loose event categories, is achieved, because multivariate methods exploit correlations between observables. The combination of several variables, all of them with some signal-to-background discrimination power, into one classifier is a powerful tool towards the discovery of rare physical processes. In this thesis Boosted Decision Trees (BDTs) are used as multivariate classifier. This method is implemented in the Toolkit for Multivariate Data Analysis (TMVA) [99], which is part of the ROOT data analysis framework [136]. The final BDT output distributions are then used for the final signal extraction.
The analysis of the $H \rightarrow \tau\tau$ decay distinguishes three channels: $\tau_l\tau_l$, $\tau_l\tau_\mathrm{h}$ and $\tau_\mathrm{h}\tau_\mathrm{h}$ defined by the τ decay products (see Section 5.3). As illustrated in Figure 5.5a the $\tau_l\tau_\mathrm{h}$ and $\tau_l\tau_l$ channel consist of two ($\tau_\mu\tau_\mathrm{h}$, $\tau_e\tau_\mathrm{h}$) and three ($\tau_e\tau_e$, $\tau_e\tau_\mu$, $\tau_\mu\tau_\mu$) sub-channels, respectively. These sub-channels are defined by the flavour of the light lepton (electron or muon).
In this chapter, the multivariate analysis of the $H \rightarrow \tau_\ell\tau_h$ channel is presented. It is based on datasets from the data taking periods in 2011 and 2012 at $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV, respectively. The datasets correspond to an overall integrated luminosity of 4.5 fb^{-1} and 20.3 fb^{-1}. This analysis entered the $H \rightarrow \tau\tau$ combination (see Chapter 8) published by ATLAS [8].

[1] This strategy is referred to as cut-based analysis in the following. The cut-based analysis of the $H \rightarrow \tau\tau$ channel is briefly discussed in Section 8.4.

7.1 Data and Monte Carlo samples

The analysis is based on data which is recorded with the ATLAS detector during the Run 1 of the LHC at a proton-proton center-of-mass energy of $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV, respectively. After applying certain data-quality requirements (see Section 7.2.1), the datasets correspond to an integrated luminosity of 4.5 fb^{-1} and 20.3 fb^{-1}. In this section, the Monte Carlo signal and background samples used to analyse the $H \to \tau_\ell \tau_h$ channel are introduced, while in the subsequent section a simplified way to simulate $H \to \tau_\ell \tau_h$ background processes is presented, which will become of interest in future Run 2 analyses.

7.1.1 Monte Carlo samples of the $H \to \tau_\ell \tau_h$ analysis

For the Higgs signal the three most dominant production processes (see Section 2.3) are considered: ggF, VBF and VH. The ggF and VBF processes are simulated by the POWHEG [137–140] event generator, which uses the CT10 [38] parametrisation of the parton distribution functions (PDF), interfaced to PYTHIA8 [44]. The normalisation of the VBF process is obtained using NLO QCD and electroweak (EW) corrections [141, 142] in combination with an approximate NNLO in QCD [48]. For the gluon-gluon fusion process the normalisation is calculated at NNLO in QCD [143–148] with a soft-gluon resummation up to NNLL [149] as well as NLO EW corrections [150, 151]. The VH production process is simulated using the PYTHIA8 event generator with the CTEQ6L1 parametrisation of the PDFs [39]. Its normalisation is taken from calculation at NNLO in QCD [152] using NLO EW radiative corrections [153].
For Higgs bosons produced in ggF the shape of the generated distribution of the transverse momentum is corrected such that it matches the distribution from a calculation at NNLO with NNLL corrections derived by the HRES2.1 [154] program. The calculation considers the effect of a finite top and bottom mass [154, 155]. It also makes use of dynamical renormalisation and factorisation scales $\mu_R, \mu_F = \sqrt{m_H^2 + p_T^2}$ (see Section 2.2.1 and Section 2.2.2). Events with at most one jet and with at least two jets at particle level are reweighted separately. The transverse momentum distribution of signal events with at least two jets are corrected to match the MINLO HJJ predictions [156]. The reweighting is performed in a way such that the inclusive transverse momentum spectrum of the Higgs boson and the transverse momentum spectrum of events with at least two jets match the HRES2.1 and MINLO HJJ predictions. In addition, the jet multiplicity is required to be in agreement with (N)NLO calculations from JETVHETO [157–159].
Since the NLO EW corrections for the VBF production process depend on the transverse momentum of the Higgs boson[2] the transverse momentum spectrum for Higgs bosons produced via VBF is reweighted according to the differences between POWHEG+PYTHIA and HAWK [141, 160] calculations.
The irreducible $Z \to \tau\tau$ background is modelled by the data-driven embedding method as described in Section 6.1.
All other background processes are simulated by different event generators which are interfaced either to PYTHIA [44] or HERWIG [45, 161] delivering the parton shower, hadronisation and the modelling of underlying events. In samples generated with HERWIG TAUOLA [120] is used to simulate the decay of τ leptons. The photon radiation for charged leptons is provided by PHOTOS [121]. W+jets and Z+jets samples are generated using the ALPGEN event generator [162] which employs the MLM matching scheme [163] between the hard process (calculated with LO matrix elements for up to five partons) and the parton shower. For the WW process the GG2WW [164] program is used in order to

[2] The correction increases from few percent at low transverse momentum up to about 20% at 300 GeV.

generate the loop-induced $gg \rightarrow WW$ process. While the CTEQ6L1 parametrisation of the PDFs is used for the ACERMC [165], ALPGEN and HERWIG generators the CT10 parametrisation is used for events generated with GG2WW.
The ATLAS detector response [72] is simulated for all processes by the GEANT4 [73] program (Section 4.2). Events from minimum-bias interactions are simulated using the AU2 [166] parameter tuning of PYTHIA8. The AU2 tuning contains a set of optimised parameters for parton shower, hadronsiation as well as multiple parton interactions. They are merged with the simulated signal and background processes according to the luminosity profile of the recorded data. The contributions from those pileup interactions are simulated both within the same bunch crossing as the hard scattering process and in neighbouring bunch crossings. In the final step, the simulated events are reconstructed using the same algorithms as for data.
For Z+jet events VBF-filtered Monte Carlo samples[3] are used in combination with non-filtered samples in order to increase the available statistics in the VBF category. The samples are merged by employing an emulated VBF filter, which exploits truth level information.
A list of Monte Carlo samples used for the analysis can be found in Appendix B.

7.1.2 Prospective studies of MC events simulated with ATLFASTII

The most CPU consuming part of the event simulation is the shower simulation. Since for many applications a detailed simulation of the detector response is not necessary, different algorithms have been developed in Run 1 to reduce CPU time for simulation.
The ATLFASTII algorithm [72] provides a fast simulation of the ATLAS tracking system by FAT-

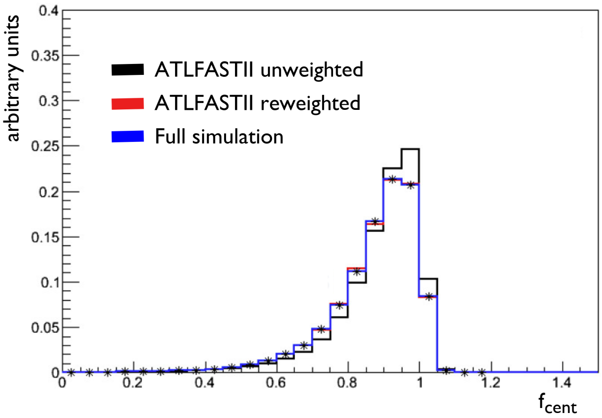
Table 7.1: BDT and likelihood τ ID efficiency at the *medium* and *tight* working point for truth-matched τ_{had} candidates in full and fast-simulated $Z \rightarrow \tau\tau$ events. The last column shows the τ ID efficiencies for fast-simulated events after application of E_T and η dependent weights.

Working point	Full simulation	Fast simulation	Fast simulation reweighted
BDT *tight*	(44.33±0.07)%	(48.37±0.07)%	(44.73±0.07)%
BDT *medium*	(66.32±0.07)%	(69.16±0.07)%	(65.92±0.07)%
Likelihood *tight*	(41.26±0.07)%	(44.53±0.07)%	(42.19±0.07)%
Likelihood *medium*	(64.91±0.07)%	(67.47±0.07)%	(65.19±0.07)%

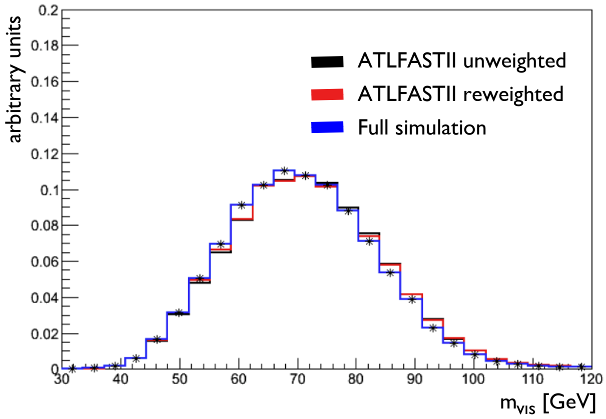
RAS [167] in combination with a shower parametrisation by FastCaloSim [168]. It is 10 times faster than the common full simulation. For that reason, the use of samples simulated with ATLFASTII are of interest for the $H \rightarrow \tau\tau$ analysis. Based on comparisons of full- and fast-simulated 8 TeV MC events it was investigated to which extent fast-simulated samples are applicable in the presented analysis.
In this context, the effect of a simplified calorimeter geometry on τ related observables is of particular interest. In Table 7.1 the efficiencies of the BDT and likelihood τ ID are compared at the *medium* and

[3] The VBF filter is applied on generator level in order to select the typical VBF signature in MC events. The loose (tight) VBF filter requires the dijet mass to exceed 200 GeV (400 GeV) and $\Delta\eta_{jj}$>2.0 (4.0). By using the VBF filter the MC samples are enriched with events revealing this signature and hence reduces the statistical uncertainties in the analysis categories.

tight working points calculated for truth-matched[4] $\tau_{\rm had}$ candidates in $Z \to \tau\tau$ samples: The ID efficiency is slightly overestimated in fast-simulated event samples. By reweighting the fast-simulated distribution of the centrality energy fraction in bins of η and E_T (Figure 7.1) a good agreement with respect to the distributions of τ ID input variables, p_T resolution and τ ID efficiency is achieved [169, 170]. Except for $t\bar{t}$ the reweighting scheme reveals a good performance for different processes containing true $\tau_{\rm had}$ candidates. Besides the study of $\tau_{\rm had}$ characteristic observables, the potential use of fast-simulated W+jets and other fast-simulated samples in the context of the $H \to \tau_\ell \tau_h$ analysis was investigated [170, 171]. For the Run 1 analysis the reduction in simulation time cannot cancel disadvantages of an additional reweighting and derivation of corresponding systematic uncertainties, so that the presented analysis is exclusively based on full-simulated processes. Nevertheless, for analyses based on data collected in Run 2, the application of fast-simulated processes should be re-investigated. Especially, since in 2014 a new shower model was implemented which includes shape fluctuations [74] improving the $\tau_{\rm had}$ modelling. Also the embedding technique (Section 6.1) might benefit, since the simulation of the $Z \to \tau\tau$ mini event is the most time consuming part in the embedding procedure. By using fast-simulated τ leptons the large CPU resources necessary for the embedding technique could be significantly reduced and thus embedded samples would get available sooner after the detection of the $Z \to \mu\mu$ collision data.



(a)



(b)

Figure 7.1: Central energy fraction $f_{\rm cent}$ and visible mass distribution for $Z \to \tau\tau$ events simulated with ATLFASTII and full ATLAS simulation.

7.2 Event selection

Events of interest are selected exploiting characteristics of Higgs production mechanisms and decay signatures discussed in Section 5.3. The selection is subdivided into several steps: A general data cleaning to guarantee the data quality, an object selection defining the physics objects to be analysed and a basic preselection which selects objects to be expected in the signal process. A final categorisation distinguishes between ggF and VBF production mode. Many efforts were made to harmonise the event selection- and in particular the object selection- between the three channels.

[4] The term *truth-matched* refers to the technique in which a reconstructed object like the $\tau_{\rm had}$ candidate is matched to a particle at generator-level.

7.2.1 Data quality

In order to ensure the quality of data collisions a data cleaning and jet cleaning is applied, i.e. rejecting events not originating from proton-proton collisions like cosmic rays: The event is required to have at least one reconstructed primary vertex with at least four associated tracks with $p_T > 400\,\text{MeV}$. The application of a GRL (see Section 3.4) ensures that the detector worked well.

7.2.2 Trigger requirements

Events are preselected using single-electron or single-muon triggers. The exact trigger thresholds depend on the lepton flavours and datasets ($\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV). For the $e\tau_{\text{had}}$ channel of the $\sqrt{s}$ =7 TeV dataset, four different electron triggers are used (EF_e20_medium[5], EF_e22_medium, EF_e22_medium1 and EF_e22vh_medium1) with increasing p_T thresholds and identification criteria in order to account for the increasing pileup. Two muon triggers are used: EF_mu18_MG for the beginning of data taking period and EF_mu18_MG_medium, which introduces a medium identification criteria to cope with the increasing pileup. For the $\sqrt{s}$ =8 TeV dataset one single trigger is used for each channel: EF_e24vhi_medium1 and EF_mu24i_tight. While the p_T thresholds of the triggers are called online thresholds the lepton p_T thresholds of the analysis are called offline thresholds. The offline thresholds are high enough above the corresponding online thresholds to ensure a fully efficient trigger. All online and offline thresholds used for both datasets and analysis channels are summarised in Table 7.2.

Table 7.2: Summary of trigger and lepton p_T thresholds.

Channel	Trigger	Online p_T	Offline p_T
$e\tau_{\text{had}}$ ($\sqrt{s}$ =7 TeV)	Single electron	$p_T(e) > 20\text{-}22\,\text{GeV}$	$p_T(e) > 25\,\text{GeV}$ $p_T(\tau_{\text{had}}) > 20\,\text{GeV}$
$\mu\tau_{\text{had}}$ ($\sqrt{s}$ =7 TeV)	Single muon	$p_T(\mu) > 18\,\text{GeV}$	$p_T(\mu) > 22\,\text{GeV}$ $p_T(\tau_{\text{had}}) > 20\,\text{GeV}$
$e\tau_{\text{had}}$ ($\sqrt{s}$ =8 TeV)	Single electron	$p_T(e) > 24\,\text{GeV}$	$p_T(e) > 26\,\text{GeV}$ $p_T(\tau_{\text{had}}) > 20\,\text{GeV}$
$\mu\tau_{\text{had}}$ ($\sqrt{s}$ =8 TeV)	Single muon	$p_T(\mu) > 24\,\text{GeV}$	$p_T(\mu) > 26\,\text{GeV}$ $p_T(\tau_{\text{had}}) > 20\,\text{GeV}$

[5] The triggers are named according to the trigger level (EF stands for the event filter trigger) and the triggered object (e for electrons, mu for muons). The object is triggered if it passes the specified p_T threshold (given by the number in the trigger name) and the isolation requirement (last part of the trigger name).

7.2.3 Object and event preselection

Events also have to pass the object preselection summarised in Table 7.3. In the object preselection quality requirements on different physics objects are defined. The subsequent event preselection listed in Table 7.4. has the aim to identify events with a $H \to \tau_\ell \tau_h$ signature (see Section 5.3). For the identification only those physics objects which are selected by the object preselection are considered. The criteria for the object preselection, e.g. the requirement on the electron ID, are looser compared to the requirements used in the event preselection.

Light leptons

Electrons and muons originating from τ decays are an essential part of the signal signature. Muons are reconstructed by the STACO algorithm (see Section 4.3.4) and have to be identified as *loose*, which implies *combined* as well as *segmented-tagged* muons. The muons exceed a transverse momentum of $p_T(\mu) > 10$ GeV and are situated in the pseudorapidity range of $|\eta| < 2.5$. Further quality criteria on the ID track which is associated with the muon candidate are also applied to achieve a precise measurement of the muon momentum and a reduction of the misidentification rate [80]. Electron candidates are reconstructed using energy clusters in the electromagnetic calorimeter which are matched to tracks in the ID (see Section 4.3.5). They have to pass the LoosePP identification, have a transverse energy of larger than 15 GeV and be within $|\eta| < 2.47$. Electrons in the transition region between barrel and endcap calorimeter, i.e. $1.37 < |\eta| < 1.52$ (crack-region), are rejected since the reconstruction and identification efficiency in this region is low.

Jets

Jets reconstructed with the anti-k_t clustering algorithm (Section 4.3.6) and a distance parameter of $R = 0.4$ are selected if they have a transverse momentum larger than 30 GeV and a pseudorapidity $|\eta| < 4.5$. In order to reduce the number of jets from additional proton-proton collisions from the same or neighbouring bunch crossing the jet vertex fraction in the 7 TeV analysis is required to be $|JVF| > 0.75$ for jets within $|\eta| < 2.4$. The conditions in the 8 TeV dataset are relaxed to $|JVF| > 0.5$ for jets with $p_T < 50$ GeV to account for the increased pileup. Contaminations from $t\bar{t}$ processes are reduced by a b-veto (see Section 4.3.6).

τ_{had} candidates

All τ_{had} candidates reconstructed according to the description in Section 4.3.7 need to pass the Medium identification and their transverse momentum must exceed $p_T > 20$ GeV. The calorimeter cluster and the leading track must satisfy $|\eta| < 2.47$. Only candidates with one ore three (charged) associated tracks are considered. The summed electric charges of these tracks must be ± 1.

Overlap removal

It is possible that the so far identified objects geometrically overlap. If two objects are found to be within $\Delta R < 0.2$, only one of them is used for the further analysis. Muons have the highest reconstruction purity and hence are always kept preferred to other objects. The overlap removal (OLR) for the other objects is performed in the following order of priority which is driven by the reconstruction purities: electron, τ_{had} candidates and jets. Objects which have lower priority are discarded, if they overlap with an object of higher priority. In the case of an overlap removal between muons and τ_{had} candidates, the transverse momentum threshold of the muon is reduced to 2 GeV in order to remove muon$\to \tau_{\text{had}}$ fakes.

Table 7.3: Summary of the selection criteria for the different physics objects. For electrons and muons the criteria are split into *object-selection* which refers to the selection used for the overlap removal and *event-selection* which summarises criteria used to select events in the analysis preselection.

Object-selection	
Object	Criteria
Muon	Combined or segmented $p_T(\mu) > 10\,\text{GeV}$ (2 GeV for OLR with τ_{had}) $\lvert\eta\rvert < 2.5$ ID quality criteria
Electron	Medium identification $p_T(e) > 15\,\text{GeV}$ $\lvert\eta\rvert < 2.47$ Not in crack-region

Event-selection	
Object	Criteria
Muon	Combined $p_T(\mu) > 26\,\text{GeV}$ ($p_T(\mu) > 22\,\text{GeV}$ for 7 TeV) $I(p_T, 40) < 0.06$ and $I(E_T, 0.2) < 0.06$
Electron	Tight identification $p_T(\mu) > 26\,\text{GeV}$ ($p_T(\mu) > 25\,\text{GeV}$ for 7 TeV) $I(p_T, 40) < 0.06$ and $I(E_T, 0.2) < 0.06$
Jet	LC topo $p_T(jet) > 30\,\text{GeV}$ $\lvert\eta\rvert < 4.5$ $\lvert JVF\rvert > 0.5$ for jets with $p_T < 50\,\text{GeV}$ ($\lvert JVF\rvert > 0.75$ for jets within $\lvert\eta\rvert < 2.4$ for 7 TeV)
τ_{had}	BDT medium identification $p_T(\tau_{\text{had}}) > 20\,\text{GeV}$ $\lvert\eta\rvert < 2.47$ One or three tracks Charge ± 1 Electron veto in $e\tau_{\text{had}}$ channel

Event preselection

In the next step, $H \to \tau_\ell \tau_h$ characteristic object combinations are selected, while part of the background discussed in Section 5.4 is already rejected. The object identification requirements are tightened after the overlap removal in order to address different background contributions and compositions, e.g. additional isolation requirements are introduced to suppress background originating from misidentified b- or c-jets. Muons are now required to pass the combined identification, while electrons have to pass the tight identification criteria TightPP. In addition, isolation criteria are introduced to reduce background contaminations due to hadronic jets. They are the same for both electrons and muons and for both datasets (7 TeV and 8 TeV): $I(p_T, 40) < 0.06$ and $I(E_T, 0.2) < 0.06$. Next, in the $e\tau_{\text{had}}$ channel, the electron veto algorithm is tightened to Medium and applied on τ_{had} candidates. The preselection requires exactly one lepton and one τ_{had} candidate to suppress events from $Z \to \ell\ell$ decays and to ensure orthogonality to the two other $H \to \tau\tau$ analysis channels (see Section 8.2 and Section 8.1). They are also required to have opposite charge and to pass the transverse momentum thresholds summarised in Table 7.2. Requiring the transverse mass $m_T(\ell, E_T^{\text{miss}})$ to be below 70 GeV already rejects a large part of the W+jets background. Further, a b-jet veto is applied to all jets with $p_T(\text{jet}) > 30$ GeV. If the event contains such a vetoed jet, the event is discarded. Finally, events are rejected, if the $m_{\tau\tau}$ mass reconstructed with the MMC does not converge. This requirement is introduced in order to be consistent with the $H \to \tau_h \tau_h$ and $H \to \tau_\ell \tau_\ell$ analysis channels. In the case of Higgs and Z boson decays, the mass is reconstructed with an efficiency of almost 100%. Since the reconstruction efficiency is lower for other background sources, this requirement mainly concerns background processes with fake τ_{had} signatures.
Due to the high statistics the preselection level is appropriate to validate the embedding procedure as discussed in Chapter 6.

Background contributions at the preselection level

Although the cut on the transverse mass rejects part of the W+jets background, the main purpose of the analysis preselection remains the selection of events with a $H \to \tau_\ell \tau_h$ signature. Figure 7.2 illustrates comparisons at preselection level between collision data and the background model (see Section 7.3) for some variables introduced in Section 5.6. They demonstrate that the used background model describes the data well within the investigated phase space region. All distributions have significant background contributions from $Z \to \tau\tau$, W+jets and processes with misidentified τ_{had} candidates. In addition, contributions from $Z \to \ell\ell$+jets, diboson and top processes are visible. Also shape differences between the different signal and background processes are visible. These differences are used to separate signal from background: Figure 7.2a shows the distribution of the invariant $\tau\tau$ mass calculated with the MMC, which peaks for the background processes at lower masses than for the signal process. The signal peak is at the high mass tail of the $Z \to \tau\tau$ background. For that reason, the modelling of the $Z \to \tau\tau$ high mass tail is of high relevance. The transverse mass (Figure 7.2b) is particularly suitable to suppress the W+jets background or same sign events, i.e. events with fake τ_{had} signatures (see also Section 7.3.6), which are expected to have a large transverse mass. For signal events, the $\Delta R(\ell, \tau_{\text{had}})$ distribution (Figure 7.2c) shows a clear correlation between the direction of the visible τ decay products, while the distribution is broader for background sources like multijet events[6]. The transverse momentum ratio in Figure 7.2d is another example for a variable which provides separation power against the $Z \to \tau\tau$ background. As expected background sources like W+jets or same sign events reveal a negative $E_T^{\text{miss}}\phi$ centrality value in Figure 7.2e. Figure 7.2f demonstrates that in the signal processes mainly small values of $\sum p_T$ are measured. The discrimination power of these and other variables will be used in Section 7.6.1 to classify the background processes in the BDT.

[6] The multijet background is included in the estimate of same sign events.

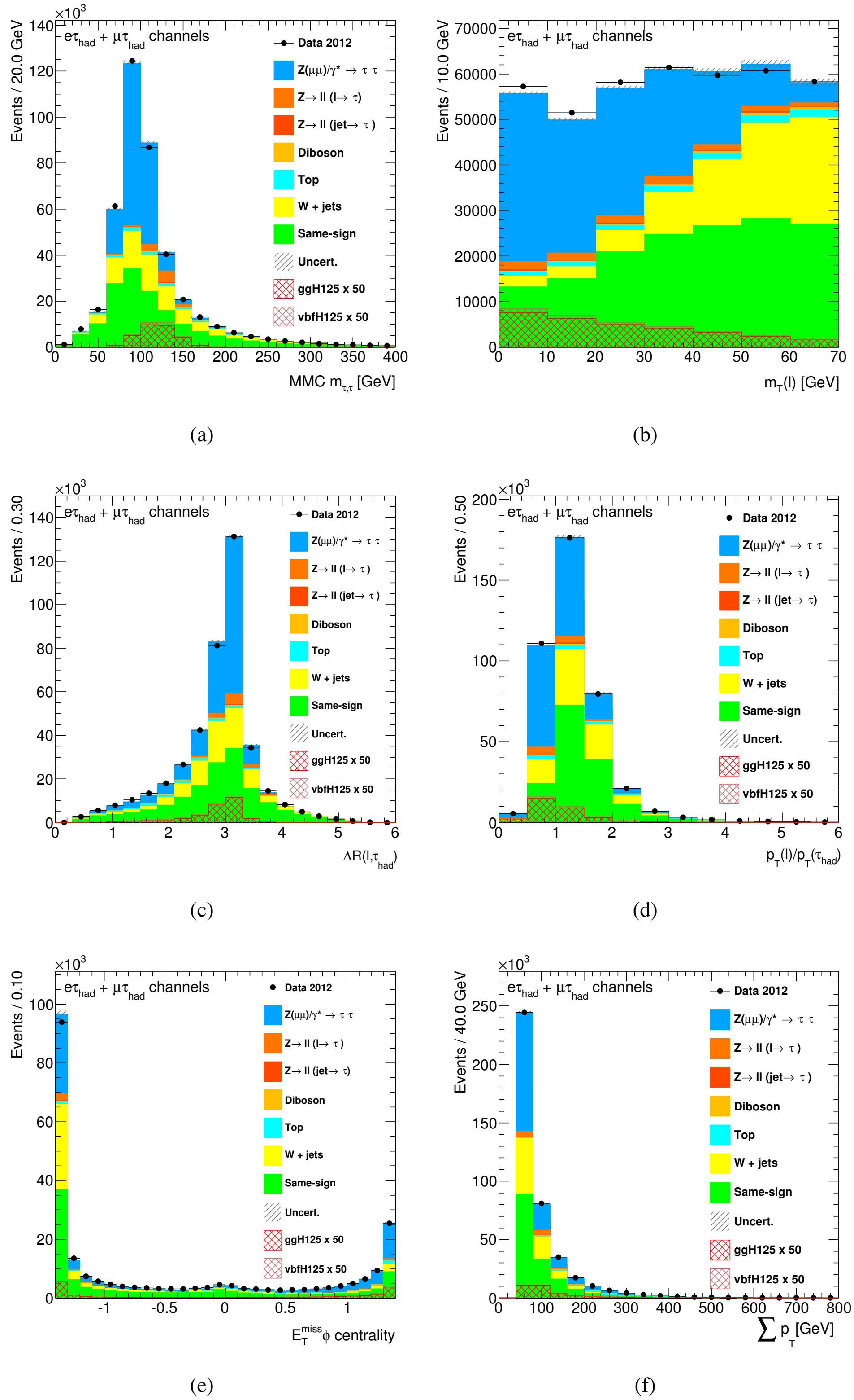


Figure 7.2: Comparisons of data with the combined background model at preselection level of the $\sqrt{s}$ =8 TeV analysis. Distributions of variables are shown which are used to separate signal from background: (a) invariant $\tau\tau$ mass calculated with the MMC, (b) transverse mass, (c) ΔR between lepton and τ_{had} candidate, (d) ratio between transverse momenta of the visible τ decay products, (e) $E_T^{\mathrm{miss}}\phi$ centrality, (f) $\sum p_T$. The background is modelled according to Section 7.3. Pure shape comparisons between signal and background processes can be found in Figure 5.16.

7.2.4 Event categorisation

In the final step, events are categorised in order to separate phase space regions which are either enriched by signal events with dominantly high-momentum Higgs bosons produced via ggF (boosted category) or events with Higgs bosons produced via the VBF mechanism. The two categories[7], summarised in Table 7.4, exploit characteristic event signatures of both production modes:

- The VBF category requires at least two jets with p_T(leading jet) > 50 GeV and p_T(subleading jet) > 30 GeV and a difference in pseudorapidity of $|\Delta\eta(j_1, j_2)| > 3.0$. Further, the visible $\tau\tau$ mass $m_{\text{vis}}^{\tau\tau}$ must exceed 40 GeV to suppress Z boson decays. This selection enhances the contributions of the VBF production compared to other mechanisms. However, due to the large cross-section of the ggF mechanism the latter still contributes significantly in this category. VBF contributes with about 64.5% to the total signal yield, ggF with 35.5%.
- Events failing the VBF selection but containing a boosted Higgs candidate with $p_T^H > 100$ GeV are collected in the boosted category which is defined to be orthogonal to the VBF category. Here, the VBF Higgs production mechanism contributes with about 18% to the total signal yield, ggF with 65.5%.

In order to significantly reduce the background contamination[8] and to improve the background-to-signal ratio in these categories, discriminating variables need to be exploited either by cutting on them - and in doing so introducing sub-categories (see Section 8.4) - or using them for the training of a multivariate discriminator (see Section 7.6).
The modelling of the discriminating variables is validated in several control regions which are constructed for each category and are enriched in one of the background processes. The exact definition of those regions is given in the following section.

Table 7.4: Summary of the $H \rightarrow \tau_\ell \tau_h$ event preselection and categorisation.

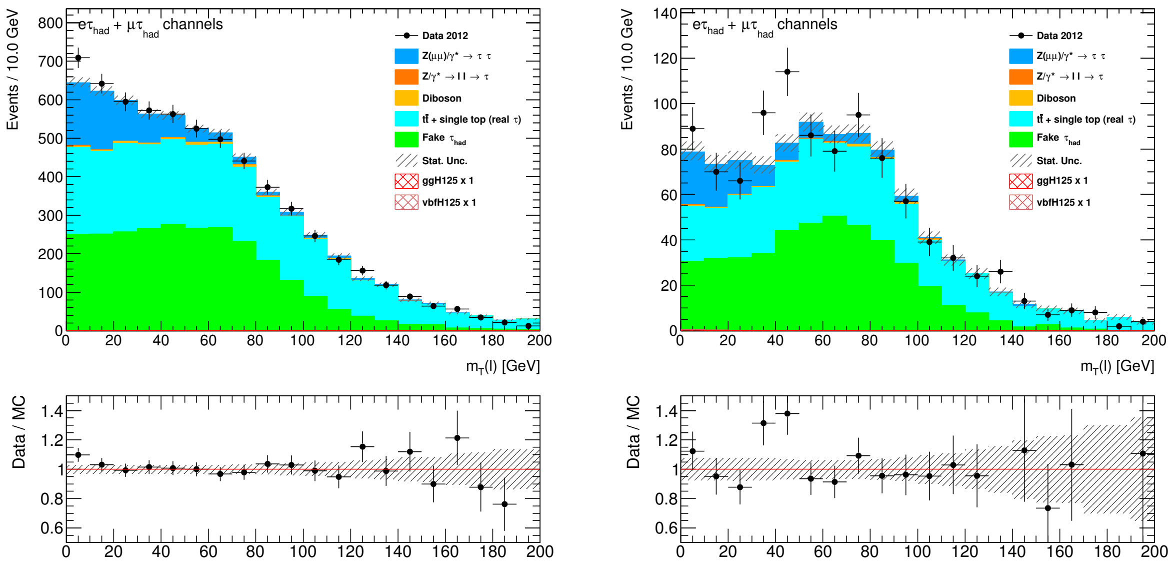
Preselection	Selection cuts
	Exactly one isolated lepton and one τ_{had} candidate with opposite charge $m_T(\ell, E_T^{\text{miss}}) < 70$ GeV Veto events with b-tagged jets with p_T(jet) > 30 GeV
Category	**Selection cuts**
VBF	At least two jets with p_T(leading jet) > 50 GeV and p_T(subleading jet) > 30 GeV $|\Delta\eta(j_1, j_2)| > 3.0$ $m_{\text{vis}}^{\tau\tau} > 40$ GeV
Boosted	Failing VBF selection $p_T^H > 100$ GeV

[7] The signal categories are also referred to as signal regions SR.
[8] Although the categorisation provides some background rejection, it is mainly used to split events into different phase space regions.

7.2.5 Control regions

For each category described above, several control regions are derived in which the contribution of a certain background is enhanced. Within these regions the background modelling is validated. They are also used to extract normalisation factors for the different background sources. Four control regions summarised in Table 7.5 are dedicated to the most dominant background processes:

- **Top control region:** For the top background two control regions are used. Both regions are defined as the corresponding signal region (boosted or VBF category) but the requirement on b-tagged jets is inverted, i.e. at least one b-tagged jet is required. While one control region requires $m_T < 70$ GeV, in the other one events have to exceed $m_T > 70$ GeV. Figure 7.3 which shows the transverse mass spectrum for both 8 TeV signal regions with inverted b-tagged requirement motivates the choice of the m_T threshold: The low-m_T region is enriched with events in which a jet is misidentified as τ_{had} candidate, while in the high-m_T region events with a real τ_{had} candidate dominate. Due to this separation the region below $m_T < 70$ GeV is used to measure the misidentification rate for jets faking the τ_{had} candidate in $t\bar{t}$ events, whereas the region with $m_T > 70$ GeV is used to validate the modelling and to derive normalisation of simulated $t\bar{t}$ events containing a real τ_{had} object (see Section 7.3.4).



(a) (b)

Figure 7.3: Distributions of the transverse mass in the boosted (left) and VBF (right) category for the 8 TeV dataset. The b-veto is inverted, i.e. at least one b-tagged jet is required. The $t\bar{t}$ and single top events which contain a true τ_{had} signature are illustrated in turquoise, while events with a fake τ_{had} signature are illustrated in green.

- $Z \to \tau\tau$ **control region:** The definition of a pure $Z \to \tau\tau$ control region is difficult due to the almost identical kinematic signature as for the signal process. An upper threshold on the reconstructed mass ($m_{\tau\tau}^{\text{MMC}} < 110$ GeV) reduces potential contaminations by signal events. The fraction of $Z \to \tau\tau$ events is further enhanced by suppressing W boson decays with a cut on the transverse mass $m_T < 40$ GeV. This control region is relevant for the validation of the embedding procedure; in particular, this region is important for the modelling of fundamental kinematic observables (see Section 7.4.1) with this data-driven method. Other validation techniques for the $Z \to \tau\tau$ background modelled with the embedding technique are described in Section 6.7.

- **Fake τ_{had} control region:** The same selection as for the signal region is used, but the requirement on the charge correlation between the τ decay products is inverted, since for misidentified jets no charge correlation is expected.
- **W+jets control region:** The W+jets control region is given by the signal region with an inverted cut on the transverse mass $m_T > 70$ GeV.

Table 7.5: Control regions of the $H \to \tau_\ell\tau_h$ analysis.

Control region	Boosted category	VBF category
Top	At least one b-tagged jet $m_T > 70$ GeV	At least one b-tagged jet $m_T > 70$ GeV
$Z \to \tau\tau$	$m_T < 40$ GeV $m_{\tau\tau}^{\mathrm{MMC}} < 110$ GeV	$m_T < 40$ GeV $m_{\tau\tau}^{\mathrm{MMC}} < 110$ GeV
Fake	Same sign τ decay products	Same sign τ decay products
W+jets	$m_T > 70$ GeV	$m_T > 70$ GeV

7.3 Background model

The relevant background processes for $H \to \tau_\ell\tau_h$ searches were introduced in Section 5.4. The $Z \to \tau\tau$ background and events with a fake τ_{had} signature are mainly estimated via data-driven methods in order to reduce systematic uncertainties. Other processes are estimated by Monte Carlo simulation. The Monte Carlo samples used in the analysis are listed in Appendix B. This section focuses on the modelling techniques used for the different background processes.

7.3.1 The $Z \to \tau\tau$ background

The irreducible $Z \to \tau\tau$ background is the dominant process. Its modelling is essential for the $H \to \tau_\ell\tau_h$ analysis. For that reason, the embedding method, described in Chapter 6, is used to model this background. The normalisation factor of the embedded events is derived separately for the $\mu\tau_{\mathrm{had}}$ and $e\tau_{\mathrm{had}}$ channel in the visible mass window 40 GeV$< m_{\mathrm{vis}} < 70$ GeV after subtracting all other background sources. In the final fit the normalisation is allowed to float freely.
$Z \to \tau\tau$ MC samples are only used for cross-checks of the embedding technique (see Section 6.7.3) and for the training of the BDTs to maximise the available statistics in the training and testing.
Due to a cut on the visible mass which needed to select $Z \to \mu\mu$ input events for the embedding technique no low-mass $Z \to \tau\tau$ events are embedded. Instead, the low-mass Drell-Yan processes below $m_{\tau\tau}^{\mathrm{MMC}} < 40$ GeV are modelled by Monte Carlo simulation.

7.3.2 The $Z \to \ell\ell$ background

The part of the $Z \to \ell\ell$ background in which a jet fakes the hadronic τ lepton is estimated by the fake factor method described in Section 7.3.5. $Z \to \ell\ell$ events in which an electron or muon fakes the τ_{had} candidate are estimated from Monte Carlo simulation. The simulated $Z \to \ell\ell$ events have to be corrected, e.g. in order to account for the electron → τ_{had} misidentification rate. Comparisons of jet-related distributions in $Z \to \ell\ell$ data and MC reveal an overall good modelling except for the dijet mass and the difference in pseudorapidity between the two leading jets. In order to correct for the mismodelling a reweighting function is derived (see Figure 7.4) in which the $\Delta\eta_{jj}$ distribution in MC is weighted to match the corresponding one in data. This reweighting method needs to be applied to all $Z \to \ell\ell$ samples with electron → τ_{had} (muon → τ_{had}) fakes. A systematic uncertainty of ± 10% is assigned to the correction which corresponds to the largest observed difference in the cut efficiency of the tight VBF selection between data and background.

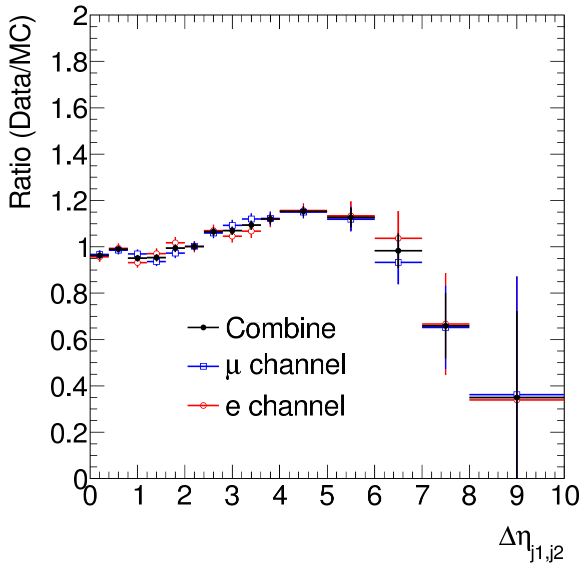


Figure 7.4: Corrections for $Z \to \ell\ell$ events as a function of the difference in pseudorapidity for the two leading jets. The reweighting function is shown for the $\mu\tau_{\mathrm{had}}$, $e\tau_{\mathrm{had}}$ and the combined channel [127].

7.3.3 The diboson background

The diboson background is taken from MC. As described in Section 6.3, the embedded $Z \to \tau\tau$ samples are contaminated by diboson events. To avoid double counting diboson events which pass the selection for embedding input events (see Section 6.1) on truth level are rejected.

7.3.4 The $t\bar{t}$ background

As for the $Z \to \ell\ell$ background two cases are distinguished: While single top and $t\bar{t}$ events in which a jet fakes the hadronic τ candidate are estimated from data as described in Section 7.3.5 events which contain a true hadronic τ lepton are taken from Monte Carlo simulation. The MC samples need to be corrected by a factor, which is derived in the high-m_T top control regions. This correction factor corresponds to the ratio between data and top Monte Carlo in the top control region after subtracting all other remaining background contributions (i.e. also the top background with a jet faking the τ_{had} candidate). The factors obtained for both categories and datasets are summarised in Table 7.6. The correction factor is allowed to float freely in the final fit. As described in Section 6.1 the overlap between $t\bar{t}$ events embedded events is removed.

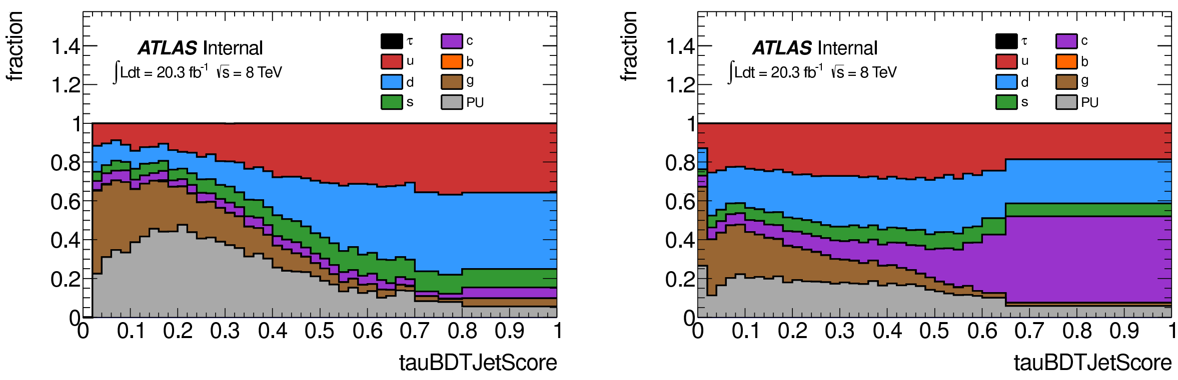
Table 7.6: Pre-fit normalisation factors and their uncertainties (defined in Section 7.7) for top background events with a real $\tau_{\rm had}$ object or an electron/muon misidentified as $\tau_{\rm had}$ candidate. The correction factors are obtained for the VBF and boosted category for the $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV dataset.

	Boosted category	VBF category
$\sqrt{s}$ =7 TeV	1.12±0.14	1.44±0.36
$\sqrt{s}$ =8 TeV	0.96±0.04	0.84±0.08

7.3.5 Fake $\tau_{\rm had}$ background

QCD, W+jets, Top (jet $\to \tau_{\rm had}$ fake) and $Z \to \ell\ell$(jet $\to \tau_{\rm had}$ fake) contribute if a jet is misidentified as $\tau_{\rm had}$ candidate. The different fake processes are estimated simultaneously in the VBF and boosted category by one single data-driven method: the so-called fake factor (FF) method. Although this method was validated in earlier analyses, for the final Run 1 analysis, it has been further optimised and extended to further fake background processes.

For the FF method a control sample is defined by events passing the selection of the VBF or boosted category but failing the $\tau_{\rm had}$ identification. These events are called anti-τ events. In addition, a lower cut on the BDT τ ID is introduced for these events in order to account for the dependence of the quark/gluon composition on the τ BDT score. This dependence is shown in Figure 7.5 for one and three prong



(a)

(b)

Figure 7.5: Fraction of one prong (left) and three prong (right) anti-τ candidates originating from the hadronisation of light or heavy quarks, gluons and from pileup jets as a function of the BDT τ ID distribution. The figures are taken from [127].

candidates: The low BDT score region is dominated by pileup jets or jets originating from gluons. The fraction of quark jets increases with higher BDT score values. Anti-τ events have to pass the requirement of 0.7xLoose. Loose is a τ BDT working point provided by the τ performance group [93]. Its exact value depends on the transverse momentum of the $\tau_{\rm had}$ candidate. The threshold for anti-τ events is a bit smaller than the Loose working point in order to achieve a good compromise between statistics and aligning the quark/gluon composition in anti-τ events with events passing the $\tau_{\rm had}$ identification of the analysis.

The number of fake background events $N_{\mathrm{bkg}}^{\mathrm{fake}}$ is given by

$$N_{\mathrm{bkg}}^{\mathrm{fake}} = \left(N_{\mathrm{data,SR}}^{\mathrm{anti}-\tau} - \sum_{\mathrm{others}} N_{\mathrm{others,SR}}^{\mathrm{anti}-\tau} \right) \cdot FF_{\mathrm{CR}} \,, \tag{7.1}$$

where $N_{\mathrm{data,SR}}^{\mathrm{anti}-\tau}$ denotes the number of anti-τ events in data in the signal region. The sum over $N_{\mathrm{others,SR}}^{\mathrm{anti}-\tau}$ includes all processes containing a true τ lepton, which does not pass the τ_{had} identification as well as events with an electron/muon faking the τ_{had} candidate. The fake factor FF_{CR} is calculated for each event in a dedicated control region according to

$$FF_{\mathrm{CR}} = \frac{N_{\mathrm{CR}}^{\text{identified-}\tau}}{N_{\mathrm{CR}}^{\mathrm{anti}-\tau}} \,. \tag{7.2}$$

$N_{\mathrm{CR}}^{\text{identified-}\tau}$($N_{\mathrm{CR}}^{\mathrm{anti}-\tau}$) denotes the number of events with one identified τ_{had} candidate (anti-τ) in the control region. FF_{CR} depends on the transverse momentum, number of tracks and on the origin of the jet $\rightarrow \tau_{\mathrm{had}}$ fake. In general, quark-initiated jets are more likely identified as hadronic τ_{had} candidates than gluon-initiated jets. Since the quark/gluon jet composition strongly depends on the specific background process, the fake factor has to be measured for the different background sources separately. The different control regions used to measure the misidentification rate FF_{CR} for different processes are summarised in Table 7.7.

When applying FF_{CR} on anti-τ data events, the actual process causing the τ_{had} fake is unknown. For that reason, a fake factor is derived which considers the relative contribution of each of the four processes to the anti-τ data sample:

$$\begin{aligned} FF_{\mathrm{CR}} &= \sum_{i=bkg} R_i \cdot FF_i \\ &= (R_{W+\mathrm{jets}} \cdot FF_{W+\mathrm{jets}}) + (R_{\mathrm{Z+jets}(j\rightarrow\tau_{\mathrm{had}})} \cdot FF_{\mathrm{Z+jets}(j\rightarrow\tau_{\mathrm{had}})}) \\ &\quad +(R_{\mathrm{Top}(j\rightarrow\tau_{\mathrm{had}})} \cdot FF_{\mathrm{Top}(j\rightarrow\tau_{\mathrm{had}})}) + (R_{\mathrm{QCD}} \cdot FF_{\mathrm{QCD}}) \,, \end{aligned} \tag{7.3}$$

where R_i denotes the relative contribution of the individual background i. The fraction of Top(jet $\rightarrow \tau_{\mathrm{had}}$) and $Z \rightarrow \ell\ell$(jet $\rightarrow \tau_{\mathrm{had}}$) are estimated from Monte Carlo simulation, whereas the other contributions, W+jets and multijet events, are estimated from data via

$$R_{W+\mathrm{jets}} = \frac{N_{W+\mathrm{jets}}}{N_{W+\mathrm{jets}} + N_{\mathrm{QCD}}} \,. \tag{7.4}$$

Table 7.7: Definition of control regions used to measure the fake factors.

Background	Control region
QCD	SR with relaxed lepton isolation, i.e. inverted track isolation and relaxed calorimeter isolation
W+jets	SR with inverted m_T cut
$Z \rightarrow \ell\ell$(jet $\rightarrow \tau_{\mathrm{had}}$)	SR with inverted dilepton veto
Top(jet $\rightarrow \tau_{\mathrm{had}}$)	SR with inverted b-jet veto and $m_T < 70\,\mathrm{GeV}$

The number of W+jets events is estimated from data in the W+jets control region extrapolated into the signal region:

$$N_{W\text{+jets}} \quad = \quad \frac{N^{\text{data}}_{\text{anti}-\tau_{\text{had}},\text{W-CR}} \cdot N^{W\text{+jets,MC}}_{\text{anti}-\tau_{\text{had}}}}{N^{W\text{+jets,MC}}_{\text{anti}-\tau_{\text{had}},\text{W-CR}}} \,, \tag{7.5}$$

whereas the number of QCD events is estimated from the data signal region after subtracting the estimated number of W+jets event and all other processes

$$N_{\text{QCD}} \quad = \quad N^{\text{data}}_{\text{anti}-\tau_{\text{had}}} - \left(N_{W\text{+jets}} + \sum_{other} N^{\text{other MC}}_{\text{anti}-\tau_{\text{had}}} \right) . \tag{7.6}$$

The contribution of multijet events R_{QCD} is estimated by

$$R_{\text{QCD}} \quad = \quad 1 - R_{W\text{+jets}} - R_{\text{Top}} - R_Z \, . \tag{7.7}$$

The measured fractions R_i for $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV are summarised in Table 7.8. For both datasets and signal categories the largest contribution arises from W+jets processes, followed by multijet events. The contributions from $Z \to \ell\ell(\text{jet} \to \tau_{\text{had}})$ and Top(jet $\to \tau_{\text{had}}$) are small in comparison.
The final fake factors are illustrated as functions of the transverse momentum in Figure 7.6. The resulting fake factors for $\sqrt{s}$ =8 TeV (Figure 7.6a and Figure 7.6b) as a function of $p_T(\tau_{\text{had}})$ vary from 0.15 (0.09) to 0.06 (0.04) for one (three) prong candidates in the boosted category. In the VBF category the factors range from 0.12 (0.08) to 0.09 (0.04). The statistics available for the derivation of the fake factors in the $\sqrt{s}$ =7 TeV dataset (Figure 7.6c and Figure 7.6d) is smaller, which results in large statistical uncertainties. Due to the low statistics the lower BDT cut for anti-τ_{had} events is dropped for the $\sqrt{s}$ =7 TeV fake factors.
The BDT cut which is part of the definition of anti-τ events is an improvement of the FF method used in earlier analyses, since it reduces the contribution of anti-τ events originating from gluon or pileup jets. Hence, their composition is closer to the composition of identified τ_{had} candidates. Compared to earlier analyses, the method is extended to the $Z \to \ell\ell$ and top background. Although the jet $\to \tau_{\text{had}}$ background is dominated by W+jets and multijet events, the contributions of $Z \to \ell\ell$ and top background are statistically large enough to estimate them also from data rather than MC. The estimation from data is favoured, since it reduces systematic uncertainties.

Table 7.8: Measured fraction R_i and its uncertainty for different background processes for $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV in the boosted and VBF category.

Background	Boosted (7 TeV)	VBF (7 TeV)	Boosted (8 TeV)	VBF (8 TeV)
$Z \to \ell\ell(\text{jet} \to \tau_{\text{had}})$	0.06±0.03	0.13±0.07	0.05±0.03	0.11±0.06
Top(jet $\to \tau_{\text{had}}$)	0.06±0.03	0.03±0.02	0.07±0.04	0.03±0.02
W+jets	0.75±0.38	0.60±0.3	0.62±0.31	0.46±0.23
QCD	0.13±0.07	0.24±0.12	0.26±0.13	0.40±0.20

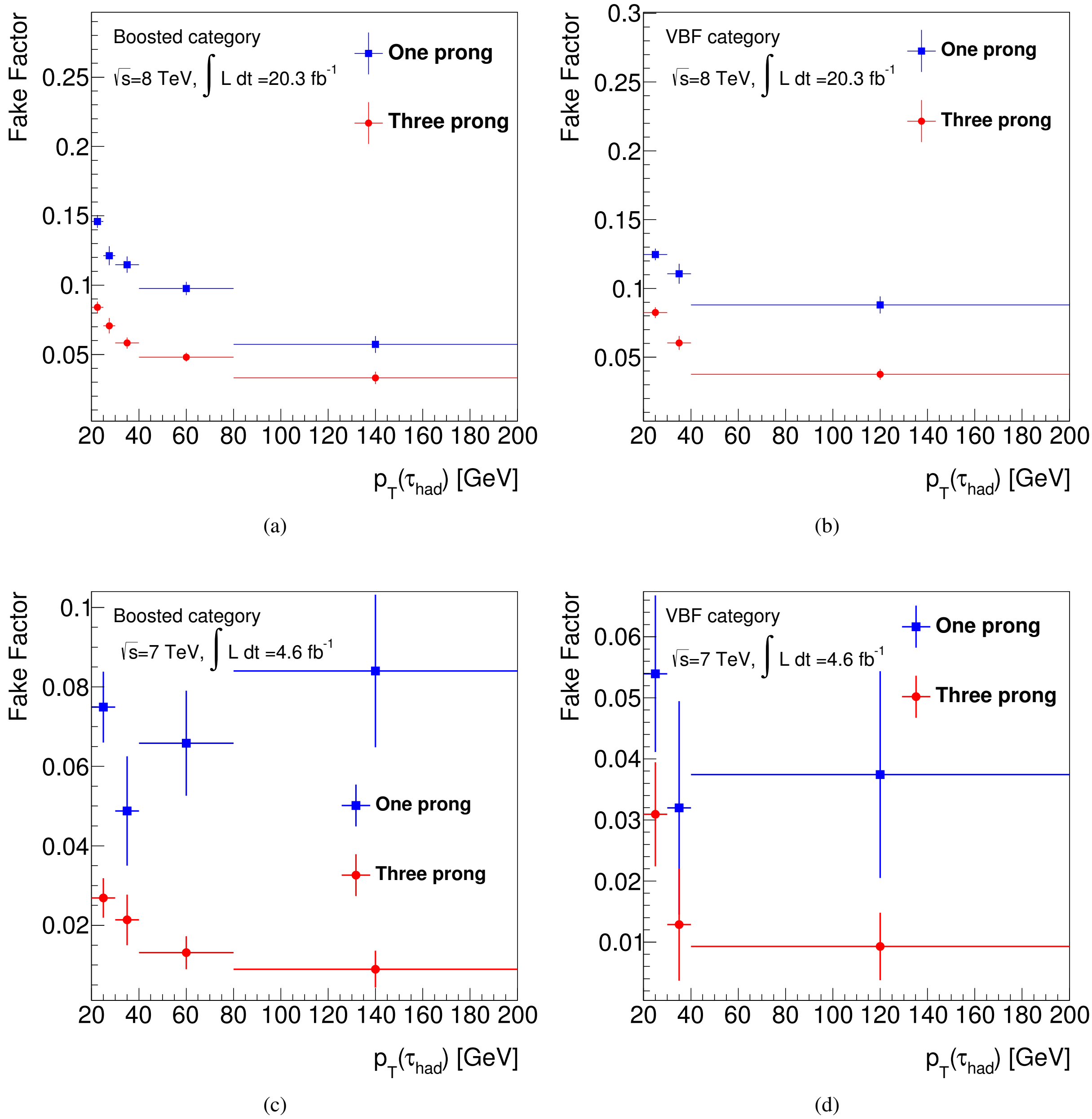


Figure 7.6: Fake factors as a function of the transverse momentum of the τ_{had} candidate. These factors are shown for the boosted (left) and VBF (right) category as well as for the $\sqrt{s}$ =8 TeV (top) and $\sqrt{s}$ =7 TeV (bottom) dataset. The functions are derived separately for one (blue) and three (red) prong candidates. Only the statistical error is displayed.

Anti-correlation between FF_{Top} and correction factor for Top (real τ_{had})

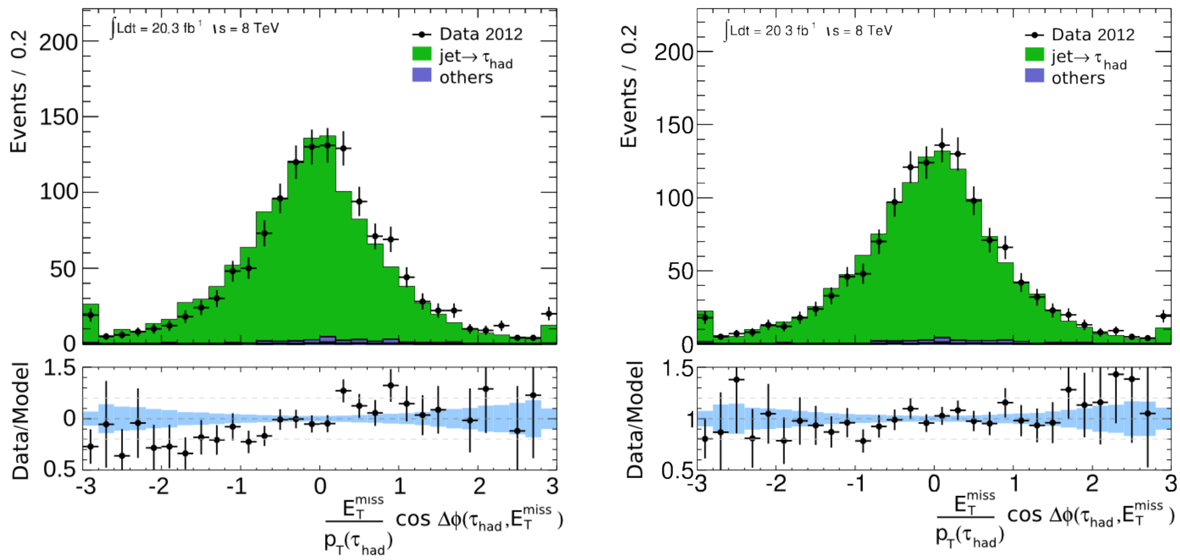
The control region used to derive the misidentification rate for top events FF_{Top} is significantly contaminated by events with a real τ_{had} object Top(real τ_{had}). These events are normalised as explained in Section 7.3.4. In order to obtain the number of misidentified τ_{had} candidates the contribution from Top(real τ_{had}) has to be subtracted. The subtraction is done without any additional normalisation correction. In principle, this treatment causes an anti-correlation between the fake factor FF_{Top} and the correction factor for Top(real τ_{had}). By floating FF_{Top} and propagate it to the combined fake background model the impact of the anti-correlation on the final likelihood fit result has been studied. The impact on the top background yield and the significance is in the per mill range and therefore negligible.

E_T^{miss} calculation in anti-τ_{had} events

In the reconstruction algorithm of E_T^{miss} anti-τ_{had} objects are not correctly calibrated. While jets misidentified as τ_{had} candidates are treated as real τ leptons, for anti-τ_{had} candidates the jet energy calibration is used which results in a mismodelling of E_T^{miss}-related variables. In particular, the projection of E_T^{miss} along the τ_{had} momentum direction

$$\frac{\vec{E}_T^{\text{miss}} \cdot \hat{p}_T(\tau_{\text{had}})}{|\vec{p}_T(\tau_{\text{had}})|} \tag{7.8}$$

is incorrectly modelled as illustrated in Figure 7.7a. For this reason, in the $\sqrt{s}$ =8 TeV analysis the E_T^{miss} is re-calculated requiring that the anti-τ_{had} object is treated as τ_{had} object. Figure 7.7b shows how the recalculation improves the modelling of the above mentioned projection. For the $\sqrt{s}$ =7 TeV analysis such a re-calculation is not available. Hence, these events are reweighted according to the distribution

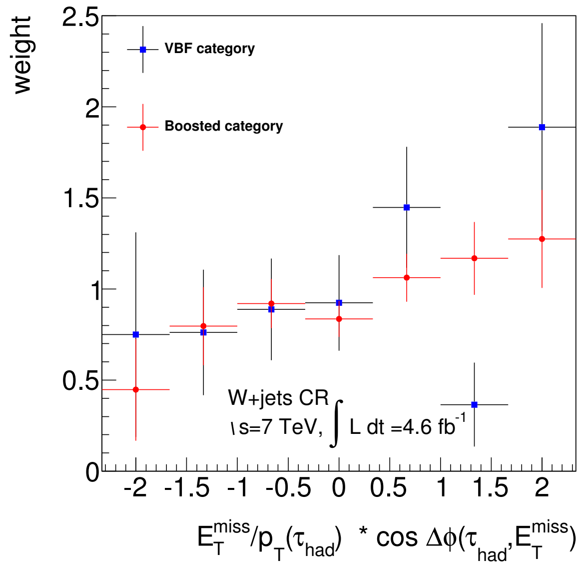


(a)

(b)

Figure 7.7: Projection of E_T^{miss} along the τ_{had} momentum direction before and after recomputing E_T^{miss} in the VBF fake τ_{had} control region.

of the $E_T^{\rm miss}$ on $\tau_{\rm had}$ projection. The weights (see Figure 7.8) are derived in the W+jets control region which contains many jets misidentified as $\tau_{\rm had}$ candidates. For the reweighting procedure a conservative uncertainty of ±100% is assigned to the weights, because the derivation of these weights is only based on events in the W+jets control region and is done in just one dimension.



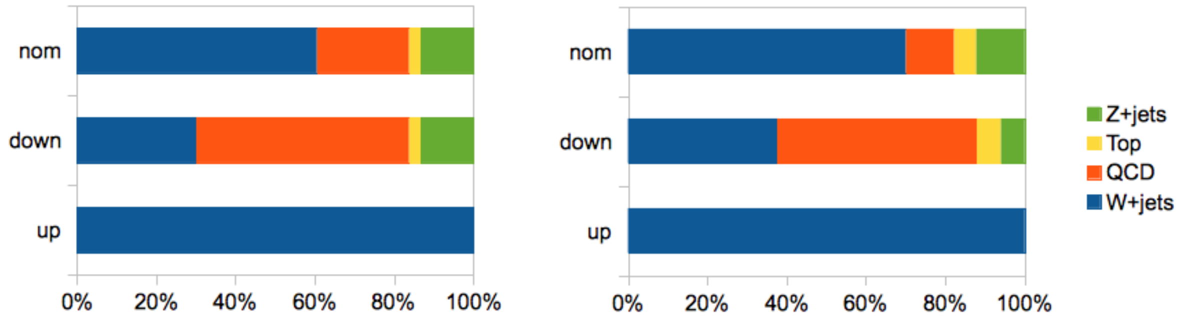
(a)

Figure 7.8: Weights and their statistical uncertainties derived in the W+jets control region. They are used to correct anti-$\tau_{\rm had}$ events in the $\sqrt{s}$ =7 TeV dataset. The weights are shown as a function of the projection of $E_T^{\rm miss}$ on the anti-$\tau_{\rm had}$ candidate for the boosted (red) and VBF (blue) category.

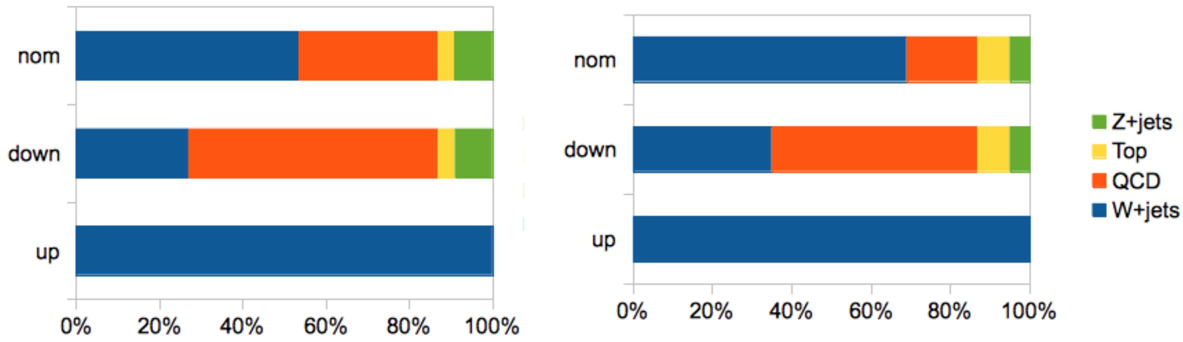
Sources of systematic uncertainties

For the fake factor method four sources of systematic uncertainties are identified corresponding to an overall impact on the background estimate of ±4.7% (±3.1%) in the boosted (VBF) category:

- **Statistical uncertainty:** The derivation of the fake factors for the different processes is influenced by the available data statistics in the different control regions. The statistical error is calculated for each factor FF_i.
- **Background composition uncertainty:** The final fake factor applied on anti-$\tau_{\rm had}$ events considers the fake composition (see Equation 7.3) which is mainly estimated from Monte Carlo simulation. To estimate the uncertainty due to the measured fraction R_i, each fraction is varied by ±50%. Figure 7.9 illustrates how the variation of R_i changes the overall fake composition for the boosted and VBF category for $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV. In the case of the +50% variation, the background composition is completely dominated by W+jets events.
- **MC closure:** In order to validate the fake factor procedure itself and to identify possible biases due to the method, a MC closure test is used in which the number of anti-$\tau_{\rm had}$ candidates in simulated control samples (W+jets, $Z \to \ell\ell$ and $t\bar{t}$) is used to estimate the number of events of the corresponding Monte Carlo sample in the signal region. The MC closure test is shown for the MMC mass distribution in Figure 7.10. For the $\sqrt{s}$ =8 TeV analysis (Figure 7.10a and



(a)



(b)

Figure 7.9: Composition of jet→ τ_{had} fakes in the VBF (left) and boosted (right) category for $\sqrt{s}$ = 7 TeV (top) and $\sqrt{s}$ = 8 TeV (bottom). The compositions are estimated with the fake factor method. They are given for nominal R_i (upper bar), the −50% down (middle bar) and the +50% up (lower bar) variation.

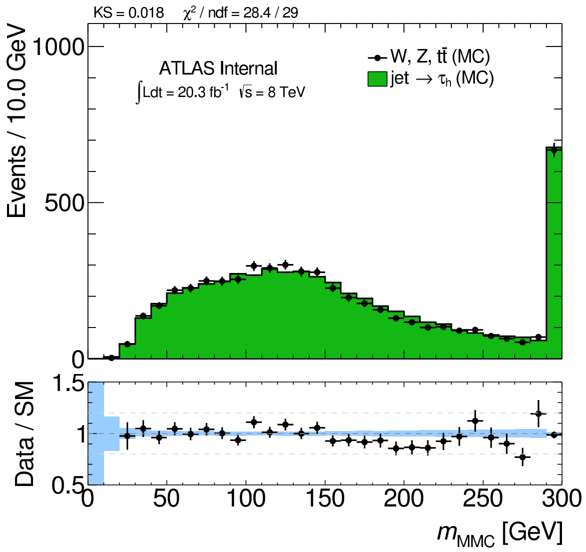
Figure 7.10b) good agreement within statistical uncertainties between expected number of events from Monte Carlo simulation and fake factor method is observed, while for the $\sqrt{s}$ =7 TeV analysis (Figure 7.10c and Figure7.10d) in particular in the boosted category a discrepancy is visible for which an uncertainty of ±10% is assigned to the fake factors. The mismodelling might be caused by potential imperfections of the E_T^{miss}-on anti-τ_{had} reweighting method described earlier, since the calculated MMC mass depends on the relative direction of E_T^{miss} and the τ_{had} candidate.

- **Closure test in fake τ_{had} control region:** In order to check for any remaining mismodelling due to the method, which is not covered by the uncertainties described so far, the fake background model (including the rest of the background model) is compared to data in the fake τ_{had} data control region (see Section 7.4.1) dominated by events with a jet misidentified as τ_{had} candidate. A good agreement is observed, so that no additional uncertainty has to be assigned.

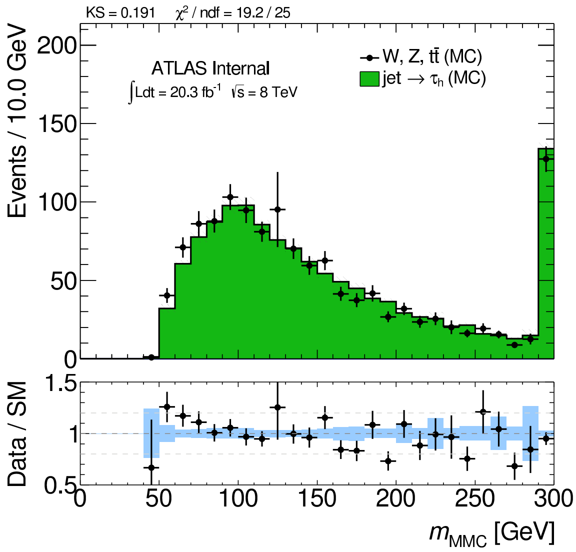
7.3.6 Estimate of fake background via the opposite sign-same sign method

In order to estimate the background due to jets misidentified as τ_{had} candidates the so-called opposite sign-same sign (OS-SS) method can be used. It has been part of the background model in earlier $H \to \tau\tau$ analyses and is in detail described in the corresponding documentations [114, 134]. This method is statistically limited due to the number of data events in the fake τ_{had} control regions. For that reason, it can only be used to estimate the background in regions with large data statistics like the preselection.

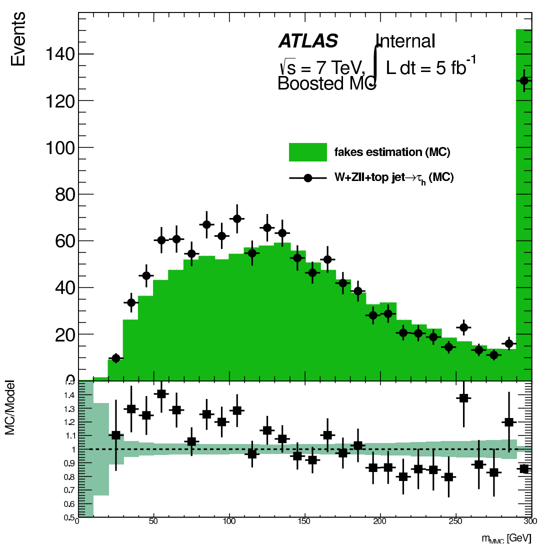
For the two signal regions, in which a rather small number of events is found, the OS-SS method is not suitable. Instead, the fake factor method is chosen as the appropriate method to estimate this background source. In the context of the presented analysis, this method is only used for the background estimate at preselection level and for the BDT training.



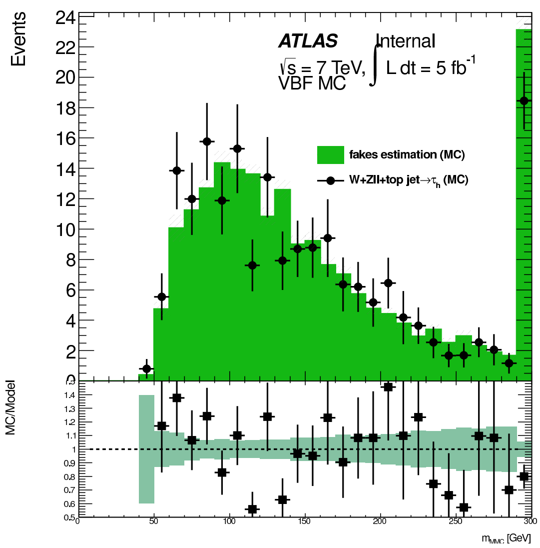
(a)



(b)



(c)



(d)

Figure 7.10: Monte Carlo closure test of the fake factor method showing the MMC distribution in the boosted (left) and VBF (right) signal region estimated from Monte Carlo (black) and fake factor method (green) for $\sqrt{s}$ =8 TeV (top) and $\sqrt{s}$ =7 TeV (bottom). The figures are taken from [127].

7.4 Validation of the background model

The background model has to be validated before making statements about background contributions in data. While a detailed validation of the BDT input variables and of the BDT score distribution with respect to the combined background model is performed in the signal and control region in Section 7.6.2, here the modelling of some basic kinematic observables is investigated.

7.4.1 Validation of the data-driven part of the background model

The larger part of the background model is determined in a data-driven way. The data-driven background model is validated in two control regions.

$Z \to \tau\tau$ background

For the embedding method several validation methods have been developed which address different aspects of the procedure. In Section 6.7 which is dedicated to more general validation methods independent of a particular use case in one of the various analyses it is confirmed that the modelling of $Z \to \tau\tau$ decays based on $Z \to \mu\mu$ data events works well. In the final validation step, the τ embedded $Z \to \mu\mu$ collision data events have to be validated within the analysis as part of the combined background model in the Z-enriched control region. Since other background processes also contribute significantly it is not a clean and stand-alone validation of the embedding method. Moreover, the complete background model involving other background estimation procedures is tested as described above.
Some examples for embedding-relevant quantities in the boosted (left-hand-side plots) and VBF (right-hand-side plots) are displayed in Figures 7.11, 7.12 and 7.13: The transverse momentum of the hadronically decaying τ lepton and visible $\tau\tau$ mass in Figure 7.11 are examples for the kinematics of the visible Z decay products. The reconstructed E_T^{miss} and $m_{\tau\tau}^{\text{MMC}}$ are shown in Figure 7.12. The transverse momentum of the Z boson and leading jet are given in Figure 7.13. Since these plots are used to validate the $Z \to \tau\tau$ background, the composition of the other background processes like fake τ leptons or diboson events are not of interest. Thus, they are summarised into one background illustrated in red. The low mass Drell Yan processes with $\tau\tau$ final states are taken from MC and shown in light blue. The τ-embedded events, illustrated in dark blue, are normalised to data in a dedicated region documented in Section 7.2.5. Within uncertainties (defined in Section 7.7), the distributions of the ATLAS data are well described by the combined background model including the dominant embedding-based $Z \to \tau\tau$ model. From these comparisons and the other validation results one can conclude, that τ-embedded data provide a reliable model of the $Z \to \tau\tau$ background.

Fake background

The fake factor method is validated in the fake τ_{had} control region which is dominated by events with a jet misidentified as τ_{had} candidate. Potential mismodelling due to this method would be visible in this region and can be addressed by assigning an additional uncertainty, covering the remaining differences. In Figure 7.14 the transverse momentum of the τ_{had} candidate and $m_{\tau\tau}^{\text{MMC}}$ are illustrated for the $\sqrt{s}$ =8 TeV VBF and boosted category. The distributions show no indication for such a mismodelling. Hence, no additional systematic is assigned.

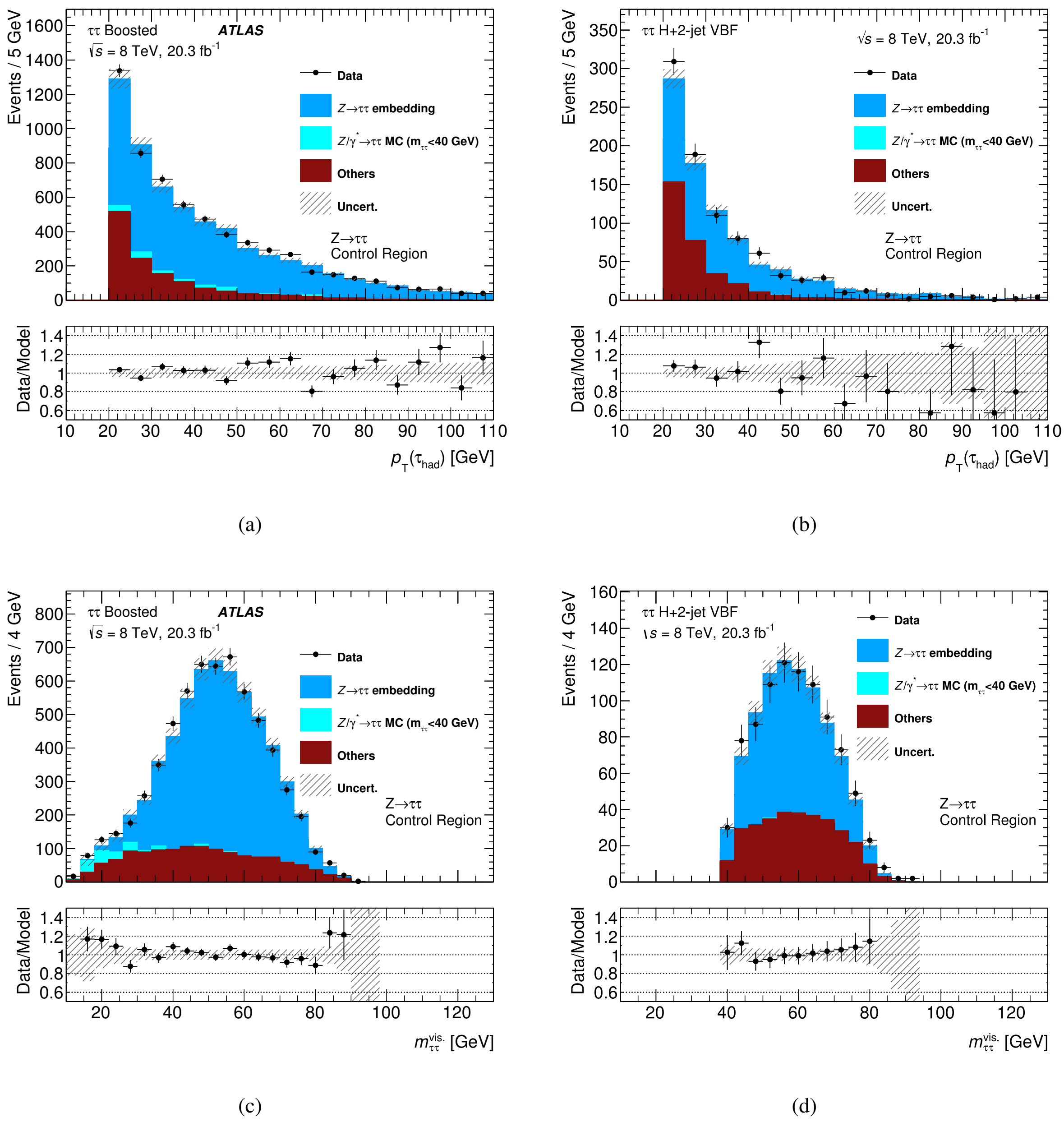


Figure 7.11: Comparison of data with the combined background model in the boosted (left) and VBF (right) Z-enriched control region. The upper (lower) plots show the transverse momentum of the τ_{had} candidate (visible $\tau\tau$ mass). The ratios show the relative difference of the data to the total background estimate. Except for the $Z \rightarrow \tau\tau$ background all other background processes are summarised under "others". The left-hand side figures are taken from [9].

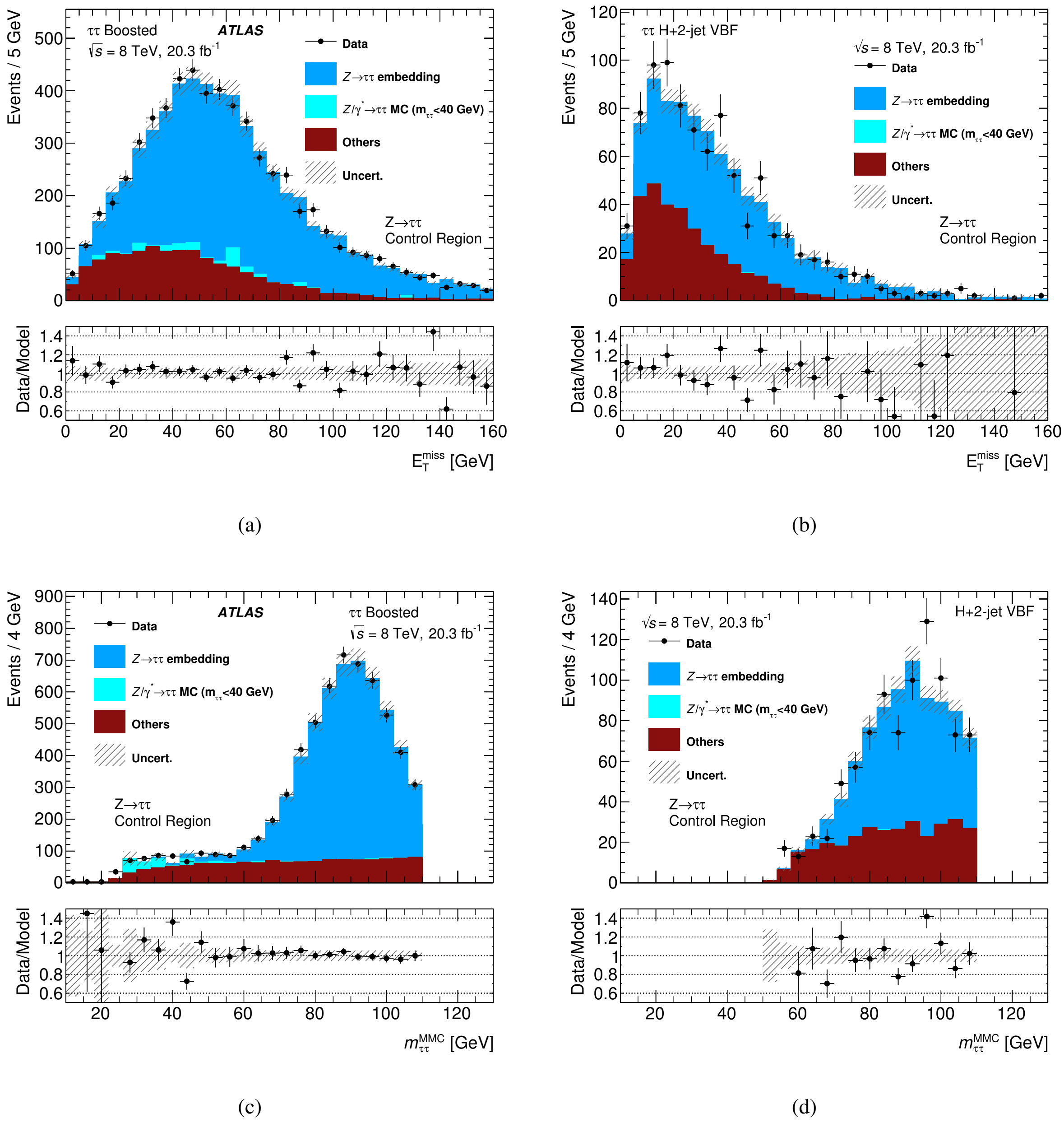


Figure 7.12: Comparison of data with the combined background model in the boosted (right) and VBF (left) Z-enriched control region. The upper (lower) plots show the missing transverse energy (invariant $\tau\tau$ MMC mass). The ratios show the relative difference of the data to the total background estimate. Except for the $Z \to \tau\tau$ background all other background processes are summarised under "others". The left-hand side figures are taken from [9].

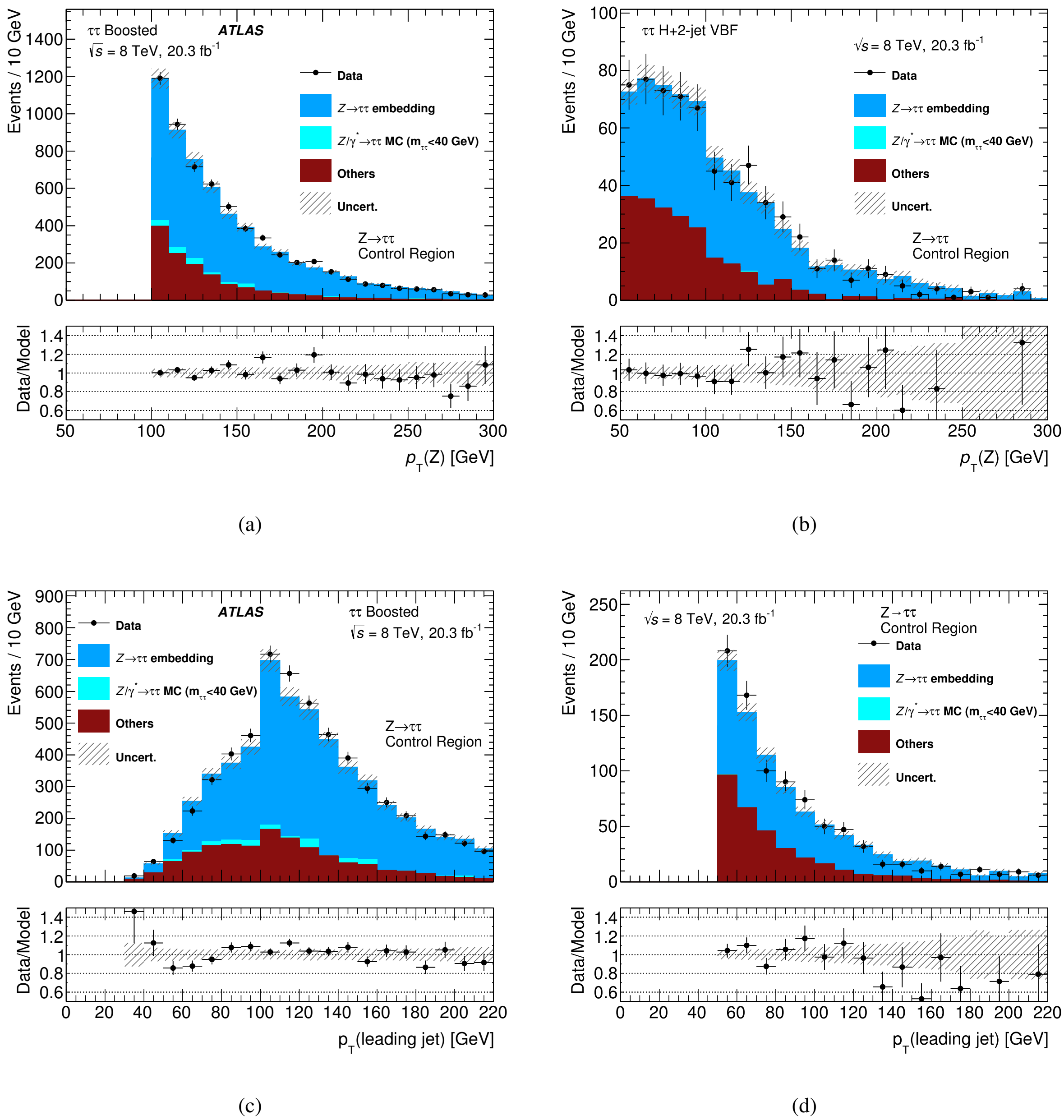


Figure 7.13: Comparison of data with the combined background model in the boosted (left) and VBF (right) Z-enriched control region. The upper (lower) plots show the transverse momentum of the Z boson (of the leading jet). The ratios show the relative difference of the data to the total background estimate. Except for the $Z \rightarrow \tau\tau$ background all other background processes are summarised under "others". The left-hand side figures are taken from [9].

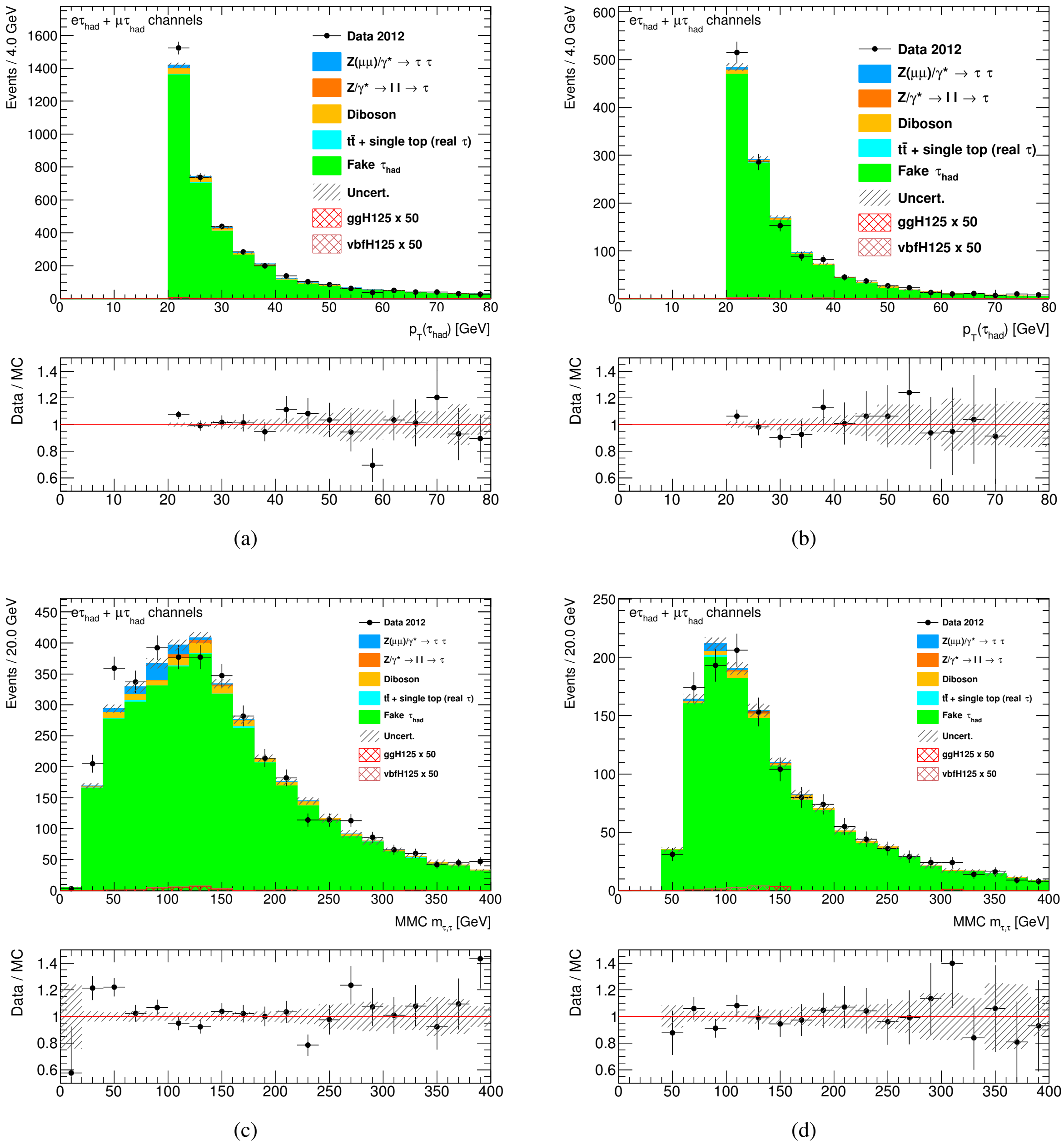


Figure 7.14: Comparison of data with the combined background model in the boosted (left) and VBF (right) fake τ_{had} control region. The ratios show the relative difference of the data to the total background estimate which is dominated by jets misidentified as τ candidates. The upper (lower) distributions show the transverse momentum of the τ_{had} candidate ($m_{\tau\tau}^{\text{MMC}}$) in the $\sqrt{s}$ =8 TeV analysis.

7.4.2 Validation of the MC simulated part of the background model

Only background processes, with small contributions in the signal regions are MC simulated. From those background processes, the top background with a true τ_{had} signature is the largest one. It is validated in the top control region which enhances the contamination of this particular background. In Figure 7.15a and Figure 7.15b (Figure 7.15c and Figure 7.15d) the transverse momentum spectrum of the light lepton (missing transverse energy) is shown for the $\sqrt{s}$ =8 TeV and $\sqrt{s}$ =7 TeV boosted category. Figure 7.16a and Figure 7.16b (Figure 7.16c and Figure 7.16d) show the same distributions in the $\sqrt{s}$ =8 TeV and $\sqrt{s}$ =7 TeV VBF category. A good agreement within uncertainties is observed.

7.4.3 Validation of the combined background model

After the validation of the background modelling in the control regions, the combined background model has to be tested in the signal categories. Figure 7.17 and Figure 7.18 show some examples of kinematic variables in the boosted and VBF signal region for the $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV analysis: In Figure 7.17a and Figure 7.17b (Figure 7.18a and Figure 7.18b) the $p_T(\tau_{\mathrm{had}})$ spectra are depicted for the $\sqrt{s}$ =8 TeV and $\sqrt{s}$ =7 TeV boosted (VBF) category. Figure 7.17c and Figure 7.17d (Figure 7.18c and Figure 7.18d) show the invariant $\tau\tau$ MMC mass. The background model agrees with the observed data within the systematic uncertainties.

7.5 Expected and observed yields in signal regions

Based on the background model described above the number of expected background events in both regions is determined and compared to the measured values in data. This comparison is done before fitting (pre-fit) and after fitting (post-fit) the BDT distribution. The pre-fit event yields for the combined background model and the individual background processes for the $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV datasets including statistical and systematic uncertainties (see Section 7.7) are summarised in Table 7.9 and Table 7.10, respectively. The observed event numbers in data match the prediction from Monte Carlo simulation in all categories within uncertainties. Since the size of the $\sqrt{s}$ =7 TeV dataset is only about a quarter of the size of the $\sqrt{s}$ =8 TeV dataset (see Table 3.1) the event yields in the $\sqrt{s}$ =7 TeV are significantly smaller. The largest contribution in the boosted category originates from the $Z \to \tau\tau$ background, while the fake τ_{had} background dominates in the VBF category of both datasets. Contributions from other background processes are small. In the boosted category, the dominant signal process is ggF, followed by VBF. Only small contributions arise from VH. As expected, in the VBF category, more signal events containing Higgs bosons produced via VBF than ggF contribute. Contributions from ggF production processes are still significant in the VBF category. Compared to the combined background yields, the number of signal events expected to enter the categories is very small. For that reason, the background has to be further suppressed in order to improve the signal-to-background ratio and to increase the signal sensitivity. This can be achieved either by introducing sub-categories by additional cuts on discriminating variables as done for the cut-based analysis (see Section 8.4) or using multivariate techniques as discussed in the following.

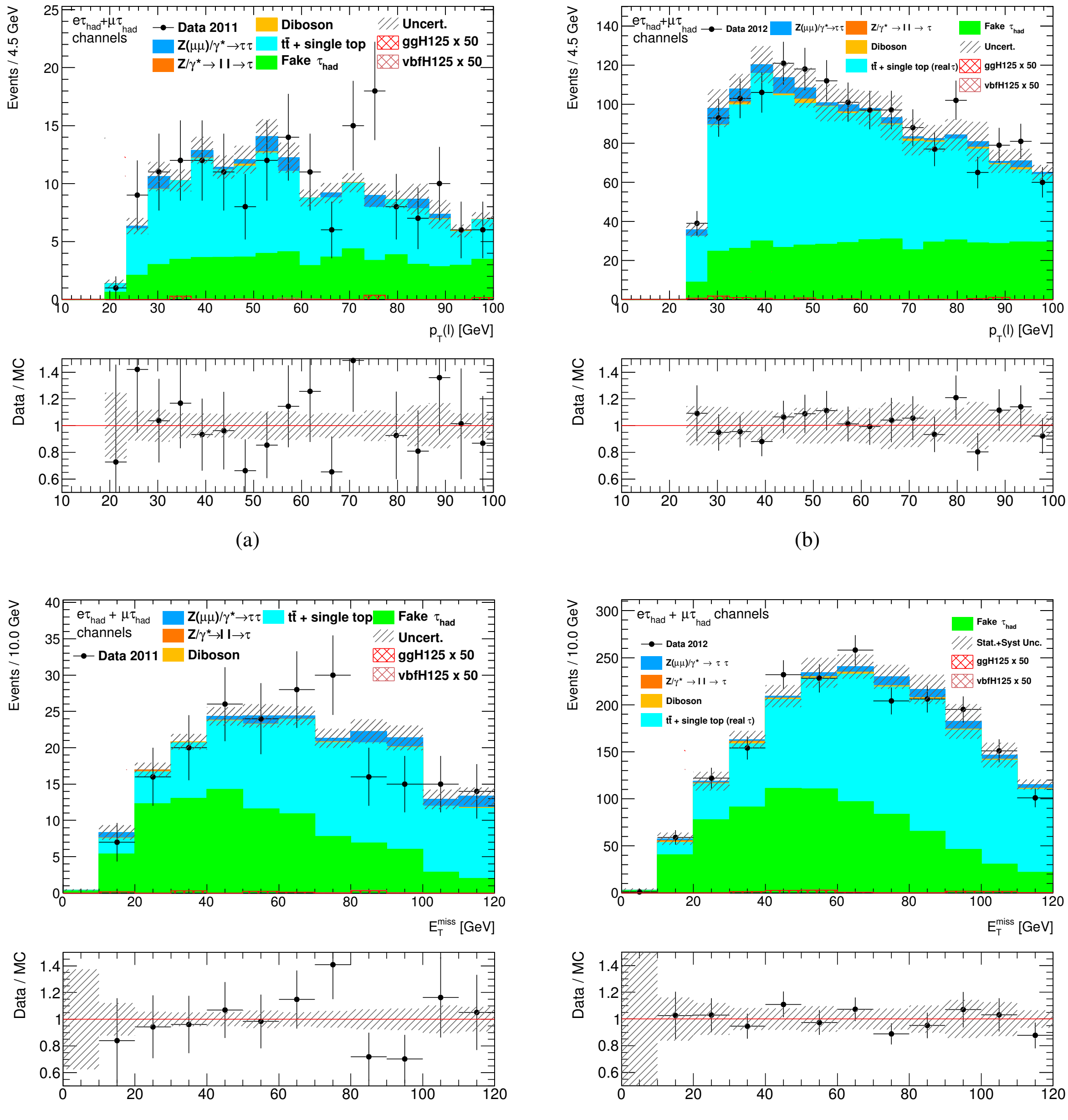


Figure 7.15: Comparison of data with the combined background model focusing on characteristic kinematic observables in the top control region of the boosted category for the $\sqrt{s}$ =7 TeV (left) and $\sqrt{s}$ =8 TeV (right) analysis. The ratios show the relative difference of the data to the total background estimate including systematic and statical errors: (a) and (b) the transverse momentum of lepton, (c) and (d) the missing transverse momentum.

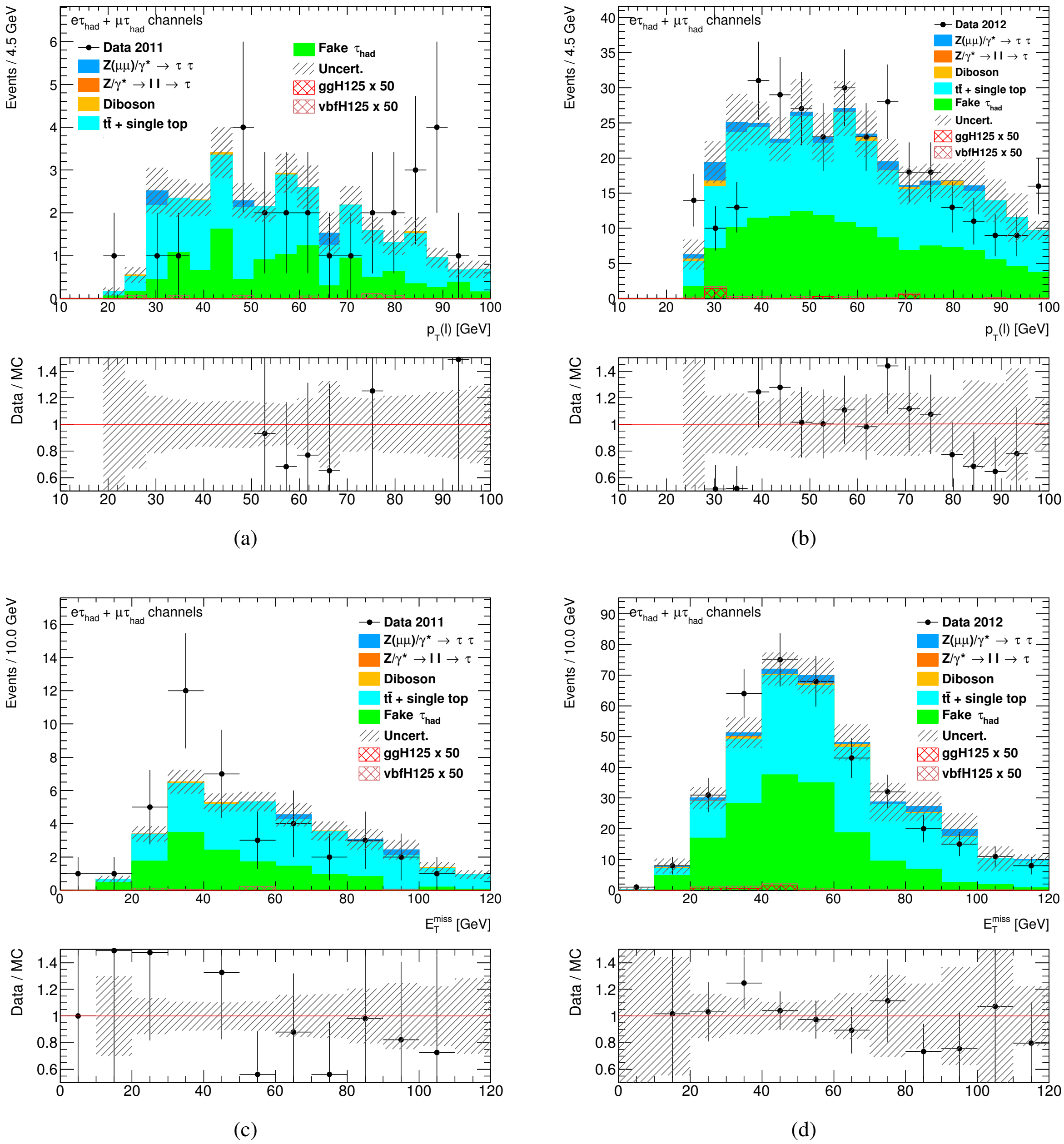


Figure 7.16: Comparison of data with the combined background model focusing on characteristic kinematic observables in the top control region of the VBF category for the $\sqrt{s}$ =7 TeV (left) and $\sqrt{s}$ =8 TeV (right) analysis. The ratios show the relative difference of the data to the total background estimate including systematic and statical errors: (a) and (b) the transverse momentum of lepton (c) and (d) the missing transverse momentum.

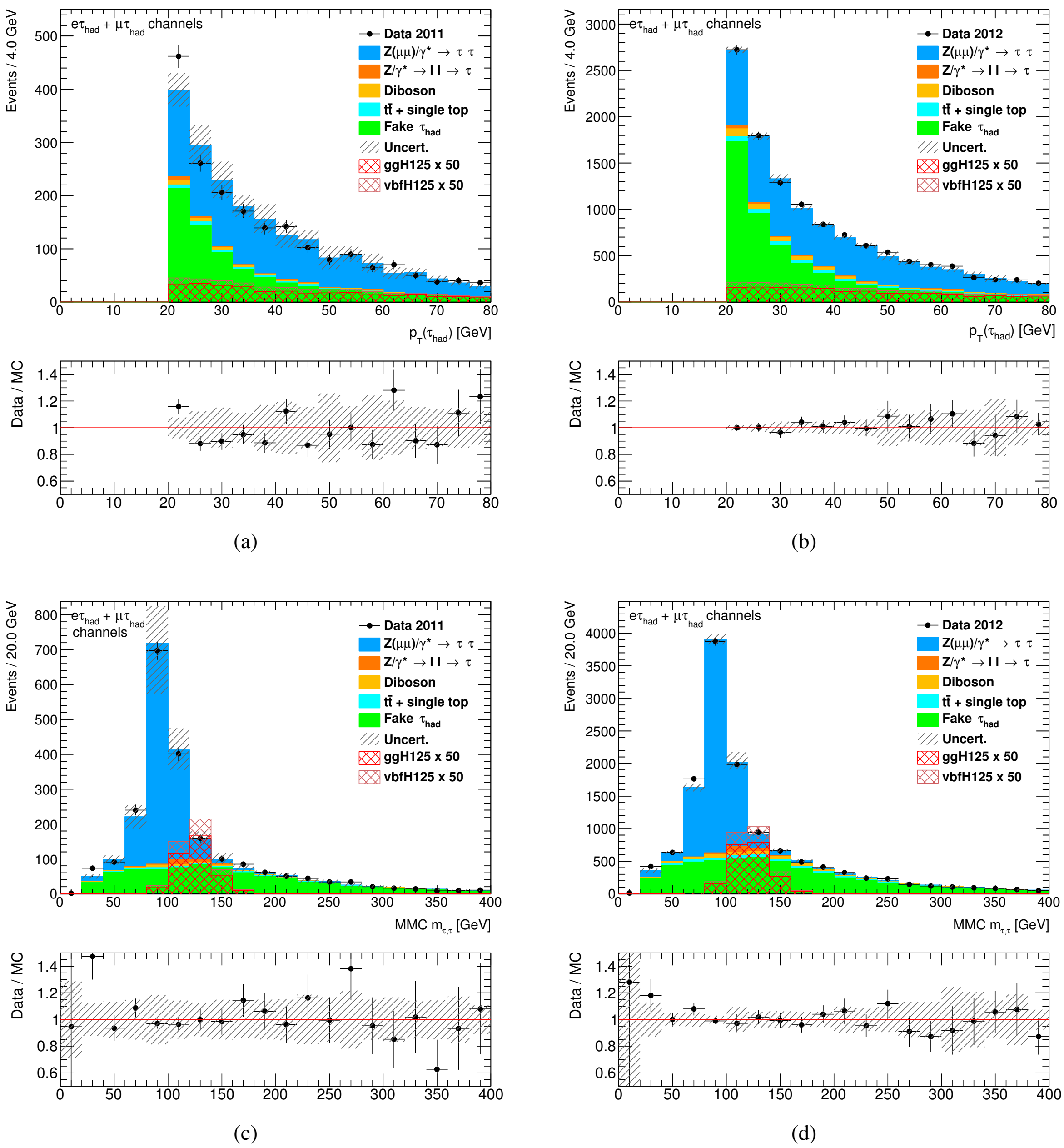


Figure 7.17: Comparison of data with the combined background model focusing on characteristic kinematic observables in the boosted signal region for the $\sqrt{s}$ =7 TeV (left) and $\sqrt{s}$ =8 TeV (right) analysis. The ratios show the relative difference of the data to the total background estimate including systematic and statical errors: (a) and (b) the transverse momentum of the τ_{had} candidate, (c) and (d) the invariant $\tau\tau$ mass calculated with the MMC.

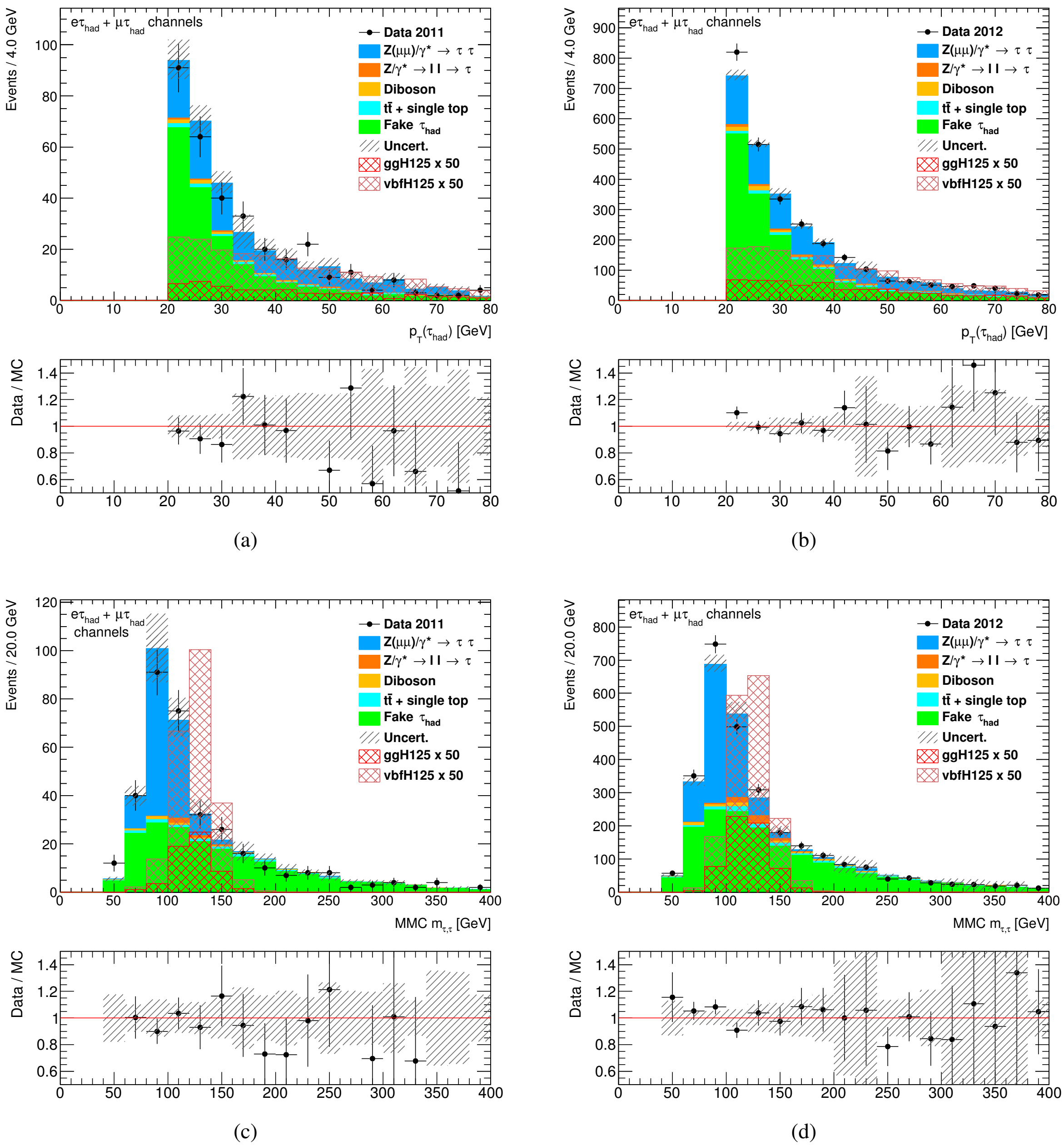


Figure 7.18: Comparison of data with the combined background model focusing on characteristic kinematic observables in the VBF signal region for the $\sqrt{s}$ =7 TeV (left) and $\sqrt{s}$ =8 TeV (right) analysis. The ratios show the relative difference of the data to the total background estimate including systematic and statical errors: (a) and (b) the transverse momentum of the τ_{had} candidate, (c) and (d) the invariant $\tau\tau$ mass calculated with the MMC.

Table 7.9: Pre-fit event yields for background and signal processes in the boosted and VBF category of the $\sqrt{s}$ =7 TeV analysis including statistical and systematic uncertainties defined in Section 7.7. The yields of the combined background model are compared to the observed yields in data.

Sample	Boosted category	VBF category
$Z \to \tau\tau$	$1231 \pm 25 \pm 201$	$136 \pm 7 \pm 37$
$Z \to \ell\ell(\ell \to \tau_{\mathrm{had}})$	$30 \pm 3 \pm 7$	$3.9 \pm 0.6^{+1.5}_{-1.2}$
Diboson	$40 \pm 1 \pm 5$	$6.5 \pm 0.4^{+1.9}_{-1.8}$
$t\bar{t}$(real τ_{had})	$51 \pm 2 \pm 5$	$10 \pm 1 \pm 2$
Fake background	$820 \pm 7 \pm 202$	$196 \pm 3^{+54}_{-54}$
VBF	$2.1 \pm 0.1 \pm 0.2$	$3.3 \pm 0.1 \pm 0.4$
ggF	$6.4 \pm 0.2^{+2.4}_{-1.9}$	$1.2 \pm 0.1^{+0.6}_{-0.5}$
WH	$1 \pm 0.05 \pm 0.11$	$0.03 \pm 0.01 \pm 0.1$
ZH	$0.51 \pm 0.03^{+0.11}_{-0.05}$	$0.02 \pm 0.01 \pm 0.02$
Total Signal	$10.0 \pm 0.2^{+2.4}_{-1.9}$	$4.6 \pm 0.1^{+0.7}_{-0.6}$
Total background	$2171 \pm 26 \pm 285$	$353 \pm 8 \pm 65$
Data	2199	349

Table 7.10: Pre-fit event yields for background and signal processes in the boosted and VBF category of the $\sqrt{s}$ =8 TeV analysis including statistical and systematic uncertainties. The yields of the combined background model are compared to the observed yields in data.

Sample	Boosted category	VBF category
$Z \to \tau\tau$	$6323{\pm}57^{+210}_{-210}$	$900{\pm}17^{+48}_{-51}$
$Z \to \ell\ell(\ell \to \tau_{\mathrm{had}})$	$213{\pm}11^{+60}_{-61}$	$56{\pm}3^{+31}_{-30}$
Diboson	$421{\pm}12^{+48}_{-48}$	$64{\pm}5^{+15}_{-14}$
$t\bar{t}$(real τ_{had})	$398{\pm}11^{+35}_{-44}$	$85{\pm}5^{+11}_{-11}$
Fake background	$5425{\pm}26^{+336}_{-336}$	$1637{\pm}13^{+72}_{-72}$
VBF	$11.0{\pm}0.1^{+0.90}_{-0.90}$	$22{\pm}0.2^{+2.47}_{-2.50}$
ggF	$40{\pm}1^{+13}_{-11}$	$12{\pm}1^{+6}_{-6}$
WH	$6.5{\pm}0.1^{+0.7}_{-0.7}$	$0.36{\pm}0.02^{+0.13}_{-0.10}$
ZH	$3.2{\pm}0.1^{+0.3}_{-0.3}$	$0.11{\pm}0.01^{+0.03}_{-1.28}$
Total Signal	$61{\pm}7.82^{+14.77}_{-13.78}$	$34{\pm}1^{+8.63}_{-9.88}$
Total background	$12779{\pm}65^{+689}_{-699}$	$2741{\pm}23^{+177}_{-178}$
Data	12952	2830

7.6 Application of BDTs in the $H \to \tau_\ell\tau_h$ analysis

While the general concept of the BDT classifier is discussed in Section 5.2, the following section is concerned with the training and evaluation of BDTs in the $H \to \tau_\ell\tau_h$ analysis.

7.6.1 The BDT training setup of the $H \to \tau_\ell\tau_h$ analysis

For both signal categories one individual BDT is trained. The training is performed on $\sqrt{s}$ =8 TeV signal and background samples. For datasets at $\sqrt{s}$ =7 TeV, no separate BDTs are trained. Instead, the BDTs trained at the higher center-of-mass are applied. Reasons, why BDTs are not re-trained for $\sqrt{s}$ =7 TeV, are elaborated on in Section 7.6.3. For the training different discriminating variables are used. They are summarised in Section 5.6.
The BDTs are trained on signal samples assuming a Higgs mass of 125 GeV. While for the VBF

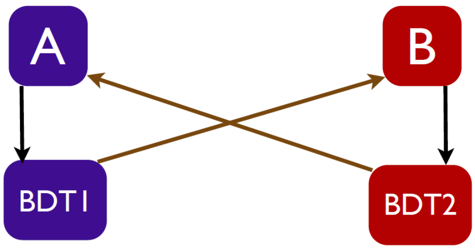


Figure 7.19: Sketch of the cross-evaluation method: BDT1 (BDT2) is trained on sample A (B). Afterwards, BDT1 (BDT2) is applied on sample B (A).

category only the VBF signal sample is considered, for the boosted category the BDT is trained for VBF, ggF and VH Higgs production mechanisms in order to improve the overall sensitivity. For the training 50% of the available statistics in background samples and 75% in signal samples is used, while the rest of the sample statistics is employed to test the BDTs. Based on these trained BDTs events are classified as signal-like or background-like within the analysis. Due to overtraining effects the BDTs must not be applied on events which were used for the BDT training. This restriction reduces the sample statistics available for physics analyses significantly. In order to be able to use 100% of the statistics for the analysis, however, a technique known as cross-evaluation is applied. This technique is illustrated in Figure 7.19: The sample is divided arbitrarily into two subsamples A and B containing the same number of events. In the next step, one BDT is trained for both samples A and B separately. The BDT trained on sample A is applied on sample B and vice versa. In the case of collision data, both BDTs are applied on one or the other half, respectively. The application of this technique assumes a similar performance of both BDTs. The actual background and signal sample size for the training are summarised in Table 7.11. For the training, the background due to jets misidentified as τ_{had} candidates is estimated by the OS-SS method (see Section 7.3.6).

BDT training variables

The BDTs are trained based on a set of variables which possess good discriminating power between signal and background processes. They are chosen from a long list of potential variables by the following procedure: For both signal categories, a list of variables is chosen. These variables must be suitable to reject background processes. Then, the number of variables is reduced by removing variables which are not well-modelled or reveal other problems. For example, variables are avoided which are strongly

Table 7.11: Sample size for different background and signal samples used for the BDT training of the boosted and VBF category.

Process	Boosted category	VBF category
$Z \rightarrow \tau\tau$	20728	34588
$Z \rightarrow \ell\ell$	868	1934
W+jets	7260	16336
$t\bar{t}$ single top	3773	8024
Diboson	1388	2376
Same Sign events	2867	7051
Total background	36884	70309
Signal (m=125 GeV)	39166	53978

correlated with other input variables. After removing variables the BDT is trained and its discrimination power is investigated. The dependence of the signal sensitivity of the analysis on the chosen variables is also studied. Variables with small impact on the BDT performance are not longer considered. After obtaining a set of basic variables new ideas for variables are tested by adding them to the BDT input list and reinvestigating the BDT performance. The variables finally used for the BDT training in the boosted and VBF category are summarised in Table 7.12.

Table 7.12: Discriminating variables for the training of the BDTs at $\sqrt{s}$ =8 TeV for the VBF and boosted category. The variable definition is given in Section 5.6.

Variable	Boosted category	VBF category
m_T	✓	✓
$m_{\tau\tau}^{\mathrm{MMC}}$	✓	✓
$\Delta R(\tau_1 \tau_2)$	✓	✓
$p_T^{\tau_1}/p_T^{\tau_2}$	✓	
$\sum p_T$	✓	
p_T^{tot}		✓
$E_T^{\mathrm{miss}}\phi$ centrality	✓	✓
$\Delta\eta(j_1, j_2)$		✓
$m_{j,j}$		✓
$\eta_{j_1} x \eta_{j_2}$		✓
$C_{\eta_1,\eta_2}(\eta_\ell)$		✓

General BDT performance studies

Different tools are employed to study the discrimination power of the trained BDTs.

- In the *importance ranking* the separation power of the variables used in the BDT training is quantified. The quantification is achieved in the following way: For each variable the nodes are identified in which the variable is used to make a decision. Based on the number of identified nodes a weight is derived. This weight accounts for the actual number of events in the corresponding node and the improvement of the Gini-Index (see Section 5.2.1). Summing up all weighted occurrences leads to a ranking value. A high ranking value means a high importance for the separation of background and signal events, while a small value is associated with a small influence of the variable on the BDT performance. Based on the ranking values the input variables can be quantitatively ordered and conclusions on their importance can be drawn. For example, if a variable has a significantly smaller ranking value than the other BDT input variables, it might be removed from the BDT training without affecting the BDT performance.

- Besides the variable ranking the investigation of correlations between the input variables are of interest. If two input variables are strongly correlated which each other they do not significantly increase the discriminating power of the BDT compared to the case in which only one of those two variables is used. It is also important to check for potential mismodelling of correlations in MC. Such mismodelled correlations can bias the BDT. The correlations between input variables were investigated and no mismodelling was observed [127].

- A good classifier is characterised by a small background misclassification rate $\epsilon_{\text{background}}$ denoting the number of background events which are misclassified as signal events. At the same time the classifier should also have a small signal misclassification rate $1 - \epsilon_{\text{signal}}$, i.e. signal events which are misclassified as background. In order to visualise the signal purity $1-\epsilon_{\text{background}}$ and the signal efficiency ϵ_{signal} of the BDT the *ROC* (receiver operating characteristic) curve is used. A sketch of the *ROC* curve is illustrated in Figure 7.20. It shows the signal purity in dependence of the signal efficiency. The best theoretical integral value is given by +1 meaning all signal and background events are correctly classified as signal and background, respectively.

The above mentioned tools, importance rankings, variable correlations and *ROC* curve, are used in this analysis in order to optimise the BDT training.

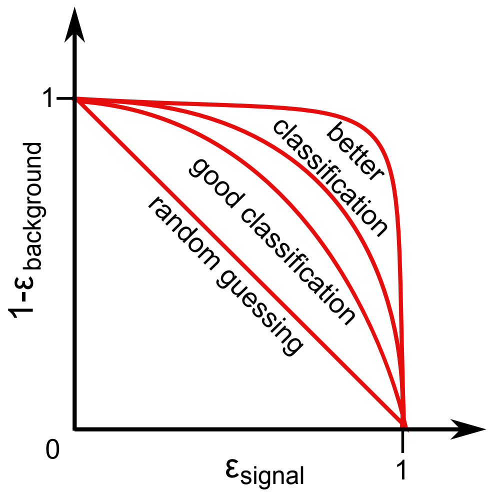


Figure 7.20: Sketch of a ROC curve. The ROC curve shows the signal purity (1-$\epsilon_{\text{background}}$) dependent on the signal efficiency ϵ_{signal}. A good classifier increases the integral below the red curve.

7.6.2 Validation of BDT score distribution and BDT input variables

Besides the above mentioned performance studies the modelling of each BDT input variable has to be validated. They are validated by comparing the estimated distributions obtained with the combined background model in Section 7.3 with the results obtained from collision data in both signal regions as well as in corresponding control regions. In the comparisons the uncertainties discussed in Section 7.7 are considered. Only a small collection of validation plots is presented in Figure 7.21, Figure 7.22 and Figure 7.23. All other validation plots can be found in Appendix C.1.

Figure 7.21a, Figure 7.22a and Figure 7.23a show the angular separation between the visible τ decay products in the boosted signal region and the corresponding fake and top control regions. The difference between background model and data for small $\Delta R(\ell\tau_{\text{had}})$ in the fake τ_{had} control region has no visible impact on the result in the signal region. This variable is well-modelled in the signal region. Good agreement is also observed in the top control region. The distance $\Delta R(\ell\tau_{\text{had}})$ is mainly used in the BDT training to discriminate signal from background processes with fake τ_{had} signatures. The modelling of the transverse momentum ratio in the boosted category (Figure 7.21b, Figure 7.22b, Figure 7.23b) agrees with the observed distribution in data. As shown before at preselection level (see Figure 7.2d) this variable has some discrimination power against the $Z \to \tau\tau$ background. Examples for the VBF category are depicted in Figure 7.21c, Figure 7.22c and Figure 7.23c. They show the $\ell - \eta$ centrality $C_{\eta_1,\eta_2}(\eta_\ell)$. As expected for signal events, the lepton lies centrally between the two jets. At small values of $C_{\eta_1,\eta_2}(\eta_\ell)$ mainly background events contribute. Figure 7.21d, Figure 7.22d and Figure 7.23d show the modelling of the difference in pseudorapidity between the two leading jets in the VBF signal and corresponding control regions. The comparisons of all BDT input variables show a good agreement between data and the background model.

In the next step, the modelling of the BDT score distribution, from which the sensitivity of the analysis is extracted, has to be validated. The validation of the BDT score is subdivided into two steps: In a first *blinded* validation the low BDT score region is investigated in which the signal contribution is negligible. Such a validation has the advantage that the analysis is *blind* with respect to the signal and hence the analysis strategy is not biased due to observations made in sensitive regions. The BDT score value up to which the BDT score distribution is *unblinded* is defined as the value at which the signal-to-background ratio exceeds 30%. After ensuring a well-modelled low BDT score region, in the next step, the other bins are *unblinded* and validated as well.

The *unblinded* BDT score distributions are summarised in Figure 7.24. Good agreement between background model and data is observed. Also in the control regions, the modelling of the BDT score distributions is in agreement with data. The BDT score distributions for the control regions are summarised in Appendix C.2.1. Figure 7.24 demonstrates that the background due to fake τ_{had} candidates dominates in the low BDT score regions and hence separates well from signal processes. The BDT score also has good discrimination between fake τ_{had} background and $Z \to \tau\tau$. This is an advantage because the $Z \to \tau\tau$ normalisation can be obtained within the fit without a significant impact of the systematic uncertainties related to the fake τ_{had} background. Due to the good separation the low BDT score region contains important information about the normalisation of the various background processes.

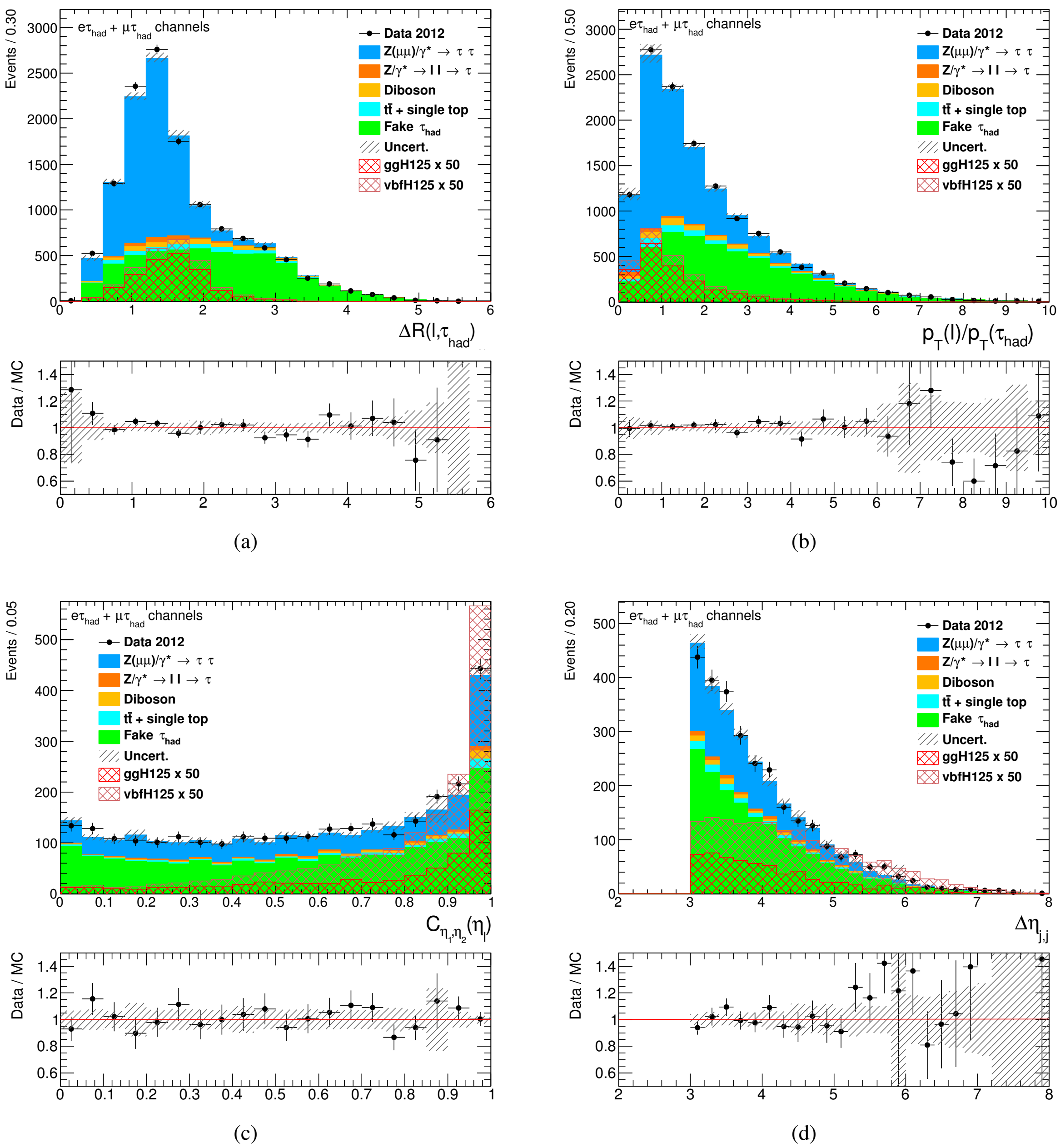


Figure 7.21: Comparison of data with the combined background model with respect to BDT input variables in the boosted (top) and VBF (bottom) signal region of the $\sqrt{s}$ =8 TeV analysis: (a) ΔR between the light lepton and the τ_{had} candidate, (b) p_T ratio between the light lepton and the τ_{had} candidate, (c) $\ell - \eta$ centrality $C_{\eta_1,\eta_2}(\eta_\ell)$ and (d) difference in pseudorapidity between the two leading jets. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

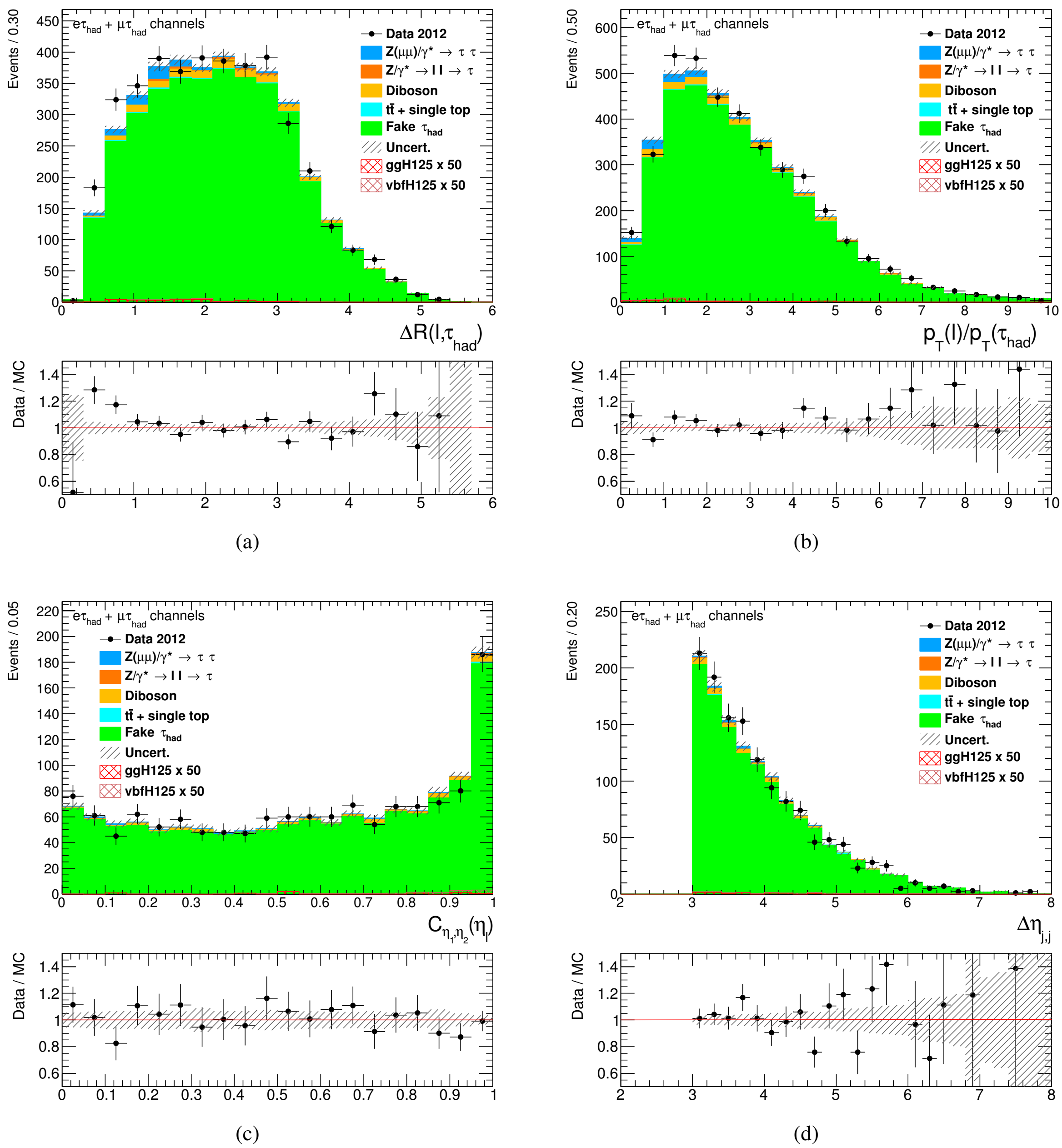


Figure 7.22: Comparison of data with the combined background model with respect to BDT input variables in the boosted (top) and VBF (bottom) fake τ_{had} control region of the $\sqrt{s}$ =8 TeV analysis: (a) ΔR between the light lepton and the τ_{had} candidate, (b) p_T ratio between the light lepton and the τ_{had} candidate, (c) $\ell - \eta$ centrality $C_{\eta_1,\eta_2}(\eta_\ell)$ and (d) difference in pseudorapidity between the two leading jets. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

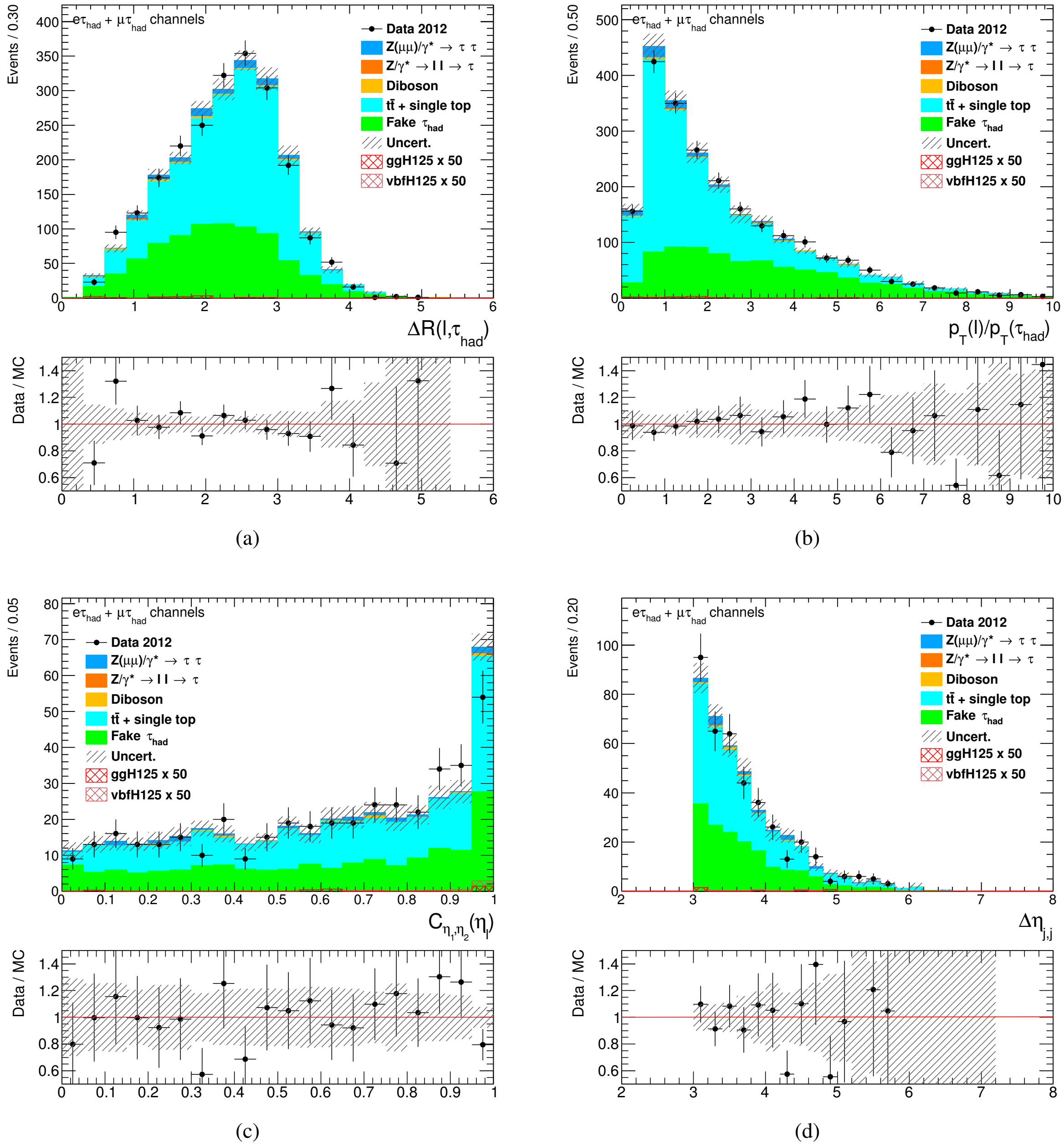


Figure 7.23: Comparison of data with the combined background model with respect to BDT input variables in the boosted (top) and VBF (bottom) top control region of the $\sqrt{s}$ =8 TeV analysis: (a) ΔR between light lepton and the τ_{had} candidate, (b) p_T ratio between the light lepton and the τ_{had} candidate, (c) $\ell - \eta$ centrality $C_{\eta_1,\eta_2}(\eta_\ell)$ and (d) difference in pseudorapidity between the two leading jets. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

In this analysis, the BDT combines two functionalities: First, it is used to improve the signal-to-background ratio. This is achieved by enriching the high BDT score region by signal events. Due to the separation power of the BDT background processes are situated more likely in the low BDT score region. Secondly, the BDT score distribution is fitted in order to extract the signal. The signal extraction based on the BDT scores in Figure 7.24 is discussed in Section 7.8.

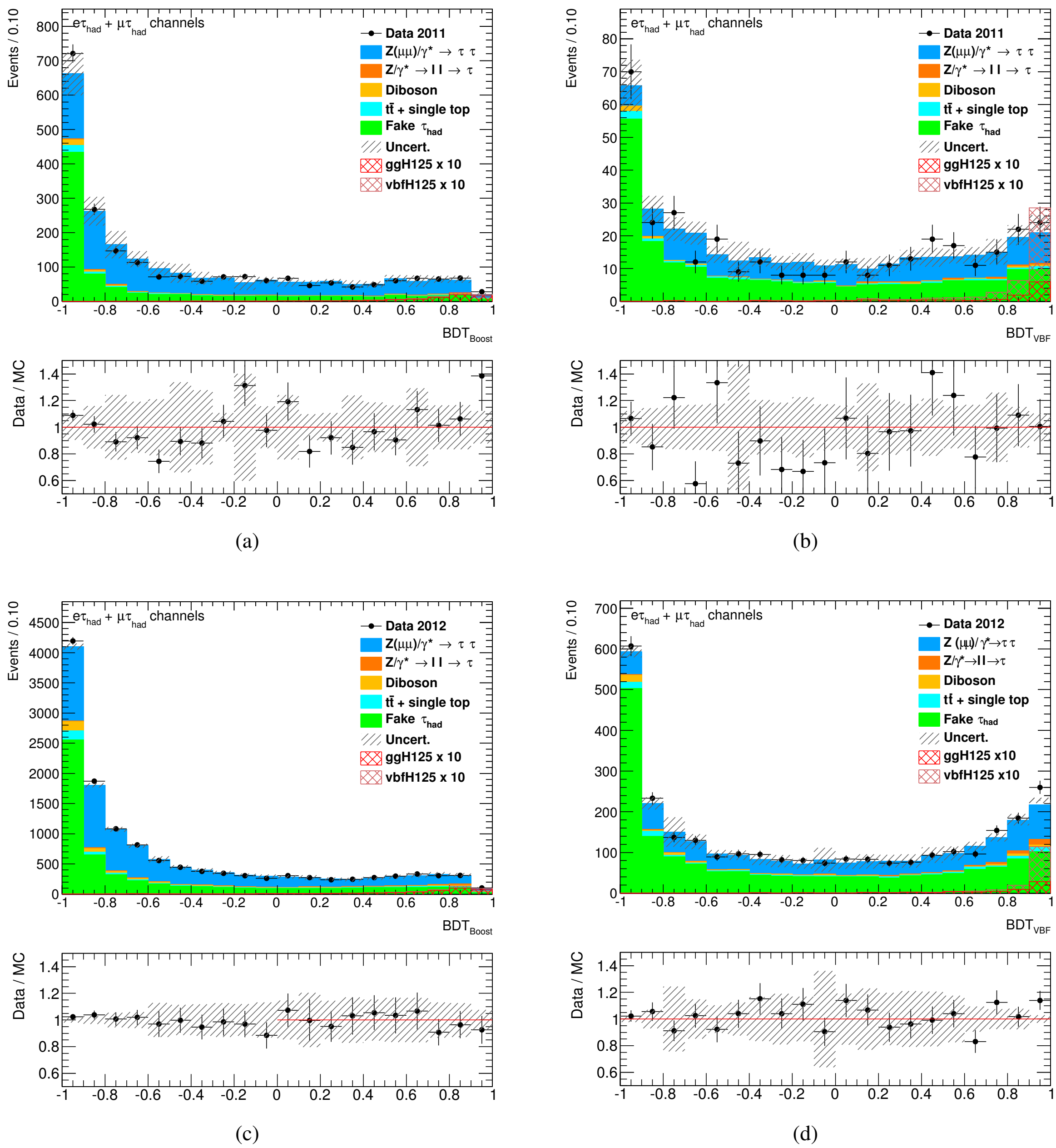


Figure 7.24: Unblinded BDT score distribution (pre-fit) for the boosted (left) and VBF (right) signal category of the $\sqrt{s}$ =7 TeV (top) and $\sqrt{s}$ =8 TeV (bottom) analysis. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

7.6.3 Application of BDTs on the $\sqrt{s}$ =7 TeV dataset

The BDT training setup used for the $\sqrt{s}$ =8 TeV analysis has been discussed in Section 7.6.1. For the $\sqrt{s}$ =7 TeV analysis, two options were investigated: First, the BDTs can be optimised for the $\sqrt{s}$ =7 TeV analysis by retraining them on the corresponding dataset. Secondly, BDTs trained on $\sqrt{s}$ =8 TeV data can be applied on $\sqrt{s}$ =7 TeV data. Since the $\sqrt{s}$ =7 TeV analysis is based on a significant smaller sample size and is carried out under different pileup conditions, it is not a-priori clear which option is the better one. In the context of this thesis, both options were studied. In the following, the two options are discussed in detail and it is argued why the second option is used in the analysis.

BDT retraining on 7 TeV data

For the training of the BDTs with $\sqrt{s}$ =7 TeV data challenges arise due to a significantly smaller sample size compared to the $\sqrt{s}$ =8 TeV analysis (see Table 7.13). In order to improve the training statistics and to avoid potential problems due to overtraining the concept of the cross evaluation is further refined. Instead of training two BDTs, $BDT1$ and $BDT2$, on split samples B and A, the samples can be split into four samples each containing 75% of the overall available samples statistics. For each of these four datasets one BDT is trained which is later applied to one of the other three data samples. Nevertheless, the training statistics is still significantly smaller compared to the $\sqrt{s}$ =8 TeV analysis.
Comparing the background composition between the boosted (VBF) signal regions of the $\sqrt{s}$ =7 TeV

Table 7.13: Number of background and signal events used in the $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV BDT training.

Category	Total background	Signal (m=125 GeV)
Boosted $\sqrt{s}$ =8 TeV	36 884	39 166
VBF $\sqrt{s}$ =8 TeV	70 309	53 978
Boosted $\sqrt{s}$ =7 TeV	7 094	6 385
VBF $\sqrt{s}$ =7 TeV	14 201	6 039

and $\sqrt{s}$ =8 TeV analysis (see Table 7.9 and Table 7.10) shows that the background composition is very similar for both datasets. Since a similar discriminating power of the BDT input variables listed in Table 7.12 is expected for the $\sqrt{s}$ =7 TeV retraining, the same input variables as for the $\sqrt{s}$ =8 TeV analysis are used.
Based on these variables the BDTs well separate signal from background processes. This separation is shown in Figure 7.25a and Figure 7.25b. In order to study the impact of the input variables on the discrimination power of the BDT importance rankings for these variables are carried out for the $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV analysis. Their results are documented in Table 7.14 and Table 7.15. The results confirm that these variables similarly contribute to the discrimination power in the $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV BDT training. The invariant $\tau\tau$ mass is ranked highest in the boosted category of the $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV analysis. Its discrimination power is smaller in the VBF category but still relevant. The invariant $\tau\tau$ mass is lower ranked in this category, since variables related to the VBF topology provide some additional separation power. Particularly in the $\sqrt{s}$ =7 TeV VBF category, the ranking value of the invariant mass differs from the other categories which might be due to the small number of events available for the BDT training. Its importance value, however, is similar to the values of other variables. Hence, the mass also contributes to the separation power of the BDT. In general, all input variables have similar values of importance. This demonstrates that all of them contribute in a compatible way to the classification performance of the BDT. The BDT is trained on a minimal set of variables in a sense that none of the chosen variables has an insignificant contribution to the

Table 7.14: Importance rankings for variables used as input for the BDT training on the $\sqrt{s}$ =7 TeV (left) and $\sqrt{s}$ =8 TeV (right) dataset in the boosted category.

Variable	Variable importance	Variable	Variable importance
$m_{\tau\tau}^{\text{MMC}}$	0.523	$m_{\tau\tau}^{\text{MMC}}$	0.218
$E_T^{\text{miss}}\phi$ centrality	0.269	$\Delta R(\tau_1\tau_2)$	0.174
$\Delta R(\tau_1\tau_2)$	0.240	$E_T^{\text{miss}}\phi$ centrality	0.169
m_T	0.200	$\sum p_T$	0.150
$p_T^{\tau_1}/p_T^{\tau_2}$	0.176	m_T	0.147
$\sum p_T$	0.011	$p_T^{\tau_1}/p_T^{\tau_2}$	0.143

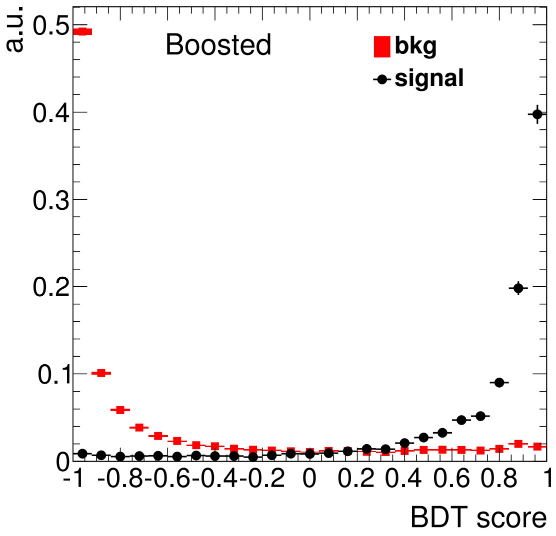
Table 7.15: Importance rankings for variables used as input for the BDT training on the $\sqrt{s}$ =7 TeV (left) and $\sqrt{s}$ =8 TeV (right) dataset in the VBF category.

Variable	Variable importance	Variable	Variable importance
$C_{\eta_1,\eta_2}(\eta_\ell)$	0.130	$\Delta\eta(j_1, j_2)$	0.134
$E_T^{\text{miss}}\phi$ centrality	0.125	$m_{\tau\tau}^{\text{MMC}}$	0.130
$\Delta\eta(j_1, j_2)$	0.122	m_T	0.119
$\Delta R(\tau_1\tau_2)$	0.113	$m_{j,j}$	0.116
$m_{j,j}$	0.111	$E_T^{\text{miss}}\phi$ centrality	0.113
$m_{\tau\tau}^{\text{MMC}}$	0.105	$\eta_{j_1} x \eta_{j_2}$	0.110
m_T	0.100	$C_{\eta_1,\eta_2}(\eta_\ell)$	0.097
p_T^{tot}	0.099	p_T^{tot}	0.091
$\eta_{j_1} x \eta_{j_2}$	0.095	$\Delta R(\tau_1\tau_2)$	0.090

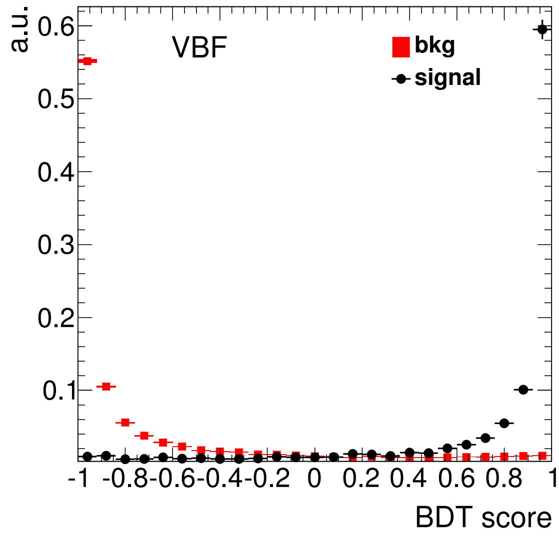
BDT performance. The largest differences with respect to the values of importance are visible in the $\sqrt{s}$ =7 TeV boosted category which might indicate that other input variables could result in an equal or better separation power.

The *ROC* integrals in Figure 7.25c and 7.25d which correspond to the BDTs trained for the $\sqrt{s}$ =7 TeV signal categories demonstrate that the separation power and the background rejection of the $\sqrt{s}$ =7 TeV BDTs are compatible with the performance of the $\sqrt{s}$ =8 TeV BDTs used in the $\sqrt{s}$ =8 TeV analysis (see Figure 7.26a and 7.26b). As for the 8 TeV analysis, the *ROC* integral of the BDT of the VBF category is larger due to additional discriminating power from VBF-jet topologies.

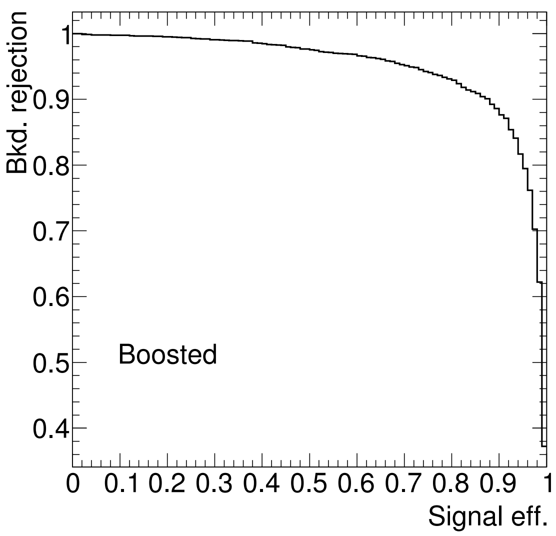
In summary, the BDT retraining on 7 TeV is possible. In order to cope with the small size of the training dataset there are some optimisations of the training setup necessary.



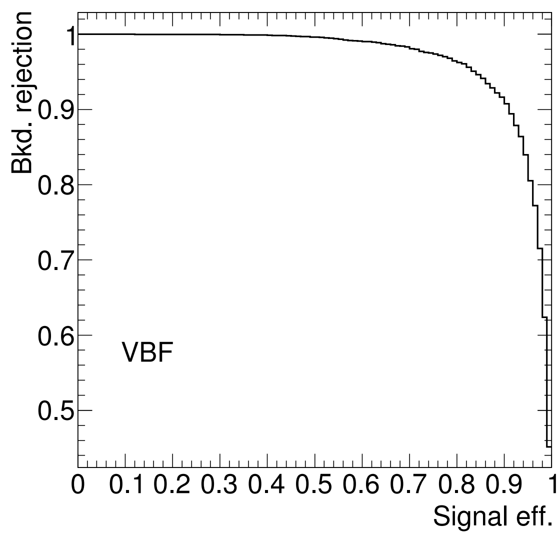
(a)



(b)



(c)

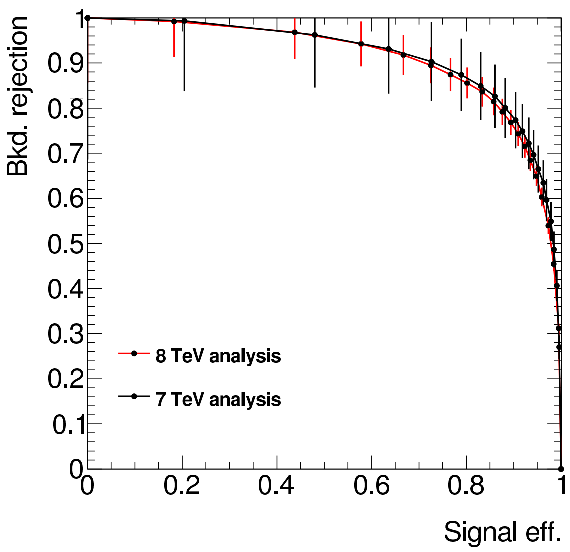


(d)

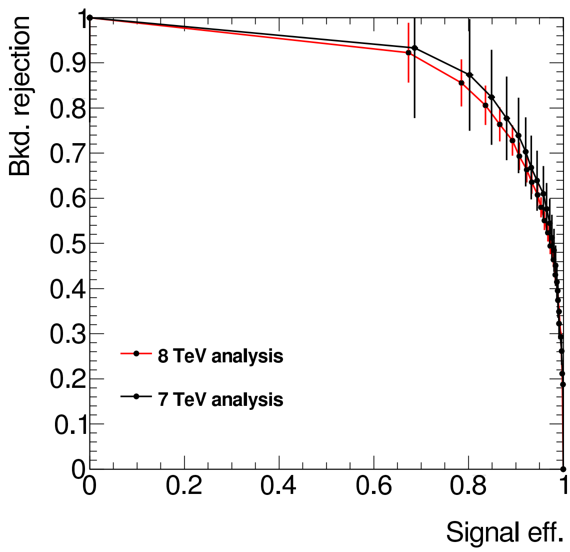
Figure 7.25: Performance of the BDTs retrained on the 7 TeV dataset: The upper plots show the BDT score distributions for signal (black) and background (red) in the boosted (left) and VBF (right) category. The lower plots show the corresponding background rejection as function of the signal efficiency for the boosted (left) and VBF (right) category.

Performance of 8 TeV BDTs on 7 TeV data

As an alternative to the retraining discussed above, the 8 TeV BDTs which are based on samples with higher statistics can be applied on 7 TeV data. The application of the 8 TeV BDTs on the 7 TeV dataset is only possible if the performance of the BDT does not significantly depend on the center-of-mass-energy. Further, the BDT performance should be independent from pileup conditions. For this reason, the BDT performance and the background composition in the BDT score distribution are investigated.
In Figure 7.26a and Figure 7.26b the *ROC* curves for the BDTs of the boosted and VBF category are compared witch each other. The *ROC* curves for the BDTs trained on $\sqrt{s}$ =8 TeV data and applied to $\sqrt{s}$ =7 TeV ($\sqrt{s}$ =8 TeV) data are illustrated in black (red). They demonstrate that the BDTs have a separation power which is similar for both categories. In addition, the BDT performance is indifferent with respect to the dataset on which the BDT is applied on. In Figure 7.26a, some slight differences between the BDT performance on the $\sqrt{s}$ =8 TeV and $\sqrt{s}$ =7 TeV dataset are visible. Similar differences are also visible when comparing the BDT retrained on $\sqrt{s}$ =7 TeV (Figure 7.25c) to the red curve in Figure 7.26a. The differences in performance might indicate that the BDT training could be further optimised for $\sqrt{s}$ =7 TeV. They are, however, small enough so that such an optimisation is not compulsory.



(a)



(b)

Figure 7.26: Background rejection as a function of the signal efficiency (*ROC* curve) for BDTs trained for the boosted (left) and VBF (right) category. The red (black) curve shows the performance of the BDT trained on $\sqrt{s}$ =8 TeV data and tested on $\sqrt{s}$ =8 TeV ($\sqrt{s}$ =7 TeV) data [127].

In Figure 7.27, the fractional background composition as a function of the BDT score is shown for both categories and both datasets. The plots demonstrate that the background composition in the BDT score distributions only slightly depends on the dataset on which the BDT is applied. Only a small increase of the fake τ_{had} background in the $\sqrt{s}$ =8 TeV dataset compared to the $\sqrt{s}$ =7 TeV dataset is visible. Since all background sources are similarly classified by the $\sqrt{s}$ =8 TeV BDT, Figure 7.27 confirms that using the $\sqrt{s}$ =8 TeV BDTs also for the $\sqrt{s}$ =7 TeV BDT analysis is an appropriate choice. Hence, for the presented $H \to \tau_\ell \tau_h$ analysis the $\sqrt{s}$ =8 TeV BDTs are used for both datasets.

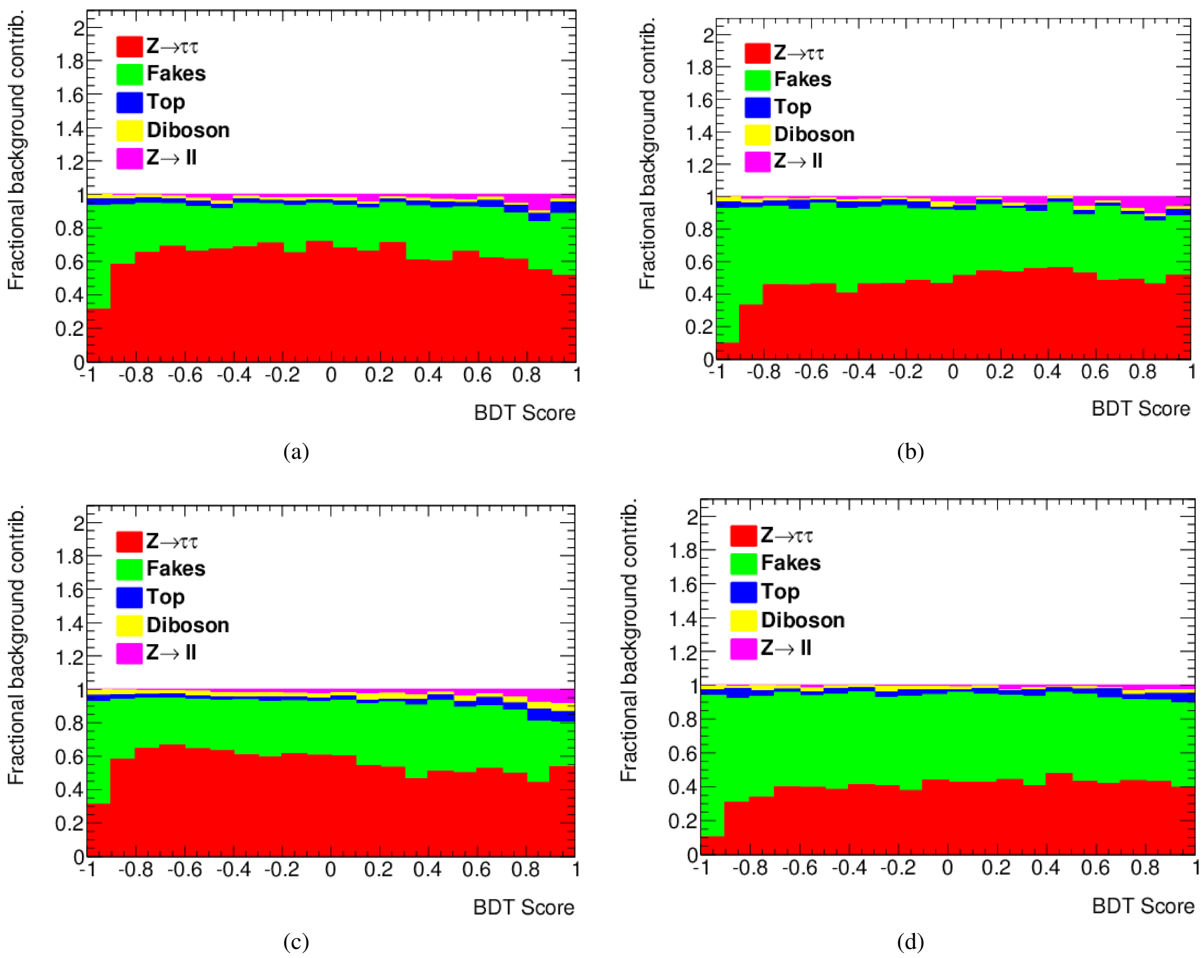


Figure 7.27: BDT score distributions for different background processes in the 7 TeV (top) and 8 TeV (bottom) analysis: The BDT score is plotted for the boosted (left) and VBF (right) category. The figures are taken from [127].

7.7 Systematic uncertainties

Systematic uncertainties arise from the experimental setup and the background modelling. In addition, the event estimate is also affected by theoretical uncertainties. Uncertainties corresponding to the performance of triggers, object reconstruction or calibration are mainly estimated from collision data. After identifying all relevant sources of uncertainties, affected objects and parameters are varied according to the corresponding estimated uncertainty. The impact of the varied parameters on the BDT score shape as well as on the background and signal event yields is evaluated in the signal categories. With respect to the variations two uncertainties are distinguishable. While shape uncertainties influence the shape of observable distributions and hence their size differs from bin to bin, normalisation uncertainties correspond to scale factors scaling the overall expected event yields in the categories up and down. Uncertainties with relevant impact and therefore considered within the signal extraction are discussed in the following.

7.7.1 Experimental uncertainties

The first group of uncertainties are experimental uncertainties which are mainly driven by uncertainties in the reconstruction and identification of physics objects, i.e. muon and electrons and τ leptons and jets. In general, uncertainties related to jets and τ_{had} candidates are bigger than for light leptons. Other uncertainties arise from trigger efficiencies, energy scales and energy resolutions of the mentioned objects.
The uncertainties on the measurement of the integrated luminosity is obtained from a calibration of the luminosity scale which is calculated from beam-separation scans as described in [172]. The final uncertainty is $\pm 1.8\%$ ($\pm 2.8\%$) for the 7 TeV (8 TeV) analysis. In the following, the relevant experimental uncertainties are briefly introduced.

Uncertainties related to muons and electrons

In order to account for resolution effects of the muon momentum and the electron energy the momentum/energy is smeared out by appropriate tools delivered by performance groups [80, 173]. The uncertainties related to the momentum/energy resolution of the light lepton are rather small in the $H \to \tau_\ell \tau_h$ analysis. This is due to the fact that the momentum and energy is measured with higher precision for muons and electrons than for τ_{had} candidates.
The efficiency of the lepton reconstruction and identification is measured via tag-and probe methods and scale factors are derived to account for efficiency differences estimated from Monte Carlo and measured in data [80, 81]. By varying the scale factors within their uncertainties their effect is estimated and propagated to the analysis.
Similar to the reconstruction and identification uncertainties, the efficiencies of the isolation requirements for muons and electrons are obtained with a tag-and-probe method. The efficiencies measured in data are used to determine correction factors for efficiencies estimated in Monte Carlo simulations. The correction factors are again varied within uncertainties in order to estimate their impact on analysis results.
Trigger efficiencies are estimated by tag and probe methods in order to get correction factors for Monte Carlo simulation, which are afterwards varied. It is the largest efficiency uncertainty for light leptons.

Uncertainties related to the τ_{had} candidate

Correction factors accounting for differences in the τ ID performance between Monte Carlo and data are estimated via tag-and-probe methods [93]. But in contrast to efficiency corrections of light leptons, the correction factors as well as their systematic variation are exclusively applied to truth-matched τ_{had} objects. Their uncertainty is of the order of 2-3% (3-5%) for one (three) prong τ decays. The amount of electrons faking the τ_{had} candidate is non-negligible and therefore has to be determined. As in the case of efficiency corrections, the derived misidentification rate differs between Monte Carlo and collision data so that is has to be corrected. The corrections are applied on τ_{had} candidates which can be matched with a true electron and varied within their uncertainties for the $\sqrt{s}$ =8 TeV analysis. For the $\sqrt{s}$ =7 TeV analysis a control region which is enriched by $Z \to \ell\ell$ processes is defined. After subtracting all background processes except for $Z \to \ell\ell$, the ratio between data and Monte Carlo events with a τ_{had} candidate matched to a true electron is determined. The uncertainty of the ratio with a size of roughly 12% is an estimate for the uncertainty on the electron $\to \tau_{\mathrm{had}}$ fake rate.
For the τ energy scale which is obtained from a fit of m_{vis} for $Z \to \tau\tau$ events in data with a precision of 2-4% a set of systematic uncertainties is varied. These uncertainties consider effects due to several components of the modelling and interpolation of true τ_{had} objects and an uncertainty special for fake τ_{had} candidates.
To estimate the uncertainty on the τ energy resolution the impact of smearing simulated τ leptons as well as changing the parton shower model is investigated. Their effect on the resolution is found to be below 1% and hence insignificant for the final discriminating variables.

E_T^{miss} related uncertainties

The calculation of E_T^{miss} is based on the energy measurement of all objects. For that reason, the energy scale uncertainties of all relevant objects, i.e. electrons, τ_{had} candidates and jets, influence the reconstructed E_T^{miss} so that it has to be recalculated after each energy scale variation [95]. Apart from these uncertainties two further uncertainties concerning the energy resolution and the E_T^{miss} soft term (see Section 4.3.8) are also taken into account. The latter is derived from $Z \to \mu\mu$ +jets Monte Carlo events and results from Monte Carlo modelling and pileup effects. It has the largest effect on the total uncertainty.

Uncertainties related to jets

The uncertainty on the jet energy resolution has to be considered for all signal and background processes which are obtained from Monte Carlo simulations, i.e. all processes except for $Z \to \tau\tau$ and background events with fake τ_{had} candidates. It is determined by an in-situ measurement which is described in [174]. Each jet is smeared by a smearing factor which corresponds to the uncertainty obtained from the in-situ measurement. For the final fit the effect of the variation is symmetrised around the nominal value.
The jet energy scale uncertainty consists of 23 independent components related to sources like pileup, modelling of insitu-jet calibration, detector-response, b-jets, flavour compositions or η-calibration. They are described in detail in [47] and [88]. In order to reduce statistical noise in the fit the significance of all components for the analysis has been studied and negligible ones removed. A further reduction of the influence of these uncertainties on the fit is achieved by using data-driven methods to estimate a large part of the background ($Z \to \tau\tau$ and fake $jet \to \tau_{\mathrm{had}}$ processes).

Impact of the experimental uncertainties on the analysis

The experimental uncertainties have different impact on background and signal event yields. The impact of the most important ones on the different event yields is summarised in Table 7.16. The numbers are derived after applying pruning and smoothing algorithms (see Section 7.8.2).
The largest uncertainties arise from the jet energy scale which range from 2%-29%. The large uncertainties related to the jet energy scale (and other uncertainties) on the event yields for the $Z \to \ell\ell(\ell \to \tau_{\text{had}})$ and diboson background result from statistical fluctuations. Their sample size is small such that its estimate depends on single events with large weights and uncertainties. Furthermore, events migrate from the boosted to the VBF category (and vice-versa) when the analysis is repeated with a varied energy scale. The jet energy resolution uncertainty ranges from 1%-21%. Since the fake τ_{had} and $Z \to \tau\tau$ background are modelled by data-driven techniques, the jet (and E_T^{miss}) related uncertainties only concern the purely MC simulated background processes. The τ energy scale has also a large impact on the event yields. Its uncertainty is of the order of 2%-18%. The lepton energy scale uncertainty is with 1%-16% the smallest uncertainty related to the energy scale. In the group of uncertainties referring to the object identification and reconstruction the uncertainty on the identified τ_{had} candidate is the largest one with a size of 2%-23%. In particular for the $Z \to \tau\tau$ background this uncertainty is large because only the visible τ decay products are taken from MC simulation (see Section 6.1). In contrast, uncertainties related to the identified light lepton are small. Their uncertainties are about or below 1%. The uncertainty related to the identification of b-jets is only relevant for $t\bar{t}$(real τ_{had}) background which is the only $H \to \tau_\ell\tau_h$ background process involving b quarks. For this background its size is about $\pm$ 12% and < 1% for all other processes. Uncertainties referring to the energy resolution and scale of E_T^{miss} are of the order of 1%-11% and are small accept for $Z \to \ell\ell(\ell \to \tau_{\text{had}})$ events due to the reason mentioned above.

7.7.2 Background estimation uncertainties

Since large parts of the background are estimated from data, additional uncertainties related to the background modelling techniques are assigned. The $Z \to \tau\tau$ background is affected by uncertainties due to the embedding procedure discussed in Section 6.6 with an impact on the $Z \to \tau\tau$ estimate of about 1.4%. Further, uncertainties related to correction factors for trigger and reconstruction efficiencies (see Section 6.4.1) are considered.
Another source of uncertainties on the background model arises due to the estimate of the fake $jet \to \tau_{\text{had}}$ background. The corresponding uncertainties are described in detail in Section 7.3.5.
For the known mismodelling of jet kinematics in $Z \to \ell\ell$ MC (see Section 7.3.5) and additional uncertainty of $\pm 10\%$ is assigned. Uncertainties, on the normalisation of the $t\bar{t}$ and single-top background have to be taken into account. Since the correction factor is allowed to float freely in the fit, contributions to the uncertainties which are mainly uncertainties on the jet energy scale and b-tagging efficiency are obtained within the fit.

7.7.3 Theoretical uncertainties

Theoretical uncertainties are derived for all signal processes and background processes which are purely simulated and normalised to theoretical cross-sections. These uncertainties arise from missing higher order QCD corrections, dependences on generator models or parton distribution functions and the modelling of underlying events and parton showers. While the uncertainties related to the simulation of background processes are in the order of $1 - 5\%$ depending on the particular process, their impact on the overall background estimate within the analysis is rather small (< 0.2%). In contrast, for the signal

Table 7.16: Impact of the most important experimental uncertainties on background and signal event yields in the $\sqrt{s}$ =8 TeV analysis after application of pruning and smoothing algorithms. The first table is related to uncertainties arising from the object identification and reconstruction. The second table summarises scale uncertainties of light leptons (LES), τ_{had} candidates (TES), jets (JES) and the energy resolution uncertainties of jets (JER). The last table lists uncertainties concerning the E_T^{miss} energy resolution and scale.

	Boosted category				VBF category			
Process	Electron	Muon	τ_{had}	b-jets	Electron	Muon	τ_{had}	b-jets
$Z \to \tau\tau$	<1%	<1%	±2.1%	-	<1%	<1%	±3.0%	-
$Z \to \ell\ell(\ell \to \tau_{\text{had}})$	±1.2%	<1%	±22%	<1%	±1.2%	<1%	±23%	<1%
Diboson	±1.2%	±1.0%	±2.0%	±1.0%	±1.1%	±1.2%	±1.8%	<1%
$t\bar{t}$(real τ_{had})	<1%	±1.2%	±3.1%	±3.5%	<1%	±1.2%	±3.4%	±12.3%
VBF	<1%	±1.2%	±3.5%	<1%	<1%	±1.2%	±3.5%	<1%
ggF	<1%	±1.1%	±3.5%	<1%	<1%	±1.2%	±3.5%	<1%
WH	<1%	±1.2%	±3.5%	<1%	<1%	±1.0%	±3.6%	<1%
ZH	<1%	±1.1%	±3.5%	<1%	<1%	±1.2%	±3.6%	<1%

	LES	TES	JES	JER	LES	TES	JES	JER
$Z \to \tau\tau$	<1%	±2.5%	-	-	<1%	±4.9%	-	-
$Z \to \ell\ell(\ell \to \tau_{\text{had}})$	±1.4%	±3.7%	±13.6%	±1.3%	±16.4%	±18%	±28.5%	±21%
Diboson	<1%	±5.0%	±8.5%	±1.3%	±2.4%	±4.0%	±16.6%	±8.5%
$t\bar{t}$(real τ_{had})	±1.0%	±4.0%	±2.5%	±2.5%	±3.1%	±3.1%	±11.8%	<1%
VBF	<1%	±2.0%	±1.5%	<1%	±1.0%	±2.2%	±6.6%	±1.0%
ggF	±1.0%	±3.5%	±5.7%	±1.0%	±2.2%	±3.3%	±18.0%	±1.0%
WH	±1.0%	±2.7%	±2.6%	<1%	±3.7%	±6.4%	±14.5%	<1%
ZH	±1.0%	±3.2%	±1.9%	<1%	±5.8%	±2.7%	±13.5%	<1%

	E_T^{miss}-scale	E_T^{miss}-resol.	E_T^{miss}-scale	E_T^{miss}-resol.
$Z \to \tau\tau$	-	-	-	-
$Z \to \ell\ell(\ell \to \tau_{\text{had}})$	±1.1%	<1%	<1%	±11%
Diboson	<1%	<1%	±1.0%	±2.0%
$t\bar{t}$(real τ_{had})	<1%	<1%	<1%	<1%
VBF	<1%	<1%	<1%	<1%
ggF	±1.1%	<1%	±1.2%	±1.4%
WH	±1.0%	<1%	±1.0%	±1.4%
ZH	<1%	<1%	±1.2%	±1.4%

theory uncertainties are highly relevant.

Uncertainties related to missing higher order QCD corrections within the cross-section calculation are estimated for the VBF and VH production mode by varying the factorisation and normalisation scale around the nominal scale $\mu_{R,F} = m_W$ [16]. These uncertainties slightly depend on the production mode or the analysis category and vary between 2-4%. For the calculation of the VBF cross-section σ_{VBF} an additional uncertainty for missing NLO electroweak corrections of 2% is used. In contrast, for the ggF production mode QCD corrections are large. At NLO their corrections are of the order of 80-100%. The uncertainties for this mode are determined by evaluating the varied factorisation and renormalisation scale $\mu_{R,F} = \sqrt{m_H^2 + p_H^2}$. The corresponding uncertainties on the cross-section in both signal categories are calculated after applying cuts on parton level which are very similar to the conditions used for the definition of the VBF and boosted signal region within the analysis [175]. Since both signal

categories are contaminated by the ggF production mode, the related uncertainty affects both categories (uncertainty impact: 24% in the boosted, 23% in the VBF category). Since the two categories are exclusive, migration effects potentially occur between the VBF and boosted category. In order to account for migration effects they are treated uncorrelated as described in [175].
The definition of the VBF category does not explicitly reject events with more than two jets. However, the input variables used for the BDT training exploit the VBF jet-topology in a way that they are able to well discriminate two-jet events from events with less or more than two jets. As a consequence, the contribution of ggF events with three jets is artificially reduced in the high BDT score region and results in a further uncertainty on the BDT shape. Since the cross-section for ggF production in association with three jets is only calculated up to LO, this uncertainty is assumed to have a potential large impact and reveal a variation of 30% in the BDT shape. Although it is of significant size and therefore considered in the fit, the impact on the analysis is smaller than 1% due to the fact that the VBF category is dominated by the VBF production mode.
Another source are uncertainties of the parton distribution functions (PDF). They are derived by evaluating acceptance differences between various PDF sets and the CT10 PDF which is used for the generation of the signal samples (see Section 7.1). A constant uncertainty of 0.9% and 1.0% (5.8% and 4.6%) for VBF (ggF) production in the boosted and VBF category is assigned which corresponds to the largest observed acceptance difference between two PDFs. Another uncertainty on the inclusive Higgs production cross-section is assigned due to the impact of the PDFs. Different generators use several models to simulate underlying events and parton showers resulting in acceptance differences. To estimate their effect the performance of POWHEG+Pythia is compared to POWHEG+Herwig for VBF and ggF production. The acceptance varies between 4% and 8% for the ggF production mode in the boosted and VBF category and between 6% and 4% for the VBF production, respectively. The BDT shape is not significantly affected by different shower models. The several generators do not only differ in their implementation of parton showers, but also in their implementation of matrix elements and their matching to parton showers. To account for acceptance differences resulting from pure generator specific properties, POWHEG+Herwig samples are compared to AMC@NLO+Herwig (MC@NLO+Herwig) for VBF (ggF) production mode. The estimated uncertainty is in the range of 2%-4%.
The theoretical uncertainties described so far also concern background processes which are estimated from simulations like diboson. They are derived for these processes in a very similar way as for the signal processes. An uncertainty related to signal process is an uncertainty on the decay branching ratio. Its estimate is described [17] and adds up to 5.7%.

7.8 Signal extraction

For the extraction of the Higgs signal from the BDT score distributions a combined likelihood fit is used. Therefore, the nominal and systematically varied BDT shapes of the signal and background processes, discussed in Section 7.3, are handed to the HistFactory program [176]. HistFactory is a part of the RootStat package and provides a method to construct parametrised probability density functions. Hence, it is able to derive limits based on binned histograms. In the case of the $H \to \tau_\ell \tau_h$ channel, four regions enter the global fit: The two signal regions and the corresponding two top control regions of the boosted and VBF category. The top control regions are included as single bins, i.e. no shape information is added to the fit. Instead, they are used to estimate the top normalisation factor within the fit. The discriminating power and signal sensitivity originate from the binned BDT distributions in both signal regions. The following section describes methods to reduce statistical noise which might be introduced by histogram bins with low event yields. Further, the construction of the likelihood fit function is explained. From the likelihood fit the nuisance parameters and signal strength are estimated.

7.8.1 Binning of the BDT distribution

The final binning used to fit the BDT shape has to be chosen in a way that its granularity allows to exploit as much shape information as possible without suffering from statistical fluctuations. In order to reach this aim the following binning optimisation procedure is carried out. The starting point is a histogram with a bin size of 0.001 units. With this bin size single events in the high BDT score region are resolved. In addition, it was proven that these bin sizes reveal the best expected significance compared to larger bin widths. The use of such a small bin width along the full BDT score region is inappropriate, because the necessary computer resources for the fit rapidly increase with a larger bin number. Further, the low BDT score region does not contribute to the signal sensitivity but to the information on the background normalisation. For that reason, a coarser binning is favoured for the low BDT score region. It is achieved by subsequently going from the highest to the next lower bins and merge them until three requirements are fulfilled: First, the number of total background events in the to-be-merged-bin has to be larger than the corresponding number in the next higher bin which ensures that the background shape continuously falls with increasing BDT score values. Further, for each relevant background process its relative statistical uncertainty must not exceed a background specific threshold. These thresholds are introduced for robustness, since too high statistical uncertainties would affect the binning optimisation. While the statistical uncertainty of the dominant $Z \to \tau\tau$ (fake τ_{had} background) must not be larger than 30% (50%), for all other considered backgrounds they are required to be smaller than 100%. Finally, in order to enable the disentanglement with respect to systematic effects and normalisation factors one of following conditions has to be satisfied:

- The total background estimate in the to-be-merged-bin is at least 50% larger than in the next higher bin.
- Moving from higher to lower BDT score bins the signal-to-background ratio decreases due to the discriminating power of the BDT. Nevertheless, it should be maximally reduced by 10% compared to the previous bin.
- If this is not the case, the ratio between the number of $Z \to \tau\tau$ events and fake τ_{had} events should change dramatically from bin to bin, i.e. more than 50%.

After the application of this algorithm on both categories and datasets the binning summarised in Table 9.5 is obtained and used for the fit.

Table 7.17: Binning of the BDT score distributions used in the boosted and VBF category of both datasets. The non-equidistant binning is obtained using a binning optimisation algorithm which allows to exploit as much shape information as possible without being sensitive to statistical fluctuations.

Category	Binning
Boosted (7 TeV)	$[-1, -0.9, -0.55, -0.05, 0.35, 0.6, 0.75, 0.817, 0.852, 0.884, 0.905, 0.926, 1.0]$
VBF (7 TeV)	$[-1, -0.75, -0.05, 0.45, 0.7, 0.828, 0.889, 0.923, 0.945, 0.97, 1.0]$
Boosted (8 TeV)	$[-1.0, -0.95, -0.75, -0.4, -0.0, 0.307, 0.56, 0.704, 0.804, 0.855, 0.904, 1.0]$
VBF (8 TeV)	$[-1.0, -0.95, -0.35, 0.35, 0.7, 0.851, 0.904, 0.936, 0.955, 0.969, 0.979, 0.988, 1.0]$

7.8.2 Pruning and smoothing algorithms

All uncertainties identified in Section 7.7 are potentially relevant for the analysis and enter the fit. The impact of these uncertainties on the event yields, however, differs between background processes: While the resulting variation is large for one background process, it might be small for the other background processes. In addition, the statistics of some of the samples is rather low in the signal regions resulting in statistically dominated variations. To prevent statistical noise from affecting the fit result all described uncertainties are considered as nuisance parameter on the normalisation only as long as their overall relative difference does not go below 0.5%. In addition, all uncertainties which are considered as possible shape variation uncertainty are pruned and smoothed before entering the fit. The pruning and smoothing algorithms are optimised in a way that they reject noise dominated variations but at the same time keep significant ones. Rejected shape uncertainties might still enter the fit as a normalisation uncertainty.
The pruning and smoothing procedure used in this analysis is subdivided into three subsequent steps:

- First, the compatibility of the downward varied shape d_i and the nominal one n_i is tested bin-by-bin using a χ^2 test

$$\chi^2 = \sum \frac{(n_i - d_i)^2}{\max(\sigma_{n_i}, \sigma_{d_i})} \,. \tag{7.9}$$

  Because the statistical uncertainties of the nominal and varied shape σ_{n_i} and σ_{d_i} are, in general, strongly correlated[9], only the largest of both is considered. The same compatibility check is carried out between the upward varied and the nominal shape. From both χ^2 distributions the p-value (see Section 7.8.6) is calculated and compared to a threshold of 0.98. If both p-values are above 0.98 the uncertainty is considered as insignificant and only used as a normalisation uncertainty. If at least one p-value is below the threshold the uncertainty passes the first pruning step.

- An additional pruning criterion is used to identify significant uncertainties. For each uncertainty and each background process the variation significance S_i is calculated according to

$$S_i = \frac{|u_i - d_i|}{\sigma_i^{all}} \,, \tag{7.10}$$

  where u_i (d_i) denotes the upward (downward) variation and σ_i^{all} the statistical uncertainty of the combined background model in bin i. The bin revealing the largest significance S_i is investigated more closely. Only if it is 10% larger than σ_i^{all}, the uncertainty is considered for this particular background as significant. Applying the pruning criterion background-wise decreases the number of shape uncertainties for background processes with low statistics.

- After pruning, smoothing and symmetrisation algorithms are applied. They are performed afterwards, because they might influence the compatibility check between the variation and the nominal shape.
  Different smoothing algorithms are implemented in the ROOT framework and help to flatten statistical fluctuations. The smoothing algorithms assume that without noise the shape of a distribution should only change slightly going from one bin to the next one. Instead of directly smoothing the BDT shape, the smoothing of the ratio of varied to nominal shape is preferred, since the BDT distribution itself might strongly vary from bin to bin, so that it might be smoothed too much. In

[9] An unknown fraction of identical events enters the nominal as well as the varied shape.

contrast, the ratio is expected to change more smoothly. The corrected varied BDT distribution is given by the nominal BDT distribution times smoothed ratio.

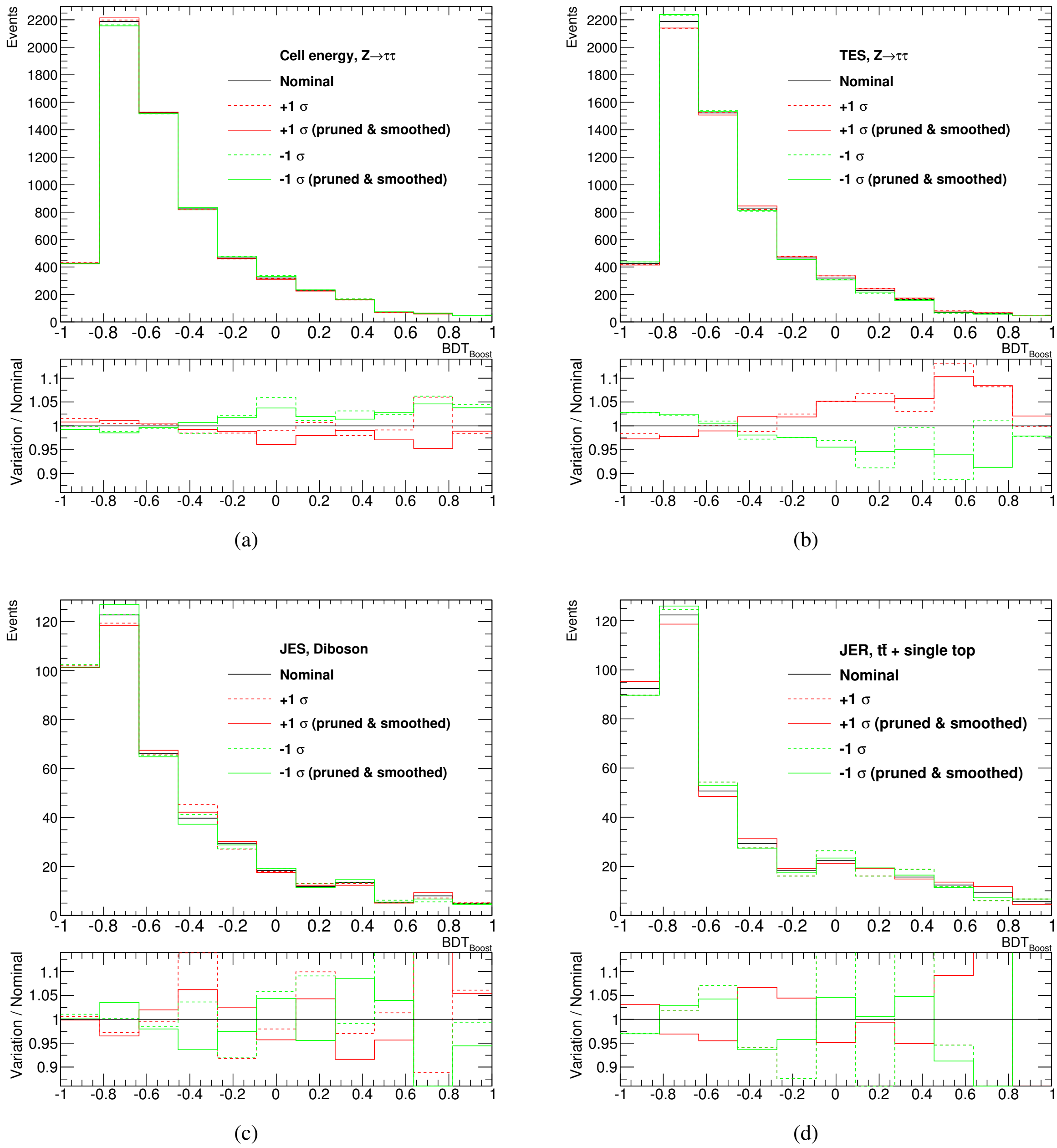


Figure 7.28: Examples for pruning, smoothing and symmetrisation effects for different systematic uncertainties. The nominal BDT score distribution in the 8 TeV boosted signal region is shown in black. The $+1\sigma$ (-1σ) variation before pruning is indicated by the dashed red (green) line. The $+1\sigma$ (-1σ) variation after applying the pruning, smoothing and symmetrisation algorithms is indicated by the solid red (green) line. The ratio plots show the $\pm 1\sigma$ systematic variation with respect to the nominal distribution: (a) the cell subtraction systematic for the $Z \to \tau\tau$ background, (b) the τ energy scale for the $Z \to \tau\tau$ background, (c) the jet energy scale for the diboson background and (d) the jet energy resolution for the $t\bar{t}$ background.

7.8.3 Symmetrisation

After smoothing, the majority of statistical fluctuations is removed. Nevertheless, it might occur that for some systematic uncertainties the upward and downward variations in some bins point in the same direction with respect to the nominal distribution. Although such one-sided variations can be a consequence of the definition of the uncertainty, they mostly originate from statistical fluctuations. The removal of these one-sided variations is favoured as they can result in multiple minima of the likelihood fit and thus disturb the minimisation process, as discussed in Section 7.8.4. For this reason, the up- and downward variation of each uncertainty is symmetrised bin-wise. Since the shape uncertainties are always symmetrised with respect to the largest variation from nominal, the symmetrisation procedure increases the overall size of uncertainties.
Examples for the influence of these pruning, smoothing and symmetrisation algorithms on shape uncertainties are shown in Figure 7.28. They show the nominal BDT score distribution in the 8 TeV boosted signal region and the $\pm 1\sigma$ variation for some selected uncertainties before and after applying these algorithms. Figure 7.28a and Figure 7.28b show the cell subtraction variation and the τ energy scale for the $Z \rightarrow \tau\tau$ background. The influence of the pruning on the jet energy scale and jet energy resolution are demonstrated in Figure 7.28c for the diboson and in Figure 7.28d for the $t\bar{t}$ background, respectively. Since the jet energy scale uncertainty is a one-sided uncertainty, it has to be fully symmetrised around the nominal distribution.

7.8.4 The maximum likelihood function

The maximum likelihood function is a common mean used in particle physics to estimate n parameters $\vec{\Theta} = \{\Theta_1, ...\Theta_n\}$ on which the probability density function $f(x; \vec{\Theta})$ (PDF) depends [177]. The PDF $f(x; \vec{\Theta})$ is a continuous function which is normalised to unity and gives the probability to measure in an event a certain value x_e for the sampling variable x. In this particular case, x_e stands for the BDT score value assigned to the observed event and $\vec{\Theta}$ for the set of nuisance parameters. For each value of x (BDT score bin) a certain number of background events B and signal events S according to the corresponding PDFs $f_B(x; \vec{\Theta})$ and $f_S(x; \vec{\Theta})$ is expected. Thereby, several signal and background processes can contribute to each of these PDFs. In general, a signal strength parameter μ is introduced. The resulting number of expected events N_{exp} for x_e is given by

$$N_{exp} = \mu S + B\,, \tag{7.11}$$

where $\mu = 0$ ($\mu = 1$) denotes the background-only (signal-plus-background) hypothesis. Although both, the signal strength and the nuisance parameters, are estimated, the signal strength is in fact of interest for the analysis, since it describes the consistency of the measurement with the SM prediction. The probability to observe N_{obs} events if N_{exp} are expected is described by a Poisson distribution:

$$P(N_{obs}|N_{exp}) = \frac{e^{-(\mu S + B)}(\mu S + B)^{N}}{N_{obs}!}\,. \tag{7.12}$$

The probability density to observe N_{obs} events at x_e hence is

$$P(x|\mu, \vec{\Theta}) = \frac{e^{-(\mu S + B)}(\mu S + B)^{N_{obs}}}{N_{obs}!} \prod_{e}^{N_{obs}} \frac{\mu S \cdot f_S(x; \vec{\Theta}) + B \cdot f_B(x; \vec{\Theta})}{\mu S + B}\,. \tag{7.13}$$

If one assumes the number of data events to be constant the probability density function can be regarded as a function of μ as well as $\vec{\Theta}$ and is referred to as likelihood function $\mathcal{L}(\mu, \vec{\Theta}|x)$. The parameters are estimated by minimising the negative likelihood function. Since it is easier to minimise a sum rather than a product, i.e. the equation

$$\begin{aligned} -\ln \mathcal{L}(\mu, \vec{\Theta}|x) &= (\mu S + B) + \ln N_{obs}! - \sum_{e=1}^{N_{obs}} \ln\left[\mu S f_S(x; \vec{\Theta}) + B f_B(x; \vec{\Theta})\right] \\ &= C - \sum_{e=1}^{N_{obs}} \ln\left[\mu S f_S(x; \vec{\Theta}) + B f_B(x; \vec{\Theta})\right] \end{aligned} \tag{7.14}$$

is used. For a binned BDT score distribution the equation can be rewritten as

$$-\ln \mathcal{L}(\mu, \vec{\Theta}|x) = C - \sum_{i}^{N_{bins}} \ln\left[\mu s_i(\vec{\Theta}) + b_i(\vec{\Theta})\right] , \tag{7.15}$$

where s_i and b_i are the expected number of signal and background events in bin i, respectively. The sum runs over all bins N_{bins} of the BDT score distribution. By adding the bins of the BDT distributions from different signal and control regions to the sum, different analysis categories can be combined. The HistFactory program [176] allows the combination of very different categories like the binned BDT score distribution in the signal region which provides shape information and the single-bin top control regions used for normalisation. The programm MINUIT [178] actually provides the solutions of the $-\ln \mathcal{L}(\mu, \vec{\Theta}|x)$ minimisation. The solutions for $(\mu, \vec{\Theta})$ minimising $-\ln \mathcal{L}(\mu, \vec{\Theta}|x)$ are called best-fit values or estimators $(\hat{\mu}, \hat{\vec{\Theta}})$. In order to estimate the variance of the best-fit values $(\hat{\mu}, \hat{\vec{\Theta}})$ the logarithmic likelihood function $-\ln \mathcal{L}(\mu, \vec{\Theta}|x)$ is minimised with respect to all parameters except for the parameter Θ_{var} for which the variance has to be determined. The change if the likelihood profile as a function of Θ_{var} is investigated. The uncertainty of Θ_{var} is derived from the values $\Theta^{'}_{var}$ for which the likelihood profile differs by $1/2$ with respect to its minimum.

7.8.5 Systematic and statistical uncertainties in the likelihood fit

The nuisance parameters estimated in the likelihood fit are normalisation factors and parametrisations of the impact of analysis relevant systematic uncertainties. The systematic uncertainties are obtained from calibration measurements before the analysis and propagated through the analysis by varying the nominal results of these measurements by $\pm 1\sigma$. In order to consider the precision of such measurements within the likelihood function and thereby constrain the nuisance parameters constrained terms are added to the likelihood function. These constrained terms are simplified likelihood function each depending on one nuisance parameter. The simplifications are necessary to reflect the calibration measurements despite of their complexity. A common simplification is a Gaussian distribution, whose expectation value is the result of the calibration measurement. Its width is represented by corresponding uncertainties. Since upward and downward variations are usually expressed in terms of relative deviations from nominal estimates, the Gaussian is centred around 0 such that a nuisance parameter of value 1 matches a shift of $\pm 1\sigma$ of the nuisance parameter. While the impact of the uncertainty within the analysis is estimated by re-analysing the data with the uncertainty varied by $\pm 1\sigma$, within the likelihood function a continuous parametrisation has to be found. Several linear and exponential interpolation algorithms are implemented in HistFactory in order to interpolate between the nominal estimate of the nuisance parameter and the corresponding $\pm 1\sigma$ shift. Those algorithms and a more detailed discussion of the incorporation of systematic uncertainties can be found in [176]. The treatment of statistical uncertainties within the likelihood is very similar to the treatment of systematic uncertainties. Again, auxiliary measurements are incorporated in constrained terms which are added to the full likelihood function. The choice of the constrained terms, however, is different. In each bin the number of expected signal and background events is estimated from the weighted number of simulated events. The sample weights depend on the individual background processes and the way used to estimate them, e.g. using fake factors or the normalisation to the measured luminosity according to the theoretical cross-section. The actual number of unweighted events is a auxiliary measurement which can be described by a Poisson distribution with an unknown mean. Multiplying the result of this measurement with the corresponding event weight results in constrained background events. Since adding such terms for each sample in each bin results in an overwhelming number of nuisance parameters, HistFactory only considers the total statistical uncertainty of all processes in each bin, i.e. only one nuisance parameter for each bin associated with the statistical component is employed. For signal processes, no statistical uncertainty is taken into account.

7.8.6 Likelihood based hypothesis tests

Hypothesis tests are important tools in particle physics to estimate the statistical compatibility of data with different signal hypotheses. In this thesis, the so called p-value is used to quantify the compatibility of the excess in data with the background-only hypothesis. This section shortly summarises the main concept of hypothesis tests as elaborated in [179] and [180].
First, a null hypothesis H_0 which is tested against an alternative hypothesis H_1 is defined. In general, H_0 denotes the background-only hypothesis ($\mu = 0$), while H_1 describes the background-plus signal hypothesis ($\mu = 1$). Then, for a given test statistics q_μ the probability density function $f(q|H_0)$, under the assumption that the hypothesis H_0 is true is determined and the p-value is calculated according to

$$p = \int_{q'}^{\mathrm{inf}} f(q|H_0)dq \,. \tag{7.16}$$

The p-value describes the probability under the assumption of H_0 to obtain an incompatibility between data and the prediction of H_0 which is equal or greater than the observed one. Often, the p-value is converted into an equivalent significance Z defined as

$$Z = \Phi^{-1}(1 - p) \,, \tag{7.17}$$

where Φ^{-1} describes the inverse of the cumulative distribution of the standard Gaussian. The background-only hypothesis is rejected and the signal therefore taken to be significant if the p-value is below a certain threshold. In general, this threshold is given by $p = 2.87 \cdot 10^{-7}$ corresponding to $Z = 5$. Evidence for the signal is claimed if Z is equal or larger than 3. In order to exclude the signal hypothesis upper limits are determined by calculating a signal event rate at which the hypothesis is rejected with a confidence level of usually 95%. As outlined in [179] several test statistics based on the profile likelihood ratio for two-sided tests, upper limits or discovery are defined. One important test statistics q_0 which also considers effects due to systematic uncertainties is used to test the compatibility between observed data and background-only hypothesis ($\mu = 0$). It is defined as

$$\begin{aligned} q_0 &= -2\ln\left(\frac{L(\mu = 0, \hat{\hat{\Theta}})}{L(\hat{\mu}, \hat{\theta})}\right) \text{ for } \hat{\mu} \geq 0 \\ q_0 &= 0 \text{ for } \hat{\mu} < 0 \,, \end{aligned} \tag{7.18}$$

where $L(\mu = 0, \hat{\hat{\Theta}})$ denotes the likelihood function minimised with μ fixed at zero, while $L(\mu, \hat{\Theta})$ is minimised as function of μ and Θ. The incompatibility between data and background-only hypothesis becomes larger with increasing q_0. The estimated signal strength is negative if the number of observed events is smaller than predicted by the background-only hypothesis. In this case, q_0 is defined to be zero in order to be still consistent with the background-only hypothesis. The derivation of the probability density function $f(q_0|\mu_{\text{assumed}})$ under an assumed signal strength μ_{assumed} is discussed in [179].

7.9 Results

The signal strength μ is obtained by a simultaneous likelihood fit of the BDT score distributions in the four signal regions (boosted and VBF category in the 7 TeV and 8 TeV analysis) and corresponding single binned top control regions. The fit does not only provide an estimate of the signal strength but also the normalisation of the $Z \rightarrow \tau\tau$ and top quark background. Further, nuisance parameters which reflect the impact of the systematic uncertainties are estimated. The post-fit BDT score distributions are shown in Figure 7.29. They demonstrate good agreement between data and the signal-plus-background model. The BDT score distribution in the 8 TeV VBF signal category indicates a clear excess of data over the background-only model in the highest bin as shown in Figure 7.29d. The corresponding fitted signal and background yields for the $\sqrt{s}$=7 TeV and $\sqrt{s}$=8 TeV analysis including event yields for the two BDT highest bins are summarised in Tables 7.18 and 7.19.
The best estimator of the signal strength μ at m_H =125 GeV is found to be

$$\hat{\mu} = 1.06^{+0.34}_{-0.34}\,(\text{stat.})\,^{+0.38}_{-0.29}\,(\text{syst.}) \pm 0.1\,(\text{theo. syst.}) \ .$$

While all uncertainties related to experimental effects or background estimation methods or summarised as *experimental systematics* (syst.), theoretical uncertainties discussed in Section 7.7.3 are referred to as *theory systematics* (theo. syst.). *Statistical uncertainties* (stat.) denotes statistical uncertainties on the total number of events in the signal and control regions. The contribution of the statistical, experimental and theoretical uncertainties are obtained by repeating the fit with experimental or theoretical uncertainties fixed to their best estimate derived from the unconditional fit. Afterwards, the uncertainty on $\hat{\mu}$ from the conditional fit is subtracted in quadrature from the overall uncertainty on $\hat{\mu}$ from the unconditional fit. The contribution of the statistical uncertainty to the overall uncertainty of $\hat{\mu}$ is of the same size as the systematic uncertainty. The estimates of the signal strength obtained from fits in the different categories and datasets are illustrated in Table 7.20. Within uncertainties all results are consistent with the $\mu = 1$ hypothesis, i.e. the Higgs boson as predicted by the SM. The overall measured signal strength is mainly driven by the measurement in the $\sqrt{s}$=8 TeV dataset. The measurement is dominated by the VBF category of the $\sqrt{s}$=8 TeV analysis. In the $\sqrt{s}$=8 TeV analysis, the signal strength measured in the VBF signal region ($\mu = 1.33^{+0.74}_{-0.59}$) is larger than the signal strength obtained in the boosted category ($\mu = 0.69^{+1.05}_{-0.69}$). Since the combined result of the $\sqrt{s}$=8 TeV analysis is driven by the more sensitive VBF category the measured signal strength is slightly above 1. Due to the smaller statistics of the $\sqrt{s}$=7 TeV dataset, the uncertainties on the measured signal strength in the $\sqrt{s}$=7 TeV analysis are larger compared to the results of the $\sqrt{s}$=8 TeV analysis. The combined $\sqrt{s}$=7 TeV result is given by $\mu = 1.0^{+1.4}_{-1.0}$, while the signal strength in the $\sqrt{s}$=7 TeV VBF category yields $\mu = 0.0^{+1.1}_{-0.0}$. However, considering the uncertainties this is still in accordance with the $\mu = 1$ hypothesis.
The excess in data is tested against the background-only hypothesis resulting in an overall observed (expected) significance of 2.5 (2.4) Gaussian standard deviations. This result is only slightly below a significance of 3σ which is necessary for an experimental evidence for the SM Higgs boson. The results

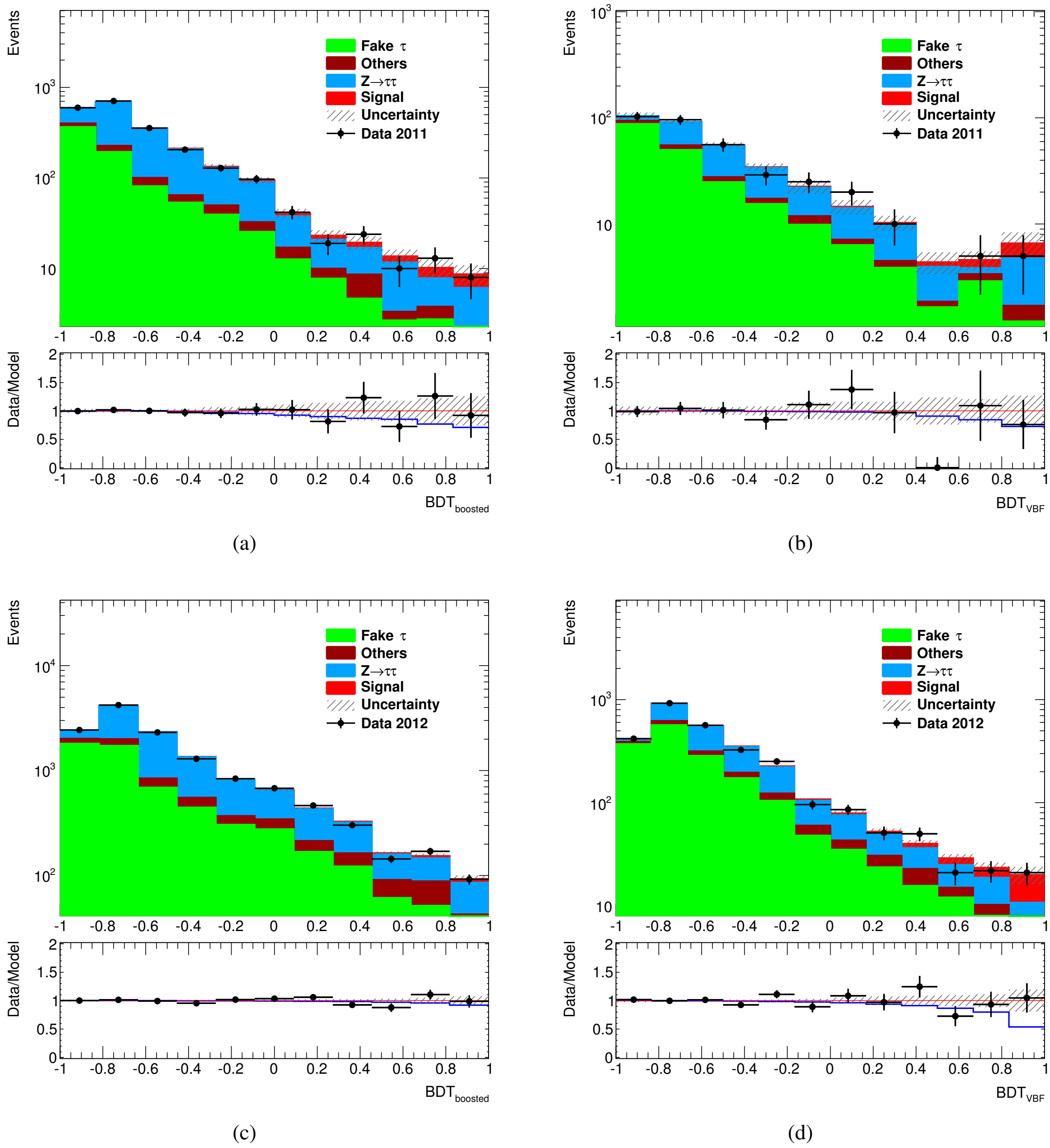


Figure 7.29: Post-fit BDT score distributions for the four signal regions. Top: BDT score for the 7 TeV boosted (left) and VBF (right) category. Bottom: BDT score for the 8 TeV boosted (left) and VBF (right) category. The signal is fitted with $\mu = 1.06$. The fitted bins are mapped to equidistant bins. The lower ratio plots show the ratio between data and the signal-plus-background model including the statistical and systematic uncertainties of the model (hatched band). The blue line illustrates the ratio between the background-only and the signal-plus-background model.

for the different categories and datasets are summarised in Table 7.21. As before, the result is dominated by the $\sqrt{s}$=8 TeV VBF category. Since in the $\sqrt{s}$=7 TeV VBF category a signal strength equal to zero is measured, the corresponding significance is negative.

Table 7.18: Post-fit event yields for background and signal processes in the $\sqrt{s}$ =7 TeV boosted and VBF category including systematic uncertainties. Besides the overall yields, the yields in the two highest bins are also shown.

Sample	Boosted category			VBF category		
	All bins	2nd last bin	Last bin	All bins	2nd last bin	Last bin
$Z \to \tau\tau$	1262±74	8.4 ±3.4	4.2± 2.1	140±12	1.6±1.2	2.1±1.1
$Z \to \ell\ell(\ell \to \tau_{\mathrm{had}})$	33±8	0.59±0.44	0.26±0.09	4.4 ±1.5	0.06±0.05	0.05±0.05
Diboson	43±6	0.10±0.10	0.21±0.20	7.3 ±1.6	0.06±0.05	0.09±0.08
$t\bar{t}$(real τ_{had})	55±14	0.27±0.14	0.72±0.43	11±5	0.37±0.32	0.41±0.39
Fake background	785±92	2.8±0.7	2.9±0.9	182±24	2.5 ±0.6	1.2±0.5
VBF	2.0±0.9	0.16±0.09	0.19±0.09	3.5±1.6	0.54±0.25	1.5±0.6
ggF	6.2±3.2	0.57±0.29	0.62±0.32	1.2±0.7	0.16±0.09	0.27±0.21
WH	1.0 ±0.5	<0.1	<0.1	<0.1	<0.1	<0.1
ZH	0.9 ±0.5	<0.1	<0.1	<0.1	<0.1	<0.1
Data	2199	13	8	349	5	5

Table 7.19: Post-fit event yields for background and signal processes in the $\sqrt{s}$ =8 TeV boosted and VBF category including systematic uncertainties. Besides the overall yields, the yields in the two highest bins are also shown.

Sample	Boosted category			VBF category		
	All bins	2nd last bin	Last bin	All bins	2nd last bin	Last bin
$Z \to \tau\tau$	6205 ±206	58.8±1.9	42±2	883±34	7.8±1.0	4.2±0.9
$Z \to \ell\ell(\ell \to \tau_{\mathrm{had}})$	215±63	14±7	10±5	50 ±23	1.4 ±0.9	0.53±0.36
Diboson	437 ±48	9.7±2.3	5.3±1.6	67±13	1.0±0.4	0.49±0.19
$t\bar{t}$(real τ_{had})	400±60	11±2	4.9±2.1	83 ±14	0.2 ±0.6	0.6±0.4
Fake background	5632±171	51±2.6	22.4±1.8	1680±50	8.3±0.9	5.3±0.7
VBF	11 ±5	1.9 ±0.9	2.1±1.0	22±10	3.3±1.6	6.9±3.3
ggF	42±22	6.6±3.3	7.3 ±3.5	13±7	0.8±0.4	0.9±0.06
WH	7.0 ±3.6	1.0±0.5	1.5±0.7	0.44±0.32	<0.1	<0.1
ZH	3.4 ±1.6	0.58±0.25	0.71 ±0.32	0.10±0.09	<0.1	<0.1
Data	12952	170	92	2830	22	21

Table 7.20: Estimate of the signal strength $\hat{\mu}$ obtained from fits in the individual signal categories (boosted and VBF) and for different datasets ($\sqrt{s}$=7 TeV and $\sqrt{s}$=8 TeV) assuming m_H =125 GeV. The last column and row show the results from the combined fit in the signal regions and for combined datasets. The error includes statistical and systematic uncertainties.

Category	7 TeV	8 TeV	7 TeV + 8 TeV
Boosted	$2.60^{+2.88}_{-2.49}$	$0.69^{+1.05}_{-0.69}$	$0.88^{+1.04}_{-0.88}$
VBF	$0.00^{+1.14}_{-0.00}$	$1.33^{+0.74}_{-0.59}$	$1.09^{+0.63}_{-0.52}$
Combined	$1.0^{+1.4}_{-1.0}$	$1.14^{+0.59}_{-0.49}$	$1.06^{+0.53}_{-0.46}$

Table 7.21: Observed (expected) local significance in terms of Gaussian standard deviations evaluated for the individual signal categories (boosted and VBF) as well as datasets ($\sqrt{s}$=7 TeV and $\sqrt{s}$=8 TeV) under the assumption of m_H =125 GeV. The last column and row show the results of the hypothesis test for the combined signal regions and for combined datasets.

Category	7 TeV	8 TeV	7 TeV + 8 TeV
Boosted	1.0(0.4)	0.7(1.0)	0.9(1.1)
VBF	-0.4(0.6)	2.5(2.0)	2.2 (2.0)
Combined	0.8(0.8)	2.5(2.3)	2.5 (2.4)

The presented $H \to \tau_\ell\tau_h$ analysis and its results enter the combined $H \to \tau\tau$ analysis published in [8]. Slight differences arise from an exclusive fit of the $H \to \tau_\ell\tau_h$ signal and control regions, while in [8] regions of all three channels are fitted simultaneously resulting in additional constraints from nuisance parameters of the other channels. These differences mainly concern the results of the less sensitive 7 TeV analysis and as consequence the combination of both datasets.

In order to estimate the influence of the various experimental and theoretical uncertainties on the results each nuisance parameter is fixed to its $\pm1\sigma$ estimate. For each fixed nuisance parameter, the fit is repeated so that the likelihood is minimised with respect to the overall signal strength μ as well as all other nuisance parameters Θ. The impact of the nuisance parameter is given by the deviation $\Delta\mu_{Global}$ of the conditional fitted signal strength $\Delta\mu(\Delta\Theta)$ from the nominal estimate obtained from the unconditional fit. The size of the changes are used to order the nuisance parameters with respect to their importance as illustrated in Figure 7.30. The post-fit and pre-fit impact on $\hat{\mu}$ are indicated by the red ($+1\sigma$ variation) and blue (-1σ variation) hatched boxes, respectively. The black points show the difference between the best-fit values and pre-fit values of the various nuisance parameters. The corresponding post-fit and nominal uncertainties are indicated by black lines and yellow bands, respectively.

The highest ranked nuisance parameter corresponds to the uncertainty on the isolation requirement used in the embedding technique (see Section 6.6) for 7 TeV data samples, followed by a nuisance parameter which describes the statistical uncertainty of the background yield in the highest BDT score bin in the 8 TeV VBF category. Further nuisance parameters reflecting statistical uncertainties of lower BDT score bins are found lower ranked. Of high importance are several components of the modelling of the jet energy scale. Slightly lower ranked are uncertainties related to the misidentification of τ_{had} candidates by electrons or jets. The highest ranked nuisance parameter associated with theory uncertainties is the uncertainty on the prediction of the $H \to \tau_\ell\tau_h$ branching ratio. Next ranked theoretical uncertainties refer to the factorisation and renormalisation scales which are used in the calculation of the cross-section of the ggF production mechanism. Besides these uncertainties, also the $Z \to \tau\tau$ and top normalisation factors which are allowed to float freely in the fit are in the list of the 50 highest ranked nuisance parameters. Like the other parameters, they are in good agreement with their pre-fit values.

The investigation of pulls of nuisance parameters between post-fit and pre-fit estimates as in Figure 7.30 is one part of the validation of the fitting procedure. Several other tests are also carried out to confirm a good behaviour of the fit. For example, ensuring that the one dimensional likelihood functions for all nuisance parameters and the signal strength has a parabolic shape around the global minimum avoids additional local minima which might otherwise be incorrectly used as best estimate for the different parameters. Another check investigates correlation between the various nuisance parameters. The correlation coefficients should be small in order to avoid some degeneracies between parameters.

7.10 Conclusion

In the presented analysis of the $H \to \tau_\ell\tau_h$ decay channel a deviation from the background-only hypothesis ($\mu = 0$) is observed which corresponds to a significance of 2.5 σ. The corresponding measured signal strength of $\mu = 1.1^{+0.5}_{-0.5}$ is consistent with the signal-plus-background ($\mu = 1$) hypothesis. This result confirms the preliminary result [135], which is based on a multivariate analysis of the $\sqrt{s}$=8 TeV dataset and yields a signal strength of $\mu = 1.4^{+0.6}_{-0.5}$.
The presented $H \to \tau_\ell\tau_h$ analysis re-analysed the $\sqrt{s}$=8 TeV dataset and included the $\sqrt{s}$=7 TeV into a combined analysis based on the complete Run 1 dataset. In particular, this analysis includes the first multivariate analysis of the $\sqrt{s}$=7 TeV dataset in the context of SM $H \to \tau\tau$ searches. For the re-analysis of the $\sqrt{s}$=8 TeV dataset the modelling of the fake and $Z \to \tau\tau$ background was further studied and optimised. The harmonisation with the $H \to \tau_\ell\tau_\ell$ and $H \to \tau_h\tau_h$ analyses was improved as much as possible. Furthermore, the latest results regarding theoretically calculations, object reconstruction and corresponding uncertainties obtained by the different performance groups were considered.
Since the $H \to \tau_\ell\tau_h$ channel has the largest branching ratio of the three $\tau\tau$ decay channels, it significantly contributes to the signal sensitivity of the combined $H \to \tau\tau$ analysis. The high sensitivity of the $H \to \tau_\ell\tau_h$ channel is essential for the performance of the combined $H \to \tau\tau$ analysis. In combination with the analysis of the $H \to \tau_\ell\tau_\ell$ and $H \to \tau_h\tau_h$ channel an evidence for the Higgs-boson Yukawa coupling to τ leptons is achieved. This combination is discussed in the following Chapter 8.

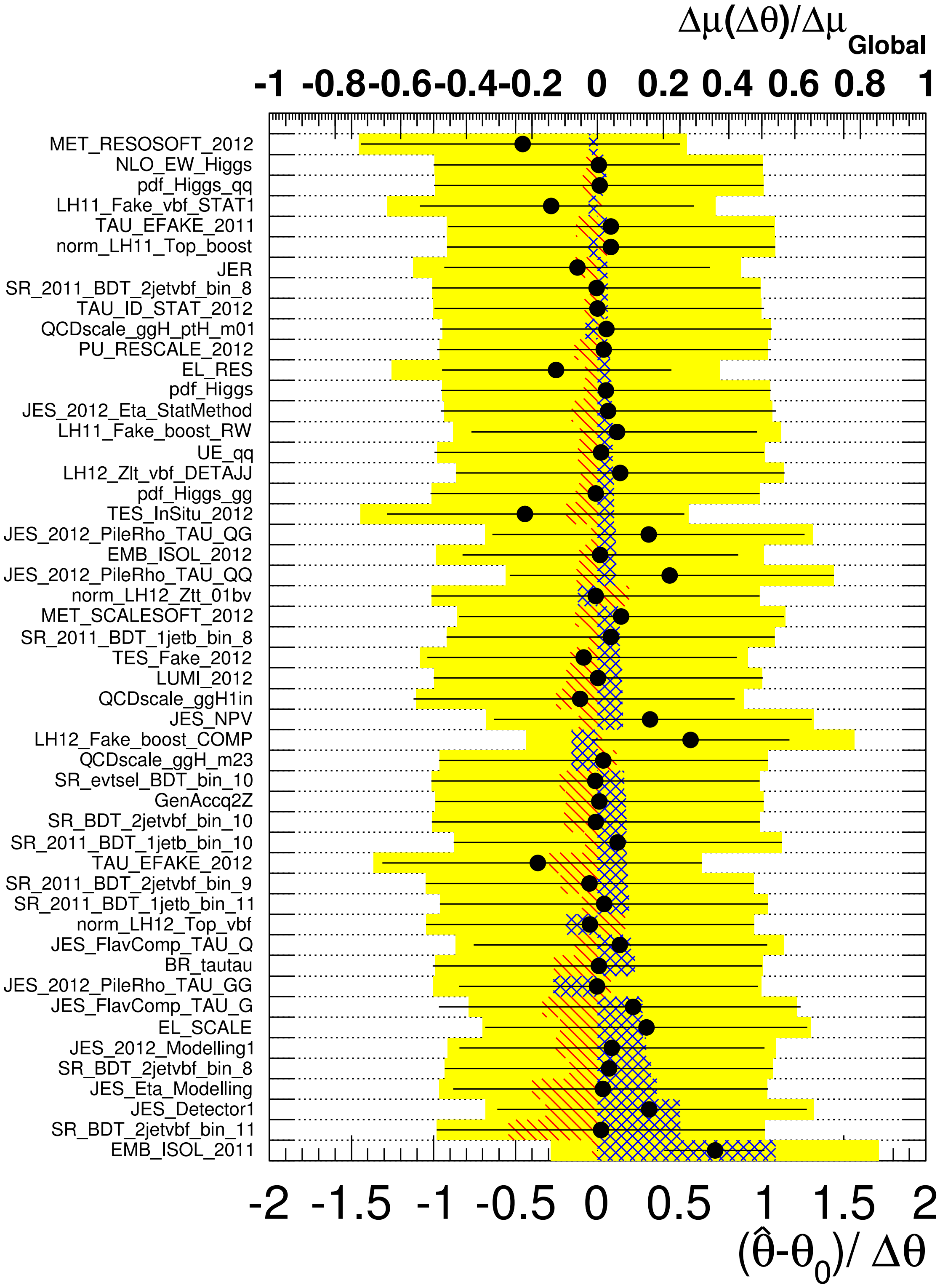


Figure 7.30: Ranking of the 50 nuisance parameters with the largest impact on the signal strength μ in the combined fit of the $H \to \tau_\ell \tau_h$ 7 TeV and 8 TeV signal regions. The black points refer to the lower axis and show the deviations of the best-fit value of the nuisance parameters $\hat{\Theta}$ from their nominal pre-fit values Θ_0. While the black lines illustrate the relative post-fit uncertainties of the individual nuisance parameters, the nominal uncertainties are given by the yellow band. The hatched boxes refer to the upper axis visualising the impact $\Delta\mu(\Delta\Theta)$ of the nuisance parameters on the measured signal strength, when fixing the individual nuisance parameters to their best post-fit value. $\Delta\mu(\Delta\Theta)$ is divided by the total error $\Delta\mu_{Global}$ on the best-fit value of the signal strength. While the red hatched boxes show the $+1\sigma$ post-fit impact, the blue hatched boxes display the -1σ pre-fit impact on $\hat{\mu}$. The labelling of the nuisance parameters is explained in Appendix E.

CHAPTER 8

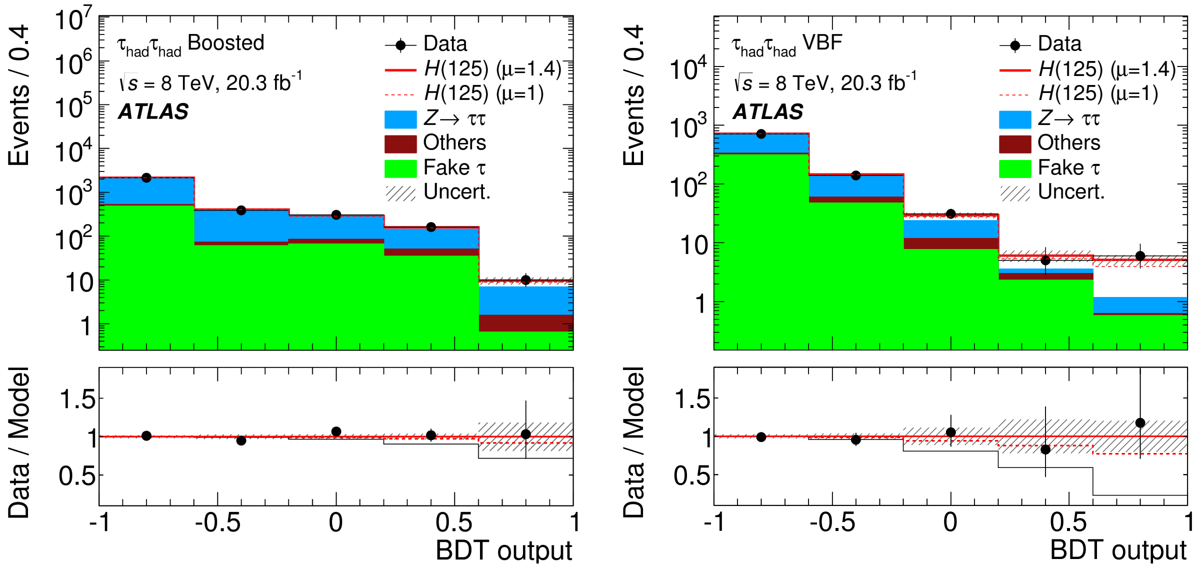
Combined analysis of the $H \to \tau\tau$ channel

The results discussed so far concerned the analysis of the $H \to \tau_\ell\tau_h$ decay channel only. Also the other two possible decay modes $H \to \tau_h\tau_h$ and $H \to \tau_\ell\tau_\ell$ were investigated by the ATLAS collaboration. In order to combine the three analysis channels the analysis strategies between these channels were harmonised. The different analyses and their combined results are published in [8].
The $H \to \tau_h\tau_h$ and $H \to \tau_\ell\tau_\ell$ analyses are based on the same dataset as the $H \to \tau_\ell\tau_h$ analysis discussed in Chapter 7. The combination led to an evidence for the Higgs boson Yukawa coupling to τ leptons. The excess of data over the expected background corresponds to an observed (expected) significance of 4.5 (3.4) standard deviations and a signal strength of $\mu = 1.43^{+0.43}_{-0.37}$ is measured which is in agreement with the SM Yukawa coupling strength.
The preliminary combination of the ATLAS with the corresponding CMS analysis [181] even results in an observation with an observed (expected) significance of 5.5 (5.0) standard deviations [182]. The measured signal strength of $\mu = 1.12^{+0.25}_{-0.23}$ is consistent with the predicted coupling strength of the SM. The observation of the Higgs boson coupling to τ leptons is a huge success of the LHC physics program in Run 1.
In this chapter the multivariate analysis of the $H \to \tau_\ell\tau_\ell$ and $H \to \tau_h\tau_h$ as well as the combined ATLAS results of all three channels documented in detail in Reference [8] are briefly summarised. In addition, the results of the cut-based $H \to \tau\tau$ analysis with ATLAS data, which is also documented in Reference [8] are discussed in Section 8.4.

8.1 Analysis of the $H \to \tau_h\tau_h$ channel

Since the branching ratio of $H \to \tau_h\tau_h$ is only slightly smaller than the branching ratio of $H \to \tau_\ell\tau_h$, a significant amount of the combined signal sensitivity originates from this channel. The channel is characterised by one isolated MEDIUM τ_{had} candidate and one isolated TIGHT τ_{had} candidate with opposite electric charge. Events are triggered by a di-τ trigger with p_T thresholds of 29 GeV (20 GeV) for the (sub-)leading τ_{had} candidate. The analysis p_T-thresholds are slightly above the online trigger thresholds. Events with electron or muon candidates are vetoed. Additional cuts on ΔR and $\Delta\eta$ between the two τ_{had} candidates reduce the dominant multijet background. The latter and $Z \to \tau\tau$ processes are the main background sources. Other electroweak processes like $Z \to \ell\ell$ only contribute at a few percent level.
A VBF and boosted signal region similar to the $H \to \tau_\ell\tau_h$ signal regions are defined. In addition, the so-called *rest-category* is defined. It contains all events passing the preselection, but failing the

requirements for the VBF and boosted signal categories. This region is used to constrain the $Z \to \tau\tau$ and mulitjet background. Its signal contamination is negligible. For both signal categories a separate BDT is trained, which is based on a similar list of discriminating variables used for the $H \to \tau_\ell\tau_h$ analysis. The post-fit BDT distributions in the two signal categories are shown in Figure 8.1. Again, the $Z \to \tau\tau$ background is estimated using embedded events, whereas the mulijet background is estimated by a template which is obtained from a fake control region. This control region is given by the signal selection criteria, but the requirements on the isolation is inverted and the two τ_{had} candidates have to have the same electric charge. In order to normalise the multijet and $Z \to \tau\tau$ background the distribution of $\Delta\eta$ between the two τ_{had} candidates is fitted after preselection, whereas the final normalisation is obtained in the global fit. All other background sources are estimated from MC simulation.
Relevant uncertainties arise from the energy calibration of jets and τ_{had} candidates, trigger efficiencies and the estimate of the fake background. Besides the signal region the $\Delta\eta$ distribution of the two τ_{had} candidates in the rest-category is included in the fit, since it provides additional discrimination power between $Z \to \tau\tau$ and multijet events.



(a) (b)

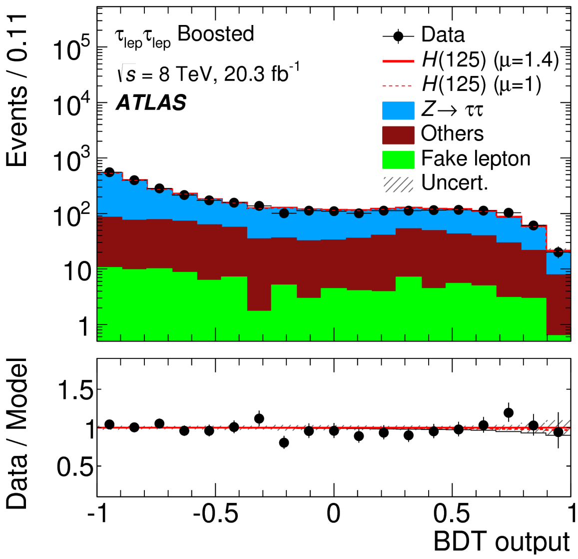
Figure 8.1: Post-fit BDT distributions for the 8 TeV dataset in the boosted (left) and VBF (right) category in the $H \to \tau_h\tau_h$ channel. The Higgs signal is plotted with a signal strength of $\mu = 1$ (dashed line) and best fitted $\mu = 1.4$ (solid line). The hatched band illustrates the statistical and systematic uncertainties. The lower ratio plots show the excess of data over the combined background model including the Higgs signal. The figures are taken from [8].

8.2 Analysis of the $H \to \tau_\ell\tau_\ell$ channel

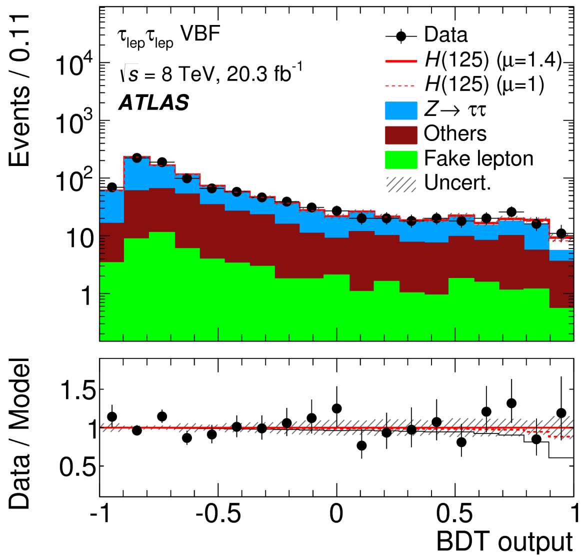
For the $H \to \tau_\ell\tau_\ell$ analysis, exactly two isolated leptons with opposite electric charge are required. All lepton flavour combinations in the final state are investigated: ee, $\mu\mu$ and $e\mu$. Leptons have to exceed a certain transverse momentum threshold. This threshold depends on the trigger conditions that selected the event. They differ between the datasets and lepton flavours. Events with additional identified τ_{had} candidates are vetoed. Additional requirements , e.g. on the invariant mass $m_{\tau\tau}^{\text{vis}}$ or on E_T^{miss}, are applied in order to further suppress background processes. In particular, for the same-flavour final states ee and $\mu\mu$ tighter selection criteria are used to reject $Z \to \ell\ell$ background processes. Events from $t\bar{t}$ decays are reduced by a b-tag veto. Besides $Z \to \ell\ell$ and $t\bar{t}$, $Z \to \tau\tau$ belongs to the dominant background processes. $Z \to \tau\tau$ decays contribute if both τ leptons decay leptonically. Contributions from $H \to WW$ are taken into account as well.

Similarly to the $H \to \tau_\ell\tau_h$ analysis the following two categories are defined: a VBF category, characterised by two jets well-separated in pseudorapidity, and a boosted category which requires $p_T(H) > 100\,\text{GeV}$. For each of these categories a separate BDT is trained. The post-fit BDT distributions are illustrated in Figure 8.2. For the BDT training a different set of variables than for the $H \to \tau_\ell\tau_h$ analysis is used, which mainly exploits the kinematics of the dilepton system.

Embedded events are used to estimate the $Z \to \tau\tau$ background contribution. The normalisation of the



(a)



(b)

Figure 8.2: Post-fit BDT distributions for the 8 TeV dataset in the boosted (left) and VBF (right) category in the $H \to \tau_\ell\tau_\ell$ channel. The Higgs signal is plotted with a signal strength of $\mu = 1$ (dashed line) and using the best fitted $\mu = 1.4$ (solid line). The hatched band illustrates the statistical and systematic uncertainties. The lower ratio plots show the excess of data over the combined background model including the Higgs signal. The figures are taken from [8].

$Z \to \tau\tau$ background is derived within the fit. The other backgrounds with two real leptons, $Z \to \ell\ell$ and $t\bar{t}$, are estimated in dedicated control regions via MC simulation. The $H \to WW$ contribution is directly taken from simulation assuming a Higgs mass of m_H =125 GeV and the corresponding cross-section predicted by the SM. Events with a misidentified lepton (W+jets, multijets and $t\bar{t}$ events) frame another set of background processes. They are estimated using a template fit technique. In this method, a control region is defined by the standard signal selection but with reverted isolation criteria for one of the two leptons. This region is enriched by misidentified leptons and can be used to estimate the shape of the differential distribution. The absolute normalisation for this background is obtained by fitting the transverse momentum spectrum of the sub-leading lepton candidate.

Uncertainties mainly arise from the jet energy scale and the b-tagging efficiency. Also uncertainties related to the embedding technique are relevant. The fit of the $H \to \tau_\ell\tau_\ell$ channel includes the signal regions as well as the $Z \rightarrow \ell\ell$ and top control regions.

8.3 Results of the combined multivariate $H \rightarrow \tau\tau$ analysis

The signal strength is extracted by a global maximum likelihood fit of the BDT distribution in all three channels and signal categories. Including information from different control regions helps to constrain the normalisation of the background. Systematics which concern all channels are treated as correlated. Assuming the Higgs mass to be m_H =125.36 GeV, the combined fit results in a signal strength of

$$\mu = 1.43^{+0.27}_{-0.26}\ (\text{stat.})\ ^{+0.32}_{-0.25}\ (\text{syst.})\ \pm 0.09\ (\text{theo. syst.})\ .$$

In Figure 8.3, the signal strengths measured in the different channels are displayed. They are consistent with the Higgs Yukawa coupling strength ($\mu = 1$) predicted by the SM. The figure shows that the measured signal strength is mainly driven by the combined analysis of the $\sqrt{s} = 8$ TeV VBF category. The systematic uncertainties are mainly composed of uncertainties on the τ and jet energy calibration and uncertainties on the background estimation methods. Statistical uncertainties are of the same size as the experimental uncertainties. The most relevant uncertainties are shown in Table 8.1. The experimental uncertainties are dominated by the jet energy scale (±0.16), followed by uncertainties related to the background normalisation (±0.12) and estimate (±0.12). Also uncertainties related to the τ energy scale (±0.07) and τ identification (±0.06) are relevant. Theoretical uncertainties on the $H \rightarrow \tau\tau$ branching ratio (±0.08) or parton shower modelling (±0.04) play an important role as well.

Table 8.1: Most relevant uncertainties and their impact on the measured signal strength [8].

Source of uncertainty	Uncertainty on $\hat{\mu}$
Signal region statistics	+0.27/-0.26
Jet energy scale	±0.13
τ energy scale	±0.07
τ identification	±0.06
Background normalisation	±0.12
Background estimate	±0.19
Branching ratio $H \rightarrow \tau\tau$	±0.08
Parton shower/underlying event	±0.04
PDF	±0.03

In Table 8.2 the expected and observed significances are summarised which are measured in the combined fit for each channel and category based on the full dataset. The combined excess corresponds to an observed (expected) significance of 4.5 (3.4) standard deviations. The excess of expected and observed data over background is illustrated in Figure 8.5a: The bins of the fitted six BDT distributions (one for each category of each channel) are ordered according to their signal-to-background ratio. In this visualisation the bins with the highest sensitivity are on the very right - independently of the category or channel. The background (grey) and signal estimates (dark red) are obtained from the unconditional fit which yields the signal strength $\mu = 1.4$. In addition, the dashed line illustrates the background estimate derived from a fit in which the signal strength is fixed to zero. The signal event yields under the signal-plus-background hypothesis ($\mu = 1$) are indicated in light red. The measured data event yields are in agreement with the predicted sum of signal and background yields.
A way to visualise the compatibility of the result with the prediction of the SM Higgs boson with mass

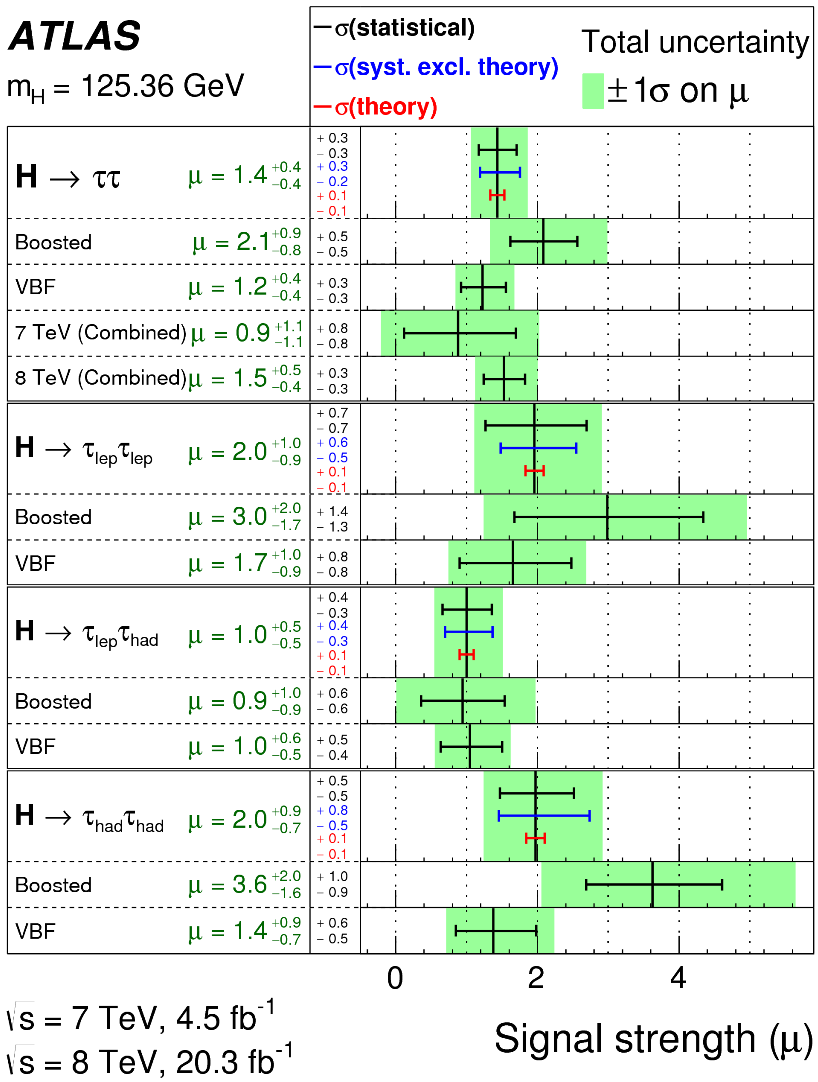


Figure 8.3: Measured best signal strength μ separated for the different channels and categories as well as the combination of the $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV. The total uncertainty on μ is shown in green, while the contributions from the statistical (experimental) uncertainty are shown at the top (middle) in black (blue). The theoretical uncertainty is illustrated in red at the bottom. The figure is taken from [8].

m_H =125 GeV is depicted in Figure 8.4. In this figure, the $\tau\tau$ invariant MMC mass is weighted by the factor ln(1+S/B). The weight enhances contributions from events which are compatible with the signal hypothesis, while other events are reduced. The reweighted distribution shows an excess of data over background at around $m_{\tau\tau}^{\mathrm{MMC}}$ =125 GeV. Including the m_H =125 GeV signal hypothesis and two alternative hypotheses (m_H =110 GeV and m_H =150 GeV) confirms that the measured excess is compatible with the m_H =125 GeV signal hypothesis.

Besides the results discussed above it is also interesting to measure the signal strength with respect to different Higgs production mechanisms. Splitting the signal into ggF production and production processes which involve couplings between the Higgs boson and gauge bosons (VBF and VH) results in

Table 8.2: Observed and expected significance for the individual categories and channels for the combined $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV analysis [8].

Channel	Category	Observed significance	Expected significance
$H \to \tau_\ell\tau_\ell$	VBF	1.88σ	1.15σ
	Boosted	1.72σ	0.57σ
	VBF+Boosted	2.40σ	1.25σ
$H \to \tau_\ell\tau_h$	VBF	2.23σ	2.11σ
	Boosted	1.01σ	1.11σ
	VBF+Boosted	2.33σ	2.33σ
$H \to \tau_h\tau_h$	VBF	2.23σ	1.70σ
	Boosted	2.56σ	0.82σ
	VBF+Boosted	3.25σ	1.99σ
Combined	VBF+Boosted	4.54σ	3.43σ

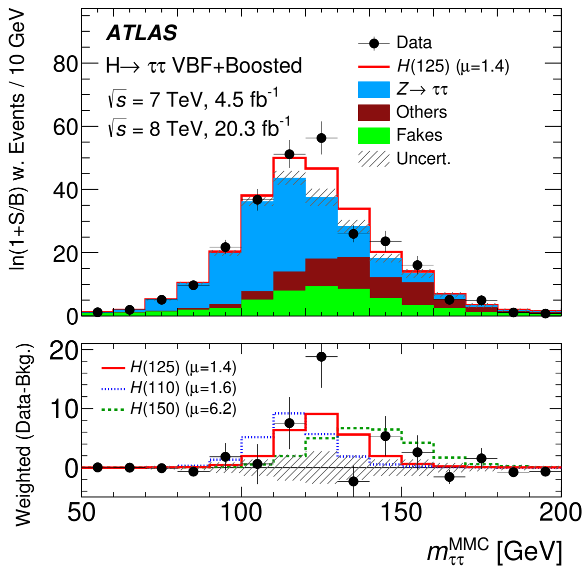


Figure 8.4: Distribution of the $\tau\tau$ invariant MMC mass for data and combined background model in the combined $H \to \tau\tau$ decay channel. The signal is normalised according to the fit result with a signal strength of $\mu = 1.4$. Each event is weighted by the factor ln(1+S/B), where S(B) denotes the predicted signal (background) events in each BDT bin. The ratio depicts the difference between weighted data and background events. The difference between the Higgs signal with mass m_H=125 GeV and background is shown in red. Signal processes with alternative mass hypotheses of m_H=110 GeV and m_H=150 GeV are shown in blue and green, respectively. The figure is taken from [8].

the measured signal strengths

$$\begin{aligned} \mu^{\tau\tau}_{ggF} &= 2.0 \pm 0.8\,(\text{stat.})\,^{+1.2}_{-0.8}\,(\text{syst.}) \ \pm 0.3\,(\text{theo. syst.}) \\ \mu^{\tau\tau}_{VBF+VH} &= 1.24^{+0.49}_{-0.45}\,(\text{stat.})\,^{+0.31}_{-0.29}\,(\text{syst.}) \ \pm 0.08\,(\text{theo. syst.})\,. \end{aligned}$$

Figure 8.5b shows the two-dimensional confidence level contours for $\mu^{\tau\tau}_{ggF}$ and $\mu^{\tau\tau}_{VBF+VH}$. Also here, the measured signal strengths are consistent with the SM prediction.

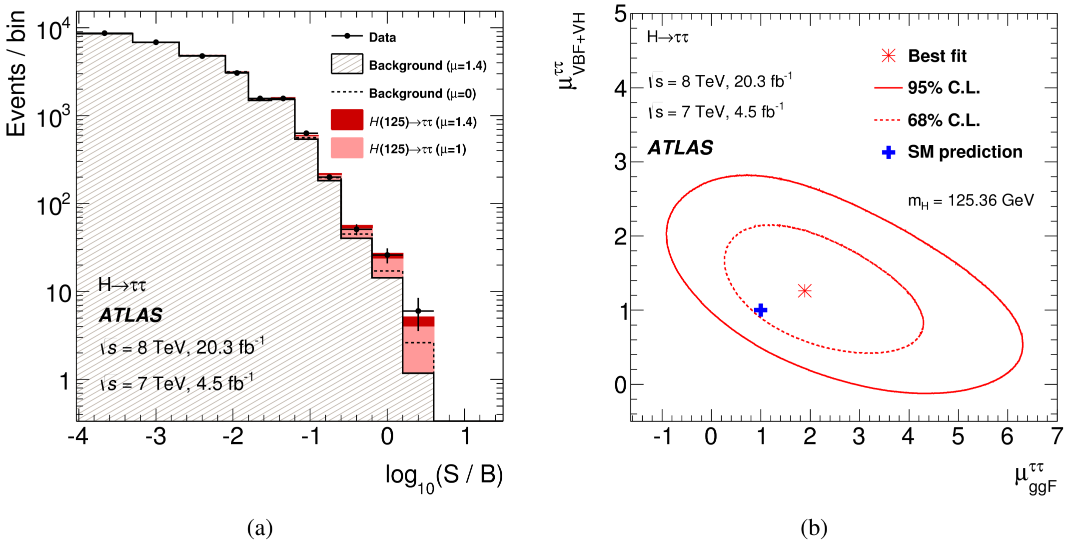


Figure 8.5: Figure (a): Event yields as a function of $\log_{10}(S/B)$ with S (B) referring to the signal (background) events. The yields are measured in the bins of the fitted six BDT distributions. Figure (b): Two dimensional 68% (dashed line) and 95% (solid line) confidence level contours for the combined $H \to \tau\tau$ analysis in the $(\mu^{\tau\tau}_{VBF+VH}, \mu^{\tau\tau}_{ggF})$ plane at m_H =125.36 GeV. The prediction from the SM is indicated by a blue plus symbol, while the best fit value to data is shown as red star. The figures are taken from [8].

8.4 Results of the cut-based $H \to \tau\tau$ analysis

As cross-check for the results of the multivariate analysis discussed above, a cut-based analysis was carried out on the $\sqrt{s}$ =8 TeV dataset and also documented in [8]. In contrast to the previously published cut-based analysis [114], the latest choice of the signal categories and the fit model enables a direct comparison of the results obtained with the cut-based and the multivariate analysis. Subcategories are introduced which yield different results for the signal-to-background-ratio, since they select different phase space regions. By considering additional phase space regions the signal sensitivity of the cut-based analysis is increased. Performing a combined fit of the $m^{\mathrm{MMC}}_{\tau\tau}$ mass distribution results in a measured signal strength of

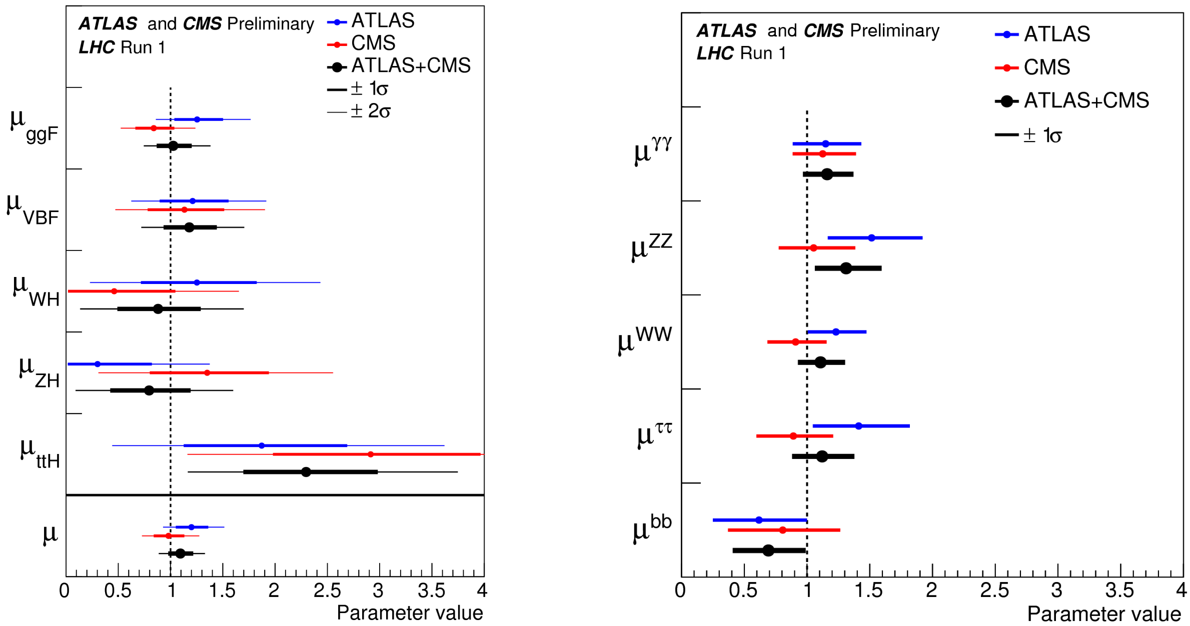
$$\mu = 1.43^{+0.36}_{-0.35}\,(\mathrm{stat.})\ {}^{+0.41}_{-0.33}\,(\mathrm{syst.})\ \pm 0.10\,(\mathrm{theo.\ syst.})$$

for m_H =125.36 GeV[1]. For m_H =125 GeV, the measured significance is 3.2σ and also provides an evidence for the $H \to \tau\tau$ decay. All in all, a good agreement of the cut-based results with the multivariate analysis is observed. In order to quantify the statistical correlation between the measured signal strengths of both analyses the so-called jackknife technique [183, 184] is applied. The correlation is calculated to be in the order of 0.55 and 0.75 for each analysis channel.
The results of the cut-based analysis are a good cross-check of the results obtained with the multivariate analysis.

[1] The mass value corresponds to the measurement of the Higgs boson mass in the $H \to \gamma\gamma$ and $H \to ZZ^* \to 4\ell$ channel published by the ATLAS collaboration [3].

8.5 Discussion

In conclusion, the combined $H \to \tau\tau$ analysis yields a signal which corresponds to an observed (expected) significance of 4.5 (3.4) standard deviations. The measured signal strength of $\mu = 1.4^{+0.4}_{-0.4}$ is consistent with the SM prediction of the Higgs boson coupling to τ leptons.



(a) (b)

Figure 8.6: Measured best combined signal strength μ and best μ for the different (a) Higgs production mechanisms and (b) Higgs decay channels for the combination (black) of ATLAS (blue) and CMS (red) results based on the datasets taken in 2011 and 2012. The uncertainties are displayed by the error bars. The figures are taken from [182].

An analysis of the $H \to \tau\tau$ decay channel published by the CMS collaboration yields similar results [181]. This analysis is based on collision data taken by the CMS detector at $\sqrt{s} = 7\,\text{TeV}$ and $\sqrt{s} = 8\,\text{TeV}$ corresponding to an integrated luminosity of $4.9\,\text{fb}^{-1}$ and $29.7\,\text{fb}^{-1}$, respectively. Assuming a Higgs boson mass of $m_H = 125\,\text{GeV}$ an excess of events over the background-only hypothesis is found resulting in an observed (expected) significance of 3.2 (3.7) standard deviations. The measured signal strength is $\mu = 0.8^{+0.3}_{-0.3}$. The preliminary combination of the $H \to \tau\tau$ analyses of both experiments yields an excess corresponding to an observed (expected) significance of 5.5 (5.0) standard deviations and a measured signal strength of $\mu = 1.1^{+0.3}_{-0.2}$ [182]. Figure 8.6 visualises this result together with the measured signal strengths for the different Higgs boson production mechanisms and for the other Higgs boson decay channels. The observation of the Higgs boson coupling to τ leptons is one of the milestones of the Run 1 physics program of the LHC. It is the first experimental observation of the coupling of the Higgs boson to fermions and therefore of high importance for the establishment of the Higgs mechanism, which gives masses to fermions.
In Run 1 the other important Higgs decay channel with a fermionic final state, the $H \to b\bar{b}$ channel, was also investigated by both collaborations. Searches in this decay channel are difficult due the direct production of $b\bar{b}$ quarks in QCD interactions. In order to cope with the overwhelming background from

multijet events the $H \rightarrow b\bar{b}$ search focuses on the VH production mechanism. The exclusive analyses of the $H \rightarrow b\bar{b}$ channel in Run 1 showed only an excess below 3σ. The ATLAS and CMS collaboration published an excess in the $H \rightarrow b\bar{b}$ channel which corresponds to a significance of 1.4σ [53] and 2.1σ [185] respectively. A combination of the $H \rightarrow \tau\tau$ and $H \rightarrow b\bar{b}$ by the CMS collaboration provided an evidence for couplings to fermions with a significance of 3.8σ [186]. Thus, despite of the difficulties in the reconstruction of the τ lepton decays and the lower branching ratio, the analysis of the $H \rightarrow \tau\tau$ channel is currently more sensitive than the one of the $H \rightarrow b\bar{b}$ channel.
Even after the observation of the Higgs boson Yukawa coupling to τ leptons, analyses of $\tau\tau$ final states remain of importance for future researches at the LHC. Studies of $\tau\tau$ final states based on Run 1 data are still ongoing. For example, an exclusive search in the $H \rightarrow \tau\tau$ decay channel for the SM Higgs boson produced in association with a vector boson was recently published by the ATLAS collaboration [102]. For this search a signal strength of $\mu = 2.3 \pm 1.6$ is measured. Further, the $t\bar{t}H(\rightarrow \tau\tau)$ decay channel is included in the search for $t\bar{t}H$ processes resulting in final states containing leptons [103]. A 5σ observation of the $H \rightarrow \tau\tau$ decay by each experiment separately and an improved sensitivity with respect to different Higgs production mechanisms are certainly goals for the near future. In Run 2, which started in May 2015 with a center-of-mass energy of $\sqrt{s}$ =13 TeV, the LHC will deliver a larger amount of collision data and the cross-section for the Higgs production will increase, so that further improved studies of the Higgs mechanism and a higher precision in the direct measurements of the Higgs boson to τ lepton coupling are possible. With increased data more signal events are available which can be used

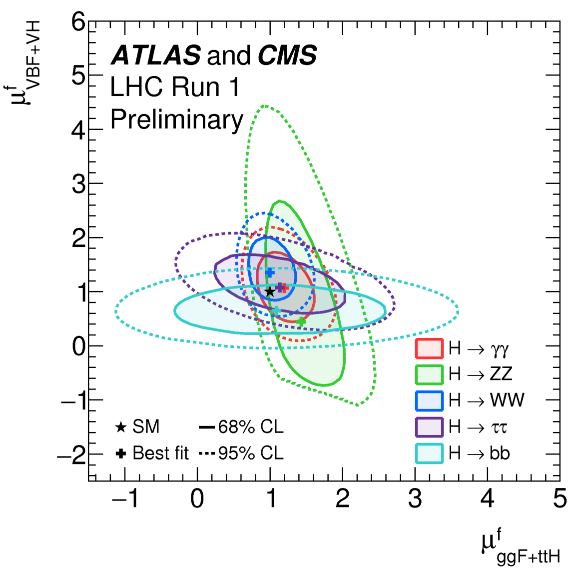


Figure 8.7: Two dimensional 68% (solid line) and 95% (dashed line) confidence level contours for the combined result of ATLAS and CMS in the $(\mu_{VBF+VH}, \mu_{ggF})$ plane and the best-fit values to the data and SM prediction. The different colours of the contours refer to the different Higgs boson decay channels. The figure is taken from [182].

to improve the separation between ggF and VBF. This improved separation will contribute to combined coupling measurements. In particular, the $H \rightarrow \tau\tau$ boson decay channel will remain one of the most sensitive Higgs channels to the VBF production. The importance of the $H \rightarrow \tau\tau$ channel for the measurement of the combined coupling is visible in Figure 8.7. It shows the coupling strengths μ_{VBF+VH} and μ_{ggF} for the different Higgs decay channels. The signal strength μ_{VBF+VH} is strongly constrained in the $H \rightarrow \tau\tau$ channel. Moreover, the $H \rightarrow \tau\tau$ channel poses one possibility to measure the spin and

parity quantum numbers of the Higgs boson [187]. For example, a first test of the CP invariance in VBF production in the $H \to \tau\tau$ decay channel based on the $\sqrt{s}$ =8 TeV dataset was recently published by ATLAS [188]. Furthermore, $\tau\tau$ final states are of importance for ongoing studies of *beyond the Standard Model scenarios* [189].
All in all, there is a strong motivation to further optimize the analysis of the $H \to \tau\tau$ decay by new analysis concepts going above the multivariate approach presented in this chapter.

CHAPTER 9

A multiple BDT analysis of the $H \to \tau_\ell \tau_h$ decay

The key role in the $H \to \tau_\ell \tau_h$ analysis presented in Chapter 7 is played by the trained BDT classifiers. These BDTs are used for two purposes: First, they are used to discriminate between signal and background events and hence increase the signal-to-background ratio. With this functionality they complement the loosely defined signal categories, which are mainly used to select events with kinematic properties associated to a Higgs boson produced via ggF (boosted category) or VBF (VBF category). Secondly, the distribution of the BDT score is fitted in order to test the compatibility between observed data and SM predictions. This multivariate strategy results in a high signal sensitivity and an evidence of the Higgs boson to τ coupling.

In contrast, the cut-based analysis (see Section 8.4) separates the background suppression from the signal extraction: Selection cuts defining the subcategories are mainly used to suppress background processes, while the distribution of the MMC mass is used in the fit. Since the cut-based analysis provides a smaller signal sensitivity, the multivariate approach is favoured.

For future analyses it would be desirable to use an ansatz which is able to combine advantages of the current cut-based and multivariate analyses. This means that an advanced multivariate strategy should be used which decouples the background suppression from the signal extraction and at the same time maintains a high signal sensitivity.

Several possible approaches were discussed during the Run 1 phase and are summarised in Section 9.1. One option is to train the BDTs without using information about the $m_{\tau\tau}^{\text{MMC}}$ mass. A phase space region with a large signal-to-background ratio is defined by the high BDT score regions. Events in this region can then be used to fit the $m_{\tau\tau}^{\text{MMC}}$ mass distribution. Since the mass has a large discrimination power which is then not exploited in this approach, this simple approach, however, leads to a significant decrease in sensitivity. This loss in sensitivity has to be recovered in some way, e.g. by tuning the BDT classifier.

In the currently used BDT classifier all background processes are treated equally. The background sources, however, are very different. For example, one can distinguish between background events with true and fake τ_{had} signatures. Due to the different origin of background events the discrimination power of the input variables (see Section 5.6) against signal depends on the particular background process.

The characteristics of the background composition can be exploited by the training of multiple BDTs. In this approach one BDT is trained for each kind of background in order to discriminate between this kind of background and the signal process. An example of such a multiple BDT approach is the 2-BDT analysis, which was investigated in the context of this thesis. The BDT classifiers distinguish between background events with true and fake τ_{had} signatures. The concept of the 2-BDT analysis and its results are discussed in the following.

9.1 Alternative approaches

Several advanced alternatives to the current multivariate analysis strategy are possible extensions to the $H \to \tau\tau$ analysis in Run 2: From simply removing the $m_{\tau\tau}^{\text{MMC}}$ from the BDT training, over BDT training at several mass points, to alternative BDTs and fits. In the following, the different approaches are briefly introduced and their advantages and disadvantages are discussed. The 2-BDT approach which is also an alternative ansatz is described separately in Section 9.2.

9.1.1 Higgs mass dependent BDT training

The BDT training described in Section 7.6 is based on the mass hypothesis m_H =125 GeV. However, it can be naturally extended to other Higgs boson mass hypotheses. By training the BDTs for Higgs masses between m_H =110 GeV and m_H =150 GeV, the expected and observed significance is measured for several mass points. This method provides by construction the same signal sensitivity at m_H =125 GeV as the analysis discussed in Chapter 7 but since the analysis is repeated for different hypotheses, additional mass sensitivity is obtained. For that reason, it is certainly an improvement over the default multivariate analysis. But this strategy has some important disadvantages. First, re-analysing the data for several mass points increases the required validation of analysis-relevant distributions in signal and control regions linearly with the number of trained BDTs. Second, high statistics signal samples for all investigated mass points are necessary. The sample production requires a large amount of computer resources.

9.1.2 BDT classifier without dependence on $m_{\tau\tau}^{\text{MMC}}$

The variables used for the BDT training (see Table 7.12) contribute similarly to the performance of the BDT. By removing the $m_{\tau\tau}^{\text{MMC}}$ from the list of input variables, the discrimination between $Z \to \tau\tau$ and signal is affected, whereas the other discriminating variables have enough discriminating power to sufficiently suppress all other background contributions. For that reason, the BDTs can be retrained without $m_{\tau\tau}^{\text{MMC}}$. As a consequence, the BDT score distribution is not directly biased with respect to one particular Higgs boson mass hypothesis so that one can cut on the BDT score distribution at a certain threshold. The threshold is given by the BDT score value above which the signal over background ratio exceeds a well-defined value e.g. 0.3. For all events, which have a larger BDT score threshold, the $m_{\tau\tau}^{\text{MMC}}$ distribution is fitted.

This method has some similarities to the τ BDT ID, in which also only events above a particular working point are used for further analyses. The advantage of this method is that the background suppression and signal extraction are disconnected. Only events in the high BDT score region and hence in a phase space region with an improved signal-to-background ratio are used in the fit. Nevertheless, removing the $m_{\tau\tau}^{\text{MMC}}$ mass from the list of training variables reduces the overall sensitivity by about 50%.

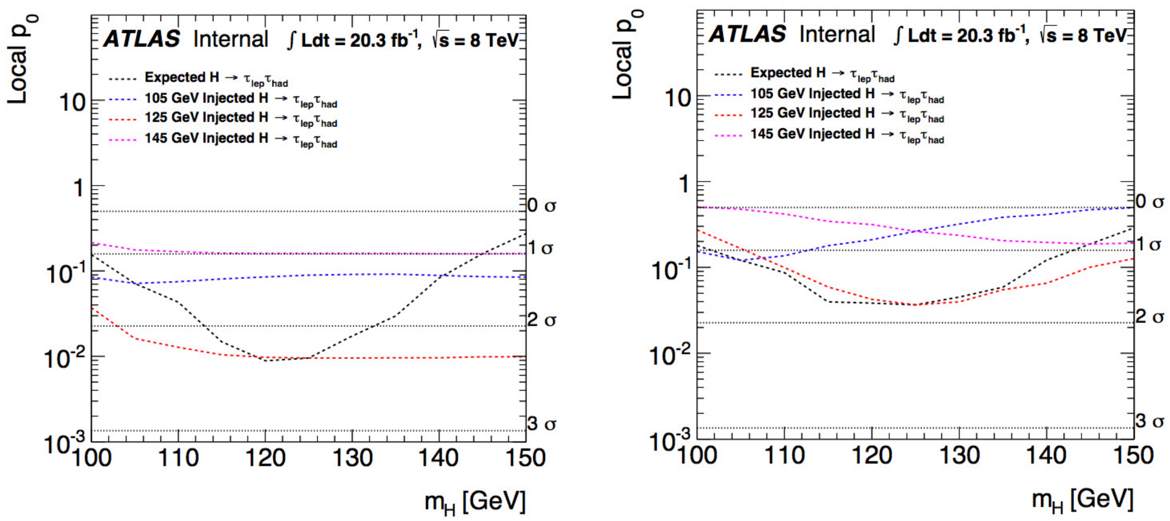
In order to further pursue this approach, a re-optimisation of the BDT input variables is required. It is questionable if such an optimisation can result in the same sensitivity as the default multivariate approach as long as all background sources are treated equally within the BDT classification.

9.1.3 Optimisation of the cut-based analysis

As mentioned earlier, the cut-based analysis discussed in Section 8.4 directly fits the $m_{\tau\tau}^{\mathrm{MMC}}$ distribution in the signal subcategories. Compared to the cut-based analysis in [114] the analysis was further optimised in order to increase its signal sensitivity. Much effort was put in the definition of subcategories with which the sensitivity of this analysis was indeed enhanced. Nevertheless, its total sensitivity is still below the sensitivity reached with the multivariate analysis.
Since the defined subcategories do not result in a compatible signal sensitivity, a multivariate approach is still favoured.

9.1.4 BDT slicing

The BDT slicing method combines the high sensitivity obtained from multivariate classifiers with the benefits of a direct fit of the $m_{\tau\tau}^{\mathrm{MMC}}$ distribution. The BDT classifier is trained without using $m_{\tau\tau}^{\mathrm{MMC}}$ as an input variable. Based on BDT score values different categories LOOSE, MEDIUM and TIGHT are defined. For each of these different BDT score regions $m_{\tau\tau}^{\mathrm{MMC}}$ is fitted separately. For this method, the list of discriminating variables used for the BDT training has to be re-optimised as well as the BDT slices need to be tuned for each channel and category. However, prospective studies [190] demonstrate that the fit of the $m_{\tau\tau}^{\mathrm{MMC}}$ distribution in slices of the BDTs described in Section 7.6.1 but without the $m_{\tau\tau}^{\mathrm{MMC}}$ as input variable already gives some mass sensitivity with a slightly smaller significance, as depicted in Figure 9.1b. The figure shows the local p-value as function of the Higgs boson mass. The black curve shows the expected significance with its minimum at around 125 GeV, while the coloured lines display different injected Higgs mass hypotheses. If a Higgs signal with mass m_H =125 GeV is injected (red curve) and the $m_{\tau\tau}^{\mathrm{MMC}}$ based BDT score distribution is fitted (Figure 9.1a) the observed sensitivity remains the same over a broad range. In contrast, if the BDT slicing ansatz is used (Figure 9.1b) the



(a) (b)

Figure 9.1: Comparison of local p_0 value as function of the Higgs mass in the $H \rightarrow \tau_\ell \tau_h$ channel for (a) fitting the BDT trained on m_H =125 GeV and (b) fitting the MMC mass in slices of the BDT output. The black curve shows the expected p_0 value, while the coloured curve show the results for different injected Higgs masses. The figures are taken from [190].

highest observed sensitivity is reached in the range of 120 GeV< m_H <130 GeV. Since the dependence of the BDT classifier on $m_{\tau\tau}^{\mathrm{MMC}}$ is removed, the overall significance is reduced.
This BDT slicing method might be a good candidate for future analysis strategies provided that the BDT classification and slices can be further optimised.

9.2 Two-BDT approach

The above suggestions represent improvements of the multivariate concept used in the Run 1 $H \to \tau\tau$ analysis. Nevertheless, they still have some disadvantages: Either they include the $m_{\tau\tau}^{\mathrm{MMC}}$ mass in the BDT training and therefore bias the BDT score distribution with respect to a certain mass hypothesis, or they result in a smaller sensitivity due to removing the mass from the BDT classifier.
In the following, the two-BDT approach is introduced. The idea is to remove the dependence of the BDT classifier on the Higgs boson mass. Since the removal causes a loss in sensitivity, the sensitivity has to be retrieved by optimising the BDT with respect to individual background sources. These background processes can be arranged in two classes: The resonant background $Z \to \tau\tau$ with the same event signature as the signal process and all other, non-resonant processes. The use of two BDTs, one trained against $Z \to \tau\tau$, the other against all other non-resonant background processes, optimises the separation between signal and background. Events which are situated in the high BDT score region of both BDT score distributions are used to fit the $m_{\tau\tau}^{\mathrm{MMC}}$ mass distribution.
As shown later, this approach results in a compatible signal sensitivity as the published multivariate analysis, while the background suppression is disconnected from the signal extraction. The two-BDT approach was originally developed for the VBF category [112]. The VBF topology has the largest discriminating power besides the $m_{\tau\tau}^{\mathrm{MMC}}$ mass. Since the VBF topology compensates for the removal of the $m_{\tau\tau}^{\mathrm{MMC}}$ mass from the BDT classifier, the VBF category is very suitable to test this approach. Although in the VBF category a signal sensitivity is achieved which is compatible to the current analysis, it was not a-priori known if this approach is also applicable to the boosted category. In this category the dominant Higgs production mechanism is the ggF production process. For that reason, the discrimination between signal and $Z \to \tau\tau$ events has to rely on other variables than those related to the VBF topology. Only if other variables can be found which recover the loss in sensitivity due to the omission of the $m_{\tau\tau}^{\mathrm{MMC}}$ mass, this approach is suitable for future analyses. In the context of this thesis, it was investigated if the two BDT-training approach is also applicable to the boosted category. Different BDT input variables were tested to improve the separation power of the two BDTs trained in the boosted category. Also variables were investigated which could be used for the classification in the VBF category. In addition, a combined analysis of the VBF and boosted category based on the 2-BDT approach was carried out.
In the following, the two BDT training in both categories for the 8 TeV dataset is introduced. Further, it is discussed how the signal is extracted and the results are compared to the results obtained with the published analyses in Section 7.9 and Section 8.4.
The presented analysis is based on the same 8 TeV data samples and event selection described in Section 7.1 and Section 7.2. The 7 TeV dataset is not considered, because of its small contribution to the overall sensitivity. Furthermore, in order to avoid overtraining the individual BDT training against different background sources requires even more statistics than the training of a single BDT.

9.2.1 Training of two BDTs

The BDT classifier discussed in Section 7.6 provides a good separation between signal and background processes. Hence, the same setup for the BDT training as discussed in Section 7.6 is used in the following. To increase the number of background events against which the BDTs are trained, the cross-

evaluation technique is extended, such that 75% of the background events are used for the training and 25% for testing.
The variables used for the BDT training in the VBF signal category are summarised in Table 9.1. The BDT trained against the $Z \to \tau\tau$ background is called *resonant* BDT below. For this BDT, variables are chosen which are suitable to distinguish between VBF-jets and jets from $Z \to \tau\tau$ events. For example, the invariant dijet mass m_{jj} should be larger than in strongly produced $Z \to \tau\tau$ events and the jets are expected to be well-separated in space ($\Delta\eta_{jj}$, $\Delta\Phi_{jj}$). The position of its visible decay products with respect to the jets (η-centrality) is also suitable for separation, since the Higgs is typically emitted between both VBF jets. The other BDT, trained in the VBF signal region, is used to separate the Higgs signal from all other processes, i.e. processes with fake $H \to \tau\tau$ signatures like W+jets, $t\bar{t}$ or multijet events. This BDT is referred to as *non-resonant* BDT in the following. It exploits input variables which are mainly suitable to distinguish between true $H \to \tau_\ell\tau_h$ signatures and fake signatures. For example, the W+jets events are rejected by using the transverse mass. In the boosted category, the separation between signal and $Z \to \tau\tau$ is difficult, since the ggF production mechanism is dominant in this category. Hence, the use of VBF topologies is no longer suitable. Since the invariant $\tau\tau$ mass is also not used there are

Table 9.1: Variables used for the BDT training of the Higgs signal with mass m_H =125 GeV against the $Z \to \tau\tau$ (resonant BDT) and all other background processes (non-resonant BDT) for the VBF category based on $\sqrt{s}$ =8 TeV data samples.

Variable	resonant BDT	non-resonant BDT
$m_T(E_T^{\text{miss}}, \ell)$		✓
$m_T(E_T^{\text{miss}}, \tau_{\text{had}})$		✓
Δp_T		✓
$p_T(\tau_{\text{had}})/p_T(\ell)$		✓
$\Delta R(\tau_{\text{had}}, \ell)$		✓
$E_T^{\text{miss}}\phi-$ centrality		✓
$x_{\tau_{\text{had}}} \cdot x_\ell$		✓
$\Delta\phi_{jj}$	✓	
m_{jj}	✓	
$\Delta\eta_{jj}$	✓	
$\eta-$ centrality	✓	
p_T^{tot}	✓	

Table 9.2: Variables used for the BDT training of the Higgs signal with mass m_H =125 GeV against the $Z \to \tau\tau$ (resonant BDT) and all other background processes (non-resonant BDT) for the boosted category based on $\sqrt{s}$ =8 TeV data samples.

Variable	resonant BDT	non-resonant BDT
$m_T(E_T^{\text{miss}}, \ell)$	✓	✓
$m_T(E_T^{\text{miss}}, \tau_{\text{had}})$	✓	
$\sum p_T$	✓	✓
$\Delta R(\tau_{\text{had}}, \ell)$		✓
$E_T^{\text{miss}}\phi$-centrality	✓	✓
$p_T(\tau_{\text{had}}) - p_T(\ell)$	✓	
$p_T(\tau_{\text{had}})/p_T(\ell)$	✓	
η(leading jet)	✓	

only few variables left which at least have some discriminating power against background events with true $H \to \tau_\ell \tau_h$ signatures.
Several variables and their performance within the BDT have been tested. The variables finally used are listed in Table 9.1 and Table 9.2. For the non-resonant BDT variables are again used which provide discrimination power against fake τ_{had} signatures. An importance ranking is performed to check the relative importance of those variables within the event classification. The results are summarised in Table 9.3 and Table 9.4 for the VBF and boosted category, respectively. In the resonant BDT of the VBF category, the η-centrality has the largest discrimination power. The omission of $m_{\tau\tau}^{\text{MMC}}$ in the training is mostly compensated by using features of VBF jets. This behaviour is expected from the performance checks of the earlier discussed BDT training in Section 5.2.1, where the $m_{\tau\tau}^{\text{MMC}}$ variable is not highest ranked (see Table 7.15). In the training of the non-resonant BDT, the transverse mass contributes most to the discrimination of the BDT. This is expected due to the dominant W+jets and multijet background. The corresponding *ROC*-curves are illustrated in Figure 9.2. The non-resonant BDT in Figure 9.2a shows a slightly better performance than the resonant BDT in Figure 9.2b. Nevertheless, the resonant BDT features an overall good background rejection. Their overall performance is compatible to the single trained BDT in the VBF category. The difficulties due to the dominance of

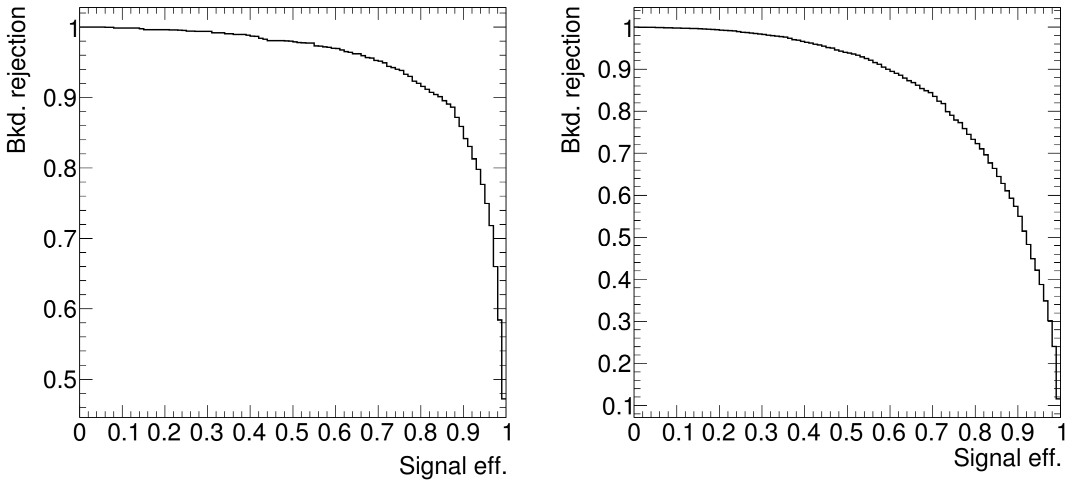
Table 9.3: Importance ranking of variables used as input for the 8 TeV training of the resonant (left) and non-resonant (right) BDT in the VBF category.

Variable	Variable importance	Variable	Variable importance
η-centrality	0.211	$m_T(E_T^{\text{miss}}, \ell)$	0.299
$m_{j,j}$	0.194	$E_T^{\text{miss}}\phi$-centrality	0.285
$\Delta\eta(j,j)$	0.168	$p_T(\tau_{\text{had}})/p_T(\ell)$	0.209
p_T^{tot}	0.046	$p_T(\tau) - p_T(\ell)$	0.187
$\Delta\phi_{jj}$	0.024	$\Delta R(\tau_{\text{had}}, \ell))$	0.155
		$m_T(E_T^{\text{miss}}, \tau_{\text{had}})$	0.063
		$x_{\tau_{\text{had}}} \cdot x_\ell$	0.003

Table 9.4: Importance ranking of variables used as input for the 8 TeV training of the resonant (left) and non-resonant (right) BDT in the boosted category.

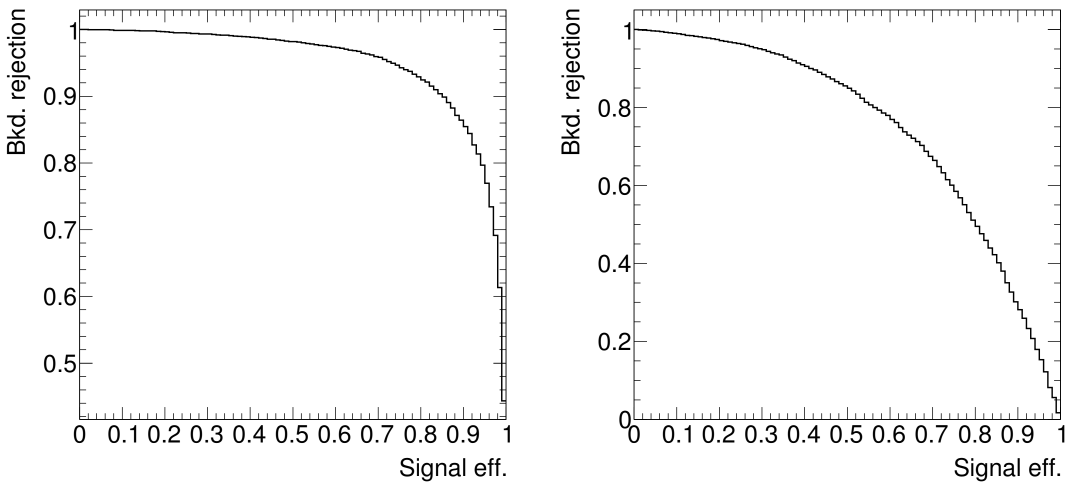
Variable	Variable importance	Variable	Variable importance
$m_T(E_T^{\text{miss}}, \tau_{\text{had}})$	0.043	$m_T(E_T^{\text{miss}}, \ell)$	0.268
$\sum p_T$	0.018	$E_T^{\text{miss}}\phi$-centrality	0.250
$m_T(E_T^{\text{miss}}, \ell)$	0.011	$\Delta R(\tau_{\text{had}}, \ell))$	0.231
$p_T(\tau) - p_T(\ell)$	0.009	$\sum p_T$	0.035
η(leading jet)	0.009		
$E_T^{\text{miss}}\phi$-centrality	0.006		
$p_T(\tau_{\text{had}})/p_T(\ell)$	0.004		

the ggF production process in the boosted category are visible in the variable ranking in Table 9.4. The variables used for the training of the non-resonant BDT have similar importance values compared to the input variables of the non-resonant BDT of the VBF category. The highest ranked variable is again the transverse mass. Variables in the ranking of the resonant BDT have rather low importance values. Since characteristics of VBF topologies cannot be exploited, the Higgs signal is difficult to separate from the $Z \to \tau\tau$ background. Since also the shape of the $m_{\tau\tau}^{\text{MMC}}$ mass which is the highest ranked variable of the single boosted BDT in Section 5.2.1 is not exploited, the resonant BDT is based on variables importance values which together own some discrimination power. The corresponding *ROC*-curve is shown in Figure 9.3b. Compared to the *ROC*-curve of the non-resonant BDT in Figure 9.3a and to the two BDTs trained in the VBF category it has the worst performance. Nevertheless, since the boosted category is the less sensitive signal category, its performance is sufficient for the further analysis.



(a) (b)

Figure 9.2: The background rejection as a function of the signal efficiency for the BDT trained on Higgs signal processes and non-resonant (left) and resonant (right) background sources, respectively, in the VBF 8 TeV signal region.

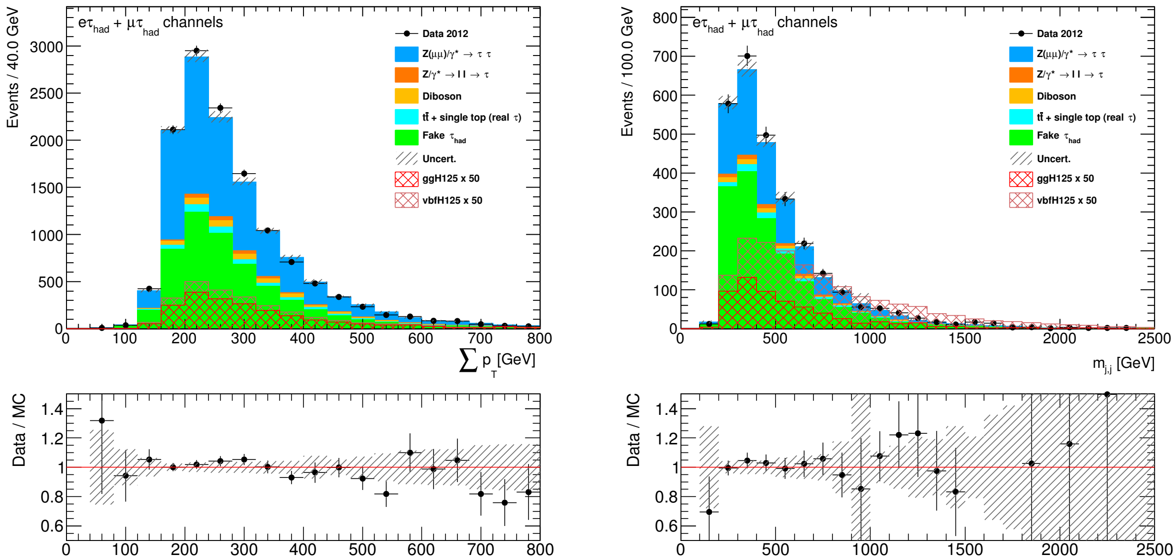


(a) (b)

Figure 9.3: The background rejection as a function of the signal efficiency for the BDT trained on Higgs signal processes and non-resonant (left) and resonant (right) background events, respectively, in the boosted 8 TeV signal region.

9.2.2 Validation of input variables and BDT distribution

The modelling of variables entering the BDT training must be validated in the signal regions as well as in the control regions (see Section 7.6.2). In Figure 9.4 the summed transverse momentum and the dijet mass are shown as examples for the BDT input variables. Further variables can be found in Appendix C.1 and D.1. While in Appendix C.1 variables which were already used in the context of the earlier presented analysis are collected, in Appendix D.1 only variables are summarised which were not yet investigated. A well-behaved modelling of these variables is observed.



(a) (b)

Figure 9.4: Two examples for variables used in the BDT training: (a) the summed transverse momentum $\sum p_T$ in the boosted category and (b) the dijet mass in the VBF category. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

Figure 9.5 shows the unblinded (non-)resonant BDT score distributions in the signal regions. The BDT score distributions of the corresponding control regions can be found in Appendix D.2. In both signal regions, the non-resonant BDT score distribution demonstrates its discriminating power against all non-resonant background sources. While contributions from background processes with fake τ_{had} signature, like W+jets, are enhanced in the low BDT score regions, the $Z \to \tau\tau$ significantly contributes in the high BDT score regions. In order to achieve a discrimination between Higgs signal and $Z \to \tau\tau$ background, the resonant BDT is employed. High BDT score values are assigned to signal events, while the resonant background is accumulated at lower values. As expected, the resonant BDT in the boosted region separated both physical processes worse. The distribution of the $Z \to \tau\tau$ background peaks at low values, whereas the distribution of the signal processes is flat over a wide range. With an appropriate cut value on the resonant BDT score distributions the $Z \to \tau\tau$ background can be essentially reduced and hence the signal sensitivity is increased, when fitting the mass distribution. To find the best thresholds for the resonant and non-resonant BDT, the fit of the mass distribution is repeated with subsequently increased cut values and their yielded signal sensitivities are compared with each other (see Section 9.2.4).

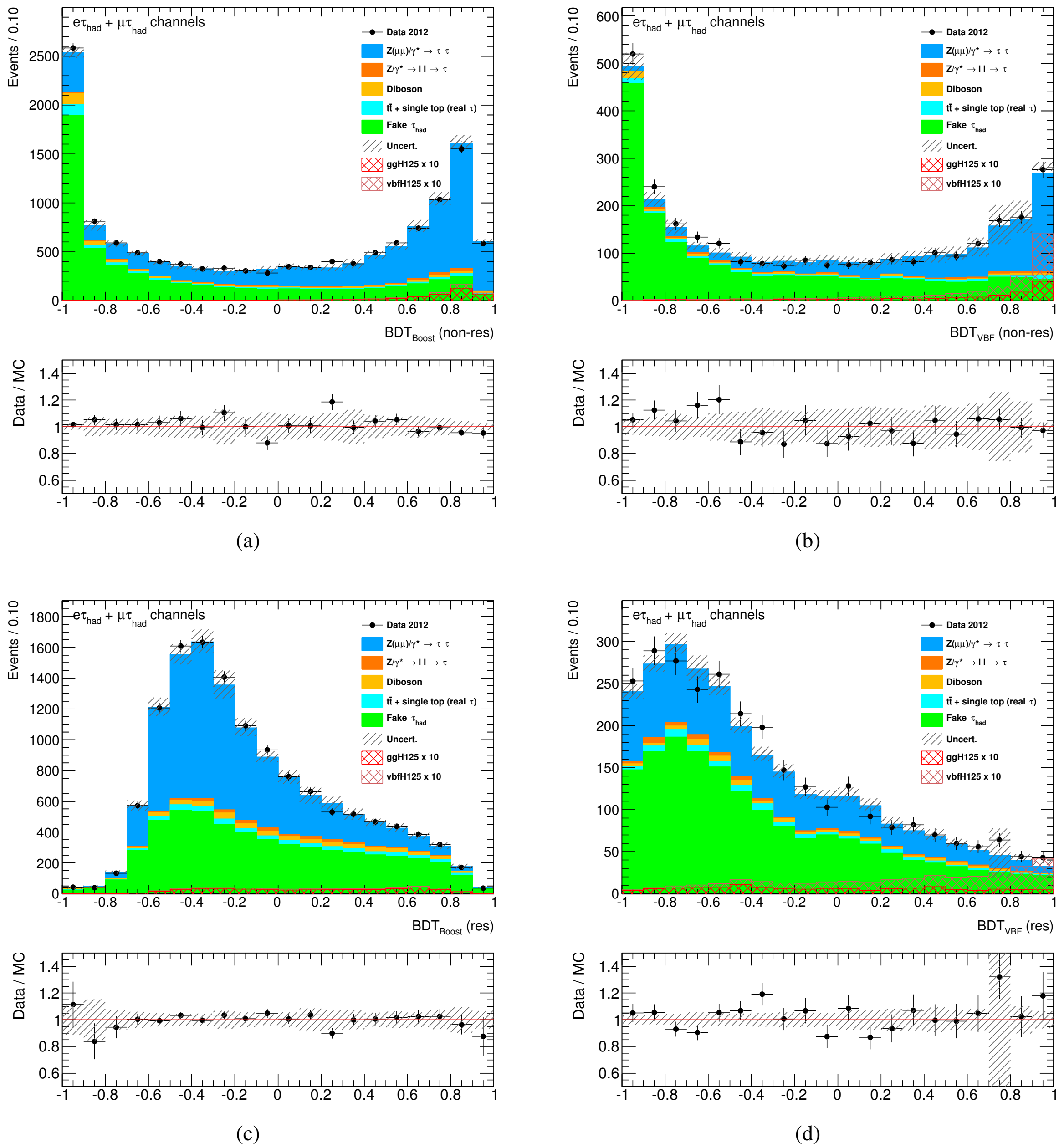


Figure 9.5: Unblinded pre-fit BDT score distribution for the boosted (left) and VBF (right) 8 TeV signal category. The upper plots show the BDTs trained against non-resonant background, like W+jets or mulitjet events, the lower plots depict the BDTs employed to separate $Z \to \tau\tau$ background from Higgs signal processes. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

9.2.3 Systematic uncertainties

The 2-BDT approach differs from the analysis in Chapter 7 in the BDT training and signal extraction, whereas the background model is the same. For that reason, all uncertainties, which are discussed in Section 7.7 are also applied here.

9.2.4 Fit model and signal extraction

The fit model including pruning and smoothing algorithms is analogue to the model described in Section 7.8: Based on the nominal and systematically varied shape of the $m_{\tau\tau}^{\mathrm{MMC}}$ mass distribution, the HistFactory machinery performs a combined likelihood fit in the boosted and VBF signal region. The signal regions are further restricted by additional cuts on the non-/resonant BDT score distribution which ensures a signal enriched phase space region (see below). The mass distribution is fitted in the range between $0\,\mathrm{GeV} < m_{\tau\tau}^{\mathrm{MMC}} < 400\,\mathrm{GeV}$ using the binning in Table 9.5. By including the corresponding top control regions the normalisation of the top background is obtained.

Table 9.5: Binning of the MMC mass distribution used for the fit of the 8 TeV dataset.

Category	Binning
Boosted	[40, 60, 80, 100, 110, 120, 130, 140, 150, 180, 200]
VBF	[40, 60, 80, 100, 110, 120, 130, 140, 150, 180, 200]

Cut optimisation

For the fit of the $m_{\tau\tau}^{\mathrm{MMC}}$ mass distribution, a phase space region enriched by signal events has to be defined. As visible from Figure 9.5, such a region is given by the combined high BDT score regions of the resonant and non-resonant BDT: While selecting the high BDT score region of the non-resonant BDT increases the signal-to-background ratio with respect to processes with fake τ_{had} signatures, an additional cut on the resonant BDT further reduces the background contamination due to $Z \to \tau\tau$ events. Due to the high signal-to-background ratio in the obtained phase space region, a fit of the $m_{\tau\tau}^{\mathrm{MMC}}$ mass distribution has large signal sensitivity.

In order to find the best BDT thresholds, the cut values are subsequently raised and the fit is repeated. Finally, the fit results are scanned for the cut values which have the highest signal sensitivity. While choosing the best cut values one should consider that slightly different cut values on the BDT scores must not result in a rapid change of the sensitivity in order to avoid local maxima due to fluctuations.

The best cut values are summarised in Table 9.6. In the non-resonant BDT score distribution, the signal is enlarged in the high BDT score bins so that high cut values are chosen. In the case of the resonant BDT, however, the signal is more flat distributed, so that a lower cut threshold is used.

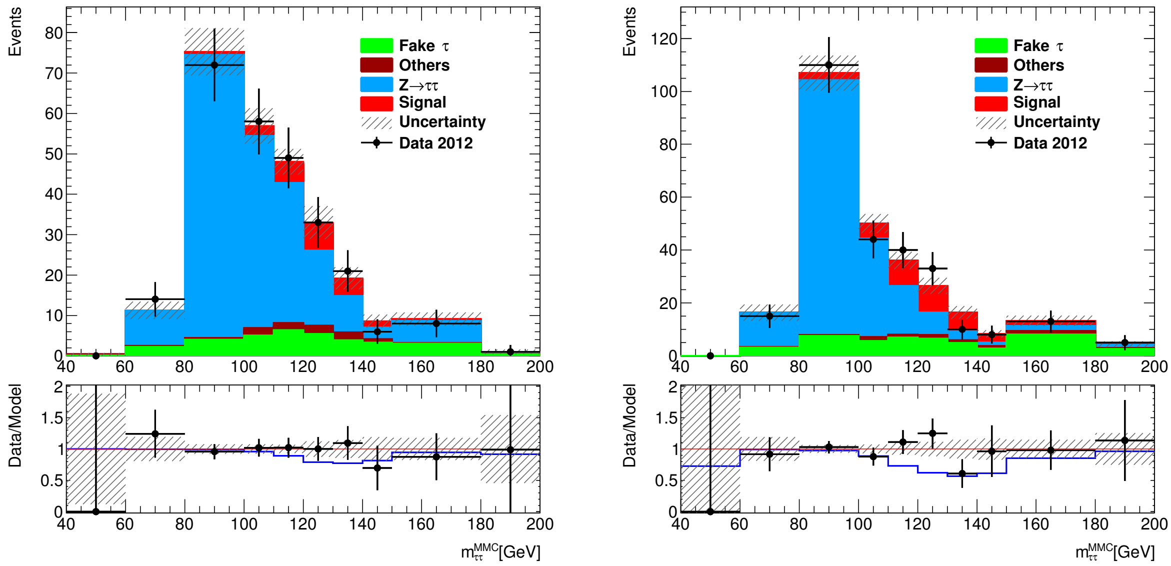
In the following the *boosted* and *VBF* category are redefined as the signal categories documented in Section 7.2.4. In addition, events have to pass the cuts summarised in Table 9.6.

Table 9.6: Cut values on the non-/resonant BDT score used for the boosted and VBF signal region. Only events above these thresholds are employed to fit the $m_{\tau\tau}^{\mathrm{MMC}}$ mass distribution.

Category	resonant BDT score	non-resonant BDT score
Boosted	0.2	0.85
VBF	-0.3	0.75

9.2.5 Results

The signal strength μ is obtained from the simultaneous likelihood fit of the $m_{\tau\tau}^{\mathrm{MMC}}$ mass distribution in the two 8 TeV signal regions and the two corresponding single binned top control regions. The post-fit mass distributions are illustrated in Figure 9.6. Within uncertainties the signal-plus-background model



(a) (b)

Figure 9.6: MMC mass distribution for the boosted (left) and VBF (right) 8 TeV signal region after the maximum likelihood fit. The signal is fitted with its best signal strength $\mu = 1.3$. The lower ratio plots show the ratio between data and the signal-plus-background model including the statistical and systematic uncertainties of the model (hatched band). The blue line illustrates the ratio between the background-only and the signal-plus-background model.

matches the observation in data. At around 125 GeV, a clear excess of data over the background-only hypothesis is observed.
The best estimator of the strength μ at m_H =125 GeV yields

$$\hat{\mu} = 1.31^{+0.32}_{-0.31}\,(\text{stat.})\ ^{+0.48}_{-0.38}\,(\text{syst.})\ \pm 0.1\ (\text{theo. syst.})\ .$$

Although the best-fit value is slightly larger than the one obtained from the fit of the BDT score distribution (see Table 7.20), within uncertainties, it is consistent with the $\mu = 1$ hypothesis. The signal strengths obtained from the single fit of the boosted and VBF region are summarised in Table 9.7. All results are consistent with the m_H =125 GeV hypothesis within uncertainties. The combined fit result is mainly driven by the VBF category. Since the measured signal strength in both categories are larger than 1, the combined signal strength is pulled in the same direction.
As before, the excess in data is tested against the background-only hypothesis. The overall observed (expected) significance yields 2.7(2.5) Gaussian standard deviations and hence is slightly larger than obtained from the default multivariate analysis in Chapter 7. The results from the separate fit of the boosted and VBF category are listed in Table 9.8. The abstinence of the MMC mass as discriminator in the BDT training is compensated by the BDT tuning with respect to individual background classes.
Finally, the importance of the nuisance parameters is illustrated in Figure 9.7. Highest ranked is the

Table 9.7: Estimate of the signal strength $\hat{\mu}$ obtained from fits of the individual 8 TeV signal categories (boosted and VBF) and their combination. The error includes statistical and systematic uncertainties.

Category	Signal strength
Boosted	$1.10^{+1.12}_{-1.10}$
VBF	$1.21^{+0.60}_{-0.51}$
Combined	$1.31^{+0.59}_{-0.50}$

Table 9.8: Observed (expected) local significance in terms of Gaussian standard deviations evaluated for the individual 8 TeV signal categories (boosted and VBF) and their combination.

Category	Significance
Boosted	0.8(0.6)
VBF	2.5(2.4)
Combined	2.7(2.5)

normalisation of the top background in the VBF category which is calculated from the fit. It is followed by nuisance parameters associated with the τ and jet energy scale. Also the cell subtraction and isolation systematic which corresponds to the dominant $Z \to \tau\tau$ background have a large impact on the measured signal strength. The highest ranked theoretical uncertainty is related to the calculated $H \to \tau_\ell \tau_h$ branching ratio. Within uncertainties the best-fit values of all nuisance parameters are in agreement with their pre-fit values.

9.3 Conclusion

The results of the 2-BDT analysis based on the $\sqrt{s}$ =8 TeV dataset are compared to the results of the 1-BDT analysis presented in Chapter 7 and the cut-based analysis in Section 8.4. The significances and measured signal strengths obtained from these three analyses are summarised in Table 9.9. All results are consistent with the Yukawa coupling strength in the SM. The uncertainties of the measured signal strengths are of the same size. While in the cut-based analysis a much lower significance for the excess of data over the background-only hypothesis is observed, both multivariate analyses yield a similar significance. Particularly the 2-BDT analysis does not only retain the sensitivity of the 1-BDT, it performs slightly better.
It is therefore shown, that the 2-BDT approach leads to a very good sensitivity not only in the VBF but also in the boosted category and in the combination of both categories. The 2-BDT approach is an excellent example of a further development of the 1-BDT approach and is certainly one of the best options for possible analysis strategies in Run 2. Also the other approaches discussed above might be able to replace the current 1-BDT approach provided that they are further optimised. Those methods are based on the same idea of combining advantages of the latest multivariate and cut-based analysis. Their particular goals, however, are different. While e.g. the Higgs mass dependent BDT training is used to increase the mass sensitivity, other approaches aim for a disentanglement of the background suppression and signal extraction or the optimization of the BDT classifier with respect to different background

sources. Approaches aiming for the latter targets remove the information on the invariant $\tau\tau$ mass from the BDT classifier and fit the mass distribution. Although the invariant $\tau\tau$ mass is not used in the BDT training it might be that the fitted mass distribution is still biased by the Higgs mass, which is assumed in the training of the BDT. Hence, if such an approach is applied in Run 2, the correlations between $m_{\tau\tau}^{\mathrm{MMC}}$ and the discriminating variables used in the BDT training have to be investigated. Depending on the potential bias of the $m_{\tau\tau}^{\mathrm{MMC}}$ distribution different questions like e.g. the mass sensitivity might be addressed in the future.

Table 9.9: Comparison of the observed (expected) local significance and the estimate of the signal strength $\hat{\mu}$ obtained from $H \to \tau_\ell \tau_h$ analyses based on the $\sqrt{s}$ =8 TeV dataset. The same dataset is analysed using different analysis strategies: The 1-BDT analysis presented in Chapter 7, the 2-BDT analysis in Section 9.2.5 and the briefly introduced cut-based analysis in Section 8.4.

	1-BDT	2-BDT	Cut-based
Signal strength	$1.1^{+0.6}_{-0.5}$	$1.3^{+0.6}_{-0.5}$	$0.7^{+0.7}_{-0.6}$
Significance	$2.5(2.3)\sigma$	$2.7(2.5)\sigma$	$1.0(1.7)\sigma$

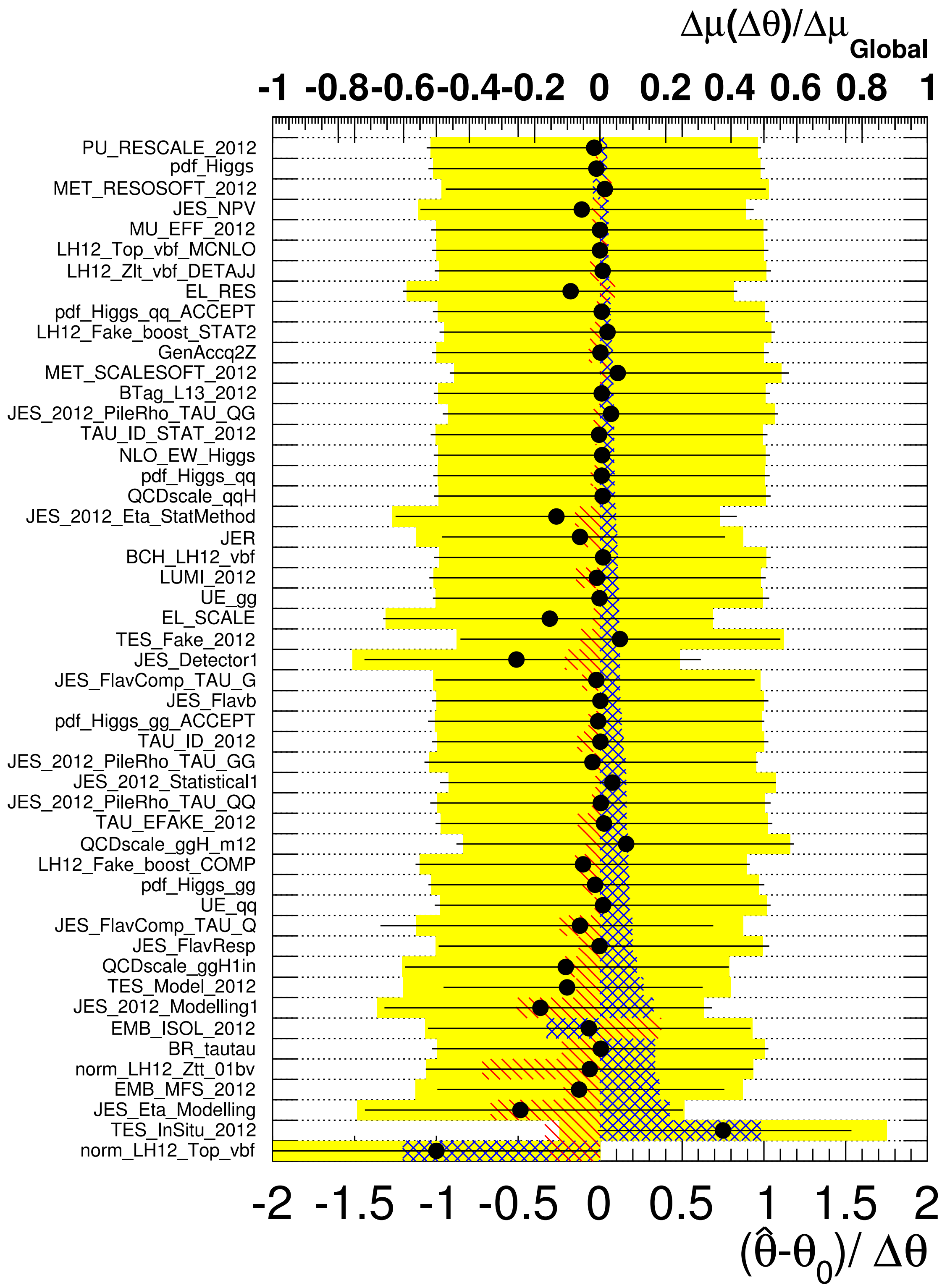


Figure 9.7: Ranking of the 50 nuisance parameters with the largest impact on the signal strength μ in the combined fit of the 8 TeV signal regions. The black points refer to the lower axis and show the deviations of the best-fit value of the nuisance parameters $\hat{\Theta}$ from their nominal pre-fit values Θ_0. While the black lines illustrate the relative post-fit uncertainties of the individual nuisance parameters, the nominal uncertainties are given by the yellow band. The hatched boxes refer to the upper axis visualising the impact $\Delta\mu(\Delta\Theta)$ of the nuisance parameter on the measured μ, when fixing the individual nuisance parameters to their best post-fit value. $\Delta\mu(\Delta\Theta)$ is divided by the total error on the best-fit value of the signal strength $\Delta\mu_{Global}$. While the red hatched boxes show the $+1\sigma$ post-fit impact, the blue hatched boxes display the -1σ pre-fit impact on $\hat{\mu}$. The labelling of the nuisance parameters is explained in Appendix E.

CHAPTER 10

Conclusions

In this thesis, a measurement of the decay of the SM Higgs boson into one hadronically and one leptonically decaying τ lepton ($H \to \tau_\ell \tau_h$) is performed based on proton-proton collision data collected with the ATLAS experiment at the LHC in 2011 and 2012. The collected data correspond to an integrated luminosity of 4.6 fb^{-1} and 20.3 fb^{-1} taken at a center-of-mass energy of $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV, respectively. The analysis of the $H \to \tau\tau$ decay in the $\tau_\ell \tau_h$ channel reveals an excess over background which corresponds to an overall observed (expected) significance for the $H \to \tau_\ell \tau_h$ observation of 2.5 (2.4) Gaussian standard deviations. The best-fit estimate of the signal strength parameter is given by

$$\hat{\mu} = 1.06^{+0.34}_{-0.34}\,(\text{stat.})\;^{+0.38}_{-0.29}\,(\text{syst.})\;\pm 0.1\,(\text{theo. syst.})$$

at m_H =125 GeV. The total uncertainty on $\hat{\mu}$ is composed of theoretical, systematic and statistical uncertainties. Systematic uncertainties mainly originate from the τ energy scale, the jet energy scale and the background estimate. The most important theoretical uncertainties arise from the prediction of the $H \to \tau_\ell \tau_h$ branching ratio.

In the presented $H \to \tau_\ell \tau_h$ analysis two signal categories, the boosted and the VBF category, are defined which are sensitive to different Higgs boson production mechanisms. The boosted category is experimentally motivated and requires a Higgs boson with high transverse momentum. In this region, Higgs signal events produced via ggF are mainly accumulated. The boosted category, however, also contains a significant amount of VBF produced Higgs boson events. The definition of the VBF category makes use of the characteristics of Higgs boson events produced via the VBF production mechanism.

In both categories, the background model is mainly data-driven. In particular, the highly relevant $Z \to \tau\tau$ background is modelled using a data-driven technique. The used technique, referred to as embedding technique, was already developed for earlier analyses. In the context of this thesis, however, a deeper understanding of lower level features of the embedding technique was gained. The size of intrinsic effects was estimated and corrections were derived. Furthermore, ways to validate this method were developed. The embedding-related studies performed in this thesis were published in [9] and were used in the presented $H \to \tau_\ell \tau_h$ analysis.

BDTs are used in order to discriminate between the signal and the background events. Based on a set of discriminating observables the BDTs are trained in the two signal categories of the 8 TeV analysis. They are then applied on both datasets in order to improve uniformity between the analyses of both datasets. Studies in this thesis investigated the performance of the BDT classifiers on the 7 TeV dataset.

A likelihood fit is simultaneously carried out in the signal and the corresponding top control regions

of the 7 TeV and 8 TeV datasets. Pruning and smoothing algorithms are used to reduce the impact of noise due to statistical fluctuations within the fit. From the simultaneous fit of the top control region, the normalisation of the top background is obtained. The fit provides an estimate of the signal strength parameter $\hat{\mu}$ mentioned above.
The presented $H \to \tau_\ell \tau_h$ analysis was part of the combined $H \to \tau\tau$ analysis of the three $\tau\tau$ decay modes in Reference [8]. The combination of $H \to \tau_\ell \tau_\ell$, $H \to \tau_h \tau_h$ and $H \to \tau_\ell \tau_h$ results in a signal strength parameter of

$$\hat{\mu} = 1.43^{+0.27}_{-0.26}\,(\text{stat.})\ ^{+0.32}_{-0.25}\,(\text{syst.})\ \pm 0.09\,(\text{theo. syst.})$$

at a mass hypothesis of m_H =125 GeV and is consistent with the SM expectation. The observed (expected) significance yields 4.5 (3.4) standard deviations. The combined result shows an evidence for the $H \to \tau\tau$ decay and hence experimentally confirms the Yukawa coupling between the Higgs boson and fermions. This result entered the preliminary combination of the ATLAS and CMS $H \to \tau\tau$ analyses [182]. In the combination an excess in data is obtained which corresponds to an observed (expected) significance of 5.5 (5.0) standard deviations. This is the first observation of the Higgs coupling to τ leptons and is one of the highlights of the LHC physics program whose emphasis lies on the experimental insight in the Higgs mechanism.
The high signal sensitivity is one of the main arguments to use multivariate methods. Since in the published analysis the BDT score distribution is fitted the background suppression and signal extraction are entangled with each other. In order to disentangle both steps several alternative approaches are discussed. Within this thesis a 2-BDT approach was tested for the $H \to \tau_\ell \tau_h$ analysis on the 8 TeV dataset: In each category one BDT is trained against the resonant $Z \to \tau\tau$ background while the other BDT is trained against all other non-resonant background sources. The $m_{\tau\tau}^{MMC}$ mass distribution is then directly fitted for events in the high BDT score region. This approach yields for the best-fit of the signal strength

$$\hat{\mu} = 1.31^{+0.32}_{-0.31}\,(\text{stat.})\ ^{+0.48}_{-0.38}\,(\text{syst.})\ \pm 0.1\,(\text{theo. syst.})$$

at a mass hypothesis of m_H =125 GeV. This is consistent with the published result. The observed (expected) significance of the excess corresponds to 2.7(2.5) Gaussian standard deviations. These results demonstrate a realisation of a multivariate analysis strategy in which the background suppression and signal extraction are disentangled. At the same time the signal sensitivity is retained because the BDT classifiers are optimised with respect to a particular background source. Thus, this approach is a significant improvement of the former analysis design so that future $H \to \tau\tau$ measurements will benefit from it in Run 2.

APPENDIX A

Production setup

For the embedding technique described in Chapter 6 the ESD format is used, because in the event re-reconstruction the missing transverse energy has to be calculated using calorimeter cells. This information, however, is not accessible in the higher level AOD format. Instead of the ESD format, samples in the DESD format are also suitable as input for the embedding procedure.

While all substeps of the embedding procedure have to run in the same ATHENA software release due to technical reasons, the condition tags and ATLAS geometries used in those steps differ. Their choice depends on the collision data or Monte Carlo samples which are processed in the particular embedding step.

Because of the increasing amount of data during Run 1 and the resulting expense of computing resources, the embedding software had to be re-organised in order to ensure an efficient production of embedded samples. The new structure of the embedding software enables to simultaneously produce embedded events for the three $\tau\tau$ final states, their corresponding systematic uncertainties and the various validation samples. Intermediate files like the $Z \to \mu\mu$ event after cell subtraction, which are needed for all embedded $\tau\tau$ final states are produced once only and reused where necessary. The re-use of those files reduces the required CPU time significantly. Figure A.1 illustrates the new workflow of the embedding software.

The embedding software was also integrated into the ATLAS JobTransforms framework [192]. In this framework multiple production tasks can be sequentially executed. As soon as the input and output file are defined the transforms identifies the relevant steps to get from the input to the output file and executes only those necessary steps. Thereby, the input and output files are not restricted to the input $Z \to \mu\mu$ events and τ-embedded events. In general, the transforms are capable to start and end at any arbitrary point of the embedding workflow. Some examples of how the production tasks can be configured are discussed in [191]. Those improvements, which are also documented in [191], were carried out in the context of this thesis and result in the integration of the embedding software into the central production system of ATLAS via the interface of the JobTransforms framework.

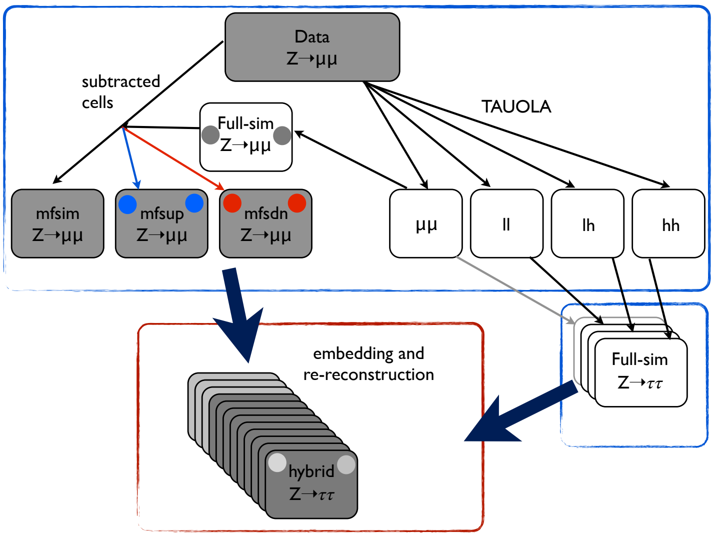


Figure A.1: The restructured embedding workflow is displayed: Starting from the same $Z \to \mu\mu$ input event the three $Z \to \tau\tau$ final states ($\tau_l\tau_l$, $\tau_l\tau_h$, $\tau_h\tau_h$) and one $Z \to \mu\mu$ decay are generated (right part of the upper blue box). The default and varied energy depositions in the simulated $Z \to \mu\mu$ decay (*mfsim*, *mfsup*, *mfsdn*) are removed (left part of the upper blue box). Afterwards, the mini-events for the three final $Z \to \tau\tau$ states are produced (lower blue box). In the last step, the simulated mini-events are merged with the nominal (varied) $Z \to \mu\mu$ data event and the resulting hybrid events are re-reconstructed (red box). The flowchart is taken from [191].

APPENDIX B

Monte Carlo samples

The following Table B.1 lists all Monte Carlo samples which are used in the analysis presented in Chapter 7.

Signal (m_H = 125 GeV)	MC generator	$\sigma \times \mathcal{B}$ [pb]	
ggF, $H \to \tau\tau$	Powheg + Pythia8	1.22	NNLO+NNLL
VBF, $H \to \tau\tau$	Powheg + Pythia8	0.100	(N)NLO
WH, $H \to \tau\tau$	Pythia8	0.0445	NNLO
ZH, $H \to \tau\tau$	Pythia8	0.0262	NNLO
Background	**MC generator**	**$\sigma \times \mathcal{B}$ [pb]**	
$W(\to \ell\nu)$, ($\ell = e, \mu, \tau$)	Alpgen+Pythia8	36800	NNLO
$Z/\gamma^*(\to \ell\ell)$, 60 GeV< $m_{\ell\ell}$ < 2 TeV	Alpgen+Pythia8	3910	NNLO
$Z/\gamma^*(\to \ell\ell)$, 10 GeV< $m_{\ell\ell}$ < 60 GeV	Alpgen+Herwig	13000	NNLO
VBF $Z/\gamma^*(\to \ell\ell)$	Sherpa	1.1	LO
$t\bar{t}$	Powheg + Pythia8	253†	NNLO+NNLL
Single top : Wt	Powheg + Pythia8	22†	NNLO
Single top : s-channel	Powheg + Pythia8	5.6†	NNLO
Single top : t-channel	AcerMC+Pythia6	87.8†	NNLO
$q\bar{q} \to WW$	Alpgen+Herwig	54†	NLO
$gg \to WW$	gg2WW+Herwig	1.4†	NLO
WZ, ZZ	Herwig	30†	NLO
$H \to WW$	same as for $H \to \tau\tau$ signal	4.7†	

Table B.1: Monte Carlo generators used to model the different signal and background processes considered in the $H \to \tau_\ell\tau_h$ analysis. In the last column, the cross-sections times branching ratios referring to processes at $\sqrt{s}$ =8 TeV are listed. Inclusive cross-sections are marked with†. In addition, for each process the perturbative order of the QCD calculation is given.

APPENDIX C

Validation plots for the multivariate analysis of the $\sqrt{s}$ =7 TeV + $\sqrt{s}$ =8 TeV datasets

C.1 BDT input variables of the $\sqrt{s}$ =8 TeV analysis

The following section illustrates all distributions of input variables used to train the BDT in the boosted and VBF signal categories of the $\sqrt{s}$ =8 TeV analysis. In addition, the same distributions are shown for the four control regions defined in Section 7.2.5.

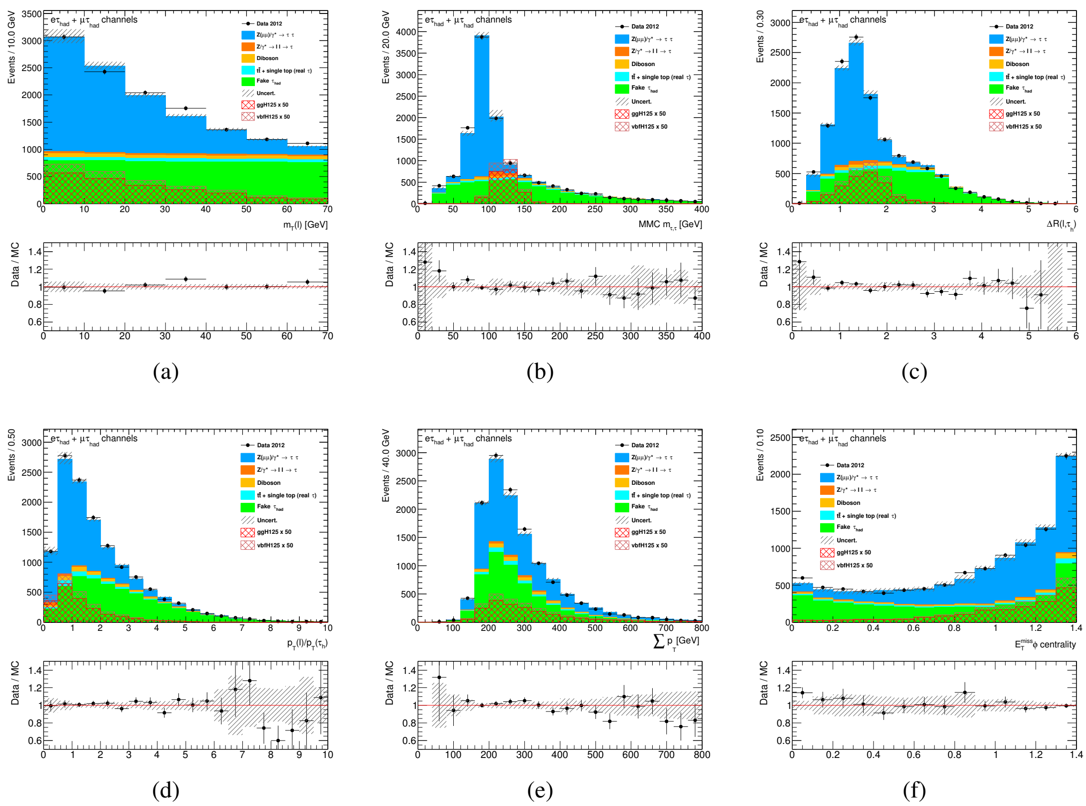


Figure C.1: Comparison of data with the combined background model with respect to BDT input variables in the boosted signal region of the $\sqrt{s}$ =8 TeV analysis: (a) transverse mass, (b) $\tau\tau$ invariant mass calculated with the MMC, (c) ΔR between lepton and τ_{had} candidate, (d) ratio between transverse masses of the τ-decay products, (e) $\sum p_T$, (f) $E_T^{\text{miss}}\phi$ centrality. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

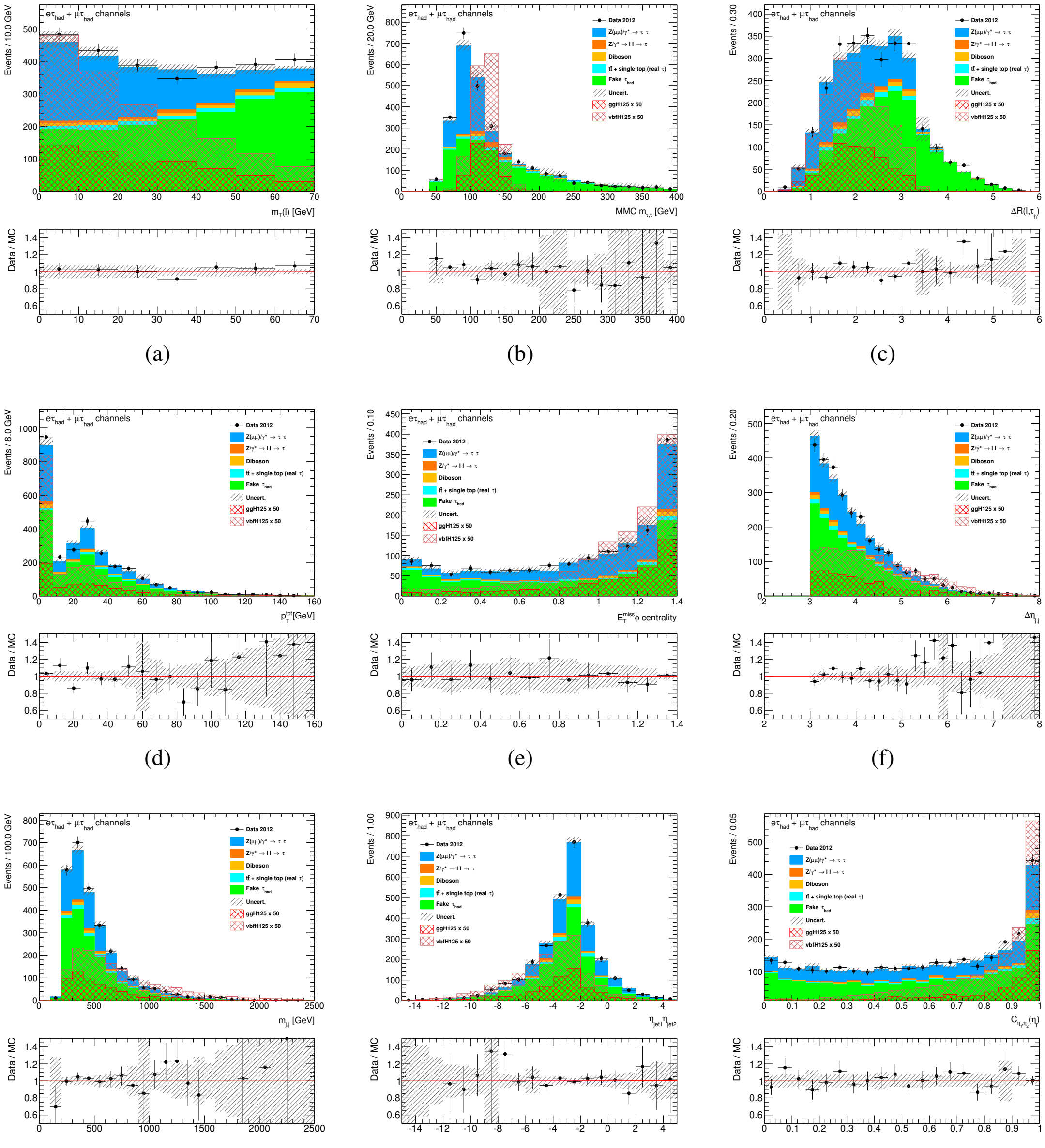
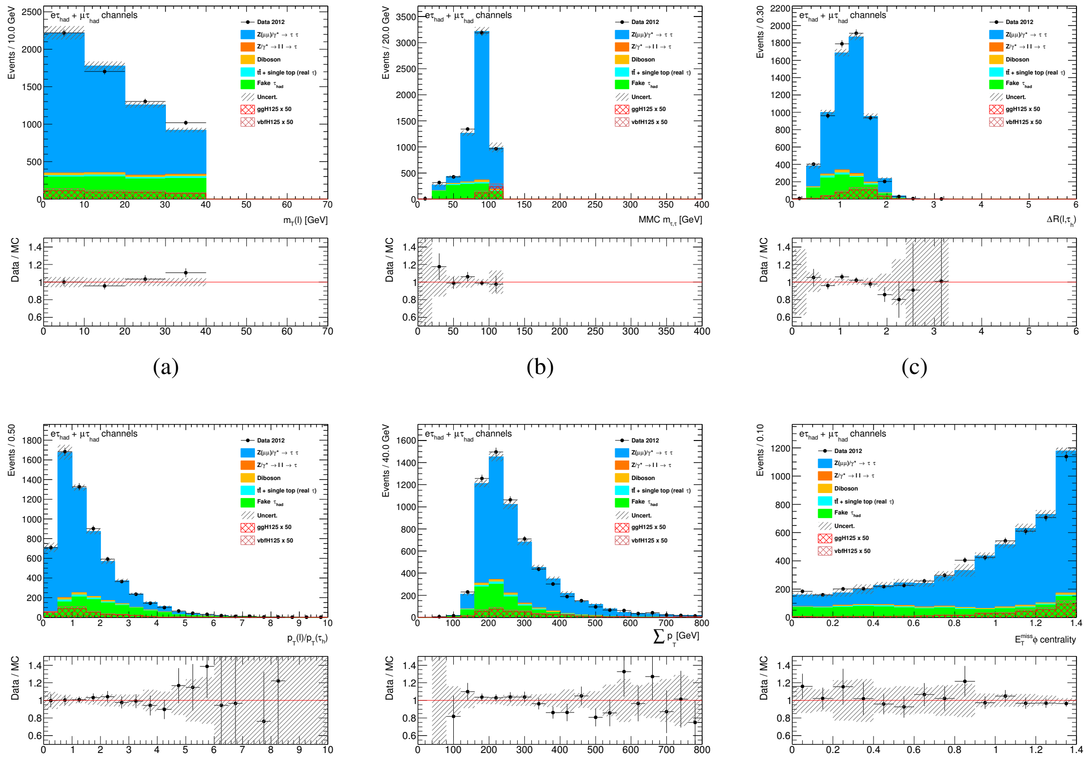


Figure C.2: Comparison of data with the combined background model with respect to BDT input variables in the VBF signal region of the $\sqrt{s}$ =8 TeV analysis: (a) transverse mass, (b) $\tau\tau$ invariant mass calculated with the MMC, (c) ΔR between lepton and $\tau_{\rm had}$ candidate, (d) p_T^{tot}, (e) $E_T^{\rm miss}\phi$ centrality, (f) difference in pseudorapidity between two leading jets, (g) dijet mass, (h) product of pseudorapidities, (i) $\ell-\eta$ centrality $C_{\eta_1,\eta_2}(\eta_\ell)$. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.



(a) (b) (c)

(d) (e) (f)

Figure C.3: Comparison of data with the combined background model with respect to BDT input variables in the boosted Z-enriched region of the $\sqrt{s}$ =8 TeV analysis: (a) transverse mass, (b) $\tau\tau$ invariant mass calculated with the MMC, (c) ΔR between lepton and τ_{had} candidate, (d) ratio between transverse masses of the τ-decay products, (e) $\sum p_T$, (f) $E_T^{\text{miss}}\phi$ centrality. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

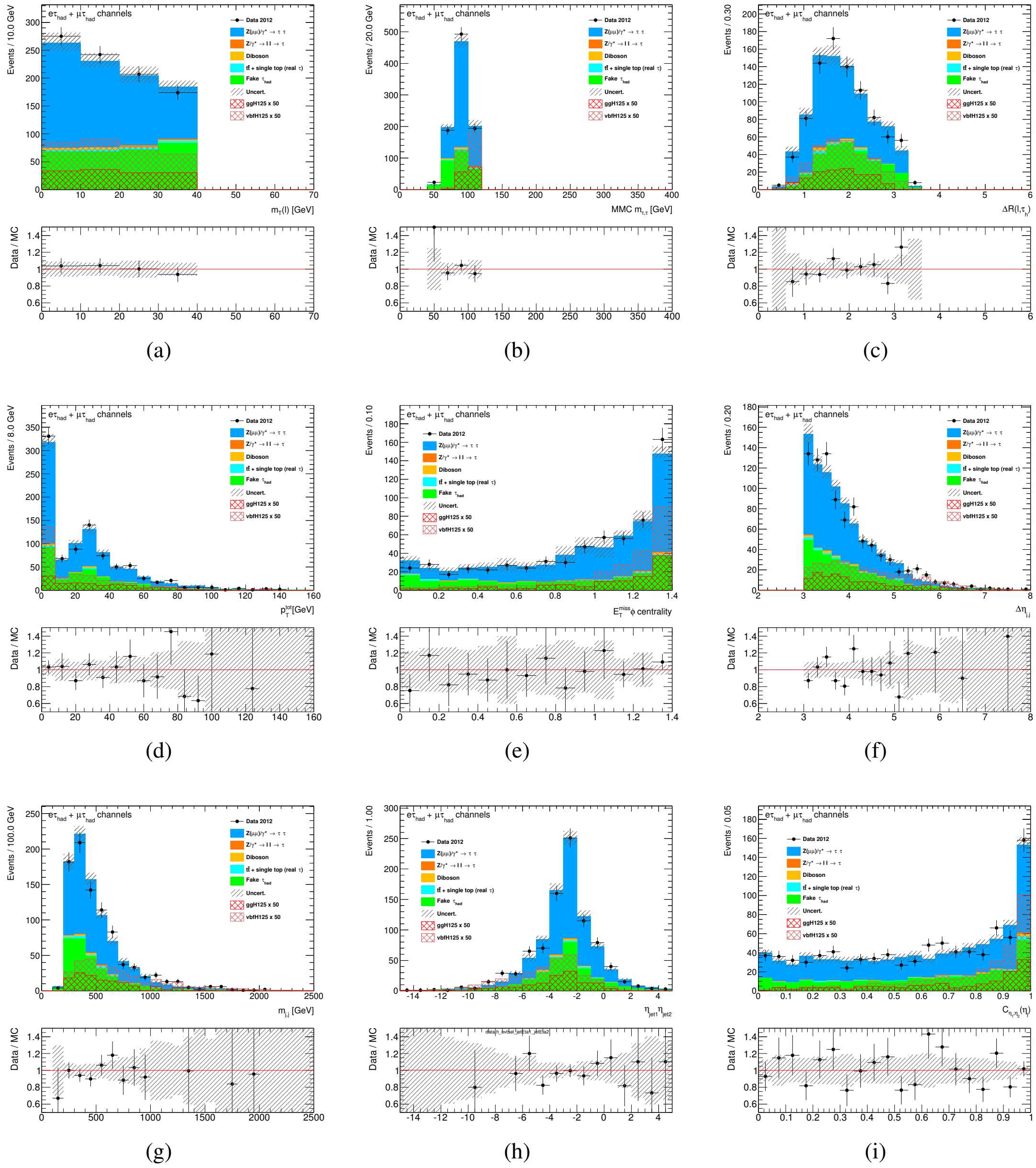


Figure C.4: Comparison of data with the combined background model with respect to BDT input variables in the VBF Z-enriched control region of the $\sqrt{s}$ =8 TeV analysis: (a) transverse mass, (b) $\tau\tau$ invariant mass calculated with the MMC, (c) ΔR between lepton and τ_{had} candidate, (d) p_T^{tot}, (e) $E_T^{\text{miss}}\phi$ centrality, (f) difference in pseudorapidity between two leading jets, (g) dijet mass, (h) product of pseudorapidities, (i) $\ell - \eta$ centrality $C_{\eta_1,\eta_2}(\eta_\ell)$. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

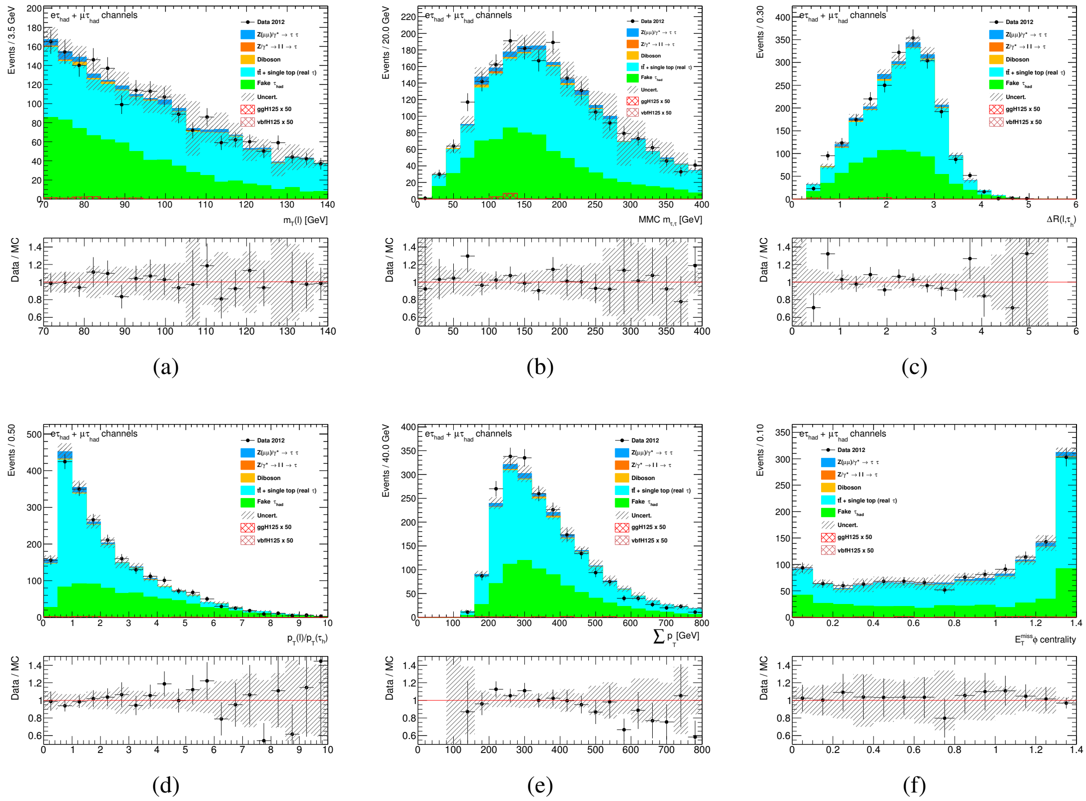


Figure C.5: Comparison of data with the combined background model with respect to BDT input variables in the boosted top-enriched region of the $\sqrt{s}$ =8 TeV analysis: (a) transverse mass, (b) $\tau\tau$ invariant mass calculated with the MMC, (c) ΔR between lepton and τ_{had} candidate, (d) ratio between transverse masses of the τ-decay products, (e) $\sum p_T$, (f) $E_T^{\text{miss}}\phi$ centrality. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

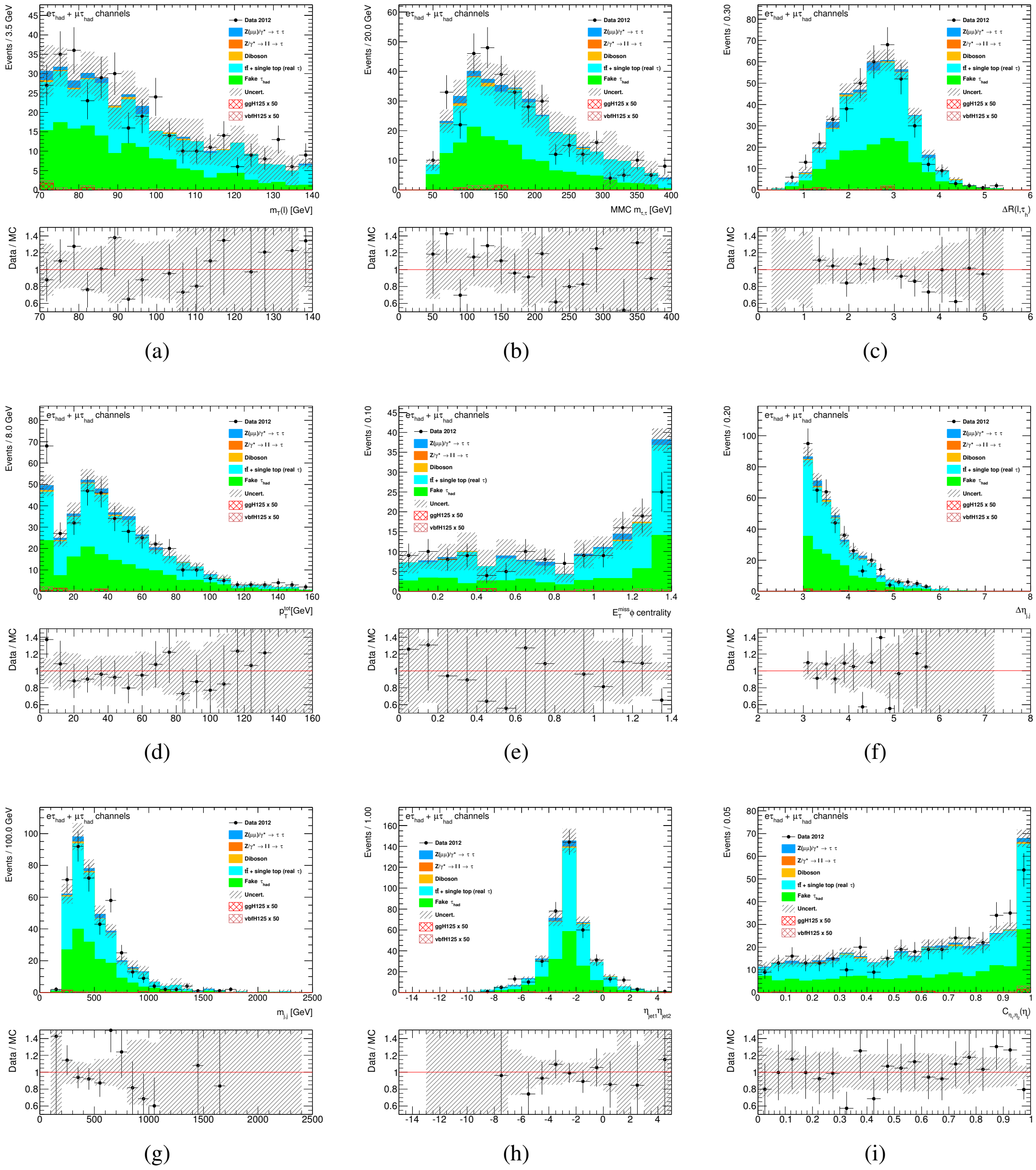


Figure C.6: Comparison of data with the combined background model with respect to BDT input variables in the VBF top-enriched control region of the $\sqrt{s}$ =8 TeV analysis: (a) transverse mass, (b) $\tau\tau$ invariant mass calculated with the MMC, (c) ΔR between lepton and τ_{had} candidate, (d) p_T^{tot}, (e) $E_T^{\text{miss}}\phi$ centrality, (f) difference in pseudorapidity between two leading jets, (g) dijet mass, (h) product of pseudorapidities, (i) $\ell - \eta$ centrality $C_{\eta_1,\eta_2}(\eta_\ell)$. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

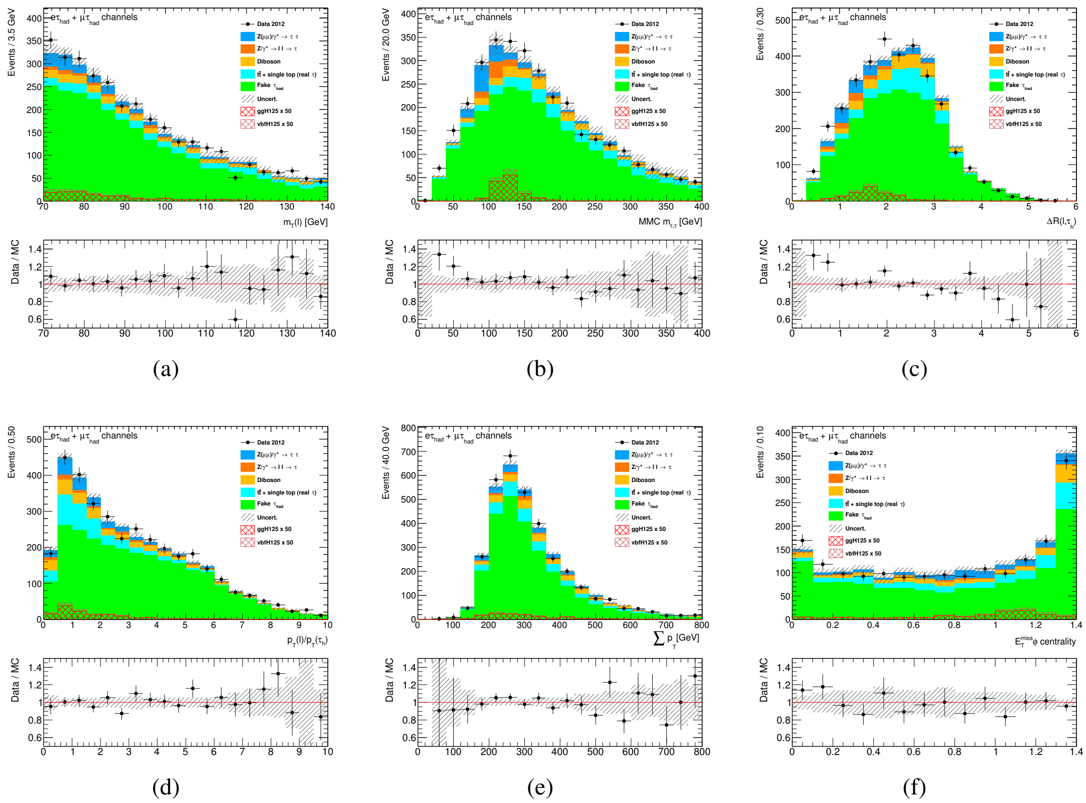


Figure C.7: Comparison of data with the combined background model with respect to BDT input variables in the boosted W-enriched region of the $\sqrt{s} = 8$ TeV analysis: (a) transverse mass, (b) $\tau\tau$ invariant mass calculated with the MMC, (c) ΔR between lepton and τ_{had} candidate, (d) ratio between transverse masses of the τ-decay products, (e) $\sum p_T$, (f) $E_T^{\mathrm{miss}}\phi$ centrality. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

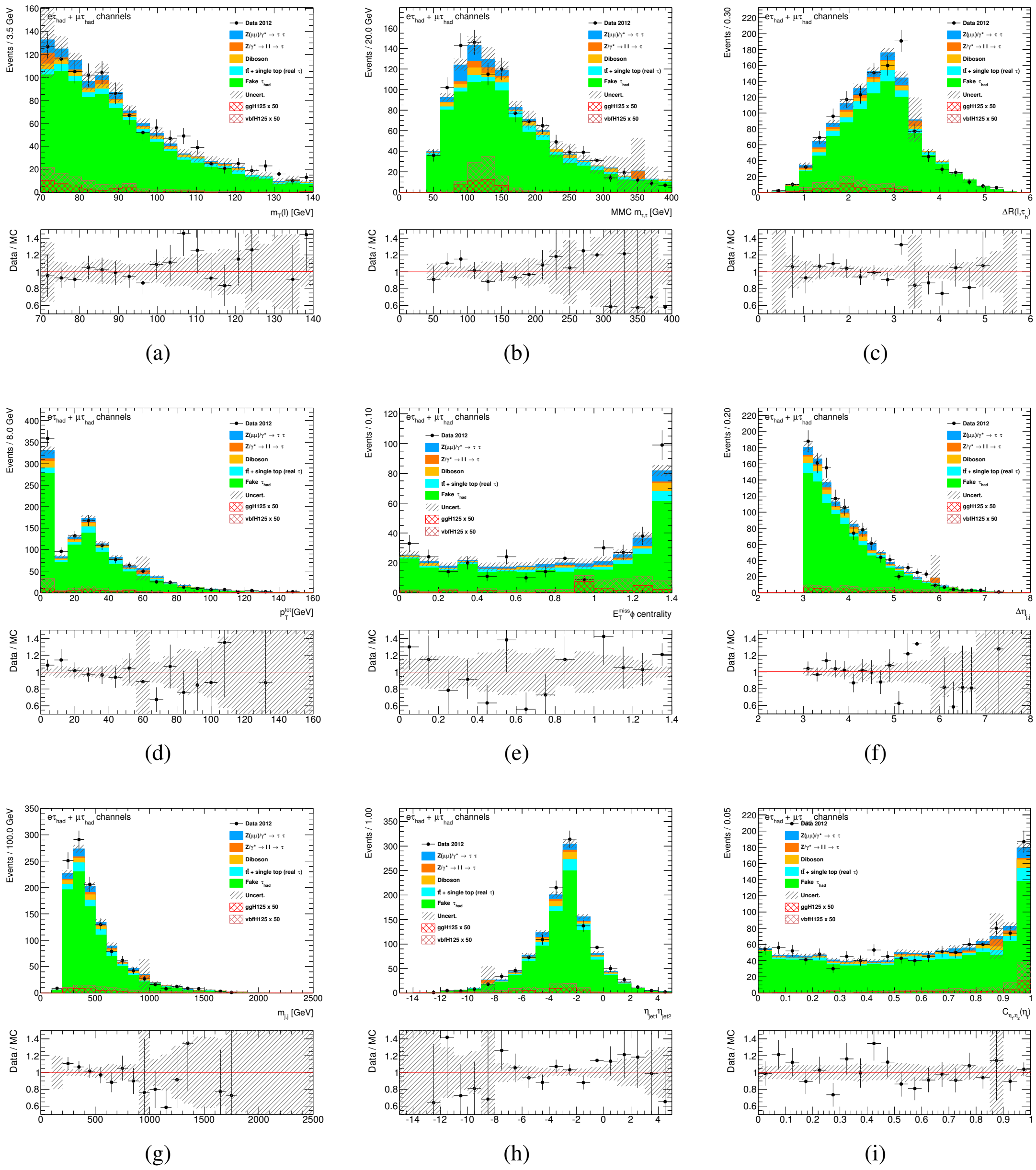
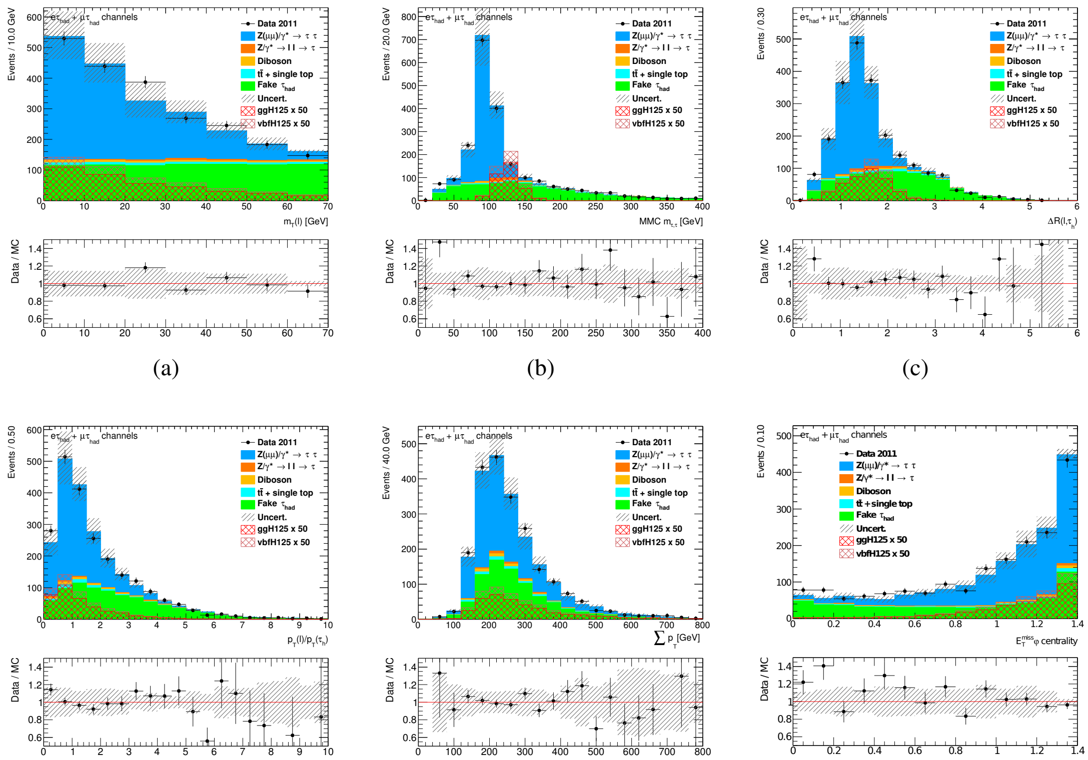


Figure C.8: Comparison of data with the combined background model with respect to BDT input variables in the VBF W-enriched control region of the $\sqrt{s}$ =8 TeV analysis: (a) transverse mass, (b) $\tau\tau$ invariant mass calculated with the MMC, (c) ΔR between lepton and τ_{had} candidate, (d) p_T^{tot}, (e) $E_T^{\text{miss}}\phi$ centrality, (f) difference in pseudorapidity between two leading jets, (g) dijet mass, (h) product of pseudorapidities, (i) $\ell - \eta$ centrality $C_{\eta_1,\eta_2}(\eta_\ell)$. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

C.2 BDT input variables of the $\sqrt{s}$ =7 TeV analysis

The same distributions shown before for the $\sqrt{s}$ =8 TeV analysis are illustrated below for the $\sqrt{s}$ =7 TeV analysis. The BDTs trained on $\sqrt{s}$ =8 TeV are evaluated for the $\sqrt{s}$ =7 TeV analysis based on those observables.



(a) (b) (c)

(d) (e) (f)

Figure C.9: Comparison of data with the combined background model with respect to BDT input variables in the boosted signal region of the $\sqrt{s}$ =7 TeV analysis: (a) transverse mass, (b) $\tau\tau$ invariant mass calculated with the MMC, (c) ΔR between lepton and τ_{had} candidate, (d) ratio between transverse masses of the τ-decay products, (e) $\sum p_T$, (f) $E_T^{\text{miss}}\phi$ centrality. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

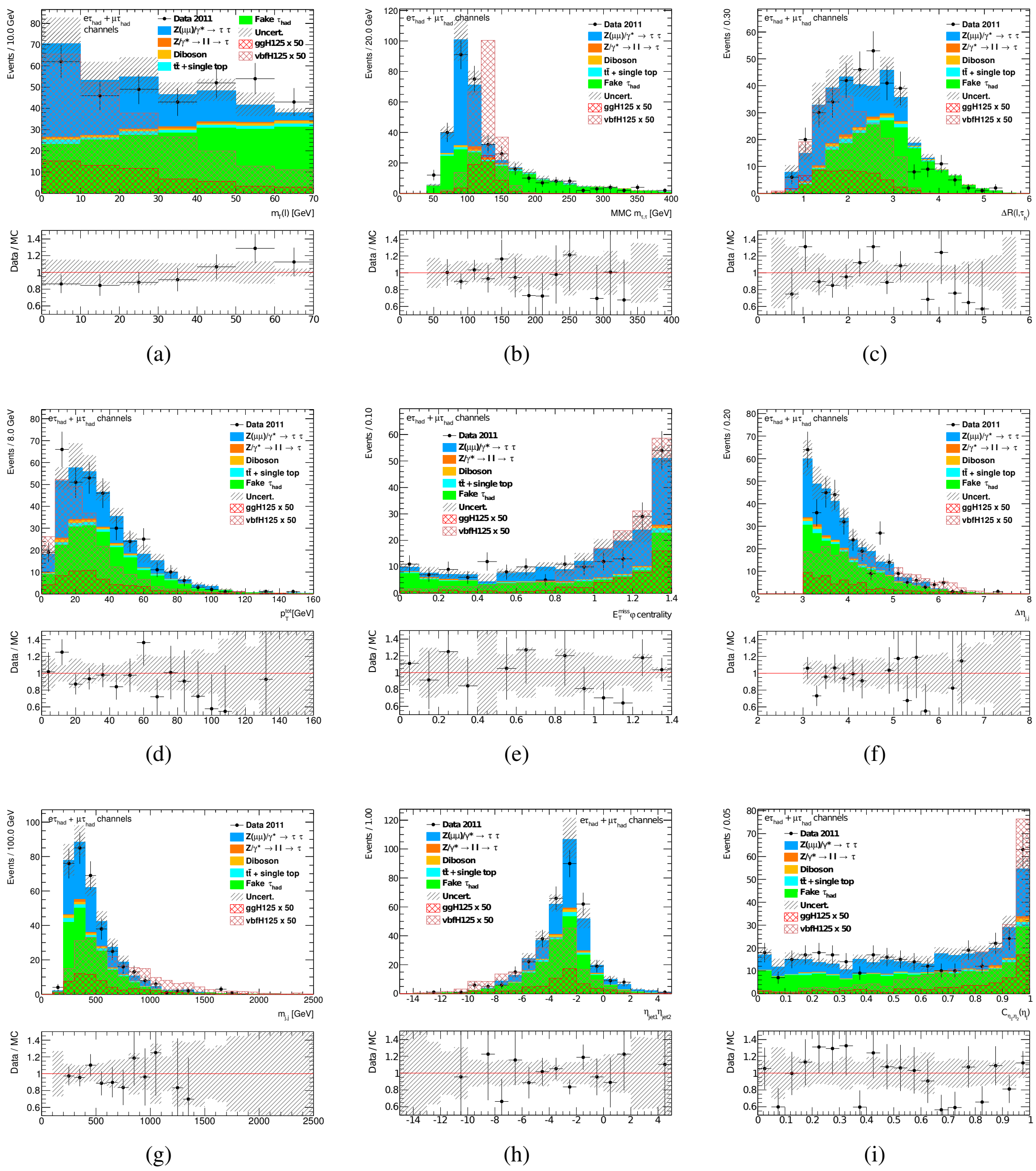


Figure C.10: Comparison of data with the combined background model with respect to BDT input variables in the VBF signal region of the $\sqrt{s}$ =7 TeV analysis: (a) transverse mass, (b) $\tau\tau$ invariant mass calculated with the MMC, (c) ΔR between lepton and τ_{had} candidate, (d) p_T^{tot}, (e) $E_T^{\text{miss}}\phi$ centrality, (f) difference in pseudorapidity between two leading jets, (g) dijet mass, (h) product of pseudorapidities, (i) $\ell - \eta$ centrality $C_{\eta_1,\eta_2}(\eta_\ell)$. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

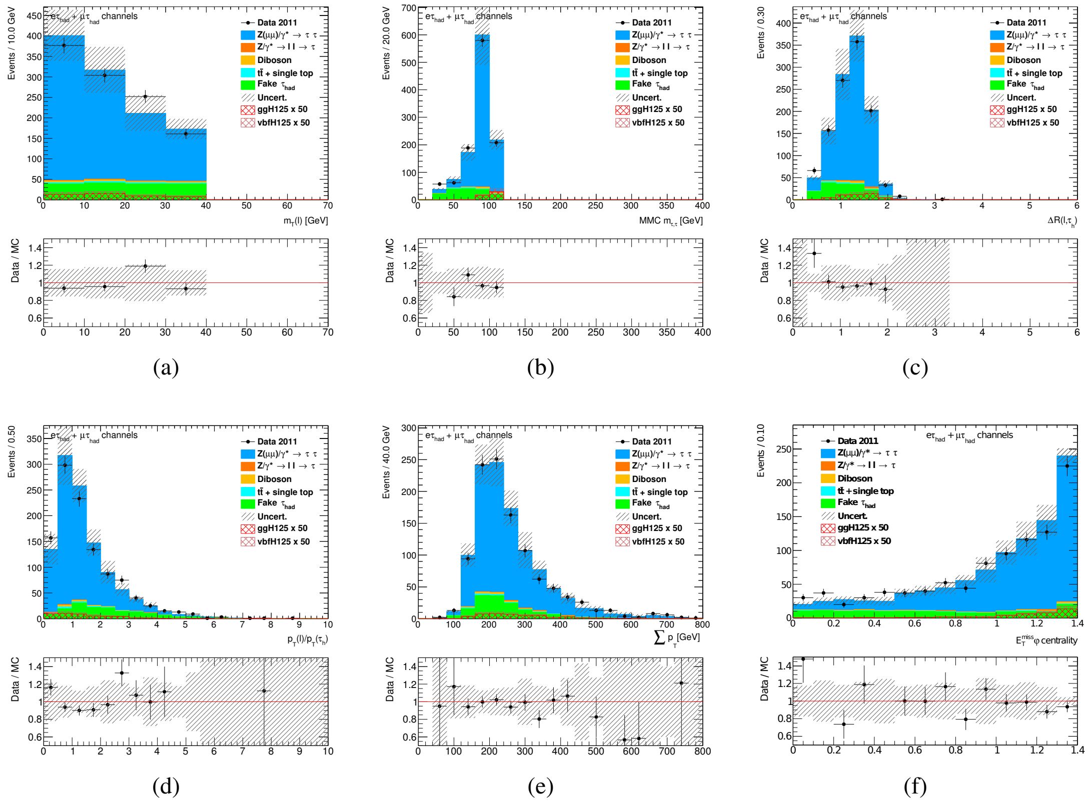


Figure C.11: Comparison of data with the combined background model with respect to BDT input variables in the boosted Z-enriched region of the $\sqrt{s}$ =7 TeV analysis: (a) transverse mass, (b) $\tau\tau$ invariant mass calculated with the MMC, (c) ΔR between lepton and $\tau_{\rm had}$ candidate, (d) ratio between transverse masses of the τ-decay products, (e) $\sum p_T$, (f) $E_T^{\rm miss}\phi$ centrality. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

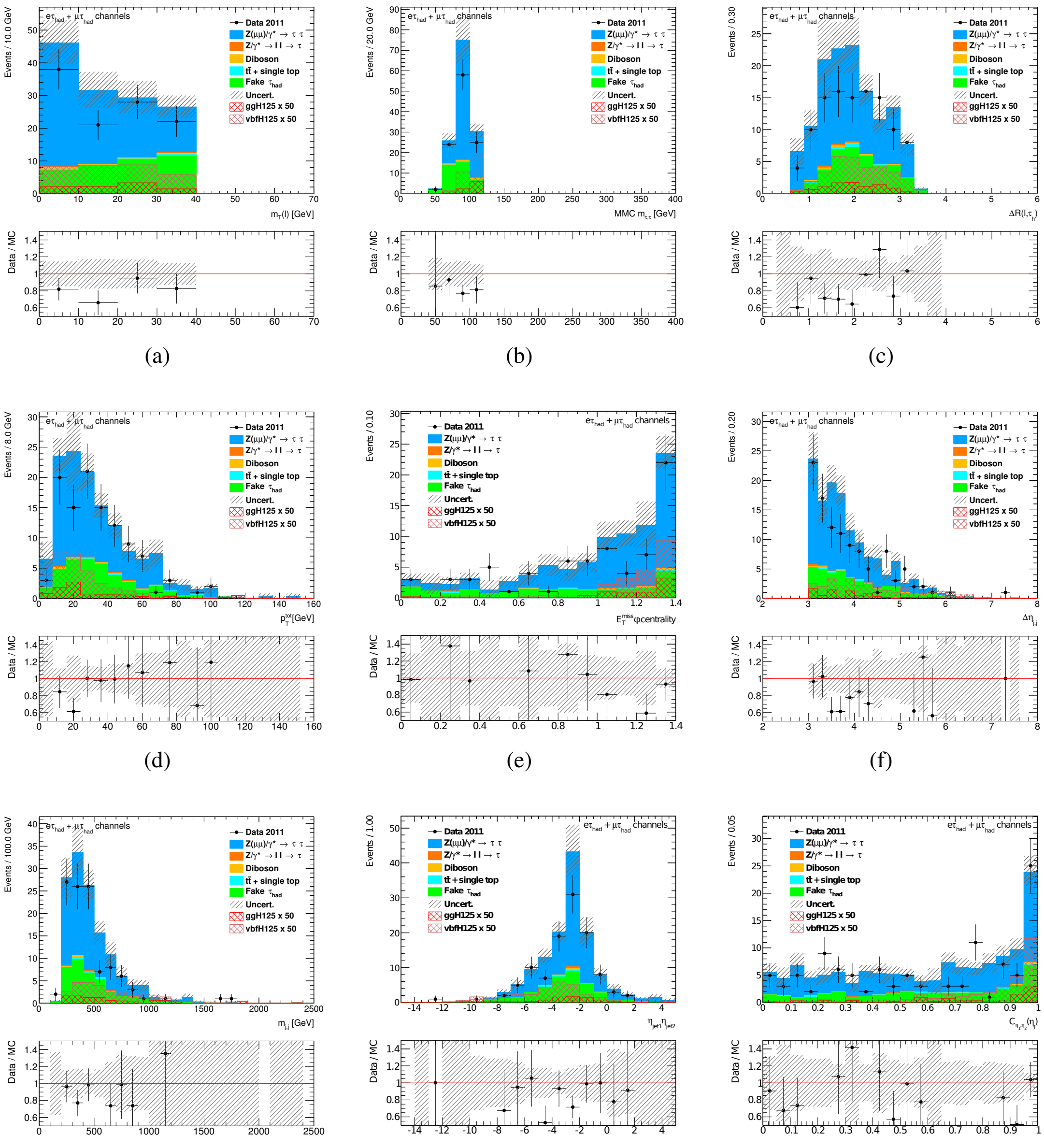


Figure C.12: Comparison of data with the combined background model with respect to BDT input variables in the VBF Z-enriched control region of the $\sqrt{s}$ = 7 TeV analysis: (a) transverse mass, (b) $\tau\tau$ invariant mass calculated with the MMC, (c) ΔR between lepton and τ_{had} candidate, (d) p_T^{tot}, (e) $E_T^{\mathrm{miss}}\phi$ centrality, (f) difference in pseudorapidity between two leading jets, (g) dijet mass, (h) product of pseudorapidities, (i) $\ell - \eta$ centrality $C_{\eta_1,\eta_2}(\eta_\ell)$. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

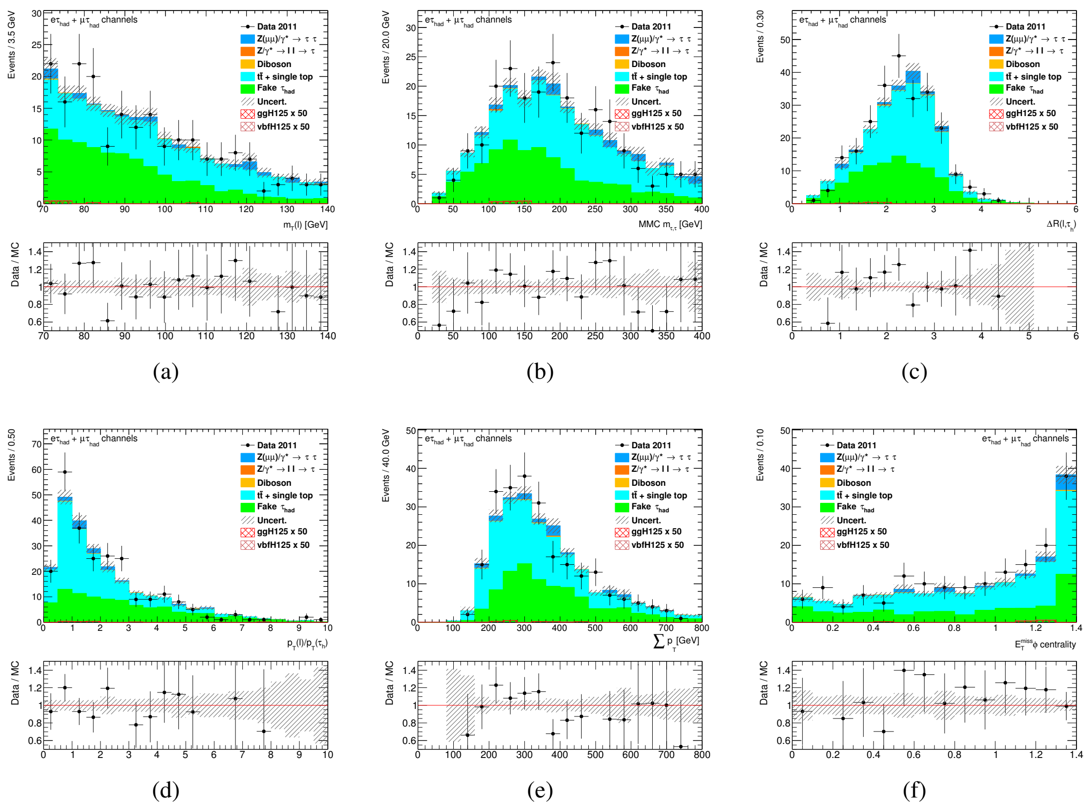


Figure C.13: Comparison of data with the combined background model with respect to BDT input variables in the boosted top-enriched region of the $\sqrt{s}$ =7 TeV analysis: (a) transverse mass, (b) $\tau\tau$ invariant mass calculated with the MMC, (c) ΔR between lepton and τ_{had} candidate, (d) ratio between transverse masses of the τ-decay products, (e) $\sum p_T$, (f) $E_T^{\text{miss}}\phi$ centrality. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

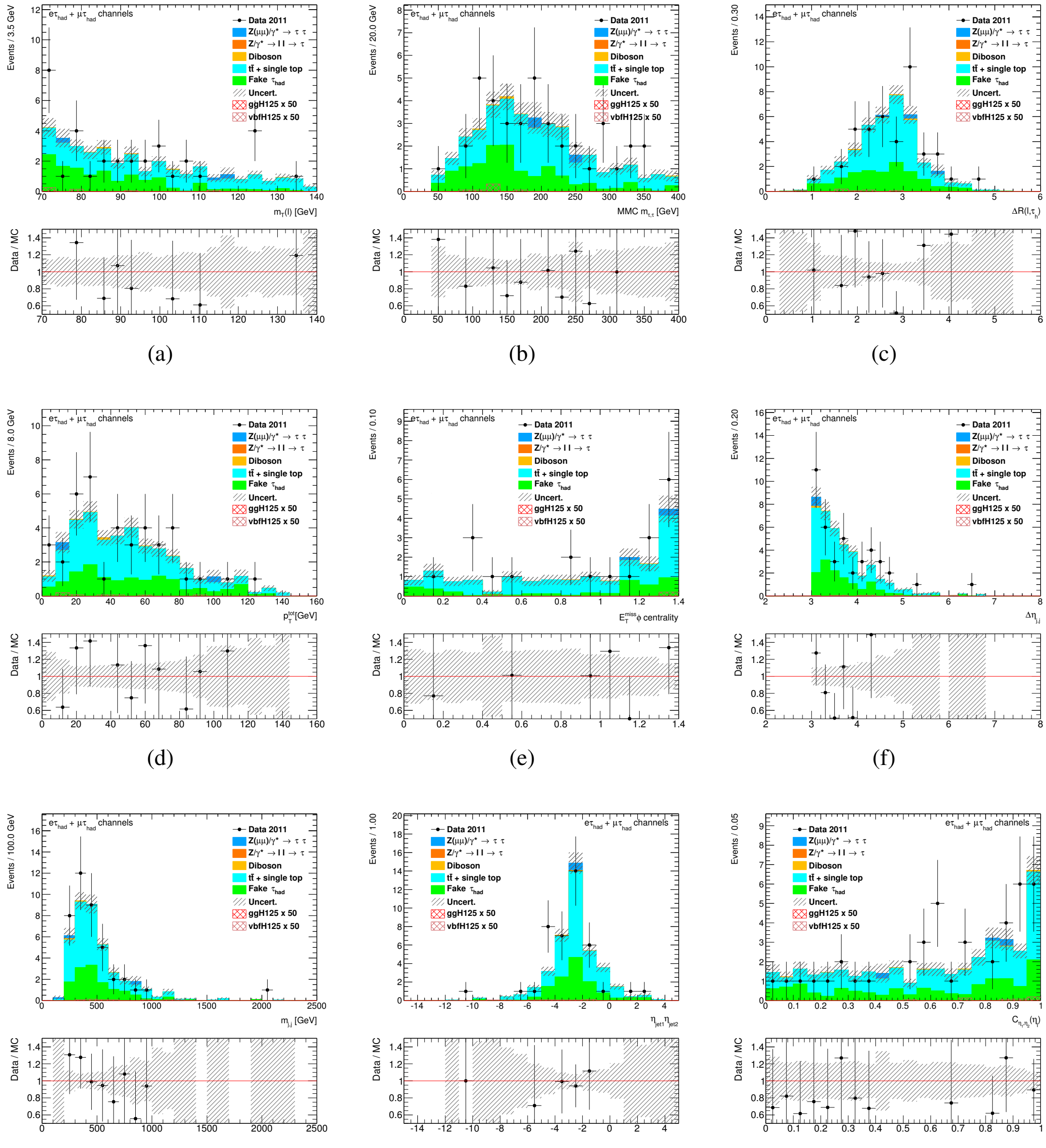


Figure C.14: Comparison of data with the combined background model with respect to BDT input variables in the VBF top-enriched control region of the $\sqrt{s}$ =7 TeV analysis: (a) transverse mass, (b) $\tau\tau$ invariant mass calculated with the MMC, (c) ΔR between lepton and $\tau_{\rm had}$ candidate, (d) p_T^{tot}, (e) $E_T^{\rm miss}\phi$ centrality, (f) difference in pseudorapidity between two leading jets, (g) dijet mass, (h) product of pseudorapidities, (i) $\ell - \eta$ centrality $C_{\eta_1,\eta_2}(\eta_\ell)$. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

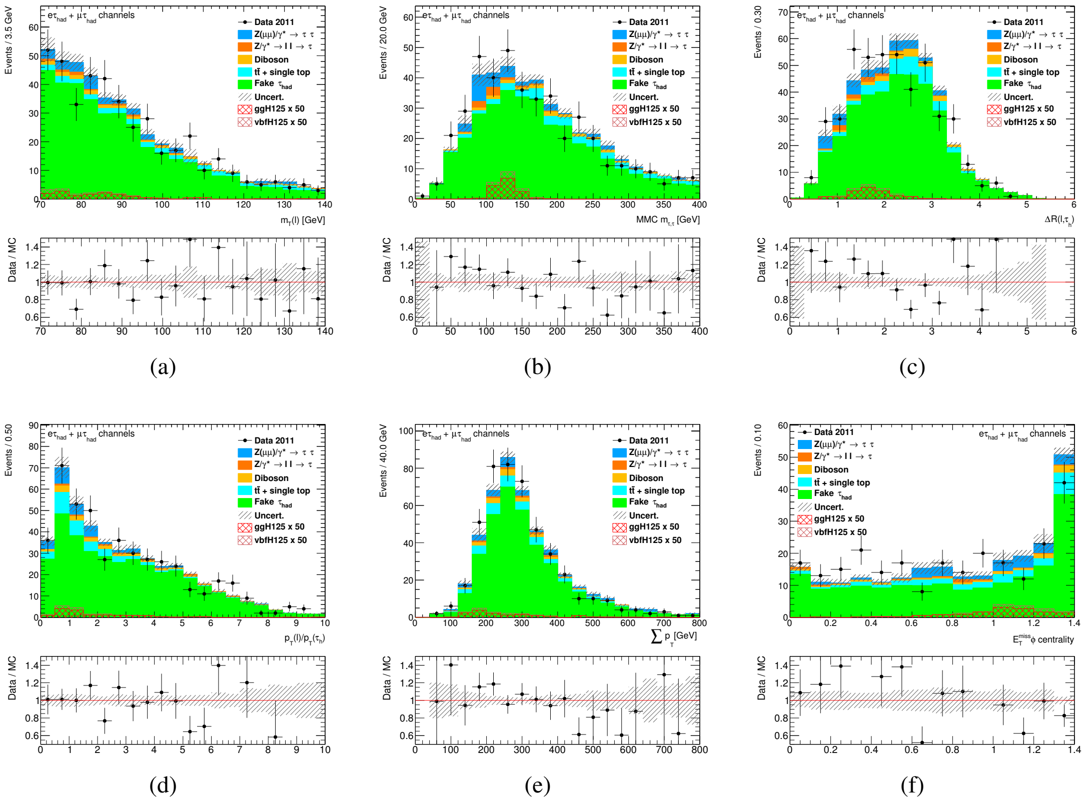


Figure C.15: Comparison of data with the combined background model with respect to BDT input variables in the boosted W-enriched region of the $\sqrt{s}$ =7 TeV analysis: (a) transverse mass, (b) $\tau\tau$ invariant mass calculated with the MMC, (c) ΔR between lepton and τ_{had} candidate, (d) ratio between transverse masses of the τ-decay products, (e) $\sum p_T$, (f) $E_T^{\text{miss}}\phi$ centrality. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

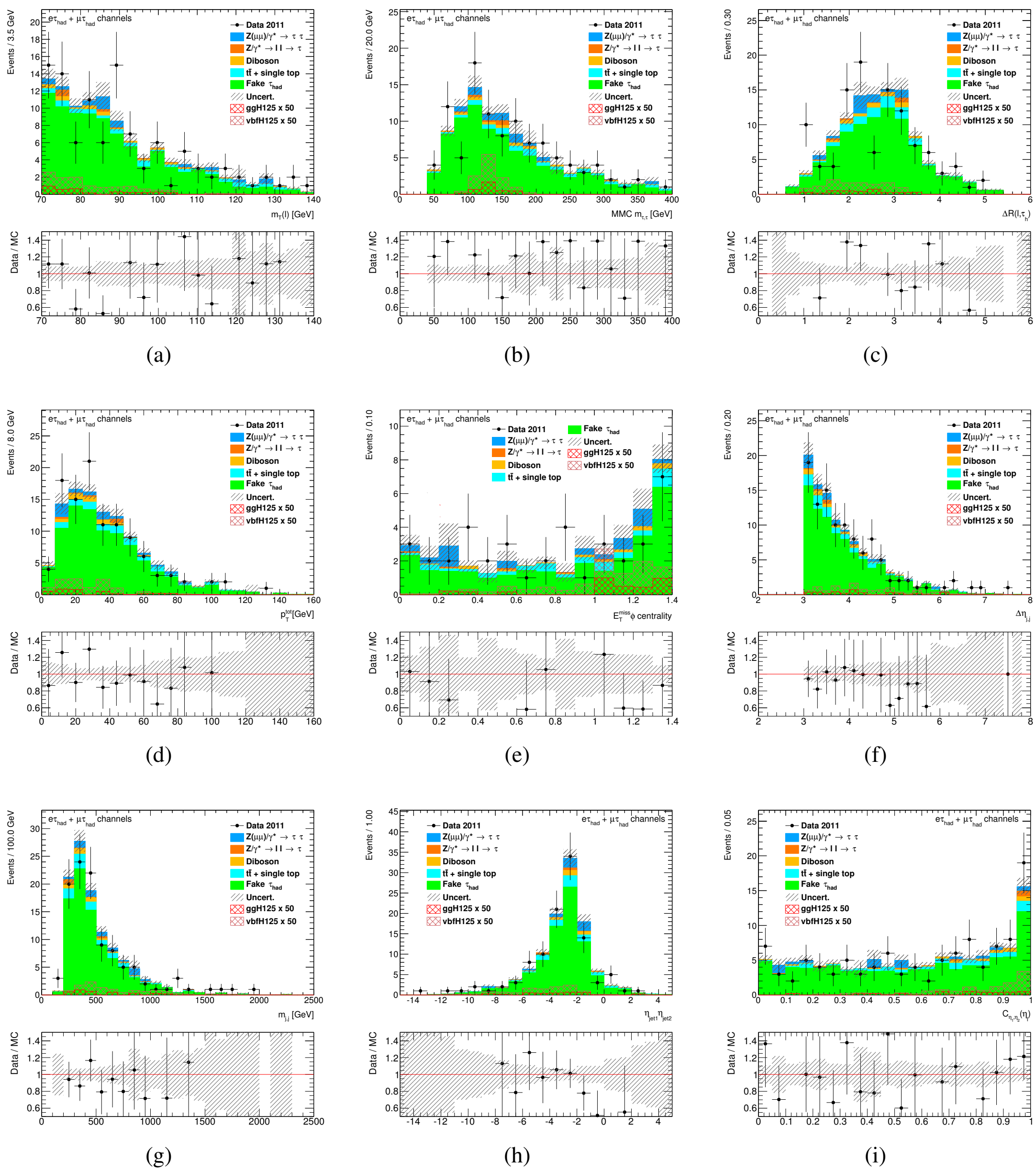


Figure C.16: Comparison of data with the combined background model with respect to BDT input variables in the VBF W-enriched control region of the $\sqrt{s}$ =7 TeV analysis: (a) transverse mass, (b) $\tau\tau$ invariant mass calculated with the MMC, (c) ΔR between lepton and τ_{had} candidate, (d) p_T^{tot}, (e) $E_T^{\text{miss}}\phi$ centrality, (f) difference in pseudorapidity between two leading jets, (g) dijet mass, (h) product of pseudorapidities, (i) $\ell - \eta$ centrality $C_{\eta_1,\eta_2}(\eta_\ell)$. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

C.2.1 Validation of the BDT score distributions

The BDT score distributions in the boosted and VBF control regions of the $\sqrt{s}$ =7 TeV and $\sqrt{s}$ =8 TeV analysis are depicted below.

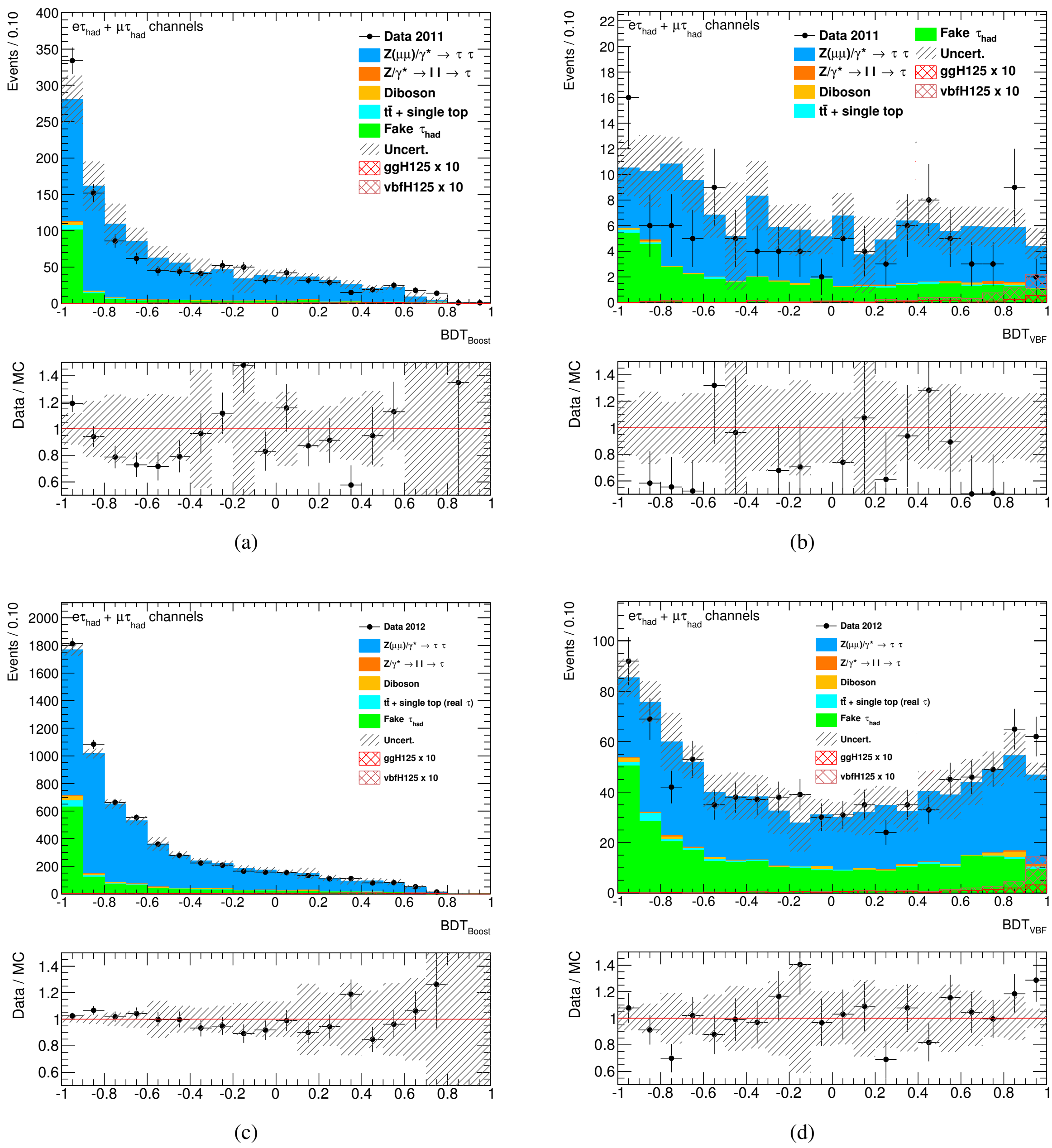
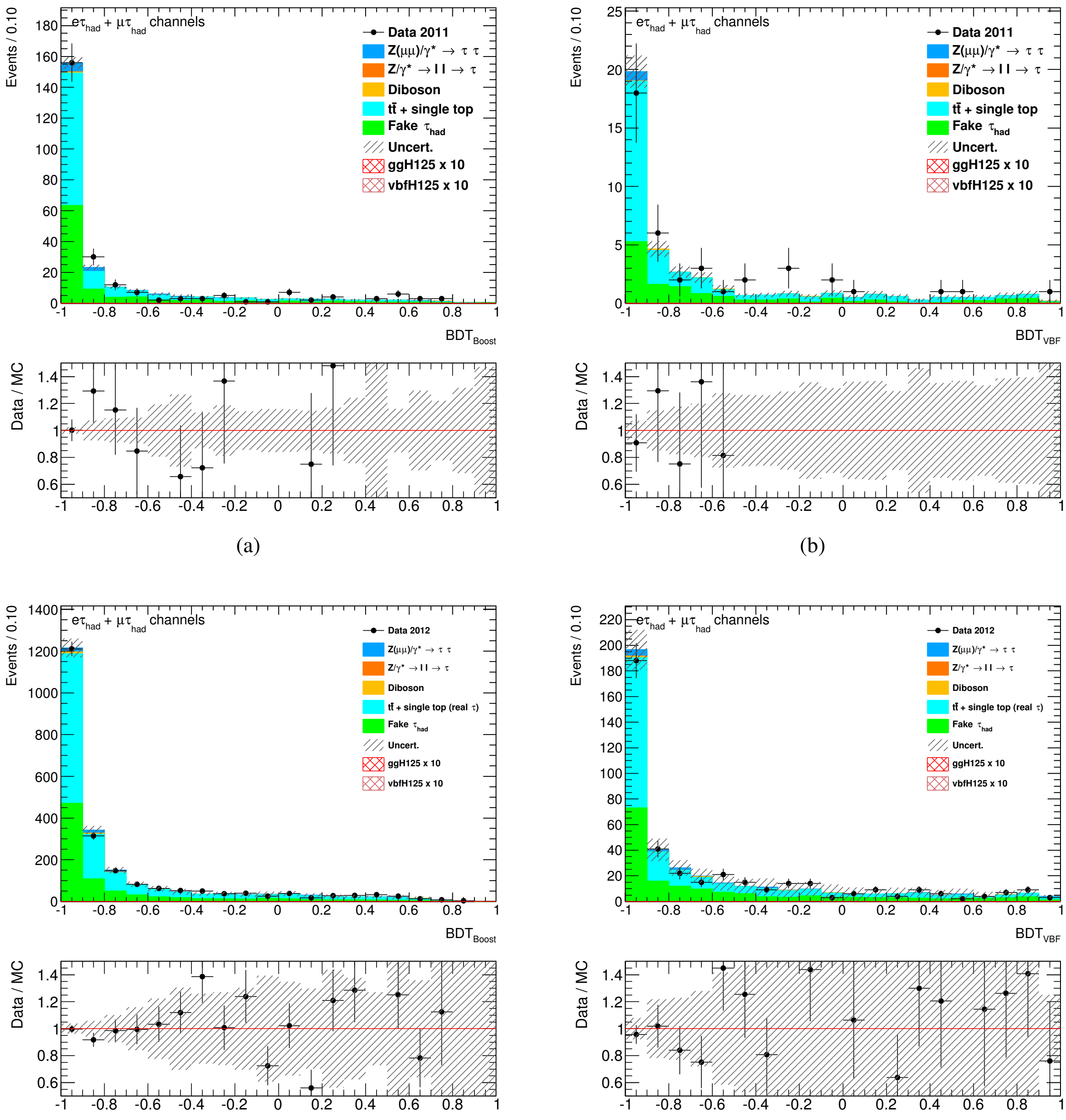
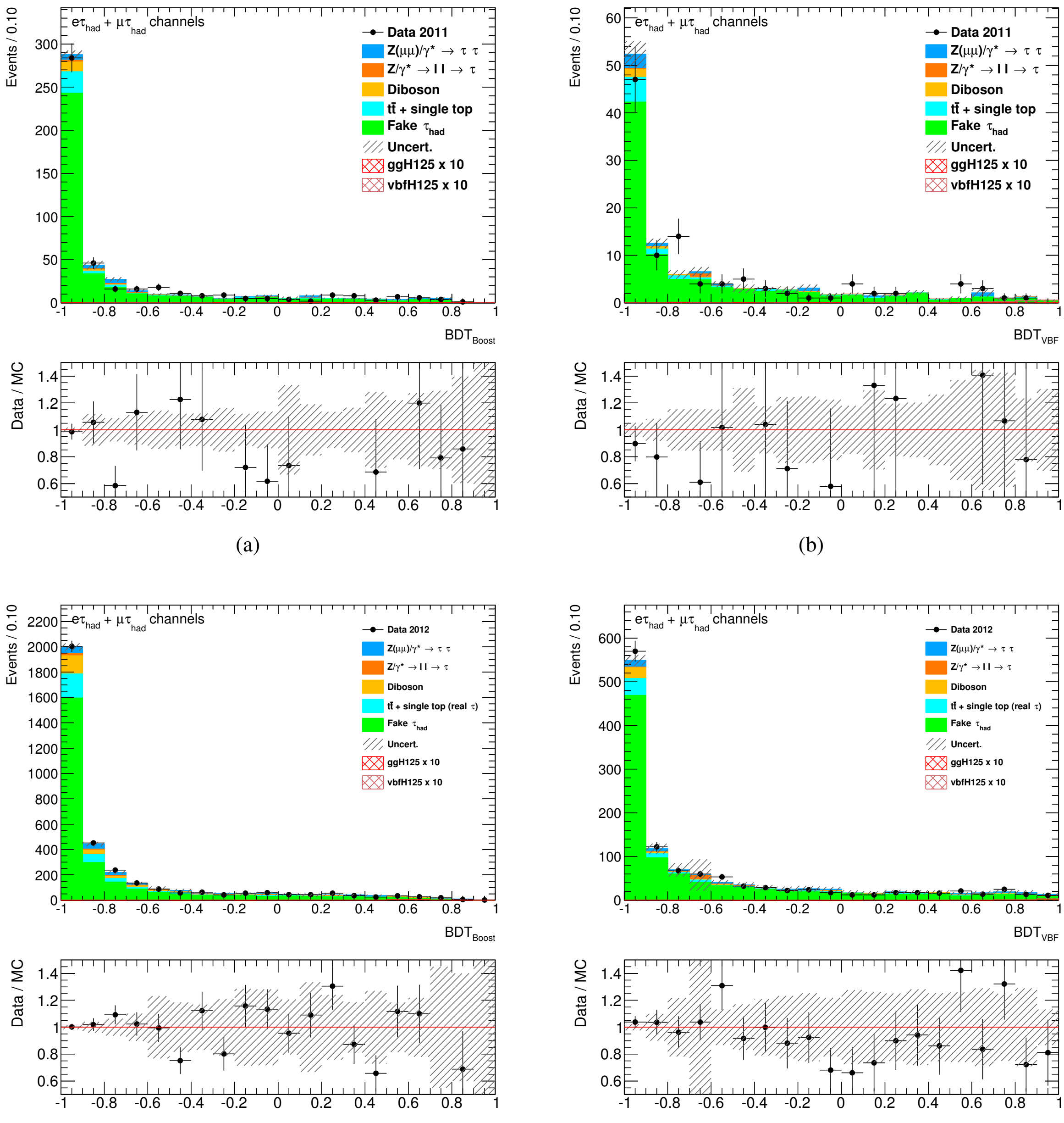


Figure C.17: Unblinded BDT score distribution (pre-fit) for the boosted (left) and VBF (right) Z-enriched control region of the $\sqrt{s}$ =7 TeV (top) and $\sqrt{s}$ =8 TeV (bottom) analysis. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.



(a) (b) (c) (d)

Figure C.18: Unblinded BDT score distribution (pre-fit) for the boosted (left) and VBF (right) top-enriched control region of the $\sqrt{s}$ =7 TeV (top) and $\sqrt{s}$ =8 TeV (bottom) analysis. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.



(a) (b) (c) (d)

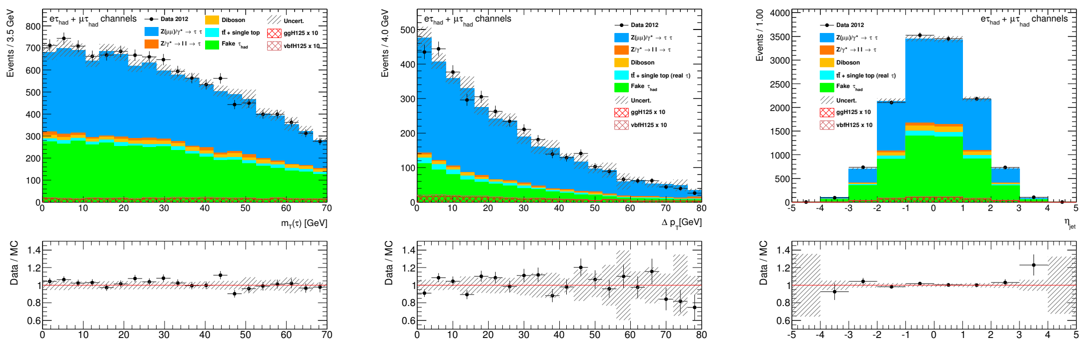
Figure C.19: Unblinded BDT score distribution (pre-fit) for the boosted (left) and VBF (right) W-enriched control region of the $\sqrt{s}$ =7 TeV (top) and $\sqrt{s}$ =8 TeV (bottom) analysis. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

APPENDIX D

Validation plots for the multiple BDT analysis based on the $\sqrt{s}$ =8 TeV dataset

D.1 Validation of discriminating variables

In the following, the variables used for the training of the optimised BDTs in Section 9.2 are displayed for the boosted and VBF signal regions as well as corresponding control regions. Due to the overlap with the analysis in Chapter 7 and to avoid redundant information, only variables, which are exclusively used for the alternative training are shown. All other variables are already summarised in Appendix C.1.



(a) (b) (c)

Figure D.1: Comparison of data with the combined background model with respect to BDT input variables in the boosted signal region of the alternative $\sqrt{s}$ =8 TeV analysis: (a) transverse mass $m_T(E_T^{\text{miss}}, \tau_{\text{had}})$, (b) momentum difference between visible τ-decay products, (c) pseudorapidity of leading jet. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

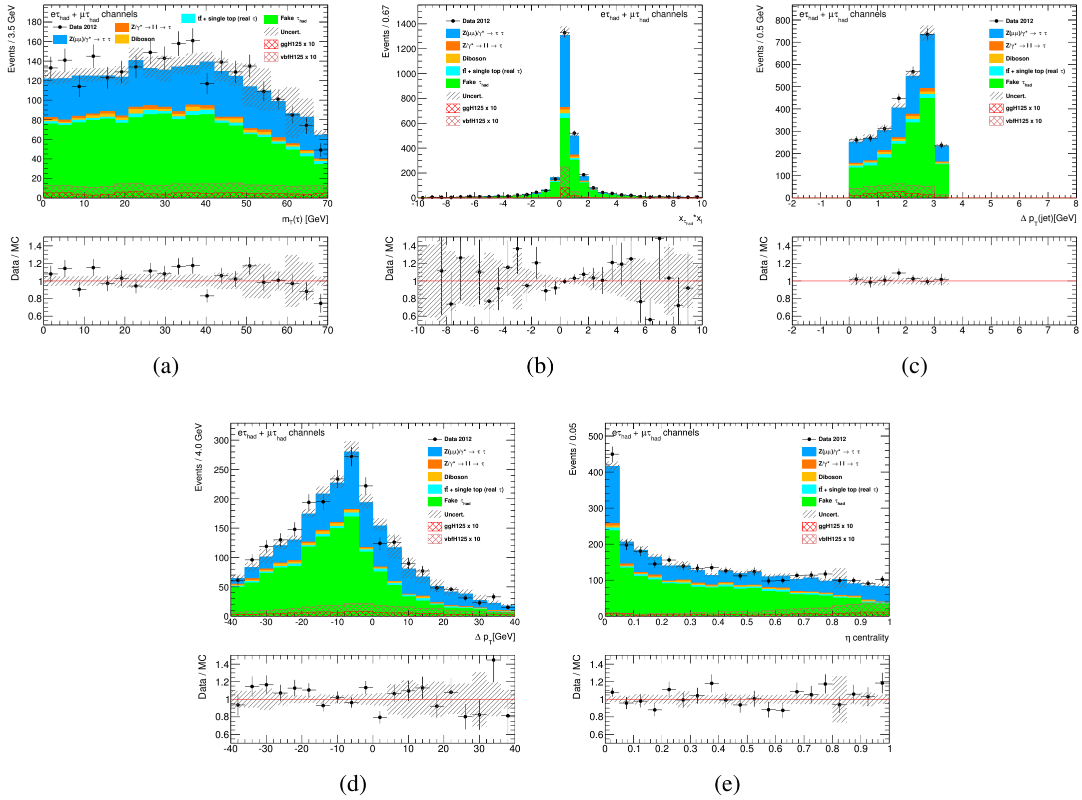


Figure D.2: Comparison of data with the combined background model with respect to BDT input variables in the VBF signal region of the alternative $\sqrt{s}$ =8 TeV analysis: (a) transverse mass $m_T(E_T^{\text{miss}}, \tau_{\text{had}})$, (b) $x_{\tau_{\text{had}}} \cdot x_\ell$, (c) $\Delta\phi$ between leading jets, (d) momentum difference between visible τ-decay products, (e) η-centrality. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

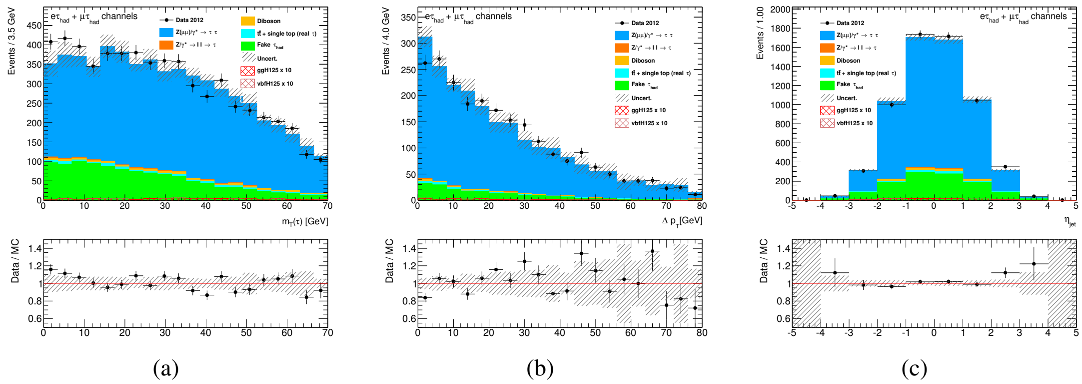


Figure D.3: Comparison of data with the combined background model with respect to BDT input variables in the boosted Z-enriched control region of the alternative $\sqrt{s}$ =8 TeV analysis: (a) transverse mass $m_T(E_T^{\text{miss}}, \tau_{\text{had}})$, (b) momentum difference between visible τ-decay products, (c) pseudorapidity of leading jet. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

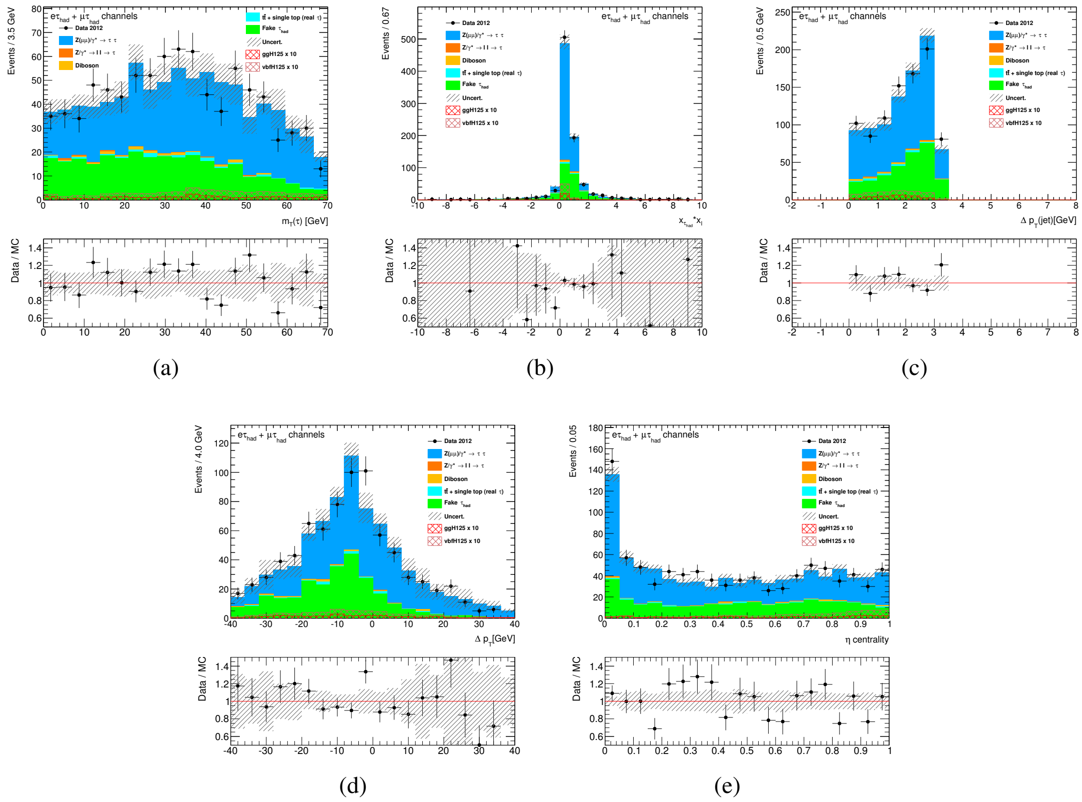


Figure D.4: Comparison of data with the combined background model with respect to BDT input variables in the VBF Z-enriched control region of the alternative $\sqrt{s}$ =8 TeV analysis: (a) transverse mass $m_T(E_T^{\mathrm{miss}}, \tau_{\mathrm{had}})$, (b) $x_{\tau_{\mathrm{had}}} \cdot x_\ell$, (c) $\Delta\phi$ between leading jets, (d) momentum difference between visible τ-decay products, (e) η-centrality. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

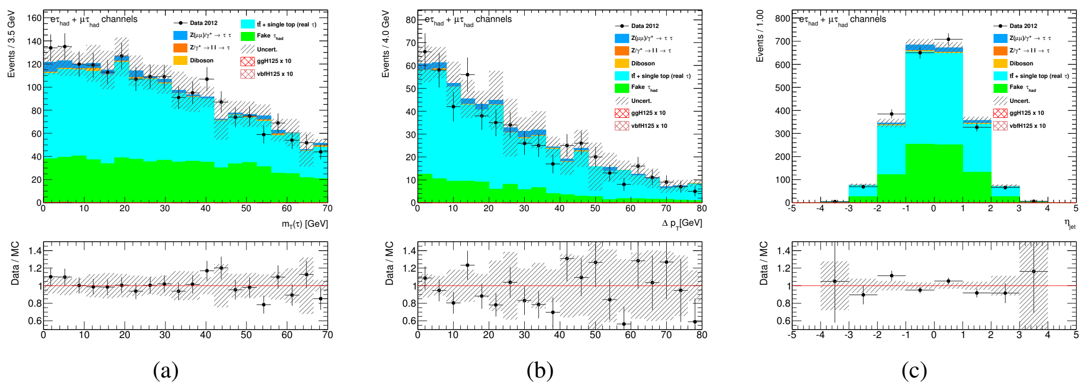


Figure D.5: Comparison of data with the combined background model with respect to BDT input variables in the boosted top control region of the alternative $\sqrt{s}$ =8 TeV analysis: (a) transverse mass $m_T(E_T^{\mathrm{miss}}, \tau_{\mathrm{had}})$, (b) momentum difference between visible τ-decay products, (c) pseudorapidity of leading jet. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

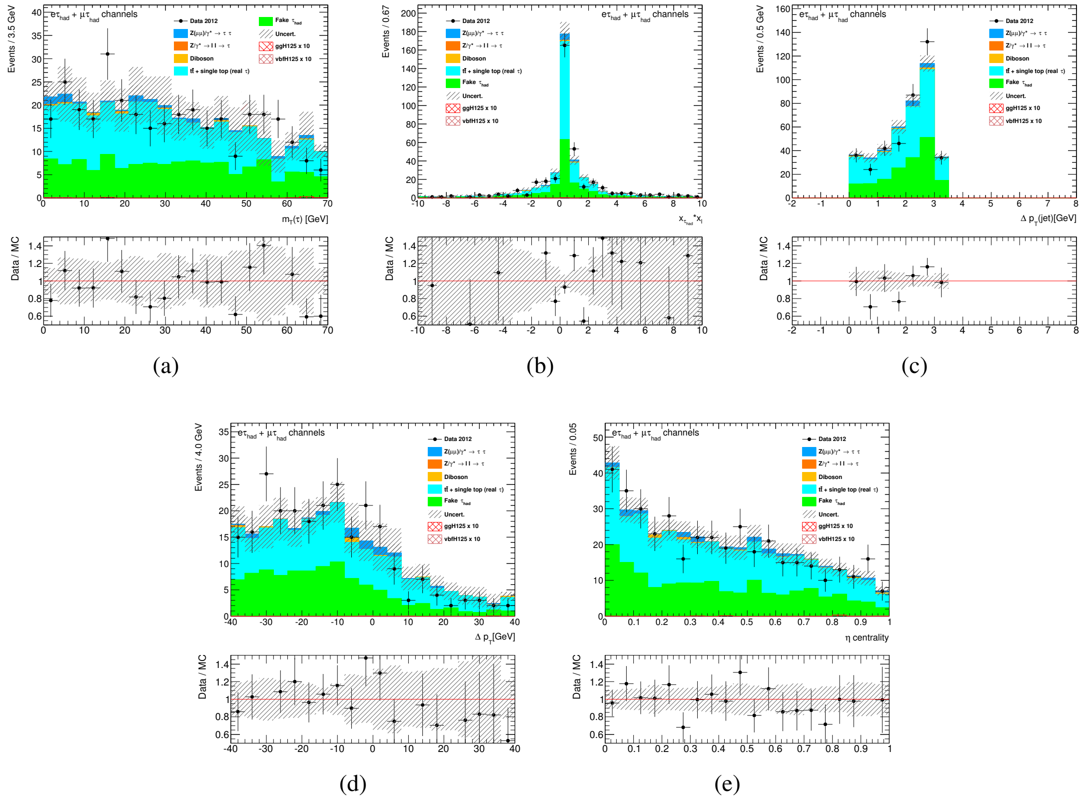


Figure D.6: Comparison of data with the combined background model with respect to BDT input variables in the VBF top control region of the alternative $\sqrt{s}$ =8 TeV analysis: (a) transverse mass $m_T(E_T^{\text{miss}}, \tau_{\text{had}})$, (b) $x_{\tau_{\text{had}}} \cdot x_\ell$, (c) $\Delta\phi$ between leading jets, (d) momentum difference between visible τ-decay products, (e) η-centrality. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

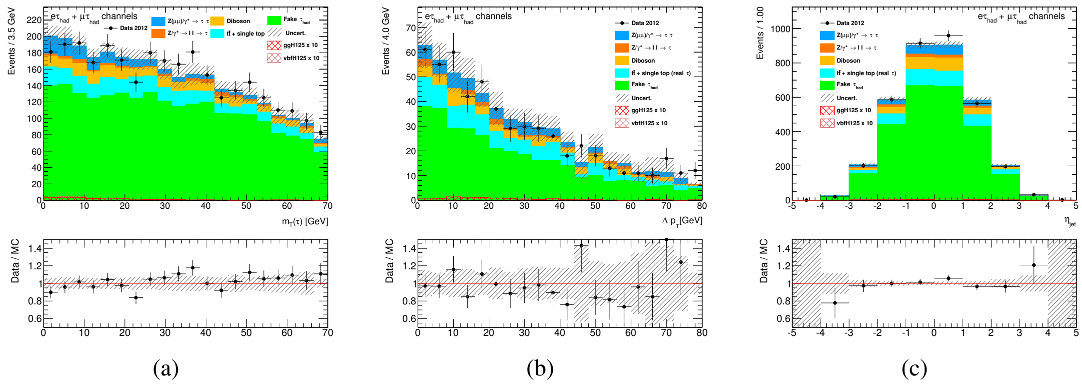
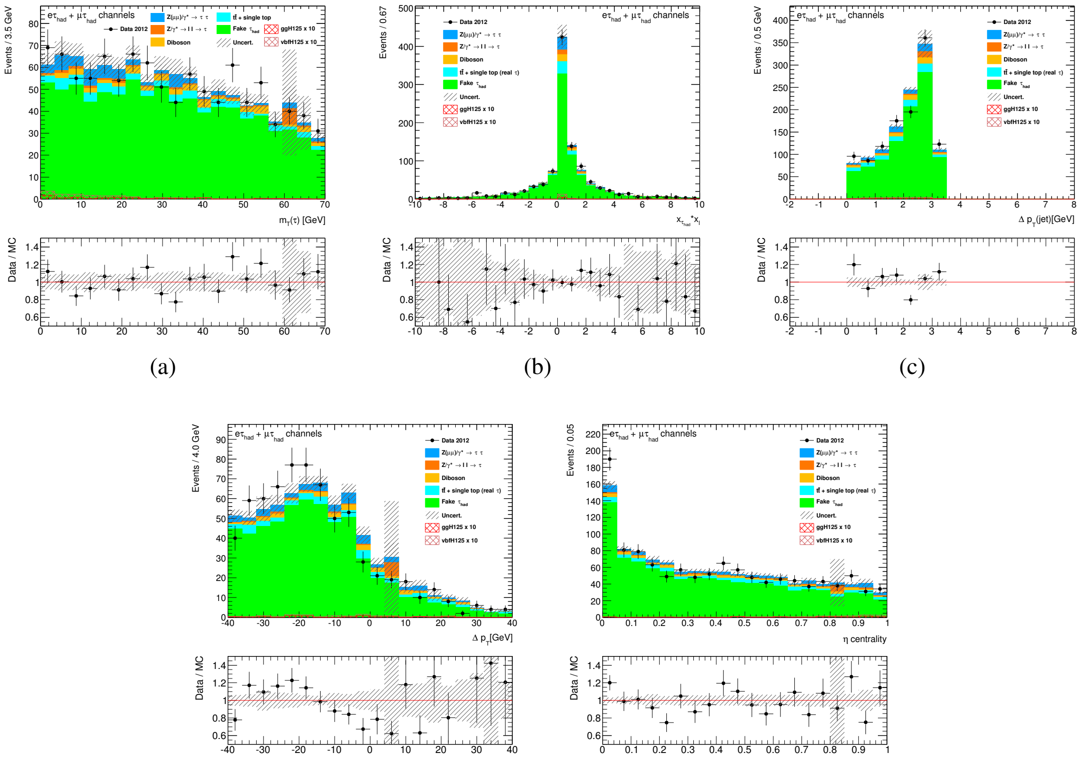


Figure D.7: Comparison of data with the combined background model with respect to BDT input variables in the boosted W-enriched control region of the alternative $\sqrt{s}$ =8 TeV analysis: (a) transverse mass $m_T(E_T^{\text{miss}}, \tau_{\text{had}})$, (b) momentum difference between visible τ-decay products, (c) pseudorapidity of leading jet. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.



(a) (b) (c)

(d) (e)

Figure D.8: Comparison of data with the combined background model with respect to BDT input variables in the VBF W-enriched control region of the alternative $\sqrt{s}$ =8 TeV analysis: (a) transverse mass $m_T(E_T^{\text{miss}}, \tau_{\text{had}})$, (b) $x_{\tau_{\text{had}}} \cdot x_\ell$, (c) $\Delta\phi$ between leading jets, (d) momentum difference between visible τ-decay products, (e) η-centrality. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

D.2 Validation of the BDT score distributions

The BDT score distributions in the boosted and VBF control regions of the optimised $\sqrt{s}$ =8 TeV analysis (Section 9.2) are depicted below.

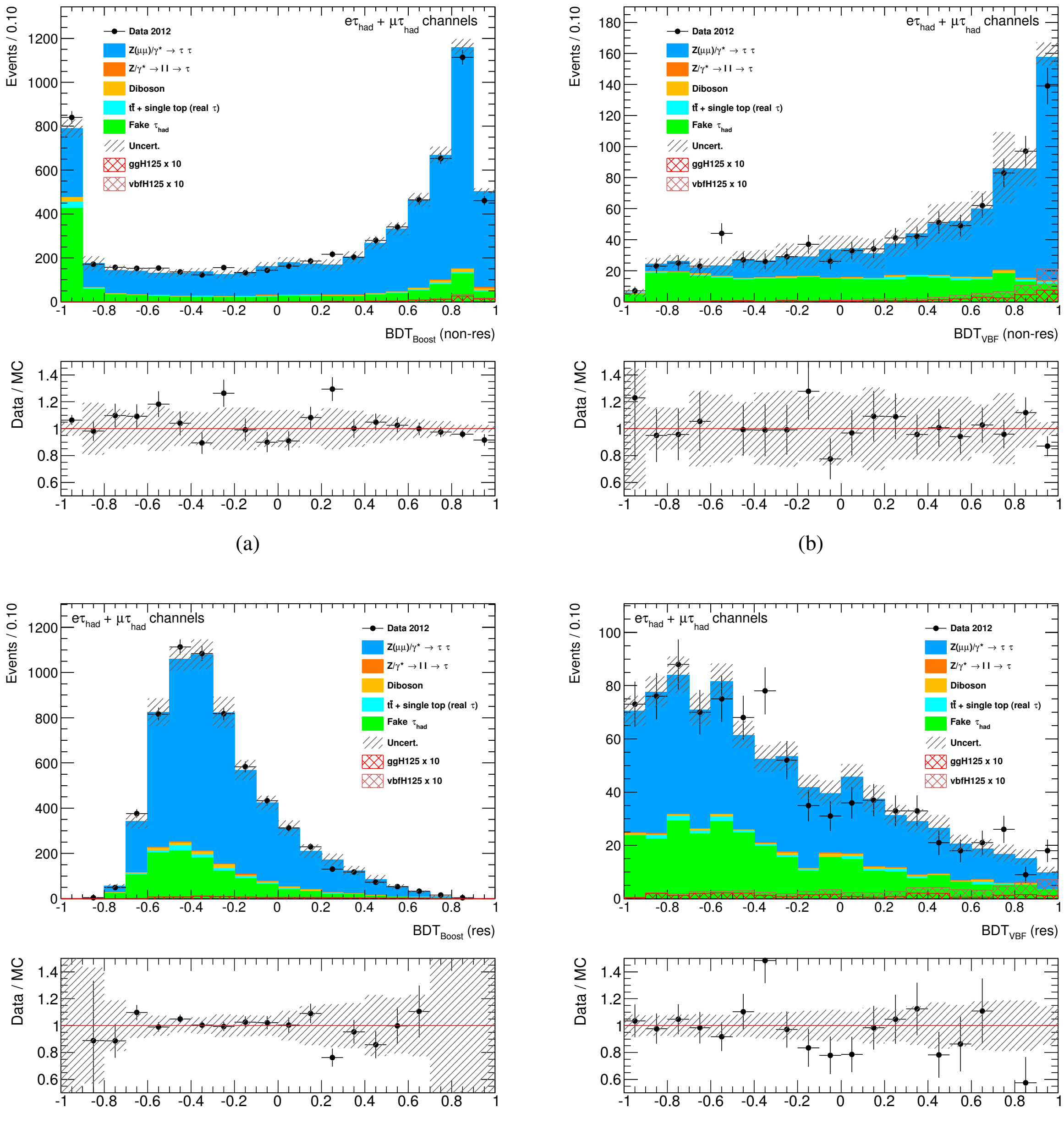


Figure D.9: Unblinded BDT score distribution (pre-fit) for the boosted (left) and VBF (right) 8 TeV Z-enriched control region. The upper plots show the BDTs trained against non-resonant background, like W+jets or mulitjet events, the lower plots depict the BDTs employed to separate $Z \to \tau\tau$ background from Higgs signal processes. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

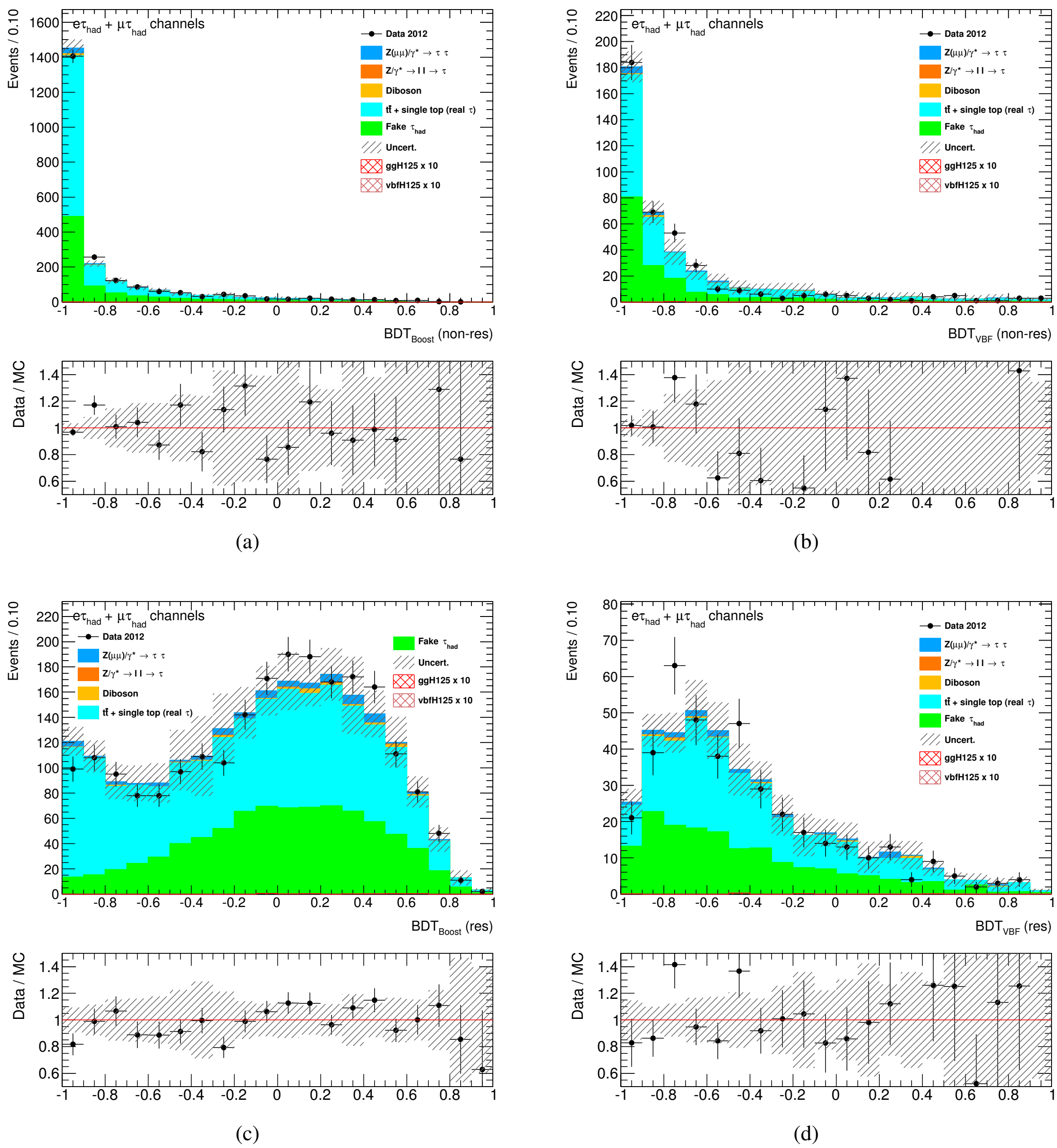


Figure D.10: Unblinded BDT score distribution (pre-fit) for the boosted (left) and VBF (right) 8 TeV top control region. The upper plots show the BDTs trained against non-resonant background, like W+jets or mulitjet events, the lower plots depict the BDTs employed to separate $Z \rightarrow \tau\tau$ background from Higgs signal processes. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

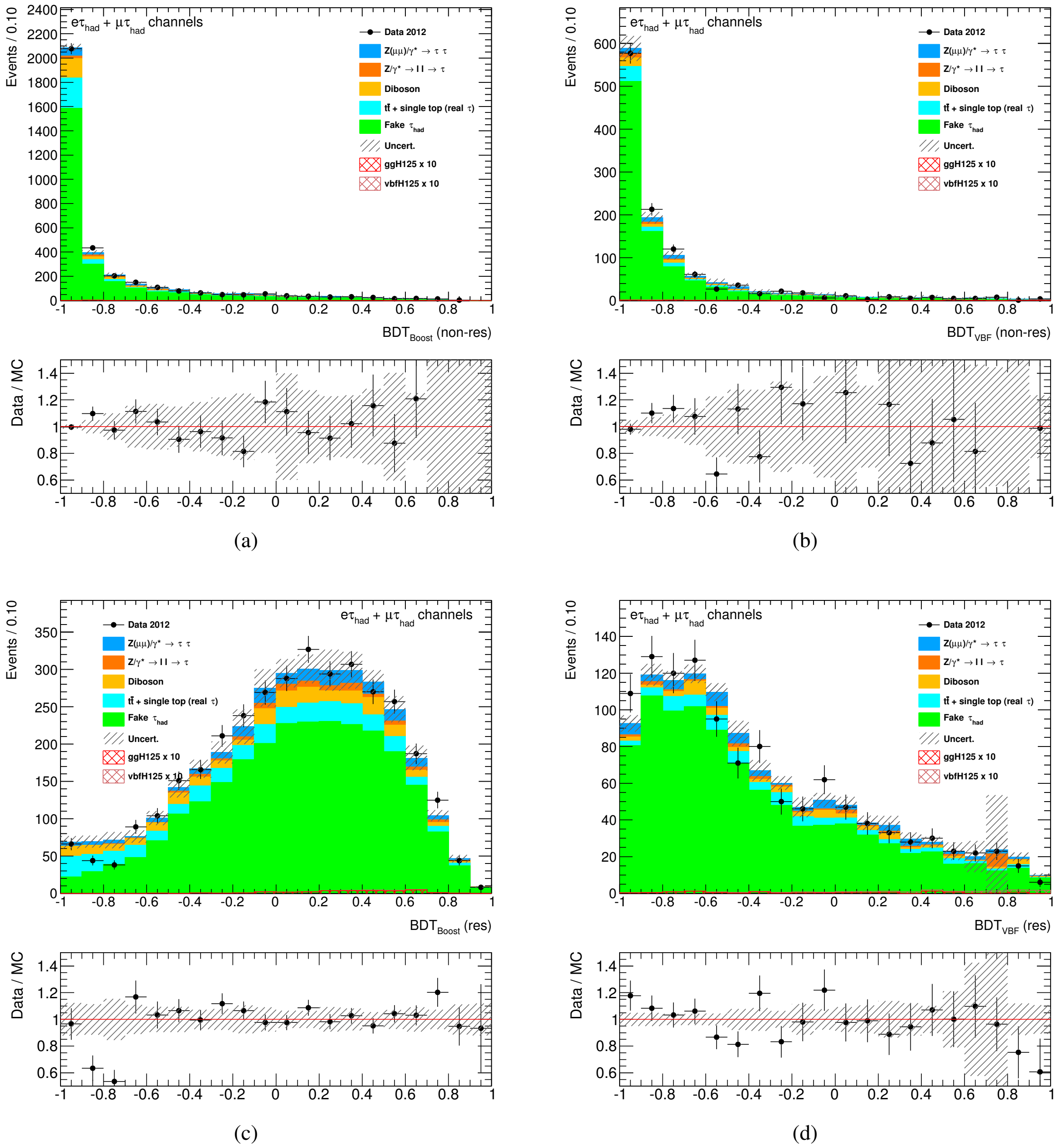


Figure D.11: Unblinded BDT score distribution (pre-fit) for the boosted (left) and VBF (right) 8 TeV W-enriched control region. The upper plots show the BDTs trained against non-resonant background, like W+jets or mulitjet events, the lower plots depict the BDTs employed to separate $Z \to \tau\tau$ background from Higgs signal processes. The ratios show the relative difference of the data to the total background estimate including systematic uncertainties in the grey error band.

APPENDIX E

Labelling of nuisance parameters

In the following, the names of the nuisance parameters illustrated in Figure 7.30 and 9.7 are explained.

- **norm_LH12_Top_vbf/boost:** Normalisation factor of the top background in the 8 TeV VBF/boosted category.
- **norm_LH12_Ztt_01bv:** Normalisation factor of the $Z \rightarrow \tau\tau$ background in the 8 TeV VBF category.
- **SR_(2011)_BDT_2jetvbf/1jetb_bin_Y:** Statistical uncertainty of the background model in the 8 TeV (7 TeV) analysis. The uncertainty refers to bin Y of the BDT score distribution in the VBF (*2jetvbf*) or boosted (*1jetb*) category.
- **LUMI_2012/2011:** Uncertainty of the measured luminosity in the 8 TeV (7 TeV) dataset.
- **EMB_ISOL_2012/2011:** Uncertainty on the $Z \rightarrow \tau\tau$ background modelled by the embedding technique in the 8 TeV (7 TeV) analysis. This systematic refers to isolation criteria which are chosen for the selection of $Z \rightarrow \mu\mu$ input data events.
- **EMB_MFS_2012/2011:** Systematic uncertainty related to the embedded $Z \rightarrow \tau\tau$ background used in the 8 TeV (7 TeV) analysis. It describes the uncertainty which arises due the subtraction of cell energies deposited by muons.
- **LH1X_Fake_boost_COMP:** Systematic uncertainty of the fake factor procedure which is related to the background composition in the 8 TeV (7 TeV) boosted category.
- **LH11_Fake_boost_RW:** Systematic uncertainty of the fake factor procedure derived from the MC closure test for the 7 TeV boosted category.
- **LH12/LH11_Fake_vbf/boost_STATY**: Statistic uncertainty of the fake factor derived for background Y (=1,2,3..) in the boosted (vbf) category of the 8 TeV (LH12) and 7 TeV (LH11) analysis, respectively.
- **JES_Detector1:** Systematic uncertainty component of the jet energy scale associated to the detector.

- **JES_Eta_Modelling:** Systematic uncertainty on the jet energy scale which refers to the η- intercalibration.
- **JES_2012/2011_Modelling1:** Component of the jet energy scale uncertainty which is associated to uncertainties arising in the calibration procedure in the 8 TeV (7 TeV) analysis.
- **JES_FlavComp_TAU_G/Q:** Systematic uncertainty component of the jet energy scale related to jets which originate from parton fragmentation (gluons G or quarks Q).
- **JES_FlavResp:** Uncertainty component of the jet energy scale which accounts for differences in the calorimeter response. The response depends on the jet flavour (quark or gluon initiated jets).
- **JES_NPV:** Component of the jet energy scale uncertainty related to pileup.
- **JES_2012_Eta_StatMethod:** Systematic uncertainty on the jet energy scale which refers to the statistical component of the η- intercalibration in the 8 TeV analysis.
- **JES_2012/2011_Statistical1:** Statistical component of the jet energy scale uncertainty in the 8 TeV (7 TeV) analysis.
- **JES_Flavb:** Systematic uncertainty component of the jet energy scale which is related to truth b-jets.
- **JES_2012/2011_PileRho_TAU_QQ/QG/GG:** Component of the pileup uncertainty of the jet energy scale in the 8 TeV (7 TeV) analysis. This component differentiates between the quark-quark (QQ), quark-gluon (QG) and gluon-gluon (GG) initial state.
- **JER:** Uncertainty on the jet energy resolution.
- **BTag_LY_2012/2011:** Systematic uncertainty component Y related to the b-tagging.
- **PU_RESCALE_2012/2011:** Uncertainties due to pileup rescaling in the 8 TeV (7 TeV) analysis.
- **EL_SCALE:** Uncertainty related to the electron energy scale.
- **EL_RES** Uncertainty of the electron energy resolution.
- **MU_EFF_2012/2011:** Combination of systematic uncertainties related to the muon trigger, reconstruction and isolation efficiency in the 8 TeV (7 TeV) analysis.
- **TAU_ID_2012/2011:** Uncertainty referring to the τ_{had} identification in the 8 TeV (7 TeV) analysis.
- **TAU_ID_STAT_2012:** Statistical component of the τ_{had} identification uncertainty in the 8 TeV analysis.
- **TAU_EFAKE_2012/2011:** Uncertainty on the rate in the 8 TeV (7 TeV) analysis which describes the misidentification probability of electrons as τ_{had}- candidates.
- **TES_Fake_2012/2011:** Systematic uncertainty of the τ energy scale which is related to events with objects misidentified as τ_{had} candidates in the 8 TeV analysis.
- **TES_InSitu_2012:** Uncertainty referring to the in-situ calibration of the tau energy scale in the 8 TeV analysis.

- **TES_Model_2012:** Uncertainty component of the τ energy scale related to the modelling of true τ leptons in the 8 TeV analysis.
- **MET_SCALESOFT_2012/2011:** Systematic uncertainty component of E_T^{miss} associated to the energy scale in the 8 TeV (7 TeV) analysis.
- **MET_RESOSOFT_2012/2011:** Systematic uncertainty component of E_T^{miss} related to the resolution and soft term in the 8 TeV (7 TeV) analysis.
- **LH12_Zlt_vbf_DETAJJ:** Uncertainty related to the reweighting technique applied on $Z \to ee/\mu\mu + jets$ events in the 8 TeV analysis.
- **BR_tautau:** Uncertainty on the calculated $H \to \tau\tau$ branching ratio.
- **GenAccq2Z:** Systematics uncertainty related to the modelling of $Z \to ee/\mu\mu + jets$. It accounts for changes in the acceptance due to the choice of a particular event generator.
- **QCDscale_ggH_XX:** Theoretical uncertainty components related to the QCD scale which originate from missing higher order QCD corrections.
- **pdf_Higgs_XX:** Theoretical uncertainty components arising from parton distribution functions.
- **UE_qq/gg:** Theoretical uncertainty related to the modelling of the underlying event.
- **NLO_EW_Higgs:** Theoretical uncertainties related to the next-to-leading order electroweak corrections.
- **LH12_Top_vbf_MCNLO:** Uncertainty which refers to the reweighting procedure of the Higgs p_T spectrum in the VBF category of the 8 TeV analysis.
- **BCH_LH12_vbf:** Uncertainty related to the BCH correction procedure.

Bibliography

[1] The ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC", *Physics Letters B* 716.1 (2012) 1 –29, ISSN: 0370-2693, DOI: 10.1016/j.physletb.2012.08.020, arXiv: 1207.7214.

[2] The CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC", *Physics Letters B* 716.1 (2012) 30 –61, ISSN: 0370-2693, DOI: 10.1016/j.physletb.2012.08.021, arXiv: 1207.7235.

[3] The ATLAS Collaboration, "Measurement of the Higgs boson mass from the $H \to \gamma\gamma$ and $H \to ZZ^* \to 4\ell$ channels with the ATLAS detector using 25 fb^{-1} of pp collision data", *Phys. Rev.* D90.5 (2014) 052004, DOI: 10.1103/PhysRevD.90.052004, arXiv: 1406.3827 [hep-ex].

[4] The CMS Collaboration, "Precise determination of the mass of the Higgs boson and studies of the compatibility of its couplings with the standard model" (2014).

[5] The ATLAS Collaboration, "Evidence for the spin-0 nature of the Higgs boson using ATLAS data", *Physics Letters B* 726.1-3 (2013) 120 –144, ISSN: 0370-2693, DOI: 10.1016/j.physletb.2013.08.026, arXiv: 1307.1432.

[6] The CMS Collaboration, "Study of the Mass and Spin-Parity of the Higgs Boson Candidate via its Decays to Z Boson Pairs", *Phys. Rev. Lett.* 110.8 (2013) 081803, DOI: 10.1103/PhysRevLett.110.081803, arXiv: 1212.6639 [hep-ex].

[7] The CMS Collaboration, "Constraints on the spin-parity and anomalous HVV couplings of the Higgs boson in proton collisions at 7 and 8 TeV", *Phys. Rev.* D92.1 (2015) 012004, DOI: 10.1103/PhysRevD.92.012004, arXiv: 1411.3441 [hep-ex].

[8] The ATLAS Collaboration, "Evidence for the Higgs-boson Yukawa coupling to tau leptons with the ATLAS detector", *JHEP* 04 (2015) 117, DOI: 10.1007/JHEP04(2015)117, arXiv: 1501.04943 [hep-ex].

[9] The ATLAS Collaboration, "Modelling $Z \to \tau\tau$ processes in ATLAS with τ-embedded $Z \to \mu\mu$ data", *JINST* 10 (2015) P09018, DOI: 10.1088/1748-0221/2015/9/P09018, arXiv: 1506.05623 [hep-ex].

[10] Higgs, Peter W., "Broken symmetries, massless particles and gauge fields", *Physics Letters* 12.2 (1964) 132 –133, ISSN: 0031-9163, DOI: 10.1016/0031-9163(64)91136-9.

[11] P. W. Higgs, "Broken Symmetries and the Masses of Gauge Bosons", *Phys. Rev. Lett.* 13 (16 1964) 508–509, DOI: 10.1103/PhysRevLett.13.508.

[12] G. S. Guralnik, C. R. Hagen and T. W. B. Kibble, "Global Conservation Laws and Massless Particles", *Phys. Rev. Lett.* 13 (20 1964) 585–587, DOI: 10.1103/PhysRevLett.13.585.

[13] F. Halzen and A. Martin, *Quarks and Leptons: An Introductory Course in Modern Particle Physics*, 1st ed., John Wiley and Sons, 1984.

[14] D. Griffiths, *Introduction to Elementary Particles*, 2nd ed., WILEY-VCH Verlag GmbH, 2008.

[15] D. G. W.N. Cottingham, *An introduction to the Standard Model of particle physics*, 2nd ed., Cambridge University Press, 2007.

[16] S. Dittmaier et al., *Handbook of LHC Higgs Cross Sections: 1. Inclusive Observables*, Geneva: CERN, 2011, DOI: 10.5170/CERN-2011-002, arXiv: 1101.0593.

[17] S. Dittmaier et al., "Handbook of LHC Higgs Cross Sections: 2. Differential Distributions", tech. rep., Geneva, 2012, DOI: 10.5170/CERN-2012-002, arXiv: 1201.3084.

[18] S Heinemeyer et al., "Handbook of LHC Higgs Cross Sections: 3. Higgs Properties" (2013), DOI: 10.5170/CERN-2013-004, arXiv: 1307.1347 [hep-ph].

[19] S. Dittmaier and M. Schumacher, "The Higgs Boson in the Standard Model - From LEP to LHC: Expectations, Searches, and Discovery of a Candidate", *Prog. Part. Nucl. Phys.* 70 (2013) 1–54, DOI: 10.1016/j.ppnp.2013.02.001, arXiv: 1211.4828 [hep-ph].

[20] *Go on particle Quest at the first CERN webfest*, https://cds.cern.ch/journal/CERNBulletin/2012/35/News Articles/1473657, CERN, 2012.

[21] D. J. Gross and F. Wilczek, "Ultraviolet Behavior of Non-Abelian Gauge Theories", *Phys. Rev. Lett.* 30 (26 1973) 1343–1346, DOI: 10.1103/PhysRevLett.30.1343.

[22] H. D. Politzer, "Reliable Perturbative Results for Strong Interactions?", *Phys. Rev. Lett.* 30 (26 1973) 1346–1349, DOI: 10.1103/PhysRevLett.30.1346.

[23] H. Fritzsch, M. Gell-Mann and H. Leutwyler, "Advantages of the color octet gluon picture", *Physics Letters B* 47.4 (1973) 365 –368, ISSN: 0370-2693, DOI: 10.1016/0370-2693(73)90625-4.

[24] S. L. Glashow, "Partial-symmetries of weak interactions", *Nuclear Physics* 22.4 (1961) 579 –588, ISSN: 0029-5582, DOI: 10.1016/0029-5582(61)90469-2.

[25] S. Weinberg, "A Model of Leptons", *Phys. Rev. Lett.* 19 (21 1967) 1264–1266, DOI: 10.1103/PhysRevLett.19.1264.

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

[27] F. Englert and R. Brout, "Broken Symmetry and the Mass of Gauge Vector Mesons", *Phys. Rev. Lett.* 13 (9 1964) 321–323, DOI: 10.1103/PhysRevLett.13.321.

[28] J. Ellis, "Outstanding questions: physics beyond the Standard Model", *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 370.1961 (2012) 818–830, ISSN: 1364-503X, DOI: 10.1098/rsta.2011.0452, eprint: http://rsta.royalsocietypublishing.org/content/370/1961/818.full.pdf.

[29] C. S. Wu et al., "Experimental Test of Parity Conservation in Beta Decay", *Phys. Rev.* 105 (4 1957) 1413–1415, DOI: 10.1103/PhysRev.105.1413.

[30] A. Barczyk et al., "Measurement of the Fermi constant by FAST", *Physics Letters B* 663.3 (2008) 172 –180, DOI: 10.1016/j.physletb.2008.04.006.

[31] K. A. Olive et al., "Review of Particle Physics", *Chin. Phys.* C38 (2014) 090001, DOI: 10.1088/1674-1137/38/9/090001.

[32] R. K. Ellis and W. J. Stirling, "QCD and collider physics", *The 1990 CERN School of Physics Mallorca, Spain, September 16-29*, 1990, URL: http://lss.fnal.gov/cgi-bin/find_paper.pl?conf-90-164.

[33] *Strangeness at LHCb*, http://imperialhep.blogspot.de/2011/08/strangeness-at-lhcb.html, Experimental particle physicists at Imperial College London, 2011.

[34] T. Plehn, *Lectures on LHC Physics*, 2nd ed., Springer Verlag Berlin Heidelberg, 20012.

[35] J. M. Campbell, J. W. Huston and W. J. Stirling, "Hard interactions of quarks and gluons: a primer for LHC physics", *Reports on Progress in Physics* 70.1 (2007) 89, URL: http://stacks.iop.org/0034-4885/70/i=1/a=R02.

[36] V. Radescu, "Hera Precision Measurements and Impact for LHC Predictions", *Proceedings, 46th Rencontres de Moriond on QCD and High Energy Interactions*, 2011 27–30, arXiv: 1107.4193 [hep-ex].

[37] A. D. Martin et al., "Parton distributions for the LHC", *Eur. Phys. J.* C63 (2009) 189–285, DOI: 10.1140/epjc/s10052-009-1072-5, arXiv: 0901.0002 [hep-ph].

[38] H.-L. Lai et al., "New parton distributions for collider physics", *Phys. Rev. D* 82 (2010) 074024, DOI: 10.1103/PhysRevD.82.074024, arXiv: 1007.2241.

[39] J. Pumplin et al., "New Generation of Parton Distributions with Uncertainties from Global QCD Analysis", *JHEP* 0207 (2002) 012, DOI: 10.1088/1126-6708/2002/07/012, arXiv: hep-ph/0201195.

[40] A. D. Martin. W.J. Stirling, R.S. Thorne, G.Watt, "Parton distributions for the LHC", *Eurpean Physical Journal C* 63.2 (2009) 189–285, DOI: 10.1140/epjc/s10052-009-1072-5, arXiv: 0901.0002.

[41] V. V. Sudakov, "Vertex parts at very high-energies in quantum electrodynamics", *Sov. Phys. JETP* 3 (1956), [Zh. Eksp. Teor. Fiz.30,87(1956)] 65–71.

[42] S. Frixione and B. R. Webber, "Matching NLO QCD computations and parton shower simulations", *Journal of High Energy Physics* 2002.06 (2002) 029, DOI: 10.1088/1126-6708/2002/06/029, arXiv: hep-ph/0204244.

[43] J. Alwall et al., "Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions", *European Physical Journal C* 53.03 (2008) 473–500, DOI: 10.1140/epjc/s10052-007-0490-5, arXiv: 0706.2569.

[44] T. Sjostrand, S. Mrenna and P. Z. Skands, "A Brief Introduction to PYTHIA 8.1", *Comput. Phys. Commun.* 178 (2008) 852, DOI: 10.1016/j.cpc.2008.01.036, arXiv: 0710.3820 [hep-ph].

[45] G.Corcella et al., "HERWIG 6: an event generator for Hadron Emission Reactions With Interfering Gluons (including supersymmetric processes)", *JHEP* 01 (2001) 097, DOI: 10.1088/1126-6708/2001/01/010, arXiv: hep-ph/0011363.

[46] Andy Buckley et al., "General-purpose event generators for LHC physics", *Physics Reports* 504.5 (2011) 145 –233, ISSN: 0370-1573, DOI: 10.1016/j.physrep.2011.03.005, arXiv: 1101.2599.

[47] The ATLAS Collaboration, "Jet energy measurement with the ATLAS detector in proton-proton collisions at $\sqrt{s} = 7$ TeV", *The European Physical Journal C* 73.3, 2304 (2013), DOI: 10.1140/epjc/s10052-013-2304-2.

[48] P. Bolzoni et al, "Higgs boson production via vector-boson fusion at next-to-next-to-leading order in QCD", *Phys.Rev.Lett* 105 (2010) 011801, DOI: 10.1103/PhysRevLett.105.011801, arXiv: 1003.4451 [hep-ph].

[49] A. Denner et al., "Standard Model Higgs-Boson Branching Ratios with Uncertainties", *Eur. Phys. J.* C71 (2011) 1753, DOI: 10.1140/epjc/s10052-011-1753-8, arXiv: 1107.5909 [hep-ph].

[50] A. Djouadi, J. Kalinowski and M. Spira, "HDECAY: a program for Higgs boson decays in the Standard Model and its supersymmetric extension", *Comput. Phys. Commun.* 108 (1998) 56–74, DOI: 10.1016/S0010-4655(97)00123-9, arXiv: hep-ph/9704448 [hep-ph].

[51] A. Bredenstein et al., "Radiative corrections to the semileptonic and hadronic Higgs-boson decays $H \to WW/ZZ \to 4$ fermions", *JHEP* 02 (2007) 080, DOI: 10.1088/1126-6708/2007/02/080, arXiv: hep-ph/0611234 [hep-ph].

[52] The ATLAS Collaboration, "Measurements of Higgs boson production and couplings in diboson final states with the ATLAS detector at the LHC", *Physics Letters B* 726.1-3 (2013) 88 –119, ISSN: 0370-2693, DOI: 10.1016/j.physletb.2013.08.010, arXiv: 1307.1427.

[53] The ATLAS Collaboration, "Search for the $b\bar{b}$ decay of the Standard Model Higgs boson in associated $(W/Z)H$ production with the ATLAS detector", *JHEP* 01 (2015) 069, DOI: 10.1007/JHEP01(2015)069, arXiv: 1409.6212 [hep-ex].

[54] The ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider", *Journal of Instrumentation* 3.08 (2008) S08003, URL: http://stacks.iop.org/1748-0221/3/i=08/a=S08003.

[55] L. Evans and P. Bryant, "LHC Machine", *Journal of Instrumentation* 3.08 (2008) S08001, URL: http://stacks.iop.org/1748-0221/3/i=08/a=S08001.

[56] The LHCb collaboration, *LHCb reoptimized detector design and performance: Technical Design Report*, Technical Design Report LHCb, Geneva: CERN, 2003, URL: http://cds.cern.ch/record/630827.

[57] *ATLAS Photos*, http://www.atlas.ch/photos/lhc.html, ATLAS Experiment, CERN, 2014.

[58] The ALICE Collaboration, "The ALICE experiment at the CERN LHC", *Journal of Instrumentation* 3.08 (2008) S08002, DOI: 10.1088/1748-0221/3/08/S08002.

[59] The TOTEM Collaboration, "The TOTEM Experiment at the CERN Large Hadron Collider", *Journal of Instrumentation* 3.08 (2008) S08007, DOI: 10.1088/1748-0221/3/08/S08007.

[60] The LHCf Collaboration, "The LHCf detector at the CERN Large Hadron Collider", *Journal of Instrumentation* 3.08 (2008) S08006, DOI: 10.1088/1748-0221/3/08/S08006.

[61] The MoEDAL Collaboration, "Technical Design Report of the MoEDAL Experiment", tech. rep. CERN-LHCC-2009-006; MoEDAL-TDR-001, Geneva: CERN, 2009, URL: http://cds.cern.ch/record/1181486.

[62] *LHC Images*, http://lhc-machine-outreach.web.cern.ch/lhc-machine-outreach/images/complex/Cern-complex.gif, 2015.

[63] *LHC@home*, http://lhcathome.web.cern.ch/projects/sixtrack/all-about-accelerators, 2015.

[64] M. Schott and M. Dunford, "Review of single vector boson production in pp collisions at $\sqrt{s}$ = 7 TeV", *Eur. Phys. J.* C74 (2014) 2916, DOI: 10.1140/epjc/s10052-014-2916-1, arXiv: 1405.1160 [hep-ex].

[65] ATLAS Collaboration, *ATLAS detector and physics performance: Technical Design Report, 1*, Technical Design Report ATLAS, Electronic version not available, Geneva: CERN, 1999, URL: http://cds.cern.ch/record/391176.

[66] The ATLAS Collaboration, *Expected performance of the ATLAS experiment: detector, trigger and physics*, Geneva: CERN, 2009, URL: https://cds.cern.ch/record/1125884.

[67] R Achenbach et al., "The ATLAS Level-1 Calorimeter Trigger", *Journal of Instrumentation* 3.03 (2008) P03001, URL: http://stacks.iop.org/1748-0221/3/i=03/a=P03001.

[68] The ATLAS Collaboration, "ATLAS first level trigger: Technical design report" (1998).

[69] *The ATLAS Collaboration*, https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResults, 2015.

[70] *The ATLAS Collaboration*, https://twiki.cern.ch/twiki/bin/view/AtlasPublic/LuminosityPublicResultsRun2, 2015.

[71] *ATHENA; the ATLAS common framework, developer guide*, http://atlas.web.cern.ch/Atlas/GROUPS/SOFTWARE/OO/z0_obsolete/architecture/General/Tech.Doc/Manual/AthenaDeveloperGuide.pdf, 2001.

[72] The ATLAS Collaboration, "The ATLAS Simulation Infrastructure", *Eur. Phys. J.* C 70 (2010) 823, DOI: 10.1140/epjc/s10052-010-1429-9, arXiv: 1005.4568 [hep-ex].

[73] S. Agostinelli et al., "Geant4– A Simulation toolkit", *Nucl. Instrum. Meth.* A 506 (2003) 250, DOI: 10.1016/S0168-9002(03)01368-8.

[74] "Performance of the Fast ATLAS Tracking Simulation (FATRAS) and the ATLAS Fast Calorimeter Simulation (FastCaloSim) with single particles", tech. rep. ATL-SOFT-PUB-2014-001, Geneva: CERN, 2014, URL: http://cds.cern.ch/record/1669341.

[75] Cornelissen et al.,
"Concepts, Design and Implementation of the ATLAS New Tracking (NEWT)",
tech. rep. ATL-SOFT-PUB-2007-007. ATL-COM-SOFT-2007-002, Geneva: CERN, 2007,
URL: https://cds.cern.ch/record/1020106.

[76] The ATLAS Collaboration, "Performance of primary vertex reconstruction in proton-proton collisions at $\sqrt{s}$ =7 TeV in the ATLAS experiment", tech. rep. ATLAS-CONF-2010-069, Geneva: CERN, 2010, URL: http://cds.cern.ch/record/1281344.

[77] E. Bouhova-Thacker et al., "Vertex reconstruction in the ATLAS experiment at the LHC", *Nuclear Science Symposium Conference Record, 2008. NSS '08. IEEE*, 2008 1720–1727, DOI: 10.1109/NSSMIC.2008.4774734.

[78] The ATLAS Collaboration, "Performance of the ATLAS Inner Detector Track and Vertex Reconstruction in the High Pile-Up LHC Environment", tech. rep. ATLAS-CONF-2012-042, Geneva: CERN, 2012, URL: http://cds.cern.ch/record/1435196.

[79] W. Lampl et al, "Calorimeter Clustering Algorithms: Description and Performance", tech. rep. ATL-LARG-PUB-2008-002. ATL-COM-LARG-2008-003, Geneva: CERN, 2008, URL: http://cds.cern.ch/record/1099735.

[80] The ATLAS Collaboration, "Measurement of the muon reconstruction performance of the ATLAS detector using 2011 and 2012 LHC proton-proton collision data", *Eur. Phys. J.* C 74 (2014) 3130, DOI: 10.1140/epjc/s10052-014-3130-x, arXiv: 1407.3935 [hep-ex].

[81] The ATLAS collaboration, "Electron efficiency measurements with the ATLAS detector using the 2012 LHC proton-proton collision data" (2014).

[82] R. Frühwirth, "Application of Kalman filtering to track and vertex fitting", *Nuclear Instruments and Methods in Physics Research Section A* 262.2-3 (1987) 444 –450, DOI: 10.1016/0168-9002(87)90887-4.

[83] T.G. Cornelissen et al., "The global χ^2 track fitter in ATLAS", *Journal of Physics: Conference Series* 119.3 (2008) 032013, URL: http://stacks.iop.org/1742-6596/119/i=3/a=032013.

[84] The ATLAS Collaboration, "Improved electron reconstruction in ATLAS using the Gaussian Sum Filter-based model for bremsstrahlung", tech. rep. ATLAS-CONF-2012-047, Geneva: CERN, 2012, URL: https://cds.cern.ch/record/1449796.

[85] M. Cacciari, G. P. Salam and G. Soyez, "The anti-k_t jet clustering algorithm", *JHEP* 04 (2008) 063, DOI: 10.1088/1126-6708/2008/04/063, arXiv: 0802.1189 [hep-ph].

[86] M. Cacciari and G. P. Salam, "Dispelling the N^3 myth for the k_t jet-finder", *Phys. Lett.* B 641 (2006) 57, DOI: 10.1016/j.physletb.2006.08.037, arXiv: hep-ph/0512210.

[87] M. Cacciari and G. P. Salam, "Pileup subtraction using jet areas", *Phys. Lett.* B 659 (2008) 119, DOI: 10.1016/j.physletb.2007.09.077, arXiv: 0707.1378 [hep-ph].

[88] The ATLAS Collaboration, "Jet energy measurement and its systematic uncertainty in proton-proton collisions at $\sqrt{s}$ = 7 TeV with the ATLAS detector", *Eur. Phys. J.* C 75 (2015) 17, DOI: 10.1140/epjc/s10052-014-3190-y, arXiv: 1406.0076 [hep-ex].

[89] The ATLAS Collaboration, "Single hadron response measurement and calorimeter jet energy scale uncertainty with the ATLAS detector at the LHC", *Eur. Phys. J.* C 73 (2013) 2305, DOI: 10.1140/epjc/s10052-013-2305-1, arXiv: 1203.1302 [hep-ex].

[90] The ATLAS Collaboration, "Pile-up subtraction and suppression for jets in ATLAS", tech. rep. ATLAS-CONF-2013-083, Geneva: CERN, 2013, URL: http://cds.cern.ch/record/1570994.

[91] The ATLAS Collaboration, "Commissioning of the ATLAS high-performance b-tagging algorithms in the 7 TeV collision data", tech. rep. ATLAS-CONF-2011-102, Geneva: CERN, 2011, URL: http://cds.cern.ch/record/1369219.

[92] The ATLAS Collaboration, "Calibration of the performance of *b*-tagging for *c* and light-flavour jets in the 2012 ATLAS data", tech. rep. ATLAS-CONF-2014-046, Geneva: CERN, 2014, URL: http://cds.cern.ch/record/1741020.

[93] The ATLAS Collaboration, "Identification and energy calibration of hadronically decaying tau leptons with the ATLAS experiment in pp collisions at $\sqrt{s}$=8 TeV", *The European Physical Journal C* 75.7, 303 (2015), DOI: 10.1140/epjc/s10052-015-3500-z.

[94] G. Nunes Hanninger, N. Wermes and J. Kroseberg, "Observation of $W \to \tau\nu_\tau$ Decays with the ATLAS Experiment", Presented 29 Apr 2011, PhD thesis, Bonn: Bonn U., 2011, URL: http://cds.cern.ch/record/1348491.

[95] The ATLAS Collaboration, "Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s}$ =7 TeV with ATLAS", *Eur. Phys. J.* C 72 (2012) 1844, DOI: 10.1140/epjc/s10052-011-1844-6, arXiv: 1108.5602 [hep-ex].

[96] The ATLAS Collaboration, "Pile-up Suppression in Missing Transverse Momentum Reconstruction in the ATLAS Experiment in Proton-Proton Collisions at $\sqrt{s}$ =8 TeV", tech. rep. ATLAS-CONF-2014-019, Geneva: CERN, 2014, URL: http://cds.cern.ch/record/1702055.

[97] The ATLAS Collaboration, "Performance of Missing Transverse Momentum Reconstruction in ATLAS studied in Proton-Proton Collisions recorded in 2012 at 8 TeV", tech. rep. ATLAS-CONF-2013-082, Geneva: CERN, 2013, URL: http://cds.cern.ch/record/1570993.

[98] *Parton luminosity and cross section plots*, http://www.hep.ph.ic.ac.uk/%7Ewstirlin/plots/plots.html, W.J. Stirling, private communication, 2016.

[99] *TMVA - Toolkit for Multivariate Data Analysis*, http://tmva.sourceforge.net/, version 4.2.0 (available as part of ROOT version 5.34/11), 2011.

[100] J. H. Friedman, "Stochastic gradient boosting", *Computational Statistics and Data Analysis* 38.4 (2002), Nonlinear Methods and Data Mining 367 –378, DOI: 10.1016/S0167-9473(01)00065-2.

[101] Y. Freund and R. E. Schapire,
"A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting",
Journal of Computer and System Sciences 55.1 (1997) 119 –139,
DOI: 10.1006/jcss.1997.1504.

[102] The ATLAS Collaboration,
"Search for the Standard Model Higgs Boson produced in association with a vector boson and decaying into a tau pair in *pp* collisions at $\sqrt{s} = 8$ TeV with the ATLAS detector" (2015),
arXiv: 1511.08352 [hep-ex].

[103] The ATLAS Collaboration, "Search for the associated production of the Higgs boson with a top quark pair in multilepton final states with the ATLAS detector",
Physics Letters B 749 (2015) 519 –541, DOI: 10.1016/j.physletb.2015.07.079.

[104] R.K. Ellis et al., "Higgs decay to $\tau^+\tau^-$: A possible signature of intermediate mass Higgs bosons at high energy hadron colliders", *Nucl.Phys* B297 (1988) 221,
DOI: 10.1016/0550-3213(88)90019-3.

[105] A.Elagin et al., "A New Mass Reconstruction Technique for Resonances Decaying to $\tau\tau$",
Nucl.Instrum.Meth A654 (2011) 481–489, DOI: 10.1016/j.nima.2011.07.009,
arXiv: 1012.4686v2 [hep-ex].

[106] *The ATLAS Collaboration, "Evidence for the Higgs-boson Yukawa coupling to tau leptons with the ATLAS detector"*,
https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PAPERS/HIGG-2013-32/,
Auxiliarly figures, 2015.

[107] The ATLAS Collaboration, "Measurement of the production cross section of jets in association with a Z boson in pp collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector",
JHEP 07 (2013) 032, DOI: 10.1007/JHEP07(2013)032.

[108] The ATLAS Collaboration, "Measurement of the Z/γ^* boson transverse momentum distribution in *pp* collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector", *JHEP* 09 (2014) 145,
DOI: 10.1007/JHEP09(2014)145, arXiv: 1406.3660 [hep-ex].

[109] The ATLAS Collaboration, "Measurement of the electroweak production of dijets in association with a Z-boson and distributions sensitive to vector boson fusion in proton-proton collisions at $\sqrt{s} = 8$ TeV using the ATLAS detector", *JHEP* 04 (2014) 031,
DOI: 10.1007/JHEP04(2014)031, arXiv: 1401.7610 [hep-ex].

[110] N. Moeser and N. Wermes, "A Sensitivity Study for Higgs Boson Production in Vector Boson Fusion in the $H \to \tau\tau \to \ell h + 3\nu$ Final State with ATLAS", Presented 18 Oct 2011,
PhD thesis, Bonn U., 2011, URL: https://cds.cern.ch/record/1540664.

[111] M. Schmitz and N. Wermes, "Higgs production in Vector Boson Fusion in the $H \to \tau\tau \to \ell\ell + 3\nu$ final state with ATLAS: a sensitivity study", Presented 2011,
PhD thesis, Bonn U., 2011, URL: https://cds.cern.ch/record/1540675.

[112] T. Schwindt and N. Wermes, "Evidence for the Higgs Boson Decay to $\tau^+\tau^-$ in the Lepton+Hadrons Final state with ATLAS", Presented 2015, PhD thesis, Bonn U., 2015, URL:
http://hss.ulb.uni-bonn.de/2015/3929/3929.html,urn:nbn:de:hbz:5n-39294.

[113] The ATLAS Collaboration, "Search for neutral MSSM Higgs bosons decaying to $\tau^+\tau^-$ pairs in proton-proton collisions at $\sqrt{s}$ =7 TeV with the ATLAS detector", *Phys. Lett.* B 705 (2011) 174–192, DOI: 10.1016/j.physletb.2011.10.001, arXiv: 1107.5003 [hep-ex].

[114] The ATLAS Collaboration, "Search for the Standard Model Higgs boson in the H to $\tau^+\tau^-$ decay mode in $\sqrt{s}$ = 7 TeV pp collisions with ATLAS", *JHEP* 09 (2012) 070, DOI: 10.1007/JHEP09(2012)070, arXiv: 1206.5971 [hep-ex].

[115] The ATLAS Collaboration, "Search for the neutral Higgs bosons of the Minimal Supersymmetric Standard Model in pp collisions at $\sqrt{s}$ = 7 TeV with the ATLAS detector", *JHEP* 02 (2013) 095, DOI: 10.1007/JHEP02(2013)095, arXiv: 1211.6956 [hep-ex].

[116] The ATLAS Collaboration, "Search for charged Higgs bosons decaying via $H^+ \to \tau\nu$ in $t\bar{t}$ events using pp collision data at $\sqrt{s}$ = 7 TeV with the ATLAS detector", *JHEP* 06 (2012) 039, DOI: 10.1007/JHEP06(2012)039, arXiv: 1204.2760 [hep-ex].

[117] ATLAS Collaboration, "Search for charged Higgs bosons decaying via $H^\pm \to \tau^\pm\nu$ in fully hadronic final states using pp collision data at $\sqrt{s}$ = 8 TeV with the ATLAS detector", *JHEP* 03 (2015) 088, DOI: 10.1007/JHEP03(2015)088, arXiv: 1412.6663 [hep-ex].

[118] The ATLAS Collaboration, "Measurement of the $W \to \tau\nu$ Cross Section in pp Collisions at $\sqrt{s}$ = 7 TeV with the ATLAS experiment", *Phys. Lett.* B 706 (2012) 276–294, DOI: 10.1016/j.physletb.2011.11.057, arXiv: 1108.4101 [hep-ex].

[119] *Development and analysis of a data-driven method to charatcterize $W \to \tau\nu$ decays in ATLAS*, Master thesis, Jessica Liebal, 2011.

[120] S. Jadach and Z. Was, "The tau decay library Tauola: Version 2.4", *Comput. Phys. Commun.* 76 (1993) 361, DOI: 10.1016/0010-4655(93)90061-G.

[121] N. Davidson, T. Przedzinski and Z. Was, "PHOTOS Interface in C++: Technical and Physics Documentation" (2010), arXiv: 1011.0937 [hep-ph].

[122] Z. Czyczula, T. Przedzinski and Z. Was, "TauSpinner Program for Studies on Spin Effect in tau Production at the LHC", *Eur. Phys. J.* C 72 (2012) 1988, DOI: 10.1140/epjc/s10052-012-1988-z, arXiv: 1201.0117 [hep-ph].

[123] A. Kaczmarska et al., "Application of TauSpinner for studies on tau-lepton polarization and spin correlations in Z, W and H decays at LHC", *Acta Phys. Polon.* B 45 (2014) 1921, DOI: 10.5506/APhysPolB.45.1921, arXiv: 1402.2068 [hep-ph].

[124] T. Przedzinski, E. Richter-Was and Z. Was, "TauSpinner: a tool for simulating CP effects in $H \to \tau\tau$ decays at LHC", *Eur. Phys. J.* C 74 (2014) 3177, DOI: 10.1140/epjc/s10052-014-3177-8, arXiv: 1406.1647 [hep-ph].

[125] The ATLAS Collaboration, "Search for neutral Higgs bosons of the minimal supersymmetric standard model in pp collisions at $\sqrt{s}$ = 8 TeV with the ATLAS detector", *JHEP* 11 (2014) 056, DOI: 10.1007/JHEP11(2014)056, arXiv: 1409.6064 [hep-ex].

[126] *Tan Wang*, Thesis in preparation, Bonn University.

[127] D. Alvarez Piqueras et al., “Measurement of the Higgs boson couplings in the $\tau\tau$ final state with the ATLAS detector - Supporting Note”, tech. rep. ATL-COM-PHYS-2014-170, Geneva: CERN, 2014, URL: https://cds.cern.ch/record/1666539.

[128] J. Apostolakis et al., “GEANT4 Physics Lists for HEP”, *Nuclear Science Symposium Conference Record, 2008. NSS ’08. IEEE*, 2008 833–836, DOI: 10.1109/NSSMIC.2008.4774655.

[129] *Reference Physics Lists*, http://geant4.cern.ch/support/proc_mod_catalog/physics_lists/referencePL.shtml, Geant4, 2016.

[130] *Private communication with Nils Ruthmann*, 2014.

[131] *Private communication with Noel Dawe*, 2014.

[132] S Menke and G. L. Usai, “Proposal for Compactification of CaloCell Objects”, tech. rep. ATL-SOFT-2004-005, Geneva: CERN, 2004, URL: https://cds.cern.ch/record/808847.

[133] *Christian Schillo*, Thesis in preparation, Freiburg University.

[134] The ATLAS Collaboration, “Search for the Standard Model Higgs boson in H$\rightarrow$ $\tau\tau$ decays in proton-proton collisions with the ATLAS detector”, tech. rep. ATLAS-CONF-2012-160, Geneva: CERN, 2012, URL: https://cds.cern.ch/record/1493624.

[135] The ATLAS Collaboration, “Evidence for Higgs Boson Decays to the $\tau^+\tau^-$ Final State with the ATLAS Detector”, tech. rep. ATLAS-CONF-2013-108, Geneva: CERN, 2013, URL: https://cds.cern.ch/record/1632191.

[136] *ROOT*, http://root.cern.ch, 2015.

[137] P. Nason, “A New method for combining NLO QCD with shower Monte Carlo algorithms”, *JHEP* 0411 (2004) 040, DOI: 10.1088/1126-6708/2004/11/040, arXiv: hep-ph/0409146 [hep-ph].

[138] S. Frixione, P. Nason and C. Oleari, “Matching NLO QCD computations with Parton Shower simulations: the POWHEG method”, *JHEP* 0711 (2007) 070, DOI: 10.1088/1126-6708/2007/11/070, arXiv: 0709.2092 [hep-ph].

[139] S. Alioli et al., “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 [hep-ph].

[140] E. Bagnaschi et al., “Higgs production via gluon fusion in the POWHEG approach in the SM and in the MSSM”, *JHEP* 1202 (2012) 088, DOI: 10.1007/JHEP02(2012)088, arXiv: 1111.2854 [hep-ph].

[141] S. M.Ciccolini A.Denner, “Strong and electroweak corrections to the production of Higgs+2jets via weak interations at the LHC”, *Phys.Rev.Lett.* 99 (2007) 161803, DOI: 10.1103/PhysRevLett.99.161803, arXiv: 0707.0381 [hep-ph].

[142] K. Arnold et al., “VBFNLO: A parton level Monte Carlo process for processes with electroweak bosons”, *Comput.Phys.Commun.* 180 (2009) 1661–1670, DOI: 10.1016/j.cpc.2009.03.006, arXiv: 0811.4559 [hep-ph].

[143] A.Djouadi, M.Spira, P.Zerwas,
"Production of Higgs bosons in proton collideres: QCD corrections",
Phys.Lett.B 264 (1991) 440–446, DOI: 10.1016/0370-2693(91)90375-Z.

[144] S. Dawson, "Radiative corrections to Higgs bosons production",
Nucl.Phys.B 359 (1991) 283–300, DOI: 10.1016/0550-3213(91)90061-2.

[145] M. Spira et al., "Higgs boson production at the LHC", *Nucl.Phys.B* 453 (1995) 17–82,
DOI: 10.1016/0550-3213(95)00379-7, arXiv: 9504378 [hep-ph].

[146] R. Harlander and W. Kilgore,
"Next-to-next-to-leading order Higgs production at hadron colliders",
Phys.Rev.Lett 88 (2002) 201801, DOI: 10.1103/PhysRevLett.88.201801,
arXiv: 0201206 [hep-ph].

[147] C. Anastasiou and K. Melnikov, "Higgs boson production at hadron colliders in NNLO QCD",
Nucl.Phys. B 646 (2003) 220–256, DOI: 10.1016/S0550-3213(02)00837-4,
arXiv: 0207004 [hep-ph].

[148] V. Ravindran, J. Smith and W. van Neerven, "NNLO corrections to the total cross section for Higgs boson production in hadron-hadron collisions", *Nucl.Phys. B* 665 (2003) 325–366,
DOI: 10.1016/S0550-3213(03)00457-7, arXiv: 0302135 [hep-ph].

[149] S. Catani et al., "Soft-gluon resummation for Higgs boson production at hadron colliders",
JHEP 0307 (2003) 028, DOI: 10.1088/1126-6708/2003/07/028,
arXiv: 0306211 [hep-ph].

[150] U. Aglietti et al., "Two loop light fermion contribution to Higgs production and decays",
Phys.Lett.B 595 (2004) 432–441, DOI: 10.1016/j.physletb.2004.06.063,
arXiv: 0404071 [hep-ph].

[151] S. Actis et al.,
"NLO Electroweak Corrections to Higgs Boson Production at Hadron Colliders",
Phys.Lett.B 670 (2008) 12–17, DOI: 10.1016/j.physletb.2008.10.018,
arXiv: 0809.1301 [hep-ph].

[152] O. Brein, A. Djouadi and R. Harlander,
"NNLO QCD corrections to the Higgs-strahlung processes at hadron colliders",
Phys.Lett.B 579 (2004) 149–156, DOI: 10.1016/j.physletb.2003.10.112,
arXiv: 0307206 [hep-ph].

[153] M. L. Ciccolini, S. Dittmaier and M. Kraemer, "Electroweak Radiative Corrections to Associated WH and ZH Production at Hadron Colliders", *Phys.Rev. D* 68 (2003) 073003,
DOI: 10.1103/PhysRevD.68.073003, arXiv: 0306234 [hep-ph].

[154] M. Grazzini and H. Sargsyan,
"Heavy-quark mass effects in Higgs boson production at the LHC", *JHEP* 09 (2013) 129,
DOI: 10.1007/JHEP09(2013)129, arXiv: 1306.4581 [hep-ph].

[155] D. de Florian et al., "Higgs boson production at the LHC: transverse momentum resummation effects in the $H \to 2\gamma, H \to WW \to \ell\nu\ell\nu$ and $H \to ZZ \to 4\ell$ decay modes",
JHEP 06 (2012) 132, DOI: 10.1007/JHEP06(2012)132, arXiv: 1203.6321 [hep-ph].

[156] J. M. Campbell, R. K. Ellis and G. Zanderighi,
"Next-to-Leading order Higgs + 2 jet production via gluon fusion", *JHEP* 10 (2006) 028,
DOI: 10.1088/1126-6708/2006/10/028, arXiv: hep-ph/0608194 [hep-ph].

[157] A. Banfi et al., "Higgs and Z-boson production with a jet veto", *Phys. Rev. Lett.* 109 (2012) 202001, DOI: `10.1103/PhysRevLett.109.202001`, arXiv: `1206.4998 [hep-ph]`.

[158] A. Banfi, G. P. Salam and G. Zanderighi, "NLL+NNLO predictions for jet-veto efficiencies in Higgs-boson and Drell-Yan production", *JHEP* 06 (2012) 159, DOI: `10.1007/JHEP06(2012)159`, arXiv: `1203.5773 [hep-ph]`.

[159] A. Banfi, P. F. Monni and G. Zanderighi, "Quark masses in Higgs production with a jet veto", *JHEP* 01 (2014) 097, DOI: `10.1007/JHEP01(2014)097`, arXiv: `1308.4634 [hep-ph]`.

[160] M. Ciccolini, A. Denner and S. Dittmaier, "Electroweak and QCD corrections to Higgs production via vector-boson fusion at the LHC", *Phys. Rev.* D 77 (2008) 013002, DOI: `10.1103/PhysRevD.77.013002`, arXiv: `0710.4749 [hep-ph]`.

[161] G. Marchesini et al., "HERWIG 5.1- a Monte Carlo event generator for simulating hadron emission reactions with interfering gluons", *Computer Physics Communications* 67 (1992) 465, DOI: `10.1016/0010-4655(92)90055-4`.

[162] M. Mangano et al., "ALPGEN, a generator for hard multiparton processes in hadronic collisions", *JHEP* 07 (2003) 001, DOI: `10.1088/1126-6708/2003/07/001`, arXiv: `hep-ph/0206293`.

[163] M. Mangano et al., "Multijet matrix elements and shower evolution in hadronic collisions: $Wb\bar{b} + n$ jets as a case study", *Nucl. Phys.* B 632 (2002) 343, DOI: `10.1016/S0550-3213(02)00249-3`, arXiv: `hep-ph/0108069`.

[164] T. Binoth et al., "Gluon-induced W-boson pair production at the LHC", *JHEP* 12 (2006) 046, DOI: `10.1088/1126-6708/2006/12/046`, arXiv: `hep-ph/0611170`.

[165] B. P. Kersevan and E. Richter-Was, "The Monte Carlo event generator AcerMC versions 2.0 to 3.8 with interfaces to PYTHIA 6.4, HERWIG 6.5 and ARIADNE 4.1", *Comput. Phys. Commun.* 184 (2013) 919, DOI: `10.1016/j.cpc.2012.10.032`, arXiv: `hep-ph/0405247 [hep-ph]`.

[166] The ATLAS Collaboration, "Summary of ATLAS PYTHIA 8 tunes", tech. rep., 2012, URL: `http://cds.cern.ch/record/1474107`.

[167] K. Edmonds et al., "The fast ATLAS track simulation (FATRAS)" (2008).

[168] The ATLAS Collaboration, "The simulation principle and performance of the ATLAS fast calorimeter simulation FastCaloSim" (2010).

[169] *ATLFAST-II for hadronic taus*, ht-tps://indico.cern.ch/event/238609/session/3/contribution/7/attachments/401697/558525/Corfu_EAI.pdf, internal communication, 2013.

[170] *Embedding and AFII for $H \to \tau\tau$ analysis*, https://indico.cern.ch/event/239856/session/2/contribution/6/attachments/404825/562594/TauWorkshop2013.pdf, internal communication, 2013.

[171] *ATLFAST-II studies for the $H \to \tau(lep)\tau(had)$ channel*, https://indico.cern.ch/event/239856/session/7/contribution/20/attachments/404817/562580/ohman_atlfast_update_corfu_2013-04-03.pdf, internal communication, 2013.

[172] The ATLAS Collaboration, "Improved luminosity determination in pp collisions at $\sqrt{s}$ =7 TeV using the ATLAS detector at the LHC" (2013), DOI: `10.1140/epjc/s10052-013-2518-3`, arXiv: `1302.4393 [hep-ex]`.

[173] The ATLAS Collaboration,
"Electron and photon energy calibration with the ATLAS detector using LHC Run 1 data",
Eur. Phys. J. C74.10 (2014) 3071, DOI: 10.1140/epjc/s10052-014-3071-4,
arXiv: 1407.5063 [hep-ex].

[174] The ATLAS Collaboration, "Jet energy resolution in proton-proton collisions at $\sqrt{s}$=7 TeV recorded in 2010 with the ATLAS detector",
The European Physical Journal C 73.3, 2306 (2013),
DOI: 10.1140/epjc/s10052-013-2306-0.

[175] I. W. Stewart and F. J. Tackmann,
"Theory uncertainties for Higgs mass and other searches using jet bins",
Phys. Rev. D 85 (3 2012) 034011, DOI: 10.1103/PhysRevD.85.034011.

[176] K. Cranmer et al.,
"HistFactory: A tool for creating statistical models for use with RooFit and RooStats",
tech. rep. CERN-OPEN-2012-016, New York: New York U., 2012,
URL: https://cds.cern.ch/record/1456844.

[177] R. Barlow, "Extended maximum likelihood",
Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment 297.3 (1990) 496 –506,
DOI: 10.1016/0168-9002(90)91334-8.

[178] F. James and M. Roos, "Minuit - a system for function minimization and analysis of the parameter errors and correlations", *Computer Physics Communications* 10.6 (1975) 343 –367,
DOI: 10.1016/0010-4655(75)90039-9.

[179] G. Cowan et al., "Asymptotic formulae for likelihood-based tests of new physics",
The European Physical Journal C 71.2, 1554 (2011),
DOI: 10.1140/epjc/s10052-011-1554-0.

[180] A. L. Read, "Presentation of search results: the CL_s technique",
Journal of Physics G: Nuclear and Particle Physics 28.10 (2002) 2693,
URL: http://stacks.iop.org/0954-3899/28/i=10/a=313.

[181] The CMS Collaboration,
"Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons",
Journal of High Energy Physics 2014.5, 104 (2014), DOI: 10.1007/JHEP05(2014)104.

[182] The ATLAS and CMS Collaboration, "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", tech. rep. ATLAS-CONF-2015-044,
Geneva: CERN, 2015, URL: http://cds.cern.ch/record/2052552.

[183] B. Efron and C. Stein, "The Jackknife Estimate of Variance",
The Annals of Statistics 9.3 (1981) 586–596,
URL: http://www.jstor.org/stable/2240822.

[184] G. Bohm and G. Zech, *Introduction to statistics and data analysis for physicists; 2nd rev. ed.*
Verlag Deutsches Elektronen-Synchrotron, Hamburg, ISBN 978-3-935702-88-1, 2014.

[185] The 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.* D89.1 (2014) 012003,
DOI: 10.1103/PhysRevD.89.012003, arXiv: 1310.3687 [hep-ex].

[186] The CMS Collaboration,
"Evidence for the direct decay of the 125 GeV Higgs boson to fermions",
Nature Phys. 10 (2014) 557–560, DOI: 10.1038/nphys3005, arXiv: 1401.6527 [hep-ex].

[187] S. Berge, W. Bernreuther and S. Kirchner,
"Prospects of constraining the Higgs bosons CP nature in the tau decay channel at the LHC",
Phys. Rev. D92 (2015) 096012, DOI: 10.1103/PhysRevD.92.096012,
arXiv: 1510.03850 [hep-ph].

[188] The ATLAS Collaboration,
"Test of CP Invariance in vector-boson fusion production of the Higgs boson using the Optimal Observable method in the ditau decay channel with the ATLAS detector" (2016),
arXiv: 1602.04516 [hep-ex].

[189] The ATLAS Collaboration, "Constraints on new phenomena via Higgs boson couplings and invisible decays with the ATLAS detector", *JHEP* 11 (2015) 206,
DOI: 10.1007/JHEP11(2015)206, arXiv: 1509.00672 [hep-ex].

[190] *Signal Extraction:*$H \to \tau_\ell \tau_h$,
https://indico.cern.ch/event/286331/contribution/2/attachments/531586/733080/Talk.pdf,
internal communication, 2013.

[191] J. Liebal et al., "Modelling $Z \to \tau\tau$ Processes with τ-Embedded $Z \to \mu\mu$ Data",
tech. rep. ATL-COM-PHYS-2014-543, Geneva: CERN, 2014,
URL: https://cds.cern.ch/record/1704931.

[192] G A Stewart et al, "ATLAS Job Transforms: A Data Driven Workflow Engine",
J. Phys:Conf Ser. 513 (2014) 032094,
URL: http://stacks.iop.org/1742-6596/513/i=3/a=032094.

Acknowledgements

Many people helped me to realise this work and I would like to express my deep gratitude for their support. First, I would like to express my deep gratitude to Prof. Norbert Wermes for his supervision of my thesis. He gave me the opportunity to work on my thesis in a lively time for particle physics and provided a lot of support. Thanks to him I was also able to travel to different places and get into contact with many interesting people.
I would like to thank my second examiner Prof. Jochen Dingfelder. During my years at the Bonn University I worked closely with members of his research groups. I appreciate the exchange in frequent meetings together with his group.
I would also like to thank Prof. Herbert Dreiner and Prof. Achim Rosch for being members of the thesis jury. I am grateful to Prof. Michael Gütschow for accepting to be the external referee of my thesis. Special thanks go to Dr. Jürgen Kroseberg, who supervised me through my PhD.
The Bonn Cologne Graduate School is acknowledged for the financial support of my doctoral thesis. It offered me the possibility to visit summer schools and interesting lectures.
Many people helped me to finalise this thesis. I thank all the people of the Physikalische Institute whom I worked with. A big thanks goes to my former Bonn colleagues Dr. Jana Kraus and Dr. Thomas Schwindt. It was a pleasure to work with them on the $H \to \tau\tau$ analysis. Due to their long experience they were a source of advice and provided help in several ways.
I would also like to thank members of the ATLAS Collaboration. In particular, I would like to thank Prof. Dr. Stan Lai and Dr. Michel Janus for the various fruitful discussions. I will keep their time as ATLAS subconvenors in good memory. Special thanks goes also to Christian Schillo with whom I worked on the embedding technique.
I was fortunate to share an office with David Hohn and Birgit Stapf who created a joyful atmosphere in the office. We had fun on so many occasions. I thank you for all the memories we created together. Of course, I thank you also for correcting my thesis.
I wish to give thanks to my parents and my *parents-in-law* whom I can rely on in all situations.
Finally, I would like to thank Ronnie for his incredible love and support. His patience and encouragement helped me to finalise my doctoral thesis.