Measurements of the CP properties of the Higgs boson couplings to tau leptons and top quarks with the CMS experiment

Albert Kenneth Dow
Imperial College London
Department of Physics

A dissertation submitted to Imperial College London
for the degree of Doctor of Philosophy

Abstract

Measurements of the $\mathcal{CP}$ structure of the Higgs boson coupling to τ leptons and top quarks using data collected at the CMS experiment during 2016, 2017 and 2018 are presented. The dataset corresponds to a total integrated luminosity of 137 fb^{-1} at a centre-of-mass energy of $\sqrt{s} = 13$ TeV. For these measurements, events with two τ-leptons in the final state are selected. The effective $\mathcal{CP}$ mixing angle for the Higgs boson to top quark coupling is measured to be $(-5^{+36}_{-37})^{\circ}$. The measurement of the $\mathcal{CP}$ mixing angle in the τ decays reports a value of $(4 \pm 17)^{\circ}$, which is the first direct measurement of $\mathcal{CP}$ properties of the Higgs boson coupling to a pair of τ-leptons. Both results are compatible with standard model expectations.

Declaration

I declare that the work in this Thesis is my own. Several individuals of the CMS collaboration have contributed directly or indirectly to work inspiring the work presented here. Where relevant, the work of others is referenced appropriately.

Chapter one introduces this Thesis and some concepts of particle physics in my own words.

Chapter two dives, in my own words, into the standard model, which the work of others.

Chapter three provides a description in my own words of the CMS detector, which others have engineered and built.

Chapter four explains the reconstruction methods designed by others at the CMS experiment, in my own words.

Chapter five to seven describe the specific methods used in the analyses to be discussed, most of which is my own work.

In Chapter five the selection criteria, simulation corrections and background methods are introduced. Where discussed in detail, these are my own work. Figures are all my own work. Corrections have been derived by several members of the CMS HTT group, to which I have collaborated too. The background methods are built on previous work by others, and have been optimised for these analyses by Daniel Winterbottom.

Chapter six describes the first analysis of this Thesis, which was initially set up by Daniel Winterbottom. Aside from items in Chapter 5 and the addition of the 2018 dataset, my contribution to this analysis consists of selection optimisation and investigations of alternative categorisation methods. I produced the final results of this analysis as presented in this Thesis, which have received endorsement from the Higgs physics conveners of the CMS experiment. These results (or some form of them) will be combined with further channels and will be part of a publication in the near future.

Chapter seven describes the second $\mathcal{CP}$ analysis. I was responsible for a significant amount of the work done on the $\tau_h\tau_h$ channel, including the optimisation of selection and categorisation methods, statistical inference and the combination with the $\tau_\mu\tau_h$ channel. I directly contributed to the synchronisation effort between collaborating groups. The analysis is published in Reference [1].

Chapter eight is a summary, in my own words, of the work discussed in this Thesis.

Not included in this Thesis is my work done on the development of a set of High-Level-Trigger algorithm sequences. I was responsible for designing the trigger sequence, optimising the trigger rate, estimating the potential efficiency, and monitoring the data collected in 2017. These triggers were included in the official trigger menu of the CMS experiment for approximately half of 2017 and all of 2018 data taking. Many of the individual trigger modules used are based on the work of others. The new trigger paths are the first kind of implementation to directly target the VBF phase space of events with Higgs bosons decaying to two τ-leptons.

Albert Kenneth Dow

Acknowledgements

I would like to thank the HEP group at Imperial College London and the STFC for giving me the opportunity of doing a PhD at your facilities. Thank you for giving me the opportunity of working at CERN in Geneva. The CMS group at Imperial consists a great team of people, many of whom I have become quite close to over the years. I have learnt a lot from many people. Thank you to colleagues at CERN who have helped and guided me in various ways. Thank you to David for the supervision and guidance, and for letting me be part of your small, yet strong team. Thank you to Danny for the guidance, advice, and help I have received. Thank you to Sasha for your contributions, especially when discussions became more theoretical. Thank you to all my friends, all the real ones out there, for their support and countless memories. Thank you to my family: thank you Eni and Oli for being the older siblings always keeping an eye on me; thank you Mamma and Papa for the unconditional support and belief in me, for motivating me, for being smart and funny. You inspire me every day. I am forever grateful.

Contents

List of Figures

List of Tables

Poca favilla gran fiamma seconda

Dante Alighieri (Paradiso I)

Chapter 1

Introduction

The standard model (SM) of particle physics is a theory that provides a description of matter and interactions between different types of matter. Not only being a beautiful theory in the way it brings together the electromagnetic and weak forces, an abundant amount of evidence for the validity of the SM has piled up through many years of research and experiments. A major milestone was the experimental discovery of the Higgs boson by the ATLAS and CMS experiments in 2012 [2, 3], which confirmed the Higgs mechanism, first proposed in the 1960s by Englert and Brout [4], Higgs [5, 6, 7], Guralnik, Hagen and Kibble [8, 9].

Nevertheless, the SM is known to fall short in explaining certain physical phenomena. The force of gravity or a dark matter candidate, for instance, is not included in the model. To this end, several models extend or build on the current description of the SM, in which new interactions and particles are present. The SM Higgs boson can be a useful probe to determine if such non-SM-like couplings exist in Nature. Since the Higgs boson is expected to be even under the inversion of charge-parity $\mathcal{CP}$, any deviation from the quantum numbers governing the $\mathcal{CP}$ structure of the Higgs boson would indicate new physics. In the realm of collider physics, these measurements are beginning to gain momentum and make significant conclusions [10, 11].

In this Thesis, the $\mathcal{CP}$ structure of the Higgs boson is investigated using events in the di-τ final state with the CMS experiment, which is a detector located in the LHC ring of the CERN laboratory. Two measurements using data collected during 2016, 2017, and 2018 are described, providing insight into the $\mathcal{CP}$ nature of the Higgs boson from different point of views. The second analysis to be discussed is documented in Reference [1].

The Thesis is structured in eight chapters. Chapter two guides the reader through the foundations of the SM. The Higgs mechanism, including the phenomenology of the Higgs boson, are discussed. This is followed by an introduction into $\mathcal{CP}$ measurements. Finally, the theory of hadron collider physics and event simulation is touched upon.

In Chapter three, the LHC and CMS detector are introduced. The CMS detector is broken down into each sub-detector to understand how each part plays a role in creating this unique machine.

Chapter four describes the reconstruction methods employed at the CMS experiment to go from raw data of tracker hits and energy deposits in the detector to collections of particles to be used for physics analysis.

Chapter five then follows with the selection criteria used by analysts at CMS. At this stage, the focus is on tailoring the list of particles to the analyses to be discussed in the later Chapters. Additionally, the background estimation methods are introduced.

Chapter six describes the analysis of the $\mathcal{CP}$ structure of the Higgs boson using its effective coupling to top-quarks. The analysis methods are discussed, leading to the extraction of the result.

Chapter seven describes the $\mathcal{CP}$ analysis using τ-lepton decays. The first result of this kind is presented herein.

Chapter eight provides a summary of the results and thereafter invites the reader to a discussion of these measurements, including the possibility of improving future analyses.

Chapter 2

Theory of the Standard Model of Particle Physics

2.1 Introduction

The standard model (SM) of particle physics is a quantum field theory that describes the fundamental particles and their interactions. It provides the best illustration of the electromagnetic, weak and strong forces that we currently have, and relies on the principle of local gauge invariance. Many predictions of the SM have been experimentally verified, most importantly the discovery of the Higgs boson in 2012 [3, 2]. Nevertheless, the SM fails to explain some phenomena in Nature, such as the existence of dark matter [12] and neutrino oscillations [13].

This Chapter will introduce the SM, its elementary particles and interactions, followed by a discussion of the Higgs mechanism that is a central theory to the SM. Furthermore, possible extensions to the SM will be introduced by elucidating $\mathcal{CP}$-violating Higgs models. Finally, this Chapter will conclude with a theoretical exploration into collider physics.

2.2 Fundamental particles

In the SM there is a set of fundamental particles and force carriers which are illustrated in Figure 2.1. There are twelve spin-$\frac{1}{2}$ fermions, five spin-1 bosons, and one spin-0 boson, excluding the three and eight color charge states that quarks and gluons respectively may possess. Fermions make up the most fundamental constituents of matter, and may be subcategorised into six quarks and six leptons known as flavours. Leptons interact via the electromagnetic and weak forces, whereas quarks may additional interact through the strong force due to their color charge. The fermions come in three generations, which are particles with identical quantum properties, except for the mass, which is largest for the third-generation fermions. The first generation, and thus the set of most elementary fermions, consists of the up- and down-quarks, the electron and its associated neutrino. Finally, for each particle there exists an anti-particle with the exact same mass, but opposite electric charge.

The second class of particles in the SM are bosons. Gauge bosons are responsible for mediating the electromagnetic, weak and strong interactions between fermions. These forces are mediated via the photon, the $W^{\pm}$ and Z bosons, and the gluon, respectively. The electromagnetic and weak interactions are unified in the SM to make up the electroweak sector. The Higgs boson, on the other hand, is the only spin-0 particle in the SM, and provides a mechanism for the existence of massive gauge bosons and fermions through spontaneous symmetry breaking.

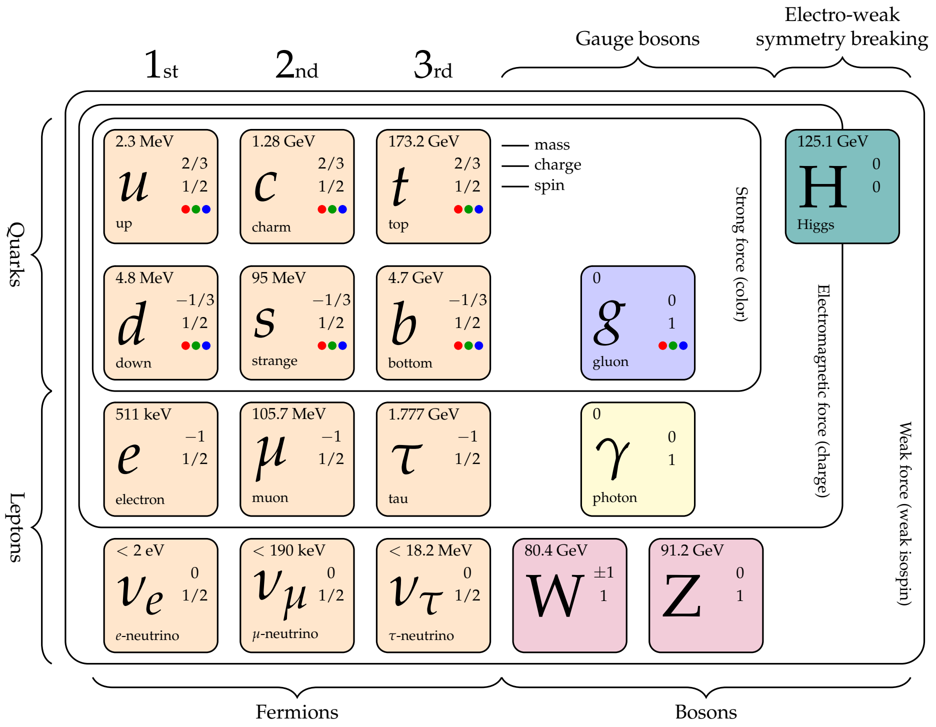

Figure 2.1: A summary of the elementary particles and interaction mediators in the SM. Six quarks and six leptons make up the total of twelve fermions. There are five gauge bosons, which include the $W^{\pm}$. The Higgs boson is the only spin-0 particle in the SM, and is responsible for giving mass to the fermions, and massive gauge bosons. The electromagnetic charges are stated in units of elementary charge. Additionally, all quarks and the gluon possess color charge, which is represented by the red, green and blue colours.

2.3 Gauge invariance

In any (local) field theory, the Lagrangian density, or simply Lagrangian, $\mathcal{L}$, may be written in space-time coordinates as a function of fields $\phi(x)$ and their derivatives as $\mathcal{L}(\phi, \partial_\mu \phi)$, where x^μ are the space-time coordinates. The equations of motions are then given by the principle of least action, which states that the path between two configurations of the system is given when the action is at an extremum. The Euler-Lagrange equation of motion for a field are then a result of this law. If an infinitesimal transformation on the $\phi(x)$ is considered through the infinitesimal parameter α as

$$\phi(x) \to \phi'(x) = \phi(x) + \alpha \Delta \phi(x) \, , \tag{2.1}$$

the transformation is considered a symmetry when it does not affect the equations of motion, which is guaranteed if the action is invariant under Equation 2.1. This implies that the Lagrangian must be invariant under Equation 2.1, which results in Noether's theorem, that states for each continuous symmetry of the Lagrangian a conserved quantity exists. Symmetries, therefore, represent underlying conservation laws and are thus the key to the construction of any Lagrangian in Nature.

This presents the central ingredient for any gauge field theory: the Lagrangian is invariant under local gauge transformations, which correspond to transformations in space-time. Through careful introduction of additional fields, a global symmetry may be altered into a local one. This will be discussed for the case of spin-$\frac{1}{2}$ particles (fermions) of mass m, which are described by the Lorentz-invariant Dirac Lagrangian as

$$\mathcal{L}_{\text{Dirac}} = \bar{\psi} \left(i\gamma^\mu \partial_\mu - m \right) \psi \, , \tag{2.2}$$

where ψ and its adjoint $\bar{\psi} = \psi^\dagger \gamma^0$ are four-component fields known as Dirac spinors, and γ^μ are a set of four 4×4 matrices that satisfy the anticommutation relations $\gamma^\mu \gamma^\nu + \gamma^\nu \gamma^\mu = 2g^{\mu\nu} \times \mathbb{1}_4$. Here $g^{\mu\nu}$ is the Minkowski flat space-time metric and $\mathbb{1}_4$ is the four-dimensional identity matrix. Performing a global phase transformation according to the unitary $U(1)$ group, the field transforms like

$$\psi(x) \to \psi(x)' = e^{ig\alpha} \psi(x) \, , \tag{2.3}$$

where g and α real, constant numbers. The Dirac Lagrangian from Equation 2.2 is invariant under such transformation due to the partial derivative of the phase resulting in the original field, times a constant. However, if the transformation were local instead, such that Equation 2.3 is re-written as

$$\psi(x) \to \psi(x)' = e^{ig\alpha(x)} \psi(x) \, , \tag{2.4}$$

the Lagrangian loses its invariance, as the derivative on the transformed field gives

$$\partial_\mu \psi(x) \to \partial_\mu \psi'(x) = e^{ig\alpha(x)} \left(\partial_\mu \psi(x) + i(g\partial_\mu \alpha(x))\psi(x) \right) \; . \tag{2.5}$$

The final term in Equation 2.5 fails at preserving the invariance under the local phase transformation. Nonetheless, local phase invariance may be achieved through redefinition of the derivative as

$$\partial_\mu \to D_\mu = \partial_\mu + igA_\mu(x) \; , \tag{2.6}$$

where a new field A_μ has been introduced. This field transforms under local phase transformation as

$$A_\mu(x) \to A'_\mu(x) = A_\mu(x) - \frac{1}{g}\partial_\mu \alpha(x) \tag{2.7}$$

and the covariant derivative $D_{\mu\nu}$ transforms as

$$D_\mu \to D'_\mu = e^{i\alpha(x)} D_\mu \psi(x) \; , \tag{2.8}$$

which is exactly way the field transforms. Through the requirement of Lorentz and gauge invariance, the Lagrangian obtains an additional term for the field $A_\mu(x)$, and may be written as

$$\mathcal{L}_{\text{QED}} = \bar{\psi} \left(i\gamma^\mu D_\mu - m \right) \psi - \frac{1}{4}(F_{\mu\nu})^2 \; , \tag{2.9}$$

where $F_{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu$. This Lagrangian describes the interacting field theory of Quantum Electrodynamics (QED). The new field $A_\mu(x)$ is the electromagnetic vector potential, which represents the photon, and is massless, since Equation 2.9 does not exhibit a term of the form $mA^\mu A_\mu$. The constant g is negative the electric charge. The Lagrangian represents an interacting theory as an interaction is present in the term $-e\bar{\psi}\gamma^\mu \psi A_\mu$. This demonstration concludes that the notion of local gauge invariance introduces a new field with a $U(1)$ symmetry that interacts with the original fermion field $\psi(x)$. The degrees of freedom of the unitary $U(1)$ group is one, which corresponds to the number of generated bosons. This symmetry group is just one of the symmetries present in the theory of the SM. The electroweak sector is a theory of gauge symmetry group $SU(2) \otimes U(1)$, whilst the strong interaction requires a $SU(3)$ symmetry.

2.4 Quantum chromodynamics

Quantum chromodynamics (QCD) is a non-Abelian gauge theory under the symmetry group $SU(3)$, meaning that the transformations are non-commutative. The symmetry group is represented by eight trace-less 3×3 matrices. In terms of these matrices, or generators t^a of the group, the covariant derivative is written as

$$D_\mu = \partial_\mu + i g_s t^a A_\mu^a \,. \tag{2.10}$$

The constant g_s is analogous to the electric charge, but for the theory of QCD, and is known as the color charge. As there are eight generators in $SU(3)$, there are eight fields A_μ^a, which correspond to eight gauge bosons known as gluons, each with a different color charge. Being a non-Abelian group is what gives the gluon its color charge and also allows it to interact with itself. This is demonstrated in the field strength as an additional term $-g_s f^{abc} A_\mu^b A_\mu^c$, where f^{abc} are structure constants of the gauge group $SU(3)$. These are present in the definition of the commutator of the generator matrices t^a as $[t^a, t^b] = i f^{abc} t^c$, such that these are completely antisymmetric. In the SM, the only fermions to possess the property of color charge are quarks. Finally, the Lagrangian of the QCD interaction may be written as

$$\mathcal{L}_{\text{QCD}} = \bar{\psi}_f \left(i\gamma^\mu D_\mu - m_f \right) \psi_f - \frac{1}{4} F_a^{\mu\nu} F_{\mu\nu,a} \,, \tag{2.11}$$

where ψ_f is the quark spinor of flavour f and m_f is the mass of that quark flavour. Conversely to QED, the QCD coupling constant g_s has the property of increasing with growing distance between the interacting particles, which is why isolated quarks are never observed in Nature. Practically, this property of color confinement means that quarks and gluons hadronise to form color-less singlets, which are detected as particle showers, or jets. The theory and practical implications of jets will be explained later in this Chapter.

2.5 Theory of electroweak interactions

A major success of the SM is the unification of the electromagnetic and weak interactions which was proposed by Glashow, Weinberg and Salam [14, 15, 16] in the mid-twentieth century, and solves issues present in Fermi's theory of weak interactions [17]. The theory starts with the covariant derivative defined for the $SU(2)$ gauge group

$$D_\mu = \partial_\mu + \frac{i}{2} g W_\mu^i \sigma^i \,, \tag{2.12}$$

where W_μ^i represents three gauge fields and σ^i are the 2×2 Pauli matrices, which are a set of three complex, Hermitian matrices that arise in Quantum Mechanics. All fermionic fields may interact via a weak interaction. An important discrete symmetry that is violated by the weak interaction is parity, which is the action of reversing the spatial coordinates of the field as

$\psi(t,\vec{x}) \to \psi(t,-\vec{x})$. The parity operator acting on a spinor is given by

$$P_{\mathrm{L,R}} = \frac{1}{2}\left(1 \mp \gamma^5\right) , \tag{2.13}$$

where $\gamma^5 = i\gamma^0\gamma^1\gamma^2\gamma^3$ and anticommutes with γ^μ. Therefore expressions like $\bar{\psi}\psi$ are even under parity, whilst $i\bar{\psi}\gamma^5\psi$ is odd under parity. The operator in Equation 2.13 may act on left- and right-handed spinors. Right-handed fermions, however, do not interact to the $\mathrm{W}^\pm$ boson, thus parity is violated. Nevertheless, the Z boson does couple to left- and right-handed particles, thus the covariant derivative in Equation 2.12 must be altered to account for this difference between the charged $\mathrm{W}^\pm$ and neutral Z bosons. Through the introduction of an additional $U(1)$ gauge symmetry, the covariant derivative may be written as

$$D_\mu = \partial_\mu + \frac{i}{2}gW^i_\mu\sigma^i + g'YB^i_\mu , \tag{2.14}$$

where Y is the weak hypercharge, and $W^i_\mu\sigma^i$ and B^i_μ are the weak isospin and hypercharge fields in this $SU(2)_L \otimes U(1)_Y$ group, respectively, where the L represents the left-handed fermion doublets. The weak hypercharge, Y, can be related to the electric charge, Q, and the third component of the weak isospin, t_3, as

$$Q = t_3 + \frac{Y}{2} . \tag{2.15}$$

The physical bosons that are observed in Nature are related to the fields from Equation 2.14 as

$$W^\pm_\mu = \frac{1}{\sqrt{2}}\left(\mathrm{W}^1_\mu \mp i\mathrm{W}^2_\mu\right) \tag{2.16}$$

$$Z_\mu = \cos\theta_\mathrm{W}\mathrm{W}^3_\mu - \sin\theta_\mathrm{W}B_\mu \tag{2.17}$$

$$A_\mu = \sin\theta_\mathrm{W}\mathrm{W}^3_\mu + \cos\theta_\mathrm{W}B_\mu , \tag{2.18}$$

where $\theta_\mathrm{W} = \arctan\frac{g'}{g}$ is the weak mixing angle. This implies a photon associated to the field A_μ, which is massless. However, the same applies to the other three gauge bosons, which have been experimentally observed to have masses as stated in Figure 2.1. The addition of mass terms of the form $-m(\bar{\psi}_L\psi_R + \bar{\psi}_R\psi_L)$ does not accommodate for this, as left and right-handed fermions transform differently and have different $U(1)$ charges, preventing the Lagrangian to be gauge invariant. A solution to this issue, and the issue of massless $\mathrm{W}^\pm$ and Z gauge bosons, is provided through spontaneous symmetry breaking and the Higgs mechanism [4, 5, 6, 7].

2.6 Sponaneous symmetry breaking and the Higgs mechanism

Spontaneous breaking of a symmetry in a quantum field theory occurs when the Lagrangian of the theory remains invariant, whilst the vacuum solutions do not. The notion of spontaneous symmetry breaking is discussed now in the context of a Lagrangian that is invariant under the local $U(1)$ transformation given in Equation 2.4:

$$\mathcal{L} = |D_\mu \phi|^2 - V(\phi) - \frac{1}{4}(F_{\mu\nu})^2 \,, \tag{2.19}$$

where the field $\phi(x)$ is a complex scalar field that interacts with itself and with an electromagnetic field, and the covariant derivative D_μ is defined as in Equation 2.6. A potential $V(\phi)$ may be chosen to be of the form

$$V(\phi) = -\mu^2 \phi^* \phi + \frac{\lambda}{2}(\phi^* \phi)^2 \,, \tag{2.20}$$

where μ^2 and λ are real positive numbers. Due to the existence of the quartic term in the potential, the minimum of occurs when

$$\langle \phi \rangle \equiv v = \left(\frac{\mu^2}{\lambda}\right)^{\frac{1}{2}} \,. \tag{2.21}$$

The field has obtained a vacuum expectation value v as the $U(1)$ symmetry has been broken spontaneously. The shape of a potential of this form is shown in Figure 2.2. The Lagrangian from Equation 2.19 can be expanded about the vacuum expectation value by expanding the

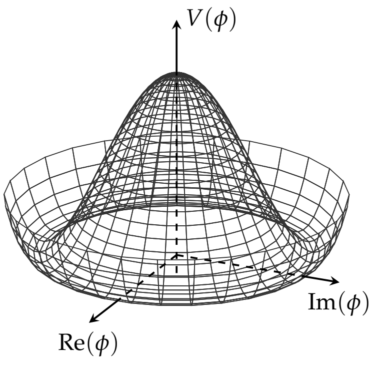

Figure 2.2: Form of the potential in Equation 2.20 that gives rise to a spontaneously broken symmetry when the complex scalar field ϕ acquires a degenerate non-zero vacuum expectation value v given in Equation 2.21.

field about its (positive) minimum as

$$\phi(x) = v + \frac{1}{\sqrt{2}} \left(\phi_1(x) + i\phi_2(x)\right) \ , \tag{2.22}$$

where ϕ_1 and ϕ_2 are real scalar fields. Using this expansion, the potential becomes

$$V(\phi) = -\frac{1}{2\lambda}\mu^4 + \frac{1}{2} \cdot 2\mu^2\phi_1^2 + \mathcal{O}(\phi_i^3) \ . \tag{2.23}$$

Now the field ϕ_1 has mass $m = \sqrt{2}\mu = \sqrt{2\lambda}v$ and ϕ_2 is a massless field. On the other hand, the term with the covariant derivative in the Lagrangian contains a term $g^2v^2A_\mu A^\mu$, such that the mass of the field can be identified as $m_A^2 = 2g^2v^2$. This is how a mass term for the gauge boson A_μ has been introduced to the theory through spontaneous symmetry breaking. The scalar particle ϕ_1 that is part of this theory turns out to be the Higgs boson, whose mass depends on the vacuum expectation value v.

The SM Higgs boson field is a scalar field that is a two-component spinor and transforms as an $SU(2)$ doublet

$$\phi(x) = U(x)\frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix} \ , \tag{2.24}$$

which in the general case is acted upon by an $SU(2)$ gauge transformation $U(x)$ that can be eliminated through another gauge transformation. $h(x)$ is a real scalar field. The potential is of the form presented in Equation 2.2, such that the Lagrangian can be written as

$$\mathcal{L} = |D_\mu\phi|^2 + \mu^2\phi^\dagger\phi - \lambda(\phi^\dagger\phi)^2 \ , \tag{2.25}$$

where the covariant derivative D_μ is the one from Equation 2.14. Acting this covariant derivative on the field $\phi(x)$ produces terms like

$$\frac{g^2v^2}{4} W_\mu^+ W^{-\mu} + \frac{(g^2 + g'^2)v^2}{8} Z_\mu Z^\mu \ . \tag{2.26}$$

The rotations in $SU(2)_L \otimes U(1)_Y$ space described in Equation 2.16 have been applied. The $W^\pm$ and Z bosons now have acquired masses, which are related by $m_W = m_Z \cos\theta_W$. Through spontaneous symmetry breaking and the Higgs mechanism, these gauge bosons have obtained their masses.

Returning to the point mentioned in Section 2.5 about missing mass terms for the quarks and leptons, the Higgs mechanism provides a solution to this issue. The Yukawa terms that describe the interaction between the Higgs field ϕ and a fermion field ψ are of the form

$$\frac{g_f}{\sqrt{2}} v \bar{\psi}_L \phi \psi_R + h.c. \,, \tag{2.27}$$

where g_f is Yukawa coupling constant for fermion f and $h.c.$ represents the Hermitian conjugate of the previous expression. Now, gauge invariant mass terms are present in the Lagrangian. Expanding the expression in Equation 2.27 yields terms of the form

$$m_f = \frac{1}{\sqrt{2}} g_f v \,. \tag{2.28}$$

A mass term for fermion f has been generated. This extends to all quarks and leptons, and the value of the Yukawa coupling constant g_f that controls the mass of the fermion is a free parameter that must be measured experimentally. Additionally, interaction terms that are proportional to the mass of the fermion arise.

The Higgs boson interaction terms with the $\mathrm{W}^{\pm}$ and Z gauge bosons are

$$m_{\mathrm{W}}^2 \left(\frac{2h}{v} + \frac{h^2}{v^2} \right) \mathrm{W}_\mu^+ \mathrm{W}^{-\mu} + \frac{m_{\mathrm{Z}}^2}{2} \left(\frac{2h}{v} + \frac{h^2}{v^2} \right) \mathrm{Z}_\mu \mathrm{Z}^\mu \,. \tag{2.29}$$

The couplings of the Higgs boson to the massive gauge bosons is proportional to their masses. The mass of the Higgs boson, as previously found, arises from the potential energy terms in the Lagrangian

$$-\lambda v h^2 - \lambda v h^3 - \frac{1}{4} \lambda h^4 \,, \tag{2.30}$$

such that the mass of this scalar particle associated to the field $h(x)$ is given by

$$m_h = \sqrt{2\lambda} v \,. \tag{2.31}$$

2.7 The Higgs boson at the Large Hadron Collider

The Large Hadron Collider (LHC) set out to discover the Higgs boson as a key ingredient in its physics programme. This was achieved in 2012 by the ATLAS and CMS collaborations when the discovery of a new boson with a mass of 125 GeV was announced [2, 3]. Additional evidence, such as the measurement of spin and parity of this new boson, confirm the compatibility with the Higgs boson as predicted by the SM [18, 10].

At the LHC, two proton beams are made to collide with each other at large values of centre-of-mass energies, $\sqrt{s}$. The Higgs boson may be produced out of these collisions. The four major production modes to consider for Higgs boson production at the LHC with proton beams of $\sqrt{s} = 13\,\text{GeV}$ are: gluon-gluon fusion (ggH), vector boson fusion (VBF), vector boson associated production (VH), and top quark-antiquark pair associated production (ttH). The most dominant production mode is through ggH as is summarised in Table 2.1. Vector boson fusion is the second most dominant production mode, with a cross-section that is one order of magnitude smaller than that of ggH, however, it proves to be one of the most significant modes. This is due to the associated quarks that are produced in the interaction, which hadronise to form jets. These jets can be used to discriminate against backgrounds, such as the QCD-induced multi-jet background activity, as they tend to be well separated in spatial coordinates or approximately back-to-back. The remaining associated Higgs boson production modes are less significant in terms of cross-section. The cross-sections for these production modes is summarised in Table 2.1 and the leading-order Feynman diagrams that represent these processes are illustrated in Figure 2.3. The Higgs boson decays almost immediately after its production [19], meaning its decay products must be detected to extrapolate back and infer on its prior presence. Decay modes with branching fraction greater than 0.2% are summarised in Table 2.2. Its decay to a $b\bar{b}$-quark pair occurs about 60% of the time, however, it is plagued by large hadronic backgrounds at the LHC, making it an experimentally very challenging Higgs boson decay channel. Nevertheless, this channel has been observed in recent years with increasing data [20, 21]. In spite of their small branching fractions, the decay modes $\gamma\gamma$ and ZZ were observed in the first discovery of the Higgs boson, as these have less background contamination and cleaner signatures [2, 3]. The H $\rightarrow \tau^+\tau^-$ process, with a branching fraction of about 6%, was discovered in 2016 [22]. Fermionic decays of the Higgs boson occur directly at tree-level and are illustrated in Figure 2.4 for a generic fermion f.

Table 2.1: Cross-sections and their associated theoretical uncertainties for the main production modes of the SM Higgs boson of mass $m_H = 125\,\text{GeV}$ at proton-proton centre-of-mass energies of $\sqrt{s} = 13\,\text{TeV}$ [19].

Production mode	Cross-section [pb]	Uncertainty [%]
ggH	48.58	5.0
VBF	3.78	2.1
W$^{\pm}$-associated H	1.37	2.0
Z-associated H	0.88	4.1
$t\bar{t}$-associated H	0.51	6.8

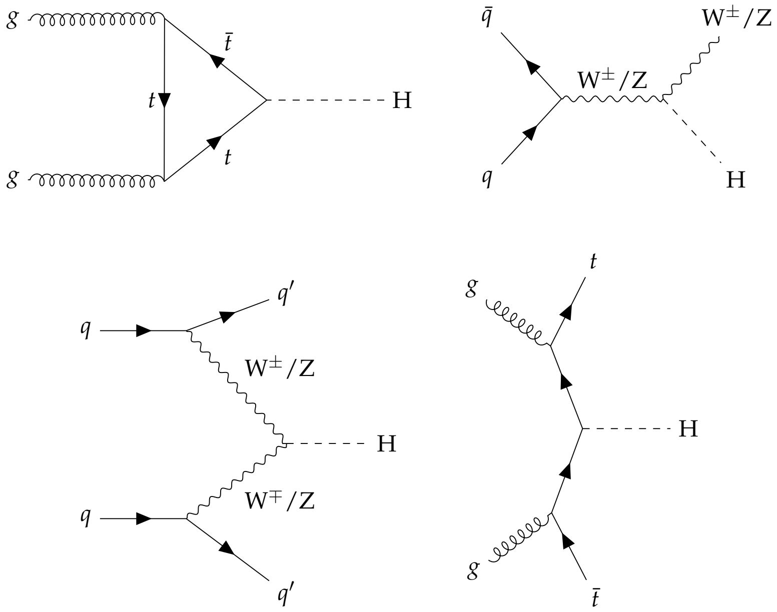

Figure 2.3: Feynman diagrams for the main SM Higgs boson production modes at leading order. From left to right, the top row illustrates the ggH and VH processes, whereas the bottom row shows the VBF and ttH processes. The ggH process involves a virtual loop of a heavy quark, which is predominantly induced through the t-quark, being the heaviest particle in the SM. Two incoming quarks radiate a vector gauge boson which merge to form a Higgs boson in the VBF mode. In the V-associated Higgs boson production process, a quark-antiquark pair annihilate to produce a vector boson that radiates a Higgs boson. Finally, the ttH process requires two incoming gluons that split into $t\bar{t}$-quark pairs where one of these pairs annihilates to form a Higgs boson.

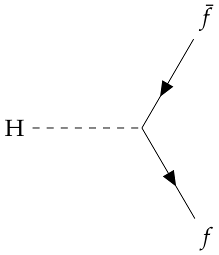

Figure 2.4: Feynman diagram for a generic fermionic decay of the Higgs boson. The Higgs boson decays into a fermion-antifermion pair, $f\bar{f}$.

Table 2.2: Decay modes with branching fraction $\mathcal{B} > 0.2\%$ for the SM Higgs boson of mass $m_H = 125\,\text{GeV}$. The $*$ on the massive gauge bosons indicates a virtual (off-shell) boson. Decays to quark and lepton pairs occur directly, whereas decays to massless particles require the presence of a virtual loop of particles [19].

Decay mode	$\mathcal{B}$ [%]
$b\bar{b}$	58.24
$\mathrm{W}^{\pm}\mathrm{W}^{\mp *}$	21.37
gg	8.19
$\tau^{+}\tau^{-}$	6.27
$c\bar{c}$	2.89
ZZ^{*}	2.61
$\gamma\gamma$	0.23

2.8 $\mathcal{CP}$ violation in the Higgs sector

Although seen as being the best description of particles and their interactions, the SM fails to describe certain observations. An example is the discovery of neutrino oscillations [13], which requires neutrinos to be massive particles. In the SM, however, the Higgs mechanism does not allow mass terms for neutrinos to be written in the Lagrangian due to the lack of right-handed neutrinos, as the Higgs boson couples only to the $SU(2)_L$ gauge. Mass-generating mechanisms have been developed theoretically [23], but still require experimental evidence. Another issue that has established itself as a key problem of the SM is the lack of explanation on the imbalance of matter and antimatter in the universe. This baryonic [1] asymmetry requires a certain amount of $\mathcal{CP}$ violation. Charge conjugation $\mathcal{C}$ and parity $\mathcal{P}$ are discrete symmetries which convert a particle into its antiparticle and a left-handed particle into a right-handed particle through the sign reversal of Cartesian coordinates, respectively. Although $\mathcal{CP}$ violation has been observed so far only in the quark sector, the amount of violation is too small to be consistent with models that explain the baryonic asymmetry observed, such as Sakharov's model of baryogenesis [24]. Strong evidence for $\mathcal{CP}$ violation in neutrino oscillations has been discovered more recently [25]. This naturally raises the question whether or not $\mathcal{CP}$ violation may be observed in other areas of the SM. Of particular interest is the Higgs sector, meaning the collection of fields and particles that give rise to the Higgs mechanism. As previously discussed in this Chapter, the Higgs boson is the only particle in the SM that is responsible for generating the masses of fermions and gauge bosons, however, extensions to the SM exist

[1] A baryon is a composite particle made up of an odd number (at least three) of valence quarks. These particles are the basis of all matter in the Universe. The lightest baryon is the proton, which contains two u and one d quark. The term *valence* quark refers to quarks that determine the quantum numbers of particles, as opposed to the unspecified number of *sea* quarks and gluons that are present in hadrons. These concepts are part of the parton model of hadrons.

which predict a Higgs sector consisting of more fields and particles. The simplest possible extension to the SM is achieved through the introduction of a second complex $SU(2)$ doublet. Models of this kind are known as two-Higgs-doublet models (2HDM) and manifest themselves through a myriad of theories [26].

In the SM, the Higgs sector comprises only the single Higgs boson, as was depicted in Figure 2.1. This boson is a spin-zero scalar particle, and is therefore adheres to $\mathcal{CP}$ symmetry. A $\mathcal{CP}$-odd spin-zero Higgs boson is not allowed in the SM, but would manifest itself as a parity-odd boson through its asymmetry under rotational transformations, which in turn can be inferred on through angular correlations of particles associated to its production or decay. In a generic 2HDM, however, the Higgs sector consists of three neutral Higgs boson and two charged Higgs bosons. In $\mathcal{CP}$-violating cases, the neutral Higgs bosons are not eigenstates of the tandem $\mathcal{CP}$ operation, as would be the case for the SM Higgs boson, but are compositions of $\mathcal{CP}$-even and $\mathcal{CP}$-odd components. To add the possibility of a $\mathcal{CP}$-violating component in the SM Lagrangian describing the interaction between the Higgs boson resonance $h(x)$ and fermions of flavour f we can write the Lagrangian [27] as

$$\mathcal{L}_{hff} = g_f \bar{\psi}_f \psi_f h + \tilde{g}_f \bar{\psi}_f i\gamma^5 \psi_f h \, , \tag{2.32}$$

where $g_f \approx \tilde{g}_f \approx g_f^{\text{SM}} = \frac{m_f}{v}$ is the Yukawa coupling constant first introduced in Equation 2.27. The second term in Equation 2.32 introduces a $\mathcal{CP}$-odd contribution to the interaction.

Alternatively, this may be rewritten in terms of the reduced Yukawa coupling, $\kappa_f = \frac{g_f}{g_f^{\text{SM}}}$ and $\tilde{\kappa}_f = \frac{\tilde{g}_f}{g_f^{\text{SM}}}$ as

$$\mathcal{L}_{hff} = \frac{m_f h}{v}(\kappa_f \bar{\psi}_f \psi_f + \tilde{\kappa}_f \bar{\psi}_f i\gamma^5 \psi_f) \, . \tag{2.33}$$

An effective mixing angle that is proportional to the ratio of $\tilde{\kappa}_f$ to κ_f can then be defined. Measuring such an angle would indicate how the Higgs boson couples to the fermion, thus, making inference on the Higgs boson's $\mathcal{CP}$ state possible. How these may be measured at the LHC will be the topic of discussion of the following section.

2.8.1 Measuring the $\mathcal{CP}$ structure of the Higgs boson at the LHC

Several analyses have provided interesting results on the $\mathcal{CP}$ state of the Higgs boson. One of these involves the H $\rightarrow$ ZZ process, where the $\mathcal{CP}$-even and $\mathcal{CP}$-odd hypotheses were tested [10]. Figure 2.5 shows the test statistic for the two hypothesis and how, at confidence levels of 95%, the pseudo-scalar Higgs boson hypothesis is rejected. This analysis exploited angles between decay planes of the Z bosons, which provide differentiation between the two $\mathcal{CP}$ states [28]. This analysis has been extended to study the VBF and VH production processes,

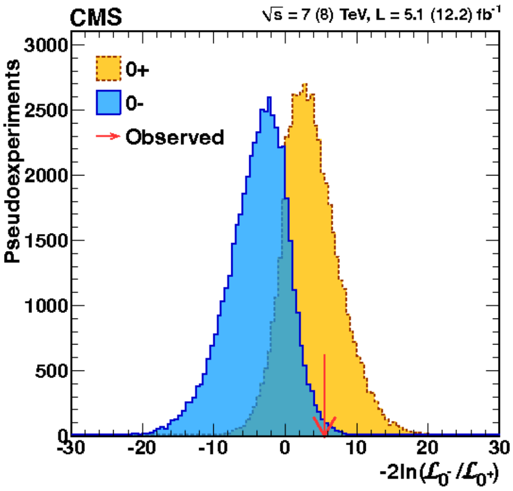

Figure 2.5: Test statistic for $\mathcal{CP}$-even and $\mathcal{CP}$-odd hypotheses in H → ZZ analysis. The $\mathcal{CP}$-odd pseudoscalar hypothesis of the Higgs boson is rejected at a confidence level of 95%. This Figure has been reproduced from [10].

as jets from the Higgs boson production vertex exhibit a topology that provides discriminating power between $\mathcal{CP}$ states [11]. Unlike the previous method used for the H → ZZ analysis, this analysis also tests $\mathcal{CP}$-violating cases where both $\mathcal{CP}$-even and $\mathcal{CP}$-odd components contribute [29]. Nevertheless, most $\mathcal{CP}$-violating models considering the coupling between the Higgs boson and vector bosons do not contain tree-level $\mathcal{CP}$-odd terms, suppressing this component heavily. The $\mathcal{CP}$-even Higgs boson couples to the vector boson V as $h\mathrm{V}^{\mu}\mathrm{V}_{\mu}$, whilst the $\mathcal{CP}$-odd component couples through a dimension five operator $h\mathrm{V}^{\mu\nu}\mathrm{V}_{\mu\nu}$, where $\mathrm{V}^{\mu\nu}$ is the field strength for boson V [27]. Therefore, on a more practical level, the observables used in these analyses will not vary as much for the $\mathcal{CP}$-even and $\mathcal{CP}$-odd hypotheses. For fermionic decay modes of the Higgs boson, on the other hand, $\mathcal{CP}$-even and $\mathcal{CP}$-odd couplings occur at tree level. A recent measurement that exploits the ttH production mode, and thus the coupling between the Higgs boson and the t-quark, has reported a rejection of the $\mathcal{CP}$-odd hypothesis at a significance level of 3.2σ [30]. For this measurement, correlations between the t-quark decay products allow the $\mathcal{CP}$ structure of the vertex to be tested. Similarly, angular correlations between the τ-lepton decay products can be used in H → $\tau^+\tau^-$ decays to provide the differentiation power between $\mathcal{CP}$ states [31, 32]. Alternatively, the H → $\tau^+\tau^-$ mode can also be used to tag events where the Higgs boson was produced through ggH with two associated jets from the vertex [27, 33]. This way the t-quark Yukawa coupling can be probed by exploiting the topology of the jets. These two analyses will be the main subjects of this Thesis and will be described in Chapters 6 and 7. The methodology employed in both analyses will be introduced in the following sections, starting with the analysis of the t-quark Yukawa coupling.

2.8.2 t-quark Yukawa coupling measurement with $gg \to \mathrm{H} + 2$ jets

Assuming the ggH process predominantly contains the t-quark loop and the Yukawa interaction to fermions is as described by Equation 2.32, the effective interaction between the Higgs boson and the gluon may be written as

$$\mathcal{L}_{hgg} = \kappa_t \frac{\alpha_s}{12\pi v} h G^a_{\mu\nu} G^{a,\mu\nu} + \tilde{\kappa}_t \frac{\alpha_s}{16\pi v} h G^a_{\mu\nu} G^a_{\rho\sigma} \epsilon^{\mu\nu\rho\sigma} , \tag{2.34}$$

where $G^a_{\mu\nu}$ is the gluon field strength tensor, α_s is the strong coupling constant from QCD [33], and $\epsilon^{\mu\nu\rho\sigma}$ is the 4-dimensional Levi-Civita symbol. Feynman diagrams containing Hgg, Hggg and H$gggg$ vertices emerge from this Lagrangian. These are illustrated in Figure 2.6, where the leftmost diagram representing Hgg contributes to the signal whilst the Hggg and H$gggg$ dilute the signal. The tensor structure of the ggH vertex is

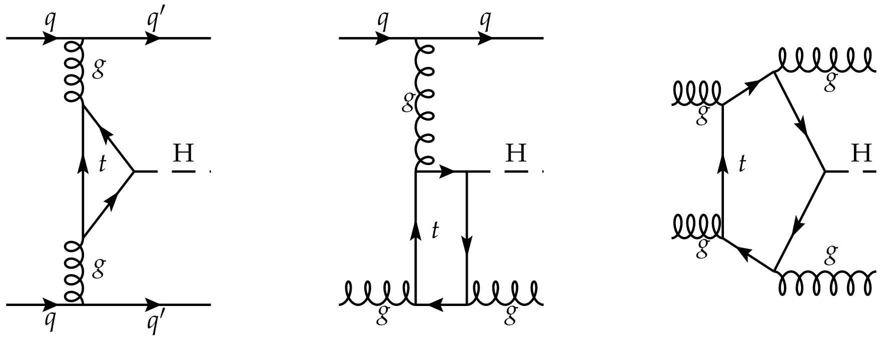

Figure 2.6: Feynman diagram for the Hgg, Hggg, and H$gggg$ processes due to the effective Lagrangian interaction described in Equation 2.34.

$$T^{\mu\nu} = a\left(q_1 \cdot q_2 g^{\mu\nu} - q_1^{\nu} q_2^{\mu}\right) + b\epsilon^{\mu\nu\rho\sigma} q_{1\rho} q_{2\sigma} , \tag{2.35}$$

where q_1 and q_2 denote the four-momenta of the gluons entering the t-quark loop, a and b are scalar form factors that are associated to the Yukawa coupling constants by the relations

$$a = \kappa_t \cdot \frac{\alpha_s}{3\pi v} \tag{2.36}$$

$$b = \tilde{\kappa}_t \cdot \frac{\alpha_s}{2\pi v} . \tag{2.37}$$

This way the $\mathcal{CP}$-mixing angle introduced in Equation 2.33 can be defined as the ratio of these form factors

$$\tan \alpha_{gg} = \frac{b}{a} , \tag{2.38}$$

with $-\frac{\pi}{2} < \alpha_{gg} < \frac{\pi}{2}$. Any mixing angle of $0 < |\alpha_{gg}| < \frac{\pi}{2}$ represents a $\mathcal{CP}$-violating case where both scalar and pseudo-scalar contributions exist in the Lagrangian.

The tensor structure of the effective ggH coupling can be analysed using the azimuthal angular separation between the two jets, $\Delta\phi_{jj} \equiv \Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2})$ in the final state that emerge at the ggH vertex. This has been shown to be a powerful discriminating variable for analyses considering HVV couplings [34], and works analogously for the case of ggH. Distributions of $\frac{d\sigma}{d|(\Delta\phi(\vec{p}_{j_1},\vec{p}_{j_2}))|}$ will provide differentiation between the two terms in the tensor structure of Equation 2.35 [35]. The observable $\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2})$ is, however, a parity-even operator, which does not provide any sensitivity to $\mathcal{CP}$-violating scenarios where both $\mathcal{CP}$-even and $\mathcal{CP}$-odd contributions are present. A parity-odd operator, which does provide this information, is the signed $\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2})$, where the sign is defined by fixing the direction along the beam axis for two counter-propagating proton-proton beams. The (normalised) distributions of $\frac{d\sigma}{d\Delta\phi(\vec{p}_{j_1},\vec{p}_{j_2})}$ are illustrated in Figure 2.7 for the $\mathcal{CP}$-even, $\mathcal{CP}$-odd and $\mathcal{CP}$-mixed cases. These demonstrate the power of the discriminating variable, prior to including effects due to detector geometry [33]. The measurement of the t-quark coupling to the Higgs boson using the ggH production mode will adopt this methodology and parameterisation to infer on the $\mathcal{CP}$ nature of the Higgs boson.

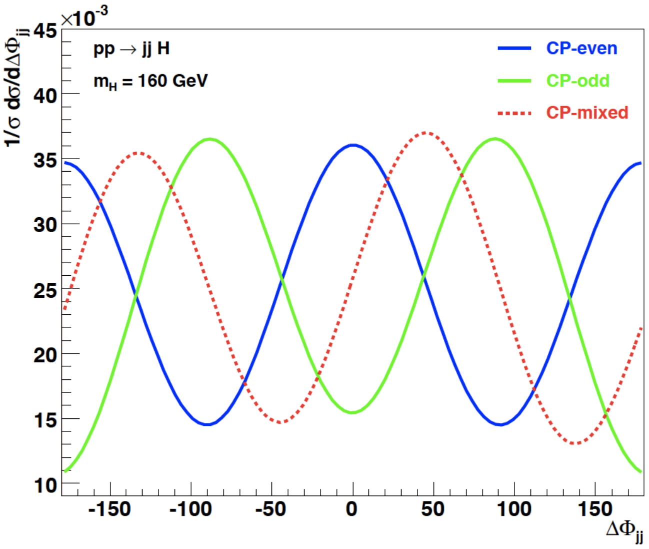

Figure 2.7: Normalised distributions of the discriminating observable $\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2})$ for $\mathcal{CP}$-even, $\mathcal{CP}$-odd and $\mathcal{CP}$-mixed cases. The $\mathcal{CP}$-mixed case occurs when the form factors defined in Equation 2.35 are $a = b \neq 0$. Note that these distributions are at parton level and have an analysis-level selections applied to enhance the separation of the different $\mathcal{CP}$ states. This Figure has been reproduced from [33].

2.8.3 τ-lepton Yukawa coupling measurement with $\mathrm{H} \to \tau^+\tau^-$ decays

The measurement of the Yukawa coupling using $\mathrm{H} \to \tau^+\tau^-$ decays involves different techniques. The nature of the production mode of the Higgs boson in this case is not relevant, as long as the Higgs boson has been produced in the pp collision. The analysis starts with the Lagrangian describing the Yukawa coupling of the Higgs boson and a fermion, where in this case the fermion is the τ lepton as described in Equation 2.33. The (tangent of the) effective mixing angle, $\phi_{\tau\tau}$, can be defined as just the ratio of reduced Yukawa couplings:

$$\tan\phi_{\tau\tau} = \frac{\tilde{\kappa}_\tau}{\kappa_\tau}\,, \tag{2.39}$$

where $-\frac{\pi}{2} < \phi_{\tau\tau} \leq \frac{\pi}{2}$, and $0 < |\phi_{\tau\tau}| < \frac{\pi}{2}$ represents any $\mathcal{CP}$ violating scenario between the purely $\mathcal{CP}$-even ($\phi_{\tau\tau} = 0$) and $\mathcal{CP}$-odd ($|\phi_{\tau\tau}| = \frac{\pi}{2}$) cases.

A difference in $\mathcal{CP}$ states of the Higgs boson is manifested in spin correlations of the τ-lepton decay products. The decay probability of a Higgs boson to a fermion pair may be written in terms of spin vectors, $\vec{s}$, of a generic fermion, f, in the fermions' rest frames as [36]

$$\Gamma(\mathrm{H} \to f\bar{f}) \propto 1 - s_{\parallel}\bar{s}_{\parallel} \pm s_{\perp}\bar{s}_{\perp}\,, \tag{2.40}$$

where the sign depends on the $\mathcal{CP}$ state of the Higgs boson, and the parallel and perpendicular directions are those of the direction of travel of the fermions. In the case of the Higgs boson decaying into two τ leptons, the perpendicular spin component has a direct effect on the angular correlations of the τ-lepton pair decay products. Due to the presence of neutrinos in decays of the τ lepton, the rest frame of the di-τ lepton pair is experimentally difficult to determine. However, switching from the Higgs boson rest frame to the rest frame of the leading charged decay products of each τ lepton leaves the physics invariant [31], and provides a good approximation.

Focusing on the simplest case where both τ leptons decay into a single charged pion $\pi^\pm$ and an undetected neutrino, the decay in the zero-momentum frame of the two pions is illustrated in Figure 2.8. This method requires precise measurement of the four-momenta of the $\pi^\pm$ and their impact parameters [31], denoted $q^\pm$ and $\lambda^\pm = (0, \vec{j}^\pm)$, respectively. The 3-dimensional impact parameter $\vec{j}^\pm$ defines the vector between the primary vertex and the closest point to the charged pions' track. The decay plane $\lambda^\pm$ spanned by the charged pions' impact parameter and momentum vector represents the true decay plane of the τ lepton, including its neutrino decay products. Since this induces the aforementioned experimental challenges, the four-vector $\lambda^\pm$ is boosted into the charged pions' zero-momentum frame. In this way a correlated decay plane of the τ lepton is reconstructed. Therefore, boosting into the corresponding frame of reference of the charged decay products, the angle between the planes spanned by the impact parameters and momenta vectors is given by

$$\phi^* = \arccos\left(\hat{\lambda}_\perp^+ \cdot \hat{\lambda}_\perp^-\right) \ , \tag{2.41}$$

where $0 \leq \phi^* \leq \pi$. This angle has the power to discriminate between $\mathcal{CP}$-even and $\mathcal{CP}$-odd Higgs boson states. A second variable can be defined as

$$\mathcal{O}_{\mathcal{CP}}^* = \hat{q}^{*-} \cdot \left(\hat{\lambda}_\perp^+ \times \hat{\lambda}_\perp^-\right) \ , \tag{2.42}$$

which is a $\mathcal{CP}$-odd and time-reversal-odd (T-odd) correlation that is sensitive to mixed $\mathcal{CP}$ scenarios. The information of two variables defined in Equations 2.41 and 2.42 can be combined into one acoplanarity angle $\phi_{\mathcal{CP}}$ as

$$\phi_{\mathcal{CP}} = \begin{cases} \phi^* & \text{if } \mathcal{O}_{\mathcal{CP}}^* \geq 0 \, , \\ 2\pi - \phi^* & \text{if } \mathcal{O}_{\mathcal{CP}}^* < 0 \, , \end{cases} \tag{2.43}$$

where the $*$ denoting the zero-momentum frame of the charged particles has been dropped. The distribution of $\phi_{\mathcal{CP}}$ will be of sinusoidal form for any $\mathcal{CP}$ state, where a phase determines its exact nature. This technique will be referred to as the impact parameter method. For decays of the τ lepton with more than one visible particle, such as $\tau^- \rightarrow \rho^- \bar{\nu}_\tau \rightarrow \pi^- \pi^0 \bar{\nu}_\tau$, an analogous method can be applied, as illustrated in Figure 2.9. The impact parameters are replaced by the four-momenta of the neutral pions π^0. The determination of $\phi_{\mathcal{CP}}$ is then performed using identical vector algebra as described for the impact parameter method. However, due to destructive interference from different polarisations of the ρ mesons, an

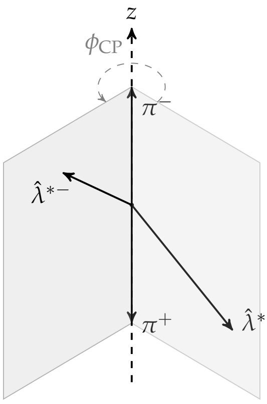

Figure 2.8: Acoplanarity angle using decay planes in the $\tau^+\tau^- \rightarrow \pi^+\pi^-\nu_\tau\bar{\nu}_\tau$. In the zero-momentum frame of the two charged pions, the decay planes spanned by the momenta and impact parameters of the charged particles are constructed.

additional observable needs to defined as

$$y^{\tau^\pm} = \frac{E_{\pi^\pm} - E_{\pi^0}}{E_{\pi^\pm} + E_{\pi^0}}\,, \tag{2.44}$$

where the E_π represents the energy of the corresponding pion in the laboratory frame. A shift of $\phi_{\mathcal{CP}} \to 2\pi - \phi_{\mathcal{CP}}$ is applied if the product of $y^{\tau^+}y^{\tau^-}$ is negative. The distribution of $\phi_{\mathcal{CP}}$ now provides sensitivity to the $\mathcal{CP}$ state of the Higgs boson.

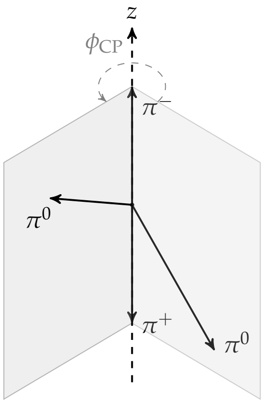

Figure 2.9: Acoplanarity angle using decay planes in the $\tau^+\tau^- \to \rho^+\rho^-\nu_\tau\bar{\nu}_\tau$. In the zero-momentum frame of the two charged pions, where one charged pion originates from the subsequent $\rho^\pm \to \pi^\pm\pi^0$ decay, the decay planes spanned by the momenta and impact parameters of the charged particles are constructed.

Since the two decay planes used in the calculation of $\phi_{\mathcal{CP}}$ are constructed separately, the aforementioned techniques can also be combined into a mixed method, where one plane is determined using an impact parameter and the other using a neutral pion [31]. This is illustrated in Figure 2.10.

2.9 Hadron collider physics and event simulation

In modern collider physics experiments, a large sample of simulated events is required to model and interpret results. Through the use of Monte Carlo techniques, events of a particular process can be simulated. These, however, ignore higher-order effects in perturbation theory, which result in infrared and ultraviolet divergences. Renormalisation techniques are applied to remove these effects.

In proton-proton collisions, where partons a and b interact and produce a final state c, $(ab \to c)$, the cross-section can be factorised into a process-dependent hard subprocess and a process-

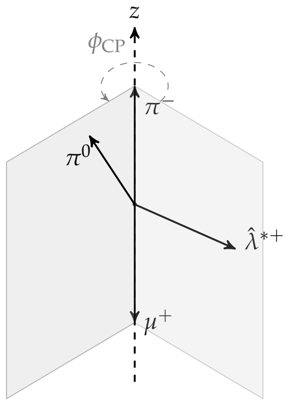

Figure 2.10: Acoplanarity angle using decay planes in the $\tau^+\tau^- \to \mu^+\rho^-\nu_\tau\bar{\nu}_\tau$.

independent normalisation part as [37]

$$\sigma = \sum_{a,b} \int_0^1 dx_a dx_b \int f_a^{p_1}(x_a, \mu_F) f_b^{p_2}(x_b, \mu_F) d\hat{\sigma}_{ab\to c}(\mu_F, \mu_R) \, . \tag{2.45}$$

In Equation 2.45, x_a is the momentum fraction of parton a with respect to the original proton, and similarly for x_b for parton b. The two factors f_a^p represent the parton distribution functions, which describe the density of the associated parton in the parent proton p. The μ_F and μ_R are the factorisation and renormalisation scale, which are unphysical quantities set to specify the cross-section. Their choice is rather arbitrary; they are usually designated specific values depending on the scale of the interaction, for instance the mass of a resonance in an s-channel scattering process. These are then varied in the simulation to determine the impact of the chosen scales. These are typically propagated as alternative event weights through an analysis in the systematic uncertainty template. The hard subprocess is contained in the process-dependent factor in Equation 2.45 as $d\hat{\sigma}_{ab\to c}(\mu_F, \mu_R)$, which is the parton-level cross-section of the production of final state c. The differential cross-section is determined by the matrix element squared that is averaged over initial spin states and colors of process $ab \to c$, scaled by the incoming parton flux. The desired accuracy of the matrix element calculation is given by the number of considered Feynman diagrams. The lowest order in vertices representing strong or electroweak interactions is known as the leading-order (LO), or tree-level, diagram. This is followed by the next-to-leading-order (NLO) diagram which considers virtual or real emission corrections. Nevertheless, to have a complete picture of the process, the impact of higher orders on the matrix element can be included using a parton shower algorithm. This type of generator provides an evolution from the scale of the hard subprocess to the low scale of $\mathcal{O}(1\,\mathrm{GeV})$. These scales are where hadronisation occurs, where particles with color are confined into bound states of neutral color. Experimentally, these are observed as hadronic showers, known as jets, which represent the parton's energy deposit as a cluster of hadrons.

The most commonly used jet clustering algorithm at modern hadron colliders, such as the CMS experiment, is the anti-k_{T} algorithm [38]. Its ability to cluster jets and not obtain any sensitivity to soft radiation makes it an attractive choice of the list of available jet algorithms. It begins by defining the distance between particle i and j as d_{ij} and the distance between particle i and the beam B as d_{iB}. Additionally, two distance metrics are defined as

$$d_{ij} = \min(p_{T,i}^{-2}, p_{T,j}^{-2}) \frac{\Delta R_{ij}^2}{R^2} \tag{2.46}$$

$$d_{iB} = p_{T,i}^{-2} \,, \tag{2.47}$$

where $p_{T,i}^{-2}$ and $p_{T,j}^{-2}$ are the transverse momenta of the particles i and j, respectively, and ΔR is the separation between particle i and j in the solid angle. The radius parameter of the jet is controlled by variable R. In an iterative fashion, the algorithm calculates d_{ij} and d_{iB} for all particles. If the minimum distance is one given by d_{ij}, then the particles i and j are clustered into one object. Conversely, if d_{iB} results in the minimum distance, then particle i is promoted to a jet and is removed from the particle list. The algorithm iterates over all particles in an event and the result is a collection of jets [38].

2.9.1 Modelling of $\mathcal{CP}$-dependent signal

In order to describe signal distributions of any $\mathcal{CP}$ scenario, three sets of samples are required. The $\mathcal{CP}$-even (SM), $\mathcal{CP}$-odd (PS), and maximally-mixed $\mathcal{CP}$ (MM) scenarios are chosen. The cross-section of a scenario of generic $\mathcal{CP}$ can be written as

$$\mu = \frac{\sigma}{\sigma_{\mathrm{SM}}} = \kappa_t^2 + \frac{\sigma_{\mathrm{PS}}}{\sigma_{\mathrm{SM}}} \cdot \tilde{\kappa}_t^2 \,, \tag{2.48}$$

where $\kappa_t = \frac{g_t}{g_t^{\mathrm{SM}}}$, $\tilde{\kappa}_t = \frac{\tilde{g}_t}{g_t^{\mathrm{SM}}}$, and σ_{PS} is the cross-section for a pseudo-scalar state when $\kappa_t = 0$ and $\tilde{\kappa}_t = 1$. This equals $\left(\frac{3}{2}\right)^2 \sigma_{\mathrm{SM}}$ as can be deduced from Equation 2.36. Similarly, the differential cross-section for any variable x can be expressed as a function of the three chosen $\mathcal{CP}$ scenarios as

$$\frac{d\sigma}{dx} = \kappa_t^2 \cdot \frac{d\sigma_{\mathrm{SM}}}{dx} + \tilde{\kappa}_t^2 \cdot \frac{d\sigma_{\mathrm{PS}}}{dx} + 2\kappa_t\tilde{\kappa}_t \cdot \frac{d\sigma_{\mathrm{int}}}{dx} \,, \tag{2.49}$$

where the last term is the interference term. The $\mathcal{CP}$ mixing angle ϕ_f can be introduced into Equation 2.49 using

$$\tan\phi_f = \frac{b}{a} = \sqrt{\frac{\sigma_{\text{PS}}}{\sigma_{\text{SM}}}} \cdot \frac{\tilde{\kappa}_t}{\kappa_t} \,. \tag{2.50}$$

Using this relation, the differential cross-section may be written as

$$\begin{aligned}
\frac{d\sigma}{dx} = \mu \cdot \Bigg(&\cos^2(\phi_f)\frac{d\sigma_{\text{SM}}}{dx} && (2.51)\\
&+ \sin^2(\phi_f)\frac{\sigma_{\text{SM}}}{\sigma_{\text{PS}}}\frac{d\sigma_{\text{PS}}}{dx} && (2.52)\\
&+ 2\cos(\phi_f)\sin(\phi_f)\frac{\sigma_{\text{SM}}}{\sigma_{\text{MM}}}\frac{d\sigma_{\text{int}}}{dx}\Bigg) \,. && (2.53)
\end{aligned}$$

The interference term can be written in terms of the difference of the chosen maximally-mixing $\mathcal{CP}$ distribution and the $\mathcal{CP}$-even and $\mathcal{CP}$-odd distributions:

$$\begin{aligned}
\frac{d\sigma}{dx} = \mu \cdot \Bigg(&(\cos^2(\phi_f) - \cos(\phi_f)\sin(\phi_f))\frac{d\sigma_{\text{SM}}}{dx} && (2.54)\\
&+ (\sin^2(\phi_f) - \cos(\phi_f)\sin(\phi_f))\frac{\sigma_{\text{SM}}}{\sigma_{\text{PS}}}\frac{d\sigma_{\text{PS}}}{dx} && (2.55)\\
&+ 2\cos(\phi_f)\sin(\phi_f)\frac{\sigma_{\text{SM}}}{\sigma_{\text{MM}}}\frac{d\sigma_{\text{MM}}}{dx}\Bigg) \,. && (2.56)
\end{aligned}$$

Finally, as the individual differential distributions are normalised to the SM cross-section, the cross-sections can be set to be equal to each other without loss of generality. Therefore Equation 2.54 can be simplified to

$$\begin{aligned}
\frac{d\sigma}{dx} = \mu \cdot \Bigg(&(\cos^2(\phi_f) - \cos(\phi_f)\sin(\phi_f))\frac{d\sigma_{\text{SM}}}{dx} && (2.57)\\
&+ (\sin^2(\phi_f) - \cos(\phi_f)\sin(\phi_f))\frac{d\sigma_{\text{PS}}}{dx} && (2.58)\\
&+ 2\cos(\phi_f)\sin(\phi_f)\frac{d\sigma_{\text{MM}}}{dx}\Bigg) \,. && (2.59)
\end{aligned}$$

An important consequence of Equation 2.57 is that the inclusive yield of the distributions remains invariant to a change in ϕ_f. As any beyond-the-SM effect can alter the cross-section, an alteration in the cross-section does not provide any conclusive evidence of a non-SM $\mathcal{CP}$ state. However, observing the differential distribution provides a direct way of determining the $\mathcal{CP}$ nature of the Higgs boson. This generic parameterisation is used in both analyses and will be re-introduced in the respective Chapters.

2.10 Summary

The SM of particle physics is a quantum field theory that offers a precise and extensively tested description of the elementary particles and their interactions. Through spontaneous symmetry breaking, the Higgs mechanism is responsible for generating massive particles, with the exception of neutrinos. The discovery of the Higgs boson in 2012 strengthened the foundations of the SM. Nevertheless, due to unanswered questions and a growing list of evidence for physics beyond the SM, several models and extensions accommodate for a Higgs sector with $\mathcal{CP}$-violating properties. These can be tested using the Higgs boson decay to a pair of τ leptons, which may probe the $\mathcal{CP}$ structure of Yukawa couplings from the Higgs boson production and decay processes. The proposed measurements study either the Yukawa coupling to the t-quark using the effective ggH production vertex in association with two jets, or the Yukawa coupling to the τ lepton using spin correlations of its decay products. Using the appropriate Monte Carlo techniques, proton-proton collisions can be simulated to perform such measurements at the CMS experiment.

Chapter 3

The LHC Complex and CMS Detector

3.1 Introduction

The Large Hadron Collider (LHC) was designed to lead the frontier of high-energy physics and to study the Standard Model and Higgs mechanism. Additionally, many interesting searches of beyond the standard model physics and rare processes have been proposed and are ongoing efforts. The LHC is the world's leading accelerator for particle physics experiments and will continue to provide insight into nuclear and particle physics for the next decades.

3.2 The LHC complex

The LHC [39] is a hadron accelerator and circular collider built in a tunnel located about 100 m underground at the *European Organization for Nuclear Research* (CERN) in Geneva, Switzerland. The 27 km circumference tunnel was previously used by the *Large Electron-Positron Collider* (LEP) accelerator. Protons are initially extracted from hydrogen gas bottles through ionisation, which removes the from the bound states in the nuclei. A linear accelerator (LINAC2) based on radio-frequency (RF) cavities accelerates these to an energy of 50 MeV. Subsequently, the protons are injected into the *Proton Synchroton* (PS) Booster, which consists of four superimposed synchrotron rings, and reach an energy of 1.4 GeV. They are further passed through the PS and the *Super Proton Synchrotron* (SPS) where they are accelerated to 450 GeV and consequently injected into the main LHC ring in two adjacent counter-rotating beam pipes. Through the use of eight RF cavities, they are accelerated to the desired centre-of-mass energy, 14 TeV being the design luminosity of the LHC machine. Protons in these beams are clustered into bunches, consisting of 2808 proton bunches at the design operation. Each of these is made up of $\mathcal{O}(10^{11})$ protons and are spaced an interval of 25 ns apart. Finally, 1232 niobium-titanium superconducting dipole magnets ensure the beams stay on track in the curved paths, which requires a cooling temperature of 1.9K to create magnetic fields of 8.3 T. Four beam crossing points are spread around the LHC ring, where the four main detectors – ATLAS [40], CMS [41], LHCb [42] and ALICE [43] – are built. A schematic illustrating the aforementioned parts of the LHC complex is shown in Fig. 3.1.

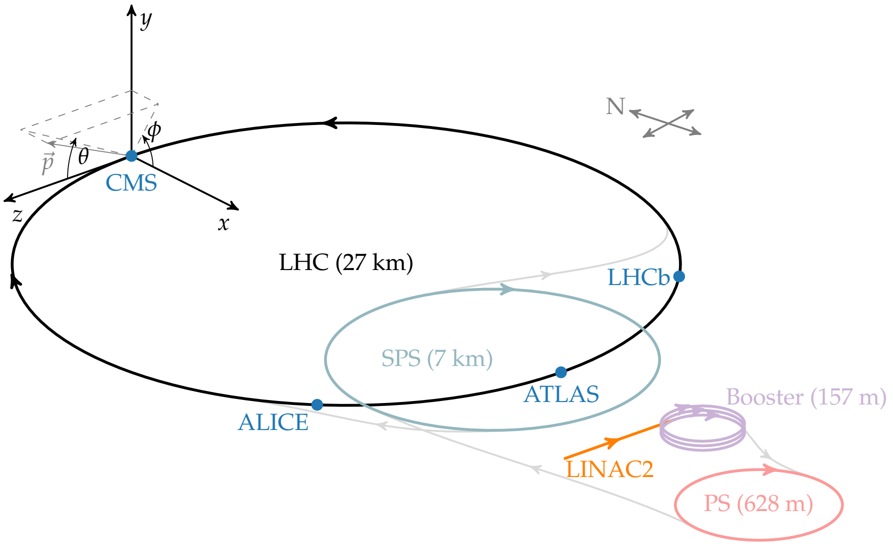

Figure 3.1: Sections of the LHC complex relevant to CMS analysts. The chain of injectors and accelerators is illustrated, leading to the LHC ring with its four main detectors. The coordinate system employed by CMS is drawn onto the schematic.

The event rate of a particular process with cross-section σ is given by

$$\mathcal{N} = L\sigma \, , \tag{3.1}$$

where L denotes the integrated luminosity of the beam delivered by the LHC. Therefore the luminosity directly determines the number of expected events, meaning a high luminosity is desirable especially for processes with low cross-section, such as the ones involving Higgs boson production. For instance, for ggH production with a cross-section of 48.5 fb^{-1} at a centre-of-mass energy of 13 TeV [19], one Higgs boson will be produced every two seconds. In terms of beam parameters, the luminosity is defined as

$$L = \frac{N_b^2 n_b f_{rev} \gamma_r}{4\pi \epsilon_n \beta^*} F \, , \tag{3.2}$$

where N_b is the number of particles per bunch, n_b is the number of bunches per beam, f_{rev} is the frequency of one revolution, γ_r is the gamma factor due to relativistic effects, ϵ_n is the normalised transverse beam emittance, β^* is the beta function at the interaction point and F is an additional factor that reduces the luminosity due to the crossing angle by the beams at their collision point. This equation holds for beams that have Gaussian shapes in the transverse direction. The integrated luminosity of physics-approved data for the 2016, 2017 and 2018 CMS data-taking runs are depicted as a function of time in Fig. 3.2. They correspond to an integrated luminosity of 35.9 fb^{-1}, 41.5 fb^{-1} and 59.7 fb^{-1}, respectively, recording a total of 137 $\mathrm{fb}^{-1}$of data available for use in physics analyses.

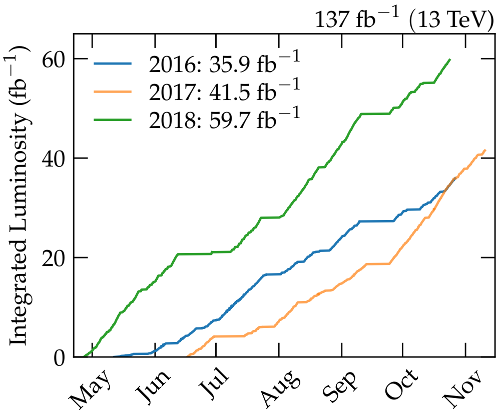

Figure 3.2: Integrated luminosity over the 2016, 2017 and 2018 data-taking period at the CMS experiment for use in physics analysis. Data is approved for physics research if all sub-detectors were functioning as expected.

3.3 CMS detector

The CMS detector was designed to investigate high-energy physics collisions at the TeV scale and shed light on the existence of the Higgs boson, which was eventually discovered in 2012 through independent searches by ATLAS [2] and CMS [3]. As illustrated in Fig. 3.3, the 12.5×10^6 kg CMS detector consists of a set of sub-detectors surrounding the beam axis around the interaction point. The centre-piece of CMS is the 3.8 T superconducting magnetic solenoid that provides a large bending power to charged-particle tracks to ensure high precision measurements of their momenta. Within the bore of the magnet coil, the inner tracker and main parts of the calorimeters are situated. Outside of the coil, several muon detectors make up the muon system that provides full coverage.

The coordinate system adopted by the CMS experiment is shown in Fig. 3.1. The origin is centred at the nominal interaction point. The x-axis points inwards towards the centre of the LHC ring, the y-axis is oriented vertically upwards, and the z-axis points tangentially outwards along the beam axis. Therefore, any transverse component of a vector quantity, such as p_{T}, is defined in the (x, y) plane. The angles ϕ and θ are defined from the x- and z-axis, respectively. Finally, a radial coordinate system that is more commonly used employs the Lorentz-invariant pseudo-rapidity which is defined as $\eta = -\ln\left(\tan\frac{\theta}{2}\right)$. Using this variable, (squared) positions can be expressed in the (η, ϕ) plane as $R^2 = \eta^2 + \phi^2$.

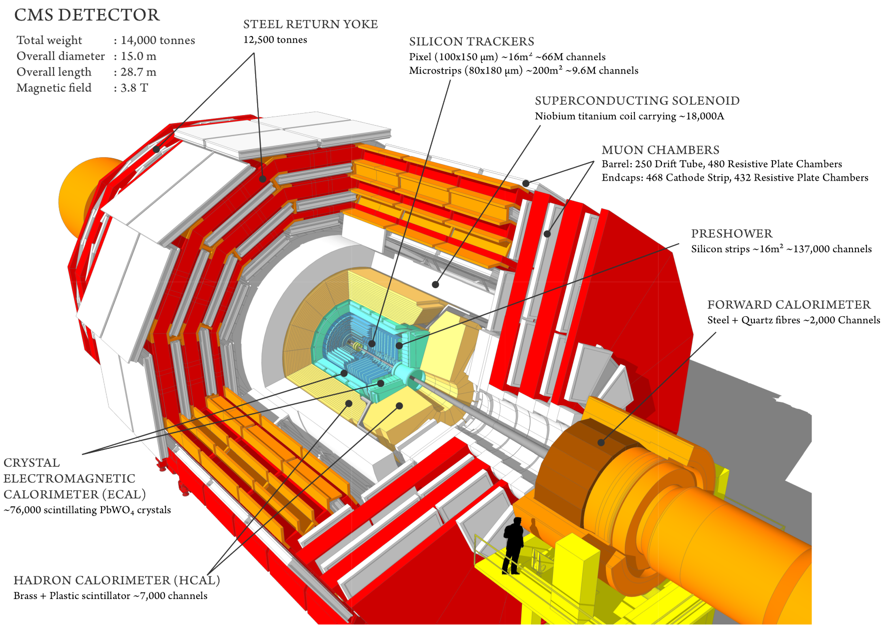
CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~16m² ~66M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 468 Cathode Strip, 432 Resistive Plate Chambers
PRESHOWER
Silicon strips ~16m² ~137,000 channels
FORWARD CALORIMETER
Steel + Quartz fibres ~2,000 Channels
CRYSTAL ELECTROMAGNETIC CALORIMETER (ECAL)
~76,000 scintillating PbWO₄ crystals
HADRON CALORIMETER (HCAL)
Brass + Plastic scintillator ~7,000 channels

Figure 3.3: A cross-sectional schematic view of the CMS detector. Figure was reproduced from [44].

3.4 Magnet

The superconducting solenoid magnet [45] is essential to allowing precise measurements of charged-particle momenta to be made, due to its large homogeneous magnetic field of 3.8 T leading to high bending power. It has a length of 12.5 m and a free bore of 6 m diameter. Stabilised reinforced niobium-titanium conducting wire is used to make up the four layer winding in the coil. To operate as a superconducting magnet, the temperature is kept at about 4.5 K using liquid helium as the cooling agent. The magnetic flux is returned through a 10^7 kg iron yoke, which is made up of five wheels and two endcaps. Within the solenoid lies the tracker, whose performance relies heavily on the magnet.

3.5 Tracker

For every pp interaction, a precise measurement of the collision point and trajectories of charged particles is desired. The 5 m long inner tracking system [46, 47] is built around the interaction point and has a diameter of 2.5 m. A high granularity and fast response sub-detector is required to cope with the $\mathcal{O}(10^3)$ particles from every bunch crossing. Additionally, the tracker needs to be resistant again radiation damage caused by the high particles flux. The solution adopted by the CMS experiment to accommodate for these challenges is to use silicon detector technology. Positions of charged particles are determined from ionisation deposits on the reverse-biased p-n junction. The curved trajectory of radius r for a particle of charge q in a magnetic field of strength B is reconstructed using a set of deposits (or hits), such that the transverse momentum is given by

$$p_{\mathrm{T}} = rqB \, . \tag{3.3}$$

The tracking system consists of two main components: silicon pixels and strips. These are illustrated in Fig. 3.4. The pixel detector 3.5, which was upgraded between the 2016 and 2017

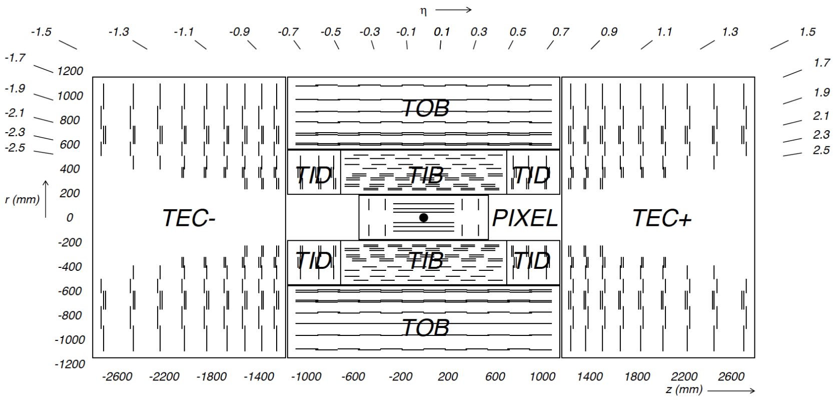

Figure 3.4: A schematic illustrating the layout of the 2016 version of the full tracker. The pixel detector has been upgraded to include an additional layer [47]

.

data-taking periods, covers a region of $|\eta| < 2.5$ and measures the 3-dimensional position of hits. Four (previously three) layers of 53 cm pixels are placed in the barrel region (BPIX) at radii 3.0 cm, 6.8 cm, 10.2 cm and 16.0 cm. Three (previously two) disks are located at the endcap regions (FPIX) at longitudinal positions of ±29.1 cm, ±39.6 cm and ±51.6 cm. The pixel detector contains a total of 124 million pixels, each with dimensions of 100 μm × 150 μm.This upgraded set-up enables seeding the track reconstruction using a collection of four hits as opposed to three hits, which has an intrinsically lower fake rate. The silicon strips, on the

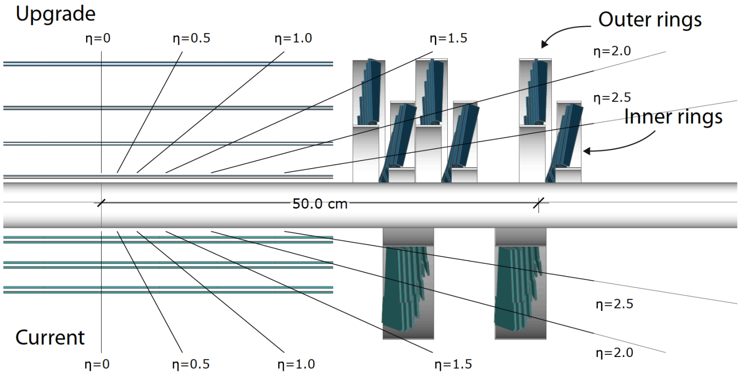

Figure 3.5: A cross-sectional schematic view of the 2016 and 2017/2018 versions of the pixel detector [47].

other hand, consists of four sub-modules and is placed further away from the beam pipe and is therefore subjected to less flux than the pixels. The strips are grouped into an inner and an outer component. The tracker inner barrel (TIB) and tracker inner disks (TID) comprises four layers in the barrel region and three disks in the endcap, respectively. The TIB layers of individual cell sizes of 10 cm × 80 µm are placed at radii ranging from 20 cm to 50 cm and covering $|z| <$65 cm, whereas the TID provides coverage in the overlap region between the inner and outer components and thus is placed further along the z-axis perpendicularly to the TIB layers. The TIB and TID detectors are enclosed by the tracker outer barrel (TOB), which consists of six layers of thickness 500 µm positioned at radii of 55 cm to 116 cm and provides longitudinal coverage up to $|z| <$118 cm. The tracker outer endcap (TEC) is positioned at each end of the TID and TOB, and comprises nine disks. These are placed at longitudinal distances of 124 cm$<$ $|z| <$282 cm and have radii of 22.5 cm to 113.5 cm. Finally, stereo configurations (forming crosshatch layout) with an angle of 100 mrad are used for the first two layers of TIB and TOB, the first two rings in TID, and the first, second and fifth disk of the TEC. In total, 9.6 million silicon strips are placed within the inner tracking system.

With the high rate of incoming particle collisions and therefore data, quick response processing is provided by an on-board chip that makes a decision on whether or not to accept the event and consequently read out the analogue signal for digitisation and further processing. This is further detailed in section 3.9.

3.6 Electromagnetic calorimeter

The electromagnetic calorimeter (ECAL) [48] is the first calorimeter system surrounding the inner tracker. Calorimeters are essential to obtain precise measurements of energy deposits of particles. Additionally, given their geometry, they allow for full coverage in the (η, ϕ) plane, such that any missing energy due to neutrinos, for instance, can be inferred. The ECAL, shown in Fig. 3.6, consists of about 76×10^3 scintillating lead tungstate ($PbWO_4$) crystals which cover a pseudo-rapidity region of $|\eta| < 3.0$. A small Molière radius[1] of 2.2 cm and short radiation length of $X_0 = 0.89$ cm are motives for the choice crystal element, as these allow for a compact design while providing a high granularity. The ECAL sub-detector is split into three regions:

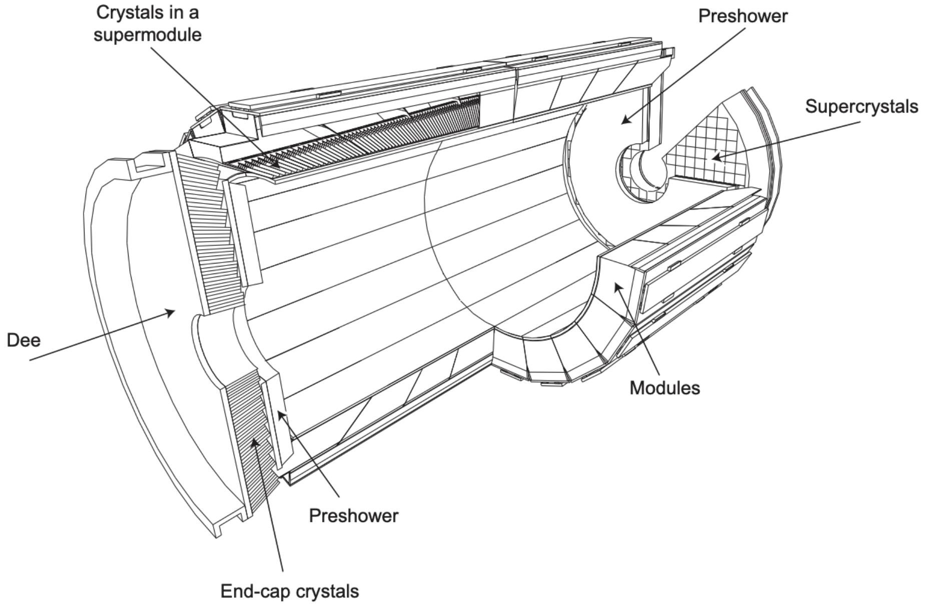

Figure 3.6: A schematic view illustrating the ECAL detector of the CMS experiment. Figure reproduced from [49]

.

the barrel (EB), endcap (EE) and preshower (ES). The EB covers a region up to $||\eta|| = 1.479$ and has crystals aligned at an average inner radius of 129 cm from the z-axis and tilted 3° with respect to the nominal interaction point, in both η and ϕ. The axis tilt ensures that particle trajectory are not aligned with any gap between crystals. The front face area of the crystals covers 0.0174×0.0174 in (η, ϕ), equivalent to 22 mm × 22 mm, and has a length of 230 mm which corresponds to $25.8X_0$. Crystals are grouped into pairs in ϕ and into clusters of five in η to create submodules. Furthermore, these are clustered into groups of ten in ϕ and four

[1] When a shower deposits energy, the radius containing 90% of this energy, on average, is known as the Molère radius.

(or five) in η to create a module. Four modules are connected together side-by-side to form a supermodule covering an angle of 20° in ϕ, with 18 supermodules in half a barrel.

The two EE sections, on the other hand, span the pseudo-rapidity region of $1.479 < |\eta| < 3.0$. They are situated at $|z| = 315.4$ cm from the nominal interaction point and comprise 7,324 crystals of length 22 cm ($24.7X_0$) with a front face area of 2.86 cm × 2.86 cm. At the front side of each EE, the ES detectors are situated, which have a thickness of 20 cm ($3X_0$). These ES detectors cover a region of $1.653 < |\eta| < 2.6$ and provide additional calorimetry to identify individual photons in the decay of neutral pions ($\pi^0 \to \gamma\gamma$), and to enhance the position resolution of electrons and photons.

The energy resolution of a supermodule in the ECAL sub-detector, $\mathrm{E}_{\mathrm{ECAL}}$, is parametrised in the following fashion using a stochastic term S, a noise term N and a constant term C:

$$\left(\frac{\sigma_{\mathrm{ECAL}}}{\mathrm{E}_{\mathrm{ECAL}}}\right)^2 = \left(\frac{S}{\sqrt{\mathrm{E}_{\mathrm{ECAL}}}}\right)^2 + C^2 + \left(\frac{N}{\mathrm{E}_{\mathrm{ECAL}}}\right)^2 . \tag{3.4}$$

The first term in Eq. 3.4 covers fluctuations in photon yield and lateral shower containment, whereas the second term estimates the contribution of leakage through crystals, inter-calibration errors and non-uniformity of longitudinal light collection. Finally, the third term is responsible for noise contributions from electronics, digitisation and pileup. A test beam of electrons with momenta between 20 GeV to 250 GeV measured these values as $S = 0.028$ GeV$^{1/2}$, $N = 0.12$ GeV and $C = 0.003$ [41].

3.7 Hadronic calorimeter

The second piece of calorimetry in the CMS apparatus is the hadronic calorimeter (HCAL) [41, 50], without which the measurement of energies of hadronic jets would not be possible. The HCAL provides coverage up to $|\eta| = 5$ and is divided into four sub-detectors: the hadronic barrel (HB), endcap (HE), outer (HO) and forward (HF) calorimeters. This sub-system is illustrated in Figure 3.7.

The HB and HE are both sampling calorimeters, covering pseudo-rapidity regions of $|\eta| < 1.3$ (HB) and $1.3|\eta| < 3.0$ (HE), which consist of brass absorber plates positioned parallel and perpendicular to the beam axis, respectively. The choice of brass for the absorber material was driven by the non-magnetic property of brass and by its short interaction length of 16.42 cm, allowing the design to remain compact. The absorber plates are layered with plastic scintillator tiles of size (0.087×0.087) in (η, ϕ) for $|\eta| < 1.6$ and (0.17×0.17) beyond $|\eta| = 1.6$. The thickness of absorber material varies from 5.82 interaction lengths to 10.6 interaction

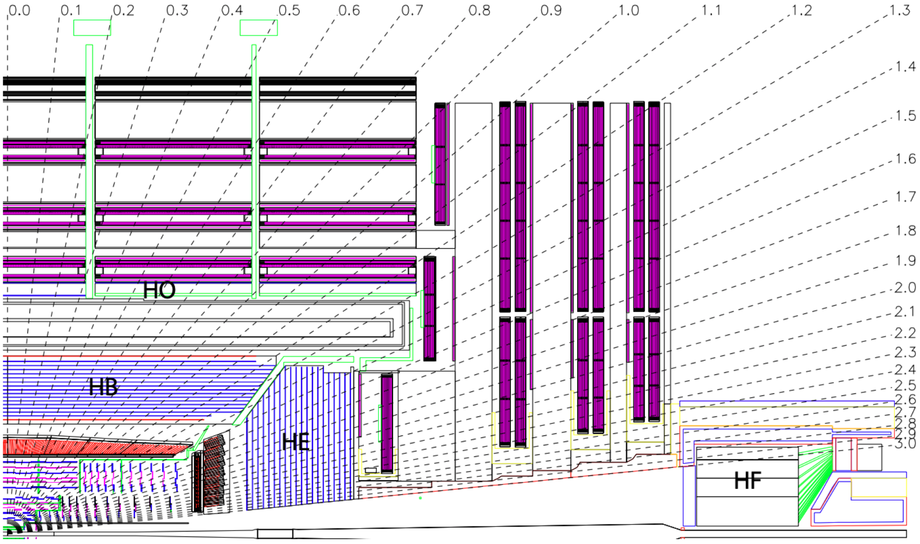

Figure 3.7: A schematic view illustrating the HCAL detector of the CMS experiment. Figure reproduced from [41]

.

length. Light is collected using wavelength-shifting fibres and hybrid photo-diodes provide the read-out for each scintillator tile.

In order to provide further containment for hadronic showers of low pseudo-rapidity, the HO is placed outside of the solenoid and extends the amount of absorber material to a minimum of 11.8 interaction lengths. The same material and read-out methods are used for this sub-detector, as was the case for the HB and HE. Studies have shown that effects of shower leakage are decreased, which lead to improved measurements of $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ [41].

The HF sub-detector uses slightly different technology, due to the large radiation flux in the forward regions. Therefore, a radiation-hard active medium is required, which is why quartz fibres inserted into a steel absorber structure was chosen. Plates with a thickness of 10 interaction length of absorber material are used for the absorber structures. Half of the quartz fibres cover the whole length of the absorber, whereas half are placed 22 cm into the HF. Hadrons will deposit approximately the same energy in both halves, however, electrons and photons will accumulate most of their energy in the first 22 cm, allowing for easier distinction between the different types of particles. The charged particles emitted from the generated showers create Cherenkov light in the quartz fibres, which are collected and read out by photo-multiplier tubes.

Similarly to the ECAL, the energy resolution is parametrised in the following manner:

$$\left(\frac{\sigma_{\mathrm{HCAL}}}{\mathrm{E}_{\mathrm{HCAL}}}\right)^2 = \left(\frac{S}{\sqrt{\mathrm{E}_{\mathrm{HCAL}}}}\right)^2 + C^2 \,. \tag{3.5}$$

In this case, the stochastic factor S and constant C has been measured to be $S = 0.943\ \mathrm{GeV}^{1/2}$ and $C = 0.084$.

3.8 Muon system

The final sub-detector found outside of the solenoid is the muon system [41, 51]. The presence of a strong magnetic field and its iron return yoke are responsible for delivering precise momentum determination and triggering capability. Together with the ability of identifying muons with a low fake rate, these are the main objectives of this sub-detector. For muon identification, three different types of technology based on gaseous detectors are used: drift tube (DT) chambers, cathode strip chambers (CSC) and resistive plate chambers (RPC). Collectively, these sub-systems cover a pseudo-rapidity range up to $|\eta| = 2.4$.

The DT chambers are placed in the barrel region and cover up to $|\eta| = 1.2$. There are in total four stations situated between the layers of flux return plates. The DT chambers consist of rectangular drift cells, each that have a cross-section of 13 mm × 42 mm and 2.4 m long sensitive wires immersed in a mixture of argon and carbon dioxide gas. As muons enter the chambers and ionise the gas, electrons drift across to the wire and the signal is then picked up. Four cells make up a superlayer, and superlayers are oriented both parallel and perpendicular to the beam axis to provide measurements of the muon position in the ϕ and the z-direction, respectively. The spacial resolution was found to be 77 µm to 123 µm in ϕ and 133 µm to 393 µm [52].

The second muon sub-system, found in the endcaps and covering $0.9 < |\eta| < 2.4$, is series of multiwire proportional chamber modules, the CSCs. These modules contain six planes of anode wire, creating six layers of gas mixture. Four CSC stations are placed between the endcap layers of the return yoke. Cathode strips provide a measurement of the ϕ coordinate, whereas the wires measure the position in the η. The resolution in the $(r - \phi)$ is 40 µm to 120 µm [53].

Finally, a set of RPCs is available in the region up to $|\eta| = 1.6$, which enhances triggering capabilities due to its fast response time. Within each RPC there are two parallel plates, creating the anode and cathode, each with a layer of gas. There are six layers in the barrel return yoke and four layers in the endcap return yoke. The signal from the ionisation of the gas is read out using aluminium strips that are parallel to beam axis. The spatial resolution is measured to be 0.78 cm to 1.38 cm, which is significantly lower than that of the other muon sub-systems.

3.9 Trigger system

As the collision rate at the LHC is about 40 MHz, collecting all the events is impossible. Therefore a way of reducing the rate of data to be saved needs to be put in place, which has been achieved using the trigger system. Any event selected for physics purposes needs to be accepted by the Level-1 (L1) hardware trigger and the High-Level-Trigger (HLT). The L1 trigger is based on a set of front-end electronics, making a fast decision within about 3 µs, whereas the HLT runs subsequently over a computing farm of processors and is therefore able to do a more sophisticated data selection decision. The rate is reduced to $\mathcal{O}(100\,\text{kHz})$ by the L1 trigger, and finally to about $\mathcal{O}(1\,\text{kHz})$.

The L1 trigger [54] is a time-multiplexed system, such that it makes use of the relevant information on the CMS detector as the signals arise. The L1 trigger is divided into a calorimeter and muon trigger 3.8. The first layer of the calorimeter trigger (Calo Trigger Layer 1) consists of FPGA cards mapping out the energy deposits in the ECAL and HCAL sub-detectors for many bunch crossings. The FPGA cards in the next layer (Calo Trigger Layer 2) receives information on a single bunch crossing for each calorimeter. Objects are subjected to basic identification algorithms and are sorted by their transverse momenta. Finally, a list of best candidates is delivered to the global trigger.

Meanwhile, hits in the three muon sub-systems are passed to the Muon Track-Finding Layer and muon tracks are reconstructed in different (η, ϕ) regions, until in the next layer (Sorting/Merging Layer) tracks are combined in ϕ. Then the global trigger combines the output from the muon trigger with the output of the calorimeter trigger, and a decision on the event is performed.

The next layer of triggering is the HLT [55], which makes use of all information from the CMS detector. 13 000 CPU nodes are employed to perform a selection based on reconstruction algorithms used also at analysis level, as opposed to the ones used at L1 triggering. Objects at HLT level are thus of better resolution and have high identification efficiencies.

3.10 Worldwide LHC Computing Grid

In total, $\mathcal{O}(10\,\text{PB})$ of data is approved by the HLT for storage, calling for a computing infrastructure to process, store and easily make available the data for analysts. The Worldwide LHC Computing Grid [56] provides this capability through a global effort by a myriad of research

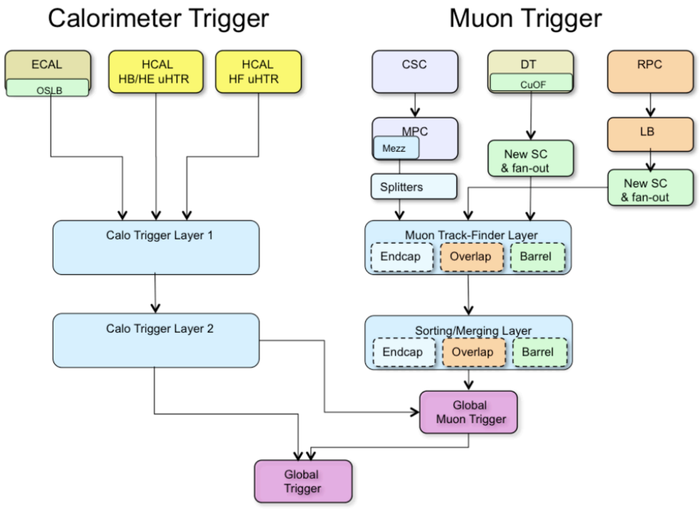

Figure 3.8: A schematic illustrating the flow of data through the Level-1 trigger [54].

institutions, which are divided into three tiers. Tier-0, where data is fully reconstructed, consists of the CERN Data Center and the Wigner Research Centre for Physics in Budapest. From this tier, data is copied and distributed to at least one Tier-1 site and, subsequently, to several Tier-2 centres, where it becomes available to analysts across the globe.

3.11 Summary

The CMS detector is part of the LHC complex, which hosts a range of particle physics experiments. The use case of the CMS detector is very varied, ranging from exotic searches of new physics beyond the standard model, to precision measurements of key interactions, such as the Drell-Yan process. Therefore the general-purpose detector is composed of multiple layers of active material to identify, track, and measure energies for all particles. The heart of the detector is the superconducting magnet. It is worth noting that for the reconstruction of a τ-lepton the full detector is required, which will become more apparent in the following Chapter. An on-side hardware trigger stores collision data classified as interesting physics events, which are filtered further by the offline trigger system. Due to the vast amount of data collected, the Worldwide LHC Computing Grid provides storage and transfer services to CMS analysts.

Chapter 4

Physics Object Reconstruction

4.1 Introduction

Having reviewed the important ingredients of the CMS detector, it is natural to discuss the reconstruction procedure of each physics object at CMS. This is necessary to understand how the candidate particles in the di-τ final states are selected and what additional selection is applied. Moreover, since τ-leptons decay into both hadronic and leptonic particles, the reconstruction becomes particularly arduous and knowledge of the full detector is required.

4.2 Tracks and vertices

The reconstruction of tracks is performed using the *combinatorial track finding* (CTF) algorithm [57], which makes use of a Kalman filter (KF) [58]. Several iterations of the CTF sequence are required to provide the final collection of reconstructed tracks. This *iterative tracking* process is split into ten different parts. The first three iterations aim to reconstruct prompt tracks originating near the pp interaction, whereas the later iterations target tracks outside of the beam-spot, the luminous area of the pp beam. After each iteration, the hits associated to a track are removed, such that the complexity of the problem is reduced, simplifying the track reconstruction process for more intricate tracks. Each iteration is performed in the following way: A seed provides the initial estimation of track candidates and their trajectories with 2 to 3 hits in the pixel detector. Through the use of a KF, extrapolation of the seed trajectories towards the expected track is performed, finding any hits that could match to the track. Another KF is employed to extract the best-fit parameters of each trajectory. Finally, further track selection is performed by applying quality cuts to reduce the fraction of fake tracks – tracks with no charged particle associated to it.

4.2.1 Electron tracking

The tracking algorithm described above can be extended to more specialised cases to facilitate the reconstruction of electrons, for instance. Electron track reconstruction efficiency suffers from radiative effects, such as bremsstrahlung and high-energy photon emission. This may lead to missing tracks and therefore an alternative adaptation of the CTF procedure can be employed. Rather than using a series of KF, a Gaussian-sum filter (GSF) can be used. This method provides a more appropriate description of the hits in the trajectory where sudden losses of energy are present. Finally, a boosted decision tree (BDT) optimises for high electron track reconstruction efficiency and low fake track rate.

4.2.2 Muon tracking

A specialised track finding procedure is also performed for the muon track reconstruction which utilises hits in the muon chambers. We can define three muon types based on the reconstruction method and quality:

standalone muon: hits in the muon system are used to seed the track finding sequence and result in a standalone-muon track.

global muon: standalone-muon tracks are combined with tracks from the inner detector, if compatible, and result in a global-muon track.

tracker muon: inner tracks with $p > 2.5\,\text{GeV}$ and $p_{\mathrm{T}} > 0.5\,\text{GeV}$ are extrapolated to the muon system and considered a tracker-muon track if a match exists.

4.2.3 Vertex reconstruction

The primary-vertex reconstruction aims at determining the positions of all interaction points in the event using the tracks reconstructed by the CTF algorithm. Tracks are selected using criteria based on the number of strips and hits associated to a track, and additional quality cuts that are consistent with the tracks being produced in the region of primary pp interactions. Clustering is performed using a deterministic annealing algorithm [59] that finds the set of tracks originating from the same interaction point. After the vertices and their associated tracks have been identified a fit is performed using the adaptive vertex fitter [60], which results in the parameter estimates of the vertex candidate, such as the position and covariance matrix, vertex quality parameters, like the number of degrees of freedom, and track weights, which indicate the probability of a track originating from the vertex. The primary-vertex reconstruction is close to 100% efficient when at least two tracks are associated to the vertex.

In some cases, the luminous region of the LHC beams, also known as the beam spot, may be used to refine the vertex reconstruction through the additional constraint. This will be used in the reconstruction of the primary vertex for the $\mathcal{CP}$ with Higgs boson decays to τ leptons.

4.3 Calorimeter clustering

The calorimeter, as discussed in Chapter 3, is responsible for detecting and measuring the energy deposits of neutral and charged particles [61]. Clustering is carried out separately in the relevant subdetectors – the barrel and endcaps in both the ECAL and HCAL and the two preshower layers. Initially cluster seeds are to be identified. These are cells with energy deposits greater than a given threshold. Additionally the deposits are required to be larger than the energy of the nearest neighbouring cells. Then topological clusters are formed using the seeds by merging cells within the cluster. The reconstruction of the clusters is then performed using a two-step iterative expectation-minimisation algorithm which is based on a Gaussian-mixture model. Energy deposits in the cells of the topological clusters are modelled as Gaussian functions in this algorithm. The parameters of these functions are determined by measuring the expected fractional energy at each cell and using this to perform a maximum likelihood fit. The expected fraction, f_{ji}, of energy, E_j, at cell position, $\vec{c}_j$, of the ith energy deposit is given by

$$f_{ji} = \frac{A_i e^{\frac{-(\vec{c}_j - \vec{\mu}_i)^2}{2\sigma^2}}}{\sum_{k=1}^{N} A_k e^{\frac{-(\vec{c}_j - \vec{\mu}_k)^2}{2\sigma^2}}}, \tag{4.1}$$

where the sum runs over the number of seed cells in the topological cluster. The parameters are determined from an analytical maximum likelihood fit with

$$A_i = \sum_{j=1}^{M} f_{ji} E_j \tag{4.2}$$

and

$$\vec{\mu}_i = \sum_{j=1}^{M} f_{ji} E_j \vec{c}_j \,, \tag{4.3}$$

where the sum is performed over all cells in the topological cluster. The fit is repeated until the expectation-minimisation algorithm reaches convergence. Thereafter, the energies and positions of the Gaussian functions are used to define the cluster parameters.

In order to describe the true energy deposited by particles in the calorimeters, a calibration procedure is performed in the hadronic and electromagnetic calorimeters.

4.4 Particle Flow

At CMS, physics objects are reconstructed using the Particle Flow (PF) algorithm [61]. This algorithm relies on combining information from the individual sub-detectors to form the electrons, muons, charged and neutral hadrons. It is especially important for the reconstruction of complex objects such as jets, hadronically decaying τ-leptons and missing transverse momenta.

The PF sequence reconstructs a particle by combining elements from different sub-detectors using the *link algorithm*. The granularity of each sub-detector and the multiplicity of particles in a unit of the solid angle present limitations to the likelihood of linking elements of a given particle. On the other hand, the probability of linking all elements is constrained by the quantity of material covered by a particle's trajectory, as a higher amount of material may result in more secondary particles and kinks in the trajectory. If two elements are linked the distance between the linked elements is used as a quality metric of the link.

A link between a track in the inner tracker and clusters in the calorimeter are produced in the following way. First the track is extrapolated from the last pixel hit position to the two layers of the preshower, the electromagnetic and hadronic calorimeters. If the extrapolated track falls within a cluster region a link is established, with the distance between the track and cluster positions defining the link distance in the $(\eta - \phi)$ plane. In the instance where several clusters are linked to the same tracks, or vice versa, the shortest distance link is saved. Photons emerging from electron bremsstrahlung are linked to GSF tracks if the extrapolated tangents to the tracks are consistent with ECAL clusters at each tracker layer. Links between tracks originating from photon conversion are formed using a dedicated conversion finder if the tracks are compatible with features of photon conversion. A link between two calorimeter clusters, for instance preshower to ECAL or ECAL to HCAL, is produced if clusters in the more granular calorimeter lies within the cluster position of the less granular calorimeter. Again, the smallest distance link is retained in the case of multiple links. Finally, for nuclear interactions, tracks are linked through a secondary common vertex if at least three tracks are present, with at most one incoming track from the primary vertex and two outgoing tracks, or three outgoing tracks. Once all links and particles are identified, the PF sequence is revisited to remove any mis-identified particles and inefficiencies. Specific particle reconstruction algorithms target the different types of physics object that are being sought and will be discussed in the following sections.

4.4.1 Muons

For global muons, additional inner tracks and calorimeter clustered in a cone of $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2} = 0.3$ around the muon are considered in order to reduce the number of muons mis-identified as hadrons. If the sum of the transverse momenta of these additional tracks and the sum of transverse energies deposited in the calorimeter clusters is less than 10% of the p_{T} of the global muon, then it is considered to be isolated:

$$\frac{1}{p_{\mathrm{T}}}\left(\left|\sum_{\substack{i\in\text{tracks}\\ \Delta R<0.3}}\vec{p}_{\mathrm{T},i}\right|+\sum_{\substack{j\in\text{clusters}\\ \Delta R<0.3}}E_{\mathrm{T},j}\right)<0.1 \tag{4.4}$$

Global muons may fail this isolation requirement and still be identified, in which case inner tracks must be matched with at least three track segments in the muon detectors or calorimeter clusters must be consistent with the muon hypothesis to reduce high-p_{T} charged hadron mis-identification.

The PF muon momentum is taken from the inner track in cases where $p_{\mathrm{T}} < 200\,\mathrm{GeV}$. For momenta higher than 200 GeV, the muon p_{T} is selected by considering the lowest χ^2 probability from each tracker fit combination: tracker only, tracker and first muon detector plane, global and global without muon detector planes of high occupancy [53].

The elements used by the PF algorithm to reconstruct and identify muon candidates are removed from the sequence to avoid multiple usage of the same elements in the reconstruction of different particle types.

4.4.2 Electrons and isolated photons

The PF reconstruction procedure proceeds with electrons, and requires knowledge of the activity in the tracker and calorimeters. Due to the high probability of bremsstrahlung by electrons in the tracker, the reconstruction of isolated photons and electrons is performed very similarly, and accurate description of the bremsstrahlung photon is required. The electron candidate is seeded by a GSF track if the associated ECAL cluster is linked to a maximum of two other tracks. The photon candidate, on the other hand, is seeded by an ECAL supercluster of $\mathrm{E_T} > 10\,\mathrm{GeV}$ that has no GSF track link. Additionally, for both types of candidates, the aggregate energy in HCAL cells that are within a distance of $\Delta R = 0.15$ must not surpass 10% of the supercluster energy. In order to properly assign the energy, all clusters linked to the supercluster or to GSF track tangents are associated to the candidate. This also covers any tracks or clusters linked to a GSF track consistent with photon conversion. The energy, after calibration, is allocated to the photon candidate, whose direction is defined by the supercluster position. Similarly, the electron candidate's energy is taken from the combined energy of the ECAL cluster and the GSF track momentum. In this case, its direction is given by the GSF track.

Finally, an additional selection is applied on electrons to ensure a high rate of correctly identified candidates. To this end, a set of BDTs is trained for the barrel and endcaps, and for isolated and non-isolated electrons separately. The input features consist of fourteen variables which include the energy radiated by the GSF track, the ratio between energy deposits in the ECAL and HCAL, the KF and GSF track χ^2, and the number of pixel hits. Conversely, photon candidates must be isolated from additional tracks and calorimeter clusters and the ratio of HCAL and ECAL energies must be compatible with typical energy distributions of photon showers.

In similar fashion as in the PF muon reconstruction, all elements used in the reconstruction of the aforementioned electrons and isolated photons are masked against being used as inputs for other particle reconstruction sequences. At this stage, the remainder of particles to be reconstructed consists of hadrons and non-isolated photons from jet fragmentation.

4.4.3 Hadrons and non-isolated photons

As the PF algorithm continues, less and less tracker hits and calorimeter clusters are available to be used for reconstruction of other particle types. With the bulk of particles in the event being reconstructed, the focus shifts towards particles emerging from hadronisation and jet fragmentation. These particles include charged hadrons (such as $\pi^{\pm}$ or protons), neutral hadrons (such as neutrons), non-isolated photons from π^0 decays, for instance, and seldom muons from early charged hadron decays.

Within tracker acceptance the charged and neutral hadrons are distinguishable. Any ECAL cluster that is not linked to a track is considered a photon. The photon hypothesis is preferred over the neutral hadron hypothesis as the jet deposits about 25% of its energy in the ECAL as photons, whereas about 3% as neutral hadrons (1% for hadronic decays of the τ-lepton due to Cabibbo-suppression). This assumption, however, does not hold anymore outside of tracker acceptance, since the energy deposited by charged and neutral hadrons amounts to about 25% in this case. Precedence is given to hadrons if ECAL clusters are linked to a HCAL cluster, otherwise the photon hypothesis stands.

Any other HCAL clusters may be linked to tracks, which may also have a link to remaining ECAL clusters. These are consistent with the single charged hadron hypothesis for each linked track. Calibration of the cluster energies is performed and the validity of the hypothesis reviewed by defining the excess deposit, δ_{T}, as the difference between the total calorimeter energy, $\mathrm{E_T} = \mathrm{E_T^{ECAL}} + \mathrm{E_T^{HCAL}}$ and the sum of the track momenta:

$$\delta_{\mathrm{T}} = \mathrm{E_T} - \left| \sum_{i \in tracks} \vec{p}_{\mathrm{T,i}} \right|, \tag{4.5}$$

where the sum runs over tracks linked to the clusters. If $\mathrm{E}_{\mathrm{T}}^{\mathrm{ECAL}} > \delta_{\mathrm{T}} > 500\,\mathrm{MeV}$, this excess is interpreted as an additional energy deposit and is assigned to a photon candidate with energy δ_{T}. The calorimeter energies are re-calibrated and the remaining excess is assigned to a neutral hadron given that $\delta_{\mathrm{T}} > 1\,\mathrm{GeV}$. The momentum and energy of the charged hadron is taken from the track, assuming that the mass is consistent with the charged-pion mass for the definition of the latter. Infrequently, it may occur that the sum of track momenta is larger than the calorimeter energy, leading to a negative excess. In such cases, a less stringent muon search, similar to the one described in section 4.4.1, is performed.

4.4.4 Jets

As described in Chapter 2, PF jets are reconstructed using the infrared- and collinear-safe anti-k_{T} algorithm with radius parameter $R = 0.4$, which is seeded by all reconstructed PF particles. Jets above 15 GeV are considered as final jets due to the unreliable reconstruction at lower momenta. The resolution of jets in the (η, ϕ) plane improves significantly by seeding jet reconstruction with PF candidates, as opposed to just using information from the calorimeters.

4.4.5 Missing transverse momentum

The presence of undetected particles, such as neutrinos, can be inferred from the imbalance in momentum perpendicular to the beam axis. This can be quantified as the missing transverse momentum, $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$, and calculated using all PF candidates in the event:

$$\vec{p}_{\mathrm{T}}^{\,\mathrm{miss,\ raw}} = -\sum_{i \in \mathrm{particles}} \vec{p}_{\mathrm{T,i}} \tag{4.6}$$

Note, Eq. 4.6 describes the uncalibrated $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$. Alternatively, there is another method in CMS that is used for reconstructing $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ which makes use of the *pileup per particle identification* (PUPPI) algorithm [62, 63], which provides a more robust description of the $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ variable against pileup, which is significant during 2016–2018 LHC data-taking operations. PUPPI $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ uses the variable α_{PUPPI}, as defined in Eq. 4.7, which is sensitive to the shape distribution differences between particles produced by jet hadronisation processes in genuine QCD mechanisms and those emerging from pileup vertices.

$$\alpha_{\mathrm{PUPPI},i} = \log \sum_{\substack{j \neq i \\ \Delta R_{ij} < 0.4}} \left(\frac{p_{\mathrm{T,j}}}{\Delta R_{ij}} \right) \begin{cases} \text{for } |\eta| < 2.5, & j \in \text{ charged PF particles from primary vertex} \\ \text{for } |\eta| > 2.5, & j \in \text{ all PF particles,} \end{cases} \tag{4.7}$$

For every neutral PF candidate $\alpha_{\mathrm{PUPPI},i}$ is determined, which is used to define χ^2_{PUPPI}.

$$\chi^2_{\mathrm{PUPPI},i} = \frac{(\alpha_{\mathrm{PUPPI},i} - \bar{\alpha}_{\mathrm{PU}})^2}{(\alpha^{\mathrm{RMS}}_{\mathrm{PUPPI},i})^2} \tag{4.8}$$

The probability that the PF particle is from pileup is defined as the cumulative distribution of χ^2_{PUPPI}:

$$w_{\mathrm{PUPPI},i} = F_{\chi^2,\mathrm{NDF}=1}\chi^2_{\mathrm{PUPPI},i} \tag{4.9}$$

The cumulative distribution function, $F_{\chi^2,\mathrm{NDF}=1}$, serves as an approximation to $\chi^2_{\mathrm{PUPPI},i}$ with one degree of freedom, NDF, for all PF particles. Thus, $w_{\mathrm{PUPPI},i} = 0$ is interpreted as a likelihood that particle i emerged from a pileup vertex, whereas $w_{\mathrm{PUPPI},i} = 1$ indicates that particle i is a product of the primary vertex interaction. These weights are calculated for each particle in the event and used to re-scale their four-momentum individually. This procedure is performed before jet clustering, such that the inputs to the $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ calculation are adjusted. Therefore the calculation of $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ using the PUPPI method is performed in the same way as described in Eq. 4.6 but with PUPPI-reweighted PF candidates.

4.5 Hadronic taus

The physics objects that require knowledge of all the previously mentioned ones are τ-leptons. These may decay both leptonically and hadronically due to their relatively high mass of 1.78 GeV. Since their decays are accompanied by neutrinos, $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ needs to be well reconstructed, too. The reconstruction of leptonic decays of the τ-lepton have been covered in sections 4.4.1 and 4.4.2. This section will outline the methods used to reconstruct and efficiently identify its decays to hadronic particles.

The major decay modes of the τ-lepton are outlined in Table 4.1. For the hadronic decays only the ones relevant to the τ_h reconstruction algorithm used in CMS, called the hadrons-plus-strip (HPS) algorithm [64], are shown. These are also illustrated in Fig. 4.1 and 4.2.

The HPS algorithm considers all jets with $p_{\mathrm{T}} > 14\,\mathrm{GeV}$ and $|\eta| > 2.5$ as possible τ_h-lepton candidates. The different decay modes tabulated in Table 4.1 are then reconstructed. Many of these have π^0 mesons in the final states, either through an intermediate ρ or a_1 resonance, or directly from the τ_h-lepton decay. The π^0 mesons decay into γ pairs, which in turn have a high probability of undergoing e^+e^- pair conversion in the tracker. The presence of the CMS magnet allows the tracks of the e^+e^- pairs to bend, leading to a separation in ϕ and η, especially the former. At this stage, in order to reconstruct the π^0 meson, all of the e/γ candidates from the τ_h-lepton decay are clustered into *strips* within some region $\Delta\eta \times \Delta\phi$. Due to the possibility

Table 4.1: Hadronic and leptonic decay modes of the τ-lepton are indicated as well as any intermediate resonance relevant to specific decays. The charged hadron, $h^{\pm}$, typically represents a $\pi^{\pm}$ or $K^{\pm}$. The approximate branching fractions, $\mathcal{B}$, are taken from the particle data group (PDG) [12].

Decay mode		$\mathcal{B}$ [%]
$\tau^- \rightarrow e^-\bar{\nu}_e\nu_\tau$		17.8
$\tau^- \rightarrow \mu^-\bar{\nu}_\mu\nu_\tau$		17.4
All leptonic decays		35.2
$\tau^- \rightarrow h^-\nu_\tau$		11.5
$\tau^- \rightarrow \rho\nu_\tau$	$\rho \rightarrow h^-\pi^0\nu_\tau$	25.9
$\tau^- \rightarrow a_1\nu_\tau$	$a_1^{1\text{pr}} \rightarrow h^-\pi^0\pi^0\nu_\tau$	9.5
	$a_1^{3\text{pr}} \rightarrow h^-h^+h^-\nu_\tau$	9.8
$\tau^- \rightarrow h^-h^+h^-\pi^0\nu_\tau$		4.8
Remaining hadronic decays		3.3
All hadronic decays		64.8

of multiple scattering and bremsstrahlung by the e/γ, the isolation efficiency of the τ_h is not always maximised if the strip size were kept fixed in terms of $\Delta\eta \times \Delta\phi$. Additionally in cases where the momentum of the τ_h-lepton is large, its constituents tend to travel in the direction of the original τ_h-lepton, resulting in the need of a smaller strip size. The standard $\mathrm{H} \rightarrow \tau\tau$ analysis selection cuts, in general, favour boosted τ_h-leptons, which improves the efficiency in this case. This motivates the use of a dynamic strip reconstruction that allows the strip size to be adjustable within a given range. The reconstruction algorithm of strips is performed in the following way:

- The e/γ candidate that has no strip associated to it yet and is highest in p_T is used to seed a new strip with its initial position centred at the seed candidate.
- The candidate with second highest p_T within the following strip window is then merged into the strip:

$$\begin{aligned}\Delta\eta &= f(p_\mathrm{T}^{e/\gamma}) + f(p_\mathrm{T}^{\text{strip}})\\ \Delta\phi &= g(p_\mathrm{T}^{e/\gamma}) + g(p_\mathrm{T}^{\text{strip}}).\end{aligned} \tag{4.10}$$

In Eq. 4.10, $p_\mathrm{T}^{e/\gamma}$ is the transverse momentum of the e/γ candidate to be merged to the strip, whereas $p_\mathrm{T}^{\text{strip}})$ is the strip's transverse momentum prior to the addition of the new candidate and is defined as the vector sum of its constituents. The functions $f(p_\mathrm{T})$ and

$g(p_{\mathrm{T}})$ have been empirically determined as

$$\begin{aligned} f(p_{\mathrm{T}}) &= 0.20 p_{\mathrm{T}}^{-0.66} \\ g(p_{\mathrm{T}}) &= 0.35 p_{\mathrm{T}}^{-0.71}. \end{aligned} \tag{4.11}$$

The strip size is limited to $\Delta\eta \times \Delta\phi = 0.15 \times 0.3$ as the upper limit and 0.05×0.05 as the lower limit.

- The central strip position is recalculated as a p_{T}-weighted average of all e/γ candidates within the strip.
- The reconstruction of the strip is complete once no more e/γ candidates are present within the $\Delta\eta \times \Delta\phi$ window. The sequence continues with the e/γ candidate highest in p_{T} which has no strip attributed to it yet.

Following the strip reconstruction, all possible τ_h-lepton decay hypotheses are constructed by merging the strip and charged particles. Then the compatibility with the following combinations of τ_h-lepton decay modes is checked: $h^{\pm}, h^{\pm}\pi^0, h^{\pm}h^{\mp}h^{\pm}$. All strips that have been reconstructed at this point are considered to be π^0 candidates if $p_{\mathrm{T}}^{\mathrm{strip}} > 2.5\,\mathrm{GeV}$. The visible mass of the τ_h-lepton candidate, $p_{\mathrm{T}}^{\tau_h}$, is required to be within a mass window consistent with the ρ or a_1 meson. The mass windows are defined in Table 4.2. For the decay to $h^{\pm}h^{\mp}h^{\pm}$, the secondary vertex must be within $\Delta z < 4\,\mathrm{cm}$ to be considered.

Table 4.2: Mass constraints on the visible hadronic constituents of the τ_h-lepton candidate mass which is constructed by combining the charged particle and strip from its decay. The change in the mass is defined in [64] and refers to the change due to the inclusion of e/γ candidates to the strip.

Decay mode	Mass window [GeV]
$h^{\pm}$	$0.3 - \Delta m_{\tau_h} < m_{\tau_h} < \sqrt{\frac{(1.3)^2 p_{\mathrm{T}}^{\tau_h}}{100}} + \Delta m_{\tau_h}$
$h^{\pm}\pi^0$	$0.4 - \Delta m_{\tau_h} < m_{\tau_h} < \sqrt{\frac{(1.2)^2 p_{\mathrm{T}}^{\tau_h}}{100}} + \Delta m_{\tau_h}$
$h^{\pm}h^{\mp}h^{\pm}$	$0.8 < m_{\tau_h} < 1.5$

A cone of size $0.05 < 3\,\mathrm{GeV}/p_{\mathrm{T}}^{\tau_h} < 0.10$ is associated to every τ_h-lepton candidate, within which the hadronic decay products of the tau must lie. In the case where several τ_h-lepton candidates are reconstructed for the same seed jet, the highest in p_{T} is chosen as the best candidate. A selection of major hadronic decay modes of the τ-lepton and the sub-detector in which its decay products interact in is illustrated in Fig. 4.1 and 4.2.

On top of HPS reconstruction, a new BDT-based decay-mode identification was designed [65]. The method trains separately on simulated events of τ_h-lepton decays with one track and three tracks, and uses input features related to the four-momentum of the τ_h-lepton decay

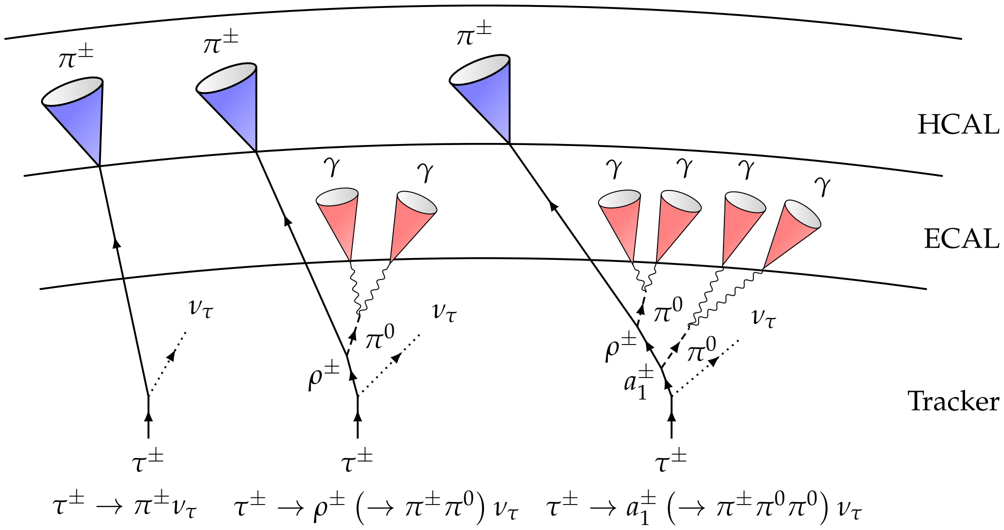

Figure 4.1: The predominant hadronic decays of the τ_h-lepton with one track. The decay products are shown, together with the relevant parts of the CMS detector. The cone shapes around the hadrons and photon candidates represent the electromagnetic and hadronic calorimeter showers. The neutrino will pass through CMS undetected and is thus indicated by a dashed line.

products, their separation in the (η, ϕ) plane, the HPS decay mode and strip properties. The most significant variables are the invariant masses of the one track or three track τ_h-lepton decay products. The training sample is a mixture of VBF and $gg\mathrm{H} \to \tau^+\tau^-$ simulated events with final state τ_h-lepton $p_\mathrm{T} > 20\,\mathrm{GeV}$. The method provides improved purity and efficiency for the main τ_h-lepton decay modes of about 15%. The motivation and impact of using this decay-mode classifier is illustrated further in Section 7.4.

In order to reduce inefficiencies and remain as close as possible to the reconstruction and selection in analyses, the PF algorithm is applied in the reconstruction of HLT objects. The τ_h-lepton decay mode algorithm (HPS) was not integrated in the τ_h-lepton HLT algorithms until 2018 data-taking. Prior to this, a cone-based approach was followed, where HLT PF jets containing at least one charged hadron seed the τ_h-lepton reconstruction [61]. The central axis of the cone is defined using the direction of the charged hadron in the jet with highest p_T. Neutral pions and at most two additional charged hadrons that are within the cone are combined to form the τ_h-lepton momentum four-vector.

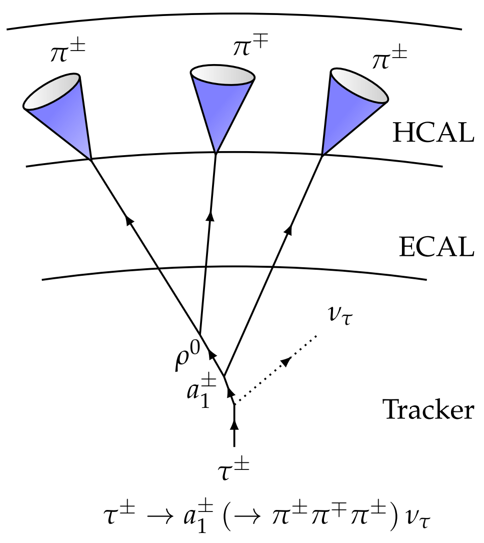

Figure 4.2: The main hadronic decay of the τ_h-lepton with three tracks. The decay products are shown, together with the relevant parts of the CMS detector. The cone shapes around the hadrons and photon candidates represent the electromagnetic and hadronic calorimeter showers. The neutrino will pass through CMS undetected and is thus indicated by a dashed line.

4.6 Mass reconstruction of full di-τ system

Due to the presence of untraceable neutrinos in the decays of the τ-lepton, the di-τ final state system contains varying amounts of missing transverse momentum. Measuring the mass of this di-τ system therefore lacks experimental accuracy, and to this end approximate methods, such as the SVFIT [66] technique, have been designed at CMS. SVFIT relies on matrix element methods and calculates the di-τ mass using the transverse momentum of the final state leptons, and the missing transverse momentum and its covariance matrix as inputs. The resulting mass distributions has a resolution of 15–20%, and enables reasonable separation between background from $Z/\gamma^* \to \tau^+\tau^-$ and Higgs boson production signals. This is highlighted in Figure 4.3. The di-τ mass is thus a key variable to separating signal events from background processes.

4.7 Summary

The physics object reconstruction is based on the particle flow algorithm which combines information from the different parts of the detector to provide the most probable particle list hypothesis and their associated four-momenta. The presence of neutrinos is inferred by the lack of transverse momentum balance in the full CMS detector system. The missing transverse

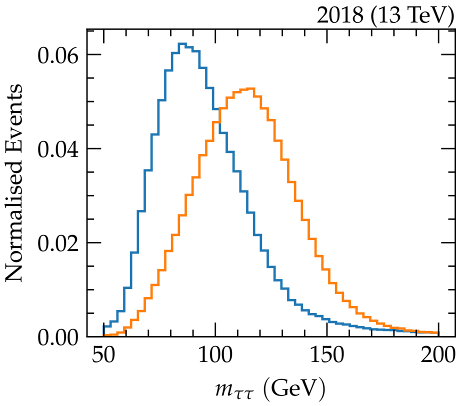

Figure 4.3: Comparison of SVFIT mass distribution for $gg \to \mathrm{H}$ and $\mathrm{Z}/\gamma^* \to \ell\bar{\ell}$ events in the $\tau_\mu\tau_h$ final state.

momentum is a key ingredient for analysis involving τ-leptons as their decay products involve at least one neutrino. The reconstruction of the hadronic τ-lepton decay products is performed using the hadron-plus-strip algorithm. Finally the full di-τ final state is reconstruction using the SVFIT algorithm, which combines lepton four-momenta and missing transverse energy using matrix element techniques. The mass of this system is a powerful variable as it separates di-τ events from Z and Higgs boson production processes.

Chapter 5

Event Selection

5.1 Introduction

As there are many common selection criteria in the analyses that will be discussed, it is convenient to outline the ingredients that are shared between them. These include any corrections and the basic selection applied to the physics objects reconstructed using the methods outlined in Chapter 4, and the procedures used to estimate the backgrounds.

5.2 Datasets

The analyses discussed in this Thesis make use of the data collected at CMS during 2016, 2017 and 2018, which corresponds to 35.9 fb^{-1}, 41.5 fb^{-1} and 59.7 fb^{-1}, respectively. In order to select events of interest for analyses where $\mathrm{H} \rightarrow \tau^{+}\tau^{-}$, a dedicated set of triggers are used. These need to be able to cover events with all combinations of electrons, muons and taus in their final states to achieve the best sensitivity in $\mathrm{H} \rightarrow \tau^{+}\tau^{-}$ interactions. The four *channels* or *final states* that are generally used in these analyses are denoted as $\tau_h\tau_h$, $\tau_\mu\tau_h$, $\tau_e\tau_h$ and $\tau_e\tau_\mu$, where the subscript on the τ-lepton represents the type of particle it decays to. Additionally, sideband regions are defined with events that are orthogonal to the signal region and may be used for cross-checking the modelling of data, for instance, or to conduct separate studies with an unbiased dataset. In $\mathrm{H} \rightarrow \tau^{+}\tau^{-}$ these usually are events with $\mathrm{Z}/\gamma^{*} \rightarrow \mu^{+}\mu^{-}$ and $\mathrm{Z}/\gamma^{*} \rightarrow e^{+}e^{-}$.

For the fully-hadronic final state, a trigger targeting two τ_h-leptons with at least 35 GeV is used. The semi-leptonic final states $\tau_\mu\tau_h$ and $\tau_e\tau_h$ use muon and electron triggers, respectively. For $\tau_\mu\tau_h$ this consists of a requirement of a single muon with p_{T} of at least 22 GeV in 2016 data-taking and 24 GeV in 2017 and 2018 data-taking. Additionally, a trigger specifically targeting a muon and hadronically decaying τ-lepton was employed with thresholds of $p_{\mathrm{T}}^{\mu} > 19\,\mathrm{GeV}$ and $p_{\mathrm{T}}^{\tau_h} > 20\,\mathrm{GeV}$ for 2016, and $p_{\mathrm{T}}^{\mu} > 20\,\mathrm{GeV}$ and $p_{\mathrm{T}}^{\tau_h} > 27\,\mathrm{GeV}$ for 2017 and 2018. Similarly, for the $\tau_e\tau_h$ final state, datasets collected with triggers requiring at least one electron or an electron and a τ_h-lepton are used. For the 2016, 2017 and 2018 datasets, the single-electron trigger requires an electron of $p_{\mathrm{T}}^{e} > 24\,\mathrm{GeV}$, $p_{\mathrm{T}}^{e} > 27\,\mathrm{GeV}$ or $32\,\mathrm{GeV}$ and $p_{\mathrm{T}}^{e} > 32\,\mathrm{GeV}$ or $35\,\mathrm{GeV}$, respectively, depending on the data-taking runs. The triggers that select an electron and τ_h-lepton candidate have p_{T} thresholds of $p_{\mathrm{T}}^{e} > 24\,\mathrm{GeV}$ and $p_{\mathrm{T}}^{\tau_h} > 20\,\mathrm{GeV}$ in 2016, and $p_{\mathrm{T}}^{e} > 24\,\mathrm{GeV}$ and $p_{\mathrm{T}}^{\tau_h} > 30\,\mathrm{GeV}$ in both 2017 and 2018. Finally, for the fully-leptonic final state involving a $\tau_e\tau_\mu$ pair the trigger p_{T} requirements are $p_{\mathrm{T}}^{e} > 12\,\mathrm{GeV}$ and $p_{\mathrm{T}}^{\mu} > 23\,\mathrm{GeV}$, or $p_{\mathrm{T}}^{e} > 23\,\mathrm{GeV}$ and $p_{\mathrm{T}}^{\mu} > 8\,\mathrm{GeV}$. The analyses discussed in later chapters focus on measurements performed solely with the most sensitive $\tau_h\tau_h$ and $\tau_\mu\tau_h$ channels.

5.3 Monte Carlo simulation

There is a vast amount of processes that are occur at the LHC and therefore need to be simulated. The proton-proton simulated collisions are used to model events with signal and background candidates[1]. As introduced in Chapter 2, Monte Carlo (MC) techniques are used to generate processes according to their matrix elements in perturbation theory, in addition to evolving the event using parton shower generators. Additionally, the detector response to the final state particles is generated to ensure a good description of data recorded at the CMS experiment.

The MC samples for background processes are produced using the following methods and tools. The $\mathrm{W}+\text{jets}$ and Drell-Yan $\mathrm{Z}/\gamma^*(\rightarrow \ell\bar{\ell})+\text{jets}$ are modelled using MadGraph5 [67] at LO accuracy. MLM jet matching [68] is applied to correctly combine inclusive events with 0-4 outgoing partons without double counting. Dedicated samples for events with exactly one, two, three or four partons are also generated to increase statistics. These are merged with the inclusive samples at analysis level using event weights. Samples to describe electroweak production of W and Z bosons are also generated at LO accuracy using MadGraph5. The single-t and $t\bar{t}$ processes are modelled at NLO accuracy using the Powheg generator [69, 70, 71]. Finally, diboson processes (WW, ZZ, WZ) are simulated using MadGraph5_aMC@NLO at NLO accuracy. The FxFx jet merging scheme is used to account for underlying processes with different parton multiplicities [72].

The parameters that were chosen to generate these are summarised in the following paragraph. The parton distribution functions NNPDF3.0 [73] and NNPDF3.1 [74] are used in the generation of 2016, 2017 and 2018 samples. Parton showering, hadronisation and decays of the τ-lepton are generated using Pythia8.2 [75]. The interactions of partons not involved in the hard subprocess of the proton-proton collision, known as the underlying event, are simulated in Pythia8.2 with the CUETP8M1 [76] tune for 2016 and CP5 [77] tune for 2017 and 2018.

[1] For $\mathrm{H}\rightarrow\tau^+\tau^-$ analyses, background events are produced from processes that have a di-τ lepton final state which does not originate from Higgs boson production.

The detector response is simulated using GEANT4 [78]. PU collisions, which are pp interactions occurring simultaneously with the interesting hard momentum transfer, are simulated using PYTHIA8.2.

5.3.1 Embedded samples

An additional method is used to model the $\mathrm{Z}/\gamma^* \to \tau^+\tau^-$ process, including any other background events with a genuine τ-lepton pair final state, and is referred to as the embedding technique [79]. This method exploits lepton universality in $\mathrm{Z}/\gamma^*(\to \ell\bar{\ell}) + \text{jets}$ events and proceeds in the following manner.

Z/$\gamma^* \to \mu^+\mu^-$ selection: $\mathrm{Z}/\gamma^* \to \mu^+\mu^-$ events in data with two global PF muon candidates are selected. These are required to pass a di-μ trigger with p_T thresholds of $p_\mathrm{T}^{\mu_1} > 17\,\mathrm{GeV}$ and $p_\mathrm{T}^{\mu_2} > 8\,\mathrm{GeV}$ for the two muons.

Z/$\gamma^* \to \mu^+\mu^-$ cleaning: energy deposits and hits due to the muon pair are removed in the event.

Z/$\gamma^* \to \tau^+\tau^-$ simulation: a $\mathrm{Z}/\gamma^* \to \tau^+\tau^-$ event is simulated, where the kinematic properties of the τ-lepton pair are fixed to those of the muon pair that was removed from the event in the previous step. The full τ-lepton simulation is then run with PYTHIA8.2.

Merging: the cleaned $\mathrm{Z}/\gamma^* \to \mu^+\mu^-$ and simulated $\mathrm{Z}/\gamma^* \to \tau^+\tau^-$ events are merged into one by combining the hits and energy deposits of the τ-lepton pair into the cleaned event from data.

Embedded events are produced individually for each data-taking era and are found in datasets referred to as embedded samples. Although designed to cover the $\mathrm{Z}/\gamma^* \to \tau^+\tau^-$ process, 1% of events selected through the requirement of the muon pair are from $t\bar{t}$ and di-boson decays, where a W or Z boson gives rise to the muons. The embedded samples are thus a simulation method which cover any type of background event where the final state is a di-τ lepton pair.

As the majority of the event is taken directly from data, this technique makes many of the issues of MC generators irrelevant and any correction that would be applied to the same objects in MC become obsolete. Corrections to MC-simulated events are discussed in Section 5.5. The main corrections that require attention for embedded events are the lepton efficiency and energy scale corrections that will be introduced later in this Chapter. Jets, which require laborious calibration for MC simulation, do not need to be corrected in embedded events, as these come from the data directly. These points are the major benefits of using embedded samples in the analysis and is why these were adopted in $\mathrm{H} \to \tau^+\tau^-$ analyses.

5.4 Physics object selection

5.4.1 Muons

After the reconstruction of PF muons, there is a contribution of misidentified muons originating from charged hadron showers reaching the muon sub-detector. Additionally, some muons may not be from *prompt* decays (of the Z and W boson) and may originate from heavy hadrons. In order to reject these muons, which are not of interest for H → $\tau\tau$ analyses, identification (ID) and isolation requirements are imposed. The muon has to be a global or tracker muon and satisfy the identification requirements listed in Table 5.1 – these correspond to the *medium* muon ID for CMS analyses and are based on the quality of the track and muon reconstruction algorithms. The pileup mitigating particle-based relative isolation $I_{\Delta\beta}$ is introduced to further remove real muons from in-flight hadron decays:

$$I_{\Delta\beta} = \frac{1}{p_{\mathrm{T}}} \left| \sum_{i\in\, \mathrm{h}^{\pm}} \vec{p}_{\mathrm{T},i} + \max\left(0, \sum_{i\in\, \mathrm{h}^{0}} \vec{p}_{\mathrm{T},i} + \sum_{i\in\, \gamma} \vec{p}_{\mathrm{T},i} - \frac{1}{2} \sum_{i\in\, \mathrm{h}^{\pm}_{\mathrm{PU}}} \vec{p}_{\mathrm{T},i} \right) \right| , \tag{5.1}$$

where the sums are over all charged hadrons ($\mathrm{h}^{\pm}$), neutral hadrons (h^{0}), photons (γ) and charged hadrons from PU vertices ($\mathrm{h}^{\pm}_{PU}$) within a cone of size $\Delta R = 0.4$ centred around the muon. The factor of 1/2 accounts for the ratio of neutral- to charged-particle ratio originating from PU vertices in the isospin limit. Finally, a selection is imposed on the transverse, d_{xy}, and longitudinal, d_z, impact parameter with respect to the primary vertex.

For muons with $p_{\mathrm{T}} > 20\,\mathrm{GeV}$, the efficiency of this ID selection is above 98%. At the same time, the mis-identification rate is below 0.1% [80].

5.4.2 Electrons

In order to identify genuine electrons efficiently from misidentified electrons from jets, photon conversions and heavy-flavour decays, requirements are imposed on the quality of the track, shower and energy deposits in the calorimeters. As opposed to muons, a more sophisticated choice is made: the selection is made on the output of a BDT with input variables related to the shape of the cluster, the quality of track fits, and the ratios between calorimeter energies. In addition, a relative isolation is defined similarly as for the muon in Eq. 5.1, however for electrons an alternative method is used for the PU subtraction based on the event-specific average energy density per unit area in the (η, ϕ) plane, ρ, and the effective area (EA) of the electron. The η-dependent effective areas are measured in bins of η and average to about 0.9.

Table 5.1: Criteria for muons identified for selection in the analysis. All muons must be reconstructed with the PF algorithm and subsequently global muons or tracker muons are used. The selection on p_T depends on the Higgs boson decay channel of interest. This will be revisited in the analysis-specific chapters. Additionally, the selection on $_{\Delta\beta}$ is loosened for the case when events are vetoed which is a channel-dependent selection performed at a later stage in the analysis.

	Tracker muon	Global muon
Fraction of valid tracker hits	-	> 0.8
Normalised global-track χ^2	-	< 3
Tracker-standalone position match χ^2	-	< 12
Kink finder χ^2	-	< 20
Segment compatibility χ^2	> 0.451	> 0.303
Additional selection		
p_T		> 15 GeV
$\lvert\eta\rvert$		< 2.4
$I_{\Delta\beta}$		< 0.15
d_{xy}		< 0.045 cm
d_z		< 0.2 cm

The relative isolation for an electron with transverse momentum p_T is defined as

$$I_{\text{EA}} = \frac{1}{p_T} \left| \sum_{i\in\,\text{h}^\pm} \vec{p}_{\text{T},i} + \max\left(0, \sum_{i\in\,\text{h}^0} \vec{p}_{\text{T},i} + \sum_{i\in\,\gamma} \vec{p}_{\text{T},i} - \rho A_{\text{eff}}\right) \right| . \tag{5.2}$$

Table 5.2: Criteria for electrons identified for selection in the analysis. All electrons must be reconstructed with the PF algorithm.

	Selection
Electron ID	True
p_T	> 25 GeV
$\lvert\eta\rvert$	< 2.1
I_{EA}	< 0.15 [2]
d_{xy}	< 0.045 cm
d_z	< 0.2 cm

For the training of the BDT, a $Z/\gamma^*(\rightarrow \ell\bar{\ell}) + \text{jets}$ MC sample generated with MADGRAPH5_AMC@NLO was used. The two training classes to be differentiated are prompt electrons and non-prompt electrons (including reconstructed electrons not matched to a generator-level electron), whereas electrons from τ-lepton decays are not used anywhere. The BDT is trained in three regions of η to optimise the efficiency for the inner barrel ($|\eta|$ < 0.8), outer barrel ($|\eta|$ <1.44) and endcaps (1.57<$|\eta|$ <2.5), and for p_{T} > 10 GeV. The chosen work point for the ID is selected as the 90% electron ID efficiency working point with about 2-5% misidentification rate. The final selection applied for electrons is listed in Table 5.2. The choice of using the 90% efficient (*loose*) working point of the electron ID together with a cut on the relative isolation I_{EA} was proposed by a study performed using 2017 data collected in the $\tau_e\tau_h$ channel. The two options studied are the *loose* and *tight* working point, which correspond to an efficiency of 90% and 80%. The metric used in the study is the asymptotic median significance, also known as the AMS:

$$\mathrm{AMS} = \sqrt{2(s+b)\log\left(1+(s/b)\right) - s}\,, \tag{5.3}$$

where s and b are the expected signal and background yields from simulation in each bin. The two working points were studied with the appropriate corrections applied for each scenario (more on this topic later, see Section 5.5. A mass window around the peak of the $\mathrm{H} \to \tau^+\tau^-$ signal was chosen using the visible (ignoring neutrinos) mass of the $\tau_e\tau_h$ pair between 50 GeV $< m_{\tau_e\tau_h} <$ 110 GeV. As illustrated in Figure 5.1, the optimum choice for the ID and isolation working points are the *loose* ID with an additional isolation cut of $I_{\mathrm{EA}} < 0.15$. In addition to this, the yields in a more VBF-like phase-space can be examined to see if the trend is similar. A selection of at least two jets with $p_{\mathrm{T}} >$ 30 GeV and invariant mass of at least 300 GeV was applied and the result is consistent with the findings in Figure 5.1: an improvement of about 10% in the significance has been found. Therefore this selection has been adopted by the H$\tau\tau$ group at CMS on signal electrons.

5.4.3 Jets

The jet collection generated using the clustering algorithm is passed through to the analysis but will be subjected to corrections before selection. Firstly, in order to suppress contribution from pileup, charged hadrons associated to pileup vertices are removed from the particle list used in the jet clustering algorithm. This is referred to as *pileup charged hadron subtraction* (CHS) [62]. Then, jet energies are calibrated in order to match to the jet energies at particle level. The corrections are derived in stages such that they are functions of the transverse momentum of the corrected four-momentum at the subsequent stage. They are applied as factors to the uncorrected four-momentum of the jet, and in the following order [81]:

Offset correction: this correction removes electronics noise and PU effects. These are estimated from the average energy density, ρ, and the jet area, A, similarly to the definition for PU subtraction for the relative electron isolation (see Equation 5.2) [82]. Contributions from the underlying event (UE) of the hard process have to be added back as

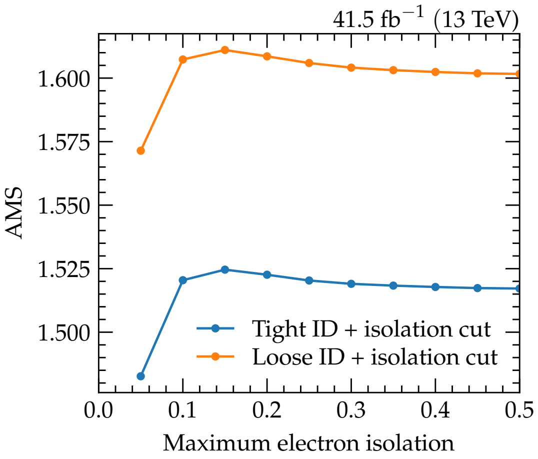

Figure 5.1: The significance of signal expected signal events with respect to background events for two working points of the electron ID selection BDT output as a function of the upper bound on the relative electron isolation. The optimum choice is found when using the *tight* working point for the ID together with an isolation cut of $I_{\mathrm{EA}} < 0.15$.

these are included in ρ. Hence the energy subtracted from the jet is $(\rho - \langle\rho_{\mathrm{UE}}\rangle) \cdot A$. The four-momentum after applying the offset correction may be written as $p'_\nu(p_{\mathrm{T}}^{\mathrm{raw}}) = C_{\mathrm{offset}}(p_{\mathrm{T}}^{\mathrm{raw}}) \cdot p_\nu^{\mathrm{raw}}$.

MC calibration and relative residual correction: the MC correction factor is to match reconstructed jet energies to generator-level energies. The correction is estimated from the ratio of the transverse momenta of reconstruction-level jets and generator-level jets, $p_{\mathrm{T}}^{\mathrm{reco}}/p_{\mathrm{T}}^{\mathrm{gen}}$, in bins of $p_{\mathrm{T}}^{\mathrm{gen}}$, where the reconstruction-level jet needs to be a maximum distance of $\Delta R = 0.25$ from the generator-level jet. The MC calibration factor is then defined as the inverse of the ratio: $C_{\mathrm{MC}} = p_{\mathrm{T}}^{\mathrm{gen}}/p_{\mathrm{T}}^{\mathrm{reco}}$.
The relative residual correction is subsequently applied and is responsible for flattening the response as a function of η. It is determined using a di-jet balancing method, where a pair of jets with $|\eta| < 1.3$ (*tag*) and with $|\eta| > 0$ (*probe*) separated by $|\Delta\phi| > 2.7$ are selected. These jets are thus approximately back-to-back, such that they should have equal, but opposite transverse momentum vectors. The balance may be defined as $B = \frac{p_{\mathrm{T}}^{\mathrm{probe}} - p_{\mathrm{T}}^{\mathrm{tag}}}{p_{\mathrm{T}}^{\mathrm{ave}}}$, where the denominator is the average p_{T} of the tag and probe jets. The correction is calculated in bins of $p_{\mathrm{T}}^{\mathrm{ave}}$. It is given as $C_{\mathrm{rel}} = \frac{2+\langle B\rangle}{2-\langle B\rangle}$. The corrected four-momentum as a function of the first-stage corrected and uncorrected four-momentum is given by $p''_\nu(p'_\nu, p_{\mathrm{T}}^{\mathrm{raw}}) = C_{\mathrm{MC}}(p'_\nu, \eta)C_{\mathrm{rel}} \cdot p_\nu^{\mathrm{raw}}$.

Absolute jet energy response correction: the purpose of this correction is to calibrate the response to flatten it in p_{T}. Using simulated events of the $\mathrm{Z} + \mathrm{jets}$ and $\gamma + \mathrm{jets}$ processes, the p_{T} distribution of the Z/γ balances exactly the p_{T} of the outgoing hadrons in the interactions at particle level. Once detector effects are included at reconstruction level, this momentum balance breaks and gives rise to $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$. The measured $p_{\mathrm{T}}^{\mathrm{miss}}$ and $p_{\mathrm{T}}^{\mathrm{Z}/\gamma}$ is then used to correct the p_{T} response of the jets. The final corrected four-momentum of the jet is then given by $p_{\nu}^{'''}(p_{\mathrm{T}}^{\mathrm{raw}}) = C_{\mathrm{abs}}(p_{\nu}^{''}) \cdot p_{\nu}^{\mathrm{raw}}$.

At this stage, the energy scales of the jets should be calibrated such that kinematic distributions in data are relatively well modelled in simulation. However, the energy resolution is not necessarily in good agreement between simulation and data, which impacts many other distributions. Similarly as for the jet energy calibration, a p_{T}-balancing method is employed such that the resolution is measured in both data and MC and a correction is derived based on the ratio of the two. The conclusions of such measurements show that the jet energy resolutions is worse in data than in simulation, such that a smearing function is applied on the four-momentum of jets in MC such that their p_{T} distributions match.

Jets are subjected to the requirements listed in Table 5.3, which corresponds to the *tight* PF jet ID at CMS and is 98% efficient. Additionally, to further reduce contamination of jets originating from pileup vertices and jets arising from calorimeter noise, the *loose* PU jet ID is applied on all jets with uncorrected $p_{\mathrm{T}} < 50\,\mathrm{GeV}$ [62]. This BDT-based discriminant uses the input features related to track and jet shape variables. As the region of discussion is outside of tracker acceptance, only the jet shape variables give the discriminating power between PU/noise jets and genuine jets. The use of this BDT in $\mathrm{H} \to \tau^+\tau^-$ analyses was motivated from a study performed on 2017 data, where jets in the region of $2.65 < |\eta| < 3.139$ are affected by noise due to problems in the ECAL sub-detector. This opened an investigation into the use of PU jet ID to see if there will be an improvement in the agreement between simulation and data for jets in the problematic η region, and not deteriorating the expected acceptance of Higgs boson signal events.

The study was performed using data in a side-band region using $\mathrm{Z}/\gamma^* \to \mu^+\mu^-$ events. These events are selected by considering muon pairs of opposite charge that are separated by at least ΔR =0.5. The muons must pass the aforementioned muon identification requirement, have $p_{\mathrm{T}} > 10\,\mathrm{GeV}$ and $I_{\Delta\beta} < 0.15$. The single-muon trigger with p_{T} threshold of 24 GeV is only required to be fired by the leading muon.

Several combinations of options have been studied to mitigate the issue in the pseudo-rapidity region: applying the PU jet ID to jets at the different existing working points defined internally by the CMS collaboration (ranging from *loose* to *medium* and *tight*) and the removal of jets with $p_{\mathrm{T}} < 50\,\mathrm{GeV}$ found in the noisy region, which defines the most extreme mitigation technique. The issue is illustrated in the top left plot of Figure 5.2, as a clear overestimation in the bin covering the problematic pseudo-rapidity region is visible in the left sub-figure.

The selection on the output score of the PU jet ID at the loose working point is defined in Table 5.4. It is applied only on jets with uncorrected p_{T} less than 50 GeV. The result after

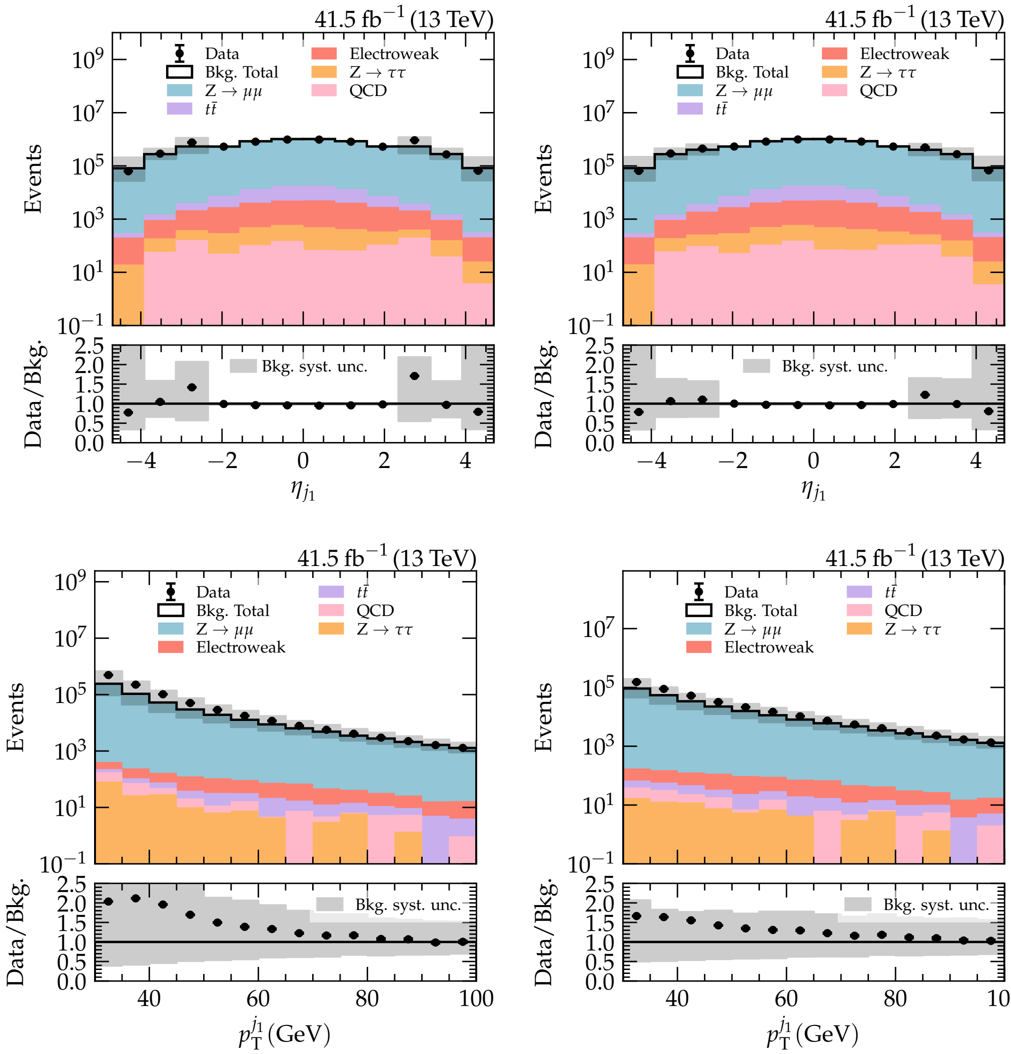

Figure 5.2: Pseudo-rapidity and p_T of the leading jet with all jets (left) and with the PU jet ID applied in the region of noise in 2017 data (right). The agreement between data and simulation is enhanced with the application of the PU jet ID.

application is shown in the top right plot of Figure 5.2, where the agreement between data and simulation is well enhanced in the problematic bin. The modelling of the jet with highest p_T (leading jet) also improves in the region of $p_T < 50\,\text{GeV}$. In addition, several key variables used in the analysis, such as the number of jets in the event and the invariant mass of the leading two jets m_{jj}, also display improved modelling after the PU jet ID is applied, as illustrated in Figure 5.5. These variables are essential for separating the different production processes of $H \to \tau^+\tau^-$ interactions: the topology of VBF events implies high m_{jj}. The loss in signal acceptance was tested and found to be only a few percent, which made this an attractive choice. Nevertheless, with the latest JEC applied, this improvement is less pronounced. It was decided to use both the cut of $p_\mathrm{T} > 50\,\mathrm{GeV}$ in the noise region and the PU jet ID inclusively in η.

The PU jet ID score was also studied in each run period of 2017 data-taking separately. In order to study this directly in data, $\mathrm{Z}/\gamma^* \to \mu^+\mu^- + 1$ jet events were split into two categories:

- **Genuine jet enriched**: jet recoils against the Z and the transverse momenta of the objects balance. This translates into the following requirements: $|\Delta\phi(p_\mathrm{T}^\mathrm{Z}, p_\mathrm{T}^j)| > 2.5$ and $0.5 < p_\mathrm{T}^j / p_\mathrm{T}^\mathrm{Z} < 1.5$.
- **Pile-up/noise jet enriched**: jet and Z are close in azimuth, that is, $|\Delta\phi(p_\mathrm{T}^\mathrm{Z}, p_\mathrm{T}^j)| < 1.5$.

In addition, to extrapolate the number of pile-up jets from the PU jet enriched region to the genuine jet region, a phase space factor of $\frac{\pi-2.5}{1.5}$ is used. The third category that is defined is the **subtracted PU** region, where the PU jet enriched distribution multiplied by the phase space factor is subtracted off the genuine jet distribution. All of this is performed in a mass window around the Z mass peak and in the pseudo-rapidity region of the ECAL noise.

In early run periods (B-D) the noise effect was less pronounced, which is also seen in the distributions of Figure 5.3. The raw PU jet ID score peaks towards the left PU/noise jets, whilst for genuine jets it peaks towards the right at positive one. The fractional content in the first five bins for the PU/noise jet enriched distribution and subtracted PU jet regions are 48% and 6%, respectively. The loose working point removes those bins, which is a low loss of genuine jets for a significant removal of PU/noise jets. In the last runs of 2017 data-taking (E-F), the noise problem was very present in the ECAL endcaps. Figure 5.4 illustrates the distributions for PU/noise and genuine jet regions as before for these run period datasets. The genuine jet histograms now peak with the PU/noise jets. The fractional contents of the first five bins are 64% of PU/noise jets and 13% for the subtracted PU distributions. This represents a 33% increase for the PU/noise jet enriched distribution and a 116% gain for the subtract PU distribution, when compared to run B. These studies confirm that the PU jet ID is a tool that can be used to discriminate against not only PU jets, but also jets due to noise in the detector.

b-tagged jets

Typically, the processes of interest with genuine τ-lepton candidates are initiated by quark or gluon jets. However, $t\bar{t}$ + jets production, where each top decays into a b-quark and a W boson, can be large backgrounds in $\mathrm{H} \to \tau^+\tau^-$ analyses. The b-quark hadronises to form a jet (b-jet) which can be utilised to reduce this background by imposing a veto on events with b-jets.

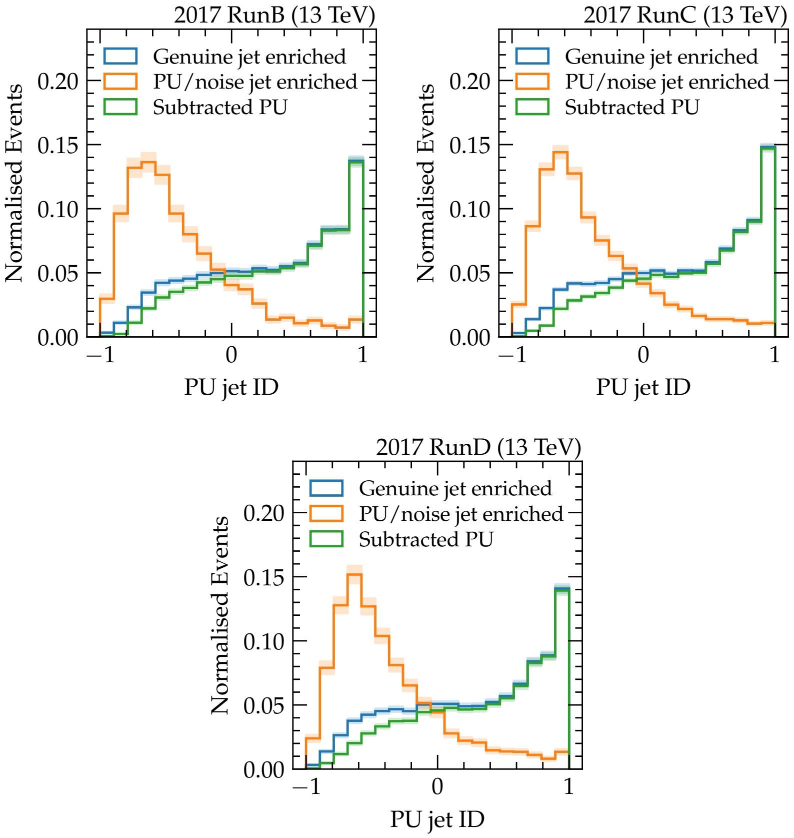

Figure 5.3: Pileup jet ID score distributions for 2017 data-taking runs B-D. The noise issue in the ECAL sub-detector is not strongly present yet. The genuine and noise jet distributions are well separable.

The tagging of b-jets relies on the properties of the tracks and vertex position of b-hadrons, therefore this is only possible for jets within the acceptance region of the tracker ($|\eta| < 2.4$). The b-hadrons have a lifetime of 1.5 ps, such that they can travel as far as 1 cm from the primary vertex. Therefore, its decay products have tracks that appear to originate from a secondary vertex, as can be seen in Figure 5.6. The secondary vertex is reconstructed using the inclusive vertex finding algorithm [83] seeded by all tracks with $p_{\mathrm{T}} > 0.8\,\mathrm{GeV}$, a 3-dimensional impact parameter greater than 50 µm and a 2-dimensional impact parameter significance of at least 1.2. The algorithm proceeds with the following steps [83]:

- *Track clustering*: tracks compatible with the seed track are grouped together depending on requirements on their spatial and angular separation.

- *Secondary vertex fitting and cleaning*: the adaptive vertex fitter, introduced in Section 4.2.3, is used to measure the position of the secondary vertex using the clustered tracks

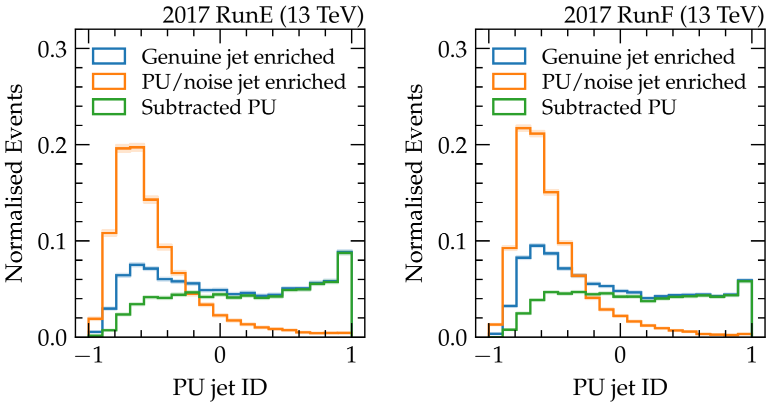

Figure 5.4: Pileup jet ID score distributions for 2017 data-taking runs E-F. The noise issue in the ECAL sub-detector is more pronounced in these datasets. The genuine jets overlap more with PU/noise jets than in earlier runs.

- *Track arbitration*: in order to settle the ambiguity of whether a track originates from the primary or secondary vertex, a track is removed as a secondary vertex track based on two conditions: if its distance to the secondary vertex and secondary vertex flight direction is larger than its absolute impact parameter and $\Delta R > 0.4$, respectively.

- *Secondary vertex refitting and cleaning*: the secondary vertex is refitted after the track arbitration process. If, after this, less than 80% of its tracks are shared with another secondary vertex and the significance of flight distance is larger than 10, the secondary vertex is kept. Finally, any secondary vertex is associated to jets if the distance between the vertex and jet axis in the (η, ϕ) plane is less than 0.3.

A multi-class deep neural network (NN) is trained on simulated $t\bar{t}$ + jets and multi-jet events with jets of $20\,\text{GeV} < p_\text{T} < 1\,\text{TeV}$ and $|\eta| < 2.4$ to discriminate jets with different hadron flavour constituents: exactly one b-hadron, at least two b-hadrons, no b-hadrons and exactly one c-hadron, no b-hadrons and at least two c-hadrons, or none of the previous combinations. Included as input features are track- and vertex-based variables, such as the flight distance significance of the secondary vertex or the track impact parameter, which have discriminating power between b-jets and light-flavour jets. In order to tag jets containing b-hadrons, the NN outputs for the two training classes with b-hadrons are summed together to result in an effective b-tag probability. This is known as the DeepCSV algorithm at CMS. The medium working point is used in $\text{H} \to \tau^+\tau^-$ analyses to consider a jet as b-tagged, which requires a combined b-tag probability of at least 0.6321, 0.4941 or 0.4184 for the 2016, 2017 or 2018 datasets, respectively. The b-tag identification efficiency for this working point is 68% for a misidentification probability of 1% [83].

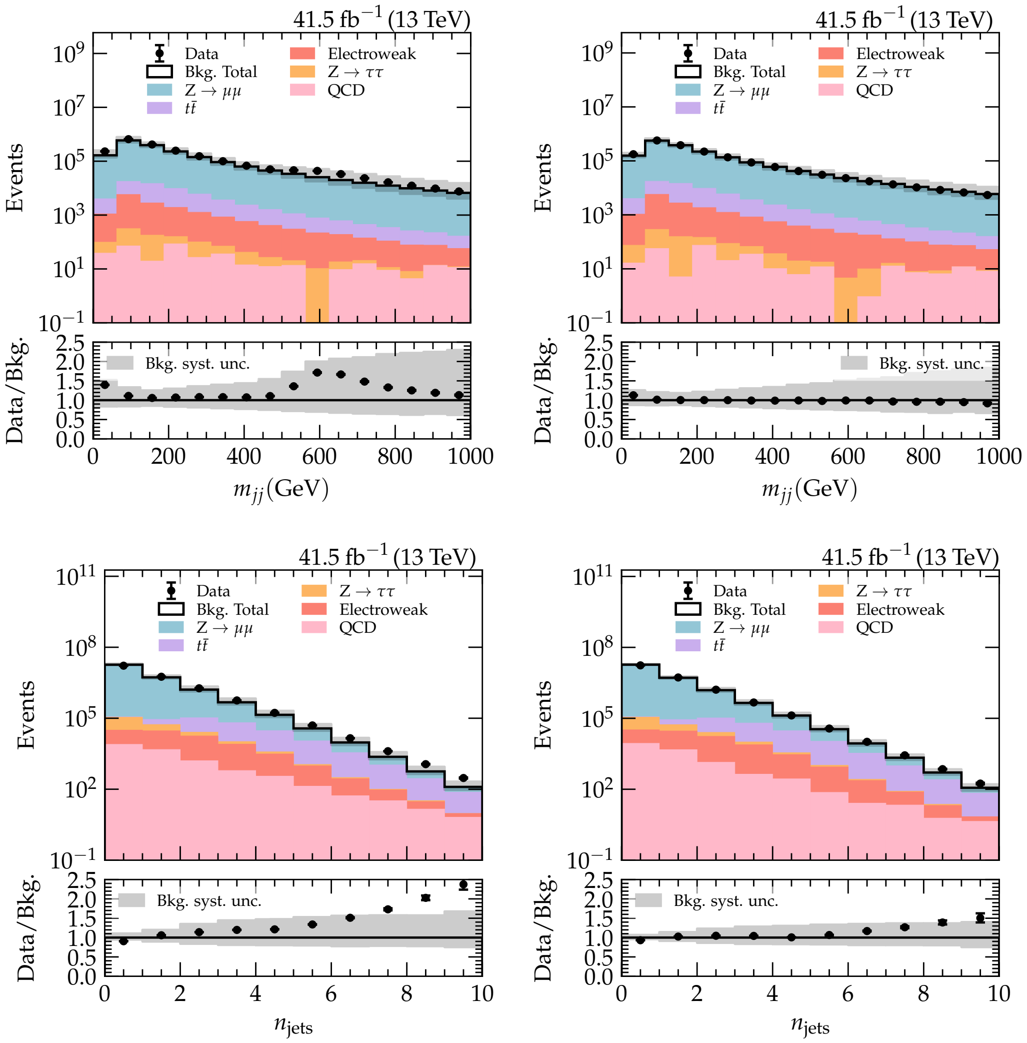

Figure 5.5: Invariant mass of leading two jets and number of jets in the event with all jets (left) and with the PU jet ID applied in the region of noise in 2017 data (right). The agreement between data and simulation is enhanced with the application of the PU jet ID.

5.4.4 Missing transverse momentum

The missing transverse momentum, as already introduced in Section 4.4.5, is an extremely important quantity to measure as it enters the calculation of the Higgs boson candidate mass in $\mathrm{H} \to \tau^{+}\tau^{-}$ analyses (see Section 4.6). The $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ measurement is affected by inefficiencies in the tracker, and non-linear response and minimum energy thresholds of the calorimeters [84]. To account for these in the measurement of $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$, the corrections applied to the jets are propagated

Table 5.3: Criteria for PF jet ID selection in the analysis. All jets are reconstructed with the anti-k_T algorithm. Jet selection depends on its position in η: central ($|\eta| < 2.4$ [or $|\eta| < 2.6$ for 2018]), near endcaps ($2.4 < |\eta| < 2.7$ [or $2.6 < |\eta| < 2.7$ for 2018]), far endcaps ($2.7 < |\eta| < 3.0$) and forward ($3.0 < |\eta| < 5.0$). This is to accommodate for the presence of the tracker in the central region, where discrimination between charged and neutral particles is available. Additionally, in forward regions the radiation levels are higher, thus requiring a tighter cut on the number of neutral particles. Multiple lines indicate the logical AND of the two selection criteria. In CMS the requirements listed are equivalent to the *tight* PF jet ID.

	Central			Near endcaps		
	2016	2017	2018	2016	2017	2018
Charged hadron energy fraction	> 0	> 0	> 0	–	–	–
Charged particle multiplicity	> 0	> 0	> 0	–	–	> 0
Charged EM energy fraction	< 0.99	–	–	–	–	–
Neutral hadron energy fraction	–	–	< 0.9	< 0.9	< 0.9	< 0.9
Neutral EM energy fraction	–	–	< 0.9	< 0.9	< 0.9	< 0.9
Constituent multiplicity	–	–	> 1	> 1	> 1	–
	Far endcaps			Forward		
	2016	2017	2018	2016	2017	2018
Neutral hadron energy fraction	< 0.98	> 0.02	–	–	< 0.99	> 0.2
Neutral EM energy fraction	< 0.98	> 0.02 < 0.99	> 0.02 < 0.99	–	< 0.9	< 0.9
Neutral particle multiplicity	> 2	> 2	> 2	> 10	> 10	> 10

Table 5.4: Selection applied on the output of the PU jet ID as a function of pseudo-rapidity range and uncorrected jet p_T. Jets passing the selection must have PU jet ID score greater than the value stated. This corresponds to the *loose* PU jet ID.

Pseudo-rapidity range	$20 < p_\mathrm{T}^\mathrm{raw} < 30\,\mathrm{GeV}$	$p_\mathrm{T}^\mathrm{raw} > 30\,\mathrm{GeV}$
$\lvert\eta\rvert < 2.5$	-0.97	-0.89
$2.5 < \lvert\eta\rvert < 2.75$	-0.68	-0.52
$2.75 < \lvert\eta\rvert < 3.0$	-0.53	-0.38
$3.0 < \lvert\eta\rvert < 5.0$	-0.47	-0.30

using

$$\vec{p}_\mathrm{T}^{\,\text{miss, corr}} = \vec{p}_\mathrm{T}^{\,\text{miss, raw}} - \sum_{i \in \text{jets}} \left(\vec{p}_\mathrm{T,i}^{\,\text{corr.}} - \vec{p}_\mathrm{T,i}^{\,\text{raw}} \right) , \tag{5.4}$$

where $\vec{p}_\mathrm{T,i}^{\,\text{corr.}}$ and $\vec{p}_\mathrm{T,i}^{\,\text{raw}}$ refer to the corrected and uncorrected transverse momentum of the i-th jet in the sum.

Table 5.5: Criteria for jet selection in the analysis. All jets are reconstructed with the anti-k_{T} algorithm and are subjected to the requirements imposed in this table. The PF jet ID criteria is listed in Table 5.3, whereas the BDT-based PU jet ID criteria are shown in Table 5.4. The requirement on the isolation variable $\Delta R(\vec{p}^{\,j}_{\mathrm{T}}, \vec{p}^{\,\tau\tau}_{\mathrm{T}})$ ensures that none of the leptons selected in the $\tau\tau$ final state overlap with any jet. The $\tau\tau$ candidate pair selection is detailed in section 5.4.6.

	Selection
PF jet ID	True
PU jet ID	True
p_{T}	$> 30\,\mathrm{GeV}$ OR ($> 50\,\mathrm{GeV}$ AND $2.65 < \lvert\eta\rvert < 3.139$)
$\lvert\eta\rvert$	< 4.7
$\Delta R(\vec{p}^{\,j}_{\mathrm{T}}, \vec{p}^{\,\tau\tau}_{\mathrm{T}})$	< 0.5

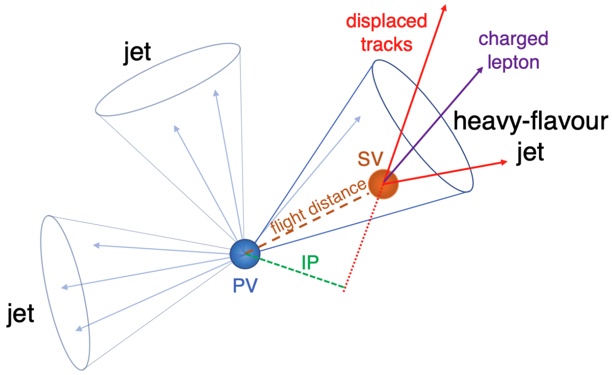

Figure 5.6: Schematic of heavy-flavour jet tagging, which is used to tag b-quark initiated jets. Figure taken from [83].

5.4.5 Hadronic taus

Following the reconstruction of the hadronic τ_h-lepton through its decay products, the misidentification rate of quark or gluon jets as τ_h-leptons is large and can result in lower signal efficiency. Introducing a requirement on the τ_h-lepton candidate isolation significantly decreases this fake τ_h-lepton background. The isolation is defined in Eq. 5.5 and is a combination of the charged and neutral particle isolation sums [64]:

$$I_{\tau_h} = \left| \sum_{\substack{i \in \text{charged} \\ \Delta R < 0.5}} \vec{p}_{\mathrm{T},i}(d_z < 0.2\,\mathrm{cm}) - \max\left(0, \sum_{\substack{i \in \gamma \\ \Delta R < 0.5}} \vec{p}_{\mathrm{T},i} - \Delta\beta \sum_{\substack{i \in \text{charged} \\ \Delta R < 0.8}} \vec{p}_{\mathrm{T},i}(d_z > 0.2\,\mathrm{cm})\right) \right| \tag{5.5}$$

In Eq. 5.5, the first term sums the transverse momenta of charged particles within a cone of size $\Delta R = 0.5$ centred around the direction of the τ_h-lepton candidate. Additionally the charged particle is required to have a longitudinal impact parameter from the τ_h-lepton vertex less than 0.2 cm to reduce contribution from PU. Similarly, the second sum is the neutral isolation for neutral particles in the same cone around the τ_h-lepton direction. The PU contribution to the neutral isolation term is estimated with the final term in Eq. 5.5, where the transverse momentum charged particles not originating from the τ_h-lepton candidate production vertex and with angular distance smaller than 0.8 from the τ_h-lepton direction are subtracted. This sum is weighted by a factor $\Delta\beta = 0.2$ to account for the different cone sizes and fraction of neutral to charged energy.

Previous methods of reducing the fake τ_h-lepton background consisted of a BDT with the aforementioned isolation variables and τ_h-lepton lifetime dependent features. The set of variables used in the MVA-based identification are tabulated in Table 5.6. In recent developments, the τ_h-lepton identification has been improved by using a more modern machine learning technique, a convolutional neural network (CNN) [85]. This CNN is trained on the simulated events for all major sources of fake τ_h-lepton backgrounds, such as $\mathrm{Z}/\gamma^*(\rightarrow \ell\bar{\ell}) + \mathrm{jets}$, $t\bar{t} + \mathrm{jets}$ and $\mathrm{W} + \mathrm{jets}$, with τ_h-lepton candidates of $20\,\mathrm{GeV} < p_\mathrm{T} < 1\,\mathrm{TeV}$, $|\eta| < 2.3$ and $d_z < 0.2\,\mathrm{cm}$. The input features to the CNN range from high-level features used during τ_h-lepton reconstruction, for instance four-momenta of particle candidates from the τ_h-lepton signal and isolation cone, and low-level features from the CMS sub-detectors, such as information from the inner tracker, calorimeters and muon system, whenever available. Grids of cells are defined in the (η, ϕ) plane: 11×11 and 21×21, each cell of size 0.02×0.02 and 0.05×0.05, for the signal and isolation cones respectively. In each cell, the highest p_T object is chosen and features are divided into three blocks: e/γ, hadrons and μ. This results in a total of over 100 000 features used globally in the CNN training. The performance with respect to the BDT-based identification shows an increase in yield of about 20% in genuine τ_h-lepton background and decrease of 23% in backgrounds with fake τ_h-lepton contribution.

5.4.6 Pair selection and additional vetoes

In the analyses of interests, the signal consists of a Higgs boson decaying to a pair of τ-leptons. Therefore in each analysis, after the baseline trigger, identification and kinematic selections, a specific set of requirements is applied, which will be further discussed in Chapters 6 and 7. Following these criteria, a set of lepton vetoes is applied, such that events with additional electrons or muons are removed. This ensures that individual channels are orthogonal to each other. More relevant for the analyses, in the $\tau_\mu\tau_h$ channel, events with two oppositely charged global muons with $p_\mathrm{T} > 15\,\mathrm{GeV}$ and $I_{\Delta\beta} < 0.3$. In order to select the most likely τ_h-lepton

Table 5.6: Input features for the BDT to identify genuine τ_h-lepton candidates and reduce the fake contribution. The first set of variables are related to the contents of the signal and isolation cones of the τ_h-lepton candidate, whereas the second group of features includes tracker information, which is sensitive to the (finite) lifetime of the τ_h-lepton.

Isolation-based variables
Terms in the τ_h-lepton isolation, defined in Eq. 5.5
Multiplicity of e/γ candidates in the signal and isolation cones with $p_{\mathrm{T}} > 0.5\,\mathrm{GeV}$
$p_{\mathrm{T}}^{\mathrm{strip,outer}} = \left\lvert \sum_{\substack{i\in e/\gamma \\ \Delta R > R_{\mathrm{sig}}}} \vec{p}_{\mathrm{T}}^{\,e/\gamma} \right\rvert$
p_{T}-weighted ΔR, $\Delta\eta$ and $\Delta\phi$ of e/γ candidates in strips
τ_h-lepton related variables
Signed 3-dimensional impact parameter of highest p_{T} track and its significance
Distance between τ_h-lepton production and decay vertices (if secondary vertex found), and its significance
p_{T}, η and HPS decay mode of τ_h-lepton candidate

candidate forming part of the dilepton pair, the most isolated τ_h-lepton is chosen. For the $\tau_h\tau_h$ channel, the two candidates are sorted according to leading τ_h-lepton isolation.

5.5 Simulation corrections

In order to ensure that all kinematic distributions are well modelled in simulation, a set of corrections is applied to the different objects. The corrections may be split into three types:

> **MC generator**: corrections due to mismodelling of the event topology, especially due to higher-order effects.
>
> **Efficiency**: corrections due to selections applied on objects. These are derived by measuring the efficiency in both data and simulation, and taking the ratio of the two to define a scale-factor as $\mathrm{SF} = \frac{\epsilon(\mathrm{data})}{\epsilon(\mathrm{sim.})}$.
>
> **Energy scale**: corrections due to differences in the response of the physical detector compared with the simulated detector response. The jet energy scale corrections have already been discussed in Section 5.4.3, whereas corrections of this type applied to leptons will be discussed here.

When *weights* are applied to a certain MC event, the expected event yield is multiplied by this weight. This means a weight of 1 does not change any kinematic distribution, whereas any weight deviating from 1 does. When going from event-level quantities to histograms, which average a quantity over a specified region (*bins*), these weights may introduce shape effects in different bins. Thus many of these corrections have associated *shape* uncertainties, which will be discussed in later chapters.

5.5.1 Pileup reweighting

The average pileup distribution depends on the conditions at data-taking, as is illustrated in Figure 5.7 for the 2016, 2017, and 2018 pileup profiles in data. As this depends heavily on the dataset, differences between data and simulation are corrected for in each dataset separately by deriving an event weight to match the mean number of pileup interactions. The distributions, including the scale-factor defined as the ratio of the pileup profile in data and in MC, are illustrated in Figure 5.8 for the 2018 data-taking run. The correction is derived in a high-statistics $Z/\gamma^*(\rightarrow \ell\bar{\ell}) +$jets sample and applied to all MC samples.

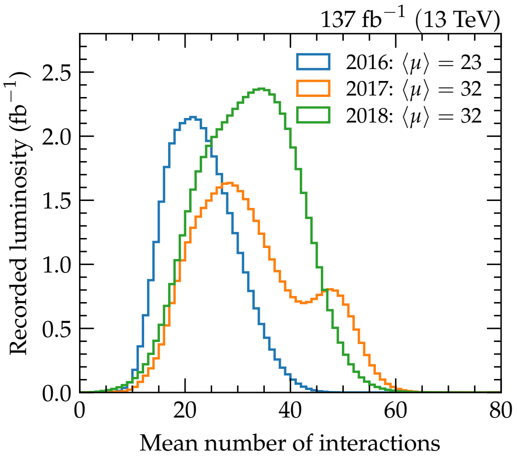

Figure 5.7: Pileup profiles in data for the 2016, 2017 and 2018 data-taking periods.

5.5.2 $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ recoil

The missing transverse momentum is corrected for a sub-set of MC samples by examining at the hadronic recoil defined as

$$\vec{U}_T = \vec{p}_{\mathrm{T}}^{\mathrm{miss}} - \sum_{i \in \mathrm{neutrinos}} p_{\mathrm{T}}^{\nu}, \tag{5.6}$$

where the sum is over all neutrinos in the event. As this cannot be directly measured using the CMS detector, events selected from the $Z/\gamma^* \rightarrow \mu^+\mu^-$ process are used. With the p_{T} of the

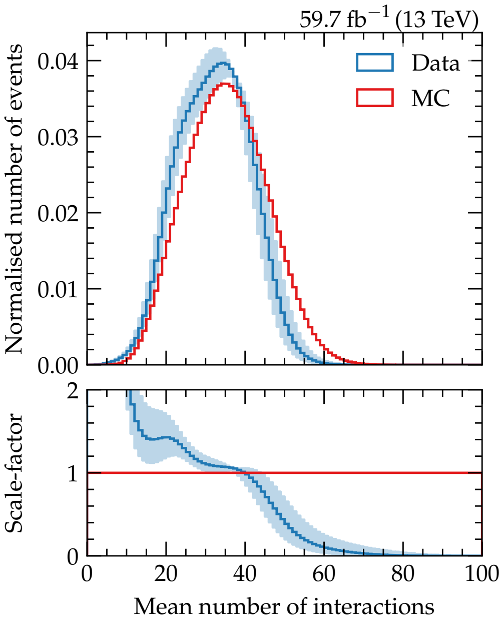

Figure 5.8: Distributions of the mean number of interactions in data and MC used to derive a scale-factor for MC. A 4.6% uncertainty variation is applied to the pileup profile in data.

di-μ system (which is equivalent to the p_{T} of the Z boson in this case) and the total momentum of recoiling jets $\vec{H}_T$, the recoil may be rewritten as

$$\vec{U}_T = -\vec{H}_T - p_{\mathrm{T}}^{\mu\mu} \tag{5.7}$$

for a leptonically decaying boson.

The perpendicular and parallel components of the recoil to p_{T}^{Z} are then fitted in both data and simulation separately using normal distributions, and the simulation is then subjected to the correction given as

$$U_{\parallel}^{\mathrm{corr.}} = \langle U_{\parallel} \rangle_{\mathrm{data}} + \left(U_{\parallel} - \langle U_{\parallel} \rangle_{\mathrm{MC}} \right) \frac{\sigma_{\mathrm{data}}(U_{\parallel})}{\sigma_{\mathrm{MC}}(U_{\parallel})} \tag{5.8}$$

$$U_{\perp}^{\mathrm{corr.}} = U_{\perp} \frac{\sigma_{\mathrm{data}}(U_{\perp})}{\sigma_{\mathrm{MC}}(U_{\perp})} \,, \tag{5.9}$$

where the mean and resolutions are found from the fitted Gaussians. The missing transverse momentum is then re-calculated taking this correction into account using Equation 5.6, where the p_{T}^{ν} is determined at generator-level.

The MC samples this correction is applied to are the $Z/\gamma^{*}(\rightarrow \ell\bar{\ell}) +$ jets, $W +$ jets and Higgs boson production samples. The $\vec{p}_{T}^{\text{miss}}$ distribution before and after applying the correction is shown in Figure 5.9.

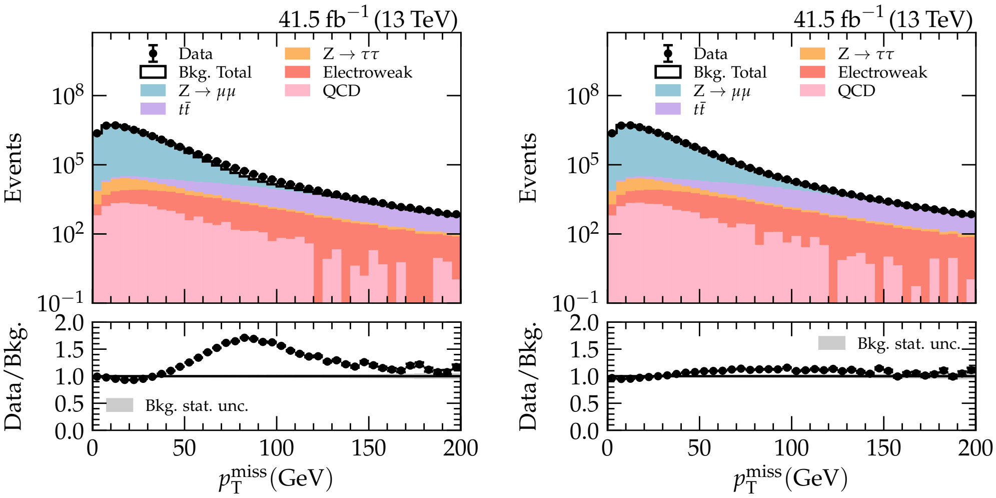

Figure 5.9: Effect of applying the recoil correction on the $\vec{p}_{T}^{\text{miss}}$ distribution in 2017 $Z/\gamma^{*} \rightarrow \mu^{+}\mu^{-}$ events. The left and right distributions show the $\vec{p}_{T}^{\text{miss}}$ before and after the correction, respectively. In this example, the large discrepancy between data and simulation in the region of 50 GeV< $\vec{p}_{T}^{\text{miss}}$ <150 GeV is largely fixed after application of the correction.

5.5.3 *t*-quark p_T reweighting

The kinematic distributions of $t\bar{t}$ events are mismodelled in simulated events. Empirically derived weights are applied to the t-quark p_T distribution to ease the disagreement between data and simulation. The weights are given by $\sqrt{\text{SF}(p_{T}^{t}) \cdot \text{SF}(p_{T}^{\bar{t}})}$, where the scale-factor, SF, is a function of the t-quark p_T and is determined from an equation with functional form $e^{a+bp_{T}+c(p_{T})^{2}}$. The coefficients were measured using RunII data and found to be $a = 0.088$, $b = 8.7 \times 10^{-4}\ \text{GeV}^{-1}$ and $c = 9.2 \times 10^{-7}\ \text{GeV}^{-2}$. The effect of the reweighting can be examined in $t\bar{t}$-production enriched region, which can selected using events in the $\tau_{e}\tau_{\mu}$ final state, based on an inversion of the cut on the topological variable D_{ζ}. Figure 5.10 shows the distribution of the p_T of the visible $\tau_{e}\tau_{\mu}$ decay products before and after the reweighting is applied.

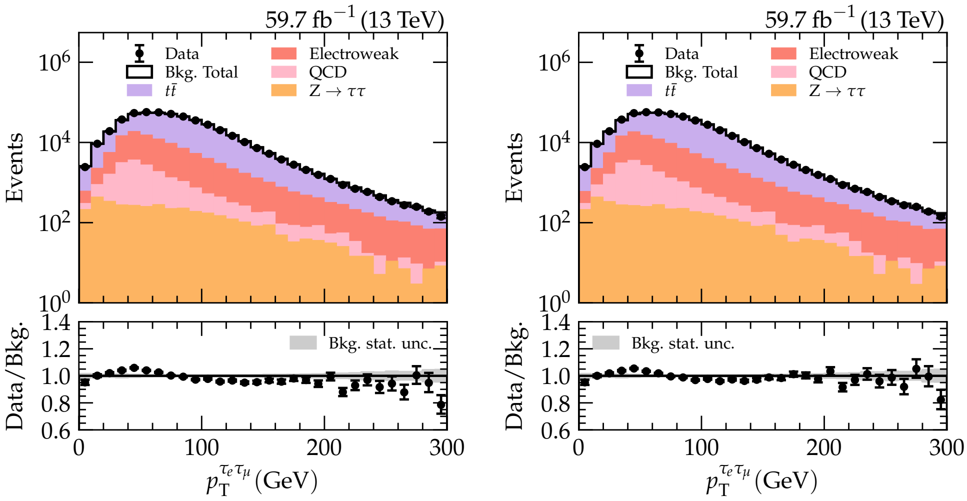

Figure 5.10: Distributions of the visible $\tau_e\tau_\mu$ p_T before (left) and after (right) applying the t-quark p_T reweighting. A $t\bar{t}$-enriched region of phase-space is selected in the $\tau_e\tau_\mu$ final state, where the correction is seen to improve the modelling of data in simulation. The grey band represents the statistical uncertainty on the background.

5.5.4 Z/γ^* p_T and mass reweighting

Events with high p_T or mass of the Z/γ^* boson in the Z/$\gamma^*(\rightarrow \ell\bar{\ell})$ + jets process are important for H → $\tau^+\tau^-$ analyses as the selection on the τ-lepton decay products prefers the phase-space with boosted Z/γ^* bosons. In the analyses, a special technique, known as τ-lepton embedding, works around these issues by using data directly, rather than simulated Z boson events. Nevertheless, in the derivation of some corrections and for the sake of a cross-check, calibrated Z/$\gamma^*(\rightarrow \ell\bar{\ell})$ + jets simluated samples are desired. However, since trends in mismodelling are present, the LO Z/$\gamma^*(\rightarrow \ell\bar{\ell})$ + jets samples are corrected by applying a reweighting as a function of the generator-level p_T and mass of the Z/γ^* boson. To this end, a correction has been derived using Z/$\gamma^* \rightarrow \mu^+\mu^-$ events in bins of the p_T and invariant mass of the di-μ system.

The weights are derived in each two-dimensional bin, i, as

$$\text{weight}_i = \frac{N_{\text{data},i} - N_{\text{MC},i}^{\text{non } Z/\gamma^*}}{N_{\text{MC},i}^{Z/\gamma^*}}\,, \tag{5.10}$$

where in the numerator the contribution of all other processes in bin i, $N_{\text{MC},i}^{\text{non } Z/\gamma^*}$, is subtracted from the number of events in data, $N_{\text{data},i}$. The denominator is the expected yield of Z/$\gamma^*(\rightarrow \ell\bar{\ell})$ + jets events in bin i as estimated from the LO samples. This correction is only to adjust the shape of the distribution, therefore the sum of the numerator across all bins and the sum of the denominator for all bins must be equal to 1:

$$\sum_i \left(N_{\text{data},i} - N_{\text{MC},i}^{\text{non Z}/\gamma^*} \right) = \sum_i N_{\text{MC},i}^{\text{Z}/\gamma^*} = 1\,. \tag{5.11}$$

Therefore the normalisation remains unaffected after the application of the weights.

An example of the final weights used for the 2018 $\text{Z}/\gamma^*(\to \ell\bar{\ell}) + \text{jets}$ sample reweighting procedure are illustrated in Figure 5.11.

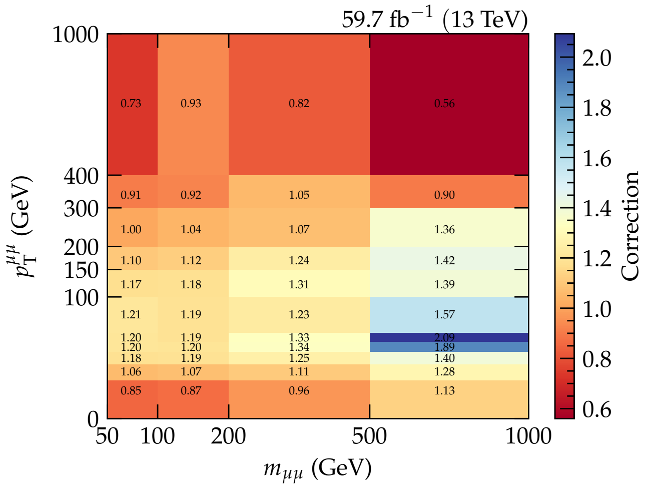

Figure 5.11: Z/γ^* p_{T}-mass correction for 2018 $\text{Z}/\gamma^*(\to \ell\bar{\ell}) + \text{jets}$ sample reweighting. The weights are derived in two-dimensional bins of di-μ p_{T} and invariant mass. Both variables are placed a square-root scale axis to facilitate the reader.

The weights are then applied as a multiplication factor to each event based on the generator-level p_{T} and mass of the Z/γ^* boson. This assumes that the reconstruction-level quantities of the di-μ p_{T} and invariant mass is similar to the generator-level ones, which is supported by the good resolution of the muon sub-detector. The impact of applying the correction is shown in Figure 5.12.

5.5.5 *b*-tagging efficiency

In order to use the b-tagging algorithm introduced in Section 4.4.4, scale-factors are applied to adjust the differences in efficiency in data and simulated events by altering the weight of selected MC events to predict the event yield in data. To achieve this, b-tagging efficiencies need to be measured for each working point used and these depend on the event kinematics

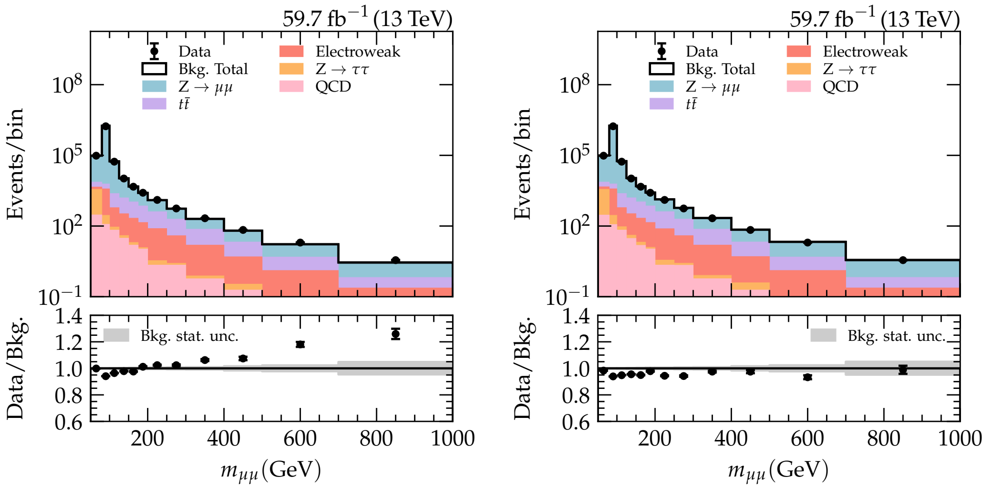

Figure 5.12: The effect of applying the Z/γ^* p_T-mass correction for 2018 $Z/\gamma^*(\to \ell\bar{\ell}) +$ jets samples. The distribution on the left shows the di-μ mass prior to the correction, whereas for the distribution on the right the reweighting procedure has been applied. The grey band represents the total statistical uncertainty on the background yield.

and selection, and the particle-level jet composition (b-tagged jet, c-tagged jet, or light-flavour-tagged jet). The efficiencies are measured in MC using $t\bar{t}$ and $Z/\gamma^*(\to \ell\bar{\ell}) +$ jets events and are displayed in Figure 5.13 for the *loose* and *medium* working point of the *DeepCSV* b-tagging algorithm. A set of probabilities is then defined using the efficiency in MC, ϵ, and scale-factor,

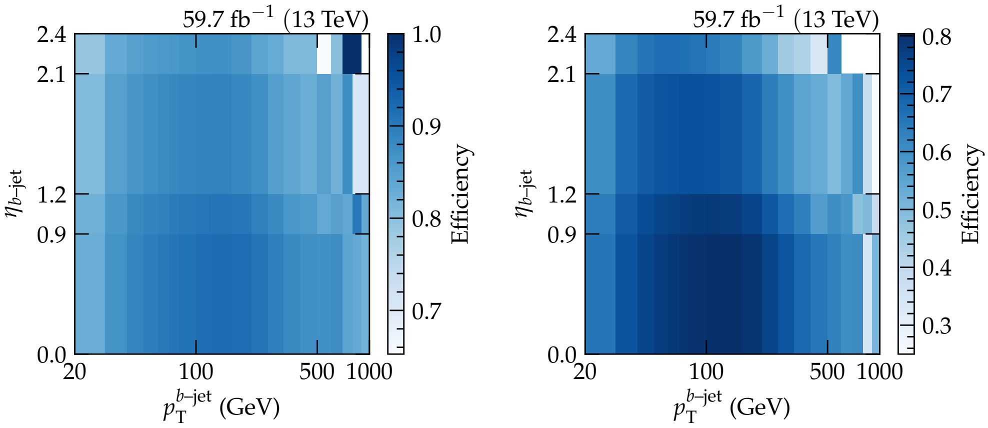

Figure 5.13: b-tag efficiencies for 2018 MC for the *loose* (left) and *medium* (right) working points of the *DeepCSV* algorithm (see Section 4.4.4). The p_T of the b-tagged jet is placed on a log-scale to facilitate the reader.

SF, in the following manner:

$$P(\mathrm{MC}) = \prod_{i \in b\mathrm{tag}} \epsilon_i \prod_{j \in \mathrm{no}\ b\mathrm{tag}} (1 - \epsilon_j) \tag{5.12}$$

$$P(\mathrm{data}) = \prod_{i \in b\mathrm{tag}} \mathrm{SF}_i \epsilon_i \prod_{j \in \mathrm{no}\ b\mathrm{tag}} (1 - \mathrm{SF}_j \epsilon_j)\ . \tag{5.13}$$

The final event weight is then derived as

$$\mathrm{weight} = \frac{P(\mathrm{data})}{P(\mathrm{MC})}\ . \tag{5.14}$$

This procedure can easily be extended to situations where multiple working points are employed, such as when using a selection based both on the *loose* and *medium* working points, which is the case for the $\mathcal{CP}$ in ggH production analysis.

5.5.6 Electron and muon efficiency

Electron and muon identification, isolation and trigger efficiencies are measured in data and simulation using the tag-and-probe technique in $\mathrm{Z}/\gamma^* \to \mu^+\mu^-$ and $\mathrm{Z}/\gamma^* \to e^+e^-$ events. They are measured separately for each data-taking year. The 2018 measurement will be the focus here, however, the technique is used analogously in 2016 and 2017 data.

The tag-and-probe method relies on identifying two candidate particles, where the nature of the chosen event is such that the tag candidate can be used to infer on the probe candidate, which is why $\mathrm{Z}/\gamma^* \to \ell\bar{\ell}$ events are used. The tag candidate must satisfy tight selections, s_T, based on identification and isolation requirements, which ensure a low misidentification rate, and fire a single-lepton trigger. The probe, on the other hand, must pass a set of very loose requirements, s_P, without the trigger selection, which keeps the probe lepton orthogonal and unbiased to the selection. The efficiency of the selection S to be measured is then given by

$$\epsilon(S|s_P) = \frac{N(S \cap s_T \cap s_P)}{N(s_T \cap s_P)}\ , \tag{5.15}$$

where $N(X)$ is the number of events passing the selection criteria X.

The tag lepton is required to pass all the kinematic, identification and isolation requirements used in the analysis. For the 2018 measurement this translates to the following for each lepton:

Electrons: pass *tight* electron ID and single-e trigger, relative isolation of $I_{\mathrm{EA}} < 0.1$, $p_{\mathrm{T}} > 36\,\mathrm{GeV}$.

Muons: pass *medium* medium ID and single-μ trigger, relative isolation of $I_{\Delta\beta} < 0.15$, $p_{\mathrm{T}} > 28\,\mathrm{GeV}$.

Additionally, the selected di-lepton pair in the $Z/\gamma^* \to \mu^+\mu^-$ or $Z/\gamma^* \to e^+e^-$ events is required to have invariant mass in the range of $65\,\text{GeV} < m_{ll} < 115\,\text{GeV}$. The efficiencies for the identification, isolation and trigger selection is measured separately. The measurements are done consecutively, such that for the isolation efficiency measurement, the probe lepton must pass the identification requirement, and for the trigger efficiency measurement, the probe lepton must pass both the identification and isolation requirements.

The measurements are done in bins of p_T and η of the probe lepton, which may pass or fail the requirement that is being measured. In order to properly count the number of events in the pass and fail regions, the invariant mass distribution is fitted. This gives a handle on the contribution of background processes to the efficiency. The signal (genuine $Z/\gamma^* \to \mu^+\mu^-$ or $Z/\gamma^* \to e^+e^-$) and background are modelled using different functional forms, depending on the efficiency being measured. The signal is modelled with a double-Voigtian PDF with the width of the Breit-Wigner PDF set to the natural width of the Z boson of 2.495 GeV, whereas the background is fitted with an error function that decays like an exponential for the identification efficiency measurement or an exponential for the isolation and trigger measurements. The turn-on regions at low invariant mass are better modelled using the error function, which is why this has been employed for the identification efficiency measurements where background contamination is usually larger than for the isolation and trigger efficiency measurements. In Figure 5.14 exemplary pass and fail regions are shown for the electron and muon identification efficiency measurements in data. Examples of the isolation and trigger efficiency measurements in simulation are shown in Figure 5.15. The efficiency, ϵ is defined as the ratio of the number of events passing the requirement to the total number of events (pass + fail). Examples of the efficiencies and final scale-factors as a function of p_T for each lepton are shown in Figure 5.16, covering the pseudo-rapidity region of $|\eta_e| < 1.0$ and $|\eta_\mu| < 0.9$ for the electron and muon measurements, respectively.

5.5.7 τ_h-lepton efficiency

The τ_h-lepton identification/isolation efficiency due to the *DeepTauID* isolation algorithm are determined for both MC and embedded simulation. Events in the $\tau_\mu\tau_h$ channel are used and the visible mass of the $\tau_\mu\tau_h$ distribution is fitted. These scale-factors are then applied to all channels using a τ_h-lepton in the pair candidate.

The trigger efficiency corrections are derived again in $Z/\gamma^*(\to \ell\bar{\ell})$ + jets events in the $\tau_\mu\tau_h$ final state, with the τ_μ-lepton candidate as the tag lepton. These were measured using dedicated μ $+\tau_h$-lepton monitoring triggers where the selections match those used in the hadronic τ-lepton trigger being measured.

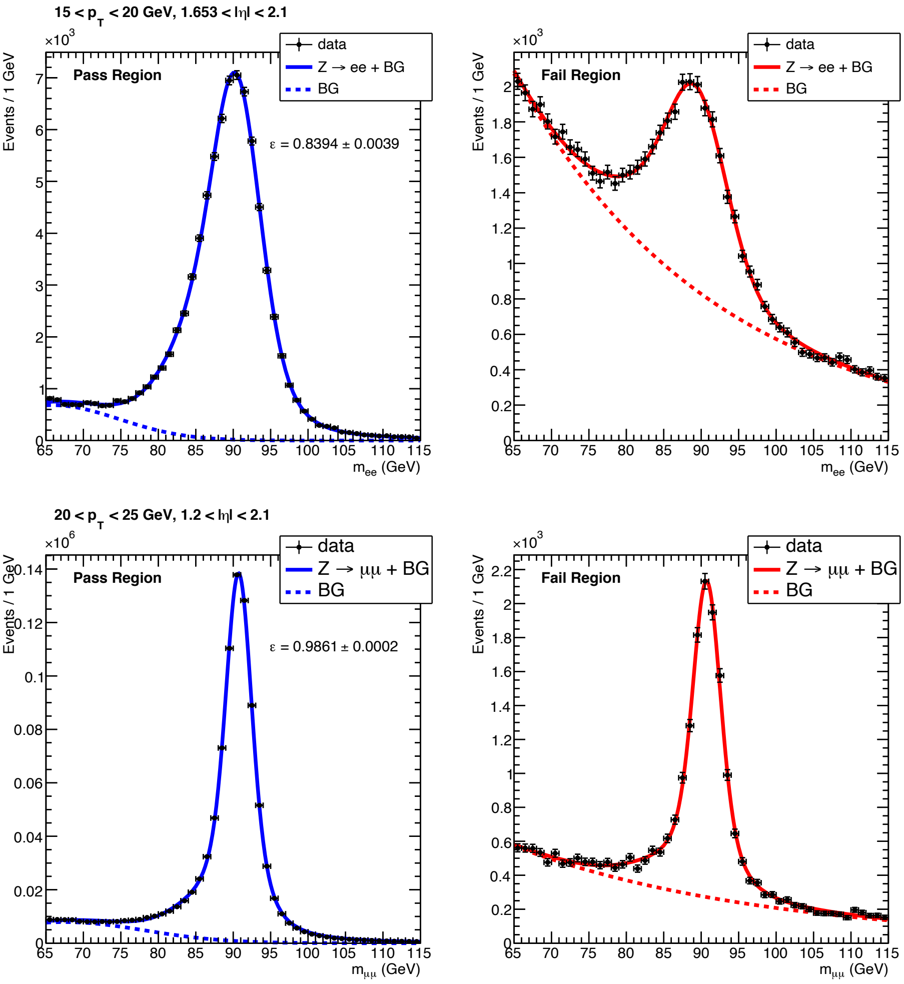

Figure 5.14: Tag-and-probe fits for identification efficiency. The top row shows fits for the electron pass and fail regions as a function of electron-pair invariant mass, whereas the bottom row is for the muon measurements.

5.5.8 τ_h-lepton energy scale

The energy scale of the τ_h-lepton is corrected for in bins of HPS decay mode. The correction is measured and applied for MC and embedded events for each year separately. They are measured in $Z/\gamma^*(\to \ell\bar{\ell}) +$jets events, where the final state leptons are a $\tau_\mu\tau_h$ pair. The correction is derived by fitting the mass distributions that are sensitive to the τ_h-lepton energy scale. The corrections are binned in HPS decay mode and are specific to each data-taking era, as listed in Table 5.7. Separate corrections are applied for MC simulated and τ-lepton

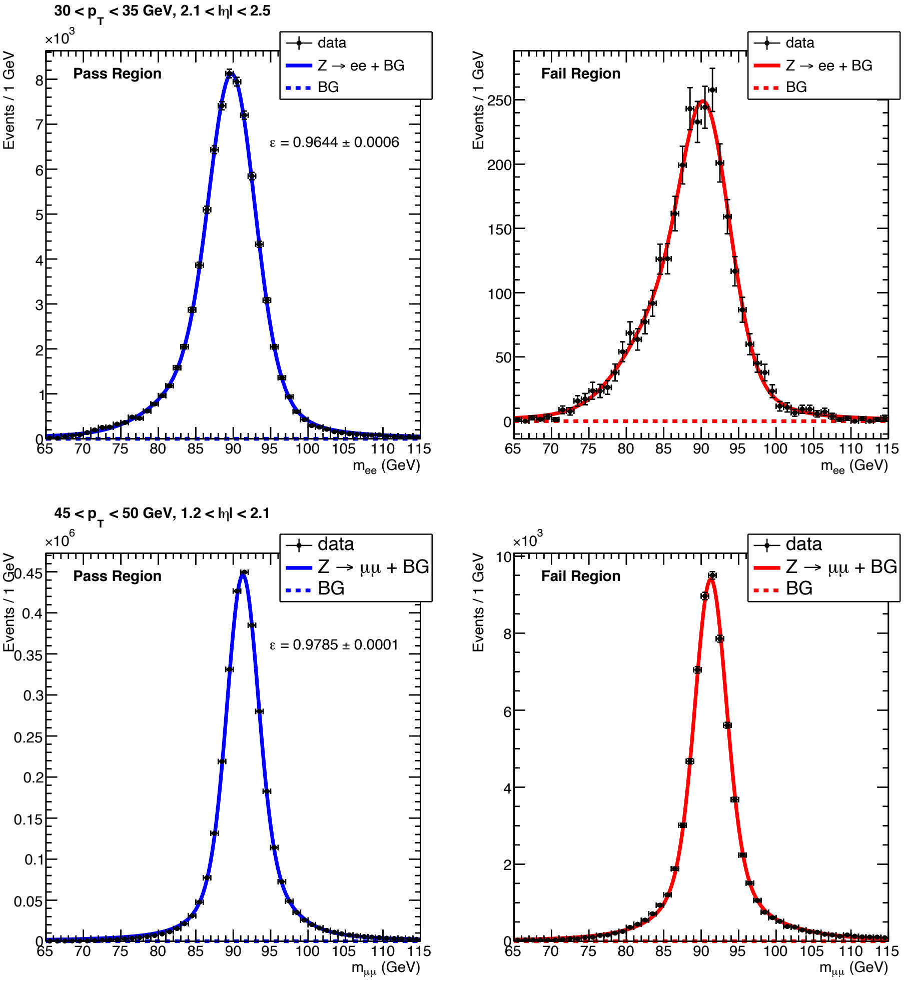

Figure 5.15: Tag-and-probe fits for isolation and trigger efficiencies. The top row shows fits for the electron pass and fail regions electron-pair invariant mass, whereas the bottom row is for the muon measurements.

embedded events from data. For 2018 events these variable distributions are illustrated in Figure 5.17.

5.5.9 $l \to \tau_h$ energy scale

In addition to the τ_h-lepton energy scale correction for genuine τ_h leptons, a shift is derived for the cases where an electron or muon fakes a τ_h lepton. As the misidentification probability is negligible for electrons and muons faking a $\tau_h \to 3h$ decay, these energy scale corrections are only measured for the h and $h\pi^0$ decay modes. The correction is derived by fitting the $\tau_\mu\tau_h$ or

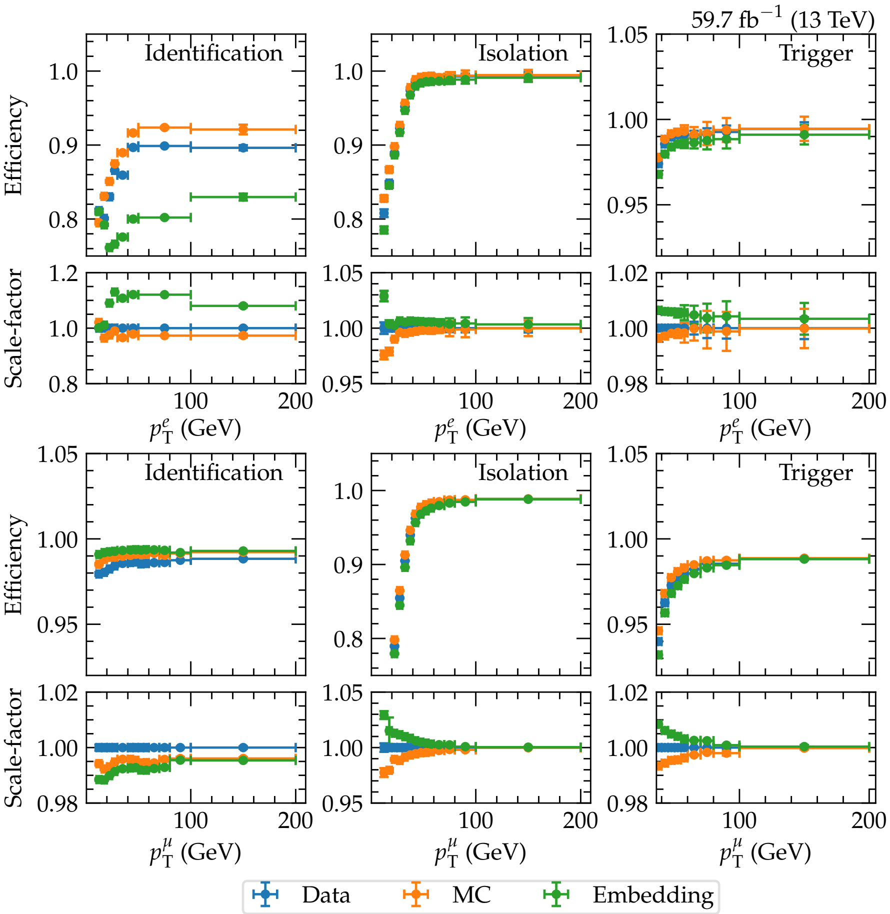

Figure 5.16: Efficiencies and scale-factors for the electron (top) and muon (bottom) identification, isolation and trigger corrections for 2018 data.

$\tau_e\tau_h$ mass distributions and deriving a shift to match the distributions with the Z mass peak. The values are given in Table 5.8.

5.5.10 $l \to \tau_h$ fake efficiency

The anti-lepton discriminators to reject background contamination on top of the τ_h-lepton ID are measured using the tag-and-probe technique with $Z/\gamma^* \to \mu^+\mu^-$ and $Z/\gamma^* \to e^+e^-$ events. The difference between the fake rates in data and simulation are accounted for using this correction.

Table 5.7: τ_h-lepton energy scale correction applied to each τ_h-lepton decay mode for the 2016, 2017 and 2018 data-taking eras. The correction is measured separately in both MC and embedded events.

Correction [%]	$\tau^\pm \to h^\pm \nu_\tau$	$\tau^\pm \to h^\pm \pi^0 \nu_\tau$	$\tau^\pm \to 2h^\pm h^\mp \nu_\tau$	$\tau^\pm \to 2h^\pm h^\mp \pi^0 \nu_\tau$
2016 MC	-0.9	-0.1	$+0.3$	-0.2
Embedded	-0.2	-0.2	-1.3	-1.3
2017 MC	$+0.4$	$+0.2$	$+0.1$	-1.3
Embedded	-0.0	-1.2	-0.8	-0.8
2018 MC	-1.6	$+0.4$	-1.2	-0.4
Embedded	-0.3	-0.6	-0.7	-0.7

Table 5.8: Lepton-to-τ_h-lepton energy scale corrections to adjust simulation for cases when electrons or muons are misidentified as τ_h leptons. For the electron, these values are stated for MC and embedded samples, delimited by a /. For muons, only the MC is corrected, as the differences between energy scales of data and embedded events are negligible.

Correction [%]	$\tau^\pm \to h^\pm \nu_\tau$	$\tau^\pm \to h^\pm \pi^0 \nu_\tau$
2016 $e \to \tau_h$	$0.7/-0.4$	$3.4/5.0$
$\mu \to \tau_h$	0.0	-0.5
2017 $e \to \tau_h$	$0.9/-2.6$	$1.2/1.5$
$\mu \to \tau_h$	-0.2	-0.8
2018 $e \to \tau_h$	$1.4/-3.1$	$1.9/-1.5$
$\mu \to \tau_h$	-0.2	-1.0

5.5.11 ggH NNLOPS reweighting

The jet multiplicity and Higgs boson p_T distributions are corrected in ggH simulation to reweight to the NNLO Higgs boson predictions [86]. These are provided for the POWHEG and MADGRAPH5_AMC@NLO samples separately, and in bins of generator-level Higgs p_T and number of jets in the event.

5.5.12 Prefiring

In the 2016 and 2017 datasets an issue is present at CMS data-taking level, which was fixed for the 2018 dataset. Due to a timing drift in the forward regions of the ECAL, L1 trigger objects became associated with previous events, leading to a decrease in efficiency during data-taking, especially for energetic objects. The probability of a jet object being prefired is found in Figure 5.18, where jets in forward pseudo-rapidity regions of about $2.0 < |\eta| < 3.2$ are the most likely to have been subjected to the prefiring issue. The weight to be applied on

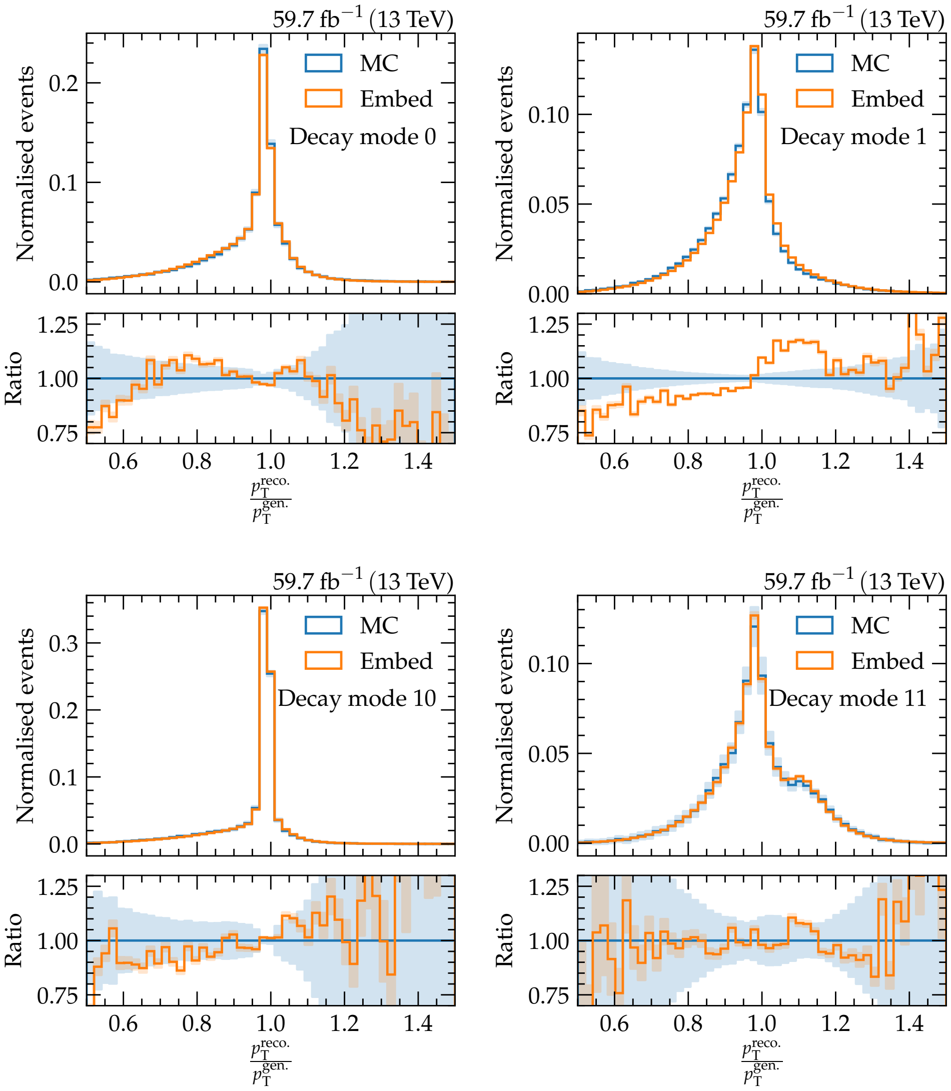

Figure 5.17: τ_h-lepton p_T resolution showing that an additional energy scale correction for embedded events is required with respect to MC events.

an event is then given as a function of prefiring probability of each jet in the event, P_i, by

$$\text{weight} = \prod_{i \in \text{jets}} (1 - P_i) \ . \tag{5.16}$$

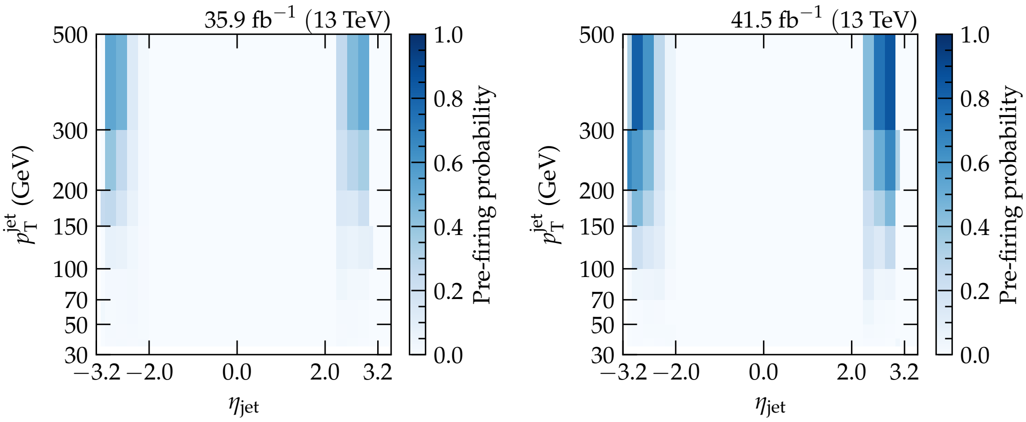

Figure 5.18: Prefiring probabilities for the 2016 and 2017 data-taking eras. These probabilities are provided in bins of jet p_T and η, and the jet p_T is placed on a square-root scale to facilitate the view of low p_T values. As the problem occurs in the forward ECAL regions, the jets in forward pseudo-rapidity have higher probabilities of having been a subject to the prefiring issue.

5.6 Background modelling

For any analysis that is pursuing a measurement of a signal embedded in a large background it essential to accurately describe the contribution of each background process. A mixture of methods based on both MC and embedded events are employed to ensure best description. Table 5.9 summarises the method used to model backgrounds depending on the type of lepton in the final state.

Table 5.9: The different experimental techniques that are deployed to estimate various background contributions

	genuine τ_μ/τ_h	jet→τ_h	lepton→τ_h
genuine τ_μ/τ_h	Embedding		
jet→τ_h	Fake Factor	Fake Factor	
lepton→τ_h	MC	Fake Factor	MC
prompt lepton	MC	Fake Factor	MC

The backgrounds that are modelled with MC are subjected to the corrections described in Section 5.5. In addition, events are weighted by

$$\frac{L_{\text{int}} \cdot \sigma \cdot \mathcal{B} \cdot w_{\text{gen}}}{N_{\text{effective}}}, \tag{5.17}$$

where L_{int} is the integrated luminosity, σ is the cross-section of the process, $\mathcal{B}$ is the branching ratio, w_{gen} is the generator weight, and $N_{\text{effective}} = \sum_{i \in \text{events}} w^{i}_{\text{gen}}$ is the effective number of events. The cross-sections used for the scaling of the $Z/\gamma^{*}(\rightarrow \ell\bar{\ell}) + \text{jets}$, $W + \text{jets}$, and $t\bar{t}$ samples are determined at NNLO precision, whilst NLO precision is used for the single-t process. For the minor backgrounds of electroweak and diboson processes, the generator-level cross-sections are used for the scaling. As introduced in Section 5.3.1, embedded samples are utilised for the modelling of the $Z/\gamma^{*} \rightarrow \tau^{+}\tau^{-}$, which is a major background in analyses studying SM Higgs boson production. As for the case of the MC, the generator weights are applied on top of the corrections in Section 5.5.

For events where a jet fakes a τ_h-lepton, MC simulation does not provide an accurate level of description. Therefore methods that use data directly are employed for these backgrounds and the *fake-factor* (FF) method provides the best description for all channels. The FF method employed for the analyses presented in this Thesis is very similar to the ones presented in earlier $H \rightarrow \tau^{+}\tau^{-}$ publications, for instance as found in Reference [87].

The FF method tries to estimate the contribution of jet $\rightarrow \tau_h$-lepton events in the signal region by extrapolating from a side-band region in data. These regions are defined using selection on the identification criteria of the τ_h-lepton: in the signal region the τ_h-lepton must pass the *medium* working point, whereas in the side-band region the requirement is to pass the loosest working point, but fail the *medium* working point. Therefore these regions are completely orthogonal, as required. The raw FF is then defined as

$$\text{FF}_i = \frac{N(\tau_h^{\text{iso}} \geq \text{medium})}{N(\tau_h^{\text{iso}} < \text{medium})} \, , \tag{5.18}$$

where the sub-script i represents the process in which the FF is being measured. The FF are measured separately in process-enriched regions for each channel, which for the $\tau_h\tau_h$ channel only comprises QCD multi-jet background, as this is the major source of jet $\rightarrow \tau_h$-lepton fake background. The process-enriched region for QCD is determined by requiring a final state lepton pair of same sign, as $H \rightarrow \tau^{+}\tau^{-}$ requires opposite sign leptons in the final state. On top of the QCD region, for the $\tau_\mu\tau_h$ and $\tau_e\tau_h$ channels the FF are also measured in $W + \text{jets}$ and $t\bar{t}$ enriched regions, which are produced using a selection on the transverse mass of the lepton, $m_{\text{T}}^{e/\mu}$, and using MC simulation, respectively. Following the determination of the raw FF, these are corrected for to account for differences between the signal region and determination region (of inverted isolation requirement). These differences may be kinematic or process-dependent differences due to the extrapolation. The individual FF are then combined into a single average using

$$\langle \text{FF} \rangle = \sum_{i \in \text{processes}} f_i \cdot \text{FF}_i \, , \tag{5.19}$$

where $f_i = \frac{n_i}{\sum_{j\in\text{events}} n_j}$ is the probability of a jet $\rightarrow \tau_h$-lepton event to originate from a certain process i from the full set of events j. These probabilities are determined in regions that only differ by the isolation requirement as well. The final combined and corrected FF is then applied in the same region to estimate the yield of events in the signal region. For the $\tau_\mu\tau_h$ and $\tau_e\tau_h$ channels, events with genuine τ_h-lepton or a $l \rightarrow \tau_h$-lepton candidates are subtracted from the final estimate using MC simulation. In the case of the $\tau_h\tau_h$ channel, two jet $\rightarrow \tau_h$-lepton candidates may be present, which represents the majority of QCD multi-jet events. However, as the FF are measured for one of the τ_h-leptons (the leading one), events with a single jet $\rightarrow \tau_h$-lepton fake are underestimated, and are thus added back using simulated events.

5.7 Summary

The physics objects that are reconstructed from pp collisions at the CMS detector are passed through a set of requirements that ensure the reconstruction efficiency remains high, whilst keeping the misidentification rate low. Due to the differences between simulation and data caused by requiring specific simulation conditions, selection criteria, and detector responses, corrections need to be measured and applied to the physics objects to ensure that the simulation correctly describes data. Even when modelling events from background processes using data, such as is done for the majority of events using the embedded samples and the fake-factor method, the signal modelling used later in the analysis to extract the physics result always comes from MC simulation. Thus, having leptons and jets that are well modelled is of great importance. All the ingredients are now laid out to proceed with the analyses, starting with the measurement of the $\mathcal{CP}$ structure of the Yukawa coupling between the Higgs boson and the τ-lepton using $gg \rightarrow \mathrm{H} + 2$ jets.

Chapter 6

Measurement of the Higgs Boson $\mathcal{CP}$ State using the ggH Production Vertex

6.1 Introduction

The measurements of the $\mathcal{CP}$ properties of the Higgs boson can be performed in multiple ways. This chapter will outline the methods employed to determine the $\mathcal{CP}$ mixing angle using the GGF production vertex in association with two jets. As described in Chapter 2, this involves the study of the Yukawa coupling of the Higgs boson to the top quark through the analysis of angular correlations between jets. The analysis uses the decays of the Higgs boson to a pair of τ leptons. The analysis presented in this Thesis consists of the two most sensitive final states: $\tau_h\tau_h$ and $\tau_\mu\tau_h$[1]. The 2016, 2017 and 2018 datasets are used in this measurement, which correspond to 35.9 fb^{-1}, 41.9 fb^{-1}, and 59.7 fb^{-1}, respectively, or a total of 137 fb^{-1}.

6.2 Signal modelling

The signal samples are modelled at NLO precision using MadGraph5_aMC@NLO according to the implementation described in Reference [88]. In addition to the set of inclusive samples with zero, one, and two outgoing partons in the matrix-element calculation, samples with exactly two outgoing partons are generated to enhance statistics in the dijet phase space. The modelling of signal for any $\mathcal{CP}$ state is then performed using the methods outlined in Section 2.9.1. As for MC simulation used for background modelling, the signal is scaled by cross-section times branching ratio using values from Reference [19].

The discriminating variable that is sensitive to Higgs boson $\mathcal{CP}$ state is the signed difference of the azimuthal angle between the two jets with leading p_{T}. In order to select the right sign, the

[1] The subscript describes the particle into which the associated τ-lepton decays: h for hadronically decaying τ-lepton, otherwise μ for $\tau \to \mu$ decays.

difference is calculated taking into account the ordering of the jets in pseudo-rapidity:

$$\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2}) = \phi(\vec{p}_{j_1}) - \phi(\vec{p}_{j_2}) \text{ , where } \eta_{j_1} < \eta_{j_2} \, . \tag{6.1}$$

The generator-level distributions of $\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2})$ for different selection criteria on m_{jj} are illustrated in Figure 6.1. As the cut on m_{jj} is increased, the distinction between the $\mathcal{CP}$ states is enhanced. The high m_{jj} phase space is therefore most sensitive for this measurement, which will become crucial for the event categorisation.

6.3 Event selection

Building on the event selection presented in Chapter 5 more specific selection is applied on each lepton in the final state pair to tailor the needs of the analysis. The particular selection applied to τ_μ or τ_h leptons in the $\tau_\mu\tau_h$ and $\tau_h\tau_h$ final states are summarised in Table 6.1.

Table 6.1: The trigger (online) and analysis-level (offline) p_T thresholds for the single muon and cross trigger for the $\tau_\mu\tau_h$ channel, and di-tau trigger for the $\tau_h\tau_h$ channel for each data-taking era considered in this analysis. The thresholds applied on the η and p_T of the trigger matched offline objects are also highlighted. The τ_h-lepton ID selection is performed with respect to the three output nodes of the *deepTauID* classifier: *tight* discrimination versus jets, *very-very-very-loose* discrimination versus electrons, and *tight* discrimination versus muons. In the $\tau_h\tau_h$ channel, the leading τ_h-lepton is required to have a tighter p_T to further reduce the jet → τ_h-lepton fake background.

	Year	Trigger requirement	Offline lepton selection		
		$p_\mathrm{T}^\mathrm{raw}$ (GeV)	p_T (GeV)	$\lvert\eta\rvert$	Isolation
$\tau_h\tau_h$	All	$\tau_h(35)$ & $\tau_h(35)$	$p_\mathrm{T}^{\tau_h^1} > 50, p_\mathrm{T}^{\tau_h^2} > 40$	$\lvert\eta^{\tau_h}\rvert < 2.1$	τ_h ID
$\tau_\mu\tau_h$	2016	$\tau_\mu(22)$, $\tau_\mu(19)$ & $\tau_h(20)$	$p_\mathrm{T}^{\tau_\mu} > 20$	$\lvert\eta^{\tau_\mu}\rvert < 2.1$	$I^{\tau_\mu} < 0.15$
			$p_\mathrm{T}^{\tau_h} > 25$	$\lvert\eta^{\tau_h}\rvert < 2.3$	τ_h ID
	2017, 2018	$\tau_\mu(24)$, $\tau_\mu(20)$ & $\tau_h(27)$	$p_\mathrm{T}^{\tau_\mu} > 21$	$\lvert\eta^{\tau_\mu}\rvert < 2.1$	$I^{\tau_\mu} < 0.15$
			$p_\mathrm{T}^{\tau_h} > 32$	$\lvert\eta^{\tau_h}\rvert < 2.3$	τ_h ID

On top of the selection presented in Table 6.1, an additional set of selection is applied in the $\tau_\mu\tau_h$ channel to further reduce the background contamination. The contribution of W + jets events to the jet → τ_h-lepton fake background can be decreased using a cut on the transverse mass of the τ_μ and $\vec{p}_\mathrm{T}^\mathrm{miss}$ system, $m_T \equiv \sqrt{2p_\mathrm{T}^{\tau_\mu} p_\mathrm{T}^\mathrm{miss}(1 - \cos\Delta\phi(\tau_\mu, \vec{p}_\mathrm{T}^\mathrm{miss})} < 50\,\mathrm{GeV}$. In addition, a veto on b-jets is employed to reduce the $t\bar{t}$ + jets background. Finally, the lepton vetoes and pair selection discussed in Chapter 5 are used to keep the different channels orthogonal and reduce the $\mathrm{Z}/\gamma^*(\to \ell\bar{\ell})$ + jets and diboson backgrounds.

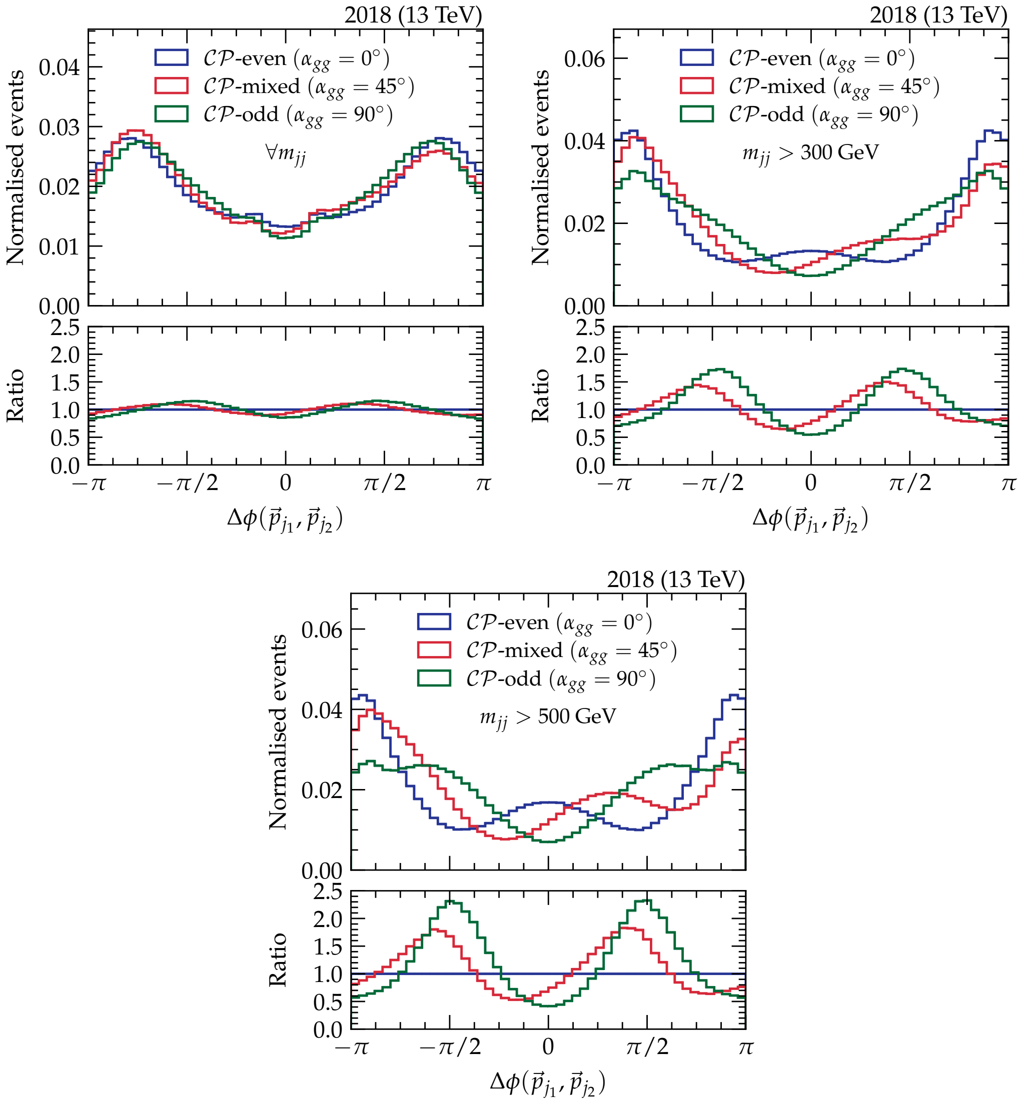

Figure 6.1: Generator-level distributions of $\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2})$ for different selection criteria on m_{jj} using the set of signal samples produced for this analysis. As the requirement on m_{jj} is increased, the separation between the illustrated distributions is enhanced. The events with $m_{jj} > 500\,\text{GeV}$ will have more $\mathcal{CP}$ discriminating power compared to events with lower m_{jj}.

6.4 Background methods

The methods employed to model the backgrounds has already been introduced in 5. In this analysis the background estimation is performed using data where possible. This results in the use of the embedded samples to model genuine $\tau\tau$ final state events largely due to a decay of Z/γ^* boson. Background events involving jet $\rightarrow$ τ_h-lepton fake candidates are estimated using the fake factor method. Smaller backgrounds, such as electroweak Z, Z/$\gamma^* \rightarrow \mu^+\mu^-$, diboson, and $t\bar{t}$ + jets are estimated using simulated events from MC.

As this analysis specifically looks for ggH production, events involving any other form of Higgs boson production mode are considered as background in this case. The ttH production mode is neglected due to its relatively small cross-section. However, VBF, WH and ZH Higgs boson production are considered, and are estimated from MC simulated events generated with POWHEG at NLO precision. The SM prediction is used for these backgrounds. This choice is motivated by the theory, as most models do not include any tree-level couplings of $\mathcal{CP}$-odd Yukawa coupling components to the heavy gauge bosons. Thus, the effect of a $\mathcal{CP}$-odd component is heavily suppressed and the kinematic distributions will be similar to those of the SM distributions. A shift in the cross-section and branching ratio is allowed and it is possible that a non-SM $\mathcal{CP}$ state will alter these parameters.

6.5 Categorisation

The selection is necessary to obtain a sample of events with objects that are well identified with high efficiency and low misidentification rate. The data at hand is now a mixture of $\mathcal{CP}$-sensitive and non-$\mathcal{CP}$-sensitive (background) events. These different types of events can be categorised into different regions of phase space using kinematic selections, such that the purity of $\mathcal{CP}$-sensitive events is increased at a cost of decreasing its efficiency. Categorisation techniques can be manually tuned to tailor towards the needs of the analysis using physics motivated selections or can be driven by machine learning methods, such as a BDT or NN, which can outperform manual selection as the choice of selection is optimised on a multivariate phase space, taking correlations between input features into account. The methods of categorisation will be discussed and compared. The simple approach, which will be discussed and motivated in Section 6.5.1, was adopted for this analysis and is used as the baseline. The machine-learning-based categorisation studies in Section 6.5.2 provide an alternative method, which can enhance the purity and efficiency of the signal categories, and thus improve the sensitivity of future measurements.

6.5.1 Cut-based categorisation

The cut-based approach for the categorisation is defined through simple selection on several kinematic variables and is based on the di-τ p_{T}, the number of jets in the event, and the mass of the two jets with leading p_{T}, which are powerful features that target the VBF-like phase space that is important for this analysis. An event of larger di-τ p_{T} generally consists of a boosted final state pair of leptons, which implies large momentum transfer at parton-level. The number of jets selection is required to select the $\mathcal{CP}$-sensitive dijet phase space of the $gg \to \mathrm{H}$ production mode. Although not directly contributing to the $\mathcal{CP}$-sensitive region, events with less than two jets are used to define additional categories which help the fit to further constrain backgrounds and systematic uncertainties. Similarly, events with low reconstructed Higgs boson p_T or low dijet mass are not discarded. The categorisation is illustrated in the diagram shown in Figure 6.2.

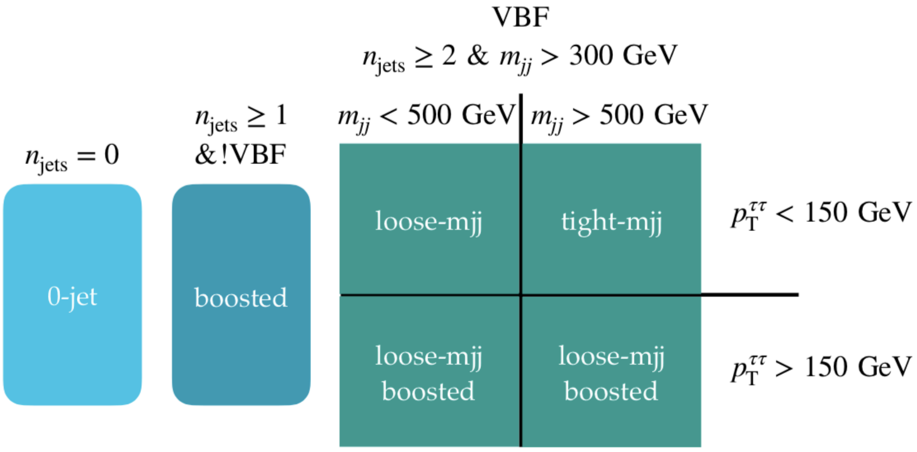
VBF
$n_\mathrm{jets} \geq 2$ & $m_{jj} > 300$ GeV
$n_\mathrm{jets} = 0$
$n_\mathrm{jets} \geq 1$ & !VBF
$m_{jj} < 500$ GeV
$m_{jj} > 500$ GeV
0-jet
boosted
loose-mjj
tight-mjj
loose-mjj boosted
loose-mjj boosted
$p_\mathrm{T}^{\tau\tau} < 150$ GeV
$p_\mathrm{T}^{\tau\tau} > 150$ GeV

Figure 6.2: Cut-based categorisation scheme using selections on the number of jets, the invariant mass of the two leading jets, and the transverse momentum of the Higgs boson. The right-most category (labelled dijet or *VBF* category) targets VBF-like events and is therefore contains the most sensitive events used for this measurement.

The choice of fit variable in the different categories is the following:

0-jet: SVFIT mass

boosted: SVFIT mass in increasing windows of Higgs boson p_T

VBF: $\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2})$ in increasing windows of SVFIT mass

6.5.2 BDT categorisation studies

For the multivariate categorisation, on the other hand, a set of features can be used simultaneously to provide optimal separation between signal and background events. The multi-variate variable can replace, for instance, the SVFIT mass used in the cut-based categorisation as this will be a high-level variable with more separating power than the most optimal variable in traditional cut-based selections. The studies have been performed using 2018 data and the final expected sensitivities of the cut-based and BDT-based approaches will be compared in Section 6.10.1.

The choice of machine learning algorithm for this study is the XGBoost package [89], which is a very robust and scalable software that is based on the stochastic gradient descent method. The problem of this study is formulated as a supervised multi-class categorisation problem, where the goal is to accurately and precisely categorise a given set of events into their true process categories. The truth information is given by its process label, which is determined from MC for signal events, or from data for embedded or jet → τ_h-lepton fake events.

Training of BDT

For the training of the set of BDTs, training samples are created using the available MC simulation, embedded samples, and data for jet → τ_h-lepton fake events. These events are required to pass a pre-selection that is similar to that defined for the analysis, as formulated in Table 6.1. Then, instead of performing the selection on the variables to categorise events into $\mathcal{CP}$-sensitive and non-$\mathcal{CP}$-sensitive events as indicated in Figure 6.2, the training samples are passed through the BDT algorithm, which finds the optimal set of selection criteria on the input features provided to the training. Each channel is trained separately due to different processes contributing to the background composition.

As the topology of the events of interest lies within the phase space of the ggH process with two associated jets, the set of events used for training are separated into two categories: *high-mjj*, with $(n_{\text{jets}} \geq 2$ and $m_{jj} > 300)$, and *low-mjj*, where $(n_{\text{jets}} \geq 2$ and $m_{jj} < 300)$ or $n_{\text{jets}} \leq 1$. While there were studies performed that included training on the low-mjj categories, these are not $\mathcal{CP}$-sensitive events, and are therefore omitted from this discussion. Instead, the 0-jet and boosted categorisation from Figure 6.2 was adopted for these events.

To make use of the full dataset in the training and provide unseen data to the trained BDT to extract expected sensitivities for the final result, the training samples are split into two sets with even and odd event number. Using the event number for this split is an arbitrary choice, but it provides a well-defined, unique event tag to split the training samples into two halves, each consisting of about 50% of the total. Therefore, for each channel, the high-mjj training samples are formed, and a training iteration is performed on all events with even event number, whilst another training is performed on all events with odd event label. During the training, however, 25% of each training sample is kept aside for cross-validation purposes. The training is monitored as a function of boosting iterations, and it terminates once the loss function does not decrease significantly anymore.

The loss function of choice, and therefore formulation of the optimisation problem, is the multinomial deviance or cross-entropy, which is a robust loss function for typical classification tasks. It is expressed mathematically as

$$H(p,q) = -\sum_x p(x) \cdot \log q(x) , \tag{6.2}$$

where $p(x)$ and $q(x)$ represent the probability of class x in the target and prediction vectors, respectively. The target is the class label supplied at training, and the prediction probability is provided after a training iteration. In order to simplify this equation, the target labels are one-hot encoded, such that a vector of $[100]$ represents class 1, $[010]$ represents class 2, and $[001]$ indicates class 3. In this manner, $p(x)$ becomes either 1 or 0.

The input features used for this study consist of kinematic variables of the dilepton pair and jet variables (except $\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2})$ as this would bias the classification to a specific $\mathcal{CP}$ state, which is not the purpose of this procedure). The BDT that was trained on events with even event label will be applied on the events with odd event number, and vice versa. The output of the multi-class BDT will be a set of scores between zero and one for each training class, which sum to one and can thus be interpreted as mutually exclusive probabilities that a given event is associated to certain process label. The event will be assigned to the class with largest score. This is illustrated in Figure 6.3.

Table 6.2: Input features to the BDT for the signal versus background classification. The variables used in the $\tau_h\tau_h$ and $\tau_\mu\tau_h$ channels are shown, where differences between the two channels are driven by the goodness-of-fit between data and simulated events, and the gain of including the feature in the channel of interest.

Feature	in $\tau_\mu\tau_h$ channel	in $\tau_h\tau_h$ channel
p_T of leading τ_h or τ_μ	✓	✓
p_T of sub-leading τ_h	✓	×
p_T of visible di-τ	✓	✓
p_T of di-τ_h + $\vec{p}_\mathrm{T}^\mathrm{miss}$	✓	✓
Transverse mass of τ_μ and $\vec{p}_\mathrm{T}^\mathrm{miss}$	✓	×
Visible di-τ mass	✓	✓
SVFit di-τ mass (using SVFit)	✓	✓
Leading jet p_T	✓	✓
Dijet p_T	✓	×
Jet multiplicity	✓	✓
Invariant mass of leading jet pair	✓	✓
$\lvert\eta\rvert$ of leading jet pair	✓	✓
$p_\mathrm{T}^\mathrm{miss}$	✓	✓

The study was performed with two formulations of the training classes:

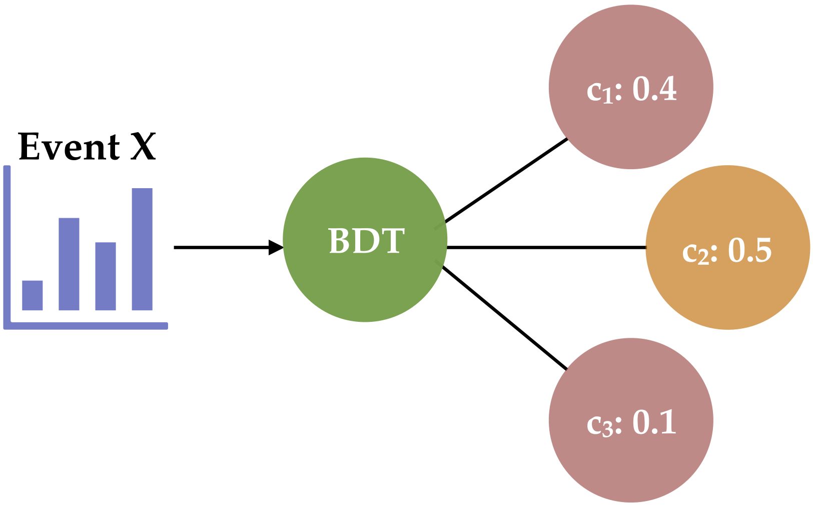

Figure 6.3: Generic BDT-based categorisation scheme to illustrate the principles behind this type of multi-classification. Event X is fed to a trained BDT model, which determines three output scores. The event is then classified into the process category with the highest BDT score (using the softmax function), which in this example is c_2. This method, therefore, effectively creates process-enriched categories.

Four classes - referred to as *split* scheme:

- signal events - consisting of simulated ggH events generated with POWHEG.
- VBF events - VBF simulated events which are treated as background in this analysis
- events with two genuine τ-leptons - from embedded samples.
- events with at least one jet $\rightarrow$ τ_h-lepton fake - from events selected with the fake-factor method.

Three classes - referred to as *merge* scheme:

- signal events - consisting of mixture of simulated ggH and VBF events.
- events with two genuine τ-leptons - from embedded samples.
- events with at least one jet $\rightarrow$ τ_h-lepton fake - from events selected with the fake-factor method.

For the first comparisons, the signal consisted of the ggH signal, and the VBF signal was added as a separate class. Thus the BDT was set-up to train ggH events against all backgrounds, including the VBF Higgs boson production. Each event is allocated its class based on the highest BDT score and is required to have at least 2 outgoing jets with invariant mass of 300 GeV. The output distributions for each high-mjj class are shown in Figures 6.4 and 6.5 for the $\tau_h\tau_h$ and $\tau_\mu\tau_h$ channels, respectively, where the data in the signal categories is blinded[2], as the final results for this analysis are not produced using this categorisation method. The background categories are unblinded as there aren't many expected signal events in each bin, however, the signal categories are left blinded. The VBF is well differentiated from the other processes, as seen in the top right plot of each Figure, whereas the ggH category is contaminated with background events.

Merging the VBF and ggH events into one training class encourages the BDT to train for the VBF-like phase space. Again, to evaluate the performance of this set-up, the events are classified into categories depending on the highest score output of the BDT. Since the analysis is only interested in using ggH events, the VBF category is still defined, but with an additional selection based on the invariant dijet mass: ggH requires $m_{jj} < 500\,\text{GeV}$, whilst VBF includes events with $m_{jj} > 500\,\text{GeV}$. Now the events that have larger sensitivity to the $\mathcal{CP}$-mixing angle will be found on the higher end of the BDT score, which are bins with higher signal purity. This is illustrated in the Figures 6.6 and 6.7 for the $\tau_h\tau_h$ and $\tau_\mu\tau_h$ channels, respectively. This way, effectively, the VBF category becomes the most sensitive one to the $\mathcal{CP}$ measurement, even though it contains a larger mixture of VBF and ggH events.

In Figure 6.1 the importance of high m_{jj} events for this analysis was demonstrated, as higher m_{jj} events contributed to larger separation between $\mathcal{CP}$ states, thus enabling a more sensitive measurement. Investigating the m_{jj} distribution as a function of the BDT score can help determining which method will provide more promising results. To this end, the BDT score distributions of the ggH and VBF categories with the additional selection of $m_{jj} > 700\,\text{GeV}$ is compared to the BDT score distribution with the nominal dijet selection of $m_{jj} > 300\,\text{GeV}$. This was done separately for each training set-up. Figure 6.8 illustrates these distributions in the $\tau_h\tau_h$ channel, including the ratio of events with $m_{jj} > 700\,\text{GeV}$ to $m_{jj} > 300\,\text{GeV}$. These distributions show that, although there are less ggH events in total in the ggH- and VBF-enriched categories, ggH events have a higher m_{jj} across all bins of the BDT score for the case when the VBF and ggH events are merged in the training.

Although the result will be determined using the simpler cut-based categorisation scheme, the expected sensitivity of the cut-based approach will be compared with the BDT-based categorisation. The trained BDT models are applied on all MC simulated events, embedded samples and observed data events to determine the performance of this categorisation technique and provide a direct comparison to the cut-based approach. The comparison will be made only between 2018 expected sensitivities to the $\mathcal{CP}$-mixing angle and rate parameter controlling the ggH production mode. Prior to this, the set of corrections applied to simulated events and associated uncertainties are discussed, as these are used as inputs to the maximum likelihood estimator used in the final fits.

[2] Blinding refers to the act of removing the observed data points from the visualisation in order to not bias the selection and therefore the final results.

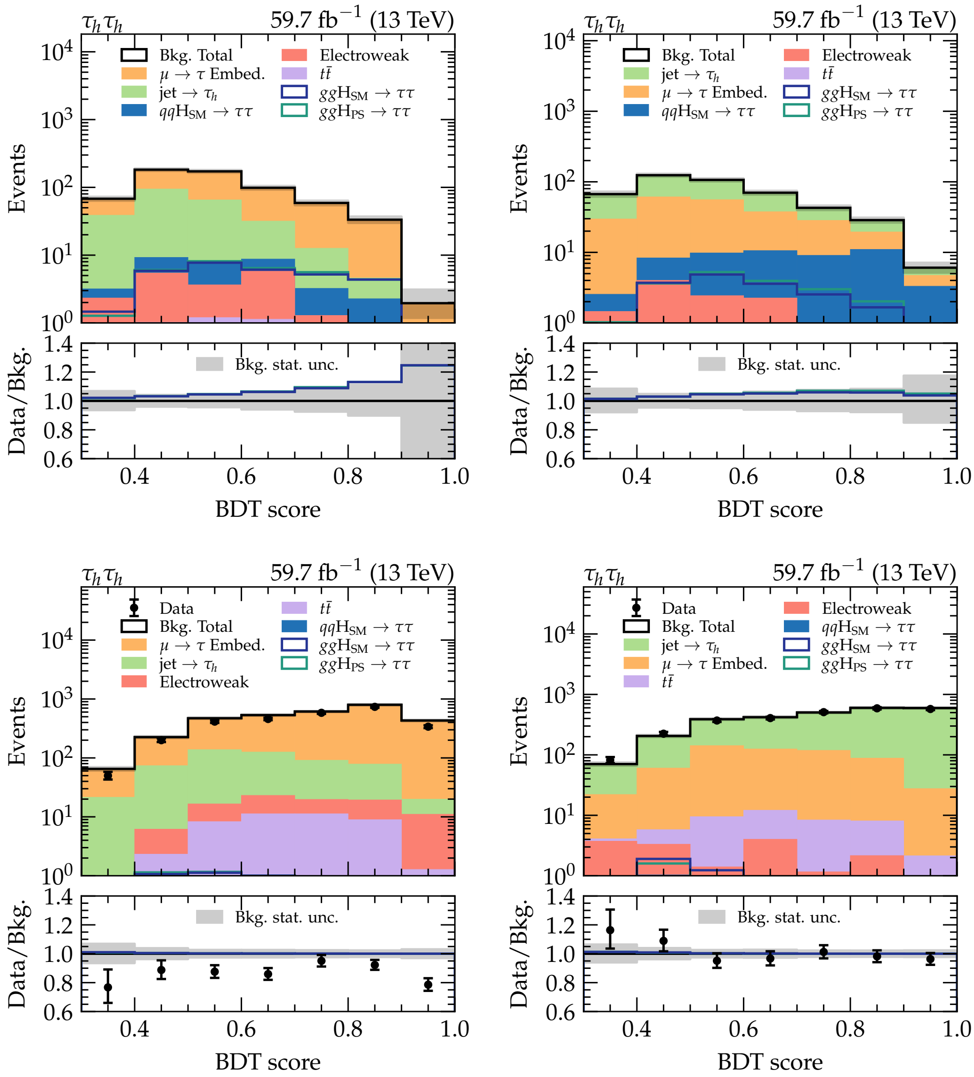

Figure 6.4: Distributions of the BDT score for the process-enriched categories as defined by the multi-class BDT when ggH and VBF are treated as separate training classes. These distributions are pre-fit and for the $\tau_h\tau_h$ channel. The top row represents the ggH and VBF categories, whereas the bottom row illustrates the genuine τ-lepton and jet-fake τ_h-lepton categories, respectively. The grey band considers only the statistical uncertainty in the background distributions.

6.6 Corrections

In order to ensure that the modelling of observed events is accurate, corrections are applied to simulated events. These object-dependent corrections are applied to the relevant type of

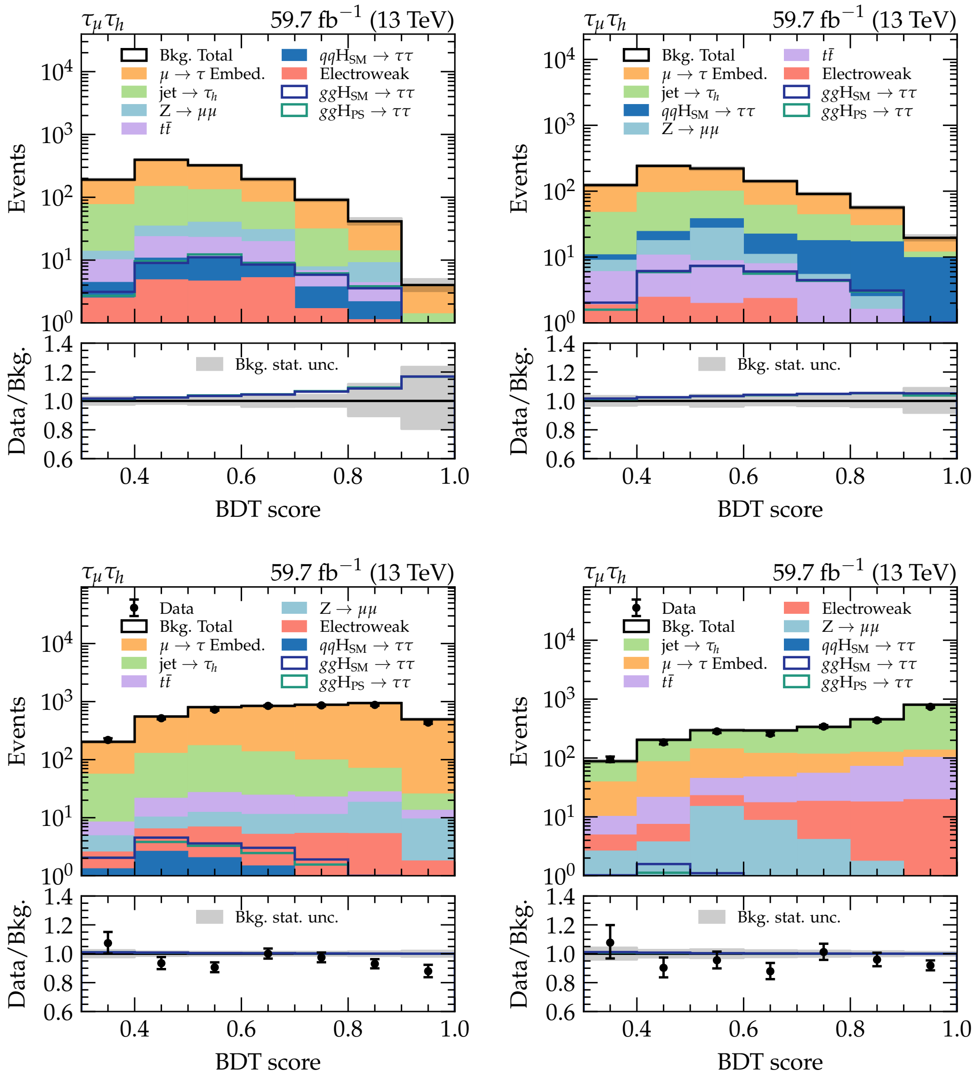

Figure 6.5: Distributions of the BDT score for the process-enriched categories as defined by the multi-class BDT when ggH and VBF are treated as separate training classes. These distributions are pre-fit and for the $\tau_\mu\tau_h$ channel. The top row represents the ggH and VBF categories, whereas the bottom row illustrates the genuine τ-lepton and jet-fake τ_h-lepton categories, respectively. The grey band considers only the statistical uncertainty in the background distributions.

simulated (MC or embedded events), and were introduced in Chapter 5. The list of corrections used in adjusting simulated events in this analysis is summarised in Table 6.3.

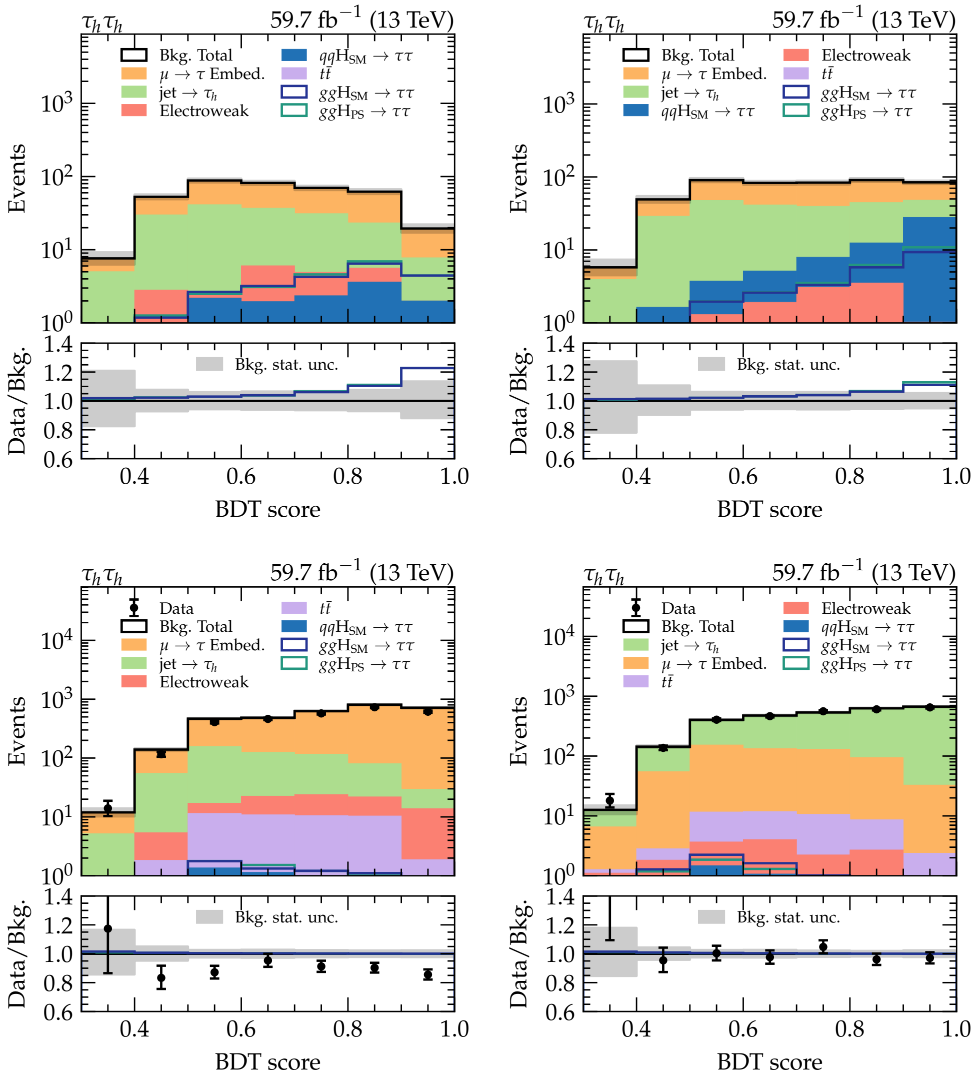

Figure 6.6: Distributions of the BDT score for the process-enriched categories as defined by the multi-class BDT when ggH and VBF are treated as a single training classes. These distributions are pre-fit and for the $\tau_h\tau_h$ channel. The top row represents the ggH and VBF categories, whereas the bottom row illustrates the genuine τ-lepton and jet-fake τ_h-lepton categories, respectively. In this case the ggH enriched category is defined through an additional selection of $m_{jj} < 500\,\mathrm{GeV}$, whilst the VBF category is the subset of events with $m_{jj} > 500\,\mathrm{GeV}$. The grey band considers only the statistical uncertainty in the background distributions.

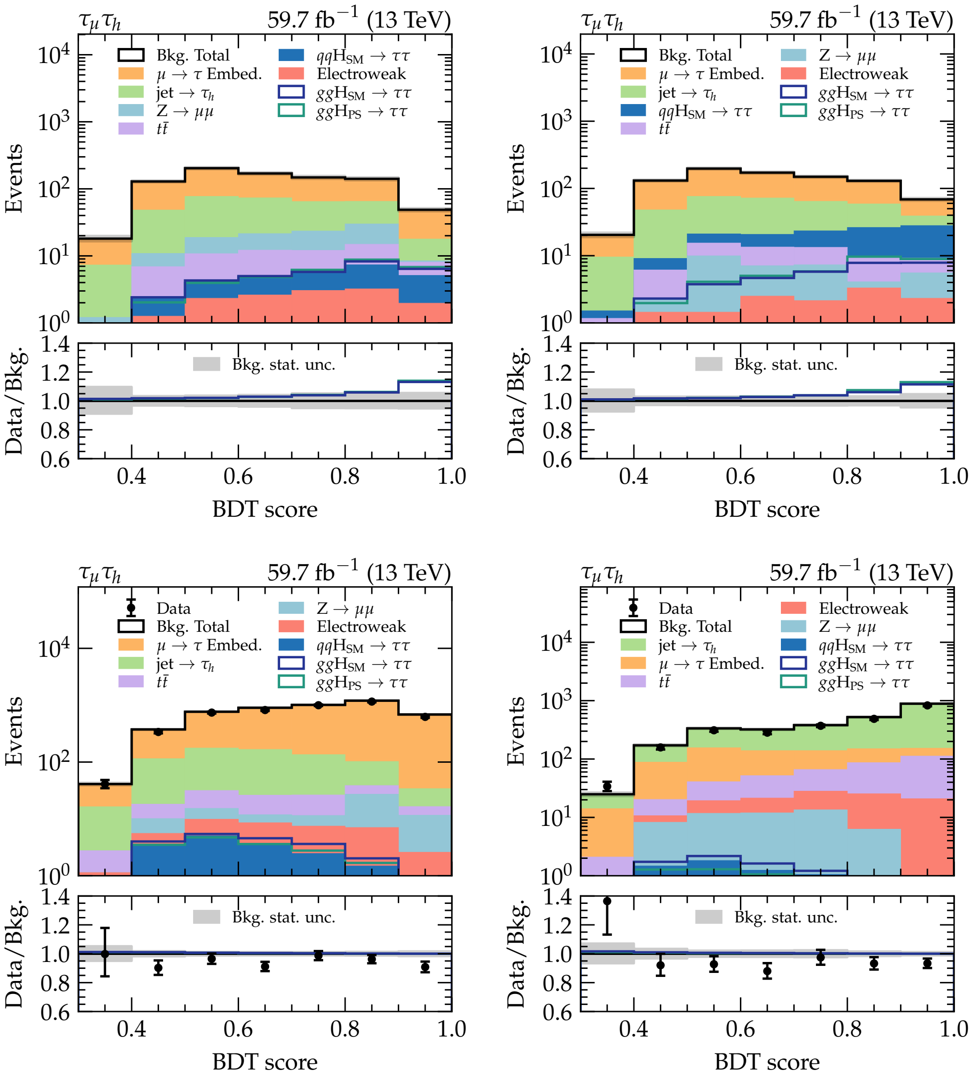

Figure 6.7: Distributions of the BDT score for the process-enriched categories as defined by the multi-class BDT when ggH and VBF are treated as a single training classes. These distributions are pre-fit and for the $\tau_\mu\tau_h$ channel. The top row represents the ggH and VBF categories, whereas the bottom row illustrates the genuine τ-lepton and jet-fake τ_h-lepton categories, respectively. In this case the ggH enriched category is defined through an additional selection of $m_{jj} < 500\,\mathrm{GeV}$, whilst the VBF category is the subset of events with $m_{jj} > 500\,\mathrm{GeV}$. The grey band considers only the statistical uncertainty in the background distributions.

6.7 Bin averaging of signal and background distributions

The distributions used in the categorisation can suffer from poorly populated bins. In order to remedy this issue, symmetries in the signal and background templates can be exploited to

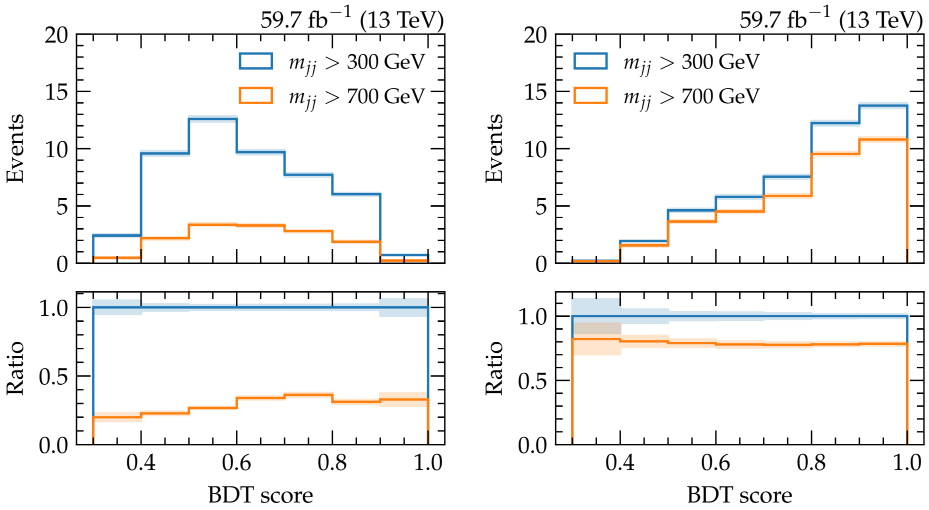

Figure 6.8: BDT score distributions of the expected ggH events in the ggH- and VBF-enriched categories (as selected by the BDT) for two definitions of the training classes: four classes (ggH, VBF, genuine, and fakes) in the left plot and three classes (ggH + VBF, genuine, and fakes) in the right plot. The selection used in the categorisation is varied between $m_{jj} > 300\,\text{GeV}$ and $m_{jj} > 700\,\text{GeV}$ to determine the ratio of high m_{jj} events with respect to lower m_{jj} events as a function of BDT score. The training where ggH is combined with VBF events produces a higher ratio of events, and high m_{jj} events are allocated higher BDT scores, which tends to imply higher signal-to-background ratio.

Table 6.3: Corrections applied to simulated events to ensure correct modelling of the observed events.

Correction	Simulation type
Lepton ID and trigger	MC + embedded
$l \to \tau_h$ fake rate	MC
Lepton energy scale	MC
$\vec{p}_\text{T}^\text{miss}$ recoil	MC
b-tagged jet efficiency	MC
Z p_T/mass reweighting	MC (Z, W, H).
t-quark p_T reweighting	MC ($t\bar{t}$)
NNLOPS p_T/n_jets reweighting	MC (ggH)

reduce the impact of statistical fluctuations in the distributions. Averaging the distributions also can help to minimise any effects that bias the measurement. The signal and background distributions are symmetric about $\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2}) = 0$ and are therefore symmetrised about that value. This is done by setting the content w of two symmetric bins, i and j, to the their mean as $w'_i = w'_j = \frac{1}{2}(w_i + w_j)$. This procedure can be applied to $\mathcal{CP}$-even and $\mathcal{CP}$-odd signal distributions and all backgrounds. For the $\mathcal{CP}$-mixed signal distribution, the distribution is anti-symmetrised, as the interference term is anti-symmetric about $\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2}) = 0$, such that symmetric bin contents of the interference term have the same magnitudes, but are opposite in sign: $|w_i|' = |w_j|' = \frac{1}{2}(|w_i| + |w_j|)$.

6.8 Systematic uncertainties

The set of systematic uncertainties relevant to this analysis may be sub-divided into two types of uncertainties: normalisation and shape uncertainties. Shape uncertainties influence the (expected) yield and shape of the signal and background distributions; normalisation uncertainties only affect the yield of the distributions, and are thus a special type of shape systematic uncertainty with a flat shape across all bins.

6.8.1 Normalisation uncertainties

Normalisation uncertainties cover systematic uncertainties related to the yields of the signal and background distributions. These types of systematic uncertainties are uncertainties on observables with only positive values, such as the luminosity, cross-section and selection efficiencies. The suite of normalisation uncertainties relevant to this measurement is now discussed, including the differences in each data-taking era (2016, 2017 and 2018) and the correlations between each year.

Luminosity: the total uncertainty on the luminosity is 2.5%, 2.3% and 2.5% for 2016, 2017 and 2018 data-taking eras, respectively. This source of uncertainty is considered partially correlated: the uncorrelated part amounts to 2.2%, 2.0% and 1.5%, whereas the correlated part is equivalent to 1.2%, 1.1% and 2.0% for 2016, 2017 and 2018, respectively.

Muon trigger efficiency: the trigger efficiency uncertainty is 2% for the muon leg. For embedded samples this implies 4%, as a di-muon trigger is used.

***b*-tagging efficiency**: this uncertainty is due to the application of the b-tagged jet veto in the $\tau_\mu\tau_h$ channel. The uncertainty applies only to $t\bar{t}$, single-t, and di-boson process yields. The b-tagging scale factors are modified within their uncertainties (measured as functions of p_{T} and $|\eta|$) and these changes are propagated to give rise to the final uncertainty on the b-tagging efficiency. These are between 1%–9% large and are uncorrelated between years.

Cross-sections of background processes: uncertainties on the cross-sections of different processes. These are all considered to be fully correlated across the three data-taking eras.

- $Z/\gamma^* \to \ell\bar{\ell}$: uncertainty of 2% to account for the uncertainty on the Z/γ^* cross-section.

- $t\bar{t}$: uncertainty of 4.2%.
- $\mathrm{W}+\mathrm{jets}$: uncertainty of 4%.
- electroweak Z: uncertainty of 4%.
- diboson and single-top: uncertainty of 5%.

Cross-sections of signal processes: uncertainties on the production processes. The size of the uncertainties are taken from Reference [19]. Additionally, for each production mode, an uncertainty on the $\mathrm{H} \rightarrow \tau^{+}\tau^{-}$ branching fraction is applied. These are all considered to be fully correlated across the three data-taking eras.

- ggH production: 3.9% and 3.2% due to variations on the QCD scale and $\mathrm{PDF}+\alpha_s$, respectively.
- VBF production: 0.4% (QCD scale) and 2.1% $(\mathrm{PDF}+\alpha_s)$.
- WH production: 1.9% (QCD scale) and 1.9% $(\mathrm{PDF}+\alpha_s)$.
- ZH production: 1.6% (QCD scale) and 1.6% $(\mathrm{PDF}+\alpha_s)$.

Yield on embedded events: an uncertainty of 4% is applied on the yield of embedded events. This uncertainty is treated as uncorrelated across years as the triggers differ.

$l \rightarrow \tau_h$ fake-rate: an uncertainty that ranges between 20%–40% is applied that mainly affects the $\mathrm{Z}/\gamma^{*} \rightarrow \ell\bar{\ell}$ yield. This uncertainties is treated as 50% correlated between the years.

6.8.2 Shape uncertainties

Shape uncertainties affect both the yield and shape of the signal and background distributions used in the extraction of the $\mathcal{CP}$-mixing angle α_{gg}. Listed below are the different sources of uncertainty that arise from an associated correction being applied to simulation. Together with the size of the uncertainty, the correlation between each data-taking era is briefly discussed.

τ_h-lepton trigger efficiency: the uncertainty on τ_h-lepton trigger efficiency depends on the p_{T} and decay mode, and is statistical in nature.

τ_h-lepton energy scale: for each decay mode, an uncorrelated uncertainty is applied. This uncertainty amounts to 0.8-1.1% and 0.2-0.5% for MC simulated and embedded events, respectively. The variations due to the energy scale are propagated to the $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$.

μ energy scale: detector-region-dependent uncertainties of 0.4% in the barrel, 0.9% in the near endcap and 1.7% in the far endcap of the muon system are applied to cover the μ

energy scale uncertainty. These are correlated between the years. The variations due to the energy scale are propagated to the $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$.

Jet energy scale: the uncertainty on the jet energy scale is provided for different detector regions and reflects the correlations between the different sources.

Jet energy resolution: uncertainties arising form the p_{T} smearing of jets in simulation to match the energy resolution of jets in data. These are treated uncorrelated between years.

t-quark p_{T} reweighting: an uncertainty of 100% is applied to $t\bar{t}$ events, defined by not applying the reweighting and applying the correction twice.

$\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ unclustered energy scale: for MC events originating from $t\bar{t}$, single-t and di-boson processes the $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ unclustered energy scale uncertainties are applied. These are processes where the $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ recoil corrections are not applied to. These are uncorrelated between years as the dominant parts are the statistical components.

$\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ recoil corrections: by varying the recoil correction parameters within their uncertainties, these are propagated to the $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ recoil corrections and used as estimations for the uncertainty. These are uncorrelated between years as the dominant parts are the statistical components.

Z/γ^* p_{T} reweighting: an uncertainty of 100% is applied to $\mathrm{Z}/\gamma^* \to \ell\bar{\ell}$ events, defined by not applying the reweighting and applying the correction twice.

Fake-factor uncertainties: these uncertainties vary depending on the region and are statistical and systematic in nature:

- statistical: fitted parameters of the fake-factors are treated as nuisances in the fit.
- systematic due to non-closure corrections.
- systematic due to extrapolation between regions.
- systematic due to the difference between observed and simulated events.

Limited statistics: limitations in the signal and background events statistics are considered using the Barlow-Beeston method [90]. A separate nuisance parameter is assigned to each bin for each process.

Table 6.4: Sources of systematic uncertainties, where the correlation, if any, between the years is indicated in the third column. The type of probability density function (PDF) is also indicated by the last column.

Uncertainty	Magnitude	Correlation	PDF
τ_h ID	p_{T}/decay-mode dependent (2–3%)	no	Gaussian
Muon reconstruction	1%.	yes	log-normal
$e \to \tau_h$ ID	5(1)% 2016(2017,2018)	no	Gaussian
$\mu \to \tau_h$ ID	20–40%	no	Gaussian
μ ID	1%	yes	Gaussian
b-jet veto	1–9%	no	log-normal
Luminosity	2.3%–2.5%	partial	log-normal
Trigger	2% for μ, p_{T}-dep. for τ_h	no	Gaussian
Embedded yield	4%	no	log-normal
$t\bar{t}$ cross-section	4.2%	yes	log-normal
Diboson cross-section	5%	yes	log-normal
Single-t cross-section	5%	yes	log-normal
$\mathrm{W} + \mathrm{jets}$ cross-section	4%	yes	log-normal
Drell-Yan cross-section	2%	yes	log-normal
Signal cross-sections	Reference [19]	yes	log-normal
Parton shower	Signal-dependent	yes	Gaussian
Renormalisation scale	Signal-dependent	yes	log-normal
Factorisation scale	Signal-dependent	yes	log-normal
t-quark p_{T} reweighing	100%	yes	Gaussian
Z/γ^* p_{T} reweighing	100%	partial	Gaussian
Prefiring (2016, 2017)	Event-dependent (0–4%)	yes	log-normal
τ_h energy scale	1% (sim), 1.5% (emb.)	no	Gaussian
$\mu \to \tau_h$ energy scale	1%	no	log-normal
Muon energy scale	0.4–2.7%	yes	Gaussian
Jet energy scale	Event-dependent	partial	Gaussian
Jet energy resolution	Event-dependent	no	Gaussian
$\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ unclustered scale	Event-dependent	no	Gaussian
$\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ recoil corrections	Event-dependent	no	Gaussian
Jet$\to \tau_h$ mis-ID	FF-dependent	partial	Gaussian
$t\bar{t}$/diboson in embedded	10%	yes	Gaussian

6.9 Statistical inference

The extraction on the $\mathcal{CP}$-mixing angle α_{gg} is performed using a simultaneous template fit, where the best-fit values occur when likelihood function $\mathcal{L}\left(\vec{\sigma}, \alpha_{gg}, \vec{\theta} \mid \text{data}\right)$ is maximised. In the likelihood function, $\vec{\sigma} = \left(\sigma_{gg\text{H}}, \sigma_{\text{VBF}}, \sigma_{\text{WH}}, \sigma_{\text{ZH}}\right)$ represents the Higgs boson production cross-section of each process considered in the analysis, α_{gg} is the $\mathcal{CP}$-mixing angle that will be measured, and $\vec{\theta}$ represents the full set of parameters accounting for the statistical, experimental and theoretical systematic uncertainties present in this measurement and are also commonly referred to as *nuisance* parameters of the fit. The likelihood function is written as a product of Poisson probabilities P in the following manner:

$$\mathcal{L}\left(\vec{\sigma}, \alpha_{gg}, \vec{\theta} \mid \text{data}\right) = \prod_{j \in \text{categories}} \prod_{i \in \text{bins}} P(n_{i,j} \mid S_{i,j}(\vec{\theta}, \alpha_{gg}) + B_{i,j}(\vec{\theta})) \times \prod_{m \in \text{nuisances}} C_m(\vec{\theta}) \,. \tag{6.3}$$

In Equation 6.3, the Poisson probabilities P correspond to the number of observations $n_{i,j}$ in bin i of the discriminating observable for category j given the expectation for the background $B_{i,j}$ and for the prediction from the signal $S_{i,j}(\vec{\theta}, \alpha_{gg}) = L \cdot \vec{\sigma} \cdot \vec{A}_{i,j}(\vec{\theta}, \alpha_{gg})$, with L as the integrated luminosity and $\vec{A}_{i,j}(\vec{\theta}, \alpha_{gg})$ is the acceptance in each bin. The nuisance constraints are parameterised by the functions $C_m(\vec{\theta})$, which a priori are taken as Gaussian PDF. While fitting for α_{gg} we require a rate parameter to scale each of production processes × the branching fraction of the $\text{H} \rightarrow \tau^+\tau^-$ decay, where the branching fraction is included this way as this analysis is not trying to disentangle alterations of the branching ratio and cross-sections. These rate parameters will be written as $\vec{\mu}^{\tau\tau} = (\mu_{gg\text{H}}, \mu_{qq\text{H}})$, where the $\mu_{gg\text{H}}$ scales the ggH cross-section and $\mu_{qq\text{H}})$ scales both the vector-boson-fusion and associated Higgs boson production cross-sections simultaneously. While extracting the best-fit value for α_{gg} these rate parameters are estimated by fit, such that they are considered unconstrained, but have a lower bound to ensure only positive values may be selected.

All of the categories from the 2016, 2017 and 2018 data-taking eras are used in the final fit, together with the uncertainty model discussed in Section 6.8, which is introduced as a constraint in Equation 6.3 in $C_m(\vec{\theta})$. The best-fit value for α_{gg} is estimated where the likelihood function is maximised. The 68% confidence level (CL), for instance, may be defined as the points on the PDF where 68% of the probability is covered. This can be expressed using the negative log-likelihood ratio as

$$-2\Delta \log \mathcal{L} = -2 \left(\log \frac{\mathcal{L}(\alpha_{gg})}{\mathcal{L}(\hat{\alpha}_{gg})} \right) \,, \tag{6.4}$$

where $\hat{\alpha}_{gg}$ is the best-fit value of the parameter of interest α_{gg}. The 68% confidence interval may be determined when $-2\Delta \log \mathcal{L} = 0.99$, whereas the 95% confidence interval lies where $-2\Delta \log \mathcal{L} = 3.84$.

In addition to the extraction of α_{gg}, the same statistical procedures may be applied to the rate parameters $\mu^{\tau\tau}_{gg\mathrm{H}}$ and $\mu^{\tau\tau}_{\mathrm{V}}$. In this case α_{gg} is estimated by the fit and the rate parameters are implemented as the parameter of interest. Fits can be performed separately for each of these. Additionally the total rate parameter modifier $\mu^{\tau\tau}$ may be extracted.

For expected results an Asimov dataset is used to perform the simultaneous fit to, rather than the actual observed data itself. The Asimov dataset is the model prediction with all parameters set to their nominal values compatible with the null hypothesis.

6.10 Results

Prior to the extraction of α_{gg}, tests are performed to validate the compatibility between data and the model. These goodness-of-fit tests are based on estimating a χ^2-statistics from the distributions and the number of degrees of freedom. Categories with p-values below a threshold of 5% are considered to have failed the goodness-of-fit test. No such values have been observed, thus the model is confirmed to be compatible with the data.

The distributions of the categories used in the fit are shown in Figures 6.9- 6.14, where the parameters are set to the best-fit values estimated by the fit. In the first couple of Figures the 0-jet and boosted categories that do not add any direct sensitivity to α_{gg} are presented. These do improve the measurement, as they help to constrain background uncertainties. The gain of including these is about 20%. The $\mathcal{CP}$-sensitive categories are illustrated in Figures 6.11- 6.14. The ggH is shown with the parameter of α_{gg} set to its measured value and, for comparison, is plotted with the pseudo-scalar ggH hypothesis, where the value of $\alpha_{gg} = 90^\circ$.

The scan of the negative log-likelihood for the fit extracting the best-fit value for α_{gg} is shown in Figure 6.15. The best-fit value is $\hat{\alpha}_{gg} = (-5^{+36}_{-37})^\circ$, which is well consistent with the SM expected value of 0. The rate parameter $\mu^{\tau\tau}_{gg\mathrm{H}}$ has been measured as $\hat{\mu}^{\tau\tau}_{gg} = 0.63^{+0.22}_{-0.21}$, whereas $\mu^{\tau\tau}_{\mathrm{V}}$ is found to be $0.97^{+0.31}_{-0.30}$, both at a confidence level of 68%. This is depicted in Figure 6.16, together with the expected sensitivities using an Asimov dataset. The total rate parameter, $\mu^{\tau\tau}$, that sets the Higgs boson production mode is shown in Figure 6.17. This parameter is a combination of the ggH and VBF + VH rate parameters, and is measured to be $0.77^{+0.13}_{-0.12}$ at 68% CL.

The dominant sources of uncertainty in the fit to data are statistical in nature. This is followed by theory uncertainties, the jet energy scale uncertainties, and the hadronic τ_h-lepton trigger efficiency.

Two-dimensional scans can be performed to observe the correlation between two parameters. In this case, the 68%, 95%, and 99.7% confidence intervals are determined when the two-dimensional negative log-likelihood satisfies $-2\Delta \ln \mathcal{L}_{2D} = 2.30, 5.99, 11.62$. Two-dimensional

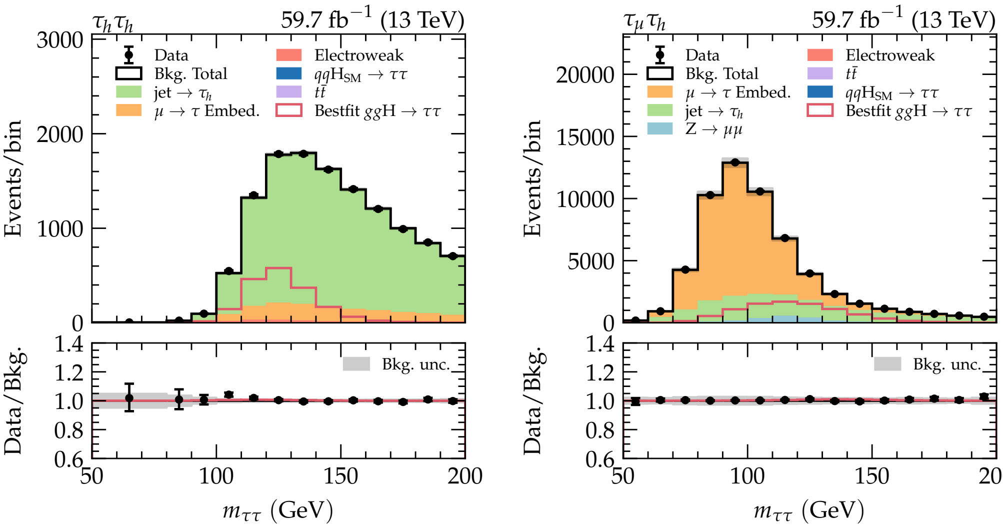

Figure 6.9: Distributions of the SVFIT mass for the 0-jet category for the $\tau_h\tau_h$ (left) and $\tau_\mu\tau_h$ (right) channels. The signal distribution is scaled by a factor of 50. The distributions are post-fit and contain the full background uncertainty model in the grey band.

fits have been extracted for the $\mu_{gg\mathrm{H}}^{\tau\tau}$ against $\mu_{\mathrm{V}}^{\tau\tau}$, as shown in Figure 6.18. A negative correlation between the two parameters has been observed, which can be explained by the categorisation methods used. The analysis is not attempting to separate ggH and VBF events in a very efficient manner, therefore, as one rate parameter is pulling the Higgs boson production rate up, the other tends to be pulled down. Another two-dimensional scan performed is for gluon-fusion rate parameter $\mu_{gg\mathrm{H}}^{\tau\tau}$ against α_{gg}, as illustrated in Figure 6.19. No strong correlation between the two parameters of interest have been observed. Finally, an interpretation in terms of the Higgs boson to t-quark Yukawa couplings is shown in Figure 6.20.

6.10.1 Comparison of expected sensitivity with BDT-based categorisation

The BDT-based categorisation scheme introduced in Section 6.5 is now directly compared to the cut-based analysis presented in the previous section. The metric will be the expected sensitivity to α_{gg} with an Asimov dataset based on the 2018 data-taking era. The $\mathcal{CP}$-sensitive categories used in the fit for the BDT method are presented in Figures 6.21 and 6.22 for the case where ggH and VBF are individual categories in the training, and Figures 6.23 and 6.24 for the training with VBF merged into the ggH class. These plots are now shown in windows of BDT score ranging between 0 and 1 for the ggH and VBF categories. It is clear from these Figures that the sensitivity to the $\mathcal{CP}$-mixing angle measurement in the case where the training events are merged arises from both the ggH and VBF category, as can be observed in the last BDT bin.

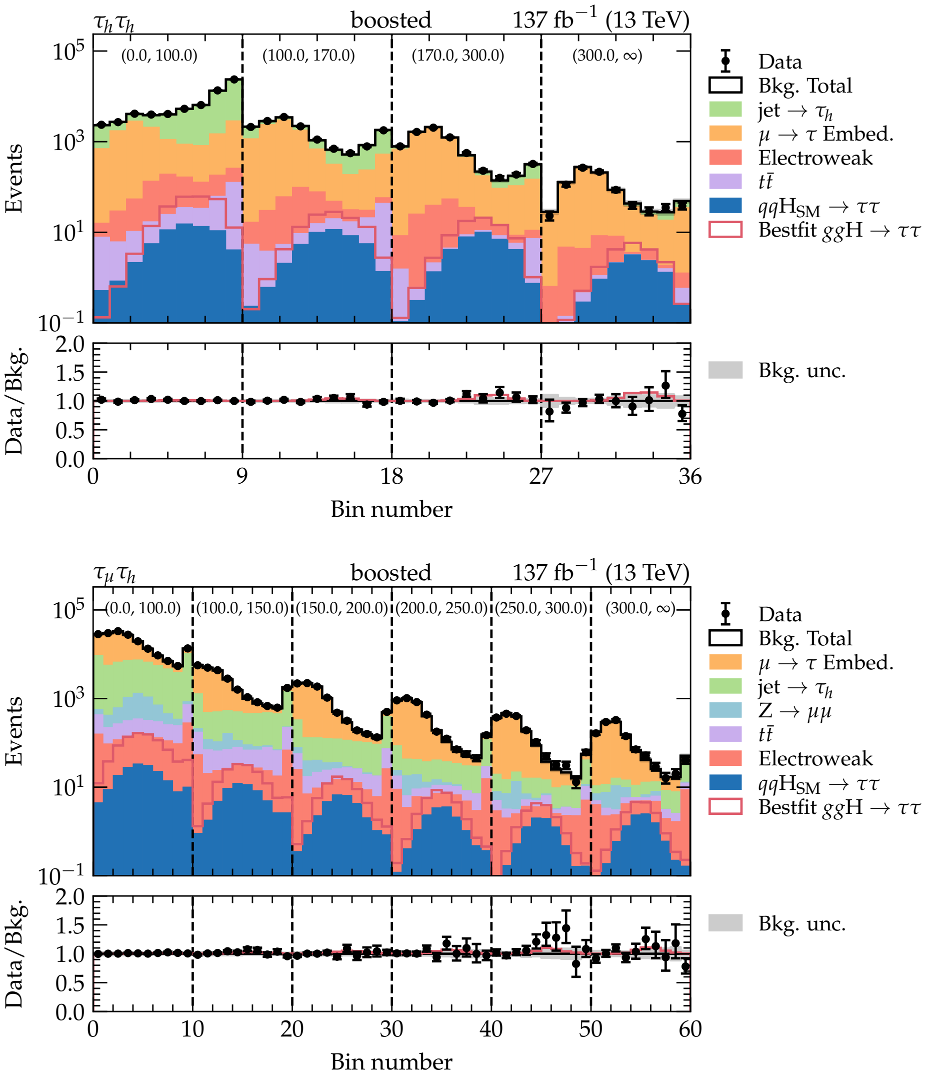

Figure 6.10: Distributions of the SVFIT mass in windows of Higgs boson p_T for the 0-jet category of the $\tau_h\tau_h$ (top) and $\tau_\mu\tau_h$ (bottom) channels. The distributions are post-fit and contain the full background uncertainty model in the grey band.

Additionally, these will be events with higher m_{jj} and thus provide more separation between different $\mathcal{CP}$ states. Therefore the expectation is that the merged-class training scenario should outperform at least the split-class training, possibly also the cut-based analysis.

The expected scans of alpha using an Asimov dataset based on predicted events under 2018 data-taking conditions are presented in Figure 6.25. From these likelihood scans one can conclude that the most sensitive measurement is expected to occur for the BDT-based cate-

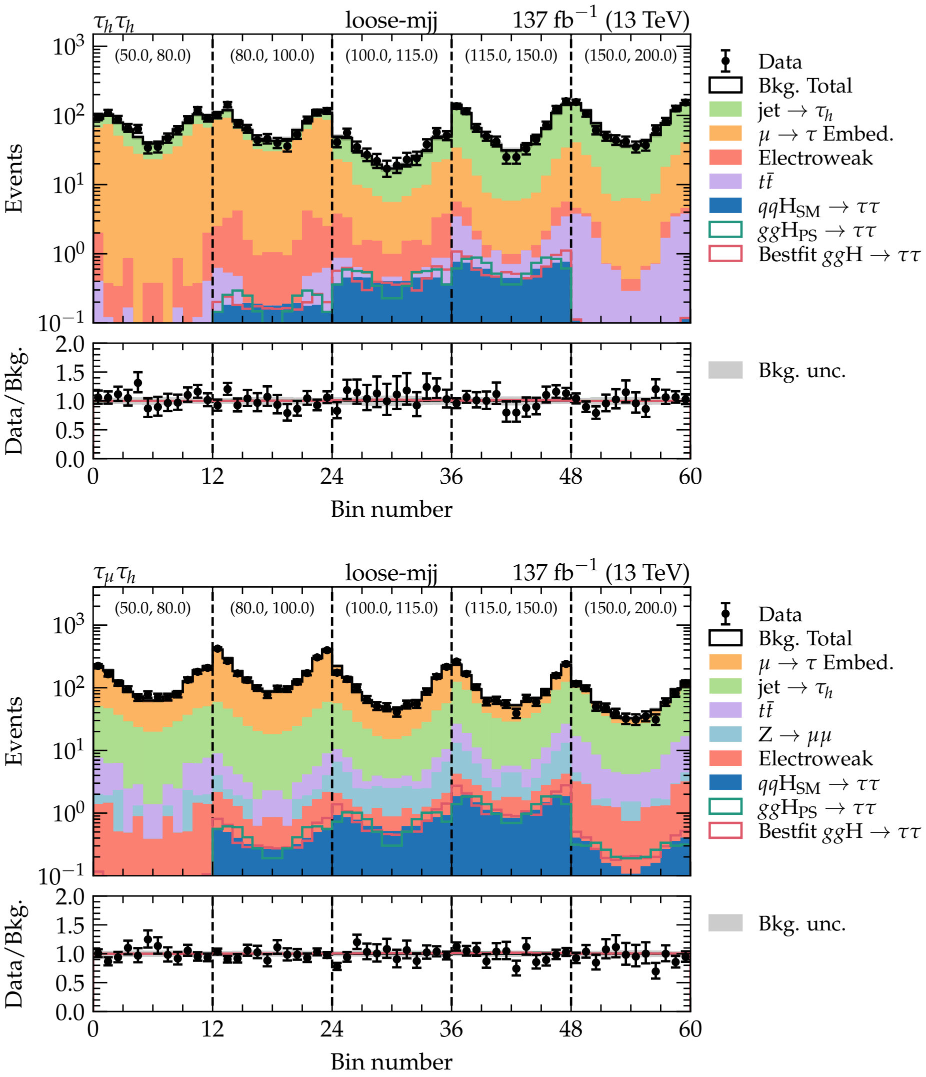

Figure 6.11: Distributions of the $\mathcal{CP}$-discriminating variable $\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2})$ in windows of SVFIT mass for the loose-mjj category of the $\tau_h\tau_h$ (top) and $\tau_\mu\tau_h$ (bottom) channels. The distributions are post-fit and contain the full background uncertainty model in the grey band.

gorisation with a single combined ggH and VBF training class, resulting in a 16% gain for the distinction between pure $\mathcal{CP}$-even and pure $\mathcal{CP}$-odd. The training with separate ggH and VBF classes produces a similar shape for the scan for mixed $\mathcal{CP}$ scenarios as the cut-based analysis, but outperforms by about 3% at the pure $\mathcal{CP}$-odd hypotheses.

Another interesting comparison that can be made is for the ggH signal strength modifier, $\mu^{\tau\tau}_{gg\mathrm{H}}$. The scan for the three cases is shown in Figure 6.26, where all categorisation methods

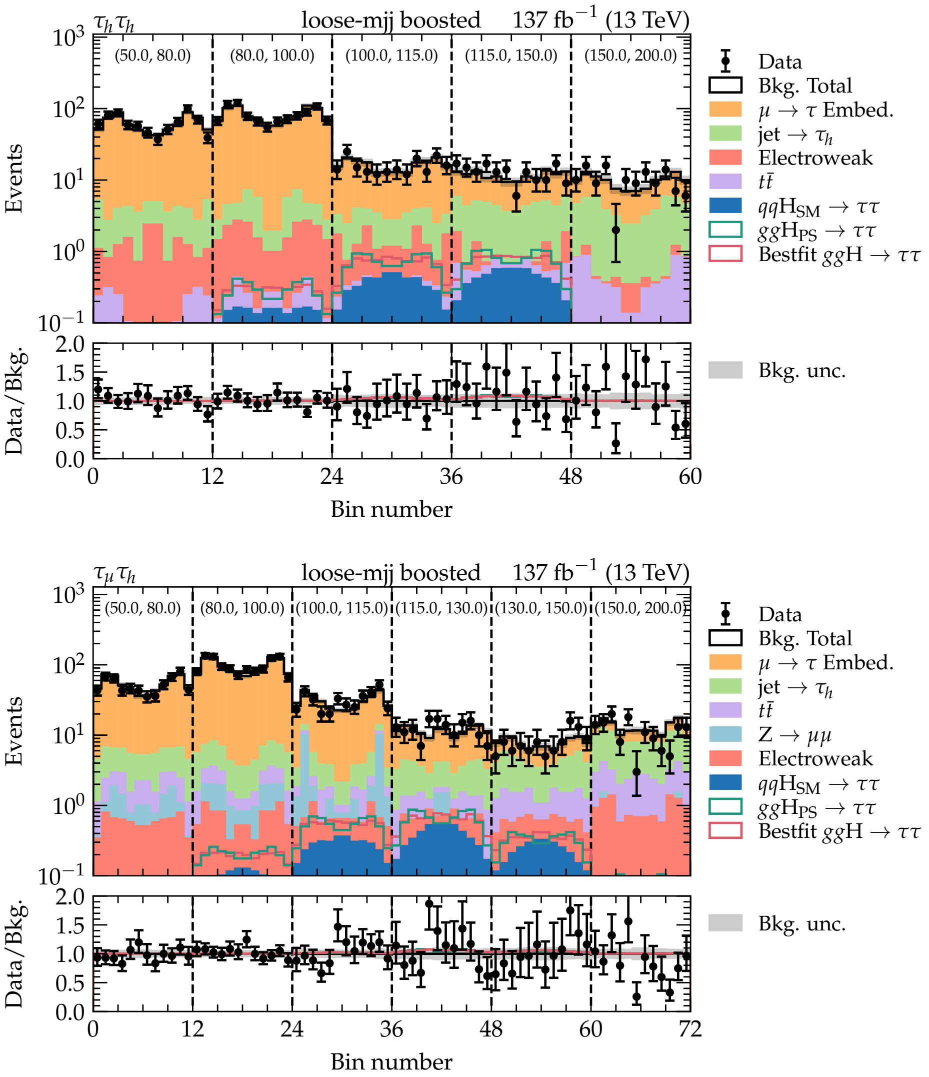

Figure 6.12: Distributions of the $\mathcal{CP}$-discriminating variable $\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2})$ in windows of SVFIT mass for the loose-mjj boosted category of the $\tau_h\tau_h$ (top) and $\tau_\mu\tau_h$ (bottom) channels. The distributions are post-fit and contain the full background uncertainty model in the grey band.

perform quite similarly at 68% CL. However, the performance gain of the BDT-based methods becomes more evident at 95% CL. A metric that can be quoted is the (expected) sensitivity of discovery, derived from comparing the $\mu^{\tau\tau}_{gg\mathrm{H}} = 0$ and $\mu^{\tau\tau}_{gg\mathrm{H}} = 1$ likelihoods. The split-class training outperforms both the cut-based and merge-class training fits. This is not unexpected, as in this case the BDT algorithm attempts to differentiate specifically between ggH and VBF events, leading to a better measurement of the $\mu^{\tau\tau}_{gg\mathrm{H}}$ signal strength. Nevertheless, the

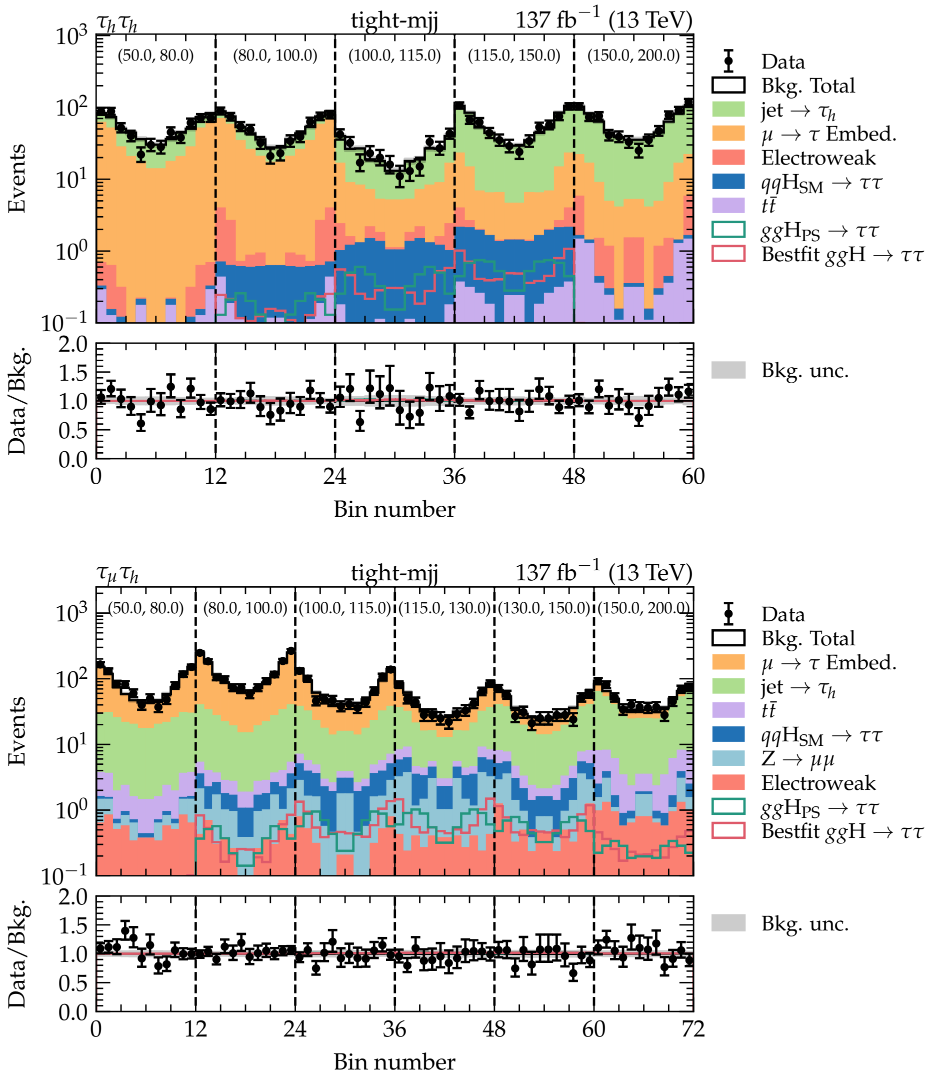

Figure 6.13: Distributions of the $\mathcal{CP}$-discriminating variable $\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2})$ in windows of SVFIT mass for the tight-mjj category of the $\tau_h\tau_h$ (top) and $\tau_\mu\tau_h$ (bottom) channels. The distributions are post-fit and contain the full background uncertainty model in the grey band.

training with merged ggH and VBF event classes results in an expected 8% gain for the $\mu^{\tau\tau}_{gg\mathrm{H}}$ measurement, which is a positive result, given it provides a 16% more precise measurement of the $\mathcal{CP}$-mixing angle. Therefore, the result can be improved in future iterations using more complex categorisation methods, such as a set of BDTs or more complicated machine learning algorithms. This study shows promising results which will be useful in other analyses, such as the next one discussed in this Thesis.

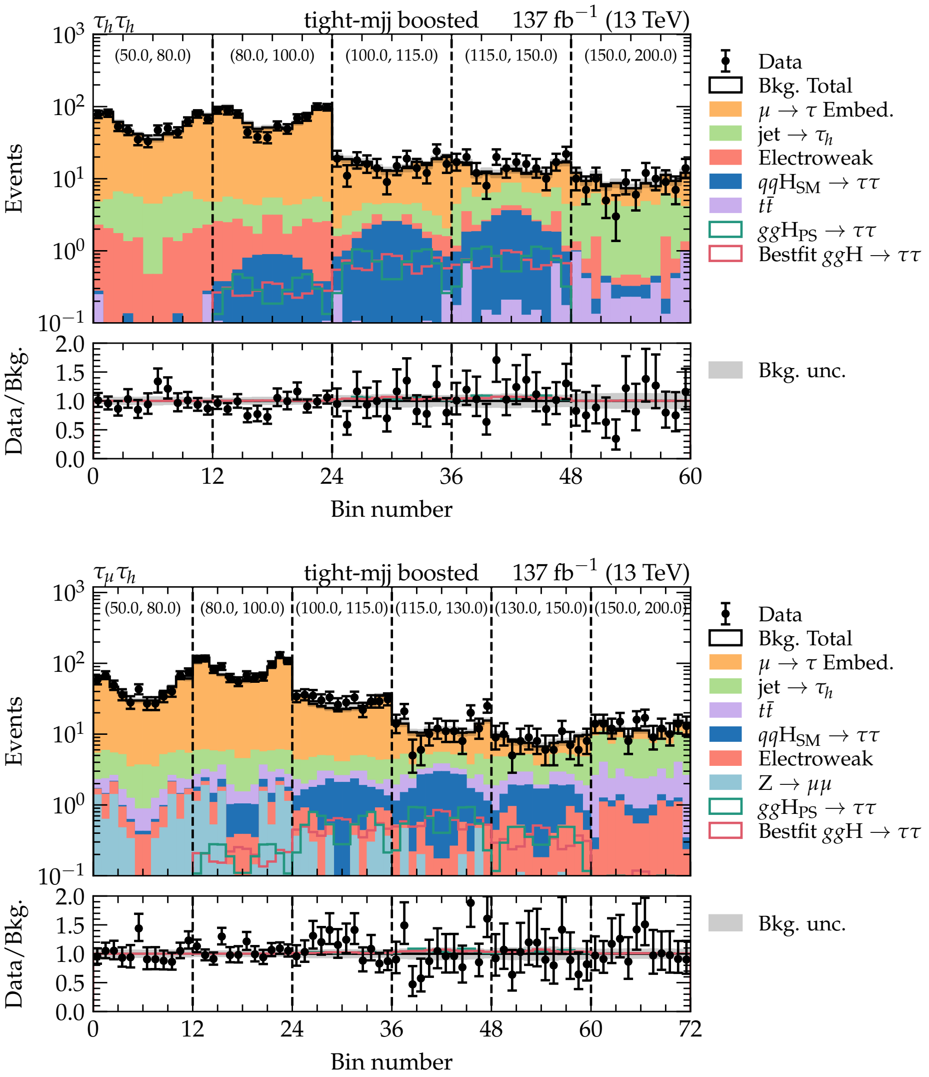

Figure 6.14: Distributions of the $\mathcal{CP}$-discriminating variable $\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2})$ in windows of SVFIT mass for the tight-mjj boosted category of the $\tau_h\tau_h$ (top) and $\tau_\mu\tau_h$ (bottom) channels. The distributions are post-fit and contain the full background uncertainty model in the grey band.

6.11 Summary

The analysis presenting a measurement of the $\mathcal{CP}$ structure of the Yukawa coupling between the Higgs boson and the t-quark through an effective coupling with gluons has been presented. This particular $\mathcal{CP}$-mixing angle, α_{gg}, is sensitive to the azimuthal separation of the two leading final-state jets in $gg \to$ H production, which is used as the final discriminating

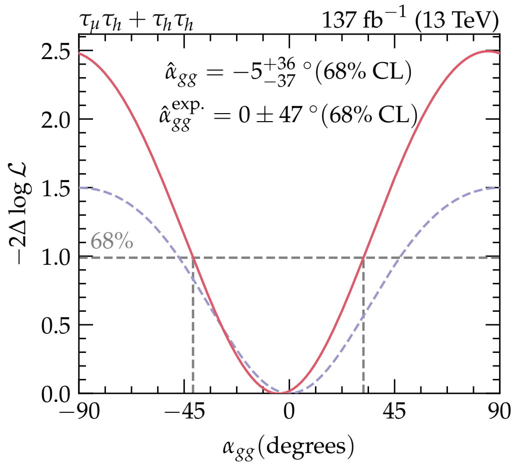

Figure 6.15: Scan of the negative log-likelihood $-2\Delta \log \mathcal{L}$ for parameter of interest α_{gg}. The expected sensitivity to α_{gg} is $(0 \pm 47)^\circ$, whereas the observed, measured value amounts to $(-5^{+36}_{-37})^\circ$ at a confidence level of 68%. The measurement is therefore well consistent with the SM prediction.

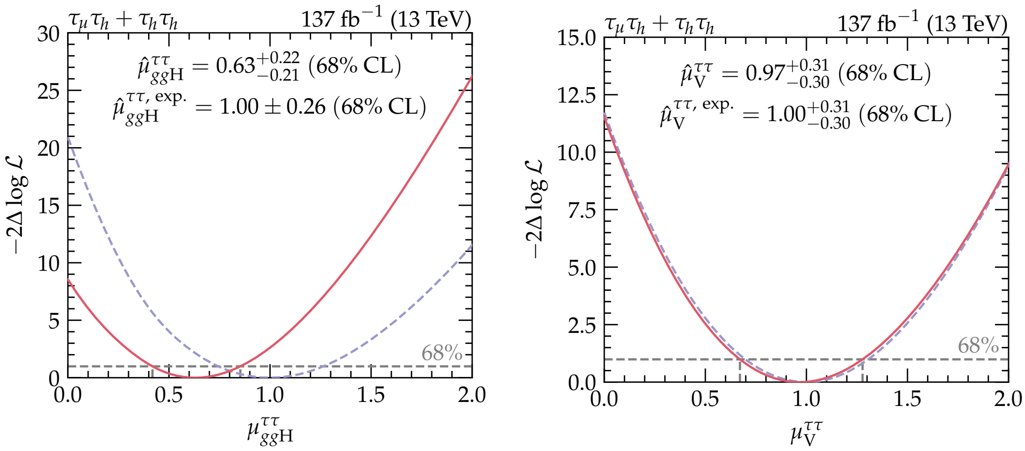

Figure 6.16: Scans of the negative log-likelihood $-2\Delta \log \mathcal{L}$ for parameter of interest $\mu^{\tau\tau}_{gg\text{H}}$ (left) and $\mu^{\tau\tau}_{\text{V}}$ (right). The expected sensitivity to $\mu^{\tau\tau}_{gg\text{H}}$ is 1.00 ± 0.26, whereas the observed value fluctuates downwards to $0.63^{+0.22}_{-0.21}$ at 68% CL. For $\mu^{\tau\tau}_{\text{V}}$, on the other hand, the expected sensitivity is $1.00^{+0.31}_{-0.30}$, and the observed value is determined to be $0.97^{+0.31}_{-0.30}$ at 68% CL.

variable. Through a simultaneous maximum-likelihood fit of all categories, the result has been obtained to be $\hat{\alpha}_{gg} = (-5^{+36}_{-37})^\circ$ at the 68% confidence level and an integrated luminosity of 137fb^{-1}, which tends towards the $\mathcal{CP}$-even hypothesis. The total rate parameter is measured as $\hat{\mu}^{\tau\tau} = 0.77^{+0.13}_{-0.12}$ at the 68% confidence level, which is in tension with the standard model prediction by only about 1.5σ. Therefore, the measurements are well consistent with a SM

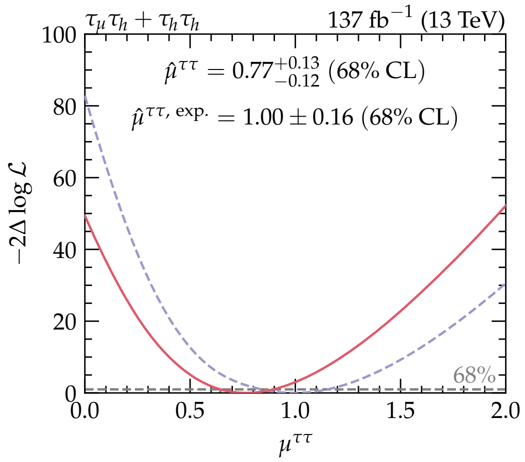

Figure 6.17: Scan of the negative log-likelihood $-2\Delta \log \mathcal{L}$ for the rate parameter scaling the total Higgs boson production cross-section $\mu^{\tau\tau}$. The expected sensitivity is 1.00 ± 0.16, whereas the observed value is $0.77^{+0.13}_{-0.12}$ at 68% CL.

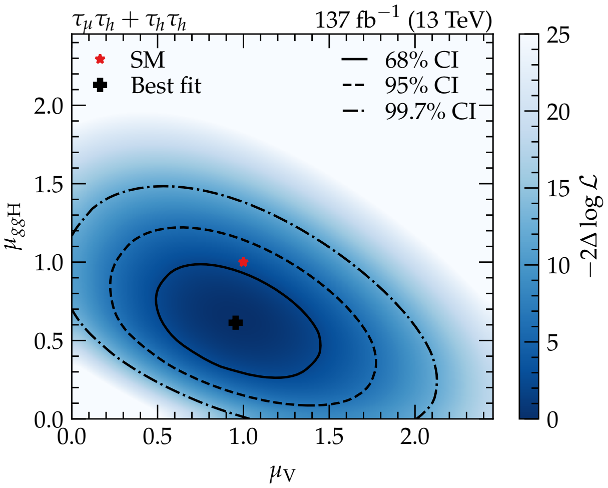

Figure 6.18: Two-dimensional parameter scan of the negative log-likelihood of the rate parameters.

Higgs boson of $\mathcal{CP}$-even nature. An alternative categorisation procedure has been presented, which relies on a set of multi-class BDTs that categorise events into most likely production mechanisms of the final-state dilepton system based on kinematic variables of the leptons and jets. This method provides $\mathcal{O}$ (10%) more stringent expected results for both the $\mathcal{CP}$-mixing angle and ggH signal strength modifier.

As is common with low statistics measurements, more data will help this measurement. This can be of the form of including more final states, such as the $\tau_e\tau_h$ and $\tau_e\tau_\mu$ channels, and more

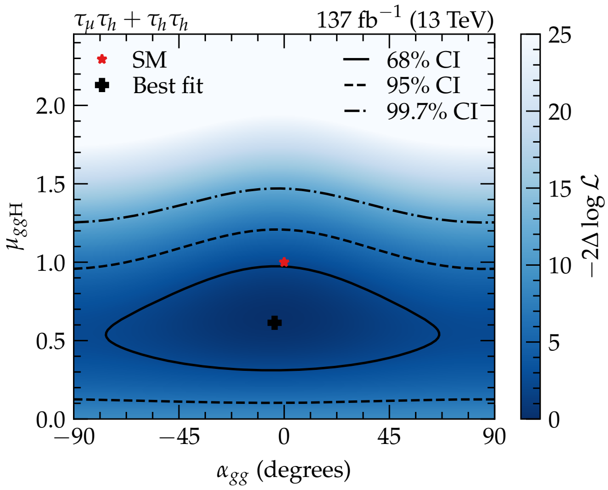

Figure 6.19: Two-dimensional parameter scan of the negative log-likelihood of the ggH rate parameter against α_{gg}. The 68%, 95%, and 99.7% CL regions are indicated as contours.

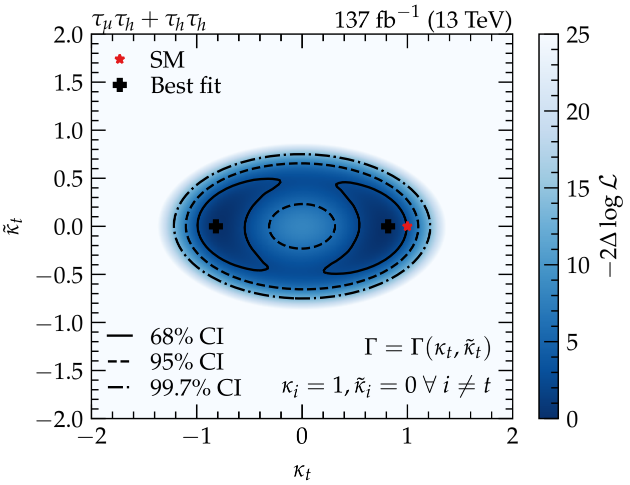

Figure 6.20: Two-dimensional scan of the negative log-likelihood of the $\mathcal{CP}$-even and $\mathcal{CP}$-odd Yukawa coupling constants κ_t and $\tilde{\kappa}_t$. The 68%, 95%, and 99.7% CL regions are indicated as contours. Higgs boson couplings to particles other than the t-quark are set to their SM predicted values.

datasets, which involves collecting more data during future collision runs. Additionally, as demonstrated, machine learning techniques can help provide more precise results.

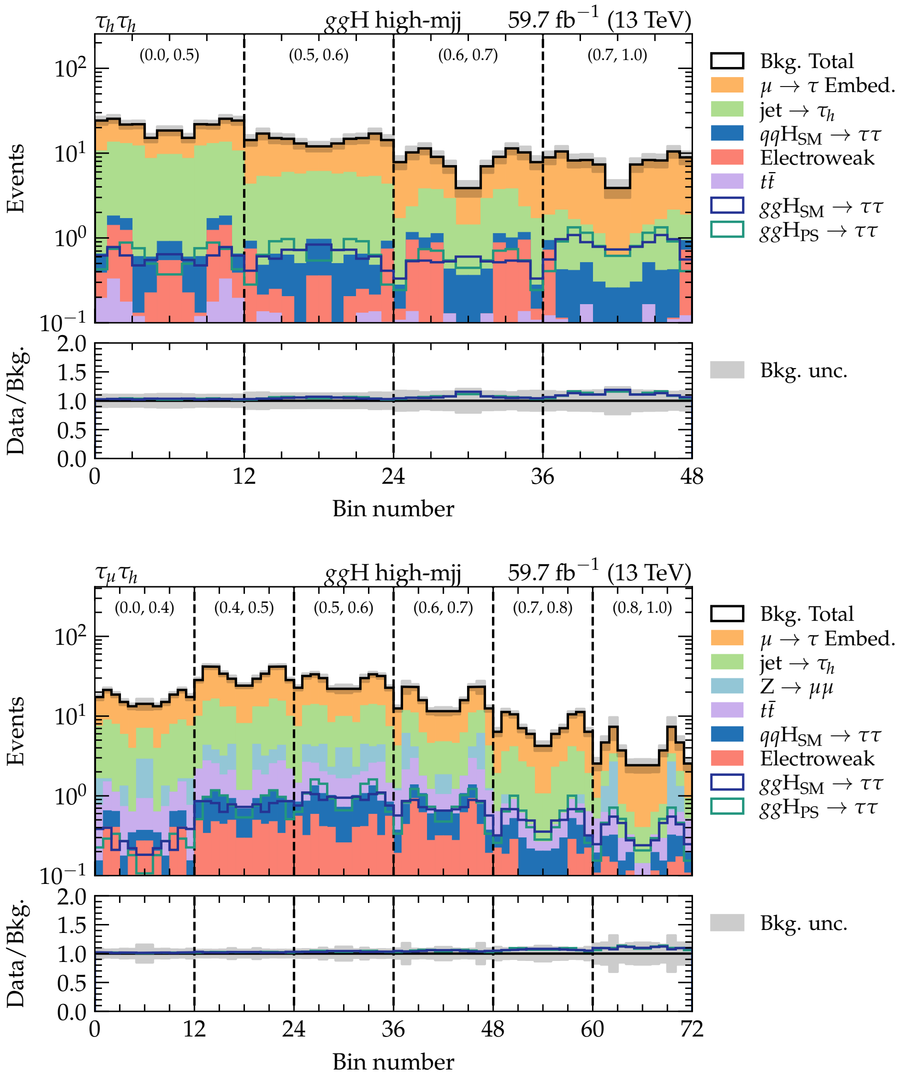

Figure 6.21: Distributions of the $\mathcal{CP}$-discriminating variable $\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2})$ in windows of BDT score for the ggH high-mjj category of the $\tau_h\tau_h$ (top) and $\tau_\mu\tau_h$ (bottom) channels for the exclusive training class definition of ggH and VBF events. The distributions are post-fit and contain the full background uncertainty model in the grey band.

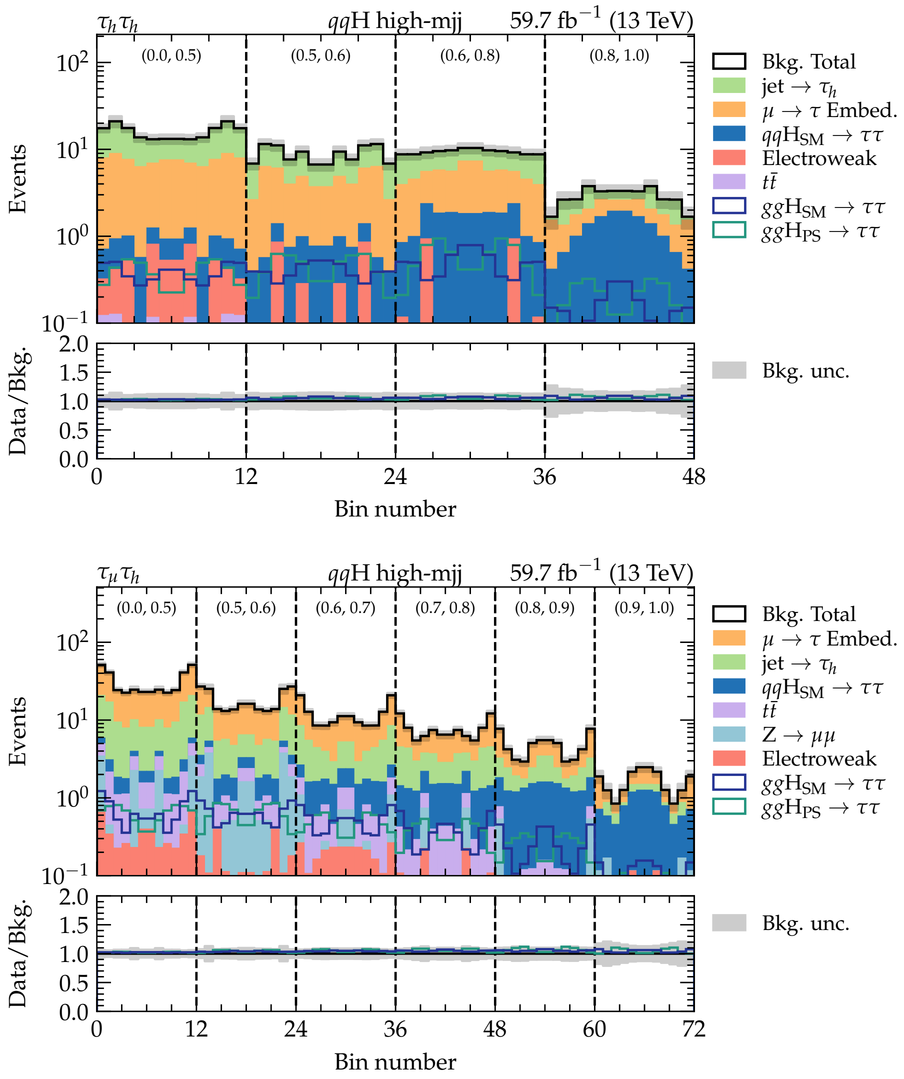

Figure 6.22: Distributions of the $\mathcal{CP}$-discriminating variable $\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2})$ in windows of BDT score for the VBF high-mjj category of the $\tau_h\tau_h$ (top) and $\tau_\mu\tau_h$ (bottom) channels for the exclusive training class definition of ggH and VBF events. The distributions are post-fit and contain the full background uncertainty model in the grey band.

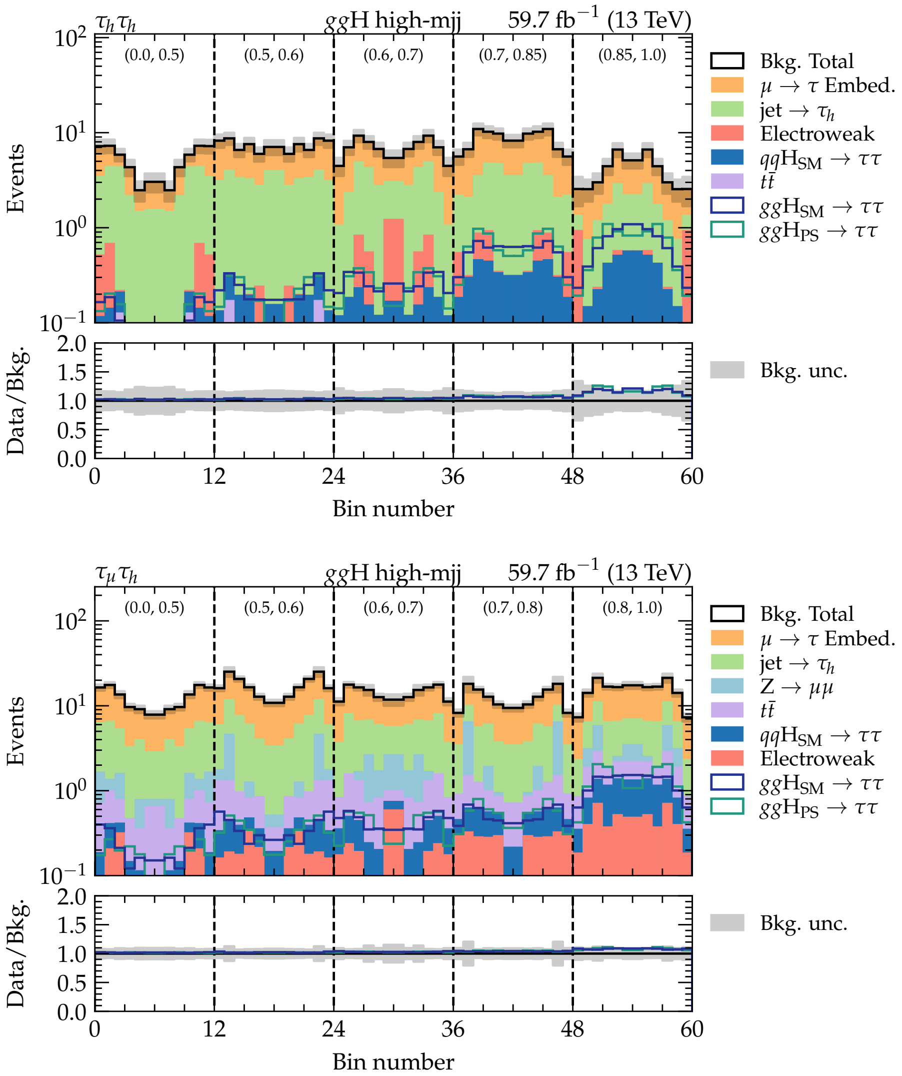

Figure 6.23: Distributions of the $\mathcal{CP}$-discriminating variable $\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2})$ in windows of BDT score for the ggH high-mjj category of the $\tau_h\tau_h$ (top) and $\tau_\mu\tau_h$ (bottom) channels for the inclusive training class definition of ggH and VBF events. The distributions are post-fit and contain the full background uncertainty model in the grey band.

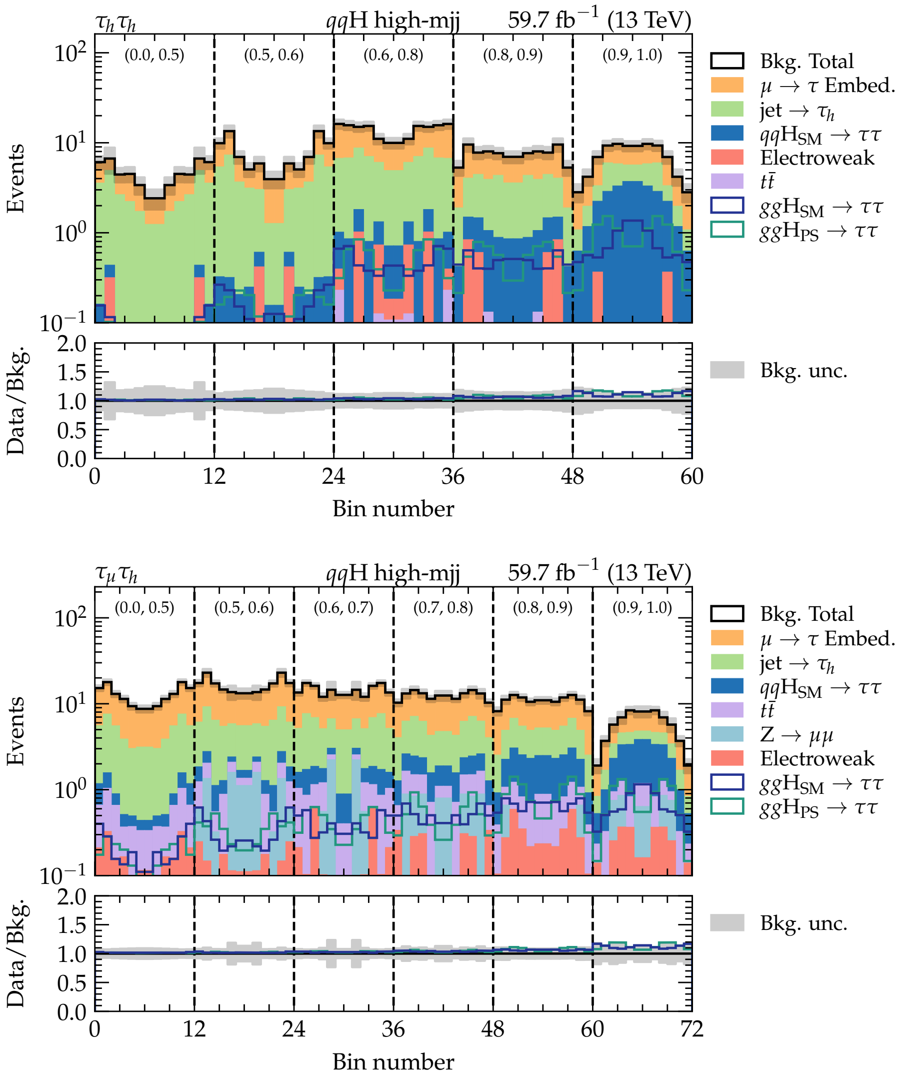

Figure 6.24: Distributions of the $\mathcal{CP}$-discriminating variable $\Delta\phi(\vec{p}_{j_1}, \vec{p}_{j_2})$ in windows of BDT score for the VBF high-mjj category of the $\tau_h\tau_h$ (top) and $\tau_\mu\tau_h$ (bottom) channels for the inclusive training class definition of ggH and VBF events. The distributions are post-fit and contain the full background uncertainty model in the grey band.

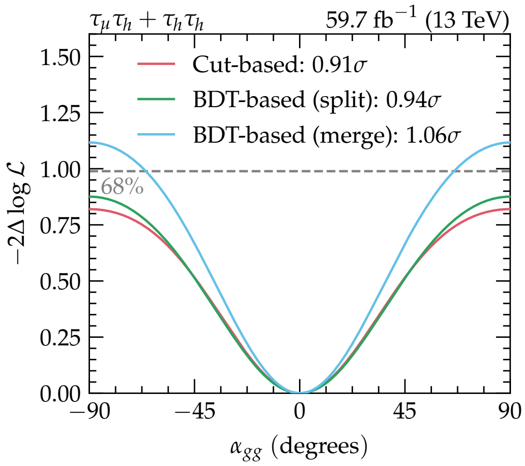

Figure 6.25: Comparison of expected sensitivities to the $\mathcal{CP}$-mixing angle α_{gg} using different categorisation methods. The categorisation attempted using a set of multi-class BDTs trained on three classes of events, where the ggH and VBF events are mixed together, performs the best.

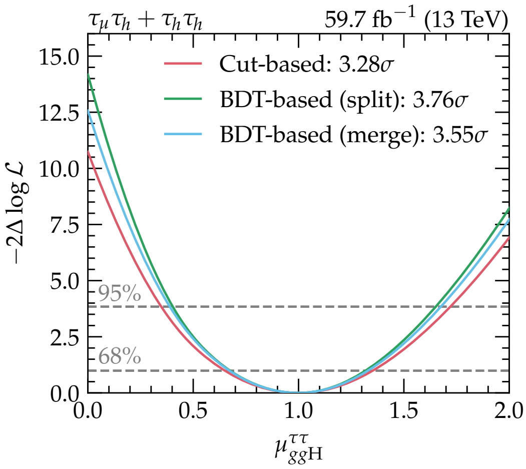

Figure 6.26: Comparison of expected sensitivities to the signal strength modifier $\mu^{\tau\tau}_{gg\mathrm{H}}$ using different categorisation methods. The categorisation attempted using a set of multi-class BDTs trained on four classes of events, where the ggH and VBF events are separate training classes, performs the best.

Chapter 7

Measurement of the Higgs Boson $\mathcal{CP}$ State using τ-lepton Decays

7.1 Introduction

The measurements of the $\mathcal{CP}$ properties of the Higgs boson can be performed in multiple ways. This chapter will outline the methods employed to understand the $\mathcal{CP}$ mixing angle using the ggH production vertex in association with two jets. As described in Chapter 2, this involves the study of the Yukawa coupling of the Higgs boson to the t-quark through the analysis of angular correlations between jets. The analysis uses the decays of the Higgs boson to a pair of τ leptons, which in turn may decay into one of the following two final states: $\tau_\mu\tau_h$ and $\tau_h\tau_h$[1]. The 2016, 2017 and 2018 datasets collected at the CMS experiment at the $\sqrt{s} = 13\,\mathrm{GeV}$ are used in this measurement, which corresponds to a total integrated luminosity of 137 fb^{-1}.

7.2 Analysis strategy

This analysis requires the reconstruction of the major decay modes of the τ-lepton outlined in Table 4.1. In principle any hadronic and leptonic decay of the τ-lepton can be used. There are different experimental methods available to reconstruct the $\mathcal{CP}$-sensitive angle, most of which rely on the geometry of the τ-lepton decay in $\mathrm{H} \to \tau^+\tau^-$ events. It is important to establish the different methods first before moving onto the steps taken to extract the best result.

In the $\tau_h\tau_h$ channel, all combinations of hadronic decays of the τ_h-lepton can be considered. From theory, we expect the $\pi^\pm\pi^\pm$ final state to provide the best separation between different Higgs boson $\mathcal{CP}$ states due to the properties of the spectral function of the direct $\tau^\pm \to \pi^\pm\nu_\tau$

[1] The subscript describes the particle into which the associated τ-lepton decays.

decay mode [91]. The hadronic decays that are considered in this analysis are the following:

$$\begin{aligned}
&\tau^{\pm} \to \pi^{\pm}\nu_{\tau} \\
&\tau^{\pm} \to \rho^{\pm}\nu_{\tau} \to \pi^{\pm}\pi^{0}\nu_{\tau}\nu_{\tau} \\
&\tau^{\pm} \to a_1^{1\mathrm{pr}\,\pm}\nu_{\tau} \to \pi^{\pm}\pi^{0}\pi^{0}\nu_{\tau}\nu_{\tau} \\
&\tau^{\pm} \to a_1^{3\mathrm{pr}\,\pm}\nu_{\tau} \to \pi^{\pm}\pi^{\mp}\pi^{\pm}\nu_{\tau}\nu_{\tau}\,.
\end{aligned}$$

Therefore, the following combinations of τ-lepton decays are used in this analysis: $\{\mu^{\pm}, \pi^{\pm}, \rho^{\pm}, a_1^{1\mathrm{pr}\,\pm}, a_1^{3\mathrm{pr}\,\pm}\} \times \{\pi^{\pm}, \rho^{\pm}, a_1^{1\mathrm{pr}\,\pm}, a_1^{3\mathrm{pr}\,\pm}\}$. The methods presented in Chapter 2 are used for the appropriate pairs of decay modes. The impact parameter method is employed for τ-lepton decays to a μ lepton or $\pi^{\pm}$ meson. For decays involving a $\rho^{\pm}$ or $a_1^{\pm}$ intermediate resonance, the neutral pion method is utilised. For $a_1^{1\mathrm{pr}\,\pm}$ decays, the neutral pion four-momenta are summed together and the $\pi^{\pm}(\pi^0\pi^0)$ system is effectively treated like a ρ decay. In the case of $a_1^{3\mathrm{pr}\,\pm}$ decays, an ambiguity exists, as two charged pions with the same electric charge are present. The strategy chosen for this analysis is to select the oppositely charged pion pair from the $a_1^{3\mathrm{pr}\,\pm}$ daughters that exhibits an invariant mass closest to the ρ^0 mass. The charged pion with opposite charge to the original τ-lepton is treated like a π^0 in the calculation of the decay plane, and the remaining charged pion with same charge as the τ-lepton is used to determine the zero-momentum frame. The mixed approach is used for cases when the impact parameter is used on the muon or $\pi^{\pm}$ meson and the neutral pion method is on the meson originating from the other τ-lepton decay.

7.3 Signal modelling

The signal samples are modelled at NLO precision with POWHEG. The Higgs boson is configured to be produced solely as a scalar Higgs boson with mass of 125 GeV. Full phase space calculations performed with HRES in the POWHEG simulation tune the p_{T} distribution of the Higgs boson to improve the modelling. The decay of the Higgs boson to a τ-lepton pair is then described using the PYTHIA8.2 generator. The τ-lepton spin correlations are not accounted for during this simulation. Event weights that can be used to reweight the signal samples to any Higgs boson $\mathcal{CP}$ state are subsequently generated using the TAUSPINNER package. These weights can be used to obtain distributions for any $\mathcal{CP}$ state. The τ-lepton polarisation effects are chosen to be modelled for the purely scalar, purely pseudo-scalar, and maximally-mixed case. These correspond to a $\mathcal{CP}$-mixing angle of 0, 90° and 45°, respectively. The signal samples for the 2016 data-taking conditions are generated with the NNPDF3.0 NLO parton distributions, whilst the 2017 and 2018 samples are produced with the NNPDF3.1 configurations. Additionally, a generator-level selection is applied to improve the statistics at reconstruction-level. This is due to the relatively high trigger requirements, which many events fail to pass at reconstruction-level. The selection is performed on the visible p_T and $|\eta|$ of the final state leptons in each channel. For the τ_μ lepton the cut is set at 10% lower than the analysis-level cuts, whilst for the τ_h the generator-level cuts are 20% lower than the analysis selection.

The discriminating variable that provides differentiation between $\mathcal{CP}$ states of the Higgs boson using the τ-lepton decays is the $\phi_{\mathcal{CP}}$ spectrum defined in Chapter 2. Examples of generator-level distributions for the $\tau_h\tau_h$ and $\tau_\mu\tau_h$ channels are illustrated in Figure 7.1. These are produced using the ggH, VBF, WH and ZH samples generated with TauSpinner and reweighted to the corresponding $\mathcal{CP}$ scenario.

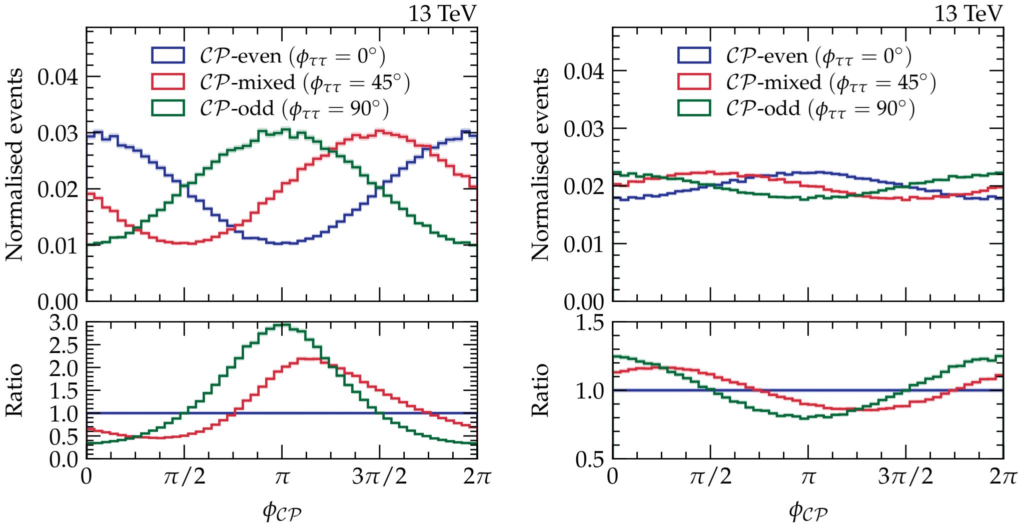

Figure 7.1: Generator-level distributions of $\phi_{\mathcal{CP}}$ using the set of signal samples produced for this analysis. The left and right distributions are for the $\rho\rho$ and $\mu\rho$ final states, respectively. The $\rho\rho$ final state clearly exhibits a large differentiation between the $\mathcal{CP}$ states, as is illustrated by the ratio of the different states with respect to the $\mathcal{CP}$-even hypothesis.

7.4 Event reconstruction

The event reconstruction methods have been outlined in Chapter 4. Since this analysis requires precise determination of impact parameter vectors, the primary vertex and impact parameter reconstruction has been modified from the nominal techniques used in CMS. Additionally, the improved BDT-based decay mode reconstruction is used as it performs better as indicated in Section 4.5.

Including all tracks of the charged decay products of the τ lepton may lead to an incorrect primary vertex position estimation due to the τ lepton's finite lifetime. Therefore, the tracks associated to the decay products are removed from the track collection created by the deterministic annealing algorithm, first introduced in Chapter 4. The adaptive vertex fitter is then utilised to refit the vertex with the modified collection of tracks, taking the beam spot position as an additional constraint to the fit. The resolution of the primary vertex is improved in the transverse plane, as illustrated in Figure 7.2. This is beneficial to this analysis, as a better vertex resolution will translate into a more precise determination of the impact parameter.

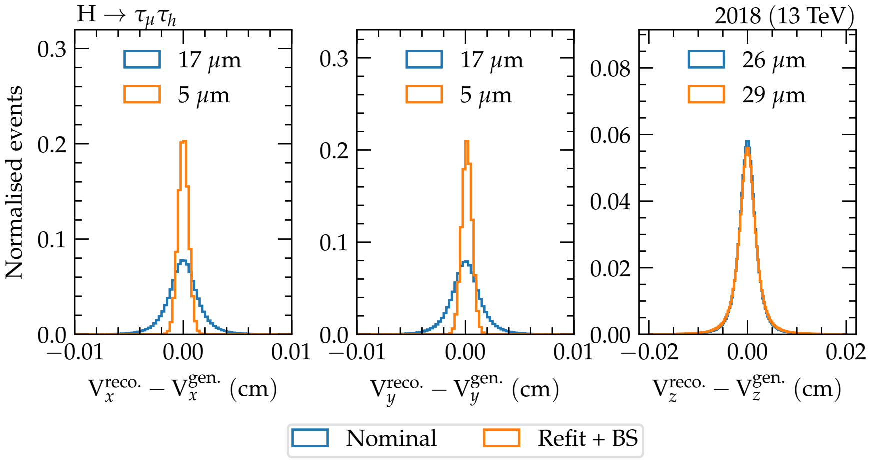

Figure 7.2: Vertex resolution for the nominal primary vertex in CMS and the refitted vertex with the beam spot constraint (BS) for the three Cartesian coordinates. The improvement is significant in the transverse plane, whilst not affecting the z-coordinate resolution.

Regarding the impact parameter reconstruction, an alternative approach to that employed in most other CMS analyses is used. The CMS impact parameter finding algorithm minimises only in the transverse plane, meaning it is two-dimensional, and no propagation of uncertainties is provided. However, the full three-dimensional impact parameter is desired. Therefore the particle's trajectory, $\vec{x}(t)$, is parameterised as a helix, and the point on the track, $\vec{x}(t'_{\text{min}})$, at which the distance to the (refitted) primary vertex $\vec{\text{V}}$ is minimised is determined. The impact parameter vector is then determined by $\vec{n} = \vec{x}(t'_{\text{min}}) - \vec{\text{V}}$. In addition, the uncertainties on $\vec{x}(t)$ and $|\vec{n}|$ are obtained analytically, such that the impact parameter significance can be defined as $S_{\text{IP}} = \frac{|\vec{n}|}{\sigma_{|\vec{n}|}}$. This significance can be used in the event selection to veto events with poorly reconstructed impact parameters.

The development of the BDT-based decay mode reconstruction for τ_h-lepton decays was motivated by the observation illustrated in Figure 7.3. These distributions show the $\phi_{\mathcal{CP}}$ angle for the $\tau_h\tau_h \to \rho\rho$ channel, where the ρ meson candidate is selected using the HPS τ_h-lepton decay mode finding. It can be seen that the presence of non-ρ mesons dilutes the amplitude of the distribution for all $\mathcal{CP}$ states. Therefore, the BDT-based decay mode is used to select each hadronic decay channel of the τ_h lepton, on top of the standard HPS τ-lepton decay mode finding.

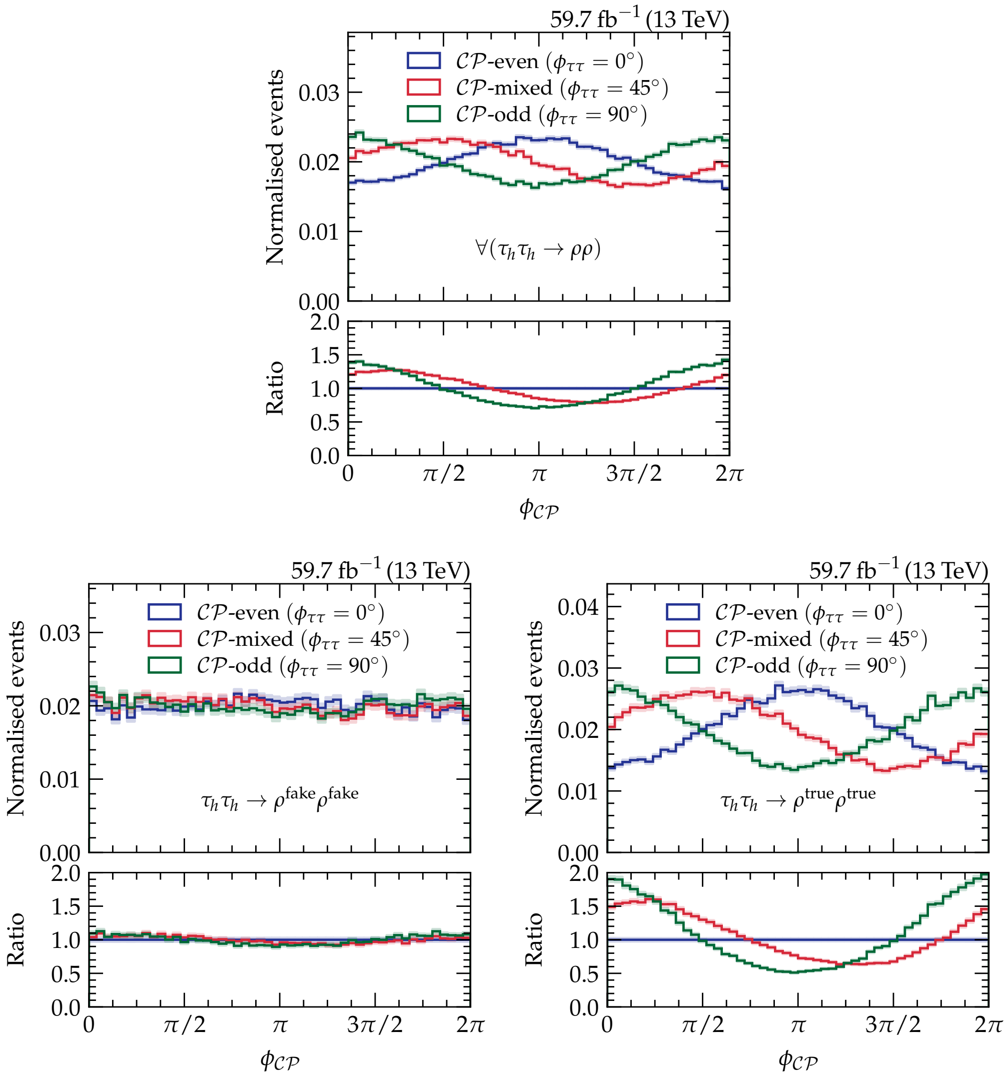

Figure 7.3: Dilution of $\mathcal{CP}$ discrimination power due to decay mode impurities using the HPS decay mode finding. The analysis-level $\phi_{\mathcal{CP}}$ distribution is shown for the $\tau_h\tau_h \to \rho\rho$ decay mode for three different selection criteria on the τ_h-lepton HPS decay modes. The top sub-figure is inclusive in ρ-meson candidates, whereas in the bottom sub-figures MC truth information is used. An improvement to the decay mode reconstruction can therefore provide larger separation between the $\mathcal{CP}$ states. To this end, the more powerful BDT-based decay mode reconstruction introduced in Section 4.5 is used in this analysis.

7.5 Event selection

As was the case for the previously discussed analysis, the selection is performed on top of the one introduced in Chapter 5. Many items are in identical to the $gg \to \mathrm{H} + 2\,\mathrm{jets}$ $\mathcal{CP}$ analysis criteria, and are summarised in the Table 7.1.

Table 7.1: The online p_T thresholds for the single muon and cross trigger for the $\tau_\mu\tau_h$ channel, and di-tau trigger for the $\tau_h\tau_h$ channel for each data-taking era considered in this analysis. The thresholds applied on the η and p_T of the trigger matched offline objects are also highlighted. The τ_h-lepton ID selection is performed with respect to the three output nodes of the *deepTauID* classifier: *tight* discrimination versus jets, *very-very-loose* discrimination versus electrons, and *tight* discrimination versus muons.

	Year	Trigger requirement	Offline lepton selection		
		$p_\mathrm{T}^\mathrm{raw}$ (GeV)	p_T (GeV)	$\lvert\eta\rvert$	Isolation
$\tau_h\tau_h$	All	$\tau_h(35)$ & $\tau_h(35)$	$p_\mathrm{T}^{\tau_h} > 40$	$\lvert\eta^{\tau_h}\rvert < 2.1$	τ_h ID
$\tau_\mu\tau_h$	2016	$\tau_\mu(22)$, $\tau_\mu(19)$ & $\tau_h(20)$	$p_\mathrm{T}^{\tau_\mu} > 20$	$\lvert\eta^{\tau_\mu}\rvert < 2.1$	$I^{\tau_\mu} < 0.15$
			$p_\mathrm{T}^{\tau_h} > 25$	$\lvert\eta^{\tau_h}\rvert < 2.3$	τ_h ID
	2017, 2018	$\tau_\mu(24)$, $\tau_\mu(20)$ & $\tau_h(27)$	$p_\mathrm{T}^{\tau_\mu} > 21$	$\lvert\eta^{\tau_\mu}\rvert < 2.1$	$I^{\tau_\mu} < 0.15$
			$p_\mathrm{T}^{\tau_h} > 32$	$\lvert\eta^{\tau_h}\rvert < 2.3$	τ_h ID

Furthermore, a selection of $m_T < 50\,\mathrm{GeV}$ and a veto on events with b-jets is applied in the $\tau_\mu\tau_h$ channel to further reduce the background contamination. On the other hand, a selection on the mass of the visible final state leptons of $m_\mathrm{visible} > 40\,\mathrm{GeV}$ is also used. Finally, the lepton vetoes and pair selection discussed in Chapter 5 are used to keep the channels orthogonal and reduce the $\mathrm{Z}/\gamma^*(\to \ell\bar{\ell}) + \mathrm{jets}$ and diboson backgrounds.

In decay modes where the impact parameter is used to determine the $\phi_{\mathcal{CP}}$ distribution, which is always the case for the μ lepton or the $\pi^\pm$ meson, a selection on the impact parameter significance of $S_\mathrm{IP} > 1.5\sigma$ is applied, which is motivated by the reasonable description of observed data using simulated events (after calibration). This selection ensure that events with poorly reconstructed impact parameters are removed.

7.5.1 Background methods

The background methods used in this analysis follow very closely the ggH analysis. The main difference arises from the use of τ_h-lepton decay products, thus fake-factors need to be remeasured in bins of decay mode. In addition, symmetries in the background distributions can be exploited to smooth these distributions. This will help to reduce the impact of statistical fluctuations. The averaging of background templates will be discussed after outlining the categorisation.

7.6 Event categorisation

The categorisation scheme follows very closely the approach that was studied in the ggH analysis in Section 6.5.2: a multi-classification-based algorithm to distinguish and isolate signal from background events. Given the reasonable performance of the BDT-based discriminator in Section 6.5.2, a similar approach was followed in this case, where the main difference arises from the absence of the need to train specifically for the di-jet phase space of ggH events. Additionally, as the analysis can use a di-τ pair from any Higgs boson decay, the signal category used in the training need not consist of just ggH or VBF processes, but can also contain events with VH production. The $\tau_h\tau_h$ channel uses a BDT for the classification task, whereas the $\tau_\mu\tau_h$ channel uses a NN. The focus of this Section will be on the $\tau_h\tau_h$ channel, however, an analogous set-up is used for the training and application in the $\tau_\mu\tau_h$ channel.

The BDT used to classify events for this analysis uses the observables in Table 7.2. The variables have been selected to provide good discrimination between processes with $\tau_h\tau_h$ final states without introducing any bias toward a specific $\mathcal{CP}$ state. Figures 7.4 - 7.6 illustrate the distributions of all input features for the inclusive selection. The ratio of the observed data and expected yields from simulation, and the ratio of the three $\mathcal{CP}$ signals with respect to the SM signal are also depicted. The simulation models the distributions well, giving confidence to the background methods used. Furthermore, the lower ratio suggests no bias towards a specific $\mathcal{CP}$ state will be established when using these variables as training features for the event classification BDT.

The training classes used in the multi-class BDT follow closely what was discussed in Chapter 6 and consist of the following three-class set-up:

signal: ggH, VBF, and VH production mode samples. These are reweighted by their cross-sections and then merged into one class. The samples used are the TauSpinner samples reweighted to the SM scenario.

genuine: background processes involving two genuine τ leptons.

jet $\rightarrow$ τ_h-lepton *fakes*: background processes involving at least one jet $\rightarrow$ τ_h-lepton fake.

Additionally, in order to ensure the imbalance in training samples for the individual training inputs, class weights that normalise the effective event numbers are derived and used in the training, in addition to the full event weight derived for each type of simulation.

Table 7.2: Input features to the BDT for the signal versus background classification. The variables used in the $\tau_h\tau_h$ channels are shown.

Feature
p_{T} of leading τ_h
p_{T} of visible di-τ
p_{T} of di-τ_h + $\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$
Visible di-τ mass
SVFit di-τ mass (using SVFit)
Leading jet p_{T}
Jet multiplicity
Invariant mass of leading jet pair
$\lvert\eta\rvert$ of leading jet pair
$p_{\mathrm{T}}^{\mathrm{miss}}$

As in the previous discussion on BDT categorisation methods, the training dataset is split into even and odd event numbers, which in turn are split into training and validation datasets. This type of cross-validation allows separate BDTs to be trained on the even and odd fold's training subset and can then be tested by applying the BDT trained on the even fold onto the odd event numbers and vice versa. This ensures that the full available statistics is used for training, while no bias is introduced. Finally, the same selection is applied to the training set as described in Chapter 5.

The hyperparameters of the BDT have been tuned using iterations of a Bayesian optimisation procedure [92]. Given a set of priors, this method explores the phase space of possible choices of hyperparameters, and finds the extremum to return a set of optimal parameters. The learning rate and maximum tree depth, which are the two hyperparameters with greatest impact on the performance, are thus set to 0.06 and 4, respectively. Lasso (L1) and ridge (L2) regularisation are also introduced to reduce the risk of over-training. Furthermore, the training terminates once the loss function of the validation set does not decrease over 50 consecutive boosting iterations.

Tests have been performed to show that there is negligible difference between the expected sensitivities obtained by training on the PS-reweighted sample instead of the SM-reweighted sample. This gives confidence that the training is independent of which CP state is trained on, and that the input variables do not bias the results.

The output of the BDT, when applied on the set of simulated MC samples, embedded events, and observed data events, is a list of scores per event, which can be interpreted as a probability of the particular event belonging to a certain class of event defined in the training. The

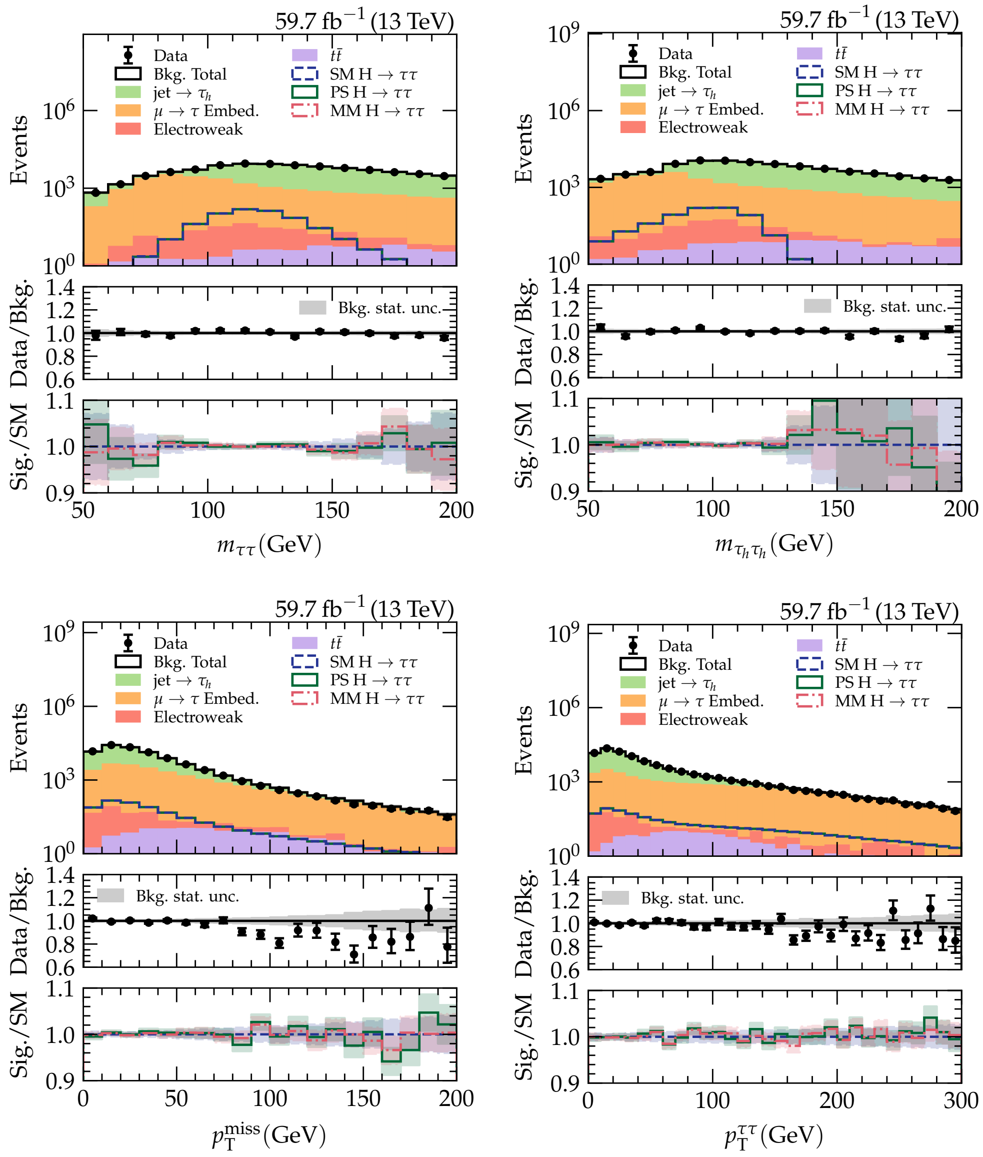

Figure 7.4: Distributions of the input features used to train the event classification BDT for the $\tau_h\tau_h$ channel. The signal is depicted for the $\mathcal{CP}$ states: standard model (SM), pseudo-scalar (PS) and maximal-mixing (MM) Higgs boson. The middle inset represents the ratio of observed data to the expected yield from simulation, whereas the lower inset shows the ratio of each signal with respect to the SM signal template. The uncertainty bands represent only statistical uncertainty.

training classes are the *signal*, *jet-fakes*, and *embedded*. Therefore, for each event, three scores are produced, whose sum is 1. The raw output scores for the 2018 training are illustrated in

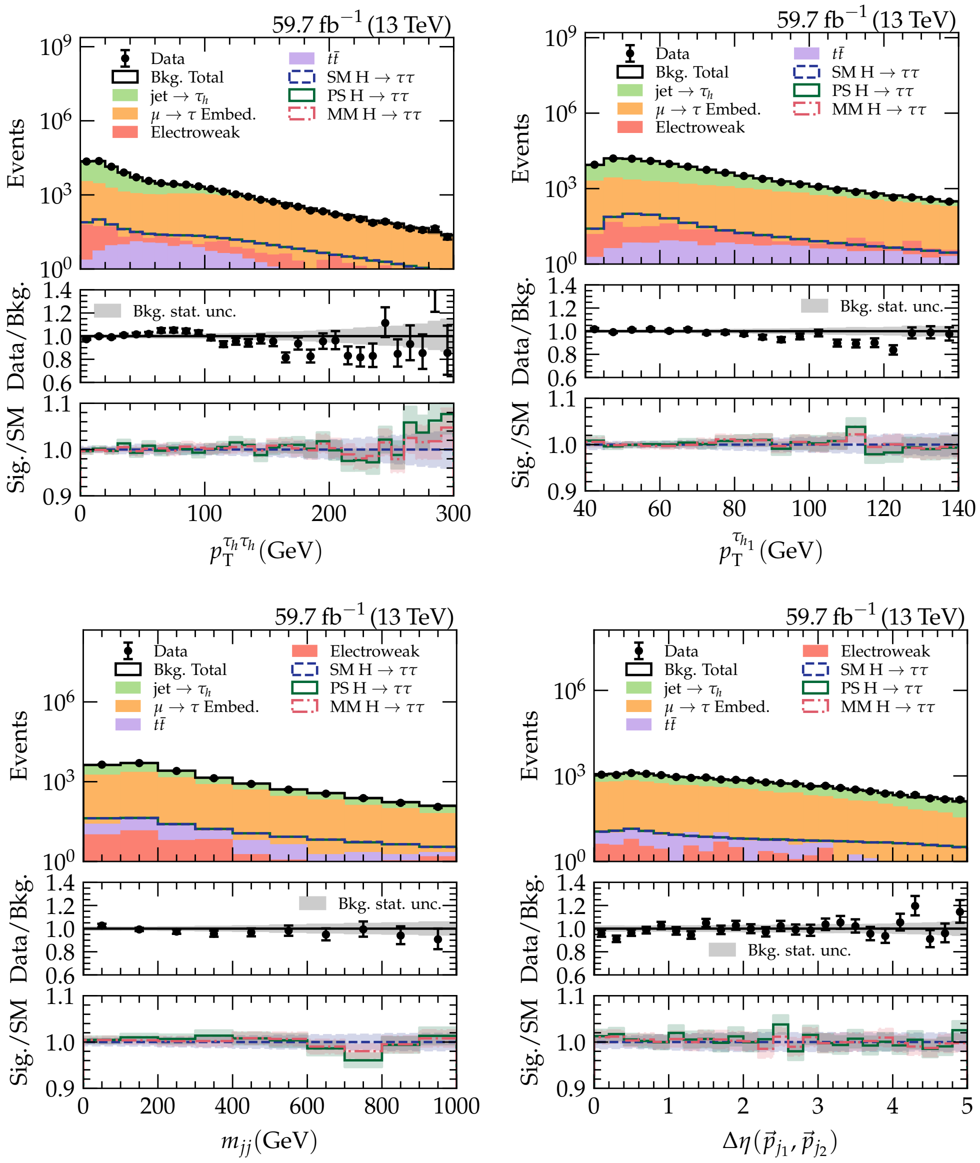

Figure 7.5: Distributions of the input features used to train the event classification BDT for the $\tau_h\tau_h$ channel. The signal is depicted for the $\mathcal{CP}$ states: standard model (SM), pseudo-scalar (PS) and maximal-mixing (MM) Higgs boson. The middle inset represents the ratio of observed data to the expected yield from simulation, whereas the lower inset shows the ratio of each signal with respect to the SM signal template. The uncertainty bands represent only statistical uncertainty.

Figure 7.7. In this Figure, the true process label is used to compare the outcome of the BDT for each class. The desired process distribution peaks close to a score of one in each of the

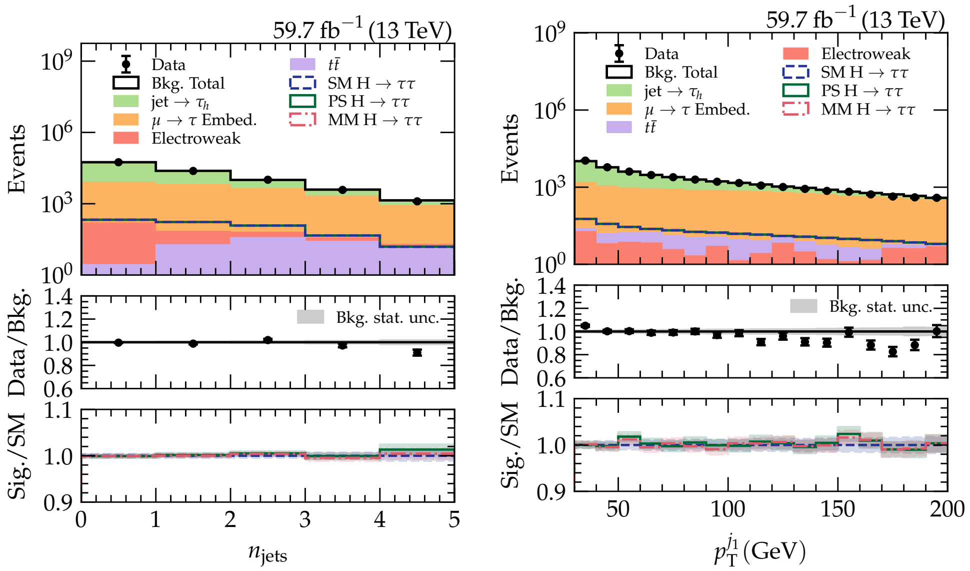

Figure 7.6: Distributions of the input features used to train the event classification BDT for the $\tau_h\tau_h$ channel. The signal is depicted for the $\mathcal{CP}$ states: standard model (SM), pseudo-scalar (PS) and maximal-mixing (MM) Higgs boson. The middle inset represents the ratio of observed data to the expected yield from simulation, whereas the lower inset shows the ratio of each signal with respect to the SM signal template. The uncertainty bands represent only statistical uncertainty.

scores, thus the results show good discrimination of the individual processes. For the purpose of categorising the events using these scores, the maximum score per event is identified and the event is assigned to the corresponding category. This means that an assigned event obtains a lowest possible score of 0.33, which is highlighted with grey lines in Figure 7.7. Using this kind of categorisation, no events are removed by a hard cut, rather the events are re-ordered into process-enriched categories. The distribution of the maximum score, or simply BDT score, is shown in Figure 7.7.

7.7 Bin averaging of signal and background distributions

As introduced in Section 6.7, averaging the signal and background distributions, whenever possible, can help to minimise the effect of statistical fluctuations. Additionally, it can minimise any effects that bias the measurement. In this analysis, backgrounds are known to be either flat in $\phi_{\mathcal{CP}}$ or symmetric about $\phi_{\mathcal{CP}} = \pi$. For instance, backgrounds with two genuine τ-leptons, such as $Z/\gamma^* \to \tau^+\tau^-$ events, are flat at particle level [93]. No experimental effects that bias or smear the distribution from its flat shape have been observed in cases where the neutral

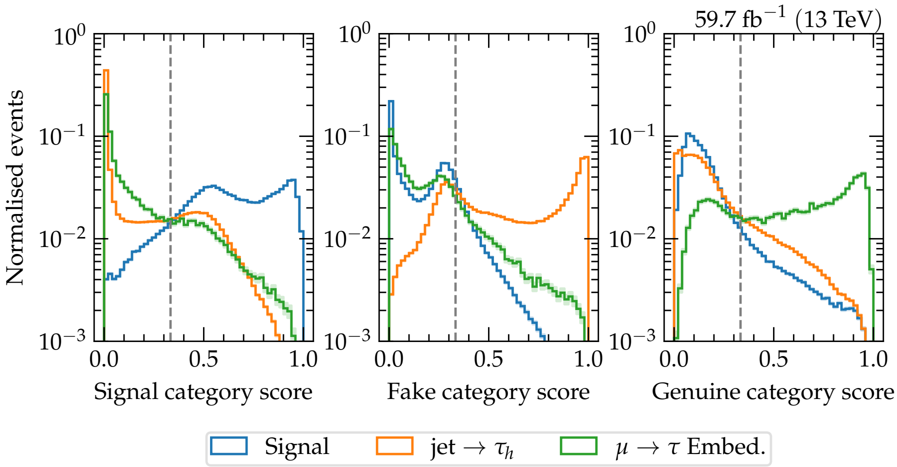

Figure 7.7: Raw BDT score distributions for each class of the multi-classifier trained on 2018 simulated and data events. For each of these distributions the true process label is used to determine the purity of each class in the individual scores. For a well performing algorithm the score distributions for its associated true class would peak at 1, with the background left behind close to 0. This is the case here, thus the BDT performs as expected.

pion method is applied. Therefore, events from these processes are flattened by merging the bins in $\phi_{\mathcal{CP}}$. For decay modes where the impact parameter method is applied on both decay products of the τ-lepton, such as the $\mu\pi$ and $\pi\pi$ channels, correlated smearing effects of the primary vertex lead to a shape effect in the $\phi_{\mathcal{CP}}$ distribution [94]. In these cases, the background distributions are symmetrised about $\phi_{\mathcal{CP}} = \pi$, which is, nonetheless, an observed symmetry of the distributions. Finally, kinematic effects lead to non-flat shapes for the jet → τ_h-lepton fake background in all decay mode categories. The distributions are symmetrised about $\phi_{\mathcal{CP}} = \pi$, as this symmetry is still valid. The symmetrisation procedure is also applied to the $\mathcal{CP}$-even and $\mathcal{CP}$-odd signal distributions, whilst anti-symmetrisation is used for the $\mathcal{CP}$-mixed signal, as shown in Section 6.7.

The effect of applying these symmetrisation procedures can be checked by producing plots of the $\phi_{\mathcal{CP}}$ distributions in the background categories (defined by the genuine and fake MVA scores) and splitting these into the different decay modes. In order to replicate the binning used in the final fit, the $\phi_{\mathcal{CP}}$ bins are also split into increasing windows of BDT or NN scores. These categories are not used in any fit, or in the final fit to determine the $\mathcal{CP}$-mixing angle, however they are instructive to look for deviations between simulation and data. To provide examples, Figures 7.8 and 7.9 illustrate these distributions for the genuine and fake categories, split into τ-lepton decay modes. No significant deviations have been observed, which gives more confidence in the validity of this symmetrisation method. Furthermore, a set of goodness-of-fit and toy studies was performed, which further confirmed no bias is introduced in the measurement of $\phi_{\tau\tau}$.

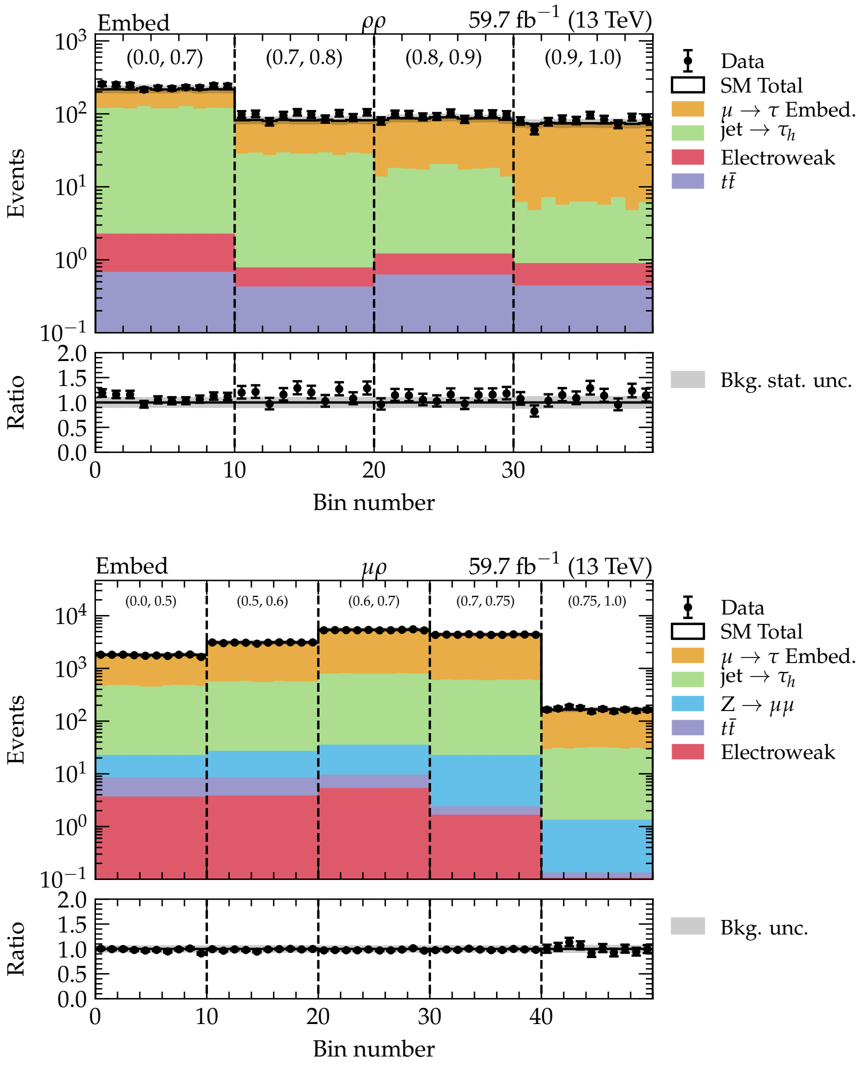

Figure 7.8: Distributions of $\phi_{\mathcal{CP}}$ in windows of MVA score for two decay modes in the genuine τ-lepton category (dominated by $Z/\gamma^* \rightarrow \tau^+\tau^-$). The distributions are not fitted, and the full uncertainty model for background events is included in the grey band.

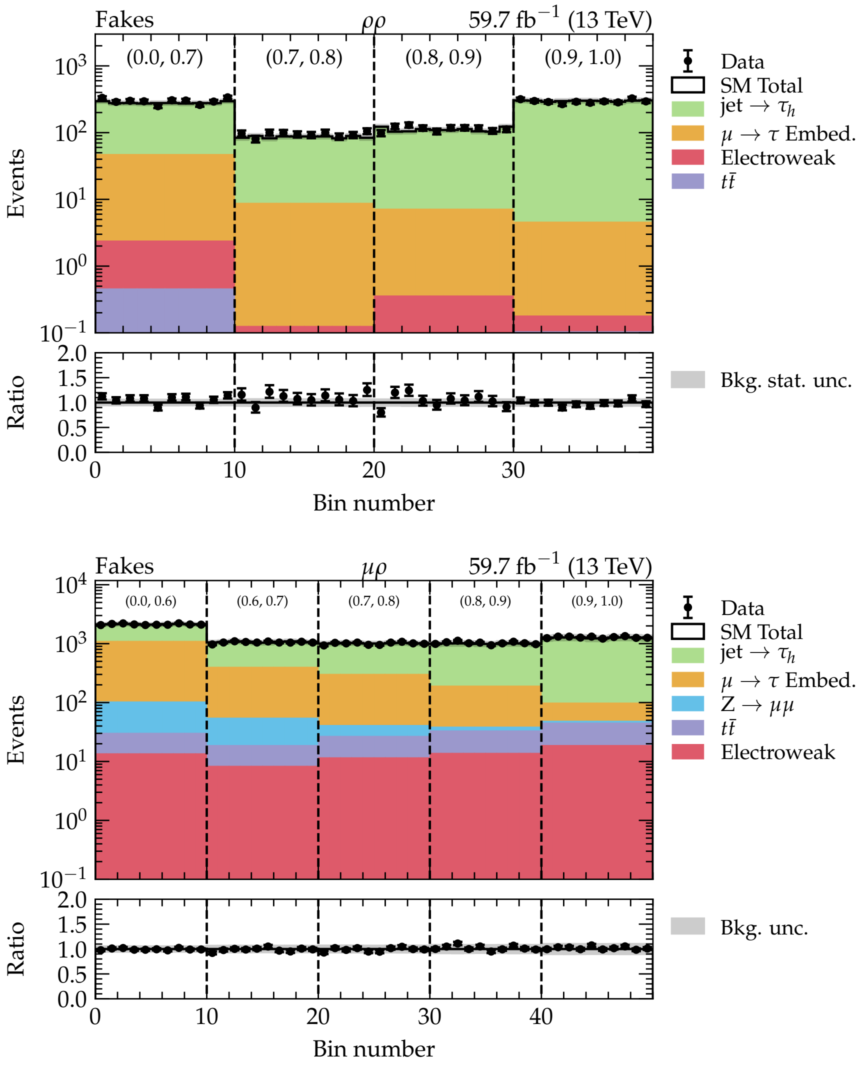

Figure 7.9: Distributions of $\phi_{\mathcal{CP}}$ in windows of MVA score for two decay modes in the jet → τ_h-lepton fake category. The distributions are not fitted, and the full uncertainty model for background events is included in the grey band.

7.8 Corrections

The set of corrections applied to simulation to improve the modelling of data in this analysis is summarised in Table 7.3. The majority of corrections are in common with the ones introduced in Section 6.6, however, it is worth highlighting here the corrections that are specific for this analysis: the τ_h-lepton ID, τ_h-lepton trigger and the impact parameter significance. These were specifically measured for this analysis as there exists a dependency on the decay mode of the

τ_h-lepton. The impact parameter significance is calibrated since a selection is performed on it to remove events with poorly modelled impact parameters, in cases when the impact parameter method is employed. The calibrated impact parameter significance for the $\tau_h^+\tau_h^- \to \pi^+\pi^-\nu_\tau\bar{\nu}_\tau$ final state is illustrated in Figure 7.10.

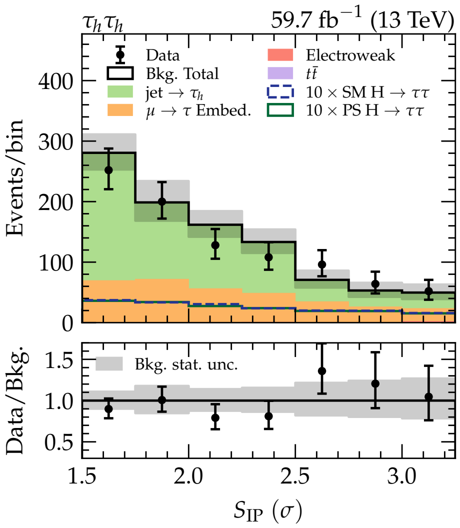

Figure 7.10: Pre-fit distribution of the calibrated impact parameter significance for events in the $\tau_h^+\tau_h^- \to \pi^+\pi^-\nu_\tau\bar{\nu}_\tau$ final state. The significance is in units of standard deviations. Events with $S_{\mathrm{IP}} < 1.5\sigma$ are discarded as these impact parameters are poorly modelled. The grey shaded area represents the statistical uncertainty only.

Table 7.3: Corrections applied to simulated events to ensure correct modelling of the observed events.

Correction	Simulation type
τ_h-lepton ID and trigger	MC + embedded
$l \to \tau_h$ fake rate	MC
τ_h-lepton energy scale	MC (genuine τ_h and $l \to \tau_h$ fakes)
e smear and scale corrections	MC + embedded
$\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ recoil	MC
b-tagged jet efficiency	MC
Z p_{T}/mass reweighting	MC (Z, W, H).
t-quark p_{T} reweighting	MC ($t\bar{t}$)
NNLOPS $p_{\mathrm{T}}/n_{\mathrm{jets}}$ reweighting	MC (ggH)
Impact parameter significance S_{IP}	MC + embedded

7.9 Systematic uncertainties

Similarly to the list of uncertainties introduced in Section 6.8, the set of systematic uncertainties takes the form of two types of uncertainties: normalisation and shape uncertainties. The list of normalisation uncertainties is identical to the uncertainties discussed in Section 6.8.1. The majority of shape uncertainties have also been covered already, the main difference being the binning in BDT-based decay mode in this analysis. This gives rise to slightly varying uncertainties on the τ_h-lepton identification and trigger efficiency, and the impact parameter significance, which was not used in the previously discussed analysis. For the τ_h-lepton identification and trigger efficiency uncertainties, these were specifically measured in bins of BDT-based decay mode as this improved the modelling of simulated events. The impact parameter significance has a conservative correction of 25%, which covers variations up to an impact parameter significance of 4σ. As the impact parameter is only used for the $\tau \to \mu$ and $\tau \to \pi$ jets, the uncertainty on the impact parameter significance is only applied to simulation in cases where one of the τ leptons jets into a μ or π. Furthermore, the fake factors differ slightly due to binning in BDT-based decay mode and a selection on the impact parameter significance. Conceptually, however, the fake factors are measured in an identical manner to the ggH $\mathcal{CP}$ analysis. The full set of systematic uncertainties is summarised in Table 7.4, indicating their magnitude and whether there is assumed to be a correlation between the data-taking eras and corresponding datasets.

7.10 Statistical inference

The extraction on the effective $\mathcal{CP}$-mixing angle $\phi_{\tau\tau}$ is performed in an analogous way as discussed in Section 6.9 using a simultaneous template fit, where the best-fit values are optimised when the likelihood function $\mathcal{L}\left(\vec{\sigma}, \phi_{\tau\tau}, \vec{\theta} \mid \text{data}\right)$ is maximised. Again, $\vec{\sigma} = \left(\sigma_{gg\text{H}}, \sigma_{\text{VBF}}, \sigma_{\text{WH}}, \sigma_{\text{ZH}}\right)$ is the Higgs boson production cross-section of each process considered in the analysis, $\phi_{\tau\tau}$ is the $\mathcal{CP}$-mixing angle in the τ-lepton Yukawa coupling, and $\vec{\theta}$ is the set of nuisance parameters in the fit. The full likelihood function is then written as in Equation 6.3, replacing α_{gg} with $\phi_{\tau\tau}$. These rate parameter modifiers are written as $\vec{\mu}^{\tau\tau} = (\mu_{gg\text{H}}, \mu_{\text{VH}})$, where the $\mu_{gg\text{H}}$ scales the ggH cross-section and μ_{VH} scales both the vector-boson-fusion and associated Higgs boson production cross-sections simultaneously. When scanning for the observed value of $\phi_{\tau\tau}$, the best-fit values of all nuisance parameters and unconstrained rate parameter modifiers are extracted as well.

As introduced in Equation 6.4, the negative log-likelihood is used to determine the best-fit value on the parameter of interest and its uncertainty at a given level of confidence. For a one-dimensional fit, the 68%, 95%, and 99.7% CL is determined when the $-2\Delta \log \mathcal{L}$ is 0.99,

Table 7.4: Sources of systematic uncertainties, where the correlation, if any, between the years is indicated in the third column.

Uncertainty	Magnitude	Correlation	PDF
τ_h ID	p_{T}/decay-mode dependent (2–3%)	no	Gaussian
Muon reconstruction	1%.	yes	log-normal
$e \to \tau_h$ ID	5(1)% 2016(2017,2018)	no	Gaussian
$\mu \to \tau_h$ ID	20–40%	no	Gaussian
μ ID	1%	yes	Gaussian
b-jet veto	1–9%	no	log-normal
Luminosity	2.3%–2.5%	partial	log-normal
Trigger	2% for μ, p_{T}-dep. for τ_h	no	Gaussian
Embedded yield	4%	no	log-normal
$t\bar{t}$ cross-section	4.2%	yes	log-normal
Diboson cross-section	5%	yes	log-normal
Single-t cross-section	5%	yes	log-normal
W + jets cross-section	4%	yes	log-normal
Drell-Yan cross-section	2%	yes	log-normal
Signal cross-sections	[19]	yes	log-normal
Parton shower	Signal-dependent	yes	Gaussian
Renormalisation scale	Signal-dependent	yes	log-normal
Factorisation scale	Signal-dependent	yes	log-normal
t-quark p_{T} reweighing	100%	yes	Gaussian
Z/γ^* p_{T} reweighing	100%	partial	Gaussian
Prefiring (2016, 2017)	Event-dependent (0–4%)	yes	log-normal
τ_h energy scale	1% (sim), 1.5% (emb.)	no	Gaussian
$\mu \to \tau_h$ energy scale	1%	no	log-normal
Muon energy scale	0.4–2.7%	yes	Gaussian
Jet energy scale	Event-dependent	partial	Gaussian
Jet energy resolution	Event-dependent	no	Gaussian
$\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ unclustered scale	Event-dependent	no	Gaussian
$\vec{p}_{\mathrm{T}}^{\mathrm{miss}}$ recoil corrections	Event-dependent	no	Gaussian
Jet$\to \tau_h$ mis-ID	FF-dependent	partial	Gaussian
$t\bar{t}$/diboson in embedded	10%	yes	Gaussian
S_{IP} in μ and π decays	25%	no	Gaussian

3.84, and 8.81, respectively. The 99.7% CL is introduced here as this analysis achieves an observed significance level of 3σ.

7.11 Results

The compatibility between data and the model is tested using goodness-of-fit tests, which estimate a statistic for each distribution used in the fit and determine a p-value. The fits suggest that the model is compatible with data. The distributions of the categories used in the fit are shown in Figures 7.11 - 7.25, where the parameters are set to the value estimated by the fit. The first set of figures shows the distributions of the background categories, which are the BDT and NN distributions for the genuine τ-lepton and jet-fake τ_h-lepton categories, respectively. The signal categories are the BDT/NN distributions of the *higgs* category, split into the different combinations of τ_h-lepton decay modes. The x-axis (labelled "Bin number") represents the binning in the cyclic discriminating variable ϕ_{CP}, ranging between 0 and 2π. The distributions are also split into windows of BDT/NN *higgs* output score, separated by the dashed lines, in order to improve the ratio of signal to background. Thus the sensitivity is largely driven by events in the rightmost BDT/NN window, where the BDT/NN selected events that are more to originate from H → $\tau^+\tau^-$ processes rather than background processes.

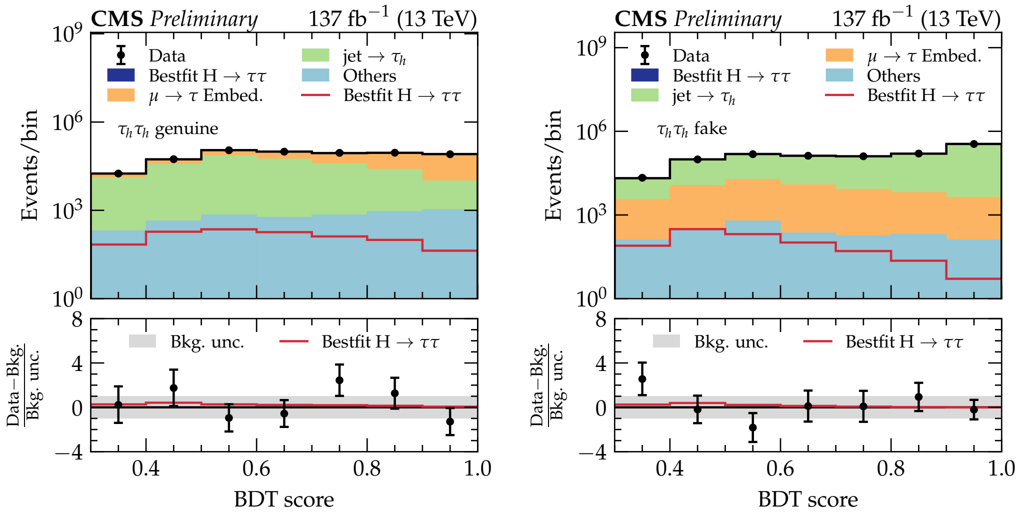

Figure 7.11: Distributions of genuine τ-lepton (left) and jet-fake τ_h-lepton (right) BDT scores for the $\tau_h\tau_h$ channel. The parameters are set to their best-fit value determined by the combined fit. The distributions are inclusive in τ_h-lepton decay mode. Figure taken from [1].

The scans of the expected contribution of each decay-mode category are displayed in Figure 7.26 and 7.27 for the $\tau_\mu\tau_h$ and $\tau_h\tau_h$ channels, respectively. The expected scans are produced with an Asimov dataset and indicate that the largest sensitivities arise from the $\mu\rho$ and $\mu\pi$ decay modes in $\tau_\mu\tau_h$ channel, and the $\rho\rho$ and $\pi\rho$ decay modes in the $\tau_h\tau_h$ channel.

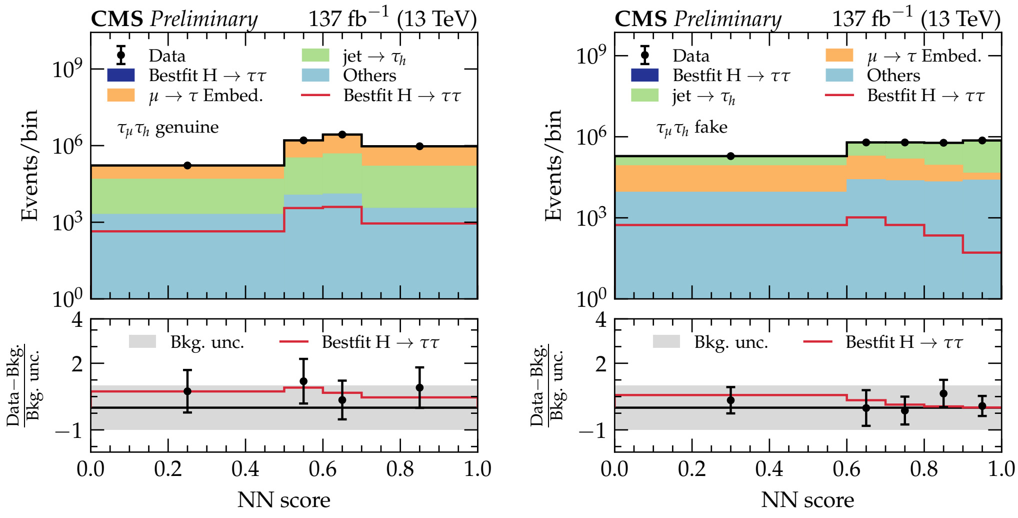

Figure 7.12: Distributions of genuine τ-lepton (left) and jet-fake τ_h-lepton (right) NN scores for the $\tau_\mu\tau_h$ channel. The parameters are set to their best-fit value determined by the combined fit. The distributions are inclusive in τ_h-lepton decay mode. Figure taken from [1].

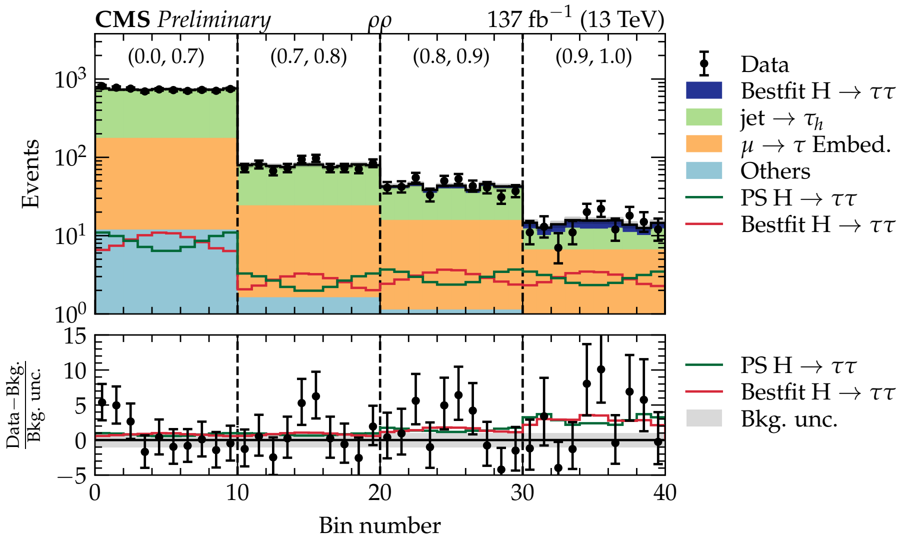

Figure 7.13: Distributions of $\phi_{\mathcal{CP}}$ in windows of BDT score for the $\rho\rho$ channel. Both the best-fit and pseudoscalar (PS) signals are shown over the background + best-fit signal stack. The x-axis labels represent the cyclic bins of $\phi_{\mathcal{CP}}$ between $(0, 2\pi)$). Figure taken from [1].

The scan of the negative log-likelihood for the fit extracting the best-fit value for $\phi_{\tau\tau}$ is shown in Figure 7.29. The measured value of $\phi_{\tau\tau}$ comes to $(4 \pm 17)^\circ$ at the 68% CL, which is well in agreement with the $\mathcal{CP}$-even hypothesis predicted by the SM. Separate fits were performed

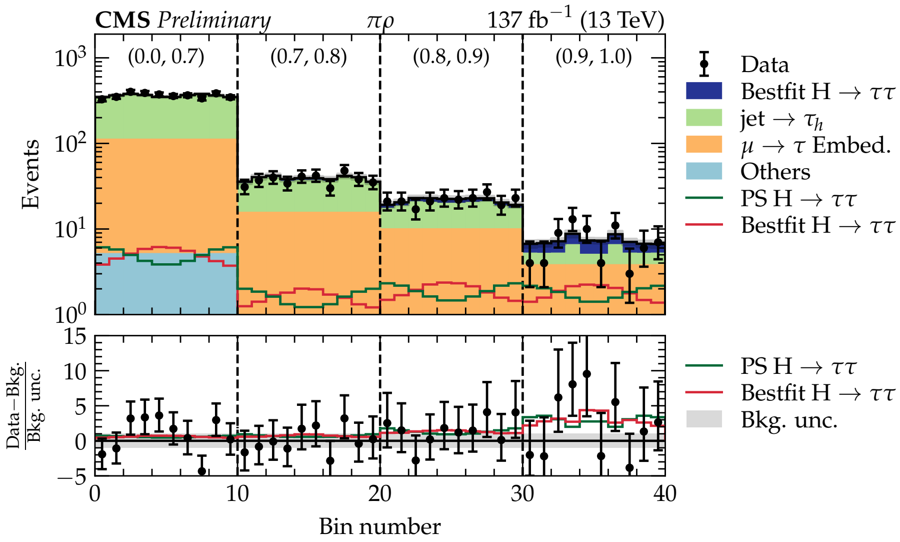

Figure 7.14: Distributions of $\phi_{\mathcal{CP}}$ in windows of BDT score for the $\pi\rho$ channel. Both the best-fit and pseudoscalar (PS) signals are shown over the background + best-fit signal stack. The x-axis labels represent the cyclic bins of $\phi_{\mathcal{CP}}$ between $(0, 2\pi)$). Figure taken from [1].

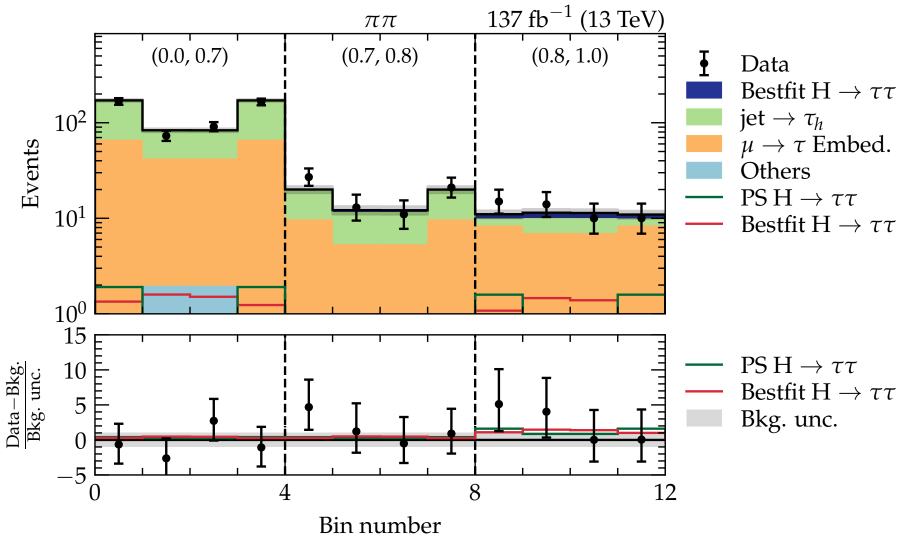

Figure 7.15: Distributions of $\phi_{\mathcal{CP}}$ in windows of BDT score for the $\pi\pi$ channel. Both the best-fit and pseudoscalar (PS) signals are shown over the background + best-fit signal stack. The x-axis labels represent the cyclic bins of $\phi_{\mathcal{CP}}$ between $(0, 2\pi)$).

where the parameter of interest was set to the ggH or VBF rate parameters. These provided observed values of $\mu^{\tau\tau}_{gg\text{H}} = 0.72 \pm 0.33$ and $\mu^{\tau\tau}_{\text{V}} = 1.02^{+0.55}_{-0.56}$. Additionally, a scan of the two-

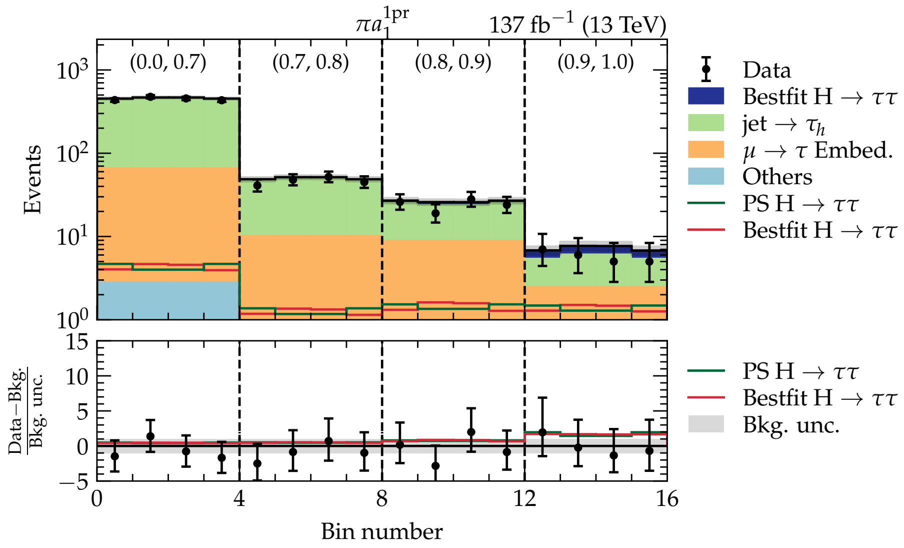

Figure 7.16: Distributions of $\phi_{\mathcal{CP}}$ in windows of BDT score for the $\pi a_1^{1\mathrm{pr}}$ channel. Both the best-fit and pseudoscalar (PS) signals are shown over the background + best-fit signal stack. The x-axis labels represent the cyclic bins of $\phi_{\mathcal{CP}}$ between $(0, 2\pi)$).

dimensional negative log-likelihood of the total branching fraction modifier $\mu^{\tau\tau}$ and $\phi_{\tau\tau}$ is performed, as illustrated in Figure 7.30. The branching fraction modifier is the ratio of the observed and SM-expected rate of H → ττ decays. The assumption made in this case is that Higgs boson couplings in production processes are set to the SM predicted values. The best-fit value for $\mu^{\tau\tau}$ is found to be 0.82 ± 0.15, which is well in agreement with the SM prediction as well. Finally, the two-dimensional scan for κ_τ against $\tilde{\kappa}_\tau$ is shown in Figure 7.31.

As for the previous analysis, the dominant sources of uncertainty in the fit to data are statistical. This is followed by the hadronic τ_h-lepton trigger efficiency and energy scale uncertainty, theory uncertainties, as is indicated in Figure 7.28. This analysis is more sensitive to the hadronic τ_h-lepton decay mode, which makes the impact of τ_h-lepton-related uncertainties on $\phi_{\tau\tau}$ more pronounced.

Finally, in Figure 7.32 the asymmetry-reweighted significance in terms of the background-subtracted data and $\mathcal{CP}$-even and $\mathcal{CP}$-odd predictions are illustrated for the three most sensitive final states used in this analysis. The reweighting is applied to provide a starker visual contrast, and supply the result with an intuitive explanation why the data favours the $\mathcal{CP}$-even model.

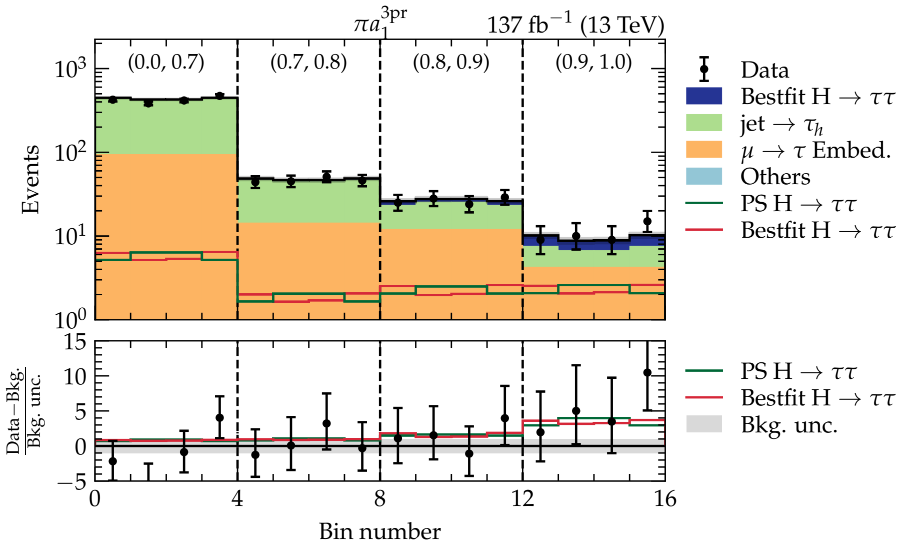

Figure 7.17: Distributions of $\phi_{\mathcal{CP}}$ in windows of BDT score for the $\pi a_1^{3\text{pr}}$ channel. Both the best-fit and pseudoscalar (PS) signals are shown over the background + best-fit signal stack. The x-axis labels represent the cyclic bins of $\phi_{\mathcal{CP}}$ between $(0, 2\pi)$).

7.12 Summary

The measurement of the $\mathcal{CP}$ state of the Higgs boson in τ-lepton decays provides a significant result in the field of Higgs physics and it is the first time that this analysis was presented. The result has been presented with data of pp collisions at $\sqrt{s} = 13\,\text{TeV}$ of 137 fb^{-1} collected at the CMS experiment. The channels considered in the analysis are the $\tau_h\tau_h$ and $\tau_\mu\tau_h$ final states, where either both the τ-lepton decays hadronically, or one hadronically and one into a muon. The effective mixing angle in the τ-lepton Yukawa coupling, $\phi_{\tau\tau}$, finds a best-fit value of $\phi_{\tau\tau} = (4 \pm 17)^\circ$ at 68% CL, which is highly consistent with the $\mathcal{CP}$-even prediction of $\phi_{\tau\tau}$ =0 provided by the SM. The pure $\mathcal{CP}$-odd (PS) prediction is rejected at 3.2σ, giving yet another strong case study for the validity of the SM. The measurements of the rate parameter modifiers results in observed values of $\mu^{\tau\tau}_{ggH} = 0.72 \pm 0.33$ and $\mu^{\tau\tau}_{V} = 1.02^{+0.55}_{-0.56}$, and $\mu^{\tau\tau} = 0.82 \pm 0.15$, which are in good agreement with respect to the SM.

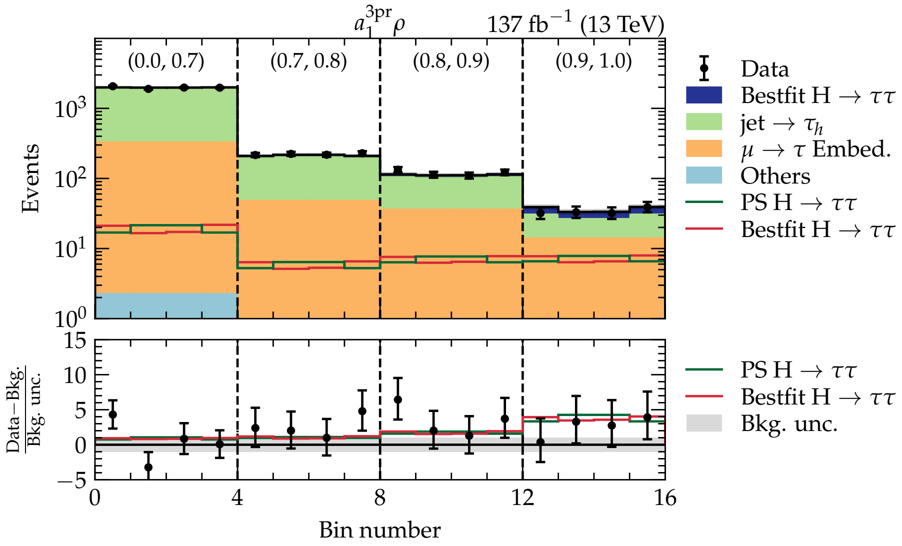

Figure 7.18: Distributions of $\phi_{\mathcal{CP}}$ in windows of BDT score for the $a_1^{3\mathrm{pr}}\rho$ channel. Both the best-fit and pseudoscalar (PS) signals are shown over the background + best-fit signal stack. The x-axis labels represent the cyclic bins of $\phi_{\mathcal{CP}}$ between $(0, 2\pi)$).

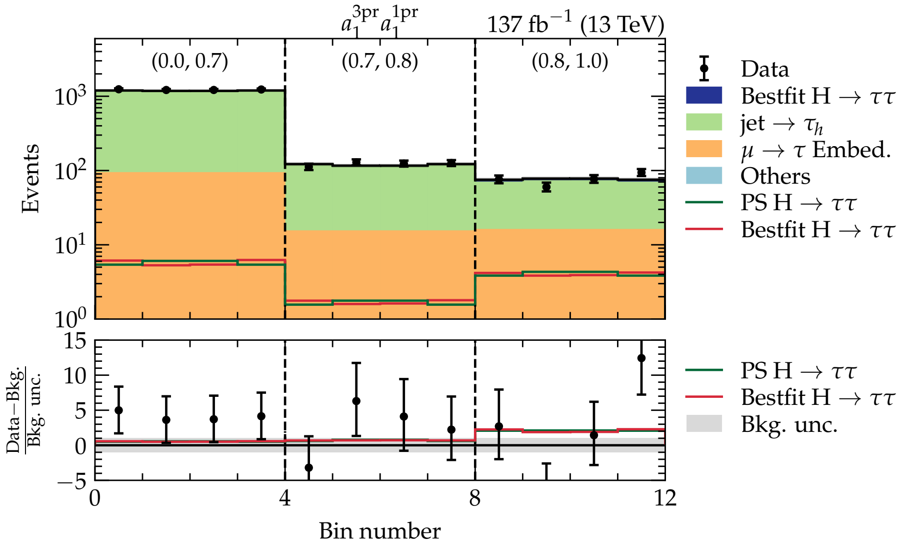

Figure 7.19: Distributions of $\phi_{\mathcal{CP}}$ in windows of BDT score for the $a_1^{3\mathrm{pr}}a_1^{1\mathrm{pr}}$ channel. Both the best-fit and pseudoscalar (PS) signals are shown over the background + best-fit signal stack. The x-axis labels represent the cyclic bins of $\phi_{\mathcal{CP}}$ between $(0, 2\pi)$).

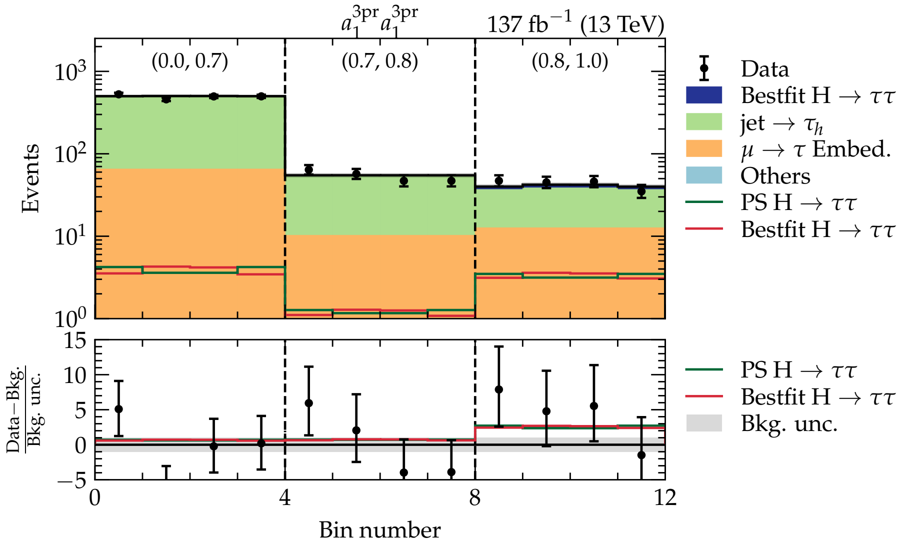

Figure 7.20: Distributions of $\phi_{\mathcal{CP}}$ in windows of BDT score for the $a_1^{3pr}a_1^{3pr}$ channel. Both the best-fit and pseudoscalar (PS) signals are shown over the background + best-fit signal stack. The x-axis labels represent the cyclic bins of $\phi_{\mathcal{CP}}$ between $(0, 2\pi)$).

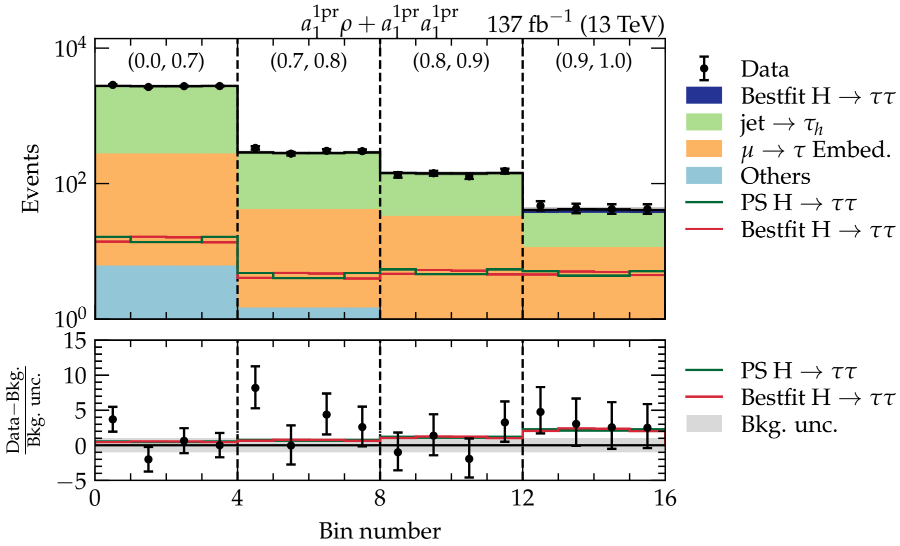

Figure 7.21: Distributions of $\phi_{\mathcal{CP}}$ in windows of BDT score for the $a_1^{1pr}\rho + a_1^{1pr}a_1^{1pr}$ channel. Both the best-fit and pseudoscalar (PS) signals are shown over the background + best-fit signal stack. The x-axis labels represent the cyclic bins of $\phi_{\mathcal{CP}}$ between $(0, 2\pi)$).

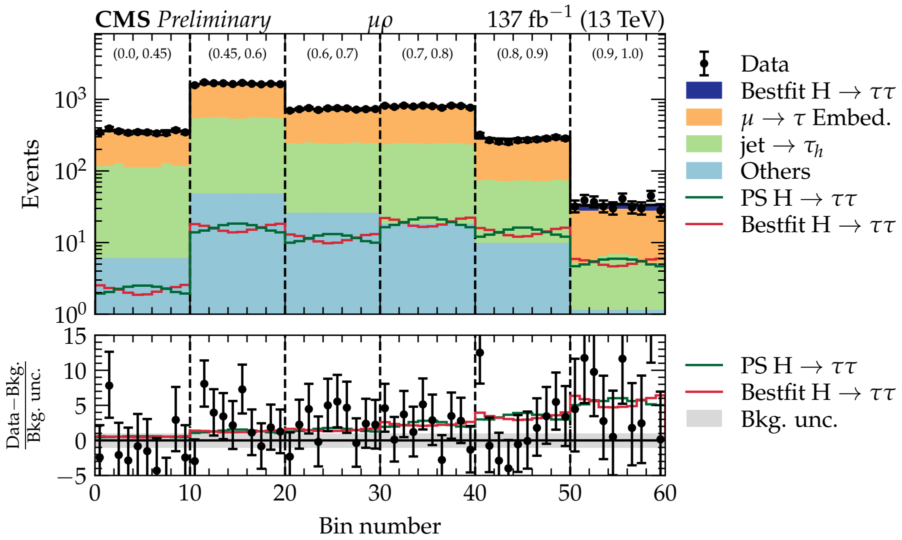

Figure 7.22: Distributions of $\phi_{\mathcal{CP}}$ in windows of NN score for the $\mu\rho$ channel. Both the best-fit and pseudoscalar (PS) signals are shown over the background + best-fit signal stack. The x-axis labels represent the cyclic bins of $\phi_{\mathcal{CP}}$ between $(0, 2\pi))$. Figure taken from [1].

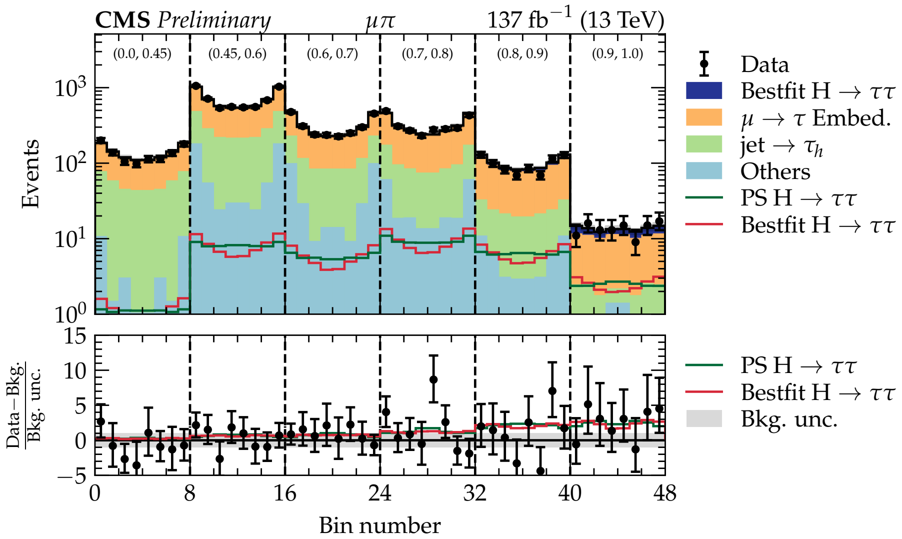

Figure 7.23: Distributions of $\phi_{\mathcal{CP}}$ in windows of NN score for the $\mu\pi$ channel. Both the best-fit and pseudoscalar (PS) signals are shown over the background + best-fit signal stack. The x-axis labels represent the cyclic bins of $\phi_{\mathcal{CP}}$ between $(0, 2\pi))$. Figure taken from [1].

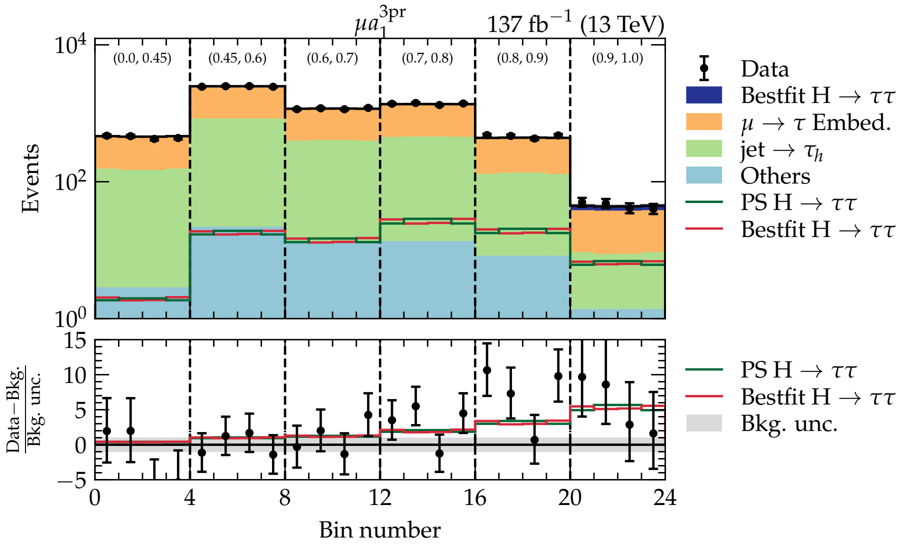

Figure 7.24: Distributions of $\phi_{\mathcal{CP}}$ in windows of NN score for the $\mu a_1^{3\mathrm{pr}}$ channel. Both the best-fit and pseudoscalar (PS) signals are shown over the background + best-fit signal stack. The x-axis labels represent the cyclic bins of $\phi_{\mathcal{CP}}$ between $(0, 2\pi)$).

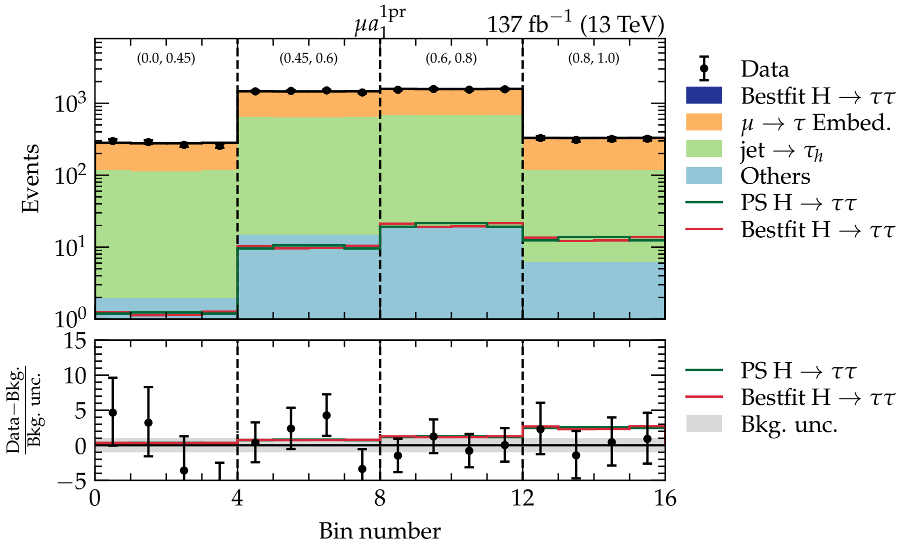

Figure 7.25: Distributions of $\phi_{\mathcal{CP}}$ in windows of NN score for the $\mu a_1^{1\mathrm{pr}}$ channel. Both the best-fit and pseudoscalar (PS) signals are shown over the background + best-fit signal stack. The x-axis labels represent the cyclic bins of $\phi_{\mathcal{CP}}$ between $(0, 2\pi)$).

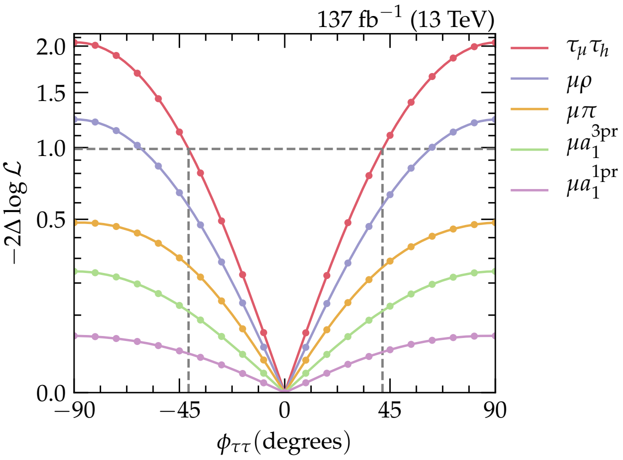

Figure 7.26: Scans of the expected negative log-likelihood $-2\Delta\log\mathcal{L}$ for parameter of interest $\phi_{\tau\tau}$ showing separately the contributions of the categories from the $\tau_\mu\tau_h$ channel.

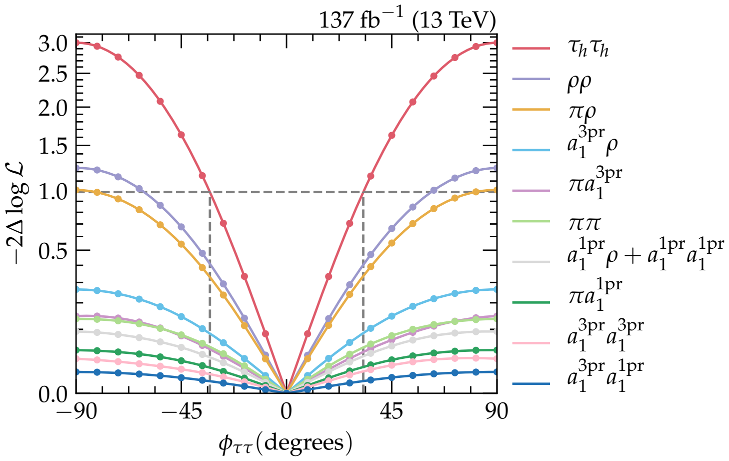

Figure 7.27: Scans of the expected negative log-likelihood $-2\Delta\log\mathcal{L}$ for parameter of interest $\phi_{\tau\tau}$ showing separately the contributions of the categories from the $\tau_h\tau_h$ channel.

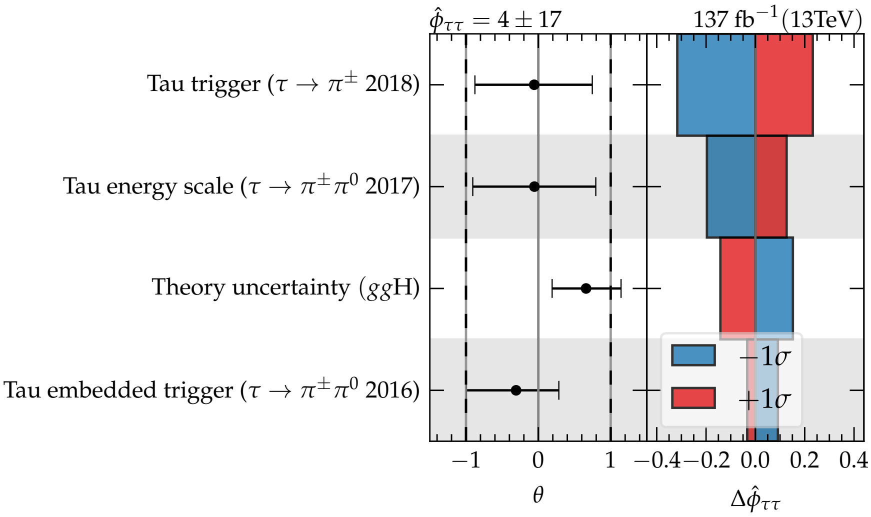

Figure 7.28: Impacts and pull distributions of the most significant sources of systematic uncertainty at their best-fit values, where θ represents the pull from the nominal value including its variation, whilst the distribution in $\Delta\hat{\phi}_{\tau\tau}$ indicates the shift induced by the nuisance parameter to the fit parameter of interest $\phi_{\tau\tau}$ when fixing the parameter to its $\pm 1\sigma$ best-fit values. The data-taking era in the parameter name indicates which dataset is considered. If the year remains unspecified, the uncertainty is correlated among all three data-taking eras. The τ_h-lepton trigger efficiency and energy scale in both embedded and MC simulated samples are amongst the dominant uncertainties, along with the theory uncertainties associated to the ggH production mode. No parameter exhibits excessive constraints from the fit or a pull larger than 1σ. The neutrinos in the τ_h-lepton decays are implied.

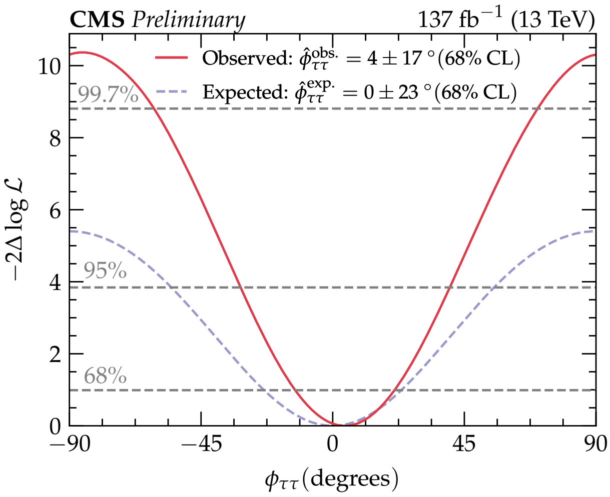

Figure 7.29: Scan of the negative log-likelihood $-2\Delta \log \mathcal{L}$ for $\mathcal{CP}$-mixing angle $\phi_{\tau\tau}$. The purple, dashed curve represents the expected result using an Asimov dataset, which gives a result of $(0 \pm 23)^{\circ}$ at the 68% CL and a significance of 2.3σ to reject the pure PS hypothesis. The red curve, on the other hand, illustrates the measured value of $\phi_{\tau\tau}$. The best-fit value is $(4 \pm 17)^{\circ}$ at the 68% CL, and the significance to reject the pure PS hypothesis is 3.2σ, as determined by the observed difference in $-2\Delta \log \mathcal{L}$ at $\phi_{\tau\tau} = 0^{\circ}$ and $|\phi_{\tau\tau}| = 90^{\circ}$. Additionally, the uncertainty at the 95% CL and 99.7% CL is observed to be $\pm 36^{\circ}$ and $\pm 66^{\circ}$, respectively. Figure taken from [1].

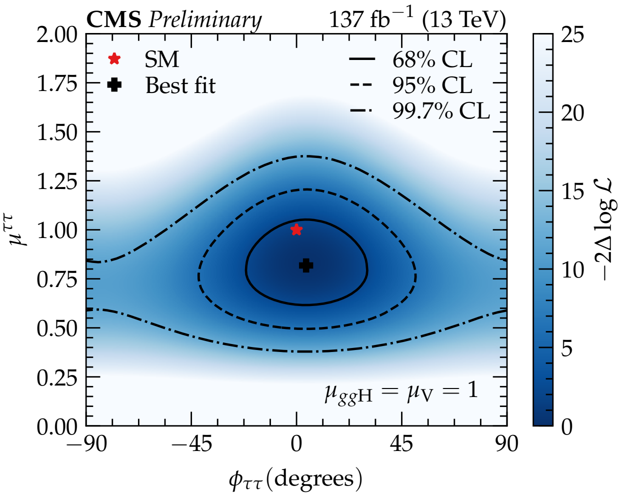

Figure 7.30: Two-dimensional scan of the negative log-likelihood of the total branching fraction modifier $\mu^{\tau\tau}$ against the $\mathcal{CP}$-mixing angle $\phi_{\tau\tau}$. The 68%, 95%, and 99.7% CL regions are indicated as contours, and the measured, best-fit value is within the 68% CL area. Higgs boson couplings to particles other than the τ-lepton are set to their SM predicted values. Good agreement with the expected value of the SM has been observed. Figure taken from [1].

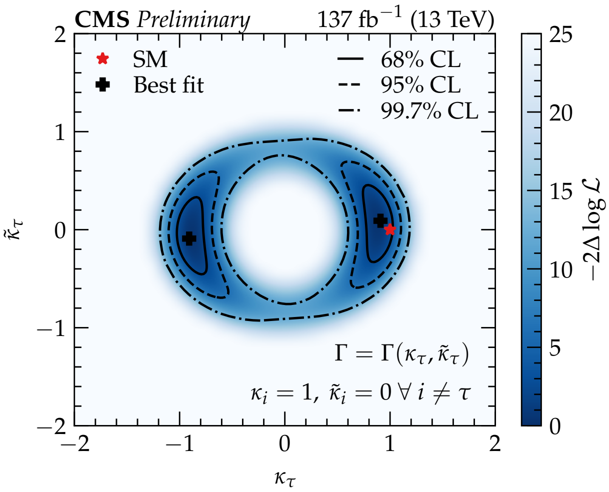

Figure 7.31: Two-dimensional scan of the negative log-likelihood of the $\mathcal{CP}$-even (κ) and $\mathcal{CP}$-odd ($\tilde{\kappa}$) Yukawa coupling constants to the τ-lepton. The 68%, 95%, and 99.7% CL regions are indicated as contours, and the measured, best-fit value is within the 68% CL area. Higgs boson couplings to particles other than the τ-lepton are set to their SM predicted values. Good agreement with the expected value of the SM has been observed. Figure taken from [1].

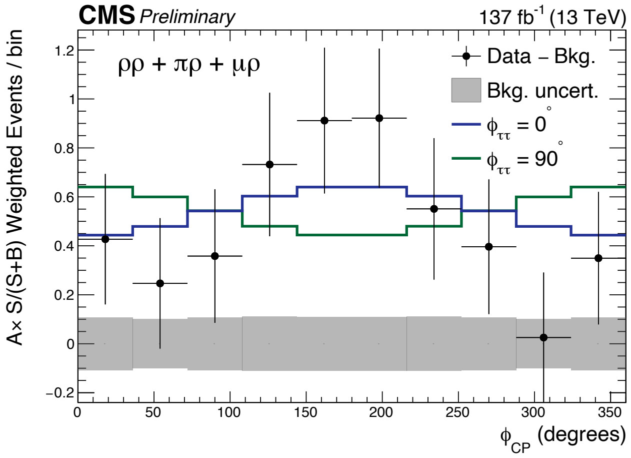

Figure 7.32: Weighted $\phi_{\mathcal{CP}}$ distribution of the three most sensitive final states considered. Event frequencies are collected from all BDT/NN distributions in the signal categories S, and the background B is subtracted from the data. A reweighting of $A \cdot S/(S+B)$ is applied, where A indicates the mean asymmetry between $\mathcal{CP}$-even and $\mathcal{CP}$-odd distributions, and is defined as $|\mathcal{CP}_i^{\text{even}} - \mathcal{CP}_i^{\text{odd}}|/(\mathcal{CP}_i^{\text{even}} + \mathcal{CP}_i^{\text{odd}})$ for bin i and normalised to the total number of bins. The uncertainty on the subtracted background is accounted for by the grey shaded area. The $\mathcal{CP}$-even model is favoured by the data. Figure taken from [1].

Chapter 8

Summary and Conclusions

Measurements of the $\mathcal{CP}$ structure of the Higgs boson couling to τ-leptons and top quarks using data of pp collisions collected by the CMS experiment at $\sqrt{s}$ =13 TeV have been presented. A total integrated luminosity of 137 fb^{-1} with the $\tau_h\tau_h$ and $\tau_\mu\tau_h$ final states was analysed. The Higgs-top Yukawa coupling measurement is parameterised with an effective $\mathcal{CP}$-mixing angle, α_{gg}, that was observed to be $(-5^{+36}_{-37})^\circ$ at 68% CL. The result is compatible with the expectation predicted in the SM, however, the uncertainty is relatively large. This measurement will benefit in the near future from the inclusion of more data from the $\tau_e\tau_h$ and $\tau_e\tau_\mu$ final states, which are expected to provide about $\mathcal{O}$ (10%) more significance to the result. A comparable improvement can be obtained through the multi-classifier categorisation scheme. The results of the presented studies indicate a $\mathcal{O}$ (10%) improvement based on an Azimov dataset. Similarly, the Higgs-τ Yukawa coupling measurement used an analogous parametrisation, where the deviation from the SM can be determined directly by measuring an effective mixing angle, $\phi_{\tau\tau}$. The best-fit value of this parameter was observed to be $(4 \pm 17)^\circ$ at 68% CL. This constitutes the first measurement of the $\mathcal{CP}$ nature of the Yukawa coupling between the Higgs boson and τ-lepton. The analysis is documented in Reference [1].

Whilst the former analysis presented here is not sensitive enough to make strong conclusions on the $\mathcal{CP}$ nature of the Higgs boson, the latter presents evidence against the pure pseudoscalar hypothesis at a significance level of 3.2σ. A similar level of significance to reject the pure $\mathcal{CP}$-odd coupling has been observed recently at the ttH and tH vertices in H $\rightarrow \gamma\gamma$ events by the CMS [95] and ATLAS [96] collaborations.

The measurement of $\phi_{\tau\tau}$ is able to exclude parts of the phase space predicted by the next-to-minimal supersymmetric model, which is a type of model that attempts to expand the SM and allow for an extended Higgs sector and several $\mathcal{CP}$ violating phases [97]. In this model, $\phi_{\tau\tau}$ is restricted to values below 27°. Thus, the presented result is able to exclude some phase space at the 68% CL.

The measurements use the same data, but are complementary in the way that one probes the Yukawa coupling at the effective $gg \rightarrow$ H production vertex, whilst the other probes the coupling at the decay side, using the $\mathrm{H} \rightarrow \tau^{+}\tau^{-}$ decay vertex, making the latter a more direct and model-independent measurement. As the dominant sources of uncertainties are statistical, accumulating and analysing more collision data will improve the precision of the measurement. Furthermore, the techniques can be improved using more advanced machine learning algorithms. It was shown that introducing BDT-based categorisation methods can significantly improve the results, if adapted properly to the task at hand. Other aspects of the analyses can be improved by following the general trend in CMS of advanced machine learning methods. Additionally, the ultimate limitation on the quality and quantity of the data arises from the CMS trigger. Improvements in the trigger implementation, such as dedicated VBF triggers, can enhance the effective number of interesting collisions recorded. By directly targeting the phase space spanned by the VBF jets associated to the Higgs boson production mode, the trigger thresholds on the τ_h leptons can be reduced to 20 GeV, which adds to the current dataset triggered at 40 GeV.

$\mathcal{CP}$ measurements are crucial to improve our understanding of the Universe, and whilst these have been around for many years in the quark and neutrino sectors, these types of measurement have only just begun to pick up in Higgs physics. It will be interesting to see in the next decade whether a deviation from the SM prediction will be observed, and what implications this may furnish.

Bibliography

[1] CMS Collaboration Collaboration, "Analysis of the CP structure of the Yukawa coupling between the Higgs boson and τ leptons in proton-proton collisions at $\sqrt{s} = 13$ TeV", Technical Report CMS-PAS-HIG-20-006, CERN, Geneva, (2020).

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

[3] CMS Collaboration, "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC", *Phys. Lett.* **B716** (2012) 30–61, doi:10.1016/j.physletb.2012.08.021, arXiv:1207.7235.

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

[5] P. W. Higgs, "Broken symmetries, massless particles and gauge fields", *Phys. Lett.* **12** (1964) 132–133, doi:10.1016/0031-9163(64)91136-9.

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

[7] P. W. Higgs, "Spontaneous Symmetry Breakdown without Massless Bosons", *Phys. Rev.* **145** (1966) 1156–1163, doi:10.1103/PhysRev.145.1156.

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

[9] T. Kibble, "Symmetry breaking in nonAbelian gauge theories", *Phys. Rev.* **155** (1967) 1554–1561, doi:10.1103/PhysRev.155.1554.

[10] CMS Collaboration Collaboration, "Study of the Mass and Spin-Parity of the Higgs Boson Candidate via Its Decays to Z Boson Pairs", *Phys. Rev. Lett.* **110** (Feb, 2013) 081803, doi:10.1103/PhysRevLett.110.081803.

[11] V. Hankele, G. Klämke, D. Zeppenfeld et al., "Anomalous Higgs boson couplings in vector boson fusion at the CERN LHC", *Phys. Rev. D* **74** (Nov, 2006) 095001,

doi:10.1103/PhysRevD.74.095001.

[12] Particle Data Group Collaboration, “Review of Particle Physics”, *Phys. Rev. D* **98** (Aug, 2018) 030001, doi:10.1103/PhysRevD.98.030001.

[13] K. Abe, J. Adam, H. Aihara et al., “Observation of Electron Neutrino Appearance in a Muon Neutrino Beam”, *Physical Review Letters* **112** (Feb, 2014) doi:10.1103/physrevlett.112.061802.

[14] S. Glashow, “Partial Symmetries of Weak Interactions”, *Nucl. Phys.* **22** (1961) 579–588, doi:10.1016/0029-5582(61)90469-2.

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

[16] A. Salam, “Weak and Electromagnetic Interactions”, *Conf. Proc. C* **680519** (1968) 367–377, doi:10.1142/9789812795915_0034.

[17] M. Peskin, “An Introduction To Quantum Field Theory”. CRC Press, 2018.

[18] G. Aad, T. Abajyan, B. Abbott et al., “Evidence for the spin-0 nature of the Higgs boson using ATLAS data”, *Physics Letters B* **726** (2013), no. 1, 120 – 144, doi:https://doi.org/10.1016/j.physletb.2013.08.026.

[19] LHC Higgs Cross Section Working Group Collaboration, “Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector”, doi:10.2172/1345634,10.23731/CYRM-2017-002, arXiv:1610.07922.

[20] M. Aaboud, G. Aad, B. Abbott et al., “Observation of H→ bb decays and VH production with the ATLAS detector”, *Physics Letters B* **786** (Nov, 2018) 59–86, doi:10.1016/j.physletb.2018.09.013.

[21] A. Sirunyan, A. Tumasyan, W. Adam et al., “Observation of Higgs Boson Decay to Bottom Quarks”, *Physical Review Letters* **121** (Sep, 2018) doi:10.1103/physrevlett.121.121801.

[22] A. Sirunyan, A. Tumasyan, W. Adam et al., “Observation of the Higgs boson decay to a pair of τ leptons with the CMS detector”, *Physics Letters B* **779** (2018) 283 – 316, doi:https://doi.org/10.1016/j.physletb.2018.02.004.

[23] P. F. Pérez and M. B. Wise, “On the origin of neutrino masses”, *Physical Review D* **80** (Sep, 2009) doi:10.1103/physrevd.80.053006.

[24] A. D. Sakharov, “Violation ofCPin variance,Casymmetry, and baryon asymmetry of the universe”, *Soviet Physics Uspekhi* **34** (may, 1991) 392–393, doi:10.1070/pu1991v034n05abeh002497.

[25] T2K Collaboration, "Constraint on the matter–antimatter symmetry-violating phase in neutrino oscillations", *Nature* **580** (2020), no. 7803, 339–344, doi:10.1038/s41586-020-2177-0, arXiv:1910.03887.

[26] G. Branco, P. Ferreira, L. Lavoura et al., "Theory and phenomenology of two-Higgs-doublet models", *Physics Reports* **516** (Jul, 2012) 1–102, doi:10.1016/j.physrep.2012.02.002.

[27] M. J. Dolan, P. Harris, M. Jankowiak et al., "Constraining CP-violating Higgs sectors at the LHC using gluon fusion", *Physical Review D* **90** (Oct, 2014) doi:10.1103/physrevd.90.073008.

[28] S. Choi, D. Miller, M. Mühlleitner et al., "Identifying the Higgs spin and parity in decays to Z pairs", *Physics Letters B* **553** (Jan, 2003) 61–71, doi:10.1016/s0370-2693(02)03191-x.

[29] A. Sirunyan, A. Tumasyan, W. Adam et al., "Constraints on anomalous HVV couplings from the production of Higgs bosons decaying to τ lepton pairs", *Physical Review D* **100** (Dec, 2019) doi:10.1103/physrevd.100.112002.

[30] CMS Collaboration Collaboration, "Measurements of $t\bar{t}$H production and the CP structure of the Yukawa interaction between the Higgs boson and top quark in the diphoton decay channel", Technical Report arXiv:2003.10866. CMS-HIG-19-013-003, CERN, Geneva, (Mar, 2020). Submitted to PRL. All figures and tables can be found at http://cms-results.web.cern.ch/cms-results/public-results/publications/HIG-19-013 (CMS Public Pages).

[31] S. Berge and W. Bernreuther, "Determining the CP parity of Higgs bosons at the LHC in the τ to 1-prong decay channels", *Physics Letters B* **671** (Feb, 2009) 470–476, doi:10.1016/j.physletb.2008.12.065.

[32] S. Berge, W. Bernreuther, and S. Kirchner, "Determination of the Higgs CP-mixing angle in the tau decay channels", *Nuclear and Particle Physics Proceedings* **273-275** (2016) 841 – 845, doi:https://doi.org/10.1016/j.nuclphysbps.2015.09.129. 37th International Conference on High Energy Physics (ICHEP).

[33] G. Klämke and D. Zeppenfeld, "Higgs plus two jet production via gluon fusion as a signal at the CERN LHC", *Journal of High Energy Physics* **2007** (Apr, 2007) 052–052, doi:10.1088/1126-6708/2007/04/052.

[34] V. Hankele, G. Klämke, D. Zeppenfeld et al., "Anomalous Higgs boson couplings in vector boson fusion at the CERN LHC", *Physical Review D* **74** (Nov, 2006) doi:10.1103/physrevd.74.095001.

[35] T. Plehn, D. Rainwater, and D. Zeppenfeld, “Determining the Structure of Higgs Couplings at the CERN Large Hadron Collider”, *Physical Review Letters* **88** (Jan, 2002) doi:10.1103/physrevlett.88.051801.

[36] M. Krämer, J. Kühn, M. L. Stong et al., “Prospects of measuring the parity of Higgs particles”, *Zeitschrift für Physik C Particles and Fields* **64** (Mar, 1994) 21–30, doi:10.1007/bf01557231.

[37] A. Buckley, J. Butterworth, S. Gieseke et al., “General-purpose event generators for LHC physics”, *Physics Reports* **504** (Jul, 2011) 145–233, doi:10.1016/j.physrep.2011.03.005.

[38] M. Cacciari, G. P. Salam, and G. Soyez, “The anti-ktjet clustering algorithm”, *Journal of High Energy Physics* **2008** (Apr, 2008) 063–063, doi:10.1088/1126-6708/2008/04/063.

[39] L. Evans and P. Bryant, “LHC Machine”, *JINST* **3** (2008) S08001, doi:10.1088/1748-0221/3/08/S08001.

[40] ATLAS Collaboration, “The ATLAS Experiment at the CERN Large Hadron Collider”, *JINST* **3** (2008) S08003, doi:10.1088/1748-0221/3/08/S08003.

[41] CMS Collaboration, “The CMS Experiment at the CERN LHC”, *JINST* **3** (2008) S08004, doi:10.1088/1748-0221/3/08/S08004.

[42] LHCb Collaboration, “The LHCb Detector at the LHC”, *JINST* **3** (2008) S08005, doi:10.1088/1748-0221/3/08/S08005.

[43] ALICE Collaboration, “The ALICE experiment at the CERN LHC”, *JINST* **3** (2008) S08002, doi:10.1088/1748-0221/3/08/S08002.

[44] T. Sakuma and T. McCauley, “Detector and Event Visualization with SketchUp at the CMS Experiment”, *J. Phys. Conf. Ser.* **513** (2014) 022032, doi:10.1088/1742-6596/513/2/022032, arXiv:1311.4942.

[45] CMS Collaboration, “The CMS magnet project: Technical Design Report”. Technical Design Report CMS. CERN, Geneva, 1997.

[46] CMS Collaboration, V. Karimäki, M. Mannelli, P. Siegrist et al., “The CMS tracker system project: Technical Design Report”. Technical Design Report CMS. CERN, Geneva, 1997.

[47] A. Dominguez, D. Abbaneo, K. Arndt et al., “CMS Technical Design Report for the Pixel Detector Upgrade”, Technical Report CERN-LHCC-2012-016. CMS-TDR-11, (Sep, 2012). Additional contacts: Jeffrey Spalding, Fermilab, Jeffrey.Spalding@cern.ch Didier Contardo, Universite Claude Bernard-Lyon I, didier.claude.contardo@cern.ch.

[48] CMS Collaboration Collaboration, “The CMS electromagnetic calorimeter project: Technical Design Report”. Technical Design Report CMS. CERN, Geneva, 1997.

[49] CMS Collaboration Collaboration, “The CMS ECAL performance with examples”, Technical Report CMS-CR-2013-430, CERN, Geneva, (Nov, 2013).

[50] J. Mans, J. Anderson, B. Dahmes et al., “CMS Technical Design Report for the Phase 1 Upgrade of the Hadron Calorimeter”, Technical Report CERN-LHCC-2012-015. CMS-TDR-10, (Sep, 2012). Additional contact persons: Jeffrey Spalding, Fermilab, spalding@cern.ch, Didier Contardo, Universite Claude Bernard-Lyon I, contardo@cern.ch.

[51] CMS Collaboration, J. G. Layter, “The CMS muon project: Technical Design Report”. Technical Design Report CMS. CERN, Geneva, 1997.

[52] CMS Collaboration, “The performance of the CMS muon detector in proton-proton collisions at $\sqrt{s}$= 7 TeV at the LHC”, *Journal of Instrumentation* **8** (nov, 2013) P11002–P11002, doi:10.1088/1748-0221/8/11/p11002.

[53] CMS Collaboration, “Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV”, *Journal of Instrumentation* **13** (June, 2018) P06015, doi:10.1088/1748-0221/13/06/P06015, arXiv:1804.04528.

[54] CMS collaboration Collaboration, “CMS Technical Design Report for the Level-1 Trigger Upgrade”, Technical Report CERN-LHCC-2013-011. CMS-TDR-12, (Jun, 2013). Additional contacts: Jeffrey Spalding, Fermilab, Jeffrey.Spalding@cern.ch Didier Contardo, Universite Claude Bernard-Lyon I, didier.claude.contardo@cern.ch.

[55] D. Trocino, “The CMS High Level Trigger”, *Journal of Physics: Conference Series* **513** (jun, 2014) 012036, doi:10.1088/1742-6596/513/1/012036.

[56] CMS Collaboration, Bayatyan, G L and Della Negra, Michel and Foà, and Hervé, A and Petrilli, Achille, “CMS computing: Technical Design Report”. Technical Design Report CMS. CERN, Geneva, 2005. Submitted on 31 May 2005.

[57] CMS Collaboration, “Description and performance of track and primary-vertex reconstruction with the CMS tracker”, *JINST* **9** (May, 2014) P10009. 80 p, doi:10.1088/1748-0221/9/10/P10009.

[58] R. E. Kalman, “A New Approach to Linear Filtering and Prediction Problems”, *Transactions of the ASME–Journal of Basic Engineering* **82** (1960), no. Series D, 35–45.

[59] K. Rose, “Deterministic annealing for clustering, compression, classification, regression, and related optimization problems”, *Proceedings of the IEEE* **86** (1998), no. 11, 2210–2239.

[60] R. Frühwirth, W. Waltenberger, and P. Vanlaer, “Adaptive Vertex Fitting”, Technical Report CMS-NOTE-2007-008, CERN, Geneva, (Mar, 2007).

[61] CMS Collaboration, “Particle-flow reconstruction and global event description with the CMS detector”, *JINST* **12** (2017) 10003, doi:10.1088/1748-0221/12/10/P10003.

[62] CMS Collaboration, “Pileup mitigation at CMS in 13 TeV data”, arXiv:2003.00503.

[63] D. Bertolini, P. Harris, M. Low et al., “Pileup Per Particle Identification”, *JHEP* **10** (2014) 059, doi:10.1007/JHEP10(2014)059, arXiv:1407.6013.

[64] CMS Collaboration, “Performance of reconstruction and identification of τ leptons decaying to hadrons and ν_τ in pp collisions at $\sqrt{s} = 13$ TeV”, *JINST* **13** (Sep, 2018) P10005. 71 p, doi:10.1088/1748-0221/13/10/P10005. Replaced with the published version. Added the journal reference and the DOI. All the figures and tables can be found at http://cms-results.web.cern.ch/cms-results/public-results/publications/TAU-16-003 (CMS Public Pages).

[65] CMS Collaboration, “Identification of hadronic tau decay channels using multivariate analysis (MVA decay mode)”,.

[66] L. Bianchini, B. Calpas, J. Conway et al., “Reconstruction of the Higgs mass in events with Higgs bosons decaying into a pair of τ leptons using matrix element techniques”, *Nucl. Instrum. Meth. A* **862** (2017) 54–84, doi:10.1016/j.nima.2017.05.001, arXiv:1603.05910.

[67] J. Alwall, M. Herquet, F. Maltoni et al., “MadGraph 5: going beyond”, *Journal of High Energy Physics* **2011** (Jun, 2011) doi:10.1007/jhep06(2011)128.

[68] J. Alwall, S. Höche, F. Krauss et al., “Comparative study of various algorithms for the merging of parton showers and matrix elements in hadronic collisions”, *The European Physical Journal C* **53** (Dec, 2007) 473–500, doi:10.1140/epjc/s10052-007-0490-5.

[69] S. Frixione, P. Nason, and C. Oleari, “Matching NLO QCD computations with parton shower simulations: the POWHEG method”, *Journal of High Energy Physics* **2007** (Nov, 2007) 070–070, doi:10.1088/1126-6708/2007/11/070.

[70] S. Alioli, K. Hamilton, P. Nason et al., “Jet pair production in POWHEG”, *Journal of High Energy Physics* **2011** (Apr, 2011) doi:10.1007/jhep04(2011)081.

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

[72] R. Frederix and S. Frixione, “Merging meets matching in MC@NLO”, *Journal of High Energy Physics* **2012** (Dec, 2012) doi:10.1007/jhep12(2012)061.

[73] R. D. Ball, V. Bertone, S. Carrazza et al., “Parton distributions for the LHC run II”, *Journal*

of High Energy Physics **2015** (Apr, 2015) doi:10.1007/jhep04(2015)040.

[74] NNPDF Collaboration, “Parton distributions from high-precision collider data”, *Eur. Phys. J. C* **77** (2017), no. 10, 663, doi:10.1140/epjc/s10052-017-5199-5, arXiv:1706.00428.

[75] T. Sjöstrand, S. Ask, J. R. Christiansen et al., “An introduction to PYTHIA 8.2”, *Computer Physics Communications* **191** (2015) 159 – 177, doi:https://doi.org/10.1016/j.cpc.2015.01.024.

[76] V. Khachatryan, A. M. Sirunyan, A. Tumasyan et al., “Event generator tunes obtained from underlying event and multiparton scattering measurements”, *The European Physical Journal C* **76** (Mar, 2016) doi:10.1140/epjc/s10052-016-3988-x.

[77] CMS Collaboration Collaboration, “Extraction and validation of a new set of CMS PYTHIA8 tunes from underlying-event measurements”, Technical Report CMS-PAS-GEN-17-001, CERN, Geneva, (2018).

[78] GEANT4 Collaboration, “GEANT4: A Simulation toolkit”, *Nucl. Instrum. Meth. A* **506** (2003) 250–303, doi:10.1016/S0168-9002(03)01368-8.

[79] CMS Collaboration Collaboration, “An embedding technique to determine genuine $\tau\tau$ backgrounds from CMS data”, Technical Report CMS-PAS-TAU-18-001, CERN, Geneva, (2018).

[80] CMS Collaboration, “Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at $\sqrt{s} = 13$ TeV”, *Journal of Instrumentation* **13** (Jun, 2018) P06015–P06015, doi:10.1088/1748-0221/13/06/p06015.

[81] T. C. collaboration, “Determination of jet energy calibration and transverse momentum resolution in CMS”, *Journal of Instrumentation* **6** (nov, 2011) P11002–P11002, doi:10.1088/1748-0221/6/11/p11002.

[82] M. Cacciari and G. P. Salam, “Pileup subtraction using jet areas”, *Physics Letters B* **659** (Jan, 2008) 119–126, doi:10.1016/j.physletb.2007.09.077.

[83] CMS Collaboration, “Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV”, *JINST* **13** (2018), no. 05, P05011, doi:10.1088/1748-0221/13/05/P05011, arXiv:1712.07158.

[84] T. C. collaboration, “Performance of the CMS missing transverse momentum reconstruction in pp data at $\sqrt{}$s = 8 TeV”, *Journal of Instrumentation* **10** (feb, 2015) P02006–P02006, doi:10.1088/1748-0221/10/02/p02006.

[85] CMS Collaboration, “Performance of the DeepTau algorithm for the discrimination of

taus against jets, electron, and muons",.

[86] K. Hamilton, P. Nason, E. Re et al., "NNLOPS simulation of Higgs boson production", *Journal of High Energy Physics* **2013** (Oct, 2013) doi:10.1007/jhep10(2013)222.

[87] CMS Collaboration Collaboration, "Measurement of Higgs boson production and decay to the $\tau\tau$ final state", Technical Report CMS-PAS-HIG-18-032, CERN, Geneva, (2019).

[88] F. Demartin, F. Maltoni, K. Mawatari et al., "Higgs characterisation at NLO in QCD: CP properties of the top-quark Yukawa interaction", *Eur. Phys. J. C* **74** (2014), no. 9, 3065, doi:10.1140/epjc/s10052-014-3065-2, arXiv:1407.5089.

[89] "KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining". Association for Computing Machinery, New York, NY, USA, (2016).

[90] R. Barlow and C. Beeston, "Fitting using finite Monte Carlo samples", *Computer Physics Communications* **77** (1993), no. 2, 219 – 228, doi:https://doi.org/10.1016/0010-4655(93)90005-W.

[91] S. Berge, W. Bernreuther, B. Niepelt et al., "How to pin down theCPquantum numbers of a Higgs boson in itsdecays at the LHC", *Physical Review D* **84** (Dec, 2011) doi:10.1103/physrevd.84.116003.

[92] E. Brochu, V. M. Cora, and N. de Freitas, "A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning", 2010.

[93] S. Berge, W. Bernreuther, and S. Kirchner, "Determination of the Higgs CP-mixing angle in the tau decay channels at the LHC including the Drell–Yan background", *The European Physical Journal C* **74** (Nov, 2014) doi:10.1140/epjc/s10052-014-3164-0.

[94] S. Berge, W. Bernreuther, and S. Kirchner, "Prospects of constraining the Higgs boson's CP nature in the tau decay channel at the LHC", *Physical Review D* **92** (Nov, 2015) doi:10.1103/physrevd.92.096012.

[95] A. Sirunyan, A. Tumasyan, W. Adam et al., "Measurements of ttH Production and the CP Structure of the Yukawa Interaction between the Higgs Boson and Top Quark in the Diphoton Decay Channel", *Physical Review Letters* **125** (Aug, 2020) doi:10.1103/physrevlett.125.061801.

[96] G. Aad, B. Abbott, D. Abbott et al., "CP Properties of Higgs Boson Interactions with Top Quarks in the ttH and tH Processes Using H→ $\gamma\gamma$ with the ATLAS Detector", *Physical Review Letters* **125** (Aug, 2020) doi:10.1103/physrevlett.125.061802.

[97] S. King, M. Mühlleitner, R. Nevzorov et al., “Exploring the CP-violating NMSSM: EDM constraints and phenomenology”, *Nuclear Physics B* **901** (Dec, 2015) 526–555, doi:10.1016/j.nuclphysb.2015.11.003.