IEKP-KA/2014-09

Karlsruhe Institute of Technology

Position Resolution and Upgrade of the CMS Pixel Detector and Search for the Higgs Boson in the $\tau^+\tau^-$ Final State

Zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
von der Fakultät für Physik des Karlsruher Instituts
für Technologie genehmigte

Dissertation

von

Dipl.-Phys. Armin Burgmeier
aus Karlsruhe

Tag der mündlichen Prüfung: 13. Juni 2014

Referent: Prof. Dr. G. Quast
Institut für Experimentelle Kernphysik

Korreferent: Prof. Dr. U. Husemann
Institut für Experimentelle Kernphysik

Contents

Introduction

The field of particle physics strives to understand the fundamental laws of nature. It describes the fundamental particles that all matter is made of and the interactions between them. Progress in the field is achieved via interplay of theory and experiment: experimental results give rise to corrections and generalizations of existing theories, which are then probed again by new experiments. The ultimate goal is to find a "theory of everything", a single theory which is able to describe all interactions in the universe, from the motion of planets to subatomic particle reactions.

The Standard Model of particle physics is an overwhelmingly successful theory. It describes three of the four known fundamental forces in a unified way and with extremely high precision. Numerous experiments have probed the Standard Model over the last decades, without finding significant discrepancies. The latest breakthrough result was the discovery of the Higgs boson, announced on July 4, 2012. The Higgs boson is a particle predicted by the Standard Model, however was not observed experimentally until very recently. This thesis focuses on the experimental search for the Higgs boson, and the comparison of the measured properties of the newly-discovered boson with the prediction of the Standard Model.

In Chapter 1, the general concepts of the Standard Model are discussed, with emphasis on the mechanism of electroweak symmetry breaking. This mechanism explains the masses of the fundamental particles in a mathematically consistent way, and ultimately leads to a prediction of the Higgs boson, a physical particle that can be created and studied in experiments.

The experimental devices for fundamental research have grown over the decades from small laboratory instruments to beamlines of many kilometers with detectors as tall as multiple-story buildings. Since creation of heavy particles requires high energies, the facilities have to be large in order to enable such energies. Only international collaborations of many hundreds or thousands of physicists can build and operate these machines. The Higgs boson was found at the Large Hadron Collider (LHC) at CERN, a 27 km ring accelerator. Chapter 2 describes both the collider and the Compact Muon Solenoid (CMS), one of the four particle detectors and which has recorded the data that are analyzed in this thesis.

Modern data analysis relies heavily on automated tools to process the recorded collision data. Furthermore, Monte Carlo techniques are used to make predictions according to the Standard Model, since analytical calculations cannot be performed. Results from the Monte Carlo simulation are compared to the recorded data to judge the compatibility of the data and the theory. In Chapter 3, the software packages used to obtain the results in this thesis are described, and the algorithms for reconstruction of particles based on the recorded CMS data are discussed.

CMS consists of several individual subdetectors. The silicon pixel detector is situated closest to the collision point, since it provides a very high spatial resolution of particle trajectories crossing the detector. However, this implies that the rate of particles per unit area is highest in the silicon pixel detector, leading to severe radiation damage over

time. Therefore, this subdetector must be replaced after several years of operation. At the end of 2016, it is planned to install an improved detector in CMS, benefiting from lessons learned while manufacturing and operating the first one. Prototypes of the new readout chip for the upgraded pixel detector have been tested in the laboratory and in beam tests. An important measure of performance is the spatial resolution in r-ϕ direction. The resolution affects both the measurement of the transverse momentum of a particle and the reconstruction of the impact parameter with respect to the primary interaction point. A good impact parameter resolution is especially helpful for reconstructing b-quarks and tau leptons. In Chapter 4, test beam measurements of the position resolution of the current and the upgraded readout chip are presented, and compared to direct measurements in CMS and to a Monte Carlo simulation.

The Higgs boson is an unstable particle that decays immediately after it was created. Its mass was measured to be around 125 GeV. At this mass, according to the Standard Model, decays to many different pairs of particles are possible, with different probabilities for each individual decay to occur. One promising channel is the decay to two tau leptons. Other channels have a higher discovery sensitivity, however the di-tau channel allows to directly probe the coupling of the Higgs boson to fermions. In Chapter 5, an overview of the CMS analysis in this channel is given.

Just like the Higgs boson, the Z boson can decay into two tau leptons as well. Since the production cross-section for Z bosons is several orders of magnitude above that for the Higgs boson, there will be many more tau leptons from Z decays than from Higgs decays. The two cannot be separated from each other easily because the particles in the final state are exactly the same, leading to the same signature in the detector. Therefore, knowing exactly the expected number of Z boson decays is crucial, so that any excess above that expectation can be attributed to the Higgs boson. This number can be obtained from Monte Carlo simulations, however, there are several sources of systematic uncertainties coming with it. In Chapter 6, the "tau embedding method" is presented, which allows to estimate the contribution from Z decays from the recorded data sample itself. Within the scope of this thesis, the method has been enhanced to reduce systematic uncertainties and to prepare it for more challenging conditions in future CMS data taking.

The tau lepton itself is not a stable particle either, but it decays into lighter leptons or hadrons. Furthermore, the Higgs boson can also be produced not only by itself, but together with other particles such as W bosons and Z bosons. This leads to many different possible final states which are analyzed separately, and then the individual results are combined on a statistical basis. Chapter 7 presents in detail the analysis in the channel where the Higgs boson is produced in association with a W boson, and both tau leptons decay into hadrons. The analysis was driven forward significantly by the work performed for this thesis.

1 The Standard Model of Particle Physics

In particle physics, physicists strive to understand the most basic building blocks of matter and the interactions between them. The gold foil experiment by Ernest Rutherford in 1909 [1] marks the start of modern particle physics. It yielded insights into the substructure of the atom, namely that it consists of a dense, charged core and surrounding electrons.

Since then, the principle of the experimental methods only changed marginally. All modern particle accelerators still perform scattering experiments with various particles, even if most machines collide two beams instead of one beam and a fixed target. This way, the composition of nuclei was probed, inelastic scattering experiments revealed the substructure of the proton, and myriads of new composite and fundamental particles have been discovered.

The results obtained in such scattering experiments are used as an input to formulate theories which not only explain the results of past experiments but which are also able to predict the outcome of future experiments. New experiments eventually verify or falsify existing theories. *Quantum field theory* (QFT), the framework modern particle physics is based on, is especially remarkable in this regard as it has been verified to an unprecedented accuracy.

There are four different fundamental forces between particles known to date. Three of them can be successfully described by QFT. Together, they are referred to as the *Standard Model of Particle Physics*, or Standard Model in short. In gauge theories, interactions are mediated via force carrier particles, also called gauge bosons. They are summarized in Table 1.1. Gravity is the only force which could not yet be consistently formulated as a QFT. It is described by the theory of *General Relativity* which does not take quantum mechanical effects into account.

While forces are mediated via gauge bosons, all matter consists of fermions. These particles come in three generations where particles in different generations only differ in their masses but have exactly the same properties otherwise. Table 1.2 summarizes the fermions. For each fermion, there is also an anti-fermion with the same properties as the fermion but opposite couplings to the Standard Model interactions.

In the following, a brief introduction to the ideas (Section 1.1) and the essential results (Sections 1.2, 1.3, 1.4 and 1.5) of quantum field theory are given, especially in the electroweak sector of the Standard Model which is most relevant for the remainder of this thesis. Section 1.6 presents experimental results and Section 1.7 concludes by briefly discussing the shortcomings of the current theory. A more complete introduction into QFT can be found for example in [2].

1.1 Quantum Field Theory

The basic idea of quantum field theory comes from classical field theory. In classical field theory, a dynamical system minimizes the action S as it propagates from one state to

Table 1.1: The four fundamental interactions. For each force also an example of an interaction based on the corresponding force is given. The graviton is not part of the Standard Model and has not yet been observed experimentally. However, if gravity can be described by a quantum field theory the corresponding force carrier would be called "Graviton".

Force	Carrier	Mass [GeV]	Range	Example
Strong	8 Gluons	0	10^{-15} m	Holding together nuclei
Weak	$W^{\pm}$ boson	80.4	10^{-18} m	Radioactive β decay
	Z^0 boson	91.2		
Electromagnetic	Photon	0	∞	Radio communication
Gravitation	(Graviton)	0	∞	Motion of planets

Table 1.2: The various fermions are categorized into leptons (top) and quarks (bottom). The quarks interact strongly, electromagnetically and weakly. They form mesons and baryons such as pions, protons or neutrons. The leptons only interact electromagnetically (if charged) and weakly. The second and third generation fermions are so heavy that they eventually decay to the first generation ones, except for the neutrinos which are approximately massless. Charge, weak isospin, and color denote the coupling to the electromagnetic interaction, the weak interaction, or the strong interaction, respectively. Only left-handed fermions have weak isospin; right-handed fermions do not interact weakly. This is discussed in detail in Section 1.3.2.

	Generation			Charge	Weak Isospin	Color
	1	2	3			
Leptons	$\begin{pmatrix} \nu_e \\ e^- \end{pmatrix}$	$\begin{pmatrix} \nu_\mu \\ \mu^- \end{pmatrix}$	$\begin{pmatrix} \nu_\tau \\ \tau^- \end{pmatrix}$	0 $-e$	$+\frac{1}{2}$ $-\frac{1}{2}, 0$	$-$ $-$
Quarks	$\begin{pmatrix} u \\ d \end{pmatrix}$	$\begin{pmatrix} c \\ s \end{pmatrix}$	$\begin{pmatrix} t \\ b \end{pmatrix}$	$+\frac{2}{3}e$ $-\frac{1}{3}e$	$+\frac{1}{2}, 0$ $-\frac{1}{2}, 0$	r, g, b r, g, b

another. S can be expressed as an integral of the Lagrangian L or the Lagrangian density $\mathcal{L}$:

$$S = \int \mathcal{L}\left(\phi, \partial_\mu \phi\right) \mathrm{d}^4 x \,. \tag{1.1}$$

The Lagrangian depends on one or more fields ϕ and their derivatives $\partial_\mu \phi$, usually composed of a kinetic term, a (rest) mass term and interaction terms. From the principle of least action, the Euler-Lagrange equations are derived:

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

In classical field theory, the fields are real or complex functions. In QFT, the fields are replaced by operators which obey the same commutation relations as the classical variables. This process is called *second quantization*. The operators are applied to states of the system which are specified by the number of particles with a certain momentum p (and spin in case of non-scalar fields). The ensemble of such states is called "Fock space".

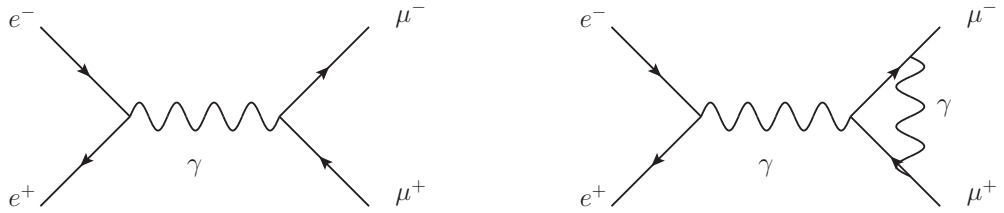

Figure 1.1: Example Feynman diagrams for the $e^+e^- \to \mu^+\mu^-$ process (e^+e^- annihilation). The first diagram shows the leading order contribution. The second diagram shows one of many next-to-leading order contributions where one additional photon is exchanged between the final state particles. Other second order contributions include photon exchange of the initial state particles or a fermion loop in the photon propagator.

As with the quantum mechanical harmonic oscillator, there exist ladder operators $a_p^\dagger$ and a_p, which create or destroy a particle with momentum p, respectively. The field $\phi(x)$ can be written in terms of a_p and $a_p^\dagger$ as a Fourier integral. For example, for a scalar field, it is given by

$$\phi(x) = \int \frac{\mathrm{d}^3 p}{(2\pi)^3} \frac{1}{\sqrt{2p^0}} \left(a_p \mathrm{e}^{ip_\mu x^\mu} + a_p^\dagger \mathrm{e}^{-ip_\mu x^\mu} \right) \tag{1.3}$$

with $p^0 = \sqrt{\vec{p}^{\,2} + m^2}$.

Physical Lagrange densities should yield the known differential equations for free particles, that is the Klein-Gordon equation for scalar fields, the Dirac equation for spin-$\frac{1}{2}$ fields and the Maxwell equations for massless spin-1 fields.

In order to calculate an observable quantity, such as a cross section or a decay width, the *transition amplitude* $\langle \mathrm{f}|H_{\mathrm{int}}|\mathrm{i}\rangle$ must be computed where $|\mathrm{f}\rangle$ and $|\mathrm{i}\rangle$ denote the initial and final states, respectively, and H_{int} is the interaction Hamiltonian which can be derived from the Lagrangian density. Eventually, observables are proportional to the magnitude squared of the transition amplitude (also known as *matrix element*).

Since, for interacting processes, such matrix elements cannot be computed analytically, one resorts to perturbation theory. Coupling constants such as the electric charge e or the weak or strong coupling constants α_W or α_S for the weak or strong interactions, respectively, are used as the perturbation parameter. The procedure of deriving this perturbation series is highly nontrivial and at this point only a reference to [2] shall be given.

Every term in the perturbation series can be assigned a schematic drawing called a *Feynman diagram* [3] which consists of propagators (possibly virtual particles with a specific momentum and spin) and vertices (interactions between particles). The translation from diagram elements to mathematical terms are known as *Feynman rules*. Propagators not connected to a vertex represent external particles in the initial or final state.

Figure 1.1 shows example Feynman diagrams for the $e^+e^- \to \mu^+\mu^-$ process of quantum electrodynamics (QED). Every vertex contributes a factor of $\sqrt{\alpha}$ to the matrix element, and therefore diagrams with many vertices are higher order in perturbation theory. This explains why the diagram on the left is the leading order diagram of the process and the one on the right is a higher order diagram with lower contribution to the matrix element.

For higher order effects, it usually happens that loops occur in Feynman diagrams. In this case it must be integrated over all possible momenta of the particles within the loop, which leads to divergences. In order to circumvent such divergences and to obtain finite numbers for observables, a procedure called *renormalization* must be applied. Renormalizability is a feature of a particular quantum field theory, and in fact the reason why no quantum field theory can be formulated for gravity is that such a theory would not be renormalizable. However, the mathematical concepts behind renormalizability are again beyond the scope of this thesis.

1.2 Quantum Electrodynamics

Quantum electrodynamics is the theory which describes all electromagnetic effects and interactions. It is the simplest of the three interactions of the Standard Model but its basic ideas are also applicable to the weak and strong interactions.

As a quantum field theory, QED is fully characterized by its Lagrangian density. The Lagrangian density for free fermions is given by

$$\mathcal{L}_{\text{Dirac}} = \bar{\psi}\left(i\partial\!\!\!/ - m\right)\psi\,, \tag{1.4}$$

where ψ is a Dirac spinor field and $\bar{\psi} = \psi^\dagger\gamma^0$. The motivation for this form of the Lagrangian density is that plugging it into the Euler-Lagrange equations leads to the well-known Dirac equation for spin-$\frac{1}{2}$ fermions.

Classical electrodynamics is a gauge theory, which means that the four-potential A_μ can be transformed as

$$A_\mu \to A'_\mu = A_\mu - \partial_\mu \Lambda\left(x\right) \tag{1.5}$$

with an arbitrary scalar field Λ. This transformation has no effect on the observable quantities $\vec{E}$ and $\vec{B}$: they are invariant under *local* gauge transformations.

This property motivates a similar invariance in quantum electrodynamics. For the spinor field ψ, a *global* phase transformation

$$\psi \to \psi' = \mathrm{e}^{i\alpha}\psi \tag{1.6}$$

vanishes when applied in the Lagrangian density. However, in classical electrodynamics, Λ may depend on space-time. If $\alpha = \alpha\left(x\right)$, i.e. it is a *local* phase transformation, then an additional term appears because of the derivative in the Lagrangian density. This additional term is nonzero and therefore breaks the gauge invariance.

In order to restore gauge invariance, the derivative ∂_μ is substituted by the *covariant derivative*,

$$D_\mu = \partial_\mu + ieA_\mu\left(x\right)\,, \tag{1.7}$$

where A_μ is a new vector field. Its transformation property under local gauge transformations can easily be derived by requiring the Lagrangian density to be gauge-invariant:

$$A_\mu \to A'_\mu = A_\mu - \frac{1}{e}\partial_\mu\alpha\left(x\right)\,. \tag{1.8}$$

This transformation is astonishingly similar to the gauge transformation of classical electrodynamics, Equation 1.5. At this point, it is easy to identify the field A_μ as the photon field.

In other words, postulation of local gauge invariance introduces the photon into the theory. The phase transformation, Equations 1.6 and 1.8, is the symmetry transformation of the U (1) group. Therefore, the theory is said to be invariant under U (1) transformations.

With the covariant derivative in place, an additional term in the Lagrangian density shows up, $e\bar{\psi}\gamma^\mu\psi A_\mu$, which describes the coupling of fermions to the field A_μ. However, the Lagrangian density needs to be extended further to fully account for the photon field:

$$\mathcal{L}_{\mathrm{QED}} = \bar{\psi}\left(i\partial\!\!\!/ - m\right)\psi - \frac{1}{4}F_{\mu\nu}F^{\mu\nu} - e\bar{\psi}\gamma^\mu\psi A_\mu\,, \tag{1.9}$$

where $F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$ is the electromagnetic *field strength tensor*. The $\frac{1}{4}F_{\mu\nu}F^{\mu\nu}$ term is the kinetic term for the photon field. A mass term along the lines of $\frac{1}{2}m^2 A_\mu A^\mu$ must not be added because it would spoil gauge invariance again. This is in perfect agreement with the observation that the photon is a massless particle.

Equation 1.9 is the full Lagrangian of QED. The Euler-Lagrange equations for A_μ lead to the inhomogeneous Maxwell equations, $\partial_\mu F^{\mu\nu} = e\bar{\psi}\gamma^\nu\psi = ej^\nu$.

1.3 Electroweak Unification

As with QED, the theory of the weak interaction is a quantum field theory. In the first part of this section, the differences between the two interactions are discussed. In the second part, gauge invariance is postulated in order to obtain the force carriers of the weak interaction, the W and Z bosons. It turns out that, in order to describe experimental results consistently, the electromagnetic and weak interactions need to be unified into a single electroweak interaction.

1.3.1 Weak Isospin and Parity Violation

The weak interaction can turn charged leptons to neutrinos or up-type quarks to down-type quarks and vice versa. This motivates a spin-like formalism where particles are arranged in doublets of Dirac fields,

$$\psi = \begin{pmatrix} \psi_\nu\,(x) \\ \psi_e\,(x) \end{pmatrix}_{\mathrm{L}}\,. \tag{1.10}$$

Particles in such a doublet are said to have *weak isospin* $T = \frac{1}{2}$ where the third component of weak isospin is $T_3 = +\frac{1}{2}$ for neutrinos and up-type quarks and $T_3 = -\frac{1}{2}$ for charged leptons and down-type quarks. T_3 can be seen as the "charge" of the weak interaction.

The index L in Equation 1.10 denotes a doublet with left-handed chirality. A fermion field can be decomposed into a left-handed and right-handed component with the projection operators $(1 \pm \gamma^5)/2$. The famous Wu experiment [4] has shown that the weak interaction only couples to left-handed fermions. Right-handed fermions form a singlet with respect to the weak interaction with $T = T_3 = 0$. Therefore, right-handed charged leptons only interact electromagnetically and right-handed neutrinos are not observed at all. The different coupling of left-handed and right-handed fermions is known as *parity violation* of the weak interaction.

1.3.2 The Glashow-Weinberg-Salam Model

For the weak interaction, similar ideas as for QED can be applied: the Lagrangian density for this fermion doublet, $\bar{\psi}\left(i\partial\!\!\!/ - m\right)\psi$, should be invariant under local gauge transformations. In contrast to the QED case, there are now three linearly independent ways to transform the phase for a doublet of complex fields, or more generally the SU (2) symmetry group. They are given by the Pauli matrices $\vec{\sigma}$, which are therefore called the generators of the SU (2) group. The transformation of the fields is given by

$$\psi \to \psi' = \mathrm{e}^{\frac{i}{2}\vec{\sigma}\cdot\vec{\alpha}(x)}\psi\,. \tag{1.11}$$

Again, to restore invariance of the Lagrangian density under this transformation, a covariant derivative is introduced:

$$D_\mu = \partial_\mu - \frac{i}{2}g\vec{\sigma}\cdot\vec{W}_\mu\,, \tag{1.12}$$

where g is a coupling constant and $\vec{W}_\mu$ are three new vector fields. It is now tempting to identify these as the W^+, the W^- and the Z^0, the carriers of the weak force found experimentally. However, it has been observed that the Z boson couples differently to neutrinos (or up-type quarks) than to charged leptons (or down-type quarks), for example by measuring the branching fractions of the Z boson.

To solve this problem, both the electromagnetic and the weak interaction must be considered together, leading to electroweak unification, or the Glashow-Weinberg-Salam (GWS) model [5, 6, 7]. In this case, the covariant derivative becomes

$$D_\mu = \partial_\mu - \frac{1}{2}ig\vec{\sigma}\cdot\vec{W}_\mu - \frac{1}{2}ig'B_\mu\,, \tag{1.13}$$

where the last term comes from the U (1) symmetry of QED. It is formally the same as Equation 1.7 where e and A_μ have been renamed to $\frac{g'}{2}$ and B_μ for reasons that will become obvious soon. The covariant derivative is plugged into the kinetic Lagrangian density term for left-handed fermion fields to obtain the couplings of the fields to the gauge bosons:

$$\begin{aligned} i\bar{\psi}D_\mu\gamma^\mu\psi &= \bar{\psi}\left(i\partial_\mu + \frac{1}{2}g\vec{\sigma}\cdot\vec{W}_\mu + \frac{1}{2}g'B_\mu\right)\gamma^\mu\psi & (1.14)\\ &= \bar{\psi}\left(i\partial_\mu + \frac{1}{2}\begin{pmatrix} g'B_\mu + gW_\mu^3 & gW_\mu^1 - igW_\mu^2 \\ gW_\mu^1 + igW_\mu^2 & g'B_\mu - gW_\mu^3 \end{pmatrix}\right)\gamma^\mu\psi\,. & (1.15)\end{aligned}$$

What can be learned from Equation 1.15 is that what actually couples to the fermion fields are not the B_μ and $\vec{W}_\mu$ fields, but linear combinations of them:

$$\begin{aligned} W_\mu^+ &= \frac{W_\mu^1 - iW_\mu^2}{\sqrt{2}} & (1.16)\\ W_\mu^- &= \frac{W_\mu^1 + iW_\mu^2}{\sqrt{2}} & (1.17)\\ A_\mu &= \frac{g'B_\mu + gW_\mu^3}{\sqrt{g^2 + g'^2}} & (1.18)\\ Z_\mu^0 &= \frac{-g'B_\mu + gW_\mu^3}{\sqrt{g^2 + g'^2}} & (1.19)\end{aligned}$$

The mixing of the B and W^3 fields explains how the coupling of the Z boson also has an electromagnetic component, and therefore also depends on the electric charge. The mixing can also be parameterized by the electroweak mixing angle, called *Weinberg angle*:

$$\tan \vartheta_W = \frac{g'}{g} . \tag{1.20}$$

Experimentally, it can be determined by cross section measurements of elastic neutrino-nucleon scattering [8] or from couplings measurements of the Z boson [9].

However, there is still one problem remaining. It has been verified experimentally that the $W^{\pm}$ and the Z^0 particles are not massless [10, 11, 12, 13]. Introducing a mass term for the gauge fields would spoil the gauge invariance, though. This inconsistency can be explained theoretically by introducing the Higgs mechanism, which is discussed in the following section.

1.4 The Higgs Mechanism

The idea is that instead of adding a mass term for the gauge bosons directly into the Lagrangian density, new fields are added. The mass terms arise from the interaction of the gauge boson fields with the new fields. This mechanism was first proposed by P. Higgs and others [14, 15, 16].

Since three masses need to be generated, the new field needs at least three degrees of freedom. The simplest way to do this is to introduce a doublet of complex scalar fields,

$$\Phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} . \tag{1.21}$$

The Lagrangian density is extended by terms which are invariant under $\mathrm{SU}(2) \otimes \mathrm{U}(1)$ transformations:

$$\mathcal{L}_\Phi = (D_\mu \Phi)^2 \underbrace{-\mu^2 \left|\Phi^\dagger \Phi\right| - \lambda \left|\Phi^\dagger \Phi\right|^2}_{-V(\Phi)} , \tag{1.22}$$

where μ has the dimensions of a mass and λ is a dimensionless constant. There cannot be any Φ^6 or higher terms because that would lead to a non-renormalizable theory (e.g. [17]). Therefore, in order to have a stable groud state, λ must be positive.

Figure 1.2 shows the form of the potential for $\mu^2 > 0$ and $\mu^2 < 0$. In the second case, the minimum of the potential is not at the origin but at

$$|\langle \Phi_0 \rangle| = \frac{v}{\sqrt{2}} = \sqrt{\frac{-\mu^2}{2\lambda}} . \tag{1.23}$$

The newly introduced parameter v is called the *vacuum expectation value*. Since the Lagrangian density is invariant under $\mathrm{SU}(2) \otimes \mathrm{U}(1)$ transformations, the vacuum expectation value will not change under such a transformation. Therefore, it is possible to find a transformation such that

$$\langle \Phi_0 \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} . \tag{1.24}$$

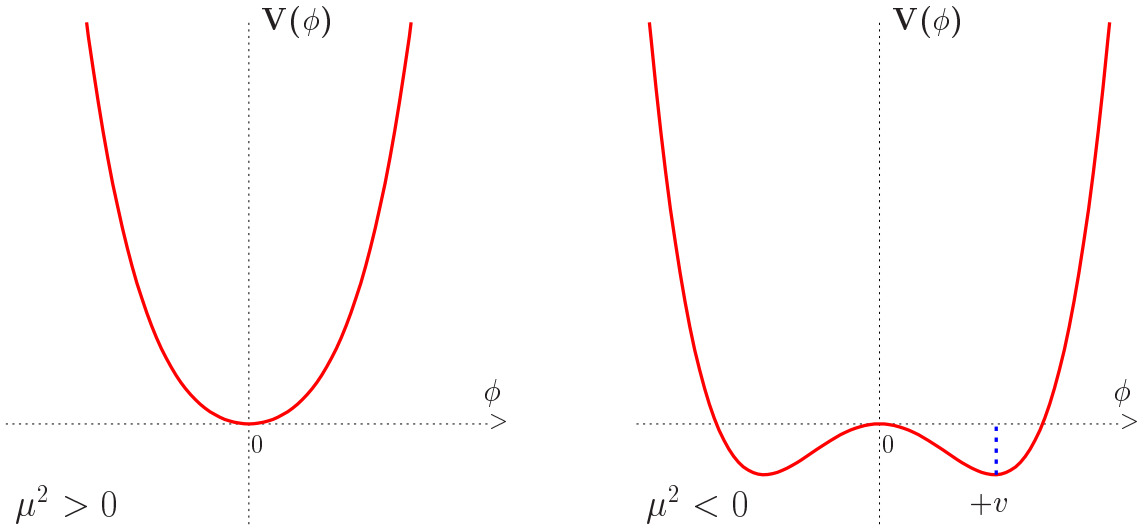

Figure 1.2: Potential of the field Φ for $\mu^2 > 0$ (left) and $\mu^2 < 0$ (right). In the second case, there is more than one ground state and the system chooses one at random. From [18].

The system will spontaneously fall into one of the many possible ground states. In the ground state, the system is located in a minimum of the potential where it is no longer invariant under SU (2) transformations. Figure 1.2 visualizes this: the potential does not change when one rotates the coordinate system around the y axis. However, if the center of the rotation is located in the minimum of the potential, then, in the $\mu^2 < 0$ case, the curve does not stay invariant. This phenomenon is called *spontaneous symmetry breaking*. The theory remains unbroken under U (1), however. This will be the reason why the photon remains massless in the following.

The gauge boson masses arise from the kinetic term in the Lagrangian density when it is evaluated at the potential minimum:

$$(D_\mu \Phi)^2 = \Phi^\dagger \left(\partial_\mu + ig\vec{W}_\mu \cdot \frac{\vec{\sigma}}{2} + \frac{1}{2} ig' B_\mu\right) \left(\partial^\mu - ig\vec{W}^\mu \cdot \frac{\vec{\sigma}}{2} - \frac{1}{2} ig' B^\mu\right) \Phi \quad (1.25)$$

$$= \frac{1}{2} \cdot \frac{1}{4} v^2 \left(\left|gW_\mu^1 - igW_\mu^2\right|^2 + \left|g'B_\mu - gW_\mu^3\right|^2\right) + \ldots , \quad (1.26)$$

where

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + H(x) \end{pmatrix} \quad (1.27)$$

has been expanded around the vacuum. The real field $H(x)$ is called the *Higgs* field. It results as a remaining degree of freedom from the original complex doublet and, by construction, vanishes at the minimum of the potential.

In Equation 1.26, terms containing H or ∂_μ have been omitted. They lead to the kinetic term for the Higgs field and to interaction terms between gauge bosons and the Higgs field. What remains are mass terms for the W bosons and the Z boson (note that $\left|gW_\mu^1 - igW_\mu^2\right|$ $= \left|gW_\mu^1 + igW_\mu^2\right|$, and therefore the mass term in Equation 1.26 accounts for both the W^+

and the W^-). There is no mass term for the photon field, so it remains massless. The masses can be directly read from Equation 1.26:

$$m_W = g\frac{v}{2}\,, \qquad m_Z = \sqrt{g^2 + g'^2}\frac{v}{2}\,. \tag{1.28}$$

This implies that the W and Z masses are not independent but related by the Weinberg angle:

$$\frac{m_W}{m_Z} = \cos\vartheta_W\,. \tag{1.29}$$

Evaluating the potential terms in the original Lagrangian leads to the mass term for the Higgs field, $-\mu^2 H^2$. The value μ is a free parameter of the theory and can be determined by measuring the mass of the Higgs boson. The value of v is known by measurements of the W boson mass and the Weinberg angle.

Other terms in the potential are proportional to H^3 and H^4. Such terms represent interactions of the Higgs boson with itself, and λ corresponds to the coupling strength. Note that λ is not independent, but given by μ and v from Equation 1.23.

1.4.1 Yukawa Interactions

The term *Yukawa interaction* refers to the interaction of a scalar field with a Dirac field. Indeed, when introducing the Higgs doublet, additional Yukawa terms must be added to the Lagrangian density that represent the interaction of the Higgs field with the fermions. For each fermion field ψ, the Yukawa interaction is given by

$$\mathcal{L}_{\text{Yukawa}} = -g\bar{\psi}\Phi\psi\,, \tag{1.30}$$

where g is the Yukawa coupling constant to the fermion in question. After the spontaneous symmetry breaking, using Equation 1.27, two terms are obtained:

$$\mathcal{L}_{\text{Yukawa}} = -gv\bar{\psi}_e\psi_e - g\bar{\psi}_e H\psi_e\,. \tag{1.31}$$

There are two lessons that can be learned from Equation 1.31. The first term can be identified with the mass of the fermion, $m = gv$. This is important, because a direct mass term as in Equation 1.4 would not be gauge-invariant under SU (2) symmetry transformations of the weak interaction. The second term corresponds to the interaction of the fermion with the Higgs field. The coupling strength g is proportional to the mass of the particle, which means that the coupling between a fermion field and the Higgs field is directly proportional to the fermion's mass. This becomes relevant for experimental searches for the Higgs boson.

The terms for up-type fermions can be obtained in a similar way with the charge-conjugated field $\tilde{\Phi} = i\sigma_2\Phi^*$. It is important to note that the theory does not make any prediction about the masses of the fermions. The Yukawa coupling g is a free parameter in the theory, and it can be different for each fermion.

1.5 Quantum Chromodynamics

Quantum Chromodynamics (QCD) is the quantum field theory of the strong interaction. QCD is a rich field, and at this point only the basic concepts and results are discussed. More in-depth material can be found e.g. in [19].

The charge of the strong interaction is called color, however this is just an analogy and has nothing to do with actual colors. Quarks are arranged in color triplets,

$$\psi = \begin{pmatrix} \psi_r \\ \psi_g \\ \psi_b \end{pmatrix} , \tag{1.32}$$

so the Lagrangian density is postulated to be invariant under SU (3) transformations. In a similar way as for the electromagnetic and weak interactions, this leads to eight gauge bosons, called gluons. A gluon carries both color and anti-color so that, due to color conservation, quarks change their color when interacting with a gluon. The leptons do not carry color charge and therefore form SU (3) singlets.

Since the gluons are color-charged themselves, they interact with each other. This is different from QED where photons do not carry electric charge, and it leads to an important consequence. The effective potential of a color-charged particle is proportional to

$$V_c(r) \propto \alpha(r) \frac{1}{r} + \beta r . \tag{1.33}$$

The first term is attributed to the color charge of quarks, and as with the electromagnetic interaction the potential diminishes at large distances. The second term, which originates from gluon self-coupling, however, leads to much energy being stored in the color field for color charges which are far apart from each other. At distances of about 1 fm, it is energetically favorable to create a new quark-antiquark pair out of the vacuum to shorten the distances between individual quarks. The conclusion of this is that no free quarks can be observed since they always arrange with other quarks or antiquarks to form color-neutral objects, called *hadrons*. This effect is called *color confinement* of QCD. Possible arrangements include mesons (color and anticolor) and baryons (red, green and blue or anti-red, anti-green and anti-blue).

The masses of mesons or baryons are usually much higher than the masses of their quark constituents. For example, the proton, which consists of two up quarks and one down quark (the *valence quarks*), has a mass of 938 MeV whereas the quarks themselves have masses around 5 MeV. The remainder of the mass is attributed to the color field between the three quarks, i.e. carried by the gluons that they constantly exchange. Gluons can also, temporarily, generate additional quark-antiquark pairs in accordance with Heisenberg's uncertainty principle. Therefore, the probability of, for instance, finding a strange or a charm quark within the proton is larger than zero. Such temporary quarks are called *sea quarks*.

The total momentum of a hadron is split among its constituents (called partons), the valence quarks, sea quarks, and gluons. The Bjorken scale variable,

$$x = \frac{p_P}{p_H} , \tag{1.34}$$

denotes the fraction of the full hadron momentum p_H that a parton carries. x is not deterministic, but every time a parton performs an interaction, its momentum can be different. One can think of the gluons constantly exchanging momentum between the quarks. This behavior can be described with *parton density functions* (PDFs). Let $f_d(x)$ be the PDF for down quarks within a proton. Then $f_d(x)\,\mathrm{d}x$ equals the probability of finding a down quark with momentum between x and $x + \mathrm{d}x$ inside the proton. The actual

parton density functions depend on the energy of the hadron. They can be measured with inelastic proton scattering experiments (e.g. [20]).

Due to the confinement, quarks or gluons cannot exist alone. When a high-energetic quark or gluon is created in particle collisions, their energy is high enough to create not only one quark-antiquark pair, but many of them. This leads to the formation of a whole bundle of hadrons all of which move into approximately the same direction. Such a bundle is called a *jet*. Since some hadrons are unstable, also leptons and photons can be part of a jet, created by the decay of unstable hadrons.

1.6 Experimental Verification

The Standard Model has been verified by hundreds of experiments. A recent example is the prediction of the top quark: after the bottom quark was discovered and it was clear that there exists a third generation of quarks, the search for the top quark has started. Eventually, it was discovered by the CDF and D0 collaborations at the Tevatron [21, 22] in 1995. It was heavier than originally expected but in full agreement with the Standard Model. In 2000, the τ neutrino was experimentally observed [23]. Also, the $W^{\pm}$ and Z^0 bosons were predicted in the GWS model before their existence has been experimentally confirmed.

The latest discovery was that of the Higgs boson in the year 2012 at the Large Hadron Collider (LHC) [24, 25]. Its mass is a free parameter of the Standard Model, and it has been measured to a value close to 125 GeV. The remainder of this thesis focuses on the measurement of the Higgs boson.

1.6.1 Higgs Boson Production

At particle colliders, the Higgs boson is produced via different mechanisms. The most important ones at hadron colliders are shown in Figure 1.3. At LEP, the Large Electron-Positron collider, and also at the Tevatron, which collides protons and anti-protons, the Higgs strahlung process (Fig. 1.3c) was the most dominant production process (at LEP with electrons instead of quarks in the initial state). The LHC (see chapter 2 for details) collides two proton beams where the probability of finding an antiquark in the initial state is much lower. Therefore, the gluon-gluon fusion process (Fig. 1.3a) is the dominant production process at the LHC.

The vector boson fusion process (Fig. 1.3b) is also very interesting at the LHC. Its production cross section is about one order of magnitude lower than for gluon fusion, however it has a very clean event topology where there are two jets in opposite hemispheres of the detector and very low activity between these two jets. This allows for a very clean separation against background processes because only few other non-Higgs processes result in a similar topology.

The production with an associated top quark (Fig. 1.3d) only plays a minor role because due to the very heavy particles in the final state the available phase space limits the production cross section. Replacing the top quarks by bottom quarks leads to a higher phase space but lower coupling to the Higgs boson.

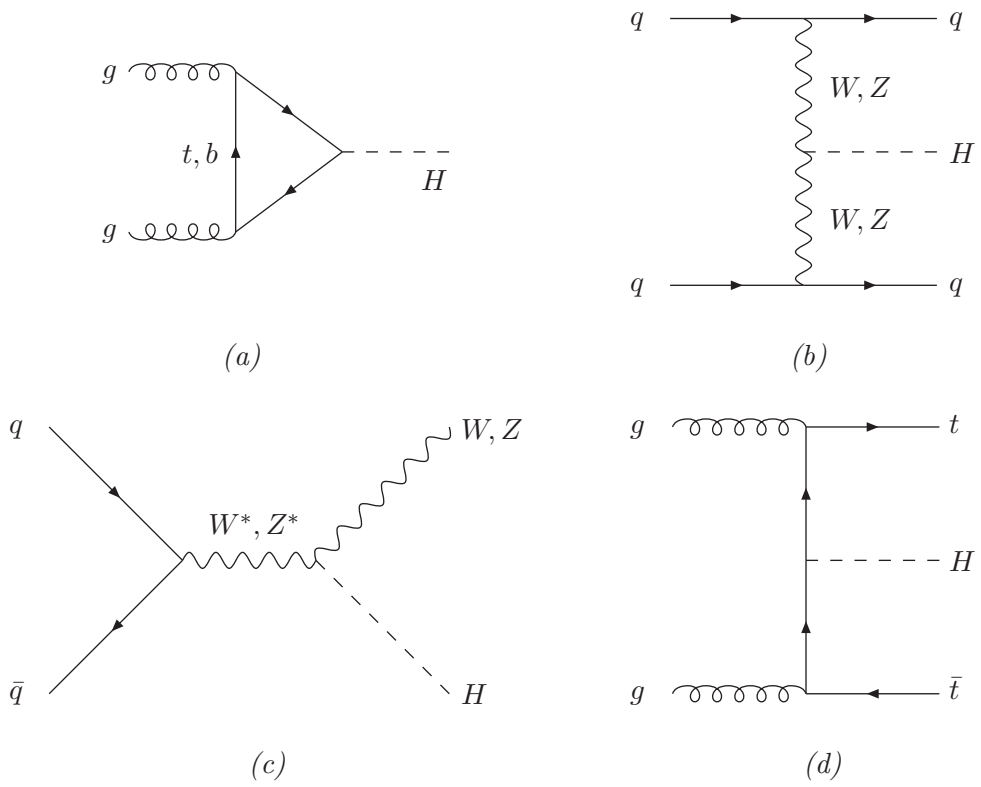

Figure 1.3: Leading order Feynman diagrams for Higgs production at a hadron collider. The processes are called (a) gluon-gluon fusion, (b) vector boson fusion, (c) Higgs strahlung and (d) quark associated production. From [26].

1.6.2 Higgs Boson Decay

The lifetime of the Higgs boson is very short ($\mathcal{O}\left(10^{-25}\,\mathrm{s}\right)$) so that it decays in almost the same instant in which it was produced. The Higgs boson dominantly decays to particles with high masses, as long as the decay is kinematically allowed. The branching ratio depends on the mass of the Higgs boson and can be seen on the left hand side of Figure 1.4.

For a mass of around 125 GeV, the most prominent decays are $H \to b\bar{b}$ and $H \to \tau^+\tau^-$. Also, the decay into two photons is possible, but suppressed, because it requires a top quark or W boson loop, since the Higgs boson does not couple to the massless photon directly. The first channels in which the Higgs boson was seen are the di-photon [28, 29] and the ZZ channels [30, 31]. Even though the branching ratios are low, the channels are more sensitive experimentally. The number of events from other processes with a similar signature in the detector (background events) is small, and the resonance mass can be reconstructed accurately.

However, all decay channels are important to verify the couplings of the Higgs boson to fermions and gauge bosons.

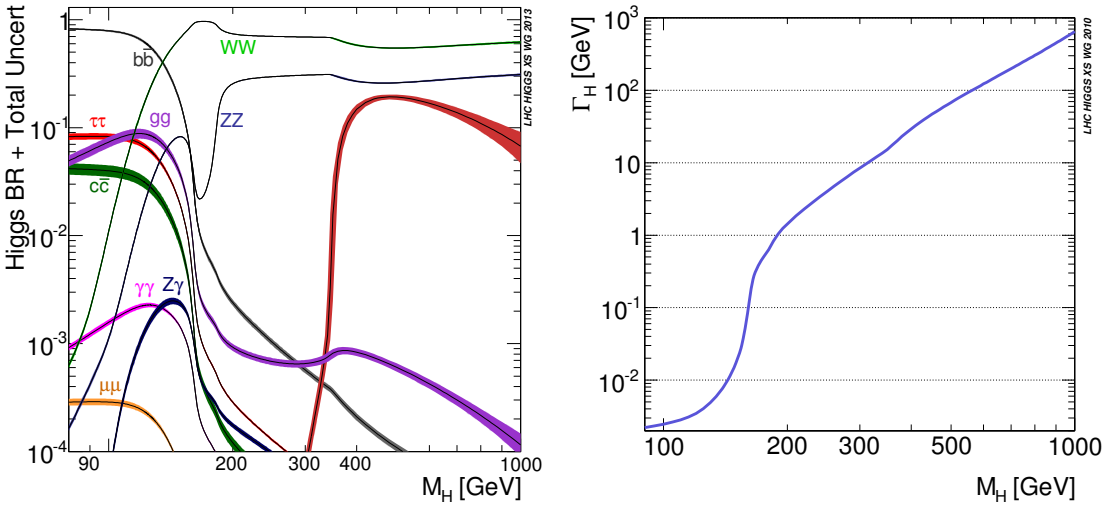

Figure 1.4: The branching ratio (left) and the decay width (right) of the Standard Model Higgs boson as a function of its mass, from [27].

1.6.3 Experimental Status

Since the Higgs boson mass fixes all other parameters in the Higgs sector, properties such as the couplings to the other particles or the spin and CP properties of the Higgs boson, are predicted by the Standard Model. The basic question is whether the boson that was recently discovered is indeed the Higgs boson as predicted by the Standard Model, or a different particle. Experimental verification of these properties is therefore crucial.

To date, the $\gamma\gamma$ [28, 29], ZZ [30, 31], W^+W^- [32, 33], and $\tau^+\tau^-$ [34, 35] final states have been observed experimentally. Furthermore, many alternating spin and CP hypotheses have been tested against the Standard Model, and in all cases the Standard Model is preferred by the data [36, 37]. Apart from making the existing measurements more precise, the goal for the foreseeable future will be to measure the $b\bar{b}$ decay as well as more rare decays such as $\mu^+\mu^-$ and $Z\gamma$. In the longer term, the measurement of the Higgs self-coupling will become interesting, possibly only at the future International Linear Collider (ILC).

1.7 Limitations of the Standard Model

It is known already that the Standard Model cannot be the most basic theory of particle physics. For once it does not include a theory of gravity, but there are also other problems that cannot be solved within the Standard Model.

The first direct observation of physics beyond the Standard Model is the discovery of neutrino oscillation at Super-Kamiokande [38] and SNO [39]. In the Standard Model, neutrinos are massless, however, neutrino oscillation can only be described if the difference of the squared neutrino masses is nonzero.

Another evidence for physics beyond the Standard Model comes from cosmology: the velocity of distant galaxies rotating around the galaxy center is higher than what is predicted with Newton's laws based on the mass of baryonic matter present within the galaxy. This

suggests that there is additional mass which cannot be seen, i.e. it does neither interact strongly nor electromagnetically, because otherwise it could be observed with telescopes. This property gives it the name "dark matter". The Standard Model does not have a particle which can describe dark matter. The only stable particle which interacts only weakly is the neutrino which is too light to account for the observed rotation curves.

The theory of supersymmetry solves this problem by introducing new stable, massive particles. Supersymmetry postulates that, for every Standard Model particle, there exist new particles (called superpartners) with different masses and whose spin differs by $\frac{1}{2}$. Dark matter could consist of such a superpartner. Furthermore, there are at least 5 Higgs bosons in supersymmetric theories.

Other shortcomings of the Standard Model include its high number of free parameters and the inability to explain why exactly there are three generations of fermions with a strict mass hierarchy.

Apart from discovery or exclusion of the Higgs boson in the full mass range, observing any signal of "New Physics" is the primary goal of the Large Hadron collider which is discussed in the next chapter.

2 The CMS Experiment at the LHC

Large facilities are needed in order to study the Standard Model experimentally under laboratory conditions. Most experiments exploit the same basic principle: accelerating particles to high energies and letting them collide with another particle beam or a fixed target. Particle detectors built around the collision region record the result of the primary particle interaction. The latest incarnation of this concept is the Large Hadron Collider (LHC) at CERN[1] near Geneva, Switzerland. This thesis presents data obtained with the Compact Muon Solenoid (CMS) experiment, which is one of the detectors at the LHC. In the remainder of this chapter, first the LHC machine and then the CMS detector are described in more detail.

2.1 The Large Hadron Collider

The LHC is the largest particle accelerator ever built. It reaches unprecedented collision energies which allow to study particle interactions in an energy regime that has not been explored before experimentally. One of its primary goals is the search for the Higgs boson. Now, that a Higgs-like particle has been found [24, 25], the precise measurement of the properties of this particle is of high priority. Another important objective is searching for unknown physics processes that are not explained by the Standard Model such as dark matter, supersymmetry, or microscopic black holes. In many of these scenarios, new observable particles in the mass range of a few 100 GeV to a TeV are predicted. Such energies are reachable at the LHC, so that these particles, if they exist, could be produced and detected. Recent reviews on these topics can be found in [40, 41, 42].

The LHC is a proton-proton collider. Two beams of protons are circulating in an underground tunnel of nearly 27 km of circumference and colliding at a center-of-mass energy of up to $\sqrt{s} = 8\,\text{TeV}$. The reason for choosing two proton beams is twofold. First, the high mass of the proton means that the energy loss due to synchrotron radiation is very small, as it scales with m^{-4}. The previous flagship accelerator at CERN, the Large Electron-Positron Collider (LEP) [43], could not reach energies higher than $\sqrt{s} = 209\,\text{GeV}$ due to the synchrotron radiation of the electrons and positrons. Second, protons are chosen over antiprotons. The Tevatron collider [44] at FNAL[2], in operation until 2011, was a proton-antiproton collider with a center-of-mass energy of $\sqrt{s} = 1.96\,\text{TeV}$. While simpler from an engineering point of view, the production rate of antiprotons would not be sufficient to provide enough particles for the collision rates anticipated at the LHC.

The LHC started operation in September 2008 when, for the first time, particles have been circulating in the machine. After an incident which damaged several of the superconducting magnets [45], the physics program started in March 2010 with a center-of-mass energy of $\sqrt{s} = 7\,\text{TeV}$. After two years of running in this configuration, the energy was increased to $\sqrt{s} = 8\,\text{TeV}$ for another year. The two following years are dedicated to a

[1] European Organization for Nuclear Research
[2] Fermi National Accelerator Laboratory

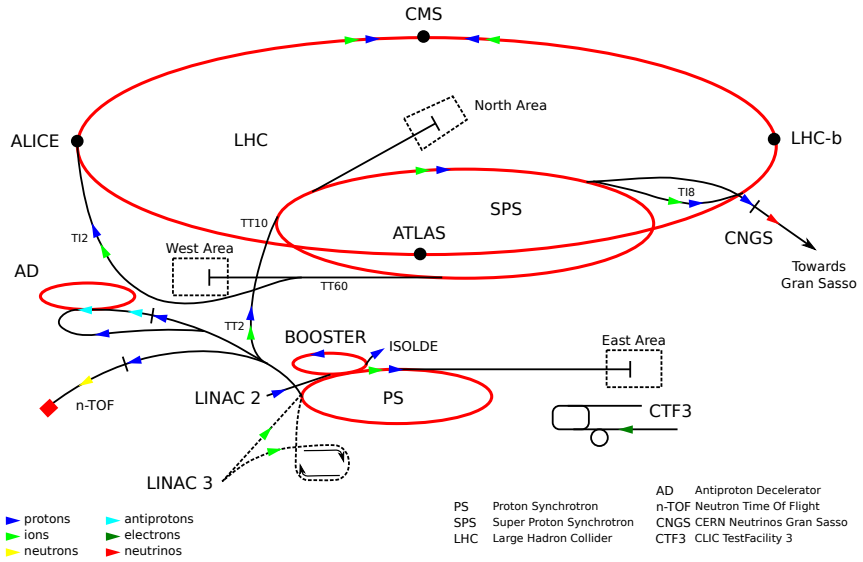

Figure 2.1: Schematic view of the particle accelerators at CERN. The LINAC II, Booster, Proton Synchrotron (PS) and Super Proton Synchrotron (SPS) are used as pre-accelerators for the LHC. Taken from [46].

shutdown after which it is expected that operation can be resumed with energies close to the design energy of $\sqrt{s} = 14\,\text{TeV}$ in 2015. For about one month in each year of operation, the LHC is colliding lead ions instead of protons. Even though the energy per nucleon is lower than for pp running, the total energy is much higher. This operation mode leads to a very hot and dense environment, in which studies about a possible new matter state called quark-gluon-plasma (QGP) are made. A QGP resembles the conditions in the early universe only microseconds after the big bang when temperatures are so high that quarks and gluons can propagate freely.

2.1.1 Particle Acceleration

Before the beam is injected into the main LHC ring, the particles are pre-accelerated by a sequence of smaller particle accelerators (injectors). First, protons are accelerated by a linear accelerator, LINAC 2, to an energy of 50 MeV. Then, they make their way through several ring accelerators: the Proton Synchrotron Booster (PSB, 1.4 GeV), the Proton Synchrotron (PS, 26 GeV), and the Super Proton Synchrotron (SPS, 450 GeV). At the energy of 450 GeV, the protons are injected into the LHC where they are accelerated to their final energy of 4 TeV (2012) and then brought to collision at the various interaction points. Figure 2.1 shows the CERN accelerator complex including the LHC and its injectors.

2.1.2 Performance of the LHC

To assess the performance of a collider, the most important quantities are its center-of-mass energy and the luminosity. The center-of-mass energy is the energy that is available to produce new particles, while the luminosity is a measure of the number of particle

Table 2.1: Important collider parameters of the LHC and Run II of the Tevatron. For the LHC, both the original design value and the best value achieved so far are reported. The LHC values are taken from [47]. The Tevatron parameters are taken from [48, 49].

Parameter	LHC		Tevatron	Unit
	Design	Achieved		
Circumference	26.7	26.7	6.28	km
Beam Energy	7000	4000	980	GeV
Number of Particles per Bunch	1.15	1.7	p: 2.9	10^{11}
			$\bar{p}$: 1.0	10^{11}
Number of Bunches	2808	1380	36	
Bunch Spacing	25	50	396	ns
Crossing Angle at IP	285	290	136	µrad
Normalized Emittance	3.75	2.5	15	mm mrad
β^* at IP1/IP5	0.55	0.6	0.28	m
Luminosity	100	77	4	$10^{32}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$

interactions per time, which is given by

$$\frac{\mathrm{d}N}{\mathrm{d}t} = \sigma \cdot L\,, \tag{2.1}$$

where σ is the cross section of the process in question and L the (instantaneous) luminosity of the collider, measured in $\mathrm{cm}^{-2}\mathrm{s}^{-1}$. Obviously, the higher the luminosity, the more particle interactions occur. While high energies allow heavy particles to be created, high luminosities allow rare interactions to be probed and a large amount of data to be accumulated, reducing statistical uncertainties.

The LHC beam is not a continuous stream of particles, but it consists of individual bunches of protons. This bunch structure allows the particle acceleration to be performed with 400 MHz radio-frequency (RF) cavities, which also correct longitudinal injection errors. In the transverse direction, the beam is kept focused by quadrupole and sextuple magnets. The major part of the LHC, however, are the 1232 blue dipole magnets which keep the beam on a circular trajectory. Each bunch of particles nominally consists of 1.1×10^{11} protons, even though higher values have been routinely reached at the expense of a larger spacing between bunches. Bunches are injected in several bunch trains of up to 144 bunches with 50 ns spacing (with the nominal value being 25 ns).

In general, lower bunch spacings allow more particle bunches in the collider, but fewer particles per bunch, since the limiting factor is the total beam current. The luminosity is proportional to the number of bunches and to the square of the number of particles per bunch. Therefore, big bunches with a large gap between them would be preferable in principle. However, fewer interactions per bunch crossing has the additional advantage of cleaner events with less activity, making them easier to analyze and to identify individual particles. The LHC experiments have been designed for a bunch spacing of 25 ns. Even though this value has not yet been reached, the experiments have been performing very well under the harsher conditions of 50 ns spacing. This is partly due to the fact that also the collision energy has not reached its design value yet.

Other important parameters that contribute to the luminosity are the emittance of the beam ϵ and the beam size at the interaction region β^*. The emittance is a measure of the spread of particles in position-momentum space, i.e. low emittance means that the particles in the beam are confined to a small region and have similar momenta compared to each other. Before two beams are brought to collision, they will be squeezed as much as possible. The β^* variable represents how well the beams can be squeezed with respect to their normal beam parameters. It can be thought of the distance from the interaction point where the beam is twice as wide as at the interaction point itself. Table 2.1 presents the most important parameters of the LHC, and, for comparison, the Tevatron.

Another quantity related to the luminosity is the *integrated* luminosity, defined as

$$\mathcal{L} = \int L \mathrm{d}t \,, \tag{2.2}$$

where the integral goes over a period of running the machine, for example one year of LHC operation. It is a measure of total data accumulated and typically measured in inverse barns[3]. The LHC has delivered more than $30\,\mathrm{fb}^{-1}$ to each of the two major experiments so far. However, only a fraction of about $25\,\mathrm{fb}^{-1}$ is usable by the experiments due to inefficiencies while recording the collisions or other technical problems during times of stable beams.

In order to measure the luminosity, the clusters in a subdetector with low occupancy are counted in events that are only triggered by two bunches crossing in the interaction region ("zero-bias" events) [50]. The number of particles observed on average is then proportional to the luminosity, where the proportionality factor corresponds to the total inelastic *pp* cross section multiplied by the geometric acceptance of the detector. This absolute calibration is obtained from a special operation mode of the LHC in which the two beams are swiped through each other in both transverse directions, known as a *van-der-Meer scan* [51].

2.1.3 The LHC Experiments

Figure 2.2 shows an aerial photograph of the LHC area with 8 points where access to the underground tunnel is possible. At four of these points, there are interaction regions where the particle beams can be brought to collision. The following four experiments are installed at these locations to record the collision events:

- ALICE: The primary goal of the ALICE experiment [53] is to study the quark-gluon plasma in heavy-ion collisions. Its main feature is a large time projection chamber for full three-dimensional track information in high-occupancy events. ALICE is located at Point 2 (P2) on the LHC ring.

- ATLAS: The ATLAS experiment is one of the two general purpose experiments [54]. It is designed to identify and study all particles produced in LHC collisions, to be prepared for any sort of new physics processes. ATLAS features both a solenoidal and a toroidal magnet with a maximum magnetic field of 2 T, to bend charged particle tracks so that their momentum can be determined by measuring the curvature. The detector is cylindrical in shape, 45 m long with a diameter of 22 m. It is installed at P1.

[3] $1\,\mathrm{b} = 10^{-28}\,\mathrm{m}^2$

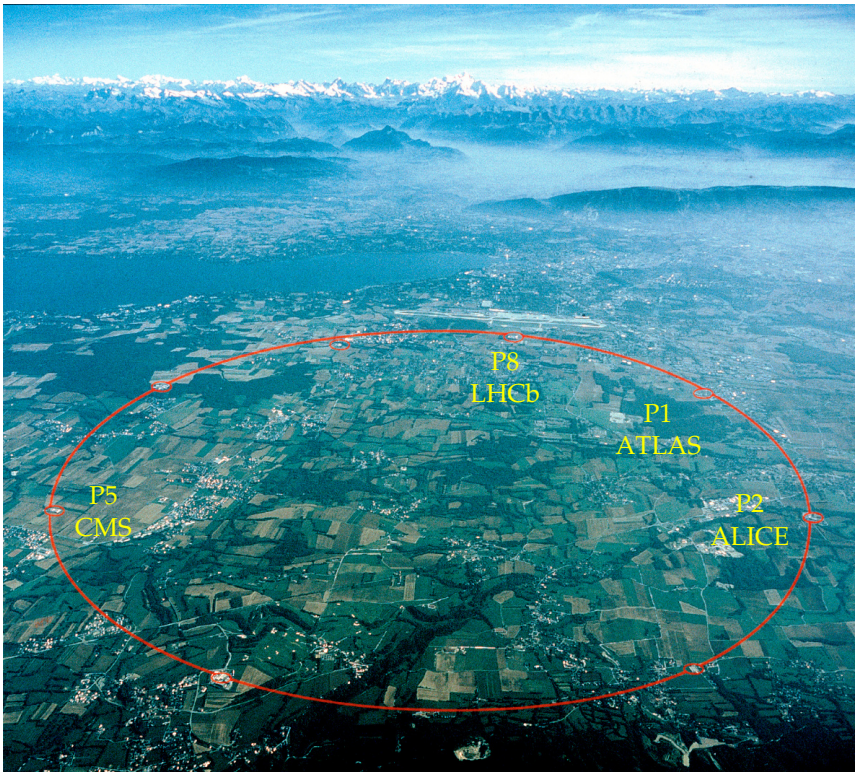

Figure 2.2: Aerial view of the area where the LHC is located. The accelerator itself is underground. There are eight locations where the tunnel can be accessed, numbered P1 to P8. Experiments have been installed at four of them. In the background, the airport of Geneva, Lake Geneva, and the Alps can be seen. Taken from [52].

- CMS: The other general purpose experiment at the LHC is the Compact Muon Solenoid (CMS) [55]. Compared to ATLAS, it is compact with a length of 21 m and a diameter of 16 m. However, it weighs 12 500 t, almost twice as much as ATLAS. CMS is characterized by a superconducting solenoid providing a 3.8 T magnetic field and an excellent performance in identifying muons. It is situated at P5.

- LHCb: The LHCb experiment [56] specializes in the study of B meson decays. Its research goals are searching for very rare decays and exploring the CP violation in the bottom system and also in the charm system. Its primary instrument is the Vertex Locator, a silicon strip detector very close to the interaction region, in order to accurately measure the secondary vertices from B meson decays. LHCb is located at P8 on the LHC ring.

In the following section, the CMS experiment is described in more detail.

2.2 The CMS Experiment

CMS is a general purpose detector consisting of several sub-detector layers around the interaction point. The silicon tracker is closest to the interaction point and provides an

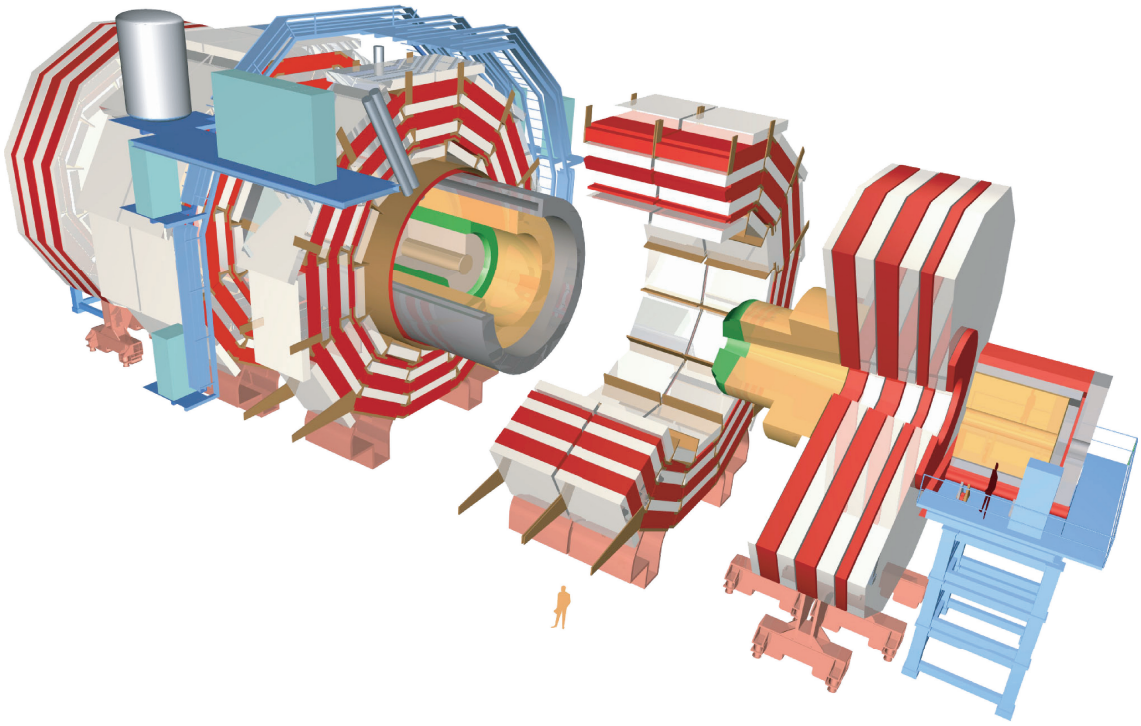

Figure 2.3: An overview of the CMS detector with its individual components. CMS consists of the central barrel part and two endcaps. For operation, the detector is fully closed. Taken from [57].

accurate tracking of charged particles. Around the tracker, there is the electromagnetic calorimeter (ECAL). It is used for stopping and measuring the energy of electromagnetically interacting particles. The hadronic calorimeter (HCAL) comes next, where hadronically interacting particles are stopped. Only outside of the HCAL is the superconducting solenoid, which provides a 3.8 T magnetic field inside the detector, and about 2.0 T outside of the coils. The inside field is very homogeneous and points in the direction of the beam, so that charged particles traversing the detector are bent in the plane transverse to the beam. From the curvature of the particle tracks, their momentum and charge can be determined. Outside of the solenoid, there is the iron return yoke, interspersed with the muon system, a variety of detector technologies to identify muons. The individual subdetectors are discussed in more detail in the following sections.

CMS is designed to instrument almost the full 4π solid angle with detectors, to maximize acceptance and to be able to draw conclusions from the total energy and momentum balance in a collision event. To this extent, CMS consists of a central barrel in which the sensitive area is layed out parallel to the beam, and of two endcaps in the forward and backward regions with the sensitive area perpendicular to the beam. Figure 2.3 shows an overview of the detector.

2.2.1 The CMS Coordinate System

The CMS coordinate system is chosen such that the origin is at the nominal interaction point in the center of the apparatus. The x axis then points toward the center of the LHC

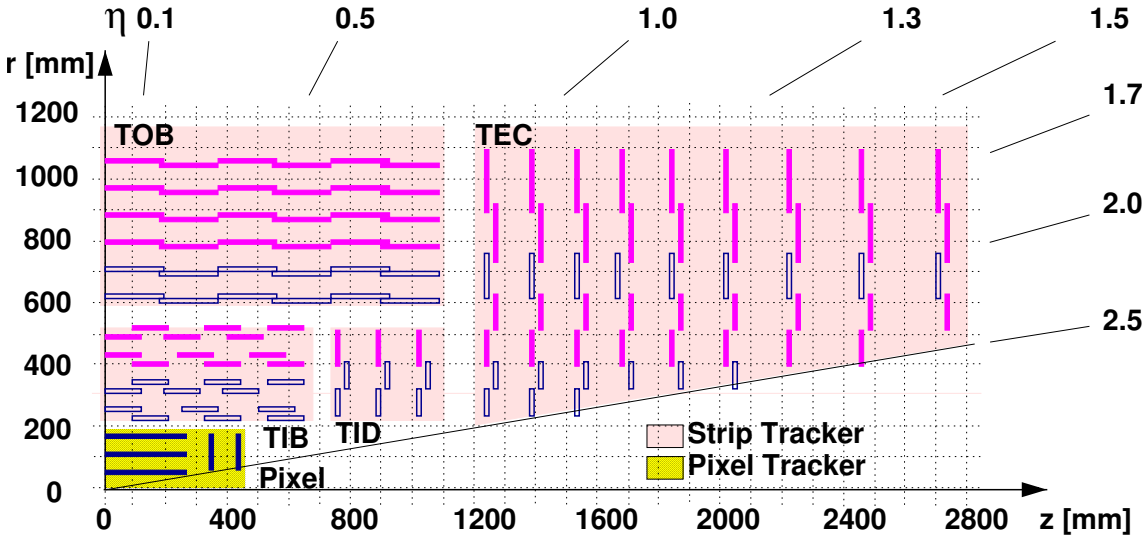

Figure 2.4: Layout of one quarter of the CMS tracking system in the r-z plane. The interaction point is in the lower left corner. Closest to the interaction point is the pixel system, followed by the strip detectors in the outer regions. Taken from [58].

ring, the y axis upwards toward the surface, and z points parallel to the counterclockwise beam direction, toward the ALICE experiment. Particle momenta are preferably given in p_{T}, η, and ϕ coordinates, however, where p_{T} is the particle's momentum projected to the plane transverse to the beam, $p_{\mathrm{T}} = \sqrt{p_x^2 + p_y^2}$, ϕ is the azimuthal angle defined in this plane with respect to the x axis, and the pseudorapidity η is related to the polar angle θ, measured from the z axis, as

$$\eta = -\log\left(\tan\left(\frac{\theta}{2}\right)\right) . \tag{2.3}$$

Therefore, a pseudorapidity of $\eta = 0$ corresponds to a particle perpendicular to the beam, while $\eta = \pm\infty$ corresponds to the beam direction. For massless particles, this is equivalent to the rapidity

$$y = \frac{1}{2}\log\left(\frac{E + p_z}{E - p_z}\right) \tag{2.4}$$

which has the property that differences in rapidity are invariant under Lorentz boosts along the z axis. This is useful in some analyses, because the momentum in z direction is random according to the parton density functions of the two colliding protons. However, this quantity depends on the energy of the particle in question and is therefore inconvenient for a global coordinate system.

2.2.2 The Silicon Tracker

The silicon tracking detector [59] is the system which is closest to the interaction point. Its objective is to measure the trajectories of charged particles. Many layers of pixel and strip modules are used to track the position of particles as they traverse the detector. The

three barrel layers and two endcap layers of the pixel detector are the innermost layers of the tracker, providing the best position resolution and a three-dimensional hit information. The pixel size is $150 \times 100\,\mu\text{m}^2$, however the position resolution is as good as $8\,\mu\text{m}$ in the most sensitive coordinate due to charge sharing between neighboring pixels combined with full analog readout. Being so close to the interaction region, the pixel sensors are exposed to very high particle fluxes, requiring a radiation hard design. In total, there are 66 million readout channels. The pixel detector is described in more detail in Chapter 4.

The strip detector covers the region behind the pixel detector, and has 9 or 10 layers. Each strip layer provides 2-dimensional hit information in r-ϕ (barrel) or z-ϕ (endcap), allowing for an accurate p_T measurement. A strip is between 7.0 cm and 12.5 cm long and the pitch between strips is between $60\,\mu\text{m}$ and $270\,\mu\text{m}$. This design allows to instrument a large region with sensitive material while keeping the number of readout channels manageable (and, therefore, also the power consumption and the cost). Figure 2.4 shows an overview of the layout of the silicon tracker. The strip tracker is divided in four parts: the Tracker Inner Barrel (TIB), Tracker Outer Barrel (TOB), Tracker Inner Disk (TID) and Tracker Endcaps (TEC). The blue, open rectangles represent a double layer of modules ("stereo" modules) whose strips are tilted with respect to each other by $\approx 100\,\text{mrad}$, providing also information in the third coordinate, even though ambiguities remain for high occupancies. The strip tracker features about 9.6 million readout channels.

The total area covered with sensitive silicon material is about $200\,\text{m}^2$, making the CMS silicon tracker the largest in the world. It covers the pseudorapidity range up to $|\eta| < 2.5$. Both the pixel and the strip tracker exploit the same detection principle: the silicon bulk is doped with low-density n implant, while one side of the sensor has a high p doping. This creates a p-n junction which, when applying high voltage in reverse direction, is depleted of free charge carriers through recombination of electrons and holes. When a charged particle crosses the depleted region, electron-hole pairs are created, making a measurable signal, which is then amplified and can be read out.

2.2.3 The Electromagnetic Calorimeter

The electromagnetic calorimeter (ECAL) [61] consists of around 80 000 lead-tungsten ($PbWO_4$) crystals. Within the crystals, charged particles emit photons via bremsstrahlung. The photons convert to electron-positron pairs when interacting with the crystal material, and the process continues until the energy of the photons is below the pair production threshold. Since the ECAL is a homogeneous calorimeter, the photons then excite the scintillating material which re-emits the absorbed energy as light with a well-defined wavelength. As the crystals are transparent for light it can fully traverse the crystal and is detected by avalanche photodiodes (APDs) in the barrel and vacuum photodiodes (VPTs) in the endcaps. The energy of the primary particle is then proportional to the total number of emitted photons.

Lead-tungsten was chosen as the scintillating material due to its sufficient radiation hardness and its short scintillation time, allowing the calorimeter to be operated within the nominal LHC bunch crossing frequency of 40 MHz. 80% of the light is emitted within 25 ns. In addition, the high density of $8.28\,\text{g/m}^3$ and short radiation length of 0.89 cm allow the construction of a very compact calorimeter that can be placed inside the solenoid. The crystals are 23 cm long (22 cm in the endcaps) to limit the leakage of electromagnetic showers behind the calorimeter. Each crystal covers a region of 0.0175×0.0175 in η-ϕ space.

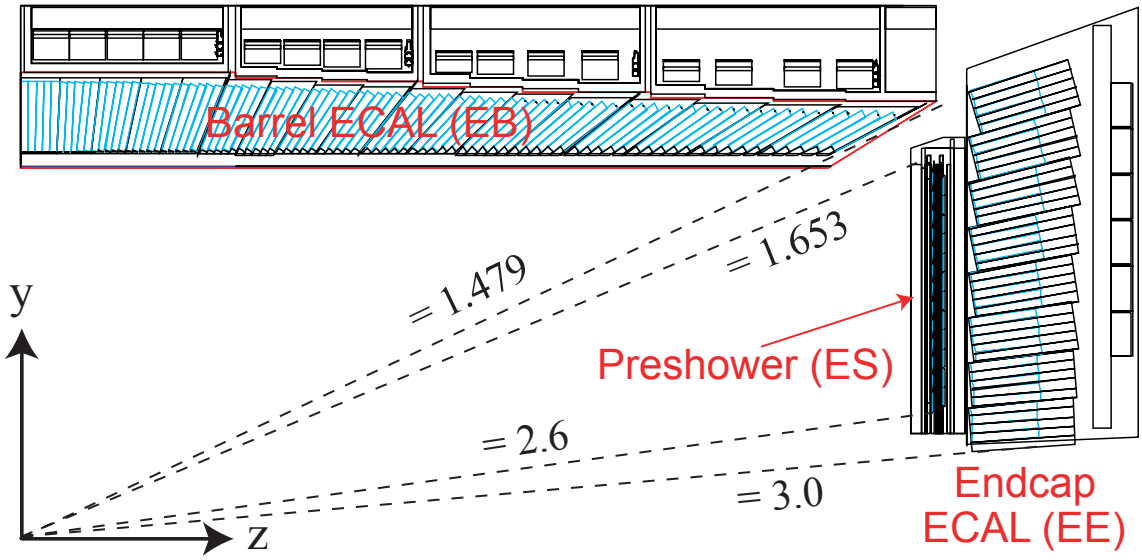

Figure 2.5: Overview of the electromagnetic calorimeter. It covers a pseudorapidity range up to 3.0. In the endcap, the indicated η regions (dashed lines) are not covered by the pre-shower detector. Taken from [60].

Figure 2.5 gives an overview of the layout of the ECAL. The preshower detector is a lead-based sampling calorimeter with 3 radiation lengths. It helps to improve the spatial resolution in the endcaps which is especially important for the reconstruction of neutral pions, since they immediately decay to two mostly collinear photons. The ECAL covers the pseudorapidity range up to $|\eta| < 3.0$.

2.2.4 The Hadronic Calorimeter

The hadronic calorimeter (HCAL) [63] is a sampling calorimeter which uses brass absorbers due to its low radiation length, again allowing for a compact design. In the barrel, each $\approx 5\,\text{cm}$ of absorber is followed by a layer of active scintillator material with a depth of 3.7 mm, for a total of 17 layers. In the endcaps, $\approx 7.9\,\text{cm}$ of absorber is used with 9.0 mm scintillators and 18 layers in total. Plastic scintillator was chosen due to its moderate radiation hardness and long-term stability. The granularity of the calorimeter is 0.087×0.087 in η-ϕ space below $|\eta| = 1.6$, and 0.17×0.17 above.

Strongly interacting particles (hadrons) interact with the matter in the absorber, producing cascades of low energetic particles. Most of the energy is deposited in the absorber material, and only some particles reach the scintillator material which emits detectable photons. Since only a fraction of the energy of the primary particle is registered, the calorimeter needs to be carefully calibrated to reconstruct the original energy, and suffers from large statistical fluctations. This also explains why the relative energy resolution of the hadronic calorimeter is worse than the one of the electromagnetic calorimeter.

The thickness in interaction lengths increases with pseudorapidity for the barrel part of the HCAL, and in the central part, hadronic showers cannot be fully stopped. Therefore, the magnet coil is used as an additional absorber and the region outside of the coil is instrumented with scintillators in the central region.

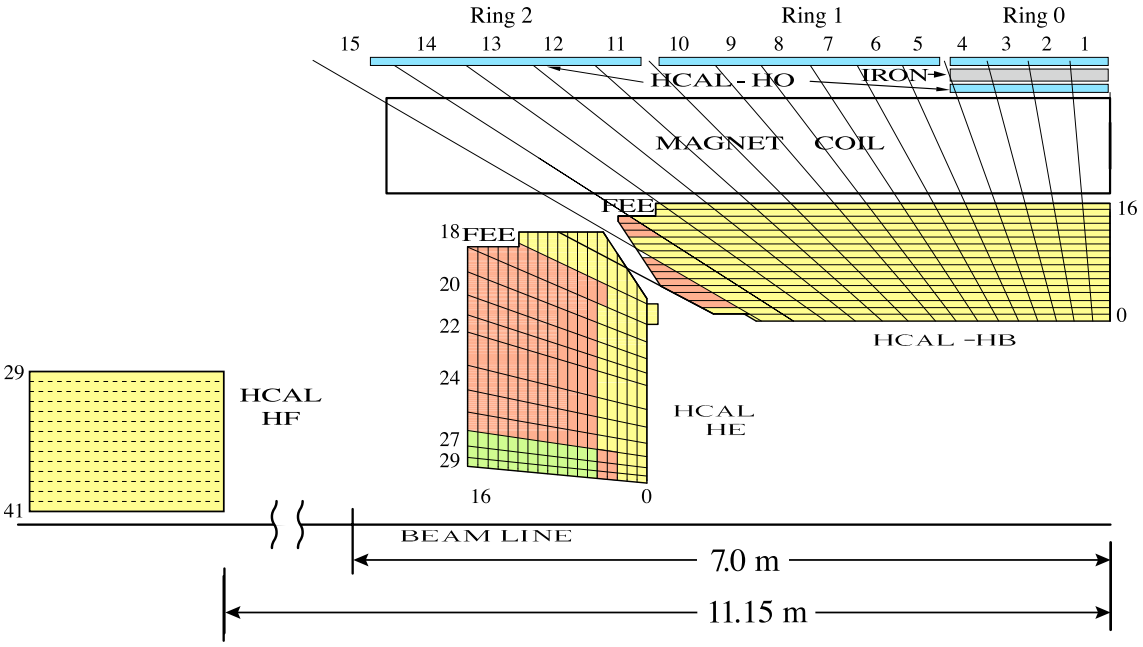

Figure 2.6: Layout of the hadronic calorimeter. The barrel and endcap cover up to $|\eta| < 3.0$ while the forward calorimeter extends the range to $|\eta| < 5.0$. In the central part, there is a calorimeter component outside of the magnet (HO). Taken from [62].

Figure 2.6 shows the layout of the HCAL including the longitudinal segmentation. Other than the barrel and the endcaps, there is also a forward region covering the pseudorapidity range $3.0 < |\eta| < 5.0$ (HF). This is especially important so that particles in the very forward direction are not counted toward the missing energy in an event. Another purpose of the HF is the measurement of the luminosity, which is proportional to the occupancy in the HF to good approximation.

2.2.5 The Muon System

Other than neutrinos, which are completely undetectable, the only particle that is not stopped in one of the two calorimeters is the muon. Due to its higher mass with respect to the electron, it does not lose much energy due to bremsstrahlung in the electromagnetic calorimeter, and since it does not interact hadronically it is not stopped in the hadronic calorimeter either. Its lifetime of 2.2 µs is so long that in particle physics experiments it can be regarded as a stable particle. Therefore, muons are detected outside of the solenoid with additional tracking detectors. The muon system covers a pseudorapidity range of $|\eta| < 2.4$.

Figure 2.7 shows an overview of the muon system in CMS [64]. It is located between the iron that serves as a return yoke for the magnetic field. In the barrel region, drift tubes are deployed, while in the endcaps cathode strip chambers are used. The detector principle for both technologies is the same: inside a gas chamber, there is a wire which is on high voltage with respect to the edge of the chamber. When a charged particle crosses the gas, it ionizes the gas, and the electrons drift toward the wire. On their way, they ionize more

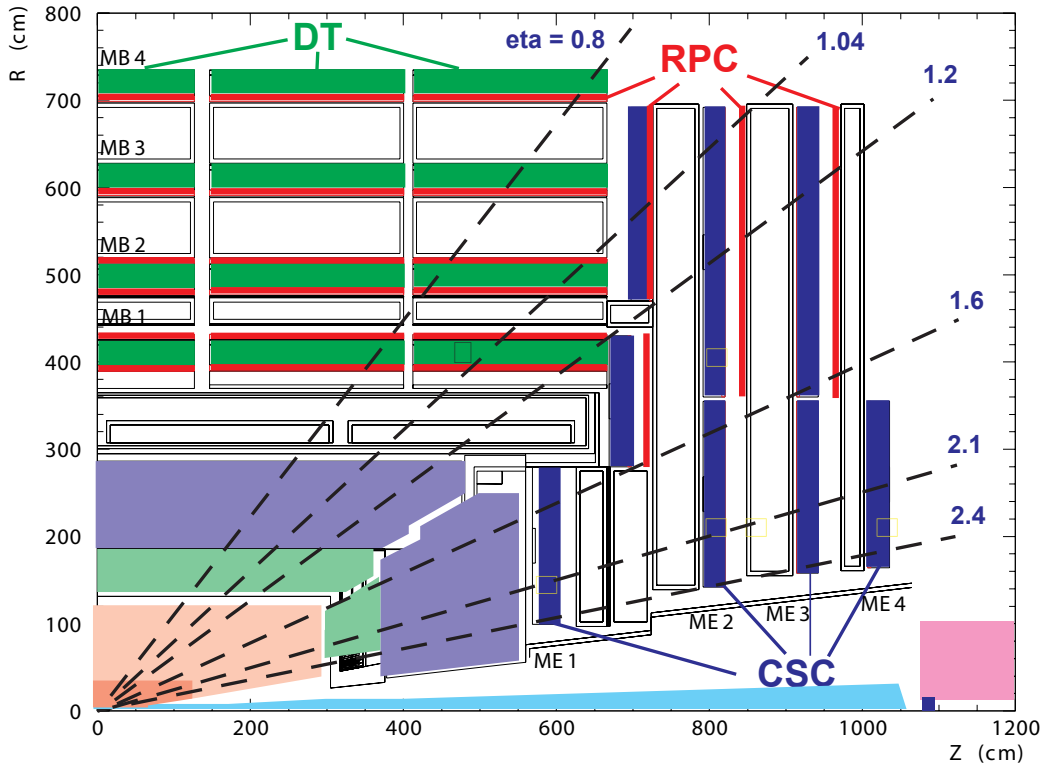

Figure 2.7: Slice in the r-z plane of the CMS detector, with a highlight on the muon system. The white rectangles represent the iron return yoke which is interspersed with the muon chambers. Three different technologies are used: drift tubes (DTs), cathode strip chambers (CSCs) and resistive plate chambers (RPCs). Taken from [60].

gas molecules, so that a cascade of electrons is produced, which then makes a measurable signal.

Both in the barrel and in the endcaps, there are also resistive plate chambers installed as a third detector technology. In this case, the voltage drop is between the two sides of the gaseous chamber which is only a few mm thick. The short drift distance leads to a prompt signal in the detector. The time resolution is of the order of 1 ns, giving valuable information for the first stage of the trigger (L1 trigger), and also to associate a muon measured in the other detectors to the correct LHC bunch crossing.

2.2.6 Data Acquisition and Trigger

When running at its design specification, there are more than 30 million collisions in CMS every second. Recording one bunch crossing event requires data of the order of 1 MiB to be archived, so the full dataset cannot be stored for later analysis. The key to reducing the amount of data is that not every single bunch crossing produces collisions worth studying. Most collisions at the LHC are soft hadronic interactions that are not relevant for most studies. Only once in a while, high-energetic jets or particles such as electrons or muons are produced. Typically such events indicate that an interesting interaction has occurred,

for example that a high-mass resonance was created for a very short amount of time. Therefore, most events are filtered out by an automatic triggering system.

The CMS trigger system consists of two stages. The first stage, the Level-1 (L1) trigger, brings the rate down from 40 MHz to about 100 kHz. It has to analyze every single bunch crossing and needs to make a decision in about 3.2 µs, limited by the hardware buffers in which the event data are kept while the trigger is running. The L1 trigger is implemented in hardware for efficiency reasons. It uses only information from the calorimeters and the muon system, but not from the tracker.

The second stage is the High Level Trigger (HLT), which reduces the data rate to $\mathcal{O}(400\,\text{Hz})$. It runs on a farm of ordinary computers with a special version of the CMS event reconstruction software, optimized for usage in the trigger. The HLT makes use of tracking information as well and, on average, has about 40 ms to make its decision. For single events the time needed might be up to 1 s, though.

CMS defines a variety of triggers for different kinds of analyses. If one of the triggers fires, the event is accepted. The energy thresholds of all triggers are arranged such that the total trigger rate does not exceed the rate that can be processed. At the end of 2012, the final trigger rate was about 400 Hz. Accepted events are transferred to the computing center at CERN for storage and a prompt reconstruction of physics objects so that physicists all around the world can analyze the data in a timely manner. Another 400 Hz was archived without prompt reconstruction, to be used for later analysis after the data taking has finished for the year ("data parking").

3 High Energy Physics Software and Computing

In today's field of High Energy Physics, there are many tasks that can only be performed by computers; the computational complexity has grown so large that these tasks cannot be carried out manually anymore. Software tools have been developed for calculation of Feynman diagrams, simulation of physics processes and particle detectors, data acquisition, storage and synchronization of large amounts of data, analysis of that data, and statistical interpretation of results. Each of these tasks is highly non-trivial and requires sophisticated software to be carried out.

The High Energy Physics community has developed a rich set of software tools and frameworks for their specific needs. Such efforts are necessary because many problems that need to be solved are unique to particle physics, and, therefore, commercial solutions are not available.

In this chapter, individual software packages and algorithms are presented which played an important role in obtaining the results in this thesis. In the following sections, Monte Carlo event generators, the ROOT data analysis framework, the CMS software framework, particle identification algorithms, and the concept of grid computing are discussed.

3.1 Monte Carlo Event Generation

When performing a particle physics experiment, it is essential to know what outcome to expect from the theory of the Standard Model. This allows to tell whether the experimental result is consistent with the theory, or whether it challenges the Standard Model. Most, if not all, observables in a collider experiment are not directly accessible in the Standard Model. While it might be possible to analytically compute the p_{T} distribution of the final state muons in $Z \to \mu^+\mu^-$ events, the situation is different when additional experimental constraints, such as geometric acceptance or analysis cuts, are imposed, or when the detector response introduces additional smearing.

In practice, Monte Carlo methods are used to solve this problem [65]. Instead of an analytical calculation, individual physical processes are simulated using pseudo-random numbers. Many collision events are generated this way, simulating what would happen in the collider experiment. Each such event can be used the same way as a recorded data event, and all experimental observables are immediately accessible. Obviously, the more events are generated, the more precise the prediction will be.

There are three main steps involved before simulated events can be compared to real data taken by the detector:

1. Hard interaction: The hard interaction is typically the process of interest, for example $q\bar{q} \to Z \to \mu^+\mu^-$. Such hard interactions can usually be calculated very well in perturbation theory, since at high momentum transfers the strong coupling constant α_s takes low values and the method converges rapidly. This part of the event simulation

is called "matrix-element" level, because it corresponds to the scattering amplitude computed from the matrix element $\langle\psi_f|H_{\text{int}}|\psi_i\rangle$, where H_{int} is the interaction Hamiltonian. There are various tools available to simulate the hard interaction to leading, next-to-leading or even next-to-next-to-leading order in perturbation theory [66, 67]. Some of these tools are presented later in this section.

2. Soft interactions: In addition to the hard interaction, in proton-proton collisions, there are also various soft interactions and emissions. Since it is not the protons as a whole that enter the hard scattering process, but only the quarks or gluons inside the proton, the remaining partons can cause additional interactions, either with the hard scattering products or other proton remnants. Such interactions of proton remnants are known as the *underlying event*. There can also be additional gluon and quark radiation. Due to the strong coupling constant being high at low momentum transfers, such radiation occurs frequently and tends to be soft. This phenomenon is called *parton shower*. Finally, since only colorless objects can be observed in nature, single quarks and gluons will form hadrons, a process known as *hadronization*.

   All of these soft interaction processes have in common that they cannot be computed with classical perturbation theory. For low momentum transfers, α_s is so high that the series expansion diverges. In principle, these interactions can be computed by solving the Euler-Lagrange equations on a discretized grid of spacetime points. This approach is known as *lattice gauge theory* [68], but with today's knowledge and computing capabilities, it is not possible to describe the interactions in LHC collisions. Instead, one resorts to heuristic models and empiric parameterizations for the description of soft interactions. The parameters are tuned to fit previous experiments and also well-known processes at the LHC.

3. Detector Simulation: This last step takes smearings and geometric effects caused by the experimental apparatus into account. After the hadronization step, a list of stable particles is available, which is used to simulate the response of the detector. The output of the detector simulation can then be processed by the same reconstruction and particle identification algorithms that are also used for data recorded by the experiment. This procedure allows to make a one-to-one comparison between the theoretical prediction and the experimental result.

   Unlike the other two steps, the detector simulation is specific to the experiment. The CMS detector is modeled with GEANT 4 [69], and the detector simulation is fully integrated within the CMS software framework, discussed in Section 3.3. GEANT 4 models the traversal of particles through matter using an accurate model of the geometry of the CMS apparatus.

   The matter interactions, including detailed shower evolutions in the calorimeters, are then converted to simulated detector hits. At the next stage, the readout electronics are simulated, including electronic and thermal noise.

   Depending on the event activity, the detector simulation can take a very long time, on the order of tens of seconds per event. Therefore, another simulation, known as *FastSim*, has been developed, which is up to a factor of 1000 faster. It relies on parameterizations for time-consuming steps such as the shower development in the calorimeters. The parameters have been tuned so that the output is comparable

to the GEANT 4-based detector simulation. However, all simulated samples used throughout this thesis have been created with the full detector simulation.

In the following, various Monte Carlo Event Generator programs are presented. Each of these has strengths in different areas so that they nicely complement each other, and, taken together, provide a very accurate modeling of known particle interactions.

3.1.1 Pythia

PYTHIA [70, 71] is a general purpose Monte Carlo event generator. The current version is PYTHIA 8, which is written in C++. However, PYTHIA version 6.4, written in FORTRAN, is still very widespread, since it has proven its reliability and is very well-accepted in the particle physics community. Also in CMS, PYTHIA 6.4 is used throughout. However, the official development of PYTHIA 6.4 has been stopped and by now PYTHIA 8 provides all features that PYTHIA 6.4 had. It is therefore expected that many physicists will soon migrate to PYTHIA 8.

PYTHIA takes the initial and final state particles as an input, as well as their momentum vectors. It contains a large set of $2 \to 2$ physics processes, such as $e^+e^- \to \mu^+\mu^-$ or $pp \to e^+e^-$, on matrix-element level in leading order (LO) in perturbation theory. From these, it calculates the differential cross section $\frac{\mathrm{d}\sigma}{\mathrm{d}\Omega}$.

For simulating the soft QCD interactions, including initial and final state radiation, PYTHIA has many parameters that can be tweaked. A set of such parameters is commonly referred to as a *tune*. These parameters are determined by making the simulation fit to inclusive proton-proton collision events. The D6T [72] tune has been developed in the pre-LHC era with data from the CERN Super Proton Synchrotron (SPS) experiments and the Tevatron experiments. It is able to describe phenomena observed at these machines up to beam energies of 0.9 TeV, but it was shown to have discrepancies in the charged particle multiplicity at LHC energies [73]. In CMS, the Z2 tune was developed [74], which also fits to early LHC data.

The parton distribution function (PDF) is another parameter in PYTHIA. It describes the distribution of the quarks and gluons inside the proton as a function of their momentum fraction x. Experimental data on PDFs can be determined with proton scattering experiments, for example ep scattering at HERA, $p\bar{p}$ scattering at the Tevatron or pp scattering at the LHC. For all simulated data used in this thesis, the CTEQ6 [75] PDF set is used.

Finally, for the hadronization process, PYTHIA uses the phenomenological Lund string model [76], in which soft gluons are represented by field lines, in a similar way as in electromagnetism. The main difference is that the gluon self-interactions cause field lines to attract each other, leading to "tube"-like structures. When there is enough energy in the gluon field, a quark-antiquark pair is produced. This process is repeated until colorless hadrons are formed, accommodating confinement in QCD.

3.1.2 Madgraph

MADGRAPH is a matrix-element level event generator [77]. It is currently available in version 5, which is used by CMS. The software is written in the C++ and PYTHON programming languages. MADGRAPH implements an algorithm to find by itself all leading order Feynman diagrams contributing to a particular process, and then generates code to integrate over the

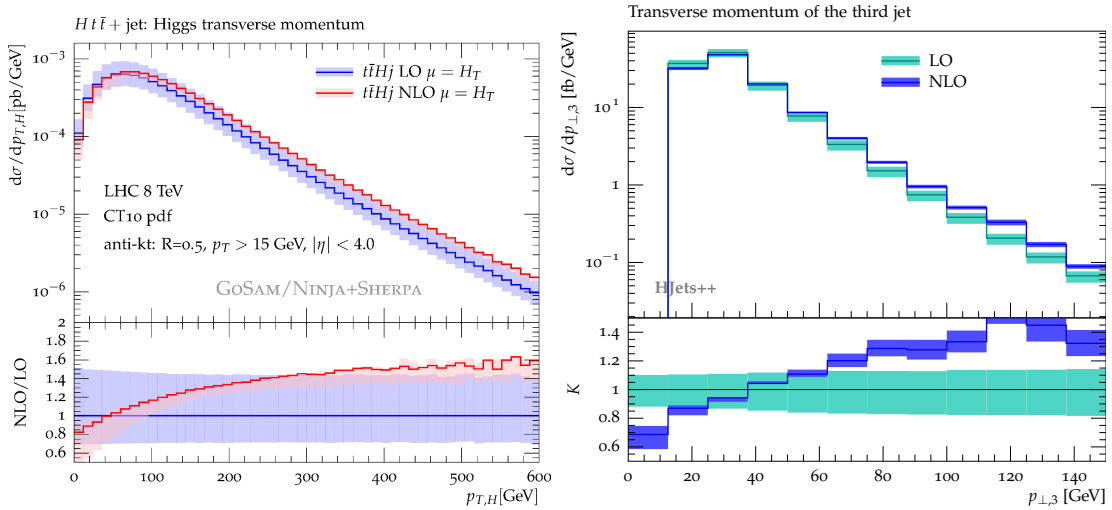

Figure 3.1: Left: Distribution of the transverse momentum of the Higgs boson in $t\bar{t}H$ production. Right: Transverse momentum of the third jet in H + 3 Jets production (VBF). In both cases, NLO effects significantly alter the shape of the distribution. From [87, 88].

phase space and to calculate a cross section. The tool MADEVENT, included in MADGRAPH, can be used for event generation. It can be interfaced to PYTHIA for simulation of the parton shower and the hadronization, which MADGRAPH does not do by itself.

MADGRAPH allows treatment of initial and final state radiation on the matrix-element level, because it can easily handle $2 \to n$ processes. In CMS, this makes it the first choice for analyses that need a very accurate modeling of jet multiplicity and jet-based variables, such as angular distributions between jets.

3.1.3 Powheg

The tool POWHEG (POsitive Weight Hardest Emission Generator) is a next-to-leading order (NLO) event generator on matrix-element level [78, 79]. It supports many Standard Model processes, including but not limited to, $Z/\gamma^* \to \ell^+\ell^-$, $W \to \ell\nu$ [80, 81], di-boson production [82], $t\bar{t}$ production [83] and Higgs production [84, 85, 86].

In CMS, POWHEG is used for simulation of processes when NLO effects in differential distributions of interest play a significant role. If NLO effects are only significant for the total cross section of the process, a LO event generator can be used and the events can be re-weighted to the NLO or even the NNLO cross section calculated with a tool such as MCFM [89]. However, in other cases, NLO effects can make a difference in the shape of an observable, which is then used to make an experimental cut. In this case, a simple event weight is not enough to correct for the effect. Instead, the events are generated at NLO matrix-element level with POWHEG. Figure 3.1 shows two examples where NLO effects are significant for the shape of two distributions.

As with MADGRAPH, POWHEG can be interfaced to PYTHIA for parton shower and hadronization.

3.1.4 Tauola and TauSpinner

TAUOLA is a software for simulating tau lepton decays [90]. In principle, PYTHIA can perform this task, too, but it only covers the most important decays, and, most importantly, it does not take into account spin correlations between the two taus. TAUOLA supports more than 20 decay modes of the tau lepton, and it takes spin correlation effects into account. It is written in FORTRAN, and can easily be interfaced to PYTHIA by instructing PYTHIA not to simulate tau lepton decays, and then using TAUOLA to do the job.

All official CMS simulations use TAUOLA for tau decays when there are tau leptons in the final state. Since version 8.150, PYTHIA also has a sophisticated tau lepton decay algorithm that is comparable to TAUOLA [91].

TAUSPINNER is a tool created by the same main authors as TAUOLA [92]. It can be used to compute event weights to enable spin correlations in a sample of events that were originally simulated without spin correlation effects. TAUSPINNER only needs the four-momenta of the tau leptons and their decay products as input. The initial quark state in, for example, $q\bar{q} \to Z/\gamma^* \to \tau^+\tau^-$ events, is reconstructed on a statistical basis from the kinematic constraints and from the proton PDF.

3.2 Data Analysis with ROOT

ROOT is a framework for large-scale data analysis [93]. It is written in C++ and can be regarded as the successor of the FORTRAN-based PAW. ROOT is not a standalone program, but a collection of common tools that are frequently needed for data analysis. These tools include, but are not limited to, data visualization, reading and writing large datasets from disk or tape, fitting models to data, calculating errors and confidence intervals, and numerical algorithms for computing mathematical functions and integrals. Most figures in this thesis have been created with ROOT.

ROOT provides classes for various objects such as vectors, graphs, histograms or n-tuples. These classes allow performing many operations on these objects with a simple interface. In the following, a brief summary is given about the most important objects in ROOT.

- Histograms: Histograms are represented by the `TH`*`X`*`Y` classes, where X is a number indicating the dimensionality and Y an identifier for the type, such as `D` for double precision floating point numbers. A histogram is an object which stores event counts in bins, modeling a differential distribution when only a finite sample of data points is available, such as from a physics experiment. Events in histograms can be weighted, and ROOT can then be instructed to also store the sum of squares of the weights in each bin, so that the Gaussian error in each bin can be computed.

  The ROOT histograms allow easy fitting of functions with many parameters to the content of the histogram. Both χ^2 and maximum likelihood fits assuming Poisson statistics in each bin can be used. For more elaborate cases, ROOT also allows to specify the function to be minimized by hand, for example to perform a maximum likelihood fit with a probability density function other than the Poissonian. The actual minimization is carried out with the MINUIT algorithm [94], however there are other minimizers available.

- Profile Plots: A profile plot is a special case of a histogram, showing the mean and RMS of one variable as a function of another. In many cases, it can replace a 2-

dimensional histogram, and it is more straight-forward to visualize and interpret. The ROOT class `TProfile` can be used in many ways like a 2-dimensional histogram. Also, a 2-dimensional histogram can be automatically converted into a profile plot when needed.

- Graphs: A ROOT `TGraph` is a series of x-y coordinate pairs. A graph can be used when one variable is plotted against another. Unlike a histogram, a graph can have asymmetric errors in both x and y. The same fitting procedures as for histograms are also available for graphs.

- Trees: A ROOT tree is a generalization of the concept of an n-tuple. Instead of only numbers, also more complex objects can be stored in a tree, such as vectors or strings. Trees are often used to store experimental data, where each entry of the tree corresponds to one dataset, for example from one collision event. Each dataset is then represented by a tuple of numbers or more complex objects.

  ROOT trees are represented by the `TTree` class which provides an interface for common operations, such as reading and filling the tree, drawing the distribution of a variable or visualizing the correlation between two variables. What makes ROOT trees very powerful, however, is that they can very efficiently be read and written to disk. ROOT easily handles trees with multiple gigabytes of data. Both raw and processed data taken by CMS is stored as ROOT trees, and it is the same for many other high energy physics experiments.

3.2.1 Multivariate analysis with TMVA

Consider a set of events, each of which can be either classified as signal or background. In a Higgs analysis, typically Higgs decay events are treated as signal, and events originating from other Standard Model processes are considered background. Classically, one performs cuts on various observables in the event to separate signal-like events from background-like events. In a *multivariate analysis* (MVA) approach, however, one tries to combine all available observables into one new variable which contains all information about background separation. In this way, it is possible to exploit the correlation between variables which is not easy to do in a cut-based approach. Sometimes, multivariate techniques are also referred to as *Machine Learning*.

The TMVA package is a framework for multivariate analyses especially designed to work well together with ROOT [95]. It provides a common interface to many different multivariate techniques such as likelihood ratio, k-nearest neighbour estimators, support vector machines, neural networks or decision trees. All methods have in common that they need to be trained. The training can only be performed on a sample where it is known whether an event is signal or background. Typically, such training samples are obtained from Monte Carlo simulation. The method then "learns" by itself which observables or combination of observables provide a good separation. Once the method has been trained, it can be applied to real events, and for each event, it produces one number which tells how signal-like or background-like the event is. This number can then be used for further analysis, and typically it is more powerful in terms of background separation than any cut-based approach.

Apart from a common interface to train and apply different MVA methods, TMVA also allows to easily compare different methods by running more than one method on the

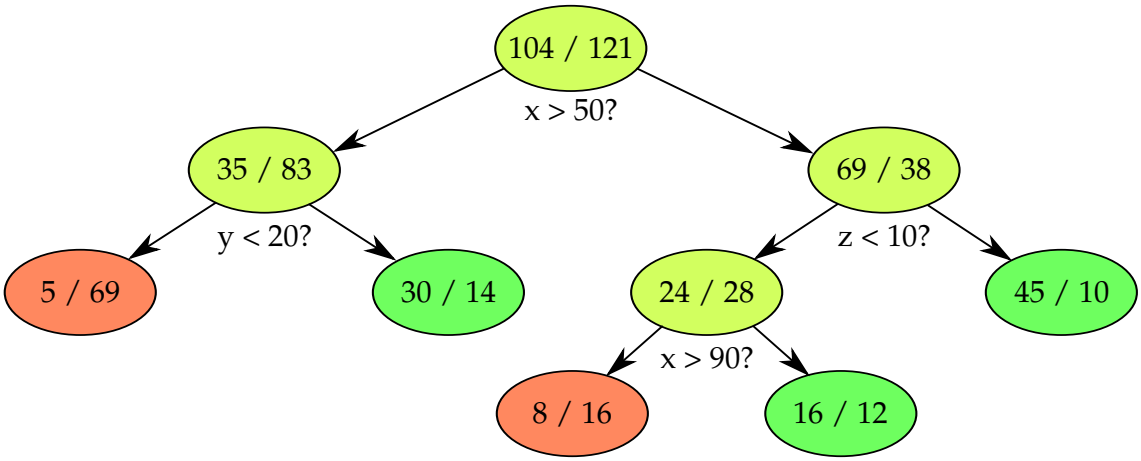

Figure 3.2: Example for a decision tree. The numbers in each node refer to the number of signal or background events in the training sample. Cuts on various variables, here x, y and z, lead from one node to the next. The leaf nodes are either mostly signal (green) or mostly background (red). A test event that ends up in a green (red) node is considered signal-like (background-like).

same set of data. It provides mechanisms to evaluate the trained methods, making sure that the method did not only learn statistical fluctuations in the training sample, known as *overtraining*. This is checked by training on one half of the sample, and applying the method to the second half. If the shape of the resulting MVA discriminant is not compatible between the two, some sort of overtraining has occurred. This can happen especially with training samples that only have a small number of events.

The input data is given to TMVA in the form of ROOT trees.

Boosted Decision Trees

A *boosted decision tree* (BDT) is an MVA method that is very widely used in CMS algorithms and analyses. Figure 3.2 shows a schematic drawing of a decision tree based on three variables x, y and z. To decide whether an event is signal-like or background-like, first the x variable is compared to the value 50. Depending on the outcome the event ends up in the node to the left or to the right. The procedure is repeated until the event ends up in a leaf node which is then either classified as background (red) or as signal (green).

The tree itself is constructed using the training sample. At each node, only one variable is used to discriminate between signal-like and background-like events. The variable which provides the best separation between signal and background events in the training sample is chosen. When a certain criterion is reached, such as a low number of training events left, or a high signal or background purity, the procedure is stopped. At each leaf node, the ratio of signal to background events decides whether the node is considered signal-like or background-like.

A *boosted* decision tree is actually not a single tree, but a collection of many trees, known as a *forest*. The first tree is constructed with the procedure as described above. Then, the boosting takes place. Boosting means that another tree is trained where those training events which were classified incorrectly in the first tree are assigned a higher weight. This way, in the boosted tree, an extra effort is made to reduce the number of misclassifications.

The boosting procedure is repeated many times; the default number in TMVA is 400. There are various ways of how the boosting itself is performed. Available choices in TMVA include adaptive boosting (AdaBoost) [96] and Gradient Boosting [97].

In order to obtain the result from the boosted decision tree, a majority vote of all trees in the forest is cast. This can be as simple as the ratio of the number of signal trees over the number of all trees.

Decision trees by themselves are subject to overtraining. A small fluctuation in the training sample can cause a decision to be made on an otherwise insignificant variable. This can be improved when making sure that at each node there are enough events available to tell a significant separation between signal and background from an insignificant one.

3.3 The CMSSW Framework

CMSSW is the CMS software framework [60]. It is used to process CMS data at all stages, including the High Level Trigger, generation of Monte Carlo events, event reconstruction, and physics analyses. The framework consists of one executable, `cmsRun`, which is steered with configuration files written in PYTHON.

CMSSW is shipped with a copy of external programs, such as C++ and FORTRAN compilers or Monte Carlo generators. This makes sure that the same version of CMSSW always uses the same version of external software, so that they are always compatible with another.

3.3.1 The Event Data Model

CMSSW processes the data by always handling one collision event at a time. There is no inter-dependency between events. This allows for a trivial parallelization of any task: a given set of events can be split into small groups which are then processed in parallel by many `cmsRun` jobs. In more recent versions of CMSSW, it is also possible that a single `cmsRun` job processes multiple events in parallel. This procedure allows to utilize many-core CPUs while, at the same time, sharing the global state that is common for all events.

The actual work is then carried out by modules which themselves are written in C++. There are the following different types of modules:

- Source: A `Source` module must be the first module executed. It creates the event, for example by reading a `.root` file, or by creating an empty event that can then be filled with a Monte Carlo generator.
- EDProducer: A `EDProducer` module can read data from the event and use it to write new data back into the event. The produced data can then be used by other modules. For example, a tracking module can read the hits in the inner tracker from the event and write back the list of reconstructed tracks.
- EDFilter: A `EDFilter` module can read data from the event to make a decision whether to filter the event or not. This can be used in physics analyses to reject events that do not fulfill certain criteria, such as two reconstructed, opposite-charge muons existing in the event.
- EDAnalyzer: A `EDAnalyzer` module can read data from the event to analyze them. It can not write data back into the event, but it can create histograms and fill ROOT trees for later analysis and inspection.

- OutputModule: An output module writes the event into some sort of storage, such as a `.root` file, in general including all data that has been produced by `EDProducer`s. However, the products to store can be explicitly specified, for example to save disk space by not storing information that is no longer needed. Usually, after the reconstruction of physics objects has been performed, the raw detector output is no longer needed.

The scheduling of the modules is arranged in paths. Each path is a linear sequence of modules which are processed one after another. When the same module appears multiple times in different paths, it is only executed once. Every module can access data that is either produced by the event source, for example by reading from a file, or by a `EDProducer` which ran before the module itself runs. If a module relies on products produced by a `EDProducer`, the producer must be present in all paths where the module is used.

When all paths have been executed, the event is written by the `OutputModule`, if there is any, and if the event was not filtered by a `EDFilter`. Afterwards, the next event is processed.

CMS datasets are organized in three major data tiers which define what kind of data is stored.

- `RAW`: The `RAW` data tier contains all the raw detector output as it was acquired by the DAQ, without further processing.
- `RECO`: The `RECO` data tier contains reconstructed objects, such as reconstructed hits in the tracker and calorimeters, reconstructed tracks, particle flow objects, jets, and so on.
- `AOD`: The `AOD` (analysis object data) data tier is a subset of `RECO` for use in physics analyses. It does not contain the low-level objects such as reconstructed hits but only high-level objects like reconstructed tracks and particles.

3.3.2 The Conditions Database

Often, a job does not only need to access data specific to an event, but needs additional information. This is the case for example for calibration settings, alignment constants and correction factors. In addition, these values can be time-dependent since the calibration might change over time. For simulated samples, completely different constants might be needed.

In CMS, these values are stored in the *Conditions Database* which is hosted at CERN and can be replicated at other sites. The actual payloads in the database have a tag and an interval of validity assigned to them, so that for different collision runs different values can be assigned. Every job can then access the database and query specific data for a given run.

3.4 Particle Reconstruction and Identification

In this section, the reconstruction and identification of physics objects at CMS is discussed. The reconstruction proceeds in several steps. In the first step, the raw detector output is converted to hits which are assigned a position in space. Adjacent hits are merged

to clusters. Next, in the inner tracker and the muon system, hits in the various layers are correlated to reconstruct particle tracks. In the third step, particle candidates are reconstructed by linking together information from different sub-detectors, such as a track in the tracker with a cluster from one or both calorimeters. Finally, such particle candidates are either recognized as individual high-energetic particles such as electrons or muons, or they are used to make composite objects like jets and hadronically decaying tau leptons.

In the following, these steps are discussed in more detail.

3.4.1 Track and Vertex Reconstruction

The reconstruction of charged particle trajectories in the inner tracker is performed with the Combinatorial Track Finder (CTF) algorithm [98]. It starts with seeds of 3 hits in the innermost layers. The initial trajectories are then propagated outwards, taking into account the magnetic field inside the CMS detector and multiple scattering at detector material. New hits from outer layers are then added iteratively. In the next layer, all hits compatible with the current trajectory are considered, and, if there are more than one, multiple candidate tracks are constructed. Ambiguities are then resolved at a later stage. The current trajectory and its error estimation is updated with the new hit using the Kalman Filter method [99]. The procedure stops when the end of the tracker has been reached or no hits are found in two consecutive layers. Eventually, the full list of hits that has been determined is fit with the least squares method, in order to obtain full information on the track parameters.

The reconstructed tracks are used to find primary interaction vertices. For this purpose, all tracks are extrapolated to the interaction region, and if multiple tracks intersect at the same point, a vertex can be reconstructed. In CMS, the *Deterministic Annealing* algorithm is implemented [100]. In this algorithm, tracks are assigned to vertices with a particular weight which can be computed from the compatibility of the track parameters with that vertex. It is allowed for one track to be part of multiple vertices.

3.4.2 Particle Flow

The *Particle Flow* (PF) approach attempts to combine information from all CMS subdetectors, i.e. silicon tracker, electromagnetic and hadronic calorimeters, and muon system [101, 102]. Any of the sub-detectors by itself cannot unambiguously identify particles: for example, all charged leptons produce a signal in the inner tracker, and both electrons and photons produce showers in the electromagnetic calorimeter. The ultimate goal of the PF algorithm is to reconstruct every individual stable particle that has crossed the detector by combining the signals in all subdetectors in an optimal way. An important property of the algorithm is that every detector signal is only attributed to one reconstructed particle, so that any form of double-counting is avoided. In the traditional approach where each detector component is evaluated individually, this has to be taken care of manually for particles that leave signatures in more than one component.

Figure 3.3 shows how the different kinds of particles leave different signatures in the detector:

- Charged Hadrons (solid green): Most hadrons in CMS are pions and kaons. They produce a track in the inner tracker, almost no signal in the electromagnetic calorimeter, and are stopped in the hadronic calorimeter.

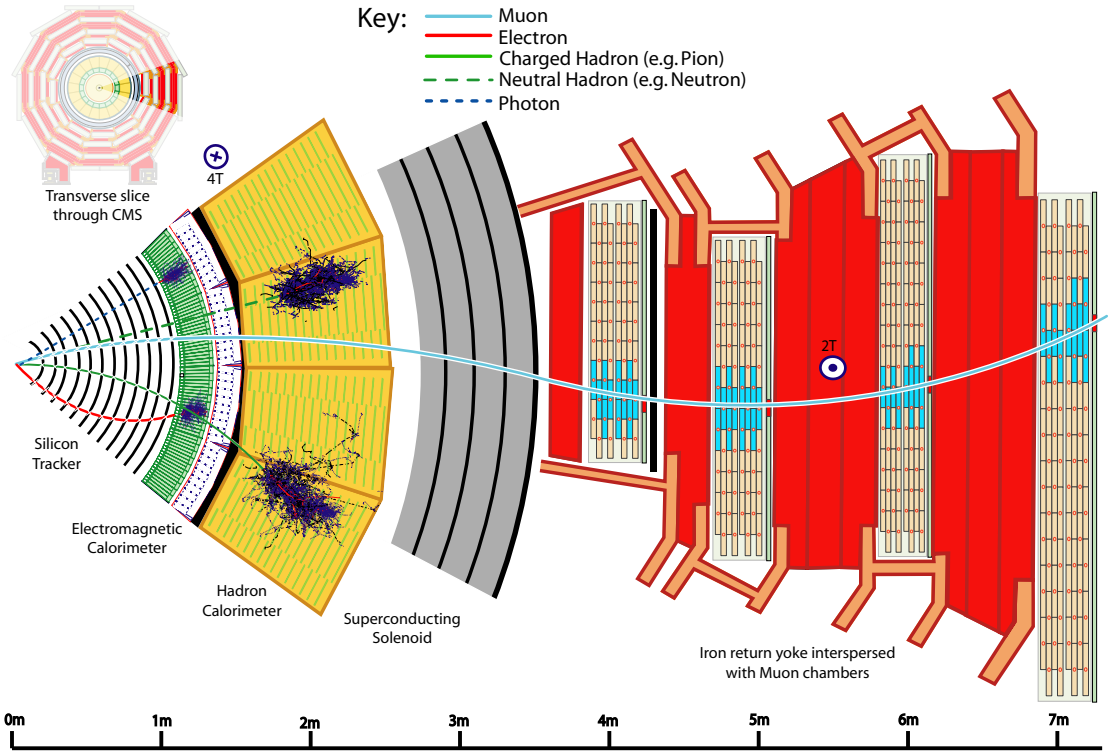

Figure 3.3: A slice through the central region of the CMS detector. It can be seen how different kinds of particles leave different signatures in the detector. The PF algorithm attempts to use information from all sub-detectors to identify individual particles. From [103].

- Neutral Hadrons (dashed green): Neutral hadrons such as neutrons and neutral kaons leave a signature only in the hadronic calorimeter.
- Photons (dashed blue): Photons do not leave a signature in the tracker but are stopped in the electromagnetic calorimeter.
- Electrons (solid red): Electrons and also positrons produce a track in the silicon tracker and are stopped in the electromagnetic calorimeter.
- Muons (solid blue): Unlike the electron, the muon is not stopped in the electromagnetic calorimeter due to its higher mass. Therefore, it is the only particle that can make its way to the outer muon chambers and every signal in the muon system is attributed to a muon.

3.4.3 Jets

High-energetic quarks and gluons in the final state manifest themselves as a bundle of particles in the same direction. In physics analyses, often the properties of the original quark or gluon are of interest. Therefore, all particles in such a jet are clustered together

and by summing up their four-momenta one obtains an estimate for the four-momentum of the quark or gluon.

There are various ways to cluster the particles that belong to a jet. Such jet algorithms are typically required to fulfill two important properties:

- Infrared Safety: A jet algorithm is called *infrared safe* when additional very low-energetic particles in the input do not alter the output of the algorithm. This requirement is due to the fact that quarks and gluons frequently emit additional soft gluons. The goal of the algorithm is to reconstruct the original parton, so no matter whether it radiated a soft gluon or not, the algorithm output must be the same.
- Collinear Safety: A jet algorithm is called *collinear safe* when it gives the same result for one high-energetic particle and for two particles that are collinear to each other where each of the particles carries a fraction of the total energy. This can easily happen, for example when calorimeter deposits are split between adjacent calorimeter cells, or when a high-energetic quark or gluon is radiated.

Traditionally, cone-based jet algorithms were used to perform the jet clustering. Starting from a so-called seed, such as the highest energetic object, all objects within a particular distance ΔR in η-ϕ-space are added to the jet, where ΔR is a parameter of the algorithm and must be optimized for the individual application. Apart from the obvious questions of how to choose the seed and how to deal with overlapping cones, many cone algorithms turn out not to be infrared or collinear safe. The Seedless Infrared-Safe Cone algorithm, SISCone [104], is an example of a cone algorithm that fulfills both requirements. Asymptotically, its runtime is $\mathcal{O}\left(N^2 \log N\right)$ when N is the number of objects to be clustered.

Recently, sequential algorithms have become more popular, since they are intrinsically infrared and collinear safe. Given a list of objects (such as particles), the algorithm works as follows. For each object i, a distance measure with respect to the beam d_i, and for every pair of objects (i, j), a distance measure between the two objects d_{ij}, is defined. In an iterative procedure, the minimum of all d_i and d_{ij} is chosen. If the chosen object is a d_{ij}, the two objects are merged together into a jet, and if it is a d_i, then the jet with index i is declared final and removed from the list. The procedure is repeated until there are no objects left. This algorithm can be implemented in $\mathcal{O}\left(N \log N\right)$.

When R is a parameter of the algorithm, p_{T}^i and p_{T}^j are the transverse momenta of objects i and j, and ΔR_{ij} is the geometrical distance between the objects in the η-ϕ plane, popular choices for d_i and d_{ij} include

- The k_{T} algorithm [105]:

$$d_i = \left(p_{\mathrm{T}}^i\right)^2,\ d_{ij} = \min\left(\left(p_{\mathrm{T}}^i\right)^2, \left(p_{\mathrm{T}}^j\right)^2\right)\Delta R_{ij}/R \tag{3.1}$$

- The Cambridge / Aachen algorithm [106, 107]:

$$d_i = 1,\ d_{ij} = \Delta R_{ij}/R \tag{3.2}$$

- The anti-k_{T} algorithm [108]:

$$d_i = 1/\left(p_{\mathrm{T}}^i\right)^2,\ d_{ij} = \min\left(1/\left(p_{\mathrm{T}}^i\right)^2, 1/\left(p_{\mathrm{T}}^j\right)^2\right)\Delta R_{ij}/R \tag{3.3}$$

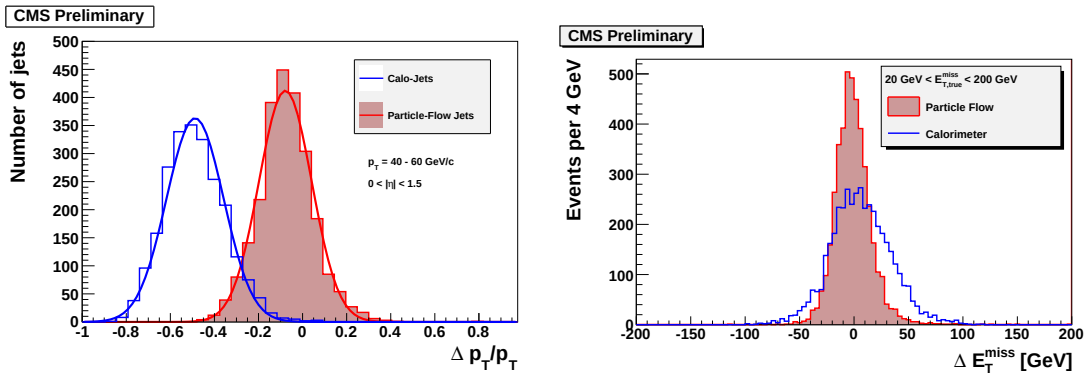

Figure 3.4: Left: Comparison between jets when constructed from calorimeter information only (blue) or from PF candidates (red). In simulated QCD multijet events the relative difference between reconstructed energy and true energy is shown for central jets between 40 GeV and 60 eV. Right: The absolute difference between reconstructed and true $E_{\mathrm{T}}^{\mathrm{miss}}$ is shown in simulated $t\bar{t}$ events. From [101].

Jets clustered with the k_{T} algorithm start with clustering the softest object first, which leads to them having an arbitrary shape, i.e. in general they are not conical. The anti-k_{T} algorithm on the other hand starts with the hardest object, which always produces conical jets and, therefore, can be regarded as a "perfect" cone algorithm.

Many of these algorithms are implemented in software in the FASTJET package [109]. In CMS, the FASTJET implementation is used and the standard jet algorithm is anti-k_{T} with $\Delta R = 0.5$. The algorithms themselves can be applied to different kinds of objects. Traditionally, jets are built from calorimeter clusters, but the algorithms work just as well for tracks. In CMS, particle candidates identified with the particle flow algorithm are used for most analyses.

Picture 3.4 shows on the left hand plot the relative difference between reconstructed and true transverse momentum for calorimeter-based and PF-based jets. It can be seen that the resolution, represented by the width of the curve, is better when combining all detector information than when only using the calorimeters. Also, the peak itself is better centered around 0, whereas for calorimeter jets, the jet momentum tends to be underestimated.

Jet Energy Corrections

The measured energy of a jet does not necessarily correspond to the true energy of the original parton. For example, Figure 3.4 shows that, especially for calorimeter based jets, the energy is underestimated in CMS. Therefore, the measured jet energy needs to be corrected before it can be used by a physics analysis. At the LHC, this calibration can be performed with $Z/\gamma^{+} \rightarrow \mu^{+}\mu^{-}$ events where the boson is produced in association with one jet [110]. The energy of the muons can be measured very precisely, and, due to momentum conservation, the transverse momentum of the associated jet must be the same. From many such events, a correction factor can be calculated to move the average reconstructed jet momentum to the true value. A similar procedure can also be performed with di-jet events and γ + jet events.

B-Tagging

In many CMS analyses, b-quarks play a special role. This includes all final states with top quarks, as the top quark immediately decays into a b-quark and a W boson before it hadronizes. In the context of Higgs analyses, b-quarks are interesting for searches of a supersymmetric Higgs boson whose production with associated b-quarks is enhanced. When produced, b-quarks immediately form B mesons which have a relatively high lifetime of the order of picoseconds. So-called b-tagging techniques attempt to exploit this fact to separate jets induced by gluons or lighter quarks from jets that originate from a b-quark (b-jets).

In CMS, the *Combined Secondary Vertex* (CSV) algorithm is used for b-tagging [111]. First, the reconstructed tracks inside the jet are used to reconstruct a secondary vertex which is several millimeters offset from the primary vertex, using an adaptive vertex fitter [112]. This is taken to be the decay vertex of the B meson. Even if no secondary vertex can be reconstructed, information can be obtained from the 3D impact parameter of the tracks with respect to the primary vertex. The CSV algorithm combines these information into a likelihood where the probability densities are taken from MC simulation. Two likelihood ratios are constructed: one which discriminates b-jets from charm-quark induced jets and one that discriminates b-jets from jets induced by a gluon or a lighter quark.

There are several working points available. For example, for a 85% identification efficiency of b-jets, the light parton misidentification rate is 10%. The performance of the b-tagging in CMS was studied in more detail with $\sqrt{s} = 8$ TeV data taken in the year 2012 in [113].

3.4.4 Missing Transverse Energy

The total momentum must be conserved in every collision event. Since, at the LHC, both beams are symmetric, the initial momentum is 0, so it must be 0 after the collision as well. In the direction of the beam, the two interacting partons can carry two different momentum fractions of the proton momentum, $x_1 \neq x_2$, and so the total momentum of the hard scattering can be nonzero in the longitudinal direction. In the transverse direction, however, the total momentum must be 0. The missing transverse energy is defined as

$$\vec{E}_{\mathrm{T}}^{\mathrm{miss}} = -\sum_{\text{particles}} \vec{p}_{\mathrm{T}}\,, \tag{3.4}$$

where the sum is over all particles in the event. When this quantity is far away from 0, it typically means that a particle has escaped the detector without being detected. Since the detector covers almost the full region around the interaction point, the only particle which can go away undetected is the neutrino which does hardly interact with matter at all.

As with jets, the $E_{\mathrm{T}}^{\mathrm{miss}}$ value can be computed with any types of objects. In CMS, PF particles are used for most analyses. The right hand plot in Figure 3.4 shows the improvement in resolution with respect to using only calorimeter information for the construction of $E_{\mathrm{T}}^{\mathrm{miss}}$.

Both jets and $E_{\mathrm{T}}^{\mathrm{miss}}$ can be very sensitive to particles from additional soft interactions in the same bunch crossing ("pile-up"). CMS has developed techniques to mitigate these which are described in more detail in Appendix A.

3.4.5 Electrons

Electrons in CMS are interacting with the tracker material before they reach the electromagnetic calorimeter. About 35% of electrons will have lost 70% of their energy due to bremsstrahlung before being stopped in the calorimeter [114]. This leads to a spread of the energy in ϕ direction. Therefore, super-clusters (clusters of clusters) are being reconstructed in the calorimeter, to not only collect the energy of the electron itself but also of all bremsstrahlung photons.

When a super-cluster has been found, it is used to seed the track reconstruction. The track reconstruction for electrons proceeds in a similar way as described in Section 3.4.1, however the Kalman Filter is replaced by a dedicated "Gaussian Sum Filter" (GSF) [115], which is used to handle non-Gaussian fluctuations from bremsstrahlung emissions which are modeled according to the Bethe-Heitler formalism [116].

A BDT is trained in order to discriminate real electrons from misidentified electrons, for example a jet which leads to a similar signature in the detector [117]. As training events, Monte Carlo simulation of the $Z/\gamma^* \to e^+e^-$ process has been used for the signal whereas inclusive Z production in data was used for the background sample. There are many variables used for discrimination, including the following types of variables:

- Pure tracking variables such as the χ^2 of the track fit or the energy loss due to bremsstrahlung.
- Pure calorimeter variables characterizing the shape of the shower, including the width of the super-cluster in η and ϕ.
- Geometrical matching between the track and the super-cluster, both at the interaction vertex and on the calorimeter surface.
- Energy matching between tracker and calorimeter, such as the ratio of the momentum measured in the tracker and the energy deposited in the calorimeter.

Figure 3.5 shows the performance of the electron identification BDT on a testing sample.

In addition to being created by the primary interaction, electrons can also be produced by highly energetic photons which undergo pair production when they interact with the tracker material. Such electrons are typically not interesting in analyses, since in the primary interaction a photon was produced. In order to reject such electrons, a conversion veto is applied, requiring that there are no missing hits in the innermost layer of the pixel detector. This rejects all conversions except those taking place at the first tracker layer or at the beampipe. Such events are rejected by pairing the electron track with other tracks and imposing geometrical constraints on the track pair, such as the opening angle in the r-z plane and the distance of closest approach.

3.4.6 Muons

Muons are reconstructed both in the inner tracker and the muon chambers. In both cases, the Kalman Filter technique as described in Section 3.4.1 is applied. The two tracks are then matched with each other, starting from the outer track which is extrapolated to the tracker volume. If a matching track in the inner tracker is found, a global fit with hits both in the inner tracker and the muon system is performed. The momentum assigned to the muon is taken from the track measurement in the inner tracker, since it has a superior

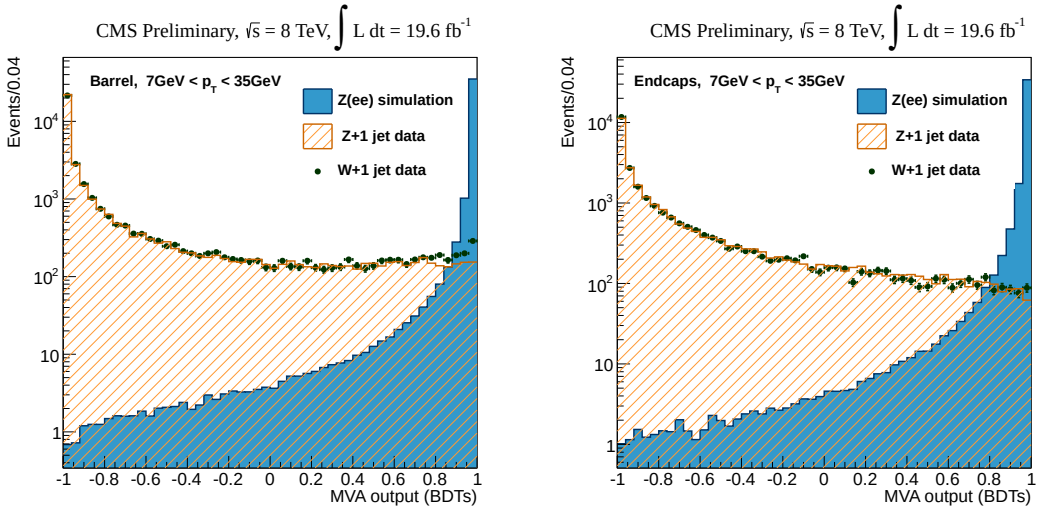

Figure 3.5: Output of the BDT for electron identification, for electrons detected in the ECAL barrel (left) and for electrons detected in the ECAL endcaps (right). The signal peaks at +1 while the background accumulates around −1. From [117].

resolution compared to the muon system. Only for high-p_{T} muons, $p_{\mathrm{T}} \gtrsim 200\,\mathrm{GeV}$, the result from the global fit is used, since the radius of the curvature of the muon track becomes too large for an accurate momentum measurement in the relatively small volume of the inner tracker.

In another approach, each track in the inner tracker is extrapolated to the muon chambers, and if at least one muon station matches the track, it is classified as a "tracker muon". The reconstruction efficiency of this method is typically higher for low-p_{T} muons, however, the reconstruction efficiency for muons within the geometrical acceptance of the detector and with $p_{\mathrm{T}} > 5\,\mathrm{GeV}$ either as a "global muon" or a "tracker muon" is higher than 99% [118].

The particle flow algorithm uses the muons reconstructed as described here to identify both isolated muons and muons in jets, and it avoids assigning muon tracks to other particles.

In order to suppress charged hadrons to be mis-identified as muons, identification of muons from cosmic rays or muons produced from pions or kaons decaying in flight while traversing the detector, additional identification criteria are applied. There are different sets of criteria for different analyses. In many Higgs analyses which feature isolated muons with reasonable high p_{T}, $p_{\mathrm{T}} \gtrsim 10\,\mathrm{GeV}$, the "tight ID" is used with the following requirements:

- The muon is reconstructed as a "global muon".
- The muon is identified by the particle flow algorithm.
- The χ^2/N_{dof} of the global track fit is less than 10.
- At least one hit in the pixel detector.
- At least 5 inner tracker layers that are used in the global track fit.

Table 3.1: Tau decay modes reconstructed by CMS. Some decays produce an intermediate vector meson which can be exploited for tau reconstruction. Charge conjugate states are implied. The neutral pions decay immediately into two photons. From [119].

Final State	Resonance	Branching Fraction [%]
$e^-\nu_\tau\bar{\nu}_e$	-	17.83 ± 0.04
$\mu^-\nu_\tau\bar{\nu}_\mu$	-	17.41 ± 0.04
$\pi^-\nu_\tau$	-	10.83 ± 0.06
$\pi^-\pi^0\nu_\tau$	$\rho(770)$	25.52 ± 0.09
$\pi^-\pi^0\pi^0\nu_\tau$	$a_1(1260)$	9.30 ± 0.11
$\pi^-\pi^+\pi^-\nu_\tau$	$a_1(1260)$	8.99 ± 0.06

- At least one muon chamber hit is used in the global track fit.
- The muon track is matched to muon segments in at least two muon stations. This requirement implies that the muon is also reconstructed as a "tracker muon".
- The transverse impact parameter with respect to the interaction vertex is less than 2 mm.
- The longitudinal impact parameter with respect to the interaction vertex is less than 5 mm.

3.4.7 Tau Leptons

Tau leptons are very short-lived particles that cannot be observed directly. They decay either into lighter leptons – electrons or muons – or hadronically. Hadronic tau decays can be mediated via very short-lived resonances and in the final state there are one or three charged hadrons (typically pions or kaons) and neutral pions. The neutral pions immediately decay into two photons, whereas the charged hadrons live long enough to be detected directly. In each tau lepton decay, there is also one or two neutrinos produced, depending on whether the decay is leptonic or hadronic. Neutrinos do not interact with the detector material and therefore escape undetected. In the following, hadronic tau decays are referred to with the symbol τ_{had}. Table 3.1 lists the most important decay modes.

Leptonic tau decays are very hard to distinguish from prompt leptons. In many cases, the event topology can provide an additional clue. For example, when there is a τ_{had} and a muon in the event, it is very likely that it originates from a $Z \to \tau^+\tau^- \to \tau_\mu + \tau_{\text{had}}$ process, since the Z boson does not decay into leptons of different flavor. Also, a significant amount of $E_{\text{T}}^{\text{miss}}$ in the event can be a hint for tau lepton decays. The lifetime of the tau lepton is $c\tau = 87\,\mu\text{m}$. This is short, but macroscopic, especially for boosted tau leptons with a high γ factor. It is very hard to exploit this short lifetime, since the impact parameter or the distance between the secondary vertex and the primary vertex is just on the edge of the tracking accuracy. Nevertheless, CMS has measured the $Z \to \tau^+\tau^-$ cross section at $\sqrt{s} = 7\,\text{TeV}$ in the $\tau_\mu + \tau_\mu$ final state [120], and this channel also contributes to the $H \to \tau^+\tau^-$ search [35].

The challenge in the reconstruction of hadronic tau decays is the similarity of the signature with quark-induced or gluon-induced jets. A hadronic tau lepton features one or

three charged particle tracks, and activity both in the electromagnetic calorimeter (from the neutral pions) and the hadronic calorimeter (from the charged pions). In CMS, the hadronic tau reconstruction is seeded by jets made from particle candidates. The "Hadron Plus Strips" (HPS) algorithm [121] then tries to reconstruct the decay mode of the tau lepton.

The HPS algorithm attempts to take into account the conversion of photons to electron-positron pairs in the tracker, which leads to a spread of energy in the calorimeter in ϕ direction, due to the magnetic field bending the electron and positron trajectories. The algorithm starts with the highest-energy electromagnetic object in the jet and adds other photons or electrons in a window of $\Delta\eta \times \Delta\phi = 0.05 \times 0.20$ to the object until there are no more candidates left. In the next step, the reconstructed strips and charged particles in the jet are combined together to make a τ_{had} candidate. The decay modes in Table 3.1 are attempted to be reconstructed, where the $\pi^{\pm}\pi^0$ and $\pi^{\pm}\pi^0\pi^0$ modes are considered together (two strips could be two neutral pions or the well separated electron and positron from one neutral pion). Allowed combinations must fulfill the constraint that the invariant mass of two objects be compatible with that of an intermediate meson (π^0, $\rho(770)$ or $a_1(1260)$). If there is more than one allowed combination, the one which leads to the higher transverse momentum of the τ_{had} candidate is chosen.

When the τ_{had} candidate has been reconstructed, further requirements are made to prevent electrons or muons to be mis-reconstructed as hadronically decaying tau leptons. The candidate is rejected when there is a track segment in the muon system or when there is signal above the noise level present in the CSC, DT or RPC modules located in the two outermost muon stations within a cone of $\Delta R = 0.5$ of the τ_{had} direction. It is also rejected if the energy in the electromagnetic and hadronic calorimeters assigned to the τ_{had} candidate is higher than 20% of the momentum of the leading track of the τ_{had} candidate.

Electrons can look very much like hadronically decaying taus with one charged track because the bremsstrahlung photons can be identified as neutral pions. In order to discriminate against electrons, a variety of variables is combined in a BDT [122]. The BDT is trained on a simulated sample of Z/γ^*, $t\bar{t}$ and $H \to \tau^+\tau^-$ events, where the background or signal events are characterized by a generator-level electron or tau lepton, respectively, within $\Delta R = 0.3$ of a reconstructed τ_{had} candidate. Depending on whether a GSF track exists that can be matched to the highest-p_{T} track of the τ_{had} or whether there are photons part of the τ_{had} candidate, a different set of variables is chosen:

- Variables that are always used:
  - p_{T} and η of the τ_{had}.
  - The ratio of energy deposits associated to the τ_{had} in the electromagnetic calorimeter over the sum of deposits in both the electromagnetic and the hadronic calorimeter.
  - The ratio of the electromagnetic and hadronic calorimeter deposits over the momentum of the highest-p_{T} track.
  - The reconstructed mass of the τ_{had}.
  - The distance in η and ϕ to the nearest crack in the electromagnetic calorimeter.
- Variables used if particle flow photons are part of the τ_{had}:
  - The p_{T}-weighted quadratic mean (RMS) of distances in η and in ϕ between all photons and the highest-p_{T} track.

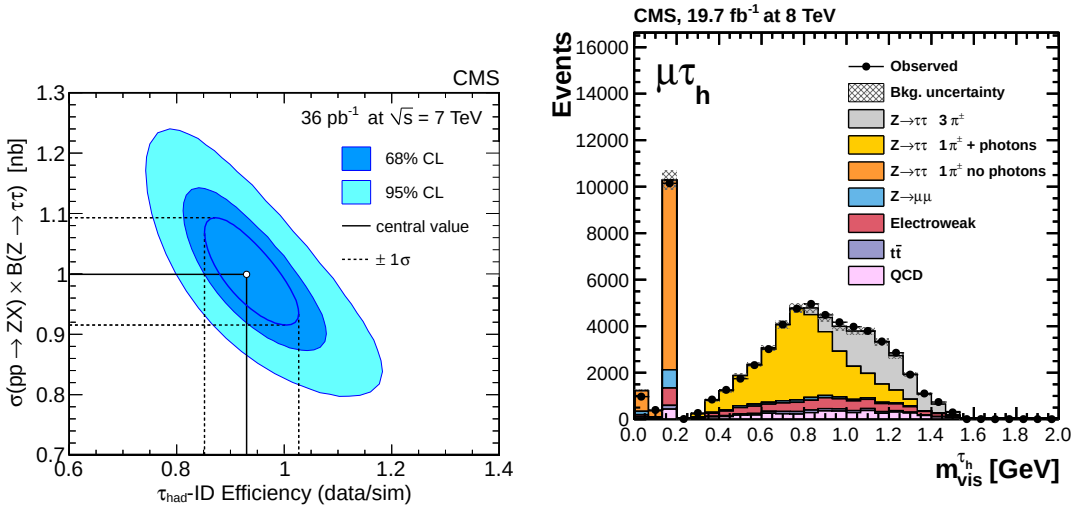

Figure 3.6: Left: Simultaneous fit of the $Z \to \tau^+\tau^-$ cross section at $\sqrt{s} = 7\,\text{TeV}$ and the data / simulation ratio of the τ_had identification efficiency. The cross section is well compatible with the theoretical prediction and the scale factor for the τ_had identification efficiency is compatible with 1. Right: Reconstructed invariant mass of the τ_had candidate in $\tau_\text{had} + \tau_\mu$ events. The peak at around 100 MeV is due to tau decays into single pions, whereas the decay into more than one observable particle creates a continuum at higher masses. From [120, 35].

  - The fraction of the total τ_had energy carried by the photons.

- Variables used if a GSF track is matched to the highest-p_T track of the τ_had.
  - $\log(p_\text{T})$ and η of the GSF track.
  - The MVA output of the PF electron MVA for the leading charged hadron. The PF electron MVA is used by the particle flow algorithm to discriminate between electrons and pions [102].
  - The χ^2/N_dof of the GSF track.
  - $(N^\text{GSF}_\text{hits} - N^\text{KF}_\text{hits})/(N^\text{GSF}_\text{hits} + N^\text{KF}_\text{hits})$, where N^GSF_hits and N^KF_hits are the number of hits in the tracks reconstructed with the Gaussian Sum Filter or Kalman Filter, respectively.

Different analyses can choose different working points, i.e. trade-offs between electron rejection and τ_had efficiency, depending on the level of backgrounds with electrons they have. CMS officially defines four working points called "loose", "medium", "tight" and "very tight". When a simple overlap between reconstructed electrons and reconstructed τ_had candidates is avoided geometrically, the "loose" working point has a τ_had efficiency of about 95% with 3.5% of electrons being misidentified as hadronically decaying tau leptons.

The full identification efficiency for hadronic tau leptons depends on the exact working points chosen. From simulation, it is estimated to be between 40% and 70% [123]. On the left hand side of Figure 3.6, a simultaneous fit of the $Z \to \tau^+\tau^-$ cross section and the data / simulation ratio of the τ_had identification efficiency is shown. This is a result from the $Z \to \tau^+\tau^-$ cross section measurement of CMS. It shows that the τ_had identification

efficiency in simulation models the data well. The plot on the right hand side shows the invariant mass of all particles associated to the reconstructed τ_{had} candidate in $\tau_{\text{had}} + \tau_\mu$ events. Since the neutrino carries away some of the momentum, the reconstructed mass is below the nominal tau lepton mass of 1.777 GeV. The three reconstructed decay modes from $Z \to \tau^+\tau^-$ events are shown individually. At the charged pion mass, there is a sharp peak because the $\tau^- \to \pi^-\nu_\tau$ process is a two-body decay. The other decays have more particles in the final state, and so the mass of the visible decay products is less constrained.

3.4.8 Lepton Isolation

When leptons are produced in decays of W, Z or Higgs bosons, there are no other particles coming from the same process. However, leptons can also be produced in a jet, for example by heavy quark decays. Also, especially in the case of hadronic tau decays, jets can be misidentified as leptons. In order to reject such events, an isolation criterion is applied to leptons, i.e. it is required that in the region around a lepton, there are no other high-energetic particles. The isolation is based on particle flow candidates and defined as

$$I_{\text{PF}}^{\text{rel}} = \frac{1}{p_{\text{T}}} \left(p_{\text{T}}^{\text{charged}} + p_{\text{T}}^{\text{neutral}} + p_{\text{T}}^{\text{gamma}} \right) , \tag{3.5}$$

where p_{T} is the transverse momentum of the lepton and $p_{\text{T}}^{\text{charged}}$, $p_{\text{T}}^{\text{neutral}}$ and $p_{\text{T}}^{\text{gamma}}$ are the scalar sum of all charged particles, neutral hadrons, or photons, respectively, inside a cone of ΔR around the lepton. For electrons and muons $\Delta R = 0.4$ is typically chosen, while for tau leptons $\Delta R = 0.5$ is a common choice.

The exact working point for the isolation variable $I_{\text{PF}}^{\text{rel}}$ is chosen slightly differently in each analysis to take into account different background sources, but typically the cut on $I_{\text{PF}}^{\text{rel}}$ is between 0.10 and 0.20 for signal leptons. For hadronic taus, the absolute isolation value is used, $I_{\text{PF}}^{\text{rel}} \cdot p_{\text{T}}$, and the working points are 0.8 GeV ("tight"), 1.0 GeV ("medium") or 2.0 GeV ("loose").

Particles from pile-up interactions can contribute to the isolation variable, even for prompt leptons. This effect is being corrected, described in more detail in Appendix A.

3.5 Grid Computing

The amount of data taken by the LHC experiments is on the order of 10 PiB per year and experiment [124]. This is more than with any other collider experiment before, and it cannot be handled by the CERN computing center alone, both in terms of data storage and in terms of CPU resources. Instead of expanding the center at CERN, it was decided to make use of all the computing infrastructure that exists already at many scientific institutes around the world. The task to process the LHC data is therefore divided amongst the participating institutes, in an effort known as the *Worldwide LHC Computing Grid* (WLCG).

3.5.1 Structure of the WLCG

The WLCG is structured in 4 different layers, or *Tiers*. The first layer, known as Tier-0, is the computing center at CERN. It receives the data directly from the experiments and stores them. In addition, it performs an initial reconstruction of the data, called prompt

reconstruction, which is used for analysis of the data shortly after they were taken. For final analyses, the reconstruction is run again later with better-known alignment and calibration constants.

This re-reconstruction is performed at the Tier-1 centers. These are typically large computing centers with a dedicated 10 Gbit/s connection to CERN [125, 126]. Apart from running the re-reconstruction, they are responsible for storing the raw detector data as well. The raw data is distributed to all the Tier-1 centers such that all the Tier-1 centers together hold a copy of the data stored at the Tier-0. Also, the reconstructed and simulated datasets are primarily stored at the Tier-1 centers. At the moment, there are 11 Tier-1 centers in the WLCG, and there is not more than one Tier-1 center per country for each experiment.

The Tier-2 centers typically have less storage capabilities than a Tier-1 center, but they provide large CPU resources. Most of the Monte Carlo event production and data analysis is run on the Tier-2 centers. While the Tier-0 and Tier-1 centers are restricted in their use, any physicist can send their jobs to the Tier-2 centers. Individual datasets can be copied from Tier-1 to Tier-2 centers so that they are available for analysis.

The Tier-3 centers are not officially part of the WLCG. It is formed by individual computers and workgroup servers. They are used for end-user analyses and visualization of analysis results, and also to provide an entry point to the grid functionality to members of the institute operating the center.

3.5.2 Grid Authentication

Before a user can access a grid service, a two-step authentication and authorization procedure needs to be performed. First, the user needs to be authenticated, i.e. it must be known to the system who they are. This task is performed using public key cryptography: every user has a certificate which is signed by a certificate authority that is well known to CERN. To obtain a certificate, the user must be a member of an institute that participates in the WLCG. A certificate request can then be made and after an institute responsible confirms the user's identity, the certificate is issued.

In the second step, it must be known to the system what a certain user is allowed to do. For example, a CMS user should not be able to use ATLAS resources. This is achieved using the concept of *virtual organization*s (VOs). For example, there is one VO for each LHC experiment. Individual users can be member of certain VOs, and the VOs have certain resources associated to them. Once the grid certificate is available, users can apply for membership in a VO which has to be confirmed by a VO responsible. After this has happened, the corresponding resources can be used.

3.5.3 Components of the WLCG

In order for the grid to work reliably, all sites need to provide the same interfaces of how to access data and how to submit computing jobs. For this purpose, every site hosts one or more *Storage Element*s (SEs) and one or more *Compute Element*s (CEs).

The task of a SE is to provide access to data stored at the site. It can be accessed with grid tools to operate on the individual files, very much the same way as on a normal filesystem. The SE itself is running a software such as dCache [127] which manages a pool of disks and tapes and dynamically optimizes their usage depending on the access pattern to the stored files. For example, files that are accessed very often are kept on disk, maybe

even replicated on more than one to provide quicker access. On the other hand, files that are rarely used are moved to tape storage.

The CE accepts computing job requests and delegates them to the actual worker nodes. However, users do not submit jobs directly to a CE of a specific site. Instead, they are sent to a workload management system (WMS), with a description of the requirements of the jobs in terms of CPU time, memory size, and the data that needs to be available on the SE in order for the job to complete. The WMS then matches these requirements with the sites that are available, and sends the job to one matching site. This makes sure that the job is only sent to a site that can actually process it correctly, and the total workload is shared between sites.

3.5.4 Analysis Workflow

The typical workflow of a physics analysis is a multi-step procedure. In a first step, the grid is used to run over events in `RECO` or `AOD` format. During this step, events that are not relevant to the analysis are filtered out, and data that are not used are dropped. For example, in an analysis studying $Z \to \mu^+\mu^-$ events, information about hadronic taus can be safely dropped. This task is performed using the WLCG. The output of this step is then on the order of gigabytes or few terabytes, so that it can be stored locally at the analyst's institute.

In the second step, the actual analysis program runs on the n-tuples, either interactively or on a local batch farm, to create histograms of interesting observables. For this step, usually many iterations are needed until the analysis is optimized and free of errors.

These histograms are then combined in the third step to make plots that can be used to draw conclusions on the physics processes, and that are used as an input to modify the analysis for the next iteration.

4 Position Resolution and Upgrade of the CMS Pixel Detector

The current pixel detector of CMS was designed to be operated at an instantaneous luminosity of $10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$. At values greater than this, the occupancy gets so high that there is significant data loss, caused by deadtime during readout and by limited buffer size on the readout chip. Furthermore, the performance deteriorates with increasing radiation damage. By the end of 2016, it is expected that the total fluence will exceed the equivalent radiation dose of 10^{15} 1 MeV neutrons per cm^2 in the first layer of the pixel barrel [128, 129]. Radiation defects lead to reduced charge collection, worsening the position resolution. It is estimated that the position resolution will be a factor of two worse at the end of 2016. Eventually, the increasing leakage current will make it impossible to operate the detector and it has to be replaced.

In order to provide optimum performance beyond 2016 and also beyond the design luminosity, an upgrade of the pixel detector is planned to be installed in CMS at the end of 2016 [130] (Phase 1 Pixel Upgrade). An improved version of the pixel readout chip improves the tracking efficiency at high luminosities and allows to be operated at higher fluences. In addition, it is planned to install a fourth layer for the pixel barrel and a third layer for the pixel endcaps. This improves the tracking efficiency and the measurement of the track parameters. The first layer of the barrel will be moved closer to the interaction point, improving the impact parameter resolution. Figure 4.1 shows a comparison of the geometry of the current configuration and the upgrade configuration of the barrel.

The upgrade of the pixel detector is essential for maintaining high tracking efficiency and low rates of misidentified tracks especially in an environment with a large number of pile-up interactions. While crucial for the reconstruction of most physics objects, this is especially important for b-jets and tau leptons. Since these objects have a macroscopic flight length, reconstruction of secondary vertices or large impact parameters can help to identify them. Both a good tracking resolution and movement of the first layer closer to the interaction point reduce the uncertainty when extrapolating to the interaction region. This allows the efficiency of b-tagging and tau identification to be improved significantly.

The emphasis in this chapter therefore lies on the measurement of the position resolution in the barrel pixel detector. The readout chip for the CMS pixel detector intended for the upgrade is studied and compared to the current version of the chip. Test beam facilities allow to study the behavior and response of particle detectors under beam conditions before they are installed in the experiment. DESY operates such a test beam facility where the measurements for this thesis have been performed. After presenting the CMS barrel pixel module and especially the silicon sensor and the readout chip in Section 4.1, the test beam facility at DESY is introduced in Section 4.2. Sections 4.3 and 4.4 discuss the measurement of the position resolution in the test beam and directly in CMS. In Section 4.5, the results are compared to a Monte Carlo simulation. Section 4.6 summarizes the findings.

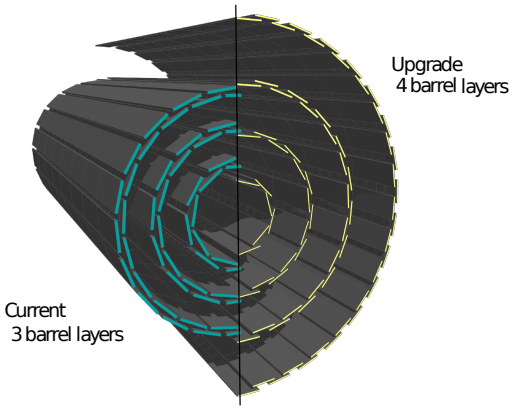

Figure 4.1: Comparison between the geometry of the current pixel barrel detector with three layers (left) and the upgraded pixel barrel detector with four layers (right). It can be seen that the innermost layer will be closer to the interaction region in the future detector. From [130].

4.1 The CMS Pixel Module

The CMS pixel barrel detector consists of multiple modules, where each module is composed of 16 readout chips, covering 66 560 pixels in total, each pixel with size 100 µm × 150 µm. There exist also half-modules of eight readout chips which are mounted at the edges of the two half-barrels. A full module is 66 mm in length and 22 mm in width, weighing 2.2 g and consuming about 2 W of power. The full barrel has a total number of 672 full modules and 96 half modules. In each layer, the modules are mounted alternating as inward-facing or outward-facing modules at slightly different radii from the center, to provide full 2π coverage in ϕ. After the upgrade, the total number of modules will grow to 1184, mostly due to additional modules required for the fourth layer. There will be only one type of module with 16 readout chips on it.

Each module consists of various components, discussed in the following.

- Silicon Sensor: The silicon is the active material where high-energetic charged particles create electron-hole pairs. It features a *pn* junction that is depleted of free charge carriers when high voltage is applied to the sensor. Induced charge carriers are transported to the edge of the sensor where they are read out by the readout chip, which is bump-bonded to the sensor. Section 4.1.1 describes the sensor in more detail.

- Readout Chip (ROC): The readout chip amplifies the signal from the silicon sensor and when a signal is detected that is above the threshold, the pixel address is stored in a data buffer together with a timestamp. If a trigger signal is received for that timestamp, the hit information in the buffers are propagated to the token bit manager. Section 4.1.2 describes the ROC in more detail.

- Token Bit Manager (TBM): The token bit manager controls the readout of all the ROCs on a module. The readout proceeds sequentially from ROC to ROC by passing around the "token bit". From the TBM the data are transferred to the front-end driver of the CMS data acquisition system via optical links [131].

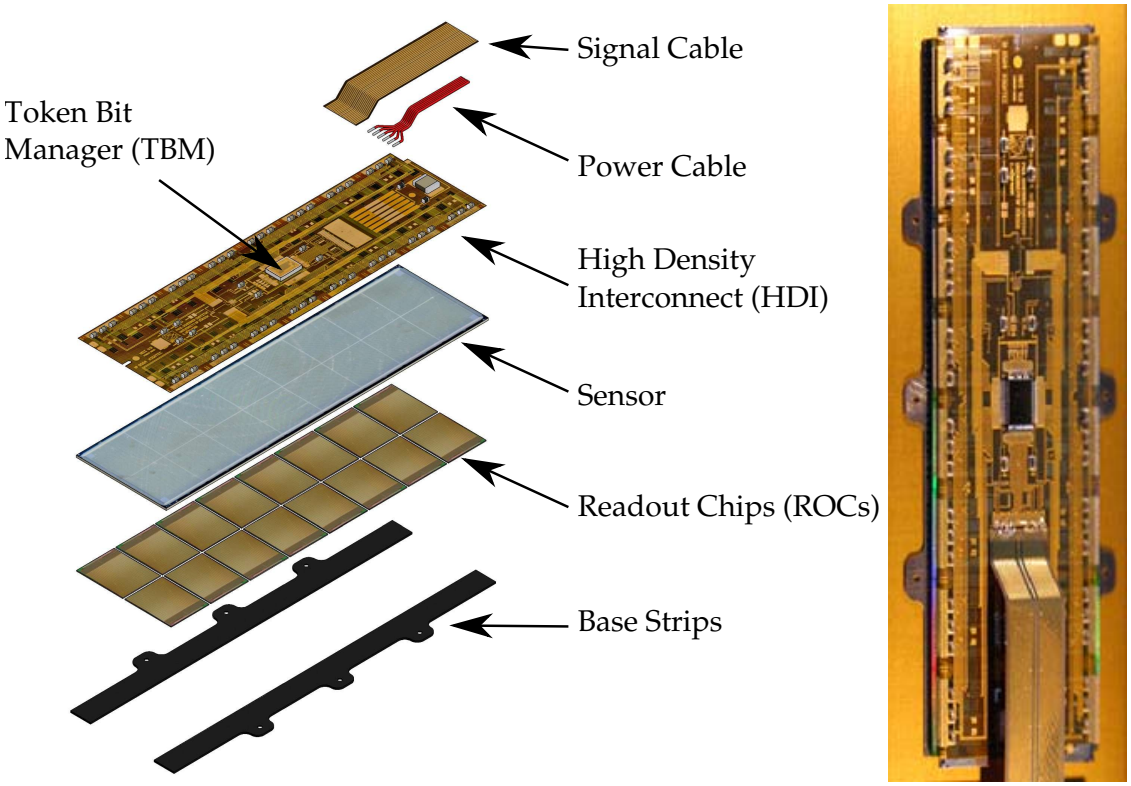

Figure 4.2: Left: Exploded view of the individual components of a CMS pixel barrel module. Right: A photograph of a module. The main structure that can be seen is the HDI. On the upper edge, the sensor sticks out a bit. From [131, 132].

- High-Density Interconnect (HDI): The high-density interconnect is a flexible circuit board which distributes the control and power signals to the readout chips and the TBM. It is mounted on the sensor opposite to the ROCs, and is wire-bonded to the ROCs at the edges.
- Signal Cable: The control and analog signals are transmitted between the TBM and the front-end driver via the signal (polyimide) cable. It is optimized for cross-talk suppression.
- Power Cable: The power cable provides the high voltage for the sensor and the analog and digital voltages for the operation of the electronics.
- Base Strips: The base strips are used to mount the module on the support structure, and they also provide the contact to the cooling system and mechanical stability to the module.

Figure 4.2 shows an exploded view of a module where all the components can be seen individually. On the right hand side, a photograph of a full module is shown.

4.1.1 The Silicon Sensor

For the silicon sensor, an n-on-n technology has been implemented: high-dose n-implant in a highly resistive n-substrate [133]. The sensor is 285 µm thick. Given the ionization rate

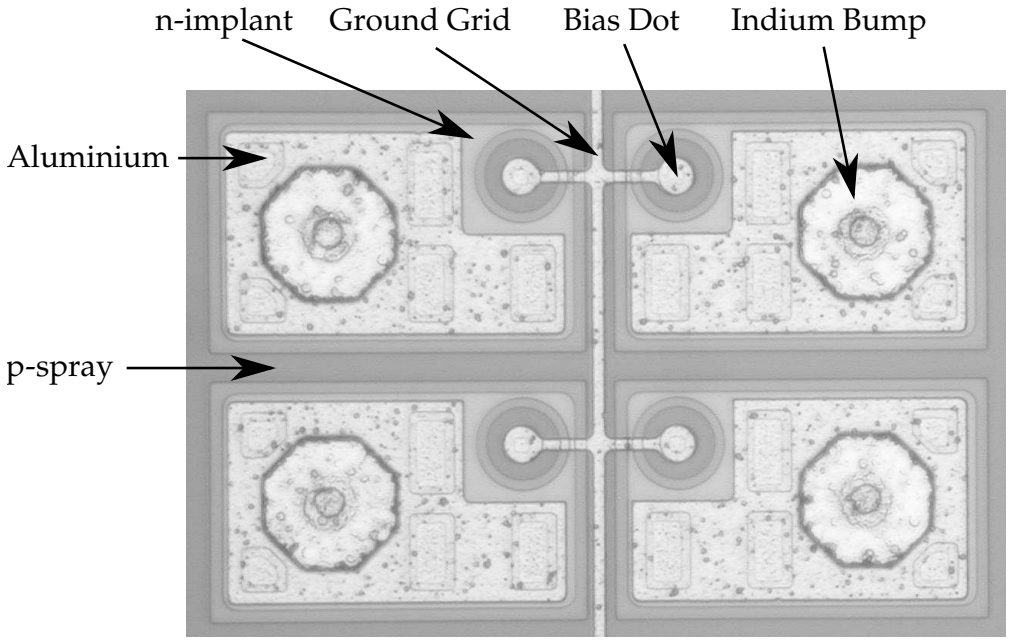

Figure 4.3: Microscope image of the front side of the pixel barrel sensor, showing 2×2 pixels. See the text for a description of the individual elements. From [135].

of minimum ionizing particles in silicon, this leads to a charge induction of ≈ 21 ke [134] as most probable value. The *pn* junction is realized with a high-dose *p*-implant at the back side of the sensor, making double-sided processing of the silicon mandatory.

Figure 4.3 shows a microscope picture of 2×2 pixels of the front side of the barrel pixel sensor. Between pixels, there is a 20 µm gap between the n-implants. This gap is filled with medium-dose *p*-implant that is sprayed over the whole sensor ("p-spray"), in order to keep the pixels isolated from each other and to keep the electric field homogeneous. The bump bonds are connected to the preamplifier of the readout chip [136]. In a corner of each pixel, there is a small *n*-implant which is separated by *p*-spray from the pixel implant. This "bias dot" is kept close to ground potential, but, since it is not connected to the preamplifier, leads to a small inefficiency in charge collection. However, in case of a bump failure, the bias dot defines the potential of the pixel via a punch-through effect between the two *n*-implants. Additionally, a structure of multiple guard rings at the back side of the sensor keeps the sensor edges close to ground potential, in order to avoid high voltage sparks across the air gap between the sensor edge and the readout chip.

The design of the sensor was deliberately chosen such that it can keep operating even after high irradiation doses of up to 10^{15} n_{eq}/cm^2. Full depletion of the un-irradiated sensor is reached around 60 V of bias voltage, however the design allows to operate at and beyond 600 V in order to keep the sensor depleted after irradiation.

The sensor is not changed for the upgrade of the pixel detector.

4.1.2 The Readout Chip

The readout chip is responsible for recording hits in individual pixels and storing them in buffers until the Level-1 trigger decision is available. It is composed of 80 rows and 52 columns, for a total of 4160 pixels per chip. The readout chip is bump-bonded to the

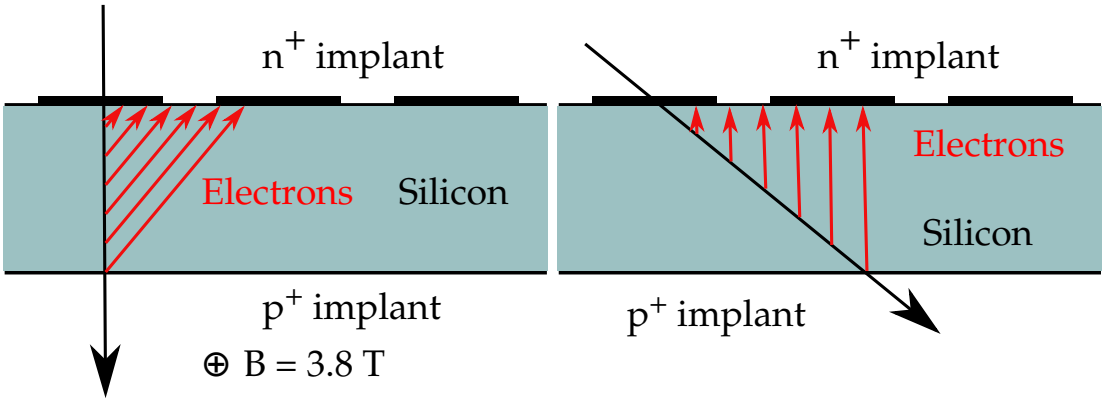

Figure 4.4: Left: Charge sharing between pixels due to Lorentz deflection of charge carriers. Right: Charge sharing without magnetic field when the track is inclined with respect to the sensor. Holes are not shown since they are not read out by the readout chip.

sensor. In row direction, the pitch is 100 µm and in column direction, the pitch is 150 µm. In CMS, the rows extend along the global ϕ direction and the columns along the global z direction. Therefore, for the transverse momentum measurement, the smaller 100 µm pitch is relevant, whereas the pseudorapidity of a track is measured in the direction with 150 µm pitch. The design of the chip is documented in [137].

Instead of simply storing a binary flag whether a pixel was hit or not, the analog pulse height is read out. With induced charges in the laboratory, a gain calibration is performed to map pulse height to signal electrons [138]. This allows to make use of charge sharing across pixels to improve the position resolution. Charge sharing occurs in r-ϕ direction because the charge carriers in the silicon are deflected by the 3.8 T magnetic field of the solenoid. The angle under which the electrons drift away with respect to the direction of the electric field is called *Lorentz angle*, and in CMS it is of the order $\tan\theta_L \approx 0.46$, corresponding to 25 degrees [134]. In z direction, inclined tracks can cross multiple pixels. Figure 4.4 illustrates the two ways of charge sharing.

Figure 4.5 shows a circuit diagram of the pixel unit cell, the electronics specific to each pixel. When a signal is present in a pixel, it is amplified by the preamplifier. The next stage is the shaper, which is AC-coupled to the preamplifier, in order to remove any voltage offset caused by leakage current. The comparator subsequently decides whether the signal is above the noise threshold, which is configurable, and in CMS operation is about 3900 electrons in the barrel [134]. If the signal is above the threshold, the sample-and-hold capacitor stores the signal and the periphery is notified.

The periphery is organized in 26 double-columns, with 160 pixels each. Upon notification from a pixel unit cell, a token bit is passed from pixel to pixel within the double column to read out each pixel. The row address and pulse height of activated pixels are stored in a data buffer together with a timestamp. When a trigger is received, the signal buffers are read out, or if no trigger arrives within the trigger latency of the CMS L1 trigger of ≈ 4 µs, the corresponding buffers can be re-used for new hits. There are 12 timestamp buffers and 32 data buffers for each double column.

Many parameters of the chip can be steered with 26 digital-to-analog converters (DACs) and three registers. The parameters include analog and digital voltages, timing of the

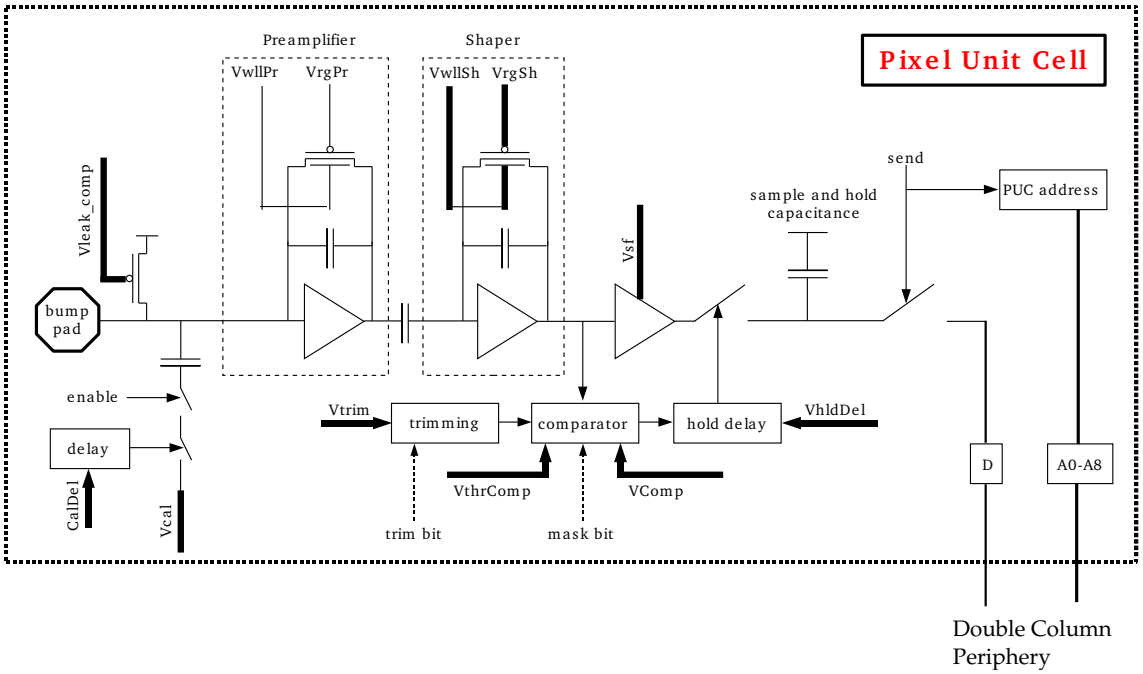

Figure 4.5: Schematic drawing of the pixel unit cell (PUC). The sensor is connected via the bump pad. Many parameters can be steered by DACs shows as thick black lines. From [139].

internal calibration signal, or the global threshold and the four trim bits that are used to make the threshold uniform over the whole readout chip. Reference [139] contains a description of all DACs.

The readout chip is a major component of the upgrade. It has been modified significantly in order to improve its performance and to allow operation at higher rates. There are three major improvements [130]:

- Reading out the pulse height and the pixel address in 6 different analog levels at 40 MHz (the LHC bunch crossing frequency) is at its limit with the current chip. In order to allow faster readout, which is required at high particle rates, a digital 160 Mbit/s link is used. There is an 8-bit analog-to-digital converter on the chip which digitizes the pulse height information at 80 MHz. Digital readout also simplifies the decoding of the signal by the front-end electronics.

- The data and timestamp buffers are subject to overflow at high particle rates, leading to data loss. At the design luminosity of $10^{34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$, corresponding to a rate of $115\,\mathrm{MHz/cm}^2$ in the layer closest to the interaction region, the current chip has a data loss rate of around 4%. The upgraded version of the chip features 24 timestamp buffers and 80 data buffers, deemed to be enough to achieve a data loss of only 0.5% at $600\,\mathrm{MHz/cm}^2$.

- The distribution of the analog signals and power within the chip leads to cross talk between pixels. This cross talk is the major contribution to the noise in a pixel and drives the threshold that has to be used for the comparator. The power distribution

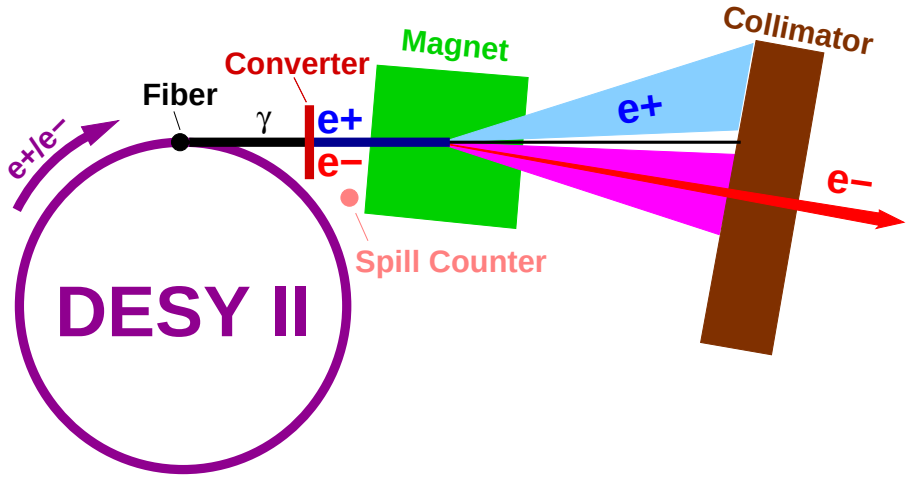

Figure 4.6: Beam generation at the DESY test beam. The e^+ or e^- beam is hitting a carbon fibre, and the bremsstrahlung photons hit a secondary copper target which converts them into electron-positron pairs. A magnet swipes the beam horizontally and allows to select the beam energy. From [141].

system has been changed substantially in the new version of the chip in order to reduce cross-talk effects. With the new chip, it is expected that the threshold can go down from almost 4000 electrons to as low as 1500 electrons. This allows to achieve a better position resolution since charge sharing can be exploited more, but it also allows to keep the chip operational after high irradiation doses when the absolute signal goes down due to trapping in the silicon.

4.2 The Test Beam at DESY

The measurements of the readout chip presented in this thesis have been performed with the test beam facility at DESY Hamburg [140]. A $1-6$ GeV electron beam is extracted from the synchrotron DESY II with a rate of up to 5 kHz and a divergence of less than 1 mrad. DESY II is an injector for the synchrotron light source PETRA III, however, during times when PETRA III does not need to be filled, the beam is available for beam tests. The electrons are accelerated and decelerated in a sinusoidal wave with a frequency of 12.5 Hz while the revolution frequency is 1 MHz.

Figure 4.6 visualizes the generation of the test beam. Electrons or positrons are injected from the LINAC II (not shown) into the DESY II synchrotron at 450 MeV [142]. Inside the synchrotron, the primary beam hits a thin carbon fibre, generating bremsstrahlung photons. These photons are converted at a secondary target (converter) into electron-positron pairs which are used for the actual test beam. A magnet spreads the particles horizontally, separating them by energy. Choosing the magnet current allows to select particles with a particular energy. The particle rate is energy dependent, since the bremsstrahlung spectrum roughly has a $1/E$ dependence. Therefore, higher energies correspond to lower rates.

Compared to other facilities, such as the PS and SPS test beams at CERN, the particle rates and energies at DESY are small. The low electron energies lead to multiple scattering even at thin materials, which must be taken into account when analyzing the data.

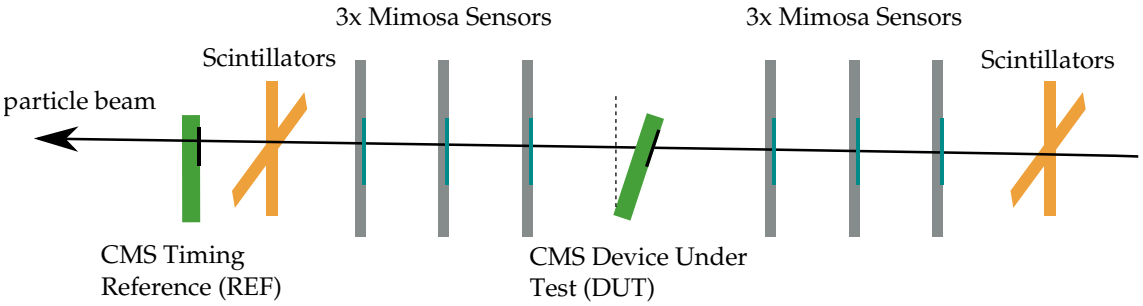

Figure 4.7: The test beam setup. The beam is coming from the right, crossing the six telescope planes and the CMS readout chip (DUT). The trigger signal is given by four scintillator detectors, two in front of the telescope and two behind. A second CMS chip is used as a timing reference (REF).

However, there is no radioactive activation, and therefore no dosimetry is required, which allows for more experimental flexibility.

4.2.1 Test Beam Setup

Apart from general functionality tests of the new readout chip, hit detection efficiency and position resolution are the two most important parameters to be measured and understood in beam tests. In order to measure these quantities, it is essential to know the true position of the beam particles in the CMS sensor. A beam telescope is used in order to obtain this information. The telescope has been developed by the AIDA collaboration using linear collider technology [143].

Figure 4.7 shows a cartoon of the test beam setup. The central elements are six MIMOSA26 pixel sensors with a pitch of 18.4 µm in both directions and 1156 × 578 pixels per sensor. The sensors are thinned to 50 µm in order to reduce multiple scattering at the telescope planes. Behind the first three planes, the device under test (DUT) is mounted, in this case a CMS pixel readout chip. The six telescope planes provide a very precise measurement of the position of beam particles, up to an accuracy of 4 µm at the DUT position. Four scintillator detectors, two mounted in front of the first plane and two behind the last plane, give the trigger signal. After receiving a trigger, both the telescope sensors and the CMS chips are read out. A four-fold coincidence of the scintillators is required to suppress triggers without an actual track. For the CMS chips, the trigger signal is sent to the testbeam hut and from there back into the measurement area, in order to delay the signal, emulating the L1 trigger latency in CMS. The 1 MHz clock of the DESY II machine is used to generate a 40 MHz signal which corresponds to the LHC bunch crossing frequency. This clock signal is fed to the CMS readout chip, emulating the LHC clock.

The readout of the MIMOSA26 sensors is very slow compared to CMS, of the order of 100 µs, which is 3 orders of magnitude above the LHC bunch crossing frequency. Therefore, several tracks will be reconstructed in the same event when operating at reasonable beam rates. In order to pick the correct track that triggered the readout (to be studied in the DUT chip), a second CMS chip is operated at the end of the telescope as a timing reference. A telescope track is only considered if it has a matching hit in the reference CMS chip. Figure 4.8 shows a photograph of the experimental setup. It can be seen that the DUT

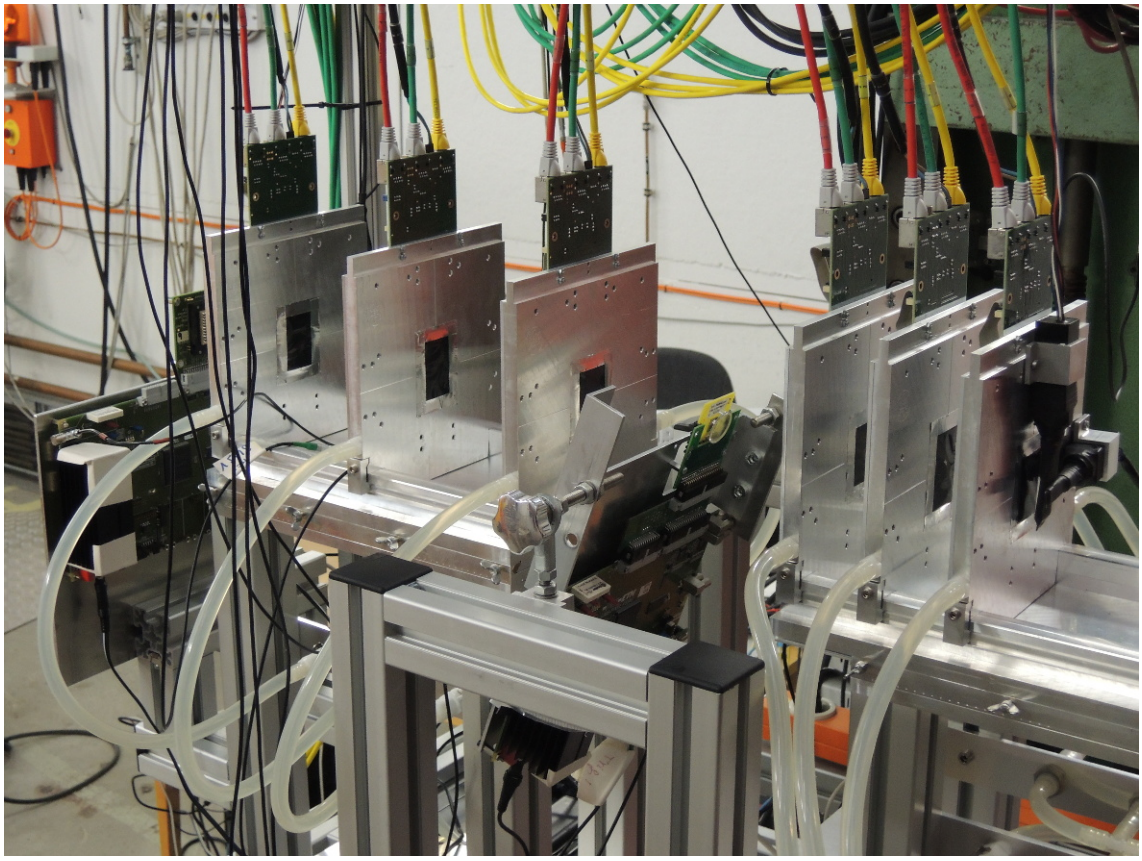

Figure 4.8: A photograph of the experimental setup. The beam comes from the right. The six telescope planes can clearly be seen. The tubes are used for cooling in order to keep the MIMOSA26 sensors at a stable temperature. The CMS chip and sensor are tilted with respect to the beam. In the very front, two scintillators are visible.

chip is inclined with respect to the particle beam, to enable charge sharing despite the lack of the 3.8 T magnetic field.

The data acquisition systems of the telescope data and the CMS data are independent. However, the telescope sends a "busy" signal while it is being read out, so that no triggers are generated during that time. This makes sure that the same events are read out by the telescope and by the CMS chips, and the data streams can be interleaved.

The hit detection efficiency is especially interesting at very high rates, on the order of $\gtrsim 100\,\mathrm{MHz/cm^2}$. Such rates cannot be achieved with the test beam at DESY. However, the position resolution can be studied very well at the DESY test beam and the following studies will concentrate on the resolution measurement.

4.3 Position Resolution in the Test Beam

In this section, the measurement of the position resolution of the CMS readout chip with the test beam at DESY is presented. It is of particular interest how the new, digital version of the readout chip performs with respect to the analog chip.

4.3.1 Analysis of Telescope Data

Before the data from the CMS readout chip can be analyzed, the tracks in the beam telescope need to be reconstructed and correlated with the hits in the two CMS readout chips. This is a multi-step procedure using the event processing framework developed for the ILC, called MARLIN[1] [144]. Each step is a MARLIN processor which reads data from an LCIO (linear collider input/output) file and can write new collections back into the file for other processors to consume. The framework is very similar to the event data model in CMS (see Section 3.3.1). Each of the following steps can be run individually, to allow easy inspection and quick turnaround in the case of problems [145]:

- Converter: The initial step consists of converting the raw data from the telescope DAQ into LCIO format.
- Clustering: This step merges adjacent hits in the individual MIMOSA26 sensors to clusters, with the assumption that such a cluster originates from a single primary particle.
- Hitmaker: The hitmaker assigns a global position to each cluster. In this step, also the offset in x and y between the various sensors is computed ("pre-alignment").
- Alignment and Tracking: Track candidates are built using a Kalman Filter [99] in one direction. The full alignment is performed with MILLEPEDE-II [146] and allows offsets in x, y, and rotations in the $x-y$ plane of each sensor to be corrected. After this step, a collection of aligned hits is available which is used to fit the final telescope tracks.
- DUT and REF alignment: In this step, the position and rotation of the two CMS readout chips with respect to the telescope are determined. A similar alignment procedure is performed with the DUT x, y and z positions, as well as 3D rotations, as free parameters. For determining the residuals, all tracks are re-fitted with the hits in the CMS DUT included. The General Broken Lines library [147] is used to describe scattering of electrons at the MIMOSA26 sensors and the CMS sensor and readout chip.
- Testbeam Analysis. In the last step, the aligned telescope tracks and CMS hits are used to study the behavior of the CMS readout chip.

4.3.2 Hit Reconstruction in the CMS chip

For the test beam analysis, hits are reconstructed in the CMS pixel detector with the center-of-gravity (CoG) algorithm. For each cluster of pixels, the hit position in local coordinates is defined as

$$\vec{x}_{\text{hit}} = \frac{\sum_i^N A_i \vec{x}_i}{\sum_i^N A_i}\,, \tag{4.1}$$

where the $\vec{x}_i$ are the integer coordinates (row and column positions) of the pixels, A_i is the charge in pixel i, and N the total number of pixels in the cluster. This method works

[1] Modular Analysis & Reconstruction for the Linear Collider

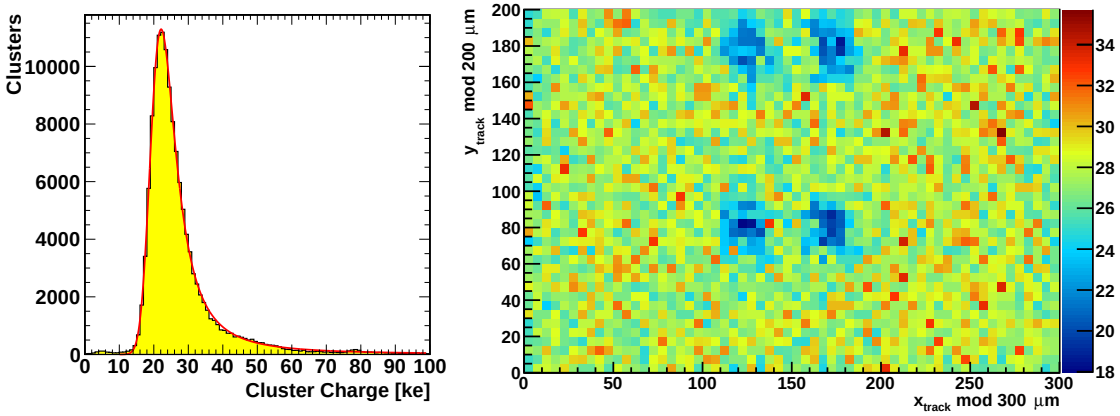

Figure 4.9: Left: The distribution of total charge in a cluster, in kilo-electrons. The fit function is a Landau distribution convoluted with a Gaussian. Right: The average charge deposited by a particle, as a function of the x and y point of incidence in a 2×2 pixels area. The position of the sensor bias dot can be clearly seen as a region of charge deficit.

very well when charge sharing predominantly occurs due to geometry, i.e. inclined tracks. In the CMS chip, this is indeed the case, and thermal diffusion is only a small factor.

Figure 4.9 shows two charge distributions in a CMS readout chip in the test beam. In this case, the sensor is not inclined with respect to the beam. On the left hand side, the cluster charge is shown for all reconstructed clusters. Clusters are not considered only if an edge pixel is involved, or the track goes through the bias dot (see below) to make sure that no charge is lost due to geometric acceptance.

The distribution is fitted with a Landau distribution that is folded with a Gaussian, where the widths of the Landau and the Gaussian, the normalization of the Landau and the most probable value of the Landau are free parameters. The convolution is necessary due to electronic noise in the readout chip, and due to a non-uniform gain over all pixels on the chip. According to the fit, the most likely charge is ≈ 21.4 ke.

On the right hand side, the average charge per cluster is shown in kilo-electrons, as a function of the impact point of the particle. A 2×2 pixel area ($300\,\mu\text{m} \times 200\,\mu\text{m}$) is depicted, and all hits outside this region are folded into it. This is possible because of the symmetry of the sensor. It can be seen how the full pixel area is responding uniformly except for the bias dot region where there is a charge deficit. Even though the deficit looks rather dramatic in this picture, in CMS it is not an issue because it is smeared out when the tracks are inclined or when the Lorentz force of the 3.8 T magnetic field deflects the charge carriers. For the charge distribution on the left hand side, however, events are removed where the track goes through the bias dot, by imposing cuts on the reconstructed track position folded into the 2×2 pixel grid (x_{mod}, y_{mod}): $x_{\text{mod}} < 105\,\mu\text{m}$ or $x_{\text{mod}} > 195\,\mu\text{m}$ and $y_{\text{mod}} < 55\,\mu\text{m}$, $95\,\mu\text{m} < y_{\text{mod}} < 155\,\mu\text{m}$ or $y_{\text{mod}} > 195\,\mu\text{m}$.

4.3.3 Position Resolution Measurement in the Test Beam

The measurement of the position resolution is presented in row direction (local y). In this direction, the pixel pitch is 100 µm. In CMS, this corresponds to the r-ϕ direction, which

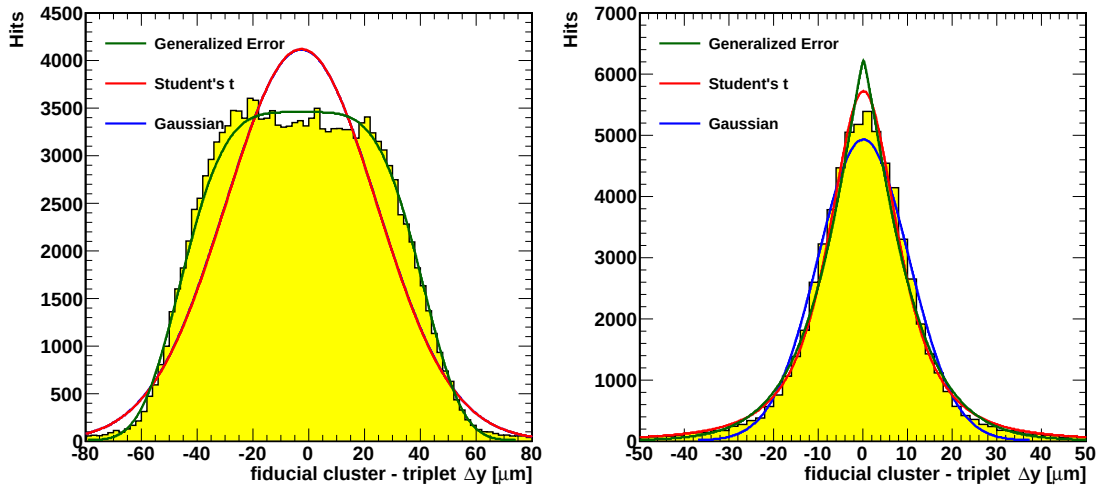

Figure 4.10: Two residual distributions, showing the difference in row direction between the reconstructed hit position in the CMS pixel chip and the prediction of the beam telescope. On the left, the distribution is shown when the CMS chip is perpendicular to the beam while on the right it is inclined by $\approx 19°$. In this case, charge sharing between pixels improves the resolution. Three different functions are fitted to the distribution to extract its width. In the left hand distribution, the Gaussian and Student's t function overlap with each other, which corresponds to a large ν parameter of Student's t function.

is relevant for the p_{T} measurement of tracks.

The Residual Distribution

The position resolution is extracted from the residual distribution of the telescope track extrapolated to the DUT position and the reconstructed hit in the DUT. Only such events are considered where no edge pixel is part of the cluster and the residual in local x (column direction) is less than 150 µm, to reduce background and contributions from multiple tracks reconstructed in the telescope that coincide with the CMS reference chip.

Figure 4.10 shows two examples of the residual distribution from runs taken in May 2013 with the digital readout chip. In the first case, the sensor is positioned such that it is perpendicular to the beam. In most cases, exactly one pixel is hit, and the center of the pixel is taken as the hit position. This results in a box distribution for the residuals. On the right hand side, the sensor is inclined with respect to the beam by 18.9°. Here, a typical track crosses two pixels and the hit can be reconstructed more precisely. The residual distribution is much narrower which means that the position resolution improves. In order to extract the width of the residual distribution, a function with the width as a free parameter is fit to the distribution. The following three functions are studied:

- Gaussian.

$$f(x) = B + \frac{A}{\sqrt{2\pi}\sigma} \cdot \exp\left(\frac{(x - x_0)^2}{2\sigma^2}\right), \tag{4.2}$$

  with four free parameters A, B, x_0 and σ. The width is taken from the σ parameter.

Table 4.1: Fit results for three functions describing the residual distributions in Figure 4.10. The quoted errors are only of statistical nature.

Function	Width 0° [µm]	Width 20° [µm]
Gaussian	27.06 ± 0.04	10.20 ± 0.04
Generalized Error	33.07 ± 0.06	7.80 ± 0.08
Student's t	26.96 ± 0.04	8.37 ± 0.04

- Generalized Error Function [148].

$$f(x) = B + \frac{A \cdot \beta}{\sqrt{8} \cdot \sigma \cdot \Gamma\left(\beta^{-1}\right)} \cdot \exp\left(\left|\frac{x - x_0}{\sqrt{2}\sigma}\right|^{\beta}\right), \tag{4.3}$$

  with five free parameters A, B, x_0, σ and β. The symbol Γ denotes the Gamma function. Note that for $\beta = 2$ one obtains the standard Gaussian. Again, the width is extracted from the σ parameter.

- Student's t Function [149].

$$f(x) = B + \frac{A}{\sigma\sqrt{\nu\pi}} \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \left(1 + \frac{1}{\nu}\left(\frac{x - x_0}{\sigma}\right)^2\right)^{-\frac{\nu+1}{2}}, \tag{4.4}$$

  with five free parameters A, B, x_0, σ and ν. The function interpolates between a Gaussian and a Breit-Wigner. For $\nu = 1$ one obtains a Breit-Wigner and for $\nu \to \infty$ the function becomes a Gaussian. As before, the width is extracted from the σ parameter.

Figure 4.10 shows the three functions being applied to the residual distributions, and Table 4.1 contains the fit results. In the first case with perpendicular tracks, only the generalized error function can describe the shape of the distribution. For a box distribution with 100 µm pitch, the standard deviation is $100\,\mu\text{m}/\sqrt{12} \approx 28.9\,\mu\text{m}$. The width of the generalized error function is $33.1 \pm 0.1\,(\text{stat.})\,\mu\text{m}$. In the case when the sensor is inclined by $\approx 19°$, the Gaussian underestimates the tail, while the generalized error function overestimates the peak of the distribution. Student's t function describes the residual distribution better and has the best χ^2/N_{dof} amongst the three. In the following, the generalized error function is used to quantify the width of the residual distribution for tilt angles below 15° (when the distribution becomes more and more box-like), and Student's t function otherwise.

The Telescope Resolution

Not only the resolution of the DUT itself contributes to the width of the residual distribution, but also the finite resolution of the beam telescope. Its resolution is estimated to be $\approx 4.5\,\mu\text{m}$ at the DUT position. This number is estimated from the compatibility of the triplets measured in the three telescope planes in front and behind the DUT after extrapolation to the DUT position. When σ_{tel} is the telescope resolution and r is the width of the residual distribution, the position resolution of the CMS detector, σ_{CMS}, is given by

$$\sigma^2_{\text{CMS}} = r^2 - \sigma^2_{\text{tel}}\,. \tag{4.5}$$

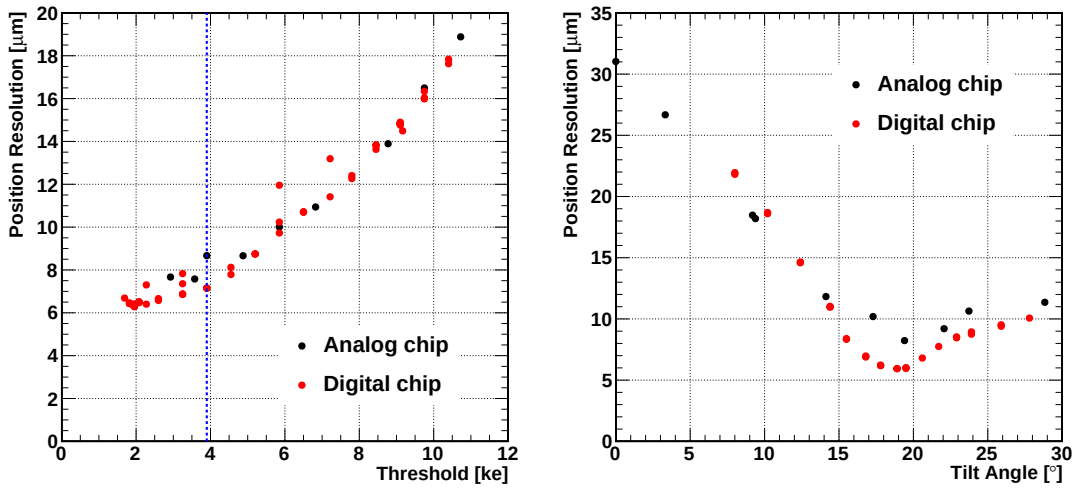

Figure 4.11: Left: Resolution as a function of the pixel threshold (see Section 4.3.3 for the exact definition). The blue dashed line is the threshold at which the CMS pixel barrel detector is currently being operated. Right: Resolution as a function of the inclination angle of the readout chip with respect to the beam. For the analog chip, the threshold is fixed at 3.25 ke and for the digital chip at 2.0 ke. The statistical uncertainty of each point is below 100 nm and corresponds to at least 50 000 tracks.

4.3.4 Measurement Results

The position resolution depends on many parameters, most importantly the readout threshold and the tilt angle with respect to the beam. The new digital readout chip allows to go to lower thresholds than the previous analog chip, and so an improvement in resolution is expected.

Figure 4.11 shows the measured resolution as a function of one of the two parameters while the other one is kept fixed. The black points show the results for the analog readout chip, taken in April and May 2012. The red points correspond to the digital chip, taken between May and August 2013. The beam energy was between 4.4 GeV and 5.6 GeV, and the bias voltage was always well above full depletion between 120 V and 150 V.

On the left hand plot, the resolution is shown as a function of the readout threshold. The tilt angle is fixed at 19.1° for the analog chip and at 19.6° for the digital chip. The blue line indicates the average readout threshold used for the CMS pixel barrel, where the position resolution is around 8 µm. In the test beam, slightly lower values are still possible for the analog chip. For the digital chip, lower values down to $\approx$ 1.5 ke are reachable. The difference in position resolution between the blue line and $\approx$ 2 ke is roughly the improvement that can be expected with the upgraded pixel detector. High threshold values give an idea of the performance of the chip after irradiation. In irradiated chips, the charge collection efficiency degrades because charge carriers are trapped inside the silicon. By increasing the threshold, the signal over threshold ratio is reduced in a similar way as it would be reduced for an irradiated sensor.

On the right hand plot, the resolution is shown as a function of the tilt angle. The readout thresholds are now fixed at nominal values of 3.25 ke or 2.0 ke for the analog or

digital chips, respectively. For very low tilt angles where there is not much charge sharing between pixels, the resolution is very similar between the two chips, because when a track hits only one pixel, the lower threshold does not make a difference. However, around the optimal angle and above, when more than one pixel is hit, the lower threshold improves the resolution. At 19°, a resolution of 6 µm is possible.

In the next two sections, the results obtained in the test beam are compared to what is measured directly in the pixel detector in CMS and Monte Carlo simulation.

4.4 Position Resolution in CMS with the Triplet Method

In this section, the measurement of the position resolution in r-ϕ direction in the CMS pixel barrel detector is presented. The measurement is performed with the *triplet method*. The results are expected to be consistent with the test beam measurements.

4.4.1 Hit Reconstruction

In CMS, the hit reconstruction is performed in a more sophisticated manner than in the test beam analysis. Cluster profiles projected in x and y are matched to templates obtained from a detailed simulation, called PIXELAV [150, 151]. The simulation is performed for varying angles of incidence and Lorentz angles. This method is especially useful for hit reconstruction in irradiated sensors, however it has been shown that there is also an improvement of the order ≈ 1 µm in resolution for unirradiated sensors [152].

4.4.2 The Triplet Method

The trajectory of a charged particle in a homogeneous magnetic field, such as inside the CMS tracker, is described by a helix. When the radius of the helix is known, two space points are enough to fix the full trajectory [153]. This fact is exploited by taking high-p_{T} tracks with hits in all three pixel layers, and the hits in two layers are used to extrapolate to the third layer (the layer under test). The curvature is taken from the full track fit, including the strip tracker. Comparing the extrapolated position with the position of the actual hit in the layer under test gives then a measure of the position resolution of the detector. The analytic solution for the intersection of the helix with the sensor plane is taken from [154].

Figure 4.12 illustrates the principle of the method. Here, the hits in layers 1 and 3 are taken to define the trajectory, together with the curvature from the full track fit. Layer 2 is then probed for its resolution. In order to quantify the resolution, two steps need to be performed. The width of the residual distribution needs to be quantified and the finite position resolution of the layers that are used to define the trajectory needs to be unfolded.

The first step is performed in a similar way as in the test beam measurement. Figure 4.13 shows the residual distributions in all three pixel barrel layers in a typical collision run, in logarithmic scale. The tails are much better described by Student's t function than by the generalized error function. Therefore, in the following, the residual width is extracted from the fit of Student's t function.

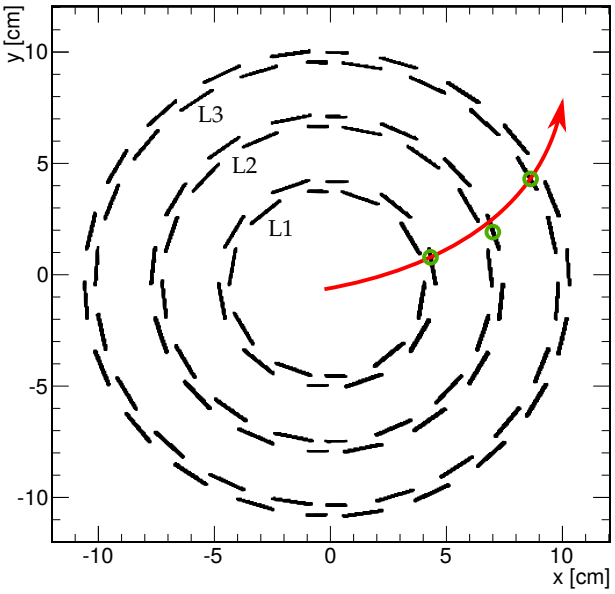

Figure 4.12: Principle of the triplet method: when the curvature of a trajector is known, two hits can be used to define the whole particle trajectory. In this case, hits in layers 1 and 3 are used, and the trajectory is interpolated to the middle layer. The difference between the interpolated position and the actual hit in the middle layer is a measure of the position resolution.

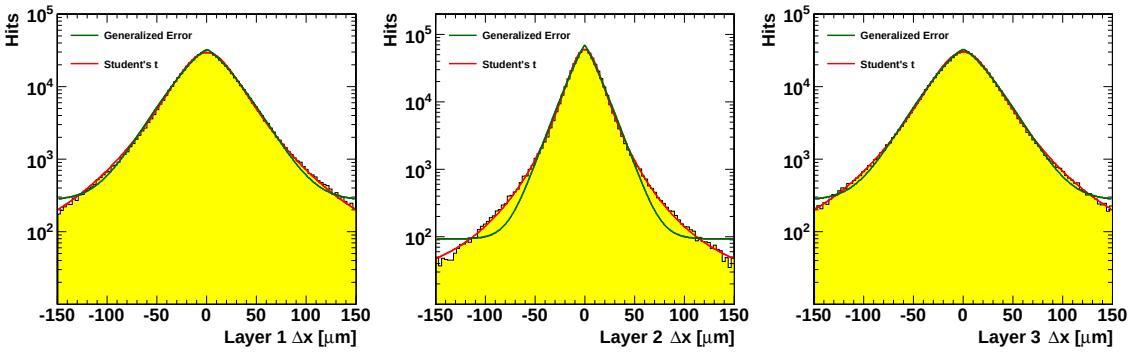

Figure 4.13: Residual distribution in Pixel Barrel Layers 1, 2 and 3 (from left to right), where the other two layers are used to define the trajectory. The distribution has significant tails which are well modeled by Student's t function, but not by the generalized error function. The data shown are from CMS run 207487.

The unfolding step starts with the error propagation to describe the width of the residual distribution when the intrinsic resolutions in the three layers are known. For high-p_{T} tracks, straight line error propagation can be assumed:

$$r^2 = \sigma_0^2 + \frac{\sigma_1^2}{4} + \frac{\sigma_2^2}{4} + \frac{d_m^2}{L_{12}^2}(\sigma_1^2 + \sigma_2^2)\,, \tag{4.6}$$

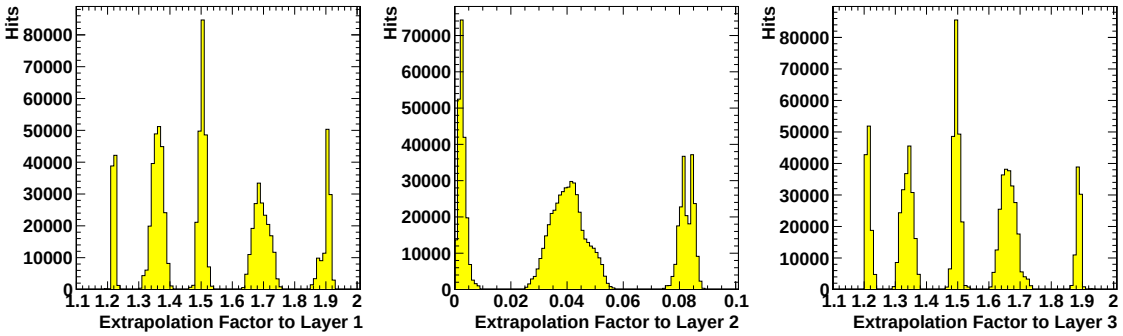

Figure 4.14: Distribution of the extrapolation factor (lever arm) for Pixel Barrel layers 1, 2 and 3 (from left to right). The different peaks originate from different combinations of a track traversing inward-facing and outward-facing modules in the different layers. The data shown are from CMS run 207487.

where σ_0 is the intrinsic resolution in the layer under test, σ_1 and σ_2 are the intrinsic resolutions of the other two layers, d_m the length of the trajectory from the layer under test to the middle between the other two layers, and L_{12} the length of the trajectory between the other two layers. The lever arm term d_m/L_{12} is crucial for the width of the distribution. It is 0 when interpolating between layers 1 and 3, and the distance between the layers is the same. When extrapolating, it becomes greater than 1 and significantly broadens the residual distribution, as visible in Figure 4.13.

In CMS, the inward-facing and outward-facing modules are mounted at slightly different radii, as can be seen on Figure 4.12. This leads to a varying distance between the hits of a track in subsequent layers, depending on the combination of inward-facing and outward-facing modules the trajectory crosses. The lever arm term d_m/L_{12} in Equation 4.6 is sensitive to the distance between the layers. Figure 4.14 shows the distribution of the lever arm factor in a normal pp collision run. The three distributions correspond to the three cases when the first, second or third layer is the layer under test, respectively, and the other two layers define the trajectory. The different peaks correspond to different combinations of a track crossing inward-facing or outward-facing pixel modules. For the resolution measurement in the following, only tracks with a lever arm factor of $d_m/L_{12} < 0.015$ for layer 2 or $1.43 < d_m/L_{12} < 1.57$ (layers 1 and 3) are used, and the lever arm factors are fixed to 0 and 1.5, respectively.

With the intrinsic resolutions in the three layers σ_1, σ_2, σ_3, and the widths of the residual distributions r_1, r_2, r_3, Equation 4.6 is used to formulate three equations for the three different layers.

$$\begin{aligned} r_1^2 &= \sigma_1^2 + \frac{\sigma_2^2}{4} + \frac{\sigma_3^2}{4} + l_1^2(\sigma_2^2 + \sigma_3^2) \\ r_2^2 &= \sigma_2^2 + \frac{\sigma_1^2}{4} + \frac{\sigma_3^2}{4} + l_2^2(\sigma_1^2 + \sigma_3^2) \\ r_3^2 &= \sigma_3^2 + \frac{\sigma_1^2}{4} + \frac{\sigma_2^2}{4} + l_3^2(\sigma_1^2 + \sigma_2^2)\,, \end{aligned} \tag{4.7}$$

where the lever arms $l_2 = 0$ and $l_1 = l_3 = 1.5$ are fixed. This is now a system of three

linear equations in the three unknowns σ_i^2. The solution is given by

$$\begin{aligned}\sigma_1^2 &= \frac{1}{9}(-r_1^2 - 10r_2^2 + 5r_3^2)\\ \sigma_2^2 &= \frac{1}{9}(-r_1^2 + 14r_2^2 - r_3^2)\\ \sigma_3^2 &= \frac{1}{9}(5r_1^2 - 10r_2^2 - r_3^2)\,.\end{aligned} \tag{4.8}$$

Unlike matrix inversion methods in other cases, this procedure is numerically stable. Typically, the method is unstable because of small fluctuations in the input data being inflated after the unfolding, especially when the unfolded distribution features sharp peaks. However, in this case, the unfolded distribution is mostly flat (the position resolutions in the three layers are expected to be very similar), and the statistical uncertainties on the widths of the residual distributions is typically less than 1 %.

This procedure allows the simultaneous determination of the position resolution in all three layers of the pixel barrel detector. When studying radiation damage effects, this is especially interesting: since the first layer is closer to the interaction region than the other two layers, it is exposed to higher fluxes and expected to degrade more quickly with radiation.

4.4.3 Measurement Results

The measurement is performed in events triggered by a di-muon trigger. This makes sure that there are two clean and isolated tracks in the event. For the analysis, all reconstructed tracks in an event with $p_\mathrm{T} > 12\,\mathrm{GeV}$, hits in all three pixel barrel layers and lever arm factors as described above are taken. The procedure is applied to all CMS runs with more than 10 000 such tracks.

Figure 4.15 shows the measured resolution as a function of integrated luminosity delivered to CMS by the LHC. Each data point corresponds to one CMS run. The error bars represent the statistical uncertainty on the width parameter that is obtained by standard error propagation of Equation 4.8, assuming the statistical uncertainties on the widths of the triplet residual distributions are uncorrelated between layers. This leads to the uncertainty on the resolution being partly correlated between layers. The correlation matrices are very similar in all runs, and the following is a typical example for the correlations between the σ_i:

$$\rho = \begin{pmatrix} 1 & -0.59 & -0.13 \\ -0.59 & 1 & -0.59 \\ -0.13 & -0.59 & 1 \end{pmatrix} \tag{4.9}$$

The trend in the two diagrams for the 7 TeV run (left) and the 8 TeV run (right) is the same: the resolution degrades over time, with sudden partial recoveries. These recoveries can be correlated to technical stops of the machine where new gain calibrations have been prepared and the thresholds have been re-optimized. The remaining degradation can be mostly attributed to the limited knowledge of the alignment parameters and the Lorentz angle. The method is very sensitive to both of these. If the Lorentz angle assumed during the hit reconstruction is different from the real one, the hit will be reconstructed at a slightly different position. In fact, the Lorentz angle has been shown to change with time due to irradiation [155]. In a similar way, if the alignment parameters are not accurate,

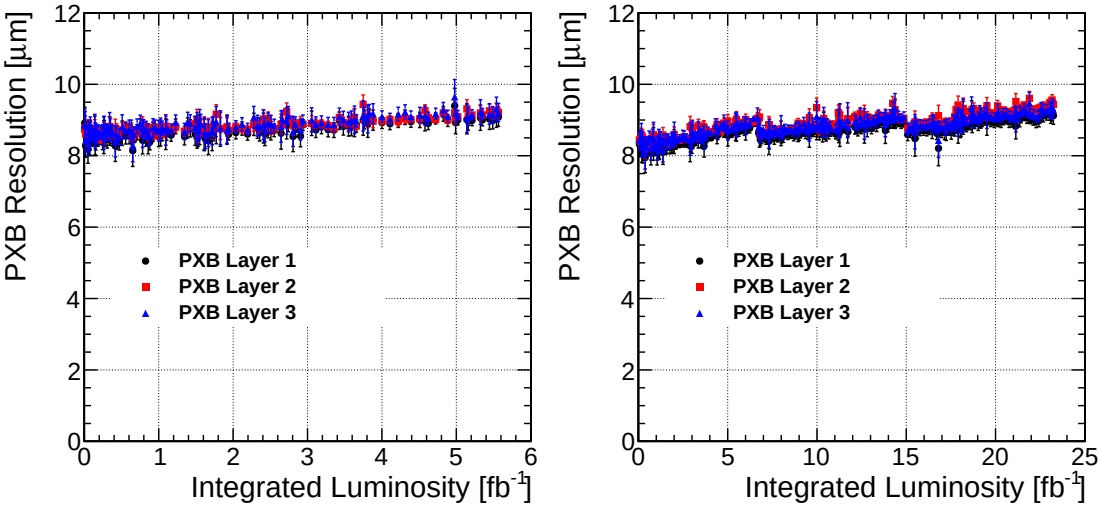

Figure 4.15: Measured position resolution in the three layers of the CMS pixel detector as a function of the delivered integrated luminosity in the 7 TeV run (left) and 8 TeV run (right) of CMS. The error bars show only the statistical uncertainties of each data point, and they are partly correlated between layers. The sudden improvements correspond to points in time where a technical stop has happened.

the residual distribution becomes broader. The alignment parameters have been obtained on the level of every single module at the beginning of the year, but throughout the year, only coarser structures such as ladders and half-shells have been re-aligned.

The effect of the Lorentz angle is shown more clearly in Figure 4.16. The mean of the triplet residual distribution in layer 2 is presented as a function of the global ϕ position of the hit. Data points in black or red represent hits in inward-facing or outward-facing modules, respectively. While overall, the data points are distributed around zero, a bias is seen for inward-facing or outward-facing modules alone. Since the electrons drift in opposite directions in the two types of modules, the Lorentz force points in the opposite direction as well. A slight mismatch of the Lorentz angle used for the hit reconstruction with respect to reality can therefore explain this effect. Around the optimal angle of $\approx 19°$, a difference of 1° in the Lorentz angle changes the hit position by 3 µm. The two plots correspond to two different CMS runs. The one on the left hand corresponds to a 8 TeV run after 0.9 fb^{-1} of data taking, and the right hand one to 21.4 fb^{-1}. The larger spread around zero in the right hand plot leads to a broadening of the residual distribution when integrating over ϕ, explaining the worse resolution in the late runs with respect to the early runs in Figure 4.15.

Overall, the resolution is between 8.0 µm and 9.5 µm for all three layers and for the full data taking period. There is no significant difference observed between the layers, which suggests that there are no radiation damage effects visible after the first three years of running. The results of the position resolution measurements are consistent with the test beam measurements discussed in Section 4.3.4.

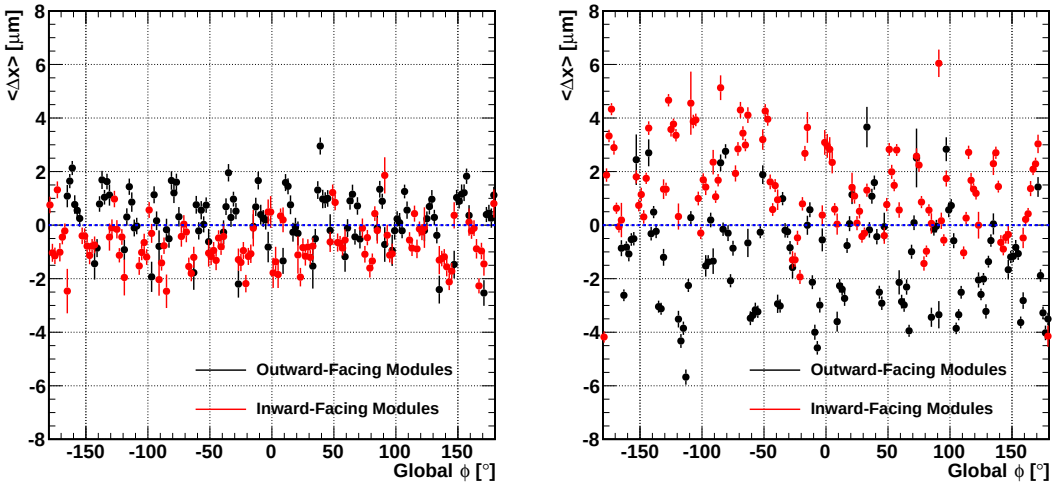

Figure 4.16: Profile plot of the mean position of the residual distribution as a function of the global ϕ position of the hit. The left hand plot shows data from run 191830 after 0.9 fb^{-1} of 8 TeV data taking. The right hand plot shows run 207487, corresponding to 21.4 fb^{-1}. It can be seen that there is a different trend for inward-facing and outward-facing modules.

4.5 Simulation

In this section, the results obtained with test beam data are compared to Monte Carlo simulation with PIXELAV [156, 151]. The procedure consists of three steps. First, the electric field inside the sensor is generated. Second, the generation of charge carriers and their transport to the edge of the sensor is simulated. In the third step, the readout chip electronics are accounted for. In the following sections, the three steps are discussed in detail.

4.5.1 Simulation of the Electric Field

The electric field is obtained with Technology Computer-Aided Design (TCAD) device simulation [157]. Only one quarter of a pixel is simulated, and a 4-fold symmetry is applied to the electric field obtained in this way. The effect of the punch-through bias dot is not included in this simulation, since it breaks the symmetry. At a tilt angle of $\approx 19°$, the effect of the bias dot can indeed be neglected, since only a small amount of charge carriers are affected by it. However, at 0°, there are tracks for which most of the generated charge carriers end up in the bias dot region.

On the left hand side, Figure 4.17 shows a 3D model of the quarter pixel sensor. The various colors represent different doping concentrations: green represents the silicon substrate with low n doping with a concentration of $10^{12}\,\text{cm}^{-3}$. Blue corresponds to high-dose p doping and red represents high-dose n doping, both amounting to $10^{18}\,\text{cm}^{-3}$ at the surface, with a steeply falling profile towards the inside of the sensor. The concentration of the p-spray is a factor of 50 lower than the high-dose dopings. The grid points of the generated mesh are chosen such that there is a high point density where there is a large doping gradient.

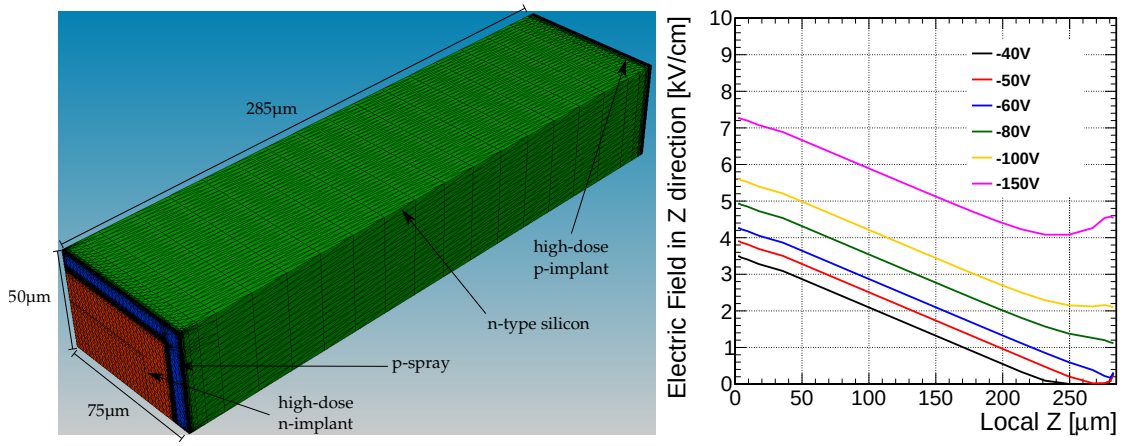

Figure 4.17: Left: A 3D model of one quarter of the pixel sensor. The various colors indicate different doping concentrations. Right: The Z (depth) component of the electric field in the middle of a pixel as a function of its depth, after device simulation with TCAD for different bias voltages.

The TCAD device simulation obtains the electrostatic potential by solving the Poisson equation and the continuity equations for electrons and holes simultaneously. The boundary conditions are chosen such that on the n-implant at the front and the p-implant at the back the potential is fixed to 0 V. At the sides of the sensor, Neumann boundary conditions are implied, i.e. $\vec{E}_{\perp} = 0$. Once the equations are solved for this configuration, the potential at the back side is changed to -150 V in small steps, where the solution of the previous step is used as initial value in the numerical solver for the next step.

On the right hand side of Figure 4.17, the depth component of the electric field is shown, as a function of the depth of the sensor. The n^+ side is at 0 and the p^+ side at 285 µm. The simulation reproduces the mostly linear behavior of the electric field, and the full depletion voltage is at ≈ 60 V, consistent with the real sensor.

4.5.2 Charge Carrier Simulation

In the next step, the PIXELAV Monte Carlo is run. The program simulates the transition of pions through silicon. However, since the ionization of pions at high energies is very close to that of electrons, it can be taken directly for the description of the DESY electron test beam. The procedure consists of three steps:

1. Charge generation: the trajectory of an incident pion is tracked through the silicon, and electron-hole pairs are generated. The energy loss $\mathrm{d}E/\mathrm{d}x$ for pions in silicon is taken from Figure 9 in [158]. The simulation of high-energetic secondary electrons (delta rays) propagating through the material is supported. The stopping power $\mathrm{d}R/\mathrm{d}E$, used to compute the range of delta rays inside the silicon, is taken from the NIST ESTAR program [159].

2. Charge transport: each electron and hole produced in the previous step is propagated through the silicon, using the electric field obtained with the TCAD simulation. The equations of motion are solved with a 6^{th} order Runge-Kutta method.

3. Charge collection: electrons that reached the front side of the sensor are counted, and a grid of pixels with size $100 \times 150\,\mu\text{m}^2$ is filled. The output of this step consists of the number of electrons in each pixel, corresponding to the input of the pre-amplifier in that pixel.

This is the most time-consuming part of the simulation. In order to speed up the processing, only every tenth electron or hole is propagated through the silicon, and the resulting number of electrons on each pixel is multiplied by 10. This allows event generation rates of $\mathcal{O}(10\,\text{Hz})$ on a single CPU core.

4.5.3 Post-Processing

In the last step, the electronics of the readout chip is accounted for by post-processing the output from the previous step, taking into account several known effects:

- Cross-Talk: The charge in each pixel contributes to the charge in adjacent pixels in row direction with the cross-talk factor f:

$$A_i' = (1-2f)A_i + fA_{i+1} + fA_{i-1}\,, \tag{4.10}$$

  where A_i is the pulse height of the ith pixel, and A_{i+1} and A_{i-1} are the neighboring pixels. The parameter f has been tuned to the data and a value of $f = 2.5\%$ was found. This effect models cross-talk between pixels due to parasitic capacitive coupling.

- Threshold: Pixels below the readout threshold are set to 0. On the readout chip, there is a small variance of effective thresholds observed between pixels. Therefore, for each pixel, the threshold is smeared with a Gaussian with a width of 100 electrons in the simulation.

- Gain: After the threshold has been applied, the total number of electrons in a pixel is smeared by a Gaussian with a width of 550 electrons. This takes into account gain variations due to non-uniform behavior of all pixels on the chip.

The steps are applied in the order in which they are presented here, which is important since they do not commute. The pixel array after correcting for the effects introduced by the readout chip is then used to apply the same algorithms for cluster finding and hit reconstruction as for the test beam analysis. The true hit position is known from the MC truth information, so the beam telescope does not need to be simulated.

4.5.4 Results

Figure 4.18 shows the cluster charge distribution in simulation (red) compared to test beam data (black) on the left hand side, for a run with the sensor perpendicular to the beam. About 100 000 events have been simulated and the total number of clusters in simulation has been scaled to match that of the data to allow for a comparison between the two shapes. Fitting a Landau distribution convoluted with a Gaussian, as described in Section 4.3.2, gives a most likely charge of 21.4 ke for the test beam data and 20.5 ke for the simulation, with negligible statistical uncertainties. The distribution in simulation is $\approx 20\,\%$ narrower than the data, but overall, fair agreement is observed.

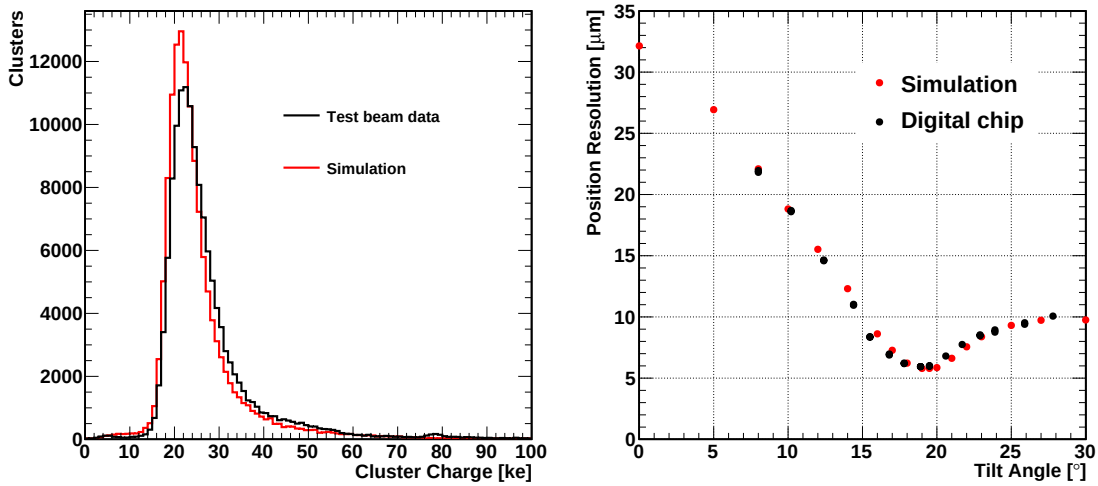

Figure 4.18: Left: Cluster charge distribution in data (black) and simulation (red). The simulation curve is a little narrower and shifted to lower charge values by ≈ 0.9 ke. Right: Position resolution as a function of the tilt angle with respect to the beam, for data and simulation.

On the right hand side, the position resolution as a function of the tilt angle is shown for simulation (red) and test beam data with the digital chip (black). The data points correspond to the same runs as shown in Figure 4.11. Each simulation point corresponds to a run with 100 000 events for that particular tilt angle. The simulation is able to describe the test beam data very well.

4.6 Summary

The upgrade of the pixel detector is an important project to maintain the physics performance of the CMS experiment in the phase of high-luminosity running. The position resolution of the pixel detector is important for accurate p_T and impact parameter measurement of tracks. This is crucial for many track-based physics objects, most importantly b-tagging and tau identification.

The DESY test beam provides a way to test prototypes of the upgraded readout chip. Other than finding potential problems with the readout chip itself that can be fixed before the final production, its behavior can be studied and compared to the previous iteration of the chip that is in operation in CMS. The AIDA beam telescope allows precise tracking of beam particles, making it possible to measure hit detection efficiency[2] and position resolution.

The main feature of the new readout chip is larger buffers to be able to record hits at high rates without data loss, and the possibility to operate at lower readout thresholds before being overwhelmed by electronic noise. Low readout thresholds in combination with charge sharing between pixels due to inclined tracks or the Lorentz force inside the CMS magnetic field allow for a much better position resolution than the pixel size alone. With

[2] Hit efficiencies of more than 99.96 % have been measured at $\mathcal{O}$ (kHz) rates in the DESY test beam (not presented here).

the current iteration of the chip, test beam measurements show that a position resolution of about 8 µm can be achieved with a pitch of 100 µm. This is consistent both with what is measured in CMS itself and with a Monte Carlo simulation of the sensor.

Prototypes of the readout chip foreseen for the upgraded pixel detector have been measured in the test beam and the results were presented in this chapter. The new, digital chip is well understood under beam conditions and resolutions down to 6 µm have been measured. This corresponds to a 25 % improvement compared to the analog chip. The results confirm the improvements made to the chip by its designers, and for the first time the gain in resolution is quantified.

5 Search for $H \to \tau^+\tau^-$ Decays

The search for the Higgs boson is one of the primary objectives of the physics program at the LHC. This milestone was achieved when the discovery was announced on July 4, 2012 [24, 25]. What has been discovered is a new resonance that is, within uncertainties, compatible to the Standard Model. More precise measurements are needed in order to verify that all predictions made by the Standard Model about its properties are realized in nature. For some of these measurements, more data than what has been collected in the first run of the LHC are needed, others however can be made already.

One such measurement is the coupling of the Higgs boson to fermions. At low Higgs masses, $m_H \lesssim 160\,\text{GeV}$, there are five major decay channels which are experimentally accessible. Initially, decays of the Higgs boson have been observed in the $\gamma\gamma$ [28, 29], W^+W^- [32, 33] and ZZ [30, 31] channels. All of these are bosonic channels. However, as outlined in Section 1.4.1, the fermions acquire their mass through Yukawa interactions instead of directly through the spontaneous symmetry breaking. Observing the coupling of the Higgs boson to fermions is therefore a direct verification of the Yukawa mechanism.

Along with the Spin and CP results [36, 37], the evidence of the newly discovered resonance coupling to fermions in the $\tau^+\tau^-$ and $b\bar{b}$ final states [34, 35, 160] was one of the most important Higgs results since the discovery was announced.

In this chapter, a brief overview of the CMS data analysis in the $\tau^+\tau^-$ final state is given. It should be pointed out that this represents the work of a large group within CMS. Furthermore, the description does by no means give complete coverage of all analysis details. For more detailed coverage it is suggested to consult the individual CMS Analysis Notes [161, 162, 163, 164, 165, 166, 167, 168] on the topic. This chapter is meant to convey the general idea of the analysis strategy before discussing two selected topics in detail – Tau Embedding and Associated Production – in the following two chapters.

In Section 5.1, the problem of mass reconstruction in $\tau^+\tau^-$ final states is briefly discussed. In Section 5.2, the dataset of recorded and simulated data used in the analysis is presented. Section 5.3 discusses the selection of candidate Higgs events. Section 5.4 presents the modeling of the background contributions in this channel and Section 5.5 discusses the major systematic uncertainties and how they are treated. Finally, in Section 5.6 the results are interpreted and quantified.

5.1 Invariant Di-Tau Mass Reconstruction

One of the most important tools for studying resonances is the reconstruction of the resonance mass. In the $\tau^+\tau^-$ channel, some momentum is carried away by the neutrinos produced in the tau decays. Since the neutrinos cannot be reconstructed, this leads to a loss of information, making it impossible to accurately determine the resonance mass: the four vectors of the reconstructed tau decay products do not sum up to the four vector of the resonance. Different techniques have been studied to still reconstruct the mass. Most try to recover some of the information carried away by the neutrinos by exploiting the

Table 5.1: The branching ratios and dominant backgrounds for the various $\tau^+\tau^-$ final states in the Higgs Search. All of these channels are studied by CMS. This table does not include the associated production channels.

Final State	Branching ratio	Dominant background
$\tau_{\text{had}} + \tau_{\text{had}}$	42%	QCD multijets
$\tau_\mu + \tau_{\text{had}}$	23%	$\text{Z}/\gamma^* \to \tau\tau$
$\tau_e + \tau_{\text{had}}$	23%	$\text{Z}/\gamma^* \to \tau\tau$
$\tau_e + \tau_\mu$	6%	$\text{Z}/\gamma^* \to \tau\tau$
$\tau_e + \tau_e$	3%	$\text{Z}/\gamma^* \to ee$
$\tau_\mu + \tau_\mu$	3%	$\text{Z}/\gamma^* \to \mu\mu$

missing transverse energy, $E_{\text{T}}^{\text{miss}}$, in the event. In CMS, a likelihood-based approach called SVfit is used to find the mass which is most compatible with the tau decay kinematics and $E_{\text{T}}^{\text{miss}}$. Appendix B describes the method in more detail and compares it to other mass reconstruction methods.

5.2 Data Used in the Analysis

The data used in this analysis was taken by CMS in the year 2011 at 7 TeV and in 2012 at 8 TeV of center-of-mass energy, with an integrated luminosity of $4.9\,\text{fb}^{-1}$ at 7 TeV and $19.7\,\text{fb}^{-1}$ at 8 TeV. The data events have been triggered by di-lepton triggers or triggers that require both a lepton and a hadronically decaying tau candidate, depending on the final state. This allows to keep the p_{T} thresholds on the individual objects as low as possible, and the acceptance high.

The data taken in the year 2010, corresponding to $36\,\text{pb}^{-1}$ at 7 TeV, are not used in the analysis since the technical effort would not be justified by the very small amount of data that would be added (only little more than 1 per mille in integrated luminosity).

The analysis also uses Monte Carlo simulation for cross-checks and estimation of contributions from background and signal processes. Most background samples, including $\text{Z}/\gamma^* \to \ell\ell$, $W^\pm \to \ell\nu$, $t\bar{t}$ production and di-boson production, were generated with MADGRAPH interfaced to PYTHIA for parton shower and hadronization. The signal samples for Higgs boson production in gluon-gluon fusion and vector boson fusion are generated at NLO with POWHEG interfaced to PYTHIA whereas the associated production of the Higgs boson with a W or Z boson or a $t\bar{t}$ pair was generated at LO with PYTHIA.

It has been found that NNLO effects are sizable for gluon-gluon fusion [169], so that the distributions obtained from the simulation have been reweighted as a function of the transverse momentum of the Higgs Boson. The correction factors were calculated using the numerical tool HRES [170].

5.3 Event Selection and Categorization

5.3.1 Event Selection

The multiplicity and flavor of final state leptons depends on the tau lepton decays and on whether there exist additional leptons from decays of W bosons, Z bosons or $t\bar{t}$ pairs produced in association with the Higgs boson. In Table 5.1, the various possible final

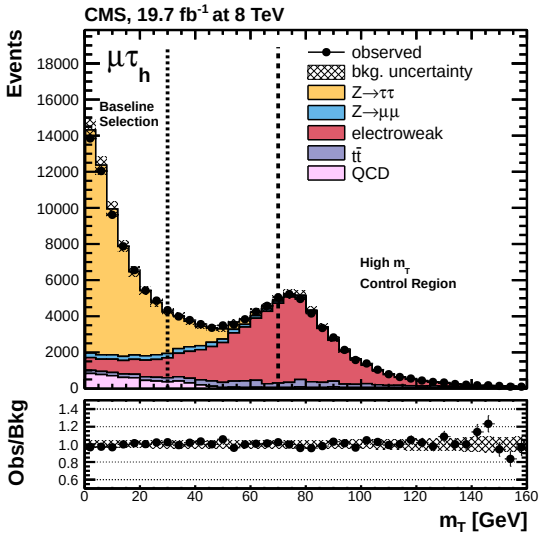

Figure 5.1: The distribution of M_{T} in the 8 TeV data in the $\tau_\mu + \tau_{\mathrm{had}}$ channel. The dashed line indicates the cut at 30 GeV. The region above 70 GeV is used for background estimation since it contains almost only W + jets events. Picture taken from [35].

states that are analyzed in CMS are summarized. In most cases, the Drell-Yan process ($\mathrm{Z}/\gamma^* \to \ell\ell$) is the dominant background. Due to the multitude of final states, many individual subchannels have to be combined, which makes the $\tau^+\tau^-$ analysis complex. The semileptonic channels, $\tau_\mu + \tau_{\mathrm{had}}$ and $\tau_e + \tau_{\mathrm{had}}$, are the most sensitive channels because of the high branching ratio. Only the fully hadronic channel has a higher branching ratio, however it is harder to separate from the QCD multijets background. In this section, a rough outline of the analysis is given without going too much into detail, mostly concentrating on these high-sensitivity channels.

Two well-reconstructed and isolated leptons are required to exist in the event, as described in Section 3.4. The two objects are required to be separated geometrically by $\Delta R > 0.5$, to avoid that the two reconstructed leptons originate from the same physical particle. Next, events with more than two well identified light leptons are rejected. More than two hadronic tau candidates are typically allowed in order not to compromise the overall acceptance. In such a case, one candidate pair, typically the one with highest scalar p_{T} sum, is chosen.

More cuts are applied to reduce background contributions. These often depend on the final state, though. For example, in many channels, the contribution from W + jets is sizable. In the $\tau_\mu + \tau_{\mathrm{had}}$ channel, this contribution comes from the W decaying into a muon and one of the jets being misidentified as a hadronic tau decay. This background can be reduced with a cut on the transverse mass M_{T} of the muon and $E_{\mathrm{T}}^{\mathrm{miss}}$, defined as

$$M_{\mathrm{T}}(\ell, E_{\mathrm{T}}^{\mathrm{miss}}) = \sqrt{2 p_{\mathrm{T}}^{\ell} E_{\mathrm{T}}^{\mathrm{miss}} \left(1 - \cos\left(\Delta\phi\right)\right)}\,, \tag{5.1}$$

where $\Delta\phi$ is the azimuthal angle between the muon and the $E_{\mathrm{T}}^{\mathrm{miss}}$ vector.

Figure 5.1 shows the distribution of the transverse mass M_{T} in the $\tau_\mu + \tau_{\mathrm{had}}$ channel. While the Higgs signal and the $\mathrm{Z}/\gamma^* \to \tau\tau$ background peak around 0, the W + jets peak

is close to the W boson mass around 80 GeV. In both the $\tau_\mu + \tau_{\text{had}}$ and the $\tau_e + \tau_{\text{had}}$ channels, the cut is chosen such that events with $M_{\text{T}} > 30$ GeV are rejected.

5.3.2 Event Weights and Scale Factors

Due to imperfections in the Monte Carlo simulation, there are differences between the data and the simulation. Sources for this can be missing higher order corrections, simplifications in the detector simulation, or the use of heuristic models to describe physical effects. Differences in observables that are important for the analysis are corrected for with event weights. For the $H \to \tau^+\tau^-$ analysis, pile-up weights and lepton efficiencies are the most important:

- Pile-Up: While the effects for multiple interactions in the same bunch crossing (pile-up) are included in the simulation, the exact distribution of the number of additional interactions is not the same in data and simulation, since the pile-up scenario for the simulation was already fixed before the data taking was complete. Therefore, a re-weighting procedure is applied to the simulated events. Each event is assigned an event weight such that the distribution of the number of additional interactions matches the data. In simulation, the number of additional interactions is known from the Monte Carlo truth information, and in data it can be estimated from the instantaneous luminosity in each bunch crossing when the total pp inelastic scattering cross section is known. For a center-of-mass energy of 7 TeV it is taken as 68.0 mb [171] and for 8 TeV it is 69.4 mb [172].

- Lepton Efficiencies: The efficiency for a lepton to activate the trigger and to pass the identification and isolation criteria outlined in Section 3.4 is not necessarily the same in data and in simulation. Typically, such differences are on the percent level. While this is very low, it is on the same order of magnitude as a possible Higgs signal in some channels, so that these effects and their uncertainties need to be understood.

  In order to correct for efficiency effects, the trigger, identification and isolation efficiencies are measured in data and simulation with a technique called *tag-and-probe* [118, 173]. This technique is especially challenging for obtaining the τ_{had} identification efficiency, since there is no background-free region. This measurement has been performed with $Z \to \tau^+\tau^- \to \tau_\mu + \tau_{\text{had}}$ events and it was found that, within 7%, the scale factor is consistent with 1 [174, 175]. For electrons and muons, the measurements are easier and have been performed as well [176, 177].

5.3.3 Event Categorization

In all sub-channels, events are split into several, non-overlapping event categories. The benefit when categorizing events is that, in some categories, the signal to background ratio is higher than when considering all events inclusively. One could make a cut instead, however in that case one would lose all events not surviving the cut. Some categories only have very few events in them so that the statistical precision is low. This is compensated by a high signal to background ratio. In order to quantify the sensitivity of the analysis and the significance of a possible excess, it is important to combine the results in all categories eventually. This will be discussed later in Section 5.5.

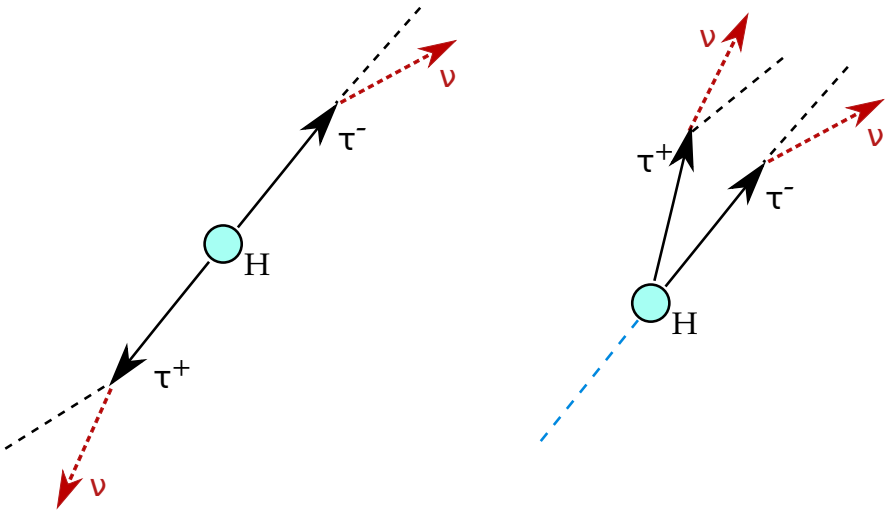

Figure 5.2: The topology of a $H \to \tau\tau$ decay when the Higgs boson is at rest (left) and when it has a high transverse momentum (right), as seen in the laboratory frame. The dashed red lines represent neutrinos from the tau decays and the black dashed lines represent the visible decay products. In the first case the two red vectors sum to approximately 0, whereas in the second case $E_{\mathrm{T}}^{\mathrm{miss}}$ will be clearly different from 0 and the direction of $E_{\mathrm{T}}^{\mathrm{miss}}$ is easier to determine.

The construction of event categories is guided by the different production mechanisms. In the $\tau_\mu + \tau_{\mathrm{had}}$ channel, the primary criterion for categorization is the number of accompanying jets in the event: events are categorized in 0-jet, 1-jet and 2-jets categories. Most events fall in the 0-jet category. The 1-jet category is motivated by the gluon-gluon fusion process when an additional gluon is radiated away, leading to the jet in the event. In this case, the Higgs boson system recoils against this jet. This improves the resolution of $E_{\mathrm{T}}^{\mathrm{miss}}$, since the only source of $E_{\mathrm{T}}^{\mathrm{miss}}$ are the two tau lepton decays and the contributions to $E_{\mathrm{T}}^{\mathrm{miss}}$ from each tau decay is not mostly canceled by the other, as it would be in the back-to-back case. Figure 5.2 illustrates this principle. The improved $E_{\mathrm{T}}^{\mathrm{miss}}$ resolution in turn facilitates the invariant mass reconstruction.

The 2-jet category is motivated by the vector boson fusion (VBF) production mechanism. In this case, the two jets are expected to be highly separated from each other in pseudorapidity, with not much additional hadronic activity between the two jets. This allows good suppression of all backgrounds, including $\mathrm{Z}/\gamma^* \to \tau^+\tau^-$. In addition, the Higgs system recoils against the two jets, so all the benefits from the 1-jet category apply here, too.

Each of these categories is divided further into sub-categories using the p_{T} of one of the tau decay products or the p_{T} of the di-tau system, defined as

$$p_{\mathrm{T}}^{\tau\tau} = \left| \vec{p}_{\mathrm{T}}^{\,L} + \vec{p}_{\mathrm{T}}^{\,L'} + \vec{E}_{\mathrm{T}}^{\mathrm{miss}} \right| , \tag{5.2}$$

where L and L' are the reconstructed tau decay products. Table 5.2 summarizes all categories in the $\tau_\mu + \tau_{\mathrm{had}}$ channel for the 8 TeV data. The categories in the other channels are not exactly identical but very similar.

Table 5.2: The different categories used in the $\tau_\mu + \tau_{\text{had}}$ channel for the 8 TeV data. The term "Rap. Gap Veto" means that there is no identified jet with pseudorapidity between the two VBF jets. It can be seen how the categories with highest signal over background ratio have a low overall event yield. The signal corresponds to the Standard Model Higgs boson with $m_H = 125\,\text{GeV}$.

Category	Definition	S/B	Background yield
VBF Tight	$N_{\text{jets}} \geq 2$, $M_{jj} > 700\,\text{GeV}$, $\Delta\eta_{jj} > 4.0$, Rap. Gap Veto, $p_{\text{T}}^{\tau\tau} > 100\,\text{GeV}$	0.165	14.7 ± 1.3
VBF Loose	$N_{\text{jets}} \geq 2$, $M_{jj} > 500\,\text{GeV}$, $\Delta\eta_{jj} > 3.5$, Rap. Gap Veto, not VBF Tight	0.058	80.1 ± 3.4
Boost	$N_{\text{jets}} \geq 1$, $\tau_{\text{had}}\, p_{\text{T}} > 45\,\text{GeV}$, $p_{\text{T}}^{\tau\tau} > 100\,\text{GeV}$, not VBF	0.013	1250 ± 34
1-Jet High	$N_{\text{jets}} \geq 1$, $\tau_{\text{had}}\, p_{\text{T}} > 45\,\text{GeV}$, not VBF, not Boost	0.011	3112 ± 61
1-Jet Low	$N_{\text{jets}} \geq 1$, $30\,\text{GeV} < \tau_{\text{had}}\, p_{\text{T}} < 45\,\text{GeV}$, not VBF	0.005	8934 ± 187
0-Jet High	$N_{\text{jets}} = 0$, $\tau_{\text{had}}\, p_{\text{T}} > 45\,\text{GeV}$	0.011	5743 ± 141
0-Jet Medium	$N_{\text{jets}} = 0$, $30\,\text{GeV} < \tau_{\text{had}}\, p_{\text{T}} < 45\,\text{GeV}$	0.002	40153 ± 1187

5.4 Background Modeling

The major background after all cuts in most cases is $\text{Z}/\gamma^* \to \tau^+\tau^-$, but also other backgrounds such as QCD multijets, W + jets and $t\bar{t}$ production are important. Yet other backgrounds are small and estimated entirely from simulation.

For the four backgrounds mentioned above, data-driven techniques are applied. The goal is to estimate the number and invariant mass shape of the background events from the data themselves, by extrapolating from sideband regions. This has the advantage that many systematic uncertainties related to the simulation do not apply, such as theory uncertainties, uncertainties related to modeling of pile-up interactions, jet or lepton energy scales, or uncertainties on the integrated luminosity. Another benefit is that the statistical precision is always comparable to that of the data in the signal region, since the integrated luminosity is the same, whereas for simulated samples one has to spend extra computing time to generate more events, so that the number of simulated events keeps up with the data.

However, often it is technically more challenging to implement such a data-driven method, and it comes with its own systematic uncertainties due to possible biases in the mass shape when extrapolating from the sideband region to the signal region. In general, the benefits outweigh the disadvantages, though. In the following, the techniques applied for the four major backgrounds are discussed.

- $\text{Z}/\gamma^* \to \tau\tau$: The Z/γ^* background is the only irreducible background. Since it has the same final state as the $H \to \tau^+\tau^-$ signal, there is no sideband region that can be used to extrapolate the yield into the signal region. In order to still estimate this background from the data, the *embedding* method is used: $\text{Z}/\gamma^* \to \mu^+\mu^-$ events are selected, and the muons are replaced by simulated tau leptons. The tau leptons are then decayed with TAUOLA and passed through the CMS detector simulation,

while the original muons are removed from the event. The $Z/\gamma^* \to \mu^+\mu^-$ selection is essentially a signal-free region, muons are much lighter than tau leptons, and therefore the Higgs coupling to muons is suppressed. The yield from the embedded sample is normalized to that from a $Z/\gamma^* \to \tau^+\tau^-$ simulated sample before splitting events in categories.

With this procedure, all event content other than the two leptons, including jets and $E_{\mathrm{T}}^{\mathrm{miss}}$ which play an important role for the event categorization and the di-tau mass reconstruction, are taken from the data. The same applies to contributions from pile-up interactions and underlying event effects. The embedding procedure is discussed in much more detail in Chapter 6.

- W + jets: The background from W + jets events is a significant background in the $\tau_\mu + \tau_{\mathrm{had}}$ and $\tau_e + \tau_{\mathrm{had}}$ channels. The way this enters the selection is that the W boson decays to a light lepton and one of the jets is misidentified as a hadronic tau decay. To estimate this background, a sideband with high transverse mass M_{T} is used. In Figure 5.1, it can be seen that the region with high M_{T} values is dominated by W + jets events. The normalization of the simulated sample of W + jets events is adjusted such that it matches the region with $M_{\mathrm{T}} > 70\,\mathrm{GeV}$. This procedure is performed separately in each of the event categories. The shape of the di-tau mass distribution is taken from the simulation.

- $t\bar{t}$ production: The $t\bar{t}$ production process is one of the major backgrounds in the $\tau_e + \tau_\mu$ channel, because it has two real leptons and, unlike the semileptonic channels, the rate of jets to be misidentified as a signal lepton is small. Again, the shape of the di-tau mass distribution is taken from the simulation. The normalization is fitted to a $t\bar{t}$-enriched control region obtained by additionally requiring b-tagged jets in the event.

- QCD multijets: In QCD multijets events, both tau decays are misidentified jets. Even though the misidentification rates are low, this background is sizable due to the high QCD multijets production cross section at hadron colliders. The QCD background is large in the $\tau_{\mathrm{had}} + \tau_{\mathrm{had}}$ channel, and it is also significant in the $\tau_e + \tau_{\mathrm{had}}$ and $\tau_\mu + \tau_{\mathrm{had}}$ channels. This background is estimated using events in which the two tau decay products have the same charge, and applying a scale factor for the different ratio of same-charge events compared to opposite-charge events. This scale factor is obtained by inverting the isolation on both reconstructed objects, to select pure QCD multijets events. Events from processes other than QCD multijets production in the isolation-inverted region are subtracted by estimating them according to the procedures described above. In practice, the method requires loosening of isolation cuts also in the signal region, in order to have enough statistical precision in the event categories with low event yields.

5.5 Systematic Uncertainties and the Global Fit

The primary goal of this analysis is to measure the quantity $\sigma_H \times \mathrm{BR}\,(H \to \tau^+\tau^-)$, where σ_H is the Higgs boson production cross section and $\mathrm{BR}\,(H \to \tau^+\tau^-)$ is the branching ratio of the Higgs boson decaying into two tau leptons. When the Higgs mass is fixed, this quantity can be calculated theoretically. The signal strength parameter, μ, is defined as

$\frac{\sigma_H \times \mathrm{BR}\left(H\to\tau^+\tau^-\right)}{\sigma_{\mathrm{SM}} \times \mathrm{BR}_{\mathrm{SM}}(H\to\tau^+\tau^-)}$, where the numerator is the measured value and the denominator is the theory prediction. Consequently, $\mu = 0$ corresponds to the case when no Higgs boson is observed in the data, whereas $\mu = 1$ represents the Standard Model Higgs boson.

The signal strength parameter is extracted from the data with a binned maximum likelihood fit, where all bins in the di-tau mass distribution (or, in the case of the $\tau_e + \tau_e$ and $\tau_\mu + \tau_\mu$ final states, a multivariate discriminant) in all categories and channels enter. All bins are treated uncorrelated and the bin content is taken to be Poisson distributed. The parameters in the fit are μ, plus additional nuisance parameters for all sources of systematic uncertainties. This allows systematic uncertainties to be constrained by the fit, and it allows profiling of the nuisance parameters, i.e. re-maximization of the likelihood with respect to the nuisance parameters when performing a likelihood scan in μ or another parameter of interest. While the signal sensitivity in the 0-jet categories is very low, critical uncertainties, such as the lepton energy scales, can be constrained by the fit in such categories with a large number of events. Since many uncertainties are correlated between categories, the more sensitive categories profit from reduced systematic uncertainties.

There are various sources of systematic uncertainties to the analysis. All systematic uncertainties are represented by nuisance parameters that can change the di-tau mass shape and/or total event yield of a process in some category. Nuisance parameters describing systematic uncertainties that alter the yield of a background or signal contribution are taken to be log-normal distributed whereas shape-altering uncertainties are considered to have Gaussian shapes [178].

The major systematic uncertainties in the analysis are channel and category specific. The high sensitivity categories have large uncertainties for background normalizations due to low statistical precision in the samples of simulated events. Especially the uncertainty on the $\mathrm{Z}/\gamma^* \to \tau^+\tau^-$ yield, coming from limited knowledge of identification efficiency and misidentification rates of hadronic tau decays, is relevant. Uncertainties due to integrated luminosity, pile-up, object identification and energy scales are typically only affecting simulated samples. Uncertainties due to extrapolation from sidebands to the signal region for backgrounds such as W + jets and QCD multijets affect those samples only. Table 2 in [35] has a complete list of systematic uncertainties in the CMS $H \to \tau^+\tau^-$ analysis.

5.6 Statistical Interpretation

In order to quantify the compatibility of the observed data with the signal-plus-background hypothesis or the background-only hypothesis, the profile likelihood ratio is used. The likelihood ratio is defined as

$$\lambda(\mu) = \frac{L\left(\mu, \hat{\hat{\theta}}(\mu)\right)}{L\left(\hat{\mu}, \hat{\theta}\right)}, \tag{5.3}$$

where μ is the signal strength parameter and L is the likelihood function, with the signal strength and all nuisances θ as parameters. The symbols $\hat{\mu}$ and $\hat{\theta}$ denote the values that maximize the likelihood, while $\hat{\hat{\theta}}(\mu)$ represents the set of nuisance parameters that maximizes the likelihood for the given μ. The profile likelihood ratio $\lambda(\mu)$ is a number between 0 and 1 which is a measure of how consistent the data are with a given signal strength.

5.6.1 Exclusion Limits

The likelihood ratio can be used to compute an *exclusion limit* for the signal strength parameter. The exclusion limit specifies what values of μ can be excluded to be realized in nature based on the observation. Given a confidence level (C.L.) of $1-\alpha$, then, with a probability of at least $1-\alpha$, the data would should an excess over the background if an excluded μ value was realized. The confidence level is typically chosen to be 95 %. The CL_S method [179] is used to calculate the exclusion limit.

Based on the profile likelihood ratio, there are multiple ways of defining a test statistic to distinguish between two predictions based on their agreement with the observed data. For setting of exclusion limits, both CMS and ATLAS use the definition

$$q_\mu = \begin{cases} -2\log\lambda(\mu) & , \quad \mu \geq \hat{\mu} \\ 0 & , \quad \mu < \hat{\mu} \end{cases} . \tag{5.4}$$

Therefore, high values of q_μ correspond to an increasing disagreement between the observation and a signal strength of μ. The piecewise definition of the test statistic makes sure that a probed signal strength μ below the best-fit to the data $\hat{\mu}$ does not represent less compatibility of the data with a signal strength of μ. Therefore, only an upper limit on the parameter μ is set.

Next, let $f(q_\mu|\mu)$ be the distribution of the test statistic q_μ when the signal strength parameter μ is fixed, and let q_μ^{obs} be the observed value of the test statistic. The integral

$$\mathrm{CL}_{S+B}(\mu) = \int_{q_\mu^{\mathrm{obs}}}^{\infty} f(q_\mu|\mu)\,\mathrm{d}q_\mu \tag{5.5}$$

then gives the probability to find a value for the test statistic to be equal or higher than the one measured. Large values of CL_{S+B} correspond to a high likelihood that the observation is compatible with a signal strength of μ. In a similar way, the integral

$$1-\mathrm{CL}_B = \int_{q_\mu^{\mathrm{obs}}}^{\infty} f(q_\mu|0)\,\mathrm{d}q_\mu \tag{5.6}$$

represents the probability to obtain a value of q_μ equal or larger than the observed one when the background-only hypothesis $\mu = 0$ was realized. The working horse for the CL_S method is the ratio

$$\mathrm{CL}_S(\mu) = \frac{\mathrm{CL}_{S+B}(\mu)}{1-\mathrm{CL}_B}, \tag{5.7}$$

which gives a measure of how well the value μ can be distinguished from the observed value $\hat{\mu}$. The upper limit on μ with a confidence level of $1-\alpha$ is defined such that

$$CL_S(\mu) = 1-\alpha\,. \tag{5.8}$$

5.6.2 Significance of an Excess

In case an excess is observed in the data, the *p-value* specifies how likely it is that a fluctuation of the background causes such an excess or an even larger one. An appropriate test statistic is then given by

$$t_0 = -2\log\lambda(0)\,. \tag{5.9}$$

High values of t_0 correspond to an increasing disagreement between observation and a signal strength of $\mu = 0$. The p-value is defined as

$$p_0 = \int_{t_0^{\text{obs}}}^{\infty} f\left(t_0|0\right) \mathrm{d}t_0 \tag{5.10}$$

where $f\left(t_0|0\right)$ is the distribution of the test statistic t_0 when $\mu = 0$.

5.6.3 The Test Statistic Distribution

In order to compute the integrals for setting an exclusion limit or computing the p-value, the distribution of the test statistic is needed. While it can be computed with some approximations [180], often a more practical way is to use Monte Carlo methods. For a fixed μ, a Poisson random number is thrown in each of the analysis bins, based on the mean number of signal and background events expected. The result of this procedure is called a *toy experiment* and it represents one possible outcome of the experiment when the true signal strength parameter has the value μ. The test statistic is computed for the toy experiment by maximizing the likelihood for the toy data, and by repeating the procedure many times one obtains the full distribution of the test statistic. Care must be taken that enough toy experiments are generated, especially when the observed value is in the tail of the distribution.

Both for exclusion limits and the p-value, in addition an expected value can be computed for reference. This value corresponds to the expectation in case that no Standard Model Higgs boson exists when calculating an exclusion limit, or that a Standard Model Higgs boson exists with $\mu = 1$ when calculating a p-value. The computation is performed with the Monte Carlo method for $\mu = 0$ or $\mu = 1$ fixed, and the median of the distribution of p-values or exclusion limits, respectively, is quoted as the expected value. The 1σ and 2σ quantiles of the distribution are used to quote the expected fluctuation around the expected value.

5.6.4 The CMS Result

The best-fit value for the signal strength parameter in the CMS $H \to \tau^+\tau^-$ analysis is $\hat{\mu} = 0.79 \pm 0.27$ for $m_H = 125\,\text{GeV}$, where the error is obtained by profiling all nuisance parameters. The p-value is shown as a function of different probed Higgs mass hypotheses in Figure 5.3. The highest p-value is observed for $m_H = 120\,\text{GeV}$, and the result is compatible with the newly discovered boson at 125 GeV. Due to the di-tau mass resolution, which is only on the order of 15-20%, the excess in the p-value plot is very broad. The right hand plot in Figure 5.3 shows the di-tau mass peak for all categories in the four major channels combined. Every event is weighted with a factor $S/(S+B)$, where S and B are the expected signal and background yields in the category of the event. In this way, the excess in the more sensitive low-yield categories is not washed out by the less sensitive categories with a higher event yield. A clear excess amounting to $3.4\,\sigma$ in the 125 GeV region can be seen, which for the first time shows direct evidence for the Higgs boson coupling to leptons.

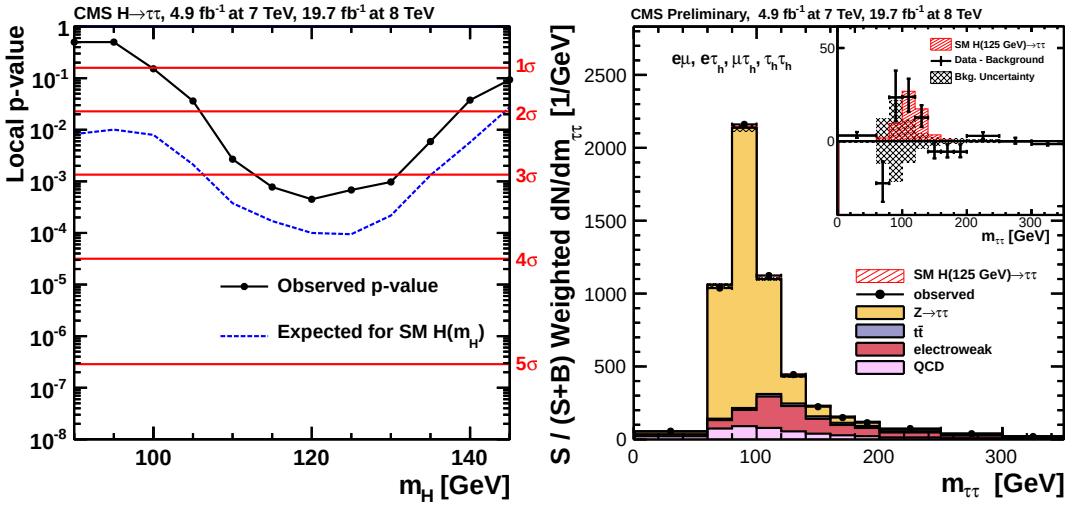

Figure 5.3: Left: The p-value seen by CMS in the $H \to \tau^+\tau^-$ channel, as a function of the hypothesized Higgs boson mass. Right: $S/(S+B)$-weighted plot of the reconstructed di-tau mass in the $\tau_{\text{had}} + \tau_{\text{had}}$, $\tau_\mu + \tau_{\text{had}}$, $\tau_e + \tau_{\text{had}}$ and $\tau_e + \tau_\mu$ channels. From [35].

5.7 Summary

In this chapter, an overview of the $H \to \tau^+\tau^-$ analysis in CMS was presented. There is evidence for the newly discovered boson to decay into $\tau^+\tau^-$, with about the rate predicted by the Standard Model. Future analyses with more LHC data will make the measurement more precise, and will also open the door for property measurements in this exciting channel. For example, the $\tau^+\tau^-$ final state is highly interesting for studying the CP quantum numbers [181]. Also, in the future, the interest in the $H \to \tau^+\tau^-$ channel will remain high, since it is the most accessible way at the LHC to directly probe the Yukawa couplings.

The $H \to \tau^+\tau^-$ analysis is very complex due to the large backgrounds and high number of subchannels and categories, making the analysis very challenging. In this chapter, many of the details of the analysis have been omitted and instead the basic concepts were introduced and discussed. This will be the base for the next two chapters which discuss two aspects of the analysis in more detail. Chapter 6 describes the technicalities and validation of the embedding technique. Chapter 7 discusses in detail the analysis in the sub-channel in which the Higgs boson is produced in association with a W boson and decays into two tau leptons both of which further decay hadronically.

6 The Tau Embedding Technique

In most $H \to \tau^+\tau^-$ analyses, the $Z/\gamma^* \to \tau^+\tau^-$ process is an irreducible background. It has the same final state as the signal, which makes it hard to efficiently suppress, and also to find a signal-free sideband region in the data that could be used to estimate the background. There are only subtle differences between the two processes:

On one side, the Higgs boson is a scalar particle with spin 0 while the Z boson has spin 1. The decay of the Higgs boson can therefore lead to different angular distributions of the decay products compared to the Z boson, depending also on the polarization of the Z boson. On the other hand, the Higgs boson is predominantly produced in gluon-gluon fusion, while the Z boson can only be produced by the fusion of a quark and an antiquark. In a pp machine such as the LHC, the antiquarks in the proton have a softer momentum spectrum than the quarks. Therefore, the Z boson will typically be boosted in the forward or backward direction. When produced in gluon-gluon fusion however, the Higgs boson is produced more centrally.

However both of these features are very hard or impossible to exploit experimentally, because the rest frame of the resonance cannot be reconstructed only from the visible tau decay products. The main discriminant between the two processes is the mass of the resonance, however the mass resolution also suffers from the energy carried away by the neutrinos from the tau decay. Therefore, the reconstructed di-tau mass peak of the Higgs signal is on top of the tail of the background from the Drell-Yan process.

In order to still obtain a signal-free sample of $Z/\gamma^* \to \tau^+\tau^-$ events from the data, the *embedding* method is used. The method exploits the fact that $Z/\gamma^* \to \mu^+\mu^-$ events can be selected with a very high efficiency and purity, and that this is a signal free region, since the coupling of the Higgs boson to muons is more than two orders of magnitude below the Higgs coupling to tau leptons. The muons are then replaced by simulated tau leptons, and the result of the simulation is merged back into the original data event. Due to lepton universality, the kinematics of the Z boson decaying to muons or tau leptons is exactly the same. With this method, some systematic uncertainties related to the simulation of the two leptons remain, such as uncertainties on lepton identification efficiencies and energy scales. However, other systematic uncertainties related to the rest of the event, including theory uncertainties on pile-up interactions and the underlying event, uncertainties on energy scales of jets and $E_{\mathrm{T}}^{\mathrm{miss}}$, and the uncertainty on the integrated luminosity, do not affect the estimation of the $Z/\gamma^* \to \tau^+\tau^-$ background.

6.1 Overview

The embedding procedure consists of multiple individual steps. First, di-muon events are selected in the data and signatures from the muons in the event are removed. Then, a separate $Z \to \tau^+\tau^-$ event is created, based on the kinematics of the original muons, and the detector simulation is run on the event. At this point, the results of the simulated detector response needs to be merged back into the original data event, and the physics object

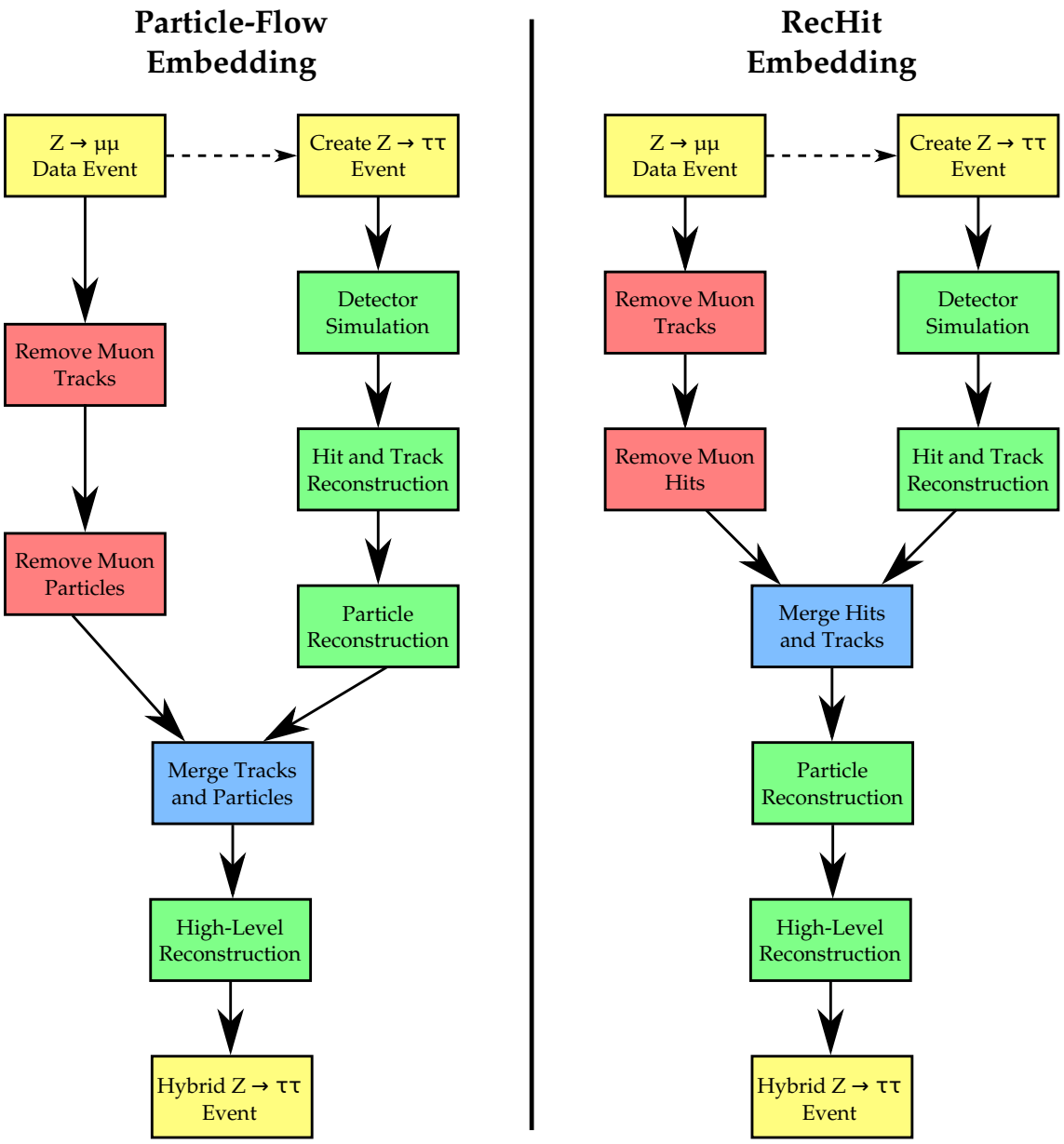

Figure 6.1: Procedure of the embedding technique for particle-flow based embedding (left) and rechit-based embedding (right). The $Z \to \tau^+\tau^-$ event is created with the same four vectors for the taus as the muons in the original event. Green boxes then correspond to production of physics objects while red boxes represent removal of objects. The merge step is indicated in blue. It can be seen how for the rechit-based embedding, the merge step takes place at a lower level in the reconstruction chain. The high-level objects, such as jets, hadronically decaying tau leptons and $E_{\mathrm{T}}^{\mathrm{miss}}$, are reconstructed from the merged event content in both cases.

reconstruction needs to be re-run. There are three different levels in the reconstruction chain where the merging can be performed:

1. Digi Level: Merge the results at the level of digitized detector output. This is the earliest step at which the data format of recorded and simulated events is identical.

2. Rechit Level: Run the track reconstruction on the separate $Z \to \tau^+\tau^-$ event, as well as the reconstruction of hits in the calorimeters and the muon system. Then merge the reconstructed hits and tracks.

3. Particle Level: Run the track and hit reconstruction, and also the particle flow algorithm on the separate $Z \to \tau^+\tau^-$ event, and merge the reconstructed particles.

Figure 6.1 visualizes the individual steps for the embedding on rechit level and on particle level. It can be easily seen how for the rechit level embedding the merging happens at a lower level than for the particle-based embedding. All three methods have their advantages and disadvantages.

The CMS $H \to \tau^+\tau^-$ analysis uses the particle-based embedding (PF embedding), since it has been well studied and commissioned in CMS already [182, 183, 184]. However, it is expected that in future LHC runs with higher luminosity and more pile-up interactions, the method will break down, since, with this method, the effect of pile-up particles on the reconstruction of the Z boson decay products is mostly neglected. This can lead to higher reconstruction efficiencies of leptons compared to the data. With the rechit level embedding (RH embedding), however, the hits from pile-up particles in the calorimeters do contribute and potentially spoil the reconstruction of leptons as they would in the data. Another benefit of the RH embedding is that observables based on calorimeter information, such as calorimeter-based $E_{\mathrm{T}}^{\mathrm{miss}}$, can be used in the analysis. Even though particle-flow based observables are usually preferred in most analyses, this can be important in special cases, for example when modeling the response of the $\ell + \tau_{\mathrm{had}} + E_{\mathrm{T}}^{\mathrm{miss}}$ cross-trigger in embedded samples[1] [185]. The RH embedding method has been pioneered internally in CMS [186], however much progress has been made since then and is presented in this thesis.

In that sense, the embedding on digi level is optimal, since it can take into account all contributions in the data event to the reconstruction of the simulated leptons. However, it is technically very challenging to implement. The main issue is that, in reality, detector components are typically misaligned, while the simulation is being run with ideal detector alignment. Especially the silicon tracker is very sensitive to misalignment, since it provides a very high spatial resolution. In the experiment, the misalignment is corrected by determining the offsets and rotations of all detector modules [58]. For the embedding procedure this is a problem, since a hit in a certain detector module in simulation would have been elsewhere in the real experiment, due to the module being oriented differently. The simulation cannot easily be run with misaligned detector components, either, because the uncertainties on the alignment parameters would lead to overlapping modules in the simulation. Additionally, there is an intrinsic problem with handling noise in the detectors, which also partly affects the RH embedding for the calorimeters, and is discussed more in Section 6.4.3.

The trigger decision can only be simulated with embedding on digi level. Otherwise, the objects required on trigger level are not merged between the original di-muon event

[1] For the L1 trigger, only calorimeter-based $E_{\mathrm{T}}^{\mathrm{miss}}$ is available

and the simulated di-tau event. While both in the PF and RH embedding the trigger simulation is run, it only uses the simulated objects to make the decision. This causes a difference in the trigger response compared to the data when the trigger requires objects to be isolated, when $E_{\mathrm{T}}^{\mathrm{miss}}$ is used to make a decision, or when the presence of pile-up would deteriorate the efficiency to find physics objects on trigger level. However, the embedded trigger response can be used to compute data-over-simulation scale factors to correct for the different trigger efficiency.

In the following, the PF embedding and the RH embedding procedures are described in detail in Section 6.2, and the two are compared and validated in Section 6.3. In Section 6.4, additional techniques to mitigate and quantify systematic effects related to the embedding method are presented. Section 6.5 concludes and gives an outlook where more work is needed in the future.

6.2 The Embedding Procedure

In this section, the individual steps to go from a $\mathrm{Z}/\gamma^* \to \mu^+\mu^-$ data event to a $\mathrm{Z}/\gamma^* \to \tau^+\tau^-$ hybrid event are discussed.

6.2.1 Selection of Di-Muon Candidates

In the first step, suitable di-muon candidate events need to be selected from the data. At this point, a very loose selection is chosen, so that the generated sample covers a large phase space and is biased as little as possible with respect to the true di-muon spectrum. The following requirements are imposed on the events:

- The double-muon trigger with a threshold of 17 GeV for one muon and 8 GeV for the other muon has fired.
- The transverse momentum of the reconstructed muon candidates is greater than 20 GeV and 10 GeV, respectively, to be consistent with the trigger requirement.
- The pseudorapidity of both muons is $|\eta| < 2.4$. The trigger is only efficient up to $|\eta| < 2.1$, however it is important to also cover the phase space beyond this because in the $H \to \tau^+\tau^-$ analysis, the maximum pseudorapidity of the hadronic tau candidates is 2.3. The lower selection efficiency is corrected with event weights.
- Both muons satisfy the identification requirements described in Section 3.4.6.
- The invariant mass of the two muons is greater than 50 GeV. This cut avoids selecting low-energetic backgrounds such as J/ψ mesons or Υ mesons. A practical reason to choose the cut at 50 GeV is that this is what the official CMS MADGRAPH $\mathrm{Z}/\gamma^* \to \ell^+\ell^-$ sample is using, and therefore allows for a consistent comparison between the embedded sample and the MADGRAPH sample. It is important that there is no upper cut on the invariant mass, because the high mass tail is the background for the Higgs signal, and therefore modeling it properly is the primary objective of the method.
- Both muons need to fulfill a loose isolation requirement, in order to reject QCD multijets background. The transverse momentum sum of the charged hadrons within a cone of $\Delta R < 0.4$ around the muon is required to be less than $0.1 \cdot p_{\mathrm{T}}$, where p_{T}

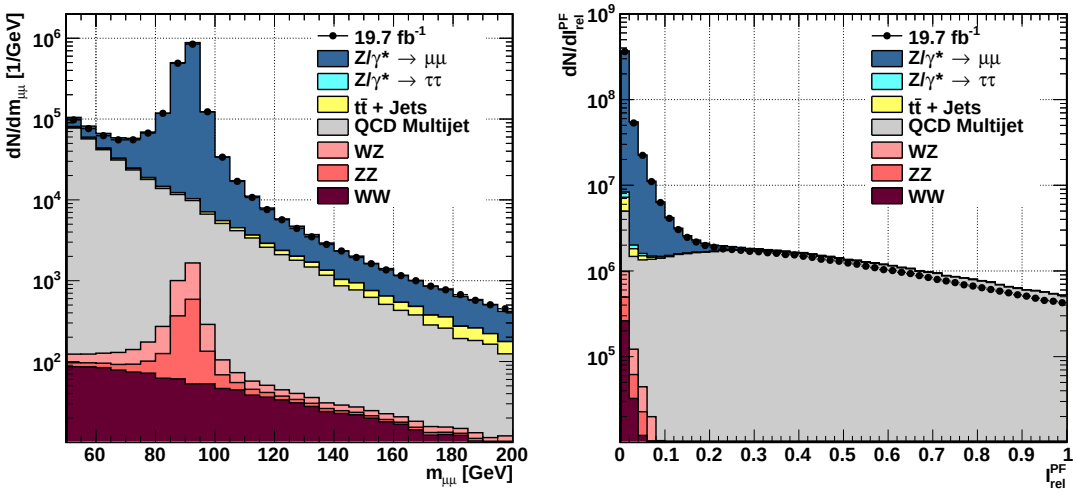

Figure 6.2: Left: The invariant mass of the selected di-muon events in the $\sqrt{s} = 8\,\text{TeV}$ dataset before rejecting the QCD multijets background. While the $Z/\gamma^* \rightarrow \mu^+\mu^-$ events are dominating, there is a significant fraction of QCD events in the sample. Right: Distribution of the relative isolation when counting only charged hadrons in the isolation cone, for one of the muons. The variable is used to reduce the QCD background.

is the transverse momentum of the muon. This cut is chosen to be loose in order to keep the bias of other observables to a minimum. Charged hadrons are chosen instead of all charged particles so that the two muons do not spoil each other's isolation in highly boosted events with mostly collinear muons.

Figure 6.2 shows the di-muon mass distribution without the isolation cut on the left, and the isolation variable on the right. The individual contributions are taken from MADGRAPH Monte Carlo simulation, except the QCD multijets background which is taken from data where both muons have the same charge. The ratio of same-sign muons to opposite-sign muons needed for the correct normalization is taken from the region with the isolation variable between 0.3 and 1.0 for both muons. Contamination from other processes in the same-sign and isolation-inverted regions is estimated with Monte Carlo simulation and subtracted on histogram level. It can clearly be seen that, without the isolation cut, there is a considerable background from QCD processes. It is therefore inevitable to apply the isolation cut.

The resulting di-muon mass spectrum after the full selection can be seen on the left hand side in Figure 6.3. The spectrum is clearly dominated by $Z/\gamma^* \rightarrow \mu^+\mu^-$. However, in certain regions of the phase space, the situation can still look different. For example, on the right hand side in Figure 6.3, the di-muon mass spectrum is shown after additionally requiring one b-tagged jet in the event. This is a typical category for searching a Higgs boson in supersymmetric models, where the coupling of the Higgs boson to b quarks is enhanced. In this case, there is a significant contamination of $t\bar{t}$ events in the selection. There are no additional cuts applied to reduce this background, to prevent introducing a bias, and to keep the efficiency high for the inclusive selection. This problem can be cured on the level of individual analyses that are suffering from it, by running the embed-

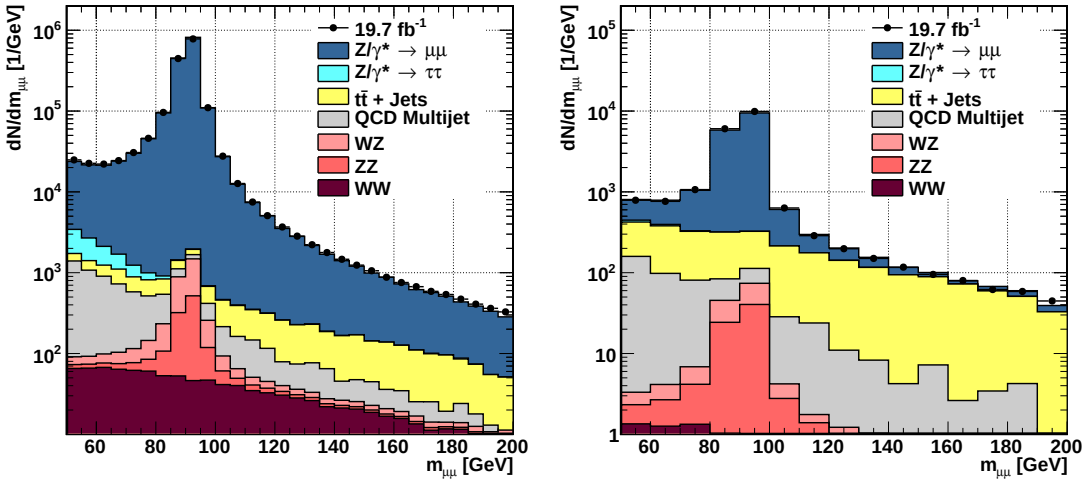

Figure 6.3: Left: The invariant mass of the selected di-muon events after the full selection. The QCD multijets background is suppressed by more than an order of magnitude. Right: The same distribution when additionally requiring a b-tagged jet in the event.

ding procedure on $t\bar{t}$ events obtained from Monte Carlo simulation, and subtracting that contribution from the data embedded sample.

Gaps in the muon system, a finite trigger turn-on curve, and the p_{T}-dependent isolation cut lead to a non-uniformity of the muon selection efficiency in p_{T} and η of the two muons. The selected di-muon sample is therefore inevitably biased as a function of transverse momentum and pseudorapidity of the muons, compared to the true generator-level distribution. Therefore, scale factors are introduced to correct for this, using MADGRAPH $\mathrm{Z}/\gamma^* \to \mu^+\mu^-$ Monte Carlo simulation. The ratio of selected di-muon events in the simulation divided by the number of generated di-muon events within the geometrical acceptance defined by the selection above is used as a scale factor. The efficiency is calculated as a function of p_{T} and η of both muons, parametrized by two 2-dimensional functions, $(p_{\mathrm{T}}^{\mu 1}, p_{\mathrm{T}}^{\mu 2}) \times (\eta^{\mu 1}, \eta^{\mu 2})$. Two 2D histograms are used as look-up tables, shown in Fig. 6.4. Possible differences between data and simulation are in the order of 1-2% [187] and are neglected.

6.2.2 Cleaning of Muon Signatures

After two good muons have been identified, their signature in the detector is removed from the event content. This includes the tracks which were identified as the muon tracks, and the particle flow candidates. For the RH embedding, in addition the hits in the muon system and the calorimeters need to be removed, such that they can be merged later with the response from the simulation of the tau leptons. In the following, the removal procedures in all sub-detectors are described.

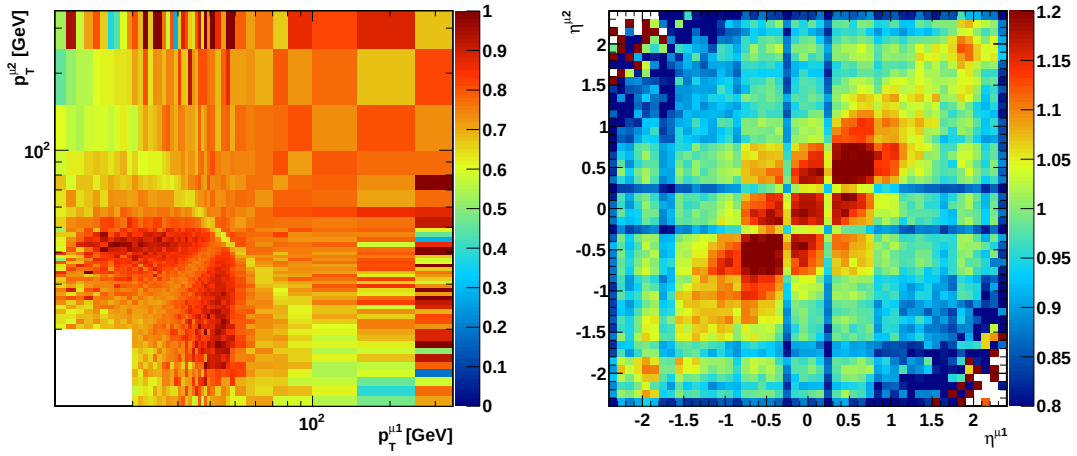

Figure 6.4: The selection efficiency for di-muon events as a function of transverse momentum (left) and pseudorapidity (right) of the two muons. The right hand histogram is already weighted with the efficiency obtained from the one on the left. This allows to simply use the product of the numbers looked up in the two histograms as the total correction factor, without correcting twice for the overall efficiency. The gaps in the muon system can clearly be seen as vertical and horizontal lines of reduced efficiency in the second plot.

Inner Tracker

Tracks in the inner tracker are matched to the muon in η-ϕ space, where the muon four-momentum is obtained from the global track fit. Often, the global track fit is consistent with the track in the inner tracker, and an unambiguous association can be made. However, in rare cases, there can be more complicated situations, so that all reconstructed tracks within an η-ϕ–cone of size $\Delta R = 0.3$ around the muon direction are considered. In case there is more than one track in the cone, tracks with transverse momentum greater than half the muon p_{T} are preferred to those with less than half the muon p_{T}. Additionally, if there is still more than one track in the cone, the track closest in ΔR to the muon is chosen. If there is more than one candidate in a cone of $\Delta R < 0.001$ and with more than 33% of the muon momentum or in a cone of $\Delta R < 0.1$ and with more than 66% of the muon momentum, all such tracks are removed. The rationale behind this is to remove all high p_{T} tracks very close to the muon direction. Multiple tracks can arise in case a muon track is reconstructed as two disjoint segments in the strip and pixel detectors, respectively.

Electromagnetic and Hadronic Calorimeters

Muons deposit only very little energy in the calorimeters. In order to remove that contribution, the track from the inner tracker is extrapolated into the calorimeter region. In this way, the calorimeter cells crossed by the muon and the path length of absorber material traversed by the muon are determined. The mean energy loss $\mathrm{d}E/\mathrm{d}x$ for muons in lead tungstate is taken from [188]. For the brass absorber material in the hadronic calorimeter, the same values are taken and a correction factor is applied. This is justified because the stopping power is only a slowly varying function of the material and the muon energy for

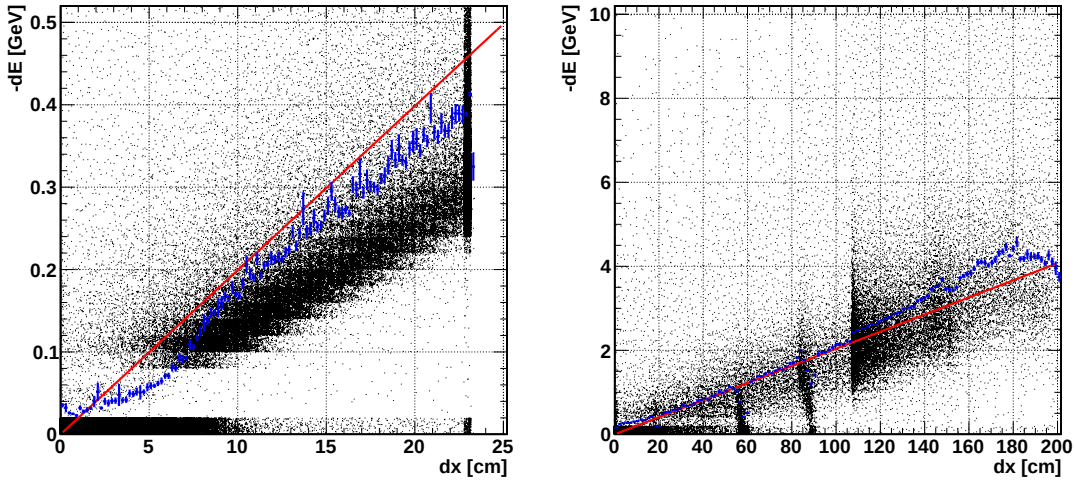

Figure 6.5: Scatter plot of the energy deposited by a muon in the ECAL barrel (left) or HCAL barrel (right). The blue data points represent a profile plot of the same data, showing the mean energy deposited and the error on the mean. The red line is the expectation from theory, without taking detector effects into account.

the muon momenta of interest. The average energy deposited in one of the calorimeters is then given by

$$\Delta E = -\frac{\mathrm{d}E}{\mathrm{d}x} \cdot \rho \cdot l \,, \tag{6.1}$$

where ρ is the density of the absorber material, l is the path length of the muon through the absorber and $\mathrm{d}E/\mathrm{d}x$ depends on the muon momentum.

The actual energy deposited by the muons has been studied with the CMS detector simulation in samples of simulated $Z \to \mu^+\mu^-$ events. No pile-up interactions were simulated, to remove possible contributions from other processes to the calorimeter response. Figure 6.5 shows a scatter plot of the energy deposited in the calorimeter versus the path length of absorber material traversed, on the left hand side for the ECAL barrel and on the right hand side for the HCAL barrel. The blue data points correspond to a profile plot showing the mean energy deposited for a particular path length, and the red line is the expectation from Equation 6.1 for the average muon momentum of the simulated event sample. The large accumulation of points at zero energy is due to the energy deposited being below the readout threshold.

In order to subtract the muon contributions from the calorimeters, the theory value from Equation 6.1 is taken, multiplied by a correction factor obtained from the ratio between simulation and expectation. Figure 6.6 shows that ratio as a function of path length for all relevant sub-detectors. Except for low path lengths where the readout threshold suppresses the detector response, the ratio is approximately flat as a function of path length in all cases. Therefore, a constant correction factor is applied to the expected value, and subtracted from the calorimeter energy. If, due to the subtraction, the energy in a calorimeter hit would be below 0, it is clamped to 0.

The calorimeter deposits are also subtracted from the missing transverse energy at the level of the level-1 trigger (L1ETM). While this does not affect the result of the simulated

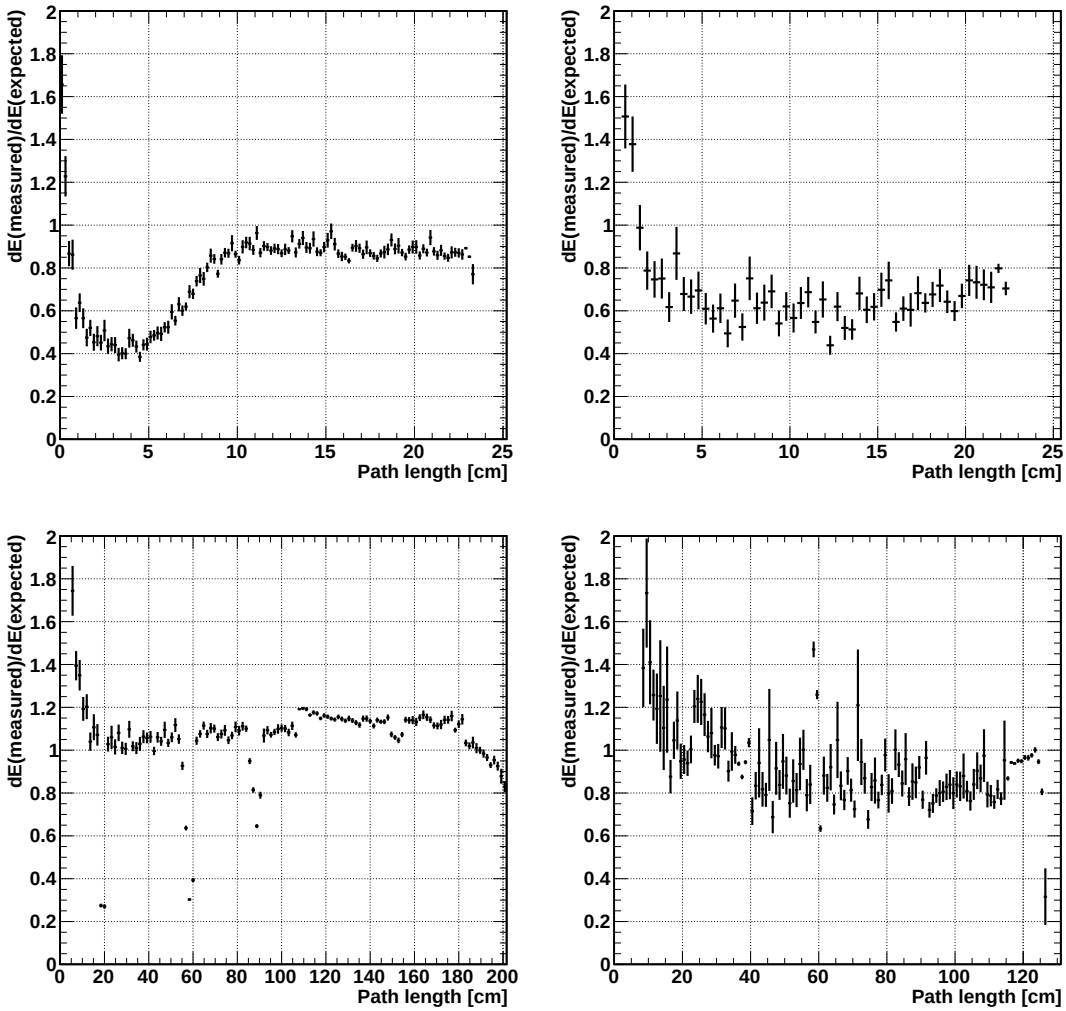

Figure 6.6: Ratio of expected over simulated energy loss for muons in the calorimeter, as a function of path length through the absorber material. The individual plots show the ECAL barrel (top left), ECAL endcap (top right), HCAL barrel (bottom left) and HCAL endcap (bottom right).

trigger decision, the corrected L1ETM value can be used to model the turn-on curve and apply event weights in the embedded sample when using a trigger that makes a decision based on $E_{\mathrm{T}}^{\mathrm{miss}}$ at level 1. The contribution of a muon to L1ETM is studied with Monte Carlo simulation. Events with only a single muon are generated with PYTHIA, where the muon transverse momentum is distributed flat between 10 GeV and 100 GeV. In such events, the only contribution to L1ETM comes from the muon. The L1ETM vector is split in the component parallel to the muon direction and the component perpendicular to it. The perpendicular component is zero on average and has a spread corresponding to the L1ETM resolution. The component parallel to the muon corresponds to the contribution of the muon to L1ETM, again smeared by the L1ETM resolution. The same approach as for the calorimeter hits is then applied: a correction factor to the theory expected value is computed based on the path length of the muon trajectory in the ECAL and

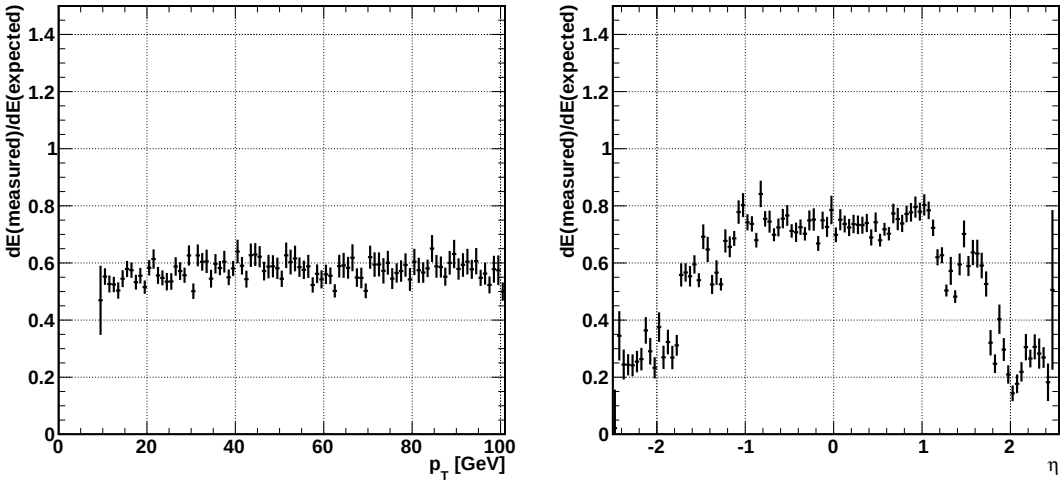

Figure 6.7: Ratio of expected over simulated energy loss as contribution to $E_{\mathrm{T}}^{\mathrm{miss}}$ on the level of the level-1 trigger, as a function of the muon transverse momentum (left) and the muon pseudorapidity (right). While the ratio is flat as a function of p_{T}, it varies in η.

Table 6.1: Correction factors to the expected mean energy loss of a muon in the various calorimeter systems and the contribution to the $E_{\mathrm{T}}^{\mathrm{miss}}$ on level-1 trigger level.

Calorimeter System	Correction Factor
ECAL Barrel	0.9
ECAL Endcap	0.9
HCAL Barrel	1.1
HCAL Endcap	0.9
HCAL Outer	0.8
L1ETM, $\eta < 1.2$	0.75
L1ETM, $1.2 < \eta < 1.7$	0.60
L1ETM, $1.7 < \eta$	0.30

HCAL. Figure 6.7 shows that the correction factor is approximately flat as a function of the transverse momentum, but it has a significant pseudorapidity dependence. As a result, different correction factors are used for different regions of pseudorapidity. The corrected theory value is then added vectorially to L1ETM, where the direction is given by the muon direction.

Table 6.1 summarizes all correction factors used for subtracting the muon contributions to the calorimeter deposits.

Muon System

In the muon system, the removal strategy depends on whether the muon has an outer track (reconstructed only in the muon system) or not. If there is an outer track, all hits from which the track was built are removed, and also all hits that are within 5 cm to the track in both η and ϕ direction. If there is no outer track, the muon trajectory reconstructed in

Table 6.2: Minimum transverse momentum cuts for the visible decay products of the generated tau leptons. The $Z \to \tau\tau$ events are generated such that the tau decay products fulfill these cuts on generator level, in order to increase the statistical precision.

Decay mode	Visible p_{T} cuts [GeV]
$e + \tau_{\mathrm{had}}$	e: 20, τ_{had}: 18
$\mu + \tau_{\mathrm{had}}$	μ: 16, τ_{had}: 18
$e + e$	leading e: 17, subleading e: 8
$\mu + \mu$	leading μ: 18, subleading μ: 8
$e + \mu$	e: 18, μ: 8 OR μ: 18, e: 8
$\tau_{\mathrm{had}} + \tau_{\mathrm{had}}$	leading τ_{had}: 30, subleading τ_{had}: 30

the inner tracker is used to extrapolate to the outer muon system, and all hits within 5 cm to the extrapolated track are removed.

6.2.3 Simulation of the Di-Tau Event

In this step, a di-tau event is generated in which the four-momenta of the tau leptons are taken from the original muons. The muon three-momenta are corrected for the higher mass of the tau lepton. The corrected magnitude of the momentum is given by

$$p^{\tau} = \sqrt{E_{\tau}^2 - m_{\tau}^2} = \sqrt{\left(\frac{1}{2}E\right)^2 - m_{\tau}^2}\,, \tag{6.2}$$

where E is the energy of the Z boson and all energies and momenta are in the rest-frame of the Z boson. In the generated di-tau event, there are no contributions from the underlying event or pile-up interactions. The tau leptons are then decayed with TAUOLA. The treatment of polarization effects is disabled in TAUOLA, since it would require the flavor and kinematics of the incoming quarks to be known. Instead, spin correlation effects are taken into account by creating an event weight with TAUSPINNER.

In order to maximize the available number of events, the tau decay simulation with TAUOLA is repeated until the visible decay products of both tau leptons are above a certain transverse momentum threshold. The rationale behind this is that visible decay products with a very low transverse momentum could not be seen on analysis level, and the event would be "lost" in this case. This is especially true for leptonic tau decays which tend to be very soft. Since the size of the embedded sample is inherently limited by the number of data events recorded, it is crucial to fully exploit the available number of $\mathrm{Z}/\gamma^* \to \mu^+\mu^-$ events.

In order to avoid creating a bias with this method, an additional event weight is assigned to every event. The weight is computed by running TAUOLA 10 000 times, and counting the fraction for which the visible decay products have a transverse momentum above the threshold. The weight is then given by the ratio of passed attempts over all attempts. The exact threshold depends on the decay channel, since on reconstruction level different thresholds are used, and, depending on the decay products, a different experimental momentum resolution is achieved. Table 6.2 lists the thresholds in all 6 channels used for the CMS $H \to \tau^+\tau^-$ analysis.

For the production vertex of the two taus, the reconstructed vertex of the two input muons is chosen. The smearing of the production vertex in the simulation is disabled.

The full CMS detector simulation is then run to generate the detector response to the di-tau event. For the RH embedding, the simulation of electronic noise in the calorimeters is disabled.

6.2.4 Reconstruction and Merging of the Event Content

After the detector simulation, part of the reconstruction chain is run on the generated di-tau event. This includes the track reconstruction in the inner tracker, and the hit reconstruction in the calorimeters and the muon system. For the PF embedding, also the particle flow algorithm is run.

In the next step, the event content is merged between the cleaned di-muon event and the di-tau event. This includes inner tracks, calorimeter hits, and muon chamber hits for the RH embedding and inner tracks and particle-flow candidates for the PF embedding.

After the merging, the particle-flow algorithm is run for the RH embedding. Then, in both cases, high-level objects are reconstructed from the merged event content. This includes jets, $E_{\mathrm{T}}^{\mathrm{miss}}$, and hadronically decaying tau candidates. For the RH embedding, also calorimeter-based jets and calorimeter-based $E_{\mathrm{T}}^{\mathrm{miss}}$ are reconstructed.

6.3 Validation

In this section, the embedding procedure is validated both for the PF embedding and the RH embedding. For the validation procedure, the method is applied to Monte Carlo $\mathrm{Z}/\gamma^* \to \mu^+\mu^-$ events and compared to Monte Carlo $\mathrm{Z}/\gamma^* \to \tau^+\tau^-$ events, both produced with MADGRAPH. Every difference observed between the two would point to a systematic effect of the embedding method itself. Distributions of various crucial observables are shown, most importantly the lepton kinematics, $E_{\mathrm{T}}^{\mathrm{miss}}$, and jet-related variables. A selection of these comparison plots with the most important variables is shown in the text, but the full selection of validation plots is available in Appendix C.

6.3.1 Muon Embedding

In order to validate the procedure technically, in a first step the two selected muons are replaced by generator-level muons instead of generator-level tau leptons. In addition, this allows to study the effect that comes from using reconstructed muons to define the generator-level objects for the embedding procedure, which effectively leads to the detector smearing to be applied twice.

The muon embedding is validated by performing a simple di-muon selection on the embedded events, and then comparing to the original MADGRAPH $\mathrm{Z}/\gamma^* \to \mu^+\mu^-$ sample. Two muons are required with $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ and $p_{\mathrm{T}} > 10\,\mathrm{GeV}$, respectively, and $|\eta| < 2.1$. The double muon trigger has to have accepted the event, and both muons are required to be identified and isolated as described in Sections 3.4.6 and 3.4.8, respectively. Finally, the invariant di-muon mass must be greater than 60 GeV.

Figure 6.8 shows the transverse momentum and the pseudorapidity of the muons, the di-muon mass and the missing transverse energy in the event. In all plots, the total number of events is normalized to the Monte Carlo sample, so that only shape differences are visible, but not differences in the overall selection efficiency. For the $H \to \tau^+\tau^-$ analysis,

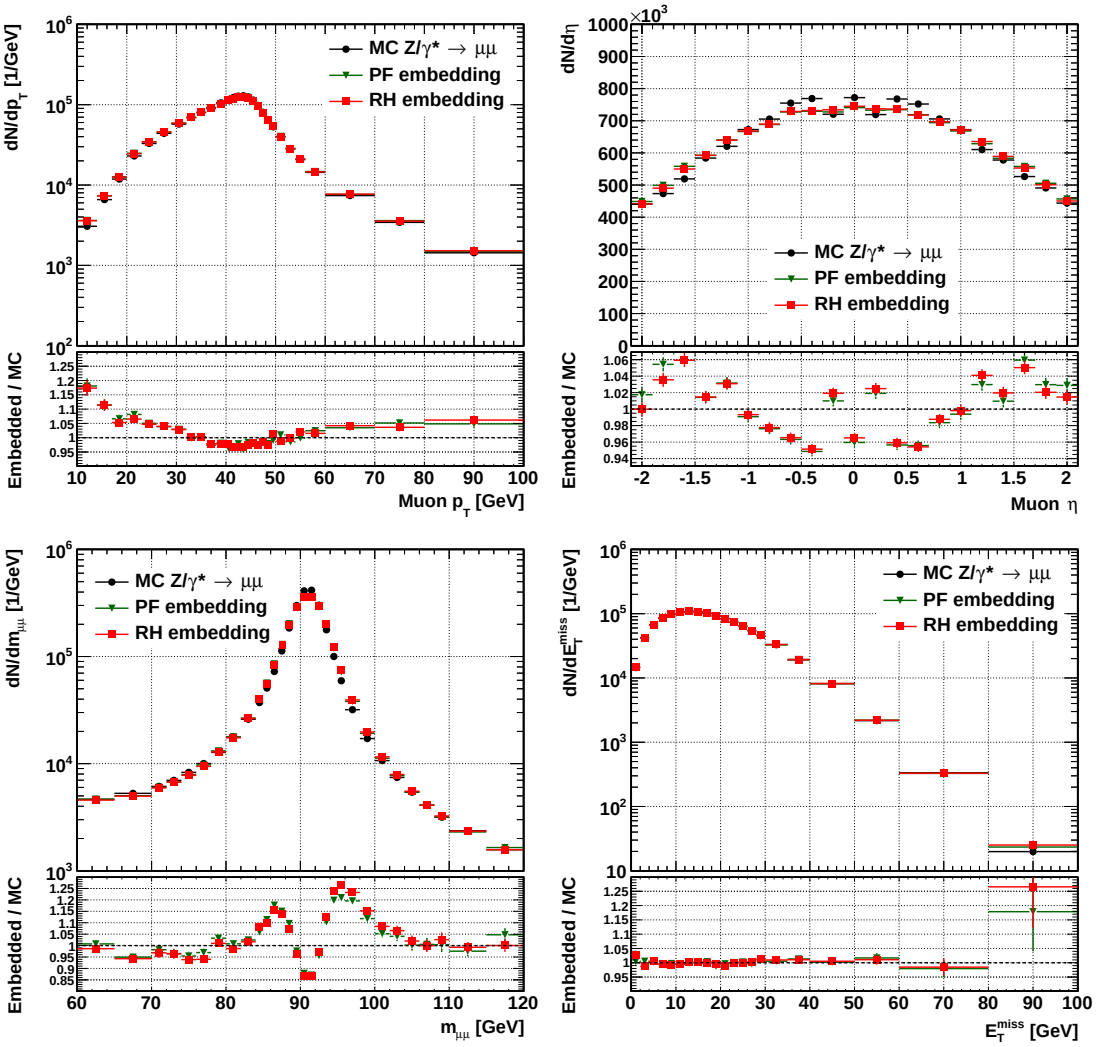

Figure 6.8: Comparison of various observables between direct MADGRAPH $Z/\gamma^* \to \mu^+\mu^-$ simulation and simulation in which the reconstructed muons were replaced by generator-level muons (muon embedding). The top left plot shows the transverse momentum of the positive muon, the top right plot shows the pseudo-rapidity of the positive muon, the lower left plot shows the invariant di-muon mass, and the lower right plot shows the particle-flow based missing transverse energy. The features in the distributions are explained in the text.

only shape differences are of interest, since the normalization of the embedded sample is performed independently.

The top left plot shows the reconstructed transverse momentum of the positive muon. The discrepancy at low p_T comes from the isolation requirement: in the embedded samples, the two original muons that were replaced by generator-level muons, are required to pass an isolation cut, as described in Section 6.2.1. This means that the embedded event tends to be "cleaner" around those muons than on average, and this is still the case after the muons are replaced by generator-level muons. Since the isolation cut is relative to the transverse momentum of the muon, there is a p_T dependency introduced by this effect

which is visible on this plot. The effect will be washed out when applying the method to tau leptons, since the visible tau momentum after the tau decay is a random variable. In addition, in Section 6.4.4, a technique is presented to (partly) avoid this effect.

The top right plot shows the muon pseudorapidity, where a good agreement is observed within 5%. In a similar way as with the isolation, in the embedded samples, the muons in the non-instrumented gap region of the muon system are already filtered out with the selection of the original muons. Therefore, there is no drop around $|\eta| \approx 0.2$ and $|\eta| \approx 1.6$ in the embedded samples.

The invariant di-muon mass is shown in the lower left plot. The mass peak is smeared out in the embedded samples, since effectively the detector effects are applied twice. As with the transverse momentum, this effect is negligible when using the method with tau leptons, since the resolution of the di-tau mass reconstruction is much worse than for muons.

The lower right plot shows the particle flow-based $E_{\mathrm{T}}^{\mathrm{miss}}$. Unlike the other observables, the $E_{\mathrm{T}}^{\mathrm{miss}}$ is sensitive to the full event content, not only the two muons. The good agreement demonstrates that also event content other than the two muons is well described by the embedding method.

It is not surprising that the performance of the PF embedded sample and the RH embedded sample is very similar for the muon embedding. Since the RH embedding mostly improves the modeling for observables determined from the calorimeters, possible improvements cannot be seen in the final state with two muons. This is different for electrons and hadronic tau decays which occur when replacing the muons by tau leptons. This case is discussed in the next section. More validation plots for the muon embedding, including observables on generator level, can be found in Section C.1 of the appendix.

6.3.2 Tau Embedding

In this section, results of the Monte Carlo validation are presented when the muons are replaced by generator-level tau leptons. The procedure is otherwise similar as in the previous section. The two embedded samples are now compared to a MADGRAPH $\mathrm{Z}/\gamma^* \to \tau^+\tau^-$ sample. The embedding procedure is validated in the $\tau_\mu + \tau_{\mathrm{had}}$ and $\tau_e + \tau_{\mathrm{had}}$ channels, since these are the most sensitive channels and together they probe all relevant physics objects.

A simple event selection is performed that is close to the selection of the CMS $H \to \tau^+\tau^-$ analysis, but not exactly the same. In particular, no event categorization is performed. First, an electron or a muon and a hadronic tau candidate are identified as discussed in Section 3.4. Other than that, the following analysis cuts are performed:

- $p_{\mathrm{T}} > 17$ (22) GeV for the muon (electron).
- $p_{\mathrm{T}} > 20$ GeV for the τ_{had} candidate.
- $|\eta| < 2.1$ for the lepton and $|\eta| < 2.3$ for the τ_{had} candidate.
- The "loose" working point of the electron identification is chosen, with the conversion rejection applied.
- Rejection against electrons and muons is required for the τ_{had} candidate.
- $\Delta R > 0.3$ between the lepton and the τ_{had} candidate.

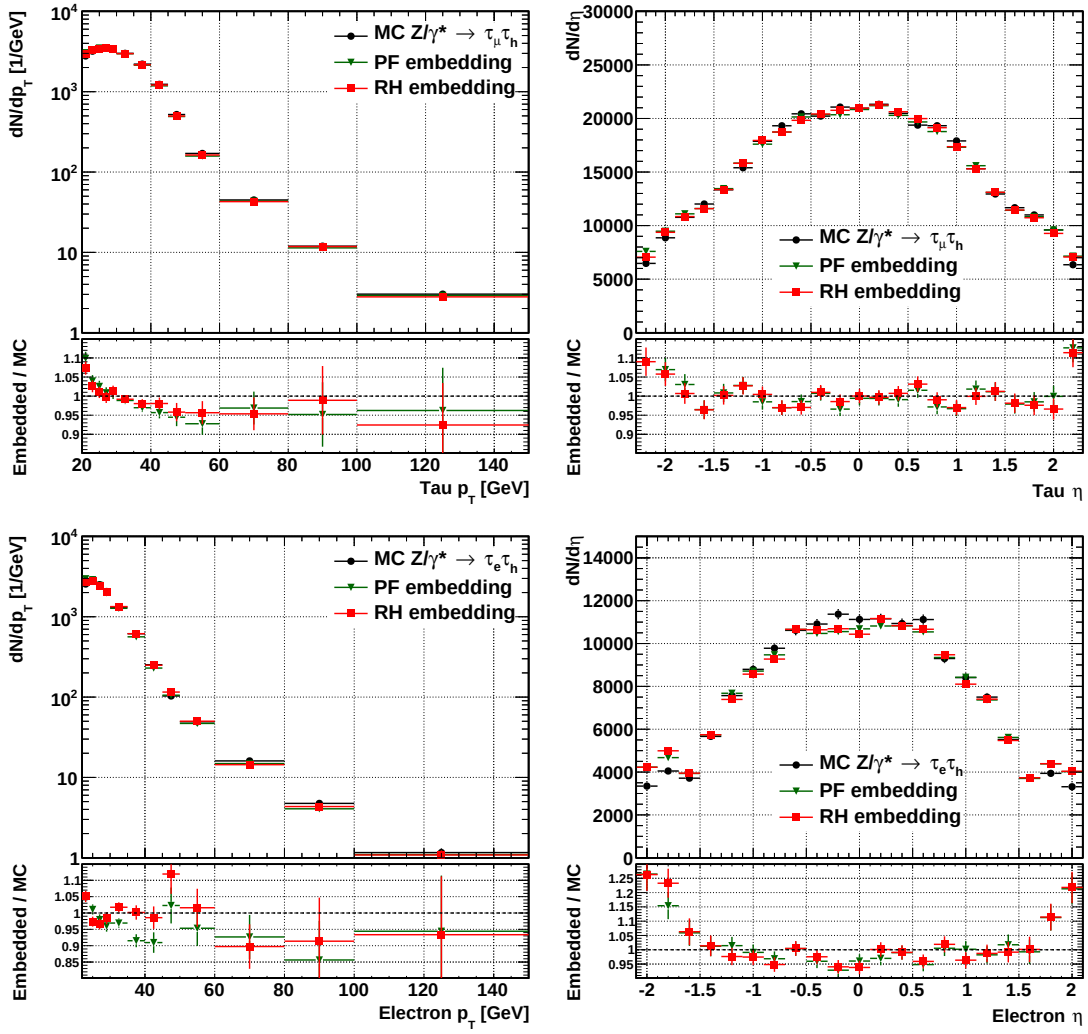

Figure 6.9: Comparison between $Z/\gamma^* \to \tau^+\tau^-$ simulation and simulated embedded events. The two top plots show the transverse momentum and pseudorapidity of the τ_{had} candidate in the $\tau_\mu + \tau_{\text{had}}$ final state, and the two bottom plots show the transverse momentum and pseudorapidity of the electron in the $\tau_e + \tau_{\text{had}}$ final state.

- The transverse mass M_{T} between the lepton and the $E_{\text{T}}^{\text{miss}}$ vector is required to be less than 30 GeV.

In addition, both the reconstructed lepton and the τ_{had} candidate are required to be geometrically matched within $\Delta R < 0.3$ to the visible decay products of the tau leptons on generator level. This allows to determine identification and reconstruction efficiencies. Finally, no trigger is required, since the trigger response is not correctly modeled in the embedded events. In practice, one would determine scale factors corresponding to the trigger efficiency and apply them as a weight to the embedded sample. However, this step is not needed for a purely simulation-based study.

Figure 6.9 shows kinematic variables of the τ_{had} candidate in $\tau_\mu + \tau_{\text{had}}$ events and of the

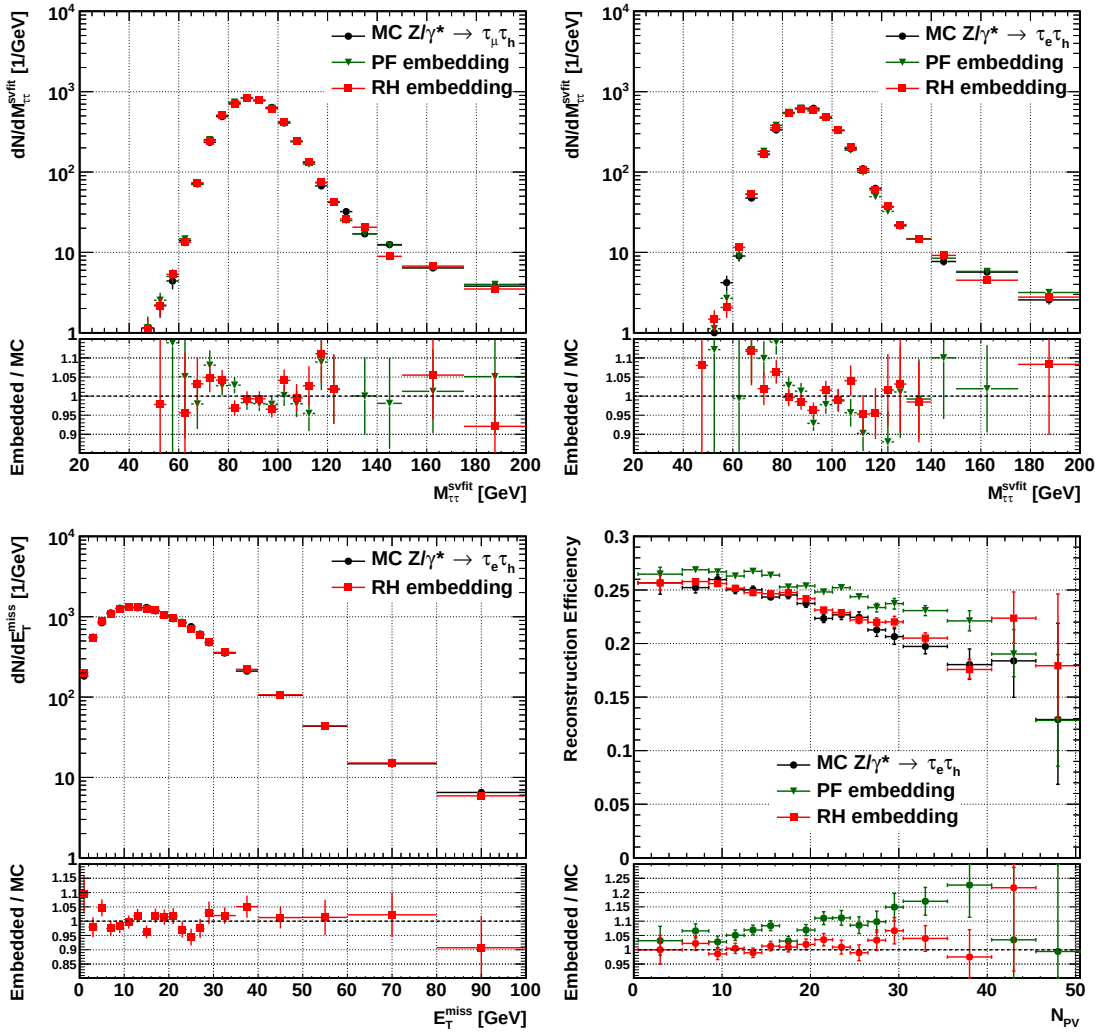

Figure 6.10: Comparison between $Z/\gamma^* \rightarrow \tau^+\tau^-$ simulation and simulated embedded events. The two top plots show the reconstructed di-tau mass in the $\tau_e + \tau_{\text{had}}$ and $\tau_\mu + \tau_{\text{had}}$ final state. The two bottom plots show the calorimeter-based $E_{\text{T}}^{\text{miss}}$ and the selection efficiency in the $\tau_e + \tau_{\text{had}}$ channel.

electron in $\tau_e + \tau_{\text{had}}$ events. The two top plots show the transverse momentum and the pseudorapidity of the τ_{had} candidate in the $\tau_\mu + \tau_{\text{had}}$ final state, and the two bottom plots show the transverse momentum and the pseudorapidity of the electron in the $\tau_e + \tau_{\text{had}}$ final state. The RH embedding clearly improves the modeling of both transverse momentum distributions compared to the PF embedding. In the pseudorapidity distributions, an excess can be seen in the forward region for the τ_{had} candidate and even more clearly for the electron. The reconstruction efficiency at high η is higher in the embedded samples because the signature in the ECAL is cleaner. For the PF embedding, contributions from pile-up are missing when the electron candidate is reconstructed, and in the RH embedding the noise suppression is applied before contributions from the electron and the pile-up are merged. The latter is discussed in more detail in Section 6.4.3. At the moment, this is an inherent systematic effect of the embedding method in CMS, however.

The most important observable for the $H \to \tau^+\tau^-$ analysis is the reconstructed di-tau mass, since it provides the best separation between the Z boson and the Higgs boson. Figure 6.10 shows the distribution for the two final states at the top. Other than the modeling of the peak area, also the high mass tail is important since this is where a signal from the Higgs boson would show up. The plot on the lower left shows the calorimeter-based $E_{\mathrm{T}}^{\mathrm{miss}}$ in the $\tau_e + \tau_{\mathrm{had}}$ final state. This observable can only be modeled with the RH embedding. The lower right plot shows the efficiency of the event selection as a function of the number of reconstructed primary vertices in the $\tau_e + \tau_{\mathrm{had}}$ channel. The event selection efficiency is defined as the ratio of the number of events passing the event selection over the number of generated events within the detector acceptance. It can be seen that the PF embedding is more efficient than both the RH embedding and the non-embedded Monte Carlo. This is due to the fact that contributions from pile-up interactions do not affect a large part of the physics object reconstruction as they do in reality, since the event content merging happens only very late in the reconstruction chain. There is also a trend visible showing that for a large number of reconstructed vertices, corresponding to a large number of pile-up interactions, the discrepancy in selection efficiency is getting bigger. This plot is a good indication that the PF embedding, while still providing an adequate modeling of the $Z \to \tau^+\tau^-$ background, is at its limit under the current conditions. Therefore, for the 13/14 TeV run of the LHC with even higher pile-up expected, switching to the RH embedding will be crucial for the Higgs analysis in the $\tau^+\tau^-$ channel.

Again, more validation plots for the tau embedding, including observables on generator level, can be found in the appendix in Section C.2.

6.4 Systematic Studies

The embedding method is subject to various additional systematic effects and uncertainties that are not present with plain Monte Carlo simulation or other background estimation methods. Individual effects are studied in the following sections, and their impact on important distributions is quantified. For this purpose, only the RH embedding technique is used.

For some effects, possible solutions are proposed or studied. For the CMS $H \to \tau^+\tau^-$ analysis, however, none of these was implemented and the embedded samples were used as described in the previous section.

6.4.1 Muon Radiation

Both muons used for the embedding can undergo final state radiation (FSR), and emit an additional photon. Since the tau lepton is much heavier than the muon, it does not radiate a photon as often. Therefore, in embedded events, there is more final state radiation than in regular $\mathrm{Z}/\gamma^* \to \tau^+\tau^-$ events, originating from the radiation of the initial muons. This leads to two systematic effects:

- The photon carries away some energy that is not recovered when only reconstructing the di-muon system. This leads to a bias in the transverse momentum and invariant mass distributions. Figure 6.11 shows the invariant mass of the two leptons on generator level for $\mathrm{Z}/\gamma^* \to \mu^+\mu^-$ and $\mathrm{Z}/\gamma^* \to \tau^+\tau^-$, on the left hand side before final state radiation and on the right hand side after final state radiation.

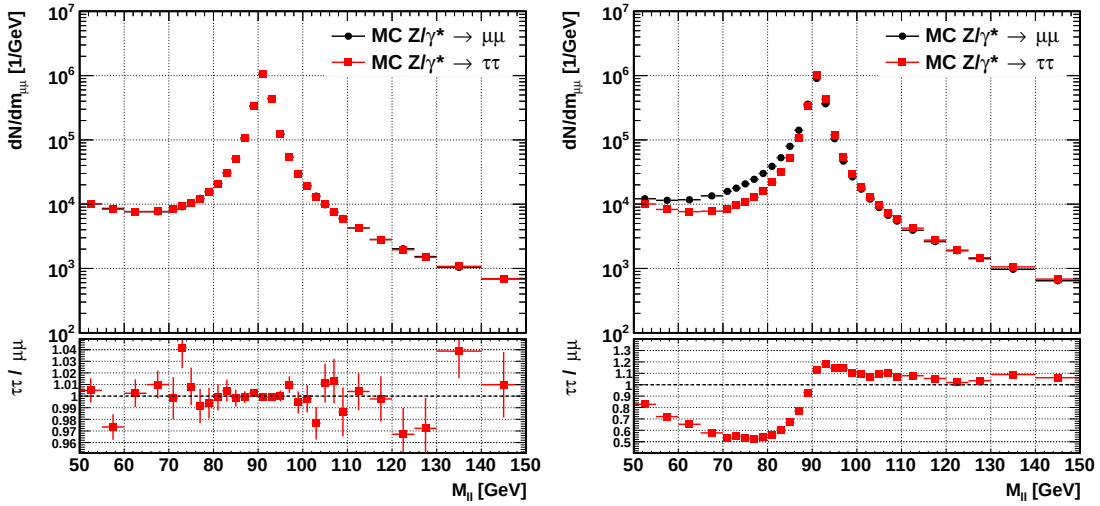

Figure 6.11: Invariant di-lepton mass on generator level in $Z/\gamma^* \to \mu^+\mu^-$ and $Z/\gamma^* \to \tau^+\tau^-$. The left hand side shows the distribution before final state radiation and the right hand side shows it after final state radiation. It can be seen how the muons are much more affected by FSR than the tau leptons.

- The emitted photons are often collinear to the muons. While this can be a handle to reconstruct the photons from FSR, it also poses an additional contribution to the lepton isolation variable. This can lead to a different efficiency for passing the isolation requirement in the $H \to \tau^+\tau^-$ analysis.

In order to mitigate the effect of the muon radiation, an attempt can be made to reconstruct the FSR photons attributed to the muon. Within the scope of the $H \to ZZ$ analysis in CMS, a tool has been developed to reconstruct FSR photons [31]. To be accepted as FSR, a reconstructed particle-flow photon must either have a transverse momentum $p_{\mathrm{T}}^{\gamma} > 2\,\mathrm{GeV}$ and be within $\Delta R < 0.07$ of the muon, or it must satisfy $p_{\mathrm{T}}^{\gamma} > 4\,\mathrm{GeV}$, be within $\Delta R < 0.5$ of the muon, and be isolated. The photon is considered isolated when the relative p_{T} sum of charged hadrons, neutral hadrons and other photons within a cone of $\Delta R < 0.3$ of the photon is below 1.0. In the $H \to ZZ$ analysis, this algorithm has been found to have an efficiency of $\approx 50\%$ to find FSR photons, and a purity of $\approx 80\%$.

In the following, the effect from muon radiation is studied on reconstruction level in $\tau_\mu + \tau_{\mathrm{had}}$ events. Two observables are chosen that are potentially sensitive to muon FSR: the SVfit di-tau mass and the relative photon isolation around the muon. Other than the default RH embedding, two more categories of embedded events are tested. The first category contains only those embedded events, for which there is no photon originating from the muon on generator level in the event. This corresponds to the Monte Carlo truth, and the FSR effect is completely removed. In the second category, those events are removed in which a FSR photon as described above has been reconstructed.

Figure 6.12 shows the two observables for the four chosen categories. The SVfit di-tau mass can be seen on the left hand side. Not much is left from the significant bias seen on generator level in Figure 6.11. This is due to the smearing that the decay of the tau leptons introduces. However, a small trend toward lower masses can be observed in the standard embedded sample (red) which is cured when omitting the events with muon radiation on

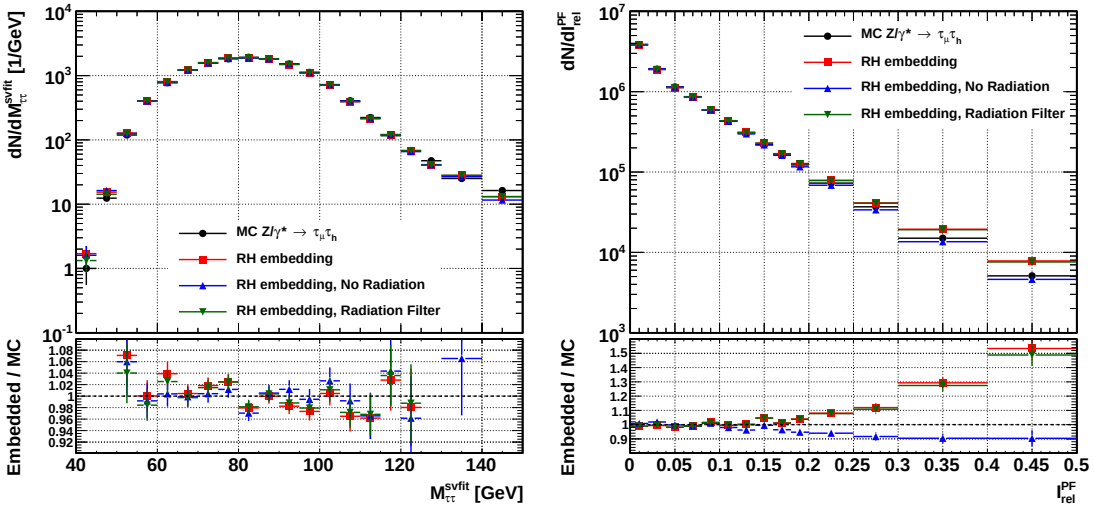

Figure 6.12: Effect of muon FSR on reconstruction level, in $\tau_\mu + \tau_{\text{had}}$ events. The left hand plot shows the SVfit di-tau mass, and the right hand plot shows the p_{T} sum of photons in a cone around the muon, divided by the muon p_{T}.

generator level (blue). The muon radiation filter on reconstruction level (green) improves the situation slightly, but does not cure the trend. The photon isolation variable depicted on the right hand side shows a good agreement between all samples up to ≈ 0.15. The reason for this is that only few events radiate high-energetic photons that cause a bin migration in this plot. The steeply falling spectrum causes the effect to be mostly visible in the tail. In the analysis, a cut on the combined isolation value is applied, which is a combination of the photon isolation with the charged and neutral hadron isolation. Therefore, only those events with a low photon isolation value survive the selection, where the radiation effect is negligible.

Overall, the effect of muon FSR is very small on reconstruction level. It can be taken into account by introducing a systematic uncertainty on the scale of the reconstructed di-tau mass and is in the order of 1%. The muon radiation filter improves the situation slightly, but does not cure the effect. However, it has not been optimized for the use case of tau embedding, and it is possible that with a different set of parameters it could do a better job.

6.4.2 Spin Correlations

The tau spin correlation effects are validated by reproducing the distributions from Figure 6 in Ref. [92]. For this purpose, $Z \to \tau^+\tau^-$ events are studied on generator level where both tau leptons decay to a charged pion or kaon, and a neutrino. In a first step, the embedding procedure is applied where the momenta of the generated muons are used to define the momenta of the embedded tau leptons. Furthermore, events in which one of the two muons has radiated a photon are skipped. This step allows to validate the spin effects without any other systematic effects coming from the original muons.

Figure 6.13 shows the z_s variable, which is constructed from the energy fractions of the visible decay products of both tau leptons, and the visible mass of the two tau leptons. The black points show the MADGRAPH $Z/\gamma^* \to \tau^+\tau^-$ Monte Carlo sample, where the tau

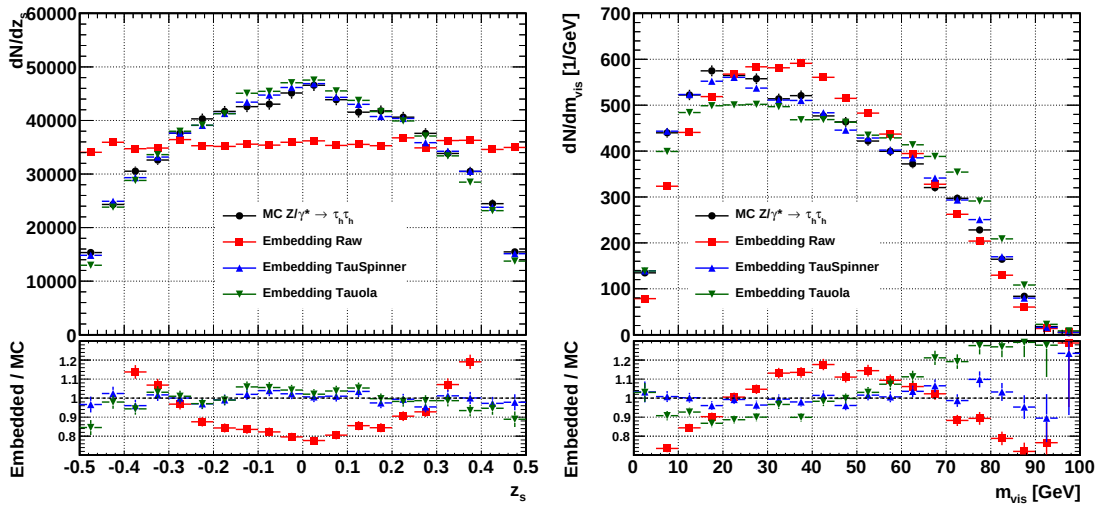

Figure 6.13: The standard CMS $Z/\gamma^* \to \tau^+\tau^-$ Monte Carlo simulation is compared to various configurations of embedded samples produced from $Z/\gamma^* \to \mu^+\mu^-$ events on generator level. Left: The z_s variable as described in Ref. [92]. The variable is sensitive to the spin correlation of the two tau leptons. Right: The visible mass of the tau decay products.

decays have been simulated with TAUOLA. The other data points correspond to embedded samples produced from MADGRAPH $Z/\gamma^* \to \mu^+\mu^-$ simulation. The red points have the polarization in TAUOLA switched off. The blue points also correspond to tau polarization turned off in TAUOLA, but event weights computed with TAUSPINNER applied. This is the default configuration also used in all other validation plots in this chapter. Finally, the green points correspond to TAUOLA with polarization turned on, and no spin weights applied.

In both distributions, it can be clearly seen that no spin correlations are visible when neither TAUOLA nor TAUSPINNER are used for taking spin effects into account. While TAUOLA does compute the spin effects, it is missing information from the initial quark state to produce the correct result. The disagreement can be seen especially in the visible mass distribution. With the weights computed with TAUSPINNER, however, the original distributions are reproduced correctly.

In the second step, reconstruction effects are taken into account. The reconstructed muon momenta are replaced by tau leptons and simple acceptance cuts are applied on the tau leptons on generator level, to account for the geometric acceptance for selecting di-muon events. Furthermore, events in which a muon has radiated a photon are not filtered out anymore. This corresponds to an embedded event sample as one would generate it from the detector data. The following acceptance cuts are performed:

- $p_\mathrm{T} > 20\,\mathrm{GeV}$ for the leading tau lepton.
- $p_\mathrm{T} > 10\,\mathrm{GeV}$ for the subleading tau lepton.
- $|\eta| < 2.1$ for both tau leptons.
- $M_{\tau\tau} > 50\,\mathrm{GeV}$, where $M_{\tau\tau}$ is the invariant di-tau mass.

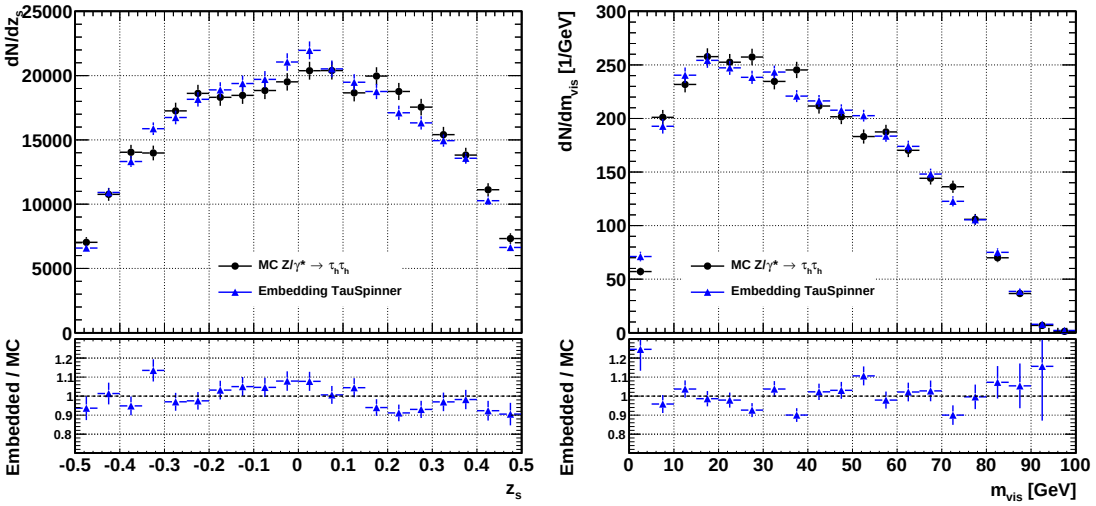

Figure 6.14: The two spin-sensitive variables on generator level, when considering reconstruction effects for the two muons of the embedded sample. The detector effects introduce a small bias to the tau spin correlations.

Also, the weight to correct for the di-muon selection efficiency is applied to the embedded sample.

Figure 6.14 shows the two observables on generator level after the steps discussed above have been performed. On the embedded sample, a very small bias can be seen in the z_s distribution which has been found to primarily originate from the acceptance cuts. Section C.3 in the appendix presents the breakdown of the three individual steps in detail. However, it can well be expected that after the tau reconstruction it will get mostly smeared out and become insignificant.

6.4.3 Calorimeter Noise

When the detector simulation is run for the di-tau event in the RH embedding procedure, the simulation of electronic and thermal noise in the calorimeters is turned off, in order to avoid applying the noise twice. While this is in principle the correct thing to do, together with the zero-suppressed readout of the calorimeter, it leads to another effect: since the noise suppression and the merging of the calorimeter hits in the embedded sample do not commute, the noise suppression cuts erase cells with noise before the signal from the di-tau event is embedded.

Figure 6.15 illustrates this effect. A 3×3 array of calorimeter cells is shown, and each point could be interpreted as a registered photon in the photomultiplier of the calorimeter. Red dots correspond to instrumental noise, while blue dots represent a signal, for example from an electron. The left hand side of the figure shows what happens in the data when there is an electron in the detector: The noise and the signal are merged, and they cannot be told apart in the calorimeter. The noise suppression is applied by imposing a cut on the total number of photons in a cell. What is left is the electron signal, with a little noise on top. What happens in the embedding case, however, is that first the noise suppression is applied, basically removing all the noise in the data. The electron signal is then merged, and what is left is the electron signal without noise on top.

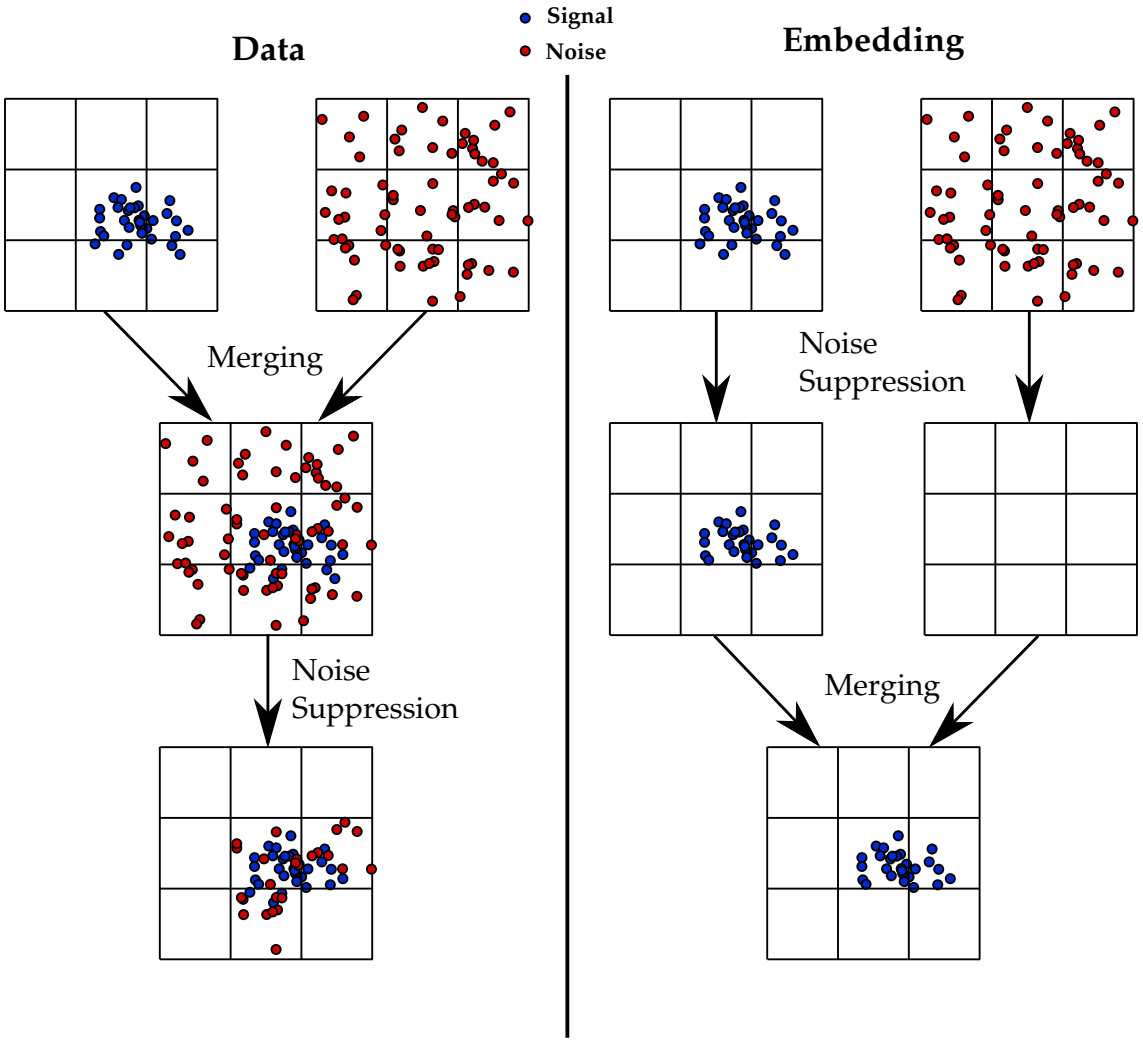

Figure 6.15: Sketch of the effect of calorimeter noise and noise suppression cuts when merging a simulated signal into an otherwise empty region. The left hand side represents the ideal scenario, however due to the zero-suppressed readout of the calorimeter, the right hand side is what happens in embedded events when the calorimeter noise simulation is disabled. A signal in the calorimeter tends to be cleaner than in the data.

The zero suppression of the calorimeter is performed on a very low level, to reduce the data rate from the detector. It is therefore not trivial to switch it off for the embedding procedure. In total, the effect leads to a cleaner signal in the embedded event. Mitigating the effect by enabling the simulation of calorimeter noise in the embedding procedure improves the modeling of regions with a real signal, however it also leads to too much noise in other detector regions.

Figure 6.16 shows on the left the electron identification efficiency as a function of the pseudorapidity of the electron. At high pseudorapidities, the identification efficiency is better in embedded events than in the pure Monte Carlo simulation. The electromagnetic calorimeter has a lower signal to noise ratio in the endcaps, which explains why a dis-

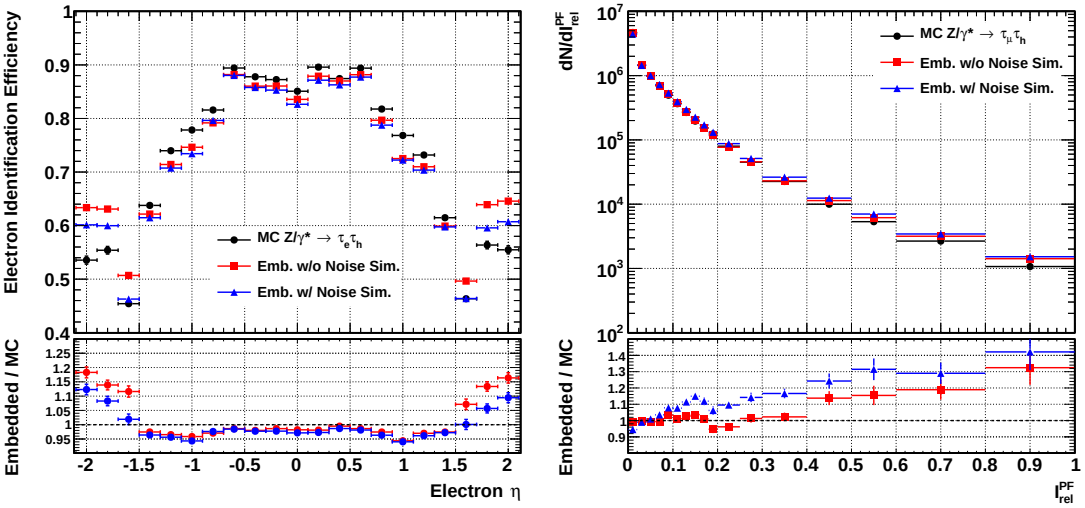

Figure 6.16: Comparison between the simulation of calorimeter noise enabled and disabled in the embedding procedure. Left: Electron identification efficiency as a function of η, in $\tau_e + \tau_{\text{had}}$ events. Right: Combined particle-flow based relative isolation of the muon, in $\tau_\mu + \tau_{\text{had}}$ events.

crepancy is seen especially in the forward region. However, the situation improves when enabling the simulation of calorimeter noise. The plot on the right hand side shows the combined relative particle-flow isolation for the muon in the $\tau_\mu + \tau_{\text{had}}$ final state. It can be seen that enabling the calorimeter noise leads to higher isolation values. This is evidence that there is too much noise close to the muon, which contributes to the particle reconstruction in the isolation cone.

More studies are needed, however, to better understand the effect. Monte Carlo simulation can help to understand whether disabling the zero-suppression for the calorimeter would mitigate the problem entirely, especially in terms of the electron identification efficiency. It might then be possible to design a special di-muon trigger which enables full calorimeter readout, or at least in the region of the triggered muons. Another approach would be to enable the noise simulation in the embedded sample only in a small region around the generator-level objects that are embedded. This could improve the situation for the electron identification while leaving the rest of the event with the normal amount of noise.

6.4.4 Muon Momentum Vector Transformation

The selection of the two muons which are subsequently replaced by tau leptons introduces a bias into the modeling of the $Z/\gamma^* \to \tau^+\tau^-$ background. There are three primary sources for this bias:

- The selection criteria for the two muons give preference to muons in certain regions of the detector. For example, there are fewer events with the muons in the non-instrumented regions of the muon system, and due to the isolation requirement the muons tend to be in "cleaner" regions of the detector with fewer contributions from pile-up interactions or other activity around them. This effect can be seen most

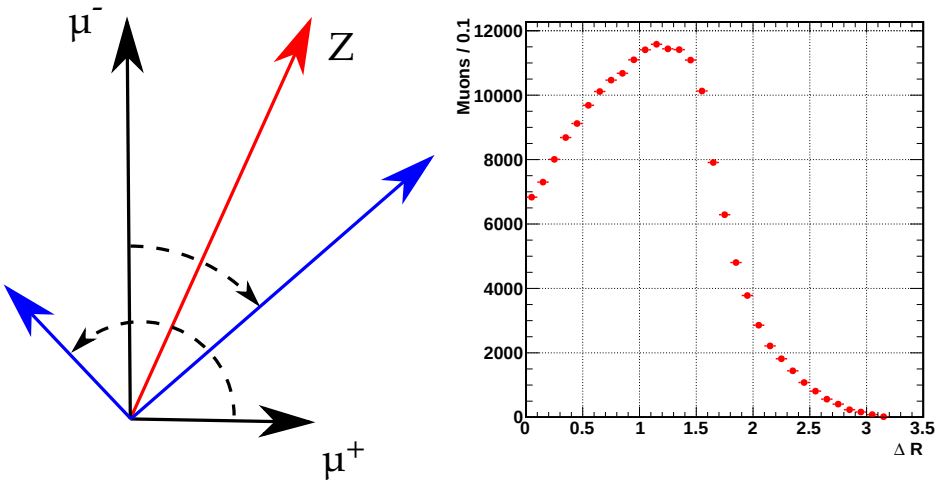

Figure 6.17: Left: Sketch of how the muon four-momenta are transformed by mirroring on the Z boson axis in the r-ϕ plane. Right: Distribution of geometrical distance in η-ϕ space of the transformed muons and the closer of the two original muons.

prominently in the muon p_{T} spectrum when replacing muons with muons, as depicted in the top left plot in Figure 6.8.

- A bias can come from the removal of the signatures of the original muons. While the matching of inner tracks and hits in the muon system can typically be removed without ambiguities, the calorimeter energy can only be subtracted on a statistical basis. On an event-by-event level, a muon can deposit more or less energy than the the mean energy that is subtracted.

- Radiated photons are typically geometrically close to the muon, and contribute to the isolation value. This effect can be seen on the right hand side in Figure 6.12.

In order to reduce this bias, the four-vectors of the two muons can be transformed before they are replaced by generator-level leptons. However, the decay kinematics of the Z boson must be invariant under such a transformation, so that the decay is still correctly modeled. The simplest case would be an arbitrary rotation around the Z boson momentum axis in the Z boson rest frame, however this is not invariant due to the polarization of the Z boson. Instead, a mirror operation on the plane defined by the Z boson momentum axis and the incoming proton beam can be used. The Z decay has been confirmed to be invariant under such a transformation with leading-order Monte Carlo simulation using PYTHIA and MADGRAPH. More information about the confirmation of the invariance can be found in Section C.4 of the appendix.

In the laboratory frame, the mirror operation corresponds to swapping the sides of the two muons with respect to the Z boson axis in the r-ϕ plane. This transformation is depicted on the left hand side in Figure 6.17. The two original muons are shown in black, and taken together they define the momentum axis of the Z boson. The blue vectors correspond to the muon momenta after the transformation when they have been mirrored on the other side of the Z boson axis. Both the transverse momentum and the pseudorapidity in the laboratory frame are not altered by the transformation, but only

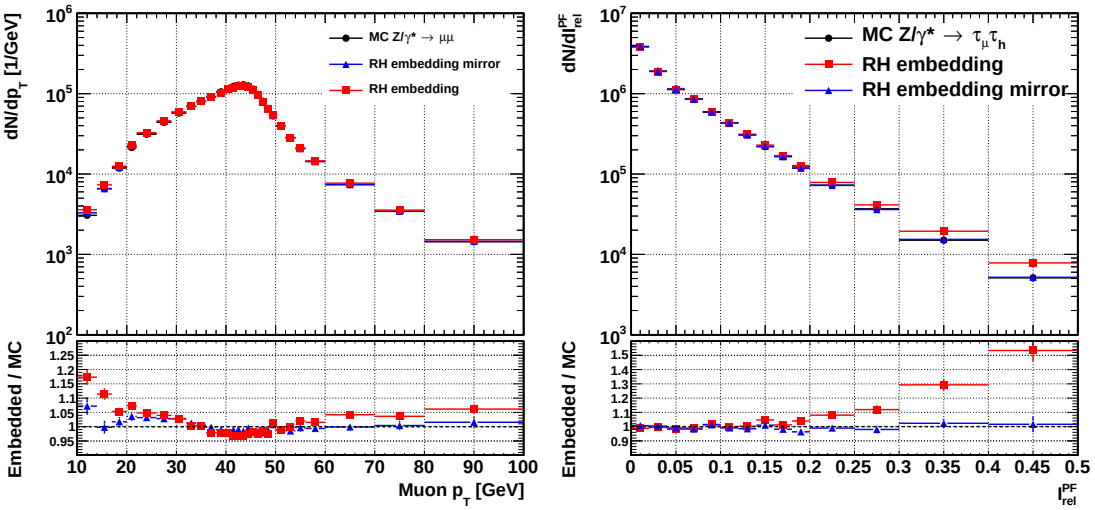

Figure 6.18: Comparison of the embedding technique with and without the mirror transformation applied to the muon four-vectors. Left: Muon transverse momentum. Right: Muon relative photon isolation.

the azimuthal angle changes. This property makes sure that there are no events migrating from outside the detector acceptance to within the acceptance, or vice versa.

Figure 6.18 shows the improvements when the mirror transformation is applied. The RH embedded sample created from $Z/\gamma^* \to \mu^+\mu^-$ simulation is compared to the direct Monte Carlo simulation. The plot on the left hand side shows the muon transverse momentum when replacing muons with muons. The bias at low p_T which comes from the isolation requirement on the original muons is much reduced. The right hand plot shows the relative photon isolation in the $\tau_\mu + \tau_{had}$ final state when replacing muons with tau leptons. The tail coming from FSR photons spoiling the muon isolation is completely cured.

6.5 Conclusions and Future Work

In most $H \to \tau^+\tau^-$ analyses, the $Z/\gamma^* \to \tau^+\tau^-$ process is the major background. The embedding method presented in this chapter is the only way to partly estimate it from the data themselves. This helps to significantly reduce the systematic uncertainties, for example, coming from the modeling of the underlying event and pile-up interactions, and also additional jets in the event from both soft and hard processes. The embedding on particle level is currently used in CMS, and has been successfully applied to data [35]. While the particle-based embedding provides an accurate modeling of the $Z/\gamma^* \to \tau^+\tau^-$ process, there is evidence that, under the conditions as they were in the first run of the LHC, it is at its limit. This can be seen in Figure 6.10, which shows the selection efficiency of events in the $\tau_e + \tau_{had}$ channel as a function of the number of primary vertices. The efficiency of selecting embedded events is too high compared to the reference sample, especially at a high number of particle interactions.

In the work presented here, a new embedding technique based on rechit level is introduced. Its goal is to provide solutions to some of the shortcomings of the particle-based method, especially in the light of run 2 of the LHC with much harsher pile-up conditions.

In addition, the rechit-based embedding is also able to model purely calorimeter-based observables. The method has been extensively validated with Monte Carlo simulation, and it shows performance as good as or better than the particle-based embedding in most observables. Additional systematic effects, including modeling of spin correlations, muon final-state radiation, and zero suppression of detector noise have been studied and their effect on important observables has been quantified. A transformation of the muon four-vectors before the particle replacement improves the bias coming from the reconstruction and identification of the primary muons.

More work is required, however, to improve the method and further reduce systematic effects that it introduces, and to expand the scope of the method. The mitigation of effects introduced by the calorimeter noise and muon FSR can certainly be improved. It might also be possible to extend the merging of reconstructed hits to the inner tracker. Instead of merging reconstructed tracks, one can instead merge the aligned hits, and re-run the track finding. This would allow to better take into account possible ambiguities during the track finding, without requiring a model of the misaligned tracker in the simulation.

Even though, in CMS, the embedding procedure is only used to obtain the efficiency for the category selections and the shape of the di-tau mass distribution, it can also be used to obtain the normalization. The decay rate of the Z boson to muons and tau leptons is almost the same, apart from the slightly smaller phase space for tau leptons. However, this method requires a very good understanding of all selection efficiencies for the di-muon event, and also possible differences in efficiencies in the selection in the Higgs analysis. Early work with 2010 data and simulation has proven the feasibility of the method [18, 183]. Another possibility not discussed here is to model other kinds of backgrounds with the embedding method, such as top-quark pair production. Modeling the $W \to \tau\nu$ process with embedded $W \to \mu\nu$ events can be useful for charged Higgs searches [189].

7 $H \to \tau^+\tau^-$ Produced in Association with a W Boson

At the LHC, the dominant production processes for the Higgs boson are the gluon-gluon fusion and vector boson fusion processes. Figure 7.1 shows on the left hand side the Higgs boson production cross section in various production modes as a function of the mass of the Higgs boson at $\sqrt{s} = 8$ TeV [27]. It can be seen that the associated production with a W or Z boson is roughly a factor 2 smaller than the vector boson fusion at $m_H = 125$ GeV. The associated production process is also known as "Higgsstrahlung", since, at leading order, it is produced by a Higgs boson radiated from a vector boson. The right hand side of Figure 7.1 shows the Feynman graph for this process, where V can be either a W or a Z boson.

However, while the production cross section is lower, additional leptons from the W boson and Z boson decays bring an advantage to the analysis. Many backgrounds are greatly reduced with respect to the Higgs boson production in gluon-gluon fusion when requiring additional leptons in the event. Especially the dominant $\mathrm{Z}/\gamma^* \to \tau^+\tau^-$ process is no longer irreducible. Additionally, triggering of candidate events is simplified. Since the additional leptons are from a W or Z decay, they tend to be harder than light leptons from tau lepton decays, allowing higher trigger thresholds.

In CMS, the $H \to \tau^+\tau^-$ analysis in the associated production channels is split into three categories: ZH, WH semileptonic and WH hadronic. The ZH analyses cover the $\tau_e + \tau_\mu$, $\tau_e + \tau_{\mathrm{had}}$, $\tau_\mu + \tau_{\mathrm{had}}$ and $\tau_{\mathrm{had}} + \tau_{\mathrm{had}}$ decays of the tau leptons from the Higgs decay, while the WH semileptonic covers the $\tau_e + \tau_{\mathrm{had}}$ and $\tau_\mu + \tau_{\mathrm{had}}$ decays. The dedicated WH hadronic analysis covers the $\tau_{\mathrm{had}} + \tau_{\mathrm{had}}$ decay. All three analyses are combined to obtain the final result on associated production in the $H \to \tau^+\tau^-$ channel, and the associated production result is combined with the other production modes [35]. With the current dataset, the associated production channels alone are not sensitive to the Standard Model Higgs boson, and therefore exclusion limits are set on the production cross section times branching ratio of the Higgs boson.

In the remainder of this chapter, the analysis in the WH hadronic channel is presented in detail. In Section 7.1, the selection of candidate events is described. The major background to the search in this channel is coming from reducible backgrounds with misidentified τ_{had}. The estimation of this background from the data is described in Section 7.2. The irreducible backgrounds are estimated with Monte Carlo simulation. A multivariate discriminant is used to reduce the background further, based on its different event topology. This procedure is described in Section 7.3. The background modeling after the multivariate selection is verified in a control region and in the simulation, shown in Section 7.4. Systematic uncertainties have only a minor effect on the analysis since it is statistically limited. They are discussed in Section 7.5. The statistical interpretation is performed in bins of the di-tau visible mass after the multivariate selection, presented in Section 7.6. This procedure allows for a good discrimination against both the reducible backgrounds as well as the irreducible WZ production. The SVfit algorithm for reconstructing the di-tau mass cannot be

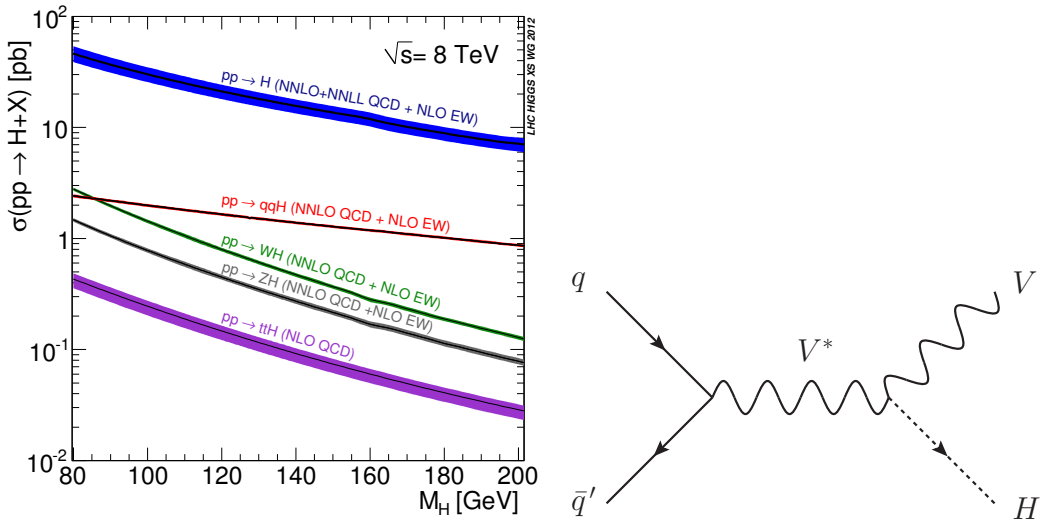

Figure 7.1: Left: Higgs boson production cross section for the different production mechanisms, as a function of the Higgs boson mass. The WH associated production is shown in green. Right: Feynman diagram for Higgs boson production in association with a W or Z boson. From [27].

used, since the neutrino from the leptonic W boson decay gives an additional contribution to the missing transverse energy $E_{\mathrm{T}}^{\mathrm{miss}}$.

7.1 Event Selection

The analysis uses both the 4.9 fb^{-1} of CMS data taken at $\sqrt{s} = 7\,\mathrm{TeV}$ and the 19.7 fb^{-1} of data at $\sqrt{s} = 8\,\mathrm{TeV}$. The WH hadronic analysis is divided in two subcategories, depending on the decay of the W boson. The channels with the W boson decaying into an electron and a muon are analyzed. This leads to a final state in which there are two hadronically decaying tau leptons and one light lepton (electron or muon).

7.1.1 Trigger Selection

The first step in the event selection is the choice of the trigger that is required to have accepted the event in question. A split strategy is chosen here: for the $\mu + \tau_{\mathrm{had}} + \tau_{\mathrm{had}}$ channel, a single-muon trigger with a transverse momentum threshold of 24 GeV is used. This is possible because the muon from the W boson decay is typically hard. A study has shown that a cross-trigger that requires one muon plus one hadronically decaying tau lepton on trigger level would not increase the acceptance, due to the additional inefficiency caused by the tau leg of the trigger. In addition, not requiring tau decays on the trigger level means that there is no bias on the tau leptons coming from the trigger, especially for the tau isolation. This allows to define a control region later which makes use of τ_{had} candidates where the isolation requirement is inverted.

In the $e + \tau_{\mathrm{had}} + \tau_{\mathrm{had}}$ channel, a cross-trigger is chosen which requires an electron and a hadronically decaying tau lepton on trigger level. The transverse momentum thresholds of the trigger vary with the instantaneous luminosity in the different data taking periods.

Table 7.1: Summary of the trigger configuration used for the WH hadronic analysis.

Thresholds	Run Range	Integrated Luminosity [fb^{-1}]
Muon Channel at $\sqrt{s} = 7$ TeV		
μ: 24 GeV	160431 - 180252	4.955 ± 0.109
Muon Channel at $\sqrt{s} = 8$ TeV		
μ: 24 GeV	190456 - 208686	19.711 ± 0.512
Electron Channel at $\sqrt{s} = 7$ TeV		
e: 15 GeV, τ_{had}: 15 GeV	160431 - 163869	0.216 ± 0.005
e: 15 GeV, τ_{had}: 20 GeV	165088 - 173198	1.779 ± 0.039
e: 20 GeV, τ_{had}: 20 GeV	173236 - 180252	2.980 ± 0.066
Electron Channel at $\sqrt{s} = 8$ TeV		
e: 20 GeV, τ_{had}: 20 GeV	190456 - 193621	0.870 ± 0.023
e: 22 GeV, τ_{had}: 20 GeV	193834 - 208686	18.835 ± 0.490

Especially for the 7 TeV dataset, no single-electron trigger is available with a reasonably low transverse momentum threshold.

Table 7.1 gives an overview of the triggers used for the analysis.

7.1.2 Offline Object Selection

After an event was accepted by one of the triggers, a primary vertex and three leptons need to be reconstructed and well-identified within the event to be considered a signal candidate.

From all reconstructed vertices in the event, the vertex with the highest $\sum p_{\text{T}}^2$ is chosen, where the sum is over all tracks associated to the vertex and p_{T} is the transverse momentum of such a track. This vertex is taken to be the hard-scatter vertex, while other vertices are interpreted as vertices from pile-up interactions. The chosen vertex is then required to have more than 7 degrees of freedom during the vertex fit, to ensure a good measurement of the vertex properties. It is required to be within 24 cm from the nominal interaction point in Z direction and within 2 mm in the transverse plane.

In the next step, a light lepton (electron or muon) candidate is identified, assumed to originate from the W boson decay. The detailed identification requirements are specified in Sections 3.4.5 or 3.4.6, respectively. The kinematic acceptance cuts are chosen to be $p_{\text{T}} > 24$ GeV and $|\eta| < 2.1$ in both cases. These acceptance cuts are mostly constrained by the available triggers. The longitudinal and transverse impact parameters of the lepton tracks with respect to the chosen primary vertex must be less than 0.2 cm and 0.045 cm, respectively. The relative particle-flow based isolation for both leptons must be lower than 0.1, except for the electron in the barrel region ($|\eta| < 1.479$), where the cut is loosened to 0.15.

The two hadronically decaying tau leptons are reconstructed with the HPS algorithm as described in Section 3.4.7. One of the two τ_{had} candidates is then required to have $p_{\text{T}} > 25$ GeV and the other one $p_{\text{T}} > 20$ GeV. The acceptance in pseudorapidity is defined by $|\eta| < 2.3$. The longitudinal impact parameter of the highest-p_{T} track of the τ_{had} candidates must be less than 0.2 cm with respect to the primary vertex, to ensure that all leptons were created in the same interaction.

It is required that the two tau leptons have opposite charge with respect to another, since both are expected to originate from the Higgs boson which is a neutral particle. This requirement implies that there is one tau lepton which has opposite charge with respect to the selected light lepton (OS τ_{had}), and one which has the same charge (SS τ_{had}). Different working points of the tau identification algorithm are used for the two taus, depending on their charge with respect to the light lepton, since different background processes lead to different misidentification rates for the two tau leptons. For example, in $Z/\gamma^* \to e^+e^-$ events, one of the electrons could be misidentified as a tau lepton, and an additional jet in the event as the other tau lepton. In this case, the OS τ_{had} candidate is much more likely to be an electron in reality than the SS τ_{had}.

The tau identification working points have been optimized by scanning 3×3 working points at a time, performing the full analysis. The working point which yields the best expected exclusion limit in the case that the Standard Model Higgs boson does not exist has been chosen. First, the isolation working point has been optimized in this way, and then the electron rejection. The procedure has been performed independently for the electron channel and the muon channel. The muon rejection working point has been found not to alter the expected sensitivity of the analysis. The results of the optimization procedure is presented in Section D.4 of the appendix.

For the SS τ_{had} candidate, the medium working point for the isolation was chosen, and the loose working point for the electron rejection. For the OS τ_{had} candidate, the choice depends on the channel. In the muon channel, the loose isolation working point and the loose electron rejection are chosen. In the electron channel, however, this tau lepton is more likely to be misidentified, and therefore the medium isolation working point and the tight electron rejection have been shown to give the best results. In addition, in the electron channel, the OS τ_{had} candidate is required to be matched to the τ_{had} candidate found by the final stage of the trigger. The matching is performed geometrically in η-ϕ space by requiring $\Delta R < 0.3$. This effectively means that, while in principle both τ_{had} candidates can cause the trigger to accept the event, only events in which the OS τ_{had} was found by the trigger are chosen in the analysis. This requirement is needed for the background estimation which is discussed in Section 7.2.

7.1.3 Topological Selection

After the selection of the candidate leptons, additional requirements are imposed on the event topology. The purpose of these is twofold. First, they are an additional measure to further suppress backgrounds. Second, they sort out events which are used in other CMS analyses, avoiding overlap between channels when combining the WH hadronic analysis with other Higgs searches within CMS. The following additional requirements on the event are made:

- No Lepton Overlap: All three signal leptons are required to be separated by $\Delta R > 0.5$ in η-ϕ space. This makes sure that two candidate objects are not actually the same physical particle (for example, an electron which is also reconstructed as a τ_{had}). The cut also ensures that one lepton cannot be found within the isolation cone of another.

- B-Jet Veto: Events with a b-tagged jet are rejected. This requirement reduces the background from top-quark pair production. A jet is considered b-tagged if it has

Table 7.2: Cross sections and branching ratios of the signal process and the most important backgrounds processes. For comparison, also the cross section for Higgs boson production in gluon-gluon fusion is given. An ℓ in the branching ratio column corresponds to a decay to either an electron or a muon, but not both. Numbers taken from [27, 190, 191, 119]

Process	$\sigma_{7\,\mathrm{TeV}}$ [pb]	$\sigma_{8\,\mathrm{TeV}}$ [pb]	Branching Ratio	
WZ ($M_{\ell\ell} > 12\,\mathrm{GeV}$)	26.6	32.4	0.36 %	($W \to \ell\nu$, $Z \to \tau\tau$)
ZZ ($M_{\ell\ell} > 12\,\mathrm{GeV}$)	10.4	12.8	0.11 %	($Z \to \ell\ell$, $Z \to \tau\tau$)
Z/γ^* + jets ($M_{\ell\ell} > 50\,\mathrm{GeV}$)	28222	32442	3.36 %	($Z \to \tau\tau$)
W + jets	96648	111905	10.8 %	($W \to \ell\nu$)
$t\bar{t}$ + jets	165	225	1.16 %	($t \to \ell\nu b$, $\bar{t} \to \tau\nu\bar{b}$)
H (125)	15.1	19.27	6.32 %	($H \to \tau\tau$)
WH (125)	0.579	0.705	0.68 %	($W \to \ell\nu$, $H \to \tau\tau$)

$p_{\mathrm{T}} > 20\,\mathrm{GeV}$, $|\eta| < 2.4$ and passes the tight working point of the combined secondary vertex tagger described in Section 3.4.3. This working point corresponds to a misidentification rate of light parton jets as b-jets of $\approx 0.1\,\%$ with a b-jet identification efficiency of $\approx 50\,\%$ [113].

- Extra Muon Veto: Events are vetoed if there exists an additional muon in the event with $p_{\mathrm{T}} > 10\,\mathrm{GeV}$, $|\eta| < 2.1$ and relative particle-flow based isolation less than 0.3. The extra muon must also be reconstructed both as a global muon and a particle-flow muon, and its longitudinal impact parameter with respect to the primary vertex must be less than 0.2 cm. This avoids overlap between this analysis and the search for the ZH process, which has four leptons in the final state.

- Extra Electron Veto: In a similar way, events with an extra electron in the event are rejected. Such a veto electron must satisfy $p_{\mathrm{T}} > 10\,\mathrm{GeV}$, $|\eta| < 2.5$, and the electron identification as described in Section 3.4.5. The relative particle-flow based isolation must be less than 0.3 and the longitudinal impact parameter with respect to the primary vertex must be less than 0.2 cm.

- $\mathbf{M_T}$ cut: The transverse mass M_{T} between the light lepton and $E_{\mathrm{T}}^{\mathrm{miss}}$ as defined in Equation 5.1 is required to be greater than 30 GeV. While this reduces the signal acceptance by $\approx 10\,\%$, the cut also reduces the $\mathrm{Z}/\gamma^* \to \ell^+\ell^-$ backgrounds significantly, and it avoids overlap with the inclusive $\tau_e + \tau_{\mathrm{had}}$ and $\tau_\mu + \tau_{\mathrm{had}}$ channels. These channels do not have a veto for additional τ_{had} candidates in the event.

7.1.4 Combinatorial Selection

After the full selection, it can happen that there is more than one ℓ-τ_{had}-τ_{had} triplet in the event which fulfills all selection criteria. In this case, the triplet with the highest product of the transverse momenta of the three candidates is chosen. This choice is justified by the fact that misidentified τ_{had} candidates have a more steeply falling p_{T} spectrum.

7.2 Background Estimation

There are both reducible and irreducible backgrounds in the analysis. The irreducible background consists of WZ and ZZ pair production, where 3 leptons in the final state can occur if both bosons decay leptonically. The contribution from ZZ production is very small, due to the lower production cross section compared to WZ, the lower branching ratio of the Z boson to decay to leptons compared to the W boson, and the rejection of events with extra electrons or muons. The background from WZ production is sizeble, however it is important to note that this background is not as dominant as the $\mathrm{Z}/\gamma^* \to \tau^+\tau^-$ process is for the inclusive Higgs search, due to the production cross sections of the involved processes. Table 7.2 lists the cross sections and branching ratios for the signal process and the most important backgrounds.

The WZ and ZZ backgrounds are taken from a Monte Carlo simulation with MADGRAPH. In principle, it would be possible to use the embedding technique by selecting WZ events with two muons from the Z boson, and replacing them by taus. However, due to the low WZ production cross section, the overall number of available embedded events would be small, and can easily be superseded with Monte Carlo event generation. In addition, in this analysis, the major background is not WZ production but the reducible backgrounds. Therefore, the WZ and ZZ processes are estimated with Monte Carlo simulation.

The reducible backgrounds are estimated directly from the data. This is especially important since many of the reducible backgrounds have very high cross sections but only a very low acceptance rate, so that a very large number of events would need to be simulated in order to estimate the reducible backgrounds from the simulation. Furthermore, the selected events would come from a highly exclusive region in the phase space, and it would require additional studies to confirm that these regions are well modeled. The major backgrounds are those, in which one or both τ_{had} are misidentified quark or gluon jets. The following processes contribute significantly to the reducible background:

- $\mathrm{Z}/\gamma^* \to \tau\tau$ + 1 jet: In this case, one of the tau leptons decays hadronically and the other one leptonically. The additional jet is misidentified as the other τ_{had}.
- $\mathrm{Z}/\gamma^* \to \ell\ell$ + 1 jet: Here, one of the two leptons is misidentified as a τ_{had}. While the rate for muons to be misidentified as τ_{had} is very low, this background is much more significant in the electron channel. The additional jet is again misidentified as the other τ_{had}.
- $t\bar{t}$ pair production: The top quark almost always decays into a W boson and a b-quark. The two W bosons then decay further into a light lepton and tau lepton, which are oppositely charged. The second τ_{had} comes from one of the two b-jets being misidentified.
- $W \to \ell\nu$ + 2 jets: The W boson decays to a light lepton, and two additional jets are misidentified as tau leptons. This background is the dominant contribution in both the electron and the muon channel, due to the high W boson production cross section, and the relatively high rate for a jet to be misidentified as a τ_{had}.
- QCD multijets: In QCD multijets events, again two jets are misidentified as a τ_{had}. The lepton can either come from a heavy quark decay, or from another jet that is misidentified as an electron or a muon.

The various backgrounds can be divided in two groups: one where one of the τ_{had} is real and the other is misidentified (the backgrounds with only one additional jet), and the other group where both τ_{had} are misidentified jets. However, in both groups it is important to note that the SS τ_{had} is always misidentified. For the two-jet backgrounds this is trivial, and for the 1-jet backgrounds there is an opposite-sign pair of real leptons, one of which is the light lepton and the other one is one of the two τ_{had}. Therefore, the other τ_{had}, which is a misidentified jet, has always the same charge as the light lepton. This property is crucial for the background estimation. In principle, there are also other backgrounds, such as $Z/\gamma^* \to \tau\tau$ where both tau leptons decay hadronically and the light lepton is a misidentified jet, or a heavy quark decay. However, due to the low rate of jets being misidentified as light leptons, these backgrounds are so small that they can be neglected. Their contribution is much smaller than the systematic uncertainty on the estimation of the backgrounds with a misidentified τ_{had}.

7.2.1 The Fake Rate Method

The *fake rate method* is used to estimate the contribution of all backgrounds for which the SS τ_{had} is a misidentified jet. As pointed out above, this is the vast majority of backgrounds. The working principle of the method is simple: first, a criterion is chosen which discriminates the signal-enriched region from a background-enriched region. For the case of taus being misidentified jets, such a criterion is the τ_{had} isolation: tau leptons for which there is few activity within the isolation cone are likely to be real taus, while taus with large activity are more likely to be part of a jet. Next, let f be the probability for a quark or gluon jet, which passes all the other τ_{had} identification criteria, to also pass the τ_{had} isolation requirement. If f is known, the number of misidentified τ_{had} in the signal region can be estimated from the number of jets in the region where the τ_{had} isolation is inverted, and is given by

$$N_{\text{sig}} = N_{\text{non-iso}} \times \frac{f}{1-f}\,, \tag{7.1}$$

where N_{sig} is the number of events in the signal region and $N_{\text{non-iso}}$ is the number of events in the region when inverting the isolation requirement on the SS τ_{had}. The rate f is also known as "fake rate" or "misidentification rate".

At this point, it becomes clear why the OS τ_{had} must be matched to the trigger object. Since the trigger imposes an isolation requirement on the τ_{had} leg, it would create a bias in the number $N_{\text{non-iso}}$ otherwise.

7.2.2 The Jet to Tau Misidentification Rate

The remaining ingredient now is the measurement of the fake rate f. This measurement is performed in the data as well. The idea behind the measurement is again simple: events from a well-known process are selected, such as $Z/\gamma^* \to \mu^+\mu^-$. In such events, there are no genuine tau leptons. Therefore, all reconstructed τ_{had} objects are known to be misidentified. It can then be determined from the reconstructed τ_{had} objects that pass all τ_{had} identification criteria. The fake rate is given by the ratio of τ_{had} objects that pass the τ_{had} isolation requirement over all reconstructed τ_{had} objects.

The fake rate is not constant, however, but it depends on many factors, such as the transverse momentum and pseudorapidity of the jet, the total number of jets in the event or

the number of pile-up interactions. The measurement is therefore performed as a function of the p_{T} of the reconstructed τ_{had} and in three different pseudorapidity regions defined by $|\eta| < 0.8$, $0.8 < |\eta| < 1.6$ and $|\eta| > 1.6$. The fake rate is measured in two different regions, one that is dominated by W + jets events, and one that is dominated by Z + jets events. This covers both reducible background categories, with either one or both τ_{had} misidentified. The definition of the two regions are chosen such that they are topologically similar to the corresponding backgrounds, but avoid event overlap with the signal region.

The Z + jets dominated region is defined as follows:

- The event is triggered by a single muon trigger
- Two oppositely charged muons identified as described in Section 3.4.6, with particle-flow based relative isolation lower than 0.1.
- The geometric acceptance is given by $p_{\mathrm{T}} > 20\,\mathrm{GeV}$ for the leading muon and $p_{\mathrm{T}} > 10\,\mathrm{GeV}$ for the subleading muon, and $|\eta| < 2.1$.
- The invariant mass of the two muons must be within 10 GeV of the nominal Z boson mass.

The W + jets dominated region is defined as follows:

- The event is triggered by a single muon trigger (muon channel) or a single electron trigger, or an electron-plus-M_{T} trigger (electron channel).
- One electron or muon is identified as described in Section 3.4.5 or 3.4.6, respectively, with particle-flow based relative isolation lower than 0.1.
- The geometric acceptance is given by $p_{\mathrm{T}} > 24\,\mathrm{GeV}$ and $|\eta| < 2.1(2.5)$ for the muon channel (electron channel).
- The transverse mass M_{T} between the lepton and $E_{\mathrm{T}}^{\mathrm{miss}}$ is higher than 40 GeV.

For both regions, the following common set of additional requirements is imposed:

- The longitudinal impact parameter of the chosen leptons with respect to the primary vertex must be lower than 0.2 cm, and the transverse impact parameter lower than 0.045 cm.
- There is no b-tagged jet in the event, where the "loose" working point of the combined secondary vertex algorithm [113] is deployed. The b-jet is only accepted if it is separated by $\Delta R > 0.3$ from either muon or electron.
- There is no other muon in the event with $p_{\mathrm{T}} > 15\,\mathrm{GeV}$, $|\eta| < 2.1$, and the same impact parameter and isolation cuts as for the signal muons or electrons.
- There is no other electron with $p_{\mathrm{T}} > 15\,\mathrm{GeV}$ and $|\eta| < 2.5$, and the same impact parameter and isolation cuts as for the signal muons or electrons.

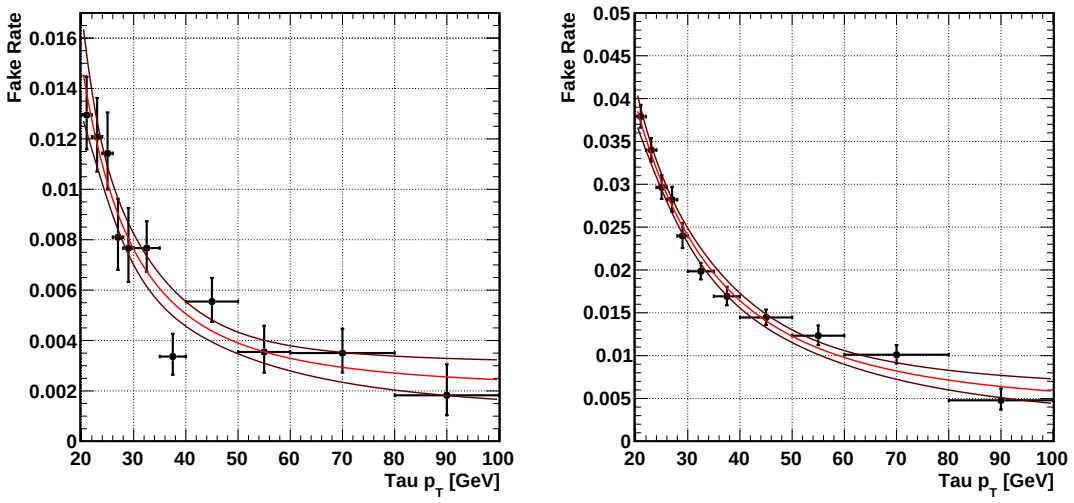

Figure 7.2: Measurement of the jet $\rightarrow \tau_{\text{had}}$ misidentification rate in the W + jets-dominated region (left) and the Z + jets-dominated region (right). The plots are made for jets in the central detector region ($|\eta| < 0.8$) and shows the misidentification rate as a function of the p_{T} of the τ_{had} candidate. The dark red lines are the $\pm 1\sigma$ errors on the fit result.

In the case of the Z + jets dominated region, all reconstructed τ_{had} candidates are evaluated for the fake rate measurement. In the W + jets case, however, only events with at least two τ_{had} candidates are considered. Since the fake rate method is used with the SS τ_{had}, the two τ_{had} candidates in the W + jets region are also required to have the same charge as the lepton from the W boson. This is necessary since the fake rate is different for jets that have an opposite charge compared to the W boson than jets with the same charge as the W boson, as illustrated in Appendix D.1.2. Note that this also implies that the two τ_{had} candidates have the same charge, which avoids events overlapping with the signal region of the analysis. In case there are more than two τ_{had} candidates in the event, all pairs of τ_{had} candidates are evaluated which fulfill all requirements, and each τ_{had} candidate that is part of such a pair is used exactly once for the fake rate measurement. This way, double counting of τ_{had} candidates is avoided while at the same time considering all τ_{had} candidates in the event.

The measurement is performed separately in the 7 TeV and 8 TeV datasets due to the different pile-up conditions. Two example measurements for the muon channel are depicted in Figure 7.2, showing the fake rate in the 8 TeV dataset as a function of the τ_{had} candidate p_{T} in the central region $|\eta| < 0.8$. The error bars indicate the 68 % Clopper-Pearson confidence interval [192]. In order to interpolate between the data points and compensate for statistical fluctuations, the tail of a Landau function is fitted to the data. The fit has three free parameters: The most probably value of the Landau, the width of the Landau, and an additive constant. Fitting an exponential model instead gives very similar results. The two dark red curves in the figure give the 1σ error which is obtained by propagating the errors on the fit parameters to the fitted function, according to

$$\Delta f = J^{\text{T}} V J \,, \tag{7.2}$$

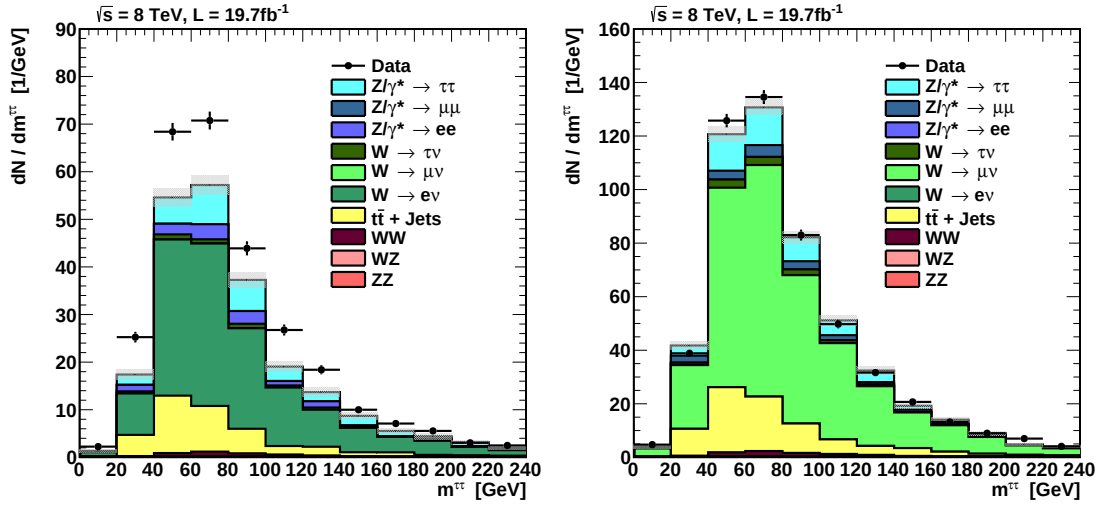

Figure 7.3: Data-to-simulation comparison of the di-tau visible mass when the isolation requirement of the τ_{had} candidate with the same charge as the light lepton is inverted. The contribution from QCD multijets processes are not shown. The plot is used to obtain the ratio of the number of events with one misidentified τ_{had} candidate to the number of events with two misidentified τ_{had} candidates. The left hand side shows the electron channel and the right hand side the muon channel.

where V is the covariance matrix of the fit parameters and J is the gradient vector (Jacobian) of the fitted function.

All fake rates for the two channels, the two data taking periods, the two measurement regions and the three pseudorapidity bins can be found in Appendix D.1.1. Additional systematic studies are presented in Section D.1.2 of the appendix. A closure test with Monte Carlo events is shown in Section 7.4.2.

In principle it can happen that events with genuine tau leptons contaminate the two measurement regions, in which case the measured fake rate would be higher than it really is. Processes that lead to such a contamination can be WZ and ZZ di-boson production, or $t\bar{t}Z$ production. However, the cross section of these processes is very small compared to the inclusive W and Z boson production, so that their contribution is negligible. This has been confirmed by comparing the measured fake rates with (uncontaminated) Monte Carlo simulation of W and Z production, where very good agreement has been found.

7.2.3 Fake Rate Weights

One important feature that can be seen in Figure 7.2 is that the two fakes rates are very different from each other: the fake rate measured in the W + jets-dominated region is significantly lower than the one measured in the Z + jets-dominated region. The reason for this is that the fake rate is different for quark-induced and gluon-induced jets. The quark-to-gluon ratio changes as a function of the number of jets in the event, and, in the case of the W + jets-dominated region, also depends on the relative charge of the jet with respect to the lepton from the W boson decay.

Table 7.3: Ratio of events with both τ_{had} candidates misidentified (W-like) and only the SS τ_{had} candidate misidentified (Z-like) in the region where the isolation of the SS τ_{had} candidate is inverted. The quoted uncertainties are only of statistical nature.

Dataset	W-like events	Z-like events	Ratio W-like [%]
Muon Channel 7 TeV	1233 ± 44	543 ± 13	69.4 ± 1.0
Muon Channel 8 TeV	7320 ± 125	3316 ± 70	68.9 ± 0.7
Electron Channel 7 TeV	759 ± 35	348 ± 11	68.6 ± 1.3
Electron Channel 8 TeV	4000 ± 93	1769 ± 55	69.3 ± 1.0

For a single event in the isolation-inverted region, it is not possible to decide whether it is a W-like (both τ_{had} misidentified) or a Z-like (one τ_{had} misidentified) event, and therefore it is not clear which of the two fake rates to apply. In order to solve this issue, the Monte Carlo simulation is used to determine the contribution of the two types of events in the region where the SS τ_{had} isolation is inverted. Figure 7.3 shows the visible di-tau mass distribution for the 8 TeV dataset for both the electron and the muon channel. The contribution from QCD multijet processes is not shown, due to insufficient statistical precision of the available Monte Carlo samples. Instead, the difference between the simulation and the data is taken to be from QCD multijets. Table 7.3 contains the number of events in the two categories of backgrounds, where QCD and W + jets are considered W-like, and Z/γ^* + jets, $t\bar{t}$ and WW di-boson production are considered Z-like. The ratio between the two is very similar in all four cases. Therefore, for the background estimation, Equation 7.1 is modified to read

$$N_{\text{sig}} = N_{\text{non-iso}} \times \left(r_W \times \frac{f_W}{1 - f_W} + (1 - r_W) \times \frac{f_Z}{1 - f_Z} \right) , \tag{7.3}$$

where f_W and f_Z are the fake rates measured in the W + jets-dominated and Z + jets-dominated regions, respectively, and r_W is taken to be 0.7 due to the expected ratio of W-like events and Z-like events in the whole sample.

In events with more than two τ_{had} candidates it can happen that the event is used both as a signal candidate and for the background estimation. In order to avoid this, the primary ℓ-τ_{had}-τ_{had} triplet in an event is already chosen as described in Section 7.1.4 before the isolation cuts on the two τ_{had} candidates are made, but after all other selection criteria. Then, depending on the isolation of the two τ_{had} candidates in the primary triplet, the event is either used as a signal candidate, for the background estimation, or, if the OS τ_{had} candidate does not pass the isolation, as an event in a control region which is discussed further in Section 7.4.

7.3 Reducible Background Suppression

Figure 7.4 shows the visible di-tau mass distribution in the 8 TeV dataset of both channels. In both cases, the reducible background dominates. It consists mostly of W + jets events, which can be seen in Figure 7.3. In order to reduce this background further, a multivariate discriminator is constructed.

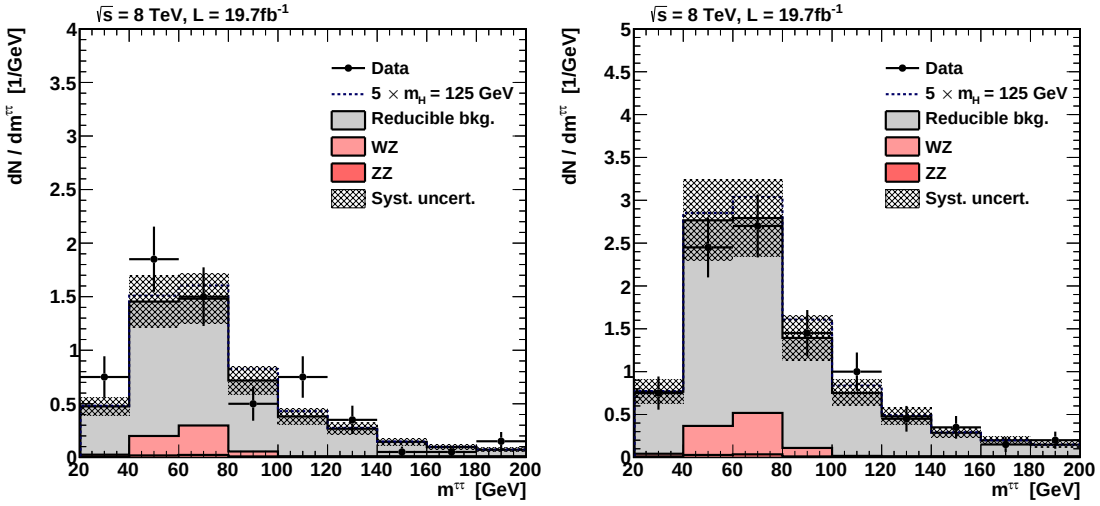

Figure 7.4: Distribution of the visible di-tau mass in the electron channel (left) and the muon channel (right) in the 8 TeV dataset. The reducible background dominates in both cases. The systematic uncertainty shown is correlated between bins and discussed later in Section 7.5.

7.3.1 BDT Training

The following topological variables have been found to provide discrimination power between signal events and the reducible background:

- Leading hadronic tau p_{T}
- Subleading hadronic tau p_{T}
- The magnitude of the $E_{\mathrm{T}}^{\mathrm{miss}}$ vector
- The separation in η-ϕ-space of the two hadronic tau candidates
- The vectorial sum of the transverse momenta of the two hadronic tau candidates, divided by their scalar sum.

Most of these variables exploit the fact that the p_{T} spectrum for quark and gluon jets is falling more steeply than that of genuine tau leptons from a Higgs decay, and also that the angular distribution is different: in the case of the Higgs boson, the two hadronic tau candidates recoil against the W boson and are therefore boosted, while, in W + jets events, they are more likely to be back-to-back. The visible di-tau mass is not used as a variable, since this variable discriminates mostly between WZ and WH events, and because the statistical interpretation of the result after the multivariate selection is performed in bins of the visible di-tau mass.

The five variables are combined into a single discriminator with a BDT as introduced in Section 3.2.1. The signal events for the training are taken from a Monte Carlo simulation of WH production where the W boson decays to a lepton on generator level and the Higgs boson decays to two hadronically decaying tau leptons. Since the event topologies are the same in the muon and electron channel, and since the BDT is mostly sensitive to

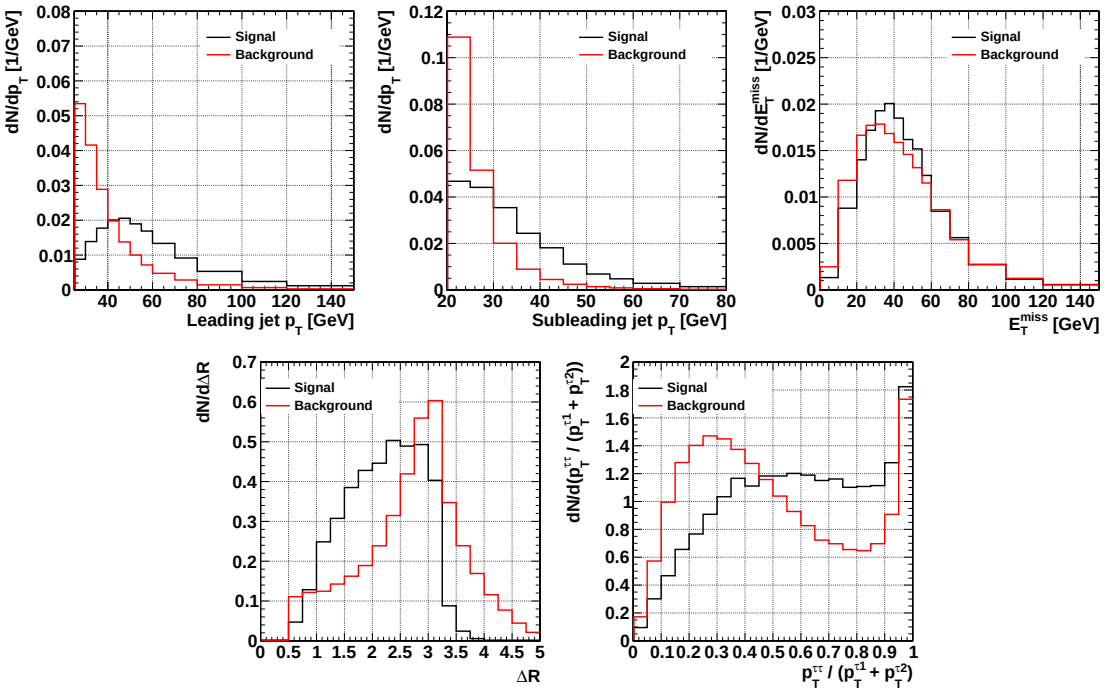

Figure 7.5: Distributions of the five input variables for the topological BDT, in the training sample. In all variables, significant discrimination power between signal and background can be observed. The signal and background distributions are normalized to unit area.

the topology, events from both channels are used to train the BDT. Also, signal samples with Higgs masses between 110 GeV and 145 GeV are combined, since the BDT input variables are not very sensitive to the mass of the Higgs boson. The training sample for background events consists of data events in which the isolation cut was inverted for both of the τ_{had} candidates. In this way, events with one lepton and two jets are selected. This region is dominated by W + jets and QCD multijets events. The background training events are weighted with a factor $f_W/(1-f_W)$ for both τ_{had} candidates, where f_W is the misidentification rate measured in the W + jets enriched region. This procedure makes sure that the τ_{had} p_{T} spectra of the training events corresponds to the events entering the signal region where both τ_{had} candidates pass the isolation cut. Only events in the muon channel are used in the background training sample, since in the electron channel the trigger already imposes an isolation requirement on one of the two τ_{had} candidates, and would therefore introduce a bias.

Figure 7.5 shows the distributions of the five input variables in the full training sample. The events from the 7 TeV and 8 TeV data taking periods are combined, since the discrimination power of the BDT output does not improve significantly when splitting them up. In this way, the total number of training events can be enhanced, improving the statistical precision of the training. In total, 5257 signal events and 165 950 background events are used in the training.

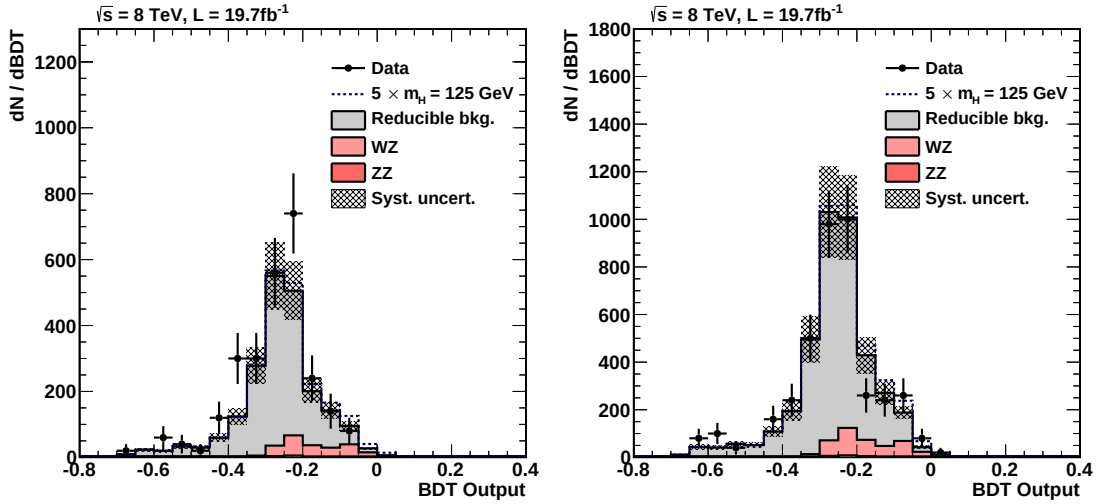

Figure 7.6: Distributions of the BDT output for reducible background reduction in the electron channel (left) and muon channel (right) in the 8 TeV dataset.

7.3.2 BDT Output

Figure 7.6 shows the output of the BDT in the 8 TeV dataset. Signal-like events peak on the right hand side of the spectrum, and background-like events on the left hand side. The BDT cannot discriminate between the irreducible WZ background and the signal, however this discrimination is performed later using the visible di-tau mass.

In order to find the best BDT value to cut on, a scan was performed on the BDT selection ranging from output values of -0.050 to -0.200, in steps of 0.005. The figure of merit is the expected exclusion limit for a Higgs mass of 125 GeV, taking into account all systematic uncertainties as discussed in Sections 7.5 and 7.6. The analysis was performed separately for the muon channel and the electron channel, but combining the 7 TeV and the 8 TeV dataset. In both cases, the best exclusion limit was obtained when cutting at BDT > -0.170. With this working point, around 60 % of the signal is retained while only 13 % of reducible background events pass the cut. Section D.4 in the appendix presents the full optimization procedure.

7.3.3 Reducible Background Composition after the BDT Selection

The BDT is trained to suppress mostly W-like events, due to the choice of the background training sample. Therefore, after the cut on the BDT discriminant, the ratio of W-like events to Z-like events changes.

Table 7.4 shows the updated ratio of W-like to Z-like events. Due to the changed ratio, the estimation of the reducible background needs to be adjusted, since the misidentification rates are different for W-like and Z-like events. The factor r_W in Equation 7.3 is therefore set to 0.6 when estimating the reducible background in a region where the cut on the BDT discriminant has been applied.

Table 7.4: Ratio of events with both τ_{had} candidates misidentified (W-like) and only the SS τ_{had} candidate misidentified (Z-like) in the region where the isolation of the SS τ_{had} candidate is inverted. Only events which have passed the cut on the BDT discriminant are shown. The quoted errors are only of statistical nature.

Dataset	W-like events	Z-like events	Ratio W-like [%]
Muon Channel 7 TeV	163 ± 19	147 ± 6	52.6 ± 3.4
Muon Channel 8 TeV	1279 ± 60	938 ± 37	57.7 ± 1.9
Electron Channel 7 TeV	116 ± 15	86 ± 5	57.5 ± 3.9
Electron Channel 8 TeV	649 ± 44	480 ± 29	57.5 ± 2.9

7.4 Validation and Control Regions

In order to verify that both the di-tau invariant mass and the BDT discriminant are well modeled with the fake rate method, cross-checks are performed in simulated events and in a control region in the data.

7.4.1 W + Jets Control Region

When inverting the isolation cut of the OS τ_{had} candidate, the event selection is dominated by W-like events with both τ_{had} candidates misidentified. In addition, the region is free of signal events as well as events from the irreducible backgrounds. This selection serves as an ideal control region for the reducible background estimation. The estimation of the event yield in the control region is performed in exactly the same way as in the signal region: the isolation of the SS τ_{had} candidate is also inverted and the events are weighted with the misidentification rate. Since this region is completely dominated by W-like events, only the fake rate measured in the W + jets enriched region is used for the event weights. In the electron channel, the single electron trigger is used to avoid the bias that comes from the trigger isolation of the $e + \tau_{\text{had}}$ cross-trigger. Even though in parts of the 7 TeV dataset the threshold on the electron transverse momentum is very high (up to 80 GeV), there are enough events available to populate the region.

Figure 7.7 shows the visible di-tau mass distribution in the W + jets control region for the 8 TeV dataset. Reasonable agreement is observed.

7.4.2 Monte Carlo Closure Test

Unfortunately, the available simulated samples do not have enough events to test the background estimation method by simply applying it to the Monte Carlo sample. However, the W + jets control region defined in the previous section features a large number of events and can be applied to a MADGRAPH W + jets simulated sample. In this case, also the misidentification rate is measured not in the data but in the same sample of Monte Carlo events. This makes this verification effectively a closure test of the fake rate method in the simulation, since there cannot be any contamination from other physics processes.

Figure 7.8 shows the visible di-tau mass in the 8 TeV MADGRAPH W + jets sample for both channels. The good agreement in both distributions verifies that the fake rate method models the background correctly.

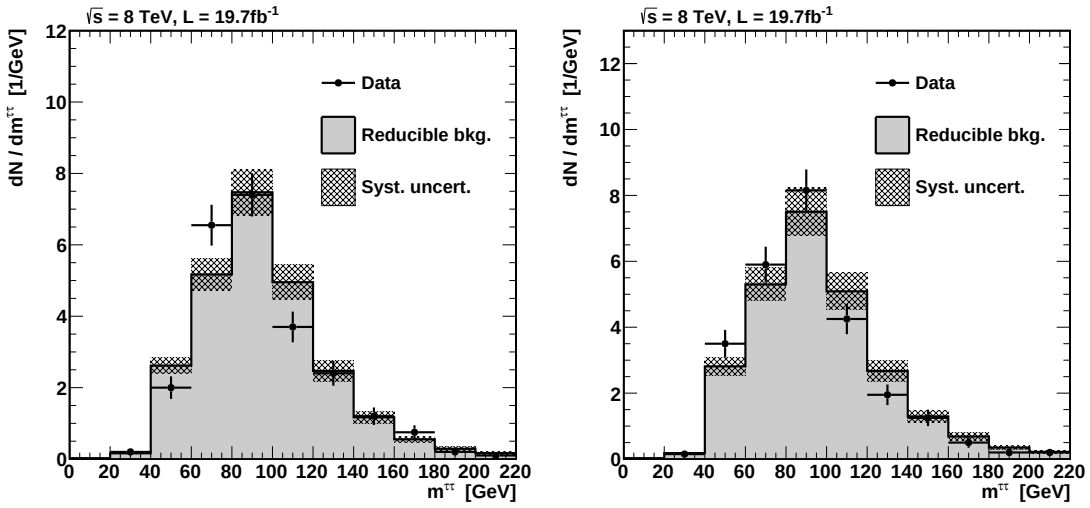

Figure 7.7: Visible di-tau mass distribution in the W + jets control region in the 8 TeV dataset. Both the electron channel (left) and the muon channel (right) are shown.

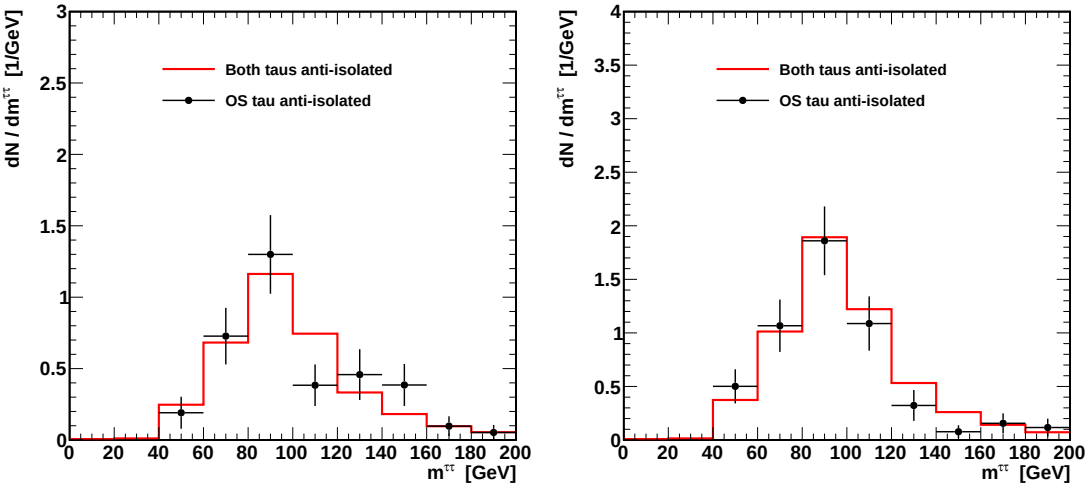

Figure 7.8: Visible di-tau mass distribution in the W + jets control region in MADGRAPH W + jets simulation. Both the electron channel (left) and the muon channel (right) are shown.

7.5 Systematic Uncertainties

There are several sources of systematic uncertainties to the analysis. Most systematic uncertainties are related to the Monte Carlo simulation of signal and irreducible background events. Other sources are related to the estimation of the reducible background. Each source of systematic uncertainty corresponds to a nuisance parameter in the final fit and affects the overall event yield or the di-tau mass shape of either the simulated samples or the reducible background estimation. When not noted differently, nuisances are uncorrelated between the 7 TeV and 8 TeV datasets, as well as the electron and muon channel. The following systematic uncertainties are considered:

- Theory: Theoretical uncertainties on the Higgs and diboson production cross sections of 4 % are assigned due to the uncertainty on the parton distribution function (PDF) and another 4 % due to the QCD renormalization scale as outlined in [193]. These uncertainties are fully correlated between both channels and data taking periods.

- Integrated luminosity: The uncertainty on the integrated luminosity is 2.2 % for the 7 TeV dataset [194] and 2.6 % for the 8 TeV dataset [50]. The luminosity uncertainties are correlated between the two channels but not between the two data taking periods.

- Lepton trigger and identification efficiency: Scale factors are applied to correct for differences in the trigger efficiency and the identification efficiency of electrons and muons between the data and the simulation [177, 176, 174, 175]. The uncertainties on the scale factors are used as systematic uncertainties. This leads to a 2 % systematic uncertainty for electrons and muons, and 6 % for each τ_{had}. Conservatively, the uncertainties for both τ_{had} candidates are taken as correlated, for a total uncertainty of 12 % due to the τ_{had} identification efficiency. In the electron channel, another 2 % is added linearly due to the τ_{had} component of the cross-trigger.

- Tau energy scale: A 3 % uncertainty is assumed on the τ_{had} energy scale [121]. This uncertainty is propagated into the visible di-tau mass distribution by running the whole analysis on the same sample with changed τ_{had} energy. This procedure models both the shape and the yield uncertainty due to the uncertainty on the τ_{had} energy scale. The yield difference is around $\approx 8\,\%$.

- $E_{\text{T}}^{\text{miss}}$ scale: The same procedure as for the τ_{had} energy scale is applied for a 10 % uncertainty on the $E_{\text{T}}^{\text{miss}}$ scale [195]. This leads to a difference of $\approx 3\,\%$ in the event yield. The $E_{\text{T}}^{\text{miss}}$ scale uncertainty is correlated between the electron and the muon channel.

- Lepton Vetoes: The systematic uncertainty on the additional lepton vetoes comes from propagating the uncertainties on the single lepton identification efficiencies. The number of events failing the veto cut is varied up and down according to the systematic uncertainty, and the relative difference in the number of passing events is taken as a systematic uncertainty. An uncertainty of 1 % is found for the muon channel and 4 % for the electron channel.

- Reducible Background: The uncertainty on the reducible background consists of three parts: First, the statistical uncertainty on the fitted fake rate function is

Table 7.5: Summary of all systematic uncertainties considered in the analysis, including the contribution to which they apply.

Systematic Source	Uncertainty	Contribution
Theory	5.7 %	Simulation
Luminosity	2.2 % (7 TeV)	Simulation
	2.6 % (8 TeV)	
Electron and Muon ID	2 %	Simulation
τ_{had} ID	12 % (Muon channel)	Simulation
	14 % (Electron channel)	
τ_{had} energy scale	$\approx$ 8 % (shape-altering)	Simulation
$E_{\text{T}}^{\text{miss}}$ energy scale	$\approx$ 3 % (shape-altering)	Simulation
Lepton veto	1 % (Muon channel)	Simulation
	4 % (Electron channel)	
Fake rate function fit	5 % - 20 % (shape altering)	Reducible bkg.
Reducible bkg. composition	10 %	Reducible bkg.
Reducible bkg. normalization	10 %	Reducible bkg.

propagated to the shape and yield of the reducible background estimation. This contribution is of the order of 5-10% for the 8 TeV data and 10-20% for the 7 TeV data. Second, the ratio of W-like events to Z-like events in the reducible background is allowed to vary by 10 %, based on the numbers in Table 7.4 and to allow for possible mis-modeling by the simulation. Since the di-tau mass shape of the reducible background does not vary when using the misidentification rate obtained in the W + jets enriched region or the Z + jets enriched region (see Section D.1.2 in the Appendix), this leads to a 10 % normalization uncertainty on the reducible background yield. Third, a 10 % uncertainty on the normalization of the reducible background is imposed, to account for any differences seen in the W + jets control region. This accounts also for non-statistical uncertainties in the measurement of the misidentification rate.

Table 7.5 summarizes all systematic uncertainties. Overall, the largest uncertainties come from the reducible background estimation and the 12 % uncertainty due to the τ_{had} identification efficiency for all simulated samples. However, since the overall event yield in this analysis is very low, the result is not very sensitive to systematic uncertainties. The statistical uncertainty from the limited data sample dominates the significance of the analysis with the LHC run 1 data. For the same reason, the systematic uncertainties are not constrained or pulled significantly by the fit. In particular, there is no pull above half a standard deviation.

7.6 Results

The di-tau visible mass in the two channels and the two data taking periods is shown in Figure 7.9. It can be seen that the reducible background has been considerably reduced with the multivariate selection and the signal to background ratio is enhanced. The signal peaks at a higher mass than the irreducible backgrounds in the di-tau mass spectrum. Table 7.6 shows the corresponding event yields, integrated over the mass spectrum. The

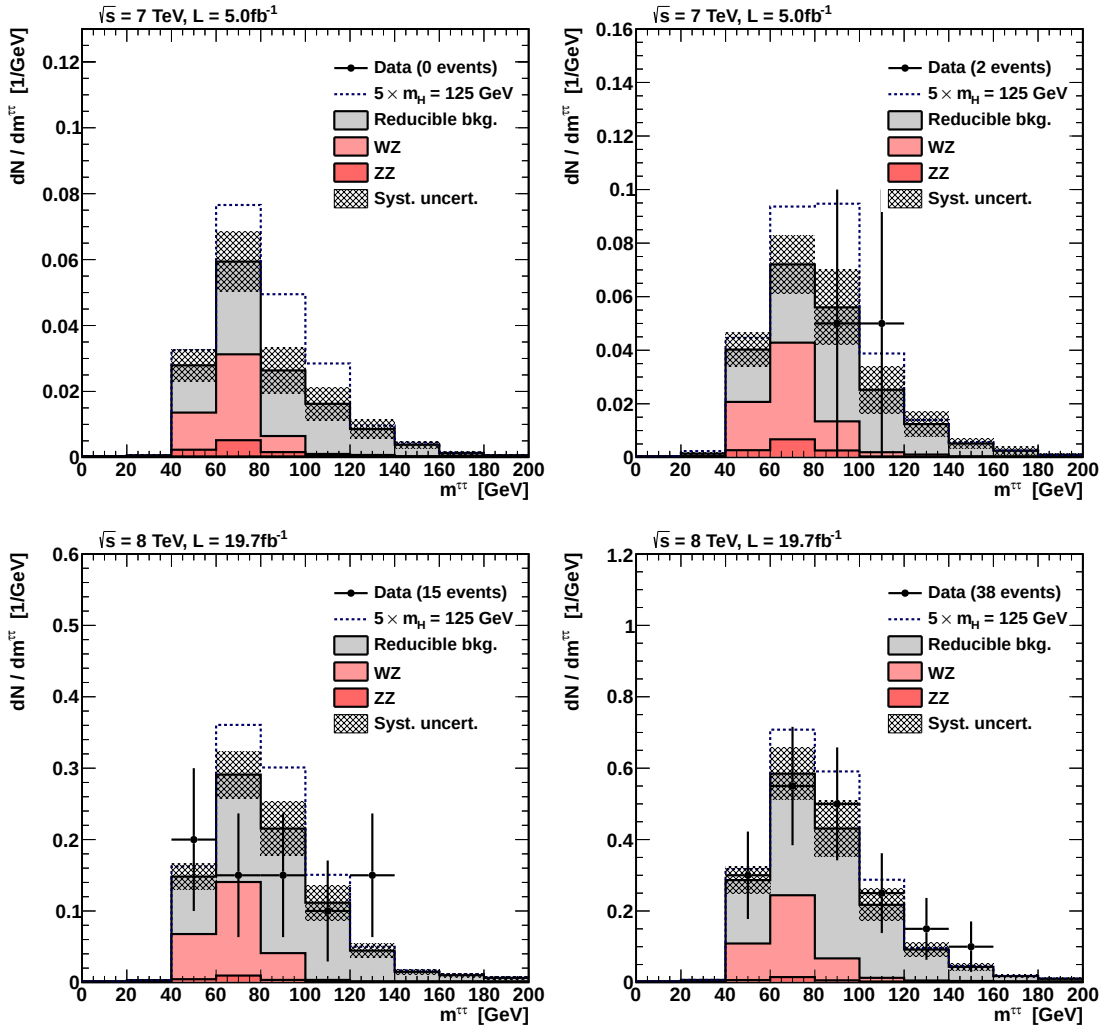

Figure 7.9: Visible di-tau mass distributions in the electron channel (left) and the muon channel (right), for both the 7 TeV dataset (top) and the 8 TeV dataset (bottom).

signal contains also contributions from the ZH and $t\bar{t}H$ processes, contributing about 15 % of the total signal. The signal from the Higgs decaying into two W bosons is negligible. It can be observed that the acceptance in the electron channel is lower than in the muon channel. There are many little factors contributing to this, such as a lower identification efficiency when compared to muons, the additional trigger inefficiency for the τ_{had} component of the trigger, and tighter τ_{had} identification working points.

Since no significant excess of the data over the background expectation is observed, the results are interpreted in terms of exclusion limits for the Standard Model Higgs boson. For this procedure, the CL_S method as described in Section 5.6.1 is used, with a binned maximum-likelihood fit in the di-tau visible mass. Figure 7.10 shows the exclusion limit as a function of the hypothesized Higgs boson mass, which ranges from 90 GeV to 145 GeV in 5 GeV steps. The exclusion limit is shown separately for the electron channel and the muon channel. The red line corresponds to the expected exclusion limit in case no

Table 7.6: Integrated event yields for the WH hadronic analysis. The quoted uncertainties are only statistical uncertainties.

	Electron Channel		Muon Channel	
Process	7 TeV	8 TeV	7 TeV	8 TeV
Signal ($m_{\mathrm{H}} = 125\,\mathrm{GeV}$)	0.24 ± 0.01	0.88 ± 0.06	0.36 ± 0.04	1.57 ± 0.08
Reducible Background	1.84 ± 0.15	11.90 ± 0.40	2.73 ± 0.18	25.23 ± 0.60
WZ	0.89 ± 0.07	4.78 ± 0.23	1.37 ± 0.08	8.33 ± 0.32
ZZ	0.20 ± 0.01	0.37 ± 0.02	0.25 ± 0.01	0.60 ± 0.02
Background	2.92 ± 0.16	17.02 ± 0.47	4.33 ± 0.19	34.10 ± 0.68
Data	0	15	2	38

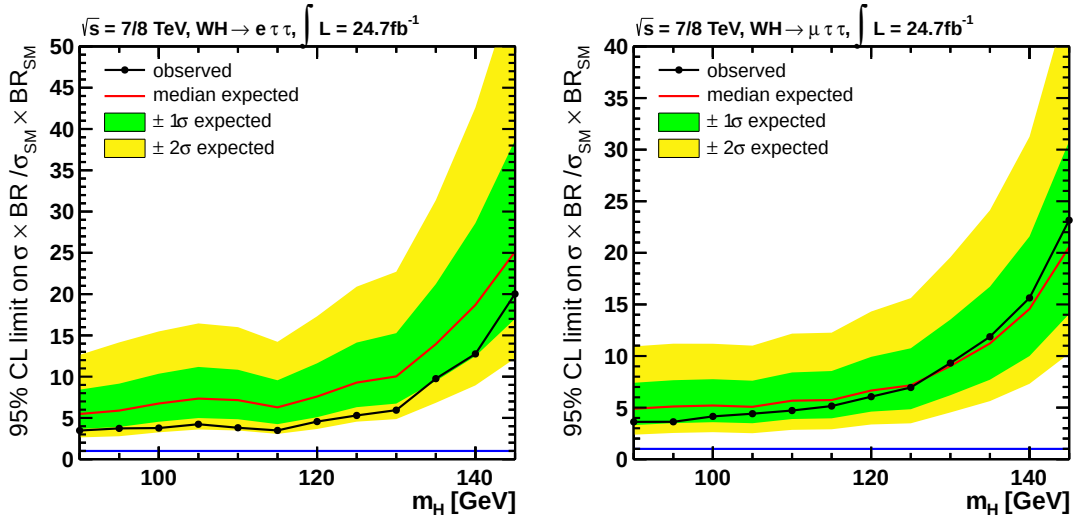

Figure 7.10: Exclusion limit as a function of the hypothesized Higgs boson mass, in the region from 90 GeV to 145 GeV. The left hand plot shows the result for the electron channel and the right hand plot for the muon channel.

Standard Model Higgs boson exists, and the green and yellow bands are the expected deviation around the expected limit. The solid black line is the value observed in the data. The y axis represents the 95 % C.L. limit on the signal strength modifier μ, so that values below 1 correspond to an exclusion of the Standard Model Higgs boson at 95 % C.L.

Figure 7.11 shows the combination of the electron and muon channels on the left hand side. At $m_H = 125\,\mathrm{GeV}$, a signal strength modifier above $\mu = 3.7$ is excluded by this channel, with an expected exclusion limit of 5.4. While this is not sensitive to the Standard Model itself, the result excludes WH production cross sections or $H \to \tau^+\tau^-$ branching ratios significantly higher than the Standard Model. The observation corresponds to a $1\,\sigma$ underfluctuation, since fewer events than expected from the background are observed. The full range of probed Higgs boson masses shows the same systematic underfluctuation, due to the poor resolution of the di-tau mass. This leads to a high correlation between the upper limits for different mass points.

The plot on the right hand side shows a comparison of the expected limit for the three associated production analyses in CMS. The expected limit is used to compare the sensi-

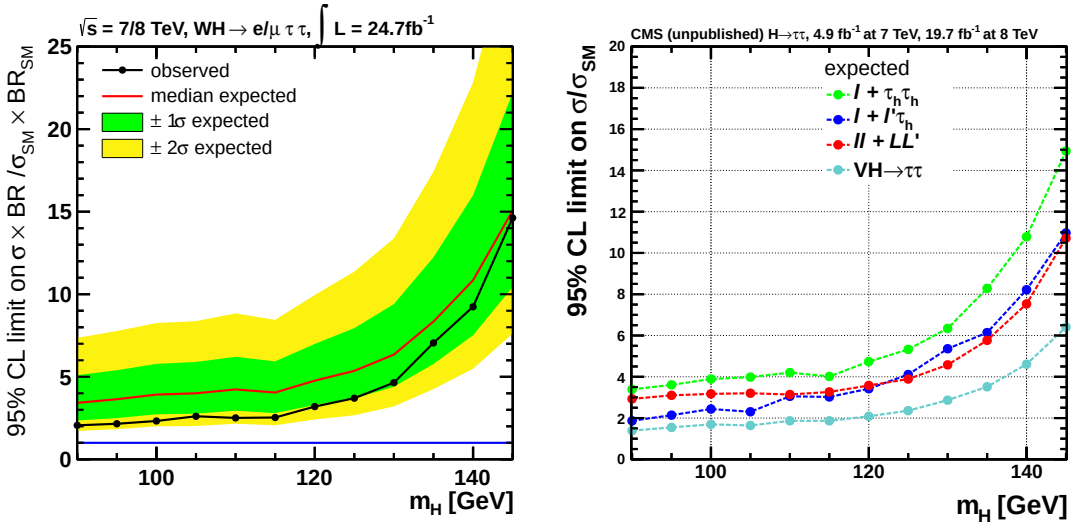

Figure 7.11: Exclusion limit as a function of the hypothesized Higgs boson mass. The left hand plot corresponds to the combination of the electron and the muon channel in the WH hadronic final states. The right plot shows the expected limit of all associated production channels in CMS, and their combination. The green line corresponds to the analysis presented in this chapter, the blue one is the WH semileptonic channel and the red line corresponds to the ZH channels. The turquoise line is the combination of the three channels. From [196].

tivity of the three channels, since it is unaffected by statistical fluctuations in the data. The other two channels are more sensitive than the one presented in this thesis, because in those channels there is only one τ_{had} candidate and at least two light leptons. Even though the branching ratio for a tau lepton to decay hadronically is higher than for a decay to light leptons, the higher identification efficiency and lower misidentification rates for electrons and muons outweigh the lower branching ratio. The expected (observed) upper limit for the combination of all channels is $\mu = 2.3$ (2.1), which is already very close to Standard Model sensitivity. The contribution of the hadronic channel to the overall combination is $\approx 20\,\%$.

This is the first analysis at the LHC searching for the Higgs boson in associated production in the di-tau final state and a leptonic decay of the associated boson. In particular, there is no comparable result to date from the ATLAS collaboration. At the Tevatron, CDF has performed a search for associated Higgs boson production in the $\tau^+\tau^-$ channel with $6.2\,\text{fb}^{-1}$. Their expected (observed) upper limit is 23.3 (26.5) times the Standard Model value for $m_H = 125\,\text{GeV}$ [197]. The D0 collaboration has searched for the $\mu + \tau_{\text{had}} + \tau_{\text{had}}$ final state and set an expected (observed) upper limit at 13.0 (19.4) times the Standard Model value with an integrated luminosity of $8.6\,\text{fb}^{-1}$ [198]. The analysis presented here supersedes the Tevatron results by a factor 2 - 10.

7.6.1 Summary and Outlook

Overall, the data collected in the first run of the LHC are not enough to observe the $H \to \tau^+\tau^-$ decay in the associated production channels. For the second run of the LHC, the associated production channels will become more interesting. The larger production

cross section at 13 TeV or 14 TeV, together with higher integrated luminosities, will provide enough data to become sensitive to the Standard Model Higgs boson. Due to a higher number of pile-up interactions, it is also expected that the misidentification rate of quark and gluon jets as τ_{had} candidates rise, leading to more background contributions. This effect can be partly mitigated by improvements in the τ_{had} identification and reconstruction. For example, it is envisaged to make use of the finite lifetime of the tau lepton (impact parameter and secondary vertex) in the τ_{had} reconstruction for the next LHC run [199].

The associated production mechanism has also theoretical and experimental advantages compared to the dominant gluon-gluon process, some of which might turn out to play an important role in the LHC run 2. For example, theory uncertainties are low compared to the gluon-gluon fusion process, and the channels can contribute to the bosonic coupling measurement. In addition, it is likely that the trigger thresholds will have to be raised for the conditions expected in LHC run 2. In this case, especially the fully hadronic channel can benefit from the additional leptons when the Higgs boson is produced together with a W boson or Z boson, keeping the acceptance high. A very interesting topic for the next LHC run are spin and CP studies in the $H \to \tau^+\tau^-$ channel, which is indeed best observed in the fully hadronic channel [181].

Summary and Conclusions

The discovery of the Higgs boson was a major scientific breakthrough. It has started a quest to precisely measure its properties, in order to find out whether the newly discovered particle is consistent with the theory, or whether it is a first glimpse at physics beyond the Standard Model. This thesis focuses on the coupling of the Higgs boson to tau leptons. Evidence for such a coupling has been found by both the ATLAS and CMS experiments at the Large Hadron Collider.

Tau leptons have a macroscopic lifetime, which allows to make use of the impact parameter, or, in the case of decays into three charged particles, secondary vertices, to identify and reconstruct tau decays. This procedure is only feasible when the detector has a high impact parameter resolution. In CMS, the silicon pixel detector is the crucial instrument for the impact parameter measurement. Both the intrinsic position resolution and the distance of the detector to the interaction point are relevant, and both will be improved with the upgraded detector. Furthermore, the additional layer in both the barrel and the endcaps improves the determination of the track parameters. Under current conditions, a position resolution in r-ϕ direction in the barrel region of 8 µm has been measured, with a pitch of 100 µm. This is consistent between Monte Carlo simulation, test beam measurements and a direct measurement with CMS data. Test beam measurements at DESY presented in this thesis show that for the upgraded version of the readout chip, an improvement of up to 25 % is possible, corresponding to a resolution of 6 µm. The R&D phase of the upgrade is now essentially complete, and in the remaining time until the upgraded detector is installed at the end of 2016, the modules will be produced and assembled.

CMS sees evidence of $H \to \tau^+\tau^-$ decays with more than $3\,\sigma$ at a Higgs boson mass of 125 GeV, in agreement with the Standard Model. The dominant background comes from the decay of the Z boson into a tau lepton pair, and the more precise this background is known, the more significant the signal. It is therefore estimated from the data sample itself as much as possible, to avoid systematic uncertainties inherent to the simulation. This thesis presents the tau embedding method with which muons in measured $\mathrm{Z}/\gamma^* \to \mu^+\mu^-$ events are replaced by simulated tau leptons. The method has been improved compared to the previously established version by merging the simulated part of the event with the data event at a lower level of the reconstruction chain. Applying the procedure on event samples generated with Monte Carlo simulation allows to validate the method. Differential distributions in many observables are reproduced correctly, and a clear improvement with respect to the previous method is visible. This makes the method ready for the next run of the LHC, where a higher number of pile-up interactions is expected. In addition, a technique is proposed to reduce systematic effects coming from the selection of di-muon events by transforming their four-vectors before performing the particle replacement. Other sources of systematic uncertainties, such as photon radiation from the original muons and the noise thresholds in the calorimeters, have been studied and quantified. It is also shown that the tau embedding method preserves the spin correlations between the two tau leptons, making the method a useful tool for studying the spin and CP properties of the Higgs boson in the di-tau final state.

There are multiple production mechanisms for the Higgs boson. Amongst others, the Higgs boson can be produced in association with a W boson or a Z boson. While the production cross section for this process is small compared to other mechanisms, additional leptons in the final state originating from the decay of the extra boson simplify the analysis from an experimental point of view. The background is reduced by a large factor, and the triggering of candidate events is more efficient, especially in the case of fully hadronic tau decays. In this thesis, an analysis is presented where the W boson decays into a light lepton and a neutrino, and both tau leptons decay hadronically. No excess in the data is seen and therefore an exclusion limit is set: the product of production cross section and branching ratio is less than 3.7 times the value predicted by the Standard Model at 95 % confidence level for a Higgs boson mass of 125 GeV. The expected exclusion limit is 5.4 times the Standard Model when no Higgs boson exists, which is a $1\,\sigma$ deviation from the observation. Therefore, this channel alone is not sensitive to the Standard Model Higgs boson. However, when combined with the other channels where the Higgs boson is produced together with a W boson or a Z boson, the expected (observed) exclusion limit is at 2.3 (2.1) times the Standard Model value, and at low masses around 100 GeV it is sensitive already. Combined with the other production methods, the channels contribute to the $3.4\,\sigma$ evidence in the $\tau^+\tau^-$ channel. The associated production channels are also interesting in the light of future LHC runs. Since the production cross section is so small, the analysis is entirely statistically limited, so that it profits considerably from a higher integrated luminosity. In the next LHC run, it is likely that the trigger thresholds have to be raised to accommodate the higher instantaneous luminosity. Due to the additional leptons in the final state, the associated production channels are less sensitive to higher trigger thresholds. This is especially relevant for fully hadronic tau decays, which are most sensitive to the spin and CP properties of the Higgs boson, a very interesting topic in the upcoming LHC run.

The results presented in this thesis represent a significant contribution to the CMS $H \to \tau^+\tau^-$ analysis published in [35]. Both the particle-based embedding technique for estimating the dominant $\mathrm{Z}/\gamma^* \to \tau^+\tau^-$ background and the analysis in the associated Higgs production channels are crucial parts of the published work. Furthermore, important groundwork for future analyses has been laid within the scope of this thesis. The associated production channels will become more interesting in future LHC runs due to their experimental benefits. An improved resolution of the readout chip foreseen for the CMS barrel pixel upgrade has been confirmed in test beam measurements, which eventually will directly improve the identification of hadronically decaying tau leptons.

The Higgs boson is not the only research topic at the LHC. The ultimate goal is to find answers to the questions which cannot be explained by the Standard Model. While studying the Higgs sector in detail is a very promising approach, also direct searches for new particles and phenomena are underway. With collision energies close to the design energy of 14 TeV, a previously unexplored region becomes accessible with run 2 of the LHC. Eventually, the findings will determine the design and concept of next-generation facilities.

A Pile-Up Mitigation

In physics analyses, one is often only interested in jets and leptons that originate from the hard-scatter process. However, there can also be particles coming from other pp interactions within the same bunch crossing ("pile-up"). In the data taking in 2012, on average there were about 20 interactions per bunch crossing in CMS. The techniques presented in this section aim to reduce the effect of pile-up interactions on the identification of physics objects. Especially the lepton isolation, the jet clustering and the $E_{\mathrm{T}}^{\mathrm{miss}}$ reconstruction are very sensitive to pile-up.

All pile-up mitigation techniques have in common that they rely on the silicon tracker to extrapolate the tracks to the interaction region. The interaction region is about 24 cm long in the direction of the beams, so that the primary vertices of different collisions will have a macroscopic distance between each other. This allows a track to be classified as a track coming from a pile-up vertex or the primary interaction vertex.

A.1 Lepton Isolation

In order to remove particles from pile-up interactions in the isolation cone, the isolation formula from Equation 3.5 is modified slightly as follows:

$$I_{\mathrm{PF}}^{\mathrm{rel}} = \frac{1}{p_{\mathrm{T}}} \left(p_{\mathrm{T}}^{\mathrm{charged}} + \max\left(0, p_{\mathrm{T}}^{\mathrm{neutral}} + p_{\mathrm{T}}^{\mathrm{gamma}} - \Delta\beta\right)\right) , \tag{A.1}$$

where, for the charged particles, only those particles are counted whose track has a longitudinal impact parameter of less than 2 mm from the primary interaction vertex. In order to remove pile-up contributions from the neutral isolation components, the $\Delta\beta$ variable is defined as

$$\Delta\beta = 0.5 \cdot p_{\mathrm{T}}^{\mathrm{PU}} , \tag{A.2}$$

where $p_{\mathrm{T}}^{\mathrm{PU}}$ is the sum over all charged particles in the isolation cone which have a longitudinal impact parameter greater than 2 mm. This is exactly the contribution of pile-up to the charged part of the isolation. The factor 0.5 is the expected ratio of charged particles to neutral particles in jets [200]. For hadronic taus it is instead chosen to be 0.4576, to make the τ_{had} identification efficiency a flat function of additional interaction vertices.

A.2 Jets

The pile-up jet identification attempts to identify jets which are mostly clustered from particles coming from pile-up interactions.

A BDT is used to discriminate between hard-scatter and pile-up jets [201]. It is trained on simulated $Z \to \mu^+\mu^-$ events where the truth information is available on generator level. Apart from track extrapolation, also the jet shape is used to discriminate between pile-up and hard-scatter jets. Pile-up jets typically have their constituents somewhat evenly

distributed within the jet, while hard jets are often very collimated. These two ideas are behind many of the following variables that are used in the training of the BDT:

- The number of charged and neutral particles, respectively.
- The jet momentum three-vector in p_{T}, η and ϕ.
- The ratio of the scalar p_{T} sum of all charged particles from the hard-scatter vertex over the scalar p_{T} sum of all charged particles in a jet.
- The ratio of the scalar p_{T} sum of all charged particles from a pile-up vertex over the scalar p_{T} sum of all charged particles in a jet. Note that this does not necessarily contain the same information as the previous variable, since there can be charged particles not assigned to any vertex at all.
- The following quantities which characterize the jet shape:

$$\langle \Delta R \rangle = \frac{1}{p_{\mathrm{T}}^{\mathrm{jet}}} \sum_{\mathrm{particles}} p_{\mathrm{T}} \cdot \Delta R \tag{A.3}$$

$$p_{\mathrm{T}}^{\Delta R}(X) = \frac{1}{p_{\mathrm{T}}^{\mathrm{jet}}} \sum_{X \leq \Delta R \leq X+0.1} p_{\mathrm{T}}\,, \tag{A.4}$$

  where ΔR is the distance in η-ϕ space between the charged particle and the jet momentum vector, and the sums go over all particles in the jet (in the second case only for particles with ΔR between the two values). The variable $p_{\mathrm{T}}^{\Delta R}(X)$ is evaluated for $X = 0.0, 0.1, 0.2, 0.3, 0.4$.

The working point is chosen such that a 95% identification efficiency is achieved for hard-scatter jets with $p_{\mathrm{T}} > 25\,\mathrm{GeV}$.

A.3 Missing Transverse Energy

A BDT regression [95] is performed to compute a correction to the reconstructed $E_{\mathrm{T}}^{\mathrm{miss}}$, both its angle and its magnitude [201]. It is trained in simulated $Z/\gamma^* \to \mu^+\mu^-$ events where the transverse momentum of the Z boson is denoted as $\vec{q}_{\mathrm{T}}$ and the sum of all other particles, the *hadronic recoil*, as $\vec{u}_{\mathrm{T}}$. With this definition, $\vec{q}_{\mathrm{T}} + \vec{u}_{\mathrm{T}} + \vec{E}_{\mathrm{T}}^{\mathrm{miss}} = 0$ holds. A first BDT is now trained to find a correction to $\vec{u}_{\mathrm{T}}$ so that it matches the direction of $-\vec{q}_{\mathrm{T}}$. The second BDT is then trained to match the true magnitude of $\vec{q}_{\mathrm{T}}$. This is appropriate because in $Z/\gamma^* \to \mu^+\mu^-$ there is no true $E_{\mathrm{T}}^{\mathrm{miss}}$, since there are no neutrinos produced.

For the training of the BDTs, 5 different $E_{\mathrm{T}}^{\mathrm{miss}}$ variables are used, each with a different set of PF candidates in the sum of Equation 3.4.

- All particles (the normal $E_{\mathrm{T}}^{\mathrm{miss}}$ definiton).
- All charged particles whose track is linked to the vertex of the hard scattering by requiring that the longitudinal impact parameter is less than 2 mm.
- All charged particles whose track is linked to the vertex of the hard scattering, and all neutral particles that are part of a jet that has passed the pile-up jet ID.

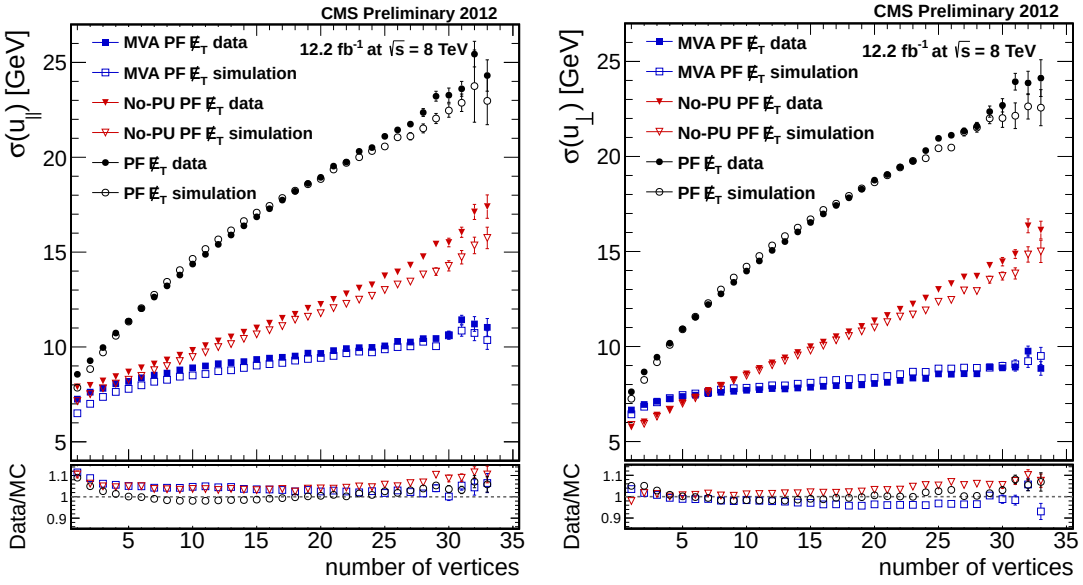

Figure A.1: Resolution of the particle flow $E_{\mathrm{T}}^{\mathrm{miss}}$ and MVA $E_{\mathrm{T}}^{\mathrm{miss}}$ in percent as a function of the number of interactions in simulated $Z \to \mu^+\mu^-$ events. The left hand side shows the $E_{\mathrm{T}}^{\mathrm{miss}}$ component that is parallel to the Z boson direction and the right hand side shows the component of $E_{\mathrm{T}}^{\mathrm{miss}}$ that is transverse to the Z boson direction. From [201].

- All charged particles whose track is not linked to the vertex of the hard scattering, and all neutral particles that are part of a jet that has failed the pile-up jet ID.
- All charged particles whose track is linked to the vertex of the hard scattering, and all neutral particles (including the ones not clustered in a jet), minus the neutral particles that are part of a jet that has failed the pile-up jet ID.

The input variables to the BDTs are then:

- The angle and magnitude of $\vec{u}_{\mathrm{T}}$ for all of the 5 $E_{\mathrm{T}}^{\mathrm{miss}}$ definitions.
- The scalar sum of transverse momenta of all particles contributing to $E_{\mathrm{T}}^{\mathrm{miss}}$ for all the 5 $E_{\mathrm{T}}^{\mathrm{miss}}$ definitions.
- The three-momenta of the two highest-p_{T} jets in the event.
- The number of reconstructed primary vertices.

The output of the BDT regression gives a new $E_{\mathrm{T}}^{\mathrm{miss}}$ estimate and will be referred to hereafter as "MVA $E_{\mathrm{T}}^{\mathrm{miss}}$" and is used for $E_{\mathrm{T}}^{\mathrm{miss}}$ reconstruction in all $H \to \tau^+\tau^-$ analyses in CMS. Figure A.1 shows the $E_{\mathrm{T}}^{\mathrm{miss}}$ resolution for the standard PF $E_{\mathrm{T}}^{\mathrm{miss}}$ and the MVA $E_{\mathrm{T}}^{\mathrm{miss}}$ in simulated $Z/\gamma^* \to \mu^+\mu^-$ events as a function of the number of reconstructed vertices. The MVA $E_{\mathrm{T}}^{\mathrm{miss}}$ improves the $E_{\mathrm{T}}^{\mathrm{miss}}$ resolution significantly, especially in the case of high pile-up.

B Invariant Di-Tau Mass Reconstruction

This section discusses different methods to reconstruct the invariant mass of a di-tau system at the LHC. The following sections give a brief overview of available mass estimators.

B.1 Visible Mass

The *visible mass* is constructed only from the visible tau decay products. If p^{vis} and p'^{vis} denote the four vectors of the reconstructed decay products then the visible mass is simply defined as

$$M_{\text{vis}}^2 = \left(p^{\text{vis}} + p'^{\text{vis}}\right)^2 . \tag{B.1}$$

Since the momentum carried away by the neutrinos is completely ignored by this method, the resulting mass does not have a peak at the true mass of the resonance but it is shifted to lower values. However, the method can still be useful for separating a $H \to \tau^+\tau^-$ signal from $Z \to \tau^+\tau^-$ background.

B.2 Collinear Approximation Mass

The *collinear approximation* attempts to reconstruct the four-vector of the tau lepton [202]. It makes the following two assumptions:

1. The neutrino(s) generated in the tau lepton decay are collinear to the visible decay products, i.e. the 3-momenta point in the same direction in the laboratory frame.

2. The only genuine $E_{\text{T}}^{\text{miss}}$ in the event is due to the neutrino(s) from the tau decays.

The applicability of the first assumption depends on the topology of the event. For tau leptons that are nearly in rest this is typically not true, however for tau leptons from decays of heavy resonances, such as Higgs bosons or Z bosons, this is a reasonable assumption. This is even more the case when the Higgs or Z particles themselves are boosted, i.e. have high momenta in the laboratory frame.

The second assumption is also fulfilled for pure Z boson or Higgs boson decays. An example where this is not the case would be associated production with a $W^\pm$ boson, in $W^\pm Z$ or $W^\pm H$ events. In this case the decay of the $W^\pm$ boson introduces another neutrino, which leads to an additional $E_{\text{T}}^{\text{miss}}$ contribution not coming from a tau decay. The experimental $E_{\text{T}}^{\text{miss}}$ determination is a challenge, especially in environments with high pile-up. The left hand side of Figure B.1 illustrates the two assumptions.

Formally, let E and E' denote the energies of the two tau leptons in the laboratory frame, and E^{vis} and E'^{vis} correspond to the energies of the visible decay products. Then $x_1 = E^{\text{vis}}/E$ and $x_2 = E'^{\text{vis}}/E'$. The first assumption can now be written as the following

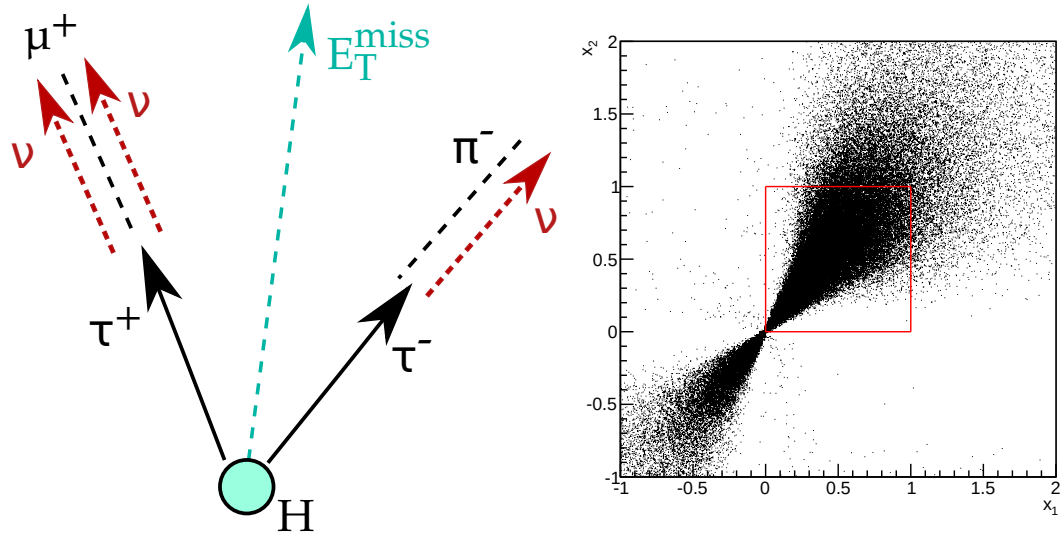

Figure B.1: Left: Assumptions of the collinear approximation in a $H \to \tau^+\tau^- \to \tau_{\text{had}} + \tau_\mu$ event: the neutrinos from the tau decays have the same direction as the visible decay products, and $E_{\text{T}}^{\text{miss}}$ is solely due to the neutrinos. Right: Distribution of the energy fractions of the visible tau decay momenta in simulated $Z \to \tau^+\tau^- \to \tau_{\text{had}} + \tau_\mu$ events in CMS. The red frame represents physical solutions.

relation for the transverse momenta $\vec{p}_{\text{T}}$ and $\vec{p}\,'_{\text{T}}$, and the momenta of the visible decay products $\vec{p}_{\text{T}}^{\,\text{vis}}$ and $\vec{p}\,'^{\,\text{vis}}_{\text{T}}$:

$$\vec{p}_{\text{T}} = x_1 \cdot \vec{p}_{\text{T}}^{\,\text{vis}}\,, \qquad \vec{p}\,'_{\text{T}} = x_2 \cdot \vec{p}\,'^{\,\text{vis}}_{\text{T}} \tag{B.2}$$

The second assumption can be written as

$$\vec{p}_{\text{T}} + \vec{p}\,'_{\text{T}} = \vec{p}_{\text{T}}^{\,\text{vis}} + \vec{p}\,'^{\,\text{vis}}_{\text{T}} + \vec{E}_{\text{T}}^{\,\text{miss}} \tag{B.3}$$

Plugging B.2 into B.3 yields a system of equations for x_1 and x_2. The solution is given by

$$x_1 = \frac{p_x^{\text{vis}} p_y'^{\,\text{vis}} - p_y^{\text{vis}} p_x'^{\,\text{vis}}}{p_y'^{\,\text{vis}} p_x^{\text{miss}} - p_x'^{\,\text{vis}} p_y^{\text{miss}} + p_x^{\text{vis}} p_y'^{\,\text{vis}} - p_y^{\text{vis}} p_x'^{\,\text{vis}}} \tag{B.4}$$

$$x_2 = \frac{p_x^{\text{vis}} p_y'^{\,\text{vis}} - p_y^{\text{vis}} p_x'^{\,\text{vis}}}{p_x^{\text{vis}} p_y^{\text{miss}} - p_y^{\text{vis}} p_x^{\text{miss}} + p_x^{\text{vis}} p_y'^{\,\text{vis}} - p_y^{\text{vis}} p_x'^{\,\text{vis}}}\,. \tag{B.5}$$

This allows the invariant di-tau mass to be calculated according to

$$M_{\text{coll}}^2 = \left(p + p'\right)^2 = \left(\frac{p^{\text{vis}}}{x_1} + \frac{p'^{\,\text{vis}}}{x_2}\right)^2\,. \tag{B.6}$$

The collinear approximation only yields a physical solution for $0 < x_1 < 1$ and $0 < x_2 < 1$. If this is not the case, either one of the two assumptions is spoiled (typically the first), or the system of equations is degenerate, or nearly degenerate. The latter happens if the two tau leptons are back-to-back to each other. Even if they are only very close to back-to-back, the solution is very unstable numerically, and small deviations in the measured

momenta can lead to large differences in x_1 and x_2. In inclusive $Z \to \tau^+\tau^-$ decays, only about 50% of the events fulfill these criteria. The right hand side of Figure B.1 shows the distribution of x_1 and x_2 for simulated $Z \to \tau^+\tau^- \to \tau_{\text{had}} + \tau_\mu$ events in CMS.

The distribution of the collinear approximation mass typically shows a peak at the right mass, but also suffers from a long non-Gaussian tail which makes it hard to discriminate between Higgs and Z events.

B.3 Missing Mass Calculator

The *missing mass calculator* (MMC) technique attempts to reconstruct the full di-tau system without the shortcomings of the collinear approximation [203]. It has been successfully used with CDF data and is the standard technique for di-tau mass reconstruction in ATLAS [34].

There are between six and eight unknowns in the description of the di-tau system, depending on the tau decay channels. The unknowns are the x, y and z momentum components of the system of neutrinos from each tau decay, and the invariant mass of the neutrino system. In the case of hadronic tau decays, there is only one neutrino, and the invariant mass is fixed to 0. For leptonic tau decays, however, there are two neutrinos and their invariant mass can be different from 0.

With the available observables, i.e. the momenta of the visible decay products and the two components of $E_{\text{T}}^{\text{miss}}$, there are only four constraints, given by the compatibility of the mass of the tau decay products with the mass of the tau lepton, and by the compatibility of the momenta of the tau decay products with $E_{\text{T}}^{\text{miss}}$. These constraints can be formulated as follows:

$$p_x^{\text{miss}} = p_x^{\nu} + p_x'^{\nu} \tag{B.7}$$

$$p_y^{\text{miss}} = p_y^{\nu} + p_y'^{\nu} \tag{B.8}$$

$$\begin{aligned}(M^{\tau})^2 = {} & (M^{\nu})^2 + (M^{\text{vis}})^2 + 2\sqrt{(p^{\text{vis}})^2 + (M^{\text{vis}})^2}\sqrt{(p^{\nu})^2 + (M^{\nu})^2} \\ & - 2p^{\text{vis}}p^{\nu}\cos\Delta\theta^{\nu\nu}\end{aligned} \tag{B.9}$$

$$\begin{aligned}(M'^{\tau})^2 = {} & (M'^{\nu})^2 + (M'^{\text{vis}})^2 + 2\sqrt{(p'^{\text{vis}})^2 + (M'^{\text{vis}})^2}\sqrt{(p'^{\nu})^2 + (M'^{\nu})^2} \\ & - 2p'^{\text{vis}}p'^{\nu}\cos\Delta\theta'^{\nu\nu}\end{aligned} \tag{B.10}$$

where p^{ν}, p'^{ν}, M^{ν} and M'^{ν} are the momenta or invariant masses of the two neutrino systems (corresponding to up to 8 unknowns), p^{vis}, p'^{vis}, M^{vis} and M'^{vis} are the momenta and invariant masses of the visible tau decay products, and $M^{\tau} = M'^{\tau} = 1.777\,\text{GeV}$. Further, $\Delta\theta^{\nu\nu}$ is the polar angle between the momentum of the neutrino system and the momentum of the visible decay products.

The solution space is now 2-, 3- or 4-dimensional, but the various solutions do not have the same probability to be realized in nature, due to the tau decay kinematics. The MMC approach uses the distribution of the geometric separation between the visible and the invisible tau decays, $\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\phi)^2}$, in the following way: the probability density function $\mathcal{P}(\Delta R; p)$ is obtained with PYTHIA and TAUOLA for each tau momentum p. Then, the solution space is scanned in a grid to find the solution for which the likelihood $\mathcal{L} = \mathcal{P}(\Delta R; p) \cdot \mathcal{P}(\Delta R'; p')$ is maximized.

The method still assumes that the only source of $E_{\mathrm{T}}^{\mathrm{miss}}$ in the event are the neutrinos from the tau decay, i.e. without adaption it cannot be used for $W^{\pm}Z$ or $W^{\pm}H$ events. This can still cause a problem in events with back-to-back topology, where small mismeasurements of $E_{\mathrm{T}}^{\mathrm{miss}}$ can cause large differences in the output, and lead to unwanted tails. In order to compensate for this, the solution space is artificially inflated by scanning in two more variables, $\Delta p_x^{\mathrm{miss}}$ and $\Delta p_y^{\mathrm{miss}}$, which represent mismeasurements of $E_{\mathrm{T}}^{\mathrm{miss}}$. The likelihood is then amended by the term $\mathcal{P}\left(\Delta p_x^{\mathrm{miss}}, \Delta p_y^{\mathrm{miss}}\right)$, given by

$$\mathcal{P}\left(\Delta p_x^{\mathrm{miss}}, \Delta p_y^{\mathrm{miss}}\right) = \exp\left(-\frac{1}{2}\begin{pmatrix} \Delta p_x^{\mathrm{miss}} \\ \Delta p_y^{\mathrm{miss}} \end{pmatrix} V^{-1} \begin{pmatrix} \Delta p_x^{\mathrm{miss}} \\ \Delta p_y^{\mathrm{miss}} \end{pmatrix}\right) \tag{B.11}$$

where V is the covariance matrix of the $E_{\mathrm{T}}^{\mathrm{miss}}$ measurement. This quantity is typically available on an event-by-event basis, and the uncertainties of all physics objects that were used in the $E_{\mathrm{T}}^{\mathrm{miss}}$ measurement contribute to it.

The MMC method combines the advantages of the visible mass and the collinear approximation in the sense that it provides a physical result for every event, produces a peak at the mass value of the resonance and avoids long tails in the distribution.

B.4 SVfit

The idea behind the *SVfit* method is very similar to that of the MMC. It is used by CMS for di-tau mass reconstruction [35].

In the SVfit method, the tau decays are parameterized differently than in the MMC method. For each tau decay, there are 3 unknown parameters $\vec{a} = (x, \phi, M^{\nu})$, where x is the energy fraction of the visible tau decay products in the laboratory frame, ϕ is the azimuthal angle of the tau lepton momentum in the laboratory frame and M^{ν} is the invariant mass of the neutrino system (0 for hadronic decays). This is the same number of parameters than in the MMC method (after introduction of $\Delta p_x^{\mathrm{miss}}$ and $\Delta p_y^{\mathrm{miss}}$).

The SVfit method now maximizes the probability for the resonance having a certain mass, which is found by integrating over the likelihood for all possible tau decays that lead to that mass:

$$\mathcal{P}(M_{\tau\tau}^{i}) = \int \delta\left(M_{\tau\tau}^{i} - M_{\tau\tau}\left(p^{\mathrm{vis}}, p'^{\,\mathrm{vis}}, \vec{a}, \vec{a}\,'\right)\right) f(p^{\mathrm{vis}}, p'^{\,\mathrm{vis}}; \vec{E}_{\mathrm{T}}^{\mathrm{miss}}, \vec{a}, \vec{a}\,') \,\mathrm{d}\vec{a}\,\mathrm{d}\vec{a}\,' \,, \tag{B.12}$$

where $M_{\tau\tau}\left(p^{\mathrm{vis}}, p'^{\,\mathrm{vis}}, \vec{a}, \vec{a}\,'\right)$ is the di-tau invariant mass that corresponds to the given values of the unknowns, f is the likelihood which is composed of three terms. Two terms correspond to the tau decay matrix elements for the two tau leptons. For the hadronic tau decay, which is a two-body-decay, the matrix element is derived from the two-body phase space taken from [119]. It is given by

$$\mathcal{L} = \frac{\mathrm{d}\Gamma}{\mathrm{d}x\,\mathrm{d}\phi} \propto \frac{1}{1 - (M^{\mathrm{vis}})^2/(M^{\tau})^2} \,, \tag{B.13}$$

defined in the physically allowed region $(M^{\mathrm{vis}})^2/(M^{\tau})^2 < x < 1$. For the leptonic tau decays, which are three-body decays, the matrix element is taken from [204] and has the form

$$\mathcal{L} = \frac{\mathrm{d}\Gamma}{\mathrm{d}x\,\mathrm{d}\phi\,\mathrm{d}M^{\nu}} \propto \frac{M^{\nu}}{4(M^{\tau})^2}\left(\left((M^{\tau})^2 + 2(M^{\nu})^2\right) \cdot \left((M^{\tau})^2 - (M^{\nu})^2\right)\right) \,, \tag{B.14}$$

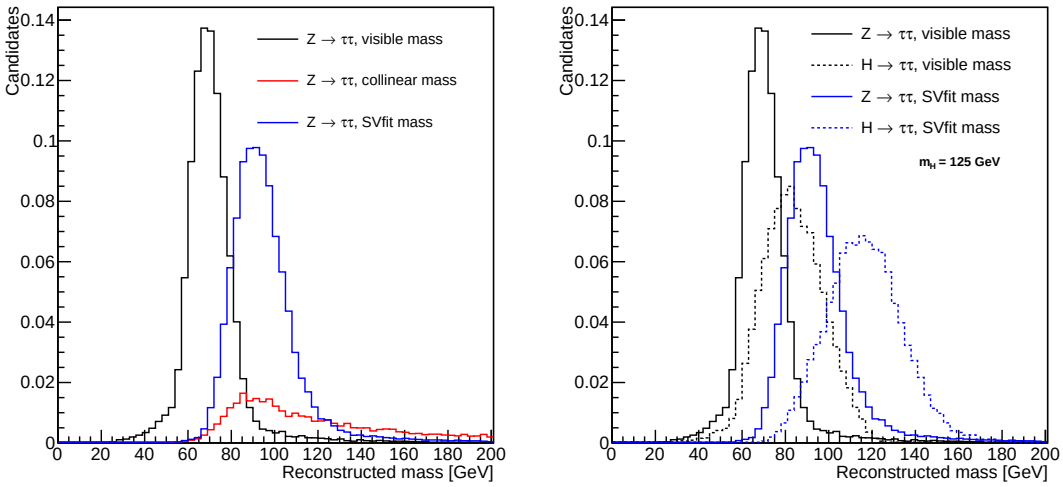

Figure B.2: Left: Comparison of the visible mass, the collinear approximation mass, and the SVfit mass in simulated $Z \to \tau^+\tau^-$ events. It can be seen how the collinear approximation leads to a long tail and how many events are rejected. Right: Mass separation between Z and H for the visible mass (black) and the SVfit mass (right).

defined in the physically allowed region $0 < x < 1$ and $0 < M^\nu < M^\tau\sqrt{1-x}$.

The third term in the likelihood is the same term used in the MMC method for the $E_{\mathrm{T}}^{\mathrm{miss}}$ uncertainty, given by Equation B.11.

Summarizing, the main differences between the MMC and the SVfit methods are:

- SVfit uses fully differential likelihoods for the tau decay matrix element, whereas the MMC uses Monte Carlo simulation to obtain ΔR distributions. It is not clear, however, that the ΔR variables provides the full information about the tau decay kinematics.
- If there is a certain combination of unknowns $\vec{A}$ which leads to mass M_A, and combinations $\vec{B}_1$, $\vec{B}_2$, and $\vec{B}_3$, all of which lead to mass M_B, then the MMC would always choose M_A if its likelihood is higher than any of $\vec{B}_1$, $\vec{B}_2$ or $\vec{B}_3$. SVfit, however, would choose M_B if the sum of the probabilities of $\vec{B}_1$, $\vec{B}_2$, or $\vec{B}_3$ being realized is higher than the probability of $\vec{A}$ being realized. In other words, MMC finds the most likely four momenta for the two tau leptons while SVfit finds the most likely invariant mass, and the two are not necessarily equivalent.

B.5 Comparison of the methods

In Figure B.2, a comparison between the different mass reconstruction methods is shown. On the left hand side, the distribution of the visible mass, the collinear approximation mass and the SVfit mass are shown for simulated $Z \to \tau^+\tau^-$ events. While the visible mass underestimates the true mass and the collinear approximation has a long tail to the right and rejects many events, the SVfit mass combines the advantages of both methods.

The performance of the MMC (not shown here) is very similar to the one of SVfit. On the right hand plot, a comparison between $Z \to \tau^+\tau^-$ and $H \to \tau^+\tau^-$ events is depicted. The SVfit mass improves the mass separation between the two, leading to a better background rejection in $H \to \tau^+\tau^-$ analyses.

C Embedding Validation

This appendix contains additional material that contributes to the validation of the embedding procedure discussed in Chapter 6.

C.1 Muon Embedding

In this section, additional validation plots are shown when replacing the reconstructed muons by generator-level muons instead of generator-level tau leptons. The selection of di-muon events is then performed es described in Section 6.3.1. Each plot shows four curves: The standard $Z/\gamma^* \to \mu^+\mu^-$ Monte Carlo sample in black, the PF embedded sample in green, the RH embedded sample in red and the RH embedded sample with the muon four-vector transformation as described in Section 6.4.4 in blue.

Figure C.1 shows various observables on generator level, for events where both generator-level muons are within the di-muon acceptance, defined as the following:

- $p_T > 20\,\text{GeV}$ for the leading muon, $p_T > 10\,\text{GeV}$ for the trailing muon.
- $|\eta| < 2.4$ for both muons
- $M_{\mu\mu} > 50\,\text{GeV}$, where $M_{\mu\mu}$ is the invariant mass of both muons.

The normalization is performed such that the number of events within the acceptance is the same as for the $Z/\gamma^* \to \mu^+\mu^-$ Monte Carlo sample. The differences in the distributions are in general very small. All differences are due to a bias coming from the initial selection of $Z/\gamma^* \to \mu^+\mu^-$ events, and from the smearing introduced by detector effects for the embedded samples. A correction for the $Z/\gamma^* \to \mu^+\mu^-$ selection as a function of the muon transverse momenta and pseudorapidities is already applied as described in Section 6.2.1. In these distributions, there is no significant difference between PF embedding and RH embedding by construction, since there is only a difference between the two in the reconstructed objects of embedded events.

Figure C.2 shows the same observables on reconstruction level. The discrepancy of the muon p_T distribution is discussed in Section 6.2.1 and in Section 6.4.4 for the case with the mirror transformation. Figure C.3 shows additional observables on reconstruction level which also include other event content than the two muons. The lower right plot shows the p_T of the positive muon before the isolation requirement. It can be seen that the effect at low p_T is gone, and the distribution resembles the distribution on generator level from Figure C.1.

The reconstructed di-muon mass in Figure C.2 shows a discrepency for the embedded sample with the mirror transformation at low mass. This effect is coming from the fact that for the embedded muons, the simulation of FSR has been disabled in order to avoid applying the FSR twice. Therefore, the isolation efficiency is "too good" with the mirror transformation, since because of the transformation, the FSR photons from the original

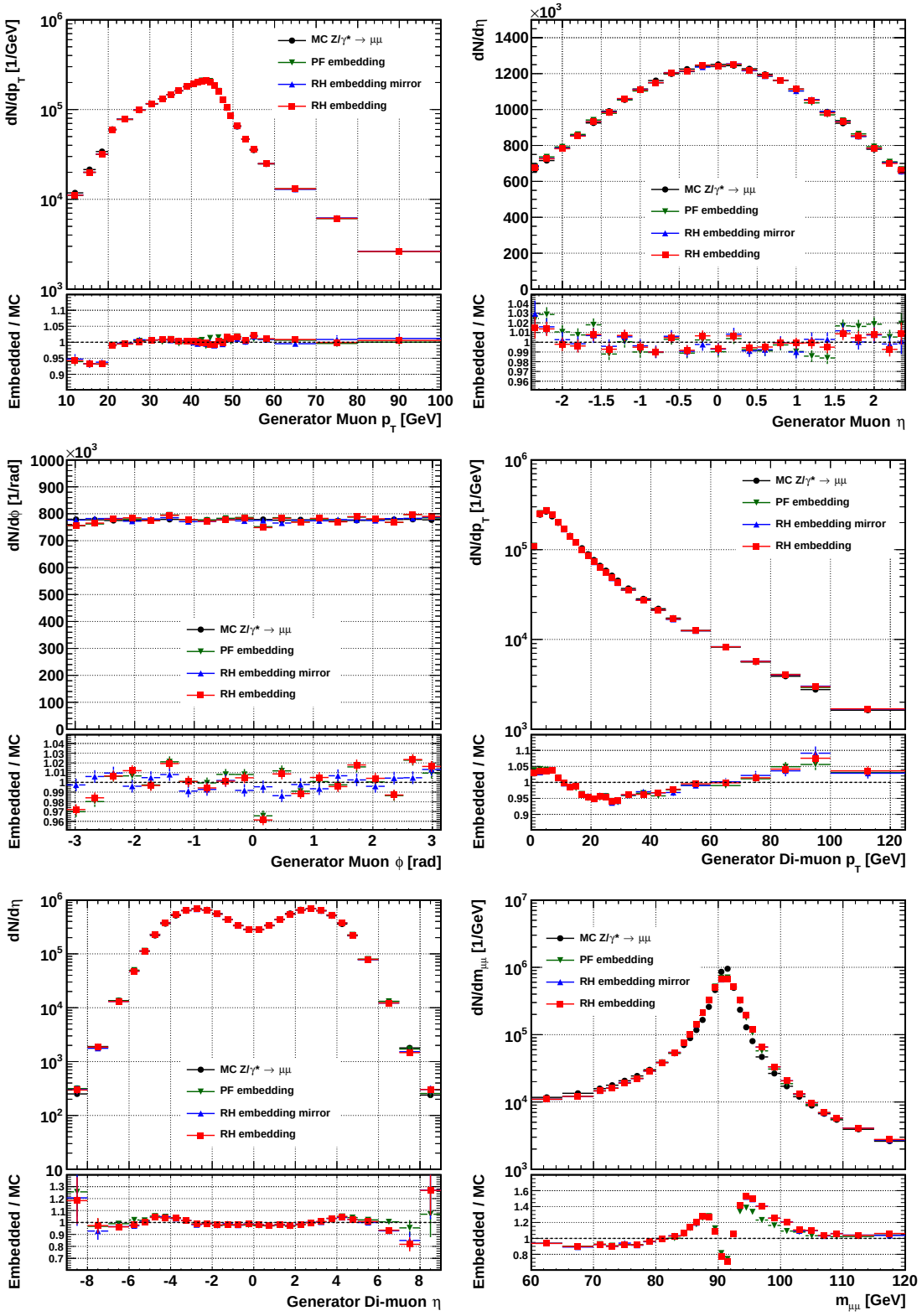

Figure C.1: Positive muon p_{T} (top left), η (top right) and ϕ (center left), and di-muon p_{T} (center right), η (bottom left) and mass (bottom right), on generator level.

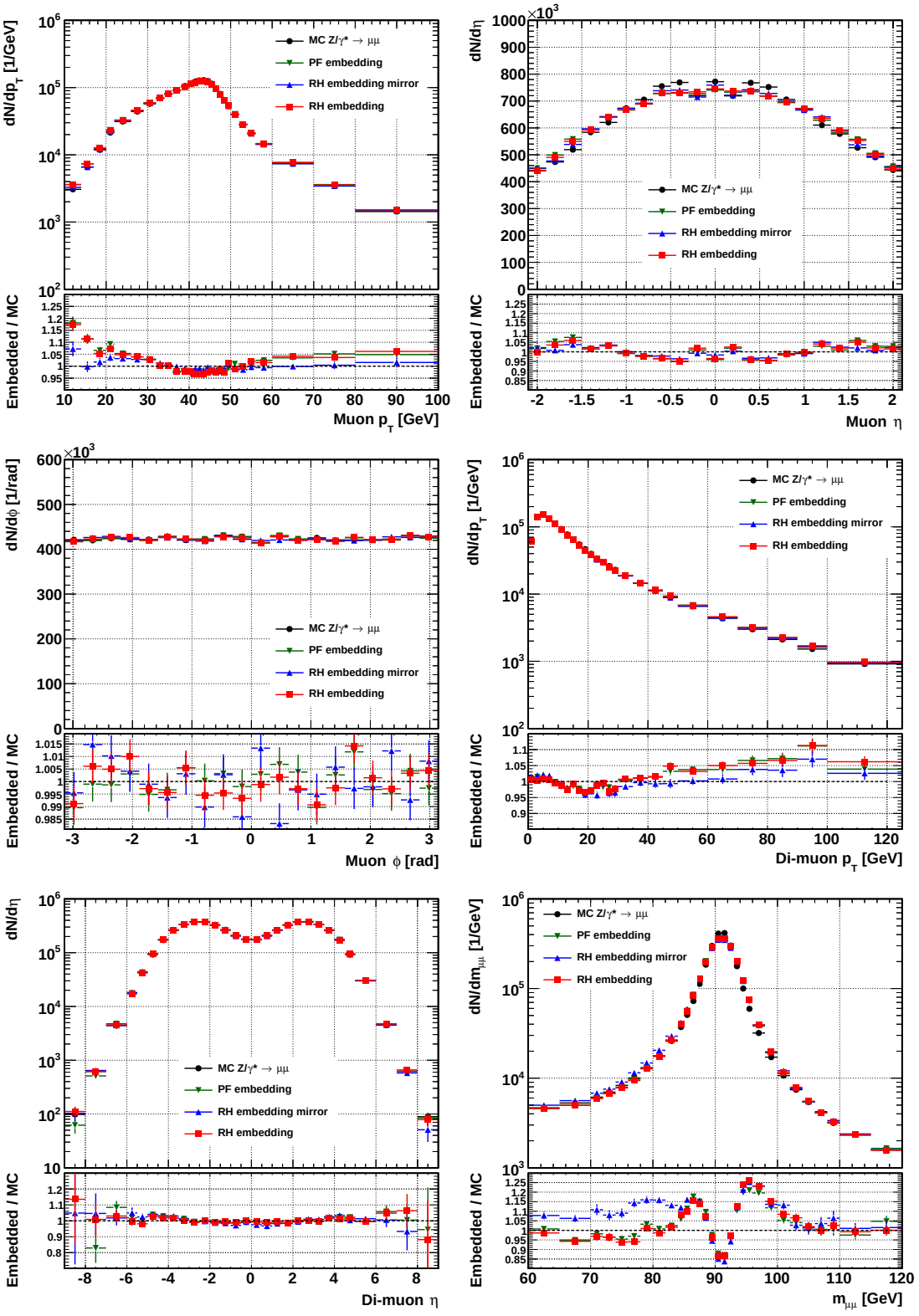

Figure C.2: Positive muon p_{T} (top left), η (top right) and ϕ (center left), and di-muon p_{T} (center right), η (bottom left) and mass (bottom right), on reconstruction level.

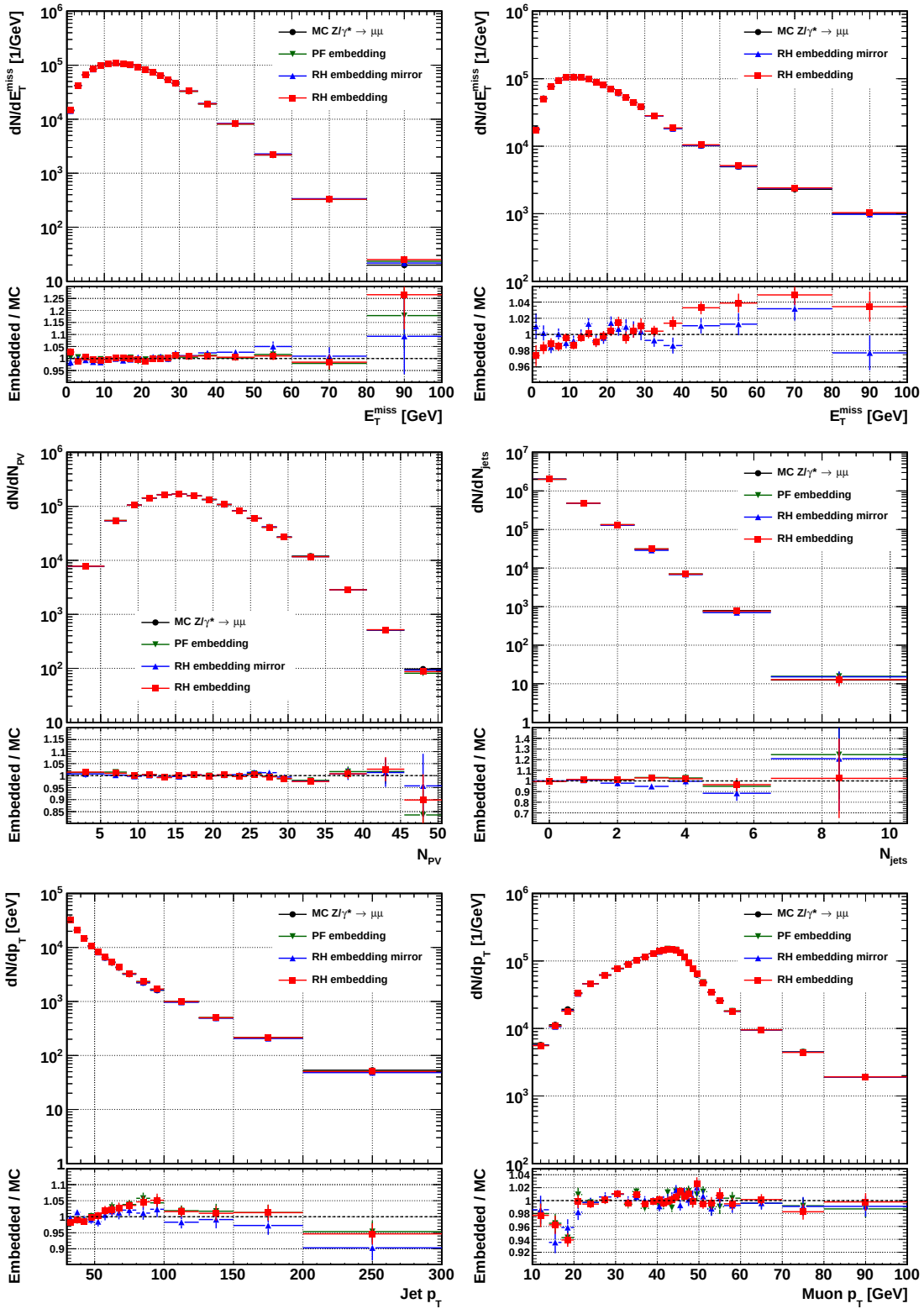

Figure C.3: PF-based $E_{\mathrm{T}}^{\mathrm{miss}}$ (top left), Calorimeter-based $E_{\mathrm{T}}^{\mathrm{miss}}$ (top right), number of primary vertices (center left), jet multiplicity p_{T} (center right), leading jet p_{T} (bottom left) and positive muon p_{T} without isolation cuts (bottom right).

muons do not contribute to the isolation sum. The effect shows up at low masses since the isolation cut is p_{T} dependent and the values at lower masses are correlated to low p_{T} muons. In the tau embedding, this effect is gone for two reasons: first, tau leptons radiate much less than muons, and second, for the tau embedding, the FSR of the embedded tau leptons is enabled.

C.2 Tau Embedding

In this section, additional validation plots are shown when replacing the reconstructed muons by generator-level tau leptons. The event selection is described in Section 6.3.2. Each plot again shows four curves: The standard $Z/\gamma^* \to \tau^+\tau^-$ Monte Carlo sample in black, the PF embedded sample in green, the RH embedded sample in red and the RH embedded sample with the muon four-vector transformation as described in Section 6.4.4 in blue.

C.2.1 The $\tau_\mu + \tau_{\text{had}}$ Final State

Figures C.4 and C.5 show important observables on generator level in the $\tau_\mu + \tau_{\text{had}}$ final state. Only kinematic acceptance cuts on generator-level have been applied on the p_{T} and η of the visible tau decay products. When not noted otherwise, the full kinematic variables of the taus are plotted and not only the visible fraction.

These plots are very similar to the generator-level validation plots for the muon embedding, since no reconstruction steps have been performed. Other than the different acceptance cuts, the major difference to the muon embedding plots is coming from final state radiation (FSR) of the muons. In the muon embedding, the FSR was simply switched off to obtain a correct modeling of the muon radiation, since the original muons have undergone radiation already. The same is the case now for the embedded samples, however in the reference $Z/\gamma^* \to \tau^+\tau^-$ simulation, the effect of FSR is reduced significantly since the tau leptons do not radiate as much. The effect is mostly visible in the di-tau mass distribution on the top left of Figure C.5: other than the additional smearing due to the muon reconstruction effects, the low mass tail is significantly enhanced in the embedded samples due to the muon radiation effects. The radiation effect is also visible, to a much lesser extent, in the distribution of the visible mass on the top right. It is discussed in more detail in Section 6.4.1.

Another feature can be seen in the distribution of the number of primary vertices in the bottom right of Figure C.5. The trend visible in the distribution shows the pile-up dependence of the di-muon selection efficiency. It is lower at a large number of primary vertices. This effect is partly mitigated on reconstruction level when the events for which pile-up effects deterioate the identification of physics objects are also filtered out in the plain $Z/\gamma^* \to \tau^+\tau^-$ simulation. This can be seen in the bottom right plot of Figure C.7.

Figures C.6, C.7 and C.8 show several observables on reconstruction level. Good agreement between the $Z/\gamma^* \to \tau^+\tau^-$ simulation and the embedded samples is observed in most of them. Figure C.8 shows the muon identification efficiency on the left, which is better in the embedded samples than in the simulation. The reason for this is that events with high pile-up and lower identification efficiency are already sorted out in the embedded sample due to the selection of the original di-muon events, as discussed in the previous paragraph. The overall event selection efficiency on the bottom right shows a slight trend for the PF embedding while the RH embedding is flat and closer to the simulation. This effect is even enhanced in the $\tau_e + \tau_{\text{had}}$ final state, which is discussed in the next section.

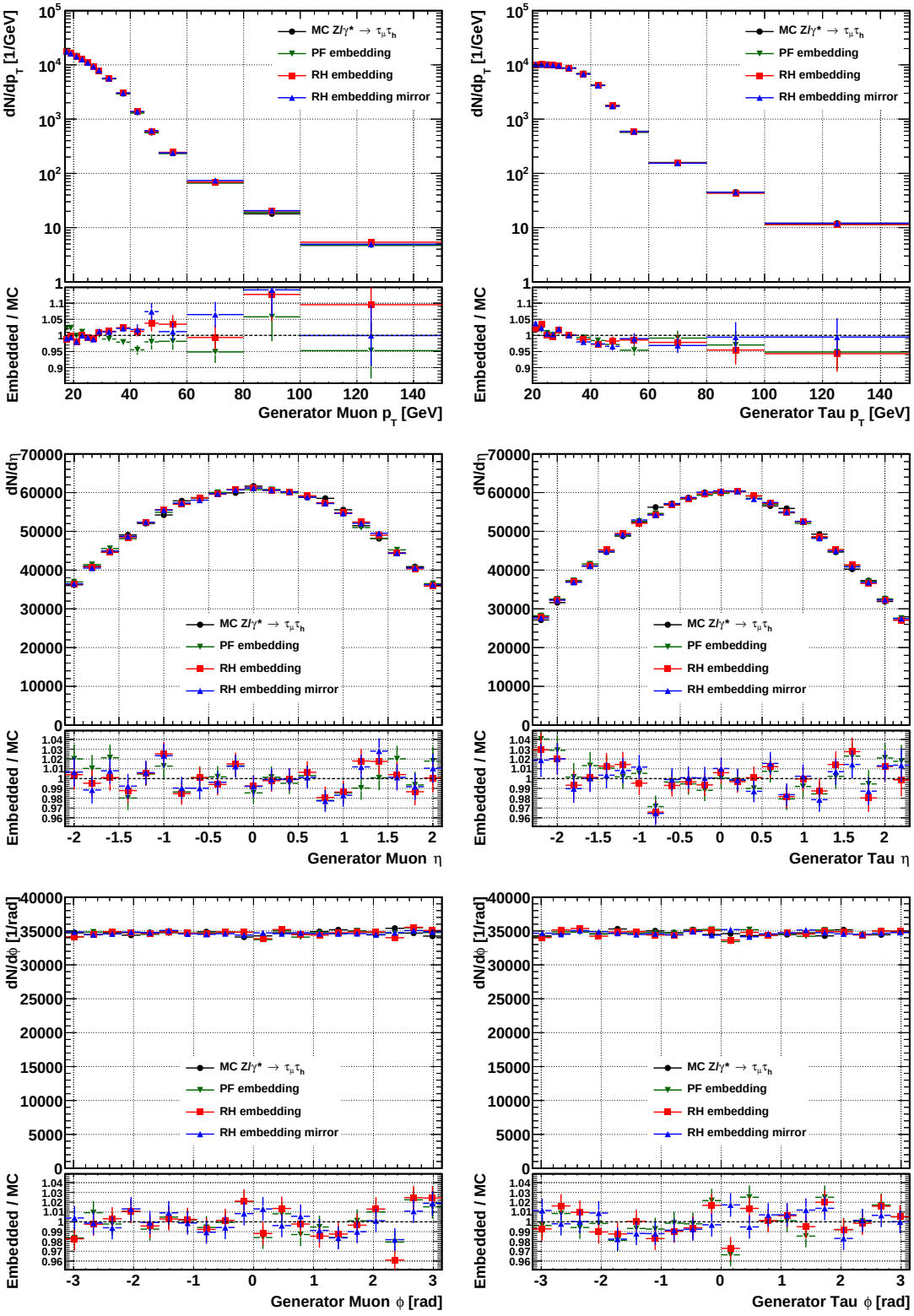

Figure C.4: τ_μ p_T (top left), η (center left) and ϕ (bottom left), and τ_had p_T (top right), η (center right) and ϕ (bottom right), on generator level.

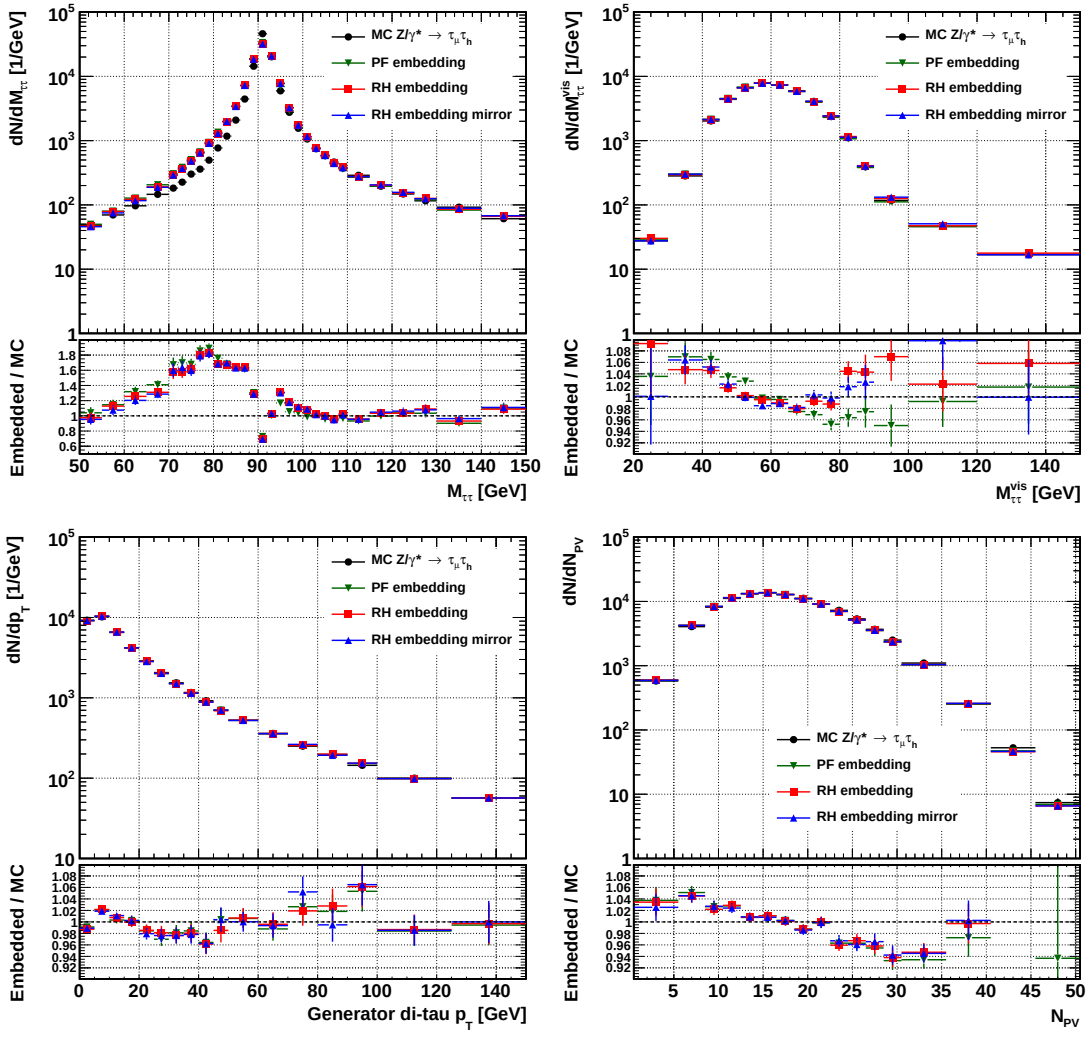

Figure C.5: Di-tau mass (top left), visible di-tau mass (top right), di-tau p_{T} (bottom left) and the number of reconstructed vertices (bottom right), on generator level.

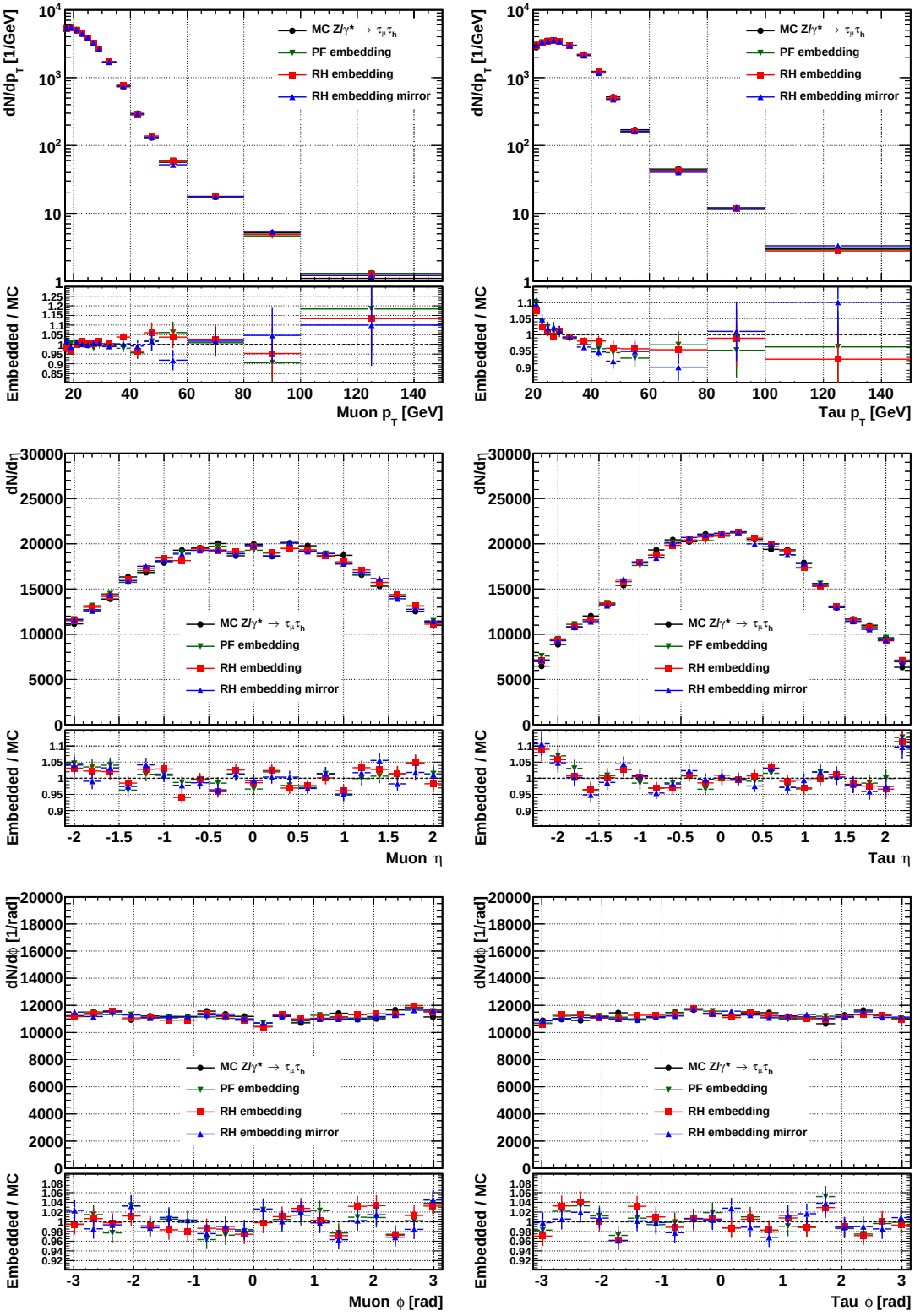

Figure C.6: Muon p_{T} (top left), η (center left) and ϕ (bottom left), and τ_{had} p_{T} (top right), η (center right) and ϕ (bottom right), on reconstruction level.

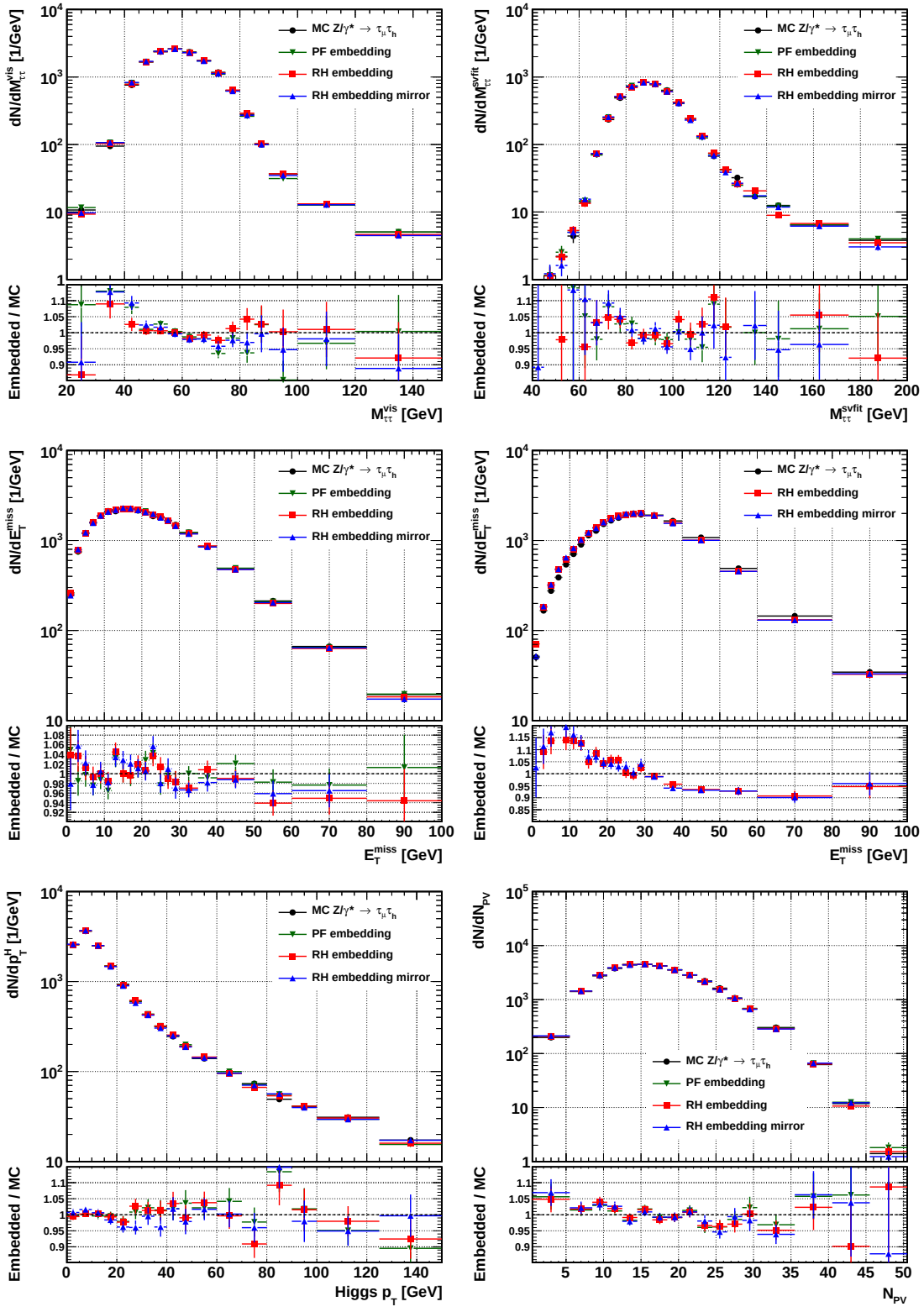

Figure C.7: Di-tau visible mass (top left), SVfit mass (top right), PF-based $E_{\mathrm{T}}^{\mathrm{miss}}$ (center left), Calo-based $E_{\mathrm{T}}^{\mathrm{miss}}$ (center right), Higgs p_{T} (bottom left) and the number of reconstructed vertices (bottom right) on reconstruction level.

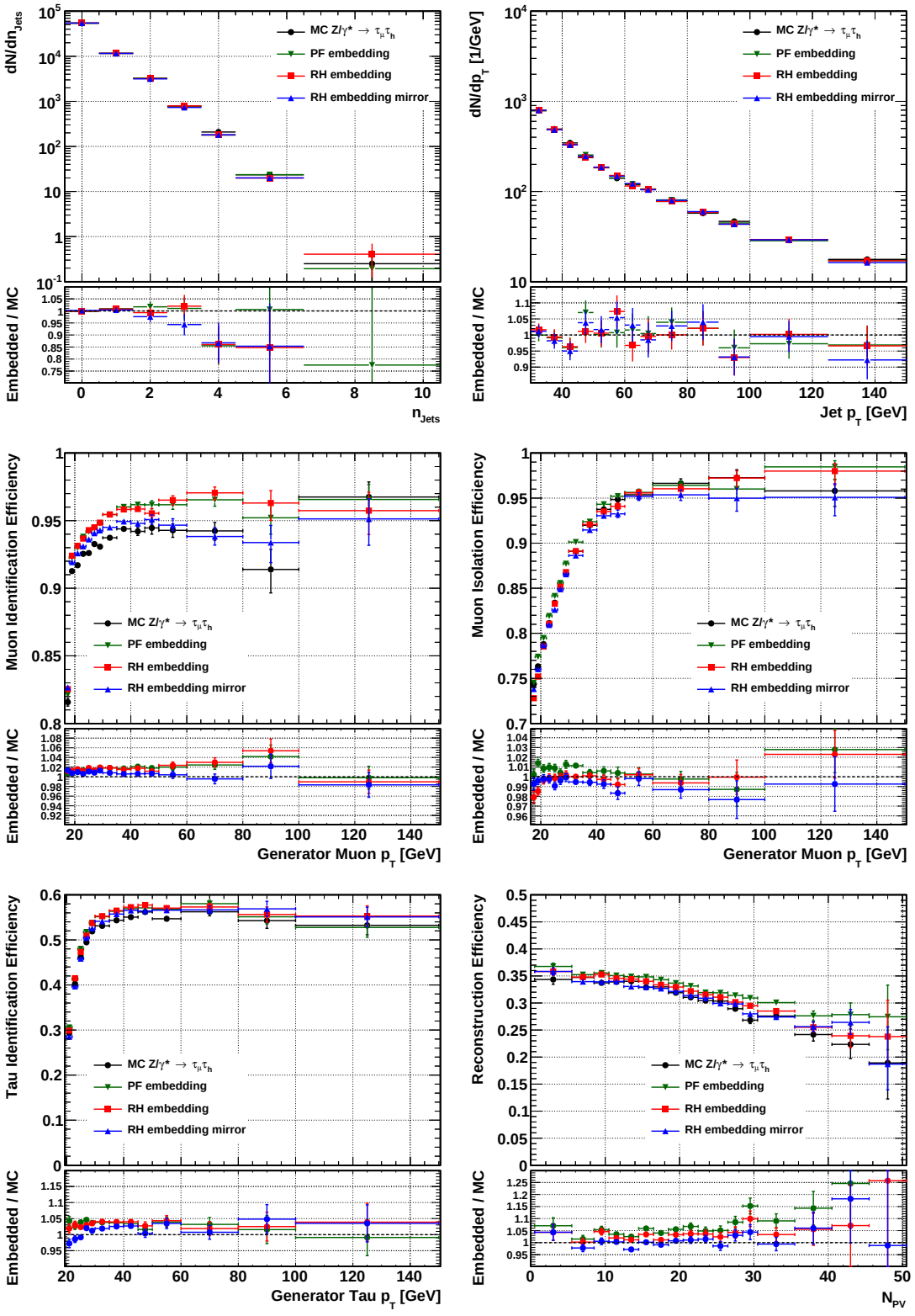

Figure C.8: Number of jets (top left), Leading jet p_{T} (top right), muon identification efficiency (center left), muon isolation efficiency (center right), τ_{had} identification efficiency (bottom left) and total event selection efficiency (bottom right).

C.2.2 The $\tau_e + \tau_{\text{had}}$ Final State

Figures C.9, C.10, C.11, C.12 and C.13 show the same distributions as in the $\tau_\mu + \tau_{\text{had}}$ case for the $\tau_e + \tau_{\text{had}}$ final state. The conclusions are very much the same. There is one additional effect that the electron identification is too efficient in the embedded samples at high η. This effect is discussed in the main text in Sections 6.3.2 and 6.4.3.

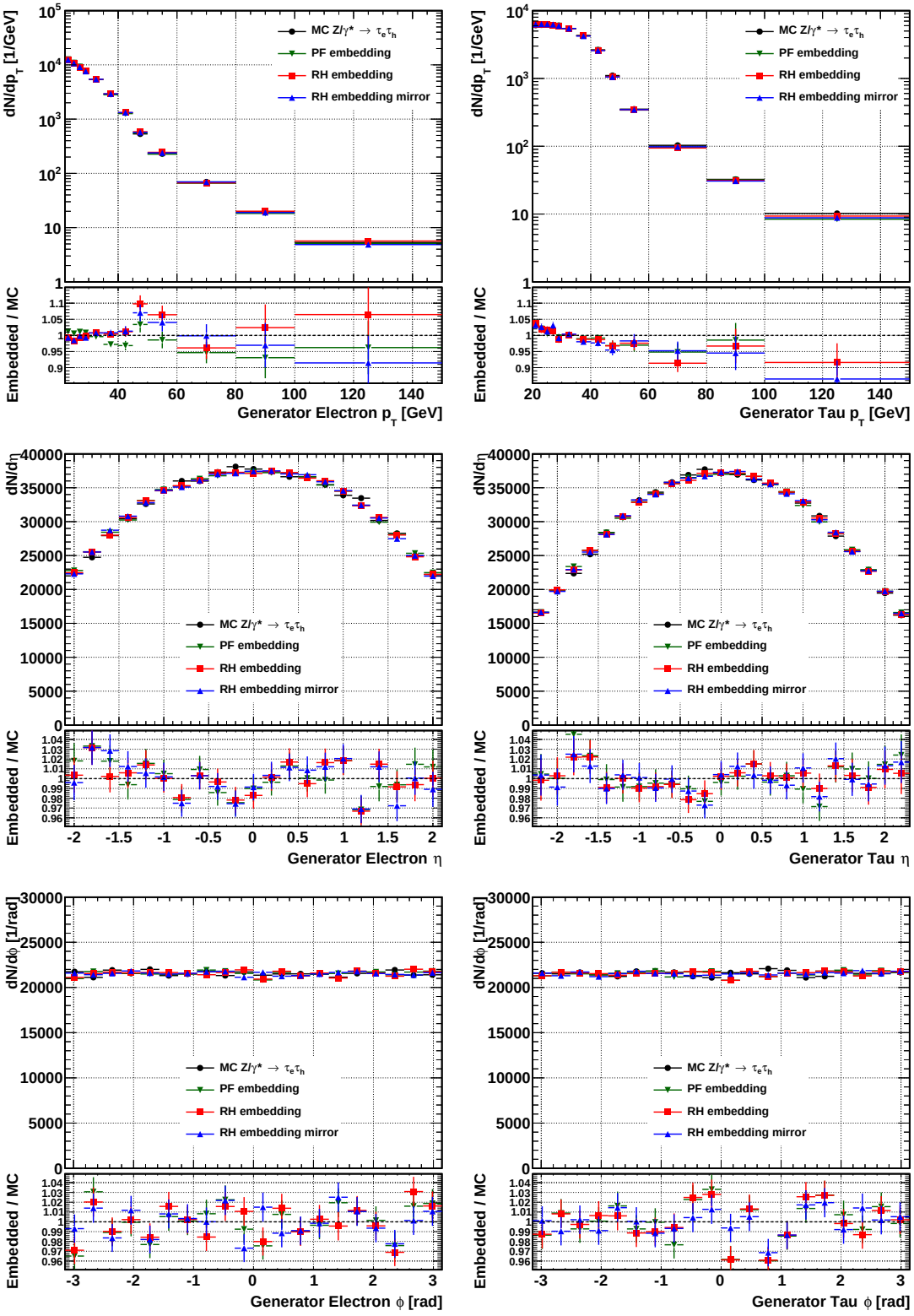

Figure C.9: τ_e p_{T} (top left), η (center left) and ϕ (bottom left), and τ_{had} p_{T} (top right), η (center right) and ϕ (bottom right), on generator level.

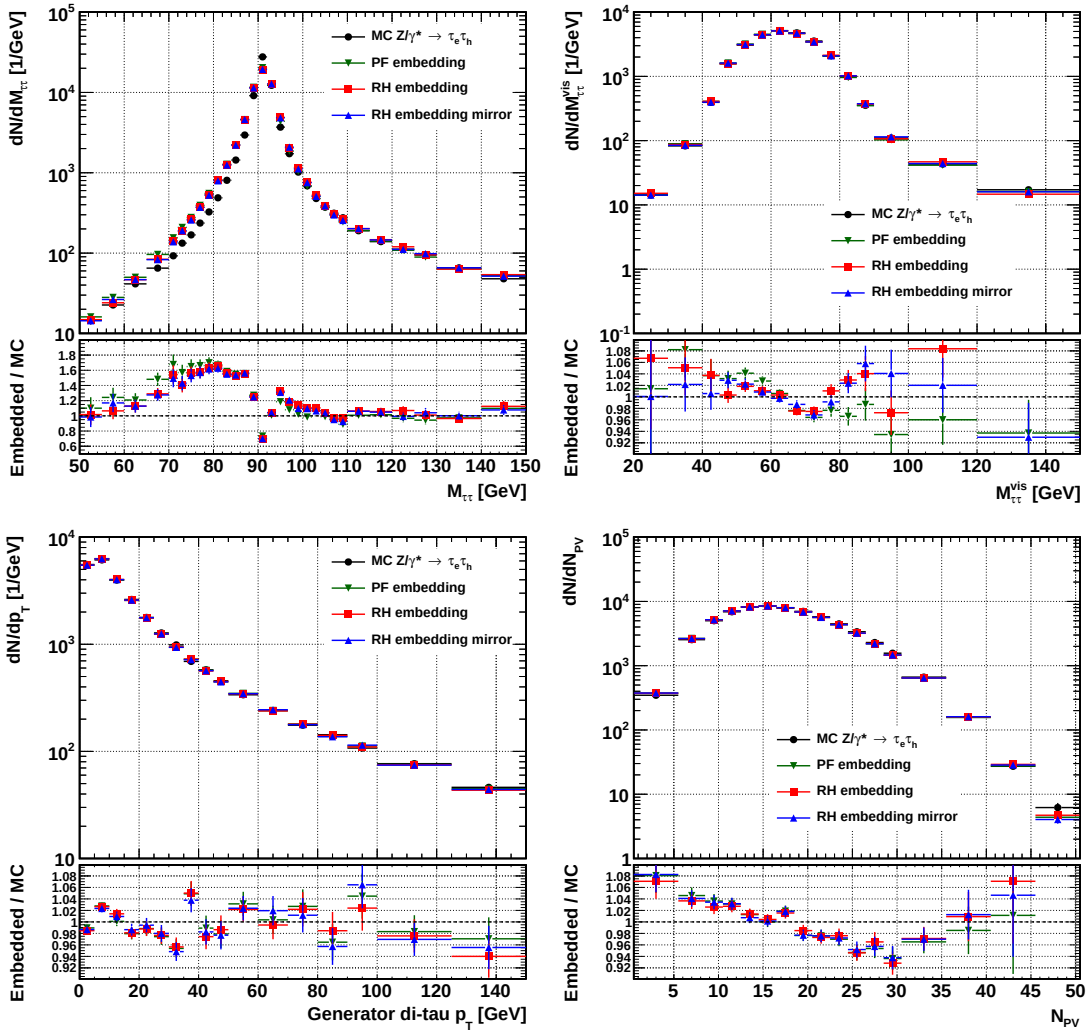

Figure C.10: Di-tau mass (top left), visible di-tau mass (top right), di-tau p_{T} (bottom left) and the number of reconstructed vertices (bottom right), on generator level.

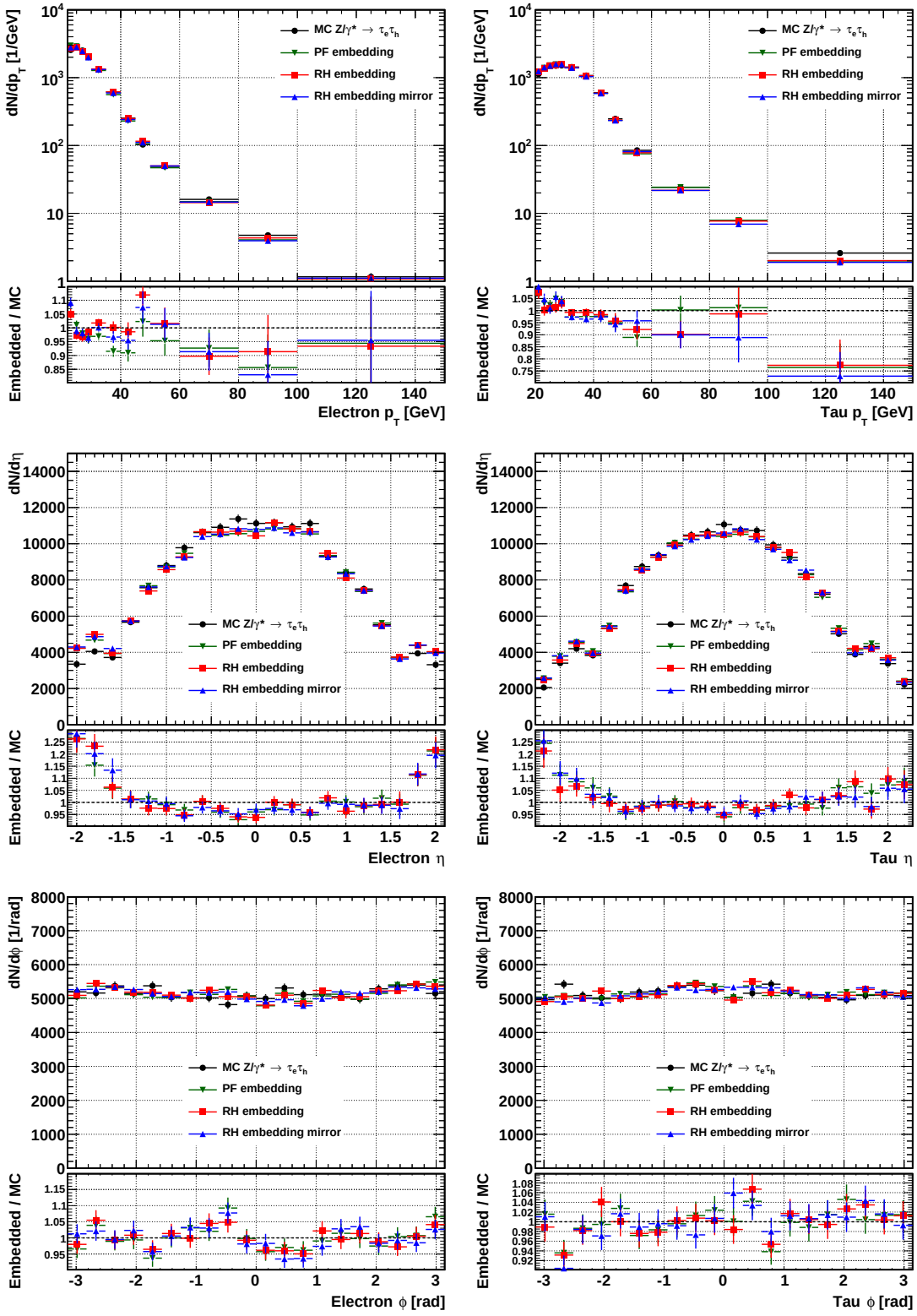

Figure C.11: Electron p_{T} (top left), η (center left) and ϕ (bottom left), and τ_{had} p_{T} (top right), η (center right) and ϕ (bottom right), on reconstruction level.

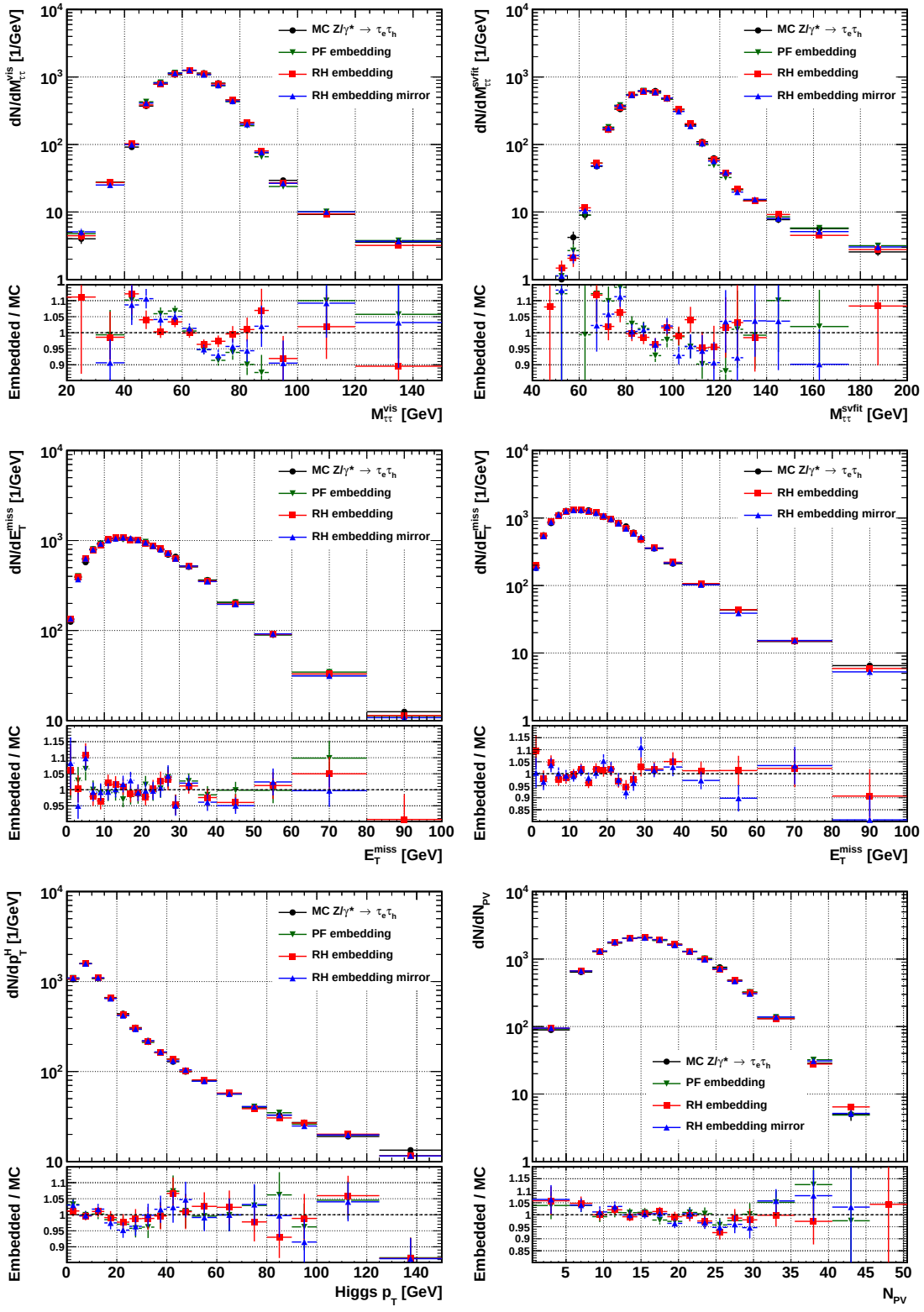

Figure C.12: Di-tau visible mass (top left), SVfit mass (top right), PF-based $E_{\mathrm{T}}^{\mathrm{miss}}$ (center left), Calo-based $E_{\mathrm{T}}^{\mathrm{miss}}$ (center right), Higgs p_{T} (bottom left) and the number of reconstructed vertices (bottom right) on reconstruction level.

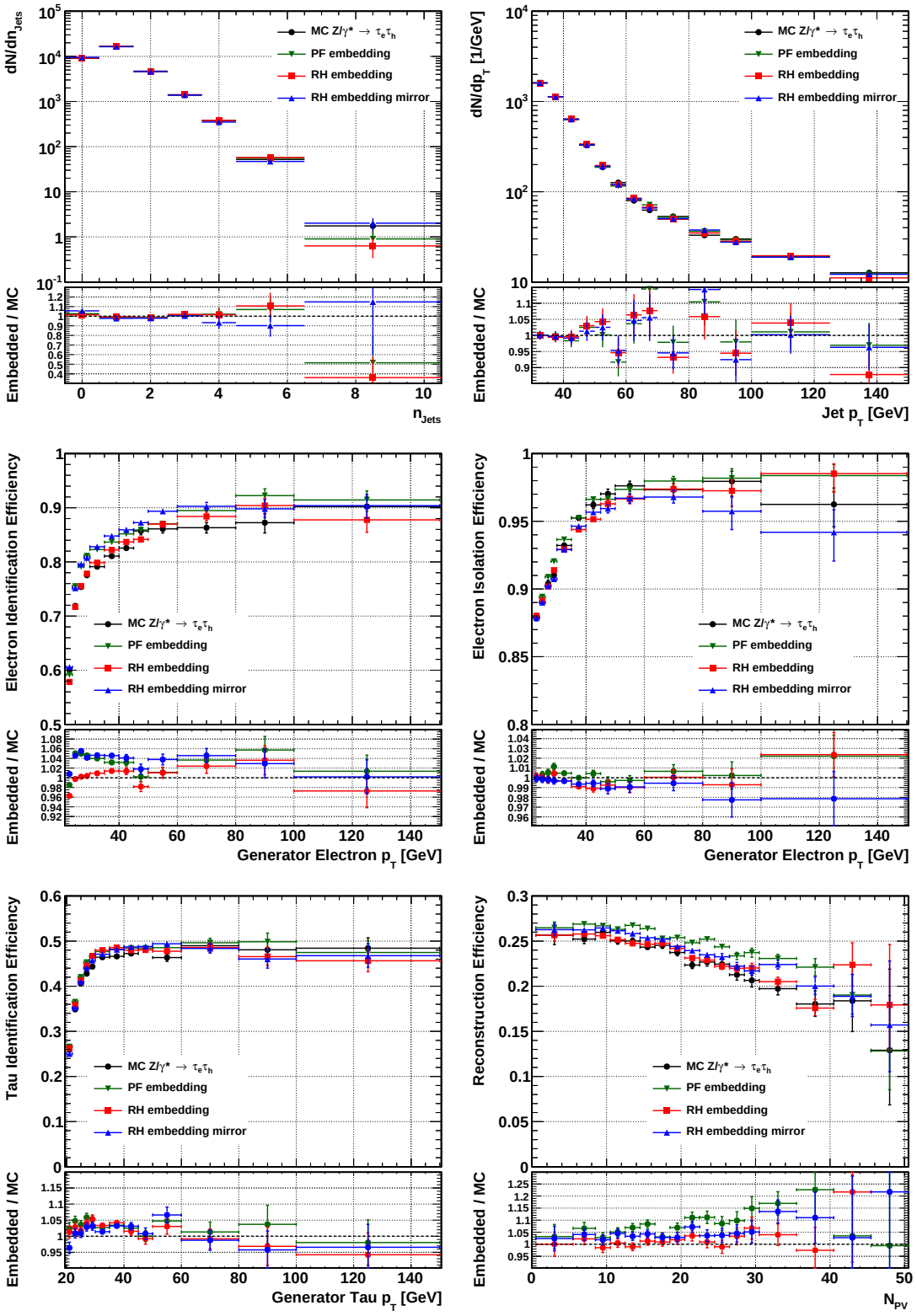

Figure C.13: Number of jets (top left), Leading jet p_{T} (top right), electron identification efficiency (center left), electron isolation efficiency (center right), τ_{had} identification efficiency (bottom left) and total event selection efficiency (bottom right).

C.3 Individual Effects in Spin Correlations

In Section 6.4.2, the modeling of the tau spin correlations in embedded samples with TAUSPINNER is discussed. In single pion decays, the z_s variable defined in [92] and the invariant mass of the two pions are sensitive to the spin correlations. Figures C.14 and C.15 show the two variables, comparing $Z/\gamma^* \to \tau^+\tau^-$ simulation where TAUOLA was used for the tau decays and an embedded sample for which the spin effects are modeled with TAUSPINNER. The top left plot corresponds to Figure 6.13 in the main text: the generator-level muons are replaced by tau leptons, no acceptance cuts on the original muons, and events in which one of the muons has an FSR photon on generator level are skipped. In this case, very good agreement is observed. The remaining plots then show how the modeling of the spin correlation behaves when lifting the three idealizations one-by-one.

On the top right plot, simple acceptance cuts are introduced on generator level: $p_T > 20\,\text{GeV}$ for the leading muon, $p_T > 10\,\text{GeV}$ for the subleading muon, and $\eta < 2.1$ for both muons. On the lower left plot, the reconstructed muons are replaced instead of the generator-level ones, and the weight for the di-muon selection efficiency discussed in Section 6.2.1 is applied. Finally, on the lower right plot, all di-muon events are taken, including the ones in which one or both of the original muons have radiated a photon. This last case corresponds to how events can be selected in the detector data, and therefore how the method would be applied in reality. These last two plots are reproduced from Figure 6.14 in the main text.

The two figures show that the acceptance cuts introduce a very small bias in the z_s distribution, but the other two effects do not significantly alter the shapes.

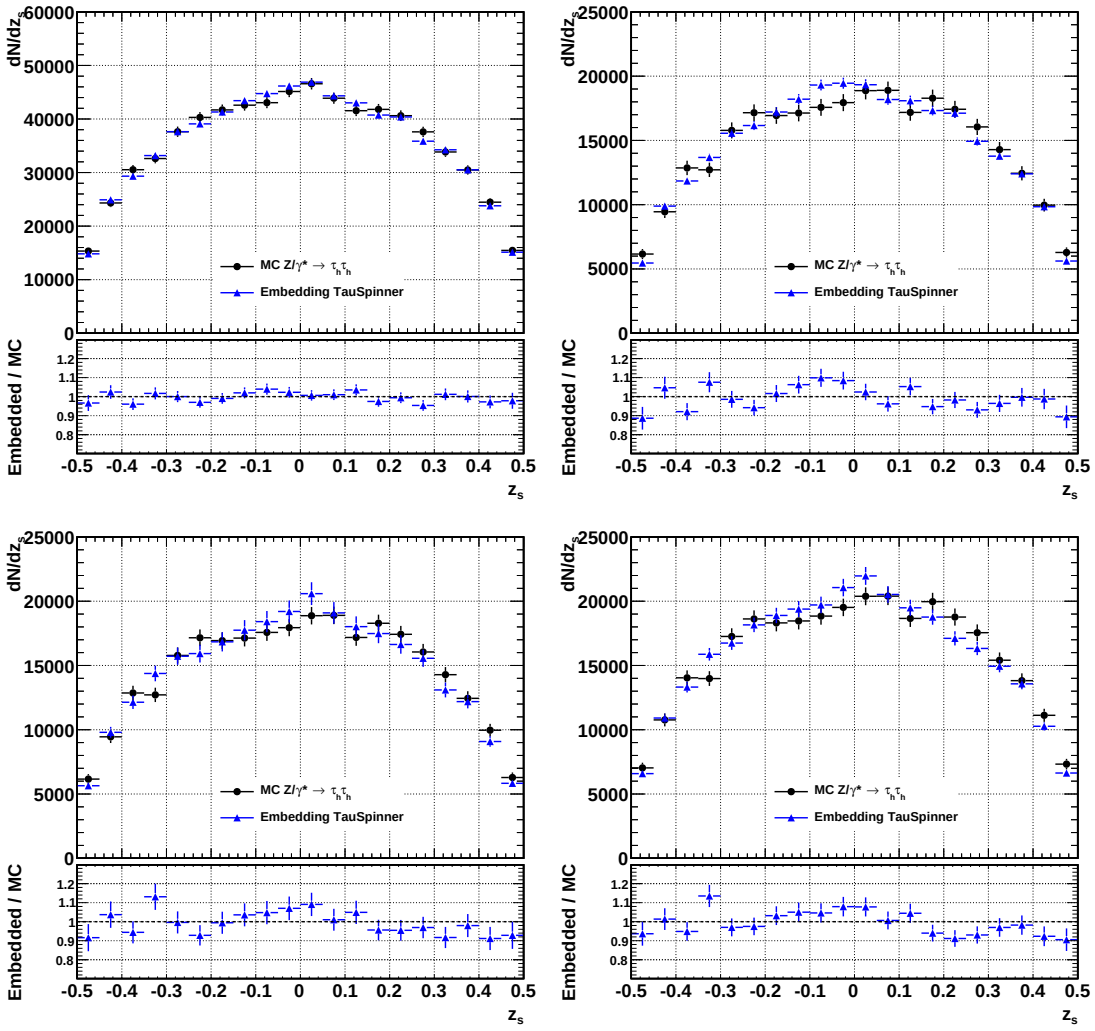

Figure C.14: Comparison of the z_s variable in direct Monte Carlo simulation and the embedded sample. The four plots correspond to four different configurations where the top left plot is an idealized environment on generator level and the bottom right corresponds to what one would see in the detector data. See the text for details.

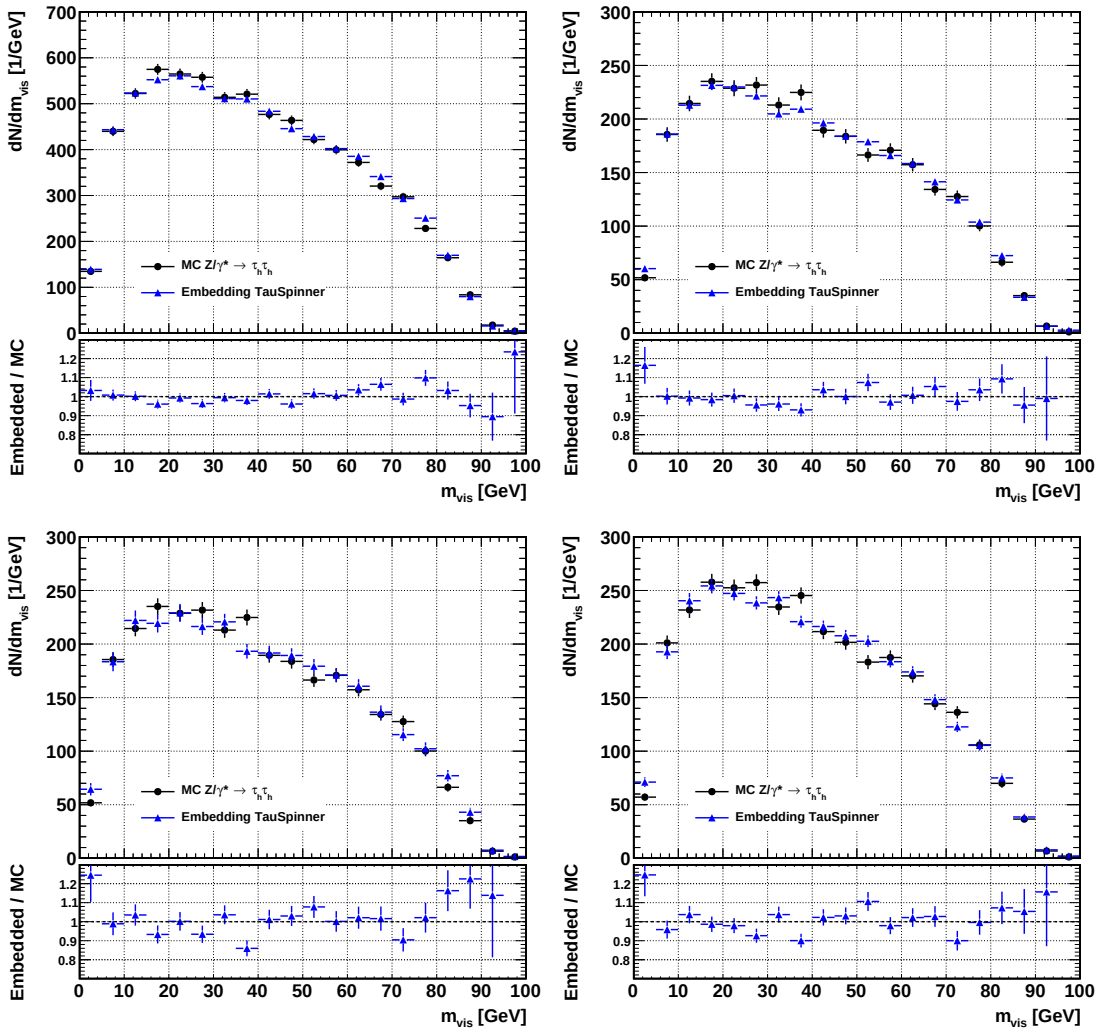

Figure C.15: Comparison of the invariant di-pion mass in direct Monte Carlo simulation and the embedded sample. The four plots correspond to four different configurations where the top left plot is an idealized environment on generator level and the bottom right corresponds to what one would see in the detector data. See the text for details.

C.4 Z Decay Invariance under Mirror Transformation

In this section, the invariance of the Z boson decay under the mirror transformation described in Section 6.4.4 is shown using MADGRAPH $Z/\gamma^* \to \mu^+\mu^-$ simulation. For this purpose, no acceptance cuts have been performed. Figure C.16 shows the azimuthal angle of the positive muon around the Z boson axis in the Z rest frame on generator level, where $\phi = 0$ is taken to be the axis of the incoming proton beam. The mirror operation then corresponds to the transformation $\phi \to -\phi$.

The top left plot is made inclusively with the whole sample. A slight modulation in the azimuthal angle can be seen which comes from the polarization of the Z boson. In a $H \to \mu^+\mu^-$ Monte Carlo sample, this plot was confirmed to be flat. Due to the polarization, an arbitrary rotation around ϕ cannot be performed while not altering the Z decay, however, it can be seen that the mirror transformation is working. The other three plots correspond to three different regions of phase space: $0.5 < \eta^\mu < 1.5$ (top right), $p_{\mathrm{T}}^{\mu} > 40\,\mathrm{GeV}$ (bottom left) and $p_{\mathrm{T}}^{Z} > 40\,\mathrm{GeV}$ (bottom right). While the exact magnitude of the ϕ modulation changes, in all cases the distribution is invariant under the mirror transformation.

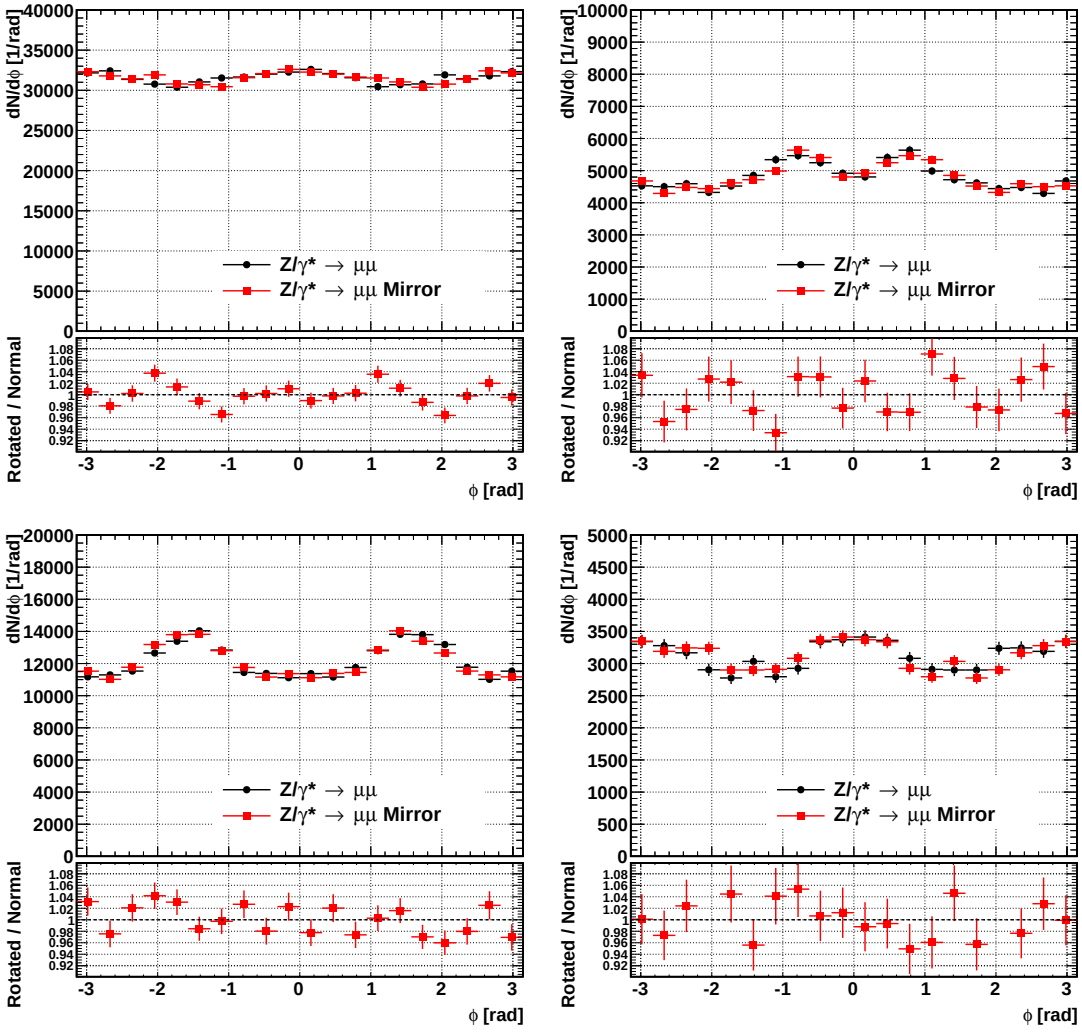

Figure C.16: The azimuthal angle of the positive muon around the Z boson axis in the Z rest frame. The different plots correspond to different regions of phase space. Top left: inclusive sample, top right: $0.5 < \eta^{\mu} < 1.5$, bottom left: $p_{\mathrm{T}}^{\mu} > 40\,\mathrm{GeV}$, bottom right: $p_{\mathrm{T}}^{Z} > 40\,\mathrm{GeV}$.

D Supporting Material for the WH Analysis

D.1 Fake Rate Measurement

In this section, additional studies to the measured fake rates are presented.

D.1.1 Measured Fake Rates

All fake rate functions used for the background estimation in thee analysis are presented here. The fake rates are measured separately in the 7 TeV and 8 TeV datasets, and also separately for the electron and the muon channel. In all cases, a Landau distribution with the most probable value and the width with an additional constant as free parameters are fitted to the distribution.

Figures D.1, D.2, D.3 and D.4 show the 24 measurements for the two channel in the two data taking periods and three pseudorapidity regions.

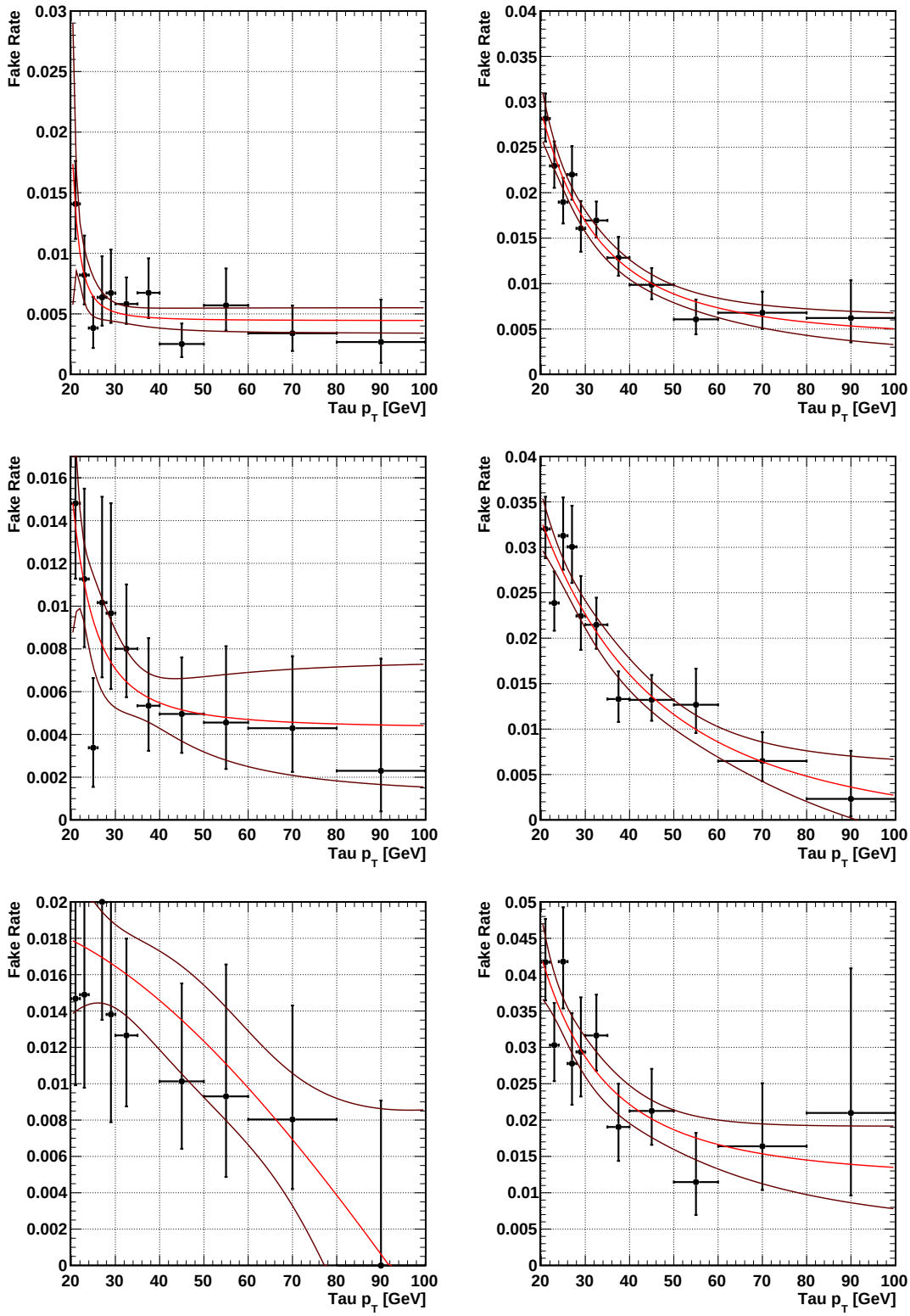

Figure D.1: Fake Rate functions in the electron channel measured in the W + jets enriched region (left) and the Z + jets enriched region (right) in the 7 TeV dataset. The top row shows the eta region $|\eta| < 0.8$, the center row $0.8 < |\eta| < 1.6$ and the bottom row $|\eta| > 1.6$.

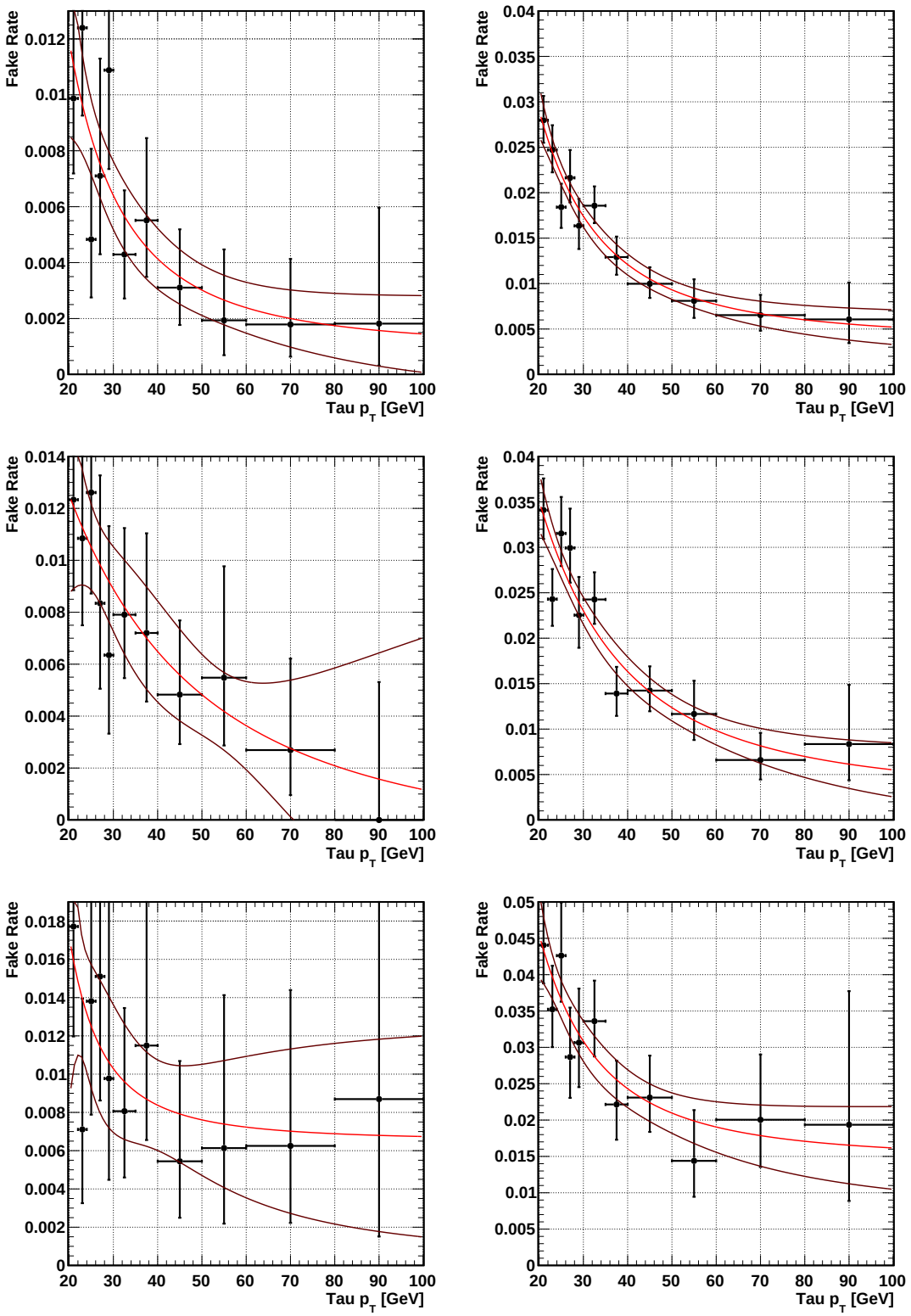

Figure D.2: Fake Rate functions in the muon channel measured in the W + jets enriched region (left) and the Z + jets enriched region (right) in the 7 TeV dataset. The top row shows the eta region $|\eta| < 0.8$, the center row $0.8 < |\eta| < 1.6$ and the bottom row $|\eta| > 1.6$.

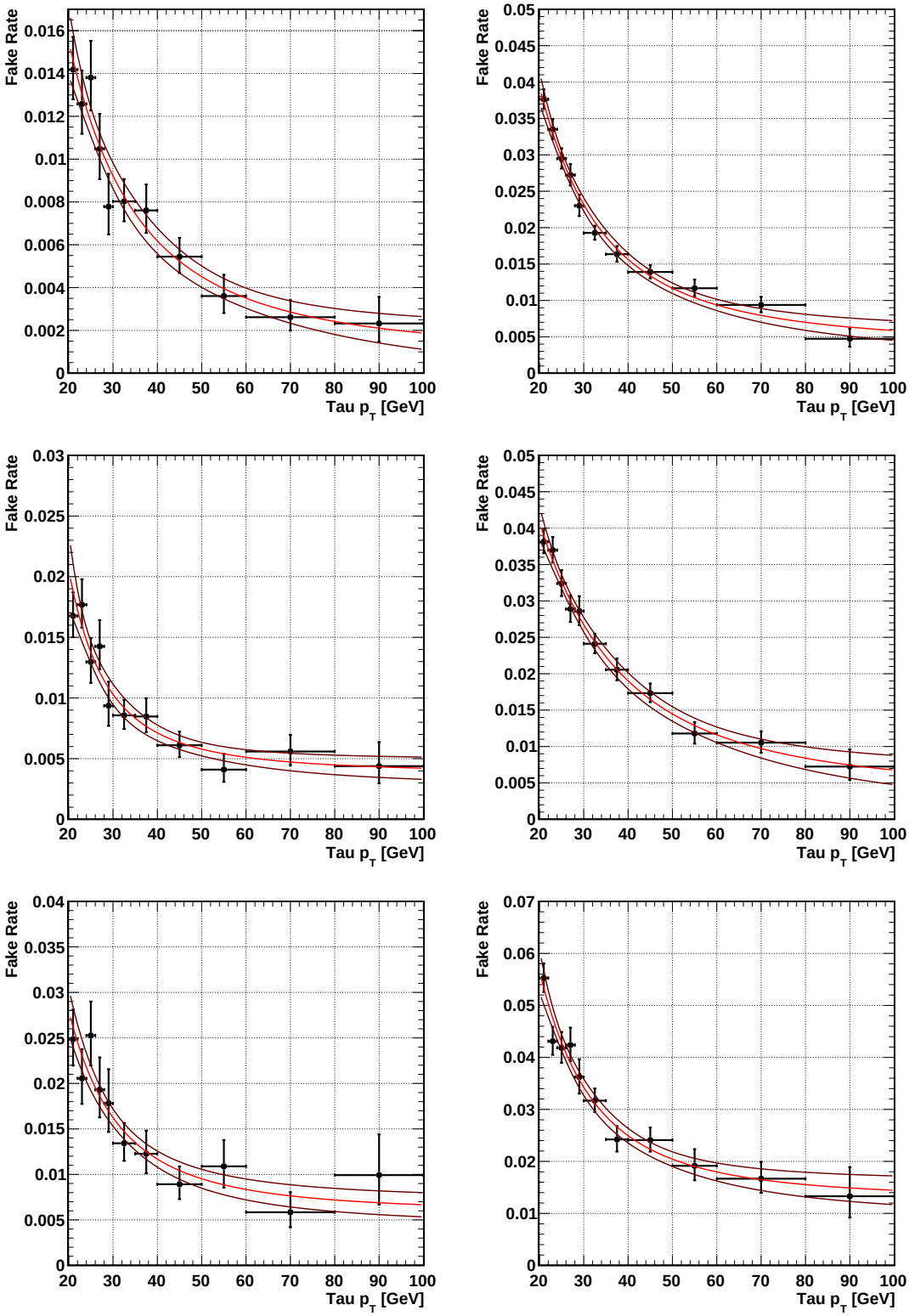

Figure D.3: Fake Rate functions in the electron channel measured in the W + jets enriched region (left) and the Z + jets enriched region (right) in the 8 TeV dataset. The top row shows the eta region $|\eta| < 0.8$, the center row $0.8 < |\eta| < 1.6$ and the bottom row $|\eta| > 1.6$.

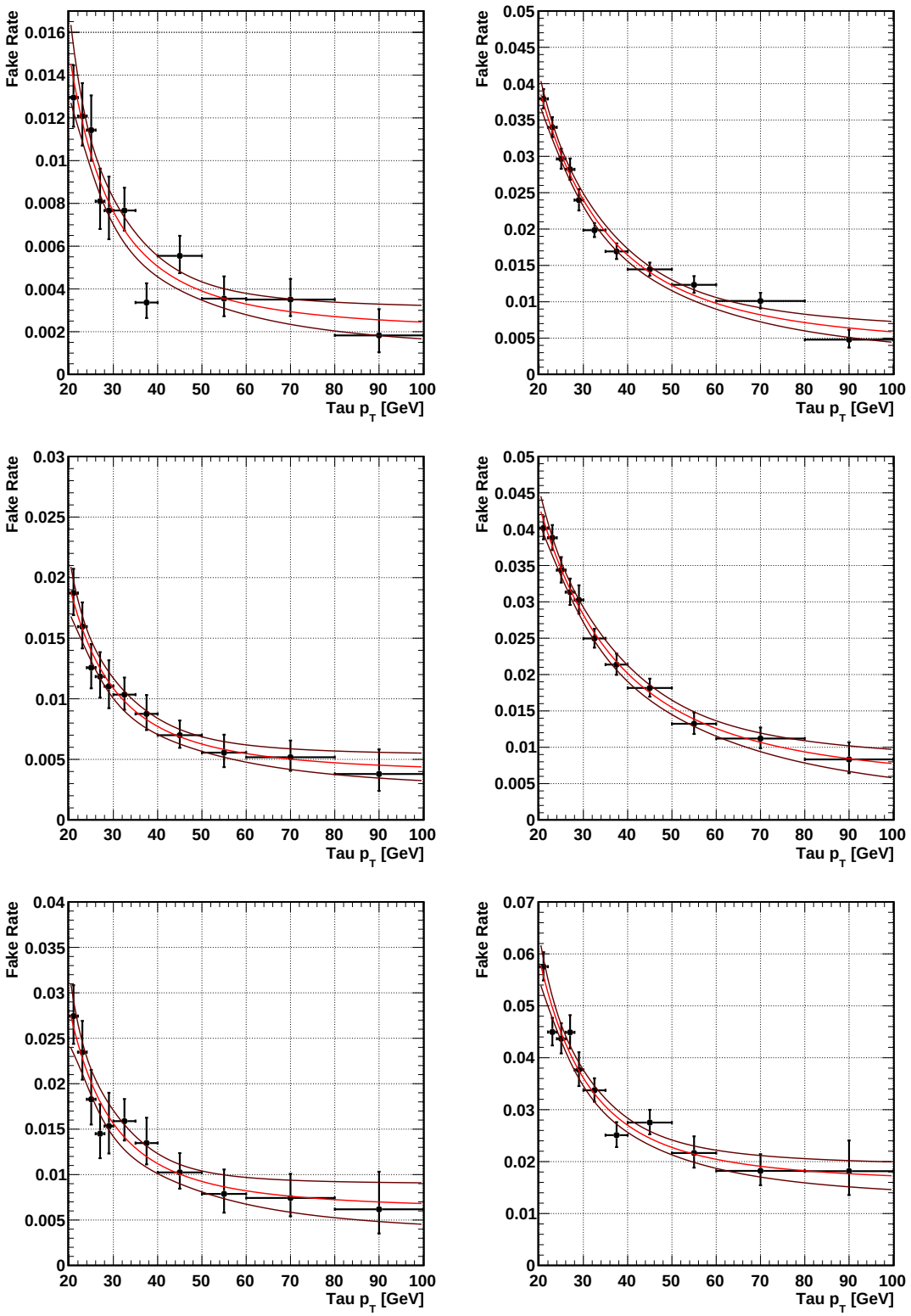

Figure D.4: Fake Rate functions in the muon channel measured in the W + jets enriched region (left) and the Z + jets enriched region (right) in the 8 TeV dataset. The top row shows the eta region $|\eta| < 0.8$, the center row $0.8 < |\eta| < 1.6$ and the bottom row $|\eta| > 1.6$.

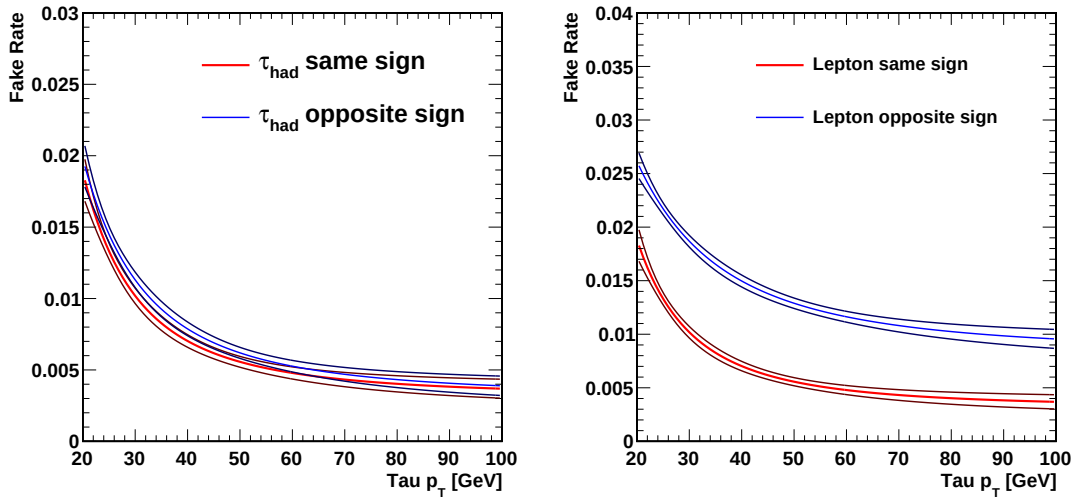

Figure D.5: Comparison of the fake rates in the W + jets enriched region for the muon channel in the 8 TeV dataset. Left: The two τ_{had} candidates same sign compared to the two τ_{had} candidates opposite sign. Right: The lepton from the W boson same sign as the two τ_{had} candidates compared to the lepton from the W boson opposite sign.

D.1.2 Additional Studies

In this section, additional studies with respect to the fake rate measurement are performed. All these studies are shown for the muon channel with the 8 TeV dataset and the three pseudorapidity regions combined.

Relative Charge

In the W + jets enriched region, there are two τ_{had} candidates required, as discussed in Section 7.2. The two τ_{had} candidates must have the same charge, in order to avoid overlapping events with the signal region. It is therefore crucial to show that the fake rate is not different for two opposite sign τ_{had} candidates. Figure D.5 compares these two fake rates on the left hand side, and good agreement within the statistical uncertainty is observed. In the analysis, the τ_{had} candidate which has the same charge as the lepton from the W boson is used for the background estimation. Therefore, for the fake rate measurement, the same requirement is made. The right hand side compares the fake rate with the case when the two τ_{had} candidates are opposite sign to the lepton from the W boson decay. It is shown that the fake rate heavily depends on the relative charge of the lepton and the τ_{had} candidates.

Jet Multiplicity

It can be seen on the left hand side in Figure D.6 that the fake rate in the W + jets enriched region is significantly smaller than the one in the Z + jets enriched region. The reason for this is twofold: first, the difference coming from the charge of the W boson with respect to the τ_{had} candidates as discussed in the previous paragraph, and second the different jet multiplicity in the two cases. On the right hand side, these two differences are

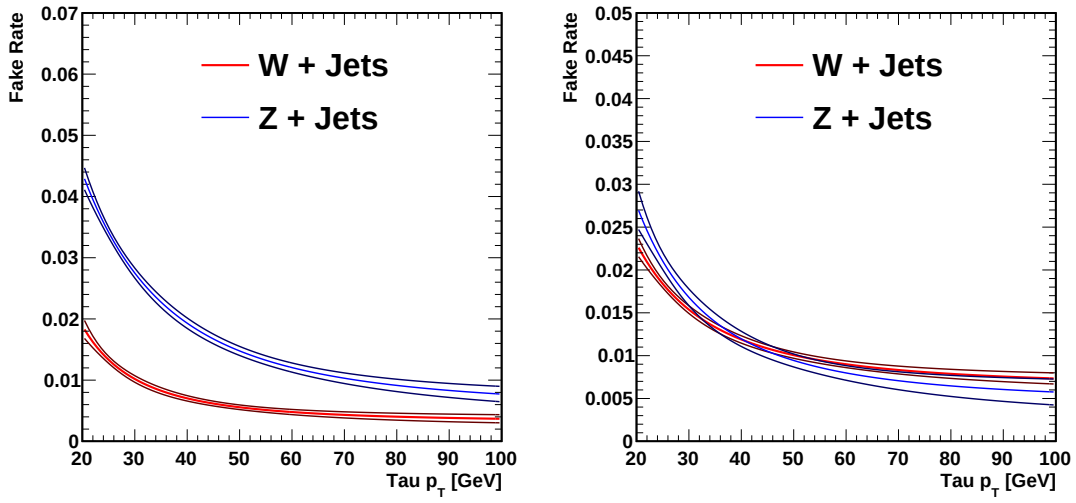

Figure D.6: Comparison of the fake rates in the W + jets enriched region and the Z + jets enriched region for the muon channel in the 8 TeV dataset. Left: The fake rates as used in the analysis. Right: For the W + jets fake rate, both opposite-sign and same-sign configurations are allowed for the lepton from the W decay and the τ_{had} candidates. For the Z + jets fake rate, at least two τ_{had} candidates with $p_{\text{T}} > 20$ GeV are required.

Table D.1: Gluon fractions in the MADGRAPH Monte Carlo simulation for the measured fake rates.

Configuration	Gluon Fraction
W + 2 jets, W and τ_{had} same sign	39 %
W + 2 jets, no sign requirement	35 %
Z + 1 jet	23 %
Z + 2 jets	36 %

removed by accepting events with both opposite-sign and same-sign W bosons in the W + jets case, and requiring at least two τ_{had} candidates in the Z + jets case, as it is already in the W + jets region. In this case, the two fake rates are comparable.

The difference is caused by a different fraction of quark-induced jets and gluon-induced jets in the sample. In general, quark-induced jets have a higher fake rate than gluon-induced ones. This hypothesis was confirmed with Monte Carlo simulation, where the highest-p_{T} parton (quark or gluon) in a cone of $\Delta R < 0.3$ around the jet was studied on generator level. The gluon fractions found are listed in Table D.1. As can be seen, after the modification to the W + jets region and the Z + jets regions to bring the fake rates into agreement, also the gluon fractions are very similar.

The difference in the fake rate as a function of the number of jets is illustrated more clearly in Figure D.7. The number of jets is given by the number of reconstructed τ_{had} candidates with $p_{\text{T}} > 20$ GeV.

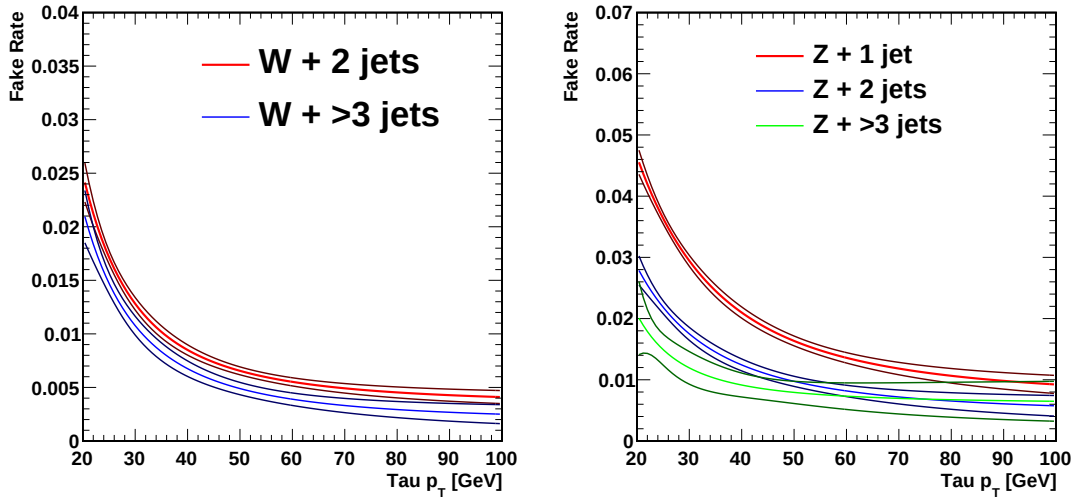

Figure D.7: Comparison of the fake rates in the W + jets enriched region and the Z + jets enriched region for the muon channel in the 8 TeV dataset, as a function of the number of tau-like jets in the event. Left: W + jets enriched region. Right: Z + jets enriched region.

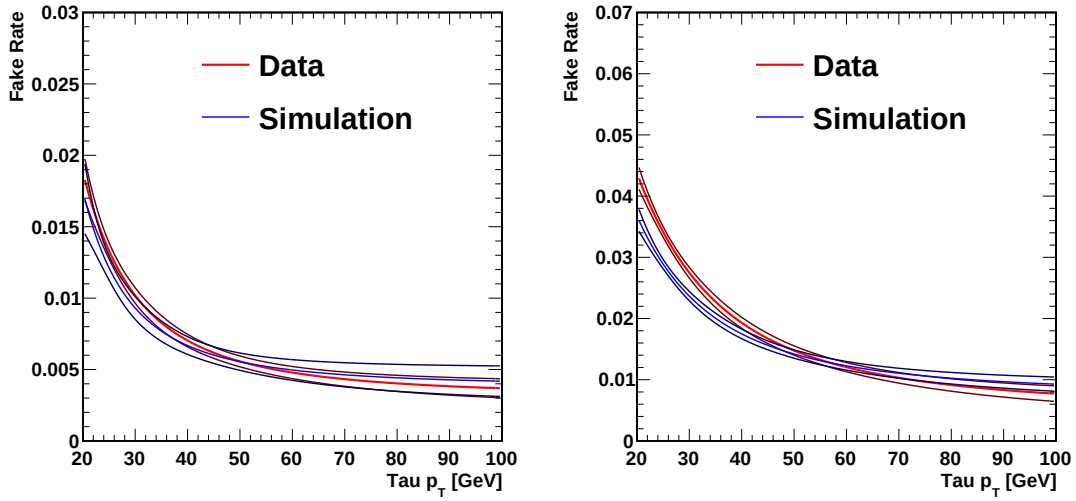

Figure D.8: Comparison of the fake rates in the W + jets enriched region and the Z + jets enriched region for the muon channel in the 8 TeV dataset. The two plots compare the agreement in the data and in MADGRAPH Monte Carlo simulation. Left: W + jets enriched region. Right: Z + jets enriched region.

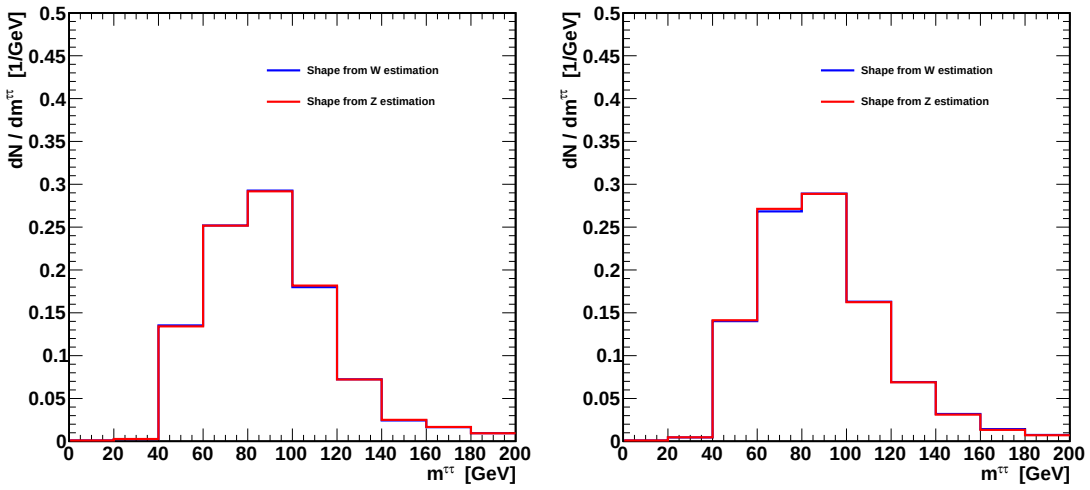

Figure D.9: Comparison of the shape of the visible mass distribution of the reducible background, when estimated only with the fake rate measured in the W + jets enriched region (blue) or the Z + jets enriched region (red). No significant difference is observed. Left: Electron channel. Right: Muon channel.

Comparison of Data and Simulation

Figure D.8 shows a comparison between the fake rates measured in data and MADGRAPH Monte Carlo simulation in the two measurement regions. Reasonable agreement is observed. This test also confirms that no major source of real tau leptons is present in the data which could bias the measurement towards larger fake rates.

Visible Mass Shape from the two Measurement Regions

Figure D.9 shows the shape of the visible mass distribution of the reducible background, normalized to unit area. A comparison is made between the fake rate obtained in the W + jets enriched region and the Z + jets enriched region. No significant difference is observed, justifying the choice of using an uncertainty on the total event yield due to the imprecise knowledge of the reducible background composition.

D.2 Reducible Background Composition

Figures D.10, D.11 and D.12 show the composition of the reducible background, compared to the data. The region where the isolation of the SS τ_{had} candidate is inverted is shown. These events are used for the background estimation after being weighted with the fake rate. The difference between the simulation and the data is taken to be due to QCD multijet processes. Figure D.10 shows the distribution of the BDT discriminant, Figure D.11 that of the visible di-tau mass after all cuts except the BDT discriminant, and Figure D.12 shows the visible di-tau mass after all cuts including the BDT discriminant. The numbers in Tables 7.3 and 7.4 in the main text are taken from the integral of these distributions.

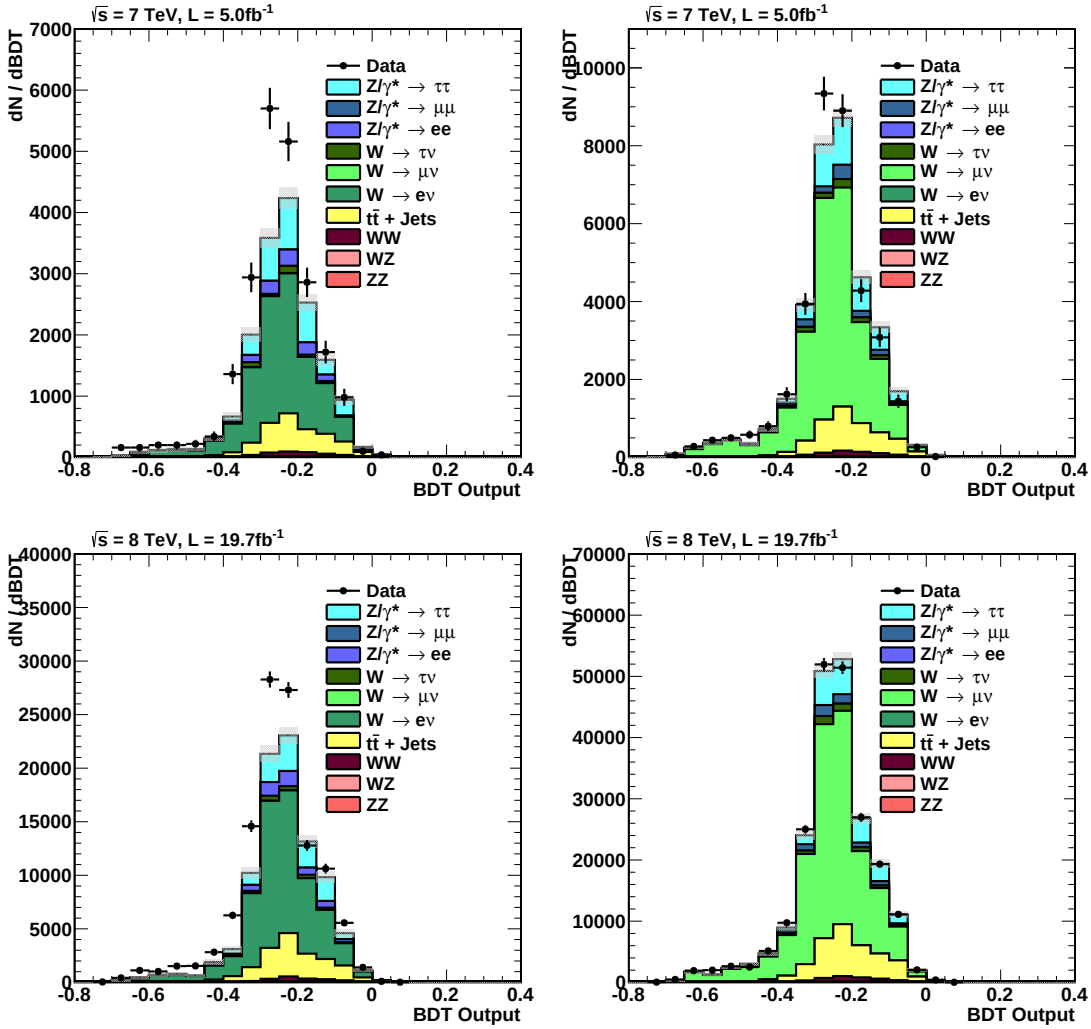

Figure D.10: BDT distribution in the region where the SS τ_{had} candidate is inverted. The data is compared to the MC simulation. The difference between the data and the simulation is due to QCD multijet processes which are not included in the simulation. The four plots show the electron channel on the left and the muon channel on the right, and the 7 TeV dataset on the top and the 8 TeV dataset on the bottom.

D.3 Distributions of BDT Variables

In this section, the distributions of the input and output variables of the BDT discriminant are shown. Figures D.13 shows the variables for the electron channel in the 7 TeV dataset, Figure D.14 for the muon channel in the 7 TeV dataset, Figure D.15 for the electron channel in the 8 TeV dataset, and Figure D.16 for the muon channel in the 8 TeV dataset. The p_{T} of the leading τ_{had} candidate is shown in the top left, the p_{T} of the subleading τ_{had} candidate on the top right, the $E_{\text{T}}^{\text{miss}}$ in the event in the center left, the ΔR between the two τ_{had} candidates in the center right, and the ratio of the vectorial p_{T} sum over the scaral p_{T} sum of the two τ_{had} candidates on the bottom left. The output of the BDT

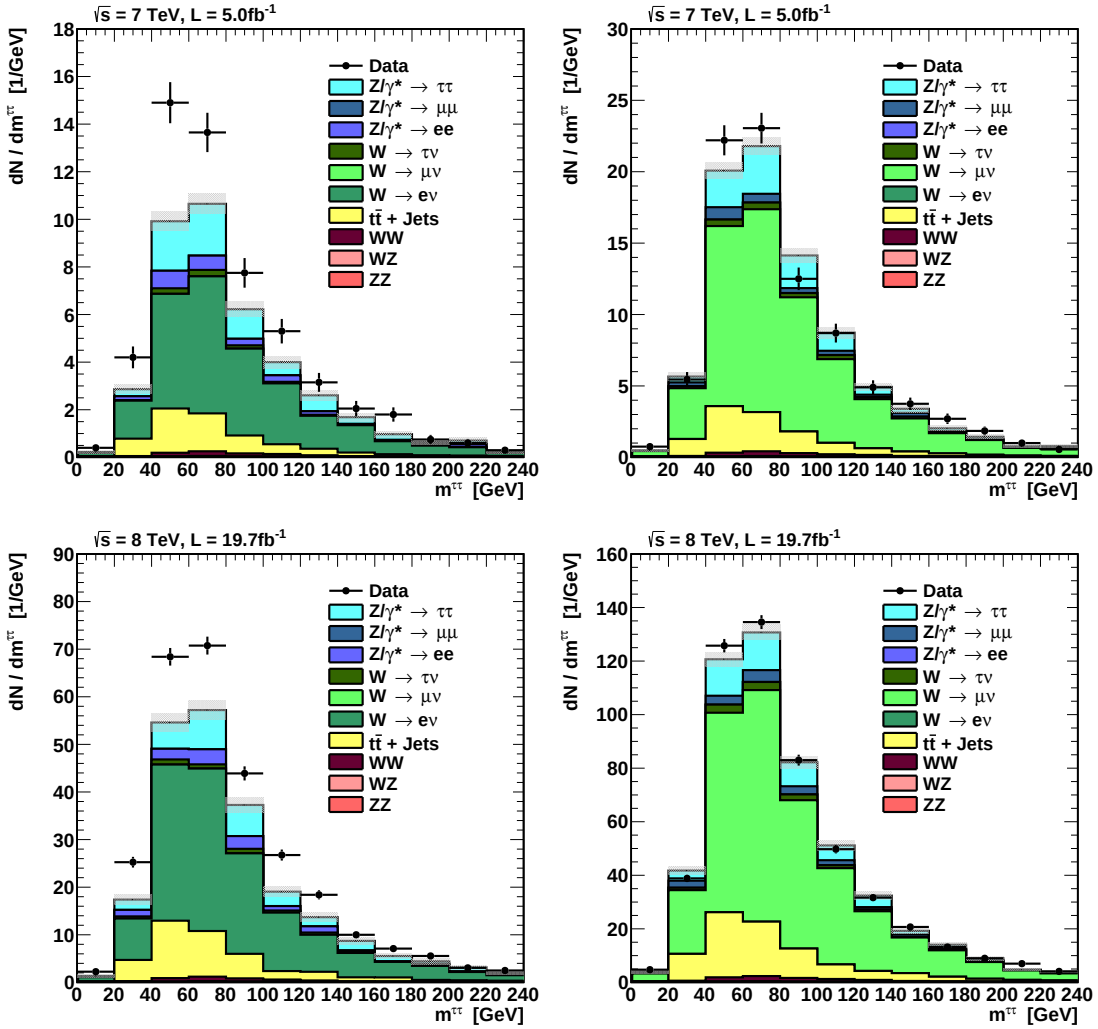

Figure D.11: Visible di-tau mass distribution before the cut on the BDT discriminant in the region where the SS τ_{had} candidate is inverted. The data is compared to the MC simulation. The difference between the data and the simulation is due to QCD multijet processes which are not included in the simulation. The four plots show the electron channel on the left and the muon channel on the right, and the 7 TeV dataset on the top and the 8 TeV dataset on the bottom.

discriminant is shown on the bottom right.

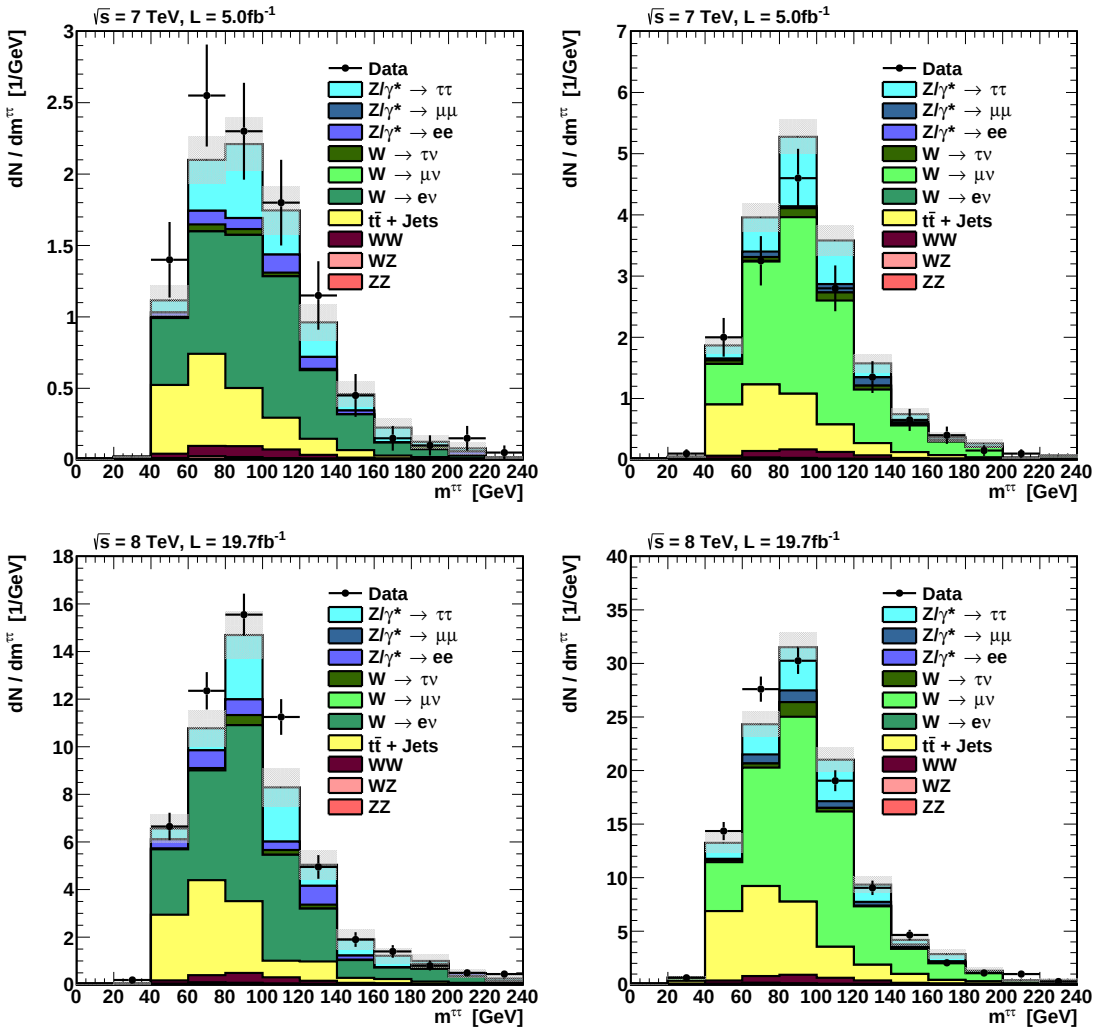

Figure D.12: Visible di-tau mass distribution before the cut on the BDT discriminant in the region where the SS τ_{had} candidate is inverted. The data is compared to the MC simulation. The difference between the data and the simulation is due to QCD multijet processes which are not included in the simulation. The four plots show the electron channel on the left and the muon channel on the right, and the 7 TeV dataset on the top and the 8 TeV dataset on the bottom.

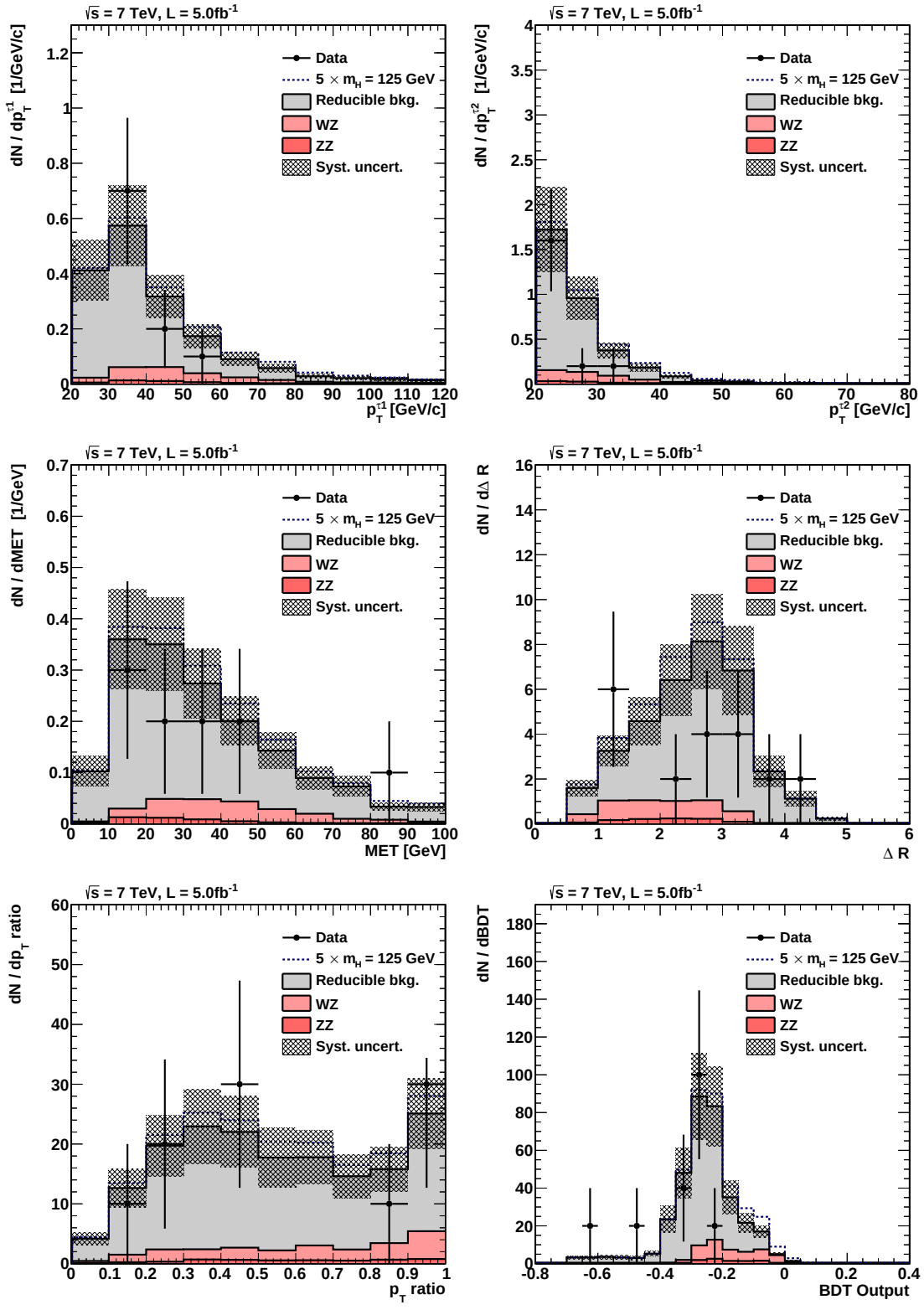

Figure D.13: BDT input and output Variables for the electron channel in the 7 TeV dataset.

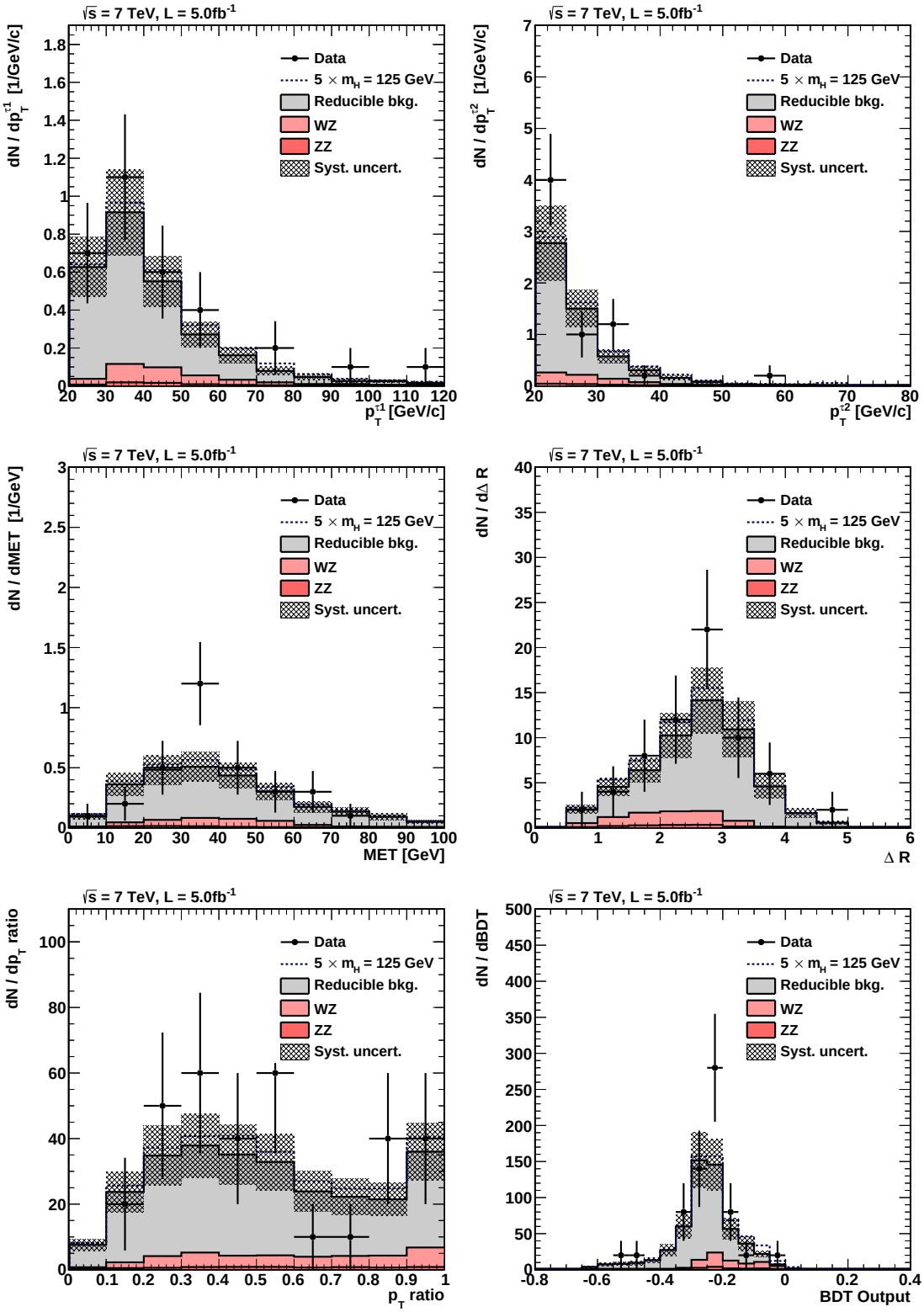

Figure D.14: BDT input and output Variables for the muon channel in the 7 TeV dataset.

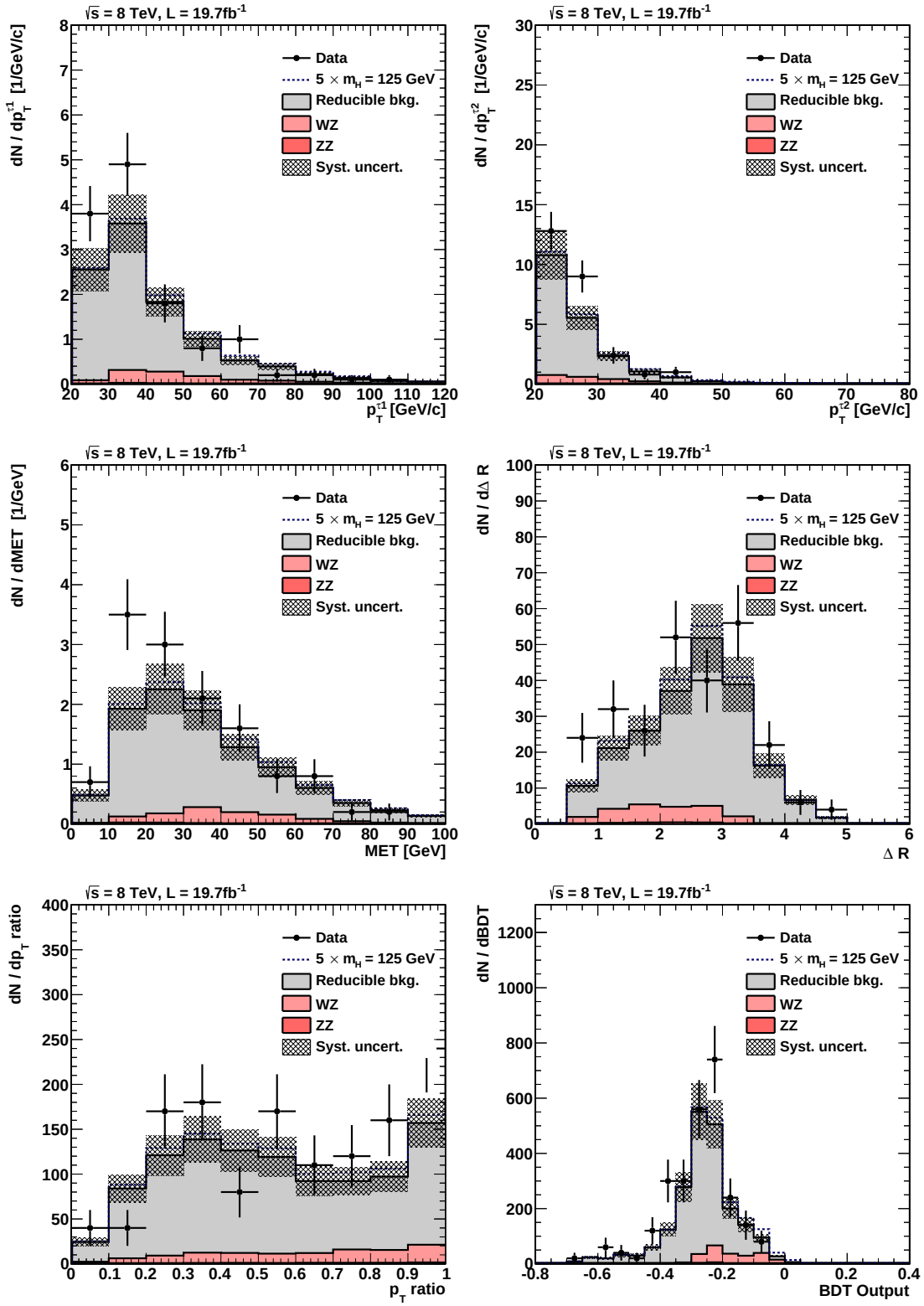

Figure D.15: BDT input and output Variables for the electron channel in the 8 TeV dataset.

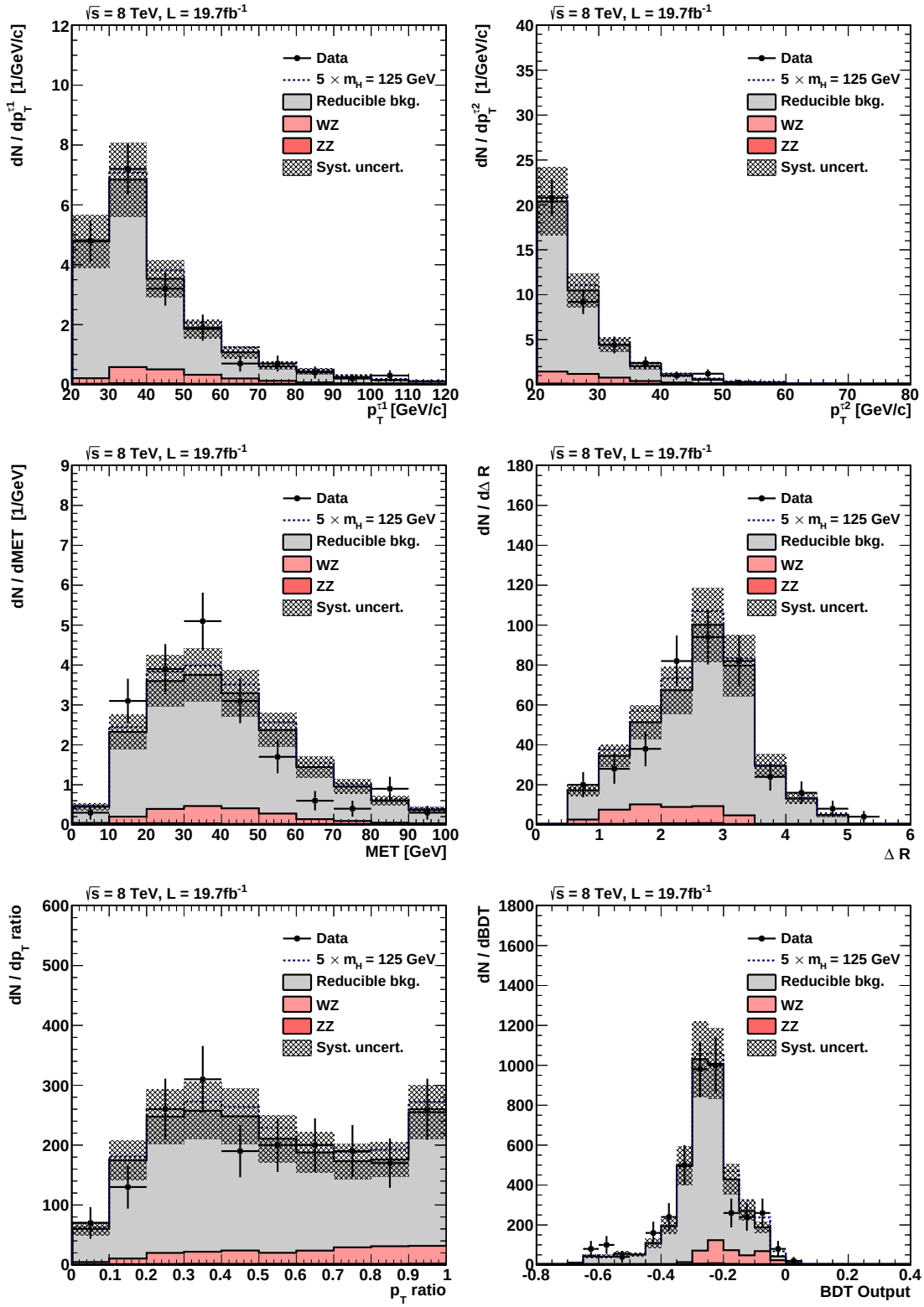

Figure D.16: BDT input and output Variables for the muon channel in the 8 TeV dataset.

Table D.2: Optimization results for the τ_{had} isolation working points. The chosen working points are highlighted. The numbers represent the expected upper limit for the signal strength parameter of a Higgs boson with mass $m_H = 125\,\text{GeV}$ at 95 % C.L.

		Expected Exclusion Limit	
OS τ_{had}	SS τ_{had}	Electron Channel	Muon Channel
Loose	Loose	17.1	11.2
Loose	Medium	15.3	**10.2**
Loose	Tight	15.2	10.2
Medium	Loose	16.4	11.1
Medium	Medium	**14.6**	10.5
Medium	Tight	14.6	10.5
Tight	Loose	16.6	11.2
Tight	Medium	14.7	10.6
Tight	Tight	14.5	10.7

D.4 Analysis Optimization

In this section, the optimization of the most sensitive working points of the analysis is presented. These include the isolation for the two τ_{had} candidates, the electron rejection for the two τ_{had} candidates and the cut on the BDT discriminant.

The available working points are scanned by performing the full analysis, and the working points which lead to the best expected exclusion limit are taken.

D.4.1 τ_{had} Isolation

Table D.2 shows the result for the optimization of the τ_{had} candidate isolation. In this case, the electron rejection working points are not yet optimized, and the cut on the BDT discriminant was replaced by a cut on the transverse momentum of the two τ_{had} candidates at 45 GeV and 30 GeV, respectively. In both the electron channel and the muon channel it can clearly be seen that the SS τ_{had} candidate plays a significant role while the isolation for the OS τ_{had} candidate is less affected. The reason for this is that in all reducible backgrounds the SS τ_{had} candidate is a misidentified jet, while there are some reducible backgrounds with a genuine OS τ_{had}.

D.4.2 τ_{had} Electron Rejection

Table D.3 shows the optimization for the electron rejection working points. The procedure is the same as for the optimization of the τ_{had} isolation. In this case, the optimized isolation working points from the previous step are implemented already, however the BDT discriminant is not yet used. It can be seen that the results for the electron channel and the muon channel are very different. In the muon channel, the probability of electrons being misidentified as τ_{had} are low, and therefore, the loose working point can be used in order to maximize the acceptance. In the electron channel, however, the OS τ_{had} candidate can be a misidentified electron in $Z/\gamma^* \to e^+e^-$ events, and therefore a tight working point must be chosen for this τ_{had} candidate. The SS τ_{had} candidate, however, can be left loose.

Table D.3: Optimization results for the τ_{had} electron rejection working points. The chosen working points are highlighted. The numbers represent the expected upper limit for the signal strength parameter of a Higgs boson with mass $m_H = 125\,\text{GeV}$ at 95 % C.L.

		Expected Exclusion Limit	
OS τ_{had}	SS τ_{had}	Electron Channel	Muon Channel
Loose	Loose	13.8	**10.2**
Loose	Medium	13.9	10.8
Loose	Tight	14.7	11.0
Medium	Loose	13.7	10.8
Medium	Medium	13.8	11.7
Medium	Tight	14.6	12.0
Tight	Loose	**13.4**	11.3
Tight	Medium	13.7	12.3
Tight	Tight	14.4	12.7

The results do not vary significantly when varying the working point of the muon rejection. Therefore, the tight muon rejection is used for both τ_{had} candidates in both channels.

D.4.3 BDT Discriminant

Table D.4 lists the result for the optimization of the cut on the BDT discriminant. It was scanned from -0.050 to -0.200 in steps of 0.005. In both the muon channel and the electron channel the minimum was found at a very similar value, and therefore in both channels the value -0.170 is chosen.

Table D.4: Optimization results for the cut on the BDT discriminant. The chosen working points are highlighted. The numbers represent the expected upper limit for the signal strength parameter of a Higgs boson with mass $m_H = 125\,\text{GeV}$ at 95 % C.L.

	Expected Exclusion Limit	
Cut on BDT Discriminant	Electron Channel	Muon Channel
-0.050	15.4	10.7
-0.055	13.9	9.8
-0.060	13.4	9.5
-0.065	13.6	9.3
-0.070	13.3	9.2
-0.075	12.8	9.0
-0.080	12.2	8.4
-0.085	11.9	8.4
-0.090	11.8	8.3
-0.095	11.8	8.3
-0.100	12.0	8.4
-0.105	12.0	8.3
-0.110	11.7	7.8
-0.115	11.7	7.9
-0.120	11.5	7.8
-0.125	11.3	7.8
-0.130	11.1	7.9
-0.135	11.0	7.9
-0.140	11.0	8.0
-0.145	10.8	8.0
-0.150	10.8	8.1
-0.155	10.6	8.0
-0.160	10.7	8.1
-0.165	10.6	8.2
-0.170	**10.6**	**7.7**
-0.175	10.5	7.7
-0.180	10.7	7.7
-0.185	10.7	7.7
-0.190	10.7	7.8
-0.195	10.7	7.8
-0.200	10.7	7.8

List of Figures

List of Tables

Bibliography

[1] Ernest Rutherford. The Scattering of the Alpha and Beta Rays and the Structure of the Atom. *Proceedings of the Manchester Literary and Philosophical Society IV*, pages 18–20, 1911.

[2] Michael E. Peskin and Dan V. Schroeder. *An Introduction To Quantum Field Theory (Frontiers in Physics)*. Westview Press, 1995.

[3] Peter Dunne. Looking for consistency in the construction and use of Feynman diagrams. *Physics Education*, 36(5):366, 2001.

[4] Chien-Shiung Wu et al. Experimental Test of Parity Conservation in Beta Decay. *Physical Review*, 106:1413, 1957.

[5] S.L. Glashow. Partial Symmetries of Weak Interactions. *Nucl.Phys.*, 22:579–588, 1961.

[6] Steven Weinberg. A Model of Leptons. *Phys.Rev.Lett.*, 19:1264–1266, 1967.

[7] Abdus Salam. Weak and Electromagnetic Interactions. Originally printed in *Svartholm: Elementary Particle Theory, Proceedings Of The Nobel Symposium Held 1968 At Lerum, Sweden*, Stockholm 1968, 367-377.

[8] J. Horstkotte, A. Entenberg, R. S. Galik, A. K. Mann, H. H. Williams, W. Kozanecki, C. Rubbia, J. Strait, L. Sulak, and P. Wanderer. Measurement of neutrino-proton and antineutrino-proton elastic scattering. *Phys. Rev. D*, 25(11):2743–2761, Jun 1982.

[9] ALEPH Collaboration, DELPHI Collaboration, L3 Collaboration, OPAL Collaboration, SLD Collaboration, LEP Electroweak Working Group, SLD Electroweak Group, SLD Heavy Flavour Group. Precision electroweak measurements on the Z resonance. *Phys.Rept.*, 427:257–454, 2006, arxiv:hep-ex/0509008.

[10] UA1 collaboration. Experimental observation of isolated large transverse energy electrons with associated missing energy at $\sqrt{s} = 540$ GeV. *Physics Letters B*, 122(1):103 – 116, 1983.

[11] UA2 collaboration. Observation of single isolated electrons of high transverse momentum in events with missing transverse energy at the CERN pp collider. *Physics Letters B*, 122(5–6):476 – 485, 1983.

[12] UA1 Collaboration. Experimental Observation of Lepton Pairs of Invariant Mass Around 95 GeV/c^2 at the CERN SPS Collider. *Phys.Lett.*, B126:398–410, 1983.

[13] UA2 Collaboration. Evidence for $Z^0 \to e^+e^-$ at the CERN anti-p p Collider. *Phys.Lett.*, B129:130–140, 1983.

[14] F. Englert and R. Brout. Broken symmetry and the mass of gauge vector mesons. *Phys. Rev. Lett.*, 13(9):321–323, Aug 1964.

[15] Peter W. Higgs. Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.*, 13(16):508–509, Oct 1964.

[16] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble. Global conservation laws and massless particles. *Phys. Rev. Lett.*, 13(20):585–587, Nov 1964.

[17] A. Zee. *Quantum Field Theory in a Nutshell.* Princeton University Press, March 2003.

[18] M. Zeise. *Study of Z Boson Decays into Pairs of Muon and Tau Leptons with the CMS Detector at the LHC.* PhD thesis, Karlsruhe Institute of Technology, Germany, 2011.

[19] Andreas S. Kronfeld and Chris Quigg. Resource Letter: Quantum Chromodynamics. *Am.J.Phys.*, 78:1081–1116, 2010, arxiv:1002.5032.

[20] H1 and ZEUS Collaboration. Combined Measurement and QCD Analysis of the Inclusive $e^{\pm}p$ Scattering Cross Sections at HERA. *JHEP*, 1001:109, 2010, arxiv:0911.0884.

[21] CDF Collaboration. Observation of Top Quark Production in $p\bar{p}$ Collisions with the Collider Detector at Fermilab. *Phys. Rev. Lett.*, 74(14):2626–2631, Apr 1995.

[22] D0 Collaboration. Observation of the top quark. *Phys.Rev.Lett.*, 74:2632–2637, 1995, arxiv:hep-ex/9503003.

[23] DONUT. Observation of tau-neutrino interactions. *Phys. Lett.*, B504:218–224, 2001, arxiv:hep-ex/0012035.

[24] ATLAS Collaboration. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys.Lett.*, B716:1–29, 2012, arxiv:1207.7214.

[25] CMS Collaboration. Observation of a new boson with mass near 125 GeV in pp collisions at $\sqrt{s} = 7$ and 8 TeV. *JHEP*, 1306:081, 2013, arxiv:1303.4571.

[26] CMS Collaboration. Figures from the CMS physics Technical Design Report, Volume II: Physics Performance. CMS Collection., Jun 2006.

[27] LHC Higgs Cross Section Working Group. Handbook of LHC Higgs Cross Sections: 3. Higgs Properties. *CERN-2013-004*, CERN, Geneva, 2013, arxiv:1307.1347.

[28] ATLAS Collaboration. Measurements of the properties of the Higgs-like boson in the two photon decay channel with the ATLAS detector using 25 fb^{-1} of proton-proton collision data. Technical Report ATLAS-CONF-2013-012, CERN, Geneva, Mar 2013.

[29] CMS Collaboration. Updated measurements of the Higgs boson at 125 GeV in the two photon decay channel. Technical Report CMS-PAS-HIG-13-001, CERN, Geneva, 2013.

[30] ATLAS Collaboration. Measurements of the properties of the Higgs-like boson in the four lepton decay channel with the ATLAS detector using 25 fb^{-1} of proton-proton collision data. Technical Report ATLAS-CONF-2013-013, CERN, Geneva, Mar 2013.

[31] CMS Collaboration. Measurement of the properties of a Higgs boson in the four-lepton final state. 2013, arxiv:1312.5353.

[32] ATLAS Collaboration. Measurements of the properties of the Higgs-like boson in the $WW^{(*)} \to \ell\nu\ell\nu$ decay channel with the ATLAS detector using 25 fb^{-1} of proton-proton collision data. Technical Report ATLAS-CONF-2013-030, CERN, Geneva, Mar 2013.

[33] CMS Collaboration. Measurement of Higgs boson production and properties in the WW decay channel with leptonic final states. 2013, arxiv:1312.1129.

[34] ATLAS Collaboration. Evidence for Higgs Boson Decays to the $\tau^+\tau^-$ Final State with the ATLAS Detector. Technical Report ATLAS-CONF-2013-108, CERN, Geneva, Nov 2013.

[35] CMS Collaboration. Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons. 2014, arxiv:1401.5041. Accepted for publication by JHEP.

[36] CMS Collaboration. Study of the Mass and Spin-Parity of the Higgs Boson Candidate Via Its Decays to Z Boson Pairs. *Phys.Rev.Lett.*, 110:081803, 2013, arxiv:1212.6639.

[37] ATLAS Collaboration. Evidence for the spin-0 nature of the Higgs boson using ATLAS data. *Phys.Lett.*, B726:120–144, 2013, arxiv:1307.1432.

[38] Super-Kamiokande. Evidence for oscillation of atmospheric neutrinos. *Phys. Rev. Lett.*, 81:1562–1567, 1998, arxiv:hep-ex/9807003.

[39] SNO. Measurement of the charged current interactions produced by B-8 solar neutrinos at the Sudbury Neutrino Observatory. *Phys. Rev. Lett.*, 87:071301, 2001, arxiv:nucl-ex/0106015.

[40] Manuel Drees and Gilles Gerbier. Mini-Review of Dark Matter: 2012. 2012, arxiv:1204.2373.

[41] Nathaniel Craig. The State of Supersymmetry after Run I of the LHC. 2013, arxiv:1309.0528.

[42] Seong Chan Park. Black holes and the LHC: A Review. *Prog.Part.Nucl.Phys.*, 67:617–650, 2012, arxiv:1203.4683.

[43] J. Drees. Review of final LEP results, or, A Tribute to LEP. *Int.J.Mod.Phys.*, A17:3259–3283, 2002, arxiv:hep-ex/0110077.

[44] Paul D. Grannis and Melvyn J. Shochet. The Tevatron collider physics legacy. *Annual Review of Nuclear and Particle Science*, 63(1):467–502, 2013, http://www.annualreviews.org/doi/pdf/10.1146/annurev-nucl-102212-170621.

[45] M Bajko et al. Report of the Task Force on the Incident of 19th September 2008 at the LHC. Technical Report CERN-LHC-PROJECT-Report-1168, CERN, Geneva, Mar 2009.

[46] WikiMedia Commons. `http://commons.wikimedia.org/wiki/File:Cern-accelerator-complex.svg`.

[47] M. Lamont. The First Years of LHC Operation for Luminosity Production. 2013. Proceedings to IPAC '13.

[48] A. Valishev. Tevatron accelerator physics and operation highlights. *Conf.Proc.*, C110328:37–40, 2011, arxiv:1202.5525.

[49] Tevatron Run II Parameters. `http://www-ad.fnal.gov/runII/parameters.pdf`.

[50] CMS Collaboration. CMS Luminosity Based on Pixel Cluster Counting - Summer 2013 Update. Technical Report CMS-PAS-LUM-13-001, CERN, Geneva, 2013.

[51] Simon van der Meer. Calibration of the effective beam height in the ISR. Technical Report CERN-ISR-PO-68-31, CERN, Geneva, 1968.

[52] AC Team. The scale of the LHC. Vue aérienne du CERN avec le tracé du tunnel LHC. `http://cds.cern.ch/record/42370`, Jan 2001.

[53] ALICE Collaboration. The ALICE experiment at the CERN LHC. *Journal of Instrumentation*, 3:S08002, 2008.

[54] ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. *Journal of Instrumentation*, 3:S08003, 2008.

[55] CMS Collaboration. The CMS experiment at the CERN LHC. *Journal of Instrumentation*, 3:S08004, 2008.

[56] LHCb Collaboration. The LHCb Detector at the LHC. *Journal of Instrumentation*, 3:S08005, 2008.

[57] CMS Collaboration. Detector Drawings. CMS Collection, Mar 2012.

[58] CMS Collaboration. Alignment of the CMS Silicon Tracker during Commissioning with Cosmic Rays. *Journal of Instrumentation*, 5:T03009, 2010, arxiv:0910.2505.

[59] CMS Collaboration. *The CMS tracker system project: Technical Design Report.* Technical Design Report CMS. CERN, Geneva, 1997.

[60] CMS Collaboration. *CMS Physics: Technical Design Report Volume 1: Detector Performance and Software.* Technical Design Report CMS. CERN, Geneva, 2006.

[61] CMS Collaboration. *The CMS electromagnetic calorimeter project: Technical Design Report.* Technical Design Report CMS. CERN, Geneva, 1997.

[62] CMS Collaboration. Performance of the CMS hadron calorimeter with cosmic ray muons and LHC beam data. *Journal of Instrumentation*, 5:3012–+, March 2010, arxiv:0911.4991.

[63] *The CMS hadron calorimeter project: Technical Design Report.* Technical Design Report CMS. CERN, Geneva, 1997.

[64] *The CMS muon project: Technical Design Report.* Technical Design Report CMS. CERN, Geneva, 1997.

[65] M.A. Dobbs, S. Frixione, Eric Laenen, K. Tollefson, H. Baer, et al. Les Houches guidebook to Monte Carlo generators for hadron collider physics. pages 411–459, 2004, arxiv:hep-ph/0403045.

[66] Andy Buckley, Jonathan Butterworth, Stefan Gieseke, David Grellscheid, Stefan Höche, et al. General-purpose event generators for LHC physics. *Phys.Rept.*, 504:145–233, 2011, arxiv:1101.2599.

[67] M. L. Mangano and T. J. Stelzer. Tools for the Simulation of Hard Hadronic Collisions. *Annual Review of Nuclear and Particle Science*, 55:555–588, December 2005.

[68] Andreas S. Kronfeld. Twenty-first Century Lattice Gauge Theory: Results from the QCD Lagrangian. *Ann.Rev.Nucl.Part.Sci.*, 62:265–284, 2012, arxiv:1203.1204.

[69] S. Agostinelli et al. GEANT4: A Simulation toolkit. *Nucl.Instrum.Meth.*, A506:250–303, 2003.

[70] Torbjörn Sjöstrand, Stephen Mrenna, and Peter Z. Skands. PYTHIA 6.4 Physics and Manual. *JHEP*, 0605:026, 2006, arxiv:hep-ph/0603175.

[71] Torbjörn Sjöstrand, Stephen Mrenna, and Peter Z. Skands. A Brief Introduction to PYTHIA 8.1. *Comput.Phys.Commun.*, 178:852–867, 2008, arxiv:0710.3820.

[72] CDF Collaboration. Charged jet evolution and the underlying event in proton-antiproton collisions at 1.8 TeV. *Phys. Rev. D*, 65:092002, Apr 2002.

[73] CMS Collaboration. Charged particle multiplicities in pp interactions at $\sqrt{s} = 0.9$, 2.36, and 7 TeV. *Journal of High Energy Physics*, 2011(1):1–38, 2011.

[74] CMS Collaboration. Measurement of the Underlying Event Activity at the LHC with $\sqrt{s} = 7$ TeV and Comparison with $\sqrt{s} = 0.9$ TeV. *JHEP*, 1109:109, 2011, arxiv:1107.0330.

[75] J. Pumplin, D.R. Stump, J. Huston, H.L. Lai, Pavel M. Nadolsky, et al. New generation of parton distributions with uncertainties from global QCD analysis. *JHEP*, 0207:012, 2002, arxiv:hep-ph/0201195.

[76] B. Andersson, G. Gustafson, G. Ingelman, and T. Sjöstrand. Parton fragmentation and string dynamics. *Physics Reports*, 97(2–3):31 – 145, 1983.

[77] Johan Alwall, Michel Herquet, Fabio Maltoni, Olivier Mattelaer, and Tim Stelzer. MadGraph 5 : Going Beyond. *JHEP*, 1106:128, 2011, arxiv:1106.0522.

[78] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. A general framework for implementing NLO calculations in shower Monte Carlo programs: the POWHEG BOX. *JHEP*, 1006:043, 2010, arxiv:1002.2581.

[79] Stefano Frixione, Paolo Nason, and Carlo Oleari. Matching NLO QCD computations with Parton Shower simulations: the POWHEG method. *JHEP*, 0711:070, 2007, arxiv:0709.2092.

[80] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. NLO vector-boson production matched with shower in POWHEG. *JHEP*, 0807:060, 2008, arxiv:0805.4802.

[81] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. Vector boson plus one jet production in POWHEG. *JHEP*, 1101:095, 2011, arxiv:1009.5594.

[82] Tom Melia, Paolo Nason, Raoul Rontsch, and Giulia Zanderighi. W+W-, WZ and ZZ production in the POWHEG BOX. *JHEP*, 1111:078, 2011, arxiv:1107.5051.

[83] Simone Alioli, Sven-Olaf Moch, and Peter Uwer. Hadronic top-quark pair-production with one jet and parton showering. *JHEP*, 1201:137, 2012, arxiv:1110.5251.

[84] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. NLO Higgs boson production via gluon fusion matched with shower in POWHEG. *JHEP*, 0904:002, 2009, arxiv:0812.0578.

[85] Paolo Nason and Carlo Oleari. NLO Higgs boson production via vector-boson fusion matched with shower in POWHEG. *JHEP*, 1002:037, 2010, arxiv:0911.5299.

[86] John M. Campbell, R. K. Ellis, Rikkert Frederix, Paolo Nason, Carlo Oleari, et al. NLO Higgs Boson Production Plus One and Two Jets Using the POWHEG BOX, MadGraph4 and MCFM. *JHEP*, 1207:092, 2012, arxiv:1202.5475.

[87] F. Campanario, T.M. Figy, S. Plätzer, and M. Sjödahl. Electroweak Higgs plus Three Jet Production at NLO QCD. *Phys.Rev.Lett.*, 111:211802, 2013, arxiv:1308.2932.

[88] Hans van Deurzen, Gionata Luisoni, Pierpaolo Mastrolia, Edoardo Mirabella, Giovanni Ossola, et al. NLO QCD corrections to Higgs boson production in association with a top quark pair and a jet. *Phys.Rev.Lett.*, 111:171801, 2013, arxiv:1307.8437.

[89] John M. Campbell and R. K. Ellis. MCFM for the Tevatron and the LHC. *Nucl.Phys.Proc.Suppl.*, 205-206:10–15, 2010, arxiv:1007.3492.

[90] S. Jadach, Z. Was, R. Decker, and J. H. Kühn. The tau decay library TAUOLA: Version 2.4. *Comput.Phys.Commun.*, 76:361–380, 1993.

[91] Philip Ilten. Tau Decays in Pythia 8. 2012, arxiv:1211.6730.

[92] Z. Czyczula, T. Przedzinski, and Z. Was. TauSpinner Program for Studies on Spin Effect in tau Production at the LHC. *Eur.Phys.J.*, C72:1988, 2012, arxiv:1201.0117.

[93] Rene Brun and Fons Rademakers. ROOT — an object oriented data analysis framework. *Nucl.Instrum.Meth.*, A389(1–2):81 – 86, 1997.

[94] F. James and M. Roos. Minuit: A System for Function Minimization and Analysis of the Parameter Errors and Correlations. *Comput.Phys.Commun.*, 10:343–367, 1975.

[95] Andreas Höcker, Peter Speckmayer, Jörg Stelzer, Jan Therhaag, Eckhard von Törne, and Helge Voss. TMVA: Toolkit for Multivariate Data Analysis. *PoS*, ACAT:040, 2007, arxiv:physics/0703039.

[96] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119 – 139, 1997.

[97] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 10 2001.

[98] CMS Collaboration. CMS tracking performance results from early LHC operation. *European Physical Journal C*, 70:1165–1192, December 2010, arxiv:1007.1988.

[99] R. Frühwirth. Application of Kalman filtering to track and vertex fitting. *Nucl.Instrum.Meth.*, A262:444–450, 1987.

[100] E. Chabanat and N. Estre. Deterministic annealing for vertex finding at CMS. pages 287–290, 2005. Proceedings to CHEP '04.

[101] CMS Collaboration. Particle-Flow Event Reconstruction in CMS and Performance for Jets, Taus, and MET. Technical Report CMS-PAS-PFT-09-001, 2009.

[102] CMS Collaboration. Commissioning of the Particle-flow Event Reconstruction with the first LHC collisions recorded in the CMS detector. Technical Report CMS-PAS-PFT-10-001, 2010.

[103] David Barney. CMS outreach, `https://cms-docdb.cern.ch/cgi-bin/PublicDocDB/ShowDocument?docid=4172`.

[104] Gavin P. Salam and Gregory Soyez. A Practical Seedless Infrared-Safe Cone jet algorithm. *JHEP*, 0705:086, 2007, arxiv:0704.0292.

[105] S. Catani, Yuri L. Dokshitzer, M.H. Seymour, and B.R. Webber. Longitudinally invariant K_t clustering algorithms for hadron hadron collisions. *Nucl.Phys.*, B406:187–224, 1993.

[106] Yuri L. Dokshitzer, G.D. Leder, S. Moretti, and B.R. Webber. Better jet clustering algorithms. *JHEP*, 9708:001, 1997, arxiv:hep-ph/9707323.

[107] M. Wobisch and T. Wengler. Hadronization corrections to jet cross-sections in deep inelastic scattering. 1998, arxiv:hep-ph/9907280.

[108] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The Anti-k(t) jet clustering algorithm. *JHEP*, 0804:063, 2008, arxiv:0802.1189.

[109] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. FastJet User Manual. *Eur.Phys.J.*, C72:1896, 2012, arxiv:1111.6097.

[110] The CMS collaboration. Determination of jet energy calibration and transverse momentum resolution in CMS. *Journal of Instrumentation*, 6(11):P11002, 2011.

[111] CMS Collaboration. Identification of b-quark jets with the CMS experiment. *JINST*, 8:P04013, 2013, arxiv:1211.4462.

[112] R. Frühwirth, W. Waltenberger, and P. Vanlaer. Adaptive vertex fitting. *J.Phys.*, G34:N343, 2007.

[113] CMS Collaboration. Performance of b tagging at $\sqrt{s} = 8$ TeV in multijet, ttbar and boosted topology events. Technical Report CMS-PAS-BTV-13-001, CERN, Geneva, 2013.

[114] S. Baffioni, C. Charlot, F. Ferri, D. Futyan, P. Meridiani, et al. Electron reconstruction in CMS. *Eur.Phys.J.*, C49:1099–1116, 2007.

[115] R. Frühwirth. Track fitting with non-gaussian noise. *Computer Physics Communications*, 100(1–2):1 – 16, 1997.

[116] H. Bethe and W. Heitler. On the stopping of fast particles and on the creation of positive electrons. *Proceedings of the Royal Society of London*, A146:83–112, 1934.

[117] CMS Collaboration. Electron performance with 19.6 fb^{-1} of data collected at $\sqrt{s} = 8$ TeV with the CMS detector. CMS-DP-2013/003, 2013.

[118] The CMS Collaboration. Performance of CMS muon reconstruction in pp collision events at $\sqrt{s} = 7$ TeV. *Journal of Instrumentation*, 7:2P, October 2012, arxiv:1206.4071.

[119] Particle Data Group. Review of Particle Physics (RPP). *Phys.Rev.*, D86:010001, 2012.

[120] CMS Collaboration. Measurement of the Inclusive Z Cross Section via Decays to Tau Pairs in *pp* Collisions at $\sqrt{s} = 7$ TeV. *JHEP*, 1108:117, 2011, arxiv:1104.1617.

[121] CMS Collaboration. Performance of τ-lepton reconstruction and identification in CMS. *Journal of Instrumentation*, 7(01):P01001, 2012.

[122] L. Bianchini, I. Naranjo Fong, and C. Veelken. Development of a new tau identification discriminator against electrons. CMS AN-2012/417.

[123] CMS Collaboration. Tau Performance Plots for 2012. CMS-DP-2013-012, 2013.

[124] Christoph Eck, J Knobloch, Leslie Robertson, I Bird, K Bos, N Brook, D Düllmann, I Fisk, D Foster, B Gibbard, C Grandi, F Grey, J Harvey, A Heiss, F Hemmer, S Jarp, R Jones, D Kelsey, M Lamanna, H Marten, P Mato-Vila, F Ould-Saada, B Panzer-Steindel, L Perini, Y Schutz, U Schwickerath, J Shiers, and T Wenaus. *LHC computing Grid: Technical Design Report. Version 1.06 (20 Jun 2005).* Technical Design Report LCG. CERN, Geneva, 2005.

[125] P. Marcu, D. Schmitz, A. Hanemann, and S. Trocha. Monitoring and visualisation of the large hadron collider optical private network. In *Roedunet International Conference (RoEduNet), 2010 9th*, pages 316–321, 2010.

[126] E. Bos, E. Martelli, P. Moroni, D. Foster, et al. LHC Tier-0 to Tier-1 High-Level Network Architecture. Technical Report LHC Optical Private Network (LHCOPN), Geneva, 2005. `https://twiki.cern.ch/twiki/pub/LHCOPN/LHCopnArchitecture/LHCnetworkingv2.dgf.doc`. Retrieved: 2014-01-28.

[127] P. Fuhrmann. dCache, the Overview. `http://www.dcache.org/manuals/dcache-whitepaper-light.pdf`, Retrieved: 2014-01-31.

[128] E. Fretwurst, N. Claussen, N. Croitoru, G. Lindström, B. Papendick, U. Pein, H. Schatz, T. Schulz, and R. Wunstorf. Radiation hardness of silicon detectors for future colliders. *Nucl.Instrum.Meth.*, A326(1–2):357 – 364, 1993.

[129] E. Fretwurst, H. Feick, M. Glaser, C. Gössling, E.H.M. Heijne, A. Hess, F. Lemeilleur, G. Lindström, K.H. Mählmann, A. Rolf, T. Schulz, and C. Soave. Reverse annealing of the effective impurity concentration and long term operational scenario for silicon detectors in future collider experiments. *Nucl.Instrum.Meth.*, A342(1):119 – 125, 1994.

[130] CMS Collaboration. CMS Technical Design Report for the Pixel Detector Upgrade. Technical Report CERN-LHCC-2012-016. CMS-TDR-11, CERN, Geneva, Sep 2012.

[131] E. Bartz. The 0.25μm token bit manager chip for the CMS pixel readout. *Conf.Proc.*, C05091210:25, 2005.

[132] S. König, Ch. Hörmann, R. Horisberger, S. Streuli, and R. Weber. Assembly of the CMS pixel barrel modules. *Nucl.Instrum.Meth.*, A565(1):62 – 66, 2006.

[133] Y. Allkofer, C. Amsler, D. Bortoletto, V. Chiochia, L. Cremaldi, S. Cucciarelli, A. Dorokhov, C. Hörmann, R. Horisberger, D. Kim, M. Konecki, D. Kotlinski, K. Prokofiev, C. Regenfus, T. Rohe, D. A. Sanders, S. Son, M. Swartz, and T. Speer. Design and performance of the silicon sensors for the CMS barrel pixel detector. *Nucl.Instrum.Meth.*, A584:25–41, January 2008, arxiv:physics/0702092.

[134] CMS Collaboration. Commissioning and performance of the CMS pixel tracker with cosmic ray muons. *Journal of Instrumentation*, 5(03):T03007, 2010.

[135] W. Erdmann. The CMS pixel detector. *International Journal of Modern Physics*, A25(07):1315–1337, 2010.

[136] Petra Merkel. Experience with mass production bump bonding with outside vendors in the CMS FPIX project. *Nucl.Instrum.Meth.*, A582:771–775, 2007.

[137] H. C. Kästli, M. Barbero, W. Erdmann, C. Hörmann, R. Horisberger, D. Kotlinski, and B. Meier. Design and performance of the CMS pixel detector readout chip. *Nucl.Instrum.Meth.*, A565:188–194, September 2006, arxiv:physics/0511166.

[138] D Kotliński. Status of the CMS pixel detector. *Journal of Instrumentation*, 4(03):P03019, 2009.

[139] S. Dambach. *CMS pixel module optimization and B meson lifetime measurements.* PhD thesis, ETH Zurich, 2009.

[140] T. Behnke, E. Garutti, I. M. Gregor, T. Haas, U. Kötz, I. Melzer-Pellmann, N. Meyners, J. Mnich, and F. Sefkow. Test Beams at DESY. EUDET-Memo-2007-11, 2007.

[141] DESY Testbeam Website. `http://testbeam.desy.de`.

[142] M. Huning and M. Schmitz. Recent Changes to the e^- / e^+ Injector (Linac II) at DESY. 2008. Proceedings of LINAC 2008, Canada.

[143] I. Rubinsky. An EUDET/AIDA pixel beam telescope for detector development. *Physics Procedia*, 37(0):923 – 931, 2012. Proceedings of the 2nd International Conference on Technology and Instrumentation in Particle Physics (TIPP 2011).

[144] The MARLIN software. `http://ilcsoft.desy.de/portal/software_packages/marlin`. Retrieved: 2014-02-07.

[145] I. Rubinsky. EUTelescope. Offline track reconstruction and DUT analysis software. EUDET-Memo-2010-12, 2010.

[146] Volker Blobel and Claus Kleinwort. A New method for the high precision alignment of track detectors. pages URL–STR(9), 2002, arxiv:hep-ex/0208021.

[147] Claus Kleinwort. General broken lines as advanced track fitting method. *Nucl.Instrum.Meth.*, 673A(0):107 – 110, 2012.

[148] Mahesh K. Varanasi and Behnaam Aazhang. Parametric generalized gaussian density estimation. *The Journal of the Acoustical Society of America*, 86(4):1404–1415, 1989.

[149] R. A. Fisher. Applications of "Student's" distribution. *Metron*, 5:90–104, 1925.

[150] V. Chiochia, M. Swartz, D. Fehling, G. Giurgiu, and P. Maksimovic. A novel technique for the reconstruction and simulation of hits in the cms pixel detector. In *Nuclear Science Symposium Conference Record, 2008. NSS '08. IEEE*, pages 1909–1912, Oct 2008.

[151] CMS Collaboration. Pixel Hit Reconstruction with the CMS Detector. 2008, arxiv:0808.3804.

[152] A. Burgmeier and D. Pitzl. Measurement of the Pixel Barrel Resolution with the Triplet Method. CMS AN-2013/356.

[153] Veikko Karimaki. Effective circle fitting for particle trajectories. *Nucl.Instrum.Meth.*, A305:187–191, 1991.

[154] Johannes Gassner. Messung der Ortsauflösung des H1-Siliziumvertexdetektors. Diplomarbeit, ETH Zurich, Switzerland, 1996.

[155] Nazar Bartosik. Simultaneous alignment and Lorentz angle calibration in the CMS silicon tracker using Millepede II. Technical Report CERN-CMS-CR-2013-343, CERN, Geneva, Oct 2013.

[156] M. Swartz, V. Chiochia, Y. Allkofer, D. Bortoletto, L. Cremaldi, et al. Observation, modeling, and temperature dependence of doubly peaked electric fields in irradiated silicon pixel sensors. *Nucl.Instrum.Meth.*, A565:212–220, 2006, arxiv:physics/0510040.

[157] Synopsys Sentaurus Device. `http://www.synopsys.com/Tools/TCAD/DeviceSimulation/Pages/SentaurusDevice.aspx`.

[158] H. Bichsel. Straggling in Thin Silicon Detectors. *Rev.Mod.Phys.*, 60:663–699, 1988.

[159] NIST ESTAR Program. `http://physics.nist.gov/PhysRefData/Star/Text/ESTAR.html`.

[160] CMS Collaboration. Search for the standard model Higgs boson produced in association with a W or a Z boson and decaying to bottom quarks. 2013, arxiv:1310.3687.

[161] A.J. Gilbert et al. Search for the Standard-Model Higgs boson decaying to tau pairs in proton-proton collisions at $\sqrt{s} = 7$ and 8 TeV. CMS AN-2013/206.

[162] A. J. Gilbert et al. Physics Objects in the Higgs to Tau Tau Analysis. CMS AN-2013/188.

[163] J. Swanson et al. Search for Higgs to Tau Tau in the Muon-Tau and Electron-Tau Channels. CMS AN-2013/178.

[164] V. Dutta et al. Search for Higgs to Tau Tau in the Electron-Muon Channel . CMS AN-2013/190.

[165] R. A. Manzoni et al. Search for the Higgs boson decaying into TauTau in the full hadronic channel. CMS AN-2013/189.

[166] A. Raspereza et al. Search for Higgs boson decays to tau pairs in the di-muon and di-electron channels. CMS AN-2013/192.

[167] M. Verzetti et al. Search for a Standard Model Higgs boson decaying to tau pairs produced in association with a W or Z boson. CMS AN-2013/187.

[168] Takahashi Y. et al. Theoretical uncertainty for the Higgs production via VBF and Gluon Fusion process. CMS AN-2013/262.

[169] Robert V. Harlander and William B. Kilgore. Next-to-next-to-leading order Higgs production at hadron colliders. *Phys.Rev.Lett.*, 88:201801, 2002, arxiv:hep-ph/0201206.

[170] D. de Florian, G. Ferrera, M. Grazzini, and D. Tommasini. Higgs boson production at the LHC: transverse momentum resummation effects in the $H \to 2\gamma$, $H \to WW \to \ell\nu$ and $H \to ZZ \to 4\ell$ decay modes. *JHEP*, 1206:132, 2012, arxiv:1203.6321.

[171] CMS Collaboration. Measurement of the inelastic proton-proton cross section at $\sqrt{s} = 7$ TeV. *Phys.Lett.*, B722:5–27, 2013, arxiv:1210.6718.

[172] CMS Pile-up Twiki Page. `https://twiki.cern.ch/twiki/bin/view/CMS/PileupJSONFileforData`.

[173] N. Adam, J. Berryhill, V. Halyo, A. Hunt, and Mishra K. Generic Tag and Probe Tool for Measuring Efficiency at CMS with Early Data. CMS AN-2009/111.

[174] V. Chiochia, M. Verzetti, C. Veelken, M. Klute, M. Chen, A. Savin, and S. Gennai. Measurement of tau identification efficiency with 2011 CMS data. CMS AN-2011/200.

[175] M. Chen, V. Chiochia, S. Gennai, M. Klute, A. Savin, C. Veelken, and M. Verzetti. Measurement of hadronic tau identification efficiency in 2011 run B data. CMS AN-2011/514.

[176] Muon Scale Factors from the Muon POG. `https://twiki.cern.ch/twiki/bin/viewauth/CMS/MuonReferenceEffs`. Retrieved: 2014-01-08.

[177] Electron Scale Factors from the EGamma POG. https://twiki.cern.ch/twiki/bin/view/CMS/MultivariateElectronIdentification#Recommended_Working_Points_With. Retrieved: 2014-01-08.

[178] J. S. Conway. Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra. *ArXiv e-prints*, March 2011, arxiv:1103.0354.

[179] Bernhard Mistlberger and Falko Dulat. Limit setting procedures and theoretical uncertainties in Higgs boson searches. 2012, arxiv:1204.3851.

[180] G. Cowan, K. Cranmer, E. Gross, and O. Vitells. Asymptotic formulae for likelihood-based tests of new physics. *European Physical Journal C*, 71:1554, February 2011, arxiv:1007.1727.

[181] S. Berge, W. Bernreuther, B. Niepelt, and H. Spiesberger. How to pin down the CP quantum numbers of a Higgs boson in its tau decays at the LHC. *Phys.Rev.*, D84:116003, 2011, arxiv:1108.0670.

[182] T. Früboes. *Search for neutral Higgs boson in $\tau\tau \to \mu\tau_{\text{jet}}$ final state in the CMS experiment.* PhD thesis, National Centre for Nuclear Research, Warsaw, Poland, 2013.

[183] Armin Burgmeier. Data-driven Estimation of Z^0 Background Contributions to the Higgs Search in the $H \to \tau^+\tau^-$ Channel with the CMS Experiment at the LHC. Diplomarbeit, Karlsruhe Institute of Technology, Germany, 2011.

[184] M. Bluj, A. Burgmeier, T. Früboes, G. Quast, and M. Zeise. Modelling of $\tau\tau$ final states by embedding τ pairs in $Z \to \mu\mu$ events. CMS AN-2011/020.

[185] N. Daci. *Electron selection and search for the Higgs boson decaying into tau leptons pairs with the CMS detector at the LHC.* PhD thesis, Université Paris Sud - Paris XI, 2013.

[186] A. Burgmeier, T. Früboes, and C. Veelken. Modelling backgrounds via Embedding technique applied on recHit level. CMS AN-2013/073.

[187] CMS Collaboration. Muon ID and Isolation Efficiencies in 2012 RunAB. *CMS Performance Note*, 2012/025, 2012.

[188] Don Groom. Atomic Nuclear Properties. http://pdg.lbl.gov/2013/AtomicNuclearProperties. Retrieved: 2014-03-10.

[189] CMS Collaboration. Search for a light charged Higgs boson in top quark decays in pp collisions at $\sqrt{s} = 7$ TeV. *JHEP*, 1207:143, 2012, arxiv:1205.5736.

[190] R. Chierici, F. Cossutti, G. G. C. Retuerto, S. Padhi, F. Stoeckli, and Silvano Tosi. Standard Model Cross Sections for CMS at 7 TeV. https://twiki.cern.ch/twiki/bin/viewauth/CMS/StandardModelCrossSections.

[191] P. Lenzi, S. Padhi, G. G. C. Retuerto, F. Wuerthwein, T. Seva, and H. Y. Tong. Standard Model Cross Sections for CMS at 8 TeV. https://twiki.cern.ch/twiki/bin/viewauth/CMS/StandardModelCrossSectionsat8TeV.

[192] C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26:404–413, 1934.

[193] S. Dittmaier, S. Dittmaier, C. Mariotti, G. Passarino, R. Tanaka, et al. Handbook of LHC Higgs Cross Sections: 2. Differential Distributions. 2012, arxiv:1201.3084.

[194] CMS Collaboration. Absolute calibration of the luminosity measurement at cms: Winter 2012 update. Technical Report CMS-PAS-SMP-12-008, CERN, 2012.

[195] CMS Collaboration. Missing transverse energy performance of the CMS detector. *Journal of Instrumentation*, 6:9001, September 2011, arxiv:1106.5048.

[196] Evidence for the 125 GeV Higgs boson decaying to a pair of τ leptons. Additional material to the analysis CMS-HIG-13-004. `https://twiki.cern.ch/twiki/bin/view/CMSPublic/Hig13004PubTWiki`. Retrieved: 2014-05-17.

[197] CDF Collaboration. Search for the standard model higgs in the $\ell\nu\tau\tau$ and $\ell\ell\tau\tau$ channels. Technical Report CDF Conference Note 10500, FNAL, 2011.

[198] D0 Collaboration. Search for Higgs boson production in trilepton and like-charge electron-muon final states with the D0 detector. *Phys.Rev.*, D88(5):052009, 2013, 1302.5723.

[199] Christian Veelken. Tau Id Developments in CMS. `https://indico.desy.de/materialDisplay.py?contribId=5&materialId=slides&confId=9586`. Talk given at the 12th workshop of the tautau Analysis Working Group, DESY, Hamburg, Germany.

[200] CMS Collaboration. Commissioning of the Particle-Flow reconstruction in Minimum-Bias and Jet Events from pp Collisions at 7 TeV. Technical Report CMS-PAS-PFT-10-002, 2010.

[201] CMS Collaboration. MET performance in 8 TeV data. Technical Report CMS-PAS-JME-12-002, CERN, Geneva, 2013.

[202] R. K. Ellis, I. Hinchliffe, M. Soldate, and J.J. van der Bij. Higgs Decay to tau+ tau-: A Possible Signature of Intermediate Mass Higgs Bosons at the SSC. *Nucl.Phys.*, B297:221, 1988.

[203] A. Elagin, P. Murat, A. Pranko, and A. Safonov. A New Mass Reconstruction Technique for Resonances Decaying to di-tau. *Nucl.Instrum.Meth.*, A654:481–489, 2011, arxiv:1012.4686.

[204] B.K. Bullock, K. Hagiwara, and A.D. Martin. Tau polarization and its correlations as a probe of new physics. *Nuclear Physics B*, 395(3):499 – 533, 1993.

Acknowledgments

First and foremost, I sincerely thank my advisor Prof. Günter Quast. He gave me the freedom to work independently and to explore my own ideas, yet, when I needed him, he was always there to share his incredible wisdom and experience. I am grateful to Prof. Ulrich Husemann for agreeing to co-referee this thesis and many constructive comments on the draft.

I thank my supervisors at DESY, Dr. Alexei Raspereza for the Higgs analysis and Dr. Daniel Pitzl for the pixel upgrade efforts. They supported me in many ways, did not get tired of answering my questions and always shared their knowledge with me.

The results in this thesis would not have been possible without the help from many smart people at other institutes from whom I have learned a lot, especially Dr. Christian Veelken, Dr. Adrian Perieanu, Prof. Alexei Safonov and Jeffrey Roe.

I am grateful to my colleagues Luigi Calligaris, Gregor Hellwig, Simon Spannagel and Raphael Friese for reading thesis drafts and providing valuable feedback. Special thanks go to Dr. Manuel Zeise who introduced me to the world of high energy physics and who was a pleasure to work with.

I would like to thank my family and friends for the support and for encouraging me in everything I do. In particular, my gratitude goes to my fiancée, Georgiana Ogrean, who cheers me up when I feel bad, who gave me a new perspective on myself, who arguably makes the best ice cream in the world, and, more than anything else, to whom I am grateful for being who she is.