TECHNISCHE UNIVERSITÄT WIEN

ÖSTERREICHISCHE AKADEMIE DER WISSENSCHAFTEN

HEPHY INSTITUT FÜR HOCHENERGIEPHYSIK

DISSERTATION

Data-Driven Background Modeling for Precision Studies of the Higgs Boson and Searches for New Physics with the CMS Experiment

Ausgeführt zum Zwecke der Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

eingereicht von

MSc Janík Walter Andrejkovič
Matrikelnummer 11840883
ausgeführt am Atominstitut der Technischen Universität Wien (E141)
in Zusammenarbeit mit dem Institut für Hochenergiephysik (HEPHY)
der Österreichischen Akademie der Wissenschaften (ÖAW)
unter der Leitung von
Univ. Prof. Dipl.-Phys. Dr.rer.nat Jochen Schieck
eingereicht an der Fakultät für Physik der
Technischen Universität Wien

Wien, 18. Juli 2022

(Unterschrift Verfasser/in) (Unterschrift Betreuer/in)

Zusammenfassung

Der große Hadronen-Speicherring (LHC) am CERN hat in den Jahren von 2016 bis 2018 eine nie da gewesene Menge an hochenergetischen Proton-Proton-Kollisionen erzeugt. Daten, welche vom Compact-Muon-Solenoid-Experiment (CMS) in dieser Zeit aufgezeichnet und zertifiziert wurden, entsprechen $137\,\text{fb}^{-1}$ und werden in dieser Arbeit analysiert. Kollisionsdaten werden für Präzisionstests des Standardmodells (SM) der Teilchenphysik verwendet und auch für Suchen nach neuen Phänomenen, welche über das SM hinausgehen.

Die in dieser Arbeit vorgestellte Präzisionsmessung befasst sich mit dem SM Higgs-Boson im Zusammenhang mit seinem Zerfall in ein Paar von Tau-Leptonen. Zudem behandeln alle hier vorgestellten Suchen nach neuen Teilchen deren Zerfall in Leptonen. In jeder Proton-Proton-Kollision werden, in Prozessen der starken Wechselwirkung, jedoch zahlreiche Teilchen erzeugt. Dass in solchen Kollisionen Teilchen falsch identifiziert werden, ist unumgänglich und derartige Verwechslungen führen zu Untergründen, welche das Signal in Mess- und Suchregionen stören. Solche Untergründe müssen abgeschätzt werden, wobei sich eine genaue Abschätzung auf Simulationsbasis als schwierig gestaltet. Aus diesem Grund wurden besondere datenbasierte Methoden entwickelt, welche die Anzahl an falsch-identifizierten Objekten in Mess- und Suchregionen besser ermitteln können. In dieser Arbeit werden solche datenbasierte Methoden vorgestellt und deren Anwendung in mehreren Analysen vorgeführt.

Für den Zerfall des SM Higgs-Bosons in zwei Tau-Leptonen wird eine Messung von Signalstärken in bis zu zwölf kinematischen Regionen vorgestellt. Eine Signalstärke ist ein Maß für die relative Stärke eines Signals in Bezug auf die jeweilige Erwartung des SMs. Die Messung der Signalstärke im Zusammenhang von Higgs-Boson-Produktion mittels Vektor-Boson-Fusion ergibt einen Wert von $0.81^{+0.17}_{-0.16}$, jene im Zusammenhang von Higgs-Boson-Produktion mittels Gluon-Gluon-Fusion einen Wert von $0.67^{+0.20}_{-0.18}$. Des Weiteren werden in dieser Arbeit zwei Suchen nach weiteren Higgs-Bosonen, welche in ein Paar von Tau-Leptonen zerfallen, vorgestellt. In beiden Fällen wird kein Überschuss gegenüber den Erwartungen des SMs beobachtet und es werden Modell-unabhängige obere Schranken auf das Produkt von Produktionswirkungsquerschnitt und Zerfallsbreite des Higgs-Bosons in zwei Tau-Leptonen berechnet. Eine Suche nach Leptoquarks der dritten Generation mit zwei Tau-Leptonen im Endzustand wird ebenfalls präsentiert. Alle vorgestellten Analysen verwenden die in dieser Arbeit gezeigte Methode zur Bestimmung der Anzahl an Untergrundprozessen von falsch-identifizierten Leptonen. Schließlich wird in dieser Arbeit ebenfalls eine Suche nach, in Paaren produzierten, leichten Top-Squarks besprochen, mit dem Ziel eine neu entwickelte, datenbasierte Untergrundbestimmungsmethode in dieser Analyse anzuwenden.

Abstract

The Large Hadron Collider (LHC) at CERN has produced an unprecedented amount of high-energy proton-proton collisions during the years 2016 to 2018. Data recorded and certified by the Compact Muon Solenoid (CMS) experiment from this time period amount to $137\,\mathrm{fb}^{-1}$ and are analyzed in this thesis. Collision data are used to perform precision tests of the Standard Model (SM) of particle physics and to search for new phenomena beyond this model.

The presented precision measurement focuses on the SM Higgs boson decaying into a pair of tau leptons. Moreover, all presented searches for new particles target their decay into leptons. Proton-proton collisions, however, lead to abundant particle formation governed by the strong interaction. Particle misidentifications in collisions with such numerously produced particles are unavoidable and pose a background in measurement and search regions alike that needs to be estimated. Modeling particle misidentification precisely by means of simulation turns out to be difficult. Thus, dedicated data-driven methods are developed to assess the amount of misidentification entering measurement and search regions. In this thesis, such data-driven methods are presented and their application to several analyses is demonstrated.

For the SM Higgs boson decay into two tau leptons, a measurement of signal strength modifiers, quantifying the signal size relative to the respective SM expectation in up to twelve kinematic regions, is presented. Results of more inclusive signal strengths yield $0.81^{+0.17}_{-0.16}$ and $0.67^{+0.20}_{-0.18}$ for Higgs boson production via vector boson fusion and gluon-gluon fusion, respectively. Furthermore, two searches for additional Higgs bosons, decaying into a pair of tau leptons, are presented in this thesis. No excess over the SM expectation is observed and model-independent upper limits on the production cross section times branching ratio to tau pairs are set in both cases. In addition, a search for third-generation leptoquarks in a di-tau final state is presented. All these analyses adopt the data-driven background estimates shown in this thesis to evaluate contributions from misidentified leptons. Lastly, a search for pair-produced light top squarks is reviewed in the context of applying a newly developed data-driven background estimation technique.

Contents

Chapter 1

Introduction

The Standard Model (SM) of particle physics describes the interactions between fundamental particles. It was developed in the second half of the twentieth century and has led the way to several discoveries of new particles. Worth mentioning are the discovery of the W and Z bosons [1], the top quark [2] and most recently the Higgs boson (H), which was discovered in 2012 by the ATLAS[1] [3] and CMS[2] [4] Collaborations. Interactions described by the SM span a very wide range of energy scales and theoretical predictions as well as experimental measurements can reach an extremely high precision [5, 6]. Despite this success, it is known that the SM is incomplete and will fail to describe correctly phenomena at very high energies. Several theoretical models exist that incorporate the SM and attempt to provide answers to the open questions of the high-energy physics community like for example the nature of dark matter. Ultimately, gaining knowledge about which of these models is realized in Nature has to be obtained from experimental analyses.

The physics program of the Large Hadron Collider (LHC) at the European Organization for Nuclear Research (CERN) encompasses several precision measurements of the Higgs boson as well as searches for new particles predicted by different beyond the Standard Model (BSM) extensions. At LHC experiments, conventionally Higgs boson analyses are performed according to the different final states the Higgs boson can decay to. One of those final states is given by the Higgs boson decaying into a pair of tau (τ) leptons, $H \to \tau\tau$. A measurement of differential signal strength modifiers, which quantify the signal sizes relative to the respective predictions from the SM, for this particular final state is presented in this thesis. Conducting a precision measurement in the context of the $H \to \tau\tau$ analysis is mainly possible due to the unprecedented amount of data collected by the CMS experiment during the years of data-taking from 2016 to 2018. However, it is crucial to precisely evaluate background contributions to the analysis to extract the best possible physics result. One major background contribution is made up by processes where recorded particles get misidentified as leptons, e.g. tau leptons in case of the SM $H \to \tau\tau$ analysis. The main focus of this thesis is to estimate such contributions. For this purpose, methods which (mostly) rely on data are exploited and optimized to best match the analysis setup. While the effort started in the context of the SM $H \to \tau\tau$ analysis, the developed method can be adjusted to estimate the contributions from misidentified leptons also in other analyses. As such, the presented method is successfully used to search for heavy Higgs bosons decaying into a pair of tau leptons. The large amount of data collected in 2016 to 2018 allows putting more stringent limits on BSM theories, predicting such heavier Higgs bosons. In addition, searches of more exotic BSM particles with di-

[1] A Toroidal LHC ApparatuS (ATLAS)
[2] Compact Muon Solenoid (CMS)

tau final states can make use of the developed background estimation method as is also showcased for one such analysis in this thesis. Another application of the method, where recorded particles are misidentified as light leptons, i.e. electrons or muons, is outlined as a last example in this thesis. In summary, the main link between the presented analyses is the background estimation method which was developed me.

Apart from analysis work, I was involved in several qualification tasks. One of them was to evaluate and report effects of software changes on the reconstruction of hadronically decaying tau leptons. Another task concerns the development of analysis code to monitor a detector calibration quantity. This detector calibration study is briefly presented in Chapter 4. Results of the study are used in the reconstruction of data serving the vast majority of physics studies of the CMS Collaboration. The rest of thesis is organized as follows. Chapter 2 serves as introduction to the relevant parts of the SM and BSM theories. An overview of the experimental setup, the recording and reconstruction of collision data is given throughout Chapter 3. Higgs boson measurements are an essential part of this thesis and therefore some relevant phenomenological aspects are discussed in Chapter 5. Furthermore, in Chapter 5 all presented analyses are reviewed in terms of experimental signatures of the respective signal processes and experimental challenges are discussed. All analyses with a di-tau final state have common analysis ingredients and share the same kind of background processes that overlay the signal. This is discussed in more detail in Chapter 6. Relevant analysis ingredients used in a search of BSM particles in a single-lepton final state are introduced in Chapter 7. Chapter 8 represents the main part of this thesis, describing in detail the methods developed to estimate contributions from misidentified leptons. Results from all analyses are presented throughout Chapter 9 in separate sections. A summary with concluding remarks is given in Chapter 10.

Results of this thesis have contributed to the following publicly available documents:

Measurements of Higgs boson production in the decay channel with a pair of τ leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV. 4 2022. arXiv:2204.12957

Armen Tumasyan et al. Search for a heavy Higgs boson decaying into two lighter Higgs bosons in the $\tau\tau$bb final state at 13 TeV. *JHEP*, 11:057, 2021. arXiv:2106.10361, doi:10.1007/JHEP11(2021)057

CMS Collaboration. Searches for additional Higgs bosons and vector-like leptoquarks in $\tau\tau$ final states in proton-proton collisions at $\sqrt{s} = 13$ TeV. 2022. URL: http://cds.cern.ch/record/2803739

The presented work has been carried out in close collaboration with analysis teams from HEPHY[3], KIT[4] and MTA-ELTE[5]. I developed and validated the data-driven method for background estimation described in detail in Chapter 8. This method was successfully applied to several physics measurements performed in collaboration with other CMS members [10, 11, 12, 13, 14, 15]. In order to present the full value of the data-driven method, results of the various experimental analyses, partly performed by other CMS members, are shown here for completeness.

[3]Institute of High-Energy Physics of the Austrian Academy of Sciences
[4]Karlsruhe Institute of Technology
[5]Hungarian Academy of Sciences and the Eötvös Loránd University

Chapter 2

The Standard Model and Beyond

If I could remember the names of all these particles, I'd be a botanist.

Enrico Fermi, 1901 to 1954

The SM of particle physics is a theory that aims at describing physics at its most fundamental level by expressing how *elementary* particles *interact* with each other. With ever higher energies achieved in particle physics experiments, it has been possible to probe particle substructures on smaller and smaller length scales. Current length scales that can be probed are in the order of 10^{-19} m. All particles with no known substructure are termed elementary. Currently, four fundamental interactions (forces) are known which are the electromagnetic force, the gravitational force, the weak force and the strong force. The SM describes all of them with exception of the gravitational force. However, on the subatomic length scales that can be currently probed, the gravitational force is negligible because its intensity is about 25 orders of magnitude lower than the weakest of the forces described by the SM. Nevertheless, at a certain scale – referred to as the Planck scale – the gravitational force will be comparable to the other forces and the SM will fail to describe physics at this scale. The Planck length for instance is in the order of 10^{-35} m and hence still far away from being reached with current particle physics experiments.

Two types of elementary particles are distinguished in the SM:

Fermions: Fermions obey the Fermi-Dirac statistics and are spin-½ particles. They make up the fundamental building blocks of matter and are thus also referred to as *matter* particles.

Bosons: Bosons follow the Bose-Einstein statistics and have an integer-valued spin. Spin-1 particles are also called *vector* bosons or force carriers[1] because they are exchanged between interacting matter particles. The Higgs boson is the only scalar elementary particle, having spin-0.

Both theoretical and experimental advances have improved our understanding of Nature, especially during the last hundred years. This has led to the formulation of the SM as a quantum field theory (QFT). Each elementary particle is viewed as an excitation of the corresponding quantum field. The Lagrangian formalism is used in QFT because the dynamics of the theory is encoded in the Lagrangian (density), $\mathcal{L}$. Equations of motion can be retrieved by means of the Euler-Lagrange equations. Lagrangian densities depend on quantum fields as well as on their derivatives with respect to Minkowski spacetime,

[1] Note that the graviton – a hypothetical particle mediating the gravitational force – has spin-2. Thus, a force carrier does not necessarily have to have spin-1.

thus incorporating our knowledge of quantum mechanics and special relativity. Furthermore, the SM is a so-called *gauge* theory. This means that the Lagrangian has to be invariant under gauge transformations coming from an underlying symmetry of Nature. It turns out that by imposing the gauge-invariance of the Lagrangian locally at each point in spacetime, the force carriers appear as a consequence of the symmetry requirement. This is why the force carriers are also referred to as *gauge* bosons. By Noether's theorem, symmetries of the Lagrangian are directly linked to conserved currents in QFTs[2].

This chapter is organized as follows. Section 2.1 gives a more detailed overview of the particle content of the SM as well as the possible interactions between these particles described in the QFT formalism. Special emphasis is given in Section 2.1.4 to the mechanism of spontaneous symmetry breaking which leads to the postulation of the Higgs boson. In Section 2.2, several shortcomings of the SM are pointed out. They motivate the search for new particles predicted in theories BSM. Section 2.3 concludes this chapter by giving an overview of BSM theories studied in this thesis.

Throughout this thesis, natural units are used

$$\hbar = c = 1 \ . \tag{2.1}$$

Hence, energies, momenta and masses have the same unit. Typical for high energy physics is the usage of electron-Volt (eV) to measure these quantities, where

$$1\,\mathrm{eV} = 1.602 \times 10^{-19}\,\mathrm{J} \ . \tag{2.2}$$

Furthermore, Einstein's summation convention is used in this chapter, meaning that repeated indices are summed over

$$a_i b^i \equiv \sum_i a_i b^i \ . \tag{2.3}$$

2.1 Particles and Interactions

An important concept of the SM are charges and conserved currents which result by imposing gauge invariance on the SM Lagrangian under certain symmetries. For instance, particles with the same electric charges repel each other. This interaction can be viewed as an exchange of a mediator particle between the two repelling particles. The mediator of electromagnetism is the photon (γ). Figure 2.1 illustrates the repulsion of two particles with the same electric charge, where approaching initial state particles move away from each other – defining the final state – after interacting with each other. In the figure the flow of electric charge is highlighted in red and is conserved at any moment in time. The change of momentum is due to the exchange of a photon between the two particles. The intermediate state is not observable and it can be easily shown by four-momentum conservation that the exchanged photon is not real. Since the photon is electrically neutral, no charged current can flow between the interacting particles.

Every fundamental interaction has an associated charge. Fermions carrying a non-zero value of that charge can interact with each other via exchange of intermediate vector bosons mediating the corresponding force. In such a case, it is said that the fermion couples to the vector boson. There are a total of twelve fermions in the SM, half of which are *quarks* and half of them are *leptons*. The difference between quarks and leptons is that quarks do interact via the strong force while leptons do not. Quarks and leptons are grouped in three *families* or *generations*, whereby the difference among the families

[2]This statement also applies to classical field theories.

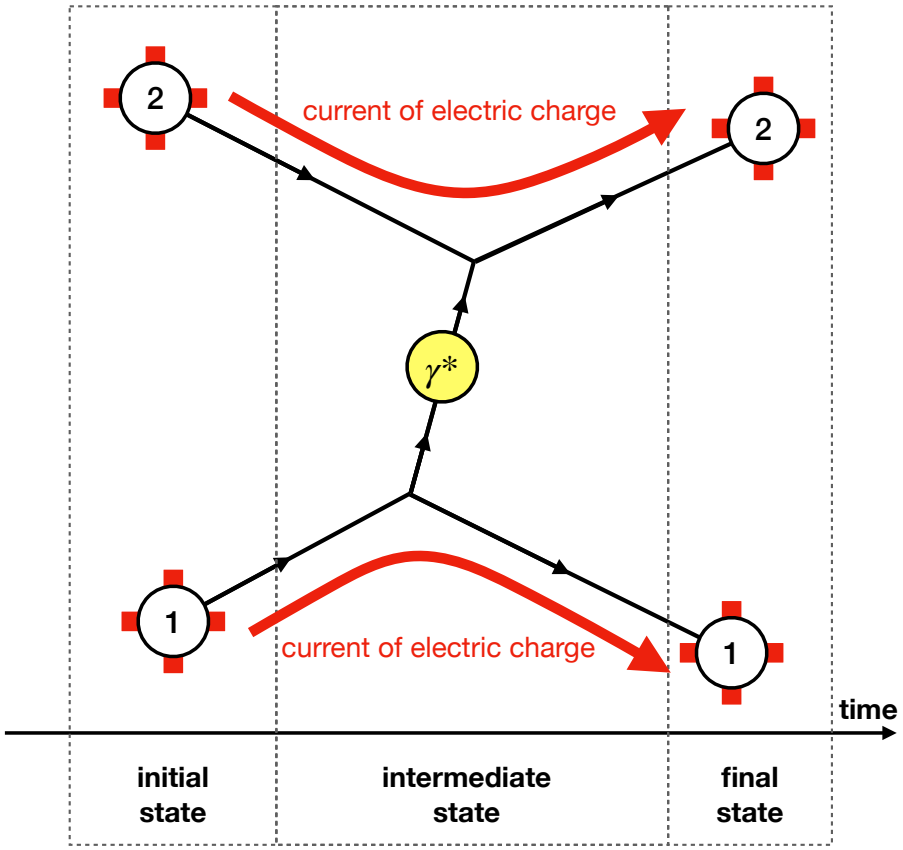


Figure 2.1: This figure illustrates schematically, how two charged particles repel each other. The two particles carry the same charge and are labeled with "1" and "2", respectively. Time flows from left to right. Momentum directions are indicated with black arrows. Initial and final state consist of the two charged particles, whereby the momentum directions in the final state have changed with respect to the initial state, i.e. the charged particle interacted. The interaction is mediated by a photon which is the fundamental force carrier of electromagnetism. Photons can be exchanged between all particles carrying electric charge. The exchange of the photon defines an intermediate state. The photon is said to be *virtual* as indicated by the asterisk because it acquires a mass to guarantee energy-momentum conservation. Virtual particles are also referred to as *off-the-mass-shell* or simply *off-shell* since they do not obey the relation $m^2 = E^2 + |\mathbf{p}|^2$, where m, E and $\mathbf{p}$ are the mass at rest, energy and momentum of the particle, respectively.

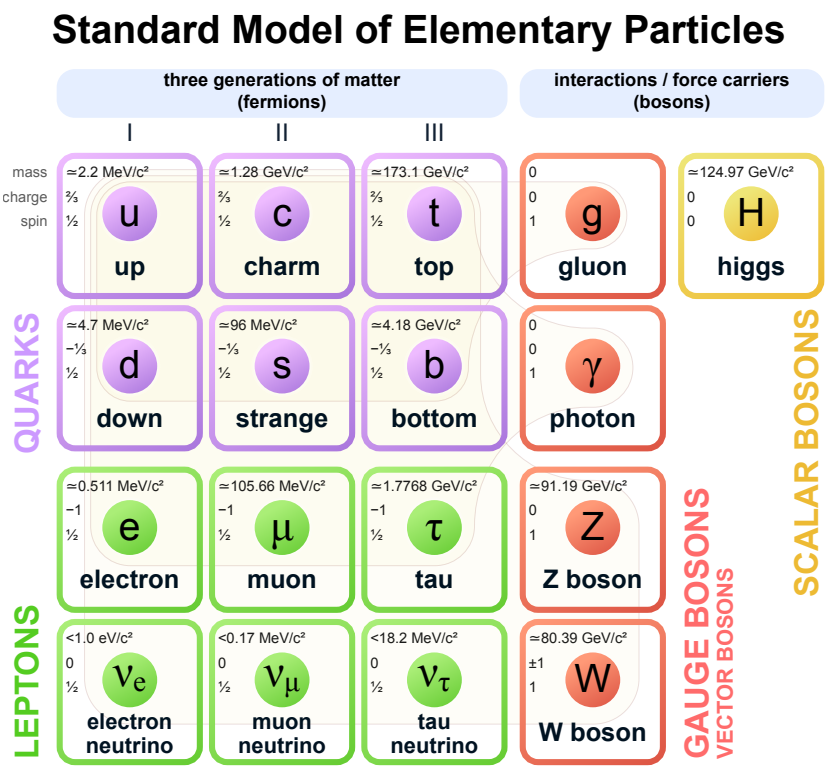


Figure 2.2: This figure is taken from [16] and shows the particle content of the SM. Mass, electric charge and spin are displayed for each elementary particle. More details about each particle are given in the text.

is the mass. Fermions of the first family are the lightest and those of the third one are the heaviest. The reason behind the copies of heavier elementary particles is unknown. Particle content together with the mass, electric charge and spin of the SM is pictorially summarized in Figure 2.2. Within the first generation of fermions are the up quark (u), the down quark (d), the electron (e^-) and the electron neutrino (ν_e). Note that quarks have fractional electric charges of $+2/3$ and $-1/3$. Quarks form bound states by constant exchange of gluons (g), the mediators of the strong interaction. The fundamental charge of the strong interaction is the *color* charge. Unlike the electric charge (e), the color charge comes in three variations – red, green and blue. Furthermore, the gluons themselves have a color charge and can interact with each other. As a consequence, it is impossible to separate two quarks from each other. Single isolated quarks have never been observed in Nature due to a phenomenon called *color confinement*. This means that only *color-neutral* states are observable.

One possibility of combining colors to arrive at a color-neutral state is by taking three quarks with color red, green and blue, respectively. Bound states of three quarks are called *baryons*. The baryon (uud) has an electric charge of unity (see the fractional electric charges of quarks in Figure 2.2) and represents the quark content of the proton. Neutrons are (udd) systems. Protons and neutrons form atomic nuclei and by adding electrons, all the elements of the periodic table can be formed. Thus, all the matter we see around us is made of fermions of the first generation.

For each matter particle of the SM, there exists also an antimatter particle. Antifermions have the same mass as their fermionic counterpart but differ in quantum numbers. For example the anti-electron is called positron (e^+) and has an electric charge of $+1e$. Another quantum number is the *lepton number* (L) and *baryon number* (B).

Leptons have $L = 1$ and their antiparticles $L = -1$. Quarks have $B = 1/3$ and antiquarks $B = -1/3$. Antiquarks carry anti-color and thus open up a second possibility to form color-neutral final states – namely by combining color and anti-color. Quark-antiquark systems are called *mesons*. The system $(\mathrm{u}\bar{\mathrm{d}})$ for instance, has an electric charge of $+1e$ and forms the π^+-meson. In principle more exotic quark-combinations such as tetraquarks [17] and pentaquarks [18] are allowed and are studied in dedicated experiments.

Coming back to the particle content of the second and third generation fermions, the strange quark (s), charm quark (c), the muon (μ^-) and the muon-neutrino (ν_μ) reside within the second generation. The third generation contains the top quark (t), the bottom quark (b), the tau lepton (τ) and the tau-neutrino (ν_τ).

Adding up the masses of the quarks making up the proton (uud) and comparing it to the proton mass of $\approx 1\,\mathrm{GeV}$ seems to be in contradiction. Barely one percent of the proton mass is due to quark masses. The remaining 99% is stored in form of binding energy that holds the quark system together. Comparing the masses of the vector bosons in Figure 2.2, it turns out that gluons and photons are massless while the W and Z bosons – the mediators of the weak interaction – are massive. As will be discussed in Section 2.1.3, it is impossible to add mass terms for the gauge bosons to the Lagrangian in a straight-forward way without breaking gauge invariance. A solution to this problem was proposed in the 1960's and uses the notion of *spontaneous symmetry breaking*. As a result a new particle was predicted – the Higgs boson – and its discovery in 2012 marks a big milestone in completing the Standard Model of elementary particles. In addition, not only the mass of W and Z bosons, but also fermion masses can be introduced by coupling them to the Higgs boson field as detailed in Section 2.1.4. The following theoretical discussion of the SM is largely based on [19, 20].

2.1.1 The Strong Interaction

The QFT describing the strong interaction – acting among quarks and gluons – is called quantum chromodynamics (QCD) [21]. The gauge symmetry group of the QCD-part in the SM Lagrangian is SU(3). Since the fundamental charge of the strong interaction is **c**olor, the underlying symmetry group is marked with a subscript "C"

$$\mathrm{SU}(3)_\mathrm{C} \ . \tag{2.4}$$

Let f be the *flavor* index running over all quark types, $f \in \{\mathrm{u}, \mathrm{d}, \mathrm{s}, \mathrm{c}, \mathrm{t}, \mathrm{b}\}$. Thus, a quark field of flavor f can be represented by

$$q_f = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} \ , \tag{2.5}$$

where $f_{1,2,3}$ are the quark fields for the three colors red, green and blue, respectively. Being a spin-½ fermion, a single quark field is described by the Dirac free field Lagrangian

$$\mathcal{L}_\mathrm{Dirac} = \overline{q}_f(\mathrm{i}\gamma^\mu\partial_\mu - m)q_f \ , \tag{2.6}$$

where γ^μ are the Dirac matrices and ∂_μ is the partial derivative with respect to the four spacetime coordinates, $\mu = 0, 1, 2, 3$. Furthermore, the mass of the quark field is denoted with m and $\overline{q}_f = {q_f}^\dagger\gamma^0$ being the Dirac adjoint.

Imposing gauge invariance on the Lagrangian of Equation (2.6) means to transform the quark field according to

$$q_f{}' = Uq_f \ , \tag{2.7}$$

where $U \in \mathrm{SU}(3)_\mathrm{C}$ and demanding that $\mathcal{L}'_\mathrm{Dirac} = \mathcal{L}_\mathrm{Dirac}$. The matrix U can be written by means of the exponential map as

$$U = \exp\left(-\mathrm{i}g_\mathrm{s}\alpha^a\lambda_a \cdot \frac{1}{2}\right) \ , \tag{2.8}$$

where the minus sign is convention and g_s and α^a are real numbers. The index a runs over the dimension of $\mathrm{SU}(3)_\mathrm{C}$, which is eight, i.e. $a = 1, 2, \cdots, 8$ and λ_a are represented by the *Gell-Mann* matrices.

So-called *local* gauge invariance is imposed by letting α^a be dependent on the spacetime coordinates (x)

$$\alpha_a \equiv \alpha_a(x) \ . \tag{2.9}$$

Under a local gauge transformation of the quark field, the Dirac Lagrangian is not invariant anymore. However, invariance can be restored by introducing new fields G^a_μ to the Lagrangian by defining the covariant derivative

$$D_\mu = \partial_\mu + \mathrm{i}g_\mathrm{s}\frac{\lambda_a}{2}G^a_\mu \ , \tag{2.10}$$

and replacing with it the partial derivative ∂_μ in Equation (2.6). It turns out that G^a_μ are the eight gluon fields that mediate the strong force. In order to ensure local gauge invariance under $\mathrm{SU}(3)_\mathrm{C}$, the gluon fields have to transform according to

$$G^{a\prime}_\mu = G^a_\mu + \partial_\mu\alpha^a(x) + g_\mathrm{s}f^a{}_{bc}\alpha^b(x)G^c_\mu \ , \tag{2.11}$$

where f_{abc} are real numbers that satisfy the commutation relation

$$\left[\frac{\lambda_a}{2}, \frac{\lambda_b}{2}\right] = \mathrm{i}f_{ab}{}^c\frac{\lambda_c}{2} \ . \tag{2.12}$$

In order to define the QCD Lagrangian, the propagation of the newly introduced gluon fields has to be added to the Dirac Lagrangian in Equation (2.6). This is achieved by adding a kinetic energy term

$$-\frac{1}{4}G^{\mu\nu}_a G^a_{\mu\nu} \ , \tag{2.13}$$

with the gluon tensor $(G^{\mu\nu}_a)$ defined as

$$G^a_{\mu\nu} = \partial_\mu G^a_\nu - \partial_\nu G^a_\mu - g_\mathrm{s}f^a{}_{bc}G^b_\mu G^c_\nu \ . \tag{2.14}$$

The QCD Lagrangian reads

$$\mathcal{L}_\mathrm{QCD} = \sum_f \left(\mathrm{i}\bar{q}_f\gamma^\mu\partial_\mu q_f - m\bar{q}_f q_f - g_\mathrm{s}\bar{q}_f\gamma^\mu\frac{\lambda_a}{2}q_f G^a_\mu\right) - \frac{1}{4}G^{\mu\nu}_a G^a_{\mu\nu} \ . \tag{2.15}$$

The three terms inside the sum can be interpreted in the following way. The first term describes the free propagation of the quark fields q_f, while the second represents the mass term. Imposing local gauge invariance under $\mathrm{SU}(3)_\mathrm{C}$ and introducing the covariant derivative (see Equation (2.10)), results in an interaction between gluon and quark fields encoded in the third term inside the sum. The strength of this interaction is quantified with g_s which is typically re-parametrized by

$$\alpha_\mathrm{s} = \frac{g^2_\mathrm{s}}{4\pi} \ , \tag{2.16}$$

the *strong coupling constant*. To be more precise, α_s is a *running* coupling constant[3], i.e. it varies between different energy scales. At high energy scales, α_s takes up small values and quarks and gluons are quasi-free. Contrary, for lower energies, corresponding to large length scales, α_s becomes large. This behavior of α_s is ultimately a consequence of the self-interaction of the gluon fields. Cubic and quartic self-interactions appear upon expansion of the last term in Equation (2.15) which describe the propagation of the gluon fields. Note that gluons must be massless since adding a term of the form

$$m^2 G^a_\mu G^\mu_a \ , \tag{2.17}$$

would violate local gauge invariance under $\mathrm{SU(3)_C}$ of $\mathcal{L}_{\mathrm{QCD}}$.

2.1.2 Lie Groups and Representations

It is worth to pause for moment and elaborate a bit on the group theoretical aspects the SM is based on. Some of them have already appeared in the previous section. The gauge symmetry groups of the SM directly impact the resulting QFTs. One important consequence is that emerging QFTs based on local gauge symmetries are unitary. This is crucial to preserve inner products and thus the probability interpretation of quantum mechanics. More details on using group theory in QFTs can be found in references [22, 23] upon which also the following discussion is based.

Every gauge symmetry group, G, used in the formulation of the SM is a *Lie group*, i.e. it has an infinite number of elements and is at same time a differential manifold. Any group element, $U \in G$, can be written as shown in Equation (2.8), namely

$$U = \exp\left(\mathrm{i}\theta^a T_a\right) \mathbb{1} \ , \tag{2.18}$$

where θ^a are numbers and T_a are called the (group) generators. These group generators form a *Lie algebra* ($\mathfrak{g}$) which can be thought of as the tangent space at the identity ($\mathbb{1}$) of the corresponding Lie group. The Lie algebra is equipped with a map called *Lie bracket*

$$\begin{aligned} &\mathfrak{g} \times \mathfrak{g} \to \mathfrak{g} \\ &(u, v) \mapsto [u, v] \ . \end{aligned} \tag{2.19}$$

Furthermore, the Lie bracket has to satisfy the Jacobi identity

$$[u, [v, w]] + [v, [w, u]] + [w, [u, v]] = 0 \ . \tag{2.20}$$

For the group generators, the Lie bracket yields

$$[T_a, T_b] = \mathrm{i} f_{ab}{}^c T_c \ , \tag{2.21}$$

where f_{abc} are known as the structure constants. In case of $\mathrm{SU(3)_C}$ the generators are given by the Gell-Mann matrices and the Lie bracket is given by the commutation relation shown in Equation (2.12). Lie groups with vanishing structure constants are called *Abelian*, those with $f_{abc} \neq 0$ are *non-Abelian*.

Let us focus for the rest of the discussion on the Lie group $G = \mathrm{SU}(N)$. The prescription that maps each element, $U \in \mathrm{SU}(N)$, to an $N \times N$ unitary matrix with determinant 1, i.e. the identity map, defines a representation on the vector space spanned by vectors

[3] The strong coupling constant is not the only example of a running coupling constant. The fine structure constant appearing in quantum electro-dynamics or the weak coupling constant – quantifying the strength of processes involving the weak interaction – are other examples. Further note, that different running coupling constants have a different functional dependence on the probed energy scale.

of length N. This representation is called the *defining* or *fundamental* representation. For example, the quark field in Equation (2.5) is in the fundamental representation of $\mathrm{SU}(3)_\mathrm{C}$. The fundamental representation is commonly denoted by **N**, e.g. the quark field, q_f, resides in the **3**-representation of $\mathrm{SU}(3)_\mathrm{C}$. Anti-quarks follow the transformation rule

$$\overline{q}_f \to \overline{q}_f U^\dagger \ , \tag{2.22}$$

corresponding to the *anti-fundamental* representation $\overline{\mathbf{3}}$. Quark fields are referred to as $\mathrm{SU}(3)_\mathrm{C}$-*triplets* because they are written as three-component vector in the (anti-) fundamental representation.

Another important representation of $\mathrm{SU}(N)$ is the *trivial* representation denoted as **1**. In this case all elements $U \in \mathrm{SU}(N)$ are mapped to the identity. In other words, all fields in the trivial representation stay invariant under a $\mathrm{SU}(N)$ transformation. For example, lepton fields do not carry color charge and are thus not affected by the strong interaction. Thus, from a group theoretical point of view, they transform as *singlets* in the **1** representation of $\mathrm{SU}(3)_\mathrm{C}$.

A last representation of $\mathrm{SU}(N)$, relevant for the SM, is the *adjoint* representation. Here, the representation acts directly on its Lie algebra viewed as a vector space. Since $\mathrm{SU}(3)_\mathrm{C}$ has 8 generators, its adjoint representation has dimension eight and is denoted by **8**. It turns out that the gauge fields transform in the adjoint representation, e.g. the eight gluon fields in case of $\mathrm{SU}(3)_\mathrm{C}$.

2.1.3 The Electroweak Interaction

The electroweak interaction is a unified theory describing both the electromagnetic and the weak force based on the underlying symmetry group $\mathrm{SU}(2) \times \mathrm{U}(1)$. This theory was developed by Glashow [24], Weinberg [25] and Salam [26], who were attributed the physics Nobel prize for this achievement in 1979. While electromagnetic interactions are mediated by massless photons, the weak interaction is mediated by the massive W and Z bosons (see Figure 2.2). The range of the electromagnetic force is infinite. In contrast, the weak interaction is of short range because the W and Z bosons are massive. The weak interaction is the only known interaction to violate *parity*[4] [27]. Furthermore, the charge conjugation parity (CP) symmetry is also violated in weak interactions [28]. Charge conjugation transforms a particle into its anti-particle and vice versa. The unification of electromagnetism and the weak interaction – being very different at first sight – is reviewed in the following.

Parity violation in so-called *weak charged current* interactions – mediated by $\mathrm{W}^\pm$ bosons – is incorporated into the theory by splitting each fermion field (ψ) into left-handed (ψ_L) and right-handed (ψ_R) *chirality* states. Left-handed fermion fields form doublets with respect to the SU(2) symmetry of the weak interaction, while right-handed fermions transform as singlets. Therefore, the underlying symmetry group of the weak interaction carries an extra subscript "L" and will be denoted from now on with

$$\mathrm{SU}(2)_\mathrm{L} \ . \tag{2.23}$$

The operators projecting ψ onto its different chirality states are given by

$$\begin{aligned} \psi_\mathrm{L} &= P_\mathrm{L}\psi = \frac{1}{2}(1-\gamma^5)\psi \\ \psi_\mathrm{R} &= P_\mathrm{R}\psi = \frac{1}{2}(1+\gamma^5)\psi \\ \psi &= \psi_\mathrm{L} + \psi_\mathrm{R} \ , \end{aligned} \tag{2.24}$$

[4] Parity is the symmetry transformation inverting the spatial coordinates.

where $\gamma^5 = \mathrm{i}\gamma^0\gamma^1\gamma^2\gamma^3$ is the product of Dirac matrices. The fermion doublets are then given by

$$\begin{pmatrix}\nu_\mathrm{e}\\ e\end{pmatrix}_\mathrm{L}, \begin{pmatrix}\nu_\mu\\ \mu\end{pmatrix}_\mathrm{L}, \begin{pmatrix}\nu_\tau\\ \tau\end{pmatrix}_\mathrm{L}, \begin{pmatrix}u\\ d'\end{pmatrix}_\mathrm{L}, \begin{pmatrix}c\\ s'\end{pmatrix}_\mathrm{L}, \begin{pmatrix}t\\ b'\end{pmatrix}_\mathrm{L} , \tag{2.25}$$

where the quark fields that are eigenstates of the weak interaction differ from the mass eigenstates. The convention chosen here is that weak and mass eigenstates coincide for up-type quarks but are rotated for down-type quarks. Therefore, down-type quark fields are denoted with an extra prime as superscript. Details of this rotation are discussed later in this section.

A new quantum number – the *weak* isospin T – is introduced with projection $T_3 = \pm 1/2$. For example, ν_eL is a weak isospin *up* state with $T_3 = +1/2$. All right-handed fermions are weak isospin singlets ($T = 0$)

$$e_\mathrm{R}, \mu_\mathrm{R}, \tau_\mathrm{R}, u_\mathrm{R}, d'_\mathrm{R}, c_\mathrm{R}, s'_\mathrm{R}, t_\mathrm{R}, b'_\mathrm{R} . \tag{2.26}$$

Right-handed neutrinos are not part of the SM because they transform as singlets with respect to every gauge symmetry and thus do not interact by any force described by the SM.

In order to simplify the notation,

$$\Psi^i = \begin{pmatrix}\psi^i\\ \psi'^{,i}\end{pmatrix}_\mathrm{L} = \begin{pmatrix}\psi^i_\mathrm{L}\\ \psi'^{,i}_\mathrm{L}\end{pmatrix} \tag{2.27}$$

covers all combinations shown in Equation (2.25), where $i = 1, 2, 3$ runs over the three generations of fermions. Similarly, ψ^i_R and $\psi'^{,i}_\mathrm{R}$ are used to refer to the fields in Equation (2.26). The free-field Dirac Lagrangian (see Equation (2.6)) reads for the weak interaction

$$\begin{aligned}\mathcal{L}_\mathrm{Dirac} = \sum_{l,q}\sum_i &\overline{\Psi^i}(\mathrm{i}\gamma^\mu\partial_\mu)\Psi^i + \overline{\psi^i_\mathrm{R}}(\mathrm{i}\gamma^\mu\partial_\mu)\psi^i_\mathrm{R} + \overline{\psi'^{,i}_\mathrm{R}}(\mathrm{i}\gamma^\mu\partial_\mu)\psi'^{,i}_\mathrm{R}\\ &- m^i(\overline{\psi^i_\mathrm{R}}\psi^i_\mathrm{L} + \overline{\psi^i_\mathrm{L}}\psi^i_\mathrm{R}) - m'^{,i}(\overline{\psi'^{,i}_\mathrm{R}}\psi'^{,i}_\mathrm{L} + \overline{\psi'^{,i}_\mathrm{L}}\psi'^{,i}_\mathrm{R}) ,\end{aligned} \tag{2.28}$$

where the masses of up-type and down-type fermions are denoted as m and m', respectively. The summation runs over all leptons (l) and quarks (q) of all three generations $i = 1, 2, 3$. From the second line of Equation (2.28), it is obvious that mass terms in the Dirac Lagrangian are not gauge invariant since they involve left-handed and right-handed fields which transform differently under $\mathrm{SU}(2)_\mathrm{L}$. In order to proceed with the isodoublet model of the weak interaction, fermions are treated as massless and it will be discussed in Section 2.1.4 how fermion masses can be re-introduced. Local gauge invariance under $\mathrm{SU}(2)_\mathrm{L}$ is imposed by transforming the left-handed fields according to

$$\begin{aligned}&\Psi^{i'} = U\Psi^i \text{ with}\\ &U \in \mathrm{SU}(2)_\mathrm{L}\\ &U = \exp\left(-\mathrm{i}g_\mathrm{w}\alpha^a(x)\sigma_a\frac{1}{2}\right) ,\end{aligned} \tag{2.29}$$

where σ_a ($a = 1, 2, 3$) are the Pauli matrices (see also Equation (2.8)). The generators of SU(2) are given by $\sigma_a/2$ satisfying the commutation relations

$$\left[\frac{\sigma_a}{2}, \frac{\sigma_b}{2}\right] = \mathrm{i}\epsilon_{ab}{}^c\frac{\sigma_c}{2} , \tag{2.30}$$

with ϵ_{abc} – the Levi-Civita symbol – being the structure constants. Three new gauge fields (W_μ^a) are introduced and added to the covariant derivative

$$D_\mu = \partial_\mu + \mathrm{i}g_\mathrm{w}\frac{\sigma_a}{2}W_\mu^a \ , \tag{2.31}$$

with g_w being the coupling strength of the weak interaction. Inserting Equation (2.31) into Equation (2.28) and neglecting the mass terms, results in

$$\mathcal{L}_\mathrm{Dirac} = \sum_{l,q}\sum_i \overline{\Psi^i}(\mathrm{i}\gamma^\mu\partial_\mu)\Psi^i + \overline{\psi_\mathrm{R}^i}(\mathrm{i}\gamma^\mu\partial_\mu)\psi_\mathrm{R}^i + \overline{\psi_\mathrm{R}^{\prime,i}}(\mathrm{i}\gamma^\mu\partial_\mu)\psi_\mathrm{R}^{\prime,i} - g_\mathrm{w}\overline{\Psi^i}\gamma^\mu\frac{\sigma_a}{2}\Psi^i W_\mu^a \ , \tag{2.32}$$

where the last term describes interactions of left-handed fermions and the three gauge fields W_μ^a. Two of the gauge fields can be related to the physical $\mathrm{W}^\pm$ bosons via

$$W_\mu^\pm = \frac{1}{\sqrt{2}}(W_\mu^1 \mp \mathrm{i}W_\mu^2) \ . \tag{2.33}$$

However, the remaining field, W_μ^3, can not be identified as the Z boson. It would have the same vector-axial structure as the charged current interaction mediated by $\mathrm{W}^\pm$ bosons which contradicts with experimental observations of neutral current weak interactions [29, 30]. In conclusion, $\mathrm{SU(2)_L}$ can not be the gauge group of weak interaction, motivating the work of unifying it with another interaction.

The QFT describing the electromagnetic force is termed quantum electrodynamics (QED). Its Lagrangian is given by

$$\mathcal{L}_\mathrm{QED} = \mathrm{i}\overline{\psi}\gamma^\mu\partial_\mu\psi - m\overline{\psi}\psi - e\overline{\psi}\gamma^\mu Q\psi A_\mu - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} \ , \tag{2.34}$$

where ψ is a fermion field and A_μ represents the photon field. The electric charge is denoted by e, the fermion mass by m and Q is the quantum number associated to the electromagnetic interaction indicating what fraction of the electric charge the fermion described by ψ carries. The electromagnetic field tensor is given by

$$F_{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu \ . \tag{2.35}$$

Photon-fermion interactions are described by the third term in Equation (2.34) which are incorporated by the covariant derivative

$$D_\mu = \partial_\mu + \mathrm{i}eQA_\mu \ , \tag{2.36}$$

based on a $\mathrm{U(1)_{em}}$ symmetry. Unlike the gauge theories behind the strong and weak interaction, QED is an Abelian theory. As a consequence, photons do not interact with each other.

The electromagnetic current for an electron field (e) is defined as

$$j_\mathrm{em}^\mu = \overline{e}\gamma^\mu Q e \ . \tag{2.37}$$

The e-ν_e-part of the third component charged weak current is given by

$$j_3^\mu = {}^1\!/_2(\overline{\nu_\mathrm{eL}}\gamma^\mu)\nu_\mathrm{eL} - \overline{e_\mathrm{L}}\gamma^\mu e_\mathrm{L} \ . \tag{2.38}$$

The difference of these two currents is

$$j_\mathrm{em}^\mu - j_3^\mu = -\frac{1}{2}\overline{\begin{pmatrix}\nu_\mathrm{e}\\ e\end{pmatrix}_\mathrm{L}}\gamma^\mu\begin{pmatrix}\nu_\mathrm{e}\\ e\end{pmatrix}_\mathrm{L} - \overline{e_\mathrm{R}}\gamma^\mu e_\mathrm{R} \ . \tag{2.39}$$

Equation (2.39) can be interpreted as a new current acting on all fields, i.e. it is acting on weak isospin doublets and singlets. The doublet is weighted with a factor $-1/2$ and the electron singlet with -1. These appropriate weights can be obtained by extending $\mathrm{SU(2)_L}$ with a U(1) symmetry group. This is not $\mathrm{U(1)_{em}}$ from QED (see Equation (2.34)) but a new group $\mathrm{U(1)_Y}$, where Y is the *weak hypercharge*. In terms of the current of the weak hypercharge, j^μ_Y, Equation (2.39) can be re-written as

$$j^\mu_{\mathrm{em}} - j^\mu_3 = \frac{1}{2} j^\mu_Y \ , \tag{2.40}$$

which translates into a relation[5] between Q, T_3 and Y

$$Q = T_3 + \frac{Y}{2}. \tag{2.41}$$

The new $\mathrm{U(1)_Y}$ group is described by the Lagrangian

$$\mathcal{L}_Y = \mathrm{i}\overline{\psi}\gamma^\mu D_\mu \psi - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} \ , \tag{2.42}$$

whereby the mass terms are dropped. The covariant derivative in Equation (2.42) is defined by

$$D_\mu = \partial_\mu + \mathrm{i} g_{\mathrm{Y}} \frac{Y}{2} B_\mu \ , \tag{2.43}$$

introducing a coupling of fermion fields with the gauge field B_μ of strength g_{Y}.

Since both, W^3_μ and B_μ couple to neutrinos, they can not be identified with the photon field A_μ because neutrinos do not have electric charge. However, a rotation of the fields W^3_μ and B_μ defines correctly the physical fields related to the Z boson and the photon. This rotation is parametrized with the Weinberg angle (θ_{w}) and is given by

$$\begin{pmatrix} A_\mu \\ Z_\mu \end{pmatrix} = \begin{pmatrix} \cos\theta_{\mathrm{w}} & \sin\theta_{\mathrm{w}} \\ -\sin\theta_{\mathrm{w}} & \cos\theta_{\mathrm{w}} \end{pmatrix} \begin{pmatrix} B_\mu \\ W^3_\mu \end{pmatrix} \ . \tag{2.44}$$

Working out the math reveals

$$e = g_{\mathrm{Y}} \cos\theta_{\mathrm{w}} = g_{\mathrm{w}} \sin\theta_{\mathrm{w}} \ , \tag{2.45}$$

meaning all couplings $(e, g_{\mathrm{w}}, g_{\mathrm{Y}})$ to the gauge bosons within the unified electroweak theory $\mathrm{SU(2)_L \times U(1)_Y}$ are determined once two parameters are known – for example e and θ_{w}.

Collecting the pieces described in this section, the Lagrangian of the electroweak interaction can be formulated. It is given by

$$\begin{aligned} \mathcal{L}_{\mathrm{EWK}} &= \sum_{l,q} \sum_i \left(\overline{\Psi^i} (\mathrm{i}\gamma^\mu D^{(\mathrm{L})}_\mu) \Psi^i + \overline{\psi^i_{\mathrm{R}}} (\mathrm{i}\gamma^\mu D^{(\mathrm{R})}_\mu) \psi^i_{\mathrm{R}} + \overline{\psi'^{,i}_{\mathrm{R}}} (\mathrm{i}\gamma^\mu D^{(\mathrm{R})}_\mu) \psi'^{,i}_{\mathrm{R}} \right) + \mathcal{L}^{\mathrm{gauge}}_{\mathrm{EWK}} \\ \mathcal{L}^{\mathrm{gauge}}_{\mathrm{EWK}} &= -\frac{1}{4} W^a_{\mu\nu} W^{\mu\nu}_a - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} \ . \end{aligned} \tag{2.46}$$

Different covariant derivatives are used depending on whether they act on weak isospin doublets or singlets

$$\begin{aligned} D^{(\mathrm{L})}_\mu &= \partial_\mu + \mathrm{i} g_{\mathrm{w}} \frac{\sigma_a}{2} W^a_\mu + \mathrm{i} g_Y \frac{Y}{2} B_\mu \\ D^{(\mathrm{R})}_\mu &= \partial_\mu + \mathrm{i} g_Y \frac{Y}{2} B_\mu \ . \end{aligned} \tag{2.47}$$

[5] This relation looks identical to the Gell-Mann–Nishijima formula known from the quark model. This is because much of the group theoretical consideration apply also to the electroweak unification.

The relation between the gauge bosons W^a_μ and B_μ, and the physical fields of charged $W^\pm$ bosons, Z boson and photon are given in Equations (2.33) and (2.44). The term $\mathcal{L}^{\text{gauge}}_{\text{EWK}}$ in Equation (2.46) accounts for the free propagation of the electroweak gauge boson fields. Bosonic tensors are given by

$$\begin{aligned} W^a_{\mu\nu} &= \partial_\mu W^a_\nu - \partial_\nu W^a_\mu - g_{\text{w}} \epsilon^a{}_{bc} W^b_\mu W^c_\nu \\ B^a_{\mu\nu} &= \partial_\mu B^a_\nu - \partial_\nu B^a_\mu \ . \end{aligned} \tag{2.48}$$

Re-writing $\mathcal{L}^{\text{gauge}}_{\text{EWK}}$ in terms of the physical fields, reveals possible trilinear (ZW^+W^-, γW^+W^-) and quadrilinear (ZZW^+W^-, $\gamma\gamma W^+W^-$, γZW^+W^-, $W^+W^-W^+W^-$) couplings among them.

In summary, the electroweak Lagrangian shown in Equation (2.46) correctly encapsulates QED and the weak interaction. It accounts for the free propagation of fermions of all three generations, the free propagation of the physical bosons $W^\pm$, Z and γ, and the interaction between fermions and bosons as well as interactions among the bosons. Mass terms for bosons or fermions would both violate gauge invariance of $\mathcal{L}_{\text{EWK}}$ and are thus dropped from the formalism. This is however in contradiction with experimental evidence, observing massive fermions as well as W and Z bosons. How massive particles are restored, is subject of the next section.

2.1.4 Electroweak Symmetry Breaking

In QFT, a system has a *spontaneously broken* symmetry if the Lagrangian describing its dynamics is invariant under these symmetry transformations but its ground state is not. In 1964, Englert and Brout [31], Higgs [32] and Guralnik, Hagen and Kibble [33] proposed independently a mechanism to generate gauge boson masses based on spontaneous symmetry breaking. The mechanism will be referred to as Brout-Englert-Higgs (BEH) mechanism in the following. In terms of the underlying symmetry group of the SM, the BEH mechanism can be viewed as the following transition

$$\text{SU}(3)_{\text{C}} \times \text{SU}(2)_{\text{L}} \times \text{U}(1)_{\text{Y}} \rightarrow \text{SU}(3)_{\text{C}} \times \text{U}(1)_{\text{em}} \ . \tag{2.49}$$

Hence, after spontaneous symmetry breaking, the photon – as well as the gluons – should remain massless and only the W and Z bosons should acquire a mass.

The BEH mechanism is realized upon addition of a new field, ϕ. This field must have non-vanishing weak isospin and weak hypercharge in order to account for electroweak symmetry breaking as shown in Equation (2.49). A simple choice is in form of a complex weak isospin doublet

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + \text{i}\phi_2 \\ \phi_3 + \text{i}\phi_4 \end{pmatrix} \ , \tag{2.50}$$

where ϕ_i $(i = 1, 2, 3, 4)$ are real scalar fields. The Lagrangian of spontaneous symmetry breaking of the electroweak theory is given by

$$\begin{aligned} \mathcal{L}_{\text{BEH}} &= (D^{(\text{L})}_\nu \phi)^\dagger (D^{\nu(\text{L})} \phi) - V(\phi) \\ V(\phi) &= \mu^2 \phi^\dagger \phi + \lambda (\phi^\dagger \phi)^2 \quad \mu^2 < 0, \quad \lambda > 0 \ . \end{aligned} \tag{2.51}$$

Since ϕ carries both weak isospin and weak hypercharge, the covariant derivative $(D^{(\text{L})}_\nu)$ from Equation (2.47) is used above. It is worth mentioning that $\mathcal{L}_{\text{BEH}} + \mathcal{L}^{\text{gauge}}_{\text{EWK}}$ (see Equation (2.46)) is gauge invariant. The real parameters of the Higgs boson potential $(V(\phi))$ are chosen such that the potential is bound from below $(\lambda > 0)$ and such that the

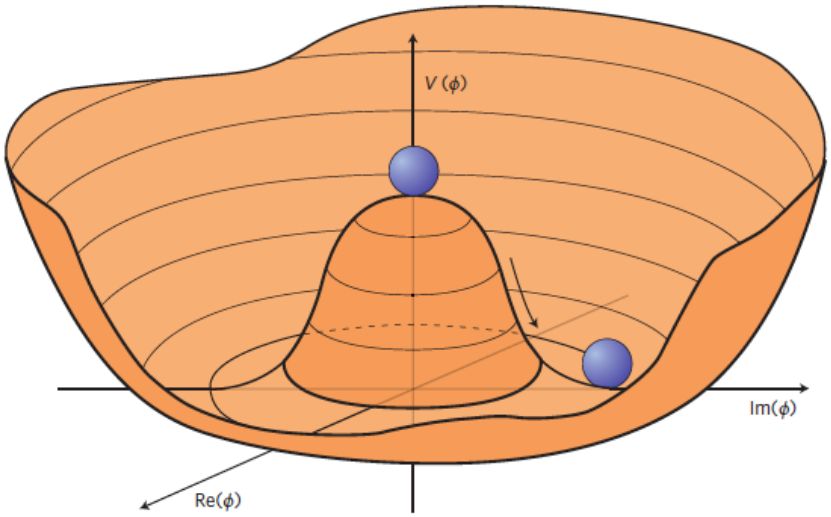


Figure 2.3: The figure is taken from [34] and shows the Higgs boson potential $V(\phi)$ (see Equation (2.51)). While the potential remains invariant under $U(1)$ transformation, the ground state (blue ball) is not. In fact, there are infinitely many possibilities to choose a ground state, whereby one is chosen *spontaneously*.

ground state is not invariant under $\mathrm{SU(2)_L \times U(1)_Y}$ transformations ($\mu^2 < 0$) anymore. The minimum of the potential is realized if

$$\phi^\dagger \phi = \frac{-\mu^2}{2\lambda} \equiv \frac{v^2}{2} \ , \tag{2.52}$$

where the field, ϕ, is treated classically and v is the so-called vacuum expectation value (vev). A QFT version of the above equation is given by

$$\langle 0|\phi^\dagger \phi|0\rangle = \frac{-\mu^2}{2\lambda} \equiv \frac{v^2}{2} \ . \tag{2.53}$$

Figure 2.3 shows a sketch of the Higgs boson potential.

There are infinitely many ground states which are positioned along a circle of the potential. Upon a specific choice of the ground state the $\mathrm{SU(2)_L \times U(1)_Y}$ symmetry is broken as shall be demonstrated in the following. A simple choice of the ground state is

$$\phi_{\text{ground}} = \begin{pmatrix} 0 \\ \frac{v}{\sqrt{2}} \end{pmatrix} \ . \tag{2.54}$$

This particular choice of the ground state satisfies Equation (2.53) and vanishes for the charged ϕ^+-component, meaning that the ground state is electrically neutral. It is straightforward to check that ϕ_{ground} is not invariant under *global* $\mathrm{SU(2)_L \times U(1)_Y}$ transformations. Invariance means

$$\begin{gathered} \exp\left(-\mathrm{i}\alpha\Lambda\right)\phi_{\text{ground}} = \phi_{\text{ground}} \Rightarrow \Lambda\phi_{\text{ground}} = 0 \\ \Lambda \in \{\frac{\sigma_1}{2}, \frac{\sigma_2}{2}, \frac{\sigma_3}{2}, \frac{Y}{2}\} \ , \end{gathered} \tag{2.55}$$

but instead the following holds

$$\begin{aligned}
\frac{\sigma_1}{2}\phi_{\text{ground}} &= \frac{1}{2}\begin{pmatrix}0 & 1\\ 1 & 0\end{pmatrix}\begin{pmatrix}0\\ \frac{v}{\sqrt{2}}\end{pmatrix} = \begin{pmatrix}\frac{v}{2\sqrt{2}}\\ 0\end{pmatrix} \neq 0\\
\frac{\sigma_2}{2}\phi_{\text{ground}} &= \frac{1}{2}\begin{pmatrix}0 & -\mathrm{i}\\ \mathrm{i} & 0\end{pmatrix}\begin{pmatrix}0\\ \frac{v}{\sqrt{2}}\end{pmatrix} = \begin{pmatrix}-\mathrm{i}\frac{v}{2\sqrt{2}}\\ 0\end{pmatrix} \neq 0\\
\frac{\sigma_3}{2}\phi_{\text{ground}} &= \frac{1}{2}\begin{pmatrix}1 & 0\\ 0 & -1\end{pmatrix}\begin{pmatrix}0\\ \frac{v}{\sqrt{2}}\end{pmatrix} = \begin{pmatrix}0\\ -\frac{v}{2\sqrt{2}}\end{pmatrix} \neq 0\\
\frac{Y}{2}\phi_{\text{ground}} &= \frac{1}{2}\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}\begin{pmatrix}0\\ \frac{v}{\sqrt{2}}\end{pmatrix} = \begin{pmatrix}0\\ \frac{v}{2\sqrt{2}}\end{pmatrix} \neq 0\ .
\end{aligned} \tag{2.56}$$

However, for the electric charge $Q = {}^{\sigma_3}\!/_2 + {}^{Y}\!/_2$ (see Equation (2.41)), it follows

$$Q\phi_{\text{ground}} = \frac{1}{2}(\sigma_3 + Y)\phi_{\text{ground}} = \frac{\sigma_3}{2}\phi_{\text{ground}} + \frac{Y}{2}\phi_{\text{ground}} \overset{\text{Eq. 2.56}}{=} 0\ , \tag{2.57}$$

i.e. ϕ_{ground} respects the $\mathrm{U}(1)_{\text{em}}$ symmetry and hence the symmetry breaking

$$\mathrm{SU}(2)_{\mathrm{L}} \times \mathrm{U}(1)_{\mathrm{Y}} \to \mathrm{U}(1)_{\text{em}}\ . \tag{2.58}$$

In a next step, the Higgs boson doublet, ϕ (see Equation (2.50)), can be expanded around the ground state. For that, let

$$\phi_3 = v + h(x)\ , \tag{2.59}$$

where $h(x)$ has a vanishing vev

$$\langle 0|h(x)|0\rangle = 0\ . \tag{2.60}$$

The Higgs boson doublet takes then the following form

$$\phi = \frac{1}{\sqrt{2}}\exp\left(\frac{\mathrm{i}\xi^a(x)\sigma_a}{v}\right)\begin{pmatrix}0\\ v + h(x)\end{pmatrix}\ , \tag{2.61}$$

which upon linear expansion[6] is equivalent to Equation (2.50). The fields $\xi^a(x)$ are massless and a consequence of the Goldstone theorem [35]. However, they do not correspond to physical fields. This can be easily seen by applying the local gauge invariance principle and choosing

$$U = \exp\left(-\frac{\mathrm{i}\xi^a(x)\sigma_a}{v}\right) \tag{2.62}$$

by setting the fields $\alpha^a(x)$ in Equation (2.29) appropriately. As a result the Higgs boson doublet looks like

$$\phi = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\ v + h(x)\end{pmatrix}\ . \tag{2.63}$$

This particular choice is referred to as the *unitary* gauge. The lower component of the Higgs boson doublet in Equation (2.63) defines

$$H(x) \equiv v + h(x)\ , \tag{2.64}$$

having a vev of

$$\langle 0|H(x)|0\rangle = v\ . \tag{2.65}$$

[6] The identification is given by $\xi^1 = \phi_2$, $\xi^2 = \phi_1$ and $\xi^3 = -\phi_4$.

This means, the field H is present in the vacuum of the SM and it is constant. Its quanta are represented by excitations of the vacuum with the field h associated to the particle called *Higgs boson*.

Inserting Equation (2.63) into Equation (2.51), the following expression can be obtained for the Lagrangian

$$\begin{aligned}
\mathcal{L}_{\text{BEH}}+\mathcal{L}_{\text{EWK}}^{\text{gauge}} = & \\
& + \frac{1}{2}\left(\partial_\mu h \partial^\mu h + 2\mu^2 h^2\right) \\
& - \frac{1}{4}\left(W_{\mu\nu}^-\right)^\dagger W^{-\mu\nu} + \frac{1}{2}\left(\frac{g_\text{w} v}{2}\right)^2 \left(W_\mu^-\right)^\dagger W^{-\mu} \\
& - \frac{1}{4}\left(W_{\mu\nu}^+\right)^\dagger W^{+\mu\nu} + \frac{1}{2}\left(\frac{g_\text{w} v}{2}\right)^2 \left(W_\mu^+\right)^\dagger W^{+\mu} \\
& - \frac{1}{4} Z_{\mu\nu} Z^{\mu\nu} + \frac{1}{2}\left(\frac{g_\text{w} v}{2\cos\theta_\text{w}}\right) Z_\mu Z^\mu \\
& - \frac{1}{4} A_{\mu\nu} A^{\mu\nu} \\
& + \frac{g_\text{w}^2 v}{2} h W_\mu^- W^{+\mu} + \frac{g_\text{w}^2}{4} h^2 W_\mu^- W^{+\mu} + \frac{g_\text{w}^2 v}{4\cos^2\theta_\text{w}} h Z_\mu Z^\mu + \frac{g_\text{w}^2}{8\cos^2\theta_\text{w}} h^2 Z_\mu Z^\mu \\
& + \frac{\mu^2}{v} h^3 + \frac{\mu^2}{4v^2} h^4 \ .
\end{aligned} \tag{2.66}$$

The following conclusions can be drawn from Equation (2.66):

- Applying the Euler-Lagrange equation to line one of Equation (2.66), yields the Klein-Gordon equation where the Higgs boson mass can be identified as

$$m_\text{H} = \sqrt{2}|\mu| \ . \tag{2.67}$$

- Applying the Euler-Lagrange equation to line two of Equation (2.66), gives the Procca equation describing a massive particle with spin-1. The mass of the negatively charged W boson can be identified as

$$m_{\text{W}^-} = \frac{g_\text{w} v}{2} \ . \tag{2.68}$$

  In fact, the positively charged W boson has the same mass when looking at line three in Equation (2.66)

$$m_{\text{W}^+} = \frac{g_\text{w} v}{2} \equiv m_\text{W} \ . \tag{2.69}$$

  Similarly, the Z boson mass can be inferred from line four in Equation (2.66)

$$m_\text{Z} = \frac{g_\text{w} v}{2\cos\theta_\text{w}} \ . \tag{2.70}$$

- As expected, line five in Equation (2.66) shows no mass term related to the photon field A_μ and hence

$$m_\gamma = 0 \ . \tag{2.71}$$

- In the sixth line of Equation (2.66), trilinear and quadrilinear interaction between the Higgs boson and massive gauge boson fields appear. In the last line of Equation (2.66), Higgs boson self-couplings appear. The coupling of the Higgs boson to the W bosons is given by

$$\lambda_\text{HWW} \overset{\text{Eq. 2.66}}{=} \frac{g_\text{w}^2 v}{2} \overset{\text{Eq. 2.69}}{=} \frac{2}{v} m_\text{W}^2 \ , \tag{2.72}$$

and similarly for the coupling to the Z boson

$$\lambda_{\mathrm{HZZ}} \overset{\text{Eq. 2.66}}{=} \frac{g_{\mathrm{w}}^2 v}{4cos^2\theta_{\mathrm{w}}} \overset{\text{Eq. 2.70}}{=} \frac{2}{v} m_{\mathrm{Z}}^2 \ . \tag{2.73}$$

In both cases the coupling of the Higgs boson to the massive gauge boson is proportional to their mass square.

It is instructive to count the number of degrees of freedom before and after electroweak symmetry breaking. At the beginning, there were four massless gauge bosons (3 W_μ^a and 1 B_μ) with each having two possible polarizations. The Higgs boson doublet (see Equation (2.50)) adds another four degrees of freedom $(\phi_1, \phi_2, \phi_3, \phi_4)$, making in total $4 \cdot 2 + 4 = 12$ degrees of freedom before symmetry breaking. After spontaneous symmetry breaking, there are three massive bosons (W^+,W^-,Z) with three possible polarization each. There is the photon field, with only two possible polarization because it is massless. Last but not least, there is one scalar massive field, making in total $3 \cdot 3 + 1 \cdot 2 + 1 = 12$ degrees of freedom. No loss of degrees of freedom occurs. Thanks to the choice of the unitary gauge, the massless Goldstone modes are shifted into the three gauge bosons W^+, W^-, and Z via their longitudinal polarization. At no point of the whole derivation gauge invariance had to be given up.

Moreover, it is possible to bring back fermion masses by coupling left-handed and right-handed fields by means of the Higgs boson doublet without breaking gauge invariance anymore. It can be shown [19] that the Lagrangian takes the following form

$$\mathcal{L}_{\mathrm{yuk}} = \sum_{f'} -g_{f'} \overline{\Psi_{\mathrm{L}}} \phi \psi'_{\mathrm{R}} - g_{f'} \overline{\psi'_{\mathrm{R}}} \phi^\dagger \Psi_{\mathrm{L}} + \sum_{f} -g_f \overline{\Psi_{\mathrm{L}}} \phi^c \psi'_{\mathrm{R}} - g_f \overline{\psi'_{\mathrm{R}}} \phi^{c\dagger} \Psi_{\mathrm{L}} \ , \tag{2.74}$$

with the charge conjugate Higgs boson doublet

$$\phi^c = \mathrm{i}\sigma_2 \phi^* \ . \tag{2.75}$$

The sums run of down-type fermions ($f' = e, \mu, \tau, d', s', b'$) and up-type fermions ($f = \nu_{\mathrm{e}}, \nu_\mu, \nu_\tau, u, c, t$), respectively. This type of coupling is called *Yukawa* coupling in courtesy of the physicist Hideki Yukawa who studied such interactions with the aim of describing the nuclear force. Weinberg first applied Yukawa's formalism to Higgs-fermion couplings [25]. It can be shown that $\mathcal{L}_{\mathrm{yuk}}$ is gauge invariant under $\mathrm{SU(2)_L} \times \mathrm{U(1)_Y}$ if the following identity of the hypercharges (y) holds

$$y_\phi = y_\Psi - y_{\psi'_{\mathrm{R}}} = -1 - (-2) = 1 \ . \tag{2.76}$$

Hence, the Higgs boson doublet must have weak hypercharge $y_\phi = 1$ and therefore – according to Equation (2.41) – the electric charges of the weak isospin states must be +1 and 0, justifying retrospectively the choice at the beginning (see Equation (2.50)). Upon electroweak symmetry breaking, the Lagrangian in Equation (2.74) produces the desired mass terms for fermions and takes up the form

$$\mathcal{L}_{\mathrm{yuk}} = \sum_{f'} -m_{f'} \overline{\psi'} \psi' - \frac{m_{f'}}{v} \overline{\psi'} \psi' h + \sum_{f} -m_f \overline{\psi} \psi - \frac{m_f}{v} \overline{\psi} \psi h \ . \tag{2.77}$$

Hence, the Higgs boson coupling to fermions (g_F) is related to the fermion mass via

$$g_F = \frac{\sqrt{2}}{v} m_F \qquad F \in \{f, f'\} \ . \tag{2.78}$$

Unlike the gauge boson couplings (see Equations (2.72) and (2.73)), the fermion coupling is only linearly dependent on the fermion mass. However, the fermion masses are not predicted and present free parameters of the theory. Therefore, measuring the Higgs boson coupling to fermions at the LHC directly tests the validity of the Yukawa theory.

Finally, the SM Lagrangian can be written down

$$\mathcal{L}_{\mathrm{SM}} = \mathcal{L}_{\mathrm{QCD}} + \mathcal{L}_{\mathrm{EWK}} + \mathcal{L}_{\mathrm{BEH}} + \mathcal{L}_{\mathrm{yuk}} \ , \tag{2.79}$$

where $\mathcal{L}_{\mathrm{QCD}}$ is given in Equation (2.15), $\mathcal{L}_{\mathrm{EWK}}$ in Equation (2.46), $\mathcal{L}_{\mathrm{BEH}}$ in Equation (2.51) and $\mathcal{L}_{\mathrm{yuk}}$ in Equation (2.77). It is possible to write $\mathcal{L}_{\mathrm{SM}}$ in terms of physical mass eigenstates instead of the weak eigenstates as demonstrated in this chapter. For quarks, the relation between weak and mass eigenstates is encoded in the Cabbibo-Kobajashi-Maskawa (CKM) matrix [36, 37]

$$\begin{pmatrix} d' \\ s' \\ b' \end{pmatrix} = \begin{pmatrix} V_{\mathrm{ud}} & V_{\mathrm{us}} & V_{\mathrm{ub}} \\ V_{\mathrm{cd}} & V_{\mathrm{cs}} & V_{\mathrm{cb}} \\ V_{\mathrm{td}} & V_{\mathrm{ts}} & V_{\mathrm{tb}} \end{pmatrix} \begin{pmatrix} d \\ s \\ b \end{pmatrix} = V_{\mathrm{CKM}} \begin{pmatrix} d \\ s \\ b \end{pmatrix} \ . \tag{2.80}$$

Due to the quark mixing encoded in V_{CKM}, it is possible that in charged current interactions $\mathrm{W}^{\pm}$ change the flavor of the involved quarks. The weak interaction is the only interaction where this is possible. Since V_{CKM} is close to diagonal, transitions among the same generation are favored. The complex phase of V_{CKM} is the source of $\mathcal{CP}$ violation in the weak interaction [37]. For neutrinos, the mixing of weak and mass eigenstates, (ν_1, ν_2, ν_3), is given by

$$\begin{pmatrix} \nu_{\mathrm{e}} \\ \nu_{\mu} \\ \nu_{\tau} \end{pmatrix} = \begin{pmatrix} V_{\mathrm{e}1} & V_{\mathrm{e}1} & V_{\mathrm{e}1} \\ V_{\mu 2} & V_{\mu 2} & V_{\mu 2} \\ V_{\tau 3} & V_{\tau 3} & V_{\tau 3} \end{pmatrix} \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix} = V_{\mathrm{PMNS}} \begin{pmatrix} \nu_1 \\ \nu_2 \\ \nu_3 \end{pmatrix} \ , \tag{2.81}$$

where V_{PMNS} is the Pontecorvo-Maki-Nakagawa-Sakata (PMNS) matrix [38].

2.2 Shortcomings of the SM

The SM is a very predictive theory and describes many processes which have been observed and analyzed by many experiments. An example of the predictive power are the postulation of the W and Z bosons which were later discovered [39, 40, 41, 1]. Another example is the prediction of the top quark in the context of the quark model together with a prediction of its mass from electroweak data [2]. The discovery of the top quark [42] with a mass consistent with the prediction marks another success of the SM. Figure 2.4 compares experimental measurements and theoretical predictions of different processes produced in high-energy collisions at the LHC. The y-axis represents a measure of the rate of these processes and spans many orders of magnitude. Over this large range, the SM and experimental observations agree and no significant deviation has been observed.

Nevertheless, the SM is not the ultimate theory. Some of its shortcomings are:

- The SM does not include the gravitational force. Its description by means of a renormalizable QFT that can be treated perturbatively has not been achieved.
- Measurements of rotational velocities of galaxies further away from the galactic center suggest that there is in fact more matter than observable [44]. This new kind of matter is termed *dark matter* and is clearly not made up of matter particles from the SM.

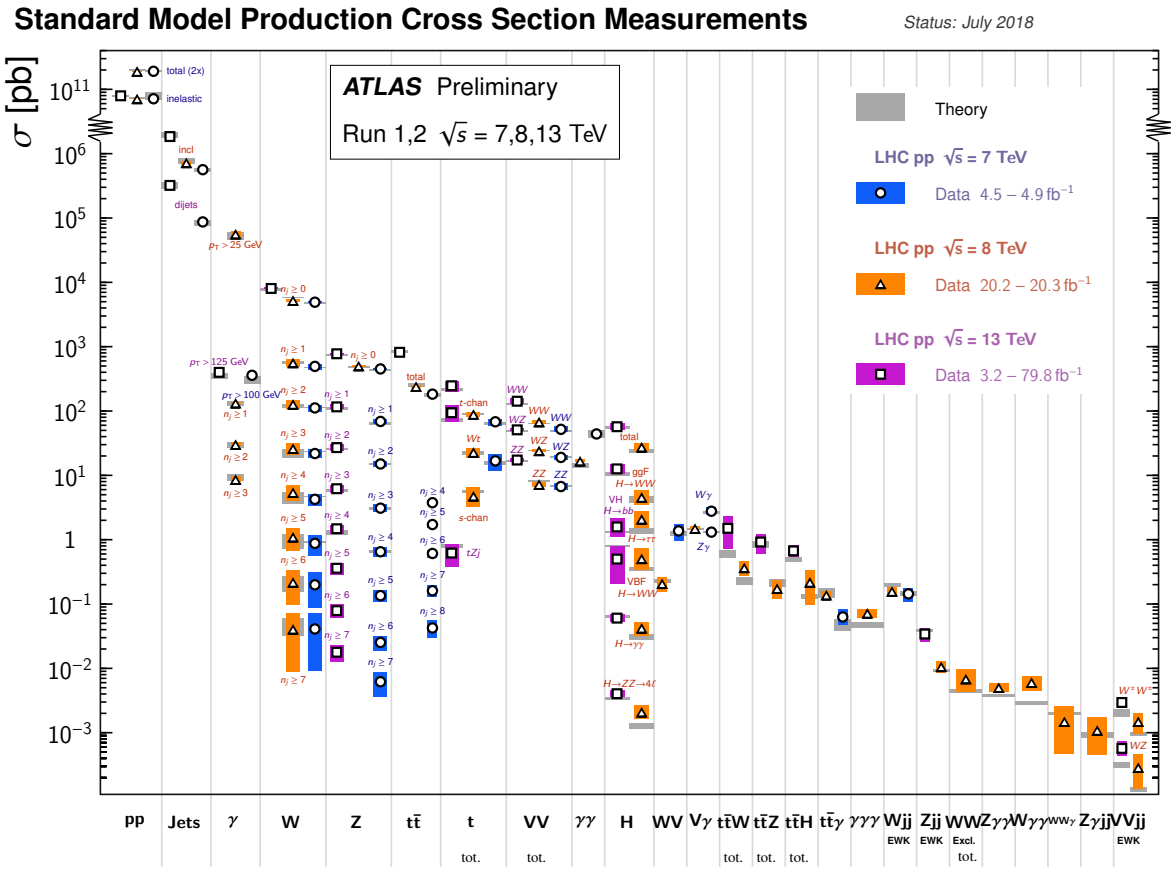


Figure 2.4: Production cross sections for different SM processes are shown and compared to the corresponding theoretical prediction. The plot is taken from [43].

- Cosmological observations [45] are consistent with an expanding universe driven by an unknown form of energy which is referred to as *dark energy*. The SM does not provide a source of such an energy.

- The fact that not all matter and antimatter has annihilated in an early stage of the universe and we observe planets, stars and galaxies made of matter, implies a matter-antimatter asymmetry. Such an asymmetry can be partially explained by $\mathcal{CP}$ violation in the quark sector of weak interactions encoded in the complex phase of the CKM matrix [46]. Similarly, the complex phase of the PMNS matrix – which remains to be determined by experiments – is a source of $\mathcal{CP}$ violation [47]. However, our present knowledge of the $\mathcal{CP}$ violation does not seem to fully explain the observed baryonic asymmetry [48].

- One of the free parameters of the SM is the QCD vacuum angle θ. Its current measured upper bound is 10^{-10} [49], suggesting that it might be zero. An implication of a vanishing vacuum angle is that QCD preserves exactly the $\mathcal{CP}$ symmetry. This problem is referred to as the *strong $\mathcal{CP}$ problem* because the SM does not provide an explanation for a small or even zero value of θ.

- Higher order quantum corrections to the Higgs boson mass, m_{H}, grow quadratically up to an energy, where new physics starts to play a role. It is important to note that corrections of fermions and bosons carry different signs. Hence, an observed Higgs boson mass of 125 GeV – much smaller than the magnitude of the corrections – can only be explained if these corrections cancel each other almost perfectly. This problem is referred to as the *naturalness* problem of the Higgs sector [50].

- Further open questions are among others

  - Why are there three generations of fermions?
  - Why do the fermion masses span so many orders of magnitude?
  - Why is the gravitational force so weak compared to the other interactions?
  - Why do the gauge couplings not meet at a certain energy scale indicating a unification of all forces?
  - How are the free parameters of the SM connected and is there some underlying pattern? In total there are 19 parameters of the SM and an additional 7 or 8 parameters dependent on whether neutrinos are treated as Dirac or Majorana particles.

2.3 Extending the SM

There are many ideas how the SM can be extended to provide answers to (some of) the open questions listed in Section 2.2. An introductory overview of different BSM theories is for example given in [51]. In this section, the focus lies on one such theoretical model called supersymmetry (SUSY). In Section 2.3.1, the basic concepts and motivation behind SUSY are presented. Special emphasis is given on the Higgs sector since this topic is relevant for the work presented in Section 9.3.1. Furthermore, the theoretical grounds for light stop searches are discussed in this section. Lastly, an introduction to leptoquarks is given.

2.3.1 Supersymmetry

Supersymmetry is a BSM theory which has been heavily studied since the 1970s when Wess and Zumino [52] laid the foundation by identifying the specific renormalization characteristics of four-dimensional supersymmetric field theories. An additional symmetry is introduced resulting in a link between fermions and bosons. As such, SUSY can connect QFTs which are based on local gauge transformation with general relativity [53]. Furthermore, SUSY could provide solutions to the fine-tuning problem of the Higgs boson mass, allows unification of the gauge coupling constants and predicts new particles serving as dark matter candidates. In what follows, a basic introduction focusing on relevant aspects of SUSY is given, following the presentation of this topic in [54].

A strong argument motivating SUSY is its solution to the fine-tuning of the Higgs boson mass, which is one of the shortcomings of the SM discussed in Section 2.2. Corrections to the Higgs boson mass from fermions (F) with a coupling to the Higgs boson field (g_F) are given by

$$\Delta m_{\mathrm{H}} = -\frac{|g_F|^2}{8\pi^2}\Lambda_{\mathrm{UV}}^2 + \cdots \ , \tag{2.82}$$

where Λ_{UV} is the energy scale at which new physics alters the SM. If it is assumed that this energy scale is the Planck scale, then the corrections Δm_{H} are 30 orders of magnitude larger than the required value for m_{H}. Contributions to the Higgs boson mass correction coming from a complex scalar particle, S, are given by the formula

$$\Delta m_{\mathrm{H}} = +\frac{g_S}{16\pi^2}\Lambda_{\mathrm{UV}}^2 + \cdots \ , \tag{2.83}$$

where g_S denotes the coupling of a massive scalar to the Higgs boson field. Comparing Equations (2.82) and (2.83) with each other allows coming to the conclusion that if each fermion of the SM was complemented with *two* scalar bosons with $g_S = |g_F|^2$, the corrections would cancel each other perfectly. It can be shown that this neat cancellation can be enforced at any order in perturbation theory by imposing a new kind of symmetry between fermions and bosons. This symmetry is called *supersymmetry*.

Supersymmetric transformations turn a fermionic state into a bosonic state and vice versa. In a light-weight formula this can be expressed as

$$\begin{aligned} Q\,|\text{fermion}\rangle &= |\text{boson}\rangle \\ Q\,|\text{boson}\rangle &= |\text{fermion}\rangle \ , \end{aligned} \tag{2.84}$$

where Q is the generator of supersymmetric transformations. By the Haag-Łopuszanski-Sohnius theorem [55], the possible forms SUSY can take and still be a QFT are highly restricted. In order to embed the SM correctly, the supersymmetric generators have to satisfy the following conditions

$$\begin{aligned} \{Q, Q^\dagger\} &= P^\mu \\ \{Q, Q\} = 0 &= \{Q^\dagger, Q^\dagger\} \\ [P^\mu, Q] = 0 &= \left[P^\mu, Q^\dagger\right] \ , \end{aligned} \tag{2.85}$$

where P^μ is the generator of spacetime translations. Enforcing local gauge invariance under supersymmetric transformation induces gravity in so-called *super gravitational theories* [53]. It should be stressed that Equation (2.85) does not show the spinor indices and is just a light-version of the exact (anti-)commutation relations derived in [54]. The commutation and anti-commutation rules formulated in Equation (2.85) define a *supersymmetric algebra*.

Single-particle states of a supersymmetric theory form *supermultiplets*[7]. Each super-

[7] Supermultiplets are irreducible representations of the supersymmetric algebra.

particle type	spin-0	spin-½	spin-1	$SU(3)_C$	$SU(2)_L$	$U(1)_Y$
(s)quark	$\tilde{Q}_L = \begin{pmatrix} \tilde{q}_L^u \\ \tilde{q}_L^d \end{pmatrix}$	$Q_L = \begin{pmatrix} q_L^u \\ q_L^d \end{pmatrix}$	–	**3**	**2**	1/3
	$\tilde{q}_R^{u,*}$	$\overline{q}_R^u$	–	$\overline{\mathbf{3}}$	**1**	-4/3
	$\tilde{q}_R^{d,*}$	$\overline{q}_R^d$	–	$\overline{\mathbf{3}}$	**1**	2/3
(s)lepton	$\tilde{L} = \begin{pmatrix} \tilde{\nu}_{l,L} \\ \tilde{l}_L \end{pmatrix}$	$L = \begin{pmatrix} \nu_{l,L} \\ l_L \end{pmatrix}$	–	**1**	**2**	-1
	$\tilde{\nu}_{l,R}^*$	$\overline{\nu}_{l,R}$	–	**1**	**1**	0
	$\tilde{l}_R^*$	$\overline{l}_R$	–	**1**	**1**	2
Higgs / Higgsino	$H_u = \begin{pmatrix} H_u^+ \\ H_u^0 \end{pmatrix}$	$\tilde{H}_u = \begin{pmatrix} \tilde{H}_u^+ \\ \tilde{H}_u^0 \end{pmatrix}$	–	**1**	**2**	1
	$H_d = \begin{pmatrix} H_d^0 \\ H_d^- \end{pmatrix}$	$\tilde{H}_d = \begin{pmatrix} \tilde{H}_d^0 \\ \tilde{H}_d^- \end{pmatrix}$	–	**1**	**2**	-1
gluino / gluon	–	$\tilde{g}^a$	G_μ^a	**3**	**1**	0
Wino / W boson	–	$\tilde{W}^i$	W_μ^i	**1**	**3**	0
Bino / B boson	–	$\tilde{B}$	B_μ	**1**	**1**	0

Table 2.1: This table summarizes the particle content of the MSSM. Particle names as well as the fermionic and bosonic component of each supermultiplet are shown. First the matter supermultiplets are listed, then the gauge supermultiplets. In the last three columns of the table the representations under SM gauge transformation are displayed, using the convention from Equation (2.41). Note, that here neutrinos are assumed to be Dirac particles and therefore the right-handed neutrino fields, $\overline{\nu}_{l,R}$, and their superpartners appear in this table as well. For the (s)quarks the index u runs over all up-type quarks, i.e. $q^u = u, c, t$, whereas d runs over the down-type quarks, i.e. $q^d = d', s', b'$. The index l used for the (s)lepton fields runs over the lepton flavors, i.e. $l = e, \mu, \tau$.

multiplet contains a fermionic and a bosonic component which is referred to as being each other's *superpartner*. From the commutation relations shown in Equation (2.85), it follows immediately that particles in the same supermultiplet have the same mass and transform identically under SM gauge transformations. The latter conclusion means that superpartners have the same electric charge, weak isospin and color degrees of freedom. Supermultiplets must have the same number of fermionic and bosonic degrees of freedom. It can be shown that a supermultiplet containing a SM fermion must be complemented by a complex scalar (spin-0) component. Such multiplets are called *scalar* or *matter* or *chiral* supermultiplets. For spin-1 gauge bosons, it turns out that they must be paired with spin-½ superpartners with resulting multiplets referred to as *gauge* or *vector* supermultiplets. Table 2.1 summarizes the particle content of the minimal supersymmetric extension of the Standard Model (MSSM) and its organization of scalar and vector supermultiplets. The MSSM adds the minimal amount of new particles to the SM. The details of Table 2.1 are discussed next.

The superpartners of quarks and leptons are the scalar quarks and scalar leptons, called *squarks* and *sleptons* for short, respectively. Since left-handed and right-handed fermions have different transformation properties under SM gauge transformations, they must belong to different supermultiplets and have each respective superpartners. Super-

symmetric particles are denoted with an extra tilde (˜). For example, the superpartner of the left-handed top quark (t_L), is the top squark ($\tilde{t}_L$), where the subscript is just used to link the top squark to the correct SM partner. Both, (s)quarks and (s)leptons are put into scalar supermultiplets which are often addressed by objects with a hat (ˆ) in the literature. For instance, the supermultiplet $\hat{Q}$ is defined as

$$\hat{Q} = \begin{pmatrix} (\tilde{q}_L^u, q_L^u) \\ (\tilde{q}_L^d, q_L^d) \end{pmatrix} , \tag{2.86}$$

and encompasses the left-handed quarks and their superpartners. The indices u and d run over the three generations of up and down-type quarks, respectively. Another scalar supermultiplet is formed by the Higgs boson and its superpartner, the *Higgsino*. However, there must be two weak isospin Higgs doublets for the MSSM to work, resulting in the so-called two Higgs doublet model (THDM) [56, 57]. One reason is the generation of quark masses after electroweak symmetry breaking. Within the MSSM, the Higgs doublet with $Y = 1$ gives mass to up-type quarks with electric charge $+2/3$, while the other Higgs doublet with $Y = -1$ gives mass to the down-type quarks with electric charge $-1/3$. Therefore, one notation adapted in the literature – and used in this thesis – is to label the Higgs doublets with a subscript "u" or "d" to remember to which quark type they couple and generate a mass term (see also Table 2.1). Having more than one Higgs boson in the THDM results also in more superpartners, which are referred to as Higgsinos.

SM gauge bosons are grouped and form gauge supermultiplets together with their superpartners which are called *gauginos*. The supersymmetric partners of the gluons are the gluinos ($\tilde{g}$), the superpartners of the W bosons are the *winos* ($\widetilde{W}$) and the superpartner of the B boson is the *bino* ($\widetilde{B}$). A mixture of wino and bino defines the massive *zino* ($\widetilde{Z}$) and massless *photino* ($\tilde{\gamma}$) after electroweak symmetry breaking in complete analogy to the mixing of the SM gauge bosons shown in Equation (2.44).

Particles in the same supermultiplet have the same mass as already mentioned earlier. However, none of the supersymmetric particle have been discovered as of this writing. This implies that SUSY must be a *broken* symmetry. The mass of the discovered 125 GeV Higgs boson actually gives a strong hint that SUSY must be broken. If SUSY was an exact symmetry, the mass of the lightest electrically neutral Higgs boson would be below the Z boson mass ($m_Z = 91\,\text{GeV}$). Broken SUSY implies that corrections to the Higgs boson mass in Equations (2.82) and (2.83) do not cancel exactly anymore and hence a mass larger than the one of the Z boson is again possible. However, the symmetry breaking should not be such that one ends up again in an unnatural fine-tuning of the Higgs boson mass. In the literature it is therefore stated that SUSY breaking must be *soft*. In many soft SUSY breaking scenarios, the lightest SUSY particle(s) can have masses in the TeV range and thus being accessible to be produced in proton-proton collisions at the LHC.

An important feature of the MSSM superpartners (see Table 2.1) is that they can mix after electroweak and soft SUSY breaking to form physical mass eigenstates. This motivates searches for top squarks as will be discussed further below. Only the gluino, being a color octet, does not have the appropriate quantum numbers for mixing with other superpartners. Neutral mass eigenstates are called *neutralinos* ($\tilde{\chi}_1^0$, $\tilde{\chi}_2^0$, $\tilde{\chi}_3^0$, $\tilde{\chi}_4^0$) and are mixtures of neutral Higgsinos and gauginos. They are labeled in a mass-ordered way, where $\tilde{\chi}_1^0$ is the lightest one. Similarly, the charged Higgsinos and gauginos mix and form four mass eigenstates called *charginos* ($\tilde{\chi}_1^\pm$, $\tilde{\chi}_2^\pm$).

In many SUSY models such as the MSSM, an additional symmetry – called $\mathcal{R}$*-parity* – is imposed. $\mathcal{R}$-parity conservation ensures for instance that protons do not decay via exchange of SUSY particles. The experimental lower bound on the lifetime of protons depends on the specific decay process analyzed and lies for example in the order of 10^{33}

years for protons decaying into positrons [58]. These lower bounds put strong constraints on BSM theories predicting processes allowing for proton decay. The definition of $\mathcal{R}$-parity is given by

$$p_{\mathrm{R}} = (-1)^{3(B-L)+2S} \ , \tag{2.87}$$

where B, L and S are the baryon number, lepton number and spin of the particle, respectively. It can be checked that all SM particles have $p_{\mathrm{R}} = 1$, while their superpartners have $p_{\mathrm{R}} = -1$. An important consequence of $\mathcal{R}$-parity conservation, is that the lightest supersymmetric particle (LSP) is stable. A weakly interacting electrically neutral LSP provides a well suited dark matter candidate. In the analysis presented in Section 9.3.3, the lightest neutralino is assumed to be the LSP.

The Higgs Sector of the MSSM and the NMSSM

In Table 2.1, the two Higgs doublets introduced in the MSSM are shown. In analogy to Equation (2.54), the ground states of each Higgs doublet can be chosen as

$$H_{\mathrm{u,ground}} = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v_{\mathrm{u}} \end{pmatrix} \qquad H_{\mathrm{d,ground}} = \frac{1}{\sqrt{2}} \begin{pmatrix} v_{\mathrm{d}} \\ 0 \end{pmatrix} \ , \tag{2.88}$$

where v_{u} and v_{d} are the vevs of each Higgs doublet which are related to the vev from the SM case by

$$v^2 = v_{\mathrm{u}}^2 + v_{\mathrm{d}}^2 \ . \tag{2.89}$$

Starting off with eight degrees of freedom – from the two complex isospin Higgs doublets with four degrees of freedom each – the following particle spectrum is generated after electroweak symmetry breaking. Like in the SM, three degrees of freedom take up the form of longitudinal polarizations of the now massive W and Z bosons. However, in the MSSM there are now five degrees of freedom left corresponding to five massive Higgs bosons. These Higgs bosons are denoted as

$$\mathrm{h}, \quad \mathrm{H}, \quad \mathrm{A}, \quad \mathrm{H}^+, \quad \mathrm{H}^- \ . \tag{2.90}$$

The Higgs bosons $\mathrm{H}^{\pm}$ are charged, while the others are electrically neutral. The A boson is a pseudoscalar, i.e. it transforms under $\mathcal{CP}$ transformation with a sign-flip ($\mathcal{CP}$-odd). The Higgs bosons h and H are $\mathcal{CP}$-even particles, where h is the lighter of the two Higgs bosons[8]. In fact, after electroweak symmetry breaking, the MSSM Higgs sector can be fully described at tree-level[9] by just two parameters. These two parameters are typically chosen as

$$\begin{gathered} \tan\beta = \frac{v_{\mathrm{u}}}{v_{\mathrm{d}}} \\ m_{\mathrm{A}} \ , \end{gathered} \tag{2.91}$$

where m_{A} is the mass of the pseudoscalar A. Different benchmark scenarios [59, 60] have been defined to study the MSSM Higgs sector in a systematic way. One of them is the $\mathrm{M}_{\mathrm{h}}^{125}$ scenario which is discussed in detail in [59]. It incorporates knowledge about the SM Higgs boson and identifies it with h of the MSSM

$$\mathrm{h} \leftrightarrow \mathrm{h}_{\mathrm{SM}} \ . \tag{2.92}$$

[8] Note, that the heavy neutral Higgs boson is not to be identified with the SM Higgs boson, even though the same symbol is used. In fact, it is the light neutral Higgs boson which is identified as being the discovered SM Higgs boson (see also Equation (2.92))

[9] At higher orders, the particle spectrum of the MSSM can have an influence on corrections to the Lagrangian.

Higgs boson	up-type fermions	down-type fermions
h	$\frac{\cos\alpha}{\sin\beta} \to 1$	$-\frac{\sin\alpha}{\cos\beta} \to 1$
H	$\frac{\sin\alpha}{\sin\beta} \to \frac{1}{\tan\beta}$	$\frac{\cos\alpha}{\cos\beta} \to \tan\beta$
A	$\frac{1}{\tan\beta}$	$\tan\beta$

Table 2.2: The coupling prefactors quantify how much the Yukawa couplings change in the MSSM with respect to the SM case. This table shows the values of the coupling prefactors in the decoupling limit, $m_A \gg m_Z \Rightarrow \alpha \to \beta - \pi/2$, for the M_h^{125} scenario. The mixing angle, α, parametrizes the rotation of two Higgs boson fields from the doublet making up the physical h and H fields [57]. In the decoupling limit, the couplings of h become SM-like. The Yukawa couplings of H approach those of A, where couplings to down-type fermions are enhanced by a factor $\tan\beta$ and suppressed by the inverse factor for up-type fermions. General expressions for coupling prefactors in different MSSM scenarios can be found in [61].

particle type	spin-0	spin-½	spin-1	$SU(3)_C$	$SU(2)_L$	$U(1)_Y$
Higgs / Higgsino	S	$\tilde{S}$	–	**1**	**1**	0

Table 2.3: The particle content of the NMSSM is the same as for the MSSM – summarized in Table 2.1 – with the addition of a scalar Higgs boson weak isospin singlet, S, as shown in this table.

One immediate consequence from the measured mass of the observed SM Higgs boson is that in the M_h^{125} scenario $\tan\beta > 1$ for arbitrary values of m_A. Furthermore, in the *decoupling limit* defined by

$$m_A \gg m_Z \ , \tag{2.93}$$

the coupling of Higgs bosons to down-type fermions is enhanced as can be seen from Table 2.2, motivating the search for additional neutral Higgs bosons in the di-tau final state. Enhanced couplings to down-type fermions influences also the phenomenology of MSSM Higgs boson production modes at the LHC as discussed in Section 5.2. Results from a search of $H, A \to \tau\tau$ decays are presented in Section 9.3.1.

The MSSM does not encompass all possible supersymmetric extensions of the SM. Furthermore, specific choices on MSSM parameters like the Higgsino mass have to be made. The next-to-minimal supersymmetric extension of the Standard Model (NMSSM) [62, 63] postulates an additional Higgs singlet which generates the Higgsino mass dynamically [64]. The transformation properties of the Higgs singlet are summarized in Table 2.3. Counting the degrees of freedom before electroweak symmetry breaking, there are now ten – four come from each complex Higgs doublet (H_u,H_d) and two from the Higgs singlet (S). After electroweak symmetry breaking – which works in analogy to the SM – again three degrees of freedom get absorbed in form of longitudinal polarizations of the now massive W and Z bosons. The remaining seven degrees of freedom form the following spectrum of massive physical Higgs bosons

$$h_{SM}, \quad h_S, \quad H, \quad A_1, \quad A_2, \quad H^+, \quad H^- \ . \tag{2.94}$$

Comparing to the set of Higgs bosons in the MSSM (see Equation (2.90)), there are now two pseudoscalar Higgs bosons, A_1 and A_2, and an extra scalar boson, h_S. The h_S boson can be light, even lighter than the observed SM Higgs boson (h_{SM}). A search for $H \to h_{SM}(\tau\tau)h_S(bb)$ decays is presented in Section 9.3.1 and experimental prospects are discussed in Section 5.2.

Light Top Squarks in the MSSM

In many SUSY models, the mass difference between mass eigenstates of squarks after symmetry breaking is proportional to the mass of their SM partner. As the top quark is the heaviest of all quarks, the mass splitting between the two stop mass eigenstates, $\widetilde{t}_1$ and $\widetilde{t}_2$, is the largest among all squarks. Thus, $\widetilde{t}_1$ might be the lightest squark which can be possibly produced in high-energy collision experiments. The top quark contributes most to the corrections to the Higgs boson mass (see Equation (2.82)) because of its large Yukawa coupling $g_F = g_t$, i.e. because of the large top quark mass. The top squark contribution to the correction of the Higgs boson mass is given by Equation (2.83) and involves a large Yukawa coupling $g_S = g_{\widetilde{t}}$, too. Hence, in view of the fine-tuning problem of the Higgs boson mass, a rather light top squark is favored from a theoretical point of view.

An important quantity when investigating the possible top squark decays, is the mass difference of the top squark to the LSP, taken as the lightest neutralino,

$$\Delta m = m_{\tilde{t}_1} - m_{\tilde{\chi}_1^0} \ . \tag{2.95}$$

Cosmological observations can be explained with SUSY models where the stop mass and LSP mass are nearly degenerate [65]. Hence, searches for top squarks at the LHC with low Δm – referred to as *compressed* region – are well motivated. More on the specific signal model considered in this thesis is discussed in Section 5.4, where also the experimental signatures of the signal model are discussed.

2.3.2 Leptoquarks

In attempts to extend the SM by unifying all gauge couplings and embedding the SM $\mathrm{SU(3)_C \times SU(2)_L \times U(1)_Y}$ symmetry into a larger symmetry group – referred to as a grand unified theory (GUT) – quarks and leptons reside in the same representation. Thus, bosons that couple to both leptons and quarks naturally arise in GUTs [66, 67, 68]. Such a boson is called leptoquark (LQ). More specifically, LQs carry both lepton number and baryon number, have fractional electric charges and can transform quarks into leptons and vice versa. Postulated LQs can either have a scalar (spin-0) or vector (spin-1) nature. In the context of GUTs, LQs reside in the same representation as the Higgs doublet. A plausible assumption is to expect the same enhanced Yukawa couplings to third generation fermions as in case of the Higgs boson because third generation fermions are the heaviest [69].

However, LQs appear also in other extensions of the SM. In $\mathcal{R}$-parity violating MSSM, squarks take the role of scalar LQs [70]. Furthermore, LQs are also predicted by theories based on technicolor [71, 72] and compositeness [73]. Searches for third generation LQs have recently gained a lot of interest in the high-energy physics community. Their existence could explain anomalies in B-physics decay rates as reported by several experiments [74, 75, 76, 77], amounting to a total deviation from the SM expectation by about four standard deviations [78].

As detailed in [79], there are six scalar and six vector LQ multiplets possible if the transformation under the SM gauge groups is considered as classification criterion. A specific scalar and a specific vector model is used in the search for a third generation LQ presented in Section 5.3. The two models as well as the production mechanisms at the LHC are detailed in Section 5.3, while results of the analysis are presented in Section 9.3.2.

Chapter 3

The CMS Experiment

> An experiment is a question which science poses to Nature, and a measurement is the recording of Nature's answer.
>
> Max Planck, 1858 to 1947

Large research facilities have been built to test the SM and to look for new physics. The goal of this chapter is to illuminate this experimental side. Section 3.1 presents the LHC, mankind's most powerful particle accelerator built so far. Relevant topics from accelerator physics are also discussed.

High-energy protons, delivered by the LHC, are brought to collision and recorded by several detectors. In this thesis, collisions recorded by the CMS detector are analyzed. An introduction to the CMS detector together with a discussion of all its subdetector elements is given throughout Section 3.2. The amount of produced particles in a single collision which traverse the CMS detector, is overwhelming. Technically it is impossible to store everything and thus special decision-making procedures have been put into place as discussed in Section 3.3. It is important that the physics processes behind colliding particles, production of new particles and interaction of those with the detector can be simulated. More details about simulation programs are discussed in Section 3.4. Both, simulation and collision data use the same reconstruction algorithms, discussed throughout Section 3.5, to obtain physics objects used for analysis. Some of the topics presented in this chapter are based on the reference [80], which introduces many basic concepts used in experimental high-energy physics.

3.1 The LHC Accelerator Facility

The Large Hadron Collider (LHC) is a circular particle accelerator built and operated by CERN[1] near the city of Geneva, Switzerland. The LHC is located about 100 m below ground in a tunnel with a circumference of 27 km. Starting from a hydrogen bottle, electrons are stripped away by applying an electric field and the obtained protons are pre-accelerated in different stages. First, they are accelerated by a linear accelerator, followed by the proton synchotron booster (PSB), the proton synchotron (PS) and finally by the super proton synchotron (SPS). Proton energies reach 450 GeV at the SPS, before they are injected in two separate beam pipes of the LHC. One proton beam runs clockwise and the other counter-clockwise inside the LHC. Not only protons, but also heavy nuclei (lead or xenon) are accelerated and injected into the LHC. Using dipole magnets, the

[1] Abbreviation derived from French *Conseil européen pour la recherche nucléaire*

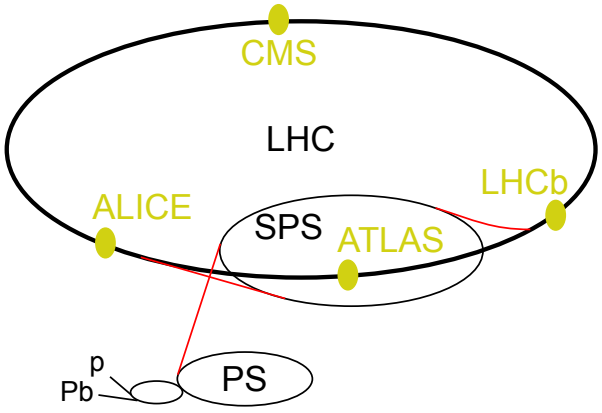


Figure 3.1: A schematic drawing of the LHC accelerator complex with its four multipurpose detector experiments is shown. Protons (p) or heavy ions (here Pp) are injected in three consecutive circular pre-accelerators - the PSB, the PS and the SPS. Relevant for this thesis is the CMS detector shown at the top. The figure is taken from [81].

particle beams inside the LHC are deflected and brought to collision at specific *interaction points*. A collision is also referred to as an *event*. Currently there are nine experiments installed at different interaction points, four of which operate one of the following large multipurpose detectors:

ALICE: A Large Ion Collider Experiment,

ATLAS: A Toroidal LHC Apparatus,

CMS: Compact Muon Solenoid and

LHCb: Large Hadron Collider beauty experiment.

Figure 3.1 shows the LHC, its pre-accelerators and the location of the four multipurpose detectors.

One life cycle of a proton beam – starting from the injection into the LHC until its depletion – is called a *fill*. Under normal circumstances, proton beams can be used for collision for about ten hours before they get dumped. During that time the beam is constantly monitored and dumped earlier in case problems arise. Data-taking periods are indexed with so-called *run numbers* by each experiment. They include time intervals without beam, where detectors record for example cosmic rays. In this thesis, results based on proton-proton collision data recorded in the years from 2015 to 2018 (Run2) by the CMS detector are presented.

An important accelerator parameter is the *center-of-mass* energy, denoted by $\sqrt{s}$. For discovery machines like the LHC, it is desired to achieve highest possible center-of-mass energies as it represents the available energy to produce undiscovered heavier particles. The center-of-mass energy of the LHC during Run2 was $\sqrt{s} = 13\,\text{TeV}$.

Due to the technology used to accelerate protons, the proton beams are not continuous but bunched. In each proton bunch there are about 10^{11} protons. There are 2808 bunches per beam with a bunch spacing of 25 ns. Hence, not the full LHC is filled with proton bunches but there are bigger gaps between bunches. The bunch spacing of 25 ns translates to a peak crossing rate of 40 MHz. However, this does not correspond to the number of collisions per second. Knowing that proton bunches travel close to the speed of light, the circumference of the LHC tunnel as well as the number of bunches in each beam, the

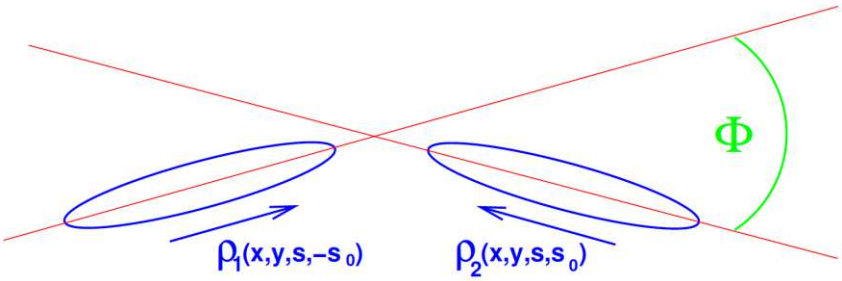


Figure 3.2: A schematic illustration of a bunch crossing is depicted. The figure is adopted from [84]. The two bunches have a particle density described by ρ_1 and ρ_2, respectively. In red the trajectories of the bunches are shown. Particle densities depend on the coordinates transverse to the trajectory (x and y) and on the longitudinal coordinate (s), measured along the bunch trajectory. The longitudinal distance between the center of the bunch and the crossing point is denoted by s_0. The crossing angle (Φ) is highlighted in green.

average crossing rate is about 30 MHz. In order to get an estimate of the number of collisions per second, the notion of *cross section* and *luminosity* have to be introduced first.

Simply speaking, the cross section is a measure of probability of a certain process to occur. The notion dates back to the first fixed target experiment such as for example Rutherford's scattering experiment [82]. It can be understood as the area a target presents to an incoming particle. Hence, cross section is an area with the unit barn [83], where one barn corresponds to $10^{-28}\,\mathrm{m}^2$. Clearly, the higher the cross section, σ_p, of a certain process p, the higher is the rate ($\mathrm{d}N_\mathrm{p}/\mathrm{d}t$) at which this process is produced in a proton-proton collision. The proportionality factor is called luminosity:

$$\frac{\mathrm{d}N_\mathrm{p}}{\mathrm{d}t} = \mathcal{L} \cdot \sigma_\mathrm{p} \ , \tag{3.1}$$

where the luminosity is denoted by $\mathcal{L}$ and measured in $\mathrm{cm}^{-2}\mathrm{s}^{-1}$. Luminosity can be thought of as the accelerator's capability to put particles in position to collide. As derived for example in [84], the luminosity of an accelerator machine colliding particle bunches head-on is given by

$$\mathcal{L} = \frac{N_1 N_2 f N_\mathrm{b}}{4\pi\sigma_x\sigma_y} \cdot S \ , \tag{3.2}$$

where N_1 and N_2 are the number of particles per bunch in each beam. The number of bunches per beam is denoted by N_b and the revolution frequency with f. Each bunch is assumed to have a density profile described by a Gaussian with a width of σ_x and σ_y, where the x-y plane is transverse to the beam axis. The so-called luminosity reduction factor, denoted by S in Equation 3.2, takes into account the crossing angle of the two colliding bunches as illustrated in Figure 3.2. Parameter values for the LHC are summarized in Table 3.1. The total inelastic proton-proton cross section is in the order of 80 mb[85]. Inserting this value together with the LHC luminosity in Equation (3.1), it follows that each second roughly 800 million inelastic collisions occur.

Integrating Equation 3.1 over time yields the number of events expected from a certain process

$$N_\mathrm{p} = \sigma_\mathrm{p} \cdot \int \mathcal{L}\mathrm{d}t = \sigma_\mathrm{p} \cdot \mathcal{L}_\mathrm{int} \ . \tag{3.3}$$

In the above equation, the *integrated* luminosity ($\mathcal{L}_\mathrm{int}$) is introduced. Integrated luminosity has a unit of inverse area and gives a measure about the amount of collisions produced over a certain time. Not all of the collisions can be recorded by CMS as will be discussed

parameter	symbol	value
number of protons per bunch	$N_1 = N_2$	1.2×10^{11}
number of bunches per proton beam	N_b	2808
revolution frequency	f	11.2 kHz
transverse Gaussian width of proton bunch density	$\sigma_x = \sigma_y$	17 µm
luminosity reduction factor	S	0.835
total crossing angle	Φ	285 µrad
center-of-mass energy	$\sqrt{s}$	13 TeV
luminosity	$\mathcal{L}$	$2 \times 10^{-34}\,\mathrm{cm}^{-2}\mathrm{s}^{-1}$

Table 3.1: A summary of important LHC parameters is given. The values are taken from [84].

in more detail in Section 3.3. Each experiment quotes the amount of data collected in terms of integrated luminosity, i.e. in units of inverse area. The integrated luminosity recorded by CMS during Run2 is shown in Figure 3.3 and amounts to roughly $150\,\mathrm{fb}^{-1}$.

Within a single bunch crossing, several proton-proton collisions occur. Since protons are composite objects, a proton-proton collision is ultimately described by the interaction of its constituents that are also referred to as *partons*. Two up type and one down type quark make up the *valence quarks* of a proton. Gluons holding the quarks inside a proton together are present, too. In addition, quark-antiquark pairs are produced due to quantum fluctuations which are referred to as *sea quarks*. Typically, one is interested in hard scattering processes, breaking up the colliding protons and producing new particles[2]. The contamination from all other proton-proton collisions within the same bunch crossing is termed pileup (PU). Figure 3.4 shows the PU distributions for the different years of data-taking of Run2. Inside the figure, the inelastic proton-proton cross section ($\sigma_\mathrm{in}^\mathrm{pp}$) is quoted which was used earlier to calculate the number of collisions per seconds (see also Equation (3.1)). Protons that undergo elastic scattering do not dissociate and there are no new particles produced. Such protons travel further along the beam pipe and are not detected by the CMS detector.

3.2 CMS Detector Subsystems

Before each detector subsystem is discussed in a dedicated section, an overview of the coordinate system and important kinematic variables is given. The origin of the CMS coordinate system is placed at the interaction point. The x-axis points to the center of the LHC accelerator, the y-axis points upward, and the z-axis points along the beam axis which runs counter-clockwise inside the LHC tunnel. This Cartesian coordinate system is illustrated in Figure 3.5. However, a modified version of spherical coordinates (r, ϕ, θ) is often used to better exploit the geometry of the CMS detector. The azimuth, $\phi \in [-\pi, \pi]$, measures the angle around the beam axis inside the x-y plane and is defined such that $\phi = 0$ is pointing in x-direction. The polar angle, $\theta \in [0, \pi]$, points along the z-axis for $\theta = 0$. Azimuth and polar angle are also depicted in Figure 3.5.

Figure 3.7 depicts a schematic overview of the CMS detector. Different layers of detector elements surround the beam pipe, which goes through the center of the detector. The central part of the detector is called the *barrel* region. It is complemented by two forward regions – one on each side with respect to the z-axis – called *endcaps*. The com-

[2] New in the sense that incoming and outgoing particles are different, not that the outgoing particles have never been produced before.

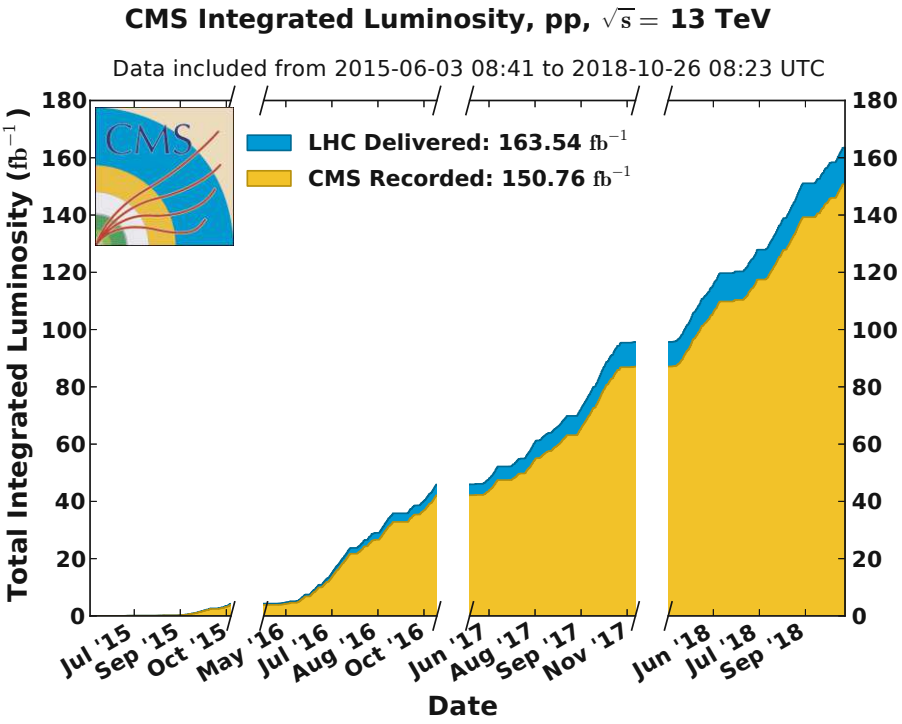


Figure 3.3: The integrated luminosity ($\mathcal{L}_{\text{int}}$) as a function of time is shown. Almost all of the collisions delivered by the LHC (shown in blue), were recorded by the CMS experiment (shown in yellow). Gaps on the x-axis indicate interruption periods used for maintenance and upgrade work. The figure is taken from [86].

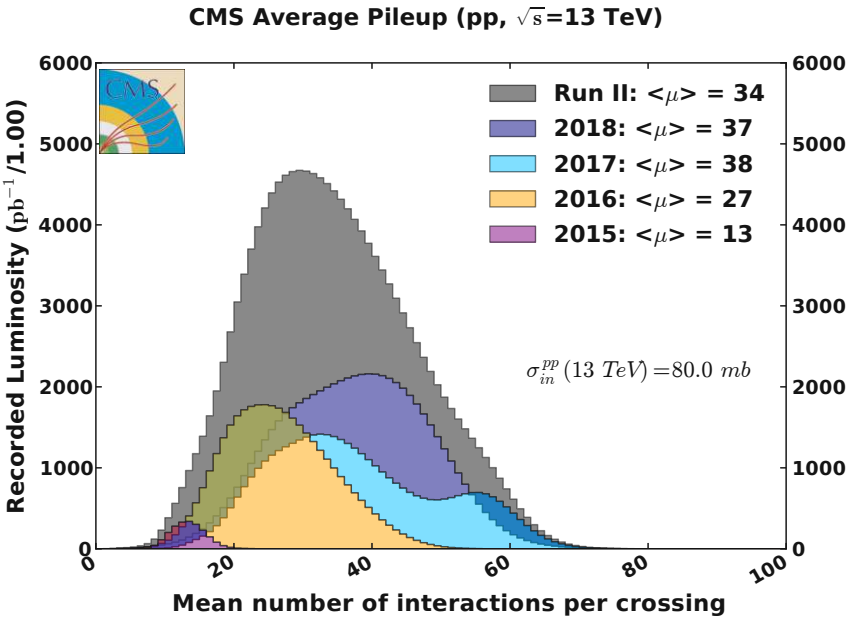


Figure 3.4: Average PU distributions are shown, i.e. distributions of the mean number of interactions (collisions) per bunch crossing, denoted by μ. The PU distribution is shown for each year of data-taking separately, as well as for the full Run2 combined. Average numbers of PU are quoted in the legend. The figure is adopted from [86].

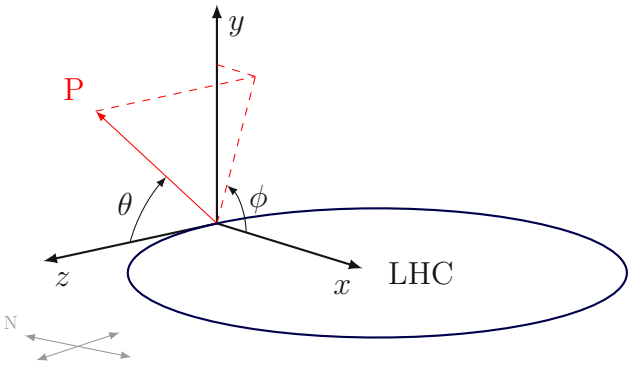


Figure 3.5: This illustration shows the CMS coordinate system. In particular, the definition of the azimuth (ϕ) and the polar angle (θ) are depicted. The figure is inspired from [87].

bination of barrel and endcaps aims at maximizing the coverage of the solid angle around the interaction point. Each detector element is briefly discussed inside the caption of Figure 3.7.

In proton-proton collisions, the proton momentum fraction carried by each of the interacting partons in not known *a priori*. Therefore, each event exhibits a different Lorentz boost along the z-axis of its center-of-mass system with respect to the detector frame. Projections of measurable quantities into the x-y plane – referred to as *transverse* plane in the following – are invariant under these longitudinal Lorentz boosts and are therefore commonly used. An important quantity is the *transverse momentum*, denoted by p_{T},

$$p_{\mathrm{T}} = \sqrt{p_x^2 + p_y^2} \,, \tag{3.4}$$

where p_x and p_y denote the x- and y-component of the particle's momentum, respectively. Another important quantity is the *transverse* mass,

$$m_{\mathrm{T}} = \sqrt{m^2 + p_{\mathrm{T}}^2} \,, \tag{3.5}$$

which is invariant under longitudinal Lorentz boosts as well. Instead of the polar angle, the *rapidity* is used because it has better properties under longitudinal Lorentz transformation, i.e. Lorentz transformation along the z-axis. The rapidity is defined as

$$y = \frac{1}{2} \ln \left(\frac{E + p_z}{E - p_z} \right) \,, \tag{3.6}$$

where E and p_z denote the energy and the z-component of the particle's momentum, respectively. Under longitudinal Lorentz transformation, the rapidity behaves additively. Consequently, the difference between the rapidities of two particles is Lorentz invariant under such transformations. In order to get an intuition of rapidity, it is useful to consider the following extreme cases. Suppose a particle is directed in the transverse plane, i.e. it does not carry any momentum in longitudinal direction ($p_z = 0$). In this case the rapidity (see Equation (3.6)) is zero. The other extreme case is when a particle is emitted in $\pm z$-direction. Neglecting the particle's mass for simplicity, it follows by inserting $E = \pm p_z$ into Equation (3.6) that the rapidity tends to infinity, i.e. $y \to \pm\infty$. Hence, rapidity encodes angular information about how much a particle is directed out of the transverse plane and interleaves it with the energy of the particle in a single coordinate. In summary, the energy and momentum of a particle can be written in terms of transverse momentum,

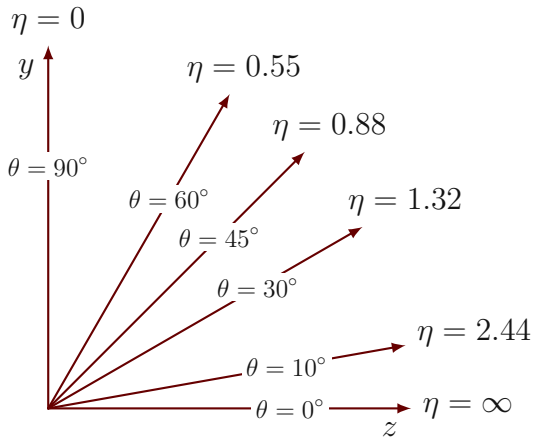


Figure 3.6: The correlation between the polar angle (θ) and the pseudorapidity (η) is illustrated with few concrete numerical examples. The figure is taken from [87].

transverse mass, rapidity and azimuth [49]

$$\begin{aligned} E &= m_{\mathrm{T}} \cdot \cosh y \\ p_x &= p_{\mathrm{T}} \cdot \cos\phi \\ p_y &= p_{\mathrm{T}} \cdot \sin\phi \\ p_z &= m_{\mathrm{T}} \cdot \sinh y \ . \end{aligned} \tag{3.7}$$

Instead of rapidity, a modified version called *pseudorapidity* (η) is commonly used in high-energy physics. Pseudorapidity is defined as the limit of rapidity (see Equation (3.6)) in case of a massless particle ($E = |\mathbf{p}|$)

$$\lim_{m\to 0} y = -\ln\tan(\theta/2) \equiv \eta \ , \tag{3.8}$$

where the identity $p_z = |\mathbf{p}| \cdot \cos\theta$ was used. Hence, pseudorapidity is directly linked to the polar angle as illustrated Figure 3.6. High-energy particles recorded by the CMS detector typically have momenta much larger than their rest mass, explaining the wide usage of pseudorapidity. For example, detector elements are divided in pseudorapidity regions and are also designed that they cover the same area in (η, ϕ)-space. The angular distance (ΔR) between two particles – labeled with 1 and 2 – is defined as

$$\begin{aligned} \Delta R &= \sqrt{\Delta\phi^2 + \Delta\eta^2} \ , \text{ where} \\ \Delta\phi &= \phi_2 - \phi_1 \text{ and} \\ \Delta\eta &= \eta_2 - \eta_1 \ . \end{aligned} \tag{3.9}$$

The angular distance is a key ingredient in the formulation of cluster algorithms as well as in definitions of particle isolations as will be described in Section 3.5. In Section 3.5.7, the reconstruction of particle bundles consisting of massive particles is discussed. Since in that case the mass of the analyzed object can not be neglected, rapidity is used instead of pseudorapidity and thus the angular distance is defined as (see also Equation (3.18))

$$\Delta R = \sqrt{\Delta\phi^2 + \Delta y^2} \ . \tag{3.10}$$

Note, that the angular distance as defined above is Lorentz-invariant with respect to boosts in z-direction. To summarize, positions of particles inside the CMS detector are expressed in terms of azimuth (ϕ) and rapidity (y) or pseudorapidity (η). Together with

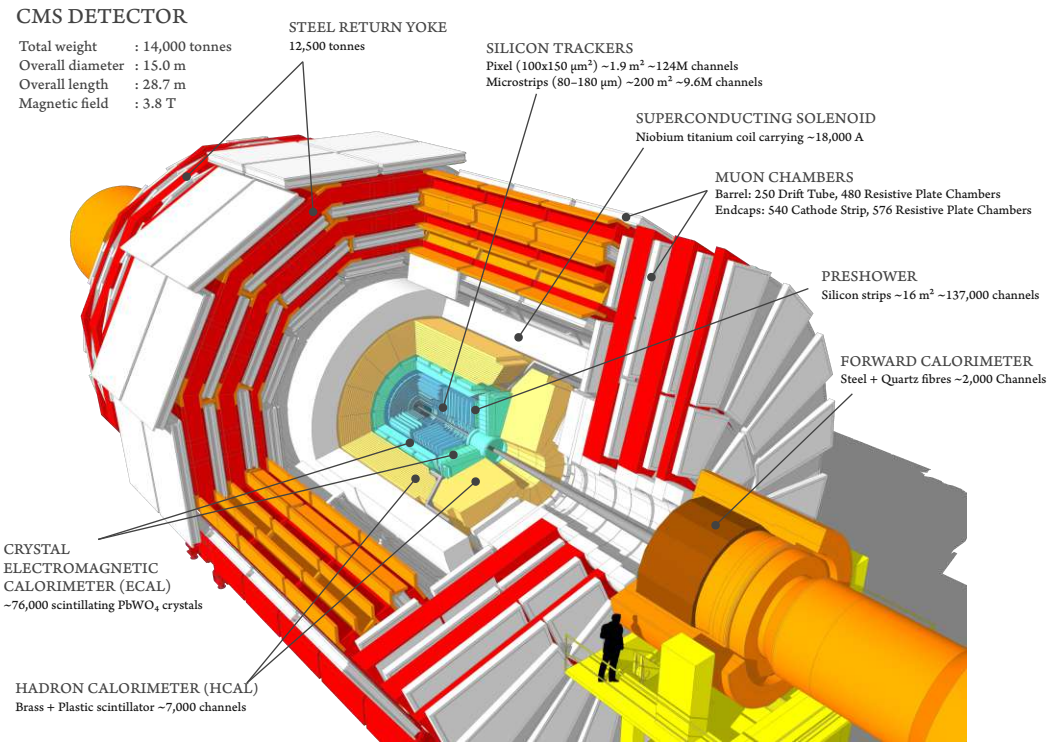


Figure 3.7: A cutaway diagram of the CMS detector is shown [93]. The beam pipe is drawn in gray and goes through the middle of the detector. Closest to the beam pipe are the silicon trackers which are used to reconstruct trajectories of charged particles. Outside the tracker volume, the calorimeters are installed. The crystal electromagnetic calorimeter measures energies of electrons, positrons and photons. Another special sub-module is the preshower detector that can very precisely distinguish between single photons and photon pairs from neutral pion decays. The hadronic calorimeter measures the energy of particles that interact via the strong force, e.g. pions. All these detector elements are placed inside a superconducting solenoid that provides a homogeneous magnetic field of 3.8 T. Inside the steel return yoke, the muon chambers are installed. Furthermore, a special forward calorimeter operates outside the CMS endcaps, extending the coverage of the detector.

the transverse momentum, every component of the momentum vector of a particle can re-written as

$$\begin{aligned} p_x &= p_\mathrm{T} \cdot \cos\phi \\ p_y &= p_\mathrm{T} \cdot \sin\phi \\ p_z &= p_\mathrm{T} \cdot \sinh\eta \ . \end{aligned} \tag{3.11}$$

In the following each detector element is described in more detail, starting from the element closest to the beam pipe and then moving further away from the beam pipe. The description will cover the most important elements and design considerations, many more details can be found in [88] or in dedicated documentations about specific detector parts and upgrades [89, 90, 91, 92].

3.2.1 Tracking System

Closest to the beam pipe is the tracking system. It detects charged particles traversing the detector material while aiming to disturb their trajectory as little as possible. From the reconstructed positions of traversing particles, the particle's trajectory can be reconstructed as described in Section 3.5.1. Since the whole tracking system is exposed to a 3.8 T magnetic field, the trajectories of charged particles are bent due to the Lorentz force.

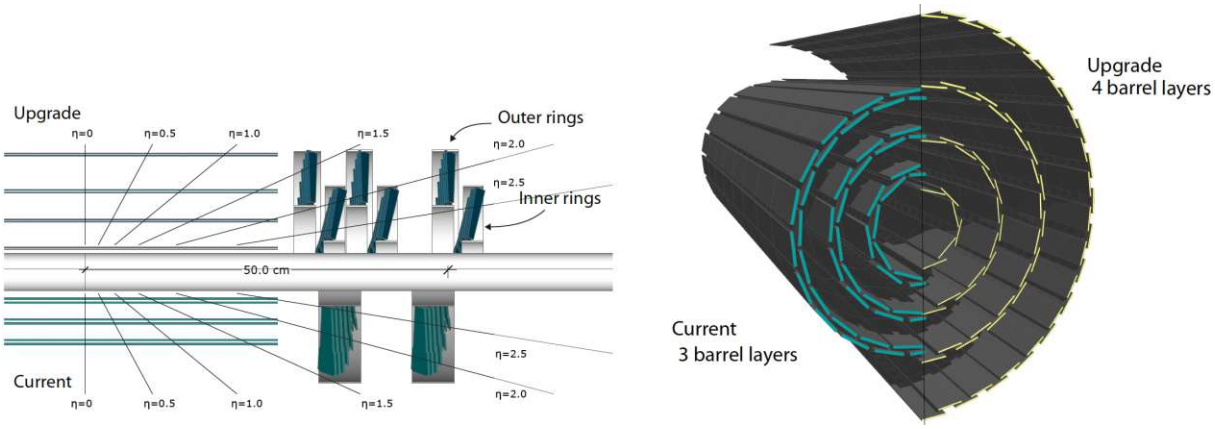


Figure 3.8: The layout of the CMS pixel detector is sketched in both of its versions during Run2 data-taking. The layout with three layers in the barrel region, labeled as "Current" in the graphic, is used for recording data in 2016. For 2017/18, the upgraded pixel detector, labeled as "Upgrade" in the graphic, has four layers in the barrel region as well as a more sophisticated geometrical setup in the endcap disks. The figure is taken from [98].

A measurement of the bending radius allows determining the transverse momentum of the particle. One distinguishes between inner and outer tracking system, the two of which are discussed below in separate sections.

Inner Tracking System - Pixel Detector

The inner tracking system is depicted in Figure 3.8. It was upgraded between the data-taking years 2016 and 2017. After the upgrade, the number of layers the barrel region was increased from three to four. Each layer consists of silicon modules, where each module is built out of rectangular readout elements with an area of about $0.01\,\mathrm{mm}^2$. These readout elements are referred to as pixels and therefore the inner tracking system is also called *pixel detector*. The endcap region of the pixel detector was also upgraded. For the 2017/18 data-taking period each endcap disk consists of inner and outer rings as shown in Figure 3.8. After the upgrade, the pixel detector counts 124 million individual pixels, before there were 66 million pixels.

The pixel detector suffers a lot from radiation damage because it is close to the collision point. Therefore, the upgrade in 2016 was on one hand needed to replace the old pixel detector which has received already a considerable radiation dose and was at the end of its design life time. On the other hand, the new pixel detector allows recording of up to four crossings of charged particles through the detector material. This is important since in 2017 and 2018 the number of simultaneous interactions per bunch crossing has increased with respect to 2016 (see Figure 3.4). The upgrade ensures a high real track reconstruction efficiency while keeping at the same time the misidentification rate low [94, 95, 96].

The pixel detector reconstructs a three-dimensional point of crossing of charged particles (hit) and reaches a precision of $10\,\mu\mathrm{m}$ in the transverse plane and a precision of $20\,\mu\mathrm{m}$ in the longitudinal direction (z-direction) [97].

Outer Tracking System - Silicon Strip Detector

The outer tracking system consists of silicon strip sensors (modules) and is shown in Figure 3.9. Different regions of the silicon strip detector are defined depending on the position of the modules. Within a radial distance from the beam direction of $200\,\mathrm{mm} < r < 550\,\mathrm{mm}$, four layers comprise the tracker inner barrel (TIB), where strips are oriented

parallel to the beam axis (z-axis). Three disks form the tracker inner disk (TID), which has radially mounted silicon strip sensors and is placed in the forward direction of the TIB. All the sensors in the TIB and the TID have a thickness of 320 µm.

The tracker outer barrel (TOB) consists of six layers surrounding both, the TIB and TID. Lastly, the tracker endcap (TEC), consisting of nine disks, is mounted in the region $|z| > 1240\,\text{mm}$. As can be seen from Figure 3.9, each disk of the TEC has different number of rings ranging from four in the outer most disks up to seven rings in disks closest to the barrel region. In these outer regions thicker sensors – measuring 500 µm – are used.

Orienting strips parallel to the z-axis in the barrel region allows measuring the (r, ϕ) positions of traversing particles. By combining two strip modules back-to-back and rotating one module with respect to the other, it is possible to simultaneously measure the (r, z) position. Such modules are called *stereo modules* and the angle of rotation between the two single modules is called *stereo angle*. Hence, stereo modules enable a three-dimensional hit position measurement. The stereo angle used in the design of the CMS strip tracker is 100 mrad. Stereo modules are shown in blue in Figure 3.9 and are also mounted inside the TID and TEC.

Modules in the TID and TEC have their strips radially oriented and measure the (ϕ, z) coordinate of traversing particles. Stereo modules enable a simultaneous measurement of (r, z) positions. Depending on the location of the strip module, the distance between two strips – called *pitch* – varies. Regions closer to the beam pipe have a strip pitch of 80 µm, while those further away can have a strip pitch up to 200 µm. As detailed in [99], the strip sensor resolution strongly depends on the pitch size. Resolutions further depend on the (η, ϕ) position of the sensor. Sensors in the TIB for instance, achieve a spatial resolution in the transverse plane of about 20 µm, whereas in the TOB it amounts to about 40 µm. The resolution achieved by stereo modules in longitudinal direction is about one order of magnitude worse.

In total more than 15 thousand silicon strip modules make up the whole silicon strip detector. The accuracy of the hit position determination is essential for a precise reconstruction of the trajectories of charged particles. Since the mounting precision is limited, track-based alignment and calibration is used to achieve a more precise knowledge of each module's position. An example of such a track-based alignment procedure will be discussed in chapter 4.

3.2.2 Calorimeters

Calorimeters are used to measure the energy deposits of particles interacting with the detector material. In order to maximize the deposits, a dense material is used to increase the probability of an interaction between the entering particle and the detector material. After each such interaction a particle looses energy and produces secondary particles. In that way, the original high-energy particle produces a whole cascade of lower energetic particles within the calorimeter. The presence of the lower energetic particles has then to be detected and read out. Two calorimeters are used in the CMS detector which are presented in the next two sections in more detail.

The Electromagnetic Calorimeter

The electromagentic calorimeter (ECAL) surrounds the CMS tracking system and measures the energy of photons, electrons and positrons, which will be collectively referred to as e/γ in the following. Figure 3.10 depicts the ECAL together with its subdivision in barrel ECAL, endcap ECAL and preshower. The barrel and endcap ECAL uses more than 75'000 scintillating lead tungsten crystals to absorb and measure the energy of e/γ.

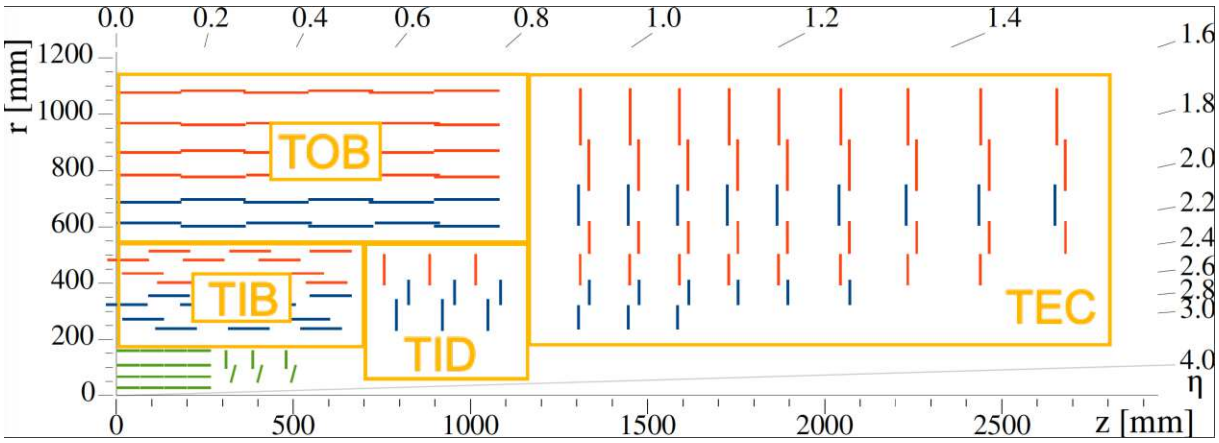


Figure 3.9: A schematic layout of the CMS tracking system [100] is shown. The tracker layout is symmetric with respect to rotation around the z-axis. The y-axis measures the radial distance from the beam direction. In green, the upgraded version of the pixel detector is shown (see also Figure 3.8). In blue and orange different silicon strip modules are depicted. Blue sensors reconstruct a three-dimensional hit position of traversing charged particles, whereas orange sensor can reconstruct only a two-dimensional hit position. The sensor thickness of modules in the TIB and TID is 320 µm and the thickness of TOB and TEC sensors is 500 µm [101]. The names of different parts of the silicon strip detector were additionally added after retrieving the image.

The crystals are very dense and induce electromagnetic showers of the high-energy impact particles. Secondary particles inside the particle shower produce electron-hole pairs within the crystal that radiate scintillation light which is then read out and translated to an electric signal by using photo-diodes. The crystal serves as both, absorbing and detection/readout material.

Lead tungsten has further features which make its use in the ECAL beneficiary. The electromagnetic shower is very narrow, characterized by a Molière radius[3] of 22 mm. Hence, the front-face of crystals is chosen to be 22×22 mm^2 and 28.6×22 mm^2 in the barrel and endcap ECAL, respectively. The radiation length[4] of lead tungsten is 89 mm. The length of the crystals is 22 cm and 23 cm in the barrel and endcap ECAL, respectively, ensuring that e/γ are fully absorbed inside the crystal. Furthermore, the light emission happens quickly. Within 25 ns, 80% of scintillation light is collected and 99% of the light is collected within 100 ns. The relative energy resolution of the ECAL is about 1.5% for energies between 10 GeV to 50 GeV and improves to 0.5% for higher energies up to 500 GeV [102]. Higher energetic particles are not guaranteed to be sufficiently absorbed within the crystal and the energy loss due to punch-through has to be taken into account, ultimately resulting in an increased energy resolution.

Neutral pions decay into a pair of photons. In order to distinguish them from high energy photons coming from the interaction point, a preshower detector is used. The preshower uses lead as absorber and silicon strips to track electrons and positrons from the induced shower.

[3]The Molière radius is defined as the cone radius of the induced shower that contains 80% of the energy.

[4]The radiation length characterizes the distance at which the incoming particle looses a certain amount of energy. For an electron, for instance, it is defined as the average distance where its energy has dropped by a factor $1/e$.

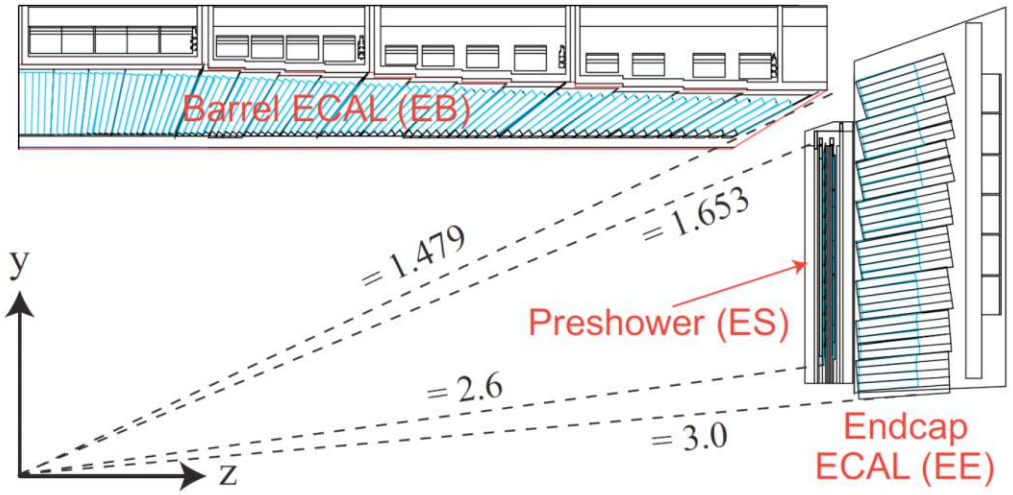


Figure 3.10: The layout of the ECAL is shown. It has a barrel part and an endcap part. In front of the ECAL endcap, a preshower detector is installed. The values next to the dashed lines indicate the respective pseudorapidities. In the region $1.479 < |\eta| < 1.653$, there is a necessary gap in the ECAL for the readout electronics. The figure is taken from [88].

The Hadron Calorimeter

The hadron calorimeter (HCAL) measures the energy of particles engaging in nuclear interactions. Examples are baryons like the proton and neutron or mesons like pions and Kaons. A schematic drawing of the HCAL is depicted in Figure 3.11. The barrel and endcap parts of the HCAL enclose the ECAL and they fill the space inside the magnetic coil. Hadrons are absorbed using layers of brass as absorbing material, inducing hadronic showers that are translated to a readout signal with interleaved plastic scintillator tiles. The stopping power is characterized by the nuclear interaction length[5]. It amounts to 5.8 and 10.6 times the interaction length for $|\eta| = 0$ and $|\eta| = 1.3$, respectively. Hadrons also deposit a small amount of energy inside the ECAL, which corresponds to roughly one nuclear interaction length.

High energetic particles can punch all the way through the HCAL. A punch-through is also possible if the hadronic shower starts late, i.e. already in some depth of the HCAL. Therefore, an outer HCAL is installed outside the magnetic coil. Furthermore, the outer HCAL is used to positively distinguish between punch-through hadrons and muons which are detected in the muon system (see Section 3.2.4). It uses the iron yoke as absorber, where in the central region, i.e. $|\eta| \approx 0$, an additional layer of iron is placed. Together these parts of the HCAL cover the pseudo-rapidity range $0 < |\eta| < 3.0$.

A forward HCAL extends the pseudo-rapidity up to $|\eta| < 5.2$, making it the most hermetic subsystem of the whole CMS detector. The large coverage is very important to precisely determine the missing transverse energy (see Section 3.5.9). Being so close to the beam line, the forward HCAL needs to be very radiation hard. Therefore, iron as absorber and quartz fibers to detect the induced Cherenkov light are used here.

The energy resolution of the HCAL can be as much as 100% for energies below 20 GeV, reaches values of 20–30% below 50 GeV and saturates at around 10% for higher energies [103].

[5]The nuclear interaction length characterizes the distance at which the incoming particle looses a certain amount of energy.

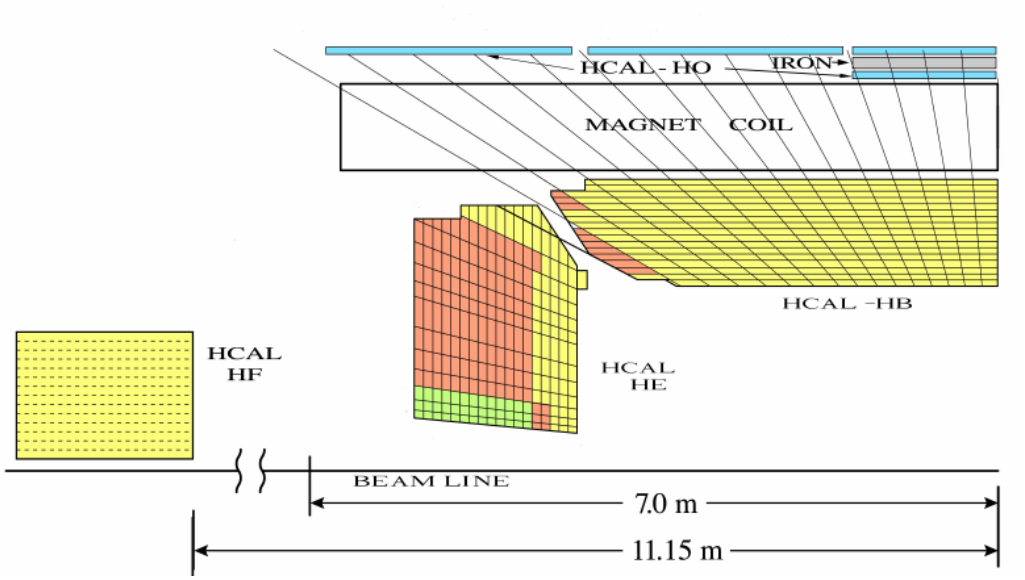


Figure 3.11: A schematic drawing of the HCAL is shown. The different parts are the barrel (HB), the endcap (HE), the outer HCAL (HO) and the forward HCAL (HF), respectively. The image is taken from [103].

3.2.3 The Superconducting Solenoid

It is already mentioned in Section 3.2.1 that the magnetic field is crucial in order to determine the transverse momentum of charged particles emerging from proton-proton collisions. The bending radius is proportional to the particles' transverse momentum and can be measured using the reconstructed hits inside the tracker. A strong magnetic field makes it possible to measure the bending for particles even with high transverse momentum. The magnetic field is measured to be 3.8 T and is oriented along the z-axis in most parts of the detector, thus resulting in a bending in ϕ-direction. The magnetic field strength is known down to a precision of 0.1% over the whole volume of the CMS detector [104]. For ultimate precision of track parameters, the variations in the magnetic field need to be taken into account, even if they are small. 12'000 tons of steal form the return yoke of the magnet and capture the magnetic field outside the solenoid. Figure 3.12 shows the magnetic field map of the CMS detector.

3.2.4 Muon Detection System

Muons leave hits inside the tracking system, but they only interact minimally with the ECAL and HCAL. Dedicated muon gas chambers are installed outside the iron return yoke of the CMS detector to measure muons as illustrated in Figure 3.13. There are three different types of gas chambers used in the CMS detector. In all of them the charge from gas ionization due to traversing muons is collected on electrodes in form of wires or plates. The choice of detector technology is driven by the magnetic field homogeneity inside the return yoke and expected muon flux.

Drift tubes (DTs) operate well in homogeneous magnetic fields and when exposed to rather low muon fluxes and hence are installed in the pseudo-rapidity range $|\eta| < 1.2$. In the endcap region, $1.2 < |\eta| < 2.4$, the magnetic field is less homogeneous and the muon flux is higher. Therefore, cathode strip chambers (CSCs) are used in that region. Both detector technologies measure a traversing muon hit position with a precision of about 0.1 mm [102]. The third detector type are resistive plate chambers (RPCs), which are installed in both barrel and endcap region. Their advantage is a very fast response

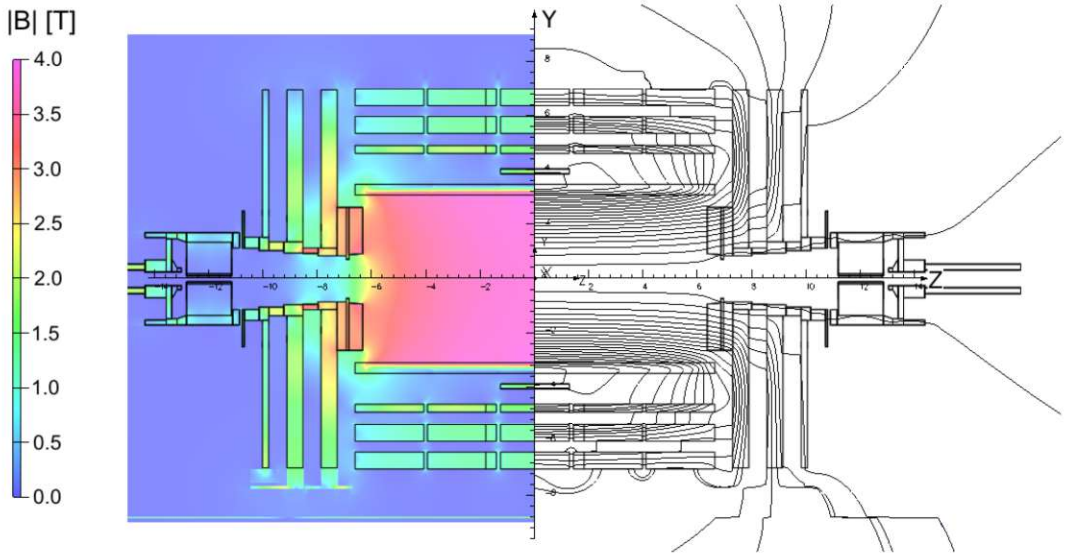


Figure 3.12: The magnetic field map on a longitudinal section of the CMS detector is shown. On the left, the absolute value of the magnetic field is displayed in a color map. On the right, magnetic field lines are shown. The figure is taken from [104].

time of roughly 1 ns, which is exploited in making fast decisions on whether to record a certain collision event or not as discussed in the next section.

3.3 Trigger System

As discussed in Section 3.1, the peak crossing rate of proton bunches at the LHC is 40 MHz. Each bunch crossing defines an event and typically consists of several proton-proton collisions. An event requires about 1 MB of storage. Hence, collecting all data would fill up to 40 TB of disk space each second. This high data rate can not be recorded. However, many collisions describe well-understood physical processes and can thus be discarded. The goal of the CMS trigger system is to reduce the data rate and save potentially interesting collisions of rare or even new physics processes. To achieve this goal, a two step approach is followed [105]. First, the so-called level-1 trigger reduces the data rate to 100'000 events per second. Afterwards, the high-level trigger (HLT) further reduces the data rate to about one thousand collisions per second that are permanently saved on disk for future analyses. In the following a brief discussion of the two-tiered trigger system is given.

The level-1 trigger uses calorimeter information and the signals coming from the muon system. Custom hardware-based algorithms estimate energies inside ECAL and HCAL as well as transverse momenta of muons. Sorting of energies inside the ECAL and HCAL is done in parallel and so does applying quality requirements on the muons from the different muon detector subsystems. Combining all information results within a fixed latency of about 4 µs in a decision of the level-1 trigger to keep or discard the collision event. If the event is kept, the full detector readout is triggered. Information from all detector elements are assembled into one event and passed to the HLT for further analysis.

The HLT runs more complex reconstruction algorithms, including information from all detector subsystems, i.e. including also tracker information. For instance, the tracker hits can be combined with the information from the muon system to get a more precise estimate of the muon's transverse momentum. Particle trajectories are reconstructed using standard algorithms – discussed in Section 3.5 – which are optimized to reduce processing

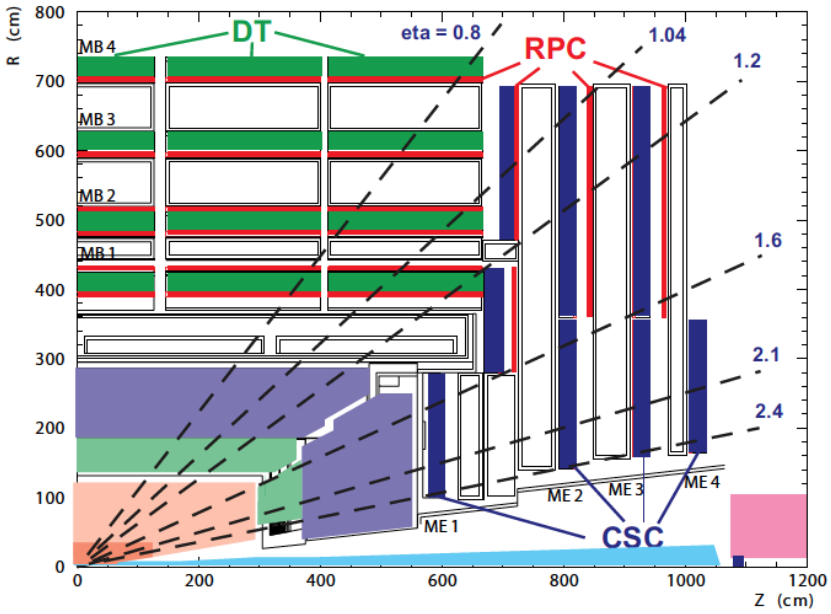


Figure 3.13: A schematic drawing of the CMS muon detection system is shown. DTs are highlighted in green and operate in the barrel region. CSCs are installed in the endcap region and are drawn in blue. Finally, RPCs are shown in red and are used in the pseudo-rapidity range $|\eta| < 2.1$. The figure is taken from [88]

time. If an event is flagged to be written to disk, all information from the detector as well as all the level-1 and HLT information that led to this decision are saved. All saved events are processed using the final reconstruction algorithms, discussed in Section 3.5.

3.4 Simulation of Proton-Proton Collisions

Previous sections discussed detector subsystems and their working principles. A vital ingredient in understanding the CMS detector and performing any kind of analysis is the simulation of proton-proton collisions. For example, to study physical processes taking place in proton-proton collisions, data acquired by the subsystems of CMS have to be reconstructed and compared with simulated collisions. Known physical processes can constitute backgrounds for new physics searches or can be used to calibrate and better understand the detector response[6]. Ultimately, simulation and collision data are reconstructed and identified using the same algorithms.

As discussed in [106], Monte Carlo (MC) techniques are well suited to model physical processes involved in proton-proton collisions. The factorization theorem [107] allows simulating these processes in different sequential steps. What is meant by that is illustrated in Figure 3.14 and will be discussed in the following.

A proton-proton collision is an involved process because the proton is a composite particle (see also Section 3.1). The so-called *parton distribution functions* represent probability densities for particular partons to carry a given fraction of the proton's momentum. For high momentum transfers, as they happen in proton-proton collisions at the LHC, frequently most of the momentum is actually carried by gluons. Hence, it is most likely

[6]The detector response describes the way electrical readout signals behave under particle-detector interactions. For example, take a detector producing a current pulse from collecting charges from ionization of the detector material. The amplitude of the pulse is correlated to the amount of ionization induced by the particle-detector interaction. If the relation is linear, one speaks of a linear detector response.

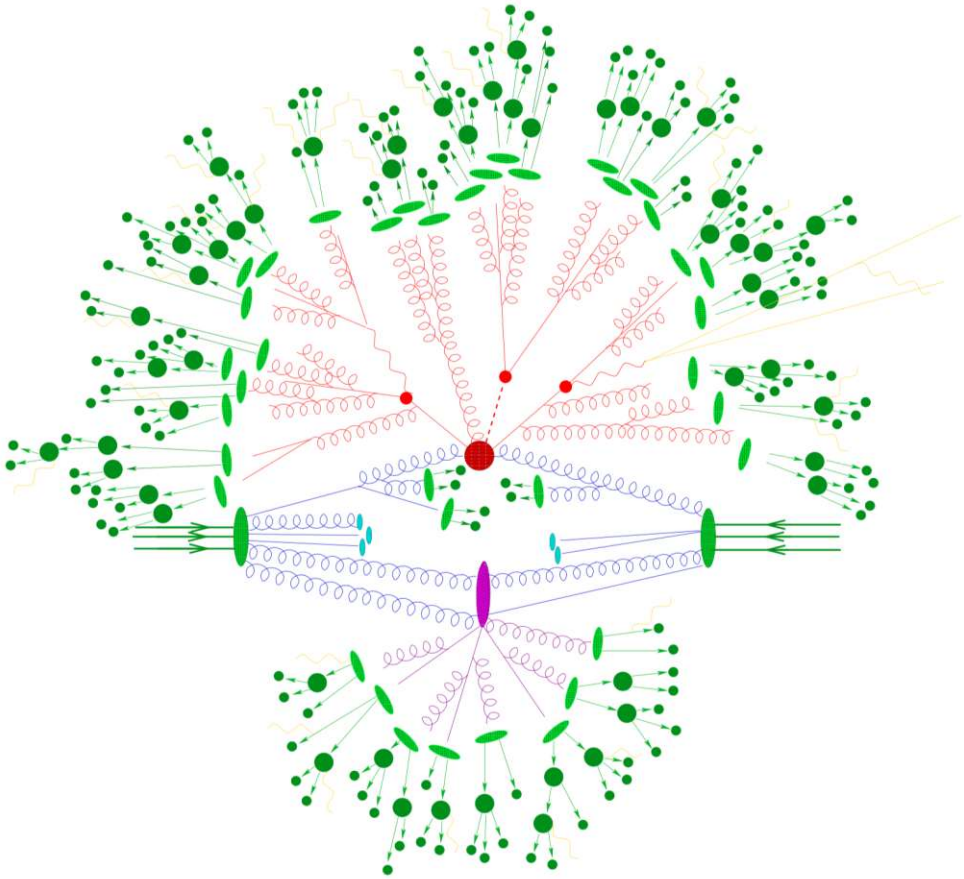


Figure 3.14: Different steps in a MC simulation of a proton-proton collision are displayed. Three lines with an arrow represent the valence quarks of the colliding protons (represented by big green ellipses). Two partons from the incoming protons interact in a hard process shown by a large red-brown circle. Before this hard interaction, both partons undergo initial state radiation. Parton showers evolve from the hard scattering process. A secondary interaction between proton remnants is drawn as big purple ellipse and other beam remnants are displayed as small blue blobs. Together they form the underlying event. Hadronization and color-neutral hadrons are indicated by small light-green ellipses. The decay of these hadrons is displayed as dark-green circles. Photon radiation is shown in yellow. The figure is taken from [108].

that two gluons interact with each other and produce (pairs of) strongly interacting particles. Within that regime, partons are asymptotically free and therefore, perturbative QCD can be used to calculate this process. Calculation up to leading-order (LO), next-to-leading-order (NLO) and higher in α_S (see Equation (2.16)) are implemented in different MC generators such as MadGraph5 (MG5) aMC@NLO [109] and POWHEG [110].

The classification of parton interactions into *soft* and *hard* depends on the momentum transfer and is subjective. In the CMS experiment, typically hard interaction processes are analyzed because they produce outgoing particles with appreciable transverse momentum or mass. In Figure 3.14, the partons involved in the hard scattering are highlighted in red. Apart from the hard scattering, the following processes can occur:

- Partons can radiate off other particles before and after QCD interaction vertices. These are termed *initial* or *final state radiation*, depending on when the radiation happened. Quarks can radiate off both gluons or photons since they carry both color and electric charge.
- It can happen that further *semi-hard* parton-parton interactions occur. This process is termed multi-particle interaction (MPI).
- Final states originating from partons not participating in any hard scattering are collectively called *beam-beam remnants*.

Since free partons do not exist in Nature, they have to form color-neutral hadrons. High-energy partons reduce their energy by splitting into other partons in a process called *fragmentation*. This energy reduction evolves a whole cascade of ever lower energy particles called *parton shower* (see also Figure 3.14). After particle energies within a parton shower have dropped to a level of about 1 GeV, partons recombine to form color-neutral hadrons in a process called *hadronization*. It is also at this energy threshold were QCD can not be treated perturbatively anymore and phenomenological models are used to model the hadronization. As *underlying event* one generally classifies the fragmentation of partons not undergoing a hard scattering. It is experimentally impossible to tell apart the different types of underlying event on an event-by-event basis. As stated in [111]: *There is only an event and one cannot say where a given particle in the event originated.*

Underlying event, parton showering and hadronization are usually modeled by the PYTHIA[112, 113] MC generator. Everything up to the level of stable hadrons happens at a length scale much shorter than the CMS detector can measure. The simulation of unstable particle decays is often performed by dedicated generators, e.g. TAUOLA [114] is used to model the hadronic decay of tau leptons. Furthermore, the interaction of all particles with the detector have to be simulated. A full detector model is implemented inside the GEANT4 [115] simulation toolkit, modeling particle interactions with the detector material. The simulation of the electronic readout is performed in a last step and is referred to as *digitization* such that both, MC based collision events and real data enter the reconstruction on the same footing. Before discussing the reconstruction of physics object, there is one last effect which needs to be simulated and taken into account. Namely, there is more than one proton-proton collision happening in one bunch crossing. However, the CMS trigger system (see Section 3.3) only acts on the collision of interest which comes from a single proton-proton collision. All extra proton-proton collisions, referred to as PU, need to be mixed into the simulated collision of interest. The number of simultaneously occurring collisions for a fixed luminosity is distributed according to a Poisson distribution. Since the luminosity varies over time, the Poisson distribution changes as well. Luminosity conditions during a LHC run are only known after the data are recorded, at this time, simulations have already completed. Therefore, a re-weighting procedure is

used to match the PU distribution of MC simulation to the one in recorded collision data (see Section 6.5.8).

3.5 Reconstruction and Identification of Physics Objects

This section explains how physics object are reconstructed and identified starting from the electronic signals of all detector subsystems. Physics objects include electrons, photons, muons and also more complicated objects such as jets, hadronically decaying tau leptons and missing transverse energy. First, the track reconstruction is explained followed by the explanation of the particle-flow algorithm. The outcome of the particle-flow algorithm is the basis for the reconstruction of further physics objects. Emphasis is given on the reconstruction and identification of hadronically decaying tau leptons because they appear in final states of many analyses presented in this thesis.

3.5.1 Track and Vertex Reconstruction

The CMS tracking system (see Section 3.2.1) detects traversing charged particles. Hit positions in each layer of the tracking system are reconstructed using the methods described in [99]. Track reconstruction means to associate hits with a charged-particle trajectory, and to estimate the trajectory parameters: momentum and position. Reconstructed hits serve as input to a combinatorial track finder (CTF) algorithm [116, 99] based on Kalman filtering (KF) [117]. Track reconstruction with the CTF algorithm can be broken down in three main components:

1. **Initial seed generation:**
   Small collections, called seeds, of hits inside the pixel and/or strip detector are formed which are compatible with a charged-particle trajectory. An initial set of track parameters based on these seeds is derived. These parameters are used to extrapolate the track to the next layer and look for compatible hits (see next step).

2. **Trajectory building:**
   The KF technique is used to iteratively propagate the charged-particle track, starting from the track seed, through each layer of the tracking system and adding compatible hits to the track. After each successful addition of a hit, the track parameters are updated.

3. **Final fitting:**
   After the track has been propagated through all the layers, the KF is used to refit and smooth the track as well as estimate the final track parameters and their corresponding uncertainties at any point of the trajectory. Quality criteria are imposed on the track to decide whether to keep the track or not. In case the track is flagged to be kept, all hits inside the track are masked and can not be used for track formation of other tracks.

The resulting track parameters are used to determine charged-particle properties like the transverse momentum or the origin of the particle track.

As detailed in [99], an iterative application of the CTF algorithm increases the track reconstruction efficiency while keeping the rate of misreconstructed tracks low, at the level of a few per cent. The difference between each iteration is given by the way the track seeds are determined and what track selection requirements are imposed. For instance, the first

iterations aim at reconstructing tracks originating at the place the hard scattering process occurred and therefore require track seeds to have three to four hits in the pixel tracker. Of course, requiring four pixel hits is only possible after the pixel detector upgrade (see Figure 3.8) and with four pixel detector layers, the efficiency of finding pixel-triplet seeds is increased. Final iterations are seeded from reconstructed hits inside the muon chambers. By using relaxed quality criteria, muon tracks can be recovered, increasing their overall reconstruction efficiency.

Reconstructed tracks are grouped together if they originate from a common intersection point, called primary vertex (PV). The grouping uses the deterministic annealing algorithm [118] and the set of all PVs is ordered according to the sum of squared transverse momenta of tracks linked to the vertex. The vertex ranked first is called *the* PV and is assumed to be the origin of the hard scattering process (see Figure 3.14). Particle tracks associated to the PV are referred to as *prompt* tracks and corresponding particles are assumed to have originated from interactions of the hard scattering process. All PVs are linked to at least two tracks and the adaptive vertex fitter [119] is used to calculate the vertex parameters, e.g. the position of the vertex. A vertex position resolution between 10 to 100 µm can be achieved, dependent on the number of associated tracks to a particular vertex [99].

3.5.2 Muon Track Reconstruction

Track reconstruction for muons uses reconstructed tracks from the pixel and strip tracker as discussed in the previous section. These tracks are referred to as *inner tracks* in the following. Furthermore, muon track reconstruction makes use of reconstructed hits inside the muon system [88, 120, 121]. Three different cases are distinguished:

Standalone: Reconstructed hits inside the muon detector's DTs and CSCs are grouped into *segments*. Segments serve as seeds for trajectory building using the KF approach. Tracks obtained after the final fitting are called *standalone-muon tracks*.

Tracker: Tracker-muon tracks start from reconstructed inner tracks with transverse momentum above 0.5 GeV and total momentum above 2.5 GeV. Inner tracks are extrapolated to the muon system and checked for compatibility with muon segments. In case of a match with at least one muon segment, quantified by the spacial distance between track and segment, the track qualifies as *tracker-muon track*.

Global: In case standalone and a tracker muon tracks are compatible with each other, the hits from all detector subsystems are refitted using the KF technique. The obtained track is declared to be a *global-muon track* which typically has better resolution of the transverse momentum than standalone or tracker muon tracks. In case a global-muon track shares the inner track with a tracker-muon track, they are merged into a single muon object candidate.

3.5.3 The Particle-Flow Algorithm – Introduction and Calorimeter Clustering

The particle-flow (PF) algorithm [122] uses as input tracks reconstructed inside the tracker systems as well as muon tracks and tries to establish links to energy deposits inside the ECAL and HCAL. A link between signals from different detector subsystems is interpreted as a particle. For example, a charged-particle track linked to an energy deposit inside the ECAL gives rise to a PF electron candidate. In that sense, the PF algorithm restores

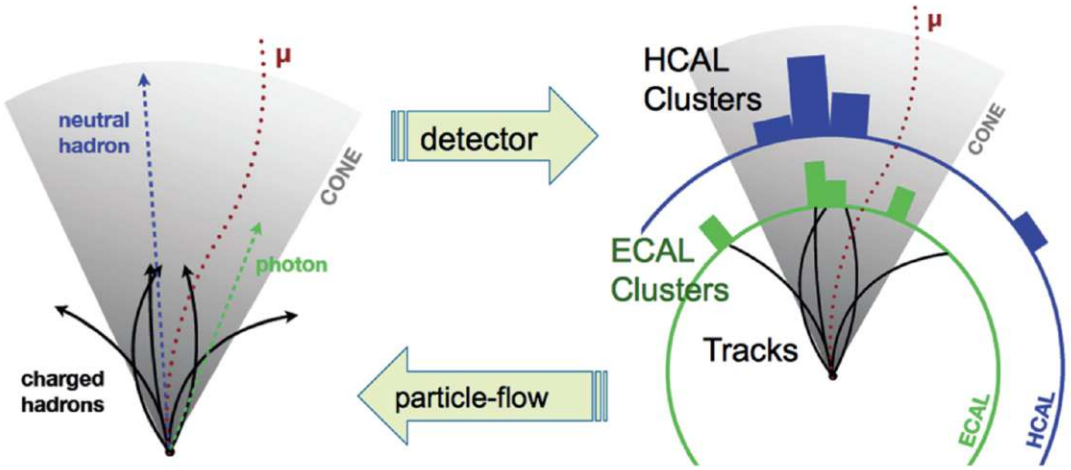


Figure 3.15: The PF algorithm clusters energies inside the calorimeters. It then looks for links between tracks and energy clusters compatible with muons and electrons. A PF electron candidate for instance is characterized by linking a charged track with an ECAL cluster. By subtracting energies of electrons and muons inside energy clusters and removing their tracks, it is possible to find neutral particles and charged hadrons in a next step. Photons are characterized by an ECAL cluster without any link to a track. The figure is taken from [123]

the particle-picture from the electronic signals recorded by the CMS detector as shown in Figure 3.15. Since absorbed particles deposit their energy over several calorimeter cells, an aggregation of theses cells has to be performed. In the following the aggregation into so-called PF clusters is discussed in more detail.

Calorimeter clustering in the context of PF event reconstruction is performed separately in ECAL barrel, endcap and preshower as well as HCAL barrel and endcap. Inside the forward HCAL (see Figure 3.11), no clustering is used and each calorimeter cell is used individually. Cluster formation starts by identifying calorimeter cells that exceed an energy threshold of typically several hundred MeV and have the highest energy deposit in their neighborhood[7]. These cells serve as seeds to the PF clustering algorithm, which in a next step aggregates neighboring cells as long as they exceed a certain energy threshold. The final cluster is referred to as *topological cluster*.

Each topological cluster consists of an ensemble of cells and typically contains several seeding cells. A Gaussian mixture model is fitted by means of an expectation-maximization algorithm to reconstruct PF clusters within a topological cluster [122]. For each seeding cell inside the topological cluster a Gaussian distribution with variable position and amplitude but fixed width is used. Final positions and energies (amplitudes) of each Gaussian after the fit serve as PF cluster parameters.

Since PF clusters are used among others to identify neutral particles like photons and neutral hadrons and measure their energy, it is crucial that the cluster energy is well calibrated. A calibration of the PF cluster energy has to be applied because the energy threshold requirement during cluster formation biases the result towards lower energies. Simulated photons for the ECAL and simulated neutral hadrons for the HCAL are used to fit analytic functions in different energy and pseudorapidity regimes. They serve ultimately as calibration functions [122]. For example, in the ECAL barrel the calibration is close to unity for large energies where threshold effects vanish, but amounts up to 20% for small energies.

[7] Energy thresholds and the definition of the neighborhood vary between ECAL and HCAL as well as inside barrel and endcap regions.

3.5.4 Electron Track Reconstruction

The basic signature an electron leaves inside the CMS detector is given by a charged-particle track followed by an energy deposit inside the ECAL. However, electrons loose a significant amount of their energy (up to 80%) in form of bremsstrahlung photons. Furthermore, bremsstrahlung photons have a 60% probability of interacting with the tracker material and produce an electron-positron pair. To reconstruct the correct momentum and energy of an electron, the changes of track bending due to energy loss of bremsstrahlung photons have to be taken into account and also the energy deposits of these bremsstrahlung photons as well as possible electron-positron pair conversions have to be collected. Figure 3.16 shows an example of an electron track reconstruction involving bremsstrahlung photons. There are two seeding strategies followed by the CMS experiment which are used for electron track reconstruction:

ECAL-based: From the position and energy of an ECAL cluster, the expected hits inside the innermost tracking layers are inferred. Starting from an ECAL cluster only works if the electron has not radiated any bremsstrahlung photons. To take bremsstrahlung photons into account, the photon ECAL cluster has to be added to the electron one. Together they form a *supercluster* (see Figure 3.16). Due to the bending in ϕ-direction of the electron-track, the supercluster formation uses a window which is narrow in η and wide in ϕ to look for potential bremsstrahlung photons. Therefore, the performance of the ECAL-based electron track reconstruction depends on its capability to build the correct superclusters. This is especially difficult for electrons with low transverse momentum. These electrons are strongly bent in the magnetic field and bremsstrahlung photons spread over a very wide ϕ-range and are easily missed by the supercluster-algorithm. Another downside of this approach is that typically several tracker hits are compatible with the ECAL cluster, resulting in increased misreconstruction rates.

Tracker-based: All tracks from the iterative CTF with a transverse momentum above 2 GeV serve as seeds for potential electrons. In case the energy loss of the electron is small, the track's χ^2 is well-behaved. If the ratio of the ECAL energy of the matched track to the track's transverse momentum is compatible with unity, the track is kept as a seed. Bremsstrahlung photons lead to the degradation of the track's χ^2 or even to missed hits along the charged-particle tracks. These charged-particle tracks are re-fitted with a Gaussian sum-filter (GSF) [124]. The GSF enables to account for sudden and substantial energy losses along a charged-particle track and hence is more suited than the KF technique for electron track reconstruction.

The addition of the track-based electron seeds increases the track reconstruction efficiency by several percent with respect to using ECAL-based tracks only.

3.5.5 The Particle-Flow Algorithm - Linking Algorithm

Particle identification with the PF algorithm is based on linking objects reconstructed within the different subdetector systems. Not all objects are checked for a link, but only nearest neighbors are considered [126]. In a last stage, the link algorithm produces PF *blocks* of elements directly linked or indirectly via some common element. More specifically, links between charged-particle tracks to calorimeter clusters are established if the extrapolated track is compatible to be within the cluster area. For electrons, GSF tracks are linked to ECAL clusters. Bremsstrahlung photons are linked by extrapolating tangents from the track to the ECAL surface looking for compatible clusters (see Figure 3.16).

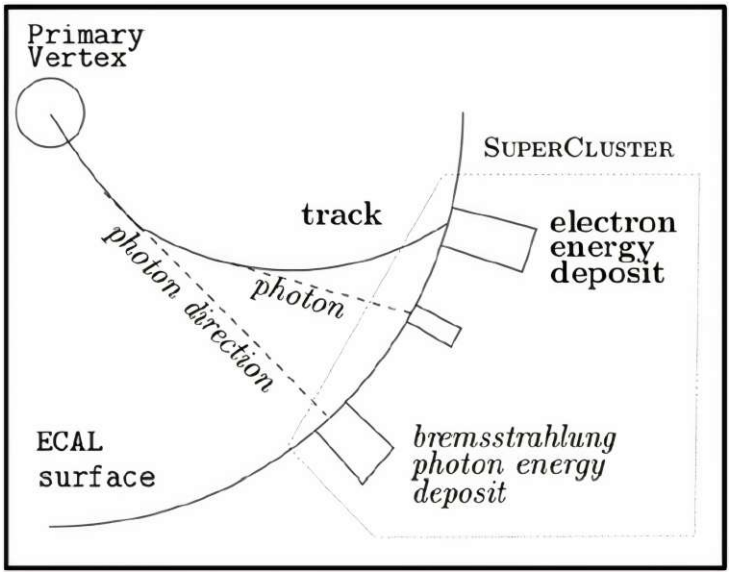


Figure 3.16: An electron originating from the PV leaves a track inside the CMS tracking system and is absorbed inside the ECAL. Two bremsstrahlung photons are also shown in the figure. These photons do not leave any tracks, indicated by a dashed line, but are absorbed by the ECAL. To retrieve the electron energy, the energy deposits of the bremsstrahlung photons are collected and merged with the electron energy deposit, forming a supercluster as detailed in the text. The figure is taken from [125].

To identify and correctly link electron-positron pairs to photons, a dedicated conversion finder [127] is used. Cluster-to-cluster links within calorimeters are also established if the cluster position of the more granular calorimeter is within the envelope of the less granular calorimeter. Mostly elements from only one particle end up in a single PF block. This is because the CMS detector

- is capable of well separating calorimeter energies of charged and neutral particles due to the strong magnetic field,
- has a fine-grained tracking system with good track reconstruction efficiency,
- has a highly segmented ECAL that can disentangle energy deposits from different particles,
- has a hermetic HCAL that has a segmentation allowing to distinguish charged and neutral hadron energy deposits and
- has a very efficient muon identification.

Identification and reconstruction of particle candidates proceed within each PF block as follows. At first, muon candidates are selected from the collection of standalone, tracker or global muon tracks using a set of loose quality criteria. Afterwards, their associated PF elements (tracks and clusters) are removed from the PF block. Next, electrons and isolated photons are identified and reconstructed followed by the removal of their corresponding PF elements. Finally, remaining clusters in the calorimeters are identified as non-isolated photons or neutral hadrons, charged hadrons in case of a link to a track. These remaining elements are used further to form more complex objects like PF jets, hadronically decaying tau leptons and missing transverse energy.

3.5.6 Muon and Electron Identification

In the first steps of the PF particle identification, muons and electron candidates are selected. Hereby, it is important to distinguish between muons and electrons from hadron decays from so called *isolated* ones. Muons or electrons from decays of electroweak bosons do not have large hadronic activity around them as opposed to those from hadron decays. There are several ways of quantifying the hadronic activity in the vicinity of a muon or electron candidate. All of the presented versions are based on an isolation cone with a radius parameter, ΔR, defined in the (η, ϕ)-plane as introduced in Equation (3.9) ($\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$). As can be seen from Figure 3.17, ΔR is the radius of a cone originating from the particle's PV. The vertex is either known from the track information or is extrapolated using calorimeter information. While it is possible to match charged hadrons ($\mathrm{h}^{\pm}_{\mathrm{PV}}$) to primary vertices, this is impossible for neutral hadrons (h^0) or photons (γ). Therefore, the addition of the contribution from neutral particles needs to be corrected for PU effects. A generic expression for the isolation sum is given by

$$I = \sum_{\Delta R} p_{\mathrm{T}}^{(\mathrm{h}^{\pm}_{\mathrm{PV}})} + \max\left(0, \sum_{\Delta R} p_{\mathrm{T}}^{(\mathrm{h}^0)} + \sum_{\Delta R} p_{\mathrm{T}}^{(\gamma)} - \sum_{\Delta R} p_{\mathrm{T}}^{(\mathrm{PU})}\right) , \tag{3.12}$$

where the transverse momentum of neutral hadrons and photons from PU is denoted by $p_{\mathrm{T}}^{(\mathrm{PU})}$.

For the muon isolation sum, the PU contribution from neutral particles is estimated via

$$p_{\mathrm{T}}^{(\mathrm{PU})} = \Delta\beta \cdot p_{\mathrm{T}}^{(\mathrm{h}^{\pm}_{\mathrm{PU}})} , \tag{3.13}$$

where $p_{\mathrm{T}}^{(\mathrm{h}^{\pm}_{\mathrm{PU}})}$ is the transverse momentum of charged particles linked to PU vertices. The factor $\Delta\beta = 0.5$ is an empirical value defined as the ratio of neutral to charged particles in inelastic proton-proton collisions. Hence, the isolation sum for muons, $I^{(\mu)}$, is also called $\Delta\beta$-corrected isolation sum and reads

$$I^{(\mu)} = \sum_{\Delta R} p_{\mathrm{T}}^{(\mathrm{h}^{\pm}_{\mathrm{PV}})} + \max\left(0, \sum_{\Delta R} p_{\mathrm{T}}^{(\mathrm{h}^0)} + \sum_{\Delta R} p_{\mathrm{T}}^{(\gamma)} - \Delta\beta \cdot \sum_{\Delta R} p_{\mathrm{T}}^{(\mathrm{h}^{\pm}_{\mathrm{PU}})}\right) . \tag{3.14}$$

Normalizing the isolation sum with respect to the muon transverse momentum, $p_{\mathrm{T}}^{(\mu)}$, defines the *relative* muon isolation

$$I_{\mathrm{rel}}^{(\mu)} = \frac{1}{p_{\mathrm{T}}^{(\mu)}} \cdot \left[\sum_{\Delta R} p_{\mathrm{T}}^{(\mathrm{h}^{\pm}_{\mathrm{PV}})} + \max\left(0, \sum_{\Delta R} p_{\mathrm{T}}^{(\mathrm{h}^0)} + \sum_{\Delta R} p_{\mathrm{T}}^{(\gamma)} - \Delta\beta \cdot \sum_{\Delta R} p_{\mathrm{T}}^{(\mathrm{h}^{\pm}_{\mathrm{PU}})}\right)\right] . \tag{3.15}$$

Specific selections on the relative muon isolation are imposed by different analyses. Furthermore, three different identification (ID) categories for muons are defined:

Loose: All PF muon candidates that are either tracker or global muons qualify as being *loose* muons.

Medium: *Medium* muons are loose muons, where at least 80% of the layers traversed in the tracking system provided a reconstructed hit. A selection on the score that quantifies how well the muon track is compatible with the segment in the muon chambers is applied. It is stricter for a tracker muon than for a global muon. However, for global muons further selections based on track χ^2 criteria are applied that quantify the quality of

- the global muon track,

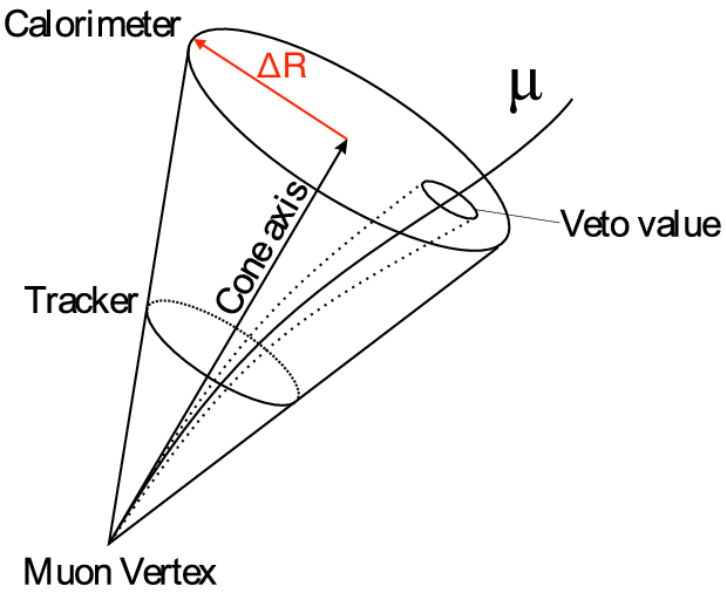


Figure 3.17: The muon track, starting from the muon vertex, passes the tracker and calorimeter layer. The isolation cone with a radius parameter ΔR (shown in red), starts at the muon vertex. In the calculation of the isolation sum, given in Equation (3.14), the muon momentum itself is not considered, indicated by the "veto value" shown in the graphic. The figure (modified) is taken from [128].

- the compatibility between tracker-only and standalone track that lead to the global track and
- evaluate the probability of a kink in the track using a dedicated kink-finding algorithm [121].

Tight: PF muon candidates that are global muons are classified as *tight* muons if they are within close proximity of the PV, have a certain minimal amount of hits in different subdetector systems (pixel and strip detector as well as in muon chambers) and possess a good track χ^2.

To identify isolated electrons, the isolation sum in Equation (3.12) is used, where the PU contribution is estimated as follows

$$\sum_{\Delta R} p_{\mathrm{T}}^{(\mathrm{PU})} = \rho \cdot A_{\mathrm{eff}}(\eta^{(\mathrm{e})}) \ . \tag{3.16}$$

In the above equation, ρ is the PU density of the collision event and A_{eff} is the effective area around the electron [129]. The effective area depends on the pseudorapidity of the electron and is a measure of the area prone to contributions from PU. The relative isolation sum for electrons reads

$$I_{\mathrm{rel}}^{(\mathrm{e})} = \frac{1}{p_{\mathrm{T}}^{(\mathrm{e})}} \cdot \left[\sum_{\Delta R} p_{\mathrm{T}}^{(\mathrm{h}^{\pm}_{\mathrm{PV}})} + \max\left(0, \sum_{\Delta R} p_{\mathrm{T}}^{(\mathrm{h}^{0})} + \sum_{\Delta R} p_{\mathrm{T}}^{(\gamma)} - \rho \cdot A_{\mathrm{eff}}(\eta^{(\mathrm{e})})\right)\right] \ . \tag{3.17}$$

Specific analysis selections are imposed on the relative electron isolation as shall be seen later. For PF electron identification, multivariate techniques based on boosted decision trees (BDTs) [130] are used. The electron ID classifier is trained on simulated $\mathrm{Z} \to \mathrm{ee}$ events and uses a large set of discriminating variables as inputs. Discriminating variables are explained in more detail in [130] and include measures of track quality and track to cluster matching. Electrons used in the di-tau analyses presented in this thesis have

to be above a certain threshold of the BDT output score. Specifically, electrons have to pass the selection threshold which has a 90% efficiency to select genuine electrons and a misidentification rate of 1% (MVA 90% ID). For the search for compressed top squarks, electrons need to pass a cut-based ID, referred to as `veto` ID, which has an average efficiency of 95% to select genuine electrons.

3.5.7 Jet Reconstruction, Identification and Tagging

Quarks and gluons produced in high-energy proton-proton collisions are not final state particles observed in the detector because of color confinement. Parton showering and subsequent hadronization produces a whole bundle of particles detectable inside the CMS detector. These bundles are called *jets* and the momentum of the jet corresponds more or less to the momentum of the original parton. First evidence for the jet structure in hadron production is discussed for example in [131]. In order to reduce the effect of PU on jet reconstruction, PF hadron candidates associated to PU vertices are removed in a method called charged-hadron-subtraction [132].

Jet clustering has evolved since the beginning of the first observation of jet structures in detectors. A sequential recombination algorithm called anti-k_T [133] is used in the CMS experiment, which is explained in the following. The anti-k_T algorithm makes use of two different distance measures:

$$\begin{aligned} d_{ij} &= \frac{\left(y_i - y_j\right)^2 + \left(\phi_i - \phi_j\right)^2}{R^2} \cdot \min\left(\frac{1}{p_{\mathrm{T},i}^2}, \frac{1}{p_{\mathrm{T},j}^2}\right) \\ d_{i\mathrm{B}} &= \frac{1}{p_{\mathrm{T},i}^2} \ . \end{aligned} \tag{3.18}$$

The radius parameter, R, is commonly chosen to be $R = 0.4$ and defines the size of the final jet. The indices i and j label all objects during cluster formation with rapidity[8] (y) and azimuth (ϕ). Initial objects are all PF charged hadron candidates, excluding those associated with PU vertices. Jet clustering then proceeds in the following steps:

1. Compute d_{ij} for all pairs of objects (i, j) as well as $d_{i\mathrm{B}}$.
2. If the smallest of all distances is of type d_{ij}, then:
   (a) Merge object i and j into a new object called *pseudo-jet*.
   (b) Recompute the distances d_{ij} between the just formed pseudo-jet and the remaining objects. Go back to step 2.
3. If the smallest distance is of type $d_{i\mathrm{B}}$, the object i is flagged as PF jet candidate and is not considered anymore in further iteration of the algorithm. Go back to step 2 if there are still unassigned objects left.

The algorithm described above will certainly stop. From the way the distances are defined (see Equation (3.18)), jets with high transverse momenta are clustered first. Another important feature, also from a theoretical point of view, of this clustering approach is that it is collinear and infrared-safe. These two concepts are illustrated in Figure 3.18 and explained in the caption therein.

Jets can be misidentified due to poor reconstruction of the track or instrumental noise within the calorimeters. In order to suppress the misidentification rate, selection criteria

[8] Note the usage of rapidity over pseudorapidity to ensure a Lorentz-invariant angular distance ΔR – using the definition of Equation (3.10) – because jets are massive objects.

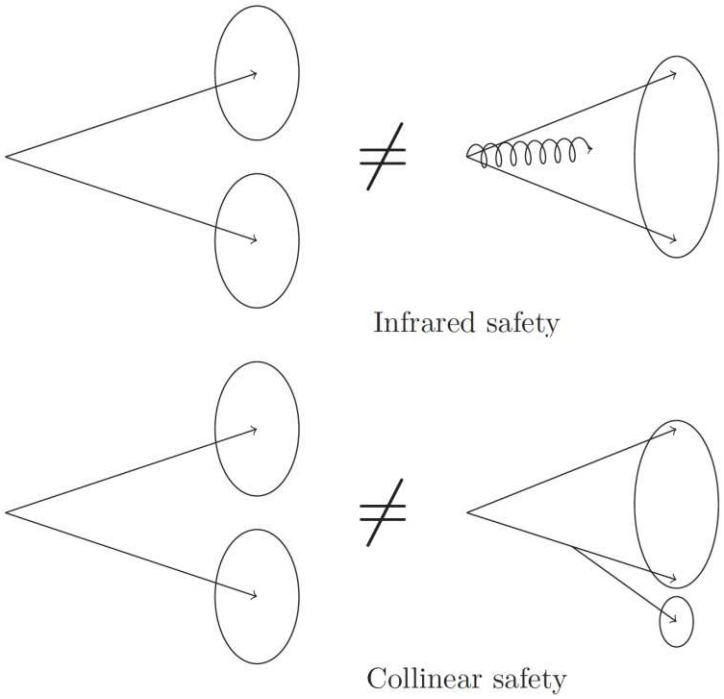


Figure 3.18: Illustration of collinear and infrared-safety taken from [80]. Top: infrared-safety means that additional soft particles (spiral line in the figure) must not change the number of jets found. On the left two jets are found indicated by the two ellipses, whereas on the right the addition of a soft (low transverse momentum) particle leads to only a single jet found. Bottom: Collinear safety means that splitting one particle in two collinear ones should not impact the result of the jet clustering.

on the number, type and energy fractions of PF candidates contributing to each jet are applied. Selections differ within pseudorapidity regions and achieve an overall efficiency above 99% to pick genuine jets while rejecting 98% of misreconstructed jets [134].

Each reconstructed jet is subject to a *jet energy calibration* [135], affecting its four-momentum. These are necessary to account for differences between true particle jets and detector-level jets reconstructed by PF and are hence applied to both, collision data and MC simulation. The overall jet energy correction consists of several parts which are applied sequentially, correcting for detector noise, pileup as well as effects of the calorimeter response. The response of the jet energy corrections, defined as the mean ratio of the reconstructed over the true jet energy, is in the order of 0.9 [122].

A lot of effort has gone into the development of so-called *jet-taggers*, which aim at further identifying the type of parton the jet has originated from. Knowing for example if a jet originates from the hadronization of a b quark is relevant when analyzing the Higgs boson decay into a pair of b quarks, which turns out to be the dominant Higgs boson decay mode (see e.g. Figure 5.2). Also the top quark decay involves a b quark leading to so-called b quark initiated jets. Tagging b quark initiated jets exploits the fact that b hadrons inside these jets have a lifetime in the order of 10^{-12} sec and hence can decay a few millimeters away from the PV. An inclusive vertex finding algorithm [136] is used to identify the position where the b hadron decays, called secondary vertex (SV).

The mass difference between b quarks and light quarks or gluons on one hand and the typically larger mass difference between b hadrons and their decay products on the other hand, lead to broader energy fluxes within the cone of a b quark initiated jet and large charged-track multiplicities within this cone. The probability to find an electron or muon inside a b quark initiated jet is around 20% and can also serve as an indicator.

decay mode	intermediate resonance	BR [%]
$\tau^- \to \nu_\tau \, e^- \, \overline{\nu}_e$		17.8
$\tau^- \to \nu_\tau \, \mu^- \, \overline{\nu}_\mu$		17.4
$\tau^- \to \nu_\tau \, h^-$		11.5
$\tau^- \to \nu_\tau \, h^- \, \pi^0$	$\rho(770)$	25.9
$\tau^- \to \nu_\tau \, h^- \, \pi^0 \, \pi^0$	$a_1(1260)$	9.5
$\tau^- \to \nu_\tau \, h^- \, h^+ \, h^-$	$a_1(1260)$	9.8
$\tau^- \to \nu_\tau \, h^- \, h^+ \, h^- \, \pi^0$		4.3
Other hadronic tau decays		3.3

Table 3.2: Possible tau decay modes are listed. Charge conjugation is implicitly assumed. The symbol $h^\pm$ denotes charged pions ($\pi^\pm$) or Kaons ($K^\pm$). The table is adopted from [140].

Two b-taggers are currently supported for Run2 analyses in CMS, DeepCSV [137] and DeepJet [138]. Both use multivariate analysis techniques based on deep neural networks to discriminate b quark initiated jets from light-quark or gluon-initiated jets. Three conditions are defined – `loose`, `medium` and `tight` – corresponding to a misidentification rate of 10%, 1% and 0.1%, respectively. For the same misidentification rate the DeepJet tagger achieves a higher efficiency in selecting genuine b quark initiated jets than the DeepCSV tagger.

3.5.8 Reconstruction and Identification of Tau Leptons

As introduced in Section 2.1 (see also Figure 2.2), the tau lepton is the heaviest of all leptons, allowing it to decay leptonically or hadronically. It is the only lepton with hadronic decay channels. The different decay channels of the tau lepton together with their corresponding decay probability are summarized in Table 3.2. The decay probability is also called branching ratio (BR). As can be seen from this table, in one third of cases the tau lepton decays leptonically into the lighter muon or electron. These muons and electrons are reconstructed using the techniques discussed earlier in this chapter. The rest of this section deals with the decay of hadronically decaying tau leptons that will be denoted as τ_h from now on. Hadronic tau decays involve charged hadrons, $h^\pm$, referred to as *prongs*, which mostly consist of charged pions. Prongs are accompanied by neutral pions in some decay channels. The tau lepton has a lifetime of 2.9×10^{-13} s [49] and decays before reaching the tracker. Hence, tau decays can only be reconstructed from their decay products but the track displacement from the PV is too small to be distinguished from prompt electrons, muons or charged particles. Since every decay involves at least one neutrino which leaves the detector undetected, it is not possible to calculate the momentum of the tau before its decay. Dedicated algorithms are employed to calculate for example the invariant mass of a di-tau system [139]. In the following the reconstruction and identification procedures of τ_h's are discussed in more detail.

The reconstruction of τ_h is based on the hadron-plus-strip (HPS) algorithm [141]. Charged hadrons with transverse momentum above 0.5 GeV consistent with originating from the PV and reconstructed by PF are considered. Neutral pions are present in more than 60% of τ_h decays and disintegrate instantly into a pair of photons. These photons convert with a high probability into electron-positron pairs inside the tracker. Due to the magnetic field, the electron-positron pair is separated in opposite ϕ-directions. In addition, both electrons and positrons can emit bremsstrahlung photons. To reconstruct the τ_h correctly, it is crucial to collect all corresponding energy deposits inside the ECAL.

These energy deposits typically lie within a narrow η- and wide ϕ-window and are therefore referred to as *strips*.

In Run2, the window size in $\Delta\eta \times \Delta\phi$ for strips is chosen dynamically [141] because:

- Remnants of nuclear interactions of charged pions from τ_{h} decays tend to have low transverse momentum and are missed in case a fixed sized $\Delta\eta \times \Delta\phi$ window is used for strip reconstruction. Furthermore, these remnants when not being part of the strip are interpreted as extra hadronic activity in the vicinity of the τ_{h} candidate, potentially leading to its rejection.
- Electron-positron pairs together with their potential bremsstrahlung photons can lie outside a fixed sized $\Delta\eta \times \Delta\phi$ window.
- In case of high transverse momenta of the τ_{h}, its decay products are more collimated and smaller strips can be used.

In the following, the dynamic strip reconstruction will be discussed. It starts from electron or photon candidates which will be denoted as e/γ in the following. The e/γ candidate with the highest transverse momentum, p_{T}, is chosen as strip seed and its pseudorapidity and azimuth are used as initial strip values

$$\begin{aligned} p_{\mathrm{T}}^{(\mathrm{e}/\gamma)} &\rightarrow p_{\mathrm{T}}^{(\mathrm{strip})} \\ \eta^{(\mathrm{e}/\gamma)} &\rightarrow \eta^{(\mathrm{strip})} \\ \phi^{(\mathrm{e}/\gamma)} &\rightarrow \phi^{(\mathrm{strip})} \ . \end{aligned} \tag{3.19}$$

Next, the e/γ candidate with the second-highest transverse momentum is added to strip if it lies within

$$\begin{aligned} \Delta\eta &= f\left(p_{\mathrm{T}}^{(\mathrm{e}/\gamma)}\right) + f\left(p_{\mathrm{T}}^{(\mathrm{strip})}\right) \ \text{and} \\ \Delta\phi &= g\left(p_{\mathrm{T}}^{(\mathrm{e}/\gamma)}\right) + g\left(p_{\mathrm{T}}^{(\mathrm{strip})}\right) \ . \end{aligned} \tag{3.20}$$

The functions f and g in Equation (3.20) are determined from simulated tau leptons and are required to collect 95% of e/γ arising from a τ_{h} decay inside the correct strip [141]. The exact functional form is given by

$$\begin{aligned} f\left(p_{\mathrm{T}}\right) &= 0.20 \cdot p_{\mathrm{T}}^{-0.66} \ \text{and} \\ g\left(p_{\mathrm{T}}\right) &= 0.35 \cdot p_{\mathrm{T}}^{-0.71} \ . \end{aligned} \tag{3.21}$$

However, the dynamic strip window defined by Equation (3.20) is bound on both ends by

$$\begin{aligned} \Delta\eta &\in [0.05, 0.15] \ \text{and} \\ \Delta\phi &\in [0.05, 0.30] \ . \end{aligned} \tag{3.22}$$

Every time an e/γ candidate is found within the $\Delta\eta \times \Delta\phi$ window, the strip coordinates get updated according to

$$\begin{aligned} p_{\mathrm{T}}^{(\mathrm{strip})} &= \sum_{\mathrm{e}/\gamma} p_{\mathrm{T}}^{(\mathrm{e}/\gamma)} \\ \eta^{(\mathrm{strip})} &= \frac{1}{p_{\mathrm{T}}^{(\mathrm{strip})}} \sum_{\mathrm{e}/\gamma} \eta^{(\mathrm{e}/\gamma)} \cdot p_{\mathrm{T}}^{(\mathrm{e}/\gamma)} \\ \phi^{(\mathrm{strip})} &= \frac{1}{p_{\mathrm{T}}^{(\mathrm{strip})}} \sum_{\mathrm{e}/\gamma} \phi^{(\mathrm{e}/\gamma)} \cdot p_{\mathrm{T}}^{(\mathrm{e}/\gamma)} \ . \end{aligned} \tag{3.23}$$

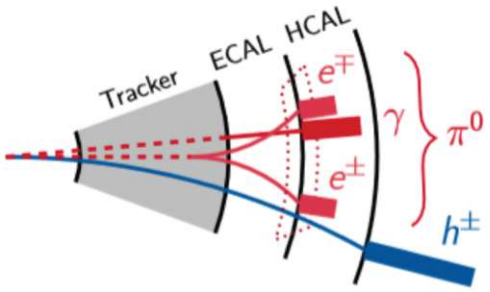


Figure 3.19: An example of a 1-prong+π^0 τ_h decay is schematically depicted. The charged hadron $h^{\pm}$ (prong) is shown in blue and leaves a track as well as energy deposits inside the HCAL. The π^0 is shown in red and decays into a pair of photons where one of the photons further decays in a electron-positron pair. All the decay products of the neutral pion are contained in a strip indicated by a red-dotted line. The figure is adopted from [142]

The strip reconstruction stops if no e/γ candidates are left inside the given $\Delta\eta \times \Delta\phi$ window. All strips with transverse momentum above 2.5 GeV are considered as π^0 candidate.

The procedure of building τ_h candidates continues by testing the compatibility of charged hadrons, i.e. prongs, with neutral pion candidates, i.e. strips, under different decay mode hypotheses. These hypotheses reflect the different decay modes listed in Table 3.2 with the addition of 2-prong decay modes to recover 3-prong τ_h candidates, where one charged-track was not reconstructed[9]. In total, the following decay mode hypotheses are tested:

1-prong : consisting of one charged hadron and no associated strips.

1-prong +π^0 : consisting of one charged hadron and one associated strip.

1-prong + 2π^0 : consisting of one charged hadron and two associated strips.

2-prongs : consisting of two charged hadrons and no associated strips.

2-prongs+π^0 : consisting of two charged hadrons and one associated strip.

3-prongs : consisting of three charged hadrons and no associated strips.

3-prongs+π^0 : consisting of three charged hadrons and one associated strip.

The mass window for each hypothesis is tailored to keep the τ_h reconstruction efficiency high and the misidentification rate from quark- or gluon-initiated jets low. Using the number of prongs ($N_{h^{\pm}}$) and the number of strips (N_{π^0}), it is useful to index the possible τ_h decay modes as

$$D_{\text{mode}} = 5 \cdot (N_{h^{\pm}} - 1) + N_{\pi^0} \ . \tag{3.24}$$

To further increase the reconstruction efficiency of genuine τ_h, it is required that the total charge adds up to ± 1 for all 3-prong decays. In addition, all prongs, as well as the strip coordinates are required to lie within a *signal cone* as illustrated in Figure 3.20 with radius parameter

$$R_{\text{sig}} = \begin{cases} 0.05 & p_T^{(\tau_h)} > 60\,\text{GeV} \\ \frac{3.0}{p_T^{(\tau_h)}[\text{GeV}]} & 30\,\text{GeV} < p_T^{(\tau_h)} < 60\,\text{GeV} \\ 0.1 & p_T^{(\tau_h)} < 30\,\text{GeV} \end{cases} \ . \tag{3.25}$$

[9] This can for example happen, when the τ_h is boosted and the different charged-particle tracks can not be resolved.

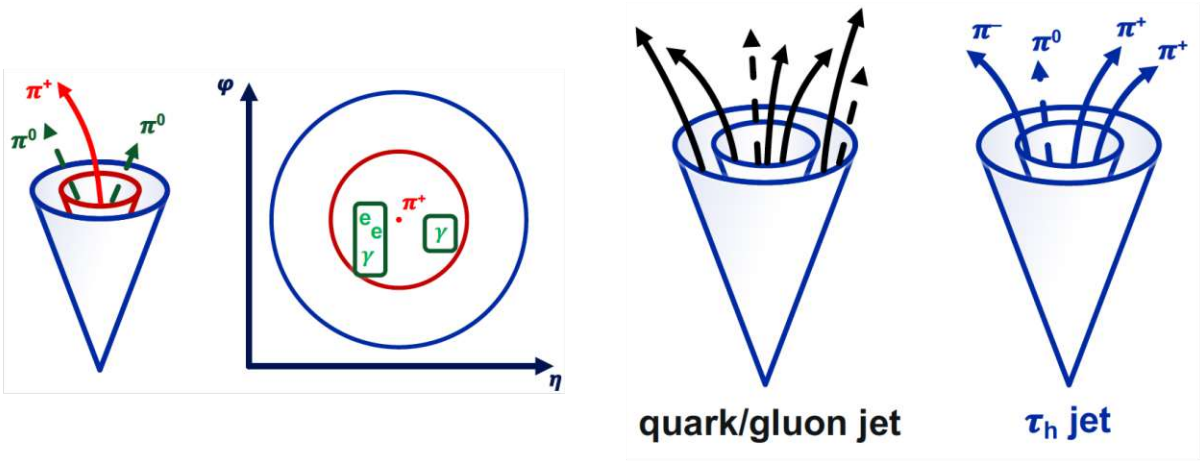


Figure 3.20: Left: The signal cone is shown in red and the isolation cone in blue. Within the signal cone, one charged pion and two neutral pions reside leading a positive identification as τ_h candidate with $D_{\text{mode}} = 2$. Right: The difference between a τ_h and quark- or gluon-initiated jets is illustrated. Quark- or gluon initiated jets are frequently misidentified as τ_h. However, they tend to have more particles between signal (inner) and isolation (outer) cone. The illustrations are taken from [143].

In case several decay mode hypotheses pass, τ_h candidates with more prongs are favored over those with higher transverse momenta. If it is still ambiguous, the decay mode with the most neutral constituents is assigned to the τ_h candidate.

The τ_h candidates determined as described above are still heavily contaminated by misidentified objects. These objects mainly consist of quark- or gluon-initiated jets, but also electrons and muons can be mis-reconstructed as τ_h candidates. In order to suppress misidentified τ_h, cone isolation sums were used and ever more evolved approaches were developed within the CMS experiment. The DeepTau classifier [144] is the current state-of-the-art τ_h discriminant based on a deep neural network. The neural network processes the information about energy deposits and tracks in terms of rectangular $\eta \times \phi$ grids containing the signal and isolation cone of the τ_h candidate. As illustrated in Figure 3.20 (right-hand side), quark- or gluon-initiated jets have typically much more activity in the region between signal and isolation cone. The $\eta \times \phi$ grids are pre-processed using convolutional layers, a well established approach in image recognition tasks. The output is combined with so-called high-level variables which include the τ_h decay mode, life time information, isolation and calorimeter energy fractions, and fed into a fully-connected feed-forward neural network. Four output nodes define the output classes the neural net is trained to distinguish, namely genuine τ_h's, quark- or gluon-initiated jets, electrons and muons misidentified as τ_h's. Figure 3.21 shows the performance of the DeepTau classifier in terms of efficiency and misidentification rate of τ_h candidates together with a comparison to previously used classifiers. For the same efficiency of selecting real τ_h, the DeepTau classifier has a reduced misidentification rate compared to previous classifiers, i.e. a better rejection power over misidentified τ_h's. The misidentification rate against jets is lower by factor two and lower up to a factor ten for the discrimination against electrons or muons. The DeepTau classifier's output value against jets, D_{jet}, is a key ingredient to the method used to estimate contributions from misidentified τ_h as discussed in Section 8.2. For the D_{jet} classifier (see top row in Figure 3.21), there are eight thresholds defined, from low τ_h identification efficiency to high τ_h identification efficiency. As the τ_h identification efficiency increases, also jet $\rightarrow \tau_h$ misidentification probability gets larger. The thresholds from lowest to highest τ_h identification efficiency are labeled as `vvvloose`,

vvloose, vloose, loose, medium, tight, vtight, vvtight.

3.5.9 Missing Transverse Energy

Due to the negligible transverse momentum of the colliding protons, the momentum imbalance in the transverse plane of the produced particles is used to infer the presence of undetected particles in a collision event. The *missing transverse momentum* denoted as $\boldsymbol{p}_{\mathrm{T}}^{(\mathrm{miss})}$ is defined as negative vectorial sum of all reconstructed PF candidates

$$\boldsymbol{p}_{\mathrm{T}}^{(\mathrm{miss})} = -\sum_{\mathrm{PF}} \mathbf{p}_{\mathrm{T}}^{(\mathrm{PF})} \ . \tag{3.26}$$

The module of $\boldsymbol{p}_{\mathrm{T}}^{(\mathrm{miss})}$ is referred to as missing transverse energy (MET). The accurate value of MET strongly depends on the CMS detector's capability to detect all particles and to reconstruct them correctly. Therefore, detector effects like energy resolution, alignment of detector elements as well as defects or blind regions of the detector impact the MET variable. Neutrinos are the only SM particles that can escape undetected. Furthermore, any new weakly interacting particles produced in proton-proton collisions will impact the MET and hence, this variable is crucial in searches for new physics beyond the SM.

The definition of MET as shown in Equation (3.26) uses PF jets. The effect of pileup is reduced by the method of charged-hadron-subtraction (see Section 3.5.7) as well as applying jet energy corrections to the MET variable. Another method developed by the CMS experiment is the pile-up-per-particle-identification (PUPPI) approach [145, 146]. The idea behind PUPPI, is to use the full collection of PF candidates and assign weights to them, reflecting the probability that they originate from the PV. Within the tracker acceptance ($|\eta| < 2.5$), it is possible to assign charged-particle tracks to the PV. All such PF candidates get assigned a weight of one, whereas all other candidates not associated to the PV get a weight of zero. Charged-particle candidates not associated to any vertex are assigned a weight of one if they are within 0.3 cm in longitudinal direction of the PV and zero otherwise. For all neutral hadrons, a weight is calculated based on the sum of transverse momenta of particles inside a cone of radius 0.4 around the neutral hadron. The more PF candidates associated with the PV reside inside this cone, the higher the probability the neutral hadron is coming from the PV and the higher the weight assigned to this neutral hadron. The missing transverse momentum using the PUPPI approach is given by

$$\boldsymbol{p}_{\mathrm{T,PUPPI}}^{(\mathrm{miss})} = -\sum_{\mathrm{PF}} \omega_{\mathrm{PUPPI}}^{(\mathrm{PF})} \cdot \mathbf{p}_{\mathrm{T}}^{(\mathrm{PF})} \ , \tag{3.27}$$

where $\omega_{\mathrm{PUPPI}}^{(\mathrm{PF})}$ denotes the assigned PUPPI-weights.

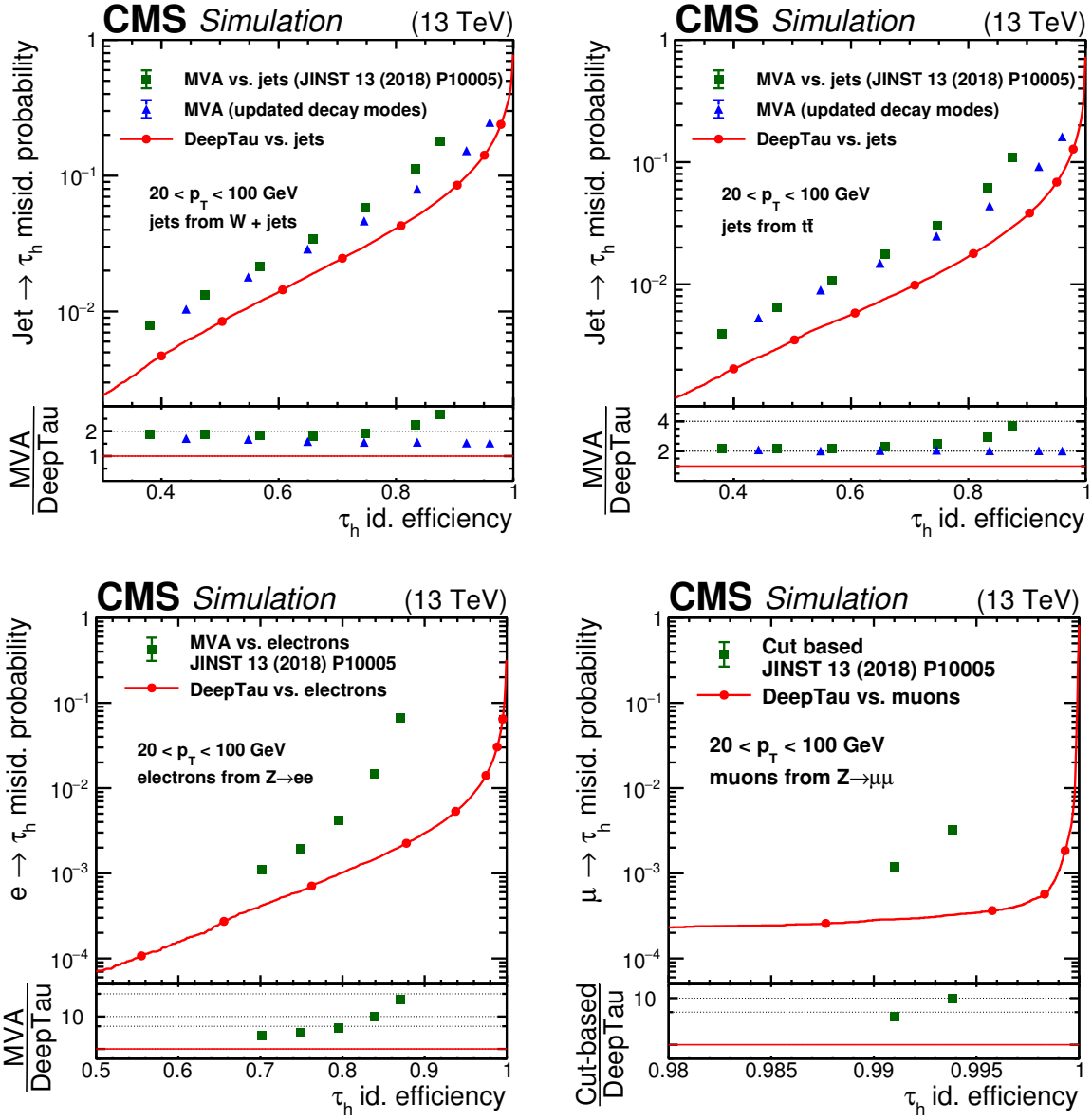


Figure 3.21: The performance of the DeepTau classifier is shown and compared to previously available discriminants. On the top, quark- or gluon-initiated jet misidentification rate versus the efficiency of identifying genuine τ_h candidates is shown. The jets on in the top left come from the W+jets process, while the jets shown in the top right plot come from $t\bar{t}$ processes. The bottom left plot shows the same for the electron misidentification and the bottom right plot for muon misidentification. In all cases the τ_h transverse momentum is below 100 GeV. The plots are taken from [144].

Chapter 4

Silicon Strip Backplane Correction

In this chapter, a track-based alignment technique, commonly referred to as the *backplane* correction, is presented. It is a correction applied to sensors in the silicon strip detector, which is presented in Section 3.2.1. The task of measuring the backplane correction included the development of a new analysis framework and was carried out in the CMS silicon strip detector project.

When a charged particle passes through a silicon strip sensor, it produces ionization. A red arrow representing a charged-particle trajectory, commonly referred to as the track, is shown in Figure 4.1 together with a schematic sketch of a silicon strip sensor. The induced charges drift to the strip plane due to the electric field applied to the sensor. They hit several silicon strips, each producing an electric signal that is read out. An ensemble of activated strips is called a *cluster*. Ideally, all charges get collected over time. A local Cartesian coordinate system (u, v, w), as illustrated in purple in Figure 4.1, is defined on each sensor. The origin of this coordinate system is located at geometric center of the sensor. The local w-direction is always oriented along the electric field applied to the sensor, i.e. perpendicular to the surface of the sensor. It follows from this definition that the silicon strips are located in the positive w-half-space, whereas the backplane is in the negative w-half-space. The local v-axis is defined parallel to the silicon strips, whereby the positive direction is defined away from the sensor's readout. The local u-axis is perpendicular to v and w, such that (u, v, w) forms a right-handed coordinate system. For sensors in the barrel region – i.e. TIB and TOB – the correspondence between local and global coordinates is as follows. The local u-direction corresponds to the global $r\phi$-direction, local v corresponds to global z and local w to global r coordinates.

Projecting the cluster's barycenter on the mid-plane of the sensor defines the reconstructed position of the particle's traversal (hit). Charge collection and reconstructed hit (rechit) position are illustrated in Figure 4.2. Inferring the charge collected from the peak position of the signal build-up is one possible readout mode the silicon sensors can be operated in. It is called PEAK mode. The signal build-up is schematically drawn on the left of Figure 4.3. However, this mode of operation is not feasible when the CMS detector records proton-proton collisions with a bunch spacing of 25 ns because the drift time from the backplane to the strips exceeds this time window. Therefore, the silicon sensors are operated in a second mode, called deconvolution (DECO) mode. As depicted on the right side of Figure 4.3, the signal pulse is sub-sampled three times to reduce the signal build-up to approximately 25 ns, which corresponds to the bunch spacing at the LHC.

In 2009, when for the first time cosmic ray data were recorded in both PEAK and DECO modes, a bias in the local w-direction between the two modes was observed. Due to the short time interval used for charge integration in DECO mode, not all the induced

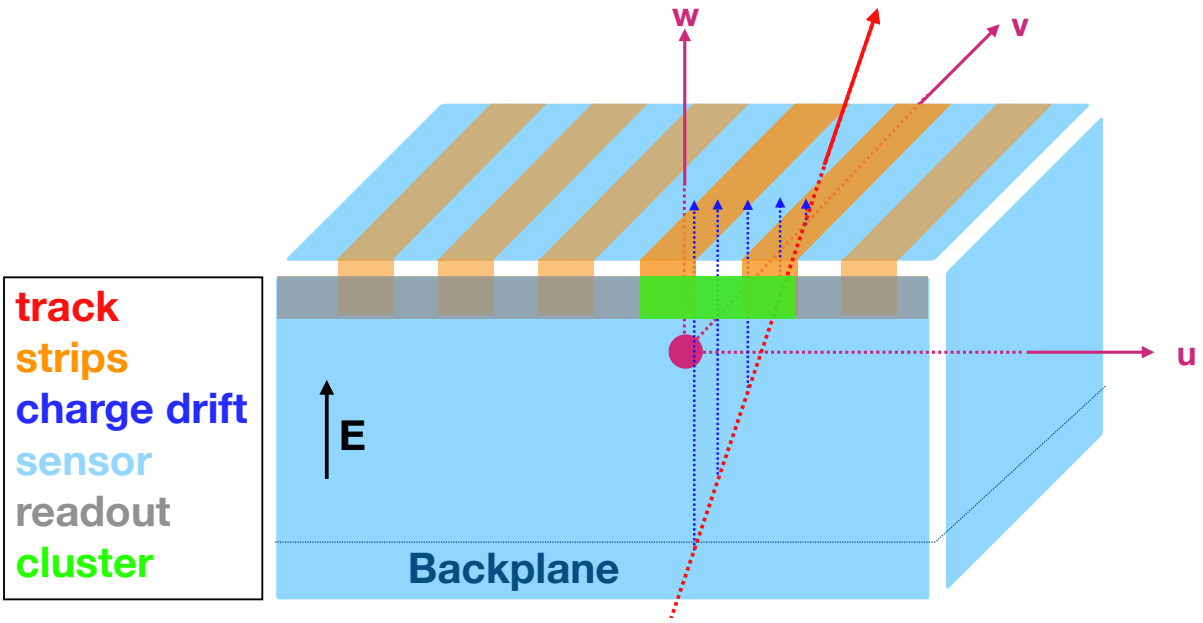


Figure 4.1: A simplified sketch of a silicon strip sensor is shown. A charged particle (red) passes through the sensor and produces ionization. The induced charges drift upward because an electric field, E, is applied as shown in the figure. The silicon strips collect the charges and produce a detectable signal that can be read out. In green, the resulting cluster of activated strips from the passing particle is shown. When measuring quantities related to the sensor, a dedicated coordinate system is used referred to as *local coordinates* u, v and w. The local v direction is parallel to the global z direction, i.e. to the beam pipe (see Figure 3.5).

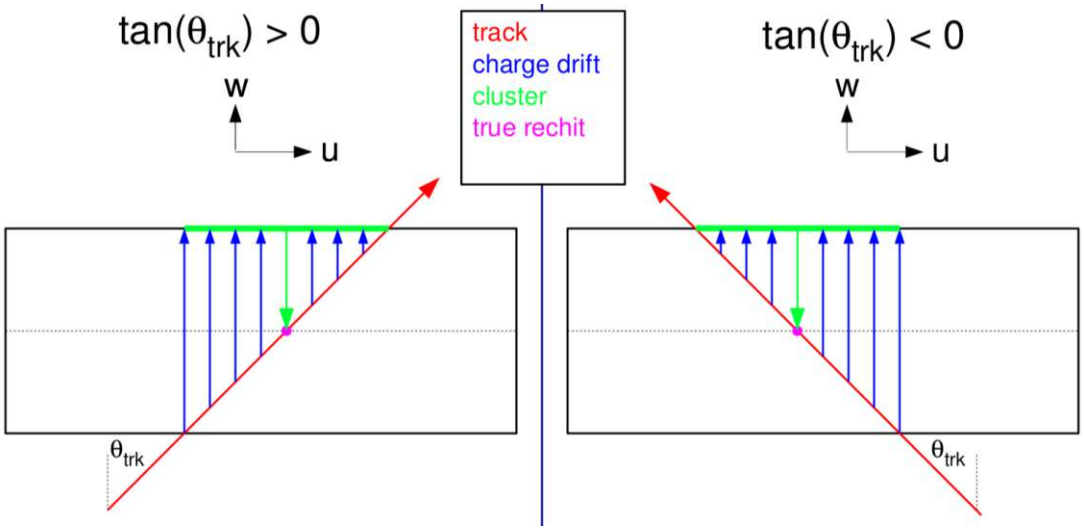


Figure 4.2: Two sensors projected on the local u-w plane are shown. The track of a passing charged particle is shown in red. Induced charges produced by a passing charged particle drift upwards along the w-axis. No magnetic field is applied and all the induced charges are collected. All strips that detect a signal form a cluster. The projection of the barycenter of this cluster onto the sensor's mid-plane matches the true passing point of the track - labeled as true rechit in the figure. Left and right distinguish the two cases whether the sign of the incident track angle (θ_{trk}) is positive or negative. This distinction becomes relevant in the case, when a magnetic field is present (see Figure 4.5). The figure is taken from [147].

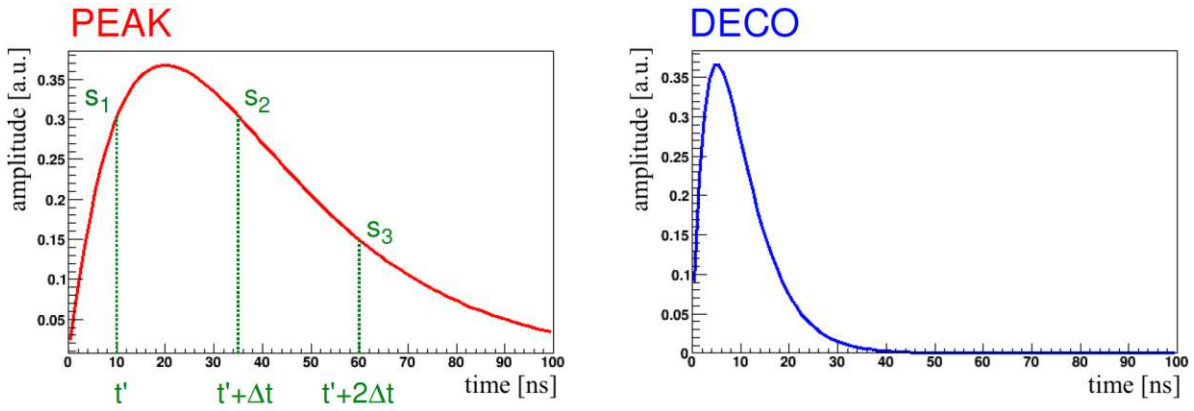


Figure 4.3: Shown are the signal pulses for the PEAK (left) and DECO (right) read-out mode, which are obtained by collecting all signals of silicon strips inside a single sensor. The y-axis measures the signal amplitude in arbitrary units. The readout time can be confined within 25 ns in DECO mode as needed when the LHC is operated with standard bunch crossings. The amplitude in DECO mode is a weighted sum of three consecutive samplings, s_i, of the PEAK amplitude with a time window of $\Delta t = 25$ ns [148]. The figure is taken from [147].

charges from the sensor's backplane are collected because the drift time exceeds the time window for charge integration. In the following, the backplane correction measurement is presented using the so-called Venturi model, which explains the observed bias in local w-direction.

Let's first ignore the magnetic field such that induced charges drift parallel to the local w-direction as shown in Figure 4.2. In DECO mode, not all charges are collected from the sensor's backplane, and the shift in local u-direction (Δu) and the shift in local w-direction (Δw) are linked via the formula

$$\Delta u = \Delta w \cdot \tan\theta_{\text{trk}} \ , \tag{4.1}$$

where θ_{trk} is the incident track angle. This situation is illustrated in Figure 4.4.

In the presence of a magnetic field, the deflection of drifting charges has to be taken into account. The deflection angle is called *Lorentz angle* and is denoted as θ_{LA}. The situation with magnetic field is depicted in Figure 4.5a. Taking the signs of the angles into account, Equation (4.1) becomes

$$\Delta u = \Delta w \cdot (\tan\theta_{\text{trk}} - \tan\theta_{\text{LA}}) \pm H_{1/2} \cdot \Delta \tan\theta_{\text{LA}} \ , \tag{4.2}$$

where half of the sensor's thickness is denoted by $H_{1/2}$. The term, $H_{1/2} \cdot \tan\theta_{\text{LA}}$, in Equation (4.2)) corrects for the fact that in DECO mode the Lorentz angle appears smaller than if all the charge would be collected. This effect is graphically illustrated in Figure 4.5b. The sign of this correction depends on whether the sensor has the local v-coordinate parallel or anti-parallel to the global z-direction. It turns out that it is a small correction and only relevant for the TOB.

Equation (4.2) describes a linear function with Δw being the slope and $\pm H_{1/2} \cdot \Delta \tan\theta_{\text{LA}}$ being the offset. Hence, the analysis strategy is to perform a straight line fit in a measurement of Δu as a function of $(\tan\theta_{\text{trk}} - \tan\theta_{\text{LA}})$. The shift Δu is calculated as difference between true and reconstructed rechit position in local u-direction (see Figure 4.5a). The measurement is done twice, once for data recorded in PEAK mode and once for data recorded in DECO mode. Finally, the backplane correction is given by the difference of the extracted value of Δw in DECO and PEAK mode. The current backplane correction values are shown in Table 4.1.

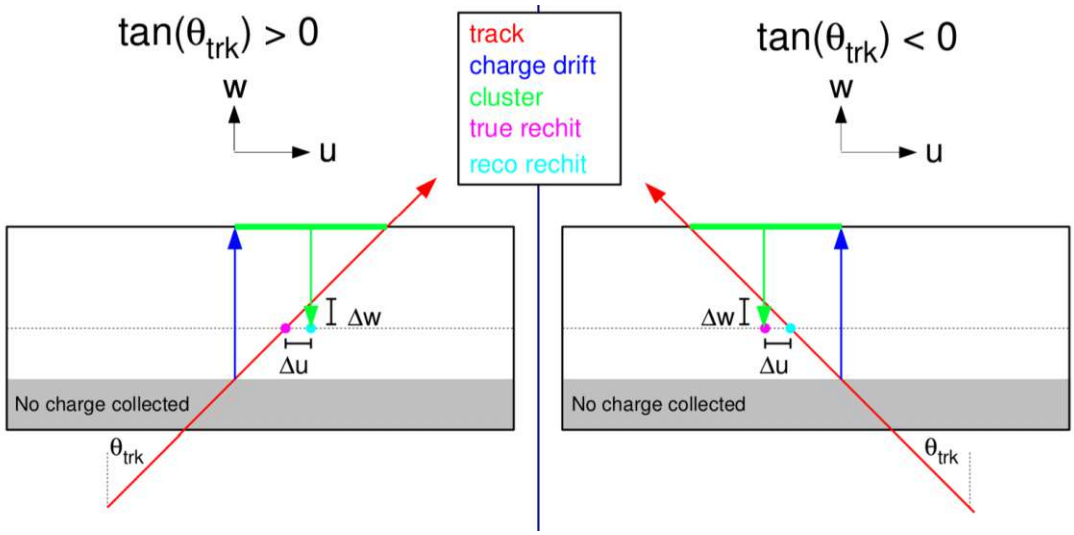
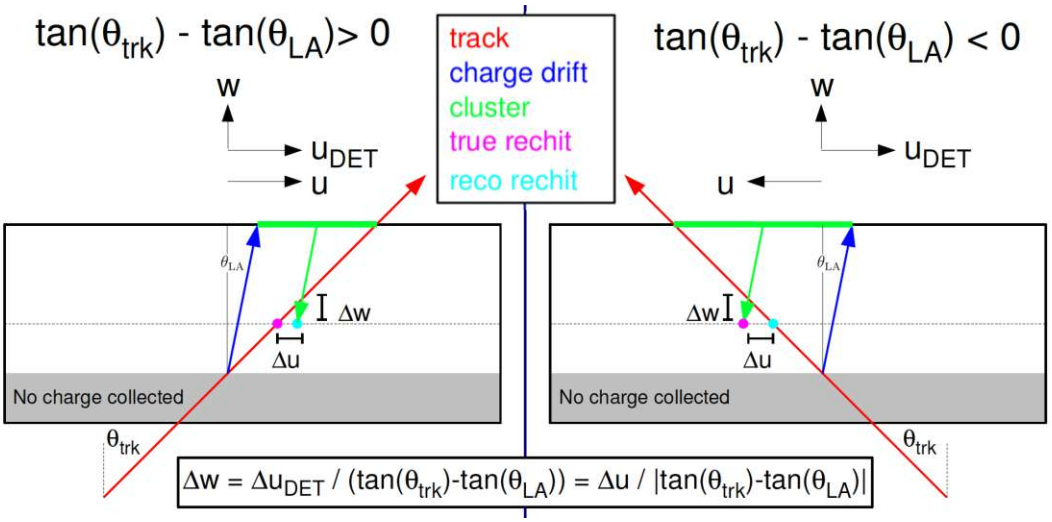


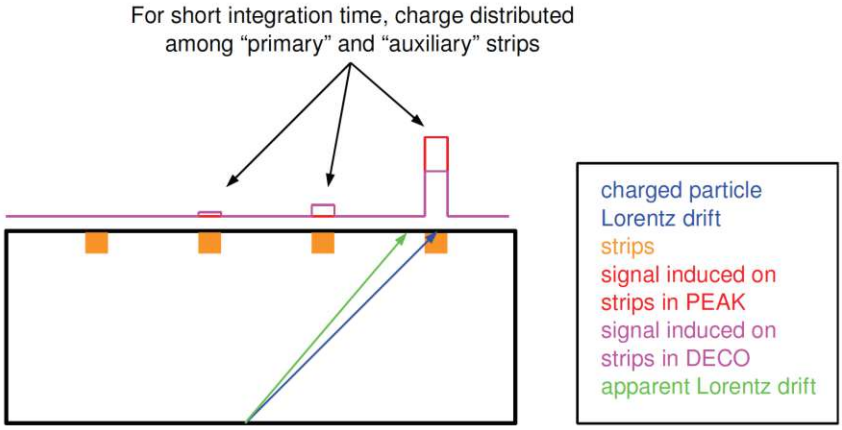
Figure 4.4: This figure illustrates what happens in DECO mode, where not all charges reach the silicon strips. The cluster size is smaller compared to Figure 4.2 and the extrapolation of the cluster barycenter does not match the true rechit position. The shifts in local u-direction and local w-direction of the mismatch rechit positions are denoted with Δu and Δw, respectively. The shifts of both directions can be related via the track angle according to the formula given in Equation (4.1). The figure is taken from [147].

correction	Δw		$\Delta \tan\theta_{\mathrm{LA}}$	
	TIB	TOB	TIB	TOB
value	0 µm	6 µm	0	−0.006

Table 4.1: Shown are the current backplane correction values. They are separately measured in TIB and TOB. As can be seen from the table, there is no correction currently needed for the TIB. Correction values are obtained from analyzing data collected in 2011 by the CMS detector. For data recorded in PEAK mode, cosmic ray data are used, whereas for the DECO mode, collision data are used.



(a) In contrast to Figure 4.4, here the charges do not drift parallel to the local w-direction, but are deflected due to the Lorentz force, which is induced by the presence of the magnetic field. The magnetic field is not explicitly drawn but points inside the plane. Left and right side of the figure distinguish between the different cases of positive and negative track and Lorentz angles and their impact on the shifts in local u- and w-direction. The figure is taken from [149].



(b) In case of the DECO readout mode, the apparent Lorentz angle is smaller. This happens because in DECO mode the charge is distributed and read out from different strips. The effect is expected to be larger in thicker sensors inside the TOB. The figure is taken from [147].

Figure 4.5: These figures illustrate the case, where sensors are operated in DECO mode, in presence of a magnetic field.

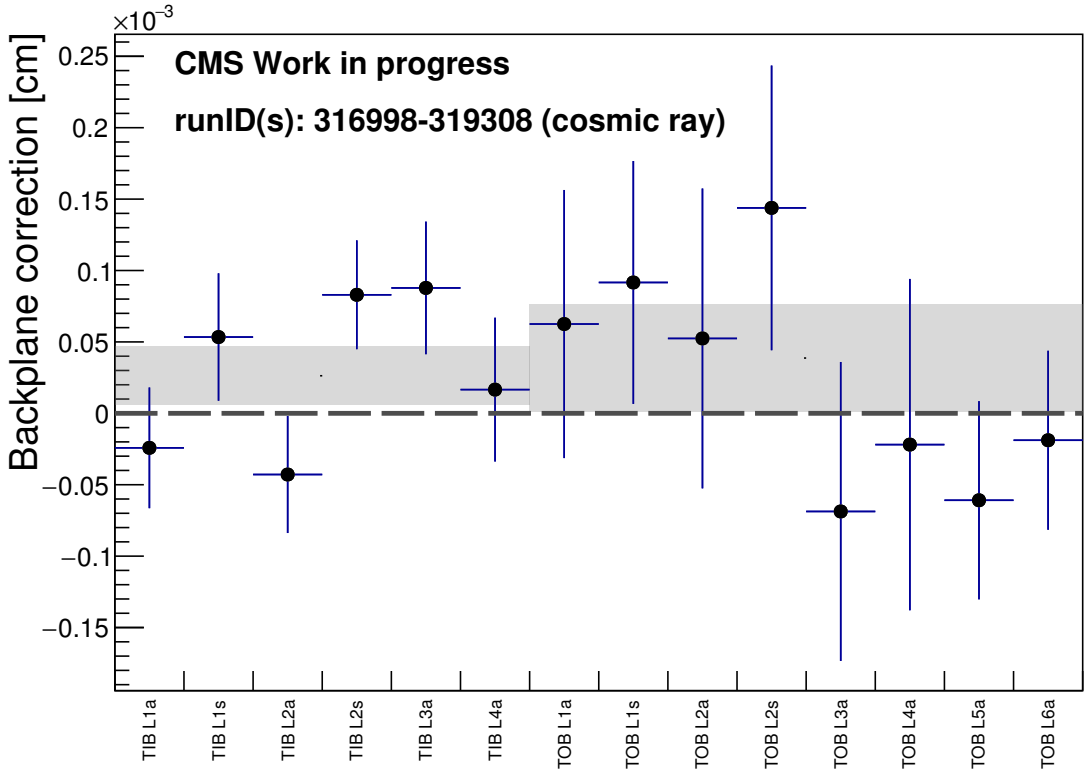


Figure 4.6: The backplane correction, $\Delta w_{\mathrm{DECO}} - \Delta w_{\mathrm{PEAK}}$, is shown per partition inside TIB and TOB. Cosmic ray data, collected in 2018, are used for the backplane measurement. The layers are abbreviated with the letter "L". The letters "s" and "a" stand for stereo and analog sensor modules, respectively. Stereo modules reconstruct 3-dimensional hit positions (see also Figure 3.9). The gray band represents the uncertainty when measurements of different layers and module types are combined for the TIB or the TOB. All uncertainties shown are purely statistical and represent one standard deviation. The dashed line is set at zero and underlines the compatibility of the individual measurements with zero.

In practice, the backplane correction values are stable over time and changes were only needed when a new timing of the silicon strip sensors was introduced. Therefore, the backplane correction values are monitored over the different data-taking periods and values consistent with zero mean no need of adapting the current values. An example of such a monitoring plot from the 2018 cosmic ray data-taking period is shown in Figure 4.6. The results are consistent with zero. While usually there is plenty of data recorded in DECO mode, data recorded in PEAK mode are scarce. In case of the plot shown in Figure 4.6, the available sample size of recorded events in PEAK mode is the limiting factor in reaching smaller uncertainties on the measured backplane correction values.

For the TID and TEC, no dedicated backplane measurement is performed, as there the correction translates in a global shift along the z-direction which is determined via other alignment procedures. For the preparation of Run 3 (2022-2025) of the LHC, starting from autumn 2021, a large volume of cosmic ray data is collected including larger amounts of data recorded in PEAK mode.

Chapter 5

Experimental Aspects of Measurements and Searches at the LHC

After introducing the SM and some of its extensions throughout Chapter 2 and presenting the CMS experiment in Chapter 3, in this chapter the focus is set on a discussion of how different signals are detected and analyzed by the CMS experiment. First, the phenomenology of the SM Higgs boson at the LHC is presented in Section 5.1. In addition, the presentation of precision measurements on the SM Higgs boson are discussed in the same section. The remaining sections discuss experimental searches for various signal processes arising from different BSM physics processes. These processes involve additional Higgs bosons in Section 5.2, leptoquarks in Section 5.3 and light top squarks in Section 5.4.

5.1 The SM Higgs Boson at the LHC

The fact that the Higgs boson mass (m_{H}) is a free parameter of the theory (see Equation (2.67)), made its search particularly challenging. The Higgs boson can decay into several different detectable final states, whereby their decay rates vary with m_{H} (see also Figure 5.2), introducing a further challenge in the search. Exclusion limits at 95% confidence level from the Large Electron-Positron Collider at CERN [150] suggest that the Higgs boson mass must be larger than 114.4 GeV. Similarly, the Tevatron Accelerator at Fermilab [151] excludes a Higgs boson with a mass in the range $147\,\mathrm{GeV} < m_{\mathrm{H}} < 180\,\mathrm{GeV}$. Despite these excluded masses,

$$\begin{aligned} m_{\mathrm{H}} &< 114.4\,\mathrm{GeV} \\ 147\,\mathrm{GeV} < m_{\mathrm{H}} &< 180\,\mathrm{GeV}\ , \end{aligned} \tag{5.1}$$

a broad range of possible masses was left. From unitarity considerations [152], an upper bound of approximately 1 TeV on m_{H} can be set. Furthermore, by combining several electroweak measurements [153], a light Higgs boson with $m_{\mathrm{H}} = 89^{+35}_{-26}$ GeV is favored by experimental data. This result assumes the SM to be the correct theory and excludes results from direct Higgs boson searches leading to the limits quoted in Equation (5.1). The outcome is not a proof that the SM Higgs boson exists, but has to be understood as a guideline as to where to look for it.

On July 4th, 2012, the ATLAS and CMS Collaborations at the LHC announced the discovery of a new boson with a mass of 125 GeV compatible with SM predictions [3, 4]. A year later, P. Higgs and F. Englert were awarded the Nobel prize in Physics *"for the theoretical discovery of a mechanism that contributes to our understanding of the origin of mass of subatomic particles, and which recently was confirmed through the discovery*

of the predicted fundamental particle, by the ATLAS and CMS experiments at CERN's Large Hadron Collider". Latest mass measurements achieve a per-mille precision [154]

$$m_{\mathrm{H}} = 125.38 \pm 0.14\,\mathrm{GeV}\ . \tag{5.2}$$

Moreover, it is confirmed that the discovered boson has zero electric charge, is a spin-0 particle and even under $\mathcal{CP}$ transformations [155, 156]. Together with the observation of different decay channels consistent with SM predictions, it is now established that the particle discovered in 2012 is very similar to the Higgs boson predicted by the SM.

5.1.1 Production and Decay Modes of the SM Higgs Boson

At the LHC, there are four main production modes of the SM Higgs boson:

1. The most dominant mode of Higgs boson production is via gluon-gluon fusion (ggF). For the SM Higgs boson, the ggF cross section is 48.6 pb [157]. The value of 48.6 pb is reached for $m_{\mathrm{H}} = 125.0\,\mathrm{GeV}$ and for a center-of-mass energy of $\sqrt{s} = 13\,\mathrm{TeV}$. In Figure 5.1e, the ggF cross section is shown in blue as a function of m_{H}. Calculations of this cross section incorporate the third next-to-leading order (N3LO) in perturbative QCD and the NLO in electroweak perturbation theory as indicated in blue in Figure 5.1e. A ggF Feynman diagram is shown in Figure 5.1a. Since the Higgs boson does not couple to massless gluons, the production via ggF has to include a loop. Contributions to the loop are from particles which have a strong coupling to the Higgs boson. In the SM, the loop contribution is dominated by the top quark. However, it should be mentioned that heavy particles predicted by BSM theories could have an influence too. The dominance of this mode stems from the large value of the gluon parton distribution function at low momentum fraction for the LHC centre-of-mass energy of $\sqrt{s} = 13\,\mathrm{TeV}$ (see also Section 3.4).

2. Higgs boson production via vector boson fusion (VBF) is the second most dominant production mode. It is shown in red in Figure 5.1e and has a cross section of 3.8 pb [157] for $m_{\mathrm{H}} = 125.0\,\mathrm{GeV}$ at $\sqrt{s} = 13\,\mathrm{TeV}$. Calculations of this cross section incorporate terms including the next-to-next-to-leading-order (NNLO) in perturbative QCD and the NLO in electroweak perturbation theory as indicated in red in Figure 5.1e. The Higgs boson production via VBF is about ten times less likely than via ggF. A Feynman diagram showing the VBF mode is shown in Figure 5.1b. The vector boson (V) can be either a W or Z boson. Experimentally, the VBF mode provides a very clean signature. The two outgoing quarks hadronize and form jets in the forward regions of the detector, having a large pseudorapidity separation. Due to this fact, the invariant di-jet mass (m_{jj}) takes typically large values.

3. The third most dominant Higgs boson production mode is in association with a W or Z boson ($\mathrm{V} \in \{\mathrm{W}, \mathrm{Z}\}$). This process is also referred to as Higgs-Strahlung (VH). For a Higgs boson mass of 125.0 GeV, the cross sections at $\sqrt{s} = 13\,\mathrm{TeV}$ are 1.4 pb [157] and 0.9 pb [157] for W and Z associated production, respectively. These cross sections are derived at NNLO in perturbative QCD and at NLO in electroweak perturbation theory as indicated in green and gray for the WH and ZH process in Figure 5.1e, respectively. Both production modes are shown in Figure 5.1e and a Feynman diagram is shown in Figure 5.1c. As shown in Figure 5.1c, a quark-antiquark pair annihilates and forms an intermediate vector boson, which radiates off a Higgs boson. Together with the VBF production mode, the VH process is exploited to gain knowledge about the coupling of the Higgs boson to vector bosons.

bosonic decay modes	BR [%]	fermionic decay modes	BR [%]
$H \to WW^*$	21.37	$H \to b\bar{b}$	58.24
$H \to gg$	8.19	$H \to \tau\tau$	6.27
$H \to ZZ^*$	2.62	$H \to c\bar{c}$	2.89
$H \to \gamma\gamma$	0.23	$H \to \mu\mu$	0.02
$H \to Z\gamma$	0.15		

Table 5.1: Tabulated values for the BRs of Higgs boson decays, possible to be observed at the LHC for a Higgs boson mass of $m_H = 125.0\,\text{GeV}$, are shown. The decay of $H \to \tau\tau$ studied in this thesis has a BR of roughly 6%. The numbers are taken from [160].

4. Higgs boson production in association with heavy quarks is the forth most dominant production mode. The top quark associated production (ttH) has a cross section of 0.50 pb [157] and the bottom quark associated production (bbH) has a cross section of 0.48 pb [157] for a Higgs boson mass of 125.0 GeV and at $\sqrt{s} = 13\,\text{TeV}$. Cross sections as a function of the Higgs boson mass are shown in Figure 5.1e for both ttH and bbH. As indicated in Figure 5.1e, the ttH cross section calculation includes terms up to NLO in perturbative QCD as well as NLO terms in electroweak perturbation theory. Terms up to NNLO in perturbative QCD in the five-flavor scheme (5FS) and terms including NLO in perturbative QCD in the four-flavor scheme (4FS) are included in the calculation of the bbH cross section[1]. Figure 5.1d shows a Feynman diagram of the heavy quark associated production. Furthermore, there is the possibility of single top quark associated production (tH). The cross section for this process is 0.07 pb [157] for $m_H = 125.0\,\text{GeV}$ at $\sqrt{s} = 13\,\text{TeV}$, including NLO terms in perturbative QCD. In all these processes, the Yukawa coupling to top and bottom quarks are directly tested.

The Higgs boson has a decay width of $3.2^{+2.8}_{-2.2}\,\text{MeV}$ [159] and hence a lifetime of approximately $2 \times 10^{-22}\,\text{sec}$. Due to this short lifetime, it is impossible for the Higgs boson to reach the CMS detector and must be identified via reconstruction of its decay products. Several different decay *modes* – also referred to as decay *channels* – can be observed at the LHC and are summarized in Figure 5.2. The respective BRs depend on the Higgs boson mass. Table 5.1 lists the BRs for a Higgs boson mass of 125.0 GeV.

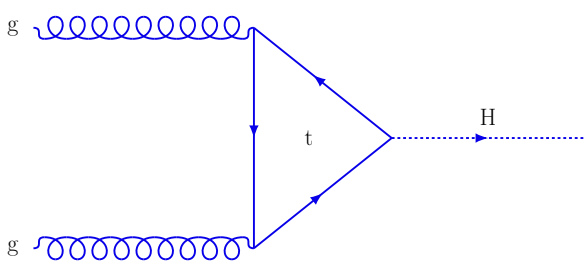
Higgs boson decay channels can be separated into bosonic and fermionic ones. Bosonic decay channels are:

$H \to WW^*$: The highest bosonic BR is taken by the decay of the Higgs boson with $m_H = 125\,\text{GeV}$ into a pair of W bosons. Each W boson decays either leptonically (⅓ of times) or hadronically (⅔ of times). Fully-hadronic final states are difficult to access experimentally since there is an overwhelming background from quark- or gluon-initiated jets originating from QCD multijet production (see also Section 6.2.2). Leptonic final states involving pairs of electrons or pairs of muons are dominated by Z boson decays (see also Section 6.2.4). Therefore, only a small fraction – about 3% – of the $H \to WW^*$ decay is experimentally accessible, namely

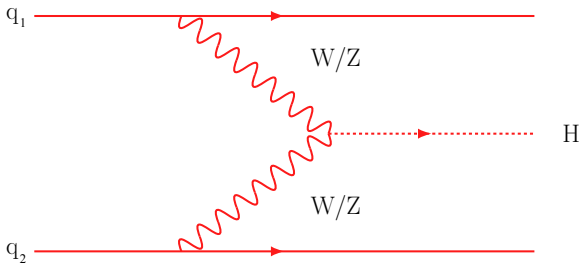
$$H \to WW^* \to e\nu\mu\nu\,. \tag{5.3}$$

Due to the presence of neutrinos in this final state, the invariant mass of the Higgs boson can not be precisely calculated and therefore the experimental sensitivity

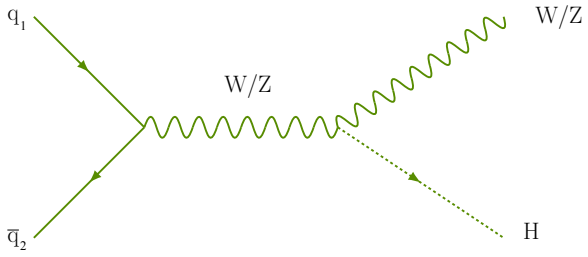
[1] For a discussion on the difference between 5FS and 4FS, see for example [158]. In short, b quarks appear only in the final state in the 4FS, whereas they are allowed also in the initial state in case of the 5FS.



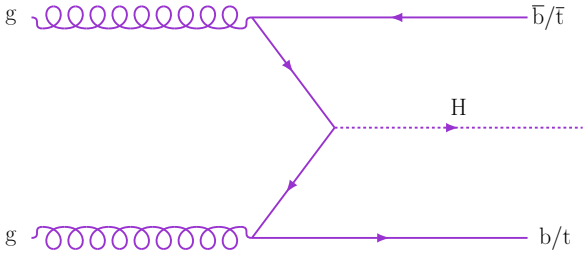
(a) gluon-gluon fusion



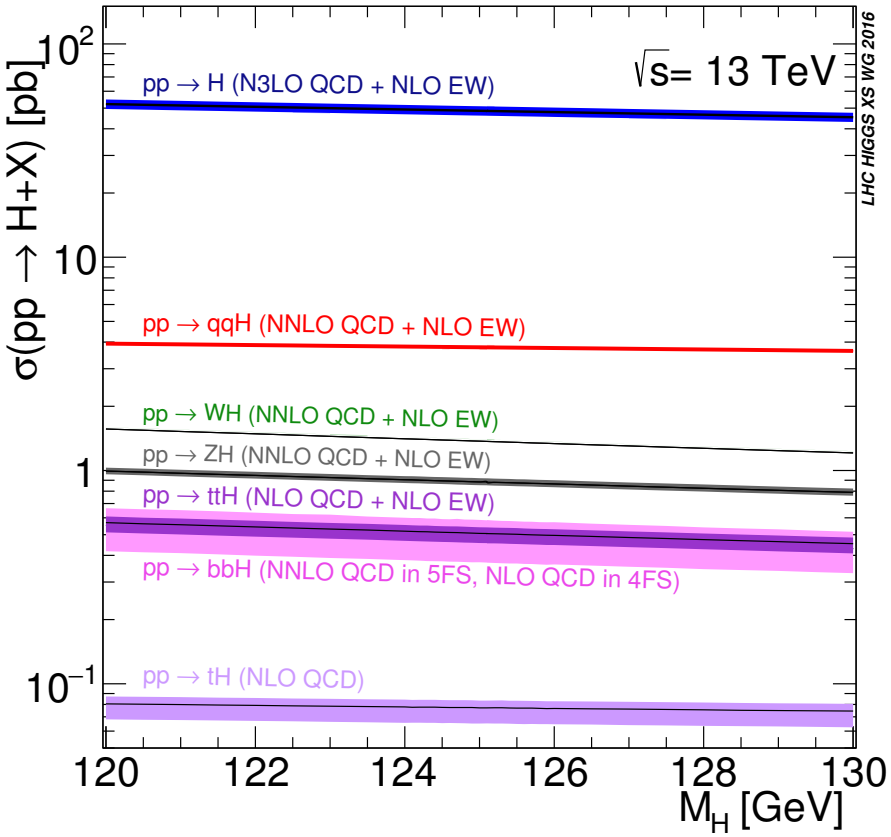
(b) vector boson fusion



(c) W or Z associated production



(d) Heavy quark (b or t) associated production



(e) SM Higgs boson production cross sections taken from [157]

Figure 5.1: Shown are different Higgs boson production modes. Figures (a)-(d) show Feynman diagrams of the different production modes – ggF, VBF, VH and ttH/bbH. In (e), cross sections of all production modes as a function of the Higgs boson mass are displayed. The center-of-mass energy is $\sqrt{s} = 13\,\text{TeV}$. The order in perturbative QCD and electroweak theory for each cross section is quoted in figure (e) for each production mode.

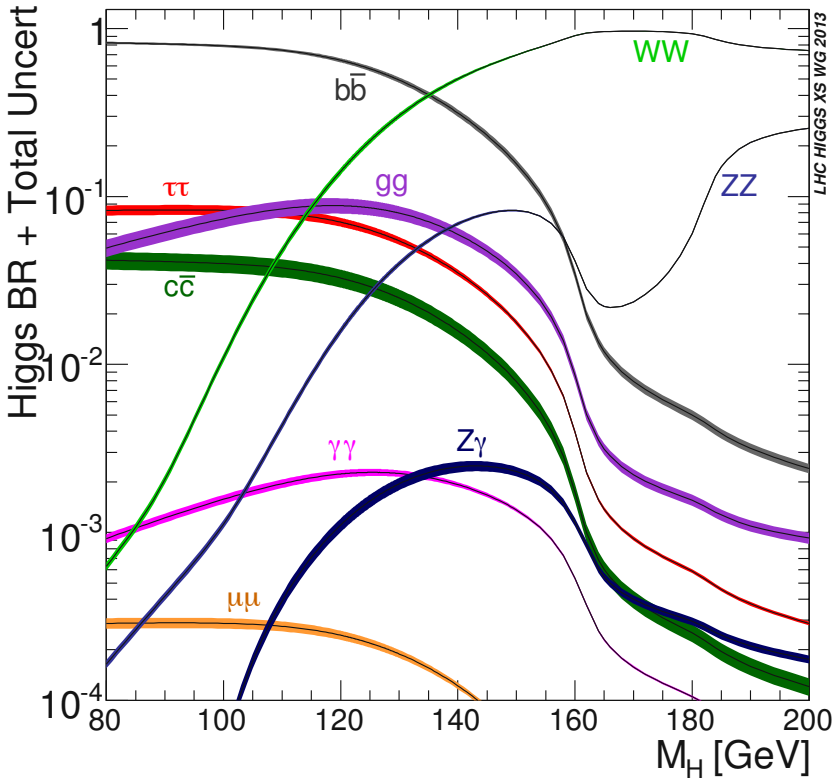


Figure 5.2: Shown are the different BRs of possible Higgs boson decays as a function of its mass. The decay of $H \to \tau\tau$ which is studied in this thesis is shown in red. The figure is taken from [160].

is smaller than for fully contained decays, like for example $H \to \gamma\gamma$ (see below). Both ATLAS and CMS experiments observed this decay and performed measurements which are compatible with SM expectations [161, 162, 163].

$H \to gg$: The Higgs boson does not couple directly to the massless gluons and thus this process involves a loop, which is dominated by heavy particles such as the top quark. The final state of this decay channel contains two gluon-initiated jets. It is extremely challenging to find such decays at the LHC because there are many more QCD multijet events (see also Section 6.2.2) produced which mimic this signal. To get a cleaner signal, the $H \to gg$ decay is searched in association with the VH production, using the leptons from the vector boson decay as indicators. In [164] a proof-of-concept of such an analysis is given. Even with the coming upgrade of the LHC, it will not be possible to observe this decay following the strategy laid out in [164]. However, by using modern machine learning techniques, experimental sensitivity can be gained as for example demonstrated in [165].

$H \to ZZ^*$: The subsequent decay of both Z bosons to electrons or muons – resulting in a four-lepton final state, is among the most sensitive decay channels. The reason why, is the rather low background contamination from other SM processes and the good resolution of the invariant mass of the four leptons. Hence, this decay channel is also referred to as the *golden* channel and it contributed to the Higgs boson discovery in 2012. The physics program of $H \to ZZ^* \to 4\ell$ of both ATLAS and CMS experiments has already advanced into a precision era, measuring differential cross sections in up to 19 kinematically different signal regions [166, 167]. All measurements are compatible with SM expectations.

$H \to \gamma\gamma$: The Higgs boson does not couple directly to the massless photon. Hence, the decay into a pair of photons requires a loop. Similarly to the ggF process (see Figure 5.1a) and the $H \to gg$ decay, this loop is dominated by heavy particles such as the top quark or the W boson[2]. Despite the low BR, a very high experimental sensitivity to this decay channel has been achieved. This is because the ECAL enables a very precise determination of the diphoton invariant mass. Together with the $H \to ZZ^* \to 4\ell$ channel, it was possible to announce the discovery of a Higgs-like boson in 2012. Also in this channel, further measurements have been conducted in up to 25 different signal regions [168, 169].

$H \to Z\gamma$: This decay has the lowest BR among the bosonic decay channels quoted in Table 5.1. It has not been observed yet. The ATLAS Collaboration however, achieved a significance of two standard deviations in a search for this decay [170].

The fermionic decay channels are the following:

$H \to b\bar{b}$: Among the fermionic decay channels, $H \to b\bar{b}$ has the largest BR. It can be used to study the Yukawa coupling of the Higgs boson to fermions. However, this decay channel suffers from large background contributions arising from QCD multijet production (see also Section 6.2.2). There are possibilities, using dedicated algorithms developed to identify b quark initiated jets to achieve an efficient selection of such events and a drastic reduction of background events. Experimental sensitivity is gained especially when searching for Higgs bosons produced via the VH process and target the leptonic decay products of the vector boson. In association with the VH production mode, this decay has been observed by both ATLAS and CMS Collaborations [171, 172, 173]

$H \to \tau\tau$: The decay channel involving a pair of tau leptons plays an important role in this thesis. Compared to $H \to b\bar{b}$, this decay channel benefits experimentally from leptonic tau decays. These leptons can be used to get a cleaner selection. However, the fact that tau leptons decay before reaching the detector and their decay includes neutrinos escaping detection, pose some challenges. The best sensitivity in this channel is obtained for ggF and VBF production modes. It has been observed by the ATLAS and CMS Collaborations [174, 175]. In Section 9.2.4, a differential measurement in twelve signal regions is presented for the $H \to \tau\tau$ decay channel. More details to how such precision measurements are presented are given in Section 5.1.2.

$H \to c\bar{c}$: This decay channel is important to test the Higgs boson coupling to second generation fermions. Experimentally, the strategy is similar to the $H \to b\bar{b}$ decay mode. However, the BR is lower and the identification of charm-induced jets is more challenging. Thus far, no observation has been made but searches by the ATLAS and CMS Collaborations are ongoing [176, 177]

$H \to \mu\mu$: The main challenge to this decay channel is the very low BR. ATLAS [178] and CMS [179] searches for this decay channel are ongoing, whereby the result from CMS gives first evidence, observing an excess of three standard deviations from the background-only hypothesis.

[2] In case of the $H \to \gamma\gamma$ decay, the loop contribution more likely involves W bosons because the respective couplings to the Higgs boson and to the photon are larger than for the top quark.

5.1.2 Precision Measurements of the SM Higgs Boson in Di-Tau Final States

As discussed in Section 2.1.4, the theory of electroweak symmetry breaking predicts the coupling of the Higgs boson to vector bosons (see Equations (2.72) and (2.73)) and also to fermions (see Equation (2.78)). These couplings determine the rate at which the Higgs boson is produced via a certain mode and the rate at which it decays. In order to achieve the most accurate results, measurements from different decay channels and from different experiments need to be combined. A consistent presentation of experimental results is thus necessary for such combinations. In addition, the sharing of statistical and systematic uncertainties has to be coordinated. All considerations include the goal of providing long-term results which can be re-interpreted and tested by the theory community. Two such developments are discussed next in more detail, one being the *kappa framework* and the other being *simplified template cross sections*.

The Kappa Framework

The kappa framework [180, 160] is a joint effort of the LHC Higgs working group [181] to define a setup where Higgs boson couplings can be studied. Design considerations of the kappa framework take into account the feasibility of experimental measurements using collision data recorded in the years from 2010 to 2012 (Run1) at the LHC. Underlying assumptions of the kappa framework are that the observed Higgs boson is the only such particle and has spin-0 and is $\mathcal{CP}$-even. Furthermore, the narrow-width approximation is applied such that cross section times BR factorizes in the following way

$$(\sigma \times \mathcal{BR})(i \to \mathrm{H} \to f) = \frac{\sigma_i \cdot \Gamma_f}{\Gamma_\mathrm{H}} \ , \tag{5.4}$$

where σ_i is the production cross section of the initial state, i. The partial decay width into the final state, f, is denoted by Γ_f, and Γ_H is the total width of the Higgs boson. In the context of the kappa framework, each component of Equation (5.4) is replaced by its SM value scaled by the square of a *coupling strength modifier*, κ

$$(\sigma \times \mathcal{BR})(i \to \mathrm{H} \to f) = \frac{\sigma_i^{(\mathrm{SM})} \kappa_i^2 \cdot \Gamma_f^{(\mathrm{SM})} \kappa_f^2}{\Gamma_\mathrm{H}^{(\mathrm{SM})} \kappa_\mathrm{H}^2} \ . \tag{5.5}$$

Hence, the *signal strength modifier* (μ_{if}) takes the following form

$$\mu_{if} \equiv \frac{\sigma \times \mathcal{BR}}{\sigma^{(\mathrm{SM})} \times \mathcal{BR}^{(\mathrm{SM})}} = \frac{\kappa_i^2 \cdot \kappa_f^2}{\kappa_\mathrm{H}^2} \ , \tag{5.6}$$

where κ_H^2 adjusts the SM Higgs boson width to take into account the effect of the modification of all κ_i's and couplings to potential new physics [182]

$$\kappa_\mathrm{H}^2 \equiv \sum_j \frac{\kappa_j^2 \cdot \Gamma_j^{(\mathrm{SM})}}{\Gamma_\mathrm{H}^{(\mathrm{SM})}} \ . \tag{5.7}$$

Numerous benchmark parametrizations are defined based on the kappa framework. A simple way, is to use just a single coupling strength modifier, κ. In such a scenario, Equation (5.6) reduces to measuring just one overall signal strength

$$\mu = \kappa^2 \ , \tag{5.8}$$

quantifying the general deviation from a SM-like Higgs boson coupling structure. However, this parametrization does not distinguish between the different mass generation mechanisms between vector bosons and fermions. Hence, another parametrization uses two coupling strength modifiers, κ_V and κ_F. Higgs boson couplings involving vector bosons are scaled with κ_V, while those involving fermions are scaled with κ_F. In Section 9.2.3, the SM compatibility of the Higgs boson couplings is evaluated in form of contour plots in the κ_V-κ_F-plane. The kappa framework is not a manifestation of QFT and is only sensitive to an overall change in rate. The approach presented next, is also sensitive to the shape of kinematic distributions.

Simplified Template Cross Sections

With the large amount of collision data acquired during Run2, many analyses presented in Section 5.1.1 moved from signal strength or coupling strength measurements to differential measurements. These differential measurements go beyond targeting specific Higgs boson production modes and use specific kinematic selections. The LHC Higgs working group developed a scheme to coordinate such efforts among experiments and between experimental and theory communities [183, 184]. This scheme is called simplified template cross section (STXS) scheme. Differential cross sections can in principle be measured as a function of one or more kinematic variables, but these distributions can be subject to important experimental and / or theoretical uncertainties. Thus, one aim of the STXS scheme is to develop a definition of regions minimizing such effects.

Several competing considerations are driving the definition of the kinematic regions – referred to as *bins* – of the STXS scheme. One main point is to have a natural extension of the signal strength measurements conducted during Run1 in more granular STXS bins. Moreover, the STXS bins are defined such that experimental sensitivities are maximized while the theory dependence is minimized. More specifically, the analysis' sensitivity is optimized to measure cross sections rather than signal strengths inside different STXS bins. Ideally, STXS bins depend only on some theoretical uncertainties and hence reduce the impact of other theoretical uncertainties. In other words, it is desirable to define STXS bins such that they encompass the kinematic region, where variations due to the relevant theoretical uncertainties are largest. Furthermore, some STXS bins are specifically designed to be sensitive to BSM effects. At the end, STXS measurements for different decay channels together with the corresponding partial decay widths can be combined and serve as input for subsequent interpretation. These interpretations can be, for example, in terms of signal strength or coupling modifiers, but also in terms of specific BSM models. Theoretical uncertainties are dealt with in this interpretation step which makes it much easier to re-interpret experimental measurements in the light of future improvements on the theoretical side. More details on the design considerations of the STXS scheme can be found in references [157, 183, 184].

In the following, different versions – referred to as *stages* – of the STXS, relevant to understand the results in Section 9.2, are presented. The STXS stage-0 scheme just distinguishes between different Higgs boson production mechanisms. One of these production mechanisms is ggF (see Figure 5.1a) which will be referred to as ggH from now on. The other relevant production mode is VBF combined with VH (see Figures 5.1b and 5.1c), where the V boson decays hadronically ($V \to qq$). The reason for this combination is that the two processes are indistinguishable in what the final states are concerned and it will be referred to as qqH in this thesis.

The STXS scheme evolved and the most recent one is the stage-1.2 scheme. Within this scheme, the ggH and qqH mechanisms are further split according to the number of

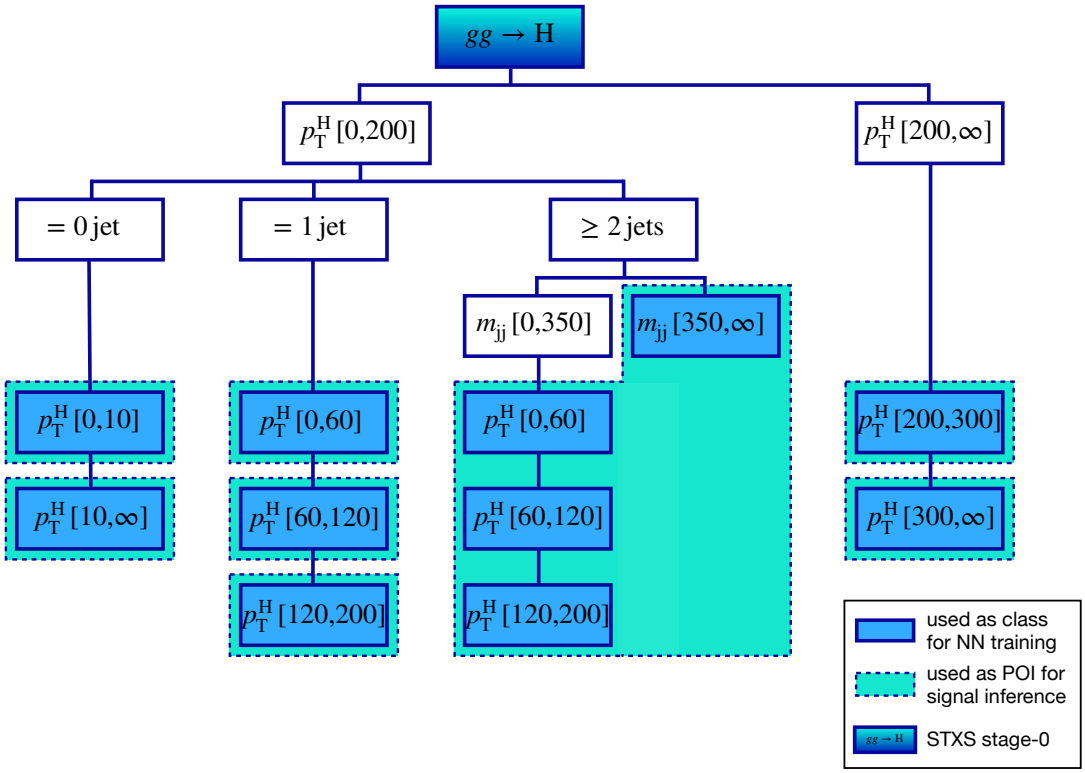


Figure 5.3: The STXS stage-0 and stage-1.2 scheme for ggH as used in the SM $H \to \tau\tau$ analysis is shown. Stage-0 does not apply any further kinematic selections and targets the ggH process as a whole. Kinematic selections on the number of jets, the invariant mass of the first two leading jets (m_{jj}) and the transverse momentum of the Higgs boson (p_T^H) define the individual STXS stage-1.2 bins. The ranges of m_{jj} and p_T^H, given in brackets, are in units of GeV. The bins used as target class by the neural network (NN) (see Section 9.2.1) are shown in filled boxes. Some of these bins are further combined in the final measurement by entering a single parameter of interest (POI) to the likelihood fit (see Section 9.2.4), indicated by dashed boxes.

jets, the invariant mass of the first two leading jets (m_{jj}) or the transverse momentum of the Higgs boson (p_T^H). Note, that the quantities used to define individual STXS bins are defined at particle-level, i.e. before detector simulation. It is allowed to merge different STXS bins, if the sensitivity of the analysis in individual bins is too small. For the SM $H \to \tau\tau$ analysis presented in Section 9.2, the reduced STXS stage-1.2 schemes are shown in Figures 5.3 and 5.4 for the ggH and qqH modes, respectively. Also other STXS schemes have been developed, including stage-1.2 schemes that target other Higgs boson production modes [185].

5.2 Searches for Additional Higgs Bosons

As stated in Table 2.2, the couplings of extra neutral Higgs bosons ($\phi = A, H$) to down-type fermions is enhanced by $\tan\beta$ in the decoupling limit of the MSSM. Practically, this limit is realized for $m_A \gtrsim 300\,\text{GeV}$, a mass range which is also supported by experimental findings [186, 187]. As a consequence, the b-associated production ($bb\phi$) becomes the most dominant production mechanism. Exemplary Feynman diagrams of b-associated production are shown in Figure 5.5. In the same figure also the ggF process ($gg\phi$) is shown. Compared to the SM case, b quark contributions to the fermion loop of $gg\phi$ are more dominant, resulting in a softer spectrum of the transverse momentum of the ϕ

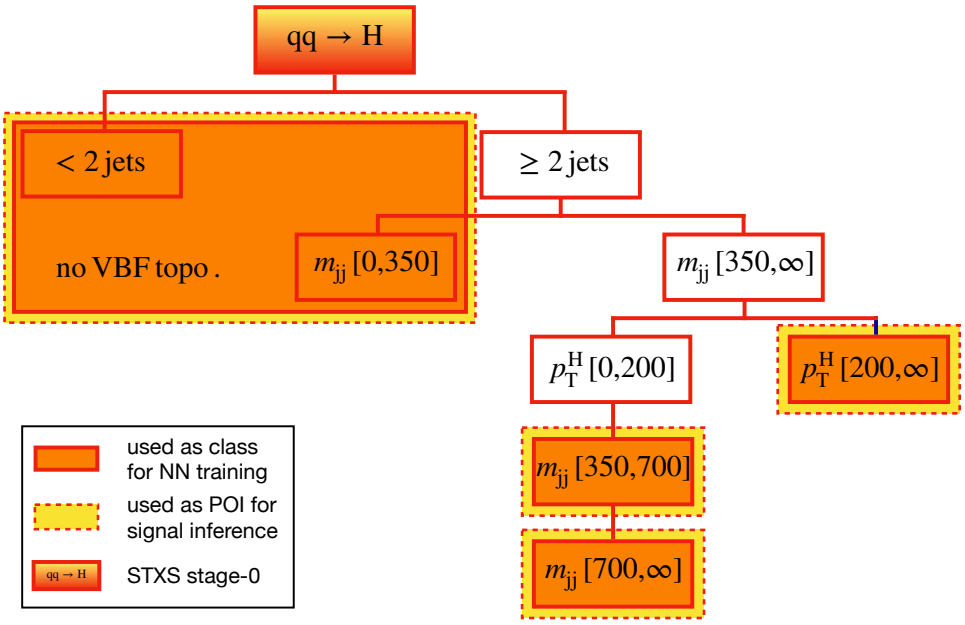
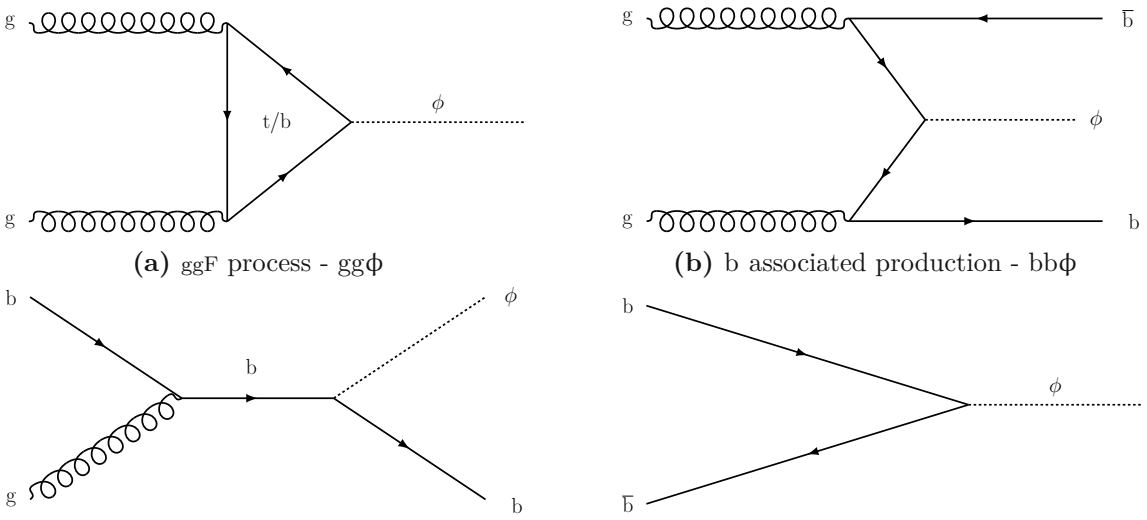


Figure 5.4: The STXS stage-0 and stage-1.2 scheme for qqH as used in the SM H $\to \tau\tau$ analysis is shown. Stage-0 does not apply any further kinematic selections and targets the qqH process as a whole. Kinematic selections on the number of jets, the invariant mass of the first two leading jets (m_{jj}) and the transverse momentum of the Higgs boson (p_T^H) define the individual STXS stage-1.2 bins. The ranges of m_{jj} and p_T^H, given in brackets, are in units of GeV. The bins used as target class by the neural network (NN) (see Section 9.2.1) are shown in filled boxes. Some of these bins are further combined in the final measurement by entering a single parameter of interest (POI) to the likelihood fit (see Section 9.2.4), indicated by dashed boxes.



(a) ggF process - ggϕ **(b)** b associated production - bbϕ

(c) ϕ radiation of b quark - bbϕ **(d)** $b\bar{b}$ annihilation - bbϕ

Figure 5.5: Shown are the dominant ϕ production modes in the context of the MSSM. In Figure 5.5a, the ggF process is depicted (see also Figure 5.1a) who's fermion loop has a significant contribution from b quarks [188]. Figures 5.5b to 5.5d show diagrams of the bbϕ process [189].

boson.

Searches for additional neutral and charged Higgs bosons by both ATLAS [186, 190] and CMS Collaborations [187, 191] exclude masses up to almost 2 TeV. The exclusion limits for high masses is driven by results from $\phi \to \tau\tau$ analyses. As motivated in [192], the addition of a scalar Higgs boson (h_S) in the context of the NMSSM (see Table 2.3) leads to an extended parameter space which is widely unconstrained by experiment. In particular, small couplings of h_S to SM particles suppress the direct production of h_S, even for small masses of h_S below the SM Higgs boson mass ($m_{h_S} < m_{h_{SM}}$). In such scenarios, the BR of the decay $H \to h_S h_{SM}$ can be large. Assuming that the BRs of h_S and h_{SM} are similar, it is expected that h_S decays dominantly into a pair of b quarks (see also Figure 5.2). To arrive at a sensitive experimental signature, the decay $h_{SM} \to \tau\tau$ is chosen because a final state with four b quark initiated jets suffers from overwhelming background contributions from QCD multijet production (see also Section 6.2.2). A final state configuration providing very good sensitivity to the signature of the NMSSM is thus given by the following decay chain

$$H \to h_S h_{SM} \to b\bar{b}\,\tau^-\tau^+ \ . \tag{5.9}$$

The NMSSM signal process from Equation (5.9) is depicted in Figure 5.6, whereby H is produced via ggF. A search for such a signal using Run2 data is presented in Section 9.3.1 and includes the following mass regions

$$\begin{aligned} 240\,\text{GeV} \leq m_H \leq 3\,\text{TeV} \\ 60\,\text{GeV} \leq m_{h_S} \leq 2.8\,\text{TeV} \ , \end{aligned} \tag{5.10}$$

under the condition

$$m_{h_S} + m_{h_{SM}} < m_H \ . \tag{5.11}$$

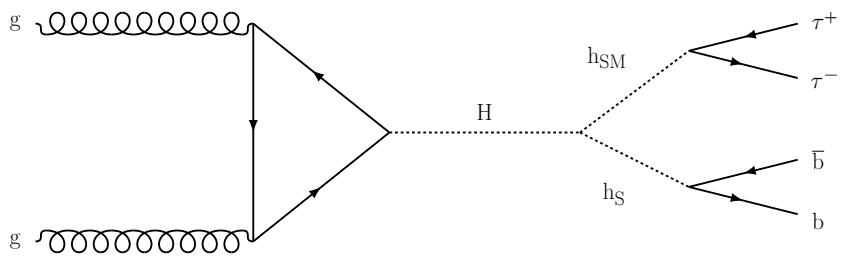


Figure 5.6: Shown is the studied signal process in the context of the NMSSM. A heavy Higgs boson (H) is produced via ggF and decays into the SM Higgs boson (h_{SM}) and a scalar Higgs boson (h_S). A final state with two b quarks and a two tau leptons is expected to have the best trade-off between high BR, while still having an experimental distinct signature.

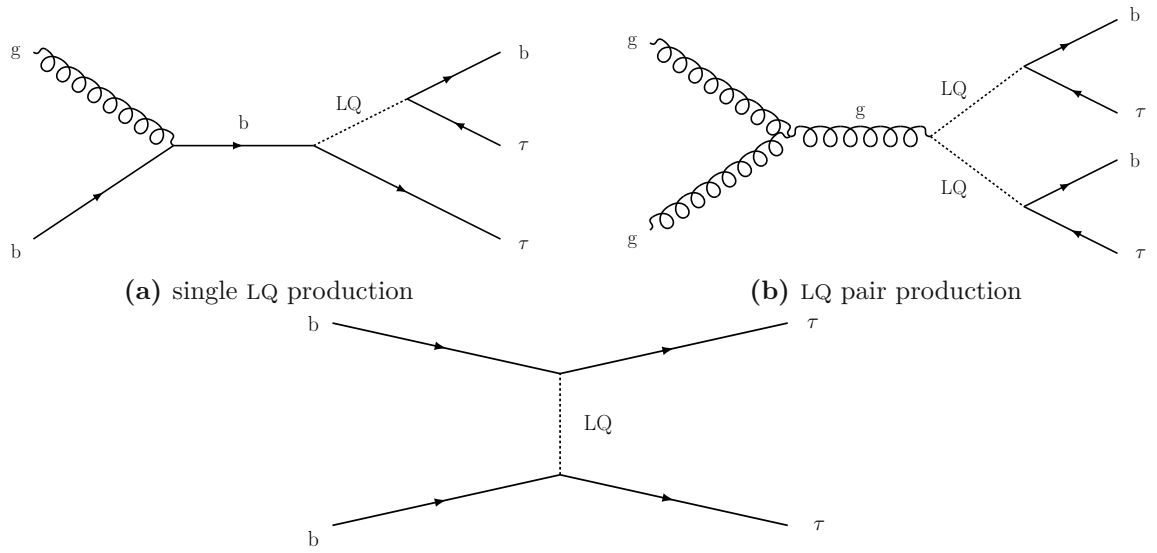
5.3 Leptoquarks at the LHC

In the search for third-generation LQs, three different signal production mechanisms are considered, single LQ production, LQ pair production and non-resonant production. All three mechanisms are depicted in Figure 5.7 As mentioned in Section 2.3.2, several LQ models exist. One model considered in this analysis, is the scalar model which is a simplified version of $\tilde{R}_2$ presented in [193]. The other model involves vector-like LQs and corresponds to the U_1 model [193].

Final states of the LQ search involve a pair of tau leptons as well as b quark initiated jets. Tau pairs involving at least one hadronic tau decay, τ_h, are the most sensitive to the signal processes. Fully leptonic channels are also considered in the analysis and are used to control the SM processes entering the analysis as backgrounds. Results presented in Section 9.3.2, focus solely on the modeling of background contributions coming from jets misidentified as τ_h. These background contributions are estimated with the FF method described in Section 8.2, demonstrating the wide range of application of the developed method.

5.4 Search for Four-Body Decays of Light Top Squarks

Another signal process investigated in this thesis is depicted in Figure 5.8, which is $\mathcal{R}$-parity conserving (see Equation (2.87)). As consequence, SUSY particles are produced in pairs and their decay chains end with a stable SUSY particle, the LSP. In this particular signal model shown in Figure 5.8, a pair of top squarks ($\tilde{t}_1\tilde{t}_1$) is produced in a proton-proton collision. The top squark is considered to be the next-to-lightest supersymmetric particle (NLSP). As mentioned in Section 2.3.1, light top squarks are well motivated from a theoretical point of view. Like in many other SUSY models, the LSP is represented by



(a) single LQ production **(b)** LQ pair production

(c) non-resonant LQ production

Figure 5.7: Different Feynman diagrams are shown for the production of LQs. The coupling of each LQ involves the tau lepton (τ) and the b quark (b). Particle symbols do not distinguished between particles and anti-particles.

the lightest neutralino ($\tilde{\chi}_1^0$). The LSP escapes the CMS detector without a trace, meaning that signal events exhibit a large imbalance in transverse momenta from the reconstructed particles, i.e. large $\boldsymbol{p}_{\mathrm{T}}^{(\mathrm{miss})}$ (see Equation (3.26)).

The mass difference (Δm) between NLSP and LSP (see also Equation (2.95)) is assumed to be below the mass of the W boson. Scenarios, where the lightest neutralino is degenerate with another superpartner have been extensively studied because coannihilation processes in the early universe can reproduce the dark matter relic density [194, 195]. From an experimental side, it is challenging to search for light top squarks at the LHC as shall be discussed in more detail in Chapter 7. This is why experimental constraints on light stop masses are weak, while the strongest limits on top squarks are at beyond the TeV scale. For

$$\Delta m < m_{\mathrm{W}} \ , \tag{5.12}$$

the top squark can decay over virtual top and W boson intermediate states

$$\begin{aligned} \tilde{\mathrm{t}}_1 &\rightarrow \mathrm{t}^* \tilde{\chi}_1^0 \\ \mathrm{t}^* &\rightarrow \mathrm{W}^* \mathrm{b} \\ \mathrm{W}^* &\rightarrow f \bar{f}' \ , \end{aligned} \tag{5.13}$$

if all other SUSY particles are decoupled, i.e. much heavier than LSP and NLSP. Equation (5.13) represents a four-body decay. Two-body decays ($\tilde{t} \rightarrow \mathrm{c}\tilde{\chi}_1^0$) are possible via flavor changing neutral current interactions. Depending on the level of flavor violation in the SUSY model, the two-body decay can be suppressed [196]. In the rest of this thesis, the top squark is assumed to decay via the four-body channel as given in Equation (5.13) with a BR of 100%. This model follows the simplified model spectra (SMS) strategy [197]. Assuming a BR of 100% in the decay under study has to be understood as a benchmark value, while realistic limits in more complex scenarios can be obtained based on the cross section limits from one or several SMS studies. Chapter 7 picks these consideration up and details on the analysis setup used to look for that particular top squark signal.

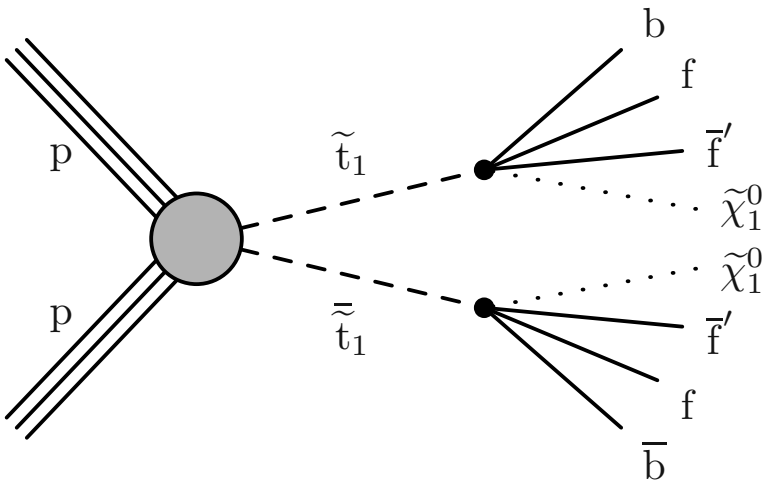


Figure 5.8: Shown is the top squark pair production at the LHC with subsequent four-body decay.

Chapter 6

Tau Pair Selection and Modeling

The final state of interest involves a pair of tau leptons. An example of such a di-tau final state, is the decay of the SM Higgs boson, symbolically written as $H \to \tau\tau$. Table 6.1 shows all possible di-tau final states with the corresponding BR. In what follows, the focus is set on final states involving at least one hadronic tau lepton decay (τ_h). According to Table 6.1, final states with at least one τ_h make up 88% of all possible decays. The statements of this chapter apply to the SM $H \to \tau\tau$ analysis and to a very large extent also to the searches for additional Higgs bosons presented in this thesis. Slight variations do exist in case of the LQ search. For the sake of clarity, the discussion focuses solely on the SM $H \to \tau\tau$ analysis, without listing differences to the other analyses.

In Section 6.1, the selection of leptons and subsequent formation of tau pairs is explained. Several different background processes also pass the selections and form valid tau pairs. These backgrounds are presented throughout Section 6.2. The modeling of the background contributions is discussed afterwards. Section 6.3 covers their modeling by means of simulation and Section 6.4 presents a data-driven method used to model certain backgrounds. The last section of this chapter discusses corrections that are needed to improve the data modeling.

6.1 Tau Pair Formation

Before selecting any electron, muon or τ_h candidate, their energy is corrected and an uncertainty of the correction is assigned. Lepton energy corrections are also referred to as lepton energy *scales* and change the four-momentum of the lepton. Muons are very precisely reconstructed compared to electrons and τ_h. Therefore, no energy scales for muons are used.

For electrons, there is not just an energy scale derived but also a resolution correction [198, 199]. Energy scale and resolution corrections, derived as functions of electron p_T and η, are applied to both data and simulation on an event-by-event basis. The corrections are deduced from a data set enriched in $Z \to ee$ events. For the energy scale, the Z peak in the invariant mass distribution (m_{ee}) in data is matched to the one in simulation.

$\tau\tau$ final state	$e\tau_h$	$\mu\tau_h$	$\tau_h\tau_h$	$e\mu$	ee and $\mu\mu$
BR [%]	23	23	42	6	6

Table 6.1: This table summarizes the different possible di-tau final states and quantifies their occurrence in terms of BRs. Neutrinos in the final state are not explicitly written. The symbol τ_h represents hadronically decaying tau leptons (see also Table 3.2).

The resolution correction adjusts the width of the simulated m_{ee} distribution to match the one in data.

Energy scales for τ_{h} are measured in four different τ_{h} decay modes. These decay modes are (see also Equation (3.24))

$$D_{\mathrm{mode}} \in \{0, \{1,2\}, 10, 11\} \text{ , i.e.}$$
$$D_{\mathrm{mode}} \in \begin{cases} \text{1-prong} \\ \text{1-prong} + \pi^0 \text{ or 1-prong} + 2\pi^0 \\ \text{3-prongs} \\ \text{3-prongs} + \pi^0 \end{cases} . \tag{6.1}$$

They amount in correcting the τ_{h} four-momentum by 0.1 to 1%. Decay modes of the the type 1-prong ($D_{\mathrm{mode}} = [0, 1, 2]$) are contaminated with electrons and muons that are misidentified as τ_{h} because they leave a similar signature inside the detector as 1-prongs and are produced abundantly in $\mathrm{Z} \to \mathrm{ee}$ and $\mathrm{Z} \to \mu\mu$ processes, respectively. Therefore, dedicated energy scales for electrons and muons being misidentified as τ_{h} are measured. For muons these corrections are in the order of 1% and for electrons they range from 1 to 5%. For misidentified electrons the corrections are split between barrel and endcap region.

The first step of the signal event selection consists in deciding on a trigger strategy. Since the final states involve electrons, muons and τ_{h}, a combination of so-called *single-lepton triggers* and *cross-triggers* is used. Single-lepton triggers look for single leptons with a transverse momentum above a certain threshold. The value of the threshold controls the trigger rate and is chosen to not overcrowd the data acquisition system. For example, the single-electron trigger, e(24), decides on keeping the event if at least one electron with transverse momentum above 24 GeV is found. Since trigger decisions have to be taken quickly, only a fast computation of transverse momenta can be performed, which results in a worse resolution than can be achieved offline[1]. The region where the trigger selection efficiency grows from 0 to 100%, when plotted against the transverse momentum calculated offline, is called *turn-on* region[2]. Cross-triggers look for two or more distinct objects within a collision event. For instance, the muon-tau cross-trigger, $\mu(20)\tau_{\mathrm{h}}(27)$, selects events with a muon and a τ_{h} with transverse momenta above 20 GeV and 27 GeV, respectively. Cross-triggers can typically afford lower thresholds in transverse momenta because the occurrence of several objects is less frequent than each single object. This allows reducing the threshold while staying within the maximally allowed trigger rate. The trigger strategy employed for the different di-tau channels and different data-taking years is summarized in Table 6.2.

Further object selections – colloquially referred to as *cuts* – are applied on events selected by the trigger conditions listed in Table 6.2 to enhance the efficiency of choosing genuine electrons, muons and τ_{h} candidates coming from the signal process. A first set of selection criteria is summarized in Table 6.3 and is discussed in the following. Impact parameter[3] cuts on d_{xy} and d_{z} are applied in order to reduce contributions from decays of (long-lived) hadrons and PU. All leptons need to pass a certain threshold of a dedicated lepton ID discriminant. For electrons, the MVA 90% ID is chosen, which is presented in

[1] In general, all kind of *online* operations have to be executed in real time. Once there are no strict time constraints, the term *offline* is used, e.g. all operation at analysis-level.

[2] In order to be in the region of 100% selection efficiency, typically a higher threshold in transverse momentum is applied offline as will be discussed later.

[3] The impact parameter is the point of closest approach to the PV. Its longitudinal component is denoted as d_{z}, whereas the transverse component is denoted as d_{xy}.

Channel	Year	Trigger strategy
$e\tau_h$	2016	$e(24)\tau_h(20)$ or $e(25)$
	2017	$e(24)\tau_h(30)$ or $e(32)$ or $e(35)$ or $e(27)$
	2018	$e(24)\tau_h(30)$ or $e(32)$ or $e(35)$
$\mu\tau_h$	2016	$\mu(19)\tau_h(20)$ or $\mu(22)$
	2017/18	$\mu(20)\tau_h(27)$ or $\mu(24)$ or $\mu(27)$
$\tau_h\tau_h$	2016	$\tau_h(35)\tau_h(35)$
	2017/18	$\tau_h(35)\tau_h(35)$ or $\tau_h(40)\tau_h(40)$

Table 6.2: This table summarizes the trigger selections applied in all tau decay channels involving at least one hadronic tau decay. Both single-lepton triggers and cross-triggers are used and connected via a logical or. The values in parentheses indicate the threshold on the transverse momentum in GeV of the corresponding object.

Section 3.5.6 along with different muon IDs. For muons, the `medium` muon ID is chosen. For τ_h candidates, the DeepTau classifier is used (see Section 3.5.8). Depending on the di-tau decay channel, different selection conditions are applied as seen in Table 6.3. Electrons and muons originating from tau decays have typically less hadronic activity in their vicinity compared to those originating from decays of jet constituents. Therefore, cuts on the relative isolation, defined in Equations (3.15) and (3.17), are applied to both electrons and muons. For electrons, a cone with the radius parameter $\Delta R = 0.3$ is used, whereas for muons $\Delta R = 0.4$ is used. Selection requirements on the transverse momentum depend on the trigger used to select the event. The transverse momentum has to be 1 GeV above the trigger threshold for electron and muon triggers and in case of a τ_h trigger, 5 GeV above the corresponding cross-trigger threshold. This choice avoids selecting events in the trigger turn-on region which is very difficult to model by simulation. Hence, the p_T cuts quoted in Table 6.3 have to be understood as being applied so that the trigger is maximally efficient. Ranges in pseudorapidity are designed such that they include the tracker acceptance and cover the isolation cones of each object.

Electrons, muons and τ_h passing the selections summarized in Table 6.3 are used to build tau-pair candidates. In order to be considered a tau pair, (τ_1, τ_2), the constituents have to be separated by $\Delta R > 0.5$. In the semi-leptonic final states – i.e. $e\tau_h$ and $\mu\tau_h$– the electron/muon is treated as τ_1 and the τ_h as τ_2. In the fully-hadronic channel $(\tau_h\tau_h)$, each tau pair is considered twice, where each τ_h within a pair is once assigned to τ_1 and once to τ_2. The pair building is performed separately for each channel and for a single collision event, multiple tau pairs can be formed which are then sorted according to the following scheme:

1. Compare the isolation of τ_1 between tau pairs. In case of the $e\tau_h$ channel, the relative isolation – defined in Equation (3.17) – is used to sort the tau pairs. In case of the $\mu\tau_h$ channel, Equation (3.15) is used instead to sort muons (τ_1) according to their relative isolation. In both cases, smaller values in relative isolation correspond to better isolated leptons. For the $\tau_h\tau_h$ channel, the isolation for τ_h is taken to be the output of the DeepTau classifier against jets, D_{jet}. The output of D_{jet} is between zero and one, with values closer to one corresponding to a more isolated τ_h. The list of all tau pairs is sorted according to the isolation of τ_1 such that the most isolated τ_1 is at the top of the list[4]. If the isolation between two pairs is similar

[4]Note that depending on the channel analyzed, τ_1 are all the same objects, i.e. either electrons, muons or τ_h's.

electron	muon	τ_h			
$d_{xy} < 0.045\,\text{cm}$ $d_z < 0.2\,\text{cm}$	$d_{xy} < 0.045\,\text{cm}$ $d_z < 0.2\,\text{cm}$	– $d_z < 0.2\,\text{cm}$			
MVA 90% ID	`medium` ID	DeepTau	$e\tau_h$	$\mu\tau_h$	$\tau_h\tau_h$
		D_{jet}	`tight`	`tight`	`tight`
		D_e	`tight`	`vvloose`	`vvloose`
		D_μ	`vloose`	`tight`	`vloose`
$I^{(e)}_{rel} < 0.1$ $\Delta R = 0.3$	$I^{(\mu)}_{rel} < 0.15$ $\Delta R = 0.4$	–			
$p_T^{(e)} > 25\,\text{GeV}$	$p_T^{(\mu)} > 20\,\text{GeV}$	$e\tau_h$		$\mu\tau_h$	$\tau_h\tau_h$
		$p_T^{(\tau_h)} > 30\,\text{GeV}$		$p_T^{(\tau_h)} > 30\,\text{GeV}$	$p_T^{(\tau_h)} > 40\,\text{GeV}$
$\lvert\eta^{(e)}\rvert < 2.1$	$\lvert\eta^{(\mu)}\rvert < 2.1$	$\lvert\eta^{(\tau_h)}\rvert < 2.4$			

Table 6.3: This table summarizes the selections imposed on the different final state objects of τ candidate decays in the SM $H \to \tau\tau$ analysis. For hadronic tau decay products, some selections vary across different di-tau final states as indicated in the table. A dash indicates that no selection is applied.

within 10^{-5}, the transverse momenta are compared in a next step.

2. Compare the transverse momenta, $p_T^{(\tau_1)}$, of all pairs. Move a pair up if the transverse momentum is larger. In case the transverse momenta are similar within $10^{-5}\,\text{GeV}$, proceed with step 3.

3. Compare the isolation values of τ_2 and do the same as in step 1. Should the isolation of τ_2 be similar within 10^{-5}, step 4 is checked.

4. Compare the transverse momenta, $p_T^{(\tau_2)}$, and move a pair up in the list if $p_T^{(\tau_2)}$ is larger.

At the end of the above procedure, the tau pair in the first place of the list is chosen to be the *signal* tau-pair candidate. Afterwards, the constituents (τ_1 and τ_2) of the signal tau pair are checked to match the physics objects responsible for triggering the event. More specifically, a matching within $\Delta R < 0.5$ between $\tau_{1/2}$ and the triggering object saved by the HLT has to be achieved in order to keep the event for further analysis.

Events enter the signal search region, if they satisfy the following (channel-dependent) requirements. Common to all final states is the requirement that signal tau pairs must have opposite signed charges of τ_1 and τ_2. Furthermore, a *third lepton* veto is imposed to avoid double counting of events among different channels. Specifically, if the event contains electrons or muons passing the "overlap veto" selection as specified in Table 6.4, it is discarded. For the semi-leptonic channels, two further selection conditions are applied. In case of $e\tau_h$ channel, oppositely charged electron pairs with $\Delta R > 0.15$ are built based on the collection labeled as "$Z \to ee$ veto" in Table 6.4. If any such pair is found, the event is rejected in order to minimize the contamination from $Z \to ee$ events. Analogously, for the $\mu\tau_h$ channel, oppositely charged muon pairs are built using the "$Z \to \mu\mu$ veto" collection defined in Table 6.4. If such muons with $\Delta R > 0.15$ exist, the event is discarded to avoid large contributions from $Z \to \mu\mu$ events. The other selection applied, concerns

electron		muon	
overlap veto	$Z \to ee$ veto	overlap veto	$Z \to \mu\mu$ veto
$d_{xy} < 0.045\,\text{cm}$ $d_z < 0.2\,\text{cm}$			
MVA 90% ID	`veto` ID	`medium` ID	`loose` ID
$I^{(e)}_{rel} < 0.3$ $\Delta R = 0.3$		$I^{(\mu)}_{rel} < 0.3$ $\Delta R = 0.4$	
$p_T^{(e)} > 10\,\text{GeV}$	$p_T^{(e)} > 15\,\text{GeV}$	$p_T^{(\mu)} > 10\,\text{GeV}$	$p_T^{(\mu)} > 15\,\text{GeV}$
$\lvert\eta^{(e)}\rvert < 2.5$		$\lvert\eta^{(\mu)}\rvert < 2.4$	

Table 6.4: For electron and muon candidates, special collections are formed which are different from those used to build signal tau pairs (see Table 6.3). For both kind of candidates, an overlap veto collection is defined, which is used to avoid double-assignments of tau pairs to several di-tau final states. Furthermore, contributions from $Z \to ee$ and $Z \to \mu\mu$ background processes are reduced by imposing that there are no electron or muon candidates in the "$Z \to ee$ veto" or "$Z \to \mu\mu$ veto" collection in case of the $e\tau_h$ or $\mu\tau_h$ channel, respectively.

$e\tau_h$	$\mu\tau_h$	$\tau_h\tau_h$
opposite-sign (OS) charges third lepton veto		
$Z \to ee$ veto	$Z \to \mu\mu$ veto	–
$m^{(\ell)}_{T,PUPPI} < 70\,\text{GeV}$		–

Table 6.5: Summary of the signal region selection cuts for the $e\tau_h$, $\mu\tau_h$ and $\tau_h\tau_h$ final states. A dash indicates that no selection is applied

the transverse mass ($m^{(\ell)}_{T,PUPPI}$)

$$m^{(\ell)}_{T,PUPPI} = \sqrt{2p_T^{(\ell)} \cdot |\boldsymbol{p}^{(miss)}_{T,PUPPI}| \left(1 - \cos\Delta\phi\left(\ell, \boldsymbol{p}^{(miss)}_{T,PUPPI}\right)\right)} < 70\,\text{GeV}, \quad \ell \in \{e, \mu\}\,. \tag{6.2}$$

The above definition of the transverse mass differs from the definition in Equation (3.5). Equation (6.2) derives from 1-to-2 particle decays, where one decay product is invisible. Invisible particles can only be estimated by means of the missing transverse momentum (see Section 3.5.9). The definition in Equation (6.2) neglects the masses of the decay products and has an endpoint at the mother particle mass. The cut of $m^{(\ell)}_{T,PUPPI} < 70\,\text{GeV}$ will become important in the description of the FF method in Chapter 8. The signal region selections are summarized in Table 6.5.

6.2 Background Composition

Different physical processes, other than $H \to \tau\tau$ decays, can pass the tau pair selection and therefore enter the signal region described in the previous section. These processes are treated as background. In order to precisely study the $H \to \tau\tau$ signal, the contribution of these backgrounds has to be estimated as accurately as possible. The relevant backgrounds to be considered are

- W boson production in association with at least one jet (W+jets),

- Processes composed uniquely of jets produced through the strong interaction, referred to as QCD multijet events.
- top quark pair production ($t\bar{t}$),
- Z boson production and
- production of W or Z boson pairs (diboson) as well as single-top quark production.

These backgrounds fall into three distinct groups of backgrounds types:

1. There are background types comprising genuine tau pairs which are indistinguishable from the $H \to \tau\tau$ signal final states. Such backgrounds are termed *irreducible*. The main contribution to this background is posed by the $Z \to \tau\tau$ decay. There is also a small fraction of $t\bar{t}$ and diboson processes with genuine di-tau final states. All these processes are estimated with the τ-embedding method described in Section 6.4.

2. Processes where at least one quark- or gluon-initiated jet is misidentified as a τ_h, form the second type of backgrounds. These are labeled as jet $\to \tau_h$ and are estimated with the fake factor (FF) method. Chapter 8 is dedicated to the FF method which forms the central piece of this thesis.

3. Everything not belonging to the first two groups is termed as *other backgrounds* (other bkg.). More specifically, it comprises processes where electrons or muons are misidentified as originating from a tau decay. Electrons and muons misidentified as τ_h, fall into this category as well.

The composition of the different backgrounds, in terms of the three categories defined above, is depicted in Figure 6.1. The composition is similar between the different data-taking years as well as between the semi-leptonic final states. In what follows, each of the background processes is discussed in more detail. During that discussion, the summary plot presented in Figure 2.4 is used to put the production rates of these background processes into perspective.

6.2.1 W+jets Background Process

For Run2 conditions, i.e. for a center of mass energy of 13 TeV, the W boson production cross section is in the order of 10^4 pb. The BR for leptonic W boson decays varies between 10 to 11%, depending on the lepton flavor $\ell \in \{e, \mu, \tau\}$. Hadronic W bosons decays make up 67% of the total decay width [49]. Higgs boson production cross sections are also shown in Figure 2.4 and are four orders of magnitude smaller than the W production cross section.

Examples of W boson productions in association with at least one jet are shown in Figure 6.2. Leptons from the leptonic W decay can be paired with a jet that is misidentified as a τ_h. Depending on the lepton flavor of the W boson decay, the W+jets process poses a background in the $e\tau_h$, $\mu\tau_h$, or $\tau_h\tau_h$ channel, respectively. However, as discussed in Chapter 8, the contribution to the $\tau_h\tau_h$ channel is small. The W+jets process falls into the category of jet $\to \tau_h$ and is estimated with the FF method.

6.2.2 QCD Multijet Production

Multijet production in QCD comprises interactions solely mediated via the strong force. The production cross section is at least six orders of magnitude above the $H \to \tau\tau$ signal

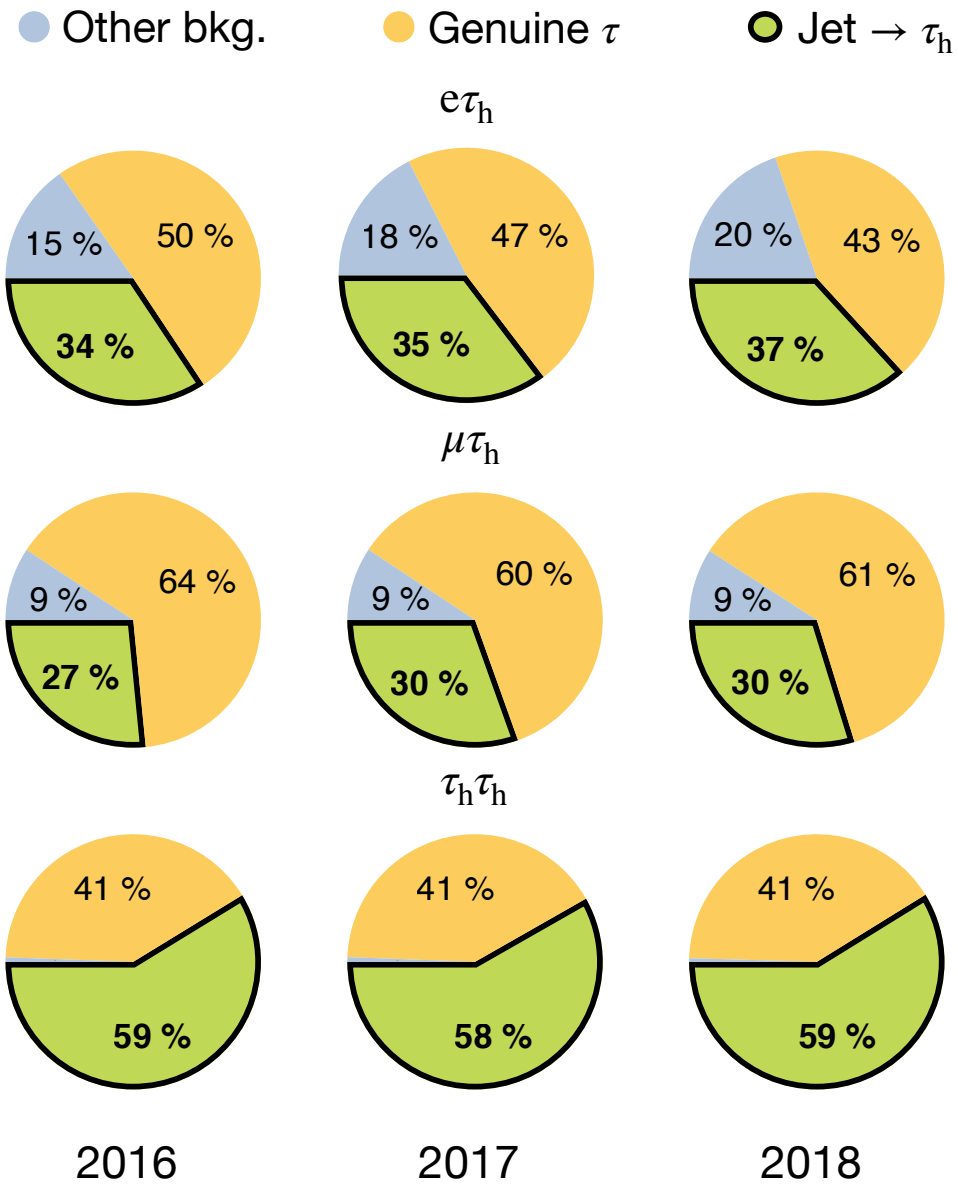
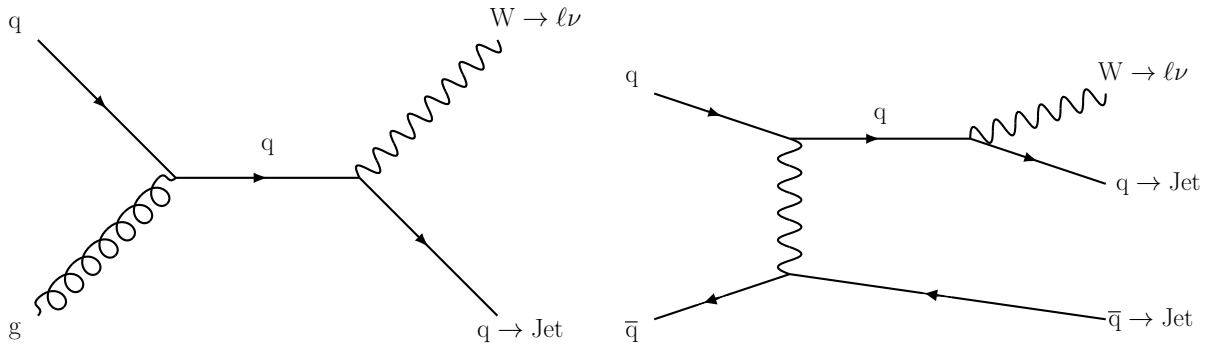


Figure 6.1: Shown are the different background contributions to the SM $H \to \tau\tau$ analysis after the selections described in Section 6.1. Backgrounds are divided in three categories, genuine τ, jet $\to \tau_h$ and "other bkg.". The τ-embedded samples estimate the contribution resulting in genuine tau pairs. Contributions from quark- or gluon-initiated jets that are misidentified as τ_h, are labeled by jet $\to \tau_h$ and are estimated by the FF method. Everything not covered by these two estimation methods falls into the third category, "other bkg.", as detailed in the text. Background compositions do not vary strongly between the different years of data taking. Furthermore, the compositions of the semi-leptonic channels – $e\tau_h$ and $\mu\tau_h$– are comparable. The $\tau_h\tau_h$ channel is dominated by jet $\to \tau_h$ processes.



(a) Example of a W+1jet process **(b)** Example of a W+2jets process

Figure 6.2: Examples of Feynman diagrams of the W+jets process are shown. In (a), a final-state quark forms a jet. If this jet is misidentified as τ_h and in combination with a leptonic W boson decay, it is possible that such an event passes the $H \to \tau\tau$ selection (see Table 6.3). In (b), a further example of a W+jets process is shown, with two jets in the final state.

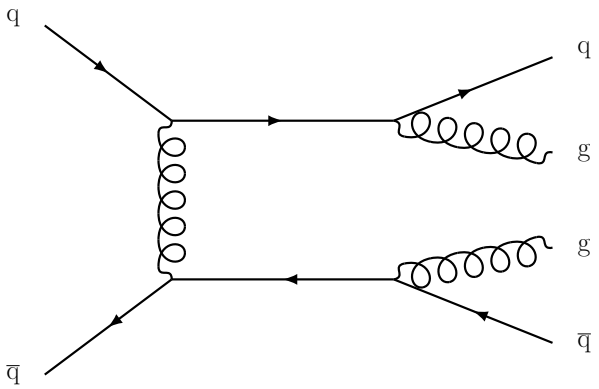


Figure 6.3: Shown is one of many possible QCD multijet processes. This particular Feynman diagram has two quarks and two gluons in its final state, leading to four jets.

process (see Figure 2.4). In Figure 6.3, one of many possible Feynman diagrams subsumed into this process is shown. Quark- or gluon-initiated jets can be misidentified as τ_h and even, less commonly, as muons or electrons. Analysis selections, as reported in Tables 6.3 and 6.5, aim at reducing these misidentification rates, most notably by requiring events to pass the `tight` threshold of the D_{jet} discriminant. However, QCD multijet events still enter the $H \to \tau\tau$ analysis because of the large production cross section. Misidentification of jets as τ_h is the dominant background in the $\tau_h\tau_h$ channel as can be inferred from Figures 6.1 and 8.2.

Due to the high production rate, a huge amount of MC simulations would be needed to match the integrated luminosity. Typically, the QCD MC samples have insufficient statistical power. Furthermore, it is difficult to model QCD multijet processes by means of simulation because the strong coupling constant takes larger values at lower energy scales (see also Section 3.4) and a purely perturbative approach is not sufficient. Many orders in perturbation theory would be needed to correctly model QCD multijet events. Instead, the FF method describes QCD multijet processes directly from data, circumventing the need of simulating this process.

6.2.3 Top Quark Pair Production

The $t\bar{t}$ production cross section at $\sqrt{s} = 13\,\text{TeV}$ is of the order of 800 pb [49]. Leading-order Feynman diagrams of this process are shown in Figure 6.4. Top quarks decay with

$t\bar{t}$ Final state	Process	BR [%]	Estimation method
dilepton			
	$\tau\tau$	1	τ embedding
	eτ	2	simulation
	$\mu\tau$	2	simulation
	ee	1	simulation
	eμ	2	simulation
	$\mu\mu$	1	simulation
lepton+jet			
	τ+jet	15	FF method
	e+jet	15	FF method
	μ+jet	15	FF method
all jets			
	jet+jet	46	FF method

Table 6.6: A $t\bar{t}$ process almost exclusively decays into $b\bar{b}W^+W^-$. Each W boson decays further leptonically or hadronically. The hadronic decays produce jets. In this table the different final states are listed together with their BRs. The $b\bar{b}$ pair is omitted as well as the neutrinos produced in leptonic decays.

a BR of 99.8% into a W boson and a b quark. Several final states are possible, following all possible W boson decay channels. Table 6.6 summarizes the different $t\bar{t}$ final states together with the estimation method used. All final states involving hadronic W boson decays, leading to potentially misidentified τ_h, are estimated with the FF method. The small part of two genuine tau leptons in the final state is estimated with the τ-embedding method. All other final states are estimated by means of simulation.

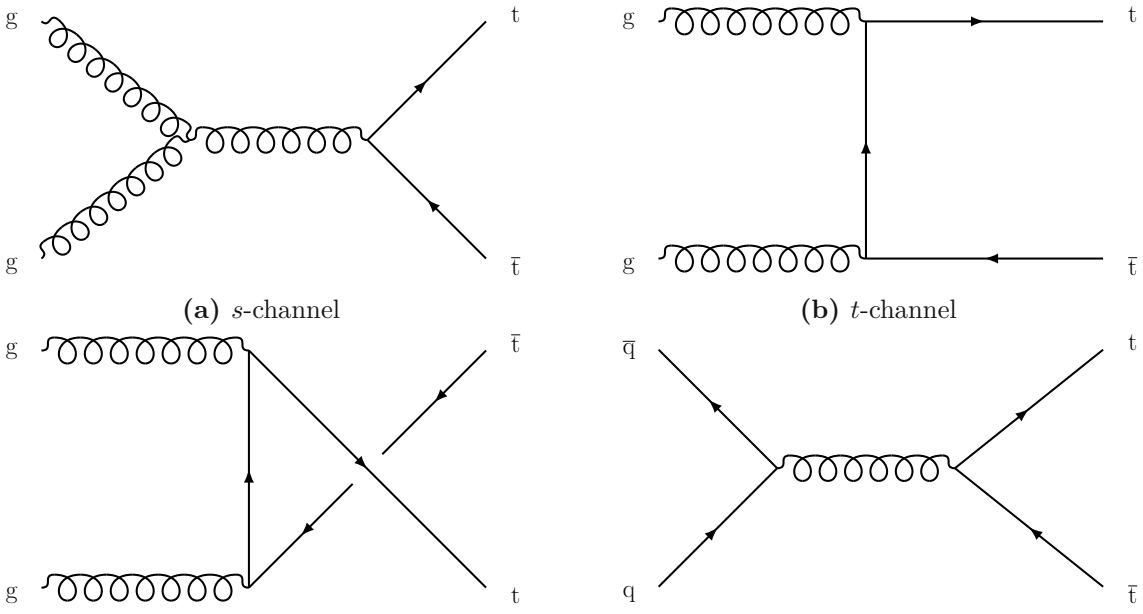
6.2.4 Z Boson Production

Lepton pairs at the LHC are dominantly produced via the Drell-Yan (DY) process shown in Figure 6.5a. The Z boson decay into a pair of tau leptons has the identical final state as the $H \to \tau\tau$ signal process. The invariant masses are different between reconstructed $Z \to \tau\tau$ and $H \to \tau\tau$ decays, being 91 GeV and 125 GeV, respectively. However, due to the neutrinos in the subsequent tau decay, the invariant di-tau mass is smeared and makes it difficult to tell these two processes apart. The $Z \to \tau\tau$ process is estimated via the τ-embedding technique (see Section 6.4).

Similar to W+jets processes, the Z boson production can be associated to one or more jets. An exemplary Feynman diagram is depicted in Figure 6.5b. The contribution Z+jets is estimated with the FF method. Despite the DeepTau selections applied on D_e and D_μ (see Table 6.3), respectively, it can still happen that an electron or muon gets misidentified as τ_h. In that case $Z \to ee$ and $Z \to \mu\mu$ processes enter the analysis inside the eτ_h and $\mu\tau_h$ channel, respectively. These lepton-τ_h misidentification processes fall into the category of "other bkg." (see also Figure 6.1) and are estimated with simulated events.

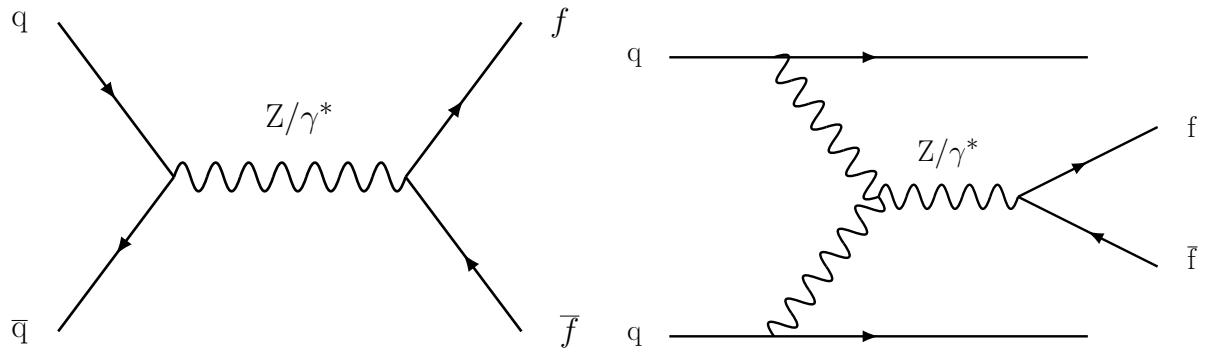
6.2.5 Diboson and Single Top Processes

Diboson processes include the production of WW, WZ and ZZ as shown in Figure 6.6. Many different final states can be formed after the subsequent W and Z boson decays



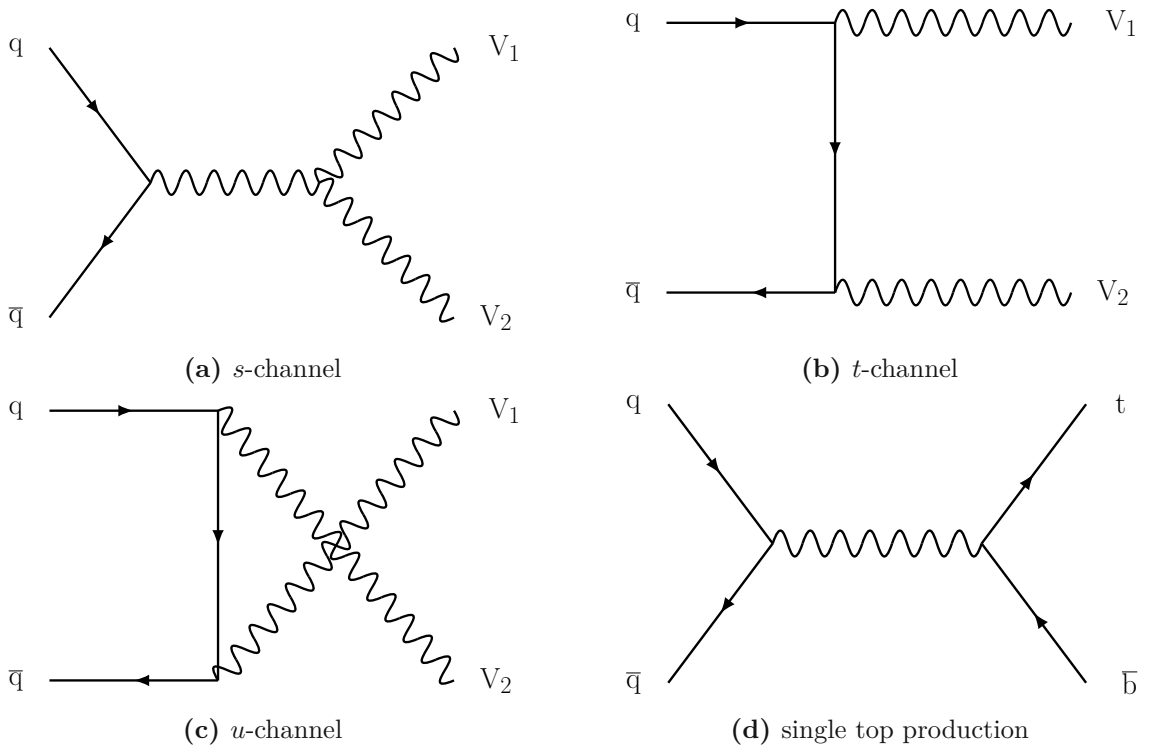
(a) s-channel **(b)** t-channel **(c)** u-channel **(d)** $q\bar{q}$ annihilation

Figure 6.4: Shown are the LO Feynman diagrams of the $t\bar{t}$ production process at the LHC. Figures (a) to (c) depict the $t\bar{t}$ production via gluon fusion in the s, t and u-channel, respectively. They make up 90% of $t\bar{t}$ events while the remaining 10% are produced via the $q\bar{q}$ annihilation process shown in (d).



(a) Drell-Yan process **(b)** Example of a Z+jets process

Figure 6.5: Shown are two examples of Z boson production and subsequent decay into a pair of fermions, $f\bar{f}$. In (a), the Drell-Yan process is shown and in (b) the production via fusion of two electroweak bosons. At the LHC the cross section for process (a) exceeds the one of process (b) by three orders of magnitude.



(a) s-channel **(b)** t-channel **(c)** u-channel **(d)** single top production

Figure 6.6: In (a) to (c), the diboson production mechanisms in the s, t and u-channel are shown, respectively. An example of a single top production is depicted in (d).

and hence, this background contributes to all di-tau final states. The small contribution of single top quark production (see Figure 6.6d) is attached for convenience to this class of backgrounds. For the rest of this thesis, diboson combined with single top processes will be simply referred to as VV. Similar as for $t\bar{t}$, the VV process contributes in all three different ways to the tau final states. Genuine tau pairs produced via VV are estimated using the τ-embedding technique and the part involving a jet that is misidentified as τ_h is estimated with the FF method. All other possibilities are covered by simulation.

6.3 Simulated Samples

In the previous section, the different background processes relevant for the $H \to \tau\tau$ analysis are discussed. Some of these backgrounds need to be estimated by means of MC simulation. This concerns all backgrounds contributing to the category "other bkg." in Figure 6.1. Furthermore, simulated samples are also needed as part of the FF method as described in Section 8.2. Table 6.7 gives an overview of the different generators used to simulate different samples. MadGraph 5 (MG5) aMC@NLO [109] and POWHEG [110] are used to model the hard interaction process. They are complemented with PYTHIA 8.2 [113] which simulates the parton showering, hadronization, and underlying event. The decay of tau leptons is also modeled using PYTHIA 8.2.

It is important to distinguish within a simulated sample the three categories introduced in Section 6.2 – i.e. genuine tau pairs, jet $\to \tau_h$ or "other bkg.". This can be achieved by *generator matching* which works as follows. For each reconstructed electron, muon or τ_h of a tau pair, an angular distance of $\Delta R = 0.2$ is used to check for a compatible object at generator level, i.e. before detector simulation. In case several generator-level objects are found, the one closest to the reconstructed electron, muon or τ_h is chosen. The truth information of the generator-level object is then used further. Different labels are

Process	includes	Generator	Precision
$t\bar{t}$	$t\bar{t}$	POWHEG	NLO QCD
W+jets	up to four add. jets	MG5	LO (NNLO) QCD, NLO electroweak
$Z \to \ell\ell$	up to four add. jets	MG5	LO (NNLO) QCD, NLO electroweak
VV	WW, WZ, ZZ	MG5 aMC@NLO	NLO QCD
	single t	MG5	NLO QCD
ggH	ggH	POWHEG	NLO (N^3LO) QCD
qqH	VBF + VH($V \to qq$)	POWHEG	NLO (NNLO) QCD

Table 6.7: This table summarizes different MC generators used to simulate different background processes (top) and signal processes (bottom) of the SM $H \to \tau\tau$ analysis. The order in perturbative QCD and electroweak theory is quoted in the last column. Values in parentheses indicate differences in the precision used to derive the respective inclusive cross section. The table is taken and changed from [12].

label	type
1	prompt electron
2	prompt muon
3	electron which is a direct decay product of a prompt tau lepton: $\tau \to$ e
4	muon which is a direct decay product of a prompt tau lepton: $\tau \to \mu$
5	hadronic tau decay $\tau \to \tau_h$
6	misidentified jet

Table 6.8: Different labels are assigned to reconstructed electron, muon and τ_h candidates depending on the matched objects at generator level. The labels and a brief explanation are summarized in this table.

assigned as summarized in Table 6.8. The genuine tau part – which is estimated with the τ-embedding method – comprises events where tau pairs carry the generator matching labels $(3, 5)$, $(4, 5)$ and $(5, 5)$ for the final states $e\tau_h$, $\mu\tau_h$ and $\tau_h\tau_h$, respectively. Events with at least one tau candidate with the label 6 from generator matching, fall into the category jet $\to \tau_h$ and are estimated with the FF method. All other possible combinations of generator-matching labels are covered by "other bkg.".

6.4 The τ-embedding Method

As discussed in Section 6.2, different background processes produce genuine di-tau final states. Most abundant is the $Z \to \tau\tau$ process which is hard to disentangle from the $H \to \tau\tau$ signal process due to the similar masses of the Z boson and the Higgs boson. Furthermore, $t\bar{t}$ and diboson production contribute to a small degree to this background. All decays to genuine tau pairs in $Z \to \tau\tau$, $t\bar{t}$ and diboson processes are mediated by the weak force which couples with the same strength to all lepton flavors in the SM. The τ-embedding technique [200] is based on this lepton universality and aims at replacing muon pairs with tau pairs. Figure 6.7 illustrates the four main steps of the τ-embedding technique:

1. First, di-muon events are selected in collision data recorded by the CMS detector. Since the CMS detector is well suited for detecting and reconstructing muons, the selection is very pure in genuine di-muon events.

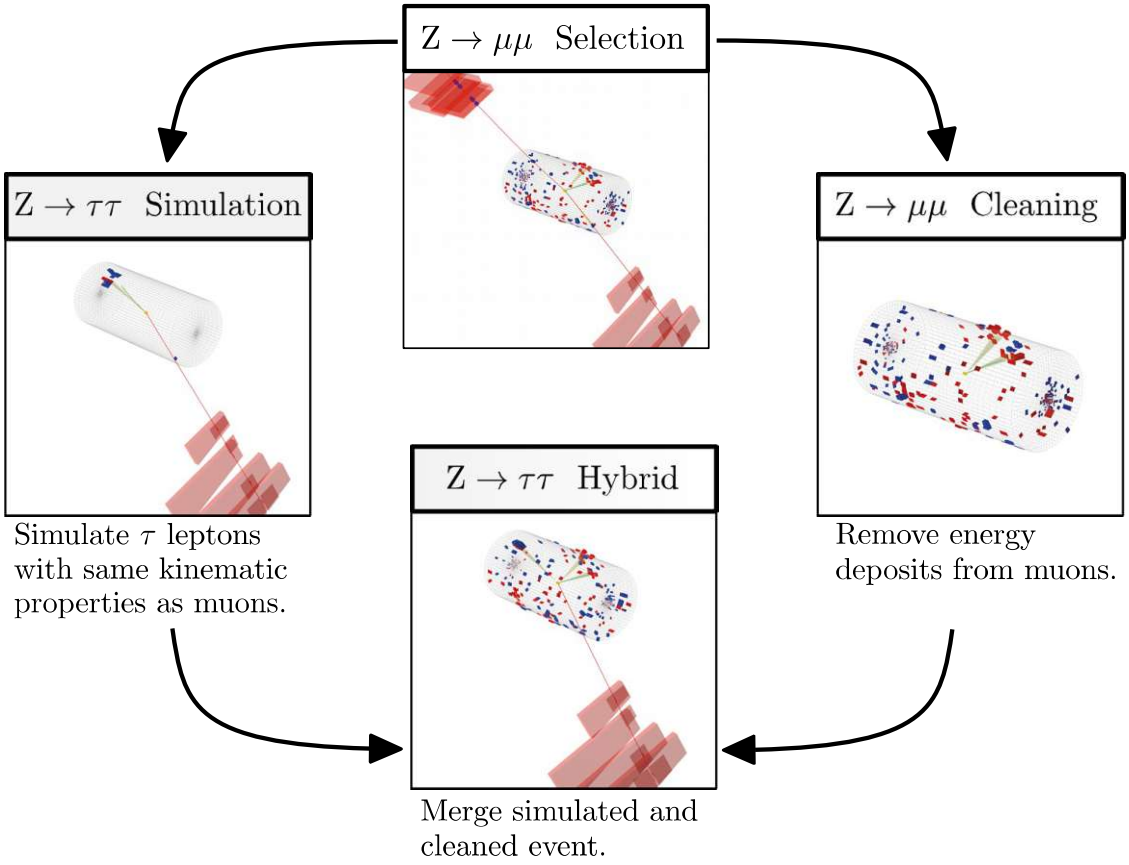


Figure 6.7: An illustration of the τ-embedding technique is shown. Starting from di-muon collision events, tracks and energy deposits of the di-muon pair are removed and replaced with the energy deposits of a simulated di-tau system. The result is a hybrid event between collision data and MC simulation. The figure is taken from [200].

2. In a step called $Z \to \mu\mu$ *cleaning*, the tracks and energy deposits in ECAL, HCAL and muon system are removed for each of the muons of the muon pair.

3. A tau pair is simulated in an otherwise empty detector with the same kinematics as the cleaned muon pair, taking into account the different masses of muons and tau leptons.

4. In the last step, the simulated tau event is added back[5] to the cleaned collision event, creating a hybrid event.

Clear advantages of the τ-embedding approach over pure MC simulation techniques are the better description of pileup, underlying event, the system of additional jets and MET. These quantities are coming directly from real collision data and no tuning and/or systematic uncertainties need to be assigned as it is the case for simulated events.

[5]Ideally, the merging step would be performed on the level of individual tracker hits and energy deposits inside the calorimeters, before executing the full reconstruction of the event. Differences in the geometry of the detector during simulation and data taking, complicate this approach. Thus, the merging is performed on the level of individual tracks and energy clusters. The triggering of the event is evaluated on the simulated tau pair. Differences in triggering efficiency and tau identification by following this approach are studied and corrected for as detailed in [200].

6.5 Corrections

Certain run conditions or problems with the detector are only known or understood after data have been recorded. In most cases, simulated samples have already been generated by this time. One such example are the pileup conditions but there are also other properties which are different between MC simulation and recorded data. In order to improve the data-to-simulation agreement, corrections are applied. Examples have already been discussed at the beginning of Section 6.1, i.e. electron scale and resolution, and τ_{h} energy scale. Since the τ decay in τ-embedded samples is simulated (see Figure 6.7), also these kind of samples are subject to a reduced set of corrections which further improve the data modeling. The different corrections employed are briefly discussed in the following.

6.5.1 Electron, Muon and τ_{h} Efficiency Scale Factors

Reconstruction, identification, isolation and trigger efficiencies are different in data and simulation because the detector response is not perfectly modeled by the MC simulation. Scale factors, defined as ratios between efficiencies in data and simulation, are measured and applied in the analysis to achieve a better data modeling. A generic expression of the efficiency to select a lepton ($l \in \{\mathrm{e}, \mu, \tau\}$) can be decomposed as

$$\epsilon_l = \epsilon_l(\mathrm{reco}) \cdot \epsilon_l(\mathrm{id} \,|\, \mathrm{reco}) \cdot \epsilon_l(\mathrm{iso} \,|\, \mathrm{id}) \cdot \epsilon_l(\mathrm{trg} \,|\, \mathrm{iso}) \ . \tag{6.3}$$

In the above equation, the reconstruction efficiency is denoted as $\epsilon_l(\mathrm{reco})$. Next, the identification efficiency ($\epsilon_l(\mathrm{id} \,|\, \mathrm{reco})$) is measured, given the successful reconstruction of the leptons. The isolation efficiency ($\epsilon_l(\mathrm{iso} \,|\, \mathrm{id})$) then depends on the selected identification condition. Lastly, the trigger efficiency ($\epsilon_l(\mathrm{trigger} \,|\, \mathrm{iso})$) is determined, dependent on the given identification condition and isolation requirement chosen in the measurement of the respective efficiencies. A common approach to measure these efficiencies is the *tag-and-probe* method (see e.g. [201]). The tag-and-probe method uses leptons from Z boson decays. A lepton is referred to as being tagged, if it passes the selections summarized in Table 6.3. A probe lepton, with opposite charge to the tag lepton and forming with the tag lepton a di-lepton pair consistent with originating from a Z boson decay, is then tested to satisfy the reconstruction, identification, isolation and trigger requirements.

For electrons, an example of the four different efficiency components from Equation (6.3) is shown in Figure 6.8. Efficiencies are measured for recorded data, simulation and τ-embedded samples as a function of transverse momentum in different pseudorapidity ranges. Ratios of data to simulation or τ-embedded samples define scale factors which are applied to simulated leptons selected in the analysis on an event-by-event basis.

For τ_{h} candidates, the efficiency is decomposed as

$$\epsilon_{\tau_{\mathrm{h}}} = \epsilon_{\tau_{\mathrm{h}}}(\mathrm{id}) \cdot \epsilon_{\tau_{\mathrm{h}}}(\mathrm{trg} \,|\, \mathrm{id}) \ . \tag{6.4}$$

The identification efficiency ($\epsilon_{\tau_{\mathrm{h}}}(\mathrm{id})$) depends on the criteria used for the discrimination against electrons, muons and jets. Hence, for the measurement of $\epsilon_{\tau_{\mathrm{h}}}(\mathrm{id})$, the same DeepTau selections are applied as used in the analysis (see Table 6.3). Additionally, independent corrections are calculated for electron and muon misidentification as τ_{h} like in case of the τ_{h} energy scale (see Section 6.1). Dedicated muon misidentification scale factors are derived and applied to $\mathrm{Z} \to \ell\ell$ ($\ell \in \{\mathrm{e}, \mu\}$) events in the $\mu\tau_{\mathrm{h}}$ channel. Similarly, the electron misidentification scale factors are measured and applied to $\mathrm{Z} \to \ell\ell$ events in the $\mathrm{e}\tau_{\mathrm{h}}$ final state.

For τ-embedded events, additional corrections of track-reconstruction efficiencies of τ_{h} candidates are employed. The track reconstruction is performed in an otherwise empty

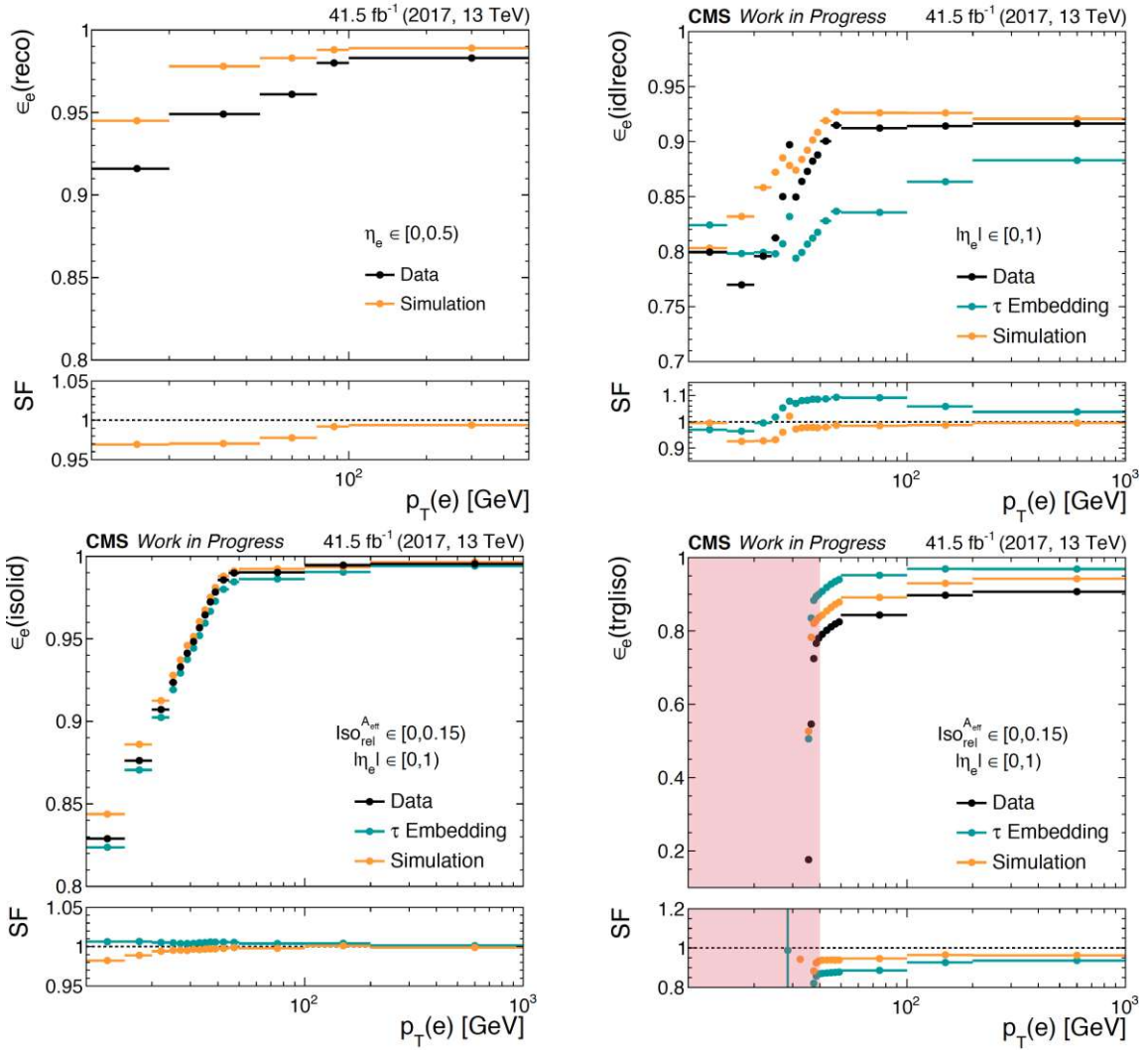


Figure 6.8: Electron efficiency measurements for all four components of Equation (6.3) are shown. All measurements use data, simulation and τ-embedded samples from the 2017 data-taking period. Efficiencies are determined as a function of the electron's transverse momentum ($p_{\mathrm{T}}(\mathrm{e})$) in different pseudorapidity ranges of the electron (η_{e}) as indicated in each plot. The reconstruction efficiency (top left) is shown for central electrons ($\eta_{\mathrm{e}} \in [0, 0.5)$). Top right and bottom row plots show the identification, isolation and trigger efficiencies, respectively. These three efficiencies are calculated inside $\eta_{\mathrm{e}} \in [0, 1)$ and the isolation selection is given by $\mathrm{Iso}_{\mathrm{iso}}^{A_{\mathrm{eff}}} \equiv I_{\mathrm{rel}}^{(\mathrm{e})} < 0.15$ (see also Equation (3.17)). The trigger efficiencies (bottom right) are determined for the e(35) trigger (see Table 6.2) and are applied if events are selected by this particular trigger. The red area marks the excluded area of the trigger turn-on curve. Ratios of data distributions to simulation or τ-embedded samples serve as scale factors (SF) used to correct simulation or τ-embedded events, respectively. All figures are adopted from [10].

Decay mode (D_{mode})	τ_{h} energy scale in simulation (τ-embedded) [%] 2016	2017	2018
0	$-1.0^{+0.7}_{-0.6}$ $(-0.2^{+0.5}_{-0.5})$	$+0.7^{+1.0}_{-0.6}$ $(+0.0^{+0.4}_{-0.4})$	$-1.6^{+0.7}_{-0.7}$ $(-0.3^{+0.4}_{-0.4})$
$\{1, 2\}$	$-0.1^{+0.4}_{-0.3}$ $(-0.2^{+0.2}_{-0.3})$	$+0.2^{+0.5}_{-0.4}$ $(-1.2^{+0.5}_{-0.2})$	$-0.3^{+0.4}_{-0.4}$ $(-0.6^{+0.4}_{-0.3})$
10	$+0.8^{+0.7}_{-0.4}$ $(-1.3^{+0.3}_{-0.5})$	$+0.2^{+0.5}_{-0.5}$ $(-0.8^{+0.4}_{-0.5})$	$-1.1^{+0.5}_{-0.5}$ $(-0.7^{+0.3}_{-0.3})$
11	$+0.1^{+1.0}_{-1.0}$ $(-1.3^{+0.3}_{-0.5})$	$-0.5^{+1.6}_{-1.0}$ $(-0.8^{+0.4}_{-0.5})$	$+0.1^{+1.1}_{-0.9}$ $(-0.7^{+0.3}_{-0.3})$

Table 6.9: Decay mode dependent energy scales for genuine τ_{h} are quoted in this table for all three years of data taking. Values outside parentheses are applied to simulated events, those inside parentheses to τ-embedded events.

detector for the τ-embedding technique (see Figure 6.7). Remnants from the cleaned muon pair might lead to migrations of the assigned decay mode of the reconstructed τ_{h} candidate. These effects are covered with dedicated correction scale factors which are measured as a function of the number of charged hadrons and the number of neutral pions [10].

6.5.2 Electron and τ_{h} Energy Scale - Revisited

The electron and τ_{h} energy scale is not only corrected for simulated samples as discussed at the beginning of Section 6.1, but also for τ-embedded samples [10]. However, electron resolution effects are neglected for τ-embedded events. Table 6.9 quotes the τ_{h} energy scales and their uncertainties for both simulation and τ-embedded samples.

6.5.3 Jet Energy and Resolution Corrections

Before any corrections are applied to jets, some of the jets need to be rejected for the 2017 data-taking period due to a problem with ECAL crystals impacted by radiation damage. Explicitly, jets with

$$\begin{aligned} p_{\text{T}}^{(\text{jet})} &< 50\,\text{GeV} \\ |\eta^{(\text{jet})}| &\in [2.65, 3.139] \end{aligned} \tag{6.5}$$

are removed. All other jets are corrected for differences in jet energy and resolution between data and simulation. Corrections involve several steps as detailed in [202]. It is worth mentioning, that τ-embedded samples are not subject to these corrections because all information about jets is taken directly from data.

6.5.4 b-tagging Efficiency

Reconstructed jets present in events with selected tau pairs have to satisfy

$$\begin{aligned} p_{\text{T}}^{(\text{jet})} &> 30\,\text{GeV} \\ |\eta^{(\text{jet})}| &< 4.7\ . \end{aligned} \tag{6.6}$$

Furthermore, identification criteria as detailed in [134] are imposed to reduce the amount of misreconstructed jets. In order for a jet to be classified as b-tagged, it has to pass the `medium` threshold of the DeepJet classifier (see also Section 3.5.7). The transverse momentum and pseudorapidity ranges for b-tagged jets differ from those stated in Equation (6.6) and are given by

$$\begin{aligned} p_{\text{T}}^{(\text{b-jet})} &> 20\,\text{GeV} \\ |\eta^{(\text{b-jet})}| &< 2.5(2.4)\ , \end{aligned} \tag{6.7}$$

where $|\eta^{(\mathrm{b-jet})}| < 2.4$ only applies to the 2016 data-taking period. A larger range of pseudorapidity is used from 2017 onward because the tracker coverage of the CMS detector was extended after the 2016 data-taking period.

Like for other efficiencies (see e.g. Section 6.5.1), the efficiency of selecting b-tagged jets is different in simulation and data. Hence, a correction needs to be measured and applied. However, the correction is not applied as event-by-event based weight. Instead, individual jets are either *promoted* to be b-tagged jets or b-tagged jets are *demoted* in a probabilistic manner that depends if the scale factor is larger[6] or smaller than one, respectively.

6.5.5 Corrections to Missing Transverse Momentum

Certain events exhibit an artificially high $\boldsymbol{p}_{\mathrm{T,PUPPI}}^{(\mathrm{miss})}$. The reasons include noise in the ECAL or HCAL or bad reconstruction of muons [203]. Based on detector-dependent criteria, such events are filtered out in the analysis.

All corrections discussed in Sections 6.5.3 and 6.5.4 related to jets are propagated to the calculation of $\boldsymbol{p}_{\mathrm{T,PUPPI}}^{(\mathrm{miss})}$ (see Equation (3.27)). In addition, processes such as VV and $\mathrm{t\bar{t}}$, i.e. without a boson resonance recoiling against a jet of the hard interaction, are corrected for effects of unclustered energy due to detector noise. However, single boson resonances, i.e. the signal process ($\mathrm{H} \to \tau\tau$) as well as $\mathrm{Z} \to \ell\ell$ and $\mathrm{W} \to \ell\nu$ processes use dedicated recoil corrections that are propagated to $\boldsymbol{p}_{\mathrm{T,PUPPI}}^{(\mathrm{miss})}$. In the following, the recoil correction is explained in more detail with the help of Figure 6.9.

In $\mathrm{Z} \to \mu\mu$ events, no genuine MET is expected. The transverse momentum of the Z boson ($\mathbf{p}_{\mathrm{T}}^{(\mathrm{Z})}$) can be reconstructed from the transverse momenta of the two muons and is perfectly balanced with the hadronic recoil ($\mathbf{H}_{\mathrm{T}}$)

$$\mathbf{p}_{\mathrm{T}}^{(\mu_1)} + \mathbf{p}_{\mathrm{T}}^{(\mu_2)} \equiv \mathbf{p}_{\mathrm{T}}^{(\mathrm{Z})} = -\mathbf{H}_{\mathrm{T}} \ . \tag{6.8}$$

A distorted momentum balance between reconstructed p_{T} of the Z boson and the hadronic recoil leads to a non-zero MET ($\boldsymbol{p}_{\mathrm{T}}^{(\mathrm{miss})}$)[7]. Such a non-zero MET is referred to as *artificial* MET ($\boldsymbol{p}_{\mathrm{T}}^{(\mathrm{miss,art})}$). The $\mathrm{Z} \to \mu\mu$ control region is thus well suited to calibrate the hadronic recoil by comparing the distributions of $\boldsymbol{p}_{\mathrm{T}}^{(\mathrm{miss,art})}$ in data and simulation. As indicated in Figure 6.9, the $\boldsymbol{p}_{\mathrm{T}}^{(\mathrm{miss,art})}$ vector is decomposed in a parallel and a perpendicular component with respect to $\mathbf{p}_{\mathrm{T}}^{(\mathrm{Z})}$. Distributions of both parallel and perpendicular components are measured for data and simulation and cumulative density functions ($F_{\mathrm{sim},\parallel}, F_{\mathrm{sim},\perp}, F_{\mathrm{data},\parallel}, F_{\mathrm{data},\perp}$) are formed as shown in Figure 6.9.

In a simulated event, genuine MET can be present due to the presence of neutrinos as shown in Figure 6.9. This can, for example, happen in the signal process ($\mathrm{H} \to \tau\tau$), where neutrinos come from the subsequent tau lepton decays. The difference between genuine and reconstructed MET is attributed to artificial MET. Artificial MET is projected on the parallel and perpendicular axes with respect to the direction of the resonance boson (R). These components are corrected using the cumulative densities as shown in Figure 6.9 Afterwards, the $\boldsymbol{p}_{\mathrm{T}}^{(\mathrm{miss})}$ vector is rebuilt using the corrected components of the artificial MET

$$\mathbf{p}_{\mathrm{T}}^{(\mathrm{miss})} = \sum \mathbf{p}_{\mathrm{T}}^{(\nu)} + \mathbf{p}_{\mathrm{T}}^{(\mathrm{miss,fake})}(\mathrm{corrected}) \ , \tag{6.9}$$

where $\sum \mathbf{p}_{\mathrm{T}}^{(\nu)}$ denotes the sum of neutrino momenta in the simulated event. Uncertainties on the recoil correction are assigned by varying the parallel and perpendicular component

[6] In that case, the b-tag efficiency is lower in simulation than in data.

[7] In the discussion of the recoil correction, the extra subscript "PUPPI" is dropped for clarity. The same recoil correction procedure can be applied to PF MET.

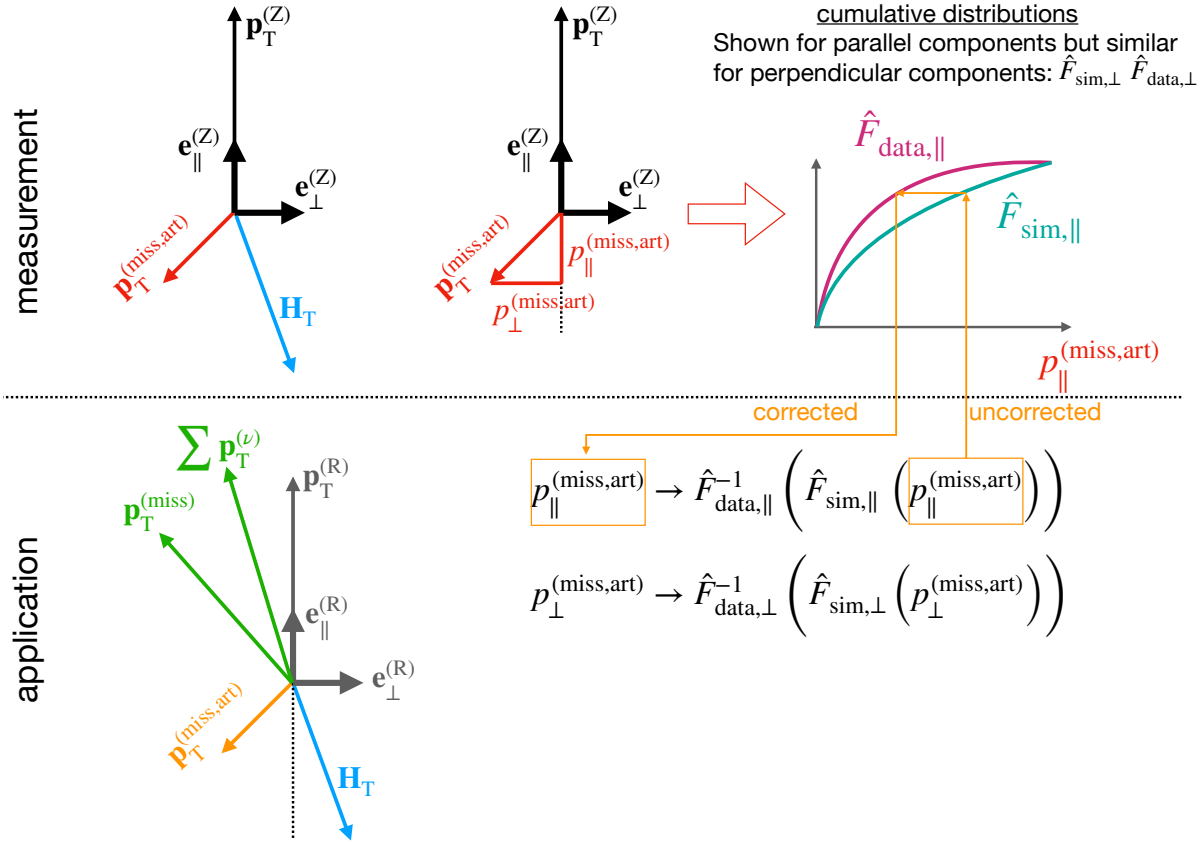


Figure 6.9: An overview of the different quantities and steps involved in the hadronic recoil correction are given in this figure. The top row illustrates the measurement step which is carried out in a $Z \rightarrow \mu\mu$ enriched region and concludes by determining cumulative probability densities, $\hat{F}$, for both parallel and perpendicular components of $\boldsymbol{p}_{\mathrm{T}}^{(\mathrm{miss,art})}$ (shown in red) in data as well as in simulation. On the bottom row, contributions of $\boldsymbol{p}_{\mathrm{T}}^{(\mathrm{miss,art})}$ in simulated events (shown in orange) are corrected using a quantile mapping procedure. Formulas of the quantile mapping are also given in the bottom row. Note, that for a general resonance (R), also genuine MET contributions due to neutrinos ($\sum \mathbf{p}_{\mathrm{T}}^{(\nu)}$) can be present in the event and need to be taken into account. More details are given in the main text.

of the hadronic recoil ($\mathbf{H}_\mathrm{T}$) with respect to the direction of $\mathbf{p}_\mathrm{T}^{(\mathrm{R})}$ and propagating the effect to $\boldsymbol{p}_\mathrm{T}^{(\mathrm{miss})}$ [10]. As concluding remark, it is worth mentioning that recoil corrections do not apply to processes estimated by data-driven techniques such as the τ-embedding and FF method.

6.5.6 Kinematic Reweighting

Inside a $\mathrm{Z} \to \mu\mu$ control region, similar to the one used for the recoil correction (see Section 6.5.5), the two-dimensional distribution of Z boson transverse momentum and invariant di-muon mass is measured in data and simulation. The ratio of the two defines the scale factors applied to $\mathrm{Z} \to \ell\ell$ events in the analysis, improving the overall data-to-simulation agreement.

The simulated p_T-spectrum of the top quark falls steeper than observed in data [204, 205]. Dedicated weights are thus applied to simulated $\mathrm{t\bar{t}}$ events in the analysis.

6.5.7 Trigger Inefficiencies

During the 2016 and 2017 data-taking years, the aging of ECAL crystals, mostly from the endcap region, was not properly taken into account in the trigger. This led to timing issues of the readout. As a consequence, some triggered events were wrongly assigned to the previous bunch crossing. The actual event of interest was then blocked because after triggering on an event, the trigger logic blocks the next two collision events from being recorded. This effect was not included during event simulation. Therefore, simulated events are weighted with the probability that such a *trigger-prefiring* has occurred that depends on the event topology. Relevant for SM $\mathrm{H} \to \tau\tau$ analysis, is the resulting loss of about 4% of events of Higgs bosons produced via VBF.

6.5.8 Pileup Reweighting

The pileup conditions change throughout a data-taking period. An apparent example is the double-peak structure of the 2017 PU distribution shown in Figure 3.4. Such run-specific changes are not taken into account at the time of event simulation. The simulation is reweighted to obtain the correct profile of true number of vertices, given the measured instantaneous luminosity distribution and total proton-proton inelastic cross section.

6.5.9 Gluon-Gluon Fusion Reweighting

Special weights, reflecting NNLO accuracy in perturbative QCD, are applied to ggH signal samples. These weights depend on p_T^H and the number of jets in the event. This is particularly relevant for the differential measurement within the STXS stage-1.2 scheme (see Figure 5.3), where the signals are also split using p_T^H and the number of additional jets.

Chapter 7

Search for Light Top Squarks

The search for light top squarks, discussed in this thesis, presents a continuous effort to constrain the mass parameters of light top squarks within a compressed mass spectrum. This chapter describes the analysis strategy as developed for previous publications by the CMS Collaboration, the most recent being [206], which serves as a basis for further developments such as presented in this and the next chapters. The selection of events and the definition of physics objects is discussed in Section 7.1. Section 7.2 summarizes the details of the used simulated samples. In Section 7.3, the SRs and control regions (CRs) are defined. The large amount of Run2 data allows defining a finer splitting of regions compared to [206], ultimately increasing the sensitivity of the analysis. At the end of this chapter, the corrections applied to simulated events are reviewed.

7.1 Data, Signal and Background Processes

The signal process depicted in Figure 5.8 ($\widetilde{\mathrm{t}}_1 \to \widetilde{\chi}_1^0 \mathrm{b} f \bar{f}'$) has the following characteristics:

1. Considerable $p_{\mathrm{T}}^{(\mathrm{miss})}$originating from the LSPs ($\widetilde{\chi}_1^0$), which escape detection.
2. Due to the compressed mass scenario ($\Delta m = m_{\tilde{t}_1} - m_{\widetilde{\chi}_1^0} < 80\,\mathrm{GeV} \approx m_{\mathrm{W}}$), the decay products have low transverse momentum.
3. There are b quark initiated jets.

The search is conducted in the single-lepton final state, i.e. one lepton, which can originate from either stop decay, is required. The lepton can be either an electron, muon or a tau lepton which decays to an electron or a muon. This final state has a smaller BR than the all-hadronic decay channel ($f\bar{f}' \leftrightarrow q\bar{q}'$). However, the all-hadronic channel suffers from a lot of background contamination from QCD multijet events, abundantly produced at the LHC (see also Figure 2.4). Furthermore, the final state jets in the all-hadronic channel are low-energetic and thus not reconstructed, ultimately limiting the sensitivity of the all-hadronic channel.

To have a chance of detecting signal events with the CMS detector, it is required that a parton undergoes inital state radiation (ISR) (see also Figure 3.14). In the presence of a high-p_{T} jet from ISR, the stop-pair system recoils against this jet and most of the extra transverse momentum is carried by the heaviest decay products, which are in this case the neutralinos, while the leptons remain having low transverse momenta. Since the neutralinos are not detected, such events result in high values of MET, which can be exploited by the trigger logic to select such potentially interesting events.

The triggers used to select events employ specific $p_{\mathrm{T}}^{(\mathrm{miss})}$and $\mathrm{H_T}$ thresholds, where $\mathrm{H_T}$ is defined as the scalar sum of all jet transverse momenta

$$\mathrm{H_T} = \sum_{\mathrm{PF\,jet}} \left| \boldsymbol{p}_{\mathrm{T}}^{(\mathrm{jet})} \right| \ . \tag{7.1}$$

Specifically, the trigger logic combines two triggers with $p_{\mathrm{T}}^{(\mathrm{miss})}$(MET) and $\mathrm{H_T}$ (HT) thresholds at 110 GeV and 120 GeV

$$\mathrm{MET}(110) \wedge \mathrm{HT}(110) \ \ \vee \ \ \mathrm{MET}(120) \wedge \mathrm{HT}(120) \ . \tag{7.2}$$

Choosing $p_{\mathrm{T}}^{(\mathrm{miss})} > 200\,\mathrm{GeV}$ and $\mathrm{H_T} > 300\,\mathrm{GeV}$, ensures that selected events are in the trigger plateau with an efficiency that is almost independent of the value of MET or $\mathrm{H_T}$, and no events from the trigger turn-on region are selected.

In the following, the object selection is discussed which aims at enhancing signal over SM background processes. An overview of the employed object selection is given in Table 7.1. The notion of *prompt*, as introduced in Section 3.5.1, and *non-prompt* leptons is central to this analysis. Prompt leptons are those that originate from the immediate formation at the PV, like it is the case for the signal process. Non-prompt leptons come mainly from decays of hadrons which originate from interactions at the PV but also from PU. To selectively pick prompt leptons, cuts on the longitudinal (d_{z}) and transverse (d_{xy}) component of the impact parameter are applied as detailed in Table 7.1. Furthermore, electrons and muons emerging from hadron decays are suppressed by imposing cuts on isolation variables. This analysis adopts a combined isolation criteria, termed hybrid isolation (HI), ensuring a more uniform selection efficiency as a function of electron or muon transverse momenta, respectively. For leptons with $p_{\mathrm{T}}^{(\ell)} > 25\,\mathrm{GeV}$, a threshold on the relative isolation ($I_{\mathrm{rel}}^{(\ell)}$) is used, exploiting a possible footprint of the lepton in the isolation cone. For transverse momenta below 25 GeV, however, a drop in lepton selection efficiency is observed using the same threshold on $I_{\mathrm{rel}}^{(\ell)}$ because the amount of energy allowed inside the isolation cone becomes very small. Therefore, for leptons with $p_{\mathrm{T}}^{(\ell)} \leq 25\,\mathrm{GeV}$, a threshold on the absolute isolation ($I^{(\ell)}$) is set. The exact thresholds are given by

$$\begin{aligned} I_{\mathrm{rel}}^{(\ell)} &< 0.2 \ \text{ for } p_{\mathrm{T}}^{(\ell)} > 25\,\mathrm{GeV} \\ I^{(\ell)} &< 5\,\mathrm{GeV} \ \text{ for } p_{\mathrm{T}}^{(\ell)} \leq 25\,\mathrm{GeV} \ . \end{aligned} \tag{7.3}$$

The different isolation variables used for electrons and muons respectively, are presented in Section 3.5.6 (see also Equations (3.15) and (3.17)). The HI ($\mathrm{HI}^{(\ell)}$) is then defined as

$$\mathrm{HI}^{(\ell)} = I_{\mathrm{rel}}^{(\ell)} \cdot \min\left(p_{\mathrm{T}}^{(\ell)}, 25\right) \ . \tag{7.4}$$

In the analysis, a selection of $\mathrm{HI}^{(\ell)} < 5\,\mathrm{GeV}$ is applied, which is equivalent to the conditions listed in Equation (7.3). The purity of the selected leptons is increased by requiring respective ID-requirements as introduced in Section 3.5.6 and detailed in Table 7.1. Kinematic selections on transverse momentum and pseudorapidity of leptons, tau leptons, jets and b-tagged jets are also summarized in Table 7.1

Standard Model processes passing the object selections from Table 7.1 pose a source of background contributions for the search of light top squarks. They can be split into four different categories:

- The largest background contribution is made up by the W+jets process (see Figure 6.2), where the W boson decays leptonically.

physics object				
jet	b-jet	τ	e	μ
	$p_T > 30\,\text{GeV}$	$p_T > 20\,\text{GeV}$	$p_T > 5\,\text{GeV}$	$p_T > 3.5\,\text{GeV}$
	$\lvert\eta\rvert < 2.4$	$\lvert\eta\rvert < 2.3$	$\lvert\eta\rvert < 2.5$	$\lvert\eta\rvert < 2.4$
–	–	–	$d_{xy} < 0.02\,\text{cm}$	
–	–	–	$d_z < 0.1\,\text{cm}$	
–	–	–	$\text{HI}^{(\ell)} < 5\,\text{GeV}$	$(\Delta R = 0.3)$
–	`medium` DeepCSV	`vloose` MVA ID	`veto` ID	`loose` ID

Table 7.1: This table lists the selections applied to physics objects in the search for light top squarks. The hybrid isolation ($\text{HI}^{(\ell)}$) is defined in Equation (7.4) in terms of relative isolation ($I_{\text{rel}}^{(\ell)}$), where $I_{\text{rel}}^{(\ell)}$ is calculated inside an isolation cone of radius $\Delta R = 0.3$ (see also Figure 3.17) as shown in this table.

- Another important background is the $t\bar{t}$ process (see Figure 6.4). Compared to the signal process, the $t\bar{t}$ process shows the most resemblance because of the presence of b quarks in the final state and neutrinos leading to MET.

- A collection of several processes contributing only to smaller extent is referred to as *rare processes*. One of the rare processes is the single top production (see Figure 6.6d). Another contribution is given by top pair production in association with a W, Z or γ boson, collectively denoted as $t\bar{t}X$. Diboson (see Figures 6.6a to 6.6c) and DY (see Figure 6.5a) production also belong to the rare processes.

- The last category is formed by events with a non-prompt or a falsely identified (fake) lepton. Non-prompt leptons originate for example from decays of B hadrons. It can also happen that a prompt lepton originating from a W+jets or $t\bar{t}$ process is lost because of its low momentum or the geometric acceptance of the CMS detector and a non-prompt lepton is selected instead. It should be noted that electrons or muons coming from tau lepton decays are considered as prompt in this analysis. Furthermore, Z+jets processes (see Figure 6.5b), where the Z boson decays into neutrinos and the jet is misreconstructed as a lepton ($Z(\rightarrow \nu\nu)$+jets) contribute to the fake leptons. Processes related to QCD multijet production (see Figure 6.3) enter the stop search whenever jets are misreconstructed as leptons. In QCD multijet events, a mismeasurement of jet energies can lead to considerable fake $p_T^{(\text{miss})}$ and thus mimic the signal process. Section 8.3 presents a way to estimate both non-prompt and fake leptons in a data-driven manner.

All preselection cuts are summarized in Table 7.2 and are discussed next. The leading jet, i.e. the jet with the highest transverse momentum, is taken as a proxy for an ISR jet. Since final state objects originating from the stop decay have low transverse momenta, requiring a large value of the transverse momentum of the ISR jet ($p_T^{(\text{ISR})}$) is in line with selecting genuine ISR jets. To reduce contributions from $t\bar{t}$, events with more than two jets with $p_T > 60\,\text{GeV}$ are rejected. Furthermore, events with two jets with $p_T > 60\,\text{GeV}$ for both jets, the azimuth angle ($\Delta\phi(j_1, j_2)$) between leading and sub-leading jet is required to be below 2.5, reducing the amount of QCD multijet events. Events with tau leptons or additional leptons with $p_T > 20\,\text{GeV}$ are also rejected. Finally, remaining events must include at least one electron or muon according to the requirements listed in Table 7.1.

preselection requirements
$p_{\mathrm{T}}^{(\mathrm{miss})} > 200\,\mathrm{GeV}$ $\quad \mathrm{H_T} > 300\,\mathrm{GeV}$ $\quad p_{\mathrm{T}}^{(\mathrm{ISR})} > 100\,\mathrm{GeV}$
Signal lepton: at least one electron or muon according to Table 7.1
3$^{\mathrm{rd}}$ jet veto: no events with $p_{\mathrm{T}}^{(\mathrm{jet}_3)} > 60\,\mathrm{GeV}$
anti-QCD cut : $\Delta\phi(j_1, j_2) < 2.5$
Tau veto: no events with hadronically decaying tau leptons with $p_{\mathrm{T}}^{(\tau)} > 20\,\mathrm{GeV}$
2$^{\mathrm{nd}}$ lepton veto: no events with $p_{\mathrm{T}}^{(\ell_2)} > 20\,\mathrm{GeV}$

Table 7.2: This table summarizes the preselection cuts applied in the search for light top squarks. Selections reduce the amount of background contamination, while retaining as much of the signal process as possible.

7.2 Simulated Samples

The main background processes, W+jets and $\mathrm{t\bar{t}}$, are simulated at LO with the MG5 aMC@NLO [109] MC event generator. Furthermore, with exception of the single top process, all other processes are also simulated with MG5 aMC@NLO using LO or NLO precision. Single top processes are simulated at NLO using POWHEG [110]. All samples use PYTHIA [113] for event hadronization and parton showering together with the CUETP8M1 [207] tune to model the underlying event. Interactions of simulated events with the detector are modeled with GEANT4.

The signal process is simulated like the main backgrounds, i.e. using MG5 aMC@NLO. The top squark decay is modeled directly in PYTHIA. A grid of *mass points* is simulated, starting with $m_{\tilde{t}_1} = 250\,\mathrm{GeV}$ and increasing in steps of 25 GeV up to $m_{\tilde{t}_1} = 800\,\mathrm{GeV}$. For each stop mass, different neutralino masses corresponding to $10\,\mathrm{GeV} \leq \Delta m \leq 80\,\mathrm{GeV}$ and separated in steps of 10 GeV, define a particular mass point. This makes in total 184 different mass points. Since the full detector simulation is the most time consuming step in the event simulation and many mass points need to be simulated, a faster approach, termed FastSim [208], is used instead[1]. Additionally, simulated signal events that are very likely not entering the analysis are filtered out. The filter is applied at the point of generation to save further computation time. Only events with $p_{\mathrm{T}}^{(\mathrm{miss})} > 80\,\mathrm{GeV}$ and $\mathrm{H_T} > 160\,\mathrm{GeV}$ are kept for further simulation steps. Depending on the mass point, the efficiency of this filter varies and needs to be taken into account in the analysis [15].

7.3 Signal and Control Regions

All selections that define SRs and CRs are applied on top of the preselection cuts summarized in Table 7.2. The definition of the signal regions aims to retain sensitivity over different Δm signal points. For small values of Δm, the final state particles exhibit very low transverse momenta. Especially b quark initiated jets have usually too low transverse momenta to pass the threshold for well reconstructed jets. Hence, in a first signal region, labeled with "1" in Figure 7.1, only events without b-tagged jets ($N_{\text{b-jet}} = 0$) are selected. This requirement reduces also the amount of $\mathrm{t\bar{t}}$ background events entering $\mathrm{SR_1}$, and consequently W+jets becomes the dominant background in this region. To reduce W+jets contributions as much as possible, leptons are restricted to the pseudo-rapidity range $|\eta^{(\ell)}| < 1.5$ and the selection of $\mathrm{H_T}$ is tightened to $\mathrm{H_T} > 400\,\mathrm{GeV}$ as well

[1] Some mass points are simulated using the the full detector model. These simulated samples are used to calibrate the FastSim performance.

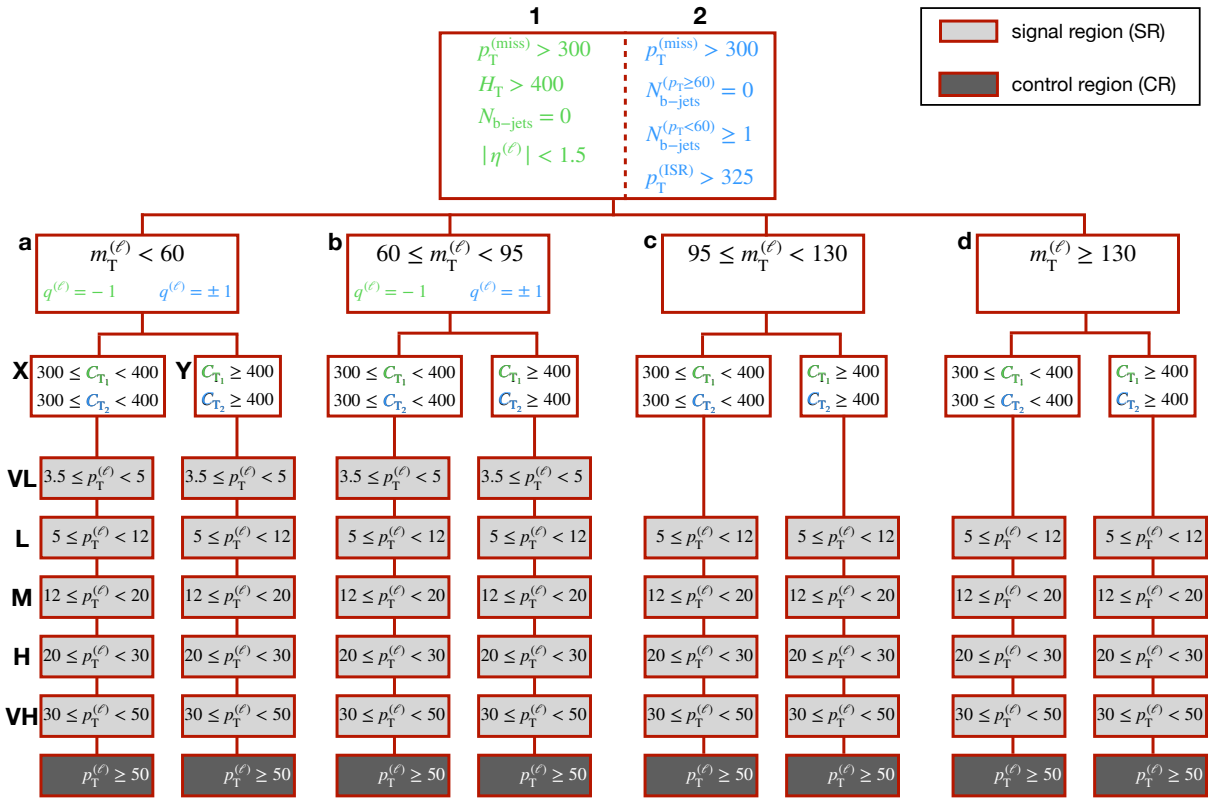


Figure 7.1: This graphic summarizes the splitting in SRs and CRs used in the search for light top squarks. Selections defining region "1" are highlighted in green and are dominated by W+jets events as can also be seen from Figure 7.2. The main backgrounds of region "2" consists of both $t\bar{t}$ and W+jets events and the specific selection cuts are emphasized in blue. Each of the two regions is split in four regions using different $m_T^{(\ell)}$ ranges, labeled "a" to "d". Each of those regions is split in two regions, applying a selection on C_{T_1} for region "1" and on C_{T_2} for region "2", respectively. The final splitting targets different $p_T^{(\ell)}$ ranges, where the low $p_T^{(\ell)}$ ranges serve as SRs ("VL","L","M","H","VH") and the last bin being the CR.

as $p_T^{(miss)} > 300\,GeV$. All mentioned selections are highlighted in green in the top box of Figure 7.1. Signal points with larger mass splitting result in harder p_T spectra of b quark initiated jets which are aggregated in a second signal region, labeled with "2" in Figure 7.1. Thus, in region "2" at least one *soft* ($p_T < 60\,GeV$) b-tagged jet is required. However, no *hard* ($p_T \geq 60\,GeV$) b-tagged jets are allowed in the event to keep the contribution from $t\bar{t}$ at a minimum. Both, $t\bar{t}$ and W+jets events form the dominant backgrounds in region "2". Contributions from $t\bar{t}$ can be reduced by tightening the cut on $p_T^{(ISR)}$ ($p_T^{(ISR)} > 325\,GeV$). All defining cuts of region "2" are highlighted in blue in the top box of Figure 7.1.

Each region is further split in four regions defined by different ranges of transverse mass ($m_T^{(\ell)}$). The definition of the transverse mass is given in Equation (6.2). In this analysis, however, MET is calculated according to Equation (3.26), i.e. contributions from PU are dealt with charged-hadron-subtraction and not the PUPPI approach. Different $m_T^{(\ell)}$ regions are labeled from "a" to "d" as shown in Figure 7.1. The choice of this variable is motivated by the discriminating power between the major background processes of this analysis, i.e. the $m_T^{(\ell)}$-distribution of W+jets events peaks at the W boson mass. Hence, W+jets events can be mainly found in the first two $m_T^{(\ell)}$-regions, "a" and "b". In this two regions the charge asymmetry of W+jets production [209][2] is exploited and only

[2] In proton-proton collisions at the LHC, more W^+ bosons than W^- bosons are produced because two

leptons with negative charge are selected (see also Figure 7.1), reducing the contribution of W+jets events. Additionally, signal processes populate different $m_{\mathrm{T}}^{(\ell)}$-regions, depending on the mass splitting, Δm. Mass points with low Δm enter the regions with lower $m_{\mathrm{T}}^{(\ell)}$, whereas mass points with large Δm are rather located in bins with large $m_{\mathrm{T}}^{(\ell)}$. Thus, splitting in different $m_{\mathrm{T}}^{(\ell)}$-regions aims at keeping the sensitivity to different mass points.

The two conditions

$$\begin{aligned} p_{\mathrm{T}}^{(\mathrm{miss})} &> 300\,\mathrm{GeV} \\ \mathrm{H_T} &> 400\,\mathrm{GeV} \end{aligned} \tag{7.5}$$

applied in region "1" can be combined to a single selection cut by defining a new variable,

$$\mathrm{C_{T_1}} = \min\left(p_{\mathrm{T}}^{(\mathrm{miss})}, \mathrm{H_T} - 100\right) \ . \tag{7.6}$$

Thus, Equation (7.5) can be compactly re-written as $\mathrm{C_{T_1}} > 300\,\mathrm{GeV}$. Similarly for region "2", the condition

$$\begin{aligned} p_{\mathrm{T}}^{(\mathrm{miss})} &> 300\,\mathrm{GeV} \\ p_{\mathrm{T}}^{(\mathrm{ISR})} &> 325\,\mathrm{GeV}\ , \end{aligned} \tag{7.7}$$

can be applied by requiring $\mathrm{C_{T_2}} > 300\,\mathrm{GeV}$, where

$$\mathrm{C_{T_2}} = \min\left(p_{\mathrm{T}}^{(\mathrm{miss})}, p_{\mathrm{T}}^{(\mathrm{ISR})} - 25\right) \ . \tag{7.8}$$

Each $m_{\mathrm{T}}^{(\ell)}$-region is further split in two regions, labeled with "X" and "Y", respectively. The region "X" refers to $300 \leq \mathrm{C_{T_1}} < 400$, whereas for "Y", $\mathrm{C_{T_1}} \geq 400$ is applied for region "1". The splitting in "X" and "Y" for region "2" works in complete analogy by replacing $\mathrm{C_{T_1}}$ with $\mathrm{C_{T_2}}$. Leptons from the signal process are expected to have low transverse momenta and therefore a final splitting according to $p_{\mathrm{T}}^{(\ell)}$ is performed. The different $p_{\mathrm{T}}^{(\ell)}$-regions that serve as SRs are labeled as "VL", "L", "M", "H" and "VH", respectively. They stand for very low, low, medium, high and very high, respectively. Only the $m_{\mathrm{T}}^{(\ell)}$-regions "a" and "b" enable the introduction of a VL-region, whereas regions "c" and "d" have too low event counts in the "VL"-region. The "VL"-region is mainly populated by signal processes with low Δm, which typically also have low $m_{\mathrm{T}}^{(\ell)}$. The "VL"-region is only filled with muons because of the applied selections shown in Table 7.1[3]. The different $p_{\mathrm{T}}^{(\ell)}$ cuts that define the separate regions are detailed in Figure 7.1. Regions with $p_{\mathrm{T}}^{(\ell)} > 50\,\mathrm{GeV}$ are expected to be vastly dominated by background events and serve as CRs as indicated in Figure 7.1. In total there are 72 SRs and 16 CRs. The published result [206] defines the CRs for $p_{\mathrm{T}}^{(\ell)} > 30\,\mathrm{GeV}$ and regions "c" and "d" are merged. A finer splitting, as presented in this section, is possible due to the large amount of available Run2 collision data.

Figure 7.2 shows the yields of the different background processes inside the different regions. In addition, three different signal mass points are overlaid. It can be seen, how signals with larger Δm values extend to higher $p_{\mathrm{T}}^{(\ell)}$ bins. In Section 8.3, it will be discussed how misidentified and non-prompt leptons are estimated from data. From Figure 7.2, it can be already inferred, that misidentified leptons from QCD multijet events are mainly expected in CRs. Another source of misidentified leptons is located in low $p_{\mathrm{T}}^{(\ell)}$ bins which can have significant contributions of events from $\mathrm{Z}(\to \nu\nu)$+jets. These bins are most sensitive for the signal as can be seen from Figure 7.2.

of the valence quarks have a positive charge.

[3] Electrons with transverse momenta below 5 GeV are excluded because their reconstruction efficiency is too low.

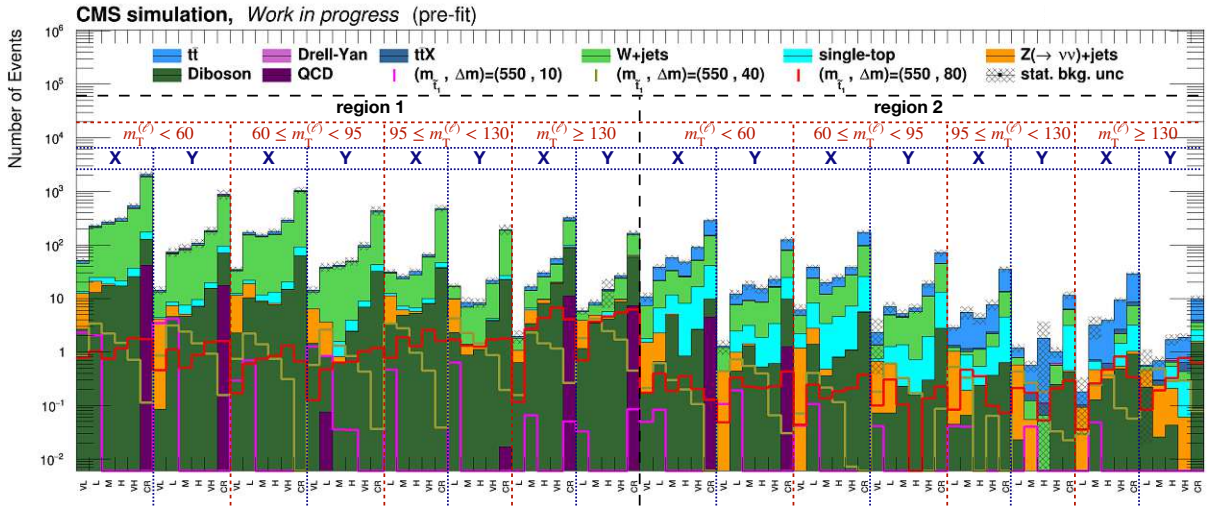


Figure 7.2: The background composition in the different regions, as defined in Figure 7.1, are shown. All backgrounds are estimated by means of MC simulation as described in Section 7.2. Three signal points with the same stop mass and different values of Δm are shown as well. The hatched area represents the statistical uncertainty on the total background yield.

7.4 Corrections to MC Simulation

As outlined in Section 6.5 for analyses with di-tau final states, also in the search for light top squarks, corrections are applied to improve the modeling of data by MC simulation. The PU reweighting is applied as described in Section 6.5.8. More specific to this analysis, quantities related to ISR receive a dedicated correction. Corrections are derived that match the ISR multiplicities in simulated $t\bar{t}$ events to the ones observed in data. These corrections are then applied to the signal, following the procedure in [210]. In general, simulated event yields are scaled down between 10 to 50%, depending on the number of ISR jets. For W+jets events, the $p_{\mathrm{T}}^{(\mathrm{ISR})}$ spectrum is being reweighted as described in [211]. Correction factors for $p_{\mathrm{T}}^{(\mathrm{ISR})} \approx 100\,\mathrm{GeV}$ are 1.18 and range down to 0.78 for $p_{\mathrm{T}}^{(\mathrm{ISR})} > 600\,\mathrm{GeV}$. Further important corrections of this analysis are selection efficiencies for leptons and b quark initiated jets, which need to cover the low momenta of these objects. How these corrections are obtained is detailed in [15].

Chapter 8

Data-Driven Background Estimation

> You don't really understand something unless you can explain it to your grandmother.
>
> Albert Einstein, 1879 to 1955

This chapter describes my main contribution to the analyses presented in this thesis. It concerns the estimation of contributions from quark- or gluon-initiated jets misidentified as leptons directly from data, using the so-called fake factor (FF) method. In case of analyses with a di-tau final state, the method models contributions from quark- or gluon-initiated jets misidentified as τ_{h}. A thorough discussion of the method as well as the associated uncertainties is given throughout Section 8.2. In Section 8.3 the focus is shifted on modeling the contributions from quark- or gluon-initiated jets that are misidentified as electrons or muons. This background enters the search region for light top squarks presented in Section 7.3.

No matter how nested the implementation of the FF method becomes or to which analysis it is applied to, it is based on well defined elementary working principles. To illustrate those, Section 8.1 serves as a gentle introduction to the FF method. Also, the necessary terminology used all through this chapter is put in place in Section 8.1.

8.1 An Illustrative Example

In this section, a short example is presented illustrating the basic working principle of the FF method. A publicly available dataset [212] is used which contains the information of hair and eye color of certain individuals. The exact numbers are not relevant and the example is chosen to have a simple starting point to explain the methodology. Let us assume we want to infer the number of individuals with brown hair and green eyes within a given group of people. In the dataset [212], 108 individuals have black hair and 286, $N(\text{brown hair})$, have brown hair. Furthermore, it is known that from all individuals with black hair, 5 have green eyes and 20 have blue eyes. The dataset comprises also individuals with other hair and eye colors than those mentioned so far. Lastly, it is known that 30% of individuals with brown hair have blue eyes, i.e. the probability, $p(\text{blue eyes}|\text{brown hair})$, of an individual having blue eyes given that it has brown hair, is 30%. All this information is summarized in Figure 8.1a and it is sufficient, if the two observables i.e. hair color and eye color are uncorrelated, to infer the number of people with brown hair and green eyes. As depicted in Figure 8.1a:

A: denotes number of individuals with green eyes and brown hair,

B: denotes number of individuals with blue eyes and brown hair,

C: denotes number of individuals with green eyes and black hair and

D: denotes number of individuals with blue eyes and black hair.

Let us define the ratio of C over D

$$\rho \equiv \frac{C}{D} = \frac{5}{20} = 0.25 \ . \tag{8.1}$$

Assuming the green-to-blue-eyes ratio (ρ) is the same within the population of individuals with brown hair and among individuals with black hair, the following equality can be written down

$$\frac{A}{B} = \frac{C}{D} \overset{\text{Eq. 8.1}}{\equiv} \rho \ . \tag{8.2}$$

The above equation can be solved for A

$$A = B \cdot \rho = N(\text{brown hair}) \cdot p(\text{blue eyes}|\text{brown hair}) \cdot \rho \ . \tag{8.3}$$

Inserting the known numbers into the above equation yields an estimate of $A = 21$. Putting Equation 8.3 into words: *The estimated number of individuals with green eyes and brown hair is calculated by weighting each individual with brown hair by the product of the probability of this individual having blue eyes and the green-to-blue-eyes ratio, ρ.*

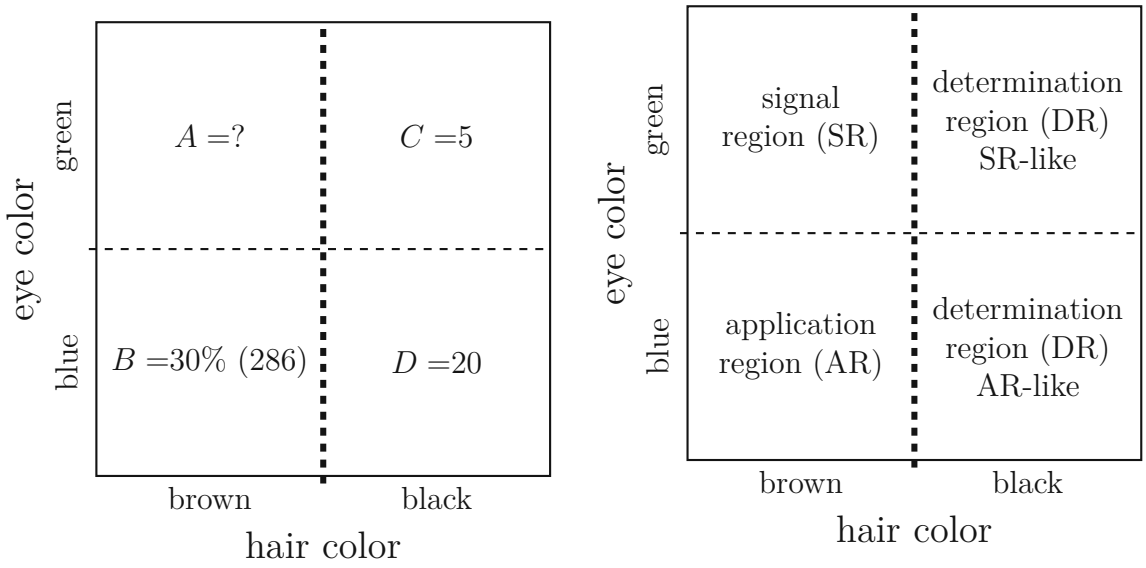
At this point let us introduce some important notation. The region of interest – denoted as A – is called signal region (SR). The region with label B is called application region (AR) because ρ gets applied to entities within this region. The region where ρ is derived is called *measurement region* or determination region (DR). One distinguishes between SR-like and AR-like DR, which correspond to the regions C and D, respectively. The naming of the different regions is summarized in Figure 8.1b.

Going through this introductory example, the following key points of the method should be kept in mind for the upcoming sections:

- The method relies on the universality of ρ, i.e. that it is the same in DR and SR/AR. In other words, hair color and eye color should be uncorrelated.
- The AR and DR must be pure and not contaminated with individuals of other hair or eye color.
- All regions in Figure 8.1b must be non-overlapping.

In short, the FF method in the context of the presented analyses has the following correspondence to this introductory example:

- The SR (A) contains misidentified leptons and using the FF method their contribution shall be estimated.
- A suitable AR, non-overlapping with the SR, needs to be defined which is dominated by misidentified leptons.
- The transfer factor (ρ) has to be measured inside a suitable DR which is pure in misidentified leptons and be non-overlapping with the SR nor the AR. Important for the success of the FF method is the fact, that ρ is measured using collision data and is also applied to recorded collision events inside the AR. As we will see, it is impossible to perfectly select misidentified leptons and there will always be a contamination from other processes entering the AR and the DR. This contamination from other processes is assessed by means of simulation and the goal is to reduce their contributions as much as possible.



(a) A summary of the numbers given in the text. **(b)** The naming of different regions is introduced.

Figure 8.1: The two graphics summarize the basic working principle of the FF method and introduce some relevant notation. Hair and eye color are shown on the x-axis and y-axis, respectively. Knowing the values inside regions B, C and D, the value of A can be inferred as explained in the text.

8.2 Estimation of the Contribution from Misidentified τ_h's in Di-Tau Final States

Searches for the SM Higgs boson or for additional MSSM Higgs bosons that decay into a pair of tau leptons necessitate a reliable method to estimate background processes where quark- or gluon-initiated jets are misidentified as τ_h. In hadron collider experiments, the overwhelming production cross section of QCD multijet events results in a non-negligible population of such events inside Higgs boson search regions. Furthermore, W+jets processes build a dominant source of the jet $\rightarrow \tau_h$ background in the semi-leptonic decay channels. As mentioned in earlier chapters, QCD multijet processes are difficult to model accurately by means of simulation. In addition, reaching higher luminosities in hadron colliders, the simulation of collision events has become more difficult with regard to pileup and underlying event simulation. Searches in extreme regions of phase space or precision measurements can easily lack a good description in terms of simulated events. *Data-driven* methods have been developed to mitigate these shortcomings. These methods aim at describing certain processes directly from recorded collision data and hence quantities like the amount of pileup are described correctly automatically. In addition, various types of corrections and parameter settings needed to calibrate MC simulation with collision data are not needed or at least less relevant in data-driven methods. The τ-embedding technique, presented in Section 6.4, is an example of a data-driven method. Another example of a data-driven technique is the FF method which is discussed in detail in this section.

Historically, the first appearance of the FF method applied to an analysis with a di-tau final state can be dated back to 2005, where it was used by the CDF Collaboration in a search for additional MSSM Higgs bosons [213]. The method was picked up by the CMS Collaboration and refined. The measurement of the $Z/\gamma^* \rightarrow \tau\tau$ cross section [214] and the search for additional MSSM Higgs bosons [187] published by the CMS Collaboration both use the FF method to estimate the jet $\rightarrow \tau_h$ background. In the SM $H \rightarrow \tau\tau$

analysis [215], the FF method is used together with the τ-embedding technique for the first time. The two methods estimate the majority of background contributions directly from data. Furthermore, the SM H $\to \tau\tau$ analysis presented in [215] uses machine learning techniques to categorize events and enhance the H $\to \tau\tau$ signal in dedicated signal categories as described in more detail in Section 9.2. The results presented in [215] are based on partial Run2 data from 2016 and 2017. In this thesis an improved FF method is presented, used for the SM H $\to \tau\tau$ analysis using machine learning techniques with data from 2016–2018 [7]. The main improvements are the following, whereas the details will become clearer in the course of this section:

- The usage of the D_{jet} classifier output to define SR-like and AR-like DRs. This exploits the better discrimination power of the DeepTau classifier compared to previous classifiers (see also Figure 3.21).
- The inclusion of a QCD multijet estimate in the FF determination of the W+jets component.
- The FF parametrization is optimized for the targeted STXS measurements.
- Technical changes for a more realistic FF uncertainty model.

The FF method estimates the contribution of quark- or gluon-initiated jets misidentified as τ_{h} and can thus be applied to the following three di-tau final states (see also Table 6.1):

$e\tau_h$: One tau lepton decays to an electron and the other decays hadronically.

$\mu\tau_h$: One tau lepton decays to a muon and the other decays hadronically.

$\tau_h\tau_h$: Both tau leptons decay hadronically.

Selections defining the SR of the SM H $\to \tau\tau$ analysis are presented in Section 6.1. Figure 6.1 shows the different background contributions present in this SR, split by the different final states listed above and years of data taking. The contribution of the jet $\to \tau_{\text{h}}$ background to the semi-leptonic final states is in the order of 30%. Within these 30%, the most dominant contribution comes from the W+jets process. For the fully hadronic final state, the jet $\to \tau_{\text{h}}$ contribution to the SR is approximately 60%. QCD multijet production is by far the dominant contribution to the jet $\to \tau_{\text{h}}$ background in this case. Fractions of individual background processes contributing to the jet $\to \tau_{\text{h}}$ background are shown in Figure 8.2.

All events inside the SR have a τ_{h} candidate passing the `tight` threshold of the DeepTau discriminant against jets, D_{jet} (see Table 6.3). At the end of Section 3.5.8, a summary of all D_{jet} thresholds is given. In the $\tau_{\text{h}}\tau_{\text{h}}$ channel, both τ_{h} candidates have to pass the `tight` threshold[1]. The output value of D_{jet} lies between zero and one, with values closer to one representing more genuine τ_{h}-like objects. By selecting the `tight` threshold for the definition of the SR, jet $\to \tau_{\text{h}}$ contributions are suppressed. When loosening this requirement, more and more jet $\to \tau_{\text{h}}$ events are acquired. Loosening the requirement on D_{jet} and rejecting events from the SR defines a region orthogonal[2] to the SR which is enriched in jet $\to \tau_{\text{h}}$ events. Figure 8.3 shows the background composition of events where tau pair candidates are required to only pass the `vloose` D_{jet} threshold

[1] Note, that in the $\tau_{\text{h}}\tau_{\text{h}}$ channel either one of the τ_{h} candidates or both can originate from a misidentified jet. Further details on this matter are presented in Sections 8.2.2 and 8.2.3

[2] Two regions are called *orthogonal* if they are non-overlapping.

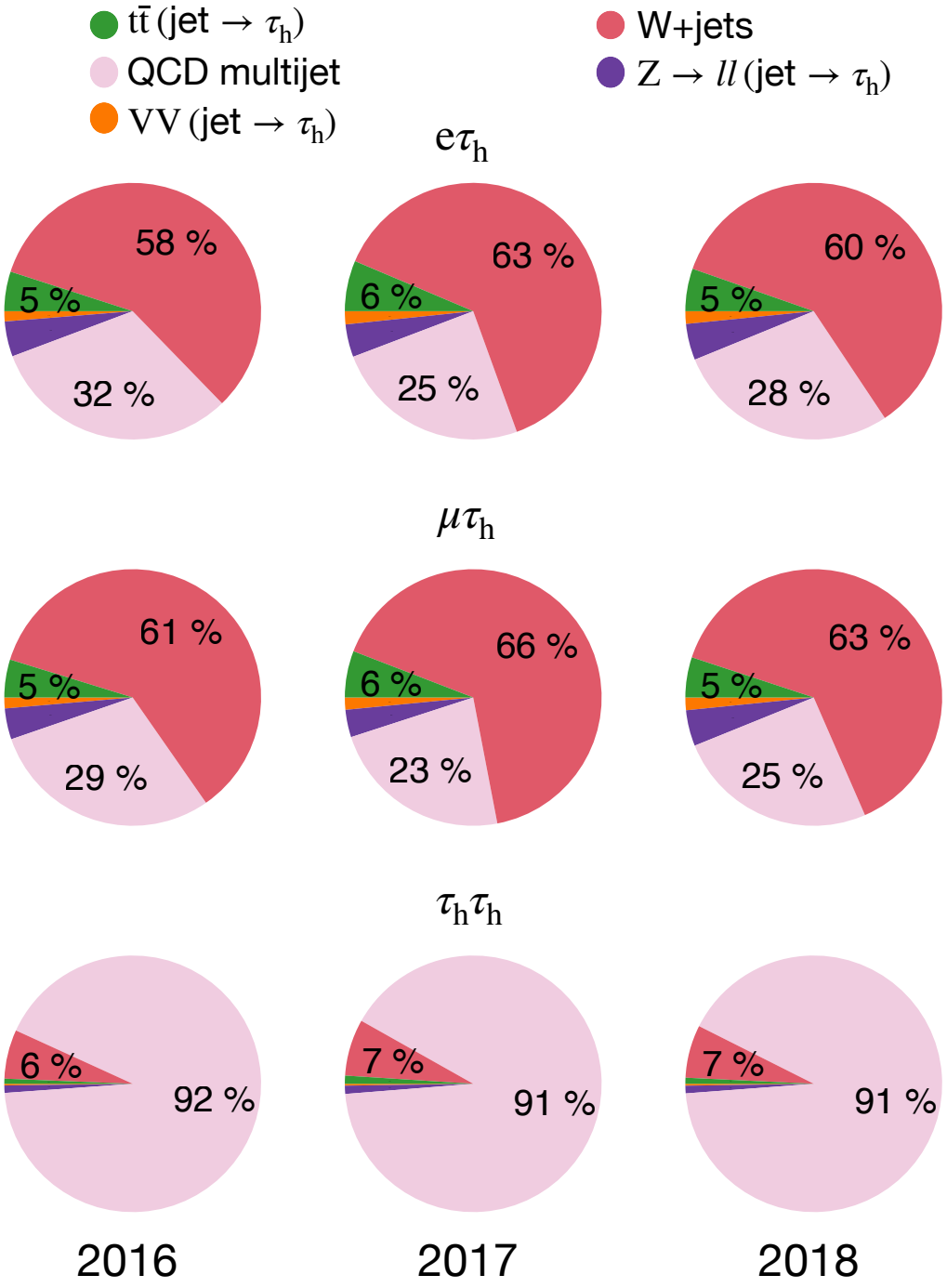


Figure 8.2: Shown are the different expected background contributions that make up the jet → τ_h process, which is highlighted with a solid line in Figure 6.1. All contributions other than QCD multijet are estimated by means of simulation, whereas the difference between these contributions to the observed data yield is attributed to the QCD multijet process. Different backgrounds are labeled in the legend at the top. The compositions do not vary much between the different years of data taking. Furthermore, the compositions of the semi-leptonic channels – $e\tau_h$ and $\mu\tau_h$– are comparable. Only relevant background contributions – being either W+jets, QCD multijet or $t\bar{t}$(jet → τ_h) – are quantified with percentages, except when their contribution is negligible. The contribution of signal events is negligible and not considered in these pie charts.

and fail the `tight` one. In case of the $\tau_h\tau_h$ channel, it is either one of the τ_h candidates failing the `tight` D_{jet} threshold and passing the `vloose` one, but not both. The different treatment of the $\tau_h\tau_h$ channel will become more clear throughout the discussions in Sections 8.2.2 and 8.2.3. Indeed, the jet $\rightarrow \tau_h$ process is dominant in every decay channel shown in Figure 8.3, comprising 80% to over 90% of all selected events, where the percentages are obtained from simulation. The idea of the FF method is to take this region as the AR and extrapolate the contribution of jet $\rightarrow \tau_h$ events from AR to SR by means of a transfer factor. This transfer factor is called *fake rate* or *fake factor* and is denoted as F_F.

Looking back at the introductory example and the illustration in Figure 8.1b, the next step lies in defining a suitable DR enriched in jet $\rightarrow \tau_h$ events, where F_F can be measured. However, this picture is complicated by the fact that different processes make up the jet $\rightarrow \tau_h$ contribution, i.e. QCD multijet, W+jets and $t\bar{t}$ (see also Figure 8.2). Each of these backgrounds have different composition in terms of gluon, light quark or b quark initiated jets which are known to have different fake rates [14]. The way this is solved, is by defining three determination regions (DRs), one for each dominant source of jet $\rightarrow \tau_h$ background. A dedicated F_F is measured inside a QCD enriched, W+jets enriched and $t\bar{t}$ enriched DR and applied as a weighted average to events in the AR. The weights correspond to the *fractions* of QCD multijet, W+jets or $t\bar{t}$ events populating the AR and are calculated using simulation only.

Fake factors depend on kinematic features of the tau pair candidates such as transverse momentum of each τ_h candidate or event properties such as the number of jets. Therefore, F_F's are parametrized in variables reflecting their most dominant dependencies as they are measured within each dedicated DR. In general, F_F's of QCD multijet events are smaller than those for W+jets events but larger than those for $t\bar{t}$ events. Misidentified τ_h's in QCD multijet events are mostly originating from gluon-initiated jets, whereas for W+jets events mostly light quark-initiated jets and for $t\bar{t}$ events heavy quark-initiated jets are misidentified as τ_h. The probability of misidentification is larger for quark-initiated jets than for gluon-initiated jets [14], explaining the observed differences in measured F_F's. Misidentification probabilities among quark-initiated jets are similar, whereas heavy quark-initiated jets with large transverse momenta are more likely to be misidentified than light quark-initiated ones [14]. The main dependence of the F_F is on the transverse momentum of τ_h ($p_T^{(\tau_h)}$), where events with higher $p_T^{(\tau_h)}$ tend to have larger F_F values.

The basic assumption of the FF method is the so-called *universality*, meaning that the F_F measured in a DR has the same dependencies as if measured inside the SR/AR which can be verified using simulated events. This assumption is also highlighted in the introductory example in Section 8.1. However, universality is not strictly given because by measuring F_F in a DR, one introduces biases since for example the jet-flavor composition changes in the DR with respect to the SR/AR. Therefore, the FF method also implements process-dependent bias corrections to partially restore universality within the method's measurement accuracy.

In summary, the F_F is measured in dedicated DRs and it is parametrized in variables reflecting its dominant dependencies. Corrections are derived improving the modeling of the jet $\rightarrow \tau_h$ contribution in variables not used for the initial parametrization. In a last step, bias corrections are calculated, accounting for possible differences between the DR, where the F_F is measured, and the AR, where the F_F is applied. The F_F derivation varies between the semi-leptonic and fully hadronic channels and is therefore discussed separately in Section 8.2.1 and Section 8.2.2, respectively. Section 8.2.3 explains how fractions of each background process inside the AR are derived and details on the F_F application. Finally, the FF uncertainty model is discussed in Section 8.2.4.

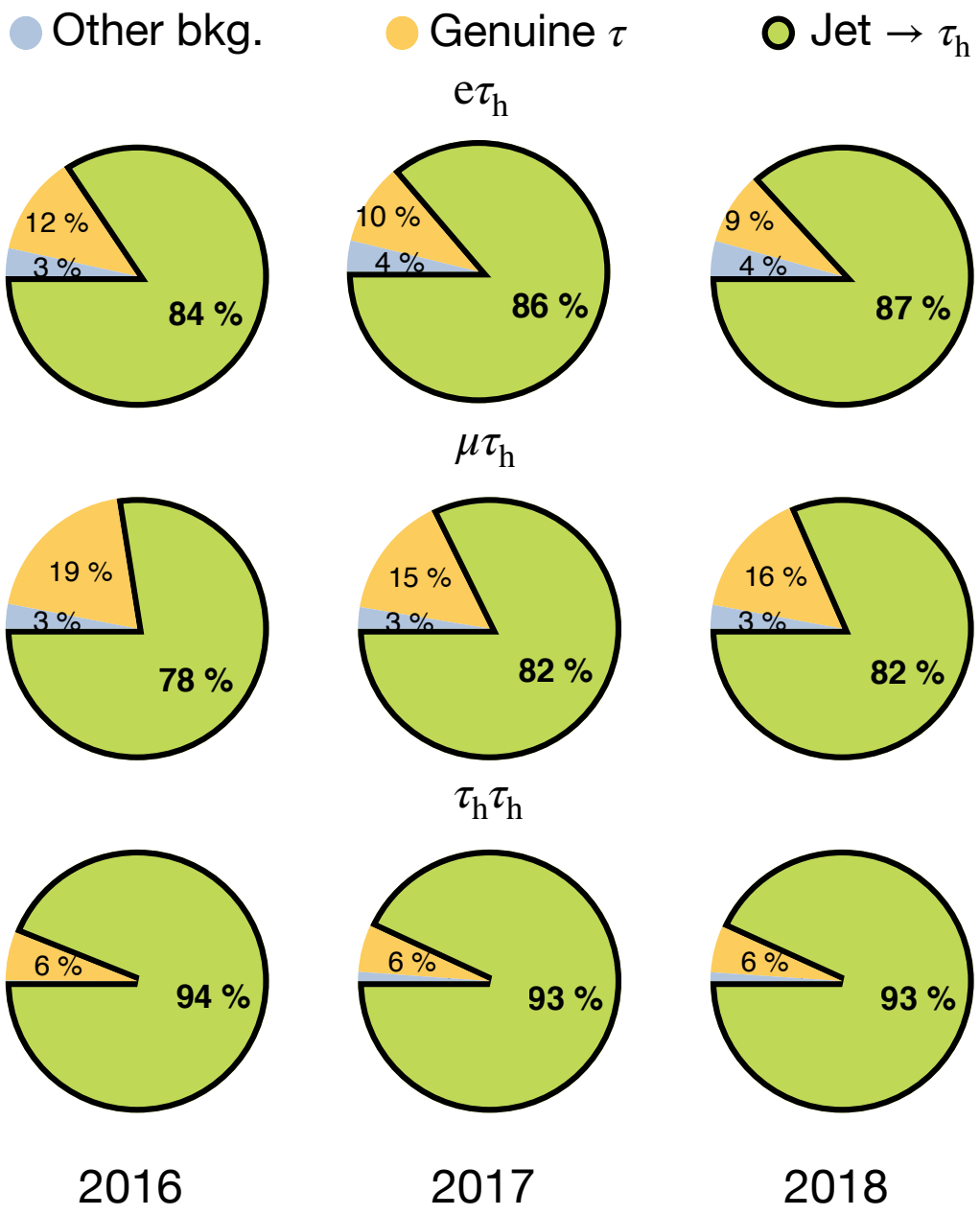


Figure 8.3: Shown are the different expected background contributions which are divided in three categories, genuine τ, jet $\rightarrow \tau_{\mathrm{h}}$ and "other bkg.". The applied selections are the same as for the $\mathrm{H} \rightarrow \tau\tau$ SR, except for the quality requirement on the τ_{h} candidates of the tau pair. Events shown in these pie charts consist of τ_{h} candidates failing the `tight` threshold of the D_{jet} discriminant and passing the `vloose` one. This threshold combination defines the AR of the FF method. The contribution resulting in genuine tau pairs is estimated by the τ-embedding technique, jet $\rightarrow \tau_{\mathrm{h}}$ processes are estimated by the FF method and "other bkg." by means of simulation (see also Figure 6.1).

8.2.1 Fake Factors in the Semi-Leptonic Final States

This section discusses the F_{F} derivation in the semi-leptonic decay channels, $\mathrm{e}\tau_{\mathrm{h}}$ and $\mu\tau_{\mathrm{h}}$. All relevant elements of the FF method are pictorially summarized in Figure 8.4 and will be covered throughout this section. The y-axis shows the D_{jet} output which is used to define the SR and the AR, both highlighted as gray boxes. All τ_{h} candidates entering the SR have to pass the `tight` condition on the D_{jet} discriminant. The AR is defined as events, where τ_{h} candidates pass the `vloose` D_{jet} condition but fail the `tight` one. The fake factor, F_{F}, is defined as

$$F_{\mathrm{F}} = \frac{N^{(\texttt{tight})}}{N^{(\texttt{vloose} \wedge \neg\texttt{tight})}} \ , \tag{8.4}$$

where

- $N^{(\texttt{tight})}$ is the number of events passing the `tight` D_{jet} condition and
- $N^{(\texttt{vloose} \wedge \neg\texttt{tight})}$ is the number of events passing the `vloose` and failing the `tight` D_{jet} condition.

The F_{F} is derived in a DR which is divided in a SR-like part and an AR-like part (see also Figure 8.2). Events inside the DR passing the `tight` D_{jet} condition fall into the SR-like category and those that pass the `vloose` D_{jet} condition but fail the `tight` one are labeled as AR-like. For each dominant component of the jet $\rightarrow \tau_{\mathrm{h}}$ processes – W+jets, QCD multijet and $\mathrm{t\bar{t}}$ – a dedicated DR is defined and shown in different colors in Figure 8.4.

The F_{F} measurement uses a parametrization in terms of the τ_{h} candidate's transverse momentum – denoted as $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$, the jet multiplicity of the event – denoted as N_{jets}, and $\Delta R^{(\ell,\tau_{\mathrm{h}})}$ – the angular distance between the τ_{h} candidate and the light lepton $\ell \in \{\mathrm{e}, \mu\}$ of the tau pair. The F_{F} measurement is indicated by the arrows with label "1" in Figure 8.4 and is repeated in each DR. Corrections to the measured F_{F} are indicated with further arrows labeled "2", "3" and "4", respectively. From looking at Figure 8.4, it becomes clear that the structures "1"-"3"/"4" are repetitive among DRs. In the following the DRs and F_{F} measurements for each of the dominant jet $\rightarrow \tau_{\mathrm{h}}$ processes are presented in a separate section.

The W+jets Fake Factor

The W+jets F_{F} is measured in a W+jets enriched DR – denoted as $\mathrm{DR}_{\mathrm{W+jets}}$– and is highlighted in red in Figure 8.4. The region $\mathrm{DR}_{\mathrm{W+jets}}$ differs from the SR in the following way:

- The transverse mass between the light lepton, $\ell \in \{\mathrm{e}, \mu\}$, and the missing transverse energy is greater than 70 GeV:

$$m_{\mathrm{T,PUPPI}}^{(\ell)} \geq 70\,\mathrm{GeV} \ . \tag{8.5}$$

  For the SR $m_{\mathrm{T,PUPPI}}^{(\ell)} < 70\,\mathrm{GeV}$ is applied (see also Equation (6.2)), thus avoiding an overlap between SR and $\mathrm{DR}_{\mathrm{W+jets}}$. Since the transverse mass distribution for W+jets processes peaks around 80 GeV – the mass of the W boson – the above selection is expected to dominantly select W+jets events.

- The number of b-tagged jets is required to be zero:

$$N_{\text{b-jet}} = 0 \ . \tag{8.6}$$

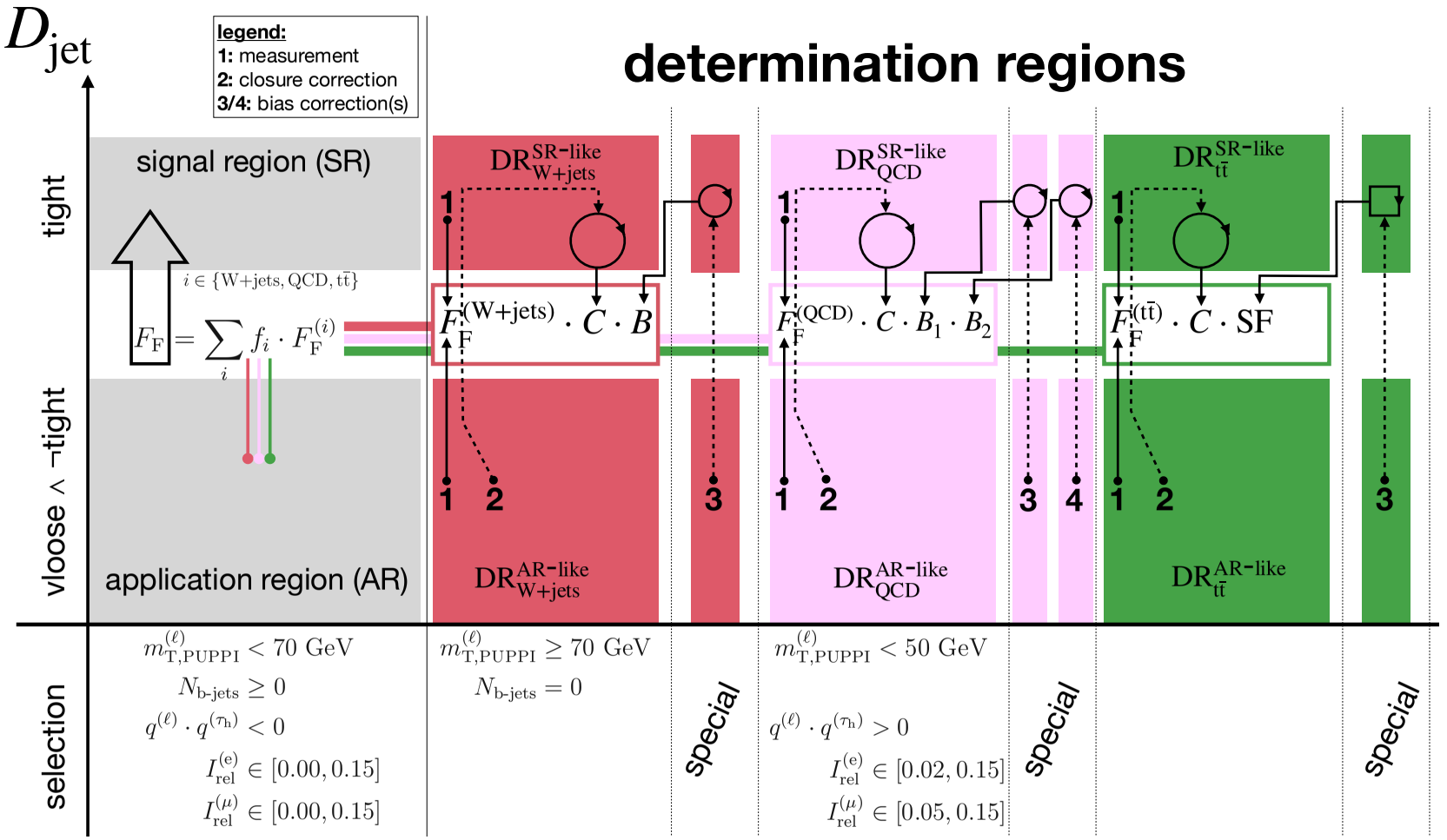


Figure 8.4: Schematic of the FF method for the semi-leptonic channels. The y-axis shows the value of D_{jet} that distinguishes SR(-like) and AR(-like) regions. Both SR and AR are shown on the left in gray boxes. In different colors the different DRs are shown for W+jets, QCD multijet and $\text{t}\bar{\text{t}}$ processes, respectively. The raw F_{F} measurement in each DR is labeled with "1", closure corrections are labeled with "2". Labels "3" and "4" are used to indicate bias corrections or scale factor (SF) correction in case of the F_{F} for the $\text{t}\bar{\text{t}}$ process. At the bottom, the different selections applied with respect to the SR are listed in order to enrich different background types. For bias and scale factor corrections, special regions are used, whose selections are given in the respective sections, where these corrections are discussed. The final F_{F} used to extrapolate events from AR to SR is a weighted sum consisting of fractions f_i and individual fake factors, $F_{\text{F}}^{(i)}$.

In the SR no requirement on the number of b-tagged jets is applied. This cut is used to reject $t\bar{t}$ events which naturally have two b quark initiated jets (see also Table 6.6).

The different selections applied with respect to the SR are also outlined at the bottom of Figure 8.4.

Figure 8.5 shows $p_T^{(\tau_h)}$ distributions inside DR_{W+jets} for the 2018 data-taking period. Note, that the distributions in Figure 8.5 are inclusive in N_{jets} and inclusive in $\Delta R^{(\ell,\tau_h)}$. A good modeling of the sum of backgrounds with respect to the observed data is achieved. As expected, DR_{W+jets} is indeed mostly populated by W+jets events. With the exceptions of two background processes, all other processes in Figure 8.5 are estimated from simulation. The first exception is the genuine tau contribution from $Z/t\bar{t}/VV \to \tau\tau$ processes, which is estimated by the τ-embedding technique (see Section 6.4). The second exception is the QCD multijet background which was neglected in previous versions of the FF method [14]. The QCD multijet background is estimated directly from data inside an altered DR_{W+jets} as follows.

The altered DR_{W+jets} has an additional same-sign (SS) charge requirement

$$q^{(\ell)} \cdot q^{(\tau_h)} > 0 \ , \quad \ell \in \{e, \mu\} \ . \tag{8.7}$$

Inside the SS DR_{W+jets} all contributions from simulation and τ-embedded samples are subtracted from data. The result of this subtraction is assigned to the QCD multijet process, whereby the difference is not allowed to reach negative values. This QCD estimate is transferred to DR_{W+jets} with a transfer factor of one, i.e. assuming that tau pairs from QCD multijet processes do not share a common origin and thus the electric charges of the tau pair are uncorrelated. In fact, this QCD multijet estimation method is a simplified version of the QCD multijet estimation used for the eμ final state of the SM $H \to \tau\tau$ analysis [7, 12, 215].

The W+jets F_F (see also Equation (8.4)) is calculated from the histograms shown in Figure 8.5. However, to incorporate the dependence on N_{jets}, the $p_T^{(\tau_h)}$ distributions are determined inside three N_{jets} categories, $N_{jets} = 0$, $N_{jets} = 1$ and $N_{jets} \geq 2$. This categorization is also motivated to align the FF method with the targeted STXS stage-1.2 measurement, which uses the jet multiplicity to define different STXS bins (see Figures 5.3 and 5.4). Furthermore, the dependence on $\Delta R^{(\ell,\tau_h)}$ is included by defining the following two categories:

- $\Delta R^{(\ell,\tau_h)} < 3$
- $\Delta R^{(\ell,\tau_h)} \geq 3$.

Hence, the $p_T^{(\tau_h)}$ histograms in Figure 8.5 are re-calculated in three N_{jets} categories times two $\Delta R^{(\ell,\tau_h)}$ categories and the FF measurement is executed in each of the six histograms as follows.

Firstly, all contributions other than W+jets are subtracted from data to obtain an estimate of the W+jets contribution directly from data. The sum of all other contributions except W+jets is denoted with

$$\sum_{!W+jets} N_{other} \ . \tag{8.8}$$

As can be seen from Figure 8.5, $\sum_{!W+jets} N_{other}$ is small compared to the observed yield inside DR_{W+jets} and is partially based on MC simulation. Formally, the W+jets F_F can

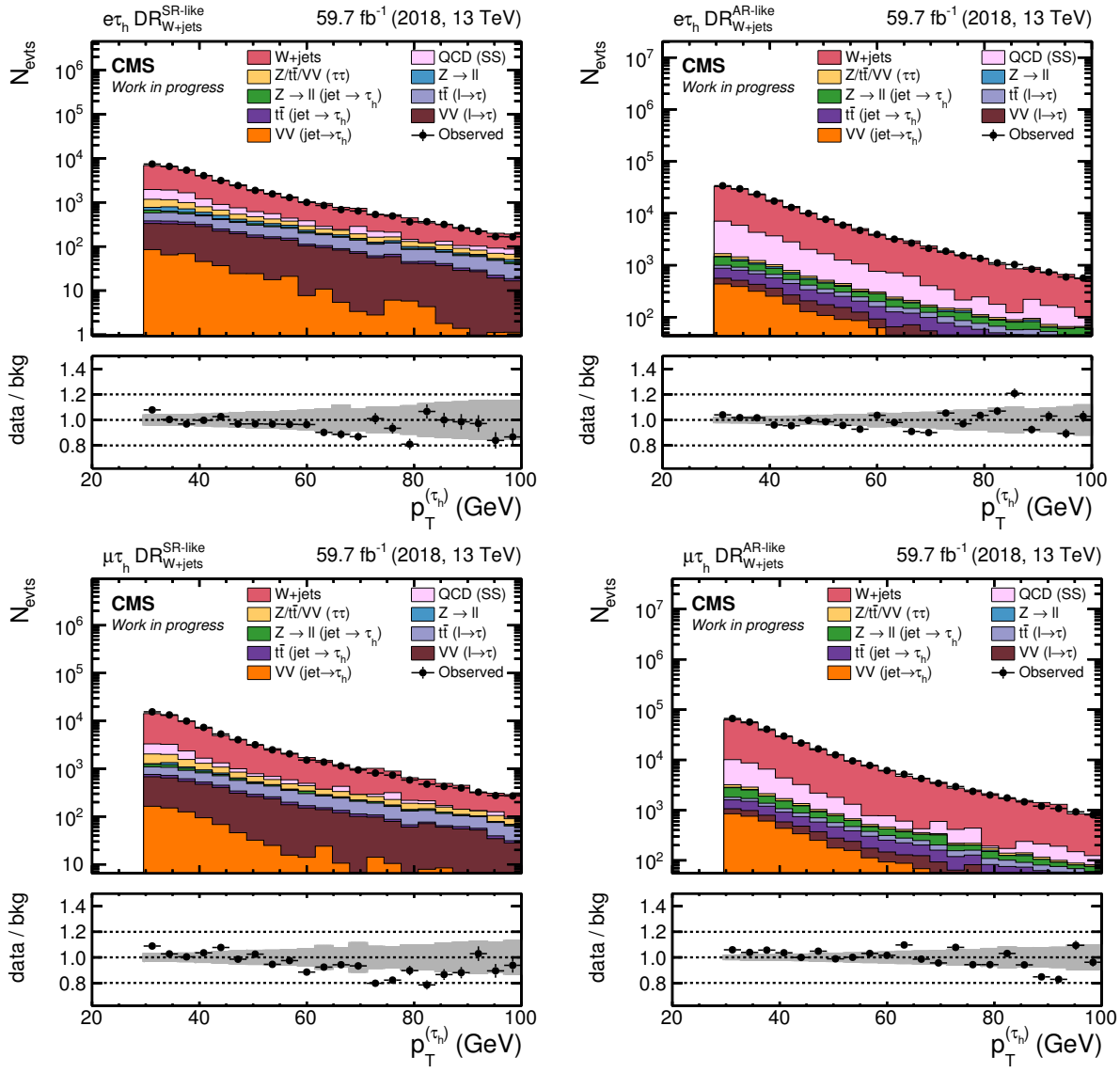


Figure 8.5: $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ distributions inside $\mathrm{DR}_{\mathrm{W+jets}}$ for the 2018 data-taking period are shown. The top row shows distributions for the $\mathrm{e}\tau_{\mathrm{h}}$ channel and the bottom row for the $\mu\tau_{\mathrm{h}}$ channel. On the left the SR-like part of $\mathrm{DR}_{\mathrm{W+jets}}$ is shown and on the right the AR-like part. In all plots the W+jets process is at the top of the stacked histograms. It is the dominating process, while all other processes combined make up a few percent of the total yield. The QCD multijet estimate is taken from a SS region as detailed in the text. Processes labeled as $\mathrm{Z/t\bar{t}}/VV\,(\tau\tau)$ are estimated by the τ-embedding technique. The ratio is calculated as observed over predicted (sum of all filled histogram). The error bars in the ratio plot represent the uncertainty on the observed contribution and the gray band reflects the uncertainty on the predicted contribution. Only statistical uncertainties are shown here.

be written in the following way

$$F_{\mathrm{F}}^{(\mathrm{W+jets})} = \frac{N_{\mathrm{data}}^{(\texttt{tight})} - \sum_{!\mathrm{W+jets}} N_{\mathrm{other}}^{(\texttt{tight})}}{N_{\mathrm{data}}^{(\texttt{vloose} \wedge \neg\texttt{tight})} - \sum_{!\mathrm{W+jets}} N_{\mathrm{other}}^{(\texttt{vloose} \wedge \neg\texttt{tight})}} \equiv F_{\mathrm{F,data}}^{(\mathrm{W+jets})}\left(p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}, N_{\mathrm{jets}}, \Delta R^{(\ell,\tau_{\mathrm{h}})}\right) , \tag{8.9}$$

where in the last step the dependencies on $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$, N_{jets} and $\Delta R^{(\ell,\tau_{\mathrm{h}})}$ are highlighted. Also highlighted in the second equality is the fact, that $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ is data-driven, i.e. it mostly depends on recorded collision data and relies only on simulation for subtracting a small contamination from other processes. The W+jets F_{F} in Equation (8.9) is also called the *raw* $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ since no corrections have been applied at this stage.

The F_{F} Parametrization

Figure 8.6 shows an example of $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ (see Equation (8.9)) as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$. The elements in this figure will be discussed in the following. The decay channel information, the N_{jets} as well as the $\Delta R^{(\ell,\tau_{\mathrm{h}})}$ category are reported on the top left of the figure. In case of Figure 8.6, a plot for the eτ_{h} channel in the $N_{\mathrm{jets}} = 0$ category with $\Delta R^{(\ell,\tau_{\mathrm{h}})} \geq 3$ is shown. Information about the year of data taking and corresponding run conditions can be found on the top right corner. For Figure 8.6, 2018 collision data – recorded at a center-of-mass energy of 13 TeV – are used which correspond to an integrated luminosity of 59.7 fb^{-1}. The measurement of $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ according to Equation (8.9) in bins of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown in black. The error bars of the measurement are asymmetric in x-direction and take into account the sample distribution within a $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ bin. The error bar to the left of each measurement point is always shorter than to right, reflecting the steeply falling $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ spectrum (see Figure 8.5). Since it is expected that F_{F}'s in general vary smoothly as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$, the measured $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ dependence is parametrized by means of a fit with a smooth function. This particularly avoids a propagation of statistical fluctuations from the F_{F} measurement to the analysis.

A linear fit model is chosen for all F_{F} parametrizations in the semi-leptonic final states. The linear[3] fit is shown in red in Figure 8.6 together with the corresponding 68% confidence level uncertainty band. The full $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ range up to 500 GeV is used for the fit. Furthermore, the χ^2 divided by the number of degrees of freedom is quoted for the linear fit in the legend. Depending on the slope of the linear fit, $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ either grows or falls towards zero for large $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$. Since $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ is applied as an event weight later, it must be avoided that $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ takes on unrealistically large or negative values for events with large $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$. This becomes even more important in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel, when a more complicated fit function is used (see Figure 8.24). One solution to this problem, is to switch to the measured value in the highest $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ bin, but this would result in a discontinuity at the bin boundary. Instead, the approach followed here is to truncate the linear fit at a certain value of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and use that value as $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ for all events with $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ above this threshold. This procedure is termed *high-p_{T} flattening* inside the legend of Figure 8.6. A threshold value of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})} = 80$ GeV is chosen universally, i.e. for every N_{jets} and $\Delta R^{(\ell,\tau_{\mathrm{h}})}$ category, and lies at a point where for each category the sample size does not exhibit large statistical fluctuations.

The uncertainty assigned to $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ is reflected by the yellow band shown in Figure 8.6, which is obtained by following resampling method. Each measurement point is

[3] It might not look as a linear fit at first sight because the x-axis in Figure 8.6 is on a logarithmic scale.

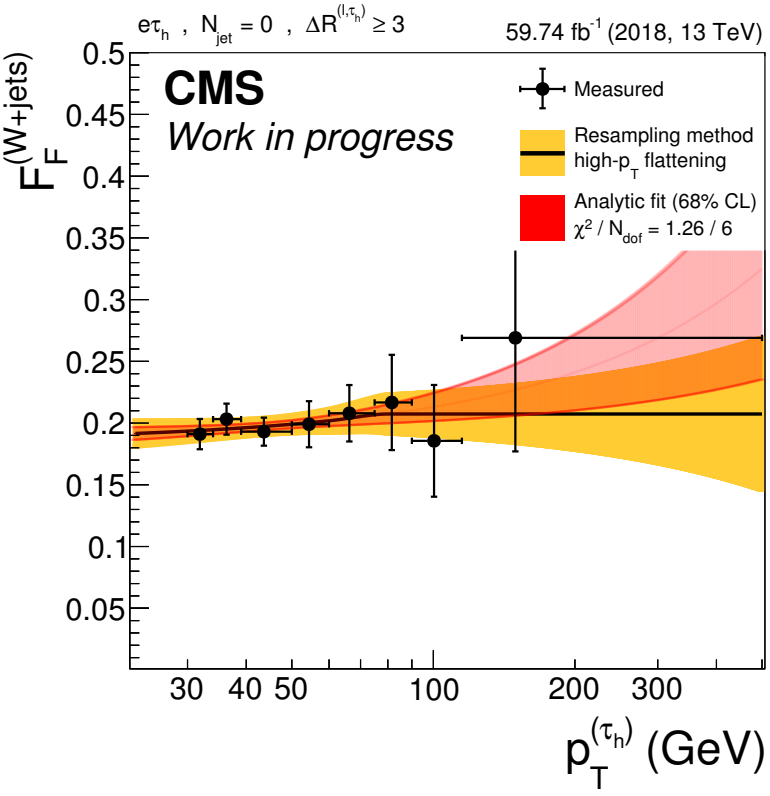


Figure 8.6: This figure shows the quantity $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$. It is derived for the eτ_{h} channel in the $N_{\mathrm{jets}} = 0$ and $\Delta R^{(\ell,\tau_{\mathrm{h}})} \geq 3$ category. Data from 2018 are used for the measurement. A linear fit is used to parametrize $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ and is shown as a solid line which is replaced by a constant at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$. The outcome of the analytic fit is shown in red. More details on the parametrization and the derivation of the uncertainty bands are given in the text.

fluctuated individually according to a Gaussian with a mean of the measured value and standard deviation corresponding to the statistical uncertainty of the measurement. The fluctuated measurement points are re-fit using a linear function. This procedure is repeated 200 times. Finally, the sample standard deviation of the ensemble of linear fits defines the yellow uncertainty band displayed in Figure 8.6. In order to reflect the uncertainty in the high-$p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ regime, the uncertainty band is inflated from the beginning of the constant continuation in a linear way.

In summary, the SM $\mathrm{H} \to \tau\tau$ analysis applies the parametrized $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ (solid black line in Figure 8.6) and its uncertainty (yellow band in Figure 8.6) enters the uncertainty model as described in Section 8.2.4. Figures 8.7 and 8.8 show $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ for all $N_{\mathrm{jets}} \times \Delta R^{(\ell,\tau_{\mathrm{h}})}$ categories for the $\mathrm{e}\tau_{\mathrm{h}}$ and $\mu\tau_{\mathrm{h}}$ channel, respectively. In Figure 8.4, the measurement of $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ is symbolically represented by the two arrows inside $\mathrm{DR}_{\mathrm{W+jets}}$ with labels "1" . The parametrized $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ will be used in the next sections to derive *closure corrections*, C, a process which is represented with a "2" in Figure 8.4.

Closure Tests inside $\mathrm{DR}_{\mathrm{W+jets}}$

In this section, the fitted parametrization of $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ (see Equation (8.9)) – developed in the previous section – is used inside $\mathrm{DR}_{\mathrm{W+jets}}$ and the effect on distributions of different variables is studied. The goal is to see how well $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ models other variables than N_{jets}, $\Delta R^{(\ell,\tau_{\mathrm{h}})}$ and $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$. For this purpose, let $\mathrm{DR}_{\mathrm{W+jets}}^{\text{AR-like}}$ be the subspace of $\mathrm{DR}_{\mathrm{W+jets}}$ with events passing the `vloose` D_{jet} condition but failing the `tight` one. The subspace of $\mathrm{DR}_{\mathrm{W+jets}}$ with events passing the `tight` D_{jet} condition, will be denoted as $\mathrm{DR}_{\mathrm{W+jets}}^{\text{SR-like}}$. This is the standard notation – used from now on – to refer to AR-like and SR-like DRs, respectively (see also Figure 8.4).

The *expected* number of W+jets events inside $\mathrm{DR}_{\mathrm{W+jets}}^{\text{SR-like}}$ is given by the numerator of Equation (8.9), i.e.

$$N_{\mathrm{exp}}^{(\text{SR-like})} = N_{\mathrm{data}}^{(\texttt{tight})} - \sum_{!\mathrm{W+jets}} N_{\mathrm{other}}^{(\texttt{tight})} \,. \tag{8.10}$$

These expected number of W+jets events can be filled in a histogram binned as a function of different variables as demonstrated in Figure 8.9. In Figure 8.9, the expected distributions are shown as black measurement points. They are compared to predicted distributions which are derived as follows. For simplicity, let us consider all *observed* events (ϵ) inside $\mathrm{DR}_{\mathrm{W+jets}}^{\text{AR-like}}$ that fall into a certain $(N_{\mathrm{jets}}, \Delta R^{(\ell,\tau_{\mathrm{h}})})$-category and a certain $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ bin such that

$$F_{\mathrm{F}}^{(\mathrm{W+jets})}\left(p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}(\epsilon), N_{\mathrm{jets}}(\epsilon), \Delta R^{(\ell,\tau_{\mathrm{h}})}(\epsilon)\right) = \text{constant} \equiv \overline{F_{\mathrm{F}}}^{(\mathrm{W+jets})} \,. \tag{8.11}$$

The expression

$$\begin{aligned}
\sum_{\epsilon} \overline{F_{\mathrm{F}}}^{(\mathrm{W+jets})} &= \overline{F_{\mathrm{F}}}^{(\mathrm{W+jets})} \cdot \sum_{\epsilon} 1 \\
&\overset{\text{Eq. 8.9}}{=} \frac{N_{\mathrm{data}}^{(\texttt{tight})} - \sum_{!\mathrm{W+jets}} N_{\mathrm{other}}^{(\texttt{tight})}}{N_{\mathrm{data}}^{(\texttt{vloose} \wedge \neg\texttt{tight})} - \sum_{!\mathrm{W+jets}} N_{\mathrm{other}}^{(\texttt{vloose} \wedge \neg\texttt{tight})}} \cdot N_{\mathrm{data}}^{(\texttt{vloose} \wedge \neg\texttt{tight})} \\
&\overset{\text{Eq. 8.10}}{=} N_{\mathrm{exp}}^{(\text{SR-like})} \cdot \frac{N_{\mathrm{data}}^{(\texttt{vloose} \wedge \neg\texttt{tight})}}{N_{\mathrm{data}}^{(\texttt{vloose} \wedge \neg\texttt{tight})} - \sum_{!\mathrm{W+jets}} N_{\mathrm{other}}^{(\texttt{vloose} \wedge \neg\texttt{tight})}} \,,
\end{aligned} \tag{8.12}$$

can be written as the product of the expected number of events and an extra factor as shown in the last line of the equality. This extra factor is bigger than one and makes

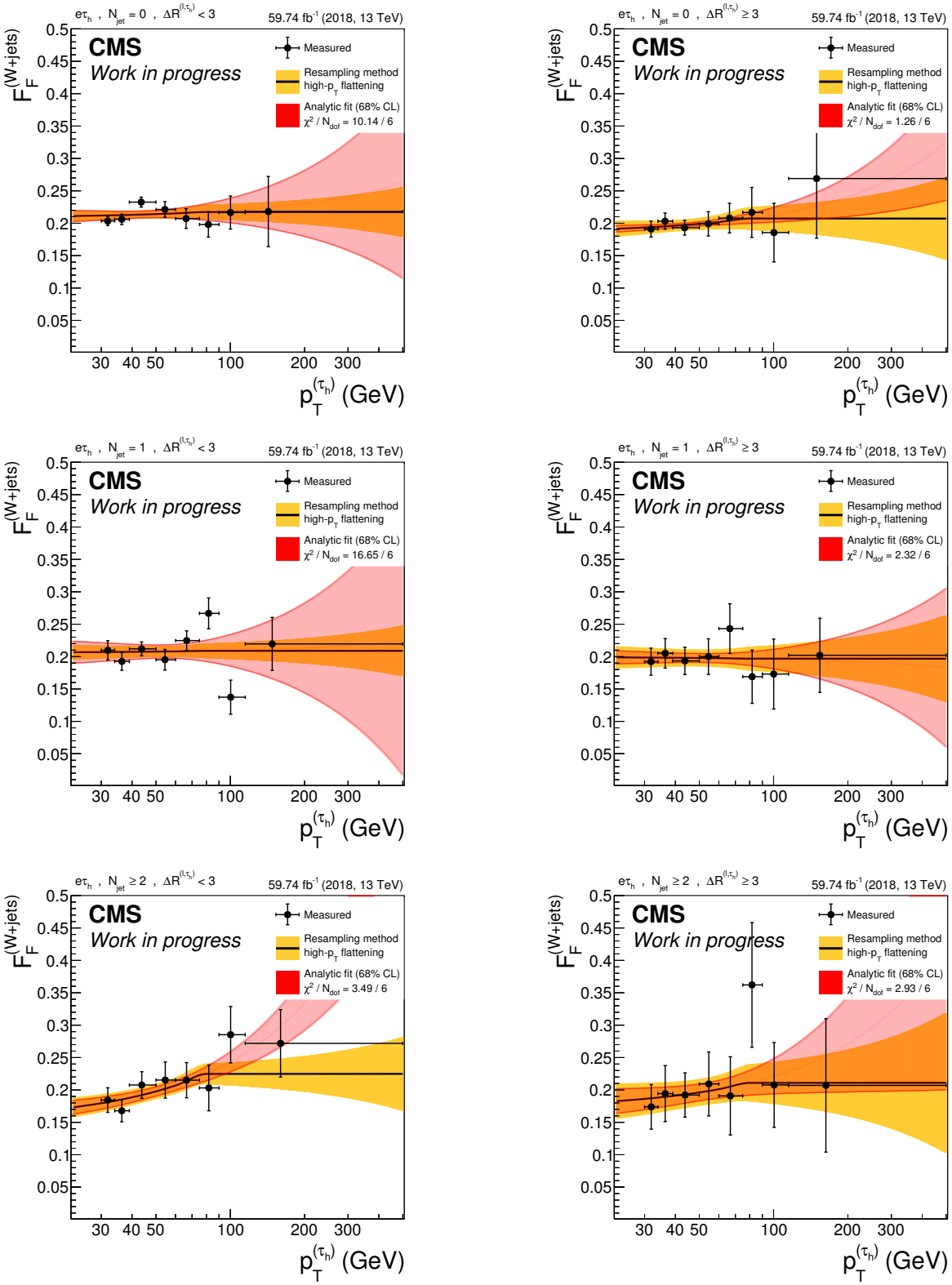


Figure 8.7: The quantity $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown for the eτ_{h} channel using 2018 data. From top to bottom, three different N_{jets} categories are displayed – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} = 1$ and $N_{\mathrm{jets}} \geq 2$. On the left, the $\Delta R^{(\ell,\tau_{\mathrm{h}})} < 3$ category is shown and on the right, the $\Delta R^{(\ell,\tau_{\mathrm{h}})} \geq 3$ category. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid black line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique, as explained in the text.

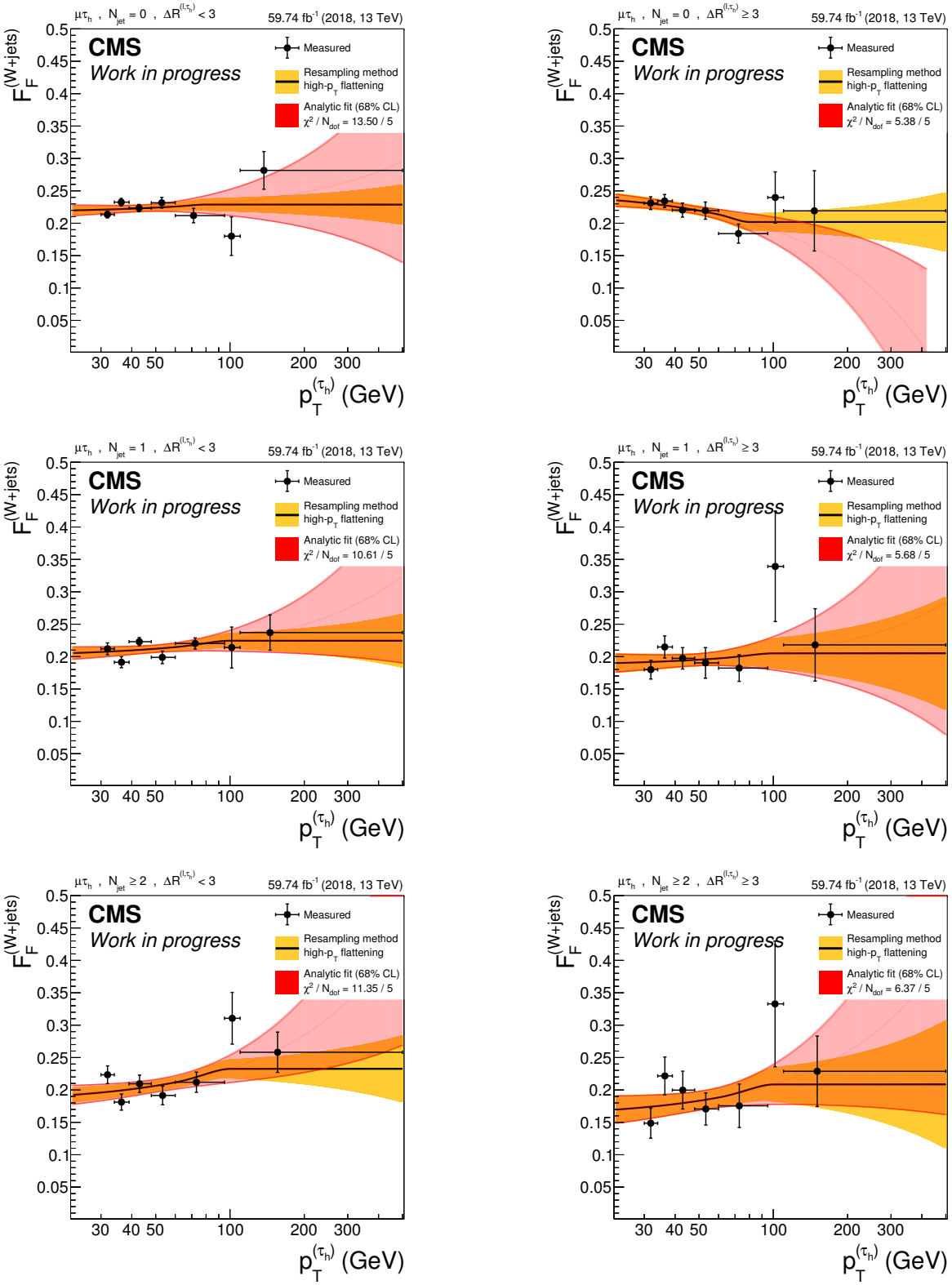


Figure 8.8: The quantity $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown for the $\mu\tau_{\mathrm{h}}$ channel using 2018 data. From top to bottom, three different N_{jets} categories are displayed – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} = 1$ and $N_{\mathrm{jets}} \geq 2$. On the left, the $\Delta R^{(\ell,\tau_{\mathrm{h}})} < 3$ category is shown and on the right, the $\Delta R^{(\ell,\tau_{\mathrm{h}})} \geq 3$ category. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid black line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique, as explained in the text.

$\sum_\epsilon \overline{F_\mathrm{F}}^{(\mathrm{W+jets})}$ larger than the expected number of events. The reason why is because not all ϵ's are W+jets events (see also right side of Figure 8.5). Hence, each observed event has to be weighted by a factor

$$p(\mathrm{W+jets}|\mathrm{DR}^{\mathrm{AR\text{-}like}}_{\mathrm{W+jets}}) = \frac{N^{(\texttt{vloose} \wedge \neg\texttt{tight})}_{\mathrm{data}} - \sum_{!\mathrm{W+jets}} N^{(\texttt{vloose} \wedge \neg\texttt{tight})}_{\mathrm{other}}}{N^{(\texttt{vloose} \wedge \neg\texttt{tight})}_{\mathrm{data}}} \,, \tag{8.13}$$

which can be interpreted as the probability of that event to originate from a W+jets process. The correct expression for the *predicted* number of events is thus

$$N^{(\mathrm{SR\text{-}like})}_{\mathrm{pred}} = \sum_\epsilon \overline{F_\mathrm{F}}^{(\mathrm{W+jets})} \cdot p(\mathrm{W+jets}|\mathrm{DR}^{\mathrm{AR\text{-}like}}_{\mathrm{W+jets}}) \,, \tag{8.14}$$

which exactly matches the expected number of events in the simplified setup. In the analysis, however, the parametrized F_F is used in Equation (8.14), which explains why the expected and predicted $p_\mathrm{T}^{(\tau_\mathrm{h})}$-distributions in Figure 8.9a do not match exactly. Note, that the probability in Equation (8.13) is binned in the variable which is investigated, e.g. as a function of m_vis in Figure 8.9d. In general, all distributions in Figure 8.9 show a good agreement. A small trend can be seen in the ratio of distributions of the light lepton transverse momentum, seen in Figure 8.9b, which motivates a correction in this variable.

Closure Correction of $F_\mathrm{F}^{(\mathrm{W+jets})}$ in $p_\mathrm{T}^{(\ell)}$

In order to improve the agreement between expected and predicted $p_\mathrm{T}^{(\ell)}$ distributions (see Figure 8.9b), the ratio of these two distributions is applied as a multiplicative correction to the raw $F_\mathrm{F}^{(\mathrm{W+jets})}$. The closure correction is denoted as

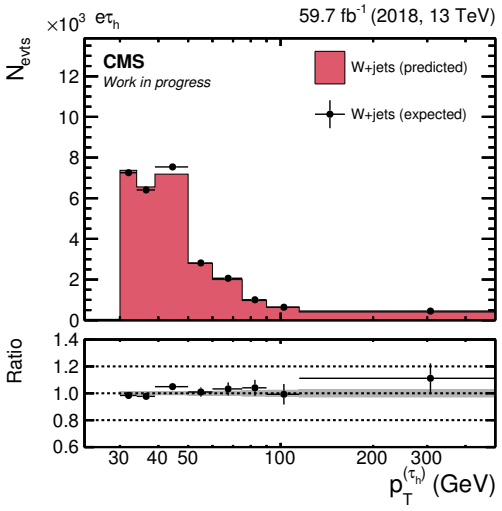
$$C^{(\mathrm{W+jets})} \equiv C^{(\mathrm{W+jets})}_{\mathrm{data}}(p_\mathrm{T}^{(\ell)}) \,, \tag{8.15}$$

where its dependence on $p_\mathrm{T}^{(\ell)}$ is highlighted. Like for the parametrization of the raw $F_\mathrm{F}^{(\mathrm{W+jets})}$, the closure correction is expected to vary smoothly as a function of $p_\mathrm{T}^{(\ell)}$. Closure corrections are smoothed with a Gaussian kernel of variable width taking into account the statistical uncertainty of neighboring measurement points. This way, no explicit choice on the functional form of the closure correction has to be made. Figure 8.10 shows the measured closure correction together with the smoothed curve which is used to retrieve the correction values. The uncertainty band is obtained by fluctuating the measurement points according to a Gaussian and redoing the smoothing. The sample standard deviation of the ensemble of smoothed curves, shown as yellow uncertainty band in Figure 8.10, serves as an uncertainty on the smoothed curve. It enters the F_F uncertainty model as discussed in Section 8.2.4. It is demonstrated in [216] that there is no significant improvement when closure corrections are measured differentially, e.g. in different N_jets categories. Therefore, for the benefit of larger sample sizes, $p_\mathrm{T}^{(\ell)}$ closure corrections are calculated inclusively in N_jets and inclusively in $\Delta R^{(\ell,\tau_\mathrm{h})}$.

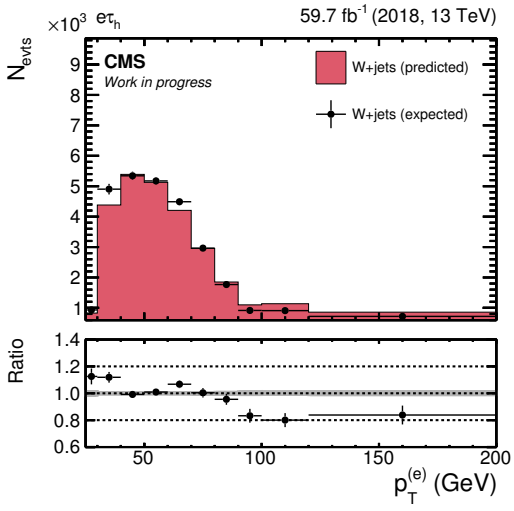
Extending Equation (8.14), by taking into account the closure correction, yields the following expression

$$N^{(\mathrm{SR\text{-}like})}_{\mathrm{pred}} = \sum_\epsilon F_\mathrm{F}^{(\mathrm{W+jets})} \cdot C^{(\mathrm{W+jets})} \cdot p(\mathrm{W+jets}|\mathrm{DR}^{\mathrm{AR\text{-}like}}_{\mathrm{W+jets}}) \,. \tag{8.16}$$

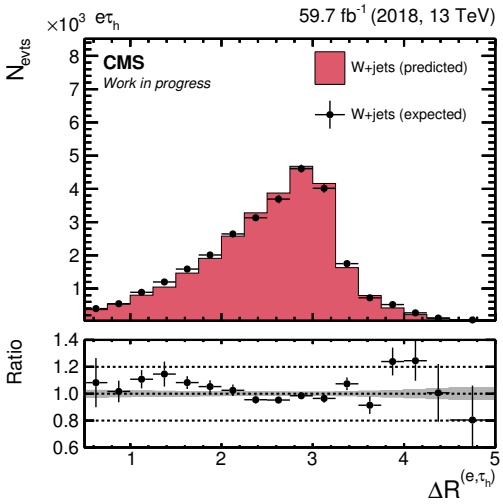
Different distributions using the above equation for the prediction are shown in Figure 8.11. Compared to Figure 8.9, an improvement in the closure of $p_\mathrm{T}^{(\ell)}$ is observed, whereas all other distributions still show a very good agreement within statistical uncertainties. The imperfect final closure in $p_\mathrm{T}^{(\ell)}$ results from using the correction value returned from the smoothed curve.



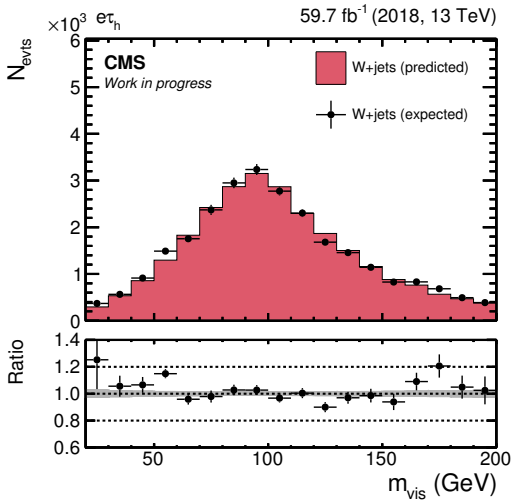
(a) τ_{h} transverse momentum



(b) electron transverse momentum



(c) ΔR between τ_{h} and electron



(d) invariant mass of visible decay products

Figure 8.9: Different closure distributions inside $\mathrm{DR}_{\mathrm{W+jets}}$ for the $\mathrm{e}\tau_{\mathrm{h}}$ channel using 2018 data are shown. The expected distribution is calculated using Equation (8.10). The prediction is obtained according to Equation (8.14), where $\overline{F_{\mathrm{F}}}^{(\mathrm{W+jets})}$ is replaced by the parametrization shown as black lines in Figure 8.7. The ratio is calculated as expected over predicted. The error bars in the ratio plot represent the uncertainty on the expected contribution and the gray band reflects the uncertainty on the predicted contribution. Only statistical uncertainties are shown.

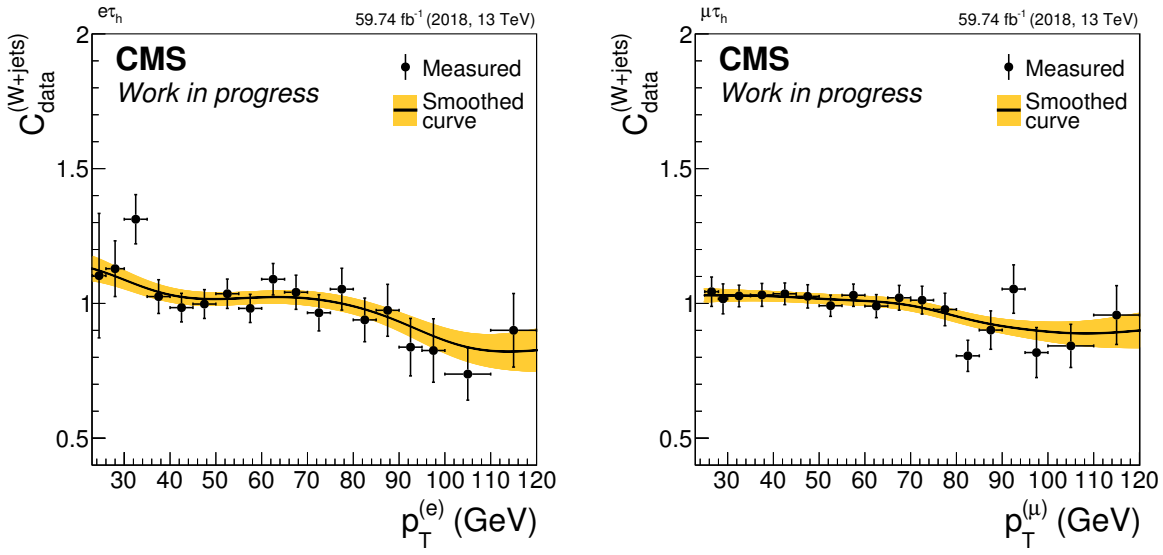


Figure 8.10: The closure corrections of $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ as a function of $p_{\mathrm{T}}^{(\ell)}$ are shown for the 2018 data-taking period. On the left, the $\mathrm{e}\tau_{\mathrm{h}}$ channel and on the right, the $\mu\tau_{\mathrm{h}}$ channel are displayed. The measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data as explained in the text.

Bias Correction of $F_{\mathrm{F}}^{(\mathrm{W+jets})}$

The missing part in the derivation of $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ is the correction, B, which is labeled with "3" in Figure 8.4. The goal of this correction is to mitigate biases introduced by applying a different $m_{\mathrm{T,PUPPI}}$ cut in the SR and inside $\mathrm{DR}_{\mathrm{W+jets}}$ (see also Equation (8.5)). To derive the bias correction, the $m_{\mathrm{T,PUPPI}}$ cut is omitted[4] such that the resulting DR, $\mathrm{DR}_{\mathrm{W+jets}}^{(\mathrm{B})}$, overlaps with the SR. Hence, it is impossible to use collision data to calculate this correction. A new $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ using *simulated* W+jets events is derived inside $\mathrm{DR}_{\mathrm{W+jets}}^{(\mathrm{B})}$ in analogy to Equation (8.9)

$$\begin{aligned} F_{\mathrm{F}}^{(\mathrm{W+jets})} &= \frac{N_{\mathrm{W+jets}}^{(\texttt{tight})}}{N_{\mathrm{W+jets}}^{(\texttt{vloose} \,\wedge\, \neg\texttt{tight})}} \\ &\equiv F_{\mathrm{F,MC}}^{(\mathrm{W+jets})}\left(p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}, N_{\mathrm{jets}}, \Delta R^{(\ell,\tau_{\mathrm{h}})}\right) \,, \end{aligned} \tag{8.17}$$

where in the last step it is emphasized that the raw $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ is derived purely from MC simulation.

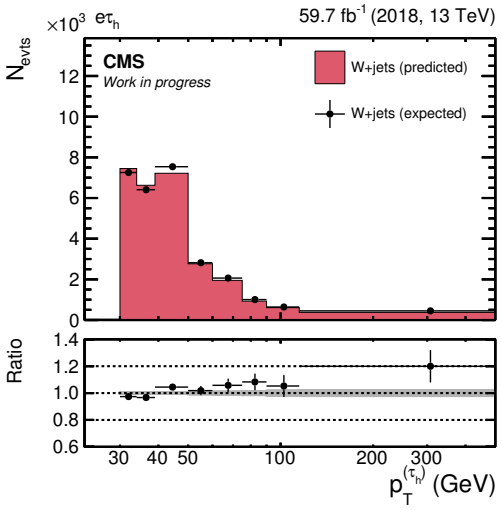
In a next step, a closure correction with respect to $p_{\mathrm{T}}^{(\ell)}$ similar to Equation (8.15) is calculated, again using solely simulated W+jets events

$$\mathrm{C}^{(\mathrm{W+jets})} \equiv \mathrm{C}_{\mathrm{MC}}^{(\mathrm{W+jets})}(p_{\mathrm{T}}^{(\ell)}) \,. \tag{8.18}$$

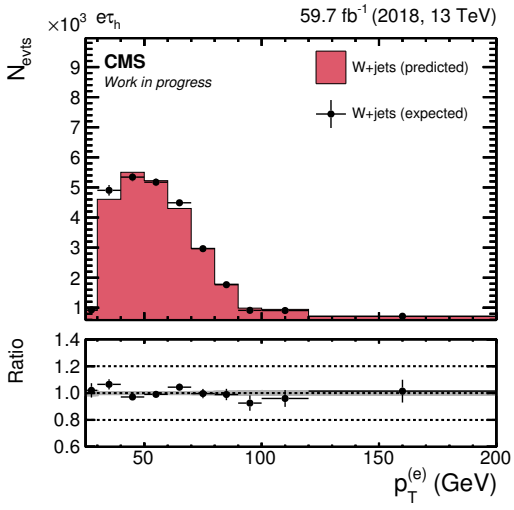
The expected number of events inside $\mathrm{DR}_{\mathrm{W+jets}}^{(\mathrm{B})\,\text{SR-like}}$ is then compared to the prediction derived using the formula

$$N_{\mathrm{pred}}^{(\text{SR-like})} = \sum_{\omega} F_{\mathrm{F,MC}}^{(\mathrm{W+jets})} \cdot \mathrm{C}_{\mathrm{MC}}^{(\mathrm{W+jets})} \,, \tag{8.19}$$

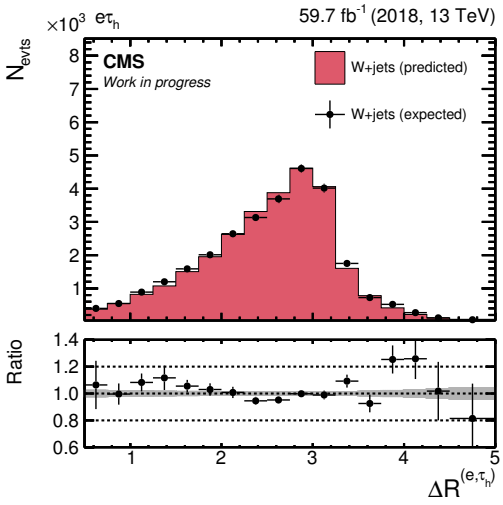
[4] Note, that the requirement of events without any b-tagged jets from Equation (8.6) is still applied.



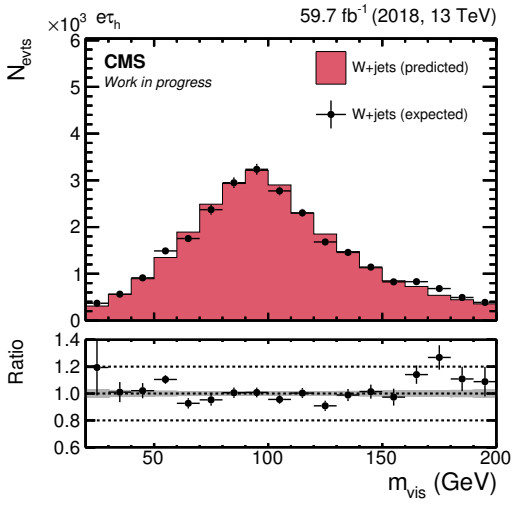
(a) τ_{h} transverse momentum



(b) electron transverse momentum



(c) ΔR between τ_{h} and electron



(d) invariant mass of visible decay products

Figure 8.11: Different closure distributions inside $\mathrm{DR}_{\mathrm{W+jets}}$ for the eτ_{h} channel using 2018 data are shown. The expected distribution is calculated using Equation (8.10). The prediction is obtained according to Equation (8.16), i.e. using the closure correction in $p_{\mathrm{T}}^{(\ell)}$ from Figure 8.10. The ratio is calculated as expected over predicted. The error bars in the ratio plot represent the uncertainty on the expected contribution and the gray band reflects the uncertainty on the predicted contribution. Only statistical uncertainties are shown.

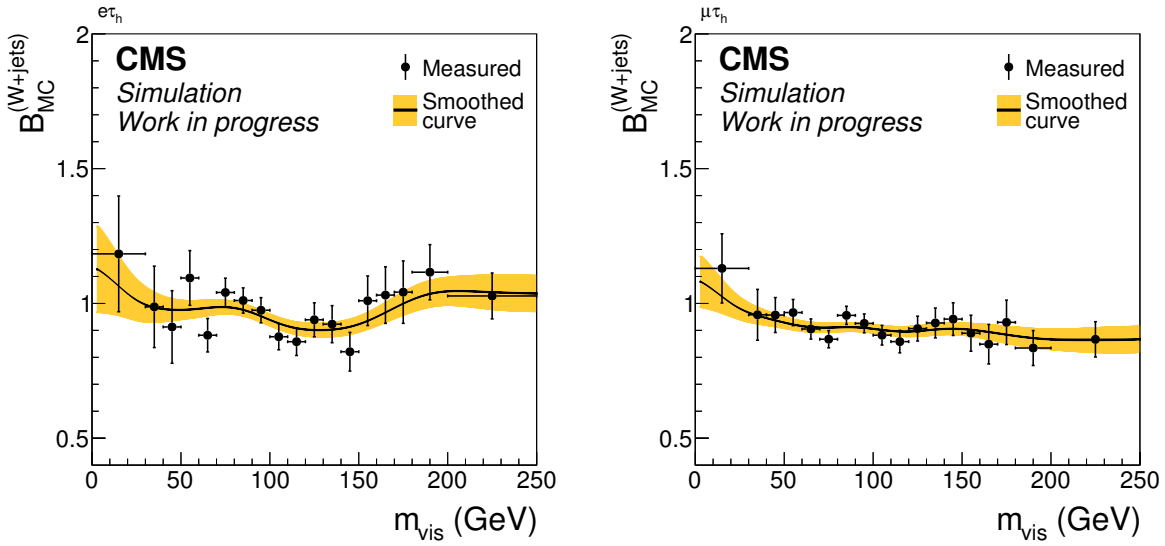


Figure 8.12: The bias corrections of $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ as a function of m_{vis} are shown, using simulated events from the 2018 data-taking period. On the left, the $\mathrm{e}\tau_{\mathrm{h}}$ channel and on the right, the $\mu\tau_{\mathrm{h}}$ channel are displayed. The measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

where the sum runs over all *simulated*[5] W+jets events (ω) inside $\mathrm{DR}_{\mathrm{W+jets}}^{\mathrm{(B)\,AR\text{-}like}}$. The ratio of the expected over predicted distribution is applied as a multiplicative correction and defines the bias correction

$$B_{\mathrm{MC}}^{(\mathrm{W+jets})}(m_{\mathrm{vis}}) \; . \tag{8.20}$$

It is calculated as a function of the invariant mass of the visible decay products from the tau pair (m_{vis}) and it is important to emphasize that the bias correction is purely based on MC simulation. The bias corrections is smoothed with a Gaussian kernel in the same manner as explained in Section 8.2.1 for the closure correction. The bias correction is shown in Figure 8.12.

Finally – using Equations (8.9), (8.15) and (8.20) – the combined W+jets F_{F} can be written as

$$F_{\mathrm{F}}^{(\mathrm{W+jets})} = F_{\mathrm{F,data}}^{(\mathrm{W+jets})}(p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}, N_{\mathrm{jets}}, \Delta R^{(\ell,\tau_{\mathrm{h}})}) \cdot C_{\mathrm{data}}^{(\mathrm{W+jets})}(p_{\mathrm{T}}^{(\ell)}) \cdot B_{\mathrm{MC}}^{(\mathrm{W+jets})}(m_{\mathrm{vis}}), \tag{8.21}$$

which is the same expression as written in a more compact form in Figure 8.4. In summary, $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ consists of a raw F_{F}, which is measured inside six different $N_{\mathrm{jets}} \times \Delta R^{(\ell,\tau_{\mathrm{h}})}$ categories. In each of these categories, the $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ distribution is fit with a linear function. The linear function is replaced by a constant for high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ values. Two multiplicative corrections are applied to the raw W+jets F_{F}. One is correcting for small non-closures in $p_{\mathrm{T}}^{(\ell)}$ inside $\mathrm{DR}_{\mathrm{W+jets}}$. Typically, the closure corrections correct the raw $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ upwards for small $p_{\mathrm{T}}^{(\ell)}$ and decrease it for high $p_{\mathrm{T}}^{(\ell)}$ (see Figure 8.10). For most of the $p_{\mathrm{T}}^{(\ell)}$ spectrum, the raw $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ is corrected by a few percent. However, corrections up to 20% are observed. The other correction reduces biases in the F_{F} measurement, coming from

[5] Therefore, no probabilities are needed in Equation (8.19) because $p(\mathrm{W+jets}|\mathrm{DR}_{\mathrm{W+jets}}^{\mathrm{(B)\,AR\text{-}like}}) = 1$ for all W+jets events, ω.

applying different kinematic selection in the DR with respect to the SR. Bias corrections take on values roughly bound by $[0.9, 1.1]$ over the whole m_{vis} spectrum (see Figure 8.12).

The QCD Multijet Fake Factor

In the following, the pink part of Figure 8.4 – representing the F_{F} derivation of QCD multijet events – is discussed. The steps "1" and "2" related to the F_{F} measurement, its parametrization and closure correction are done exactly the same as for $F_{\text{F}}^{(\text{W+jets})}$ discussed above, just in a different DR. More care is needed for the bias corrections, B_1 and B_2 (labeled with "3" and "4" in Figure 8.4), in case of the QCD multijet F_{F}. The QCD multijet F_{F} is measured in a QCD enriched DR, referred to as DR_{QCD}, differing from the SR in the following way:

- The electric charges of the selected light lepton, $\ell \in \{\text{e}, \mu\}$, and τ_{h} candidate are required to have the same sign (SS)

$$q^{(\ell)} \cdot q^{(\tau_{\text{h}})} > 0 \ . \tag{8.22}$$

  This selection enhances the contribution from QCD multijet events with respect to all other processes.

- Even after the SS requirement, up to 50% of events with well isolated leptons originate from W+jets processes. The following selection on the relative lepton isolation removes those events

$$\begin{aligned} I_{\text{rel}}^{(\text{e})} &\in [0.02, 0.15] \\ I_{\text{rel}}^{(\mu)} &\in [0.05, 0.15] \ , \end{aligned} \tag{8.23}$$

  in case of the $\text{e}\tau_{\text{h}}$ and $\mu\tau_{\text{h}}$ channel, respectively. In addition, the selection

$$m_{\text{T,PUPPI}} < 50\,\text{GeV} \ , \tag{8.24}$$

  further reduces the contribution of W+jets events to DR_{QCD}.

Figure 8.13 shows $p_{\text{T}}^{(\tau_{\text{h}})}$ distributions inside DR_{QCD}. Applying the selections described above results in a QCD-enriched DR as can be seen from Figure 8.13. Similar as for $F_{\text{F}}^{(\text{W+jets})}$, the raw QCD F_{F} ($F_{\text{F}}^{(\text{QCD})}$) is calculated according to Equation (8.9). Formally, the expression[6] for $F_{\text{F}}^{(\text{QCD})}$ is given by

$$\begin{aligned} F_{\text{F}}^{(\text{QCD})} &= \frac{N_{\text{data}}^{(\texttt{tight})} - \sum N_{\text{other}}^{(\texttt{tight})}}{N_{\text{data}}^{(\texttt{vloose} \,\wedge\, \neg\texttt{tight})} - \sum N_{\text{other}}^{(\texttt{vloose} \,\wedge\, \neg\texttt{tight})}} \\ &\equiv F_{\text{F,data}}^{(\text{QCD})}(p_{\text{T}}^{(\tau_{\text{h}})}, N_{\text{jets}}) \ , \end{aligned} \tag{8.25}$$

where in the last step its dependencies on $p_{\text{T}}^{(\tau_{\text{h}})}$ und N_{jets} are highlighted. In case of $F_{\text{F}}^{(\text{QCD})}$, three N_{jets} categories – $N_{\text{jets}} = 0$, $N_{\text{jets}} = 1$ and $N_{\text{jets}} \geq 2$ – are used and no further splitting in $\Delta R^{(\ell,\tau_{\text{h}})}$ regions is applied. Raw $F_{\text{F}}^{(\text{QCD})}$ are shown in Figure 8.14 for the 2018 data-taking period.

In exactly the same way as explained in Section 8.2.1, a closure correction is derived for $F_{\text{F}}^{(\text{QCD})}$. This step is labeled with "2" in the pink part of Figure 8.4. The predicted event distribution

$$N_{\text{pred}}^{(\text{SR-like})} = \sum_{\epsilon} F_{\text{F,data}}^{(\text{QCD})} \cdot p(\text{QCD}|\text{DR}_{\text{QCD}}^{\text{AR-like}}) \ , \tag{8.26}$$

[6] Note, that in Equation (8.25) the sum runs over all processes, including the W+jets process

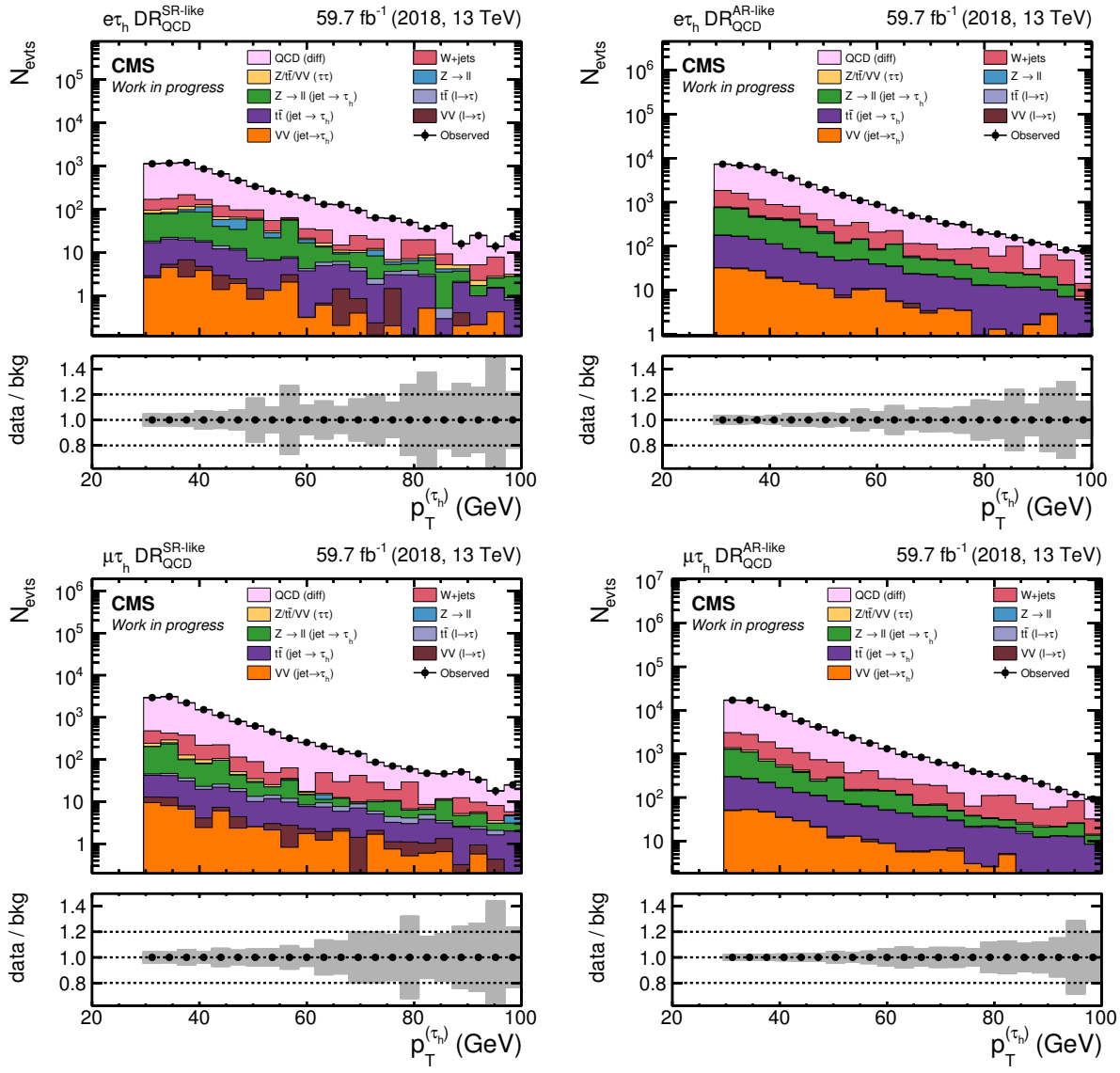


Figure 8.13: $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ distributions inside $\mathrm{DR}_{\mathrm{QCD}}$ for the 2018 data-taking period are shown. The top row shows distributions for the $\mathrm{e}\tau_{\mathrm{h}}$ channel and the bottom row for the $\mu\tau_{\mathrm{h}}$ channel. On the left the SR-like part of $\mathrm{DR}_{\mathrm{QCD}}$ is shown and on the right the AR-like part. In all plots, the QCD multijet process is at the top of the stacked histograms. The estimated yield from QCD multijet events is given by the difference between data and the sum of all other background processes. It is the dominating process while all other processes combined make up at most a few percent of the total yield. The QCD multijet estimate is simply taken as the difference between the stacked histograms and the observation. Processes labeled as $\mathrm{Z/t\bar{t}}/VV\,(\tau\tau)$ are estimated by the τ-embedding technique. Only statistical uncertainties are shown here.

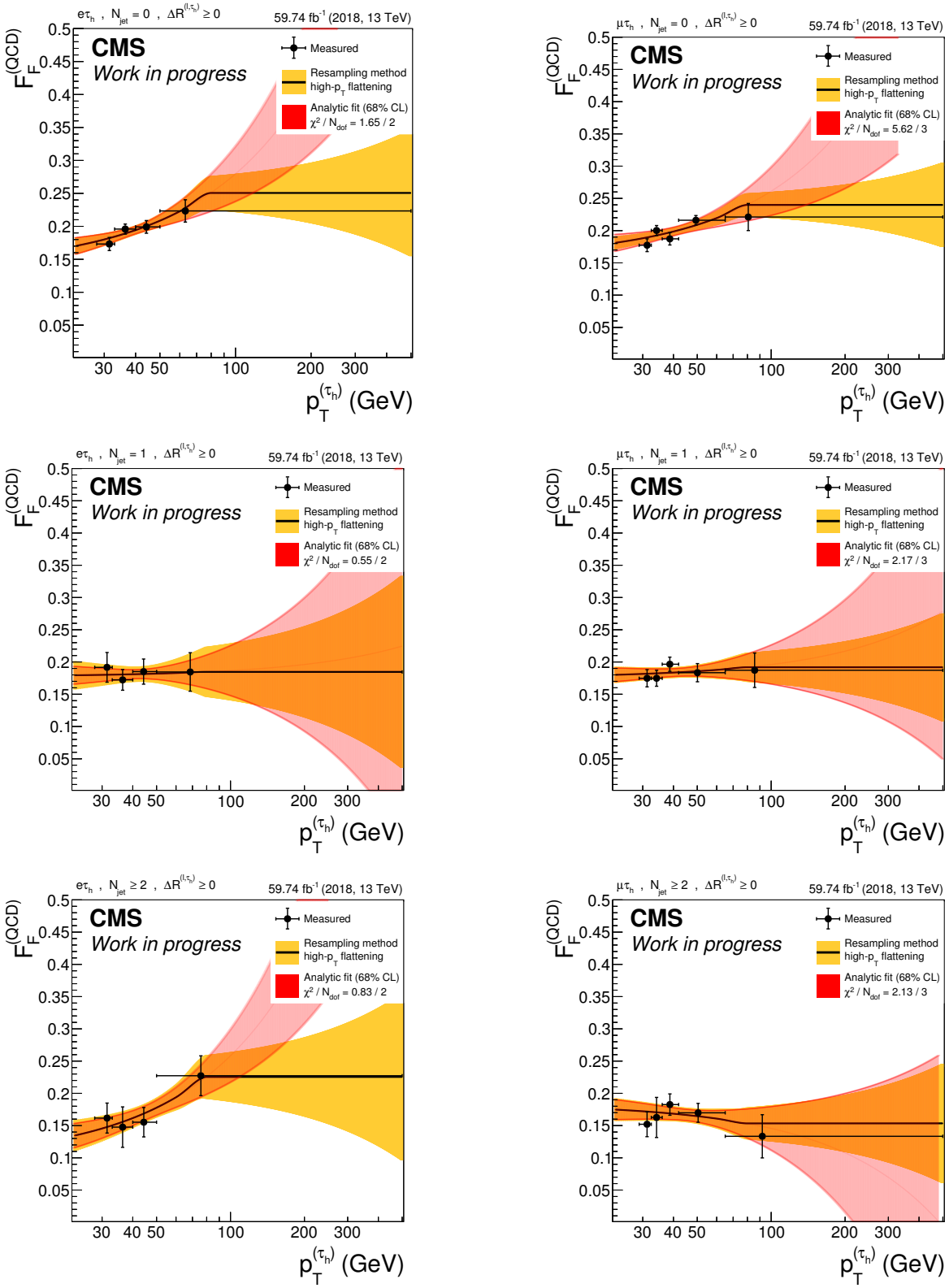


Figure 8.14: The quantity $F_{\mathrm{F}}^{(\mathrm{QCD})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown using 2018 data. On the left, distributions for the eτ_{h} channel are displayed and corresponding distributions for the $\mu\tau_{\mathrm{h}}$ channel are displayed on the right. From top to bottom, three different N_{jets} categories are displayed – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} = 1$ and $N_{\mathrm{jets}} \geq 2$. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid line is used together with its associated uncertainty band shown in yellow.

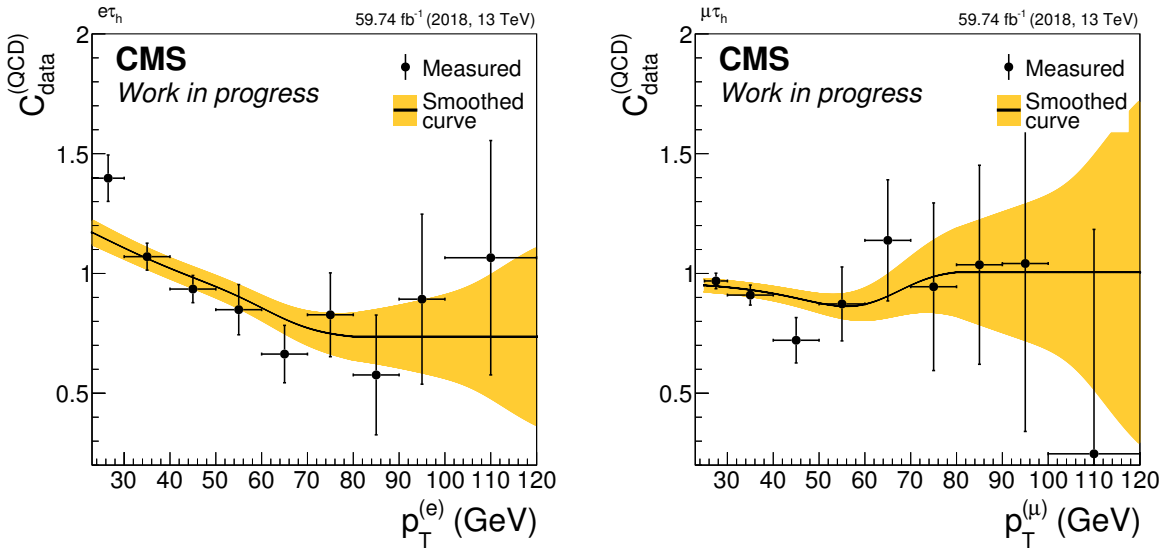


Figure 8.15: The closure corrections of $F_{\mathrm{F}}^{(\mathrm{QCD})}$ as a function of $p_{\mathrm{T}}^{(\ell)}$ are shown for the 2018 data-taking period. On the left, the $\mathrm{e}\tau_{\mathrm{h}}$ channel and on the right, the $\mu\tau_{\mathrm{h}}$ channel are displayed. The measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

is compared to the expected distribution as a function of $p_{\mathrm{T}}^{(\ell)}$. The ratio of expected over predicted $p_{\mathrm{T}}^{(\ell)}$ distribution defines the closure correction

$$C^{(\mathrm{QCD})} \equiv C_{\mathrm{data}}^{(\mathrm{QCD})}(p_{\mathrm{T}}^{(\ell)}) \ , \tag{8.27}$$

which is shown in Figure 8.15.

Bias Corrections of $F_{\mathrm{F}}^{(\mathrm{QCD})}$

The respective bias corrections, applied to $F_{\mathrm{F}}^{(\mathrm{QCD})}$, are denoted as B_1 and B_2 in Figure 8.4. They are used to correct for effects introduced by measuring $F_{\mathrm{F}}^{(\mathrm{QCD})}$ in a different region than it is applied. In the case of $F_{\mathrm{F}}^{(\mathrm{W+jets})}$, simulated W+jets events are used to derive the bias correction (see Equation (8.20)). However, QCD multijet events are difficult to simulate accurately. Therefore, several new regions – all of them orthogonal to the SR – are used to derive bias corrections directly from data. In order to ease the discussion, Figure 8.16 visualizes the different regions used.

Bias Correction: Lepton Isolation – B_1

For the derivation of the lepton isolation bias correction, a new measurement region – denoted as $\mathrm{DR}_{\mathrm{QCD,SS}}$ – is defined with the requirements:

- The electric charges of the selected light lepton, $\ell \in \{\mathrm{e}, \mu\}$, and τ_{h} candidate are required to be SS
$$q^{(\ell)} \cdot q^{(\tau_{\mathrm{h}})} > 0 \ . \tag{8.28}$$
- No restriction on the relative lepton isolation ($I_{\mathrm{rel}}^{(\ell)}$).

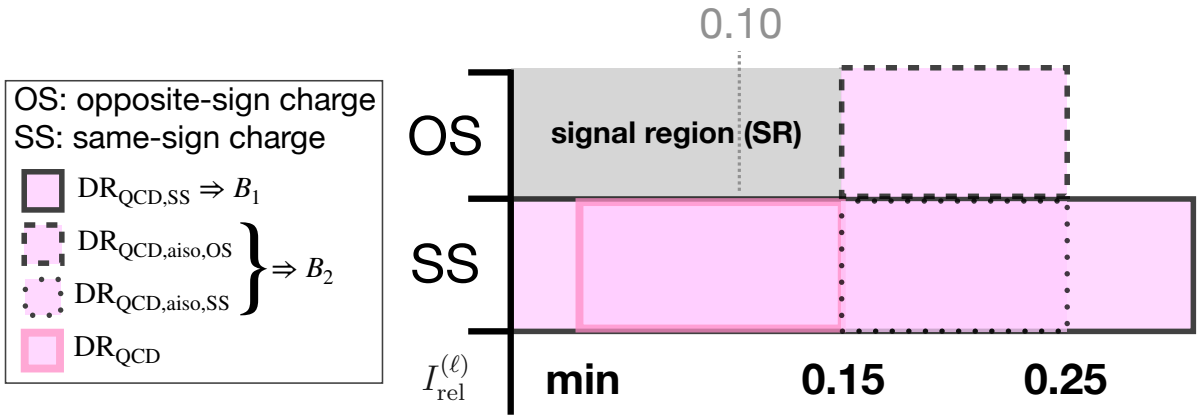


Figure 8.16: A schematic of all regions used to derive $F_F^{(QCD)}$ and its corrections is given. The figure focuses on the relevant differences between these DRs and the SR. The y-axis shows the charge requirement imposed on the tau pair and on the x-axis the relative lepton isolation is given. In gray, the SR is depicted. Different selections on $I_{rel}^{(\ell)}$ define the SR depending on the final state (see Table 6.3). Furthermore, DR_{QCD} is shown which is used to derive the raw $F_F^{(QCD)}$ and its closure correction. In case of DR_{QCD}, the definition of the lower bound on $I_{rel}^{(\ell)}$ varies between $e\tau_h$ and $\mu\tau_h$ channel (see Equation (8.23)) and is here indicated by "min". The other regions are used to derive bias corrections B_1 and B_2 (see also Figure 8.4).

Figure 8.16 depicts $DR_{QCD,SS}$ with a solid frame. The region overlaps with DR_{QCD} but is orthogonal to the SR because of the SS requirement.

First, probabilities, $p(\text{QCD}|DR_{QCD,SS}^{\text{AR-like}})$, of an event being a QCD multijet event are calculated as a function of $I_{rel}^{(\ell)}$ inside $DR_{QCD,SS}^{\text{AR-like}}$. According to Equation (8.16), the predicted number of events inside $DR_{QCD,SS}^{\text{SR-like}}$ is given by

$$N_{pred}^{(\text{SR-like})} = \sum_{\epsilon \in DR_{QCD,SS}^{\text{AR-like}}} F_{F,data}^{(QCD)} \cdot C_{data}^{(QCD)} \cdot p(\text{QCD}|DR_{QCD,SS}^{\text{AR-like}}) \ . \tag{8.29}$$

The ratio of expected over predicted $I_{rel}^{(\ell)}$ distribution inside $DR_{QCD,SS}^{\text{SR-like}}$ defines then the bias correction

$$B_{data}^{(QCD)}(I_{rel}^{(\ell)}) \equiv B_1 \ . \tag{8.30}$$

Note, that the $I_{rel}^{(\ell)}$ bias correction is calculated almost entirely from data. The $I_{rel}^{(\ell)}$ correction is shown in Figure 8.17. Over the whole $I_{rel}^{(\ell)}$ spectrum, the correction takes on values in the interval $[0.9, 1.1]$, while most of them are close to one. The larger statistical uncertainties in the low-$I_{rel}^{(\ell)}$ regime, especially visible in the $\mu\tau_h$ channel, are caused by subtracting significant contributions from W+jets processes. Uncertainties related to this subtraction are propagated to the bias correction and are included in the yellow uncertainty band shown in Figure 8.17. Furthermore, it should be mentioned that throughout the derivation of B_1 it is assumed that the F_F dependency on $I_{rel}^{(\ell)}$ is uncorrelated to the sign of the electric charges of the tau pair. Under this assumption, the bias correction measured in a SS region, can be applied in the AR (which is OS).

Bias correction: SS $\rightarrow$ OS – B_2

This correction mitigates effects introduced by measuring $F_F^{(QCD)}$ in a region with a SS requirement but applying it in an OS region. Two new regions are defined for the derivation

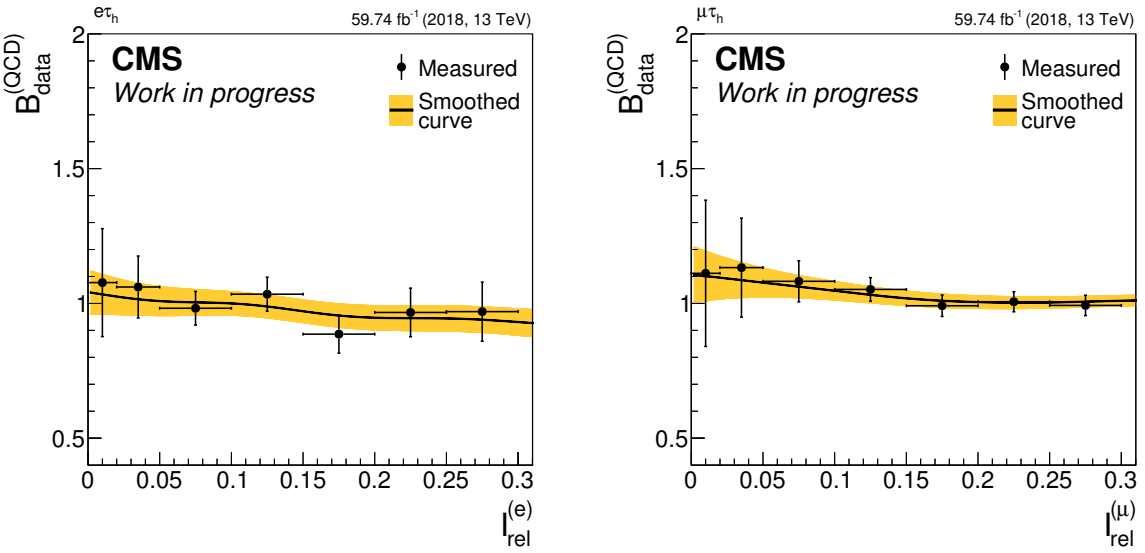


Figure 8.17: The bias corrections of $F_{\mathrm{F}}^{(\mathrm{QCD})}$ as a function of $I_{\mathrm{rel}}^{(\ell)}$ are shown for the 2018 data-taking period. On the left the $\mathrm{e}\tau_{\mathrm{h}}$ channel, and on the right the $\mu\tau_{\mathrm{h}}$ channel is displayed. The measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

of the SS $\rightarrow$ OS bias correction. One is denoted as $\mathrm{DR}_{\mathrm{QCD,aiso,OS}}$ and is shown with a dashed-lined frame in Figure 8.16. The other one is referred to as $\mathrm{DR}_{\mathrm{QCD,aiso,SS}}$ and is depicted as a box with a dotted frame in Figure 8.16. Both regions require less isolated light leptons, with

$$I_{\mathrm{rel}}^{(\ell)} \in [0.15, 0.25] \ , \tag{8.31}$$

and hence carry an extra subscript "aiso" standing for *anti-isolated* light leptons. The region $\mathrm{DR}_{\mathrm{QCD,aiso,SS}}$ has a further subscript "SS" reflecting the same-sign electric charge requirement

$$q^{(\ell)} \cdot q^{(\tau_{\mathrm{h}})} > 0 \ , \tag{8.32}$$

whereas $\mathrm{DR}_{\mathrm{QCD,aiso,OS}}$ has the OS selection applied

$$q^{(\ell)} \cdot q^{(\tau_{\mathrm{h}})} < 0 \ . \tag{8.33}$$

First, a data-driven F_{F} inside $\mathrm{DR}_{\mathrm{QCD,aiso,SS}}$ is measured in complete analogy to $F_{\mathrm{F,data}}^{(\mathrm{QCD})}$ in Equation (8.25).

$$F_{\mathrm{F,data,aiso}}^{(\mathrm{QCD})}(p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}, N_{\mathrm{jets}}) \ , \tag{8.34}$$

with the only difference being the DR. For $F_{\mathrm{F,data}}^{(\mathrm{QCD})}$ the region $\mathrm{DR}_{\mathrm{QCD}}$ is used, while for $F_{\mathrm{F,data,aiso}}^{(\mathrm{QCD})}$ $\mathrm{DR}_{\mathrm{QCD,aiso,SS}}$ is used (see also Figure 8.16). A closure correction is then derived in analogy to Equation (8.27),

$$C_{\mathrm{data,aiso}}^{(\mathrm{QCD})}(p_{\mathrm{T}}^{(\ell)}) \ . \tag{8.35}$$

The idea of the SS $\rightarrow$ OS correction is to use Equations (8.34) and (8.35), which are derived inside a SS region and apply them inside an OS region. More specifically, a prediction of QCD multijet events inside $\mathrm{DR}_{\mathrm{QCD,aiso,OS}}^{\text{SR-like}}$ is obtained by adapting Equation (8.16) to the following expression

$$N_{\mathrm{pred}}^{(\text{SR-like})} = \sum_{\epsilon \in \mathrm{DR}_{\mathrm{QCD,aiso,OS}}^{\text{AR-like}}} F_{\mathrm{F,data,aiso}}^{(\mathrm{QCD})} \cdot C_{\mathrm{data,aiso}}^{(\mathrm{QCD})} \cdot p(\mathrm{QCD}|\mathrm{DR}_{\mathrm{QCD,aiso,OS}}^{\text{AR-like}}) \ . \tag{8.36}$$

The ratio between expected and predicted distribution, binned in m_{vis}, defines the SS $\rightarrow$ OS correction

$$B^{(\text{QCD})}_{\text{data,SS}\rightarrow\text{OS}}(m_{\text{vis}}) \ . \tag{8.37}$$

The expected QCD multijet distribution inside $\text{DR}^{\text{SR-like}}_{\text{QCD,aiso,OS}}$ is calculated by subtracting contributions from observed data. These contributions are derived by means of MC simulation and τ embedding. However, the modeling of $I^{(\ell)}_{\text{rel}}$ by τ-embedded samples in this particular phase space – consisting of less isolated light leptons – is not satisfactory. Figure 8.18 illustrates the discrepancy between τ-embedded samples and MC simulation inside $\text{DR}^{\text{SR-like}}_{\text{QCD,aiso,OS}}$, where the predicted yield by τ-embedded samples is higher. The yield from τ-embedded samples can get even higher than the observed yield. In such extreme cases, the expected QCD multijet yield becomes negative. Comparing the m_{vis} distributions from τ-embedded and simulated samples in the SR however, shows a good agreement as demonstrated in Figure 8.19. A comparison of $I^{(\ell)}_{\text{rel}}$ is shown in the same figure, where a trend is clearly visible. A plausible explanation of this mis-match could lie in the τ-embedding technique itself (see Section 6.4). Since the embedding of simulated tau pairs is not done on the most fundamental level, biases in the calculation of isolation variables – as found in this particular case – can not be excluded. Consequently, for the derivation of the SS $\rightarrow$ OS correction, solely MC simulation is used for subtracting other contributions from data, i.e. τ-embedded samples are replaced by MC simulations of genuine $\text{Z}\rightarrow\tau\tau$, $\text{t}\bar{\text{t}}$ and VV events. The SS $\rightarrow$ OS correction is shown in Figure 8.20 for the 2018 data-taking period. Its values are close to one over the whole m_{vis} spectrum, while the uncertainties are in the order of 10%.

Finally, using Equations (8.25), (8.27), (8.30) and (8.37), the combined $F^{(\text{QCD})}_{\text{F}}$ can be written as

$$F^{(\text{QCD})}_{\text{F}} = F^{(\text{QCD})}_{\text{F,data}}(p_{\text{T}}^{(\tau_{\text{h}})}, N_{\text{jets}}) \cdot C^{(\text{QCD})}_{\text{data}}(p_{\text{T}}^{(\ell)}) \cdot B^{(\text{QCD})}_{\text{data}}(I^{(\ell)}_{\text{rel}}) \cdot B^{(\text{QCD})}_{\text{data,SS}\rightarrow\text{OS}}(m_{\text{vis}}) \ , \tag{8.38}$$

which is also shown in Figure 8.4. In summary, $F^{(\text{QCD})}_{\text{F}}$ consists of a raw F_{F}, which is measured inside three different N_{jets} categories. Three multiplicative corrections are applied to the raw $F^{(\text{QCD})}_{\text{F}}$. The first one is correcting for non-closures in $p_{\text{T}}^{(\ell)}$ inside DR_{QCD}. The other two corrections reduce potential biases, introduced by measuring $F^{(\text{QCD})}_{\text{F}}$ in SS region with altered $I^{(\ell)}_{\text{rel}}$ selection compared to the SR.

The $\text{t}\bar{\text{t}}$ Fake Factor

No suitably enriched $\text{t}\bar{\text{t}}$ DR with a sufficient event count which populates a similar phase space as the SR in terms of misidentified τ_{h} candidates can be identified. The contribution of $\text{t}\bar{\text{t}}$ processes to the jet $\rightarrow \tau_{\text{h}}$ background is however sub-dominant to QCD multijet and W+jets as can be seen from Figure 8.2. The idea is to not just take the $\text{t}\bar{\text{t}}$ contribution inside the SR estimated by MC simulation but measure $\text{t}\bar{\text{t}}$ F_{F} based on MC simulation and correct it on data inside a $\text{t}\bar{\text{t}}$-enriched *validation region*. The measurement of the raw $\text{t}\bar{\text{t}}$ F_{F} – labeled "1" in Figure 8.4 – and its closure correction – labeled "2" in Figure 8.4 – work the same way as for $F^{(\text{W+jets})}_{\text{F}}$ and $F^{(\text{QCD})}_{\text{F}}$ discussed in the previous sections. Due to the extremely low event count in the 0-jet category, the $\text{t}\bar{\text{t}}$ F_{F}

$$\begin{aligned} F^{(\text{t}\bar{\text{t}})}_{\text{F}} &= \frac{N^{(\texttt{tight})}_{\text{t}\bar{\text{t}}}}{N^{(\texttt{vloose} \ \wedge \ \neg\texttt{tight})}_{\text{t}\bar{\text{t}}}} \\ &\equiv F^{(\text{t}\bar{\text{t}})}_{\text{F,MC}}(p_{\text{T}}^{(\tau_{\text{h}})}, N_{\text{jets}}) \ , \end{aligned} \tag{8.39}$$

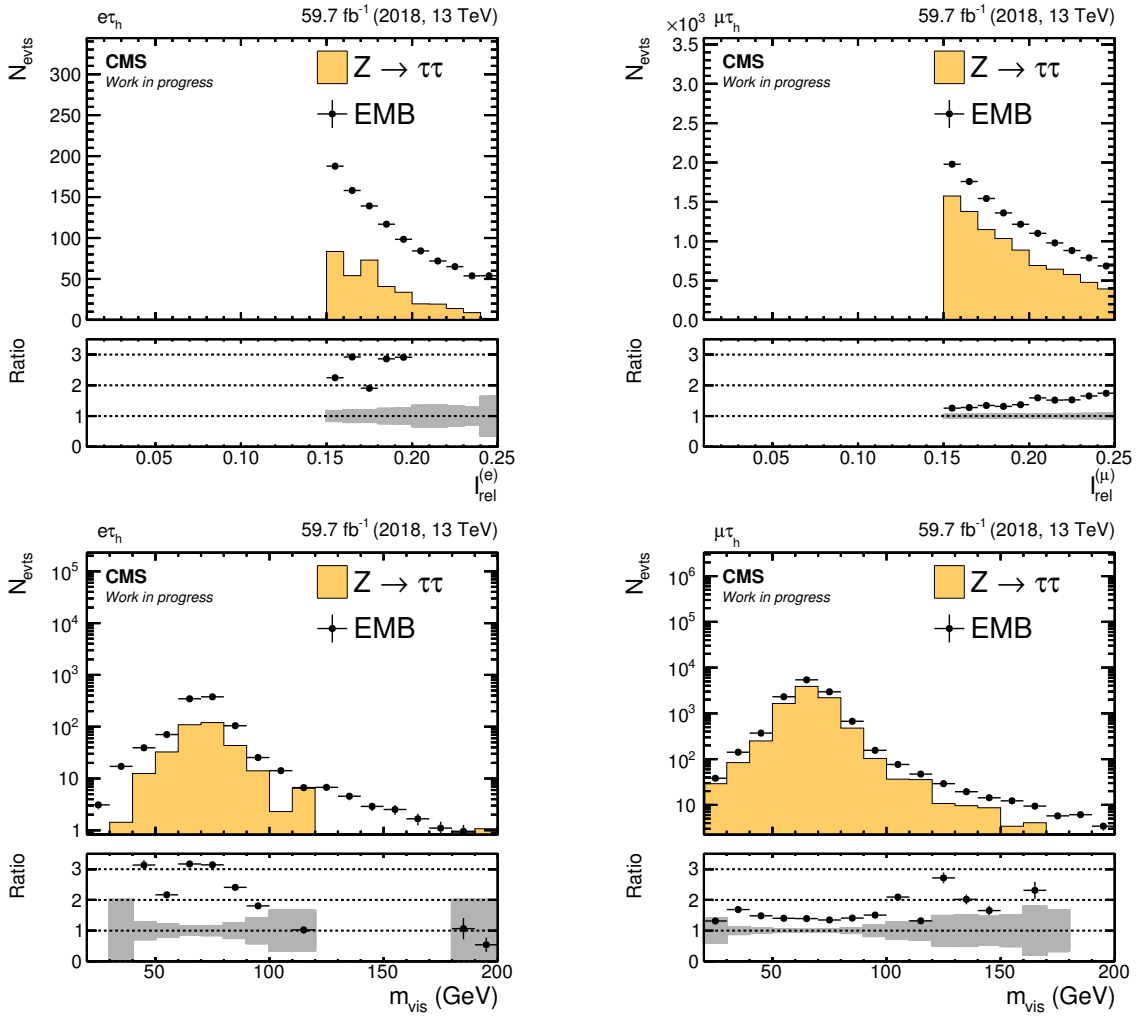


Figure 8.18: Distributions, showing the discrepancy between MC simulation ($Z \to \tau\tau$) and τ-embedded samples (EMB) inside $DR^{\text{SR-like}}_{\text{QCD,aiso,OS}}$, are presented. The top row shows the distributions of $I^{(\ell)}_{\text{rel}}$ and the bottom row the m_{vis} distributions – on the left for the $e\tau_h$ channel and on the right for the $\mu\tau_h$ channel. In all displayed distributions, τ-embedded contributions over-predict those coming from $Z \to \tau\tau$ simulation. This picture does not change when contributions of genuine $t\bar{t}$ and VV processes are taken into account by means of simulation because the dominant part of τ-embedded samples are made of $Z \to \tau\tau$ events. Only statistical uncertainties are shown here.

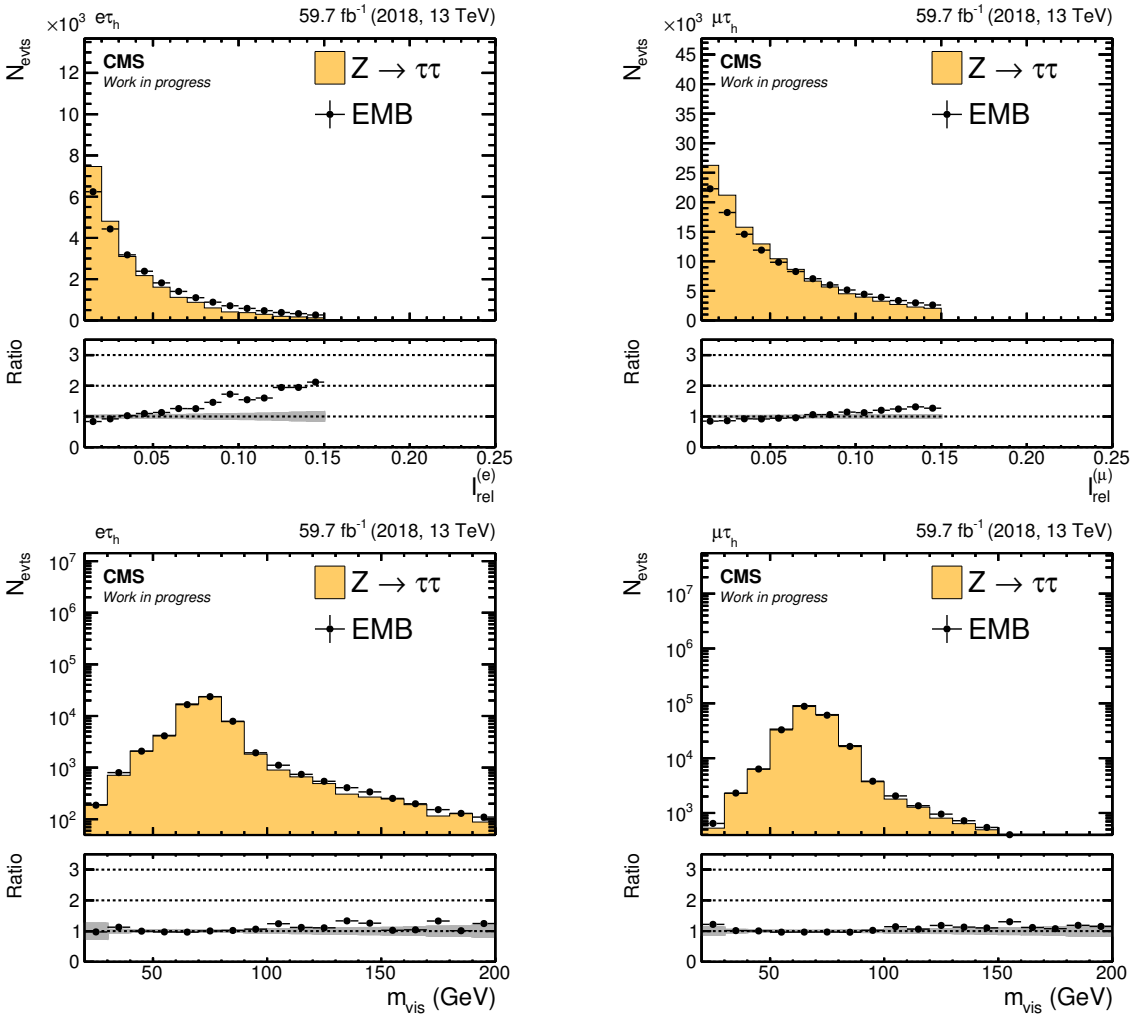


Figure 8.19: Distributions of MC simulation ($Z \to \tau\tau$) and τ embedded samples (EMB) in the SR are presented. The top row shows the distributions of $I_{\text{rel}}^{(\ell)}$ and the bottom row the m_{vis} distributions – on the left for the $e\tau_h$ channel and on the right for the $\mu\tau_h$ channel. The m_{vis} distributions show a good agreement within uncertainties between the two estimation techniques. However, a clear trend is visible in $I_{\text{rel}}^{(\ell)}$ distributions, showing that contribution estimated from τ-embedded samples over-predict those from MC simulation for higher isolation values. This picture does not change when contributions of genuine $t\bar{t}$ and VV processes are taken into account by means of simulation because the dominant part of τ-embedded samples are made of $Z \to \tau\tau$ events. Only statistical uncertainties are shown here.

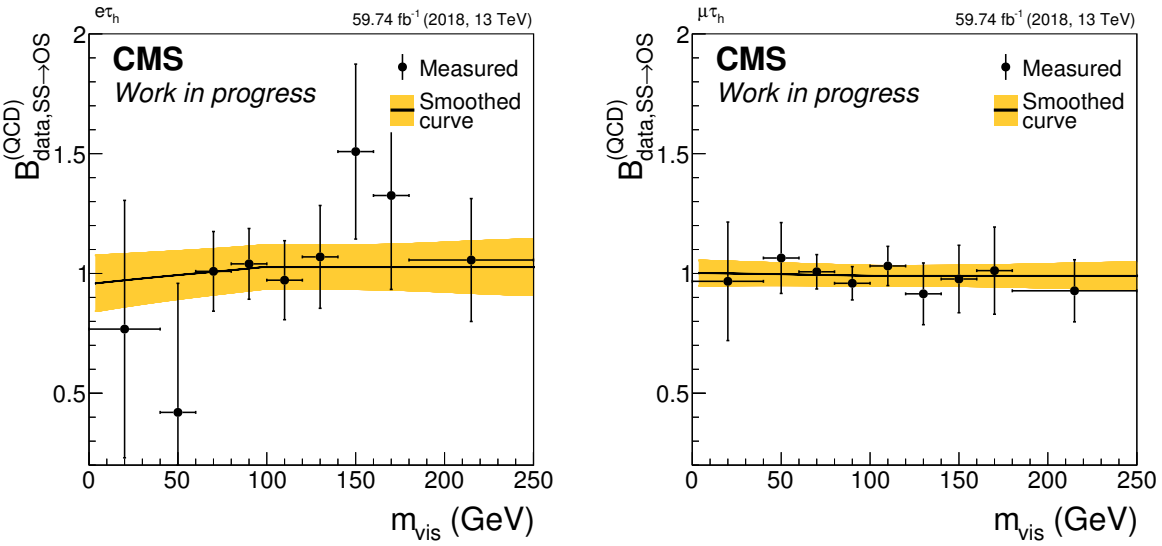


Figure 8.20: The SS → OS bias corrections of $F_{\mathrm{F}}^{(\mathrm{QCD})}$ as a function of m_{vis} are shown for the 2018 data-taking period. On the left, the eτ_{h} channel and on the right, the $\mu\tau_{\mathrm{h}}$ channel is displayed. The measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

is measured in two N_{jets} categories only, $N_{\mathrm{jets}} \leq 1$ and $N_{\mathrm{jets}} \geq 2$. The closure correction is derived and binned in m_{vis}

$$C^{(\mathrm{t\bar{t}})} \equiv C_{\mathrm{MC}}^{(\mathrm{t\bar{t}})}(m_{\mathrm{vis}}) \ . \tag{8.40}$$

Raw $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ together with the corresponding closure corrections are shown in Figure 8.21 for 2018 data and both eτ_{h} and $\mu\tau_{\mathrm{h}}$ channels, respectively.

The derivation of SF, correcting $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$, is labeled with "3" in Figure 8.4. It is derived inside a t$\bar{\mathrm{t}}$-enriched validation region that is defined via the following selection:

- At least two jets, $N_{\mathrm{jets}} \geq 2$. No restriction on N_{jets} is applied in the SR.
- At least one b-tagged jet, $N_{\text{b-tag}} \geq 1$. No selection on $N_{\text{b-tag}}$ is applied in the SR.
- Failing the third-lepton veto, i.e. at least one isolated electron and one isolated muon are required. The third lepton veto is part of the SR definition (see Table 6.5).

Even though the above selection results in a region enriched in t$\bar{\mathrm{t}}$ events, the overall event count is too low to perform a F_{F} measurement in several $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ bins and that is why $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ is derived by means of simulation in the first place. However, global fake factors, i.e. a single number, can be extracted based on simulation or in a data-driven manner. The ratio of the global F_{F}'s is used as scale factor correcting $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ as explained in the following and is expected to cover the largest part of data versus simulation discrepancies.

The simulation based F_{F} is defined as

$$F_{\mathrm{F,MC,global}}^{(\mathrm{t\bar{t}})} = \frac{N_{\mathrm{t\bar{t}}(\mathrm{jet}\to\tau_{\mathrm{h}})}^{(\texttt{tight})}}{N_{\mathrm{t\bar{t}}(\mathrm{jet}\to\tau_{\mathrm{h}})}^{(\texttt{vloose} \ \wedge \ \neg\texttt{tight})}} \ , \tag{8.41}$$

emphasizing that only those parts of simulated t$\bar{\mathrm{t}}$ events are used that are matched to the case of jet $\to \tau_{\mathrm{h}}$ (see Section 6.3). Similarly, a data-driven F_{F} can be derived inside

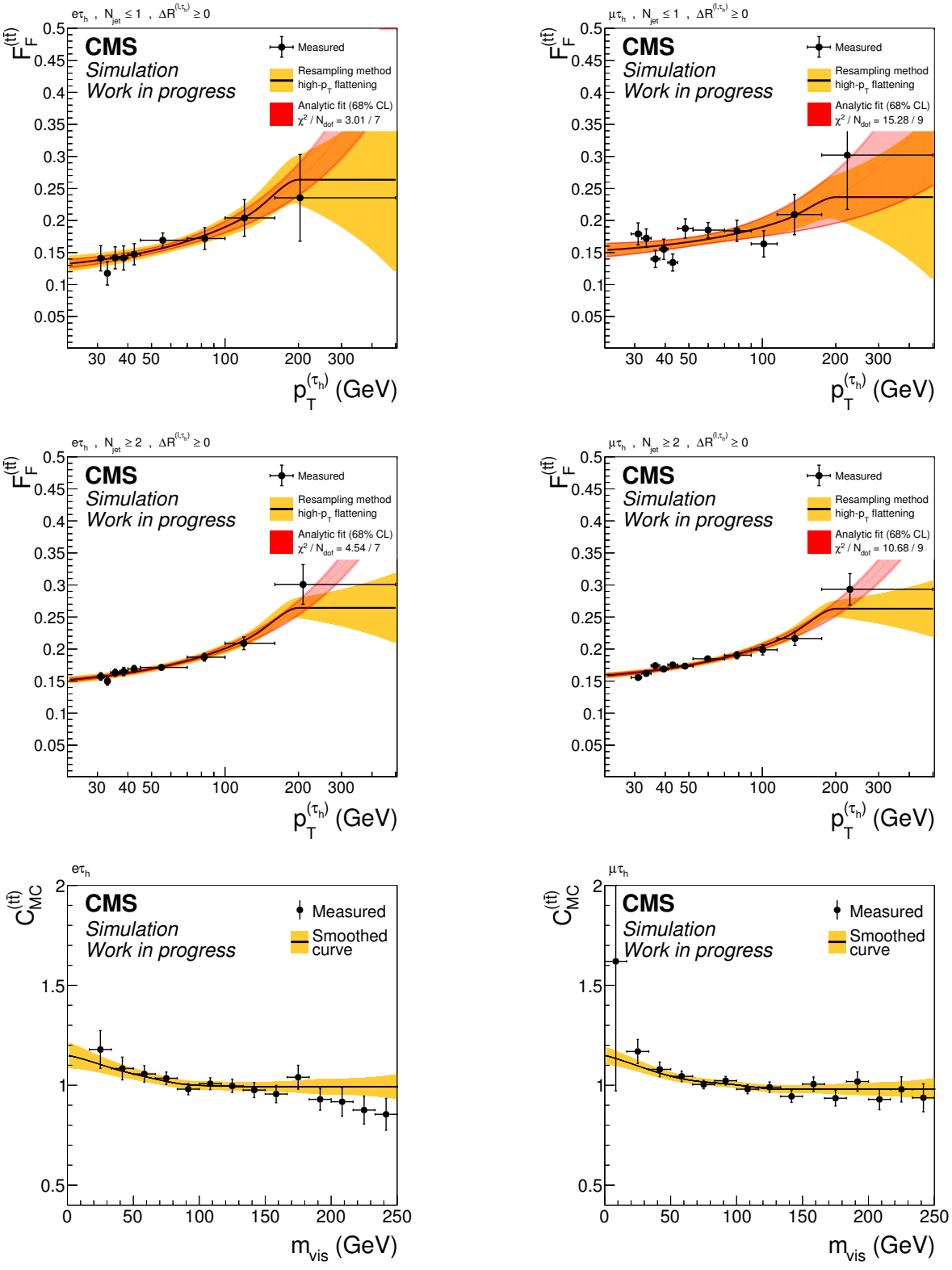


Figure 8.21: The quantity $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and its closure correction as a function of m_{vis} are shown, using simulated events from the 2018 data-taking period. On the left, distributions for the $\mathrm{e}\tau_{\mathrm{h}}$ channel are displayed and corresponding distributions for the $\mu\tau_{\mathrm{h}}$ channel are shown on the right. The top and middle row show $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ for the $N_{\mathrm{jets}} \leq 1$ and $N_{\mathrm{jets}} \geq 2$ category, respectively. In the analysis, the solid line is used together with its associated uncertainty band shown in yellow. The bottom row shows the closure correction.

$\mathrm{SF}_{\mathrm{data}}^{(\mathrm{t\bar{t}})}$	2016	2017	2018
$\mathrm{e}\tau_{\mathrm{h}}$	0.87 ± 0.15	0.56 ± 0.15	1.09 ± 0.16
$\mu\tau_{\mathrm{h}}$	1.06 ± 0.10	0.85 ± 0.10	0.92 ± 0.09

Table 8.1: The values of $\mathrm{SF}_{\mathrm{data}}^{(\mathrm{t\bar{t}})}$ – defined in Equation (8.43) – are given for all years of data taking from 2016 to 2018 and both semi-leptonic channels.

the $\mathrm{t\bar{t}}$-enriched validation region by subtracting all processes but $\mathrm{t\bar{t}}(\mathrm{jet} \to \tau_{\mathrm{h}})$ from data (see also Equation (8.9))

$$F_{\mathrm{F,data,global}}^{(\mathrm{t\bar{t}})} = \frac{N_{\mathrm{data}}^{(\texttt{tight})} - \sum_{!\mathrm{t\bar{t}}(\mathrm{jet}\to\tau_{\mathrm{h}})} N_{\mathrm{other}}^{(\texttt{tight})}}{N_{\mathrm{data}}^{(\texttt{vloose} \wedge \neg\texttt{tight})} - \sum_{!\mathrm{t\bar{t}}(\mathrm{jet}\to\tau_{\mathrm{h}})} N_{\mathrm{other}}^{(\texttt{vloose} \wedge \neg\texttt{tight})}} \,. \tag{8.42}$$

The correction scale factor is defined as the ratio of Equation (8.42) over Equation (8.41)

$$\mathrm{SF}_{\mathrm{data}}^{(\mathrm{t\bar{t}})} = \frac{F_{\mathrm{F,data,global}}^{(\mathrm{t\bar{t}})}}{F_{\mathrm{F,MC,global}}^{(\mathrm{t\bar{t}})}} \,. \tag{8.43}$$

Values of $\mathrm{SF}_{\mathrm{data}}^{(\mathrm{t\bar{t}})}$ are quoted in Table 8.1 for all years of data-taking and both semi-leptonic channels.

The combined $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ as shown in Figure 8.4 is given by

$$F_{\mathrm{F}}^{(\mathrm{t\bar{t}})} = F_{\mathrm{F,MC}}^{(\mathrm{t\bar{t}})}(p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}, N_{\mathrm{jets}}) \cdot C_{\mathrm{MC}}^{(\mathrm{t\bar{t}})}(m_{\mathrm{vis}}) \cdot \mathrm{SF}_{\mathrm{data}}^{(\mathrm{t\bar{t}})} \,. \tag{8.44}$$

In summary, $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ consists of a raw F_{F}, which is measured inside two different N_{jets} categories. Two multiplicative corrections are applied to the raw $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$. The first one is correcting for small non-closures in m_{vis} inside the SR. Raw $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ and its closure correction are calculated from $\mathrm{t\bar{t}}$ simulation. The second correction consists of a scale factor defined as ratio of data-driven and simulation-based F_{F} measured inside a $\mathrm{t\bar{t}}$-enriched validation region.

8.2.2 Fake Factors in the Fully Hadronic Final State

As can be seen from Figure 8.2, QCD multijet processes make up over 90% of the jet $\to \tau_{\mathrm{h}}$ contribution in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel. Therefore, in the fully hadronic channel only a single F_{F} is used which is derived in a QCD multijet enriched DR. This F_{F} is then not only applied to QCD multijet events, but also to W+jets and $\mathrm{t\bar{t}}$ events. The main challenge in the derivation and application of F_{F} in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel is that either of the τ_{h} or even both τ_{h} candidates can be due to a misidentified jet. Let $(\tau_{\mathrm{h}}^{(1)}, \tau_{\mathrm{h}}^{(2)})$ be the tau pair as coming from the pair selection discussed in Section 6.1. Hence, the leading τ_{h} would be $\tau_{\mathrm{h}}^{(1)}$ and $\tau_{\mathrm{h}}^{(2)}$ the sub-leading one. One would like to measure F_{F} with respect to both τ_{h} candidates. Technically this is achieved by considering each event twice, whereby the position of the τ_{h} candidates within a tau pair is switched: $(\tau_{\mathrm{h}}^{(2)}, \tau_{\mathrm{h}}^{(1)})$. The tau-ID requirement is always applied on the leading τ_{h} candidate. This way the F_{F} measurement correctly tests both hypothesis of $\tau_{\mathrm{h}}^{(1)}$ being fake, or $\tau_{\mathrm{h}}^{(2)}$ being fake. Since every event is used twice in the F_{F} derivation, each event gets weighted with an extra factor of 0.5.

The DR used to derive the QCD multijet F_{F} differs only by one criterion from the SR definition:

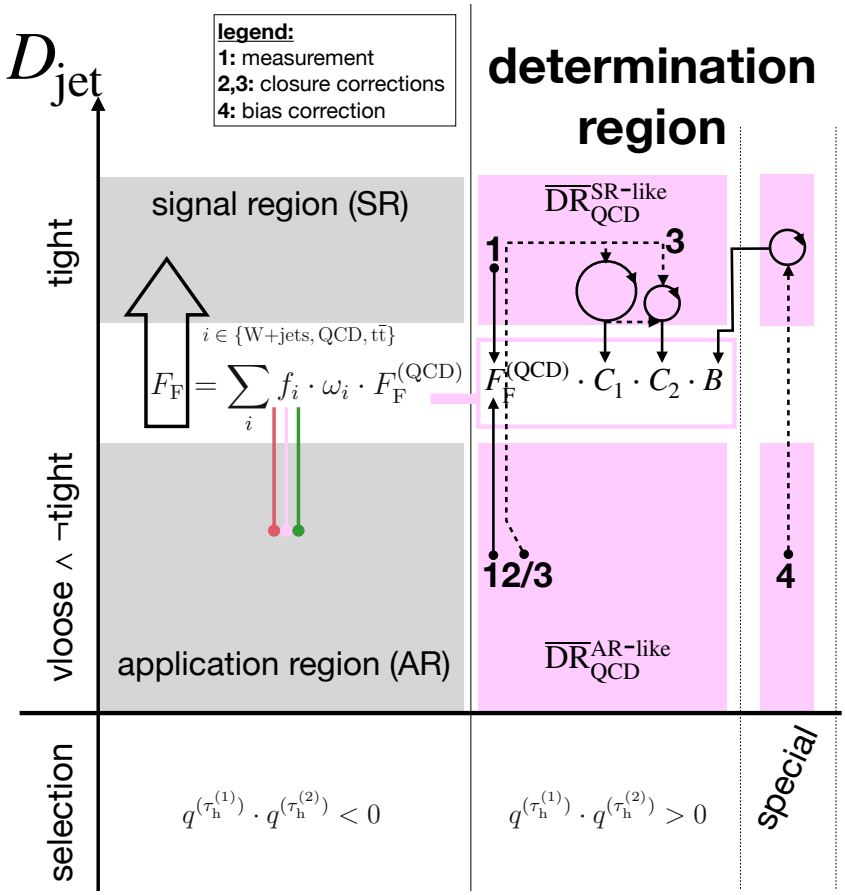


Figure 8.22: Schematic of the FF method for the fully hadronic channel. The y-axis shows the output of D_{jet} defining the SR and AR, which are shown in gray boxes. In the $\tau_{\text{h}}\tau_{\text{h}}$ channel, a single DR – denoted as $\overline{\text{DR}}_{\text{QCD}}$ – is defined. The bar in the notation is used to not confuse it with DR_{QCD} from Figure 8.4. The F_{F} measurement is labeled with "1", closure corrections are labeled with "2" and "3". Label "4" represents the bias correction. At the bottom, the different selections applied with respect to the SR are listed in order to enrich QCD multijet events. For the bias correction, a special region is used, whose selections are given in the text. The final F_{F} used to extrapolate events from AR to SR, is a weighted sum consisting of fractions (f_i), a combinatorial factor (ω_i) and $F_{\text{F}}^{(\text{QCD})}$.

- The charges of the tau pair are required to have SS

$$q^{(\tau_{\text{h}}^{(1)})} \cdot q^{(\tau_{\text{h}}^{(2)})} > 0 \,. \tag{8.45}$$

Figure 8.22 shows the FF scheme for the $\tau_{\text{h}}\tau_{\text{h}}$ channel. The DR in case of the $\tau_{\text{h}}\tau_{\text{h}}$ channel is denoted by $\overline{\text{DR}}_{\text{QCD}}$ with an extra bar to not confuse it with DR_{QCD} used for the semi-leptonic channels.

The derivation of the raw $F_{\text{F}}^{(\text{QCD})}$– labeled as "1" in Figure 8.22 – works the same way as for the semi-leptonic channels. Furthermore, the derivation of a first closure correction – labeled as "2" in Figure 8.22 – is in complete analogy to the semi-leptonic channels. In step "3", a further closure correction – labeled as C_2 in Figure 8.22 – is derived. Step "3" is thus different to the semi-leptonic channels. Step "4" represents the SS $\rightarrow$ OS correction, needed to correct for biases introduced by measuring $F_{\text{F}}^{(\text{QCD})}$ inside a SS DR and applying it to an OS AR. In the following, each of the steps is discussed in more detail.

Events are assigned to the SR-like $\overline{\text{DR}}_{\text{QCD}}$ – denoted as $\overline{\text{DR}}_{\text{QCD}}^{\text{SR-like}}$ – if both τ_{h} pass the `tight` condition on the D_{jet} discriminant. The AR-like part – denoted as $\overline{\text{DR}}_{\text{QCD}}^{\text{AR-like}}$

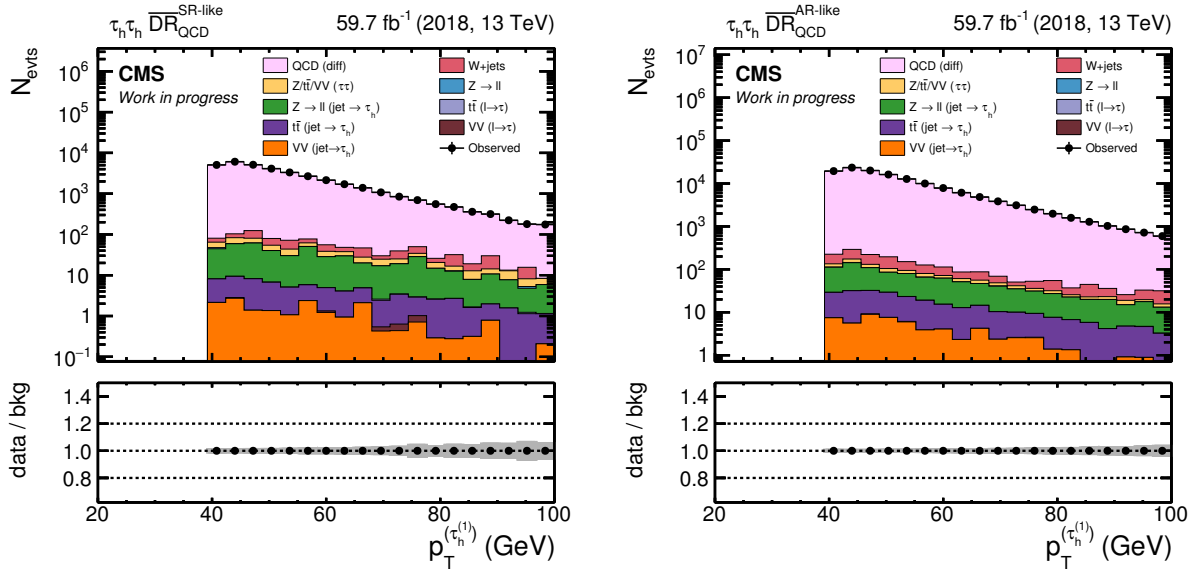


Figure 8.23: Shown are $p_{\mathrm{T}}^{(\tau_1)}$ distributions inside $\overline{\mathrm{DR}}_{\mathrm{QCD}}$ for the 2018 data-taking period. The distributions are measured for the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel. On the left, the SR-like part of $\overline{\mathrm{DR}}_{\mathrm{QCD}}$ is shown and on the right, the AR-like part. In all plots, the QCD multijet process is at the top of the stacked histograms. The estimated yield from QCD multijet events is given by the difference between data and the sum of all other background processes. It is the dominating process, while all other processes combined contribute on a sub-percent level. The QCD multijet estimate is simply taken as the difference between the stacked histograms and the observation. Processes labeled as $\mathrm{Z/t\bar{t}}/VV\,(\tau\tau)$ are estimated by the τ-embedding technique. Only statistical uncertainties are shown here.

– is defined by the sub-leading τ_{h} candidate passing the `tight` condition on the D_{jet} discriminant while the leading τ_{h} candidate fails it but passes the `vloose` condition. The transverse momentum distributions of the leading τ_{h}, $p_{\mathrm{T}}^{(\tau_1)}$, inside $\overline{\mathrm{DR}}_{\mathrm{QCD}}$ are shown in Figure 8.23, demonstrating that mainly QCD multijet events populate this region.

The raw $F_{\mathrm{F}}^{(\mathrm{QCD})}$ is calculated according to Equation (8.25) and for three different N_{jets} categories, $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} = 1$ and $N_{\mathrm{jets}} \geq 2$

$$F_{\mathrm{F}}^{(\mathrm{QCD})} = F_{\mathrm{F,data}}^{(\mathrm{QCD})}(p_{\mathrm{T}}^{(\tau_1)}, N_{\mathrm{jets}}) \ . \tag{8.46}$$

For the parametrization, a *third order* polynomial is used which is truncated at transverse momenta of the leading τ_{h} candidate of 80 GeV. The truncation is needed to avoid unrealistic F_{F} values at high p_{T} because the third order polynomial falls steeply in this region. The associated uncertainty is obtained by the same means as for all other raw F_{F}'s. After generating toys from the measurement points, the fit procedure is repeated for each toy data set and the uncertainty band is extracted from the ensemble of fitted third order polynomials. However, in the generation of the toy data set, each measurement point is allowed to fluctuate by three times the statistical uncertainty of the actual measured value. This choice is expected to cover potential biases introduced by picking a more complex fit function in case of the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel compared to the linear fit used in the semi-leptonic channels.

Following Equation (8.27), a closure correction dependent on m_{vis} is calculated. This closure correction is denoted as C_1 in Figure 8.22 and is given by

$$C_{\mathrm{data}}^{(\mathrm{QCD})}(m_{\mathrm{vis}}) \ . \tag{8.47}$$

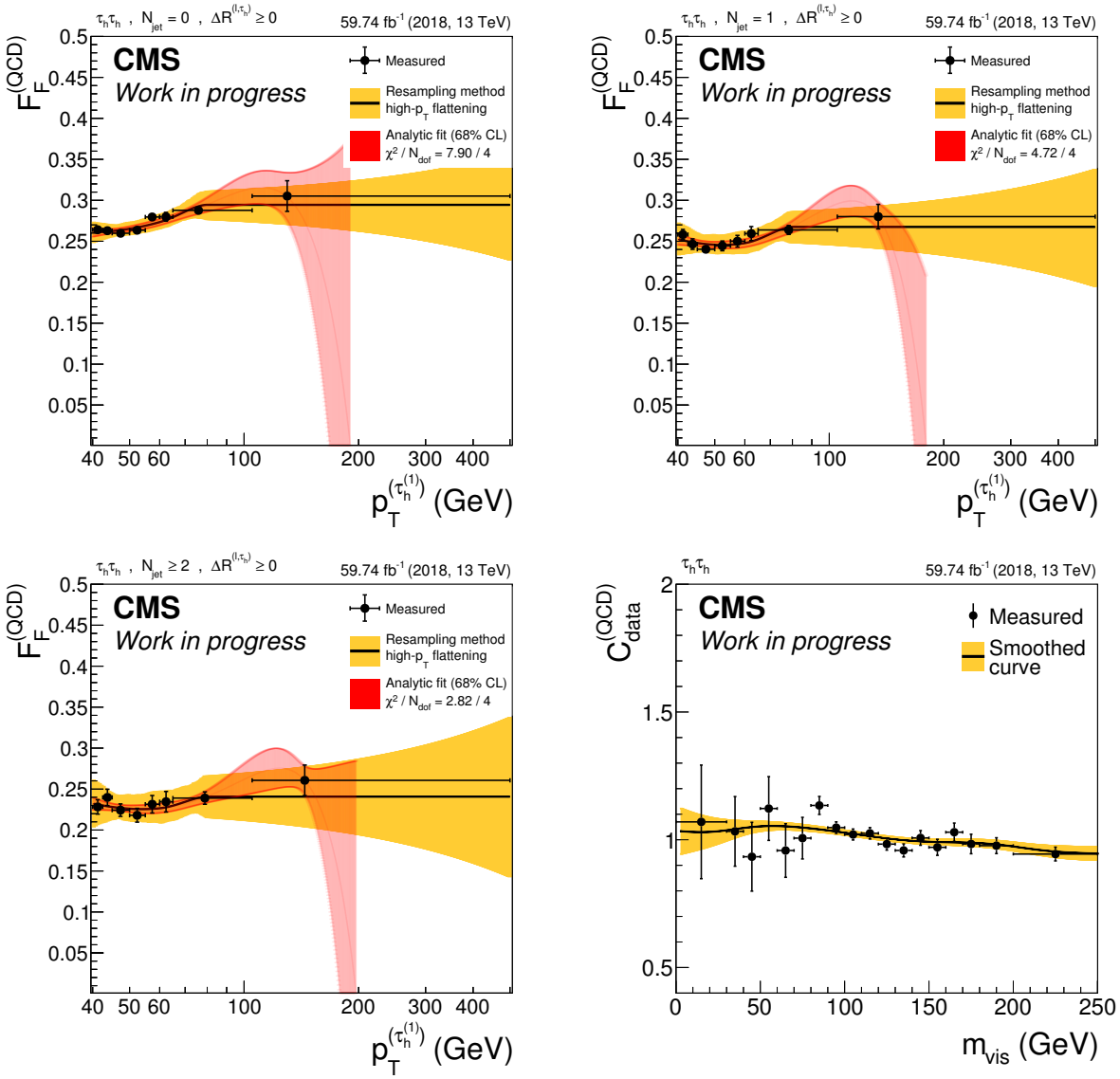


Figure 8.24: The quantity $F_{\mathrm{F}}^{(\mathrm{QCD})}$ in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel as a function of $p_{\mathrm{T}}^{(\tau_1)}$ is shown using 2018 data. The top row shows $F_{\mathrm{F}}^{(\mathrm{QCD})}$ for the $N_{\mathrm{jets}} = 0$ (left) category and for the $N_{\mathrm{jets}} = 1$ (right) category. Bottom left shows $F_{\mathrm{F}}^{(\mathrm{QCD})}$ for the $N_{\mathrm{jets}} \geq 2$ category. The F_{F} parametrization is shown as a solid line. It consists of a third order polynomial which is truncated at high $p_{\mathrm{T}}^{(\tau_1)}$ and replaced by a constant to avoid unrealistic, i.e. negative, F_{F} values at high values of $p_{\mathrm{T}}^{(\tau_1)}$. In the analysis, the solid line is used together with its associated uncertainty band shown in yellow. Lastly, on the bottom right, the closure correction as a function of m_{vis} is displayed. The measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

Figure 8.24 shows the raw $F_{\mathrm{F}}^{(\mathrm{QCD})}$ as well as the closure correction in m_{vis} using 2018 data. An additional closure correction is derived as a function of the transverse momentum of the sub-leading τ_{h} candidate, $p_{\mathrm{T}}^{(\tau_2)}$. For that, raw $F_{\mathrm{F}}^{(\mathrm{QCD})}$ and its closure correction (C_1 from Equation (8.47)) are utilized to estimate the $p_{\mathrm{T}}^{(\tau_2)}$ distribution inside $\overline{\mathrm{DR}}_{\mathrm{QCD}}^{\text{SR-like}}$ by adapting Equation (8.16)

$$N_{\mathrm{pred}}^{(\text{SR-like})} = \sum_{\epsilon \in \overline{\mathrm{DR}}_{\mathrm{QCD}}^{\text{AR-like}}} F_{\mathrm{F}}^{(\mathrm{QCD})} \cdot C_{\mathrm{data}}^{(\mathrm{QCD})} \cdot p(\mathrm{QCD}|\overline{\mathrm{DR}}_{\mathrm{QCD}}^{\text{AR-like}}) \ . \tag{8.48}$$

The ratio of the expected over predicted $p_{\mathrm{T}}^{(\tau_2)}$ distribution defines the correction. It is derived in two separate N_{jets} categories, $N_{\mathrm{jets}} = 0$ and $N_{\mathrm{jets}} \geq 1$. Highlighting the dependencies of the correction, it can be written as

$$C_{\mathrm{data}}^{(\mathrm{QCD})}(p_{\mathrm{T}}^{(\tau_2)}, N_{\mathrm{jets}}) \ . \tag{8.49}$$

Figure 8.25 (top row) shows the closure corrections for both N_{jets} categories using 2018 data. The corrections between the two N_{jets} categories clearly differ, justifying why they are measured separately.

In step "4" of Figure 8.22, the SS $\rightarrow$ OS bias correction is derived. Similar to the semi-leptonic channels, this correction is derived in an *anti-isolated* (aiso) region. For the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel, the anti-isolation is defined by events where the sub-leading τ_{h} candidate fails the `tight` condition on the D_{jet} discriminant. The anti-isolated region is split in SS and OS, denoted as $\overline{\mathrm{DR}}_{\mathrm{QCD,aiso,SS}}$ and $\overline{\mathrm{DR}}_{\mathrm{QCD,aiso,OS}}$, respectively. Figure 8.26 shows all regions used for the SS $\rightarrow$ OS correction and their relation to the SR and $\overline{\mathrm{DR}}_{\mathrm{QCD}}$. In analogy to Equation (8.34), a F_{F} is measured inside $\overline{\mathrm{DR}}_{\mathrm{QCD,aiso,SS}}$

$$F_{\mathrm{F,data,aiso}}^{(\mathrm{QCD})}(p_{\mathrm{T}}^{(\tau_1)}, N_{\mathrm{jets}}) \ . \tag{8.50}$$

Next, a closure correction is derived

$$C_{\mathrm{data,aiso}}^{(\mathrm{QCD})}(m_{\mathrm{vis}}) \ . \tag{8.51}$$

The above F_{F} together with its closure correction are applied to OS-events inside $\overline{\mathrm{DR}}_{\mathrm{QCD,aiso,OS}}^{\text{AR-like}}$. Changing Equation (8.16) appropriately yields

$$N_{\mathrm{pred}}^{(\text{SR-like})} = \sum_{\epsilon \in \overline{\mathrm{DR}}_{\mathrm{QCD,aiso,OS}}^{\text{AR-like}}} F_{\mathrm{F,data,aiso}}^{(\mathrm{QCD})} \cdot C_{\mathrm{data,aiso}}^{(\mathrm{QCD})} \cdot p(\mathrm{QCD}|\overline{\mathrm{DR}}_{\mathrm{QCD,aiso,OS}}^{\text{AR-like}}) \ , \tag{8.52}$$

The ratio between the expected and predicted distribution defines the SS $\rightarrow$ OS bias correction

$$B_{\mathrm{data,SS\rightarrow OS}}^{(\mathrm{QCD})}(m_{\mathrm{vis}}) \ . \tag{8.53}$$

The bias correction is shown in Figure 8.25 (bottom row) for the 2018 data-taking period.

The final $F_{\mathrm{F}}^{(\mathrm{QCD})}$ used for the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel consists of four multiplicative factors and is given by

$$F_{\mathrm{F}}^{(\mathrm{QCD})} = F_{\mathrm{F,data}}^{(\mathrm{QCD})}(p_{\mathrm{T}}^{(\tau_1)}, N_{\mathrm{jets}}) \cdot C_{\mathrm{data}}^{(\mathrm{QCD})}(m_{\mathrm{vis}}) \cdot C_{\mathrm{data}}^{(\mathrm{QCD})}(p_{\mathrm{T}}^{(\tau_2)}, N_{\mathrm{jets}}) \cdot B_{\mathrm{data,SS\rightarrow OS}}^{(\mathrm{QCD})}(m_{\mathrm{vis}}) \ . \tag{8.54}$$

In summary, $F_{\mathrm{F}}^{(\mathrm{QCD})}$ consists of a raw F_{F}, which is measured inside three different N_{jets} categories. In each of these categories the distribution as a function of $p_{\mathrm{T}}^{(\tau_1)}$ is fitted with a third order polynomial. The third order polynomial is replaced by a constant for high

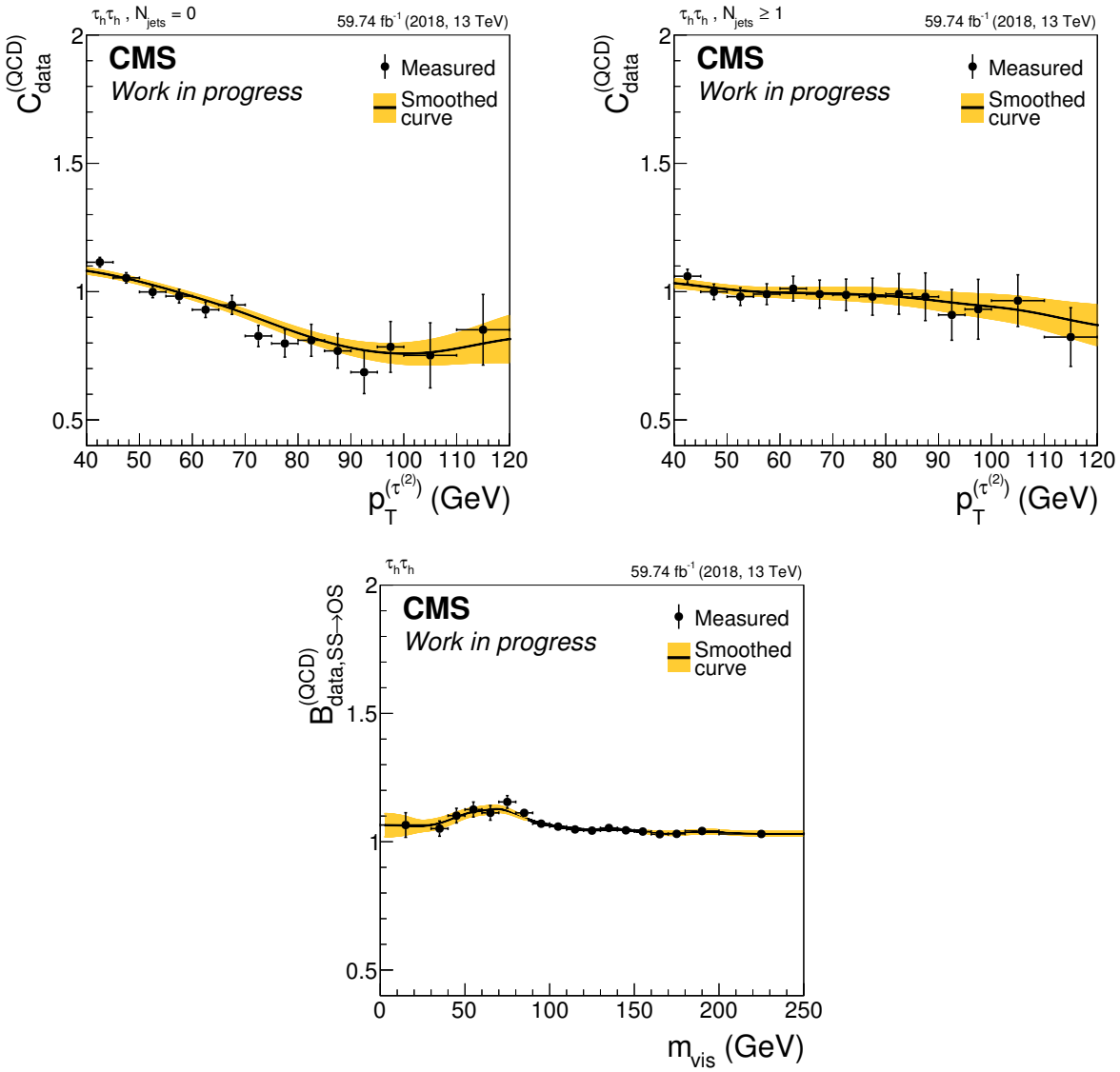


Figure 8.25: Corrections to $F_{\mathrm{F}}^{(\mathrm{QCD})}$ in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel using 2018 data are presented. The top row shows the closure correction in $p_{\mathrm{T}}^{(\tau_2)}$ derived separately in a $N_{\mathrm{jets}} = 0$ category (left) and $N_{\mathrm{jets}} \geq 1$ category (right). The bottom row depicts the SS $\to$ OS bias corrections as a function of m_{vis}. All measurements are smoothed with a Gaussian kernel of variable width and the resulting smoothed curves are used later in the F_{F} application. The uncertainty bands are obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

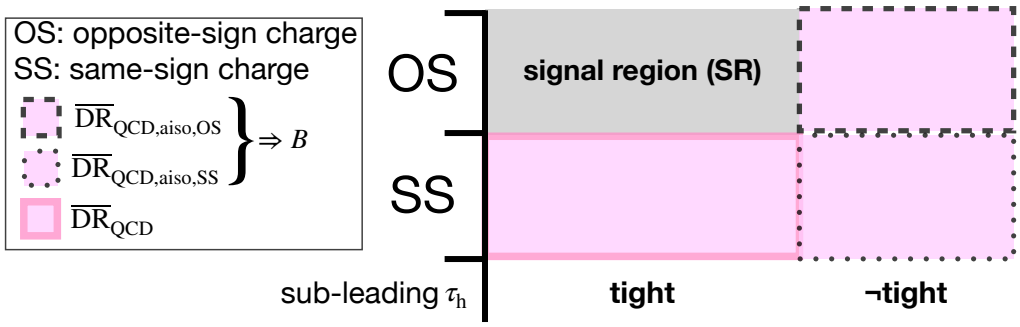


Figure 8.26: A schematic of all regions used to derive $F_{\text{F}}^{(\text{QCD})}$ and its corrections is given. The figure focuses on the relevant differences between these DRs and the SR. The y-axis shows the charge requirement imposed on the tau pair and on the x-axis the τ-ID requirement imposed on the sub-leading τ_{h} candidate. In gray, the SR is depicted. Furthermore, $\overline{\text{DR}}_{\text{QCD}}$ is shown which is used to derive the raw $F_{\text{F}}^{(\text{QCD})}$ and its closure corrections. The other regions are used to derive the bias correction, B (see also Figure 8.22).

$p_{\text{T}}^{(\tau_1)}$ values. Three multiplicative corrections are applied to the raw $F_{\text{F}}^{(\text{QCD})}$. The first one is correcting for non-closures in m_{vis} inside $\overline{\text{DR}}_{\text{QCD}}$. It amounts to a correction of $F_{\text{F}}^{(\text{QCD})}$ by just a few percent. A second closure correction as a function of $p_{\text{T}}^{(\tau_2)}$ is calculated inside $\overline{\text{DR}}_{\text{QCD}}$. This correction is measured in two different N_{jets} categories. Corrections for $N_{\text{jets}} = 0$ span a larger interval than in the case of the $N_{\text{jets}} \geq 1$ category, where the correction are typically inside $[0.9, 1.05]$. The third correction mitigates potential biases introduced by measuring $F_{\text{F}}^{(\text{QCD})}$ inside a SS region but applying it to an OS region. It takes on positive values between 1.05 and 1.15 across the whole m_{vis} range.

Fake factors and their corrections for all channels involving a τ_{h} candidate can be found in Chapter C for the 2016 and 2017 data-taking period.

8.2.3 Fake Factor Application

The application of F_{F} differs between the semi-leptonic and fully hadronic channels. Therefore, the discussion is split and the semi-leptonic case is explained first.

Each F_{F} component for the semi-leptonic channels is discussed thoroughly in Section 8.2.1. In summary, the three F_{F} components – corresponding to the dominant jet $\rightarrow \tau_{\text{h}}$ processes W+jets, QCD multijet and $\text{t}\bar{\text{t}}$ – and their dependencies are given by

$$\begin{aligned}
&\text{Eq. 8.21} \rightarrow F_{\text{F}}^{(\text{W+jets})}(p_{\text{T}}^{(\tau_{\text{h}})}, N_{\text{jets}}, \Delta R^{(\ell,\tau_{\text{h}})}, p_{\text{T}}^{(\ell)}, m_{\text{vis}}) \\
&\text{Eq. 8.38} \rightarrow F_{\text{F}}^{(\text{QCD})}(p_{\text{T}}^{(\tau_{\text{h}})}, N_{\text{jets}}, p_{\text{T}}^{(\ell)}, I_{\text{rel}}^{(\ell)}, m_{\text{vis}}) \\
&\text{Eq. 8.44} \rightarrow F_{\text{F}}^{(\text{t}\bar{\text{t}})}(p_{\text{T}}^{(\tau_{\text{h}})}, N_{\text{jets}}, m_{\text{vis}}) \ .
\end{aligned} \tag{8.55}$$

Ideally, one would like to apply $F_{\text{F}}^{(\text{W+jets})}$ to all W+jets events inside the AR, $F_{\text{F}}^{(\text{QCD})}$ to all QCD multijet events inside the AR and $F_{\text{F}}^{(\text{t}\bar{\text{t}})}$ to all $\text{t}\bar{\text{t}}(\text{jet} \rightarrow \tau_{\text{h}})$ events inside the AR. The chosen approach is to calculate the contribution of

- W+jets events,
- $\text{t}\bar{\text{t}}(\text{jet} \rightarrow \tau_{\text{h}})$ events and
- events from genuine tau or lepton to tau misidentification processes

inside the AR from simulation and τ-embedded samples. The difference between the observed data yield and the sum of the above processes is attributed to the QCD multijet process, where the difference is not allowed to become negative. Contributions from W+jets, QCD multijet and $\mathrm{t\bar{t}}$(jet $\rightarrow \tau_\mathrm{h}$) events are normalized such that they add up to one. These normalized contributions are called *fractions*[7] and are labeled as f_i in Figure 8.4. For the $\mathrm{e}\tau_\mathrm{h}$ and $\mu\tau_\mathrm{h}$ channel, fractions are shown in Figure 8.27 for the 2018 data-taking period. They are determined as a function of $m_\mathrm{T,PUPPI}$, using three different N_jets categories – $N_\mathrm{jets} = 0$, $N_\mathrm{jets} = 1$ and $N_\mathrm{jets} \geq 2$. The choice of $m_\mathrm{T,PUPPI}$ is driven by the fact that QCD multijet events dominate at low $m_\mathrm{T,PUPPI}$ and W+jets become more abundant towards higher $m_\mathrm{T,PUPPI}$ values. As for $\mathrm{t\bar{t}}$ events, one can clearly see in Figure 8.27 that they are only relevant for the category with at least two jets.

For each recorded event (ϵ) inside the AR the combined F_F

$$\begin{aligned} F_\mathrm{F} &= f_\mathrm{W+jets} \cdot F_\mathrm{F}^\mathrm{(W+jets)} + f_\mathrm{QCD} \cdot F_\mathrm{F}^\mathrm{(QCD)} + f_\mathrm{t\bar{t}} \cdot F_\mathrm{F}^\mathrm{(t\bar{t})} \\ 1 &= f_\mathrm{W+jets} + f_\mathrm{QCD} + f_\mathrm{t\bar{t}} \ , \end{aligned} \tag{8.56}$$

is determined (see also the formula shown in Figure 8.4). The combined F_F serves as extrapolation factor from AR to SR. However, it has to be taken into account that not all recorded events inside the AR are coming from the jet $\rightarrow \tau_\mathrm{h}$ process as seen in Figure 8.3. This is achieved by extrapolating the contributions from "other bkg." and τ-embedded samples from AR to SR by means of the combined F_F and subtracting them. In form of an equation, the number of predicted jet $\rightarrow \tau_\mathrm{h}$ events inside the SR is given by

$$N_\mathrm{pred}^\mathrm{(SR)} = \sum_{\epsilon \in \mathrm{AR}} F_\mathrm{F}(\epsilon) - \sum_{i \in \text{other bkg.}} F_\mathrm{F}(i) - \sum_{j \in \text{genuine } \tau} F_\mathrm{F}(j) \ , \tag{8.57}$$

where the index i runs over all events falling into the category of "other bkg." which are estimated by means of MC simulation. The index j runs over all events of the τ-embedded samples. An example of the F_F application using the above formula is presented in Figure 8.28. It shows the distributions of m_vis of the di-tau system and demonstrates the good modeling of the observed data.

For the $\tau_\mathrm{h}\tau_\mathrm{h}$ channel, the F_F application works as follows. Solely a $F_\mathrm{F}^\mathrm{(QCD)}$ is measured with the following dependencies:

$$\text{Eq. 8.54} \rightarrow F_\mathrm{F}^\mathrm{(QCD)}(p_\mathrm{T}^{(\tau_1)}, N_\mathrm{jets}, m_\mathrm{vis}, p_\mathrm{T}^{(\tau_2)}) \ , \tag{8.58}$$

because QCD multijet processes almost entirely make up the jet $\rightarrow \tau_\mathrm{h}$ contribution (see bottom row of Figure 8.2). It must be taken into account that both τ_h are potentially due to a misidentification. This impacts the definition of the AR. For the fully hadronic case, two ARs are used:

- AR_1 The leading τ_h fails the `tight` condition on the D_jet discriminant while passing the `vloose` one. The sub-leading passes the `tight` D_jet condition and is thus SR-like.
- AR_2 The roles are switched with respect to AR_1. Here the leading τ_h is SR-like and passes the `tight` condition on the D_jet discriminant. However, the sub-leading τ_h fails the `tight` D_jet condition but passes the `vloose` one.

Figure 8.29 shows the SR together with AR_1 and AR_2. Each of the τ_h candidates of the tau pair can be either real (r) or fake (f). In total, four different combinations can occur

[7] Fractions are the same objects as the probabilities used to determine closure correction inside the DRs (see for example Equation (8.13)).

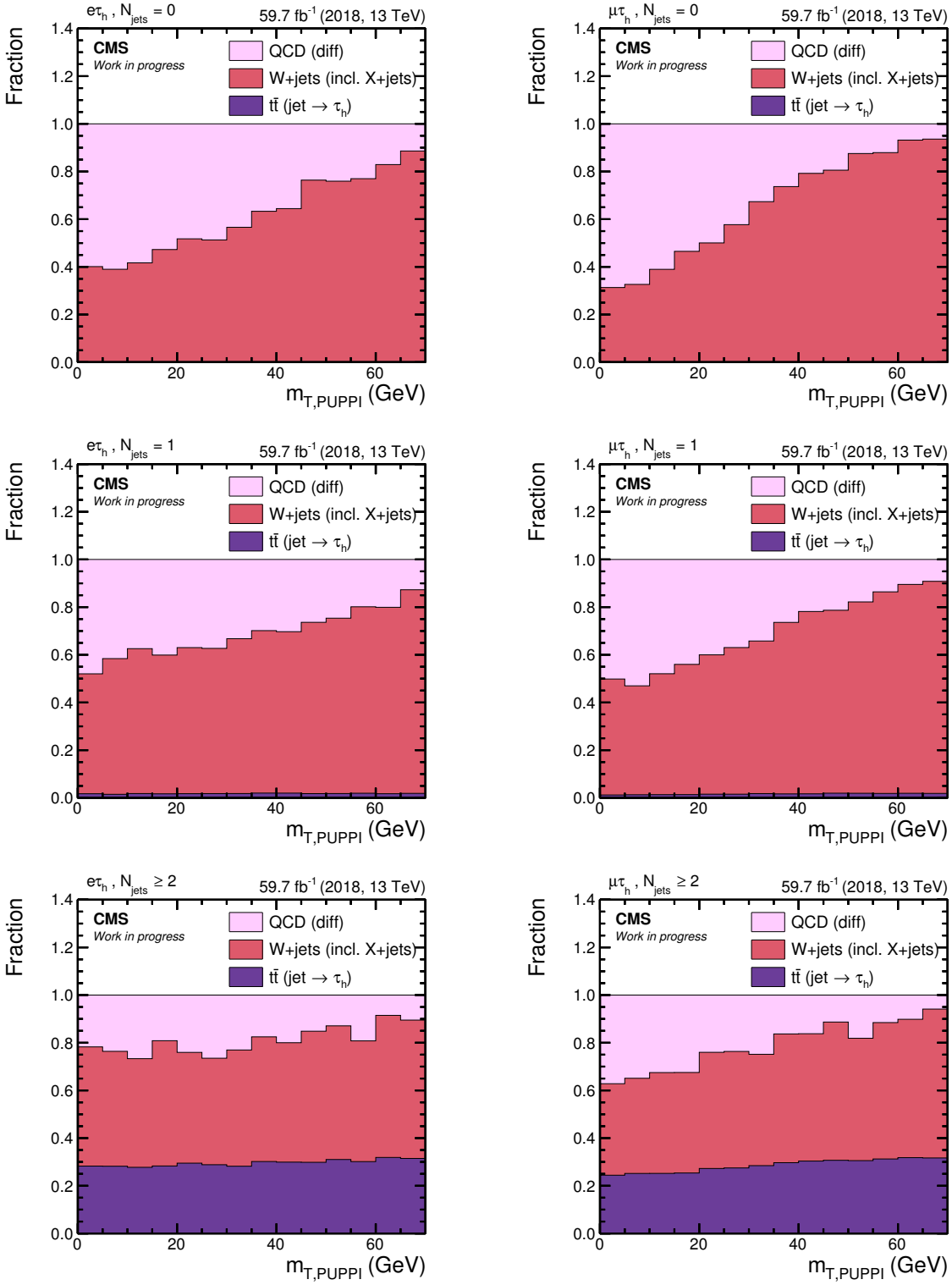


Figure 8.27: Fractions for the eτ_{h} (left) and $\mu\tau_{\mathrm{h}}$ (right) channel are shown using 2018 data, respectively. From top to bottom, three different N_{jets} categories are displayed – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} = 1$ and $N_{\mathrm{jets}} \geq 2$. The fractions are normalized such that $1 = f_{\mathrm{W+jets}} + f_{\mathrm{QCD}} + f_{\mathrm{t\bar{t}}}$ (see also Equation (8.56)). Contributions from $\mathrm{Z} \to \ell\ell(\mathrm{jet} \to \tau_{\mathrm{h}})$ and $\mathrm{VV}(\mathrm{jet} \to \tau_{\mathrm{h}})$ (see Figure 8.2) are included in the fractions of W+jets and marked as X + jets in the legend.

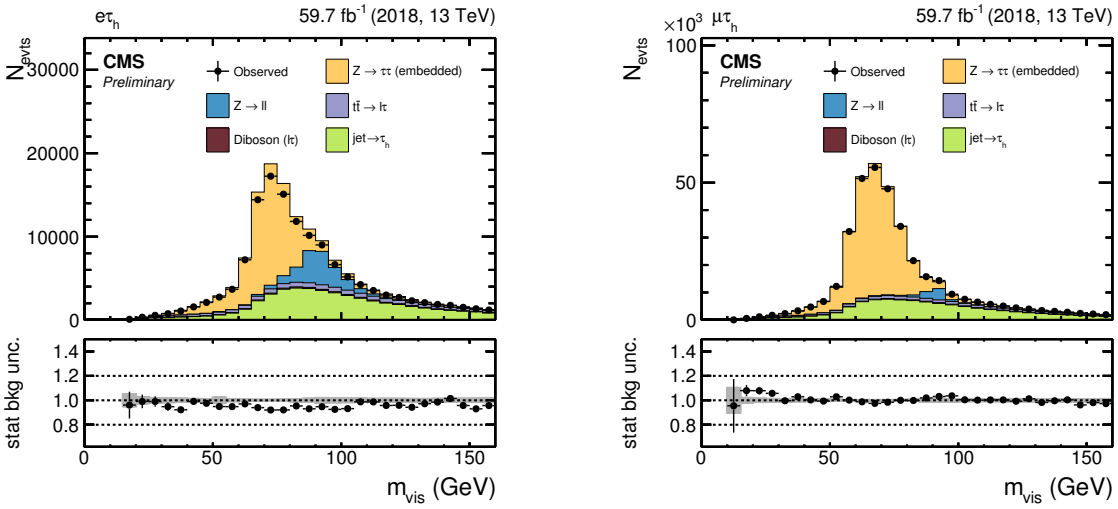


Figure 8.28: The distributions of m_{vis} are shown for the $e\tau_h$ (left) and the $\mu\tau_h$ (right) channel using 2018 data. Contributions from the jet $\rightarrow \tau_h$ process are estimated using the FF method and are obtained according to Equation (8.57). Only statistical uncertainties are shown here.

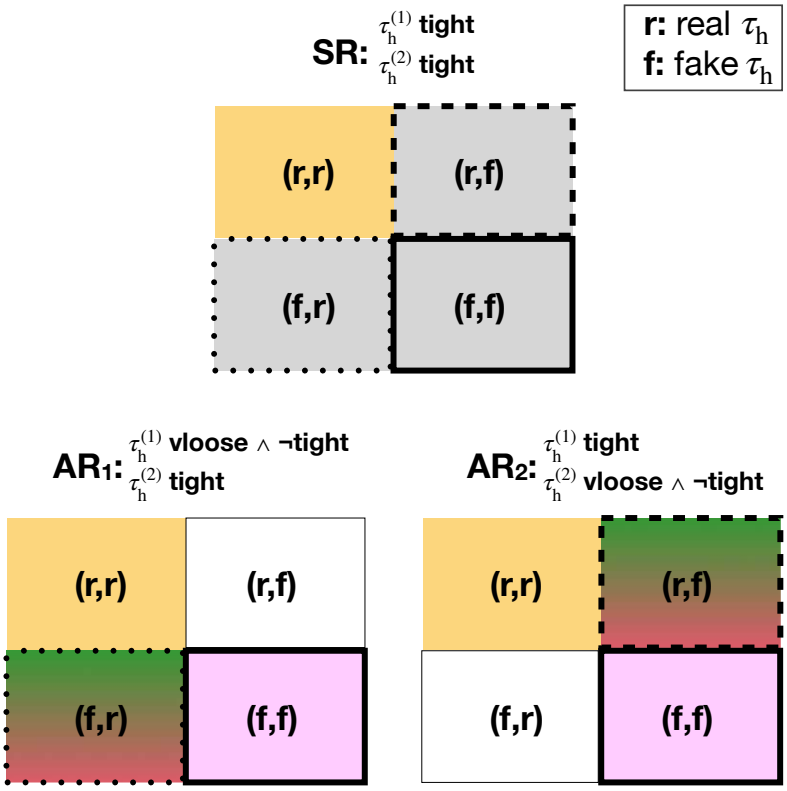


Figure 8.29: Illustration of the different combinations of jet $\rightarrow \tau_h$ in the $\tau_h\tau_h$ channel. On the top the SR is shown and at the bottom AR_1 and AR_2. For the SR the regions with at least one fake τ_h candidate are colored in dark gray. Regions of genuine tau pairs are shown in yellow. W+jets and $t\bar{t}$ processes populate dominantly the (f,r)-part of AR_1 and the (r,f)-part of AR_2. Regions with negligible contributions of jet $\rightarrow \tau_h$ are left white. QCD multijet processes – shown in pink – give rise to two fake τ_h candidates and thus populate the corresponding (f,f)-parts of both AR_1 and AR_2. As a result different combinatorial factors are applied to the different processes (see Figure 8.22).

as illustrated in Figure 8.29. For W+jets and $t\bar{t}$ processes, typically, one τ_h candidate is real while the other one being due to a misidentified jet. Almost exclusively, the real τ_h is the one passing the `tight` D_{jet} condition. Hence, extrapolating W+jets and $t\bar{t}$ processes from AR_1 to SR will populate the (f,r)-part of the SR. Similarly, extrapolating W+jets and $t\bar{t}$ processes from AR_2 to SR will populate the (r,f)-part of the SR. The F_F used for the extrapolation is $F_F^{(QCD)}$ from Equation (8.58). As indicated in Figure 8.29, the behavior for QCD multijet events is different. Tau pairs from QCD multijet production typically *both* arise from misidentified jets. This means that this process populates the (f,f)-part of AR_1 *and* AR_2. Therefore, a combinatorial factor of 0.5 is applied in the extrapolation to the SR for the QCD multijet part which takes into account that both τ_h candidates are most likely due to misidentification. In summary, this means that even though $F_F^{(QCD)}$ is applied to all processes contributing to jet $\rightarrow \tau_h$, it is weighted differently for W+jets / $t\bar{t}$ and QCD multijet events. The combinatorial weights are given by

$$\begin{aligned} \omega_{W+jets} &= 1.0 \\ \omega_{QCD} &= 0.5 \\ \omega_{t\bar{t}} &= 1.0 \ , \end{aligned} \tag{8.59}$$

and are also part of the formula quoted in Figure 8.22. Fractions ,f_i, are determined in the same way as explained above for semi-leptonic channels. Figure 8.30 shows the fractions used for $\tau_h\tau_h$ channel using data recorded in 2018. They are binned in m_{vis} and calculated for three N_{jets} categories. As can be seen from Figure 8.30, QCD multijet processes are dominant across all N_{jets} categories. The combined F_F for the $\tau_h\tau_h$ channel reads

$$\begin{aligned} F_F &= \sum_i f_i \cdot \omega_i \cdot F_F^{(i)} \\ &= f_{W+jets} \cdot F_F^{(QCD)} + f_{QCD} \cdot 0.5 \cdot F_F^{(QCD)} + f_{t\bar{t}} \cdot F_F^{(QCD)} \\ 1 &= f_{W+jets} + f_{QCD} + f_{t\bar{t}} \ . \end{aligned} \tag{8.60}$$

The F_F application works according to Equation (8.57). Figure 8.31 shows the distribution of m_{vis}. The jet $\rightarrow \tau_h$ contribution is obtained by the FF method and is shown in green. In general, a good agreement within statistical uncertainties is observed.

8.2.4 Fake Factor Uncertainty Model

The uncertainty model of the FF method takes into account the statistical uncertainties present in the derivation of the raw F_F's and its corrections. Furthermore, uncertainties on the subtracted contributions from data are taken into account as will be discussed in this section as well as uncertainties on the fractions. The FF uncertainties are an important part of the full model of uncertainties for the SM $H \rightarrow \tau\tau$ analysis that will be discussed in Section 9.2.2.

In general, uncertainties can be classified the following way. Uncertainties changing the event count with respect to the nominal value are called *normalization uncertainties*, where the nominal value is calculated according to Equation (8.57). If an uncertainty is changing the form of the nominal distribution, it is referred to as a *shape-altering uncertainty*.

Yellow uncertainty bands (see e.g. Figure 8.6) serve as a starting point to evaluate shape-altering uncertainties. The nominal value (n) is obtained retrieved from a polynomial fit or from a smoothed curve. Statistical fluctuations defined by the yellow uncertainty band at each point, x, are bound by an upper and a lower value. The interval

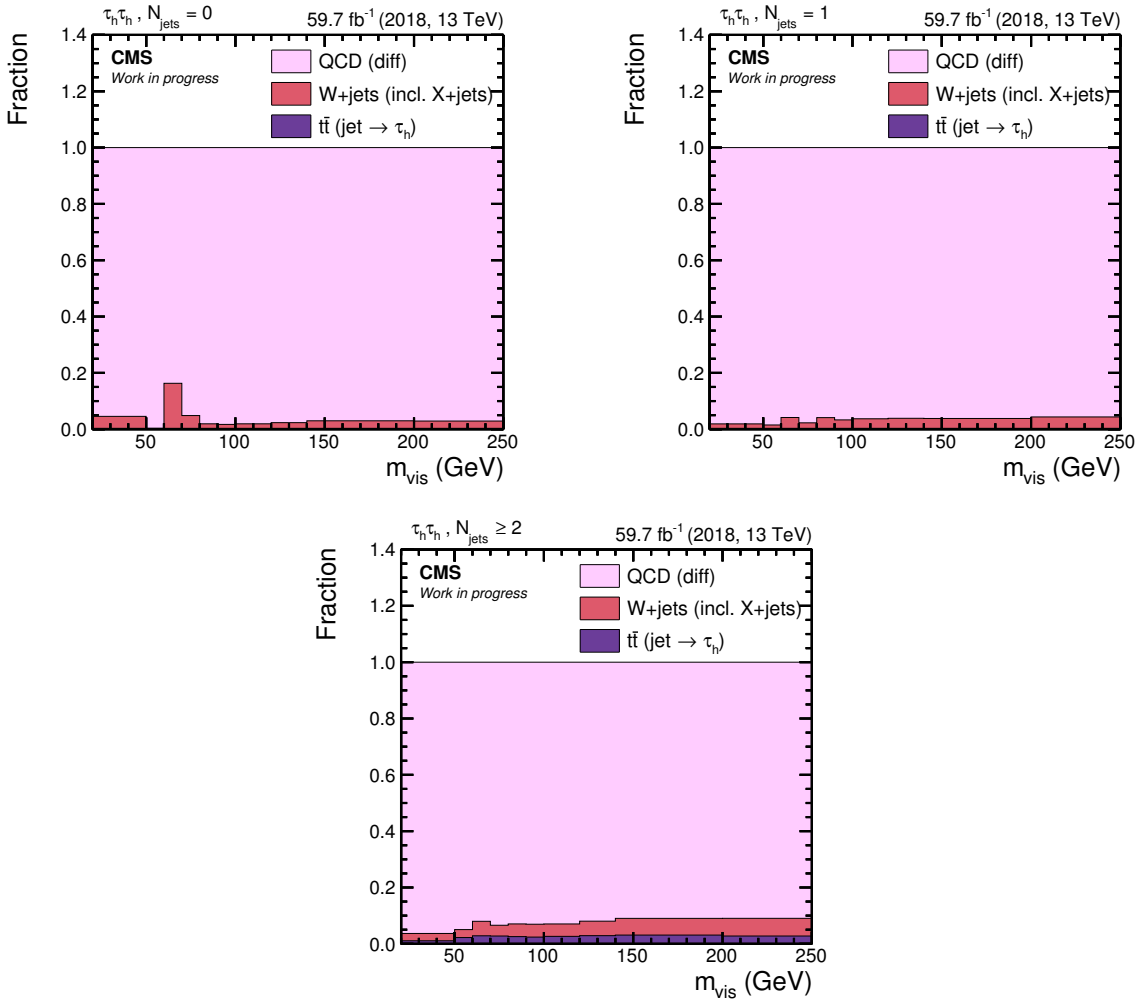


Figure 8.30: Fractions for the $\tau_\mathrm{h}\tau_\mathrm{h}$ channel are shown using 2018 data. Three different N_jets categories are displayed – $N_\mathrm{jets} = 0$, $N_\mathrm{jets} = 1$ and $N_\mathrm{jets} \geq 2$. The fractions are normalized such that $1 = f_\mathrm{W+jets} + f_\mathrm{QCD} + f_{\mathrm{t\bar{t}}}$ (see also Equation (8.60)). Contributions from $\mathrm{Z} \to \ell\ell(\mathrm{jet} \to \tau_\mathrm{h})$ and $\mathrm{VV}(\mathrm{jet} \to \tau_\mathrm{h})$ (see Figure 8.2) are included in the fractions of W+jets and marked as X + jets in the legend.

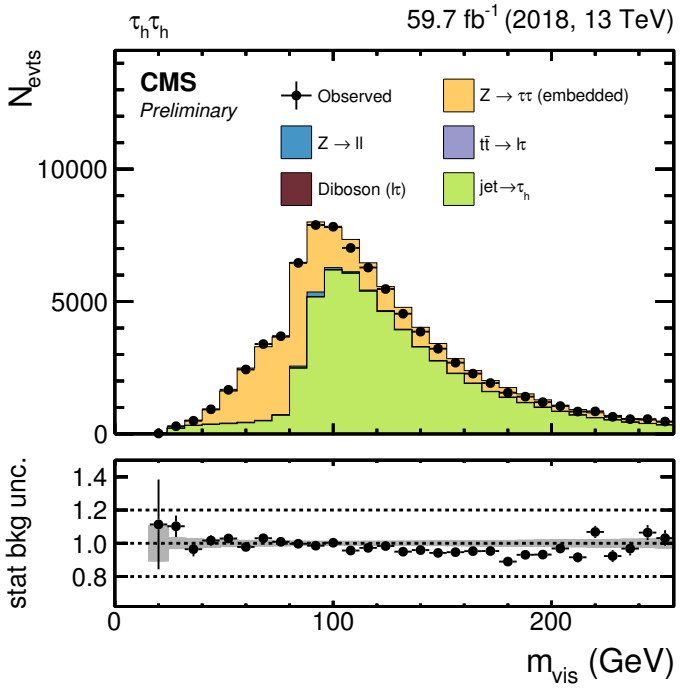


Figure 8.31: The distribution of m_{vis} is shown for the $\tau_{\text{h}}\tau_{\text{h}}$ channel using 2018 data. Contributions from the jet $\rightarrow \tau_{\text{h}}$ process are estimated using the FF method and are obtained according to Equation (8.57). Only statistical uncertainties are shown here.

between upper and lower value represents a rough approximation of a 68% confidence interval. Upper and lower boundaries are referred to as *up* and *down* variation, respectively. Hence, one can write

$$n^{+u}_{-d} \ , \tag{8.61}$$

where u and d denotes the up and down variation, respectively. Both uncertainty variations are used to define a new set of *morphed* uncertainties

$$\begin{aligned} u_{\text{morph}}(x) &= u(x)\frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} + d(x)\frac{x_{\text{max}} - x}{x_{\text{max}} - x_{\text{min}}} \\ d_{\text{morph}}(x) &= u(x)\frac{x_{\text{max}} - x}{x_{\text{max}} - x_{\text{min}}} + d(x)\frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \\ x &\in [x_{\text{min}}, x_{\text{max}}] \ . \end{aligned} \tag{8.62}$$

Equation (8.62) means that the morphed up variation, $u_{\text{morph}}(x)$, is defined by drawing a line from the lower left corner ($d(x_{\text{min}})$) of the uncertainty band towards the upper right corner ($u(x_{\text{max}})$), with the line being always inside the uncertainty band. Similarly, the morphed down variation, $d_{\text{morph}}(x)$, is defined by a line from the upper left to the lower right corner of the uncertainty band. The morphed up and down variations are shown in red and blue in Figure 8.32, respectively.

In case of semi-leptonic final states, eleven independent shape variations arise from the raw F_{F} measurement, one for each jet $\rightarrow \tau_{\text{h}}$ source (W+jets, $\text{t}\bar{\text{t}}$ and QCD multijet) and each $N_{\text{jets}} \times \Delta R^{(\ell,\tau_{\text{h}})}$ category. For each closure correction additional six variations are added. All these shape variations turn out to be small since they have a purely statistical nature and are typically derived in well-populated regions. This is reviewed again in Section 9.2.2, where the FF uncertainty model is discussed in the context of the SM $\text{H} \rightarrow \tau\tau$ analysis. Large closure or bias corrections in terms of their nominal value indicate a problem in the validity of the FF method itself. Therefore, six variations are introduced for each correction by defining a down variation, where the correction is not

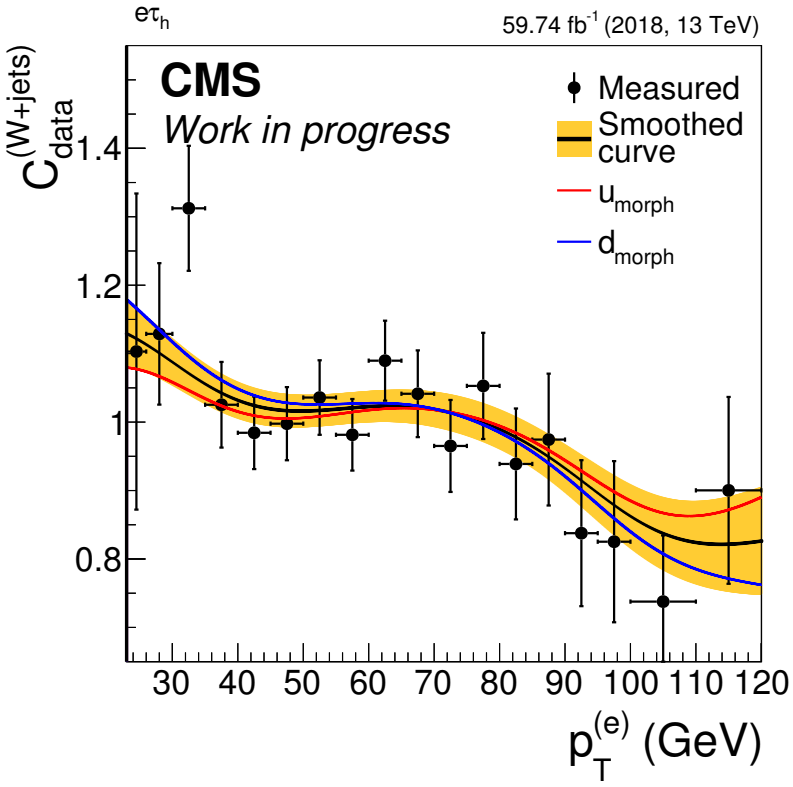


Figure 8.32: Shown is the closure correction for $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ in the eτ_{h} channel from Figure 8.10. The shape-altering up and down variations are shown in red and blue, respectively. They always lie within the yellow uncertainty band and are calculated according to Equation (8.62).

applied and an up variation, where the correction is applied twice. These variations enable to weaken or strengthen the applied correction during the signal inference step and are typically the dominant source of uncertainties, reaching values up to 10%.

Uncertainties related to the subtraction of MC simulation and τ-embedded samples in case of the data-driven $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ and $F_{\mathrm{F}}^{(\mathrm{QCD})}$ are also taken into account. These uncertainties are derived by propagating the shifts resulting from cross-section and acceptance uncertainties within each DR to the final F_{F} value and using the difference to the nominal F_{F} values as uncertainty. Similarly, the W+jets fraction inside the AR is varied according to its statistical as well as cross section and acceptance uncertainties, while keeping the sum of fraction equal to one. The total variation is in the order of $\pm$7% and is propagated to the final F_{F} values. The difference to the nominal F_{F} value is taken as uncertainty. In the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel, $F_{\mathrm{F}}^{(\mathrm{QCD})}$ is applied also to W+jets and $\mathrm{t\bar{t}}$ and two extra uncertainties of 30% are added and weighted with the fraction of W+jets and $\mathrm{t\bar{t}}$, respectively.

8.3 Estimating Background Contributions to Soft Muons and Electrons

The FF scheme employed in this section follows closely the one shown Figure 8.1b. The goal is on one hand to estimate the contributions from quark- or gluon-initiated jets that are misidentified as electrons or muons, i.e. fake leptons. On the other hand, the same measured F_{F} is applied to estimate contributions from processes with non-prompt leptons. In a first step, the DR needs to be defined which is enriched in QCD multijet events serving as a proxy for non-prompt leptons.

$p_{\mathrm{T}}^{(\ell)}$ range	electron	muon
$< 12\,\mathrm{GeV}$	PFJET(40)	μ(3)PFJET(40)
$\geq 12\,\mathrm{GeV}$	e(8)PFJET(30)	

Table 8.2: This table summarizes the triggers used to select events for the DR. Values in parentheses represent the p_{T} threshold in GeV on the respective physics object. The lowest threshold on electron transverse momenta is implemented in e(8)PFJET(30), requiring a jet with $p_{\mathrm{T}} > 30\,\mathrm{GeV}$. Since the electrons selected for analysis go down to 5 GeV in transverse momentum (see Figure 7.1), the PFJET(40) trigger is used for $p_{\mathrm{T}}^{(\mathrm{e})} < 12\,\mathrm{GeV}$. The reason for covering also the "L" region ($5\,\mathrm{GeV} \leq p_{\mathrm{T}}^{(\mathrm{e})} < 12\,\mathrm{GeV}$) with the PFJET(40) trigger is to avoid the turn on region around 8 GeV of e(8)PFJET(30). For muons, the measurement of F_{F} uses events selected by μ(3)PFJET(40) over the whole p_{T} range.

Events in the DR are selected by different triggers, dependent on the channel as summarized in Table 8.2. Triggers with low thresholds in lepton transverse momenta are chosen to select events with soft leptons, i.e. leptons with low transverse momenta like the ones present in the SR of the stop search. Events passing the respective trigger are required to have at least one jet according to the p_{T} and η requirements of the analysis, i.e. $p_{\mathrm{T}} > 30\,\mathrm{GeV}$ and $|\eta| < 2.4$ (see Table 7.1). Furthermore, only events with exactly one soft lepton are kept. Soft leptons are defined as given in Table 7.1 with the following differences

$$\begin{aligned} d_{\mathrm{xy}} &< 0.1\,\mathrm{cm} \\ d_{\mathrm{z}} &< 0.5\,\mathrm{cm} \\ \mathrm{HI}^{(\ell)} &< 20\,\mathrm{GeV}\ . \end{aligned} \tag{8.63}$$

For electrons, selection thresholds imposed on $I_{\mathrm{rel}}^{(\mathrm{e})}$, which are part of the definition of the `veto` ID, are dropped.

All triggers in Table 8.2 are *pre-scaled*. This means that they are selecting only a fraction of all events passing the trigger selection because otherwise the trigger rate would exceed its maximally allowed value. Since the simulation of events does not take trigger pre-scales into account, MC samples need to be scaled accordingly. Trigger pre-scales depend on the instantaneous luminosity of the beam, which varies over the period of data taking. In this thesis, however, a single scale factor is extracted by averaging over the whole dataset. This scale factor thus represents an effective (average) value of the trigger pre-scale. In the measurement of the effective pre-scale, yields in data and simulation are compared in a region, where electroweak processes dominate that are simulated with good precision. This is achieved by selecting events with $p_{\mathrm{T}}^{(\mathrm{miss})} > 100\,\mathrm{GeV}$, which removes a lot of QCD multijet events. The remaining events are then used to match the $m_{\mathrm{T}}^{(\ell)}$ distribution between data and simulation. Examples of $m_{\mathrm{T}}^{(\ell)}$ distributions after the fit are shown in Figure 8.33.

For the F_{F} measurement, the selection

$$\begin{aligned} p_{\mathrm{T}}^{(\mathrm{miss})} &< 50\,\mathrm{GeV} \\ m_{\mathrm{T}}^{(\ell)} &< 40\,\mathrm{GeV}\ , \end{aligned} \tag{8.64}$$

is applied, resulting in a QCD dominated region as can be seen from Figure 8.34.

The F_{F} is measured using the formula

$$F_{\mathrm{F}} = \frac{N^{(\texttt{tight})} - \sum_{\mathrm{prompt}} N_{\mathrm{MC}}^{(\texttt{tight})}}{N^{(\texttt{loose} \wedge \neg\texttt{tight})} - \sum_{\mathrm{prompt}} N_{\mathrm{MC}}^{(\texttt{loose} \wedge \neg\texttt{tight})}}\ . \tag{8.65}$$

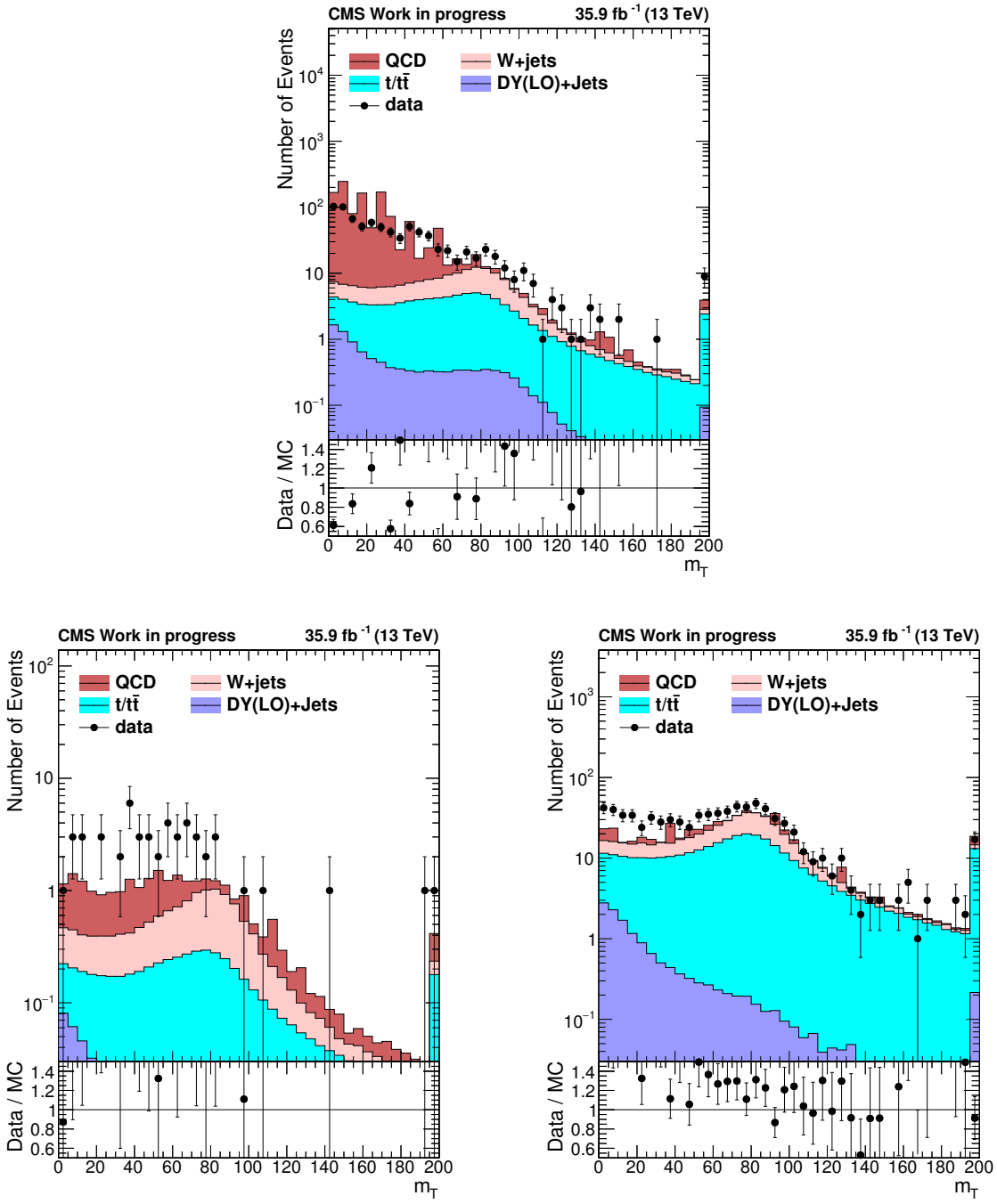


Figure 8.33: These plots show comparisons of the $m_{\mathrm{T}}^{(\ell)}$ distribution between data and simulation used to extract effective trigger pre-scales. The muon channel is depicted in the top row, selecting events with the $\mu(3)\mathrm{PFJET}(40)$ trigger and $p_{\mathrm{T}}^{(\mathrm{miss})} > 100\,\mathrm{GeV}$. The electron channel is shown in the bottom row, where on the left, the $\mathrm{PFJET}(40)$ trigger is used with $p_{\mathrm{T}}^{(\mathrm{miss})} > 90\,\mathrm{GeV}$ and on the right events pass the $\mathrm{e}(8)\mathrm{PFJET}(30)$ trigger with $p_{\mathrm{T}}^{(\mathrm{miss})} > 100\,\mathrm{GeV}$. The normalization of the $m_{\mathrm{T}}^{(\ell)}$ distributions is fit in the region $70\,\mathrm{GeV} < m_{\mathrm{T}}^{(\ell)} < 100\,\mathrm{GeV}$. In case of the bottom left plot ($\mathrm{PFJET}(40)$), the $p_{\mathrm{T}}^{(\mathrm{miss})}$ cut is relaxed to have a larger sample size and the fit range is extended to $50\,\mathrm{GeV} < m_{\mathrm{T}}^{(\ell)} < 100\,\mathrm{GeV}$.

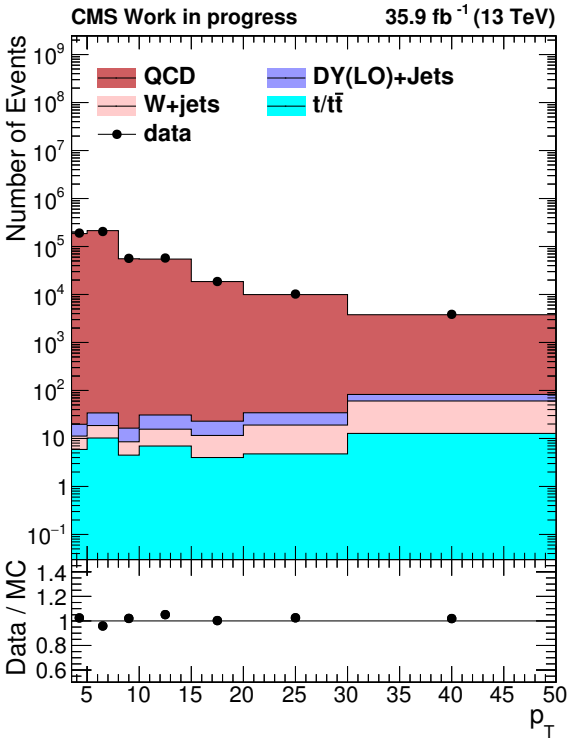


Figure 8.34: The p_{T} distribution is shown for the muon channel inside the DR. Simulated events are weighted by the derived effective trigger pre-scale for $\mu(3)\mathrm{PF}_{\mathrm{JET}}(40)$. Events satisfy the selection of Equation (8.64). This plot illustrates that each p_{T} bin is enriched in QCD multijet events.

The definition of the SR-like (`tight`) region is given by the analysis requirements (see also Table 7.1)

$$\begin{aligned} d_{\mathrm{xy}} &< 0.02\,\mathrm{cm} \\ d_{\mathrm{z}} &< 0.1\,\mathrm{cm} \\ \mathrm{HI}^{(\ell)} &< 5\,\mathrm{GeV}\ . \end{aligned} \tag{8.66}$$

The AR-like region is filled with events passing the `loose` selection criteria stated in Equation (8.63) but failing the `tight` criteria from Equation (8.66). The different regions are visualized in Figure 8.35. Contributions from prompt leptons are subtracted as indicated in Equation (8.65), leaving both non-prompt and falsely identified leptons for the measurement of F_{F} from data. Finally, F_{F} is measured separately for the electron and muon channel, as well as in different p_{T} and η bins. Two bins in η are used, separating the barrel and endcap region of the CMS detector. The binning in p_{T} is aligned with the region definition of the analysis (see Figure 7.1). Figure 8.36 shows the F_{F} maps for both, electron and muon channel. For the electron channel, the first p_{T} bin is empty since only electrons with $p_{\mathrm{T}} > 5\,\mathrm{GeV}$ are selected in the analysis. For the second and third p_{T} bin, F_{F} values are obtained from events selected by the $\mathrm{PF}_{\mathrm{JET}}(40)$ trigger, whereas for all bins with larger p_{T}, the $\mathrm{e}(8)\mathrm{PF}_{\mathrm{JET}}(30)$ trigger is used (see also Table 8.2). The quoted uncertainties are purely statistical in nature and reflect the large sample sizes available for the F_{F} measurement. Furthermore, the contamination from processes with prompt leptons is small. Varying the contribution of prompt leptons by 50% results in changes of F_{F} values in the same order as the quoted statistical uncertainties.

The presented FF method has the following benefits over the approach followed in [15, 206]:

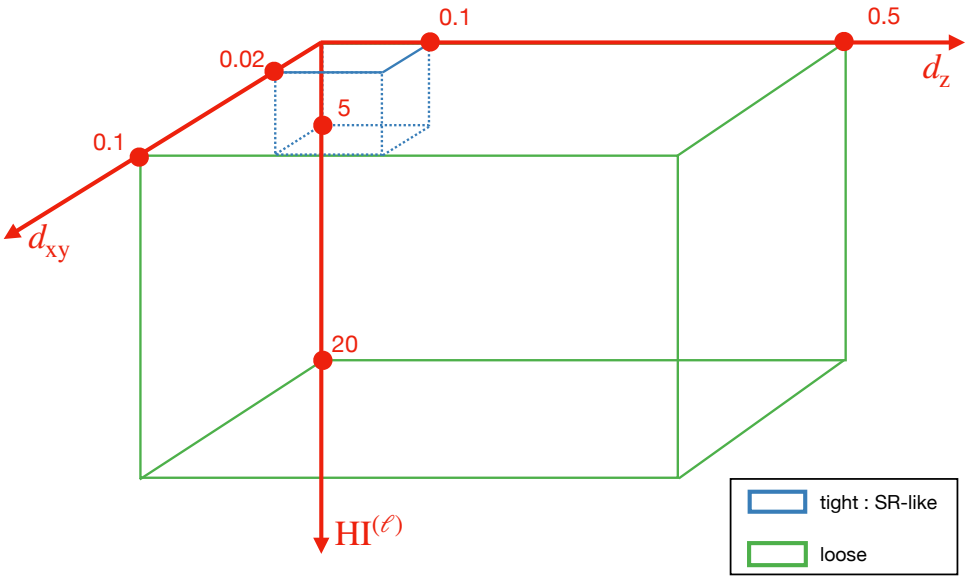


Figure 8.35: A graphical representation of the selections given in Equations (8.63) and (8.66), defining respectively the `loose` region (green) and the SR-like (blue) DR, is shown. The AR-like DR is given by the requirement `loose` $\wedge$ $\neg$`tight`. SR-like and AR-like DR are non-overlapping.

- The DR proposed here results in a more robust measurement of F_{F}. For instance, the F_{F} measurement is not dependent on the ISR reweighting (see Section 7.4) of contributions from prompt leptons.
- The purity in fake leptons is much higher in the proposed method than in its previous version.

However, the measured F_{F}'s need to be further tested for potential biases when applying them to the search for light top squarks. One such test can be in terms of a closure correction in a validation region which is similar to the SR of the stop analysis. This topic will be picked up again in Section 9.3.3.

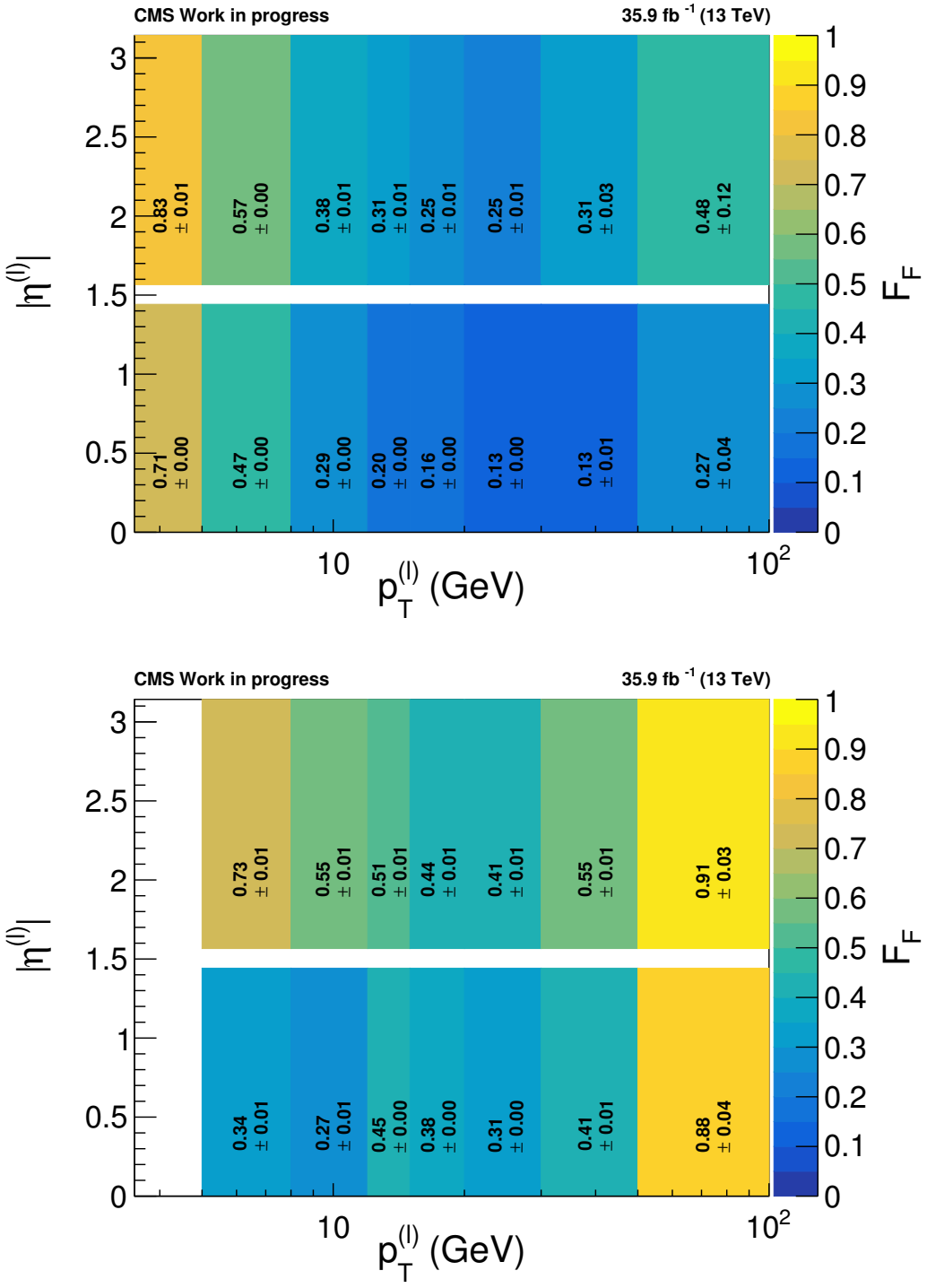


Figure 8.36: Shown are the F_{F}'s measured according to Equation (8.65) for the muon (top) and electron (bottom) channel. Different triggers are used to select the data and perform the F_{F} measurement (see Table 8.2). Quoted uncertainties are purely statistical in nature.

Chapter 9

Applications and Results

> If it [a (new) theory] disagrees with experiment, it's wrong. In that simple statement is the key to science. It doesn't make any difference how beautiful your guess is, it doesn't make even if how smart you are who made the guess, or what his[/her] name is, if it disagrees with experiment, it's wrong. That's all there is to it.
>
> Richard Feynman, 1918 to 1988

This chapter presents the results of several analyses, starting with a general introduction to statistical inference in Section 9.2. From the way the targeted signal process looks like, dedicated kinematic selections are applied, as discussed in previous chapters, to define a search or measurement region[1]. Search or measurement regions should contain a minimum amount of misidentified objects as well as signal events with the least amount of background contamination possible. Search and measurement regions are prepared for statistical inference by applying a further partitioning of events. In some of these partitions, the signal should be enriched over the background and it is possible to infer the quantities of interest about the signal from these regions. However, it is also desirable to have regions where the background is dominant. In those background dominated regions, unknowns like the normalizations of certain background processes can be inferred from data. There are two ways how the partitioning is achieved,

- by using multivariate techniques or
- in a cut-and-count based manner.

Both techniques are used in the analyses presented.

Throughout Section 9.2, the SM $H \to \tau\tau$ analysis is presented in detail. It starts by explaining the event classification, followed by a discussion of the systematic uncertainties and the presentation of the main results. Section 9.3 shows the results from searches of new particles. In all cases, the background contributions arising from jet $\to \tau_h$ are estimated with the FF method presented in Section 8.2.

9.1 Statistical Inference

All analyses presented in this thesis are based on counting experiments from a statistical point of view. Histograms provide a useful summary statistics and enter the statistical

[1] In search regions, one is looking for a new physics signal, while in measurement regions parameters of a known signal are inferred.

inference. In order to extract/measure the parameters of interest (POIs) from the selected data, the method of *binned maximum likelihood* is used as explained in the following. At the core of the binned maximum likelihood approach lies the assumption that the content of each histogram bin (i) can be modeled by a Poisson distribution if the entries (events) are statistically independent [217]

$$\mathcal{P}\left[d_i, \lambda_i\right] = \frac{\exp(-\lambda_i) \cdot \lambda_i^{d_i}}{(d_i)!} \ , \tag{9.1}$$

where d_i is the *observed* number of events in bin i, and λ_i represents the (unknown) *true* yield in this bin. The full likelihood, $\mathcal{L}$, is given by the product of the Poisson distributions over all histogram bins

$$\mathcal{L} = \prod_i \mathcal{P}\left[d_i, \lambda_i\right] \ . \tag{9.2}$$

The expected number of events can be divided into contributions from signal processes (S_i) and background processes (B_i)

$$\lambda_i = \mu \cdot S_i + B_i \ , \tag{9.3}$$

where the signal strength modifier (μ) is the POI. Hence, by varying μ during the maximum likelihood fit, the most compatible value with the data is retrieved. Clearly, a value of μ close to zero indicates an absence of the signal process and upper limits on the production cross section of the tested signal processes can be calculated. A value of μ significantly higher than zero is indicative for the presence of a signal. In case a precise signal yield prediction exists, a value of μ compatible with one is indicative of a signal matching this particular prediction. Typically, there are several background processes present within a single histogram bin. Thus, the background expectation is written as

$$B_i = \sum_b B_i^{(b)} \ , \tag{9.4}$$

where $B_i^{(b)}$ denotes the expected background contribution of process b in bin i. Similarly, the signal can originate from several sub-processes, e.g. in the context of the STXS stage-1.2 scheme (see Figures 5.3 and 5.4). In that case, the signal expectation is written as

$$S_i = \sum_s S_i^{(s)} \ , \tag{9.5}$$

where the different signal sub-processes are indexed by s. In case of several signal sub-processes, there are also several POIs introduced

$$\mu \to \mu_s \ . \tag{9.6}$$

Equation (9.3) thus reads in the most general case as

$$\lambda_i = \sum_s \mu_s \cdot S_i^{(s)} + \sum_b B_i^{(b)} \ . \tag{9.7}$$

As discussed so far, the statistical uncertainty of the observed data (d_i) is taken into account by using the Poisson model shown in Equation (9.2). However, there are also *statistical and systematic* uncertainties that affect both signal and background contributions. Therefore, both signal and background expectation are modeled as a function of those parameters ($\boldsymbol{\theta}$) which are called *nuisance* parameters

$$\begin{aligned} B_i &\to B_i(\boldsymbol{\theta}) \\ S_i &\to S_i(\boldsymbol{\theta}) \ . \end{aligned} \tag{9.8}$$

channel	classes				
$e\tau_h$ / $\mu\tau_h$	genuine τ	jet $\to \tau_h$	misc	$Z \to \ell\ell$	$t\bar{t}$
$\tau_h\tau_h$	genuine τ	jet $\to \tau_h$	misc	–	–

Table 9.1: This table summarizes the different background classes defined for NN training. In the semi-leptonic channels, five background classes are differentiated. The fully hadronic channel only discriminates among three classes.

Nuisance parameters follow a certain probability distribution, $\mathcal{C}$, and thus the full likelihood from Equation (9.2) is extended and reads

$$\mathcal{L} = \prod_j \mathcal{C}(\theta_j) \prod_i \mathcal{P}\left[d_i, \sum_s \mu_s \cdot S_i^{(s)}(\theta_j) + \sum_b B_i^{(b)}(\theta_j)\right] \equiv \mathcal{L}(\boldsymbol{\mu}; \boldsymbol{\theta}) \ , \tag{9.9}$$

where in the last step its dependence on the POIs ($\boldsymbol{\mu}$) and on the nuisance parameters ($\boldsymbol{\theta}$) is highlighted.

Nuisance parameters can purely change the yield of nominal distributions. Examples are cross section and luminosity uncertainties as well as all uncertainties associated with multiplicative corrections like efficiency corrections. All such nuisance parameters are modeled with a log-normal (lnN) distribution. In case the systematic uncertainty also changes the *shape* of the nominal distribution, the up and down variations have to be calculated in terms of their influence on the final discriminant used for signal inference. During signal inference, a polynomial function is used to interpolate between the up and down variations, while using linear extensions beyond the given up/down variations [218, 219].

9.2 SM $H \to \tau\tau$ Analysis

In Section 9.2.1, the event classification based on neural networks is explained. Systematic uncertainties, with particular emphasis on those related to the FF method, are discussed in Section 9.2.2. Results in terms of the target STXS stage-0 and stage-1.2 scheme are presented in Sections 9.2.3 and 9.2.4, respectively. A manuscript describing the results of this analysis is under journal review and available on the preprint server [7]. More details of the analysis can also be found in [12].

9.2.1 Neural-Network-Based Event Classification

The goal is to separate signal from background events inside the signal region which is defined in Section 6.1. Furthermore, going beyond a simple binary decision between signal and background, several signal and background classes should be discriminated. Each dominant background defines a background class, whereby minor backgrounds are collected in a so-called *miscellaneous* (misc) class. The different background classes are summarized in Table 9.1. For the signal, two signal classes are defined for the STXS stage-0 measurement, one aggregating ggH events and one collecting qqH events. The STXS stage-1.2 measurement comprises 15 signal classes, eleven for ggH (see Figure 5.3) and four for qqH (see Figure 5.4). The formulated multi-classification task is tackled using a deep neural network (NN) [220], schematically shown in Figure 9.1. For each decay channel a dedicated NN is trained.

Each NN is designed as a fully-connected feed-forward NN, meaning that all nodes are connected with each other in a non-circular way such that information can only flow in one

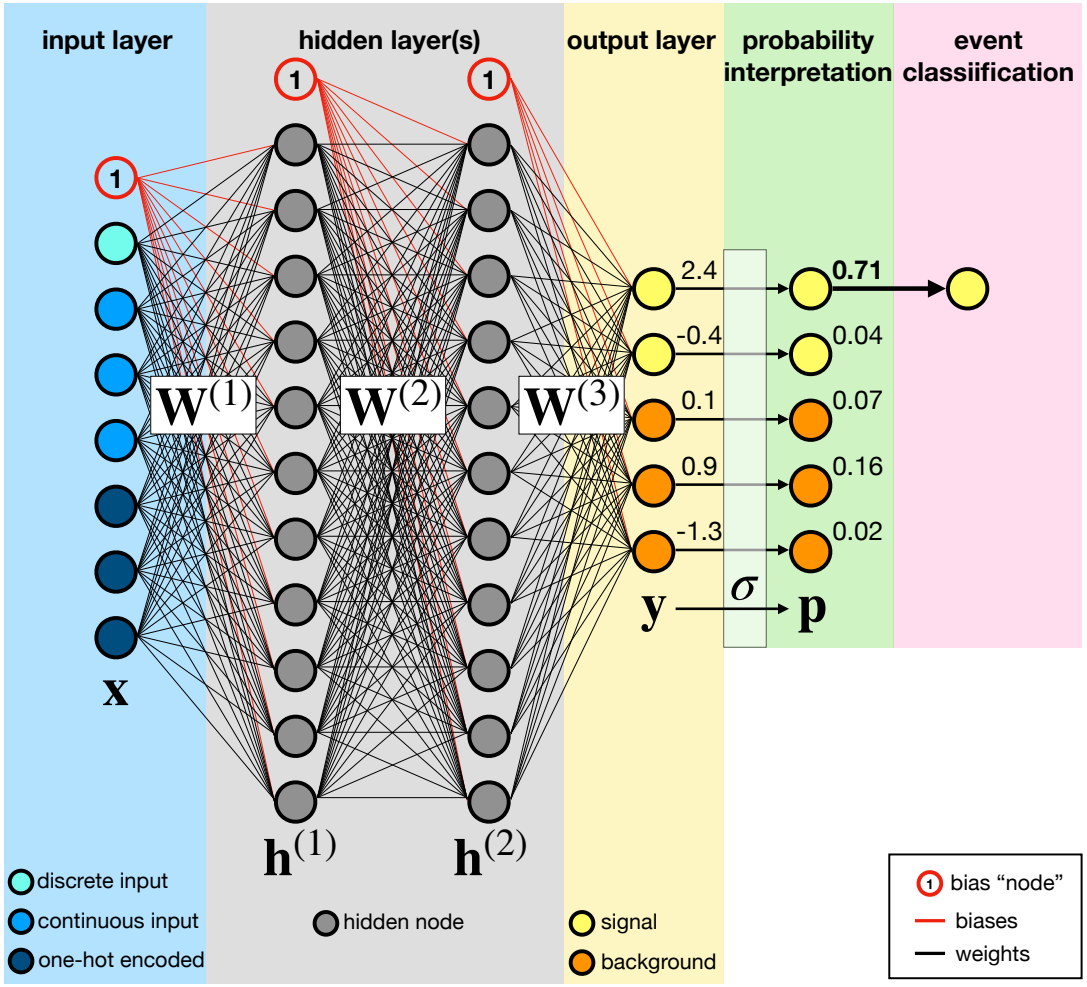


Figure 9.1: This figure summarizes the important elements of the neural network (NN) used to differentiate between signal and background classes in the context of the SM $H \to \tau\tau$ analysis. Each NN has an input layer (blue) followed by hidden layer(s) (gray) leading to the output layer (yellow). The inputs ($\mathbf{x}$) can have several different types (discrete, continuous, one-hot encoded) as indicated in the figure. All nodes from one layer are connected with the next layer, where one distinguishes between weights (black lines) and biases (red lines). The connections are represented by weight matrices ($\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{W}^{(3)}$) comprising all the parameters which are fitted during the training phase of the NN resulting in the final model used for classification. Each signal (yellow) and background (orange) class is represented by a separate node in the output layer. The value of the output is transformed using the softmax function (σ) to get values in the range $[0, 1]$. A numerical example of different output values ($\mathbf{y}$) and their transformed values ($\mathbf{p}$) is given in the figure. The event classification (red) then assigns the event to the class corresponding to the node with the highest value of p_i. More details are discussed in the main text or can be found in Table 9.2.

direction. The information flows from the input nodes, via nodes in the so-called *hidden layers* to the output nodes. In the following, the exact processing of the information by the NN is described. The output of the top node in the first hidden layer ($h_1^{(1)}$) is computed using all input nodes (x_i with $i = 1, \ldots, n_x$) as

$$h_1^{(1)} = a\left(\sum_{i=1}^{n_x} \omega_{1i} \cdot x_i + b_1^{(1)}\right) \ . \tag{9.10}$$

The *weights* (ω_{1i}) and the *bias* ($b_1^{(1)}$) in the above equation belong to the set of parameters of the NN and are represented by lines connecting pairs of nodes in Figure 9.1. A key ingredient of every NN is the usage of a non-linear *activation function*, $a(x)$. In this analysis, the hyperbolic tangent is used as activation function, i.e.

$$a(x) = \tanh(x) \ . \tag{9.11}$$

From a theoretical point of view, every NN with a multi-layer feed-forward architecture using a non-linear activation function can be used to approximate any given function [221]. The main challenge lies in monitoring the NN during the fitting of its parameters, called *training* phase, to ensure the NN is approximating the desired function, i.e. improving in classifying the inputs correctly. Instead of using the notation in Equation (9.10), it is more convenient to use a matrix notation and fill the output values of each node in vectors. Hence, the output of the first hidden layer with $n_{(1)}$ nodes can be written as

$$\begin{aligned} \begin{pmatrix} h_1^{(1)} \\ \vdots \\ h_{n_{(1)}}^{(1)} \end{pmatrix} &= a\left(\begin{pmatrix} b_1^{(1)} & \omega_{11} & \cdots & \omega_{1n_x} \\ b_2^{(1)} & \omega_{21} & \cdots & \omega_{2n_x} \\ \vdots & \vdots & \cdots & \vdots \\ b_{n_{(1)}}^{(1)} & \omega_{n_{(1)}1} & \cdots & \omega_{n_{(1)}n_x} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_{n_x} \end{pmatrix}\right) \\ \mathbf{h}^{(1)} &= a\left(\mathbf{W}^{(1)} \cdot \mathbf{x}\right) \ , \end{aligned} \tag{9.12}$$

where $\mathbf{W}^{(1)}$ is a $n_{(1)} \times (n_x + 1)$ matrix containing all weights and biases connecting the input layer with the first hidden layer. Using this matrix notation, the output vector of the second hidden layer ($\mathbf{h}^{(2)}$), having $n_{(2)}$ nodes, can be written as

$$\mathbf{h}^{(2)} = a\left(\mathbf{W}^{(2)} \cdot \mathbf{h}^{(1)}\right) \overset{\text{Eq. 9.12}}{=} a\left(\mathbf{W}^{(2)} \cdot a\left(\mathbf{W}^{(1)} \cdot \mathbf{x}\right)\right) \ , \tag{9.13}$$

where $\mathbf{W}^{(2)}$ is the $n_{(2)} \times (n_{(1)} + 1)$ matrix holding all weights and biases connecting the first hidden layer with the second hidden layer. The NN used in this analysis (see also Figure 9.1) has two hidden layers[2] which are connected to the output layer. The output vector ($\mathbf{y}$) can be written as

$$\mathbf{y} = \mathbf{W}^{(3)} \cdot \mathbf{h}^{(2)} \overset{\text{Eq. 9.13}}{=} \mathbf{W}^{(3)} \cdot a\left(\mathbf{W}^{(2)} \cdot a\left(\mathbf{W}^{(1)} \cdot \mathbf{x}\right)\right) \ , \tag{9.14}$$

where $\mathbf{W}^{(3)}$ is a $(n_y \times (n_{(2)} + 1))$-dimensional matrix, with n_y being the number of output nodes. The number of output nodes corresponds to the number of classes in the classification task. The output nodes are transformed with a different activation function, which is given in this analysis by the *softmax* function

$$\sigma_i(\mathbf{y}) = \frac{\exp(y_i)}{\sum_{j=1}^{n_y} \exp(y_j)} \ , \tag{9.15}$$

[2] In principle there can be more than two hidden layers.

	$e\tau_h$ / $\mu\tau_h$		$\tau_h\tau_h$	
	stage-0	stage-1.2	stage-0	stage-1.2
input nodes (n_x)	14+3			
hidden layers	2			
nodes per hidden layer	$n_{(1)} = 200 = n_{(2)}$			
output nodes (n_y)	7	20	5	18
trainable parameters	45407	48020	45005	47618
pre-processing	mean 0 and standard deviation 1			
weight initialization	uniform GLOROT			
bias initialization	zero			
batch size	30 events per class and year			
	630	1800	450	1620

Table 9.2: This table summarizes the NN architecture as well as information about the training setup.

where the index i labels the output nodes. Thus, the final output vector ($\mathbf{p}$) has components with values in the interval $[0, 1]$ and the sum of all components add up to one. As explained for example in [222], the softmax transformation of the output nodes makes it possible to interpret the numerical value of the node as a Bayesian probability for given input ($\mathbf{x}$) to belong to the class associated to the output node y_i

$$\mathbf{p} = \sigma(\mathbf{y}) \ . \tag{9.16}$$

The specific values used in this analysis for the number of nodes per layer are summarized in Table 9.2. Furthermore, the number of parameters of the NN are also given in Table 9.2.

From Equations (9.14) and (9.16), it follows that an NN ($\Xi^{(\lambda)}$) can be viewed as a map

$$\begin{aligned} \Xi^{(\lambda)} : \mathbb{R}^{n_x} &\to \mathbb{R}^{n_y} \\ \mathbf{x} &\mapsto \mathbf{p} \ , \end{aligned} \tag{9.17}$$

where $\Xi^{(\lambda)}$ is defined by its parameters (weights and biases)

$$\lambda = \{\mathbf{W}^{(3)}, \mathbf{W}^{(2)}, \mathbf{W}^{(1)}\} \equiv \{\lambda_p\} \ . \tag{9.18}$$

The index p in the above equation labels all parameters of the NN.

The initialization of the weights of the NN uses the GLOROT scheme [223] which promises a faster convergence to optimal NN parameters during the training phase. This is achieved by avoiding the problem of exploding gradients, i.e. avoiding large changes of the weights during the training phase (see also Equation (9.22)). Biases, however, are initialized to be zero. Furthermore, all input variables are pre-processed, except those being one-hot encoded, such that the distribution of each variable has zero mean and a standard deviation of one. This pre-processing is important because the input variables have naturally different ranges. Transverse momenta can reach a value of 100 GeV while for instance the number of jets in the event is typically a small number below ten. Without pre-processing, input features with high values would lead to a saturation of the activation function at the beginning of the NN-training. This would artificially increase the importance of such input features and thus prevent the convergence to optimal parameters of the NN.

As mentioned earlier, for each decay channel and for each STXS scheme a separate NN is trained because the number of signal and background classes vary. A *two-fold* training

is performed to exploit the whole available data set for training and application of the NN. This means, that the full set of events is split in two equal halves. Each event has an identification number which is used to easily split event samples in halves by grouping the events according to whether the event number is **e**ven or **o**dd. Let us call the two data sets D_E and D_O, respectively. An NN ($\Xi_\mathrm{E}^{(\lambda)}$) is trained using 75% of D_E. During the training, the parameters of $\Xi_\mathrm{E}^{(\lambda)}$ get updated such that more and more events in the training set get classified correctly. The remaining 25% of D_E, referred to as *validation data set*, is not used for training but to monitor if $\Xi_\mathrm{E}^{(\lambda)}$ is also performing well in classifying events from an independent data set. After the training of $\Xi_\mathrm{E}^{(\lambda)}$ has finished, it is used to classify events from D_O, the result of which is then used for statistical inference as discussed in Sections 9.2.3 and 9.2.4. The same procedure is done with a second NN ($\Xi_\mathrm{O}^{(\lambda)}$), trained and validated on data from D_O and used to classify events from D_E. With this approach, it is guaranteed that the training of the NN is not biased and evaluation happens always on an independent data set not used during the training phase. In the following, the indices of the even and odd folds will be dropped and it will be referred to as *the* NN, even though there are technically two NNs trained.

The classification task is formulated as minimization problem of the *average categorical cross entropy loss* (L_CE) which is defined as

$$L_\mathrm{CE} = -\frac{1}{N}\sum_{i=1}^{N} w^{(i)} \sum_{j=1}^{n_y} t_j^{(i)} \ln(p_j) \ , \tag{9.19}$$

where N denotes the number of events and $w^{(i)}$ encodes the importance of each event for the classification task. The indicator function ($t_j^{(i)}$) is given by

$$t_j^{(i)} = \begin{cases} 1 & \text{if event } i \text{ belongs to the class } j \\ 0 & \text{otherwise} \end{cases} \ . \tag{9.20}$$

The predicted output probability (p_j) is given by the NN output (see also Equation (9.17))

$$p_j = \left(\Xi^{(\lambda)}(\{x_i\})\right)_j \ , \tag{9.21}$$

where the NN itself depends on the set of inputs, $\{x_i\}$. After evaluation of the loss function in Equation (9.19), the parameters of the NN ($\{\lambda\}$) get updated using a gradient descent algorithm

$$\lambda_p - \eta_p \frac{\partial L_\mathrm{CE}}{\partial \lambda_p} \to \lambda_p \ , \tag{9.22}$$

where η_p is the learning rate. For this analysis the learning rate (η_p) is initialized to a value of 10^{-4} for all parameters and are altered individually according to the ADAM algorithm [224]. For large training data sets as used in this analysis, it is not advisable to use the whole training set for evaluation of the loss function, L_CE. Modern machine learning applications frequently use a subset of the training set, called *(mini)batch*, to evaluate the loss and update the parameters of the NN. In this analysis the batches are built using the *balanced batch* technique [225]. For every class in every data taking period, 30 events are randomly selected with replacement. The resulting number of events per batch are summarized in Table 9.2. The weights, $w^{(i)}$, in Equation (9.19) are normalized per class within each batch. As a consequence, each class has the same importance to the NN, in other words, there is no emphasis on correctly classifying a particular class. More details on the balanced batch approach in the context of this analysis can be found in [226].

Weights ($w^{(i)}$ in Equation (9.19)) determine how much the particular event can impact the NN output. For simulated events, weights are used for example to scale to the correct luminosity or to apply certain corrections. The training of the NN relies on labeled input data (supervised learning), i.e. the true class of an event has to be known, suggesting its limitation to the usage of simulated samples only. However, defining a dedicated background class for τ-embedded samples and one for jet $\to \tau_{\mathrm{h}}$ implies the usage of real collision data during training. For the jet $\to \tau_{\mathrm{h}}$ class, data from the AR are used and the computed event-based F_{F} (see Equations (8.56) and (8.60)) is included as a weight in the training via Equation (9.19). As discussed in Section 8.2.3, not all events in the AR are jet $\to \tau_{\mathrm{h}}$ but they are nevertheless included for the training. The subtraction of other processes than jet $\to \tau_{\mathrm{h}}$ is then performed later in the analysis, at the histogram-level. More details on the training with background classes derived from data-driven techniques used in this analysis are given in [226].

In order to avoid overtraining[3] of the NN, several regularization techniques are implemented:

- If the loss, L_{CE}, evaluated on the validation set is not decreasing for 50 epochs, the training is interrupted. This method is called *early stopping*. In the adopted balanced batch approach, an epoch is defined by 1000 batches, i.e. 1000 updates according to Equation (9.22) to the weights and biases of the NN.
- Another technique to reduce the risk of overtraining is called *dropout* [227]. The dropout rate in this analysis is 30%, meaning that in each hidden layer, 30% of nodes are *deactivated*. Which nodes are affected is determined randomly for each batch. All weights and biases associated to the deactivated nodes are not subject to any change according to Equation (9.22). It should be mentioned, that the dropout technique also helps in reducing the complexity of the NN and guides the NN to explore other connections during training.
- The weights and biases are penalized with the L2 norm by adding the sum of squares of each parameter to the loss function

$$L_{\mathrm{CE}} \to L_{\mathrm{CE}} + \delta \sum_{p} \left\| \lambda_p \right\|^2 \ , \tag{9.23}$$

  with $\delta = 10^{-5}$. This kind of regularization penalizes large parameters.

Detailed studies [228] have been carried out to reduce the set of input variables in the context of this analysis, starting from a large set of inputs. The studies make use of a metric proposed in [229] based on a Taylor expansion of the NN. It turns out that the discrimination power of an NN is not only based on single input variables, which are encoded in the first order Taylor expansion coefficients. Correlations between variables, encoded in second order Taylor coefficients, prove to be just as useful to increase the classification performance of an NN. Based on the studies documented in [228], for each channel just a single NN is trained with the same input variables. These input variables are summarized in Table 9.3 and belong to the following three types (see also Figure 9.1):

- The two input variables N_{jets} and $N_{\text{b-jet}}$ are discrete, i.e. they are integer valued.
- All variables about transverse momenta, masses, pseudorapidity and MELA [230] are continuous.

[3] Overtraining happens if the NN adapts to sample specific in contrast to generic features of the training data set. In that case, the classification performance on training data keeps increasing but the correct classification on independent events from the validation set decreases.

Variable	Description
$p_{\mathrm{T}}^{(\tau_1)}$	Transverse momentum of the leading tau of the signal tau pair
$p_{\mathrm{T}}^{(\tau_2)}$	Transverse momentum of the sub-leading tau of the signal tau pair
$p_{\mathrm{T}}^{(\mathrm{jet}_1)}$	Transverse momentum of the leading jet
$p_{\mathrm{T}}^{(\mathrm{jet}_2)}$	Transverse momentum of the sub-leading jet
$p_{\mathrm{T}}^{(\mathrm{vis})}$	Transverse momentum of the visible decay products of the signal tau pair
$p_{\mathrm{T}}^{(\mathrm{jj})}$	Transverse momentum of the first two leading jets
$m_{\text{SV-fit}}$	Invariant mass of the di-tau system including SVFIT [139]
m_{vis}	Invariant mass of the visible decay products of the signal tau pair
m_{jj}	Invariant mass of the leading two jets in the event
N_{jets}	Number of jets
$N_{\text{b-jet}}$	Number of b-tagged jets
$\Delta\eta^{(\mathrm{jj})}$	Difference in pseudorapidity of the first two leading jets
$Q^2_{\mathrm{V}_1}$	Variables based on the MELA approach [230]. Under the VBF hypothesis, they
$Q^2_{\mathrm{V}_2}$	correspond to the momentum transfer of the two vector bosons, respectively.
2016	One-hot encoded year of data taking
2017	One-hot encoded year of data taking
2018	One-hot encoded year of data taking

Table 9.3: This table lists the different input variables entering the NN. As a function of these variables, the NN assigns events to different signal and background classes.

- The information about the year of data taking is entering the NN as *one-hot* encoded. That means that three input nodes are added, one for each year of data taking (2016/17/18). A value of one is assigned to the input to which the event belongs to and a zero to the other years, e.g. an event recorded in 2017 is represented as

$$\begin{aligned} 2016 &: 0 \\ 2017 &: 1 \\ 2018 &: 0\,. \end{aligned} \tag{9.24}$$

After training of the NN, the classification is performed as indicated in Figure 9.1. Each event is assigned to the category with the highest output probability given by the NN output. Histograms are filled for each signal and background class binned in the output probability (p_j) of the NN. These histograms enter a maximum likelihood fit and are used to perform the signal inference as will be discussed in more detail throughout Sections 9.2.3 and 9.2.4.

9.2.2 Uncertainty Model

In order to extract the POIs, a binned maximum likelihood fit according to Equation (9.9) is performed. In this section, the systematic uncertainties of the SM $\mathrm{H} \to \tau\tau$ analysis entering the likelihood ($\mathcal{L}(\boldsymbol{\mu};\boldsymbol{\theta})$) as nuisance parameters ($\boldsymbol{\theta}$) are described. Special emphasis is given on uncertainties related to the FF method (see Section 8.2.4). For completeness, the uncertainties associated to simulated and τ-embedded samples are also discussed.

	source	W+jets	$t\bar{t}$	QCD multijet	total
semi-leptonic	raw F_F measurement (stat.)	6	2	3	11
	F_F correction (stat.)	2	3	1	6
	F_F correction (syst.)	2	3	1	6
	MC subtraction	1	–	1	2
	fractions	1	–	–	1
	normalization (N_{jets} binned)		3		3
	pure normalization (lnN)		2		2
	total				31
fully hadronic	raw F_F measurement (stat.)	–	–	3	3
	F_F correction (stat.)	–	–	4	4
	F_F correction (syst.)	–	–	4	4
	MC subtraction	–	–	1	1
	extra uncertainty	1	1	–	2
	normalization (N_{jets} binned)		3		3
	pure normalization (lnN)		2		2
	total				19

Table 9.4: This table summarizes the F_F-related systematic uncertainties of the SM $H \to \tau\tau$ analysis. They enter the maximum likelihood fit as nuisance parameters (see also Equation (9.9)). For the semi-leptonic channels, a total of 31 systematic uncertainties are part of the uncertainty model, whereas for the fully hadronic channel there are 19.

F_F-related Uncertainties

The measurement of each individual raw F_F as well as its corrections is subject to statistical uncertainties due to the limited sample size in the DRs, where the raw F_F or its corrections are derived. All these uncertainties are saved and propagated to the NN output. The discussion of the F_F-related uncertainties is split in two parts, where the semi-leptonic channels are discussed first and then the fully hadronic channel.

According to Table 9.4, there are eleven nuisance parameters related to the raw F_F measurement for semi-leptonic channels, listed as "raw F_F measurement (stat.)". All these nuisance parameters are determined by using the morphed variations derived from the (yellow) uncertainty bands as described in Section 8.2.4 (see also Equation (8.62) and Figure 8.32). The impact of those nuisance parameters on the final NN output is shown in Figure 9.2. As can be seen from the distributions shown in Figure 9.2, the effect of these nuisance parameters is at most 0.5% and is smaller than the statistical uncertainty in each bin of the NN output.

There are six nuisance parameters related to the corrections of the different raw F_F components which also stem from the statistical uncertainties during their derivation, reflected in the (yellow) uncertainty bands. They are referred to as "F_F correction (stat.)" in Table 9.4. Their impact on the NN output is shown in the top row of Figure 9.3 and is comparable to the statistical uncertainty within the bins of the shown histograms. Only the impact of the variations of the closure correction ($C^{(QCD)}$) of $F_F^{(QCD)}$ in the $\mu\tau_h$ channel is larger, being in the order of 1%. This is consistent with the uncertainty bands of closure corrections shown in Figure 8.15 since the uncertainty band in the $\mu\tau_h$ channel is significantly larger than the one in the $e\tau_h$ channel.

All the uncertainties discussed so far are normalized to the same yield as the nominal value, thus having a pure shape-altering effect. Uncertainties from these normalization

factors are added in quadrature and serve as normalization uncertainties. In case of the normalization factors related to the raw F_{F} measurement (see Figure 9.2), the determination of the normalization effect is done per N_{jets} category and is listed as "normalization (N_{jets} binned)" in Table 9.4. These normalization uncertainties can still have a shape-altering effect. For the normalization factors related to the corrections, just a single nuisance parameter is added which corresponds to one of the two pure normalization uncertainties listed in Table 9.4. On the bottom of Figure 9.3, these normalization uncertainties are depicted, reaching values up to 5–6%. Indeed, the nuisance parameters binned in N_{jets} exhibit also a shape-altering effect.

For each closure and bias correction an additional nuisance parameter is added. These additional nuisance parameters quantify the magnitude of each correction to the raw F_{F}, ultimately allowing the final maximum likelihood fit to adjust the applied correction. Technically this is achieved by defining the down variation as not applying the correction and the up variation by applying the correction twice as also discussed in Section 8.2.4. In Table 9.4, these nuisance parameters are referred to as "F_{F} correction (syst.)". The top row of Figure 9.4 shows the impact of these systematic variations on the NN output. From the larger y-range in the ratio plots, it is apparent that these uncertainties are the most dominant discussed so far and typically exceed the statistical uncertainties within each bin. For the $\mathrm{e}\tau_{\mathrm{h}}$ channel in 2018 data, the closure correction of $F_{\mathrm{F}}^{(\mathrm{QCD})}$ has the most dominant systematic effect which is in the order of 10%. The closure correction, depicted in Figure 8.15, applies corrections to $F_{\mathrm{F}}^{(\mathrm{QCD})}$ in the range $[0.75, 1.10]$ for the $\mathrm{e}\tau_{\mathrm{h}}$ channel, providing a consistent picture about the dominance of this nuisance parameter. For the $\mu\tau_{\mathrm{h}}$ channel, however, the correction of $F_{\mathrm{F}}^{(\mathrm{QCD})}$ is not that large as can be seen in Figure 8.15. In case of the $\mu\tau_{\mathrm{h}}$ channel, the dominant nuisance parameter is related to the bias correction of $F_{\mathrm{F}}^{(\mathrm{W+jets})}$, with an impact in the order of 7%. Comparing to Figure 8.12, which shows for the $\mu\tau_{\mathrm{h}}$ channel a bias correction of 0.9 for a wide range of m_{vis}, again explains the dominance of that particular nuisance parameter.

Two additional nuisance parameters are added according to Table 9.4, referred to as "MC subtraction", to cover the effect of subtracting events inside the DRs. One nuisance parameter is added related to the measurement of $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ and one related to $F_{\mathrm{F}}^{(\mathrm{QCD})}$ as explained in Section 8.2.4. Furthermore, one nuisance parameter is added varying the fraction of W+jets events inside the AR. Finally, one pure normalization uncertainty is added related to the subtraction of genuine τ and light lepton misidentifications (termed as "other bkg." in Figure 8.3) inside the AR. The bottom row of Figure 9.4 depicts these final nuisance parameters in the uncertainty model of the semi-leptonic channels. Their magnitude ranges up to values of 2%.

For the fully hadronic channel, there are less nuisance parameters related to the limited sample size in the derivation of the raw F_{F} and its corrections since there is only one F_{F} measured, being $F_{\mathrm{F}}^{(\mathrm{QCD})}$. The same morphing technique as explained in Section 8.2.4 is used to obtain the up and down variations of these nuisance parameters[4]. The top row of Figure 9.5 shows these nuisance parameters, which are in the sub-percent level and smaller than the statistical uncertainties in each of the bins of the shown histograms. These nuisance parameters are purely shape-altering and are normalized to the nominal yield.

In the same way as for the semi-leptonic channels, nuisance parameters are added such that they take the magnitude of the corrections and the normalization into account. These nuisance parameters are shown in the middle row of Figure 9.5. Their magnitude

[4]This is also true for the nuisance parameters related to the measurement of the raw $F_{\mathrm{F}}^{(\mathrm{QCD})}$, which uses a third order polynomial. Thus, the dependence of the up and down variations is approximated to be linear.

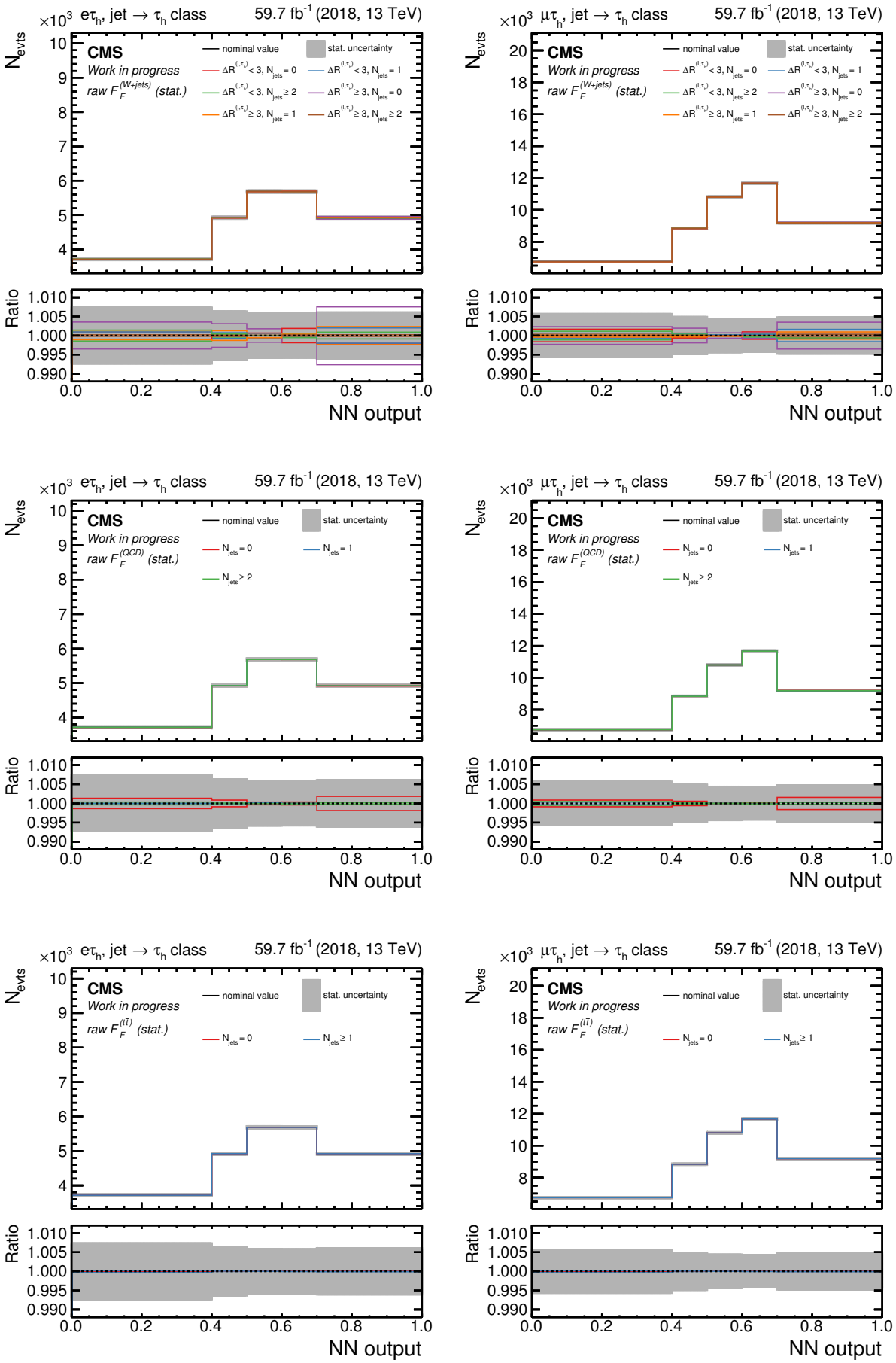


Figure 9.2: The caption is on the next page.

Figure 9.2: The number of events (N_{evts}), classified as jet $\rightarrow \tau_{\text{h}}$ by the NN, as a function of the NN output score is shown in this figure. The nominal distribution is shown as a black line and its statistical uncertainty as gray band. All distributions are from the 2018 data-taking period, the $e\tau_{\text{h}}$ channel is shown on the left and the $\mu\tau_{\text{h}}$ channel on the right. The top row shows the morphed uncertainty variations originating from the derivation of $F_{\text{F}}^{(\text{W+jets})}$ (see Figures 8.7 and 8.8). The middle row shows the morphed uncertainty variations originating from the derivation of $F_{\text{F}}^{(\text{QCD})}$ (see Figure 8.14). The bottom row shows the morphed uncertainty variations originating from the derivation of $F_{\text{F}}^{(\text{t}\bar{\text{t}})}$ (see Figure 8.21).

is in the percent level reaching values of 5–7%. The bottom row of Figure 9.5 shows the nuisance parameters covering the effect of changing the yield of events subtracted during the determination of $F_{\text{F}}^{(\text{QCD})}$ and the normalization uncertainty due to the subtraction of genuine τ and light lepton misidentification inside the AR. These two nuisance parameters are also part of the uncertainty model in the semi-leptonic channels. The normalization uncertainty due to subtraction of events is in the order of 2%, while the effect of subtraction on the derivation of $F_{\text{F}}^{(\text{QCD})}$ has only negligible effects. Lastly, two specific nuisance parameters for the $\tau_{\text{h}}\tau_{\text{h}}$ channel are also shown in the bottom plot of Figure 9.5. They are introduced to cover variations of W+jets and $\text{t}\bar{\text{t}}$ yields inside the AR, taking into account the application of $F_{\text{F}}^{(\text{QCD})}$ instead of dedicated F_{F}'s. It turns out, however, that they have a negligible effect and their impact is on the sub-percent level, well below the statistical uncertainty of the nominal yield.

Uncertainties Related to τ Embedding and Simulation

There are common uncertainties to τ-embedded and fully simulated samples regarding the reconstruction and identification of particle objects (electrons, muons and τ_{h}'s) because the tau leptons in the τ-embedded samples are also simulated. A dedicated scheme is used to introduce correlations between τ-embedded and simulated samples, where a 50% correlation between the two is applied. More details on the correlation scheme can be found in [12]. In addition, nuisance parameters can also be modeled as fully correlated or uncorrelated among processes, years or decay channels. Fully correlated processes share the same nuisance parameter whereas uncorrelated effects are modeled by separate nuisance parameters.

All corrections discussed in Section 6.5 have an associated uncertainty which is propagated to the NN output. Variations within the corresponding uncertainties can have yield and/or shape-altering effects on the NN output and are treated as nuisance parameters in the likelihood fit. Nuisance parameters that go beyond the ones discussed in Section 6.5 are theoretical cross section uncertainties as well as uncertainties on branching ratios. In accordance with the STXS stage-1.2 scheme, further nuisance parameters are introduced, modeling migrations among separate STXS bins as well as shape-only changes of single STXS bins.

The limited sample size of simulated samples, or data in case of data-driven background modeling, results in a last set of systematic uncertainties. They are modeled with the Barlow-Beeston approach [231].

9.2.3 STXS Stage-0 Results

The NN trained for the STXS stage-0 measurement has seven classes (output nodes) for the semi-leptonic channels and five classes in case of the fully hadronic channel (see Table 9.2). Two of the classes are given by the ggH and qqH signal classes as shown in Figures 5.3

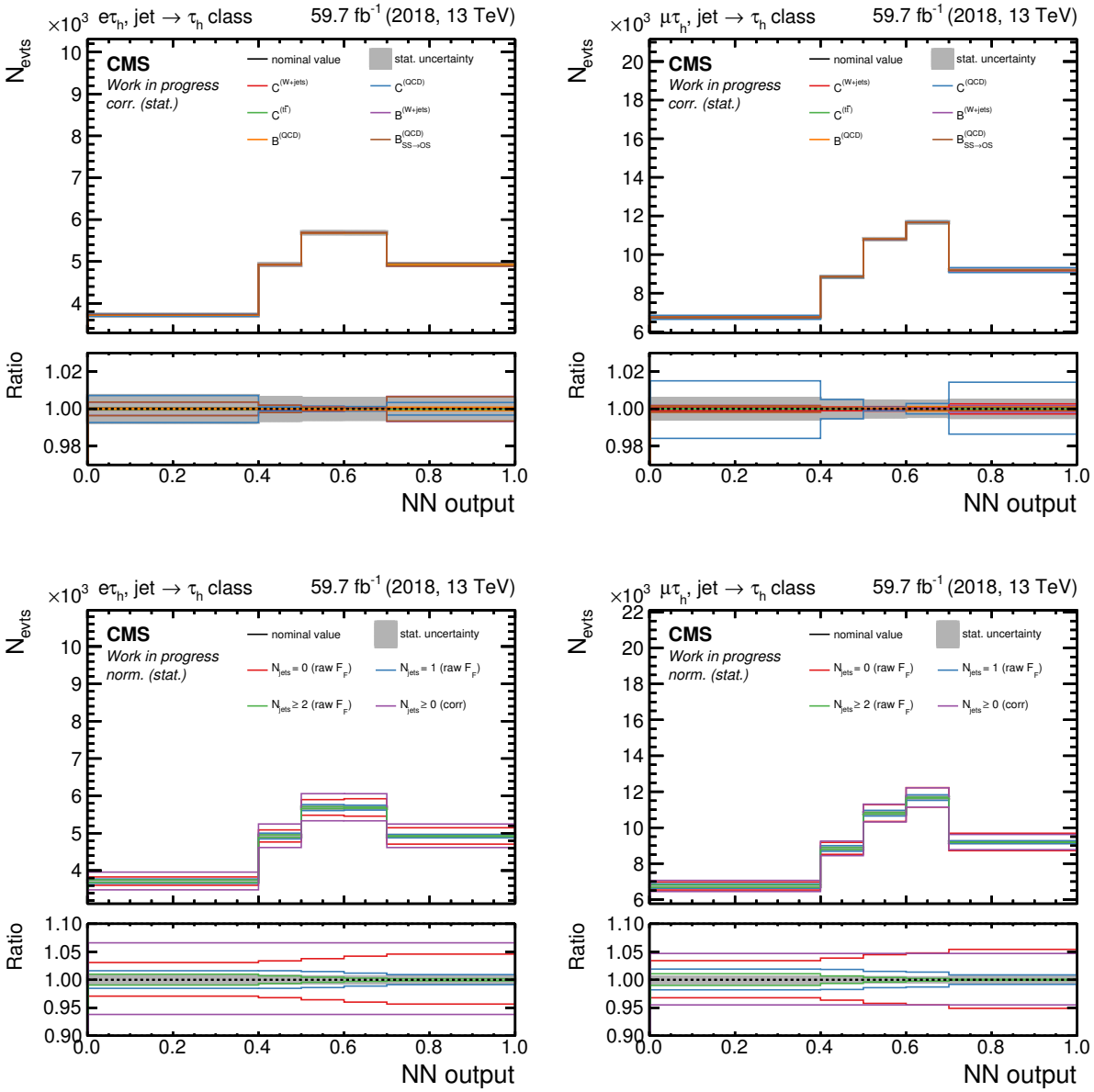


Figure 9.3: The number of events (N_{evts}), classified as jet $\rightarrow \tau_{\text{h}}$ by the NN, as a function of the NN output score is shown in this figure. The nominal distribution is shown as a black line and its statistical uncertainty as gray band. All distributions are from the 2018 data-taking period, the eτ_{h} channel is shown on the left and the $\mu\tau_{\text{h}}$ channel on the right. The top row shows the morphed uncertainty variations originating from the derivation of the corrections applied to $F_{\text{F}}^{(\text{W+jets})}$, $F_{\text{F}}^{(\text{QCD})}$ and $F_{\text{F}}^{(\text{t}\bar{\text{t}})}$. Corrections include closure (see Figures 8.10, 8.15 and 8.21) and bias (see Figures 8.12, 8.17 and 8.20) corrections. The bottom row shows the normalization uncertainties. Normalization uncertainties related to the F_{F} derivation are binned in N_{jets}, whereas there is just a single normalization uncertainty for the corrections.

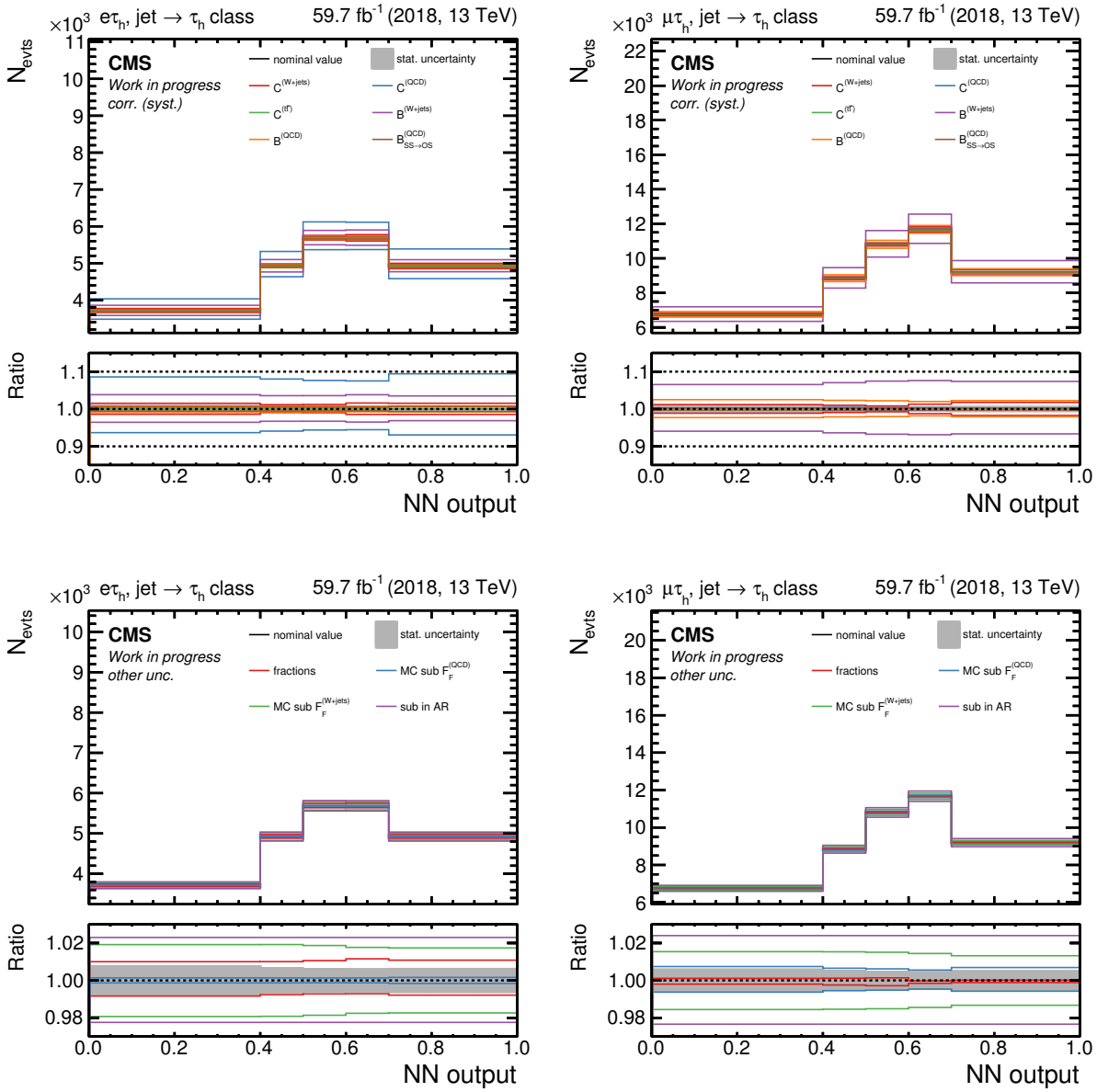
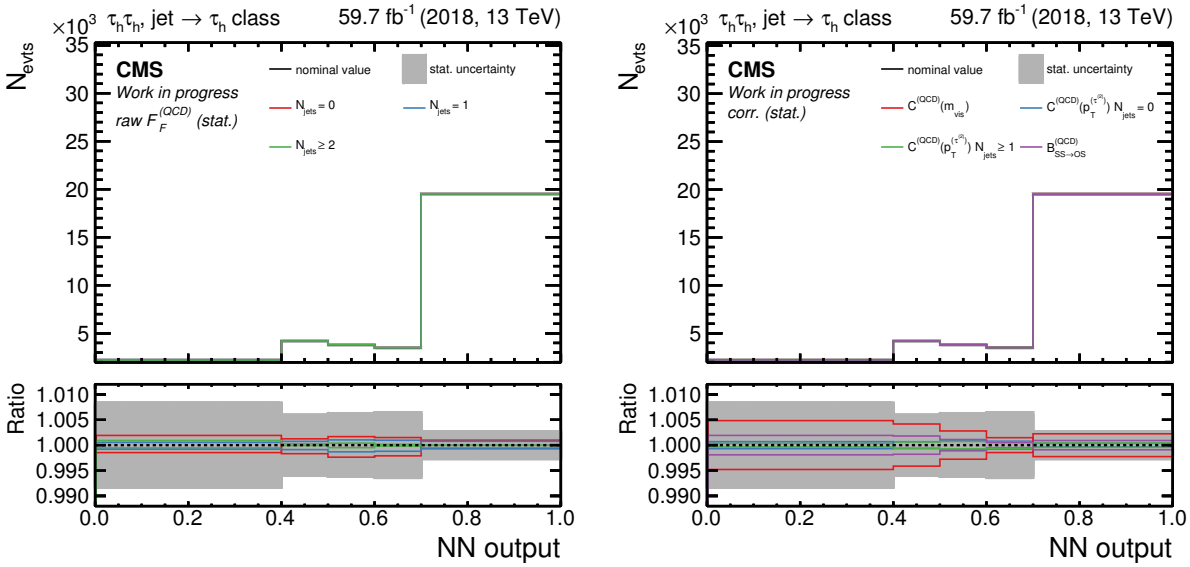
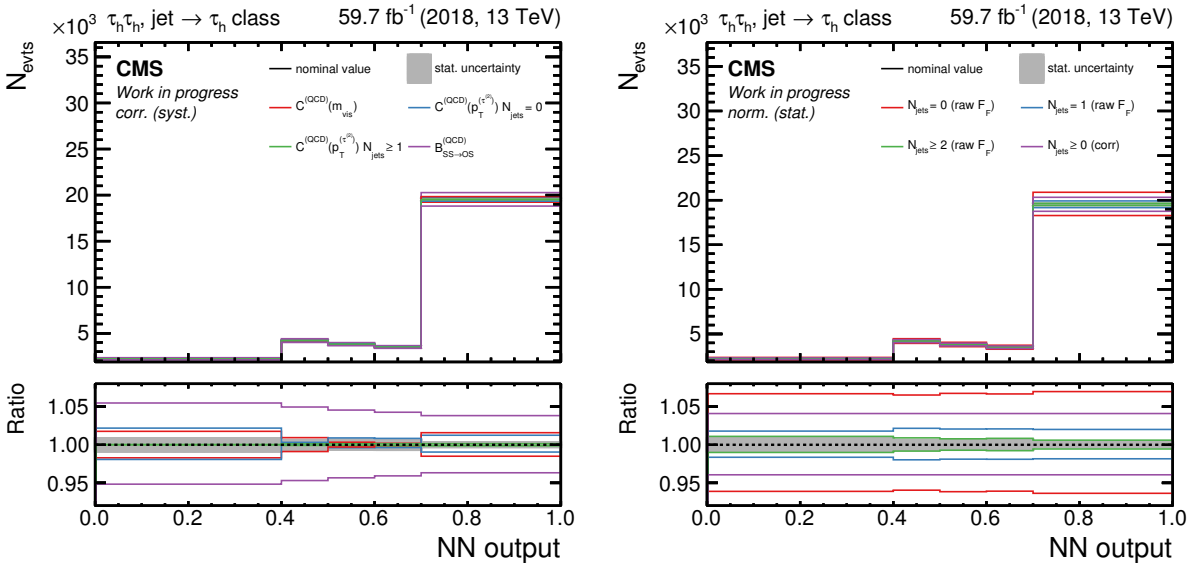


Figure 9.4: The number of events (N_{evts}), classified as jet $\rightarrow \tau_{\text{h}}$ by the NN, as a function of the NN output score is shown in this figure. The nominal distribution is shown as a black line and its statistical uncertainty as gray band. All distributions are from the 2018 data-taking period, the eτ_{h} channel is shown on the left and the $\mu\tau_{\text{h}}$ channel on the right. The top row shows the systematic variations of the different F_{F} corrections, obtained by not applying the correction or applying it twice. Some of these uncertainties are in the order of 10% and are thus the most dominant F_{F} uncertainties. The bottom row shows the shape-altering uncertainties originating from subtracting events based on simulation and τ-embedded samples in the derivation of the F_{F} or varying the fractions inside the AR (see Figure 8.27). A pure normalization uncertainty related to the subtraction of genuine and other backgrounds in the AR (see Equation (8.57)) is also shown.





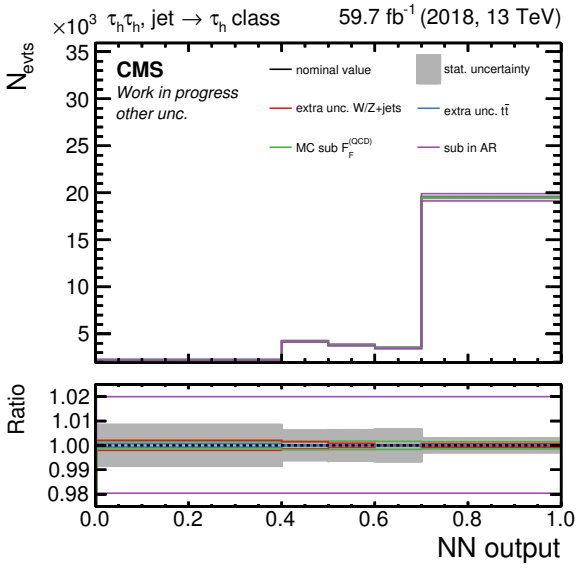


Figure 9.5: The caption is on the next page.

Figure 9.5: The number of events (N_{evts}), classified as jet $\rightarrow \tau_{\text{h}}$ by the NN, as a function of the NN output score is shown in this figure. The nominal distribution is shown as a black line and its statistical uncertainty as gray band. All distributions are from the 2018 data-taking period for the $\tau_{\text{h}}\tau_{\text{h}}$ channel. The top row shows the morphed uncertainty variations, originating from the derivation of $F_{\text{F}}^{(\text{QCD})}$ and its corrections (see Figures 8.24 and 8.25). The middle left plot shows the systematic variations of the different F_{F} corrections obtained by not applying the correction or applying it twice. The middle right plot shows the normalization uncertainties. Normalization uncertainties related to the F_{F} derivation are binned in N_{jets}, whereas there is just a single normalization uncertainty for the corrections. The bottom row shows the shape-altering uncertainties, originating from subtracting events based on simulation and τ-embedded samples in the derivation of the $F_{\text{F}}^{(\text{QCD})}$. A pure normalization uncertainty, related to the subtraction of genuine and other backgrounds in the AR (see Equation (8.57)), is also shown. Furthermore, two nuisance parameters related to the application of $F_{\text{F}}^{(\text{QCD})}$ to W+jets and $\text{t}\bar{\text{t}}$ events in the AR are shown in the bottom plot.

and 5.4, respectively. A random assignment of an event to a class (j) would happen, if the NN output probability is the same for each class

$$p_j = \frac{1}{n_y} = \begin{cases} 1/7 \text{ for semi-leptonic channels} \\ 1/5 \text{ for the fully hadronic channel .} \end{cases} \tag{9.25}$$

The probability boundaries given in Equation (9.25) present the lower bound of an event being assigned to a certain class. For each class, the distribution of events, as a function of the NN output probability that lead to their classification, is filled into a histogram. In Figure 9.6, an example is shown for events classified as jet $\rightarrow \tau_{\text{h}}$ using data from 2018. A first observation is that the dominant process within this class is indeed jet $\rightarrow \tau_{\text{h}}$. Furthermore, the higher the NN output (probability), the more enriched the corresponding bin is in jet $\rightarrow \tau_{\text{h}}$. This indicates that the NN is capable of separating jet $\rightarrow \tau_{\text{h}}$ events from other processes and aggregating them in a dedicated class.

A more condensed way of quantifying the overall success of NN classifications is by means of *confusion matrices*. An example of such a confusion matrix is shown in Figure 9.7. Each column of the confusion matrix is normalized to one, meaning that each matrix entry represents the fraction of genuine events of this class, assigned to the respective predicted class. Hence, the diagonal of the confusion matrix represents the percentage of correctly predicted events for each class. Processes like genuine τ, $\text{t}\bar{\text{t}}$, $\text{Z} \rightarrow \ell\ell$ and qqH reach percentages around 70% and more of correct classification. For the other signal class (ggH), the mis-classification is higher. From the confusion matrix, it follows that ggH events mainly get mis-interpreted by the NN as qqH or genuine τ. Similarly for jet $\rightarrow \tau_{\text{h}}$, a certain fraction of such events get mis-assigned to be genuine τ events or originating from a process summarized in the misc class. This is mainly because jet $\rightarrow \tau_{\text{h}}$ themselves are composed of several different processes and thus a correct classifications by the NN is more difficult to achieve.

As stated in the caption of Figure 9.6, the uncertainties shown include all variations before the maximum likelihood fit, modeled by nuisance parameters ($\boldsymbol{\theta}$) and summarized in Section 9.2.2. For every background class, the corresponding histogram is used in the maximum likelihood fit of $\mathcal{L}(\boldsymbol{\mu}; \boldsymbol{\theta})$ defined in Equation (9.9). There are three background classes for the $\tau_{\text{h}}\tau_{\text{h}}$ channel and five background classes for each other channel. Other channels comprise the semi-leptonic channels, $\text{e}\tau_{\text{h}}$ and $\mu\tau_{\text{h}}$, but also the fully leptonic $\text{e}\mu$ channel, which covers 6% of the di-tau final states (see Table 6.1). A description of the $\text{e}\mu$

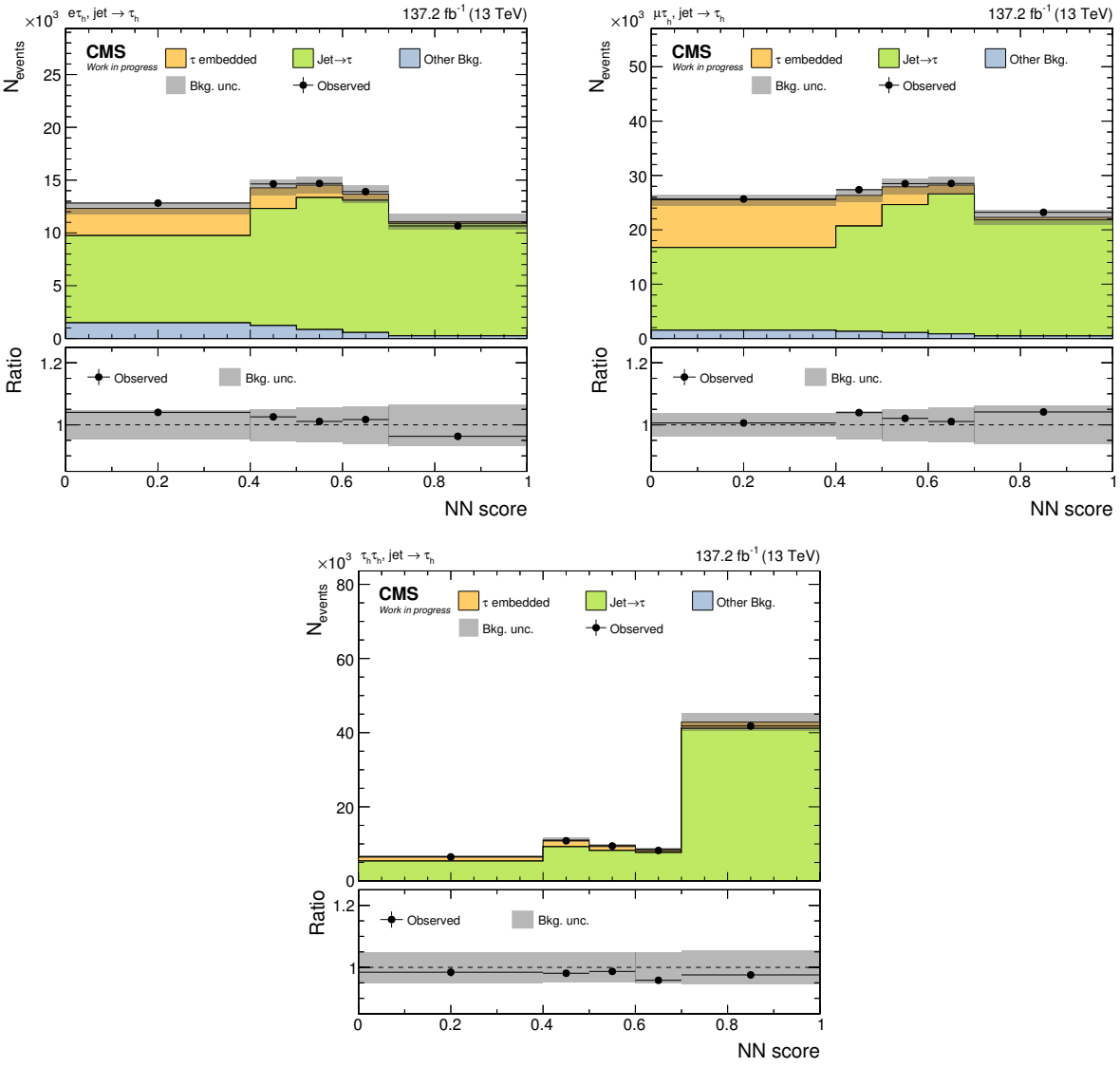


Figure 9.6: Event distributions are shown for the eτ_h, $\mu\tau_\mathrm{h}$ and $\tau_\mathrm{h}\tau_\mathrm{h}$ channel. Only events assigned to the jet $\rightarrow \tau_\mathrm{h}$ class are shown here as indicated in the title of each plot. Contributions from true jet $\rightarrow \tau_\mathrm{h}$ processes are shown in green and dominate especially in bins with a large NN output score. Note that even thought the first bin of each histogram extends down to zero, the lower bound of the NN output for classification is given by the values quoted in Equation (9.25). Uncertainties referred to as "Bkg. unc." comprise the full uncertainty model as presented in Section 9.2.2 before the maximum likelihood fit. The same plots after the maximum likelihood fit are shown in Figure B.1. The figures are taken from [12].

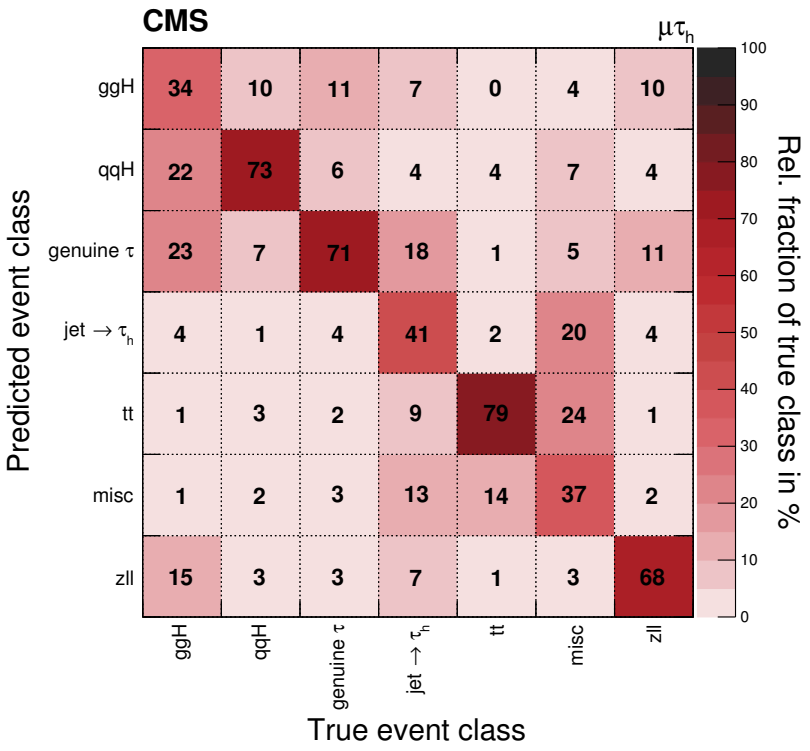


Figure 9.7: The confusion matrix for the $\mu\tau_{\mathrm{h}}$ channel using data from 2016–2018 is shown. Classification is performed with an NN targeting the STXS stage-0 scheme, i.e. using two signal classes (ggH and qqH) as shown in Figures 5.3 and 5.4. Furthermore, five background classes are defined in accordance with Table 9.1. On the y-axis, the predicted class is shown and the true class is displayed on the x-axis. The values of the matrix are such that they are normalized to one in each column. This means that each matrix entry gives the percentage of the true class that is assigned to the particular predicted class. The figure is taken from [7].

channel and the relevant backgrounds can be found in [7, 12]. Additional four histograms, one for each channel, that correspond to the signal class enter the maximum likelihood fit. These histograms combine the output score for ggH and qqH as discussed in detail in [12]. Each histogram is produced for each year of the three data-taking years. To obtain the inclusive STXS stage-0 result, a single POI is used to scale all signal processes simultaneously. In a second fit, three POIs are introduced, one that scales the ggH, one that scales the qqH and one that scales the VH contributions with respect to the SM expectation. Details on the addition of the VH production modes can be found in [7] and are not discussed in this thesis. Results of these fits are shown in Figure 9.8. In Figure 9.8, results from the NN approach presented in this thesis are labeled with NN. In addition, results from an independent analysis, using a cut-and-count approach for event classification, are also shown in Figure 9.8 and labeled as CB. Uncertainties on the best fit value of the POIs are obtained by the profile likelihood method [232]. The total uncertainty is split into four sources

- purely statistical (stat.),
- systematic uncertainty from experiment (syst.),
- systematic uncertainty from theory (theo.) and
- bin-by-bin uncertainties (bbb) due to finite sample sizes in the modeling of backgrounds and signals.

Each nuisance parameter falls into one of the above sources. Individual uncertainties are obtained by sequentially fixing all nuisance parameters to their best fit values but the ones from one source. The obtained uncertainty is subtracted in quadrature from the total uncertainty and assigned to the particular source. Since there is no large correlation among different sources expected, this approach is justified.

The measured inclusive signal strength of $\mu = 0.82^{+0.11}_{-0.10}$ is compatible with the SM expectation within two standard deviations. For the STXS stage-0 measurement, the following signal strengths are obtained

$$\begin{aligned} \mu_{\mathrm{ggH}} &= 0.67^{+0.20}_{-0.18} \\ \mu_{\mathrm{qqH}} &= 0.81^{+0.17}_{-0.16} \\ \mu_{\mathrm{VH}} &= 1.79^{+0.47}_{-0.42} \,. \end{aligned} \tag{9.26}$$

From the different uncertainty sources shown Figure 9.8, it follows that the uncertainty on the μ_{ggH} result is dominated by systematics whereas for μ_{qqH} it is statistically dominated. Furthermore, a linear correlation coefficient of -0.35 between μ_{ggH} and μ_{qqH} is in line with observing migrations of ggH events into the qqH class (see Figure 9.7). In the context of the finer granularity of the STXS stage-1.2 results discussed in Section 9.2.4, it will become even clearer which ggH topologies show the largest misclassification with qqH signal classes or the genuine τ background class.

Figure 9.8 also shows the likelihood scan where the two POIs are the coupling strength modifiers, κ_{V} and κ_{F}, as discussed in Section 5.1.2 (see also Equation (5.6)). Contours corresponding to 68% and 95% confidence intervals for the expectation, and as expected for the SM, are shown in Figure 9.8. The obtained NN result includes the expected values of $\kappa_{\mathrm{V}} = 1$ and $\kappa_{\mathrm{F}} = 1$ within the contour of 95% confidence level. The fact that the best fit values of both coupling strength modifiers are below one agrees with the measurement of μ_{ggH} and μ_{qqH}, being both smaller than one, i.e. being below the SM expectation.

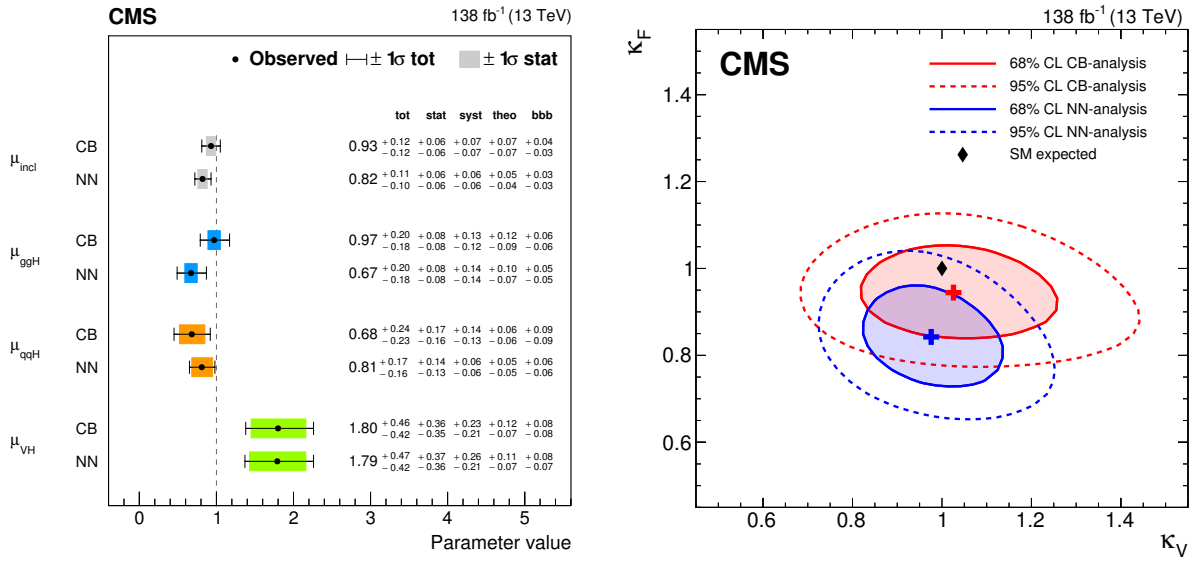


Figure 9.8: The left plot shows the results of the inclusive and STXS stage-0 measurement. Stage-0 measurements include the VH Higgs boson production mode. Details on the inclusion of the VH signal process can be found in [7] and are not covered in this thesis. Results are presented as signal strength modifiers (μ) with respect to the SM expectation. The total uncertainty is split into different sources as explained in the text. The right plot shows the STXS stage-0 results in form of κ_V-κ_F contours (see Section 5.1.2), i.e. treating the coupling strength modifiers (κ) as POIs. Both plots are taken from [7]. Results from the NN classification as presented in this thesis are labeled with NN in the plots. An independent analysis, implementing a cut-and-count approach for event categorization, is labeled with CB in the plots.

9.2.4 STXS Stage-1.2 Results

For the STXS stage-1.2 measurement, NNs are trained with 20 and 18 classes for the semi-leptonic and fully hadronic channels, respectively. The amount of background classes is the same as for the STXS stage-0 measurement. It is the amount of signal classes which is larger, targeting specific bins of the STXS stage-1.2 scheme the analysis is sensitive to. For the ggH process, there are eleven signal classes that are differentiated as shown in Figure 5.3. Four signal classes are added based on the qqH process as depicted in Figure 5.4. Figure 9.9 shows the confusion matrix for the $\mu\tau_{\mathrm{h}}$ channel and 20 different output classes. In general, the NN is sensitive to the differences between the classes and predicts the correct class with efficiencies that reach up to 80%. Further observations are the following:

- Due to the finer granularity of the ggH process in the context of the STXS stage-1.2 scheme, it is possible to better identify which event topologies get confused with qqH and genuine τ (see also Figure 9.7). In case of mis-assignments to the genuine τ class, the ggH events are coming from the 0-jet categories which makes sense because at leading order both types of processes do not have extra jets (see Figures 5.1a and 6.5a). The other misclassification happens for ggH events with at least two jets and large invariant di-jet mass ($m_{\mathrm{jj}} > 350\,\mathrm{GeV}$). These events are confused by the NN as qqH events with at least two jets and large m_{jj}, i.e. VBF event topologies. Again this is in line with the similar event topologies of the two processes.

- A prominent misclassification that stands out of the general trend of good sensitivities to different classes, is the qqH process with less than two jets or $m_{\mathrm{jj}} < 350\,\mathrm{GeV}$ in events with at least two jets. This category is labeled as "no VBF topo." in Figure 5.4. Three ggH classes with at least two jets and $m_{\mathrm{jj}} < 350\,\mathrm{GeV}$ catch about 60% of these events. As in the previous point, the similar event topologies can explain these misclassifications. However, the opposite misclassification of ggH to qqH, seems not to occur. An explanation for this observation is that the discussed ggH classes are also split in the Higgs boson transverse momentum ($p_{\mathrm{T}}^{\mathrm{H}}$). As such, the NN assigns events compatible with a certain $p_{\mathrm{T}}^{\mathrm{H}}$ rather to a more specific ggH class than to the more inclusive qqH class.

- The classification performance among the background classes, in terms of percentages of correctly assigned events (diagonal elements), is similar to the STXS stage-0 training (see Figure 9.7). Increased misclassification rates in case of STXS stage-1.2 are expected because the number of different classes has increased and so did the variety of different topological signatures.

The POIs for the maximum likelihood fit vary the signal strengths individually in different STXS stage-1.2 bins. In total, twelve POIs are introduced to the likelihood in Equation (9.9), four for qqH bins and eight for ggH bins. This combination of some of the STXS stage-1.2 bins as defined in [7] is summarized in Figures 5.3 and 5.4. As input to the maximum likelihood fit, there are 234 one-dimensional histograms binned in the NN output score. The total number accounts for individual histograms for each of the four di-tau channels ($e\mu$, $e\tau_{\mathrm{h}}$, $\mu\tau_{\mathrm{h}}$ and $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$) and the three years of data taking. Results of the signal strength modifiers are shown in Figure 9.10 and are consistent between both, the NN and the CB analysis approach. They are compatible with SM expectations everywhere but in the 0-jet STXS stage-1.2 bins. Future combinations with other decay channels will help shed light on the situation in this particular STXS bin.

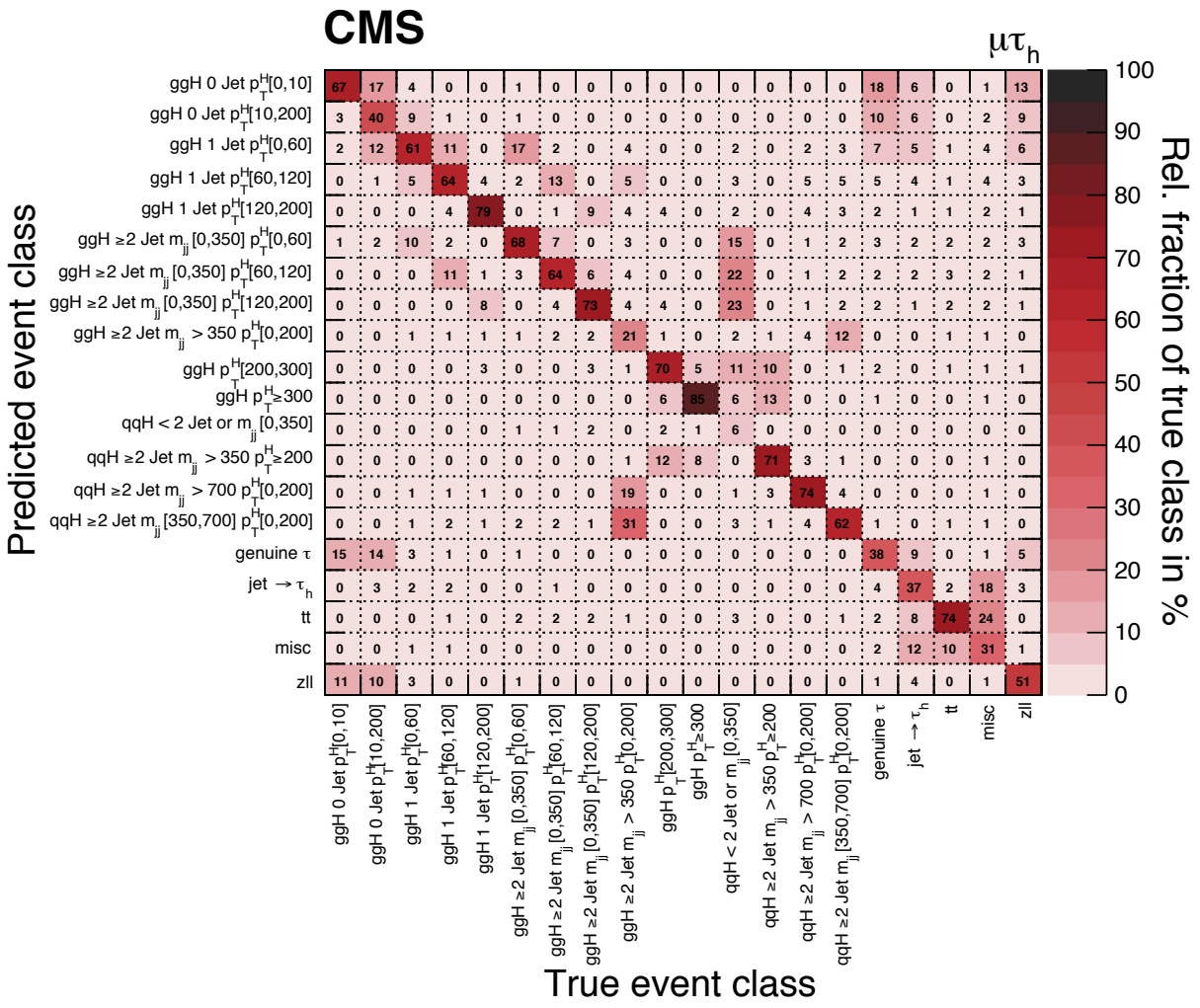


Figure 9.9: The confusion matrix for the $\mu\tau_{\mathrm{h}}$ channel using data from 2016–2018 is shown. Classification is performed with an NN targeting the STXS stage-1.2 scheme, i.e. using 15 signal classes (see Figures 5.3 and 5.4). Note, that the class "qqH $\geq$ 2 Jet $m_{\mathrm{jj}}[0, 350]$" comprises also those qqH events with less than two jets, labeled as "no VBF topo." in Figure 5.4. Furthermore, five background classes are defined in accordance with Table 9.1. On the y-axis the predicted class is shown and the true class is displayed on the x-axis. The values of the matrix are such that they are normalized to one in each column. This means that each matrix entry gives the percentage of the true class that is assigned to the particular predicted class. The figure is taken from [7].

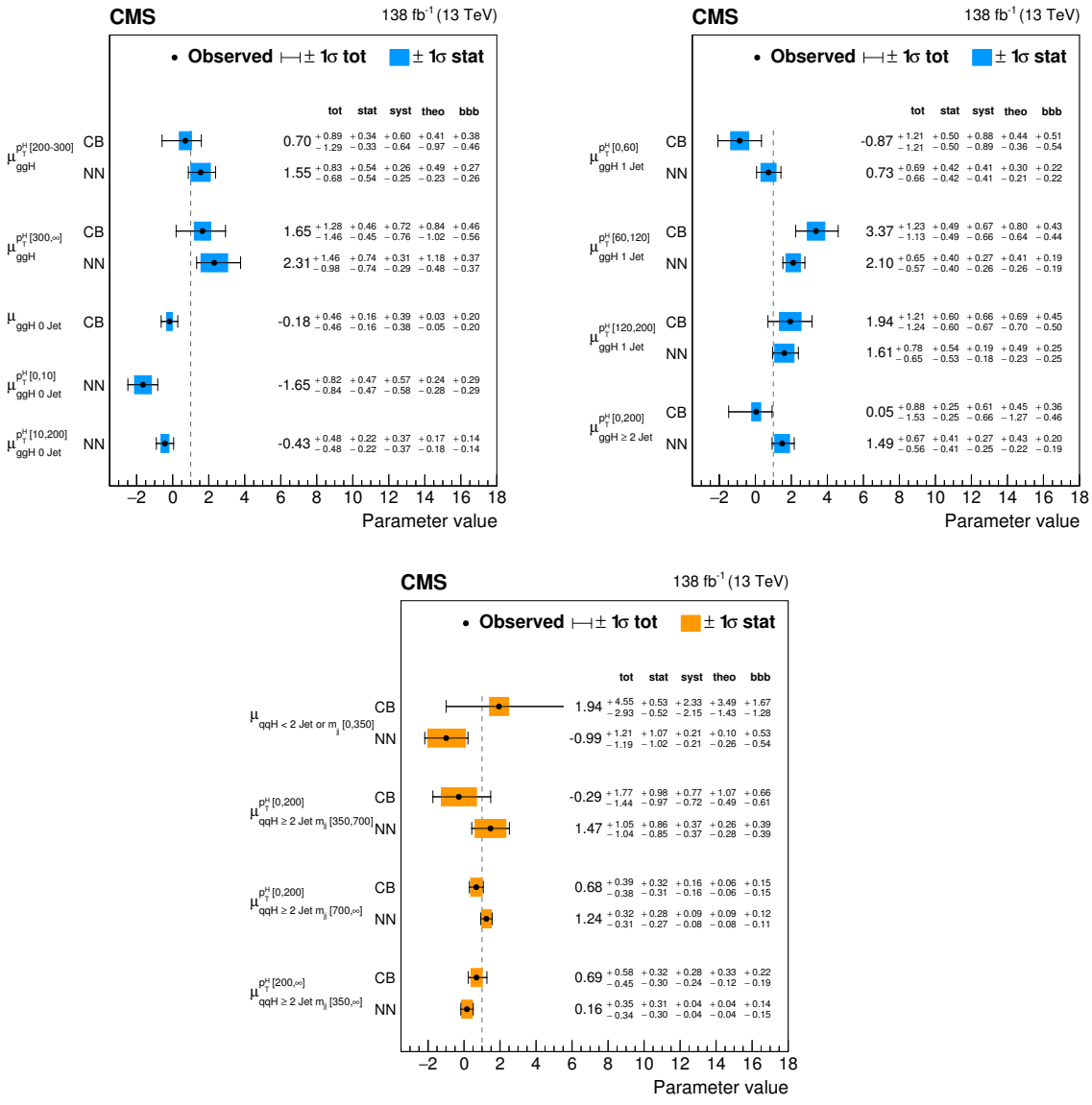


Figure 9.10: Results of the STXS stage-1.2 measurement are shown. They are presented as signal strength modifiers (μ) with respect to the SM expectation. Results of the analysis present in this thesis are labeled as NN. A reference analysis, using a cut-and-count based event categorization obtains the results labeled as CB. The top row shows signal strength modifiers for STXS stage-1.2 bins related to ggH production, the bottom row those related to qqH production. In all cases, the total uncertainty is split into different sources as explained in the main text. The plots are taken from [7].

9.3 Beyond the SM Searches

Three more analyses with di-tau final states are presented in Sections 9.3.1 and 9.3.2. All of them contain backgrounds arising from jet $\to \tau_{\mathrm{h}}$, which are estimated with the FF method explained in Section 8.2. Differences to the F_{F}'s used for the SM $\mathrm{H} \to \tau\tau$ analysis are discussed. However, further analysis related details are not covered and can be found in the quoted references. Section 9.3.3 covers some preliminary results for the search of light top squarks.

9.3.1 MSSM and NMSSM $\mathrm{H} \to \tau\tau$

The event categorization for the presented search for additional neutral Higgs bosons in the MSSM follows a cut-and-count approach which is based on the published analysis using 2016 data only [187]. Four final states are considered, $\mathrm{e}\mu$, $\mathrm{e}\tau_{\mathrm{h}}$, $\mu\tau_{\mathrm{h}}$ and $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$. Only the channels with at least one τ_{h} are discussed further. The categorization of those channels is summarized in Figure 9.11. For semi-leptonic channels, four categories are defined each, whereas for the fully hadronic channel only two categories are defined.

As final discriminant for the search of additional Higgs bosons, the total transverse mass ($m_{\mathrm{T}}^{\mathrm{tot}}$) is used, i.e. histograms binned in $m_{\mathrm{T}}^{\mathrm{tot}}$ are used for signal inference. To define $m_{\mathrm{T}}^{\mathrm{tot}}$, first the definition of the transverse mass given in Equation (6.2) has to be generalized. The transverse mass of two physics objects, o_1 and o_2, with transverse momenta $\mathbf{p}^{(o_1)}$ and $\mathbf{p}^{(o_2)}$, respectively, is defined as

$$m_{\mathrm{T}}(\mathbf{p}^{(o_1)}, \mathbf{p}^{(o_2)}) = \sqrt{2|\mathbf{p}^{(o_1)}| \cdot |\mathbf{p}^{(o_2)}| \cdot \left(1 - \cos\Delta\phi(\mathbf{p}^{(o_1)}, \mathbf{p}^{(o_2)})\right)} \,. \tag{9.27}$$

With this definition, the correspondence

$$m_{\mathrm{T,PUPPI}}^{(\ell)} = m_{\mathrm{T}}(\mathbf{p}_{\mathrm{T}}^{(\ell)}, \boldsymbol{p}_{\mathrm{T,PUPPI}}^{(\mathrm{miss})}) \,, \tag{9.28}$$

can be established and the definition of $m_{\mathrm{T}}^{\mathrm{tot}}$ can be formulated as

$$m_{\mathrm{T}}^{\mathrm{tot}} = \sqrt{\left(m_{\mathrm{T}}(\mathbf{p}_{\mathrm{T}}^{(\tau_1)}, \boldsymbol{p}_{\mathrm{T,PUPPI}}^{(\mathrm{miss})})\right)^2 + \left(m_{\mathrm{T}}(\mathbf{p}_{\mathrm{T}}^{(\tau_2)}, \boldsymbol{p}_{\mathrm{T,PUPPI}}^{(\mathrm{miss})})\right)^2 + \left(m_{\mathrm{T}}(\mathbf{p}_{\mathrm{T}}^{(\tau_1)}, \mathbf{p}_{\mathrm{T}}^{(\tau_2)})\right)^2} \,. \tag{9.29}$$

The signal tau pair (τ_1, τ_2) varies between the particular final state, i.e. $\mathrm{e}\mu$, $\mathrm{e}\tau_{\mathrm{h}}$, $\mu\tau_{\mathrm{h}}$ and $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$.

For the search of additional Higgs bosons, the likelihood of Equation (9.9) takes the following form

$$\mathcal{L}(\boldsymbol{\mu}_{\boldsymbol{s}_{\mathrm{BSM}}}; \boldsymbol{\theta}) = \prod_j \mathcal{C}(\theta_j) \prod_i \mathcal{P}\left[d_i, \sum_{s_{\mathrm{BSM}}} \mu_{s_{\mathrm{BSM}}} \cdot S_i^{(s_{\mathrm{BSM}})}(\theta_j) + \sum_{s_{\mathrm{SM}}} \cdot S_i^{(s_{\mathrm{SM}})}(\theta_j) + \sum_b B_i^{(b)}(\theta_j)\right] \,. \tag{9.30}$$

The SM Higgs boson processes ($S_i^{(s_{\mathrm{SM}})}$), associated to h of the MSSM, are treated as background. Corresponding signal processes for $S_i^{(s_{\mathrm{SM}})}$ comprise ggH and qqH. The production of a heavier neutral Higgs boson ($\phi = \mathrm{H}, \mathrm{A}$) are dominated by the production mechanisms $\mathrm{gg}\phi$ and $\mathrm{bb}\phi$ as discussed in Section 5.2 (see also Figure 5.5). Their contribution ($S_i^{(s_{\mathrm{BSM}})}$) to the likelihood in Equation (9.30) is scaled with two POIs, $\mu_{\mathrm{gg}\phi}$ and $\mu_{\mathrm{bb}\phi}$, respectively. Many nuisance parameters given by systematic uncertainties are in common with the SM $\mathrm{H} \to \tau\tau$ discussed in Section 9.2.2, where of course additional uncertainties related to the BSM signals $\mathrm{gg}\phi$ and $\mathrm{bb}\phi$ need to be included [10].

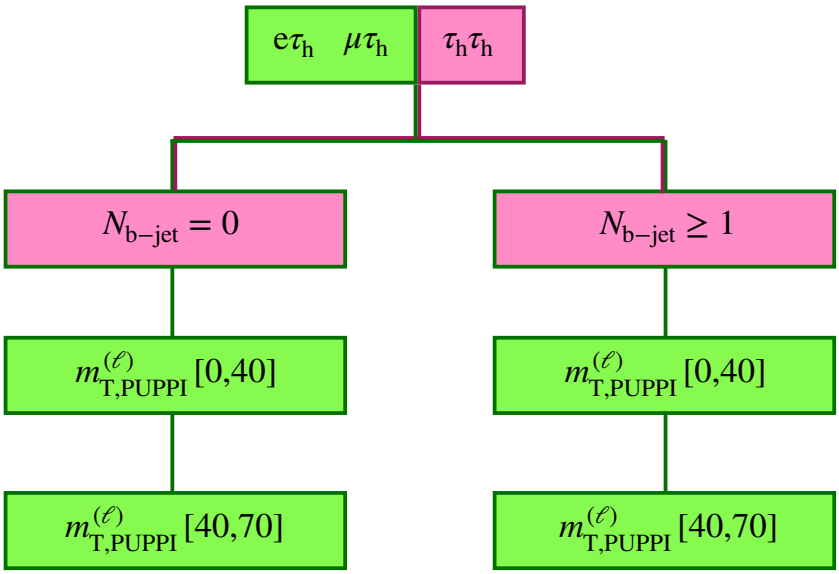


Figure 9.11: This figure represents the cut-based categorization used in the search for additional neutral Higgs bosons ($\phi = H, A$) in the context of the MSSM. For the $\tau_h\tau_h$ channel, there are two categories defined, applying different selections on $N_{\text{b-jet}}$. For the semi-leptonic channels, each $N_{\text{b-jet}}$ category is further partitioned in two sub-categories in terms of selections on $m_{T,PUPPI}^{(\ell)}$. The ranges given for $m_{T,PUPPI}^{(\ell)}$ are in units of GeV. Thus, four categories exist for each semi-leptonic channel.

Concerning the modeling of jet $\rightarrow \tau_h$, the FF method, described in Section 8.2, is used. The only difference is that the closure correction as a function of the sub-leading τ_h transverse momentum in the $\tau_h\tau_h$ channel is calculated inclusively in N_{jets} (see Figure 8.25). As a consequence there is also just a single nuisance parameter related to this correction. Other than that, the F_F-related uncertainty model is identical as discussed in Section 9.2.2 (see also Table 9.4).

Figure 9.12 shows m_T^{tot} distributions of categories with significant contributions of jet $\rightarrow \tau_h$ (shown in green). Since no excess compatible with a BSM signal is observed, model-independent upper limits on the yield are derived in a next step. This kind of result is obtained from the CL_s method [233, 234, 235]. The method uses the profile likelihood (q_μ) as test statistic [236, 237]

$$q_\mu = -2 \ln \left(\frac{\mathcal{L}(\mu, \hat{\boldsymbol{\theta}})}{\mathcal{L}(\hat{\mu}, \hat{\boldsymbol{\theta}}_{\hat{\mu}})} \right) , \tag{9.31}$$

where $\mathcal{L}(\mu, \boldsymbol{\theta})$ is given by Equation (9.30). Values with a hat ($\hat{}$) indicate the maximum likelihood estimates of the corresponding parameters from the fit to the data. The index μ in q_μ indicates that the fit to the data has been carried out for a fixed value of μ. The denominator in Equation (9.31) represents a global maximum of $\mathcal{L}(\mu, \boldsymbol{\theta})$, meaning that the POI and the nuisance parameters are fit simultaneously to the data. The test statistic in Equation (9.31) follows a certain sample distribution, $f(q_\mu)$, which differs between different hypotheses. One assumption is that there is no new signal, i.e. the signal strength parameter is zero ($\mu = 0$), which is denoted as $f(q_\mu|0)$. The other hypothesis is that there is a signal contributing to the observed data counts, i.e. $\mu > 0$. The corresponding probability density is denoted as $f(q_\mu|\mu)$. The value of CL_s is then given

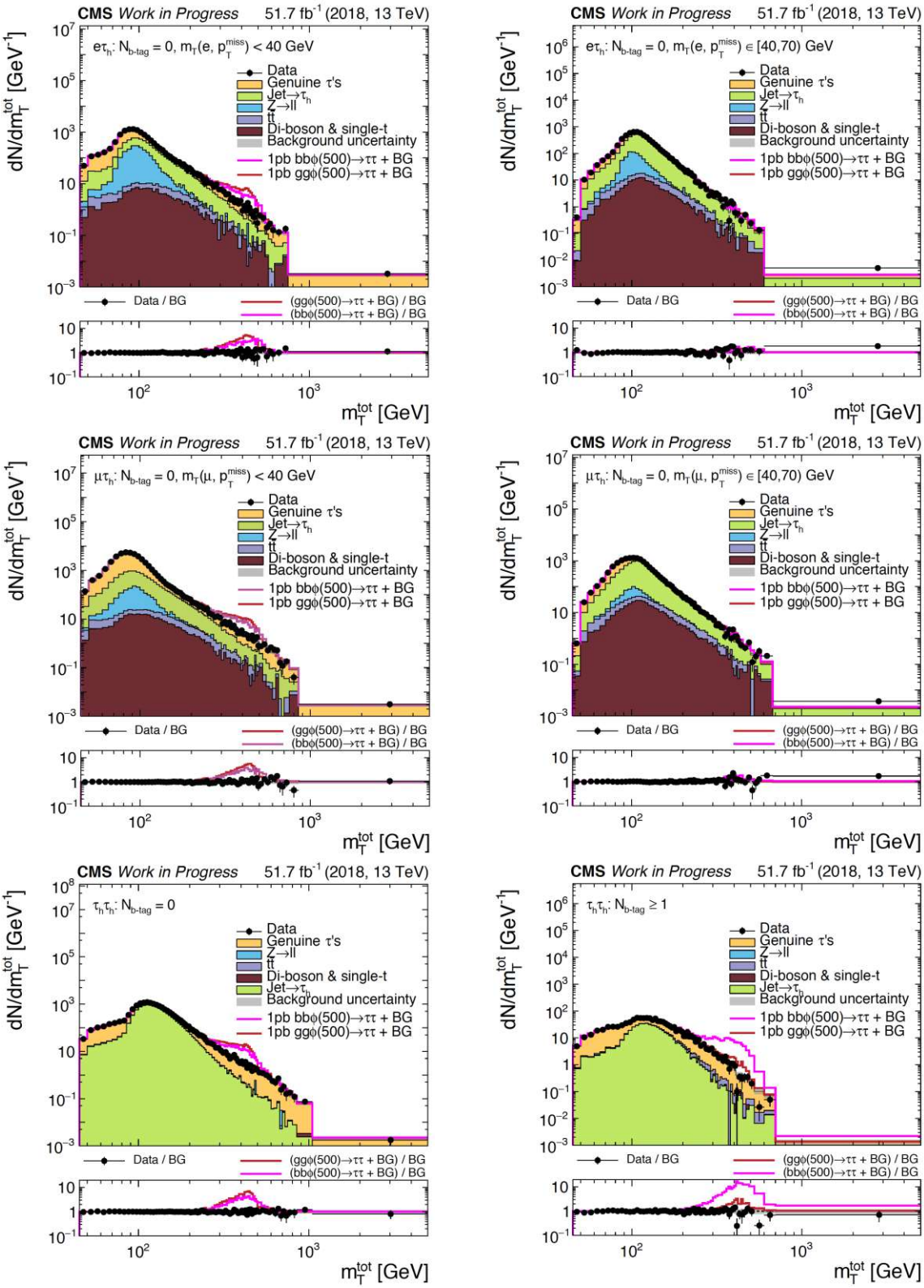


Figure 9.12: Distributions of $m_{\mathrm{T}}^{\mathrm{tot}}$ using 2018 data are shown for different final states and categories of the MSSM $\phi \to \tau\tau$ analysis. Top, middle and bottom row correspond to $e\tau_{\mathrm{h}}$, $\mu\tau_{\mathrm{h}}$ and $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel, respectively. For the semi-leptonic channels, only distributions with $N_{\text{b-jet}} = 0$ are shown, where on the left the category with $m_{\mathrm{T,PUPPI}} < 40\,\mathrm{GeV}$ and on the right with $m_{\mathrm{T,PUPPI}} \in [40, 70]\,\mathrm{GeV}$ is displayed. For the fully hadronic channel, $N_{\text{b-jet}} = 0$ is shown on the bottom left and $N_{\text{b-jet}} \geq 1$ on the bottom right. The signal processes ($\mathrm{bb}\phi$ and $\mathrm{gg}\phi$) are shown for $m_\phi = 500\,\mathrm{GeV}$, normalized to $\sigma \times \mathcal{BR}(\phi \to \tau\tau) = 1\,\mathrm{pb}$, and are added on top of the sum of background expectations. Gray uncertainty bands represent the uncertainties after the maximum likelihood fit of $\mu_{\mathrm{gg}\phi}$ and $\mu_{\mathrm{bb}\phi}$. The figures are taken from [10].

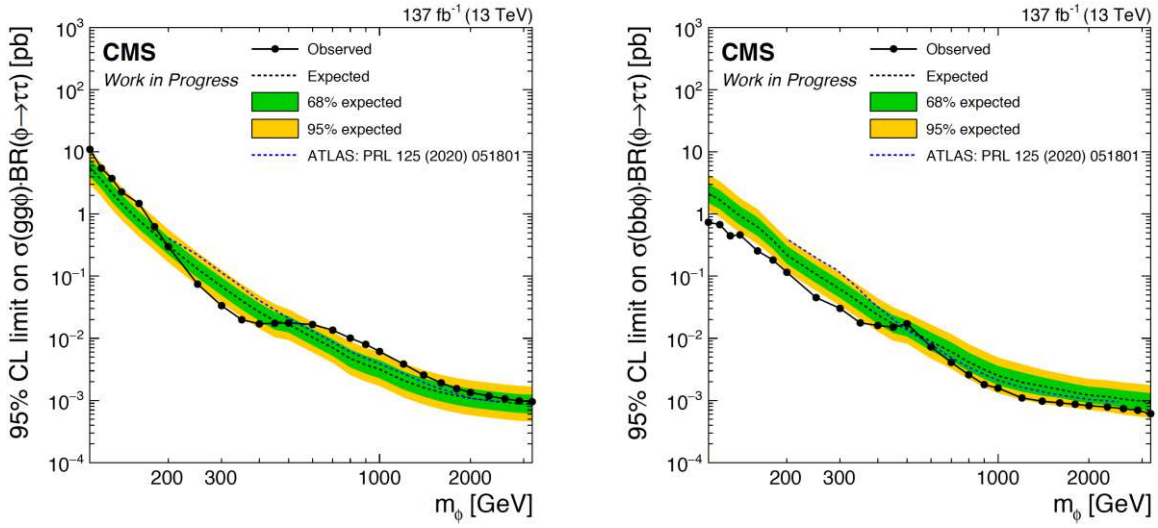


Figure 9.13: Upper limits at 95% confidence level on $\sigma \times \mathcal{BR}$ as a function of m_ϕ using data from 2016–2018 are shown. On the left the ggϕ, and on the right the bbϕ production process is shown. Expected median values in the absence of signal are shown as dashed lines, where this analysis is shown in black and the limits obtained by ATLAS are shown in blue [186]. Green and yellow bands show the central 68% and 95% expected quantiles for the upper limit (black dashed). Black dots show the observed limits, which are connected by solid black lines. The figures are taken from [10].

as a ratio of p-values of the two probability densities

$$\mathrm{CL_s} = \frac{p_\mu}{p_0} = \frac{\int_{q_\mathrm{obs}}^{\infty} f(q_\mu|\mu)\mathrm{d}\mu}{\int_{q_\mathrm{obs}}^{\infty} f(q_\mu|0)\mathrm{d}\mu} \,, \tag{9.32}$$

where q_obs denotes the value of the test statistic observed in data. Values that fall below 5%

$$\mathrm{CL_s} \leq \alpha = 0.05 \tag{9.33}$$

exclude the signal hypothesis with a confidence level of 95% $(1-\alpha)$.

For the model-independent result, only one signal strength parameter is left free floating, while the other is fit to its most compatible value with data, i.e. μ in Equation (9.32) is set either to $\mu_{\mathrm{gg}\phi}$ or $\mu_{\mathrm{bb}\phi}$. Results of upper limits for the respective cross-section times branching ratio ($\sigma \times \mathcal{BR}$) for gg$\phi$ and bbϕ are shown in Figure 9.13. Green and yellow uncertainty bands represent the central 68% and 95% expected quantiles for the upper limit[5]. The expected limits in Figure 9.13 are compared to published results of the same search by the ATLAS Collaboration [186]. Both analyses show compatible results, whereby the limits presented here are calculated over a larger range of heavy Higgs boson masses (m_ϕ). In addition, the expected limits for $m_\phi \in [110, 700]\,\mathrm{GeV}$ are stronger in the presented analysis. The data are also interpreted in a model-dependent way within the $\mathrm{M_h^{125}}$ benchmark scenario [59] mentioned in Section 2.3.1. Results of this interpretation can be found directly in [10] or in the recent publication [9] and are not discussed further here.

In the search of additional Higgs bosons in the context of the NMSSM [8, 11], many similarities to the SM $\mathrm{H} \to \tau\tau$ and MSSM $\mathrm{H} \to \tau\tau$ analysis exist. The signal process is shown in Figure 5.6 and involves a pair of tau leptons and a pair of b quarks in the final

[5] The expected limit is obtained by using generated data, referred to as *Asimov* data [238]. The Asimov data yield is defined as the expected yield corresponding to the sum all background processes.

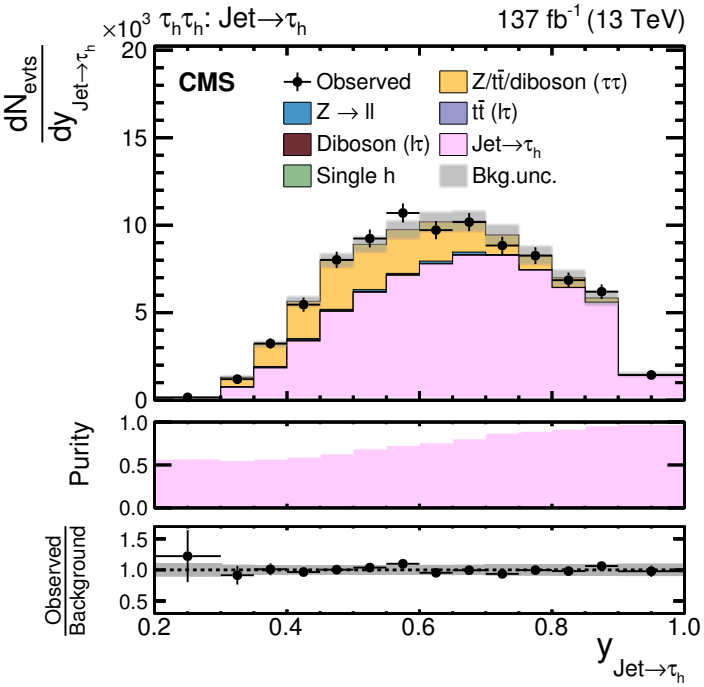


Figure 9.14: An event distribution as a function of the NN output score is displayed for the $\tau_h\tau_h$ channel and data from 2016–2018. All events contributing to this histogram are assigned by the NN to the jet $\to \tau_h$ class. The masses of the signal Higgs bosons (see also Equation (5.9)) are $m_H = 500\,\text{GeV}$ and $m_{h_S} = 110\,\text{GeV}$. Hence, all events shown here are classified by the NN trained with the signal class including this specific mass point. The gray uncertainty band represents all statistical and systematic uncertainties after the maximum likelihood fit to the signal plus background hypothesis for $m_H = 500\,\text{GeV}$ and $m_{h_S} = 110\,\text{GeV}$. The figure is taken from [8].

state. Three channels of possible tau pairs are analyzed – $e\tau_h$, $\mu\tau_h$ and $\tau_h\tau_h$. The NMSSM analysis employs an event categorization using NNs like in case of the SM $H \to \tau\tau$ analysis discussed in Section 9.2.1. One of the target classes is the jet $\to \tau_h$ class. The jet $\to \tau_h$ background is modeled by the FF method as described in Section 8.2. A notable F_F-related change is the different definition of F_F in terms of the required D_{jet} conditions. For the NMSSM analysis, the SR is defined for τ_h's that pass the `medium` threshold of D_{jet}. The AR is defined as all events that fail this `medium` threshold but pass the loosest threshold of the D_{jet} discriminant (`vvvloose`). Hence, the basic equation of the F_F derivation (see Equation (8.4)) is given by

$$F_F = \frac{N^{(\texttt{medium})}}{N^{(\texttt{vvvloose} \,\wedge\, \neg\texttt{medium})}} \,. \tag{9.34}$$

The signal classes are difficult to define in the NMSSM analysis because several signal hypotheses are tested. The mass ranges according to Equation (5.10) are split in 420 mass points. It is unfeasible to define 420 signal classes and have a sensible NN training with the amount of available simulated samples as explained in [11]. On the other hand, training 420 NNs with one signal class each is computationally not achievable. A trade-off is made and 68 signal categories are defined, where certain mass points are grouped together. Each signal category is used as signal class for a separate NN training, i.e. there are 68 NNs trained and each is specialized to separate background classes from one of the 68 signal classes.

An example of the NN output distribution is shown in Figure 9.14. A high purity of jet $\to \tau_h$ events (shown in pink), accumulated in bins with a large NN output score,

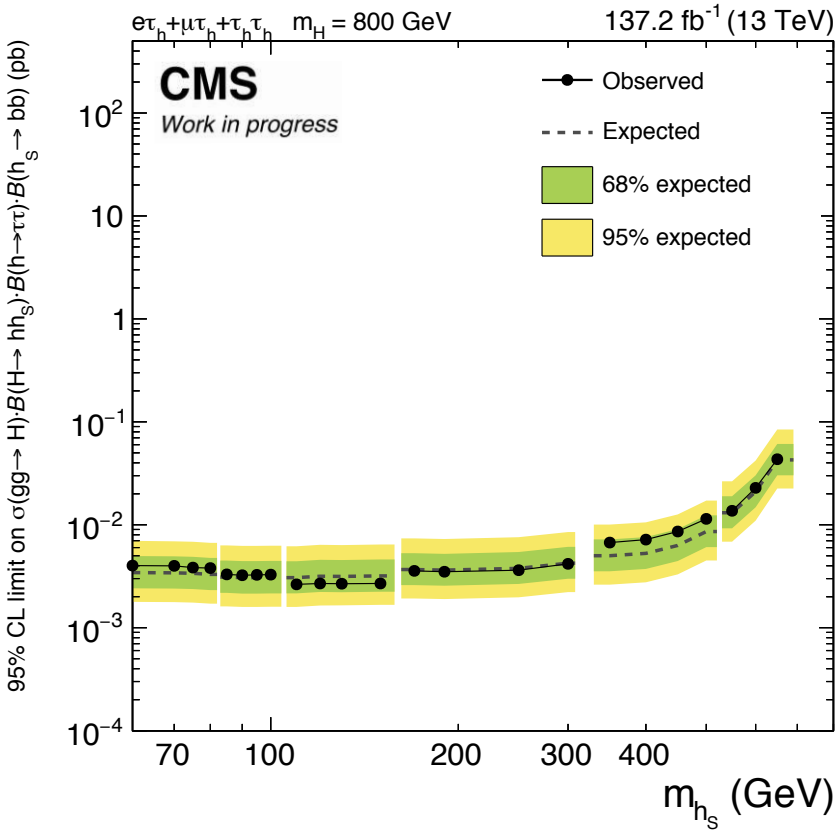


Figure 9.15: An upper limit at 95% confidence level on $\sigma \times \mathcal{BR}$ as a function of $m_{\mathrm{h_S}}$, using data from 2016–2018, is shown. The mass of the heavy Higgs boson is given by $m_{\mathrm{H}} = 800\,\mathrm{GeV}$. The expected limit in the absence of signal is shown as dashed line, whereas green and yellow bands show the central 68% and 95% expected quantiles for the upper limit, respectively. Black dots show the observed limits which are connected by solid black lines. The $m_{\mathrm{h_S}}$ region is interrupted five times, resulting in six partial limits. This is to indicate that six different NNs are used to obtain the results, which are trained for a specific signal class in terms of $(m_{\mathrm{H}}, m_{\mathrm{h_S}})$. The figure is taken from [11].

is achieved as can be seen from the middle panel of Figure 9.14. Furthermore, a good data modeling is observed. The specific NN used for classification includes the mass point $(m_{\mathrm{H}}, m_{\mathrm{h_S}}) = (500, 110)$ GeV as indicated in the caption of the figure. Following the same $\mathrm{CL_s}$ approach as for the MSSM analysis (see Equation (9.32)), upper limits on the cross section times branching ratio can be set for different mass hypotheses. An example for upper limits for the mass hypothesis $m_{\mathrm{H}} = 800\,\mathrm{GeV}$ is show in Figure 9.15.

9.3.2 Leptoquark Search

The search for third generation LQs is introduced in Section 5.3 and looks into non-resonant and resonant production of LQs (see also Figure 5.7). The event categorization applied to look for the LQ-signal, is shown in Figure 9.16. For signal inference of non-resonant production, the following discriminating variable is used

$$\begin{aligned} \chi &= \exp\left(|\Delta\eta|\right) \\ \Delta\eta &= \eta^{(\tau_1)} - \eta^{(\tau_2)}\ , \end{aligned} \tag{9.35}$$

where (τ_1, τ_2) refers to the signal tau pair. The angular variable (χ) has been extensively used to probe the substructure of quarks in di-jet analyses, e.g. in [239, 240, 241, 242, 243]. For certain BSM, including LQs, models, isotropic event distributions are expected leading

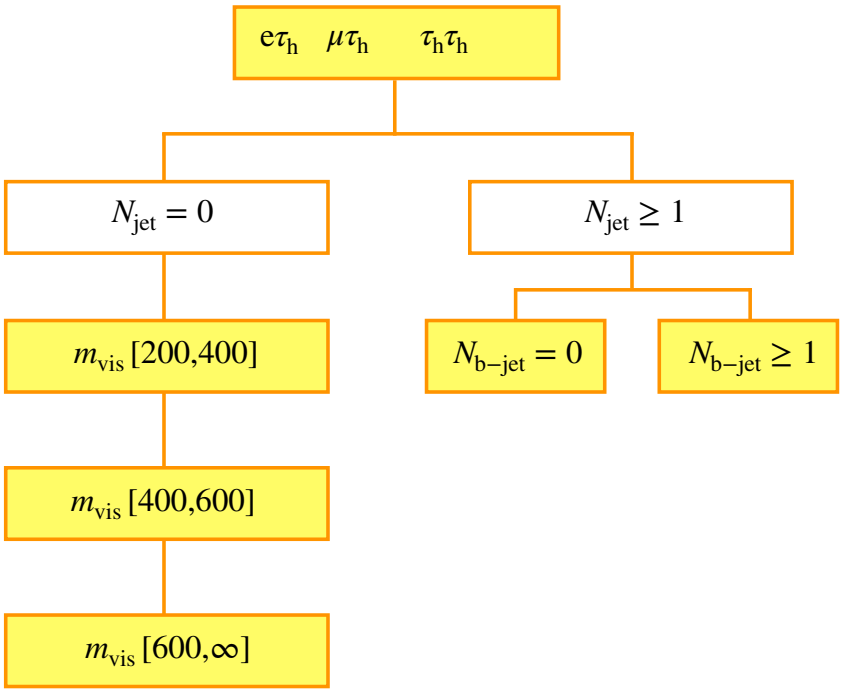


Figure 9.16: This figure represents the cut-based categorization used in the search for third generation LQs in the channels involving at least one τ_h – $e\tau_h$, $\mu\tau_h$ and $\tau_h\tau_h$. To look for non-resonant LQ production (see Figure 5.7c), events are selected with $N_{jets} = 0$ and in different m_{vis} categories. The values of the m_{vis} ranges are given in GeV. Resonant signal production (see Figures 5.7a and 5.7b), is expected to populate the categories with $N_{jets} \geq 1$, whereby a further splitting according to $N_{b\text{-jet}}$ is used. In total, five (filled boxes) signal categories are defined.

to an accumulation of signal events at low values of χ compared to known SM processes. Hence, this variable is well suited to look for LQs in di-tau events. For the resonant production, the following variable is used [244]

$$S_T^{MET} = p_T^{(\tau_1)} + p_T^{(\tau_2)} + p_T^{(jet_1)} + p_T^{(miss)} \quad , \tag{9.36}$$

being the scalar sum of transverse momenta of both signal tau leptons ($p_T^{(\tau_1)}, p_T^{(\tau_2)}$), the transverse momentum of the leading jet ($p_T^{(jet_1)}$) and MET[6].

The search in the channels involving at least one τ_h, i.e. $e\tau_h$, $\mu\tau_h$ and $\tau_h\tau_h$, follows closely the selection and tau pair formation as described in Chapter 6. As for the presented Higgs boson analyses, also this di-tau final states has contributions of jet $\rightarrow \tau_h$ processes entering the search region. This type of background is estimated with the FF method presented in Section 8.2. However, several adaptations compared to the Higgs boson analyses are applied to the FF method to better align it with the LQ search. An obvious adjustment involves the FF definition (see Equation (8.4)) which is given by

$$F_F = \frac{N^{(\texttt{medium})}}{N^{(\texttt{vloose} \wedge \neg\texttt{medium})}} \quad , \tag{9.37}$$

which is synchronized with the definition of the SR of the LQ analysis requiring signal τ_h's to pass the `medium` threshold of D_{jet}. Furthermore, the definition of N_{jets} differs in the search for LQs, where only jets with transverse momentum above 50 GeV are considered

[6] In this analysis, PU effects are treated by the charged hadron subtraction method (see also Section 3.5.9)

(compare to Equation (6.6)). Therefore, F_F are derived using this definition of N_{jets} leading to a migration of events to lower jet multiplicities with respect to the presented Higgs boson analyses. In addition, only two N_{jets} categories are used, $N_{jets} = 0$ and $N_{jets} \geq 1$, to target better the categorization scheme shown in Figure 9.16 and profit from larger sample sizes for the derivation of the raw F_F's. For the raw $F_F^{(t\bar{t})}$, only a single inclusive N_{jets} category is used. The cut on $\Delta R^{(\ell,\tau_h)}$, used to define two categories in the derivation of $F_F^{(W+jets)}$, is set to 2 because it has been studied that this definition improves the data modeling in the low χ region where the non-resonant signal is expected. Lastly, due to the inclusion of large m_{vis} values in the search for non-resonant production (see Figure 9.16), all F_F-related correction as a function of m_{vis} are extended to cover this region too. Figure 9.17 shows distributions for both discriminating variables, χ and S_T^{MET}, for all channels with at least one τ_h and two selected signal categories. The contribution from jet $\rightarrow \tau_h$ processes are shown in pink. By using these distributions in a maximum likelihood fit, the signal strengths can be inferred and upper limits can be derived according to Equation (9.32) on the production cross section times branching ratio of different LQ signal models. This analysis is currently being prepared for publication by the CMS Collaboration. Results, covering the non-resonant signal production of vector-like LQs, however, are already publicly available in [9]

9.3.3 Light Top Squarks

In this section, region histograms like the one shown in Figure 7.2 are used in a profile likelihood approach to determine expected upper limits on the cross section of top squark pair production. The branching ratio of top squarks to the studied four-body final state ($\tilde{t}_1 \rightarrow \tilde{\chi}_1^0 b f \bar{f}'$) is assumed to be 100%. Region histograms are filled with different signal mass points. The goal is to see if a gain in sensitivity can be achieved by employing the finer region splitting shown in Figure 7.1 compared to the published results [206]. The publication uses only three $m_T^{(\ell)}$ regions, merging region "c" and "d" together. In addition the "VH" SR is merged with the respective CR. Thus, in the publication in total 54 regions are defined. Before discussing the results, a brief overview of the considered systematic uncertainties is given.

Systematic Uncertainties

The normalization of the main background processes, being W+jets and $t\bar{t}$ (see Figure 7.2), in each SR is obtained from the corresponding CR in the following way. Inside each CR the sum of W+jets and $t\bar{t}$ is scaled to match the data yield after all contributions other than W+jets and $t\bar{t}$, expected from MC simulation, are subtracted from observed data. The obtained scale factors are then applied to the MC yields of W+jets and $t\bar{t}$ in each corresponding SR, e.g. the scale factor from CR1aX is applied to all SR1aX(VL-VH) (see Figure 7.1). The combined yield of W+jets and $t\bar{t}$ processes is scaled with values in the range $[0.9, 1.2]$. Uncertainties of the measured scale factors are included in the uncertainty model and are treated as nuisance parameters.

Systematic uncertainties due to the correction of ISR-related quantities (see Section 7.4) are included for W+jets and $t\bar{t}$ processes. Their respective magnitude is in the order of 15% for W+jets events and 5 to 10% for $t\bar{t}$ events. Furthermore, jet energy and resolution corrections (see Section 6.5.3) uncertainties are treated with separate nuisance parameters. The estimated uncertainty for background samples due to jet energy and resolution corrections is in the range of 1–2%. Lepton and b-tag scale factor uncertainties are also included to the uncertainty model, both being in the order

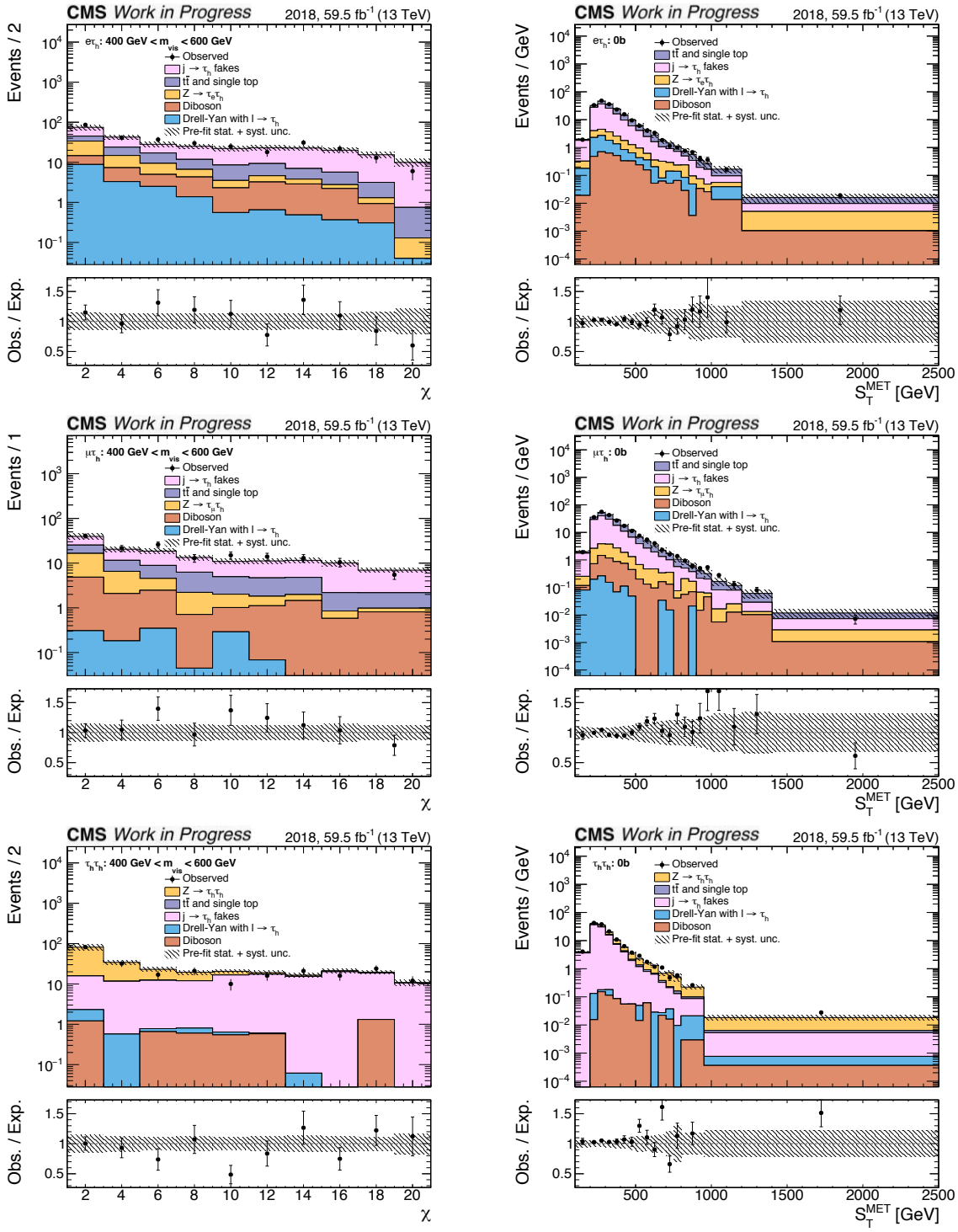


Figure 9.17: Shown are distributions of χ and $S_{\mathrm{T}}^{\mathrm{MET}}$. From top to bottom, the $\mathrm{e}\tau_{\mathrm{h}}$, $\mu\tau_{\mathrm{h}}$ and $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel using 2018 data are shown. On the left, the category with $m_{\mathrm{vis}} \in [400, 600]\,\mathrm{GeV}$ is shown. The right side depicts distributions of the category with $N_{\text{b-jet}} = 0$. The dashed uncertainty bands represent the statistical background and systematic uncertainties before maximum likelihood fit. In the lower panel, the ratio of data to expected yield is shown.

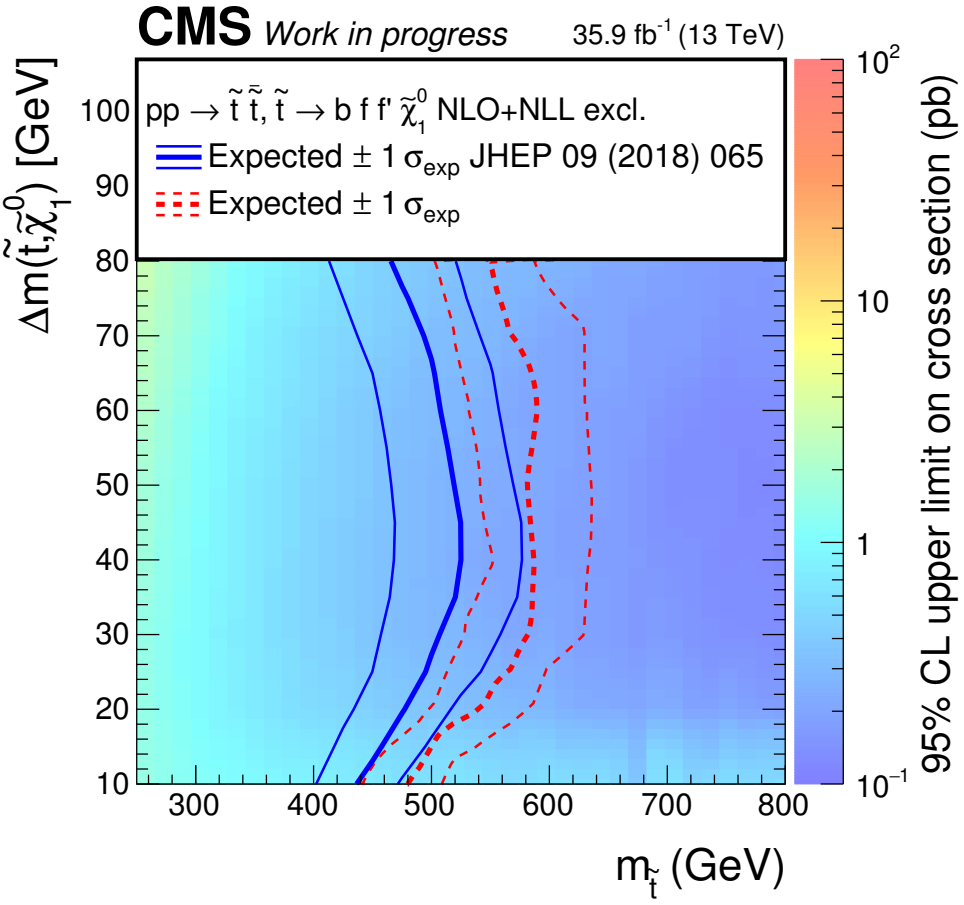


Figure 9.18: Expected exclusion limits at 95% confidence level are shown for stops in four-body decays. The BR to the four-body final state is set to be 100%. The color map shows the expected cross section limit. Expected limits are compared to published results in [206] which use a more granular region splitting than presented in this thesis.

of 1%. Lastly, limited sample sizes of simulated background contributions are the source of nuisance parameters obtained by the Barlow-Beeston approach [231].

Expected Stop Mass Limits

Upper limits on the top squark production cross section are obtained using the CL_s method and using asymptotic formulas [238]. In the limit calculation, observed data are replaced according to the expectation from simulation. The expected upper limit on the production cross section at 95% confidence level is shown as a colored map in Figure 9.18. Central ticker lines in Figure 9.18 represent excluded masses at 95% confidence level under the background-only hypothesis. Thinner lines indicate the region where 68% of the distribution of limits is contained. In blue, the expected limits are shown for the published analysis and in red for the presented analysis in this thesis. The further splitting in $m_{\mathrm{T}}^{(\ell)}$ and p_{T} regions translates to stronger exclusion limits. For all mass splittings, it is expected that stop masses with 20 to 50 GeV higher values can be excluded. This result is also encouraging in the light of extending the analysis to full Run2 collision data.

Data-Driven Background Estimation

Ultimately, the expected limit shown in Figure 9.18 should be recalculated using the FF method presented in Section 8.3. Since the F_{F} derived in this section use a DR quite

different in terms of the applied trigger logic (compare Equation (7.2) with Table 8.2), it is important to find a way to quantify potential biases introduced. One possibility is to measure closure corrections in a validation region enriched in misidentified leptons closer to the SR. A good candidate of such a region is given by inverting the anti-QCD cut in Table 7.2. Obtained non-closures can then be used as systematic uncertainties in the application of F_{F} in the analysis. The F_{F} application follows then Equation (8.57)

$$N_{\mathrm{pred}}^{(\text{SR-like})} = \sum_{\epsilon \in \mathrm{AR}} F_{\mathrm{F}}(\epsilon) - \sum_{i \in \mathrm{prompt}} F_{\mathrm{F}}(i) \ , \tag{9.38}$$

i.e. subtract extrapolated contributions from prompt leptons to the SR from those obtained by extrapolating the observed data.

Chapter 10

Conclusion and Outlook

The discovery of the Higgs boson a decade ago marks the beginning of a rich physics program followed by CMS and other experiments. During Run2, in the years 2016–2018, an unprecedented amount of proton-proton collision data, corresponding to an integrated luminosity of $137\,\text{fb}^{-1}$, has been collected by the CMS experiment. Analyzing these data allowed discovering rare decays of the Higgs boson like its decay into a pair of tau leptons, $\text{H} \to \tau\tau$. The $\text{H} \to \tau\tau$ decay is an interesting process because it allows testing the Yukawa-type coupling of the Higgs boson to fermions. As demonstrated in this thesis, studying the $\text{H} \to \tau\tau$ decay has moved from discovery to precision measurements during Run2. Precision measurements are presented in up to twelve kinematic regions defined in the context of the STXS scheme. The presented results are entering combinations of other Higgs boson decay channels as well as combinations between CMS and ATLAS results. These combined measurements will provide the ultimate sensitivity on the measurement of the Higgs boson couplings to fermions and bosons after Run2.

Special ingredients to the presented SM $\text{H} \to \tau\tau$ analysis are the use of data-driven background estimation methods and the implementation of modern machine learning tools for event classification. More specifically, up to 95% of all background contributions are estimated from data, picking up the actual beam and detector run conditions which otherwise need to be corrected for in simulated collision events. As such, the FF method presented in this thesis plays a key role. It estimates contributions from quark- or gluon-initiated jets that are misidentified as hadronically decaying tau leptons, jet $\to \tau_\text{h}$. Hence, the FF method is used in the $\text{e}\tau_\text{h}$, $\mu\tau_\text{h}$ and $\tau_\text{h}\tau_\text{h}$ final states of the $\text{H} \to \tau\tau$ decay. It is worth highlighting that the $\tau_\text{h}\tau_\text{h}$ channel is the most sensitive channel of the $\text{H} \to \tau\tau$ analysis, with the dominant background contribution coming from jet $\to \tau_\text{h}$ events.

Using data from the AR as well as τ-embedded events, improved the classification performance of the NN compared to previous versions of the SM $\text{H} \to \tau\tau$ analysis. To gain trust in the NN-based event classification, a thorough validation based on more than one thousand saturated goodness-of-fit tests has been carried out [7, 12]. However, it remains an open field of study to what extent machine learning based analysis techniques affect the reliability of statistical inference. As such, one interesting aspect is the inclusion of systematic uncertainties in the training of NNs [245], a topic relevant for measurements that are already dominated by systematic uncertainties and will thus not benefit from acquiring more collision data. An example of a measurement which is dominated by systematic uncertainties is the ggH signal strength modifier of the SM $\text{H} \to \tau\tau$ analysis, for which a value of

$$\mu_\text{ggH} = 0.67^{+0.08}_{-0.08}(\text{stat.})^{+0.14}_{-0.14}(\text{syst.})^{+0.10}_{-0.07}(\text{theo.})^{+0.05}_{-0.05}(\text{bbb})$$

is measured. The measurement of the signal strength modifier of Higgs boson production

via vector boson fusion, however, is still dominated by statistical uncertainties. For the qqH signal strength modifier of the SM $H \to \tau\tau$ analysis, a value of

$$\mu_{\text{qqH}} = 0.81^{+0.14}_{-0.13}(\text{stat.})^{+0.06}_{-0.06}(\text{syst.})^{+0.05}_{-0.05}(\text{theo.})^{+0.06}_{-0.06}(\text{bbb})$$

is measured. For the inclusive signal strength modifier, a value of

$$\mu_{\text{incl}} = 0.82^{+0.06}_{-0.06}(\text{stat.})^{+0.06}_{-0.06}(\text{syst.})^{+0.05}_{-0.04}(\text{theo.})^{+0.03}_{-0.03}(\text{bbb})$$

is obtained. A previous analysis published by the CMS Collaboration [174] measures

$$\mu_{\text{incl}} = 1.09^{+0.15}_{-0.15}(\text{stat.})^{+0.16}_{-0.15}(\text{syst.})^{+0.10}_{-0.08}(\text{theo.})^{+0.13}_{-0.12}(\text{bbb}) \ .$$

A clear reduction of uncertainties across all different sources can be observed in the presented analysis which uses collision data from 2016 to 2018, compared to this previous result. A major step in reducing systematic uncertainties consists in the transition to data-driven background estimation techniques presented in this thesis, i.e. the τ-embedding technique and the FF method, which are not used in [174]. While systematic uncertainties related to contributions from jet $\to \tau_{\text{h}}$ events range up to 15–20% in case of the analysis [174], it is shown in this thesis that the systematic variations of different FF components reach at most 10%.

In this thesis not only the application of the FF method to the SM $H \to \tau\tau$ analysis is showcased, but also its usage in the search of new BSM particles. Adaptations of the method are needed to align it with the respective analysis strategies in the searches of extra heavy Higgs bosons or third generation LQs and are discussed in this thesis. These alignments include the definition of the F_{F} itself, choosing the same identification condition as used for the SR of the respective analysis for the numerator of the F_{F}. Further adaptations concern the range of variables and definition of different categories that best match the implemented event categorization. Checking that contributions from misidentified quark- or gluon-initiated jets are well modeled in distributions of the variable used for the final statistical inference, is an important step in validating the FF method. The working principles behind the FF method are general and can be adopted also to estimate contributions from quark- or gluon-initiated jets that are misidentified as muons or electrons, which is also illustrated in the presented search for light top squarks. The development of the FF method for the search of light top squarks is, however, still in progress and certainly needs to be studied and tested in more detail.

Advances of identification algorithms, such as for example the DeepTau classifier, decrease the probability of quark- or gluon-initiated jets being misidentified as τ_{h} or light leptons. This implies however, that extremely large amounts of simulated events would be needed to precisely describe such contributions. Detector simulations alone consume 40% of all computation resources of CMS [246]. Using the data-driven FF method circumvents the dependence on a large number of simulated events and benefits from the substantial amount of data collected during Run2. With the upcoming data-taking periods of the LHC, data-driven methods can benefit even more from the amount of recorded collision data, which is expected to reach $300\,\text{fb}^{-1}$ by the end of 2030 and $3000\,\text{fb}^{-1}$ by the end of 2036 [247].

Challenges of the FF method lie in correctly assessing systematic uncertainties because individual fake factors are measured in different kinematic regions, all differing from the region where they are applied. Furthermore, it is very important to study the composition of misidentified objects in each analysis because fake factors depend on the jet flavor, i.e. whether a light-quark, heavy-quark or gluon-initiated jet is misidentified. Also in this area, machine learning techniques could help to identify for example better suited regions for

the fake factor measurement, being more enriched in the desired misidentified impostor. Another example is the parametrization of the fractions used in the FF method of the di-tau analyses, which could be addressed by means of an underlying machine learning model.

Acknowledgments

I would like to thank my supervisor, Jochen Schieck, for giving me the opportunity to carry out the research project presented in this thesis and for his help in difficult stages. A big thanks goes to Martin Flechl and Markus Spanring, who gave their best to get me started in the $H \to \tau\tau$ analysis at CMS. Many people at the institute have contributed in one way or the other to this work. Special thanks go to Wolfgang Adam and Ivan Mikulec for guiding me in my research, patiently clarifying my many confusions during discussions and very carefully proofreading this thesis. To the group of careful proofreaders I would like to add and express my gratitude to Suman Chatterjee and Roger Wolf.

I am happily looking back at my research stay at KIT which was made possible through funding from the Vienna doctoral program "particles and interactions" (DKPI). Thank you, Roger Wolf and Günter Quast for inviting me to KIT. Thanks to all (PhD) students from KIT whom I met and worked with: Janek B., Sebastian W., Artur G., Stefan W., Max B., Sebastian B., Moritz S., Simon J. and Oliver K. In particular, I could not have done all the work around the fake factor method without the help from Janek B.

Without any particular order, I would like to thank Izaak N., Koushik M., Mahmoud M. and Armin K. I appreciate the help from Roberval W. and Pieter D., who guided my in setting up the backplane correction measurement presented in this thesis. To Dennis M. R., Riccardo M. and Jan S. – thank you for your help with understanding the offline tau validation work flows.

To my office mates, Priya, Lukas and Sebastian, thank you for having always time for an interesting discussion, be it about physics, ROOT or something else. As student representative of HEPHY, I am blessed to have met many different people. I am especially grateful for the opportunity to have helped young students carrying out their master thesis at our institute. Thanks to my students, Maja (for always asking the difficult questions) and Julia, but also to other students I met – Gerhard, Tim, Tommy and Markus D. – just to name a few. There are also several people behind the scenes, who made everything around the actual work run smooth. Special thanks to the secretaries, Simone, Barbara, Nathalie and Zorica, but also to all computing administrators, who made sure that everything is up and running (almost at all time).

Let me also thank my external experts, Prof. M. Schumacher and Prof. G. Piacquadio, for accepting to read and grade this thesis.

At the end, I could not have completed this work without the moral support from my friends and family.
For the endless support of my parents and grandmother:
Danke Mama, Grossmama und Roland, ďakujem papa a Milena.
For distracting me and sharing many nice adventures with me:
Ďakujem Radko, Martin, mama Eva, Lucia, Miško, Alenka a Matúško.
Merci Oli, thanks Kevin&Cris, Hinnerk&Chrissi and Amit.
For loving me just the way I am:
Ďakujem láska moja!

Chapter 11

Acronyms

CMS Compact Muon Solenoid . . . 9
ATLAS A Toroidal LHC ApparatuS . . . 9
LHC Large Hadron Collider . . . 9
CERN European Organization for Nuclear Research . . . 9
Run2 data recorded in the years from 2015 to 2018 . . . 37
Run1 data recorded in the years from 2010 to 2012 . . . 80
DR determination region . . . 115
SR signal region . . . 115
CR control region . . . 107
AR application region . . . 115
TIB tracker inner barrel . . . 44
TOB tracker outer barrel . . . 45
TID tracker inner disk . . . 45
TEC tracker endcap . . . 45
ECAL electromagentic calorimeter . . . 45
HCAL hadron calorimeter . . . 47
HLT high-level trigger . . . 49
MC Monte Carlo . . . 50
QCD quantum chromodynamics . . . 15
QED quantum electrodynamics . . . 20
LO leading-order . . . 52
NLO next-to-leading-order . . . 52
NNLO next-to-next-to-leading-order . . . 75
PU pileup . . . 39
CTF combinatorial track finder . . . 53
PF particle-flow . . . 54
DT drift tube . . . 48
CSC cathode strip chamber . . . 48

Appendices

Appendix A

Fake Factors for the LQ Search

In this section, all F_{F}'s and their corrections are presented using the definition of Equation (9.37). They are used to estimate the jet $\rightarrow \tau_{\mathrm{h}}$ background in the search for third generation LQs presented in Section 9.3.2. In order to ease the navigation, Table A.1 summarizes the organization of the plots.

year	channel	$F_{\mathrm{F}}^{(\mathrm{W+jets})}$	$F_{\mathrm{F}}^{(\mathrm{QCD})}$	$F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$	F_{F} corrections
2016	$\mathrm{e}\tau_{\mathrm{h}}$	Fig. A.1	Fig. A.3	Fig. A.4	Fig. A.6
	$\mu\tau_{\mathrm{h}}$	Fig. A.2			Fig. A.7
	$\tau_{\mathrm{h}}\tau_{\mathrm{h}}$	–	Fig. A.5	–	Fig. A.8
2017	$\mathrm{e}\tau_{\mathrm{h}}$	Fig. A.9	Fig. A.11	Fig. A.12	Fig. A.14
	$\mu\tau_{\mathrm{h}}$	Fig. A.10			Fig. A.15
	$\tau_{\mathrm{h}}\tau_{\mathrm{h}}$	–	Fig. A.13	–	Fig. A.16
2018	$\mathrm{e}\tau_{\mathrm{h}}$	Fig. A.17	Fig. A.19	Fig. A.20	Fig. A.22
	$\mu\tau_{\mathrm{h}}$	Fig. A.18			Fig. A.23
	$\tau_{\mathrm{h}}\tau_{\mathrm{h}}$	–	Fig. A.21	–	Fig. A.24

Table A.1: This is a summary table linking all figures of raw F_{F}'s and their corrections. Fake factors are derived according to Equation (9.37).

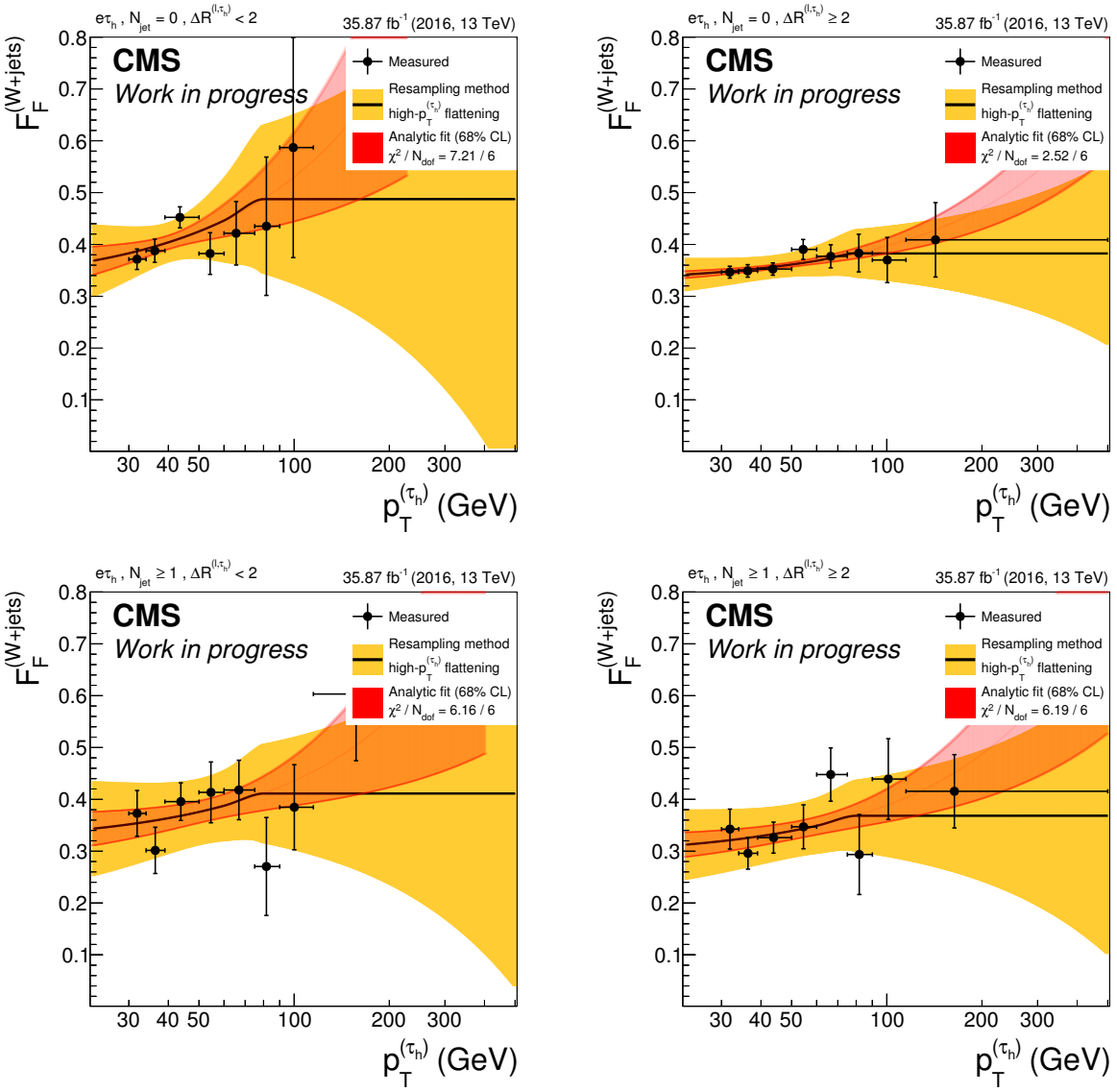


Figure A.1: The quantity $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown for the eτ_{h} channel using 2016 data. Top and bottom display the two N_{jets} categories – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} \geq 1$, respectively. On the left, the $\Delta R^{(\ell,\tau_{\mathrm{h}})} < 2$ category is shown and on the right the $\Delta R^{(\ell,\tau_{\mathrm{h}})} \geq 2$ category. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid black line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

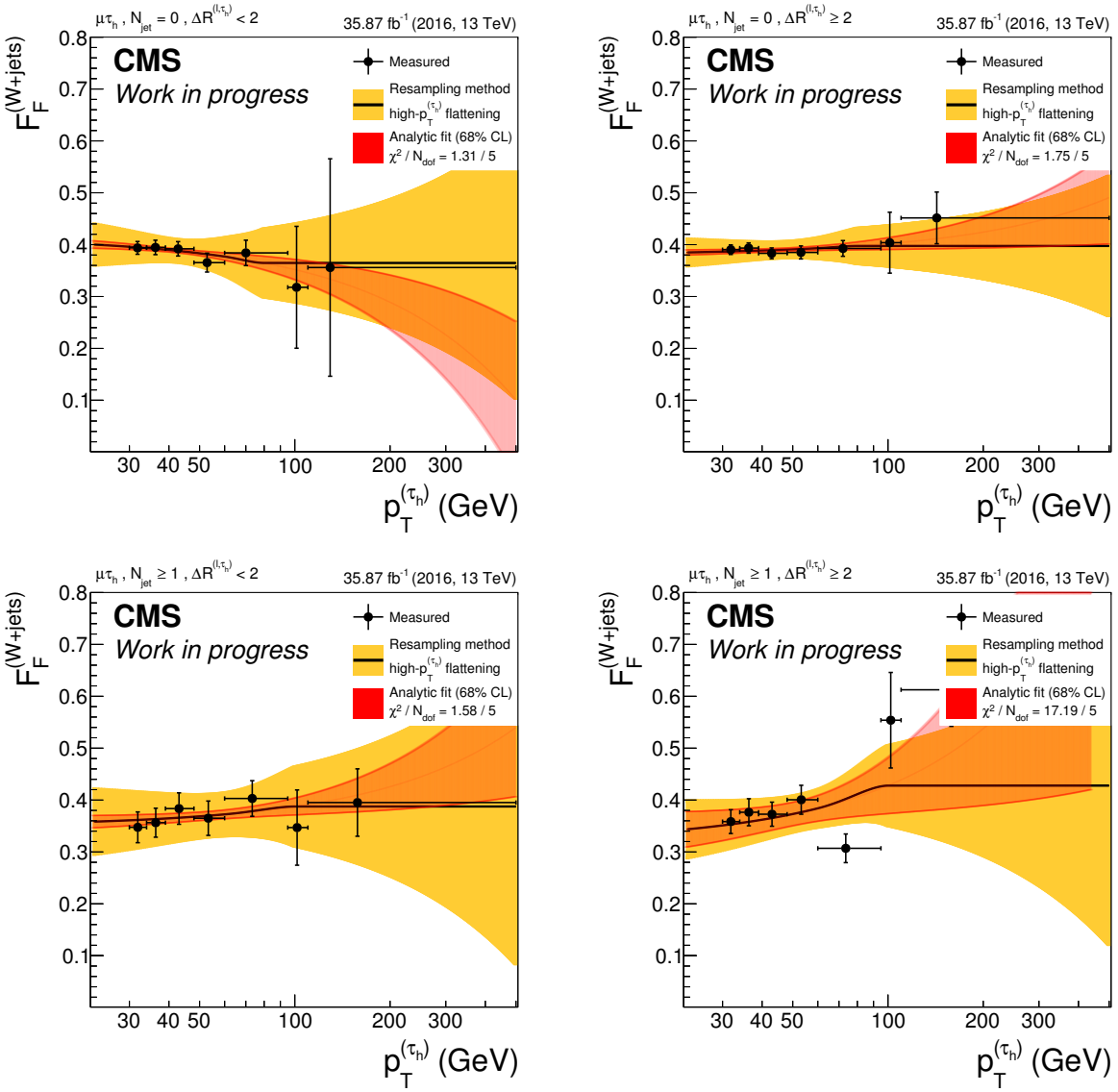


Figure A.2: The quantity $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown for the $\mu\tau_{\mathrm{h}}$ channel using 2016 data. Top and bottom display the two N_{jets} categories – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} \geq 1$, respectively. On the left, the $\Delta R^{(\ell,\tau_{\mathrm{h}})} < 2$ category is shown and on the right the $\Delta R^{(\ell,\tau_{\mathrm{h}})} \geq 2$ category. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid black line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

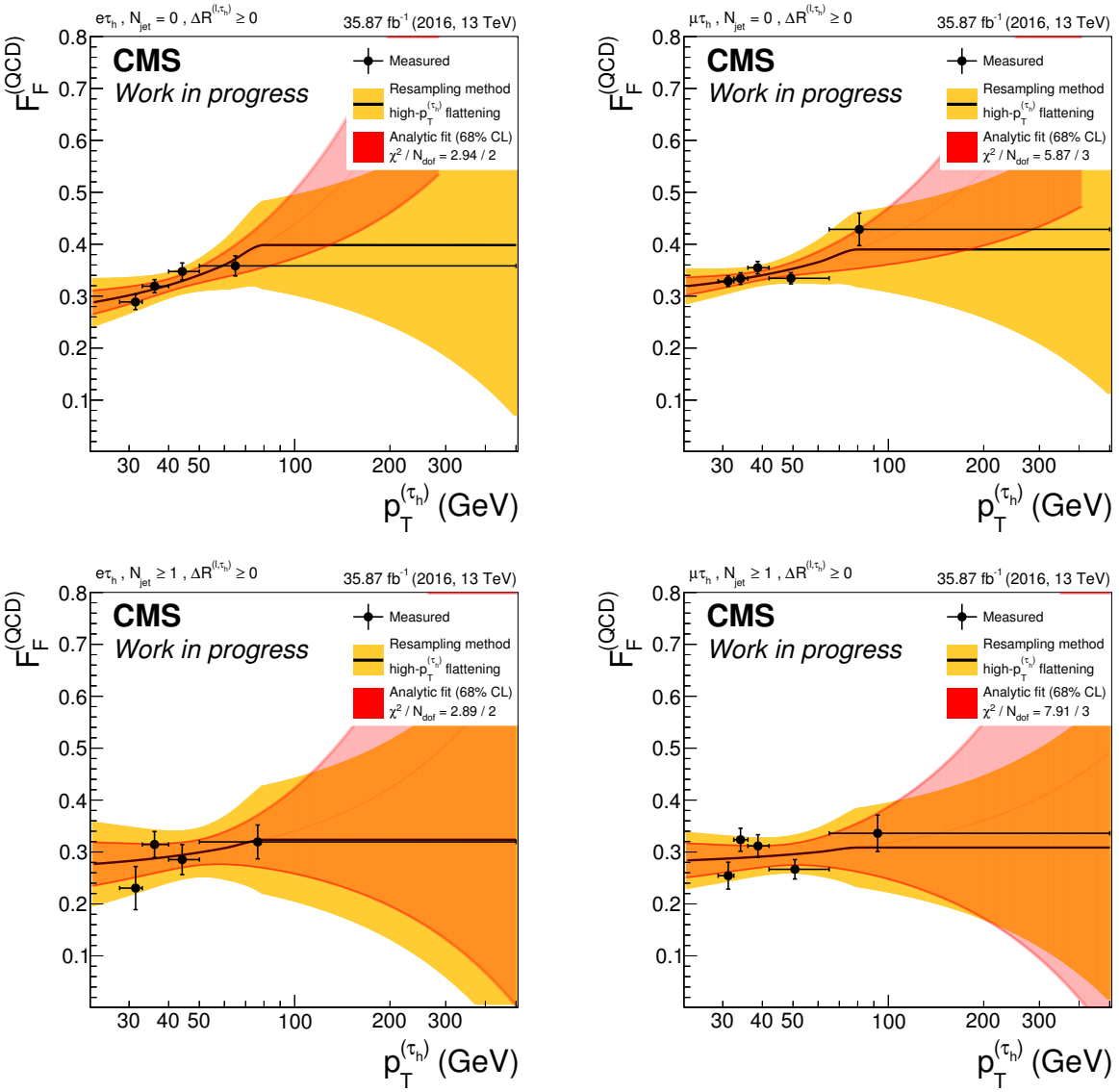


Figure A.3: The quantity $F_{\mathrm{F}}^{(\mathrm{QCD})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown using 2016 data. On the left, distributions for the eτ_{h} channel are displayed and corresponding distributions for the $\mu\tau_{\mathrm{h}}$ channel are displayed on the right. Top and bottom display the two N_{jets} categories – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} \geq 1$, respectively. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

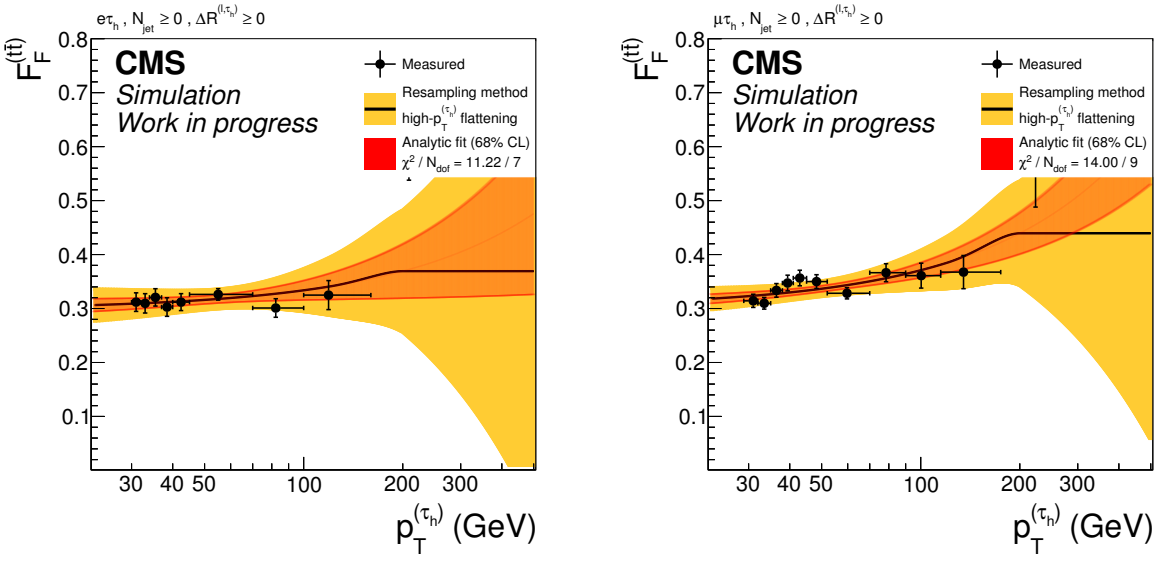


Figure A.4: The quantity $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown using simulated events from the 2016 data-taking period. On the left, the distribution for the $\mathrm{e}\tau_{\mathrm{h}}$ channel is displayed and the corresponding distribution for the $\mu\tau_{\mathrm{h}}$ channel is displayed on the right. The measurement of $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ is inclusive in N_{jets}. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

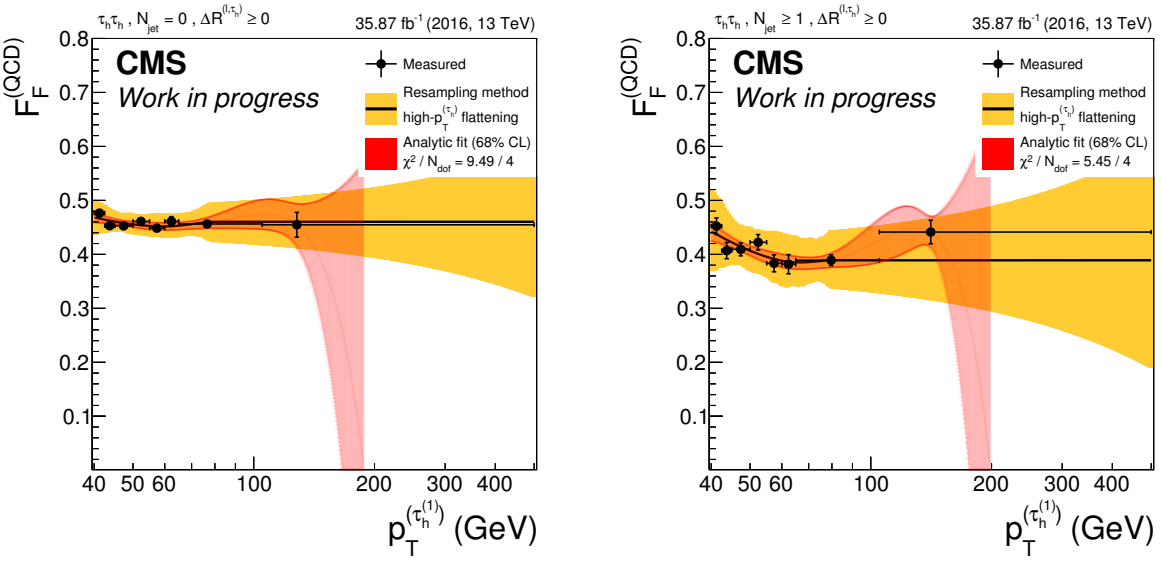


Figure A.5: The quantity $F_{\mathrm{F}}^{(\mathrm{QCD})}$ as a function of $p_{\mathrm{T}}^{(\tau_1)}$ is shown for the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel using 2016 data. Left and right display the two N_{jets} categories – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} \geq 1$, respectively. The F_{F} parametrization is shown as a solid line. It consists of a third order polynomal fit which is truncated at high $p_{\mathrm{T}}^{(\tau_1)}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid black line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

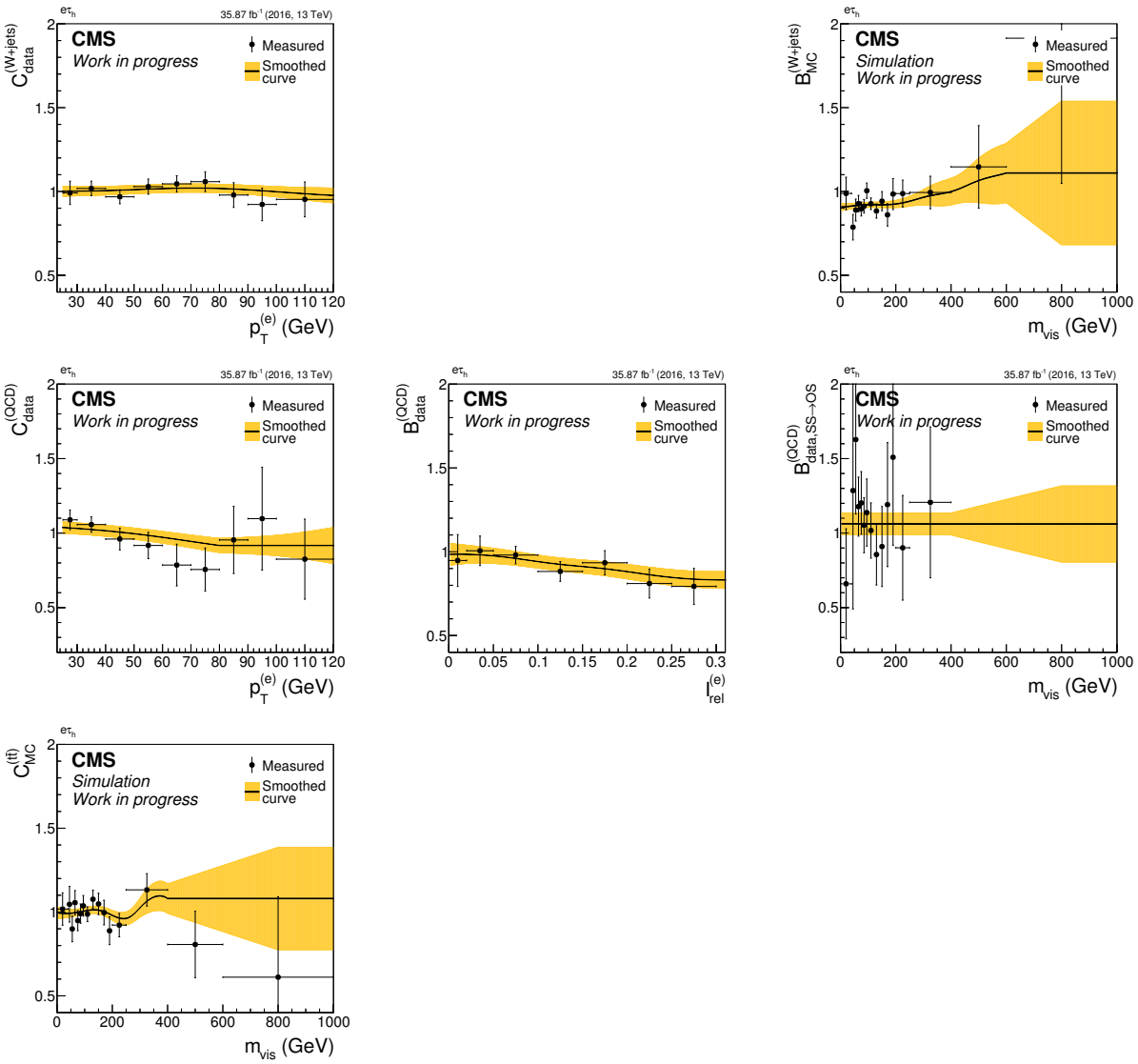


Figure A.6: Shown are all the F_{F}-related corrections in the eτ_{h} channel using 2016 data. From top to bottom, the correction applied to the raw $F_{\mathrm{F}}^{(\mathrm{W+jets})}$, $F_{\mathrm{F}}^{(\mathrm{QCD})}$ and $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ are displayed, respectively. On the left, closure corrections are shown, and in the middle and right column the bias corrections. Each correction measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

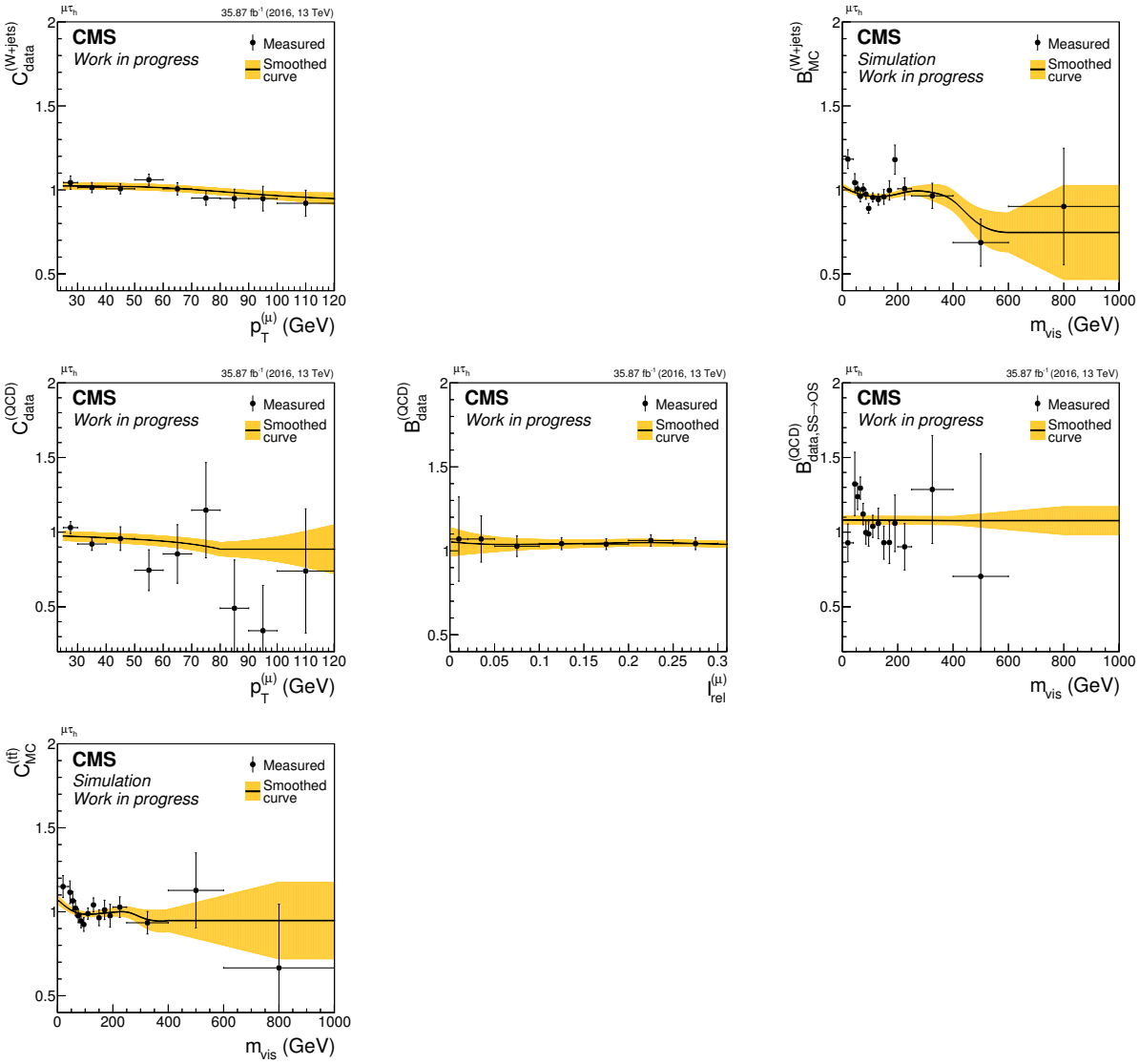


Figure A.7: Shown are all the F_{F}-related corrections in the $\mu\tau_{\mathrm{h}}$ channel using 2016 data. From top to bottom, the correction applied to the raw $F_{\mathrm{F}}^{(\mathrm{W+jets})}$, $F_{\mathrm{F}}^{(\mathrm{QCD})}$ and $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ are displayed, respectively. On the left, closure corrections are shown, and in the middle and right column the bias corrections. Each correction measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

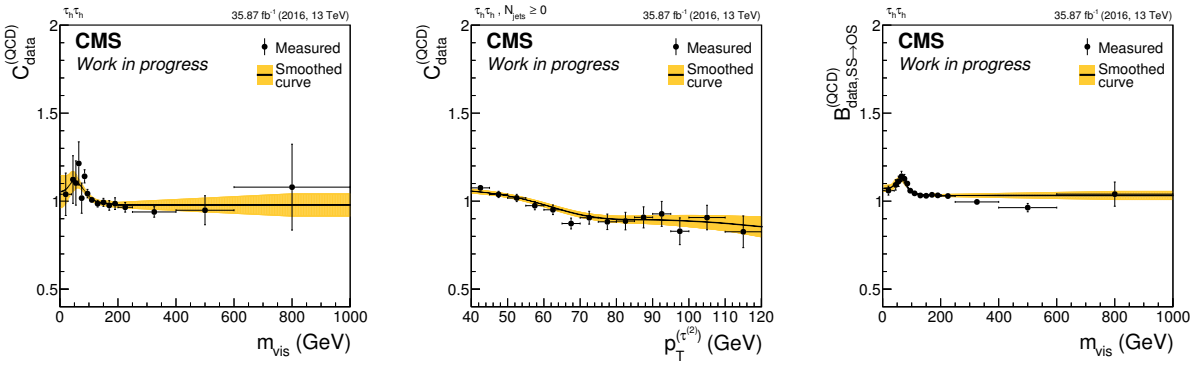


Figure A.8: Shown are all the F_{F}-related corrections in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel using 2016 data. From left to right, the two closure corrections are displayed first and then the bias correction. Each correction measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

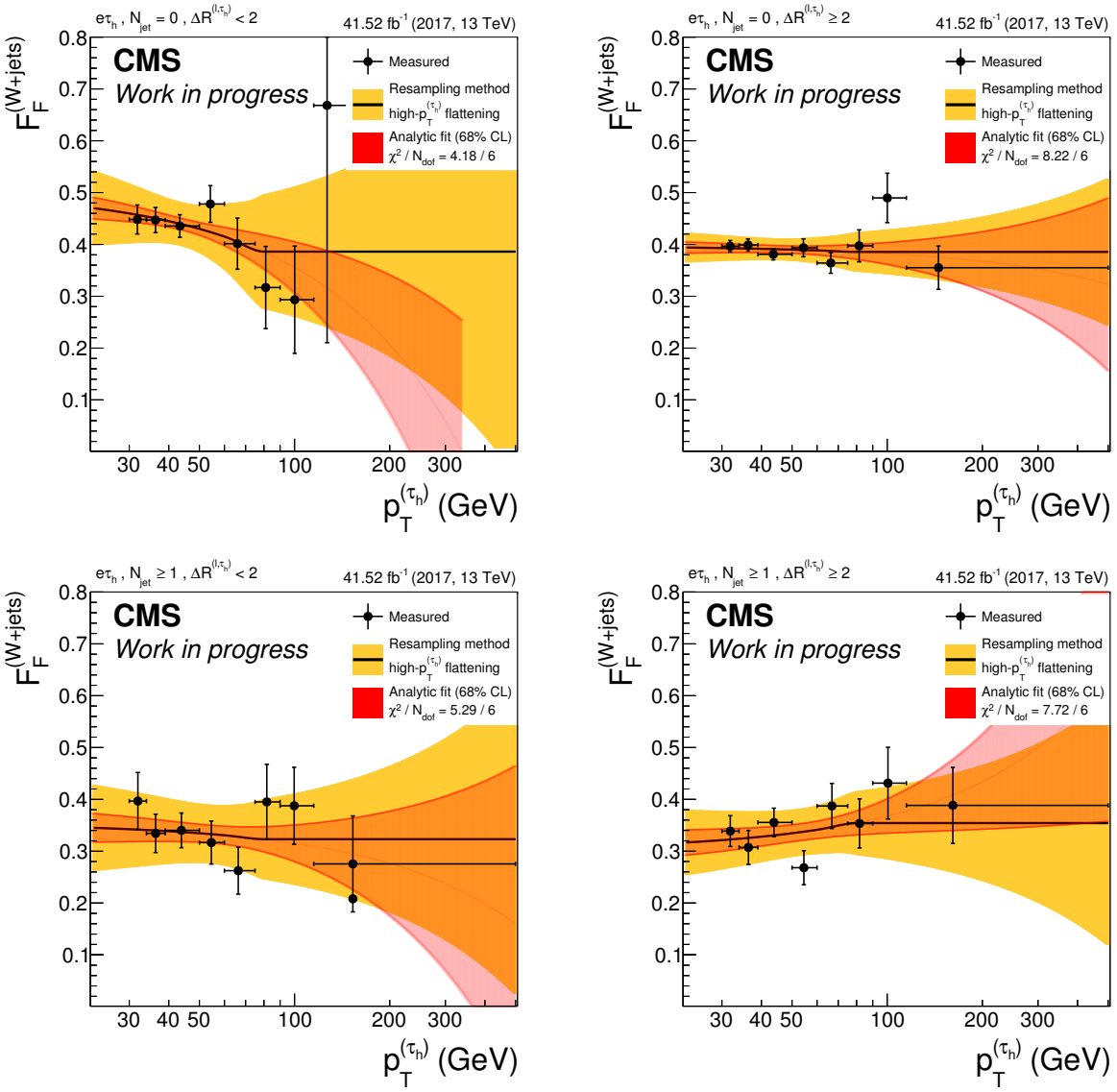


Figure A.9: The quantity $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown for the eτ_{h} channel using 2017 data. Top and bottom display the two N_{jets} categories – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} \geq 1$, respectively. On the left, the $\Delta R^{(\ell,\tau_{\mathrm{h}})} < 2$ category is shown and on the right the $\Delta R^{(\ell,\tau_{\mathrm{h}})} \geq 2$ category. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid black line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

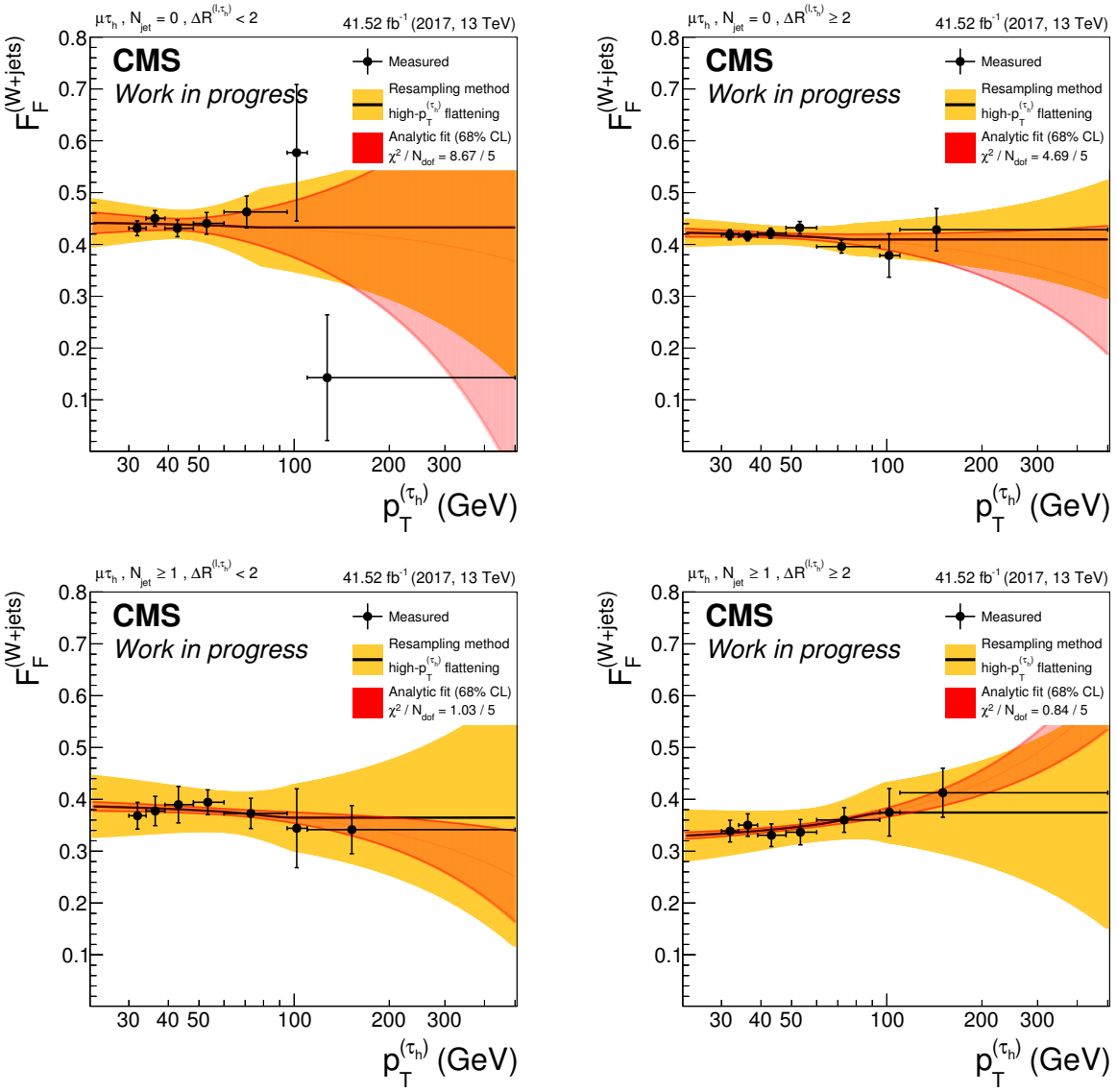


Figure A.10: The quantity $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown for the $\mu\tau_{\mathrm{h}}$ channel using 2017 data. Top and bottom display the two N_{jets} categories – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} \geq 1$, respectively. On the left, the $\Delta R^{(\ell,\tau_{\mathrm{h}})} < 2$ category is shown and on the right the $\Delta R^{(\ell,\tau_{\mathrm{h}})} \geq 2$ category. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid black line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

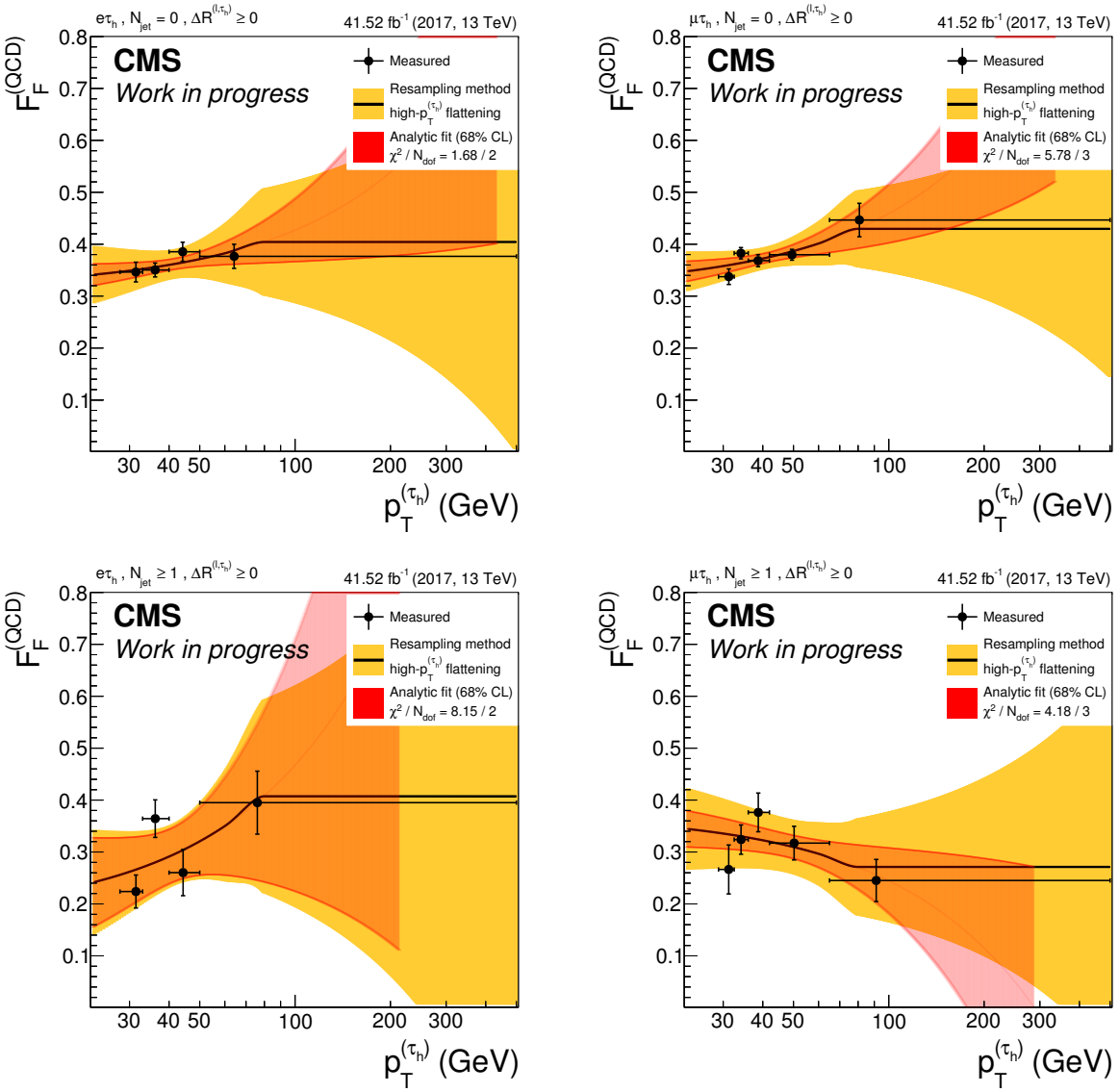


Figure A.11: The quantity $F_{\mathrm{F}}^{(\mathrm{QCD})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown using 2017 data. On the left, distributions for the $e\tau_{\mathrm{h}}$ channel are displayed and corresponding distributions for the $\mu\tau_{\mathrm{h}}$ channel are displayed on the right. Top and bottom display the two N_{jets} categories – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} \geq 1$, respectively. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

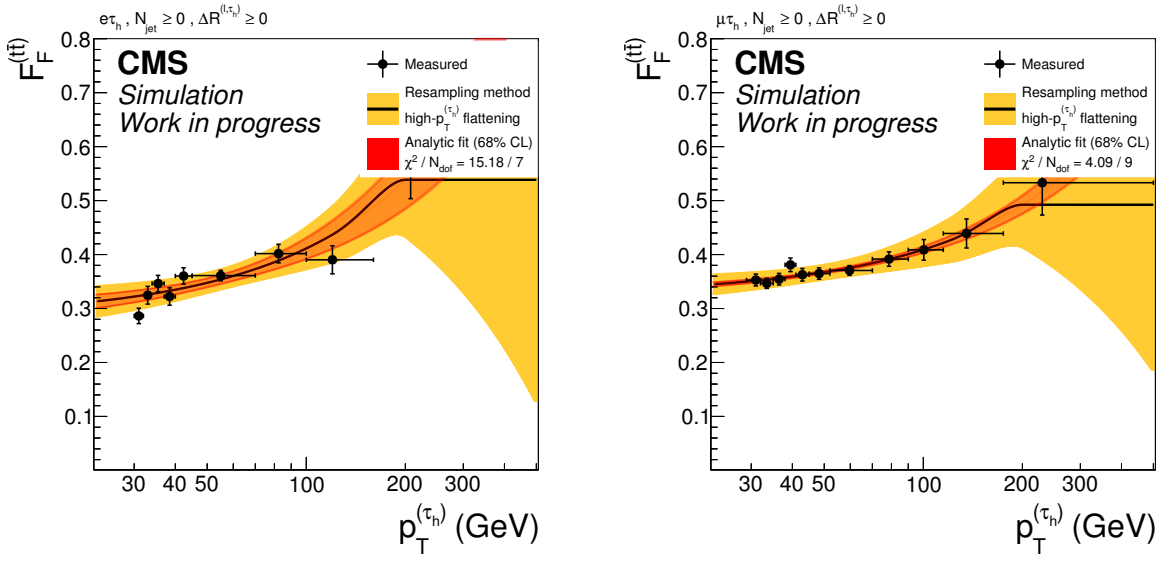


Figure A.12: The quantity $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown using simulated events from the 2017 data-taking period. On the left, the distribution for the $\mathrm{e}\tau_{\mathrm{h}}$ channel is displayed and the corresponding distribution for the $\mu\tau_{\mathrm{h}}$ channel is displayed on the right. The measurement of $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ is inclusive in N_{jets}. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

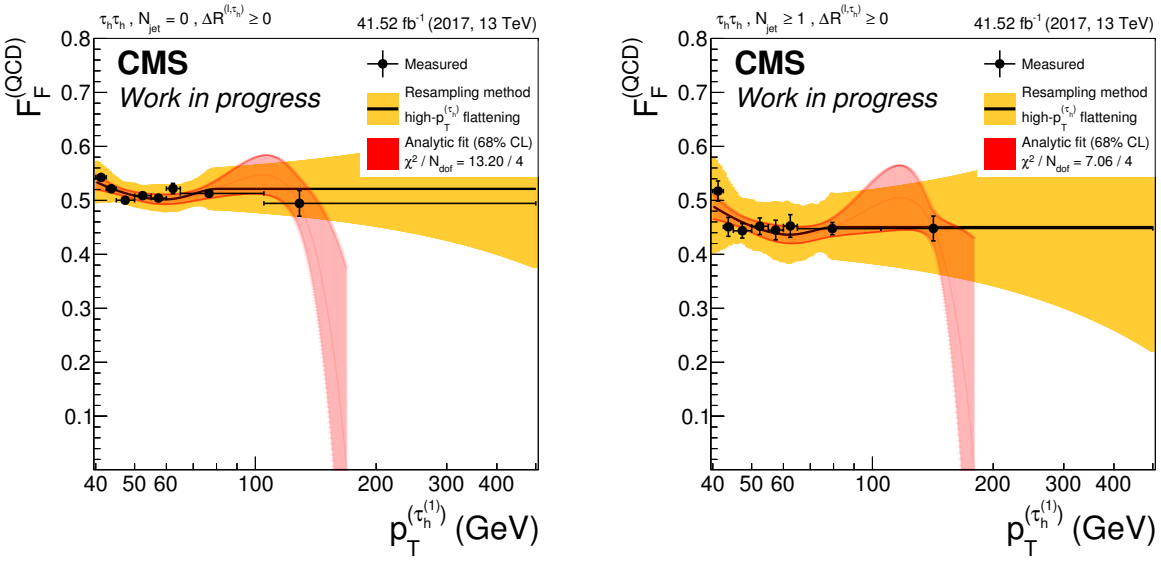


Figure A.13: The quantity $F_{\mathrm{F}}^{(\mathrm{QCD})}$ as a function of $p_{\mathrm{T}}^{(\tau_1)}$ is shown for the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel using 2017 data. Left and right display the two N_{jets} categories – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} \geq 1$, respectively. The F_{F} parametrization is shown as a solid line. It consists of a third order polynomal fit which is truncated at high $p_{\mathrm{T}}^{(\tau_1)}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid black line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

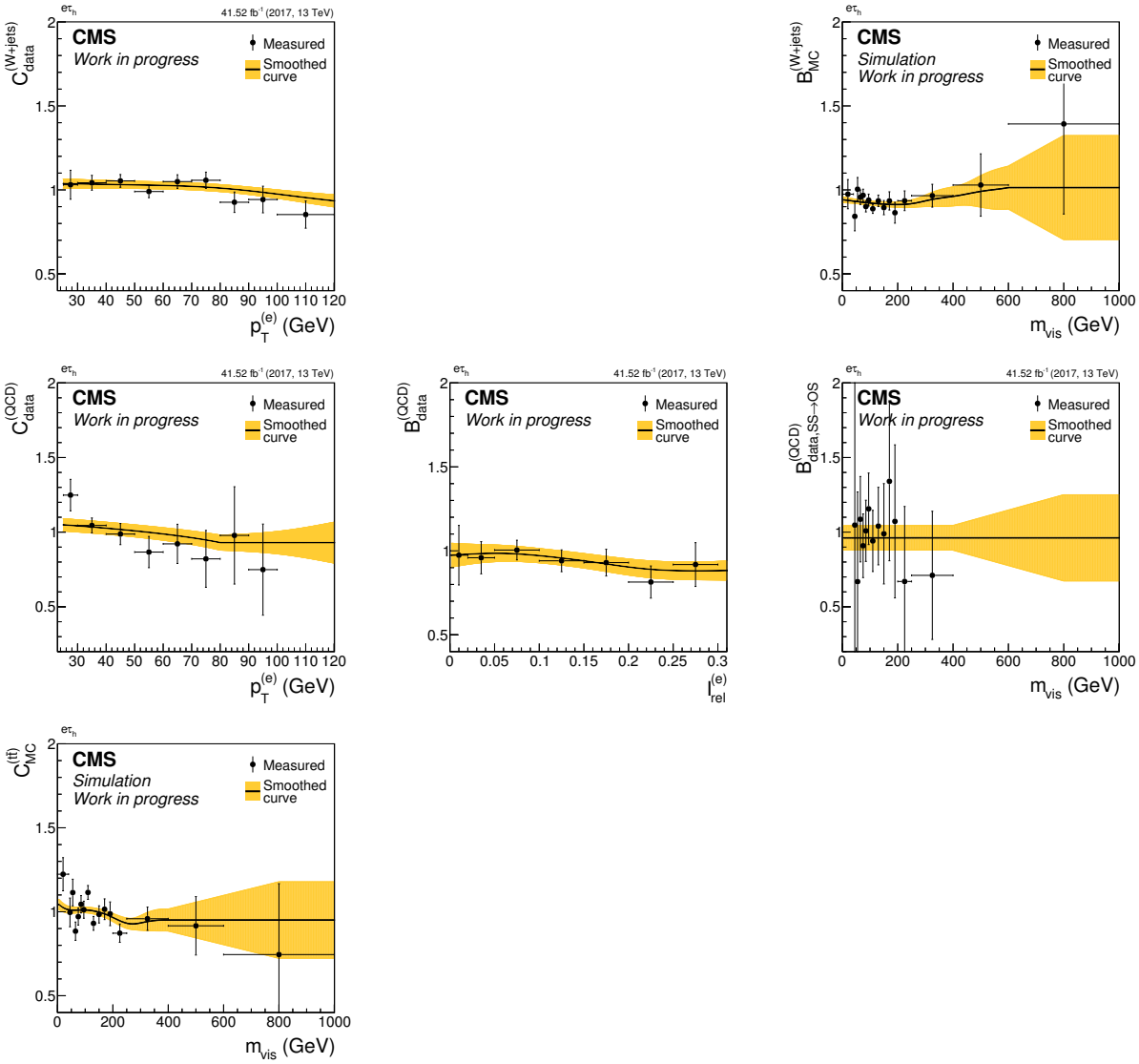


Figure A.14: Shown are all the F_{F}-related corrections in the eτ_{h} channel using 2017 data. From top to bottom, the correction applied to the raw $F_{\mathrm{F}}^{(\mathrm{W+jets})}$, $F_{\mathrm{F}}^{(\mathrm{QCD})}$ and $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ are displayed, respectively. On the left, closure corrections are shown, and in the middle and right column the bias corrections. Each correction measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

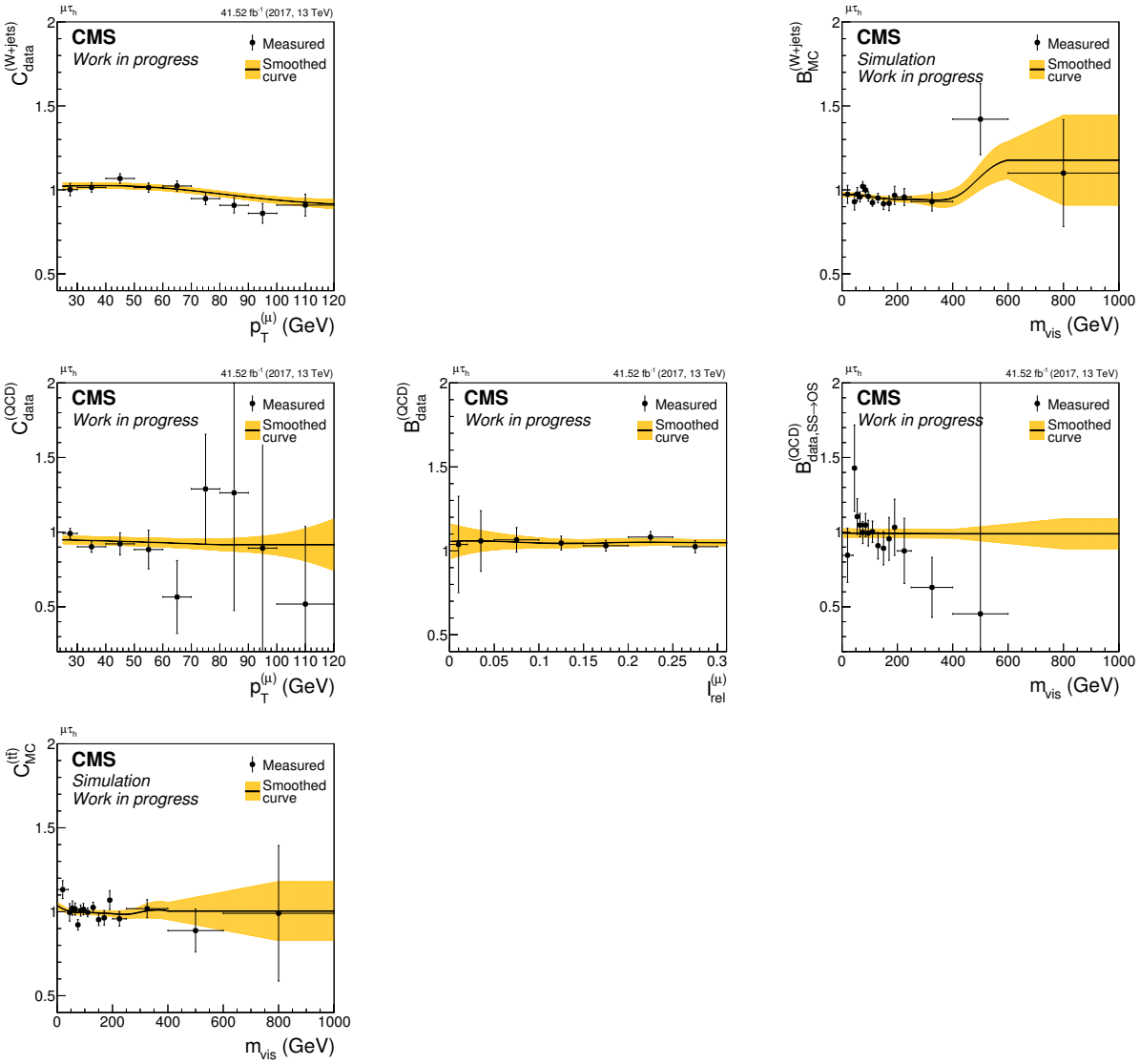


Figure A.15: Shown are all the F_{F}-related corrections in the $\mu\tau_{\mathrm{h}}$ channel using 2017 data. From top to bottom, the correction applied to the raw $F_{\mathrm{F}}^{(\mathrm{W+jets})}$, $F_{\mathrm{F}}^{(\mathrm{QCD})}$ and $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ are displayed, respectively. On the left, closure corrections are shown, and in the middle and right column the bias corrections. Each correction measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

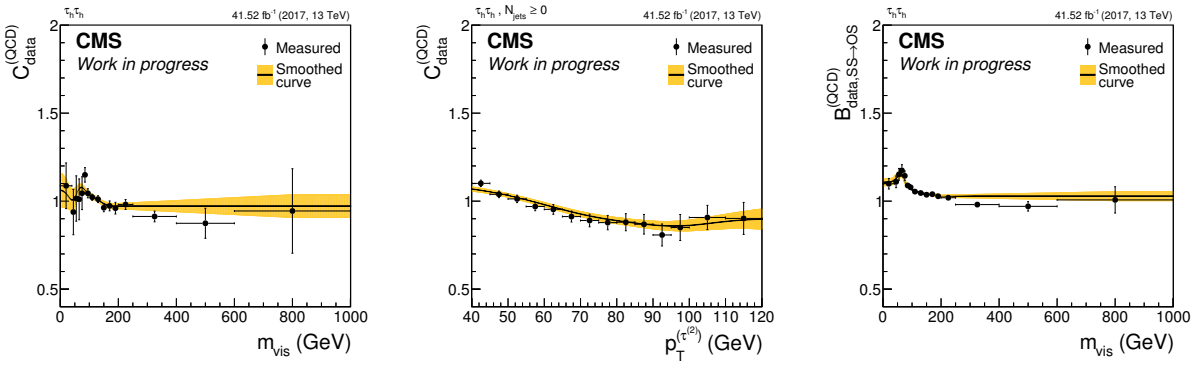


Figure A.16: Shown are all the F_{F}-related corrections in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel using 2017 data. From left to right, the two closure corrections are displayed first and then the bias correction. Each correction measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

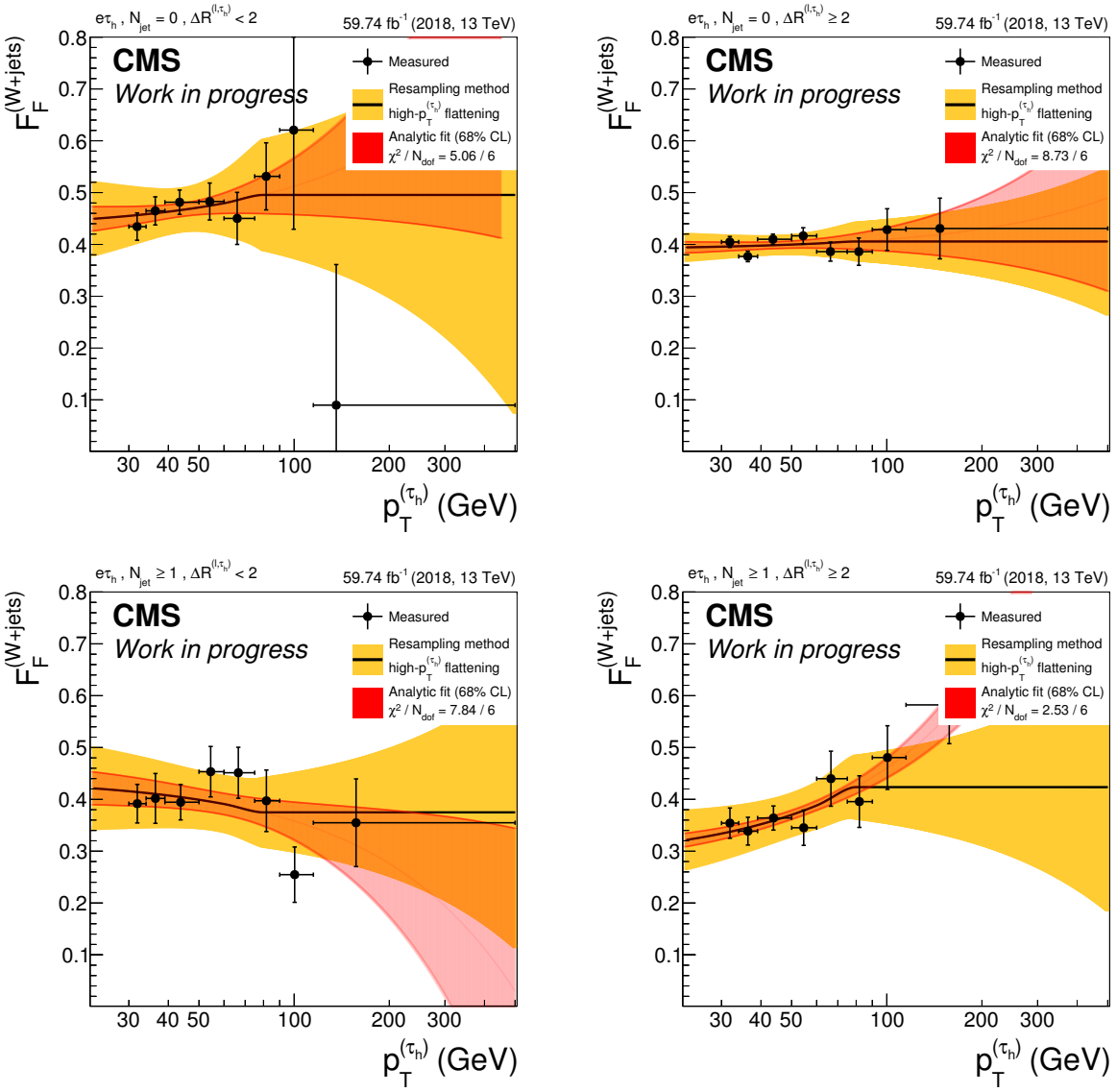


Figure A.17: The quantity $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown for the eτ_{h} channel using 2018 data. Top and bottom display the two N_{jets} categories – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} \geq 1$, respectively. On the left, the $\Delta R^{(\ell,\tau_{\mathrm{h}})} < 2$ category is shown and on the right the $\Delta R^{(\ell,\tau_{\mathrm{h}})} \geq 2$ category. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid black line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

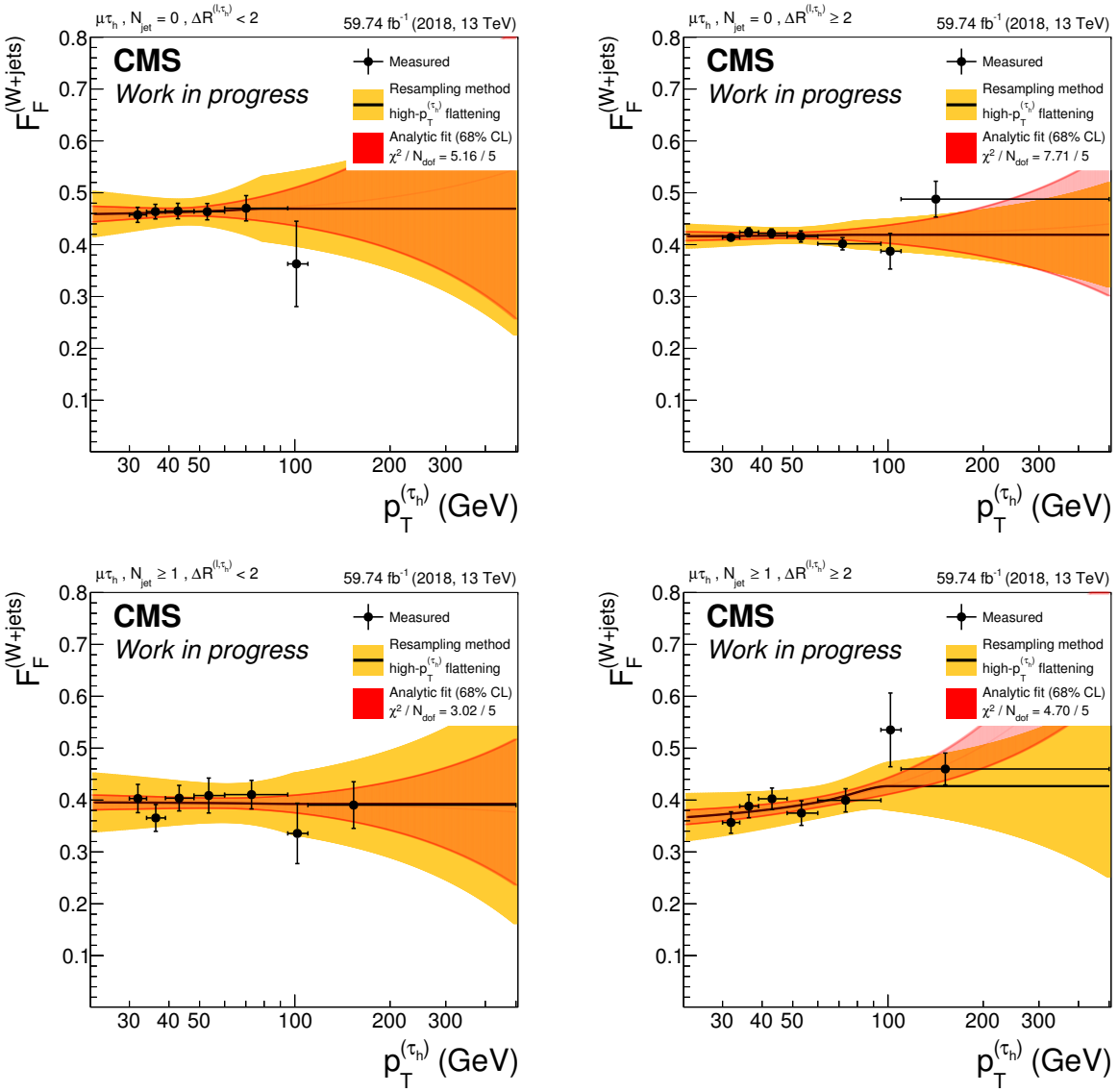


Figure A.18: The quantity $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown for the $\mu\tau_{\mathrm{h}}$ channel using 2018 data. Top and bottom display the two N_{jets} categories – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} \geq 1$, respectively. On the left, the $\Delta R^{(\ell,\tau_{\mathrm{h}})} < 2$ category is shown and on the right the $\Delta R^{(\ell,\tau_{\mathrm{h}})} \geq 2$ category. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid black line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

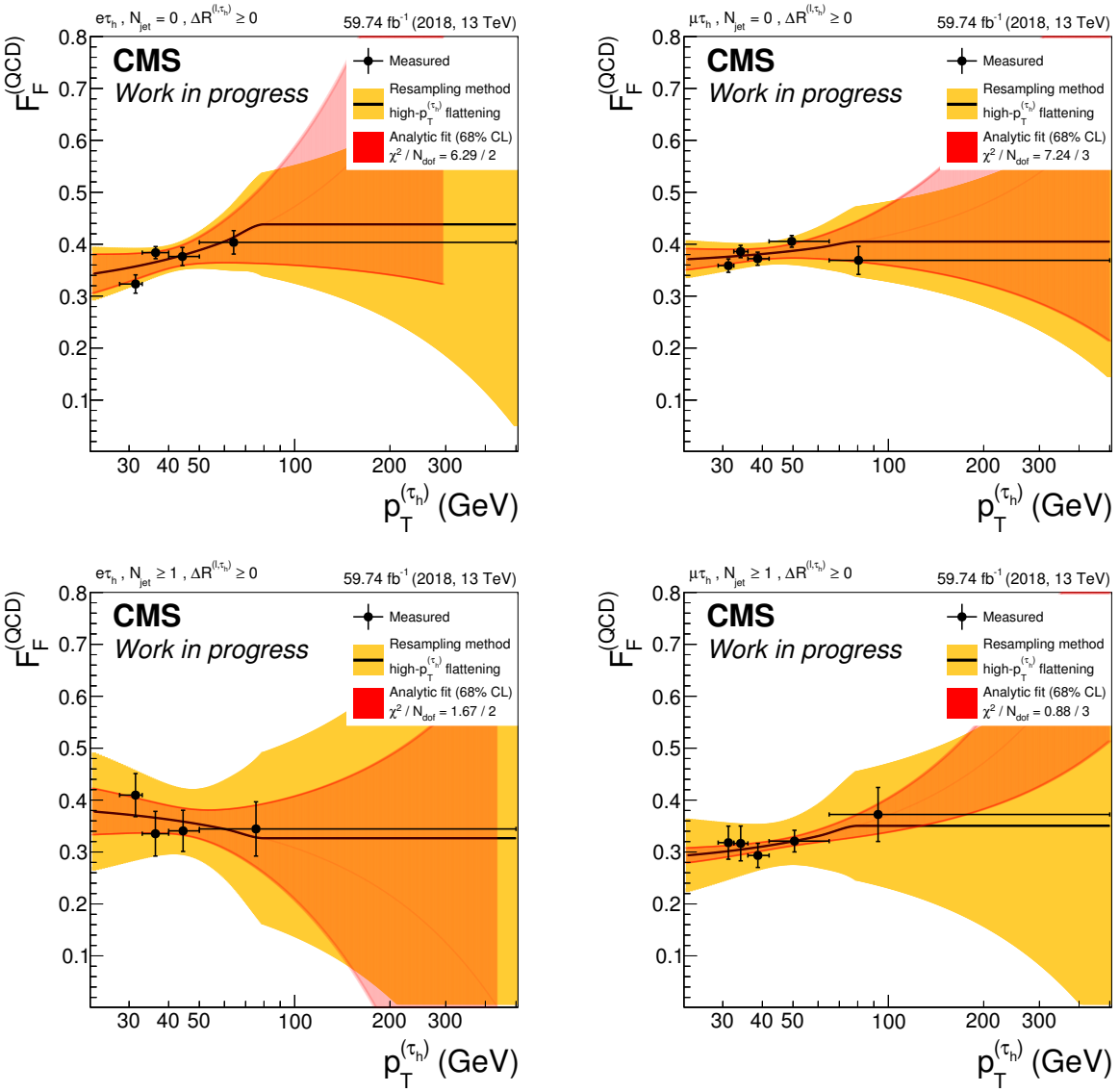


Figure A.19: The quantity $F_{\mathrm{F}}^{(\mathrm{QCD})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown using 2018 data. On the left, distributions for the eτ_{h} channel are displayed and corresponding distributions for the $\mu\tau_{\mathrm{h}}$ channel are displayed on the right. Top and bottom display the two N_{jets} categories – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} \geq 1$, respectively. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

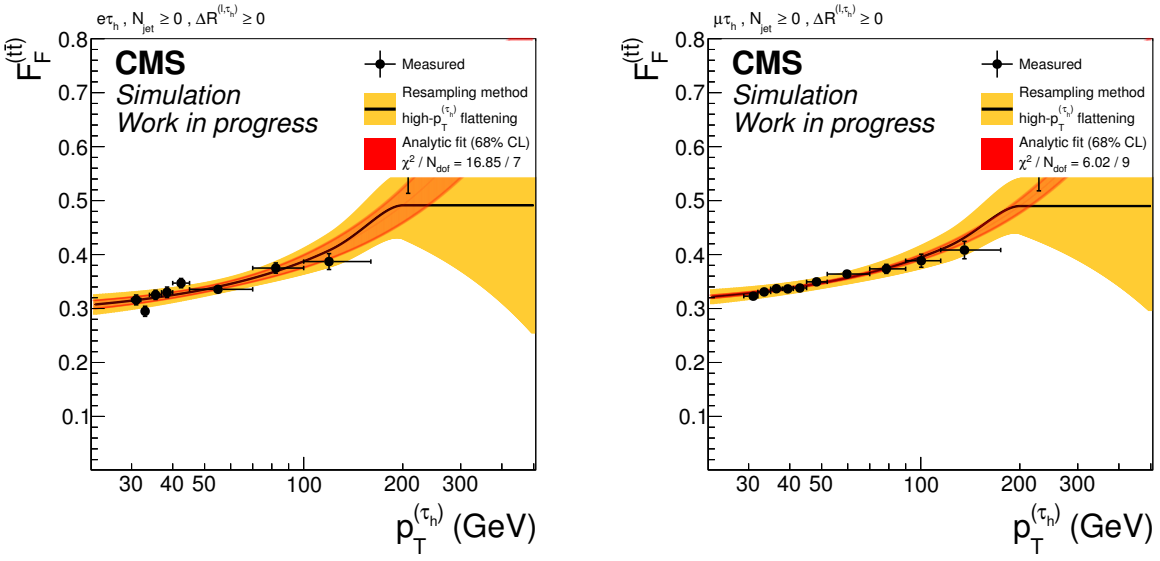


Figure A.20: The quantity $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown using simulated events from the 2018 data-taking period. On the left, the distribution for the $\mathrm{e}\tau_{\mathrm{h}}$ channel is displayed and the corresponding distribution for the $\mu\tau_{\mathrm{h}}$ channel is displayed on the right. The measurement of $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ is inclusive in N_{jets}. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

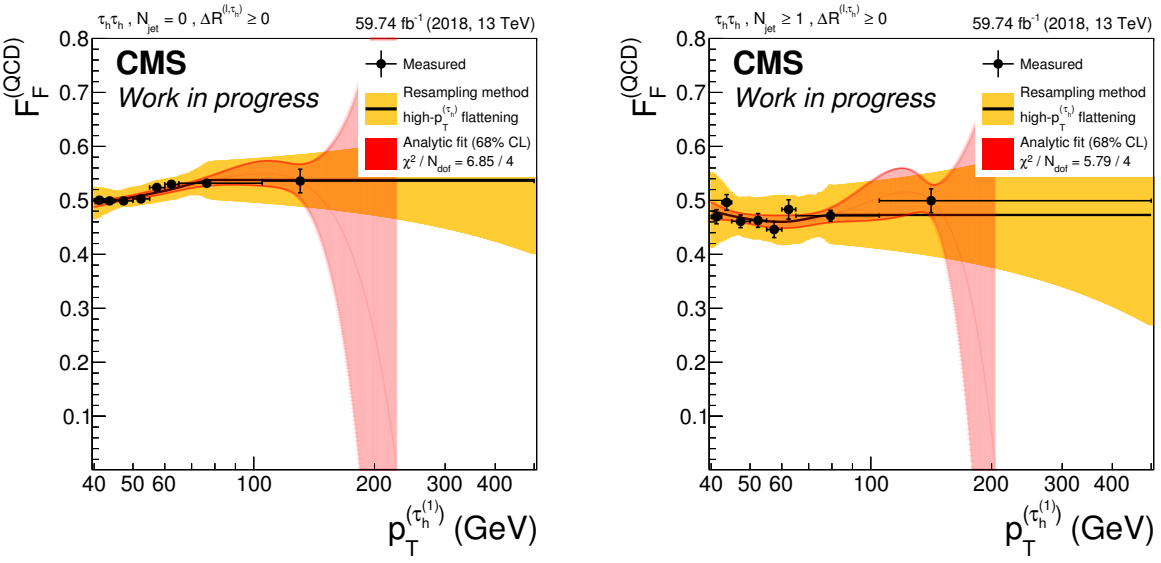


Figure A.21: The quantity $F_{\mathrm{F}}^{(\mathrm{QCD})}$ as a function of $p_{\mathrm{T}}^{(\tau_1)}$ is shown for the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel using 2018 data. Left and right display the two N_{jets} categories – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} \geq 1$, respectively. The F_{F} parametrization is shown as a solid line. It consists of a third order polynomal fit which is truncated at high $p_{\mathrm{T}}^{(\tau_1)}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid black line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

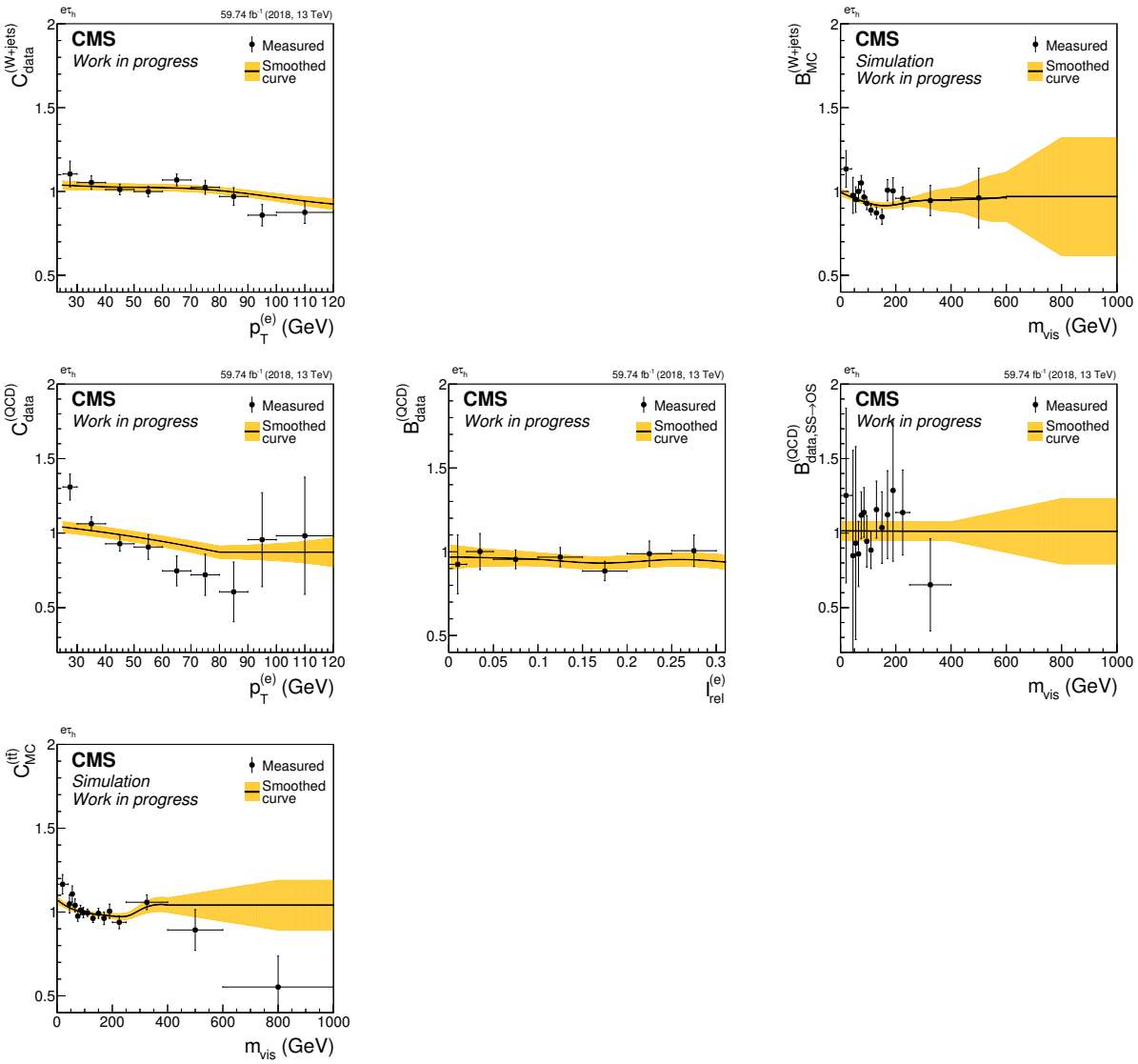


Figure A.22: Shown are all the F_{F}-related corrections in the $\mathrm{e}\tau_{\mathrm{h}}$ channel using 2018 data. From top to bottom, the correction applied to the raw $F_{\mathrm{F}}^{(\mathrm{W+jets})}$, $F_{\mathrm{F}}^{(\mathrm{QCD})}$ and $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ are displayed, respectively. On the left, closure corrections are shown, and in the middle and right column the bias corrections. Each correction measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

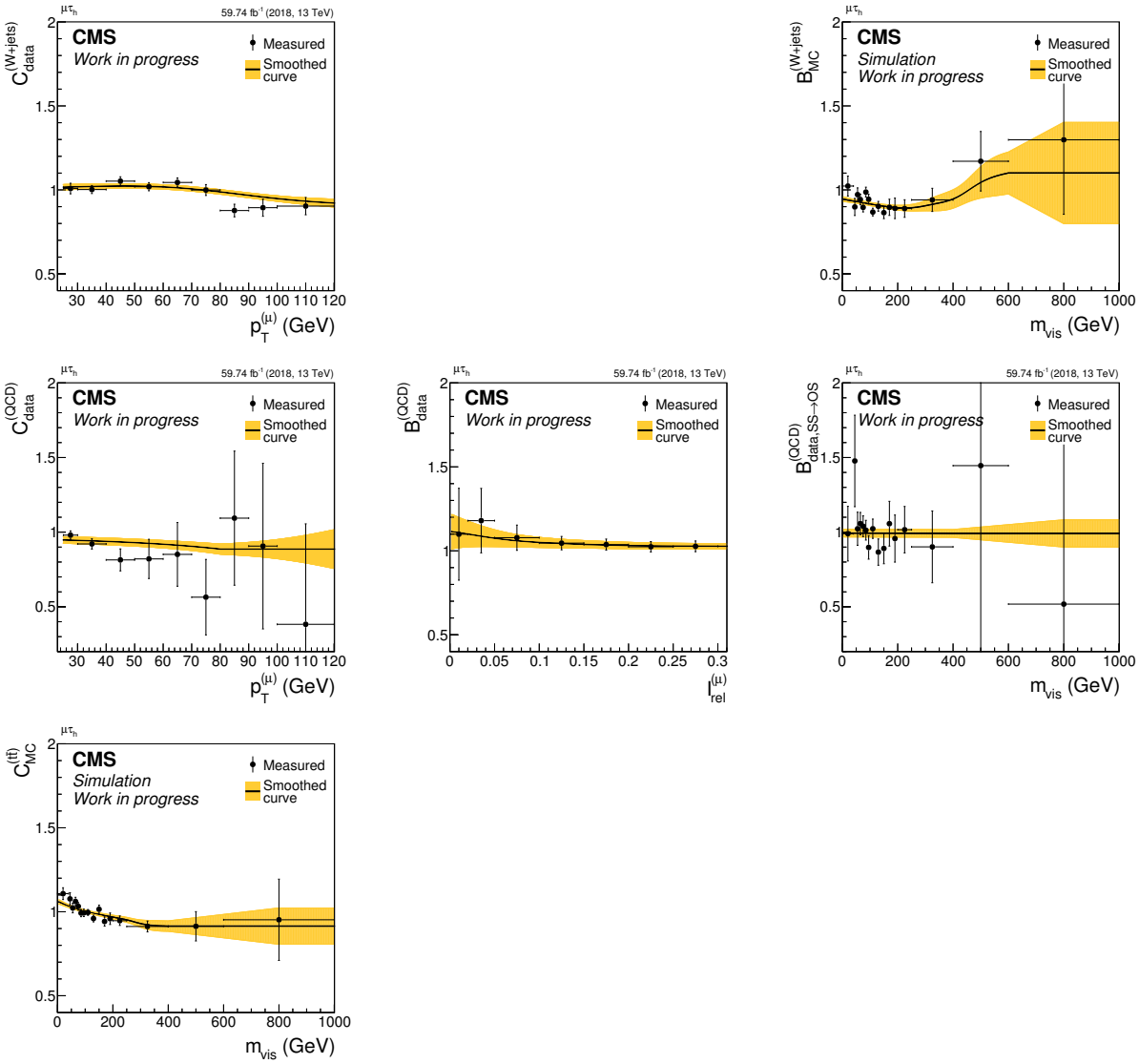


Figure A.23: Shown are all the F_{F}-related corrections in the $\mu\tau_{\mathrm{h}}$ channel using 2018 data. From top to bottom, the correction applied to the raw $F_{\mathrm{F}}^{(\mathrm{W+jets})}$, $F_{\mathrm{F}}^{(\mathrm{QCD})}$ and $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ are displayed, respectively. On the left, closure corrections are shown, and in the middle and right column the bias corrections. Each correction measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

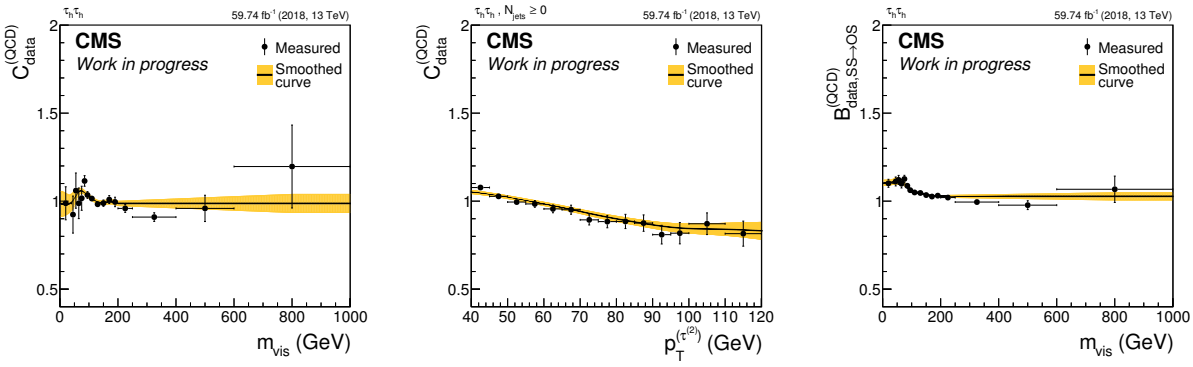


Figure A.24: Shown are all the F_{F}-related corrections in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel using 2018 data. From left to right, the two closure corrections are displayed first and then the bias correction. Each correction measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

Appendix B

Extra Material for the SM H $\rightarrow \tau\tau$ Analysis

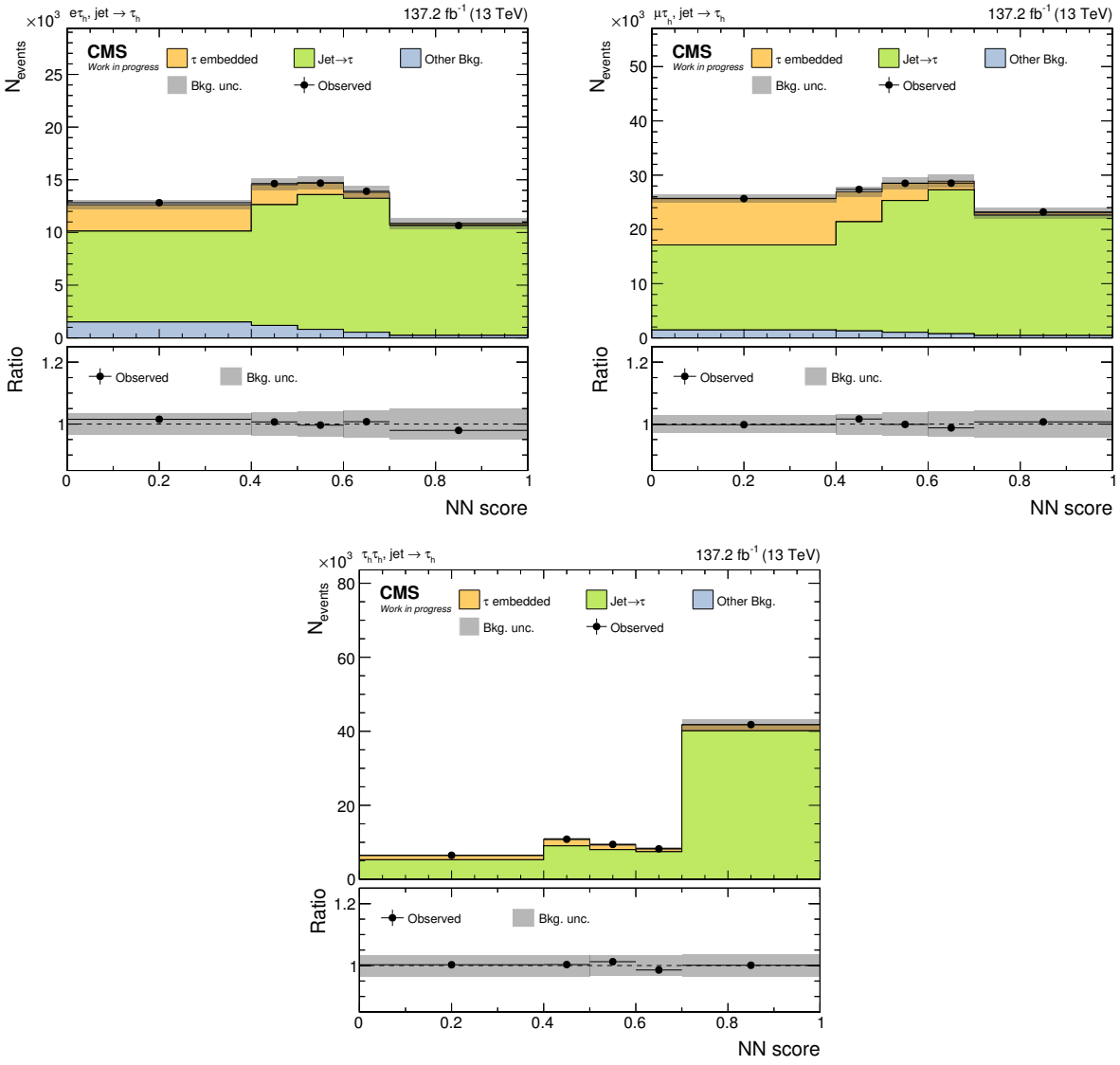


Figure B.1: Event distributions are shown for the eτ_{h}, $\mu\tau_{\mathrm{h}}$ and $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel. Only events assigned to the jet $\rightarrow \tau_{\mathrm{h}}$ class are shown here as indicated in the title of each plot. Contributions from true jet $\rightarrow \tau_{\mathrm{h}}$ processes are shown in green and dominate especially in bins with a large NN output score. Note that even thought the first bin of each histogram extends down to zero, the lower bound of the NN output for classification is given by the values quoted in Equation (9.25). Uncertainties referred to as "Bkg. unc." comprise the full uncertainty model as presented in Section 9.2.2 after the maximum likelihood fit. The same plots before the maximum likelihood fit are given in Figure 9.6. The figures are taken from [12].

Appendix C

Fake Factors for the SM H $\rightarrow \tau\tau$ Analysis

In this section, raw F_{F}'s and their corrections are presented using the definition of Equation (8.4). They are used to estimate the jet $\rightarrow \tau_{\mathrm{h}}$ background in the SM H $\rightarrow \tau\tau$ analysis, presented in Section 9.2. Control plots from $\mathrm{DR}_{\mathrm{W+jets}}$ and $\mathrm{DR}_{\mathrm{QCD}}$ are also shown. In order to ease the navigation, Table C.1 summarizes the organization of the plots. Only plots for the 2016 and 2017 years of data taking are shown because plots from 2018 can be found throughout Section 8.2.

year	channel	$\mathrm{DR}_{\mathrm{W+jets}}$	$F_{\mathrm{F}}^{(\mathrm{W+jets})}$	$\mathrm{DR}_{\mathrm{QCD}}$	$F_{\mathrm{F}}^{(\mathrm{QCD})}$	$F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$	F_{F} corrections
2016	$\mathrm{e}\tau_{\mathrm{h}}$	Fig. C.1	Fig. C.2	Fig. C.4	Fig. C.5	Fig. C.6	Fig. C.9
	$\mu\tau_{\mathrm{h}}$		Fig. C.3				Fig. C.10
	$\tau_{\mathrm{h}}\tau_{\mathrm{h}}$	–	–	Fig. C.7	Fig. C.8	–	Fig. C.11
2017	$\mathrm{e}\tau_{\mathrm{h}}$	Fig. C.12	Fig. C.13	Fig. C.15	Fig. C.16	Fig. C.17	Fig. C.20
	$\mu\tau_{\mathrm{h}}$		Fig. C.14				Fig. C.21
	$\tau_{\mathrm{h}}\tau_{\mathrm{h}}$	–	–	Fig. C.18	Fig. C.19	–	Fig. C.22

Table C.1: This is a summary table linking all figures of control plots, raw F_{F}'s and their corrections. Fake factors are derived according to Equation (8.4).

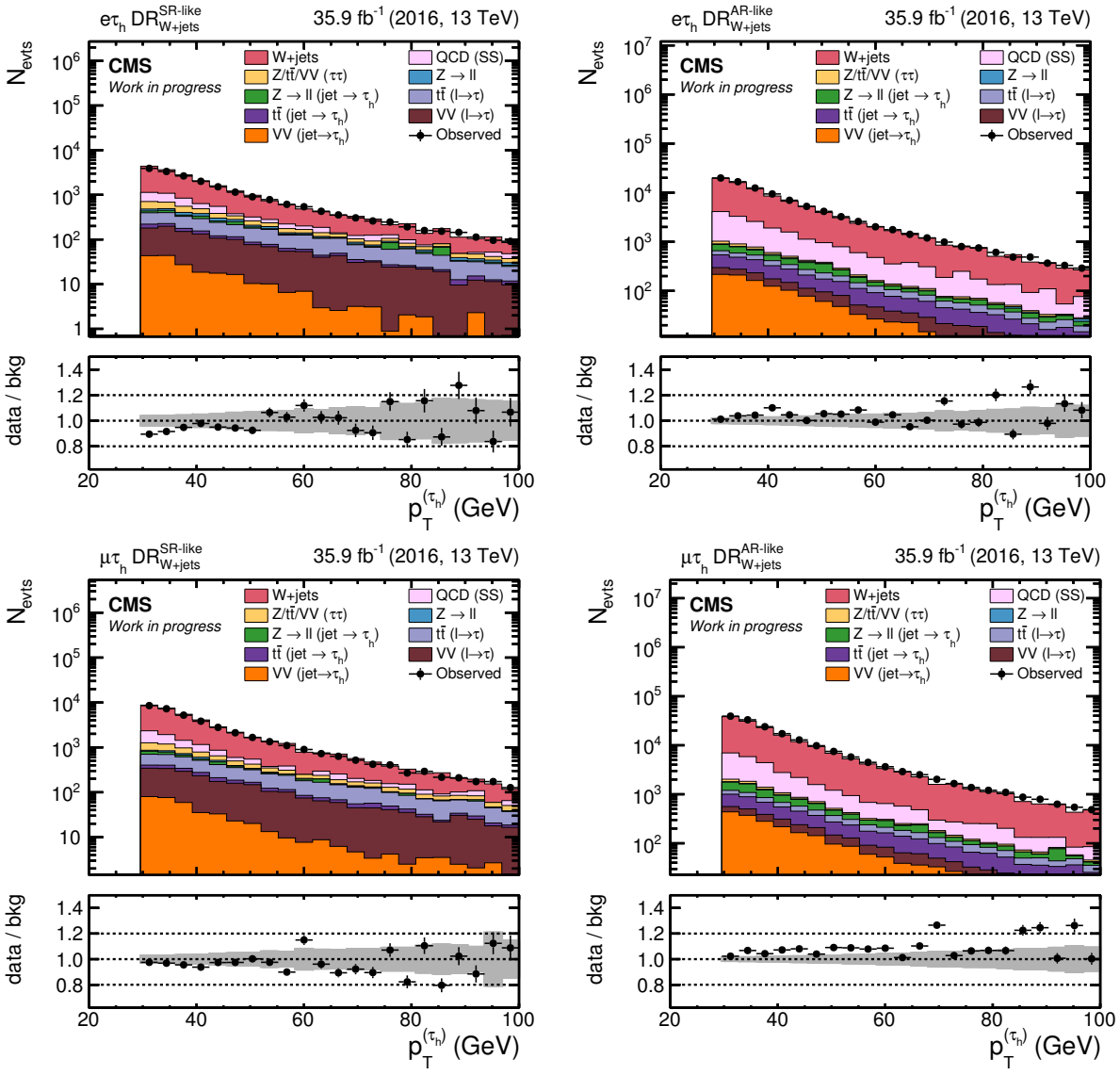


Figure C.1: $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ distributions inside $\mathrm{DR}_{\mathrm{W+jets}}$ for the 2016 data-taking period are shown. The top row shows distributions for the $\mathrm{e}\tau_{\mathrm{h}}$ channel and the bottom row for the $\mu\tau_{\mathrm{h}}$ channel. On the left the SR-like part of $\mathrm{DR}_{\mathrm{W+jets}}$ is shown and on the right the AR-like part. In all plots the W+jets process is at the top of the stacked histograms. It is the dominating process, while all other processes combined make up a few percent of the total yield. The QCD multijet estimate is taken from a SS region as detailed in the text. Processes labeled as $\mathrm{Z/t\bar{t}}/VV\,(\tau\tau)$ are estimated by the τ-embedding technique. The ratio is calculated as observed over predicted (sum of all filled histogram). The error bars in the ratio plot represent the uncertainty on the observed contribution and the gray band reflects the uncertainty on the predicted contribution. Only statistical uncertainties are shown here.

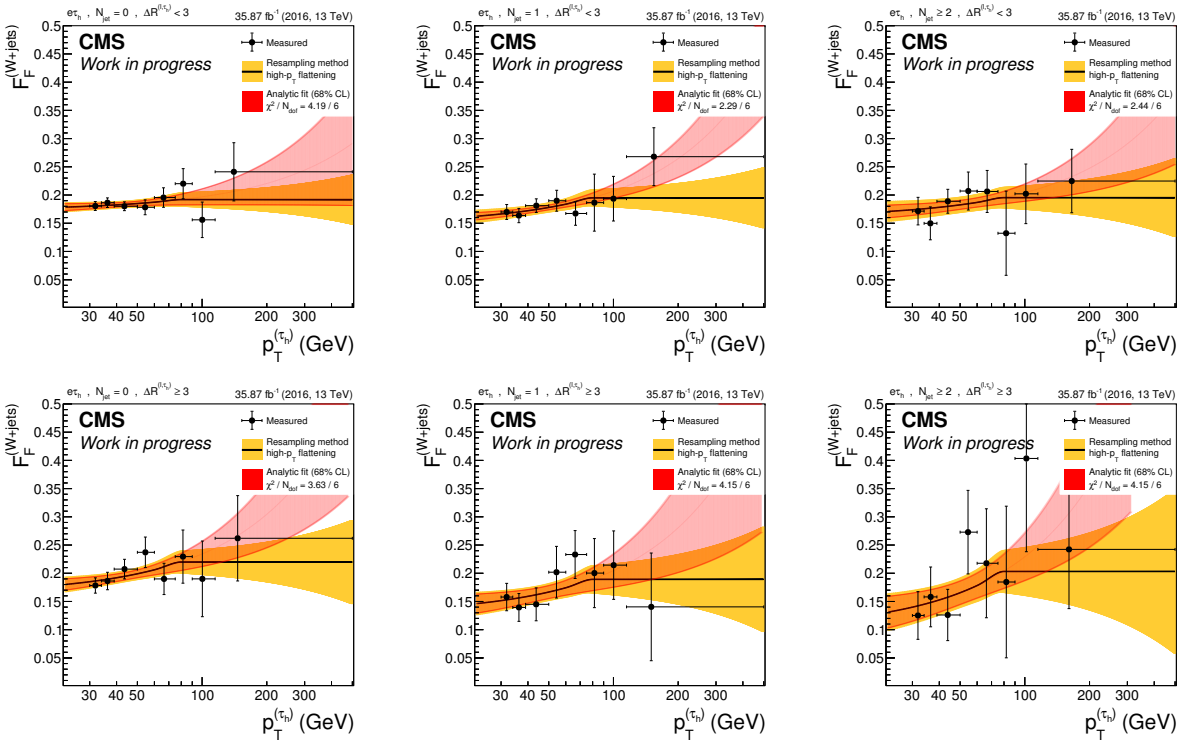


Figure C.2: The quantity $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown for the eτ_{h} channel using 2016 data. Top and bottom display the two $\Delta R^{(\ell,\tau_{\mathrm{h}})}$ categories – $\Delta R^{(\ell,\tau_{\mathrm{h}})} < 3$ and $\Delta R^{(\ell,\tau_{\mathrm{h}})} \geq 3$, respectively. From left to right, the three N_{jets} categories are displayed – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} = 1$ and $N_{\mathrm{jets}} \geq 2$, respectively. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid black line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

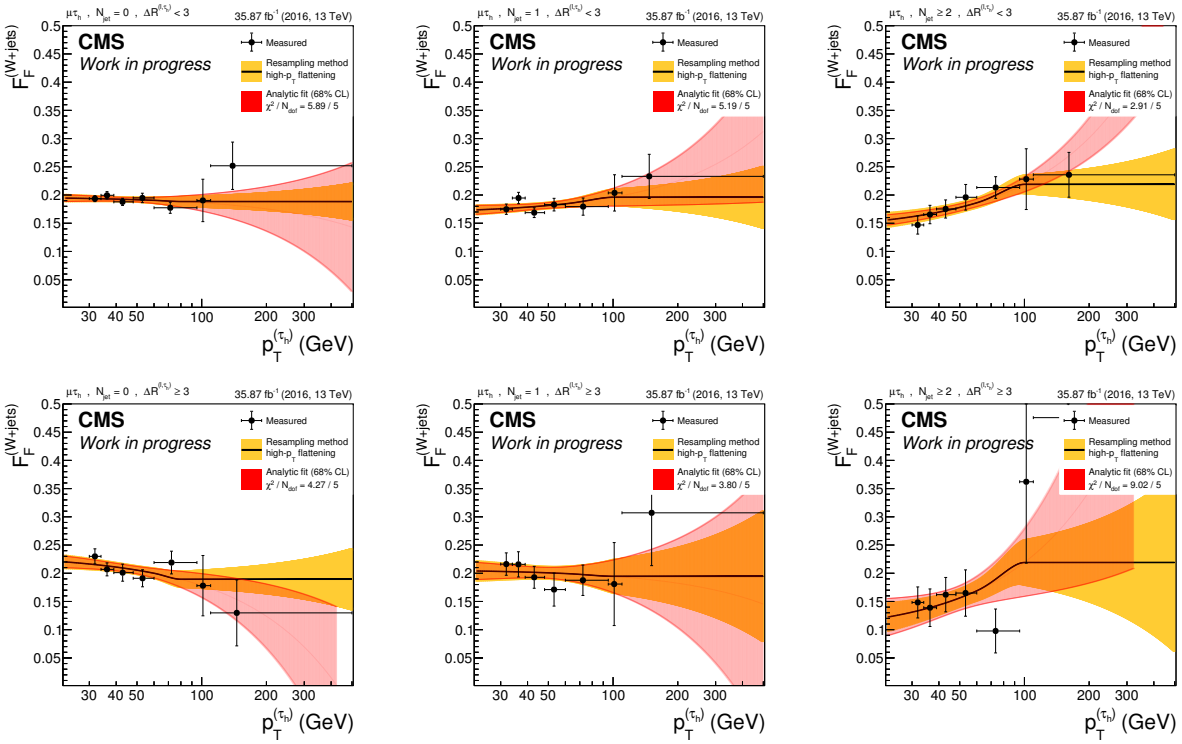


Figure C.3: The quantity $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown for the $\mu\tau_{\mathrm{h}}$ channel using 2016 data. Top and bottom display the two $\Delta R^{(\ell,\tau_{\mathrm{h}})}$ categories – $\Delta R^{(\ell,\tau_{\mathrm{h}})} < 3$ and $\Delta R^{(\ell,\tau_{\mathrm{h}})} \geq 3$, respectively. From left to right, the three N_{jets} categories are displayed – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} = 1$ and $N_{\mathrm{jets}} \geq 2$, respectively. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid black line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

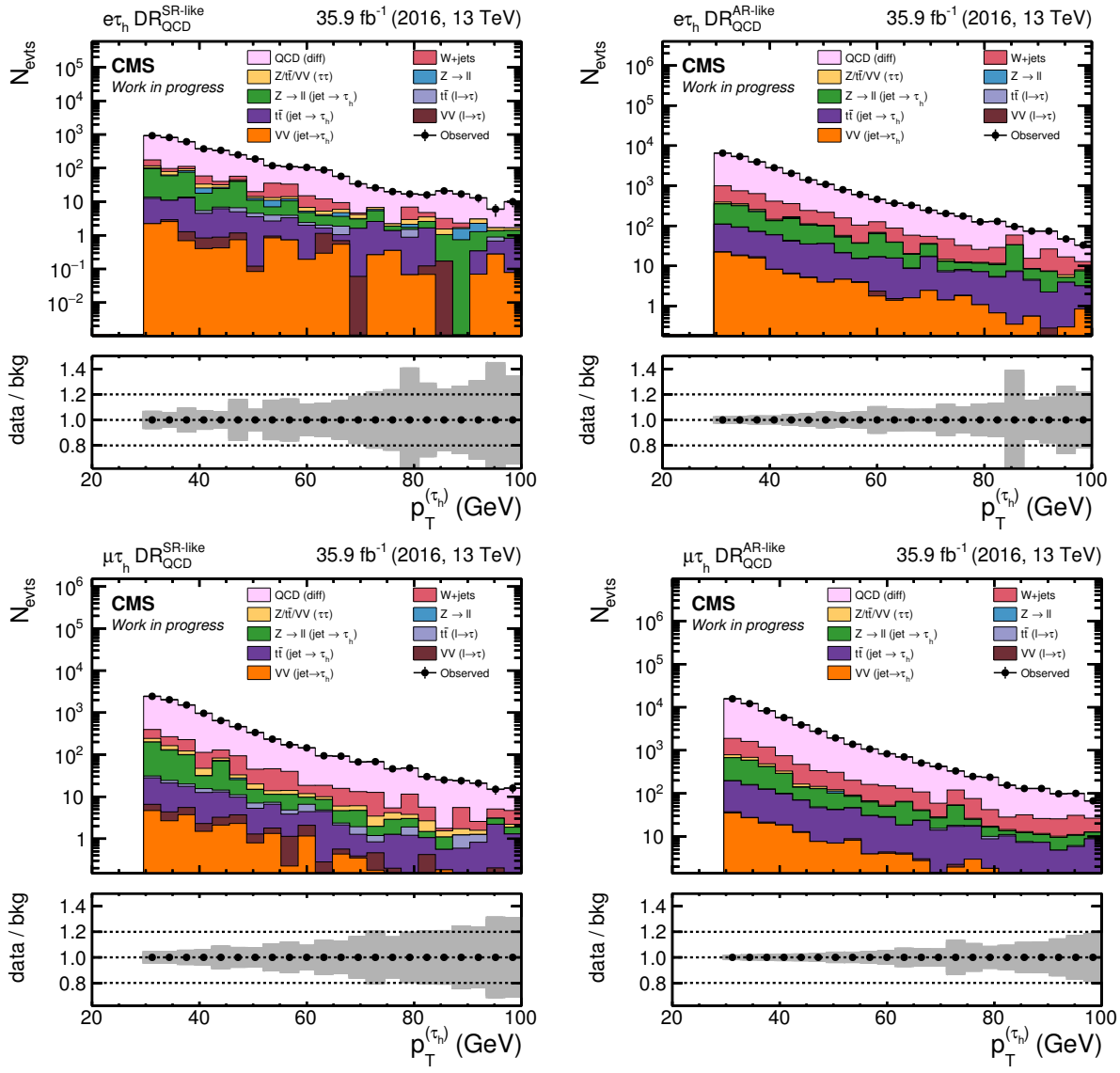


Figure C.4: $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ distributions inside $\mathrm{DR}_{\mathrm{QCD}}$ for the 2016 data-taking period are shown. The top row shows distributions for the $\mathrm{e}\tau_{\mathrm{h}}$ channel and the bottom row for the $\mu\tau_{\mathrm{h}}$ channel. On the left the SR-like part of $\mathrm{DR}_{\mathrm{QCD}}$ is shown and on the right the AR-like part. In all plots, the QCD multijet process is at the top of the stacked histograms. The estimated yield from QCD multijet events is given by the difference between data and the sum of all other background processes. It is the dominating process while all other processes combined make up at most a few percent of the total yield. The QCD multijet estimate is simply taken as the difference between the stacked histograms and the observation. Processes labeled as $\mathrm{Z/t\bar{t}}/VV\,(\tau\tau)$ are estimated by the τ-embedding technique. Only statistical uncertainties are shown here.

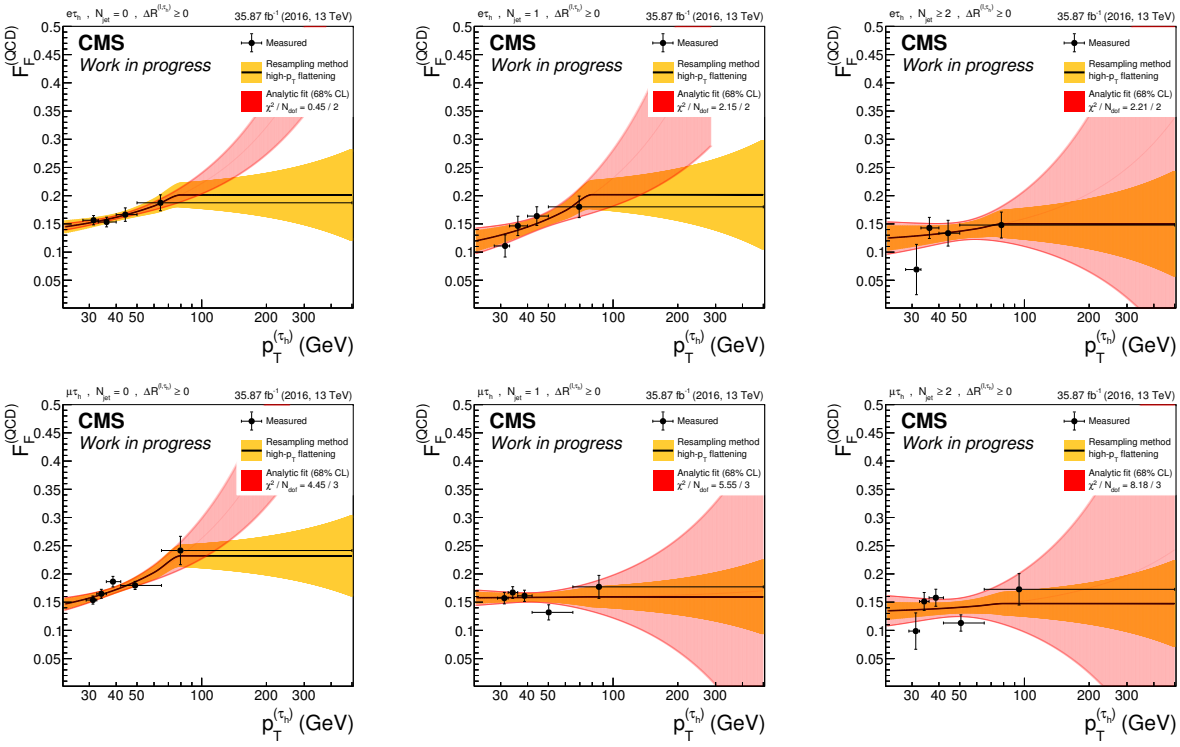


Figure C.5: The quantity $F_{\mathrm{F}}^{(\mathrm{QCD})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown using 2016 data. In the top row, distributions for the $\mathrm{e}\tau_{\mathrm{h}}$ channel are displayed and corresponding distributions for the $\mu\tau_{\mathrm{h}}$ channel are displayed on the bottom row. From left to right, the three N_{jets} categories are displayed – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} = 1$ and $N_{\mathrm{jets}} \geq 2$, respectively. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

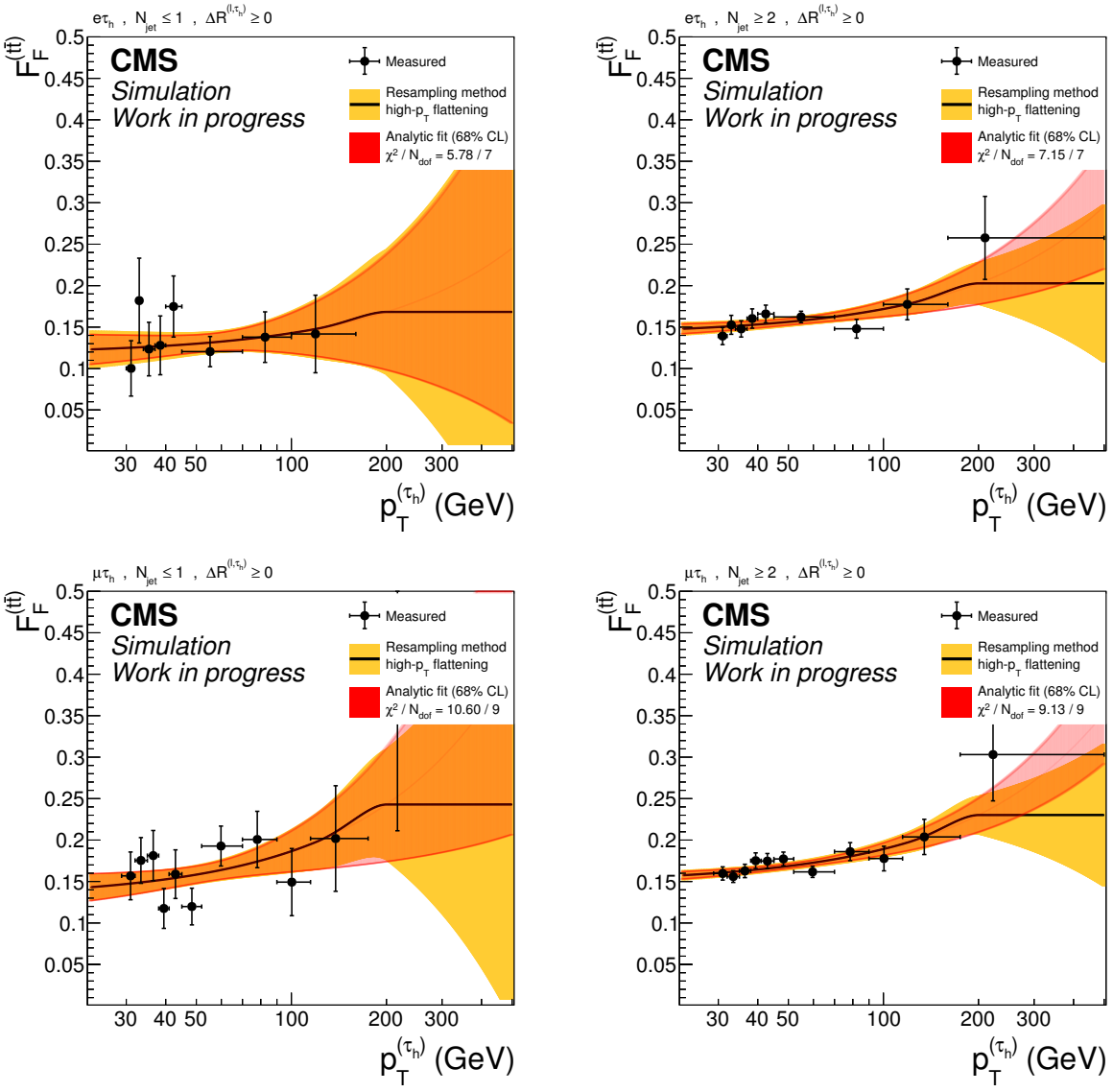


Figure C.6: The quantity $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown using simulated events from the 2016 data-taking period. The top row shows the distributions for the $\mathrm{e}\tau_{\mathrm{h}}$ channel and the corresponding distributions for the $\mu\tau_{\mathrm{h}}$ channel are displayed on bottom row. The measurement of $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ uses two N_{jets} categories, $N_{\mathrm{jets}} \leq 1$ (left) and $N_{\mathrm{jets}} \geq 2$ (right). The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

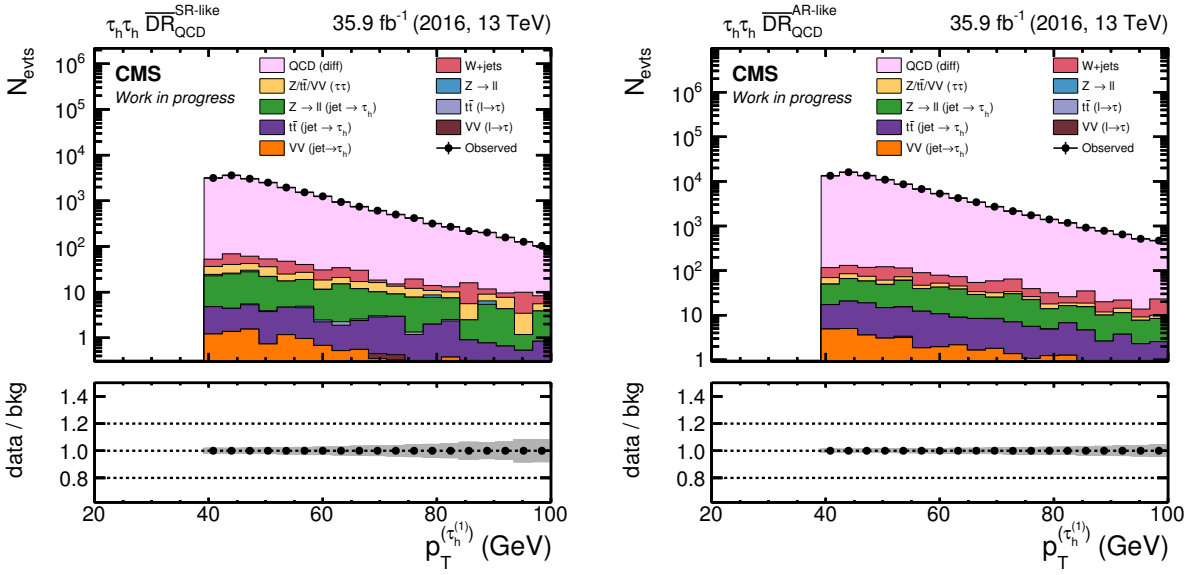


Figure C.7: Shown are $p_T^{(\tau_1)}$ distributions inside $\overline{DR}_{QCD}$ for the 2016 data-taking period. The distributions are measured for the $\tau_h\tau_h$ channel. On the left, the SR-like part of $\overline{DR}_{QCD}$ is shown and on the right, the AR-like part. In all plots, the QCD multijet process is at the top of the stacked histograms. The estimated yield from QCD multijet events is given by the difference between data and the sum of all other background processes. It is the dominating process, while all other processes combined contribute on a sub-percent level. The QCD multijet estimate is simply taken as the difference between the stacked histograms and the observation. Processes labeled as $Z/t\bar{t}/VV$ $(\tau\tau)$ are estimated by the τ-embedding technique. Only statistical uncertainties are shown here.

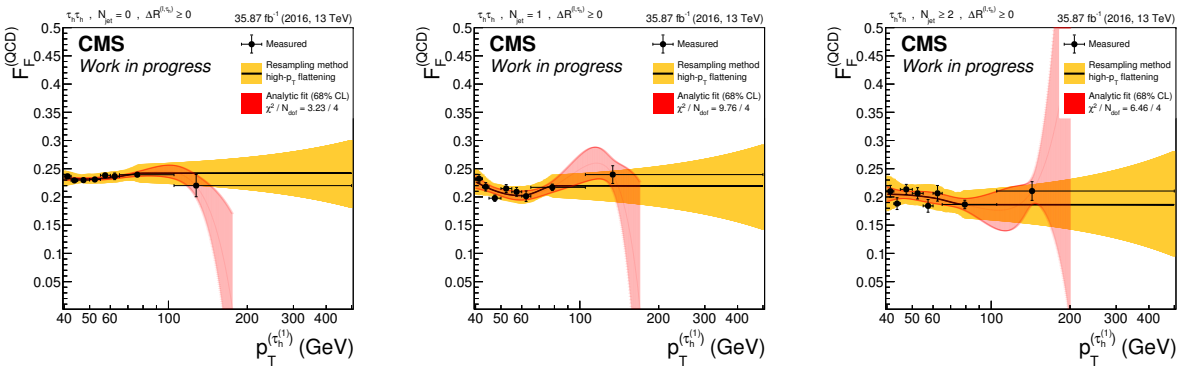


Figure C.8: The quantity $F_F^{(QCD)}$ as a function of $p_T^{(\tau_1)}$ is shown for the $\tau_h\tau_h$ channel using 2016 data. From left to right, the three N_{jets} categories are displayed – $N_{jets} = 0$, $N_{jets} = 1$ and $N_{jets} \geq 2$, respectively. The F_F parametrization is shown as a solid line. It consists of a third order polynomal fit which is truncated at high $p_T^{(\tau_1)}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid black line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

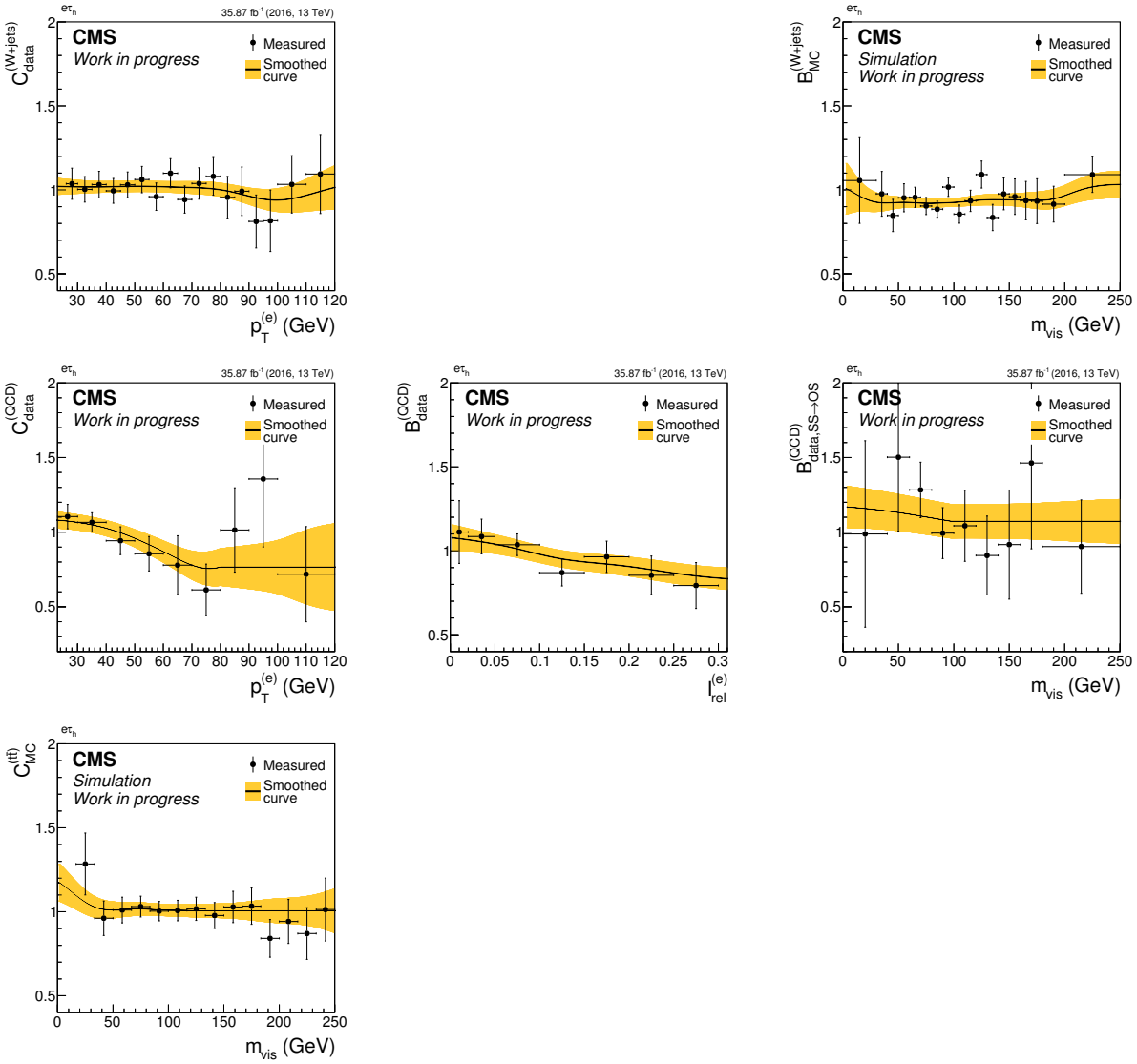


Figure C.9: Shown are all the F_{F}-related corrections in the eτ_{h} channel using 2016 data. From top to bottom, the correction applied to the raw $F_{\mathrm{F}}^{(\mathrm{W+jets})}$, $F_{\mathrm{F}}^{(\mathrm{QCD})}$ and $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ are displayed, respectively. On the left, closure corrections are shown, and in the middle and right column the bias corrections. Each correction measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

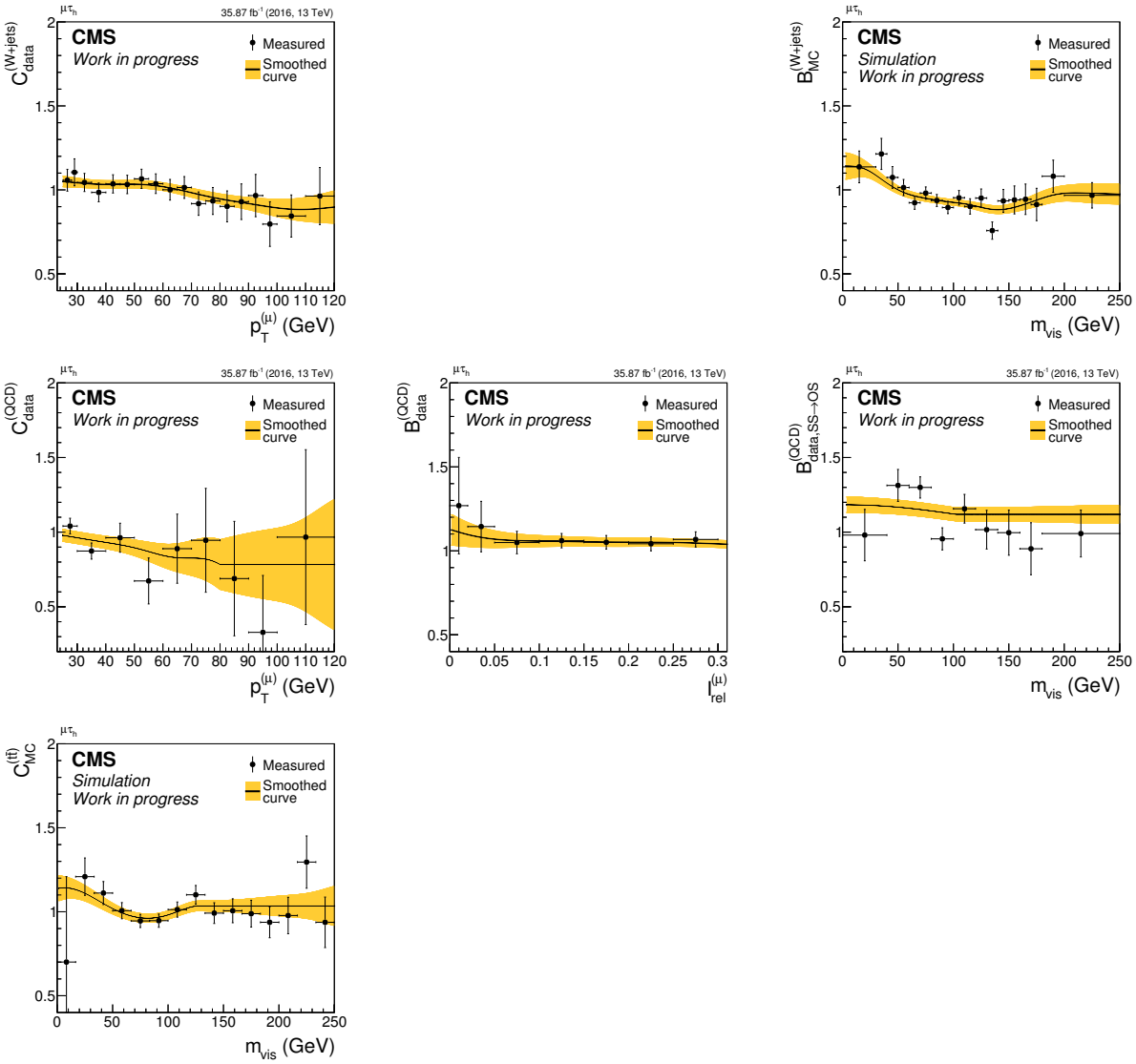


Figure C.10: Shown are all the F_{F}-related corrections in the $\mu\tau_{\mathrm{h}}$ channel using 2016 data. From top to bottom, the correction applied to the raw $F_{\mathrm{F}}^{(\mathrm{W+jets})}$, $F_{\mathrm{F}}^{(\mathrm{QCD})}$ and $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ are displayed, respectively. On the left, closure corrections are shown, and in the middle and right column the bias corrections. Each correction measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

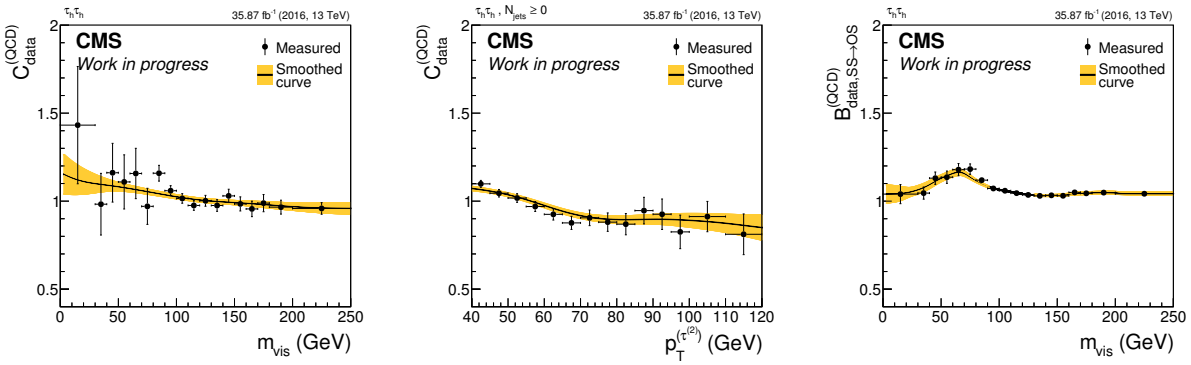


Figure C.11: Shown are all the F_{F}-related corrections in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel using 2016 data. From left to right, the two closure corrections are displayed first and then the bias correction. Each correction measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

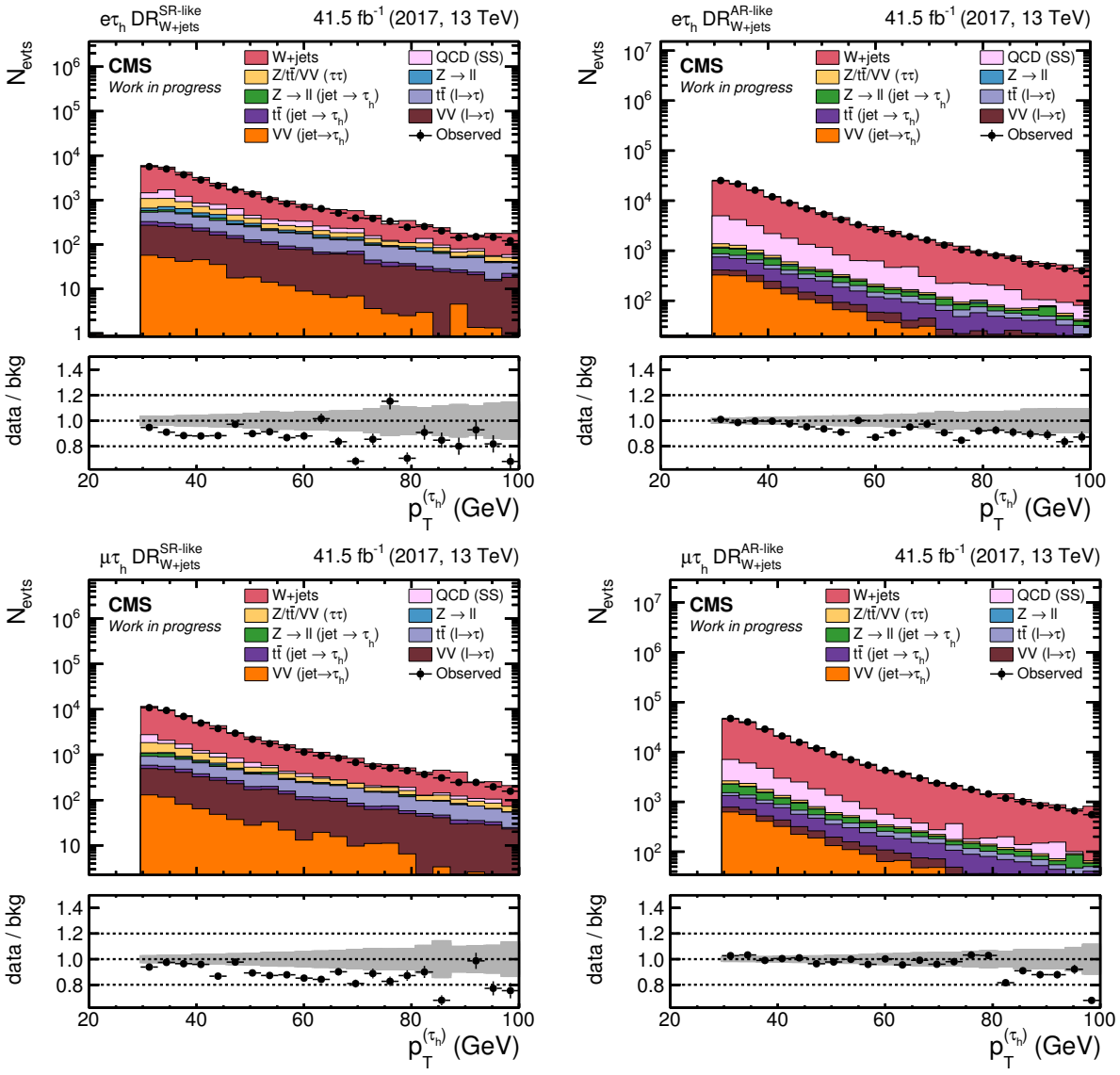


Figure C.12: $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ distributions inside $\mathrm{DR}_{\mathrm{W+jets}}$ for the 2017 data-taking period are shown. The top row shows distributions for the $\mathrm{e}\tau_{\mathrm{h}}$ channel and the bottom row for the $\mu\tau_{\mathrm{h}}$ channel. On the left the SR-like part of $\mathrm{DR}_{\mathrm{W+jets}}$ is shown and on the right the AR-like part. In all plots the W+jets process is at the top of the stacked histograms. It is the dominating process, while all other processes combined make up a few percent of the total yield. The QCD multijet estimate is taken from a SS region as detailed in the text. Processes labeled as $\mathrm{Z/t\bar{t}}/VV\,(\tau\tau)$ are estimated by the τ-embedding technique. The ratio is calculated as observed over predicted (sum of all filled histogram). The error bars in the ratio plot represent the uncertainty on the observed contribution and the gray band reflects the uncertainty on the predicted contribution. Only statistical uncertainties are shown here.

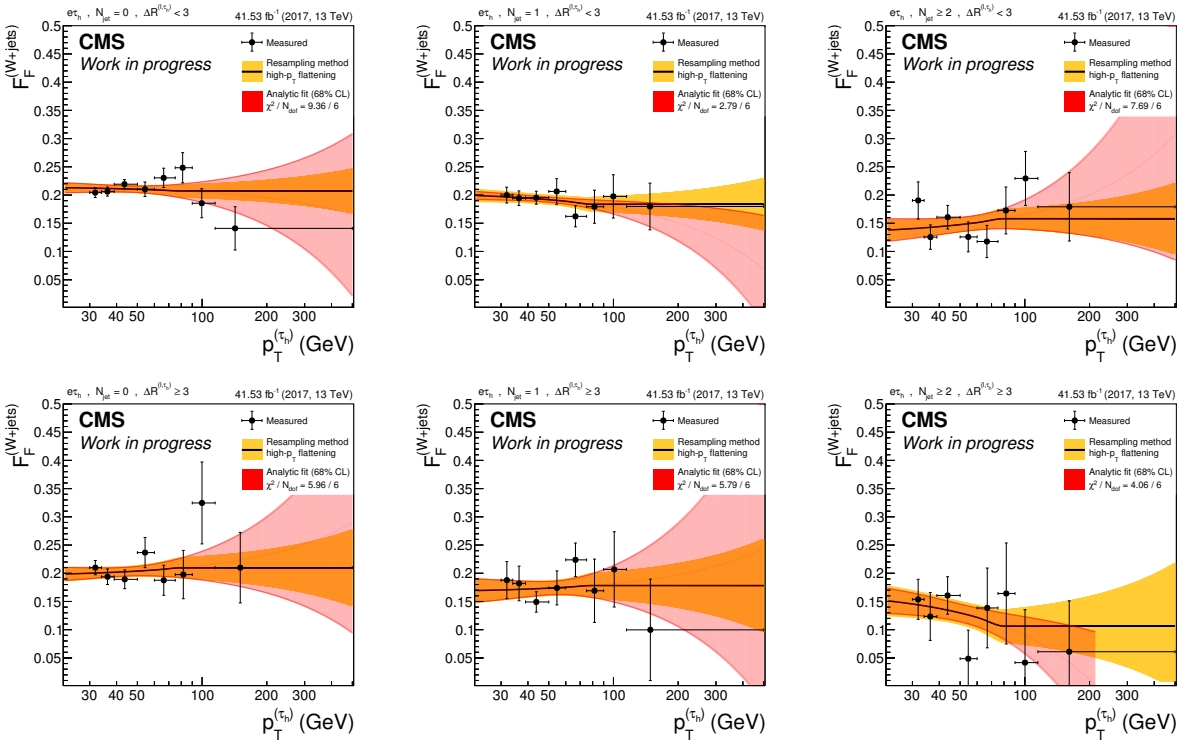


Figure C.13: The quantity $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown for the eτ_{h} channel using 2017 data. Top and bottom display the two $\Delta R^{(\ell,\tau_{\mathrm{h}})}$ categories – $\Delta R^{(\ell,\tau_{\mathrm{h}})} < 3$ and $\Delta R^{(\ell,\tau_{\mathrm{h}})} \geq 3$, respectively. From left to right, the three N_{jets} categories are displayed – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} = 1$ and $N_{\mathrm{jets}} \geq 2$, respectively. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid black line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

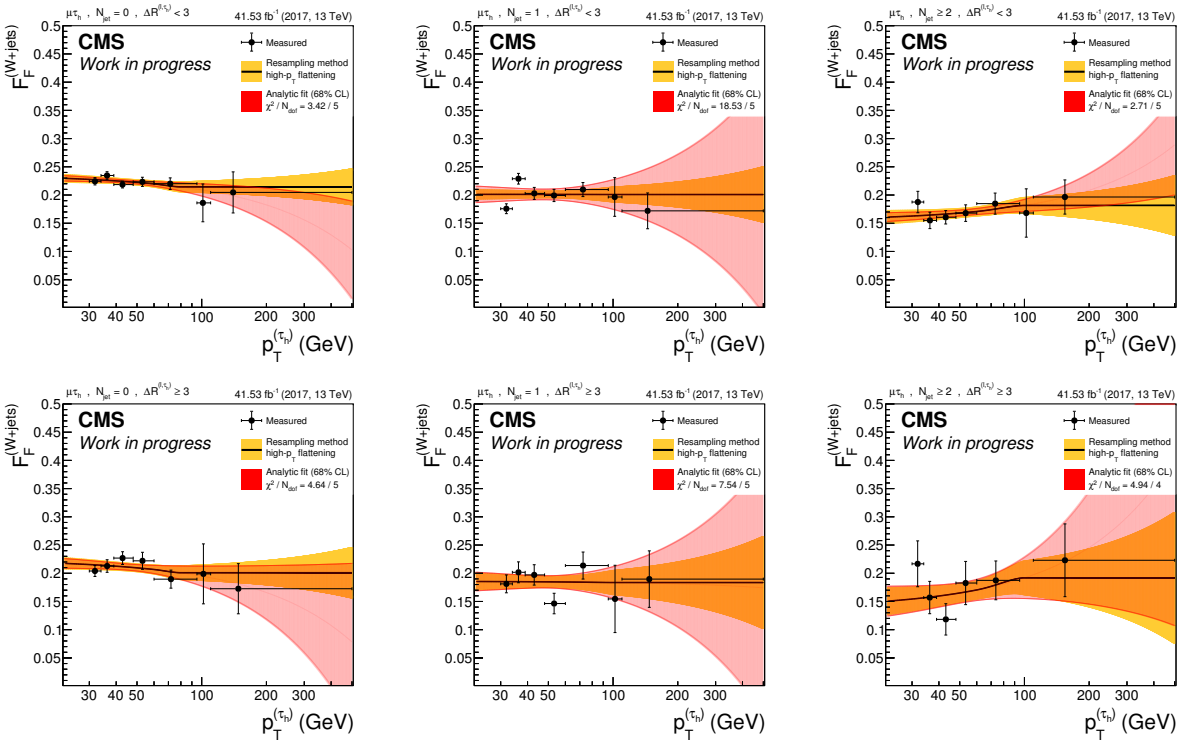


Figure C.14: The quantity $F_{\mathrm{F}}^{(\mathrm{W+jets})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown for the $\mu\tau_{\mathrm{h}}$ channel using 2017 data. Top and bottom display the two $\Delta R^{(\ell,\tau_{\mathrm{h}})}$ categories – $\Delta R^{(\ell,\tau_{\mathrm{h}})} < 3$ and $\Delta R^{(\ell,\tau_{\mathrm{h}})} \geq 3$, respectively. From left to right, the three N_{jets} categories are displayed – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} = 1$ and $N_{\mathrm{jets}} \geq 2$, respectively. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid black line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

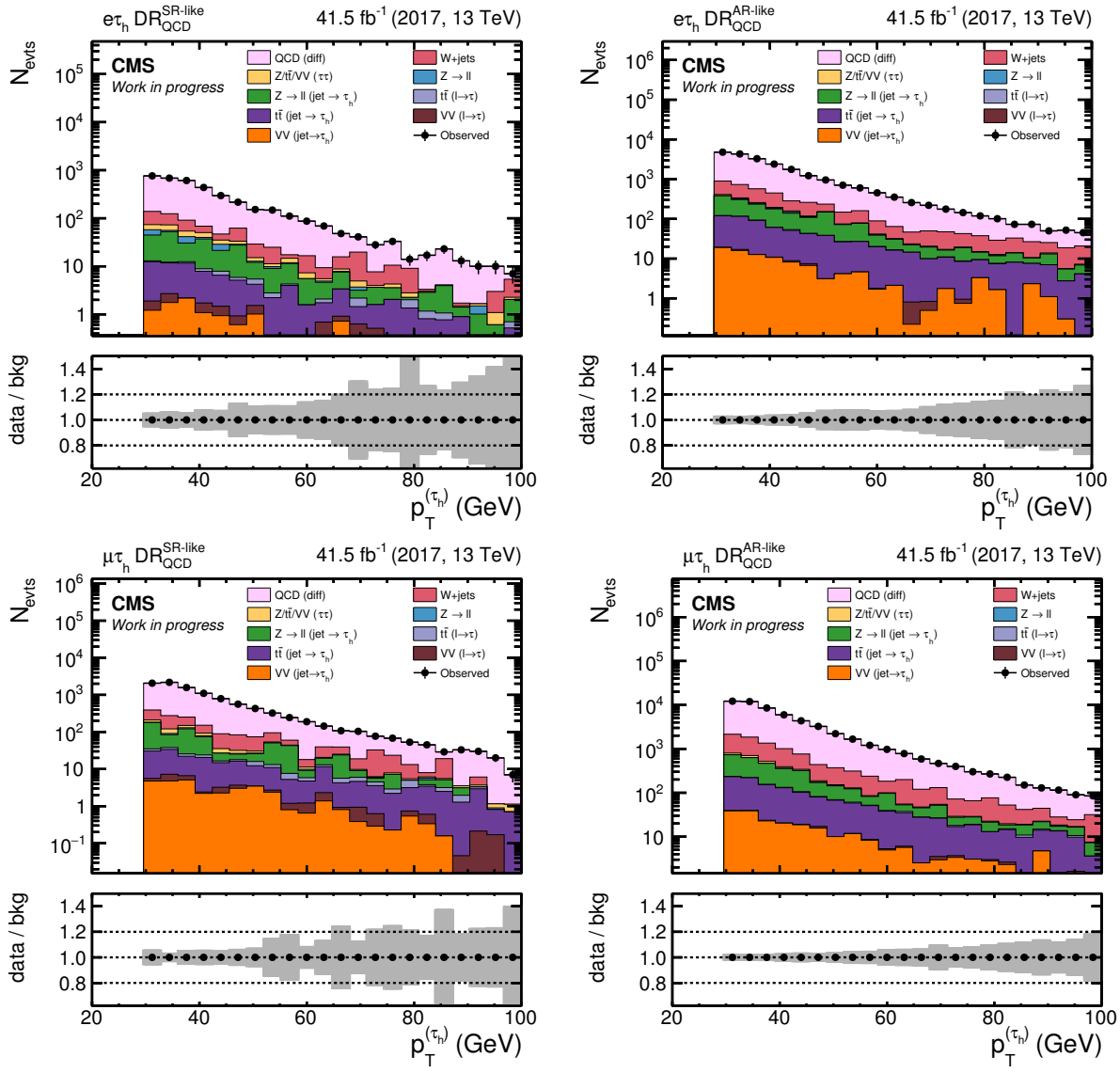


Figure C.15: $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ distributions inside $\mathrm{DR}_{\mathrm{QCD}}$ for the 2017 data-taking period are shown. The top row shows distributions for the $\mathrm{e}\tau_{\mathrm{h}}$ channel and the bottom row for the $\mu\tau_{\mathrm{h}}$ channel. On the left the SR-like part of $\mathrm{DR}_{\mathrm{QCD}}$ is shown and on the right the AR-like part. In all plots, the QCD multijet process is at the top of the stacked histograms. The estimated yield from QCD multijet events is given by the difference between data and the sum of all other background processes. It is the dominating process while all other processes combined make up at most a few percent of the total yield. The QCD multijet estimate is simply taken as the difference between the stacked histograms and the observation. Processes labeled as $\mathrm{Z/t\bar{t}}/VV\,(\tau\tau)$ are estimated by the τ-embedding technique. Only statistical uncertainties are shown here.

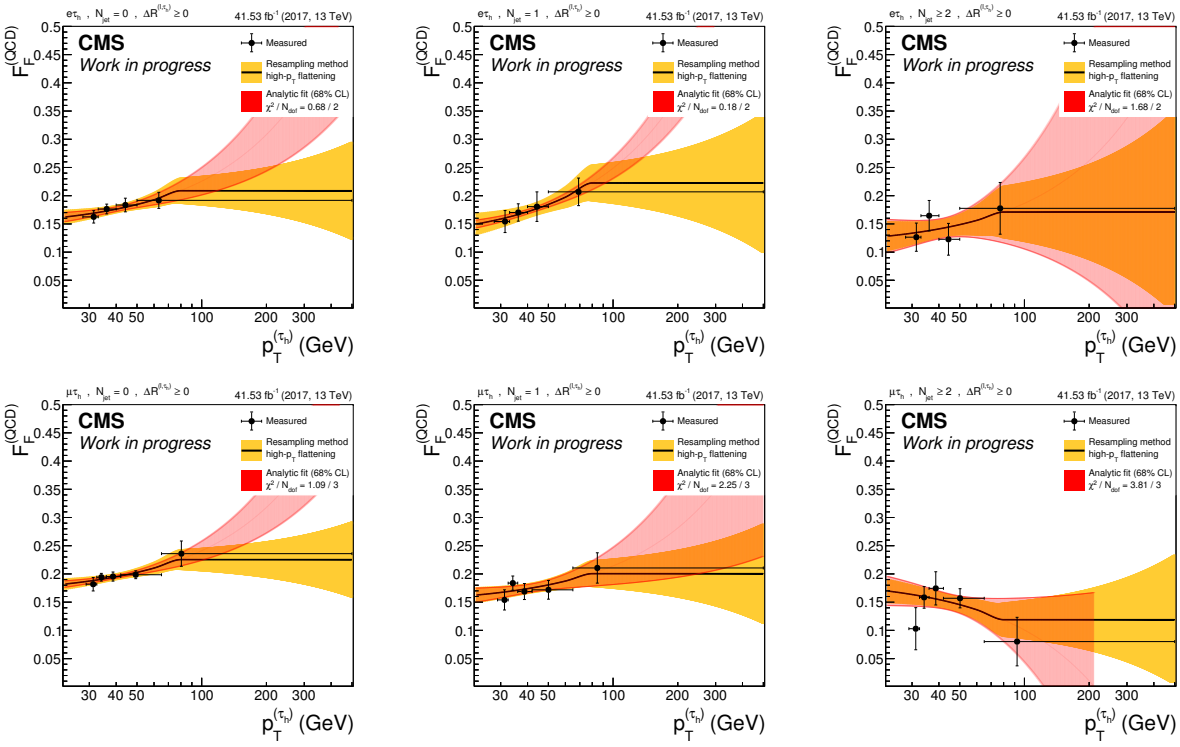


Figure C.16: The quantity $F_{\mathrm{F}}^{(\mathrm{QCD})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown using 2017 data. In the top row, distributions for the $\mathrm{e}\tau_{\mathrm{h}}$ channel are displayed and corresponding distributions for the $\mu\tau_{\mathrm{h}}$ channel are displayed on the bottom row. From left to right, the three N_{jets} categories are displayed – $N_{\mathrm{jets}} = 0$, $N_{\mathrm{jets}} = 1$ and $N_{\mathrm{jets}} \geq 2$, respectively. The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

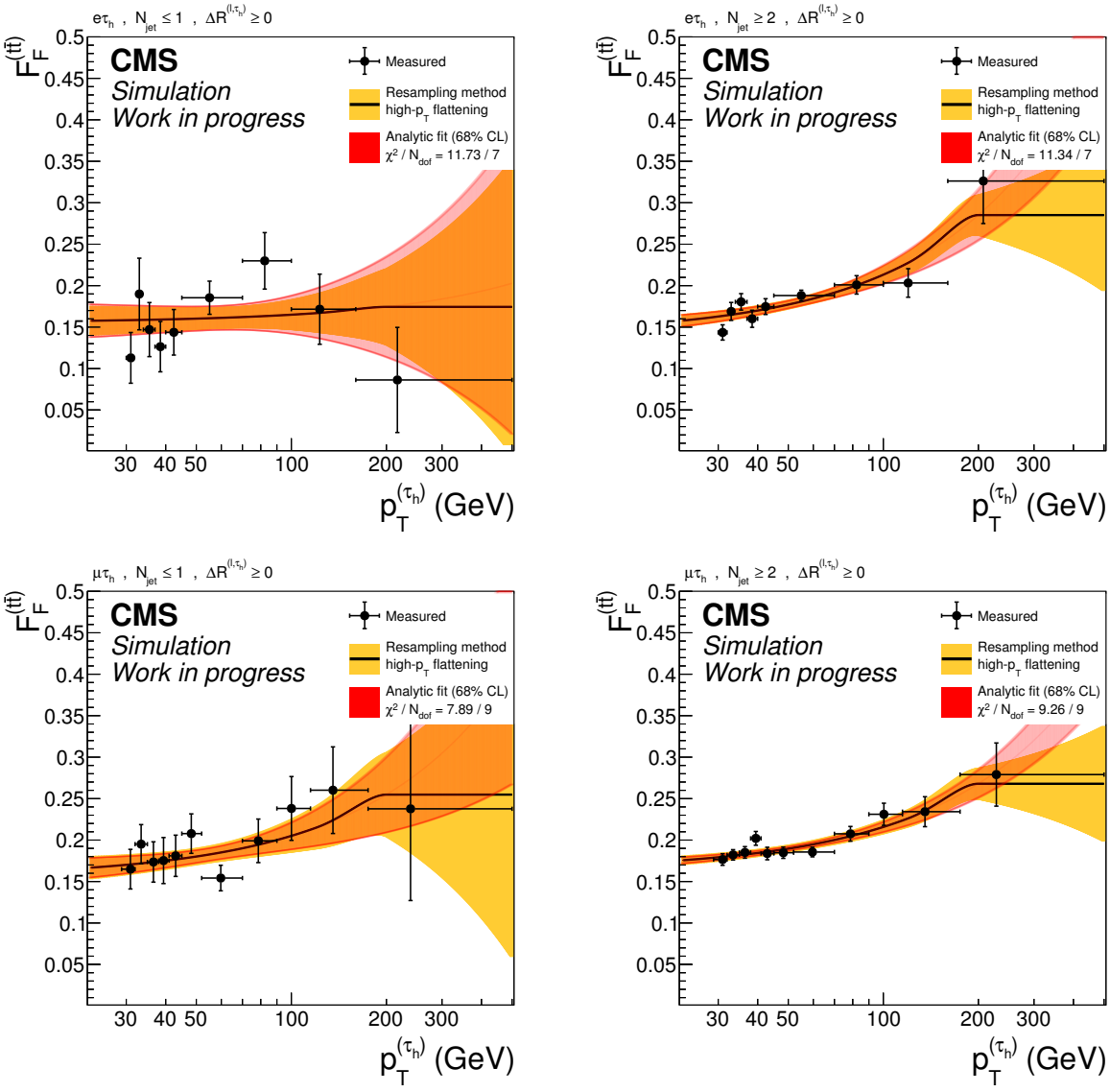


Figure C.17: The quantity $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ as a function of $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ is shown using simulated events from the 2017 data-taking period. The top row shows the distributions for the $\mathrm{e}\tau_{\mathrm{h}}$ channel and the corresponding distributions for the $\mu\tau_{\mathrm{h}}$ channel are displayed on bottom row. The measurement of $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ uses two N_{jets} categories, $N_{\mathrm{jets}} \leq 1$ (left) and $N_{\mathrm{jets}} \geq 2$ (right). The F_{F} parametrization is shown as a solid line. It consists of a linear fit which is truncated at high $p_{\mathrm{T}}^{(\tau_{\mathrm{h}})}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis the solid line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

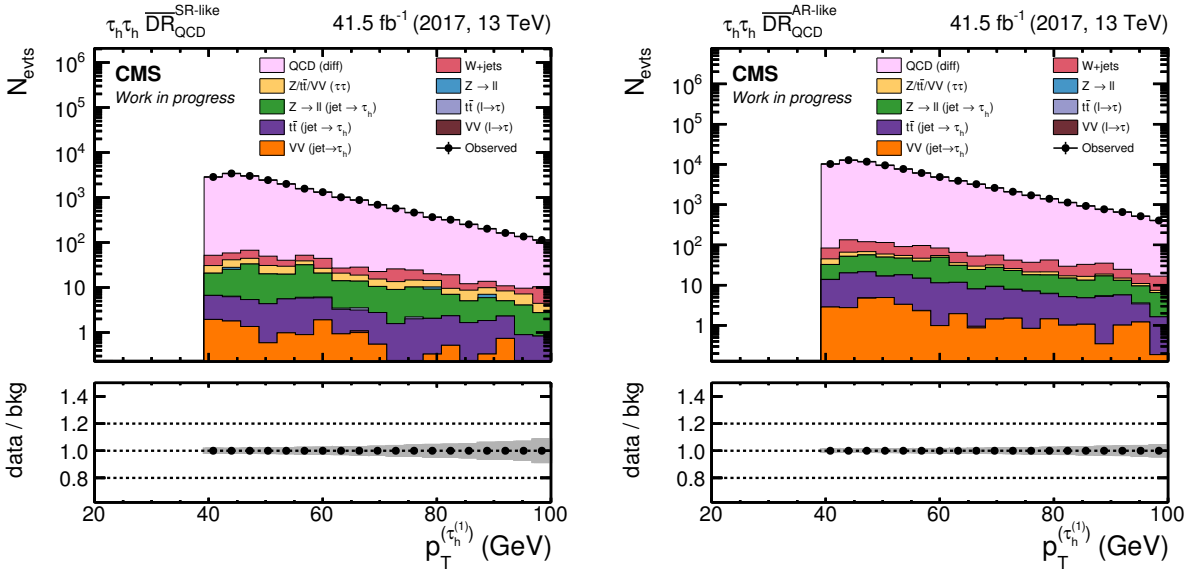


Figure C.18: Shown are $p_T^{(\tau_1)}$ distributions inside $\overline{DR}_{QCD}$ for the 2017 data-taking period. The distributions are measured for the $\tau_h\tau_h$ channel. On the left, the SR-like part of $\overline{DR}_{QCD}$ is shown and on the right, the AR-like part. In all plots, the QCD multijet process is at the top of the stacked histograms. The estimated yield from QCD multijet events is given by the difference between data and the sum of all other background processes. It is the dominating process, while all other processes combined contribute on a sub-percent level. The QCD multijet estimate is simply taken as the difference between the stacked histograms and the observation. Processes labeled as $Z/t\bar{t}/VV$ $(\tau\tau)$ are estimated by the τ-embedding technique. Only statistical uncertainties are shown here.

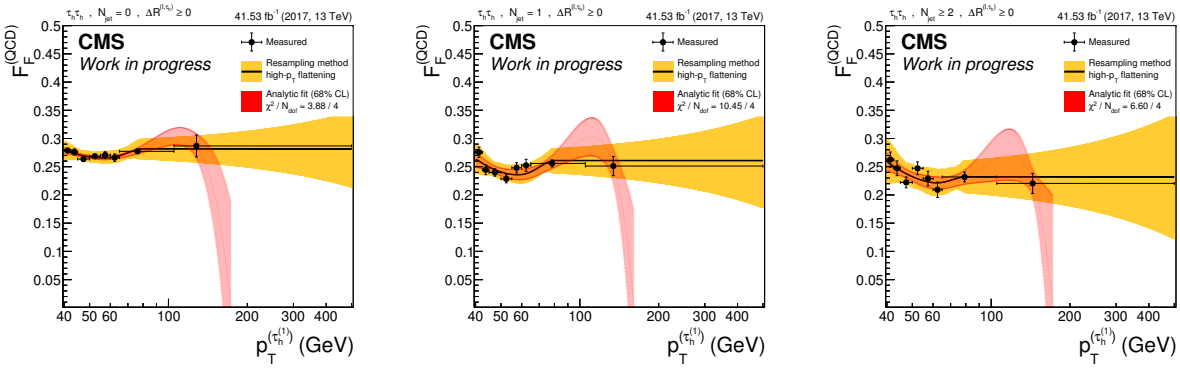


Figure C.19: The quantity $F_F^{(QCD)}$ as a function of $p_T^{(\tau_1)}$ is shown for the $\tau_h\tau_h$ channel using 2017 data. From left to right, the three N_{jets} categories are displayed – $N_{jets} = 0$, $N_{jets} = 1$ and $N_{jets} \geq 2$, respectively. The F_F parametrization is shown as a solid line. It consists of a third order polynomal fit which is truncated at high $p_T^{(\tau_1)}$ and replaced by a constant. Red uncertainty bands represent the results coming from the analytic fit. In the analysis, the solid black line is used together with its associated uncertainty band shown in yellow. The yellow uncertainty band is obtained by a resampling technique as explained in Section 8.2.1.

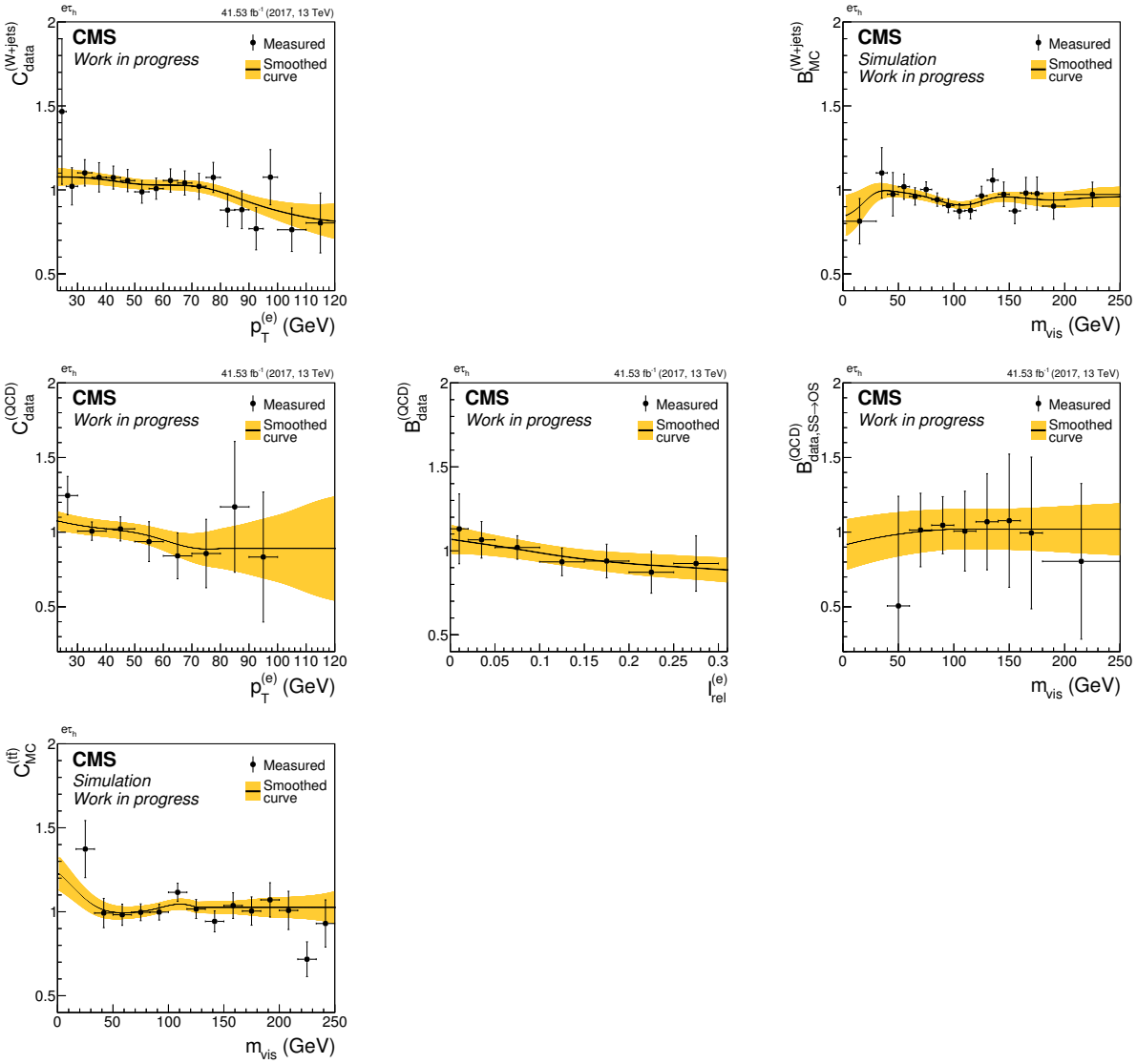


Figure C.20: Shown are all the F_{F}-related corrections in the eτ_{h} channel using 2017 data. From top to bottom, the correction applied to the raw $F_{\mathrm{F}}^{(\mathrm{W+jets})}$, $F_{\mathrm{F}}^{(\mathrm{QCD})}$ and $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ are displayed, respectively. On the left, closure corrections are shown, and in the middle and right column the bias corrections. Each correction measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

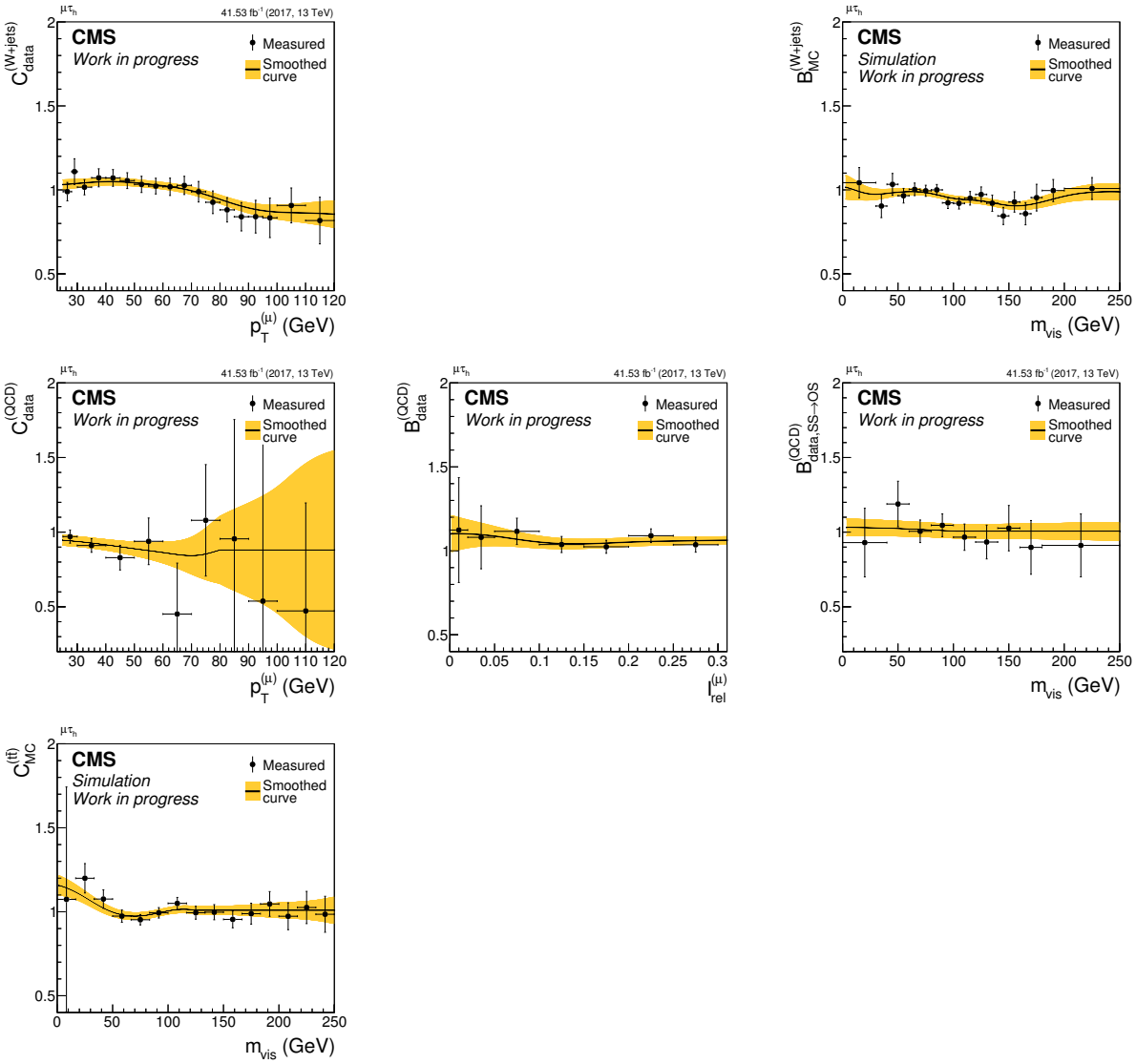


Figure C.21: Shown are all the F_{F}-related corrections in the $\mu\tau_{\mathrm{h}}$ channel using 2017 data. From top to bottom, the correction applied to the raw $F_{\mathrm{F}}^{(\mathrm{W+jets})}$, $F_{\mathrm{F}}^{(\mathrm{QCD})}$ and $F_{\mathrm{F}}^{(\mathrm{t\bar{t}})}$ are displayed, respectively. On the left, closure corrections are shown, and in the middle and right column the bias corrections. Each correction measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

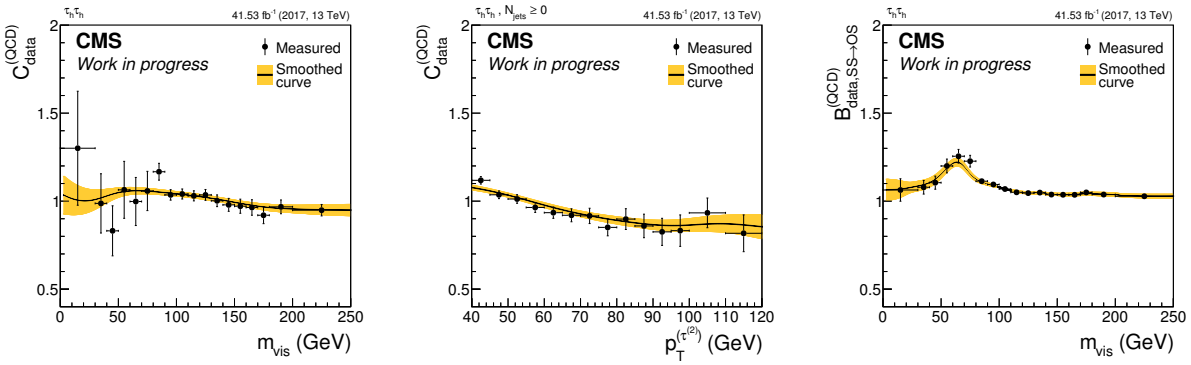


Figure C.22: Shown are all the F_{F}-related corrections in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ channel using 2017 data. From left to right, the two closure corrections are displayed first and then the bias correction. Each correction measurement is smoothed with a Gaussian kernel of variable width and the resulting smoothed curve is used later in the F_{F} application. The uncertainty band is obtained by fluctuating the measurement points and repeating the smoothing on the generated toy data.

Bibliography

[1] Luigi Di Lella and Carlo Rubbia. The Discovery of the W and Z Particles. *Adv. Ser. Direct. High Energy Phys.*, 23:137–163, 2015. doi:10.1142/9789814644150_0006.

[2] Claudio Campagnari and Melissa Franklin. The Discovery of the top quark. *Rev. Mod. Phys.*, 69:137–212, 1997. arXiv:hep-ex/9608003, doi:10.1103/RevModPhys.69.137.

[3] Georges Aad et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B*, 716:1–29, 2012. arXiv:1207.7214, doi:10.1016/j.physletb.2012.08.020.

[4] Serguei Chatrchyan et al. Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC. *Phys. Lett. B*, 716:30–61, 2012. arXiv:1207.7235, doi:10.1016/j.physletb.2012.08.021.

[5] Tatsumi Aoyama, Masashi Hayakawa, Toichiro Kinoshita, and Makiko Nio. Tenth-Order QED Contribution to the Electron g-2 and an Improved Value of the Fine Structure Constant. *Phys. Rev. Lett.*, 109:111807, 2012. arXiv:1205.5368, doi:10.1103/PhysRevLett.109.111807.

[6] Tatsumi Aoyama, M. Hayakawa, Toichiro Kinoshita, and Makiko Nio. Tenth-Order Electron Anomalous Magnetic Moment — Contribution of Diagrams without Closed Lepton Loops. *Phys. Rev. D*, 91(3):033006, 2015. [Erratum: Phys.Rev.D 96, 019901 (2017)]. arXiv:1412.8284, doi:10.1103/PhysRevD.91.033006.

[7] Measurements of Higgs boson production in the decay channel with a pair of τ leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV. 4 2022. arXiv:2204.12957.

[8] Armen Tumasyan et al. Search for a heavy Higgs boson decaying into two lighter Higgs bosons in the $\tau\tau$bb final state at 13 TeV. *JHEP*, 11:057, 2021. arXiv:2106.10361, doi:10.1007/JHEP11(2021)057.

[9] CMS Collaboration. Searches for additional Higgs bosons and vector-like leptoquarks in $\tau\tau$ final states in proton-proton collisions at $\sqrt{s} = 13$ TeV. 2022. URL: http://cds.cern.ch/record/2803739.

[10] Artur Gottmann. Global Interpretation of $\tau\tau$ Events in the Context of the Standard Model and Beyond, 2020. Presented 12 Jun 2020. URL: https://cds.cern.ch/record/2742868.

[11] Janek Bechtel. *A novel search for di-Higgs events in the $\tau^-\tau^+ + b\bar{b}$ final state in pp collisions at 13 TeV at the LHC*. PhD thesis, KIT, Karlsruhe, 2021. URL: https://publish.etp.kit.edu/record/22034.

[12] Sebastian Wozniewski. *Differential cross section measurements in the $H \to \tau\tau$ decay channel with CMS data of proton-proton collisions at the Large Hadron Collider at CERN*. PhD thesis, KIT, Karlsruhe, 2021. URL: https://publish.etp.kit.edu/record/22028.

[13] Markus Spanring. Implementation and Application of Machine Learning Techniques for the Analysis of Higgs Boson Decays to Tau Leptons with the CMS experiment. *TU Wien Bibliothekssystem*, 2019, 2019. doi:10.34726/hss.2019.50244.

[14] Johannes Brandstetter. Neutral Higgs Boson and Z Boson Decays into Pairs of Tau Leptons with the CMS Detector, Mar 2018. Presented 25 May 2018. URL: https://cds.cern.ch/record/2622431.

[15] Navid K Rad. Search for supersymmetry partners of the top quark in models with compressed mass spectra with the CMS detector. *TU Wien Bibliothekssystem*, 2018, 2018. doi:10.34726/hss.2018.56980.

[16] Courtesy to Wikipedia: "Standard Model of Elementary Particles" by MissMJ-Own work by uploader, PBS NOVA, Fermilab, Office of Science, United States Department of Energy, Particle Data Group. Particle content of the Standard Model. [Online; accessed 20-July-2021].

[17] Roel Aaij et al. Observation of the resonant character of the $Z(4430)^-$ state. *Phys. Rev. Lett.*, 112(22):222002, 2014. arXiv:1404.1903, doi:10.1103/PhysRevLett.112.222002.

[18] Roel Aaij et al. Observation of $J/\psi p$ Resonances Consistent with Pentaquark States in $\Lambda_b^0 \to J/\psi K^- p$ Decays. *Phys. Rev. Lett.*, 115:072001, 2015. arXiv:1507.03414, doi:10.1103/PhysRevLett.115.072001.

[19] Pascal Paganini. Physique des particules avancée: An introduction to the Standard Model of Particle Physics. 2020. Lectures presented at Ecole Polytechnique, Palaiseau (France).

[20] M. J. Herrero. The standard model. 1998. arXiv:hep-ph/9812242.

[21] Günther Dissertori, Ian G Knowles, and Michael Schmelling. *Quantum chromodynamics: high energy experiments and theory*, volume 115. Oxford University Press, 2003.

[22] Howard Georgi. *Lie Algebras In Particle Physics : from Isospin To Unified Theories*. Taylor & Francis, Boca Raton, 2000. doi:10.1201/9780429499210.

[23] Matthew D. Schwartz. *Quantum Field Theory and the Standard Model*. Cambridge University Press, 3 2014.

[24] S. L. Glashow. Partial Symmetries of Weak Interactions. *Nucl. Phys.*, 22:579–588, 1961. doi:10.1016/0029-5582(61)90469-2.

[25] Steven Weinberg. A Model of Leptons. *Phys. Rev. Lett.*, 19:1264–1266, 1967. doi:10.1103/PhysRevLett.19.1264.

[26] Abdus Salam and John Clive Ward. Electromagnetic and weak interactions. *Phys. Lett.*, 13:168–171, 1964. doi:10.1016/0031-9163(64)90711-5.

[27] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes, and R. P. Hudson. Experimental Test of Parity Conservation in β Decay. *Phys. Rev.*, 105:1413–1414, 1957. doi:10.1103/PhysRev.105.1413.

[28] J. H. Christenson, J. W. Cronin, V. L. Fitch, and R. Turlay. Evidence for the 2π Decay of the K_2^0 Meson. *Phys. Rev. Lett.*, 13:138–140, 1964. doi:10.1103/PhysRevLett.13.138.

[29] F. J. Hasert et al. Observation of Neutrino Like Interactions Without Muon Or Electron in the Gargamelle Neutrino Experiment. *Phys. Lett. B*, 46:138–140, 1973. doi:10.1016/0370-2693(73)90499-1.

[30] F. J. Hasert et al. Search for Elastic ν_μ Electron Scattering. *Phys. Lett. B*, 46:121–124, 1973. doi:10.1016/0370-2693(73)90494-2.

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

[32] Peter W. Higgs. Broken Symmetries and the Masses of Gauge Bosons. *Phys. Rev. Lett.*, 13:508–509, 1964. doi:10.1103/PhysRevLett.13.508.

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

[34] John Ellis. Higgs Physics. In *2013 European School of High-Energy Physics*, pages 117–168, 2015. URL: http://cds.cern.ch/record/1638469, arXiv:1312.5672, doi:10.5170/CERN-2015-004.117.

[35] J. Goldstone. Field Theories with Superconductor Solutions. *Nuovo Cim.*, 19:154–164, 1961. URL: http://cds.cern.ch/record/343400, doi:10.1007/BF02812722.

[36] Nicola Cabibbo. Unitary Symmetry and Leptonic Decays. *Phys. Rev. Lett.*, 10:531–533, 1963. doi:10.1103/PhysRevLett.10.531.

[37] Makoto Kobayashi and Toshihide Maskawa. CP Violation in the Renormalizable Theory of Weak Interaction. *Prog. Theor. Phys.*, 49:652–657, 1973. doi:10.1143/PTP.49.652.

[38] Ziro Maki, Masami Nakagawa, and Shoichi Sakata. Remarks on the unified model of elementary particles. *Prog. Theor. Phys.*, 28:870–880, 1962. doi:10.1143/PTP.28.870.

[39] G. Arnison et al. Experimental Observation of Isolated Large Transverse Energy Electrons with Associated Missing Energy at $\sqrt{s} = 540$ GeV. *Phys. Lett. B*, 122:103–116, 1983. doi:10.1016/0370-2693(83)91177-2.

[40] M. Banner et al. Observation of Single Isolated Electrons of High Transverse Momentum in Events with Missing Transverse Energy at the CERN anti-p p Collider. *Phys. Lett. B*, 122:476–485, 1983. doi:10.1016/0370-2693(83)91605-2.

[41] Dieter Haidt. The Discovery of Weak Neutral Currents. *Adv. Ser. Direct. High Energy Phys.*, 23:165–183, 2015. doi:10.1142/9789814644150_0007.

[42] S. Abachi et al. Observation of the top quark. *Phys. Rev. Lett.*, 74:2632–2637, 1995. arXiv:hep-ex/9503003, doi:10.1103/PhysRevLett.74.2632.

[43] ATLAS Collaboration. Standard Model Production Cross Section Measurements. July 2018. [Online; accessed 05-October-2021].

[44] Yoshiaki Sofue and Vera Rubin. Rotation curves of spiral galaxies. *Ann. Rev. Astron. Astrophys.*, 39:137–174, 2001. arXiv:astro-ph/0010594, doi:10.1146/annurev.astro.39.1.137.

[45] N. Aghanim et al. Planck 2018 results. VI. Cosmological parameters. *Astron. Astrophys.*, 641:A6, 2020. [Erratum: Astron.Astrophys. 652, C4 (2021)]. arXiv:1807.06209, doi:10.1051/0004-6361/201833910.

[46] Zoltan Ligeti. The CKM matrix and CP violation. *Int. J. Mod. Phys. A*, 20:5105–5118, 2005. arXiv:hep-ph/0408267, doi:10.1142/S0217751X05028624.

[47] Ivan Esteban, M. C. Gonzalez-Garcia, Michele Maltoni, Thomas Schwetz, and Albert Zhou. The fate of hints: updated global analysis of three-flavor neutrino oscillations. *JHEP*, 09:178, 2020. arXiv:2007.14792, doi:10.1007/JHEP09(2020)178.

[48] C. L. Bennett et al. First year Wilkinson Microwave Anisotropy Probe (WMAP) observations: Preliminary maps and basic results. *Astrophys. J. Suppl.*, 148:1–27, 2003. arXiv:astro-ph/0302207, doi:10.1086/377253.

[49] M. Tanabashi et al. Review of Particle Physics. *Phys. Rev. D*, 98(3):030001, 2018. doi:10.1103/PhysRevD.98.030001.

[50] James D. Wells. Higgs naturalness and the scalar boson proliferation instability problem. *Synthese*, 194(2):477–490, 2017. arXiv:1603.06131, doi:10.1007/s11229-014-0618-8.

[51] David E. Morrissey, Tilman Plehn, and Tim M. P. Tait. Physics searches at the LHC. *Phys. Rept.*, 515:1–113, 2012. arXiv:0912.3259, doi:10.1016/j.physrep.2012.02.007.

[52] J. Wess and B. Zumino. Supergauge Transformations in Four-Dimensions. *Nucl. Phys. B*, 70:39–50, 1974. doi:10.1016/0550-3213(74)90355-1.

[53] P. Van Nieuwenhuizen. Supergravity. *Phys. Rept.*, 68:189–398, 1981. doi:10.1016/0370-1573(81)90157-5.

[54] Stephen P. Martin. A Supersymmetry primer. *Adv. Ser. Direct. High Energy Phys.*, 18:1–98, 1998. arXiv:hep-ph/9709356, doi:10.1142/9789812839657_0001.

[55] Rudolf Haag, Jan T. Lopuszanski, and Martin Sohnius. All Possible Generators of Supersymmetries of the s Matrix. *Nucl. Phys. B*, 88:257, 1975. doi:10.1016/0550-3213(75)90279-5.

[56] T. D. Lee. A Theory of Spontaneous T Violation. *Phys. Rev. D*, 8:1226–1239, 1973. doi:10.1103/PhysRevD.8.1226.

[57] G. C. Branco, P. M. Ferreira, L. Lavoura, M. N. Rebelo, Marc Sher, and Joao P. Silva. Theory and phenomenology of two-Higgs-doublet models. *Phys. Rept.*, 516:1–102, 2012. arXiv:1106.0034, doi:10.1016/j.physrep.2012.02.002.

[58] H. Nishino et al. Search for Proton Decay via p —> e+ pi0 and p —> mu+ pi0 in a Large Water Cherenkov Detector. *Phys. Rev. Lett.*, 102:141801, 2009. arXiv:0903.0676, doi:10.1103/PhysRevLett.102.141801.

[59] Emanuele Bagnaschi et al. MSSM Higgs Boson Searches at the LHC: Benchmark Scenarios for Run 2 and Beyond. *Eur. Phys. J. C*, 79(7):617, 2019. arXiv:1808.07542, doi:10.1140/epjc/s10052-019-7114-8.

[60] Henning Bahl, Stefan Liebler, and Tim Stefaniak. MSSM Higgs benchmark scenarios for Run 2 and beyond: the low tan β region. *Eur. Phys. J. C*, 79(3):279, 2019. arXiv:1901.05933, doi:10.1140/epjc/s10052-019-6770-z.

[61] Mayumi Aoki, Shinya Kanemura, Koji Tsumura, and Kei Yagyu. Models of Yukawa interaction in the two Higgs doublet model, and their collider phenomenology. *Phys. Rev. D*, 80:015017, 2009. arXiv:0902.4665, doi:10.1103/PhysRevD.80.015017.

[62] Ulrich Ellwanger, Cyril Hugonie, and Ana M. Teixeira. The Next-to-Minimal Supersymmetric Standard Model. *Phys. Rept.*, 496:1–77, 2010. arXiv:0910.1785, doi:10.1016/j.physrep.2010.07.001.

[63] M. Maniatis. The Next-to-Minimal Supersymmetric extension of the Standard Model reviewed. *Int. J. Mod. Phys. A*, 25:3505–3602, 2010. arXiv:0906.0777, doi:10.1142/S0217751X10049827.

[64] J. A. Casas and C. Munoz. A Natural solution to the mu problem. *Phys. Lett. B*, 306:288–294, 1993. arXiv:hep-ph/9302227, doi:10.1016/0370-2693(93)90081-R.

[65] C. Balazs, Marcela Carena, and C. E. M. Wagner. Dark matter, light stops and electroweak baryogenesis. *Phys. Rev. D*, 70:015007, 2004. arXiv:hep-ph/0403224, doi:10.1103/PhysRevD.70.015007.

[66] H. Murayama and T. Yanagida. A viable SU(5) GUT with light leptoquark bosons. *Mod. Phys. Lett. A*, 7:147–152, 1992. doi:10.1142/S0217732392000070.

[67] Goran Senjanovic and Aleksandar Sokorac. Light Leptoquarks in SO(10). *Z. Phys. C*, 20:255, 1983. doi:10.1007/BF01574858.

[68] Paul H. Frampton and Bum-Hoon Lee. SU(15) GRAND UNIFICATION. *Phys. Rev. Lett.*, 64:619, 1990. doi:10.1103/PhysRevLett.64.619.

[69] P. Yu. Popov, A. V. Povarov, and A. D. Smirnov. Fermionic decays of scalar leptoquarks and scalar gluons in the minimal four color symmetry model. *Mod. Phys. Lett. A*, 20:3003–3012, 2005. arXiv:hep-ph/0511149, doi:10.1142/S0217732305019109.

[70] M. Heyssler, R. Ruckl, and H. Spiesberger. Leptoquark and R-parity violating SUSY processes. In *4th International Workshop on Linear Colliders (LCWS 99)*, pages 559–564, 4 1999. arXiv:hep-ph/9908319.

[71] Savas Dimopoulos. Technicolored Signatures. *Nucl. Phys. B*, 168:69–92, 1980. doi:10.1016/0550-3213(80)90277-1.

[72] Edward Farhi and Leonard Susskind. Technicolor. *Phys. Rept.*, 74:277, 1981. doi:10.1016/0370-1573(81)90173-3.

[73] Barbara Schrempp and Fridger Schrempp. LIGHT LEPTOQUARKS. *Phys. Lett. B*, 153:101–107, 1985. doi:10.1016/0370-2693(85)91450-9.

[74] Minoru Tanaka and Ryoutaro Watanabe. New physics in the weak interaction of $\bar{B} \to D^{(*)}\tau\bar{\nu}$. *Phys. Rev. D*, 87(3):034028, 2013. arXiv:1212.1878, doi:10.1103/PhysRevD.87.034028.

[75] Yasuhito Sakaki, Minoru Tanaka, Andrey Tayduganov, and Ryoutaro Watanabe. Testing leptoquark models in $\bar{B} \to D^{(*)}\tau\bar{\nu}$. *Phys. Rev. D*, 88(9):094012, 2013. arXiv:1309.0301, doi:10.1103/PhysRevD.88.094012.

[76] Ilja Doršner, Svjetlana Fajfer, Nejc Košnik, and Ivan Nišandžić. Minimally flavored colored scalar in $\bar{B} \to D^{(*)}\tau\bar{\nu}$ and the mass matrices constraints. *JHEP*, 11:084, 2013. arXiv:1306.6493, doi:10.1007/JHEP11(2013)084.

[77] Ben Gripaios, Marco Nardecchia, and S. A. Renner. Composite leptoquarks and anomalies in B-meson decays. *JHEP*, 05:006, 2015. arXiv:1412.1791, doi:10.1007/JHEP05(2015)006.

[78] Béranger Dumont, Kenji Nishiwaki, and Ryoutaro Watanabe. LHC constraints and prospects for S_1 scalar leptoquark explaining the $\bar{B} \to D^{(*)}\tau\bar{\nu}$ anomaly. *Phys. Rev. D*, 94(3):034001, 2016. arXiv:1603.05248, doi:10.1103/PhysRevD.94.034001.

[79] I. Doršner, S. Fajfer, A. Greljo, J. F. Kamenik, and N. Košnik. Physics of leptoquarks in precision experiments and at particle colliders. *Phys. Rept.*, 641:1–68, 2016. arXiv:1603.04993, doi:10.1016/j.physrep.2016.06.001.

[80] Deepak Kar. Experimental particle physics. IOP Publishing, 2019. URL: http://dx.doi.org/10.1088/2053-2563/ab1be6ch3, doi:10.1088/2053-2563/ab1be6ch3.

[81] Arpad Horvath. Drawn by Arpad Horvath with Inkscape. This W3C-unspecified vector image was created with Inkscape. CC BY-SA 2.5. [Online; accessed 16-July-2021].

[82] E. Rutherford. The scattering of alpha and beta particles by matter and the structure of the atom. *Phil. Mag. Ser. 6*, 21:669–688, 1911. doi:10.1080/14786440508637080.

[83] MG Holloway and CP Baker. How the barn was born. *Physics Today*, 25(7):9, 1972.

[84] W. Herr and B. Muratori. Concept of luminosity. In *CERN Accelerator School and DESY Zeuthen: Accelerator Physics*, pages 361–377, 9 2003. URL: http://cdsweb.cern.ch/record/603056/.

[85] G. Antchev et al. First measurement of elastic, inelastic and total cross-section at $\sqrt{s} = 13$ TeV by TOTEM and overview of cross-section data at LHC energies. *Eur. Phys. J. C*, 79(2):103, 2019. arXiv:1712.06153, doi:10.1140/epjc/s10052-019-6567-0.

[86] CMS Collaboration. Lumi Public Results - multi year plots. [Online; accessed 28-August-2021].

[87] I Neutelings. How to draw diagrams in LaTeX with TikZ, 2020. [Online; accessed 24-August-2021].

[88] G. L. Bayatian et al. CMS Physics: Technical Design Report Volume 1: Detector Performance and Software. 2006. URL: https://cds.cern.ch/record/922757.

[89] CMS Collaboration. Technical proposal for the upgrade of the CMS detector through 2020. Technical report, Jun 2011. URL: http://cds.cern.ch/record/1355706.

[90] CMS Technical Design Report for the Pixel Detector Upgrade. 9 2012. doi:10.2172/1151650.

[91] CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter. 9 2012. doi:10.2172/1151651.

[92] CMS Technical Design Report for the Level-1 Trigger Upgrade. 6 2013. URL: https://cds.cern.ch/record/1556311.

[93] Tai Sakuma. Cutaway diagrams of CMS detector. May 2019. URL: http://cds.cern.ch/record/2665537.

[94] 2017 tracking performance plots. Apr 2017. URL: https://cds.cern.ch/record/2290524.

[95] Muon tag and probe efficiency on 2016 data. Jul 2016. URL: https://cds.cern.ch/record/2203523.

[96] Muon Tracking Efficiency using Tag and Probe method for 2017 dataset. Mar 2019. URL: https://cds.cern.ch/record/2666648.

[97] Serguei Chatrchyan et al. Description and performance of track and primary-vertex reconstruction with the CMS tracker. *JINST*, 9(10):P10009, 2014. arXiv:1405.6569, doi:10.1088/1748-0221/9/10/P10009.

[98] CMS Technical Design Report for the Pixel Detector Upgrade. 9 2012. doi:10.2172/1151650.

[99] Serguei Chatrchyan et al. Description and performance of track and primary-vertex reconstruction with the CMS tracker. *JINST*, 9(10):P10009, 2014. arXiv:1405.6569, doi:10.1088/1748-0221/9/10/P10009.

[100] CMS collaboration. CMS Tracker Detector Performance Public Results. [Online; accessed 08-August-2021].

[101] Paolo Azzurri. The CMS silicon strip tracker. *J. Phys. Conf. Ser.*, 41:127–134, 2006. arXiv:physics/0512097, doi:10.1088/1742-6596/41/1/011.

[102] S. Chatrchyan et al. The CMS Experiment at the CERN LHC. *JINST*, 3:S08004, 2008. doi:10.1088/1748-0221/3/08/S08004.

[103] S. Chatrchyan et al. The CMS Experiment at the CERN LHC. *JINST*, 3:S08004, 2008. doi:10.1088/1748-0221/3/08/S08004.

[104] V. I. Klyukhin et al. The CMS Magnetic Field Map Performance. *IEEE Trans. Appl. Supercond.*, 20(3):152–155, 2010. arXiv:1110.0607, doi:10.1109/TASC.2010.2041200.

[105] CMS: The TriDAS project. Technical design report, Vol. 2: Data acquisition and high-level trigger. 12 2002. URL: https://cds.cern.ch/record/578006.

[106] Stefan Weinzierl. Introduction to Monte Carlo methods. 6 2000. arXiv:hep-ph/0006269.

[107] John C. Collins and Davison E. Soper. The Theorems of Perturbative QCD. *Ann. Rev. Nucl. Part. Sci.*, 37:383–409, 1987. doi:10.1146/annurev.ns.37.120187.002123.

[108] Frank Siegert. *Monte-Carlo event generation for the LHC*. PhD thesis, Durham U., 2010. URL: https://cds.cern.ch/record/1600005/.

[109] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H. S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The automated computation of tree-level and next-to-leading order differential cross sections, and their matching to parton shower simulations. *JHEP*, 07:079, 2014. arXiv:1405.0301, doi:10.1007/JHEP07(2014)079.

[110] Stefano Frixione, Paolo Nason, and Carlo Oleari. Matching NLO QCD computations with Parton Shower simulations: the POWHEG method. *JHEP*, 11:070, 2007. arXiv:0709.2092, doi:10.1088/1126-6708/2007/11/070.

[111] T. Affolder et al. Charged Jet Evolution and the Underlying Event in $p\bar{p}$ Collisions at 1.8 TeV. *Phys. Rev. D*, 65:092002, 2002. doi:10.1103/PhysRevD.65.092002.

[112] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. A Brief Introduction to PYTHIA 8.1. *Comput. Phys. Commun.*, 178:852–867, 2008. arXiv:0710.3820, doi:10.1016/j.cpc.2008.01.036.

[113] Torbjörn Sjöstrand, Stefan Ask, Jesper R. Christiansen, Richard Corke, Nishita Desai, Philip Ilten, Stephen Mrenna, Stefan Prestel, Christine O. Rasmussen, and Peter Z. Skands. An introduction to PYTHIA 8.2. *Comput. Phys. Commun.*, 191:159–177, 2015. arXiv:1410.3012, doi:10.1016/j.cpc.2015.01.024.

[114] S. Jadach, Z. Was, R. Decker, and Johann H. Kuhn. The tau decay library TAUOLA: Version 2.4. *Comput. Phys. Commun.*, 76:361–380, 1993. doi:10.1016/0010-4655(93)90061-G.

[115] S. Agostinelli et al. GEANT4–a simulation toolkit. *Nucl. Instrum. Meth. A*, 506:250–303, 2003. doi:10.1016/S0168-9002(03)01368-8.

[116] R. Fruhwirth. Application of Kalman filtering to track and vertex fitting. *Nucl. Instrum. Meth. A*, 262:444–450, 1987. doi:10.1016/0168-9002(87)90887-4.

[117] R. Fruhwirth. Application of Kalman filtering to track and vertex fitting. *Nucl. Instrum. Meth. A*, 262:444–450, 1987. doi:10.1016/0168-9002(87)90887-4.

[118] K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *IEEE Proc.*, 86(11):2210–2239, 1998. doi:10.1109/5.726788.

[119] R. Fruhwirth, W. Waltenberger, and P. Vanlaer. Adaptive vertex fitting. *J. Phys. G*, 34:N343, 2007. doi:10.1088/0954-3899/34/12/N01.

[120] Serguei Chatrchyan et al. The Performance of the CMS Muon Detector in Proton-Proton Collisions at $\sqrt{s} = 7$ TeV at the LHC. *JINST*, 8:P11002, 2013. arXiv:1306.6905, doi:10.1088/1748-0221/8/11/P11002.

[121] A. M. Sirunyan et al. Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV. *JINST*, 13(06):P06015, 2018. arXiv:1804.04528, doi:10.1088/1748-0221/13/06/P06015.

[122] A. M. Sirunyan et al. Particle-flow reconstruction and global event description with the CMS detector. *JINST*, 12(10):P10003, 2017. arXiv:1706.04965, doi:10.1088/1748-0221/12/10/P10003.

[123] Giovanni Petrucciani. Particle Flow reconstruction in the CMS Level-1 trigger for the HL-LHC. *EPJ Web Conf.*, 214:01019, 2019. doi:10.1051/epjconf/201921401019.

[124] W. Adam, R. Fruhwirth, A. Strandlie, and T. Todorov. Reconstruction of electrons with the Gaussian sum filter in the CMS tracker at LHC. *eConf*, C0303241:TULT009, 2003. arXiv:physics/0306087, doi:10.1088/0954-3899/31/9/N01.

[125] Marco Cipriani. Photon detection with the CMS ECAL in the present and at the HL-LHC and its impact on Higgs-Boson measurement. *Frascati Phys. Ser.*, 69:150–155, 2019. URL: https://cds.cern.ch/record/2708020.

[126] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Commun. ACM*, 18:509–517, September 1975. doi:10.1145/361002.361007.

[127] Vardan Khachatryan et al. Performance of Photon Reconstruction and Identification with the CMS Detector in Proton-Proton Collisions at sqrt(s) = 8 TeV. *JINST*, 10(08):P08010, 2015. arXiv:1502.02702, doi:10.1088/1748-0221/10/08/P08010.

[128] Hendrik Jansen. *Study of Unparticle plus Lepton Signatures at CMS*. PhD thesis, Aachen, Tech. Hochsch., 2009. URL: http://cds.cern.ch/record/1308722.

[129] Matteo Cacciari and Gavin P. Salam. Pileup subtraction using jet areas. *Phys. Lett. B*, 659:119–126, 2008. arXiv:0707.1378, doi:10.1016/j.physletb.2007.09.077.

[130] Vardan Khachatryan et al. Performance of Electron Reconstruction and Selection with the CMS Detector in Proton-Proton Collisions at $\sqrt{}$s = 8 TeV. *JINST*, 10(06):P06005, 2015. arXiv:1502.02701, doi:10.1088/1748-0221/10/06/P06005.

[131] G. Hanson et al. Evidence for Jet Structure in Hadron Production by e+ e- Annihilation. *Phys. Rev. Lett.*, 35:1609–1612, 1975. doi:10.1103/PhysRevLett.35.1609.

[132] Pileup Removal Algorithms. 2014. URL: https://cds.cern.ch/record/1751454.

[133] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-k_t jet clustering algorithm. *JHEP*, 04:063, 2008. arXiv:0802.1189, doi:10.1088/1126-6708/2008/04/063.

[134] Jet algorithms performance in 13 TeV data. Technical report, CERN, Geneva, 2017. URL: https://cds.cern.ch/record/2256875.

[135] Serguei Chatrchyan et al. Determination of Jet Energy Calibration and Transverse Momentum Resolution in CMS. *JINST*, 6:P11002, 2011. arXiv:1107.4277, doi:10.1088/1748-0221/6/11/P11002.

[136] Vardan Khachatryan et al. Measurement of $B\bar{B}$ Angular Correlations based on Secondary Vertex Reconstruction at $\sqrt{s} = 7$ TeV. *JHEP*, 03:136, 2011. arXiv:1102.3194, doi:10.1007/JHEP03(2011)136.

[137] A. M. Sirunyan et al. Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV. *JINST*, 13(05):P05011, 2018. arXiv:1712.07158, doi:10.1088/1748-0221/13/05/P05011.

[138] Performance of the DeepJet b tagging algorithm using 41.9/fb of data from proton-proton collisions at 13TeV with Phase 1 CMS detector. Nov 2018. URL: https://cds.cern.ch/record/2646773.

[139] Lorenzo Bianchini, John Conway, Evan Klose Friis, and Christian Veelken. Reconstruction of the Higgs mass in $H \to \tau\tau$ Events by Dynamical Likelihood techniques. *J. Phys. Conf. Ser.*, 513:022035, 2014. doi:10.1088/1742-6596/513/2/022035.

[140] Vardan Khachatryan et al. Reconstruction and identification of τ lepton decays to hadrons and ν_τ at CMS. *JINST*, 11(01):P01019, 2016. arXiv:1510.07488, doi:10.1088/1748-0221/11/01/P01019.

[141] A. M. Sirunyan et al. Performance of reconstruction and identification of τ leptons decaying to hadrons and ν_τ in pp collisions at $\sqrt{s} = 13$ TeV. *JINST*, 13(10):P10005, 2018. arXiv:1809.02816, doi:10.1088/1748-0221/13/10/P10005.

[142] Cristina Martin Perez. Development of τ selection techniques and search for the Higgs boson produced in association with top quarks with the CMS detector at the LHC, 2020. presented 21 Oct 2020. URL: https://cds.cern.ch/record/2745035.

[143] Izaak Neutelings. Hadronic tau reconstruction and identification performance in ATLAS and CMS. *PoS*, LHCP2020:045, 2021. doi:10.22323/1.382.0045.

[144] Armen Tumasyan et al. Identification of hadronic tau lepton decays using a deep neural network. 1 2022. arXiv:2201.08458.

[145] Daniele Bertolini, Philip Harris, Matthew Low, and Nhan Tran. Pileup Per Particle Identification. *JHEP*, 10:059, 2014. arXiv:1407.6013, doi:10.1007/JHEP10(2014)059.

[146] Albert M Sirunyan et al. Pileup mitigation at CMS in 13 TeV data. *JINST*, 15(09):P09018, 2020. arXiv:2003.00503, doi:10.1088/1748-0221/15/09/P09018.

[147] Jula Draeger. *Track based alignment of the CMS silicon tracker and its implication on physics performance*. PhD thesis, Hamburg U., 2011. doi:10.3204/DESY-THESIS-2011-026.

[148] S. Gadomski, G. Hall, T. Hogh, P. Jalocha, E. Nygard, and P. Weilhammer. The Deconvolution method of fast pulse shaping at hadron colliders. *Nucl. Instrum. Meth. A*, 320:217–227, 1992. doi:10.1016/0168-9002(92)90779-4.

[149] B. Hooberman, K. Burkett, S. Tkaczyk, and A.Venturi. Charge collection in DECO mode with the CMS silicon strip tracker. internal documentation.

[150] R. Barate et al. Search for the standard model Higgs boson at LEP. *Phys. Lett. B*, 565:61–75, 2003. arXiv:hep-ex/0306033, doi:10.1016/S0370-2693(03)00614-2.

[151] Updated Combination of CDF and D0 Searches for Standard Model Higgs Boson Production with up to 10.0 fb^{-1} of Data. 7 2012. arXiv:1207.0449.

[152] Benjamin W. Lee, C. Quigg, and H. B. Thacker. The Strength of Weak Interactions at Very High-Energies and the Higgs Boson Mass. *Phys. Rev. Lett.*, 38:883–885, 1977. doi:10.1103/PhysRevLett.38.883.

[153] Precision Electroweak Measurements and Constraints on the Standard Model. 12 2010. arXiv:1012.2367.

[154] Albert M Sirunyan et al. A measurement of the Higgs boson mass in the diphoton decay channel. *Phys. Lett. B*, 805:135425, 2020. arXiv:2002.06398, doi:10.1016/j.physletb.2020.135425.

[155] Serguei Chatrchyan et al. Study of the Mass and Spin-Parity of the Higgs Boson Candidate Via Its Decays to Z Boson Pairs. *Phys. Rev. Lett.*, 110(8):081803, 2013. arXiv:1212.6639, doi:10.1103/PhysRevLett.110.081803.

[156] Georges Aad et al. Study of the spin and parity of the Higgs boson in diboson decays with the ATLAS detector. *Eur. Phys. J. C*, 75(10):476, 2015. [Erratum: Eur.Phys.J.C 76, 152 (2016)]. arXiv:1506.05669, doi:10.1140/epjc/s10052-015-3685-1.

[157] D. de Florian et al. Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector. 2/2017, 10 2016. arXiv:1610.07922, doi:10.23731/CYRM-2017-002.

[158] Fabio Maltoni, Giovanni Ridolfi, and Maria Ubiali. b-initiated processes at the LHC: a reappraisal. *JHEP*, 07:022, 2012. [Erratum: JHEP 04, 095 (2013)]. arXiv:1203.6393, doi:10.1007/JHEP04(2013)095.

[159] Albert M Sirunyan et al. Measurements of the Higgs boson width and anomalous HVV couplings from on-shell and off-shell production in the four-lepton final state. *Phys. Rev. D*, 99(11):112003, 2019. arXiv:1901.00174, doi:10.1103/PhysRevD.99.112003.

[160] J R Andersen et al. Handbook of LHC Higgs Cross Sections: 3. Higgs Properties. 7 2013. arXiv:1307.1347, doi:10.5170/CERN-2013-004.

[161] Measurements of differential Higgs boson production cross sections in the leptonic WW decay mode at $\sqrt{s} = 13$ TeV. Technical report, CERN, Geneva, 2019. URL: https://cds.cern.ch/record/2691268.

[162] Morad Aaboud et al. Measurements of gluon-gluon fusion and vector-boson fusion Higgs boson production cross-sections in the $H \to WW^* \to e\nu\mu\nu$ decay channel in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Lett. B*, 789:508–529, 2019. arXiv:1808.09054, doi:10.1016/j.physletb.2018.11.064.

[163] Observation of vector-boson-fusion production of Higgs bosons in the $H \to WW^* \to e\nu\mu\nu$ decay channel in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. 8 2020. URL: https://cds.cern.ch/record/2728055.

[164] Linda M. Carpenter, Tao Han, Khalida Hendricks, Zhuoni Qian, and Ning Zhou. Higgs Boson Decay to Light Jets at the LHC. *Phys. Rev. D*, 95(5):053003, 2017. arXiv:1611.05463, doi:10.1103/PhysRevD.95.053003.

[165] Alexandre Alves and Felipe F. Freitas. Towards recognizing the light facet of the Higgs Boson. *Mach. Learn. Sci. Tech.*, 1(4):045025, 2020. arXiv:1912.12532, doi:10.1088/2632-2153/aba8e6.

[166] Georges Aad et al. Higgs boson production cross-section measurements and their EFT interpretation in the 4ℓ decay channel at $\sqrt{s}$ =13 TeV with the ATLAS detector. *Eur. Phys. J. C*, 80(10):957, 2020. [Erratum: Eur.Phys.J.C 81, 29 (2021), Erratum: Eur.Phys.J.C 81, 398 (2021)]. arXiv:2004.03447, doi:10.1140/epjc/s10052-020-8227-9.

[167] Albert M Sirunyan et al. Measurements of properties of the Higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 11:047, 2017. arXiv:1706.09936, doi:10.1007/JHEP11(2017)047.

[168] A. M. Sirunyan et al. Measurements of Higgs boson properties in the diphoton decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 11:185, 2018. arXiv:1804.02716, doi:10.1007/JHEP11(2018)185.

[169] Measurement of the properties of Higgs boson production at $\sqrt{s}$=13 TeV in the $H \to \gamma\gamma$ channel using 139 fb^{-1} of pp collision data with the ATLAS experiment. 8 2020. URL: https://cds.cern.ch/record/2725727.

[170] Georges Aad et al. A search for the $Z\gamma$ decay mode of the Higgs boson in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Lett. B*, 809:135754, 2020. arXiv:2005.05382, doi:10.1016/j.physletb.2020.135754.

[171] Georges Aad et al. Measurements of WH and ZH production in the $H \to b\bar{b}$ decay channel in pp collisions at 13 TeV with the ATLAS detector. *Eur. Phys. J. C*, 81(2):178, 2021. arXiv:2007.02873, doi:10.1140/epjc/s10052-020-08677-2.

[172] A. M. Sirunyan et al. Observation of Higgs boson decay to bottom quarks. *Phys. Rev. Lett.*, 121(12):121801, 2018. arXiv:1808.08242, doi:10.1103/PhysRevLett.121.121801.

[173] Morad Aaboud et al. Measurement of VH, H $\to$ b$\bar{\text{b}}$ production as a function of the vector-boson transverse momentum in 13 TeV pp collisions with the ATLAS detector. *JHEP*, 05:141, 2019. arXiv:1903.04618, doi:10.1007/JHEP05(2019)141.

[174] Albert M Sirunyan et al. Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector. *Phys. Lett. B*, 779:283–316, 2018. arXiv:1708.00373, doi:10.1016/j.physletb.2018.02.004.

[175] Morad Aaboud et al. Cross-section measurements of the Higgs boson decaying into a pair of τ-leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Rev. D*, 99:072001, 2019. arXiv:1811.08856, doi:10.1103/PhysRevD.99.072001.

[176] Georges Aad et al. Direct constraint on the Higgs-charm coupling from a search for Higgs boson decays into charm quarks with the ATLAS detector. 1 2022. arXiv:2201.11428.

[177] Direct search for the standard model Higgs boson decaying to a charm quark-antiquark pair. 2022. URL: https://cds.cern.ch/record/2802742.

[178] Georges Aad et al. A search for the dimuon decay of the Standard Model Higgs boson with the ATLAS detector. *Phys. Lett. B*, 812:135980, 2021. arXiv:2007.07830, doi:10.1016/j.physletb.2020.135980.

[179] Measurement of Higgs boson decay to a pair of muons in proton-proton collisions at $\sqrt{s} = 13\,\text{TeV}$. 2020. URL: https://cds.cern.ch/record/2725423.

[180] A. David, A. Denner, M. Duehrssen, M. Grazzini, C. Grojean, G. Passarino, M. Schumacher, M. Spira, G. Weiglein, and M. Zanetti. LHC HXSWG interim recommendations to explore the coupling structure of a Higgs-like particle. 9 2012. arXiv:1209.0040.

[181] LHC Higgs Working Group. URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHWG.

[182] J. de Blas et al. Higgs Boson Studies at Future Particle Colliders. *JHEP*, 01:139, 2020. arXiv:1905.03764, doi:10.1007/JHEP01(2020)139.

[183] J. R. Andersen et al. Les Houches 2015: Physics at TeV Colliders Standard Model Working Group Report. In *9th Les Houches Workshop on Physics at TeV Colliders*, 5 2016. arXiv:1605.04692.

[184] S. Amoroso et al. Les Houches 2019: Physics at TeV Colliders: Standard Model Working Group Report. In *11th Les Houches Workshop on Physics at TeV Colliders: PhysTeV Les Houches*, 3 2020. arXiv:2003.01700.

[185] STXS schemes. URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHWGFiducialAndSTXS.

[186] Georges Aad et al. Search for heavy Higgs bosons decaying into two tau leptons with the ATLAS detector using pp collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. Lett.*, 125(5):051801, 2020. arXiv:2002.12223, doi:10.1103/PhysRevLett.125.051801.

[187] Albert M Sirunyan et al. Search for additional neutral MSSM Higgs bosons in the $\tau\tau$ final state in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 09:007, 2018. arXiv:1803.06553, doi:10.1007/JHEP09(2018)007.

[188] Robert V. Harlander, Franziska Hofmann, and Hendrik Mantler. Supersymmetric higgs production in gluon fusion. *Journal of High Energy Physics*, 2011(2), Feb 2011. URL: http://dx.doi.org/10.1007/JHEP02(2011)055, doi:10.1007/jhep02(2011)055.

[189] M. Wiesemann, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, and P. Torrielli. Higgs production in association with bottom quarks. *Journal of High Energy Physics*, 2015(2), Feb 2015. URL: http://dx.doi.org/10.1007/JHEP02(2015)132, doi:10.1007/jhep02(2015)132.

[190] Morad Aaboud et al. Search for charged Higgs bosons decaying via $H^{\pm} \to \tau^{\pm}\nu_{\tau}$ in the τ+jets and τ+lepton final states with 36 fb^{-1} of pp collision data recorded at $\sqrt{s} = 13$ TeV with the ATLAS experiment. *JHEP*, 09:139, 2018. arXiv:1807.07915, doi:10.1007/JHEP09(2018)139.

[191] Albert M Sirunyan et al. Search for charged Higgs bosons in the $H^{\pm} \to \tau^{\pm}\nu_{\tau}$ decay channel in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 07:142, 2019. arXiv:1903.04560, doi:10.1007/JHEP07(2019)142.

[192] S. F. King, M. Mühlleitner, R. Nevzorov, and K. Walz. Discovery Prospects for NMSSM Higgs Bosons at the High-Energy Large Hadron Collider. *Phys. Rev. D*, 90(9):095014, 2014. arXiv:1408.1120, doi:10.1103/PhysRevD.90.095014.

[193] I. Doršner, S. Fajfer, A. Greljo, J. F. Kamenik, and N. Košnik. Physics of leptoquarks in precision experiments and at particle colliders. *Phys. Rept.*, 641:1–68, 2016. arXiv:1603.04993, doi:10.1016/j.physrep.2016.06.001.

[194] Antonio Delgado, Gian F. Giudice, Gino Isidori, Maurizio Pierini, and Alessandro Strumia. The light stop window. *Eur. Phys. J. C*, 73(3):2370, 2013. arXiv:1212.6847, doi:10.1140/epjc/s10052-013-2370-5.

[195] M. Adeel Ajaib, Tong Li, and Qaisar Shafi. Stop-Neutralino Coannihilation in the Light of LHC. *Phys. Rev. D*, 85:055021, 2012. arXiv:1111.4467, doi:10.1103/PhysRevD.85.055021.

[196] R. Gröber, Margarete M. Mühlleitner, E. Popenda, and A. Wlotzka. Light Stop Decays: Implications for LHC Searches. *Eur. Phys. J. C*, 75:420, 2015. arXiv:1408.4662, doi:10.1140/epjc/s10052-015-3626-z.

[197] Daniele Alves. Simplified Models for LHC New Physics Searches. *J. Phys. G*, 39:105005, 2012. arXiv:1105.2838, doi:10.1088/0954-3899/39/10/105005.

[198] Electron and Photon performance in CMS with the full 2017 data sample and additional 2016 highlights for the CALOR 2018 Conference. May 2018. URL: https://cds.cern.ch/record/2320638.

[199] Electron and photon performance in CMS with the full 2016 data sample. Mar 2017. URL: https://cds.cern.ch/record/2255497.

[200] Albert M Sirunyan et al. An embedding technique to determine $\tau\tau$ backgrounds in proton-proton collision data. *JINST*, 14(06):P06032, 2019. arXiv:1903.01216, doi:10.1088/1748-0221/14/06/P06032.

[201] Vardan Khachatryan et al. Measurements of Inclusive W and Z Cross Sections in pp Collisions at $\sqrt{s} = 7$ TeV. *JHEP*, 01:080, 2011. arXiv:1012.2466, doi:10.1007/JHEP01(2011)080.

[202] Vardan Khachatryan et al. Jet energy scale and resolution in the CMS experiment in pp collisions at 8 TeV. *JINST*, 12(02):P02014, 2017. arXiv:1607.03663, doi:10.1088/1748-0221/12/02/P02014.

[203] Albert M Sirunyan et al. Performance of missing transverse momentum reconstruction in proton-proton collisions at $\sqrt{s} = 13$ TeV using the CMS detector. *JINST*, 14(07):P07004, 2019. arXiv:1903.06078, doi:10.1088/1748-0221/14/07/P07004.

[204] Albert M Sirunyan et al. Measurement of the top quark polarization and $t\bar{t}$ spin correlations using dilepton final states in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. D*, 100(7):072002, 2019. arXiv:1907.03729, doi:10.1103/PhysRevD.100.072002.

[205] Measurement of the inclusive and differential tt production cross sections in lepton + jets final states at 13 TeV. 2016. URL: https://cds.cern.ch/record/2141097.

[206] A. M. Sirunyan et al. Search for top squarks decaying via four-body or chargino-mediated modes in single-lepton final states in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 09:065, 2018. arXiv:1805.05784, doi:10.1007/JHEP09(2018)065.

[207] Vardan Khachatryan et al. Event generator tunes obtained from underlying event and multiparton scattering measurements. *Eur. Phys. J. C*, 76(3):155, 2016. arXiv:1512.00815, doi:10.1140/epjc/s10052-016-3988-x.

[208] S. Abdullin, P. Azzi, F. Beaudette, P. Janot, and A. Perrotta. The fast simulation of the CMS detector at LHC. *J. Phys. Conf. Ser.*, 331:032049, 2011. doi:10.1088/1742-6596/331/3/032049.

[209] Chun-Hay Kom and W. James Stirling. Charge asymmetry in W + jets production at the LHC. *Eur. Phys. J. C*, 69:67–73, 2010. arXiv:1004.3404, doi:10.1140/epjc/s10052-010-1353-z.

[210] Albert M Sirunyan et al. Search for supersymmetry in multijet events with missing transverse momentum in proton-proton collisions at 13 TeV. *Phys. Rev. D*, 96(3):032003, 2017. arXiv:1704.07781, doi:10.1103/PhysRevD.96.032003.

[211] Albert M Sirunyan et al. Search for electroweak production of charginos and neutralinos in WH events in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 11:029, 2017. arXiv:1706.09933, doi:10.1007/JHEP11(2017)029.

[212] Haireyecolor: Hair and eye color of statistics students. R Package Documentation. URL: https://rdrr.io/r/datasets/HairEyeColor.html.

[213] A. Abulencia et al. Search for neutral MSSM Higgs bosons decaying to tau pairs in $p\bar{p}$ collisions at $\sqrt{s} = 1.96$ TeV. *Phys. Rev. Lett.*, 96:011802, 2006. arXiv:hep-ex/0508051, doi:10.1103/PhysRevLett.96.011802.

[214] Albert M Sirunyan et al. Measurement of the $Z\gamma^* \to \tau\tau$ cross section in pp collisions at $\sqrt{s} = 13$ TeV and validation of τ lepton analysis techniques. *Eur. Phys. J. C*, 78(9):708, 2018. arXiv:1801.03535, doi:10.1140/epjc/s10052-018-6146-9.

[215] Measurement of Higgs boson production and decay to the $\tau\tau$ final state. Technical Report CMS-PAS-HIG-18-032, CERN, Geneva, 2019. URL: https://cds.cern.ch/record/2668685.

[216] Florian Spreitzer. Estimation of the background to the Higgs tau tau signal due to hadronic jets misidentified as tau lepton decays for the CMS experiment. *TU Wien Bibliothekssystem*, 2019, 2019. doi:10.34726/hss.2019.70600.

[217] G. Cowan. *Statistical data analysis*. 1998. ISBN: 978-0-19-850156-5.

[218] Introduction to Combine. URL: https://cms-analysis.github.io/HiggsAnalysis-CombinedLimit/.

[219] J. S. Conway. Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra. In *PHYSTAT 2011*, pages 115–120, 2011. arXiv:1103.0354, doi:10.5170/CERN-2011-006.115.

[220] Simon Haykin and N Network. A comprehensive foundation. *Neural networks*, 2(2004):41, 2004.

[221] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989. URL: https://www.sciencedirect.com/science/article/pii/0893608089900208, doi:https://doi.org/10.1016/0893-6080(89)90020-8.

[222] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[223] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. URL: https://proceedings.mlr.press/v9/glorot10a.html.

[224] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, 2014. Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015. URL: http://arxiv.org/abs/1412.6980.

[225] Ryota Shimizu, Kosuke Asako, Hiroki Ojima, Shohei Morinaga, Mototsugu Hamada, and Tadahiro Kuroda. Balanced Mini-Batch Training for Imbalanced Image Data Classification with Neural Network. In *2018 First International Conference on Artificial Intelligence for Industries (AI4I)*, pages 27–30, 2018. doi:10.1109/AI4I.2018.8665709.

[226] Moritz Scham. Standard Model H $\to \tau\tau$ Analysis with a Neural Network Trained on a Mix of Simulation and Data Samples. Master's thesis, Karlsruhe Institute of Technology (KIT), 2020. URL: https://publish.etp.kit.edu/record/21993.

[227] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL: http://jmlr.org/papers/v15/srivastava14a.html.

[228] Jörger, Simon. Studies of the usage of neural networks in particle physics analyses. Master's thesis, Karlsruhe Institute of Technology (KIT), 2020. URL: https://publish.etp.kit.edu/record/21950.

[229] Stefan Wunsch, Raphael Friese, Roger Wolf, and Günter Quast. Identifying the Relevant Dependencies of the Neural Network Response on Characteristics of the Input Space. *Computing and Software for Big Science*, 2(1), Sep 2018. URL: http://dx.doi.org/10.1007/s41781-018-0012-1, doi:10.1007/s41781-018-0012-1.

[230] Andrei V. Gritsan, Raoul Röntsch, Markus Schulze, and Meng Xiao. Constraining anomalous Higgs boson couplings to the heavy flavor fermions using matrix element techniques. *Phys. Rev. D*, 94(5):055023, 2016. arXiv:1606.03107, doi:10.1103/PhysRevD.94.055023.

[231] Roger J. Barlow and Christine Beeston. Fitting using finite Monte Carlo samples. *Comput. Phys. Commun.*, 77:219–228, 1993. doi:10.1016/0010-4655(93)90005-W.

[232] Glen Cowan. Statistics for Searches at the LHC. In *69th Scottish Universities Summer School in Physics: LHC Physics*, pages 321–355, 7 2013. arXiv:1307.2487, doi:10.1007/978-3-319-05362-2_9.

[233] Alexander L. Read. Presentation of search results: The CL(s) technique. *J. Phys. G*, 28:2693–2704, 2002. doi:10.1088/0954-3899/28/10/313.

[234] Thomas Junk. Confidence level computation for combining searches with small statistics. *Nucl. Instrum. Meth. A*, 434:435–443, 1999. arXiv:hep-ex/9902006, doi:10.1016/S0168-9002(99)00498-2.

[235] Bernhard Mistlberger and Falko Dulat. Limit setting procedures and theoretical uncertainties in Higgs boson searches. 4 2012. arXiv:1204.3851.

[236] Procedure for the LHC Higgs boson search combination in Summer 2011. Technical report, CERN, Geneva, Aug 2011. URL: http://cds.cern.ch/record/1379837.

[237] Serguei Chatrchyan et al. Combined results of searches for the standard model Higgs boson in pp collisions at $\sqrt{s} = 7$ TeV. *Phys. Lett. B*, 710:26–48, 2012. arXiv:1202.1488, doi:10.1016/j.physletb.2012.02.064.

[238] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic formulae for likelihood-based tests of new physics. *Eur. Phys. J. C*, 71:1554, 2011. [Erratum: Eur.Phys.J.C 73, 2501 (2013)]. arXiv:1007.1727, doi:10.1140/epjc/s10052-011-1554-0.

[239] B. Abbott et al. High-p_T jets in $\bar{p}p$ collisions at $\sqrt{s} = 630$ GeV and 1800 GeV. *Phys. Rev. D*, 64:032003, 2001. arXiv:hep-ex/0012046, doi:10.1103/PhysRevD.64.032003.

[240] F. Abe et al. Measurement of dijet angular distributions at CDF. *Phys. Rev. Lett.*, 77:5336–5341, 1996. [Erratum: Phys.Rev.Lett. 78, 4307 (1997)]. arXiv:hep-ex/9609011, doi:10.1103/PhysRevLett.77.5336.

[241] Vardan Khachatryan et al. Measurement of Dijet Angular Distributions and Search for Quark Compositeness in pp Collisions at *sqrts* = 7 TeV. *Phys. Rev. Lett.*, 106:201804, 2011. arXiv:1102.2020, doi:10.1103/PhysRevLett.106.201804.

[242] Georges Aad et al. Search for new phenomena in dijet mass and angular distributions from pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Phys. Lett. B*, 754:302–322, 2016. arXiv:1512.01530, doi:10.1016/j.physletb.2016.01.032.

[243] Vardan Khachatryan et al. Search for quark contact interactions and extra spatial dimensions using dijet angular distributions in proton–proton collisions at $\sqrt{s} = 8$ TeV. *Phys. Lett. B*, 746:79–99, 2015. arXiv:1411.2646, doi:10.1016/j.physletb.2015.04.042.

[244] Albert M Sirunyan et al. Search for heavy neutrinos and third-generation leptoquarks in hadronic states of two τ leptons and two jets in proton-proton collisions at $\sqrt{s} = 13$ TeV. *JHEP*, 03:170, 2019. arXiv:1811.00806, doi:10.1007/JHEP03(2019)170.

[245] Stefan Wunsch and Simon Jörger and Roger Wolf and Günter Quast. Optimal Statistical Inference in the Presence of Systematic Uncertainties Using Neural Network Optimization Based on Binned Poisson Likelihoods with Nuisance Parameters. *Computing and Software for Big Science*, 5(1), Jan 2021. doi:10.1007/s41781-020-00049-5.

[246] J Apostolakis et al. HEP Software Foundation Community White Paper Working Group - Detector Simulation. 3 2018. arXiv:1803.04165.

[247] High-Luminosity Large Hadron Collider (HL-LHC): Technical Design Report V. 0.1. 4/2017, 2017. doi:10.23731/CYRM-2017-004.

List of Figures

Janík Walter Andrejkovič

Born on August 3rd, 1992
in Visp – Switzerland
Currently living in Vienna

(Scientific) Interests

- Particle Physics
- Machine Learning
- Board games
- Running

Education

2017. **Master degree in High Energy Physics**
Swiss Federal Institute of Technology (ETH)
Zurich – Switzerland
and
École Polytechnique de Paris, Paris-Saclay (EP Paris)
Paris – France

2015. **Bachelor degree in Physics**
Swiss Federal Institute of Technology (ETH)
Zurich – Switzerland

2011. **Matura**
Gymnasium Neufeld
Bern – Switzerland

Current

Oct. 2018 – Jun. 2022. **PhD student at the Institute of High Energy Physics (HEPHY) of the Austrian Academy of Sciences (ÖAW), enrolled at the Technical University (TU) of Vienna.**

Analyzing collision data, recorded by the CMS experiment at CERN.
Thesis title: *Data-driven Background Modeling for Precision Studies of the Higgs Boson and Searches for new Physics with the CMS Experiment*

Publications

with significant contributions:

The CMS Collaboration **(2022)** Measurements of Higgs boson production in the decay channel with a pair of τ leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV *arXiv, to be published in EPJC*. doi: 10.48550/ARXIV.2204.12957

The CMS Collaboration **(2021)** Search for a heavy Higgs boson decaying into two lighter Higgs bosons in the $\tau\tau$bb final state at 13 TeV *JHEP*. doi: 10.1007/JHEP11(2021)057

with minor contributions:

The CMS Collaboration **(2022)** Searches for additional Higgs bosons and vector-like leptoquarks in $\tau\tau$ final states in proton-proton collisions at $\sqrt{s} = 13$ TeV *EPJC*. url: CMS-PAS-HIG-21-001

Research Experience

Oct. 2017 – Jan. 2018. **Research internship**

University Hospital Bern (Insel)
Dynamic trajectory optimization for irradiation therapy and machine learning applications in medical physics

Mar. 2017 – Sep. 2017. **Master thesis**

Institute of Particle Physics at ETH Zurich / CERN
Thesis title: *Learning particle detectors response with boosted decision trees: The* $H \to \gamma\gamma$ *cross section example*

Languages

(Swiss) German	●●●●●
Slovak	●●●●●
English	●●●●○
French	●●●○○
Czech	●●●○○
Italian	●○○○○

Computer Skills

Python	●●●●○
C++	●●●○○
LaTeX	●●●●○
Shell/bash	●●●○○
ROOT	●●●○○
ML tools (e.g. PyTorch)	●●●○○
Grid Computing	●●●○○

Honours and Awards

Oct. 2016. **EP Paris**

Honourable Mention from the physics department of EP Paris for the research internship project (highest grade 20/20)

Apr. 2011. **Swiss Physics Olympics**

Bronze medal

Extracurricular

Oct. 2019 – Oct. 2021. **Student representative**

HEPHY

Jan. 2018 – Jun. 2018. **Military Service**

Swiss Army

Sep. 2016 – May 2017. **Library Supervisor**

ETH Zurich

Jul. 2016 – Sep. 2016. **CERN summer student**

CERN

Multivariate identification of background contributions for the H $\rightarrow \tau\tau$ fake rate method

accessible via: https://cds.cern.ch/record/2214000

Mar. 2016 – Jun. 2016. **Research internship**

Laboratoire Leprince-Ringuet (LLR)

Title: *Neutrino and meson flux simulation with FLUKA and comparison to the experiments T2K and NA61/SHINE*

Spring 2014 **Semester project**

Institute of Astronomy at ETH Zurich

Project title: *Analysis and comparison of galaxy mass profiles in shapelet and wavelet space*

Conference Talks

August 2017 **Joint Annual Meeting of the Swiss and Austrian Physical Society**

Modeling of the detector response in H $\rightarrow \gamma\gamma$ *differential cross section measurements at CMS*

Attended Schools

Sep. 2021 2nd ÖAW AI summer school, Vienna - Austria.

Sep. 2021 VDSP-ESI Winter School - Machine Learning in Physics, Vienna - Austria.

Sep. 2019 CERN School of Computing, Cluj-Napoca - Romania (5 ECTS).

Aug. 2019 1st ÖAW AI Summer school, Ligist - Austria (2 ECTS).

Lecture Notes

Jul. 2018. **Author** of the chapter about positron emission tomography as part of the physics I lecture notes for medicine students at ETH Zurich.

Sep. 2015. **Co-Author** of the lecture notes *Introduction to Particle and Nuclear Physics* for physics students at ETH Zurich.

Last updated: April, 2022.