ETP-KA/2022-20

Karlsruhe Institute of Technology

Institute of Experimental Particle Physics

# A data-driven method for Higgs boson analyses in di-τ final states for the LHC Run II and beyond

Zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

von der KIT-Fakultät für Physik
des Karlsruher Instituts für Technologie (KIT)
angenommene

Dissertation

von

M.Sc. Sebastian Brommer

aus Stuttgart

Tag der mündlichen Prüfung: 09. Dezember 2022

Erster Gutachter: Prof. Dr. Günter Quast
Zweiter Gutachter: Priv.-Doz. Dr. Roger Wolf

*Institut für Experimentelle Teilchenphysik*

## Abstract

The $\tau$-embedding is a method to estimate the contribution from events with two genuine $\tau$ leptons in the event from data. The technique uses an event-by-event approach, where two reconstructed muons are selected in data, which are replaced by two simulated $\tau$ lepton decays. The resulting $\tau$-embedded event only relies on the simulation of the $\tau$ lepton decays, while the rest of the event remains unchanged. This approach results in an improved description of the properties of jets, the underlying event, and pile-up collisions. The $\tau$-embedding method is the main estimation method for genuine di-$\tau$ backgrounds within the CMS Collaboration and has been applied in numerous Higgs boson analyses in di-$\tau$ final states over the last years. The most recent implementation of the method is described in this thesis. The method is compared with a model based on fully simulated processes in a comprehensive, realistic analysis example. More than 8 million CPU hours have been spent to produce the most recent implementation of $\tau$-embedded samples for the LHC Run II analyses. The presented studies lay the foundation for using $\tau$-embedded samples in several anticipated Higgs boson analyses in di-$\tau$ final states on the combined Run II and III data sets, which will form one of the major results of the LHC phase 1 physics program.

# Zusammenfassung

Das $\tau$-Embedding ist eine datenbasierte Methode zur Abschätzung des Beitrags von Prozessen mit zwei $\tau$-Leptonen im Ereignis. Die Methode verwendet einen ereignisbasierten Ansatz, bei dem zwei rekonstruierte Myonen in den Daten ausgewählt werden, die durch zwei simulierte $\tau$-Leptonenzerfälle ersetzt werden. Das daraus resultierende Ereignis vereint die simulierten $\tau$-Leptonenzerfälle mit einem sonst unveränderten Ereignis. Das $\tau$-Embedding führt zu einer verbesserten Beschreibung der Eigenschaften von Jets und von Pile-up-Kollisionen. Es ist die wichtigste Abschätzungsmethode für Untergründe mit zwei $\tau$-Leptonen im Endzustand innerhalb der CMS-Kollaboration und wurde in den letzten Jahren in zahlreichen Higgs-Boson-Analysen in $\tau\tau$-Endzuständen angewendet.

In dieser Arbeit wird die neueste Implementierung der Methode beschrieben. In einem umfassenden, Analysebeispiel wird die Methode mit einem Modell verglichen, das auf vollständig simulierten Prozessen basiert. Mehr als 8 Millionen CPU-Stunden wurden aufgewendet, um die neue Implementierung von $\tau$-Embedding Ergebnisse für die LHC Run II Analysen zu erzeugen. Die vorgestellten Studien legen den Grundstein für die Verwendung von $\tau$-Embedding in mehreren geplanten Higgs-Boson-Analysen in $\tau\tau$-Endzuständen auf den kombinierten Datensätzen von Run II und III, die eines der wichtigsten Ergebnisse des LHC-Phase-1-Physikprogramms darstellen werden.

# Contents

# 1 Introduction

Since its first formulation, the standard model of particle physics (SM) has proven to be an exceptional theory capable of predicting the existence of several particles long before their experimental observation. The last missing particle of the SM, the Higgs boson, was predicted by the Brout-Englert-Higgs Mechanism [1–3] in 1964 and observed by the ATLAS [4] and CMS Collaborations [5] in 2012 [6, 7]. This discovery was achieved at the Large Hadron Collider (LHC) at CERN, a 27 km long circular particle accelerator.

While most observations in particle physics can be explained within the SM, many open questions and unexplained observations remain. Among those are the description of gravity within the SM and the existence of dark matter. While the first measurement period of the LHC between 2009 and 2013 (also referred to as Run I) led to the discovery of the Higgs boson, the second period between 2015 and 2018 (Run II) was dedicated to precision measurements of the newly discovered particle and searches for new phenomena that cannot be explained by the SM alone. Physicists pursued two strategies along this path: new physics could manifest either in the existence of new particles at the highest reachable energies or in deviations of precision measurements from the SM expectation. The Higgs sector appears particularly promising in the search for physics beyond the standard model (BSM) since there is no requirement in the Brout-Englert-Higgs Mechanism that only one Higgs boson should exist.

During Run II, numerous differential measurements of all possible Higgs boson final states have been performed to investigate the properties of the observed Higgs boson. In parallel, searches for deviations from the SM expectation have been performed in many final states in energy ranges from a few GeV to several TeV. Such analyses rely on excellent measurement devices like the Compact Muon Solenoid (CMS) detector. At the same time, they rely on accurate predictions of the known SM processes to compare the experimental measurements to. Traditionally, simulation programmes using the Monte Carlo method generate these SM predictions. An important alternative to these predictions is to estimate known SM processes from carefully selected control regions in the data. Such methods have the advantage of being faster and having less need for tuning and residual corrections. Some physics processes are difficult to properly simulate, such as light-flavour, quark-, or gluon-induced jets produced in addition to the hard scattering process, the underlying event, or additional collisions occurring during the same bunch crossing called pileup.

The $Z \rightarrow \tau\tau$ process is the dominating background process in Higgs boson analyses in di-$\tau$ final states. The $\tau$-embedding method [8], the central topic of this thesis, can be used to obtain an accurate model of all non-Higgs boson SM processes with genuine $\tau$ lepton decays in the final states mostly from data. The method relies on an event-by-event approach, where events with two muons are selected. The selected muons are replaced by simulated $\tau$ lepton decays, forming a hybrid event, where most of the event content is taken from data, and only the $\tau$ lepton decays are simulated. The method was originally developed during Run I and has been in constant development by the Institute of Experimental Particle Physics (ETP) of the Karlsruhe Institute of Technology (KIT) since then. In the scope of this thesis, the $\tau$-embedding method has been successfully applied in multiple Run II Higgs boson analyses in di-$\tau$ final states. Among those are the following analyses, which have been conducted by ETP: the most accurate differential measurements of gluon-induced (ggH) and electroweak (qqH) Higgs boson production in the di-$\tau$ final states [9–11]; a search for additional Higgs bosons and vector leptoquarks in the di-$\tau$ final state in a mass range from 60 GeV to 3500 GeV [12–14]; and a search for the decay of a heavy Higgs boson (H) into two lighter Higgs bosons h and $h_S$, of which h is the observed Higgs boson with a mass of 125 GeV in the $\tau\tau$bb final state [15, 16]. In all analyses, the $\tau$-embedding method has proven to be an essential backbone and one of the most important ingredients to the success of each corresponding analysis.

In Chapter 2, a brief overview of the SM, the Higgs boson, and $\tau$ leptons physics at the LHC is given. In Chapter 3, the CMS detector and the reconstruction algorithms used by the experiment are described, followed by a summary of Higgs boson analyses in di-$\tau$ final states during Run II in Chapter 4. In the same chapter, a more detailed overview of the SM $H \rightarrow \tau\tau$ measurement is given, which serves as a proxy for the adaptations and design choices of the $\tau$-embedding method. A more detailed discussion of the method is given in Chapter 5, which includes a description of the technical details of the method, a discussion of its strengths and limitations, and a summary of the most recent production campaign and workflows. This chapter also documents the latest version of the $\tau$-embedding method after a completely overhauled and improved reconstruction of the full LHC Run II dataset. This version of $\tau$-embedded samples will be used in the upcoming Run III and Run II+III analyses. In Chapter 6, a detailed overview of all corrections required for applying $\tau$-embedded events in a typical target analysis like the SM $H \rightarrow \tau\tau$ measurement [9] is given. A comparison between the simulation of full physics processes and the $\tau$-embedding method on a subset of the data that have been analysed in [9] is performed in Chapter 7. The work presented in this chapter is supposed to serve as the foundation for the upcoming and planned Run II+III combinations of the above analyses. A Summary and conclusions are given in Chapter 8.

# 2 Theoretical Foundations

The SM is a gauge field theory to describe all fundamental particles and their interactions in a consistent formalism.

An overview of the particles which are part of the SM is given in Figure 2.1. All elementary particles can be subdivided into two groups, fermions and bosons. Fermions carry a half-integer spin, while bosons carry an integer spin. Depending on their interactions, fermions can be further separated into quarks and leptons. Leptons interact only via electroweak interactions while quarks can also interact via the strong interaction. Fermions can also be further grouped into three generations. The major difference between fermions in those generations is their mass. Matter in the world that surrounds us in our everyday life is made up of first-generation fermions. Quarks are separated into up-type quarks with an electric charge of 2/3 and down-type quarks with a charge of -1/3. Throughout this thesis, natural units with $\hbar = c = 1$ are used.

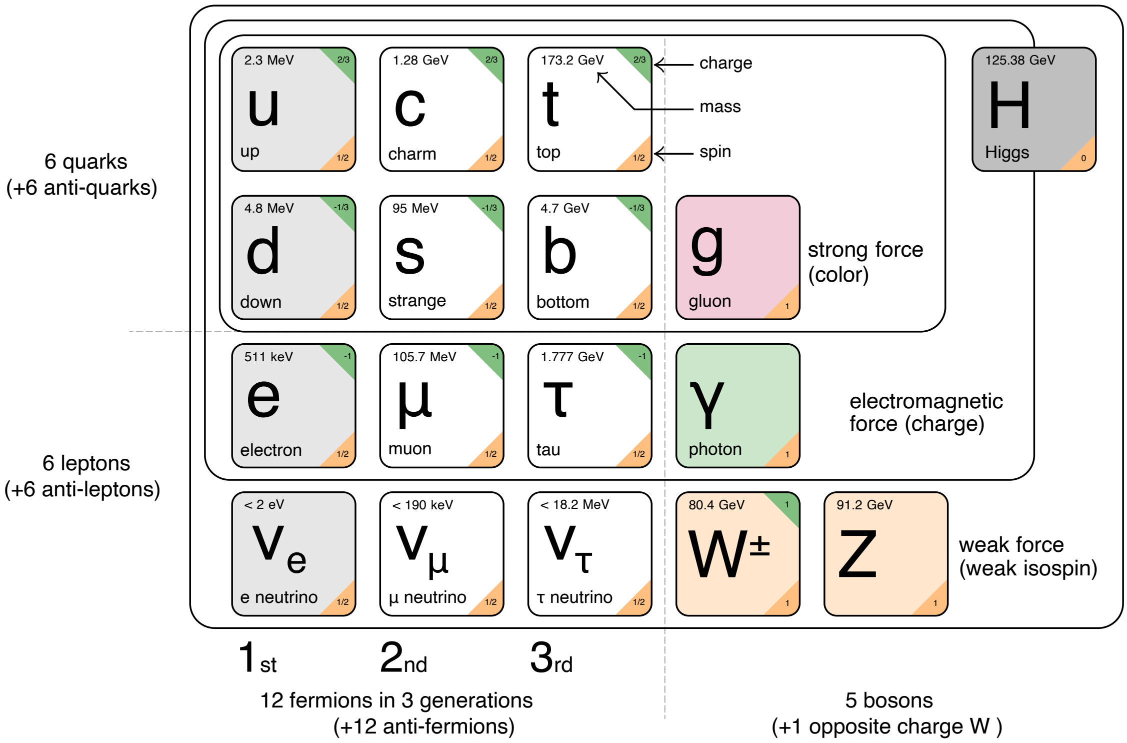


**Figure 2.1.:** Overview of all elementary particles which are part of the SM. The visualization is adapted from [17].

The different forces are each mediated by at least one gauge boson. Gluons mediate the strong force; the electroweak force is mediated by the photon, $W^+$, $W^-$, and Z bosons.

Apart from the Higgs boson, which does not mediate a fundamental force, all bosons carry spin 1.

The current formulation of the SM is based on the $\mathrm{SU}(3)_c \times \mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ symmetry group. The $\mathrm{SU}(3)_c$ group describes the strong interaction, while the $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ group describes the electroweak interactions. The latest addition to the SM is the spontaneous symmetry breaking of the electroweak interactions via the Brout-Englert-Higgs mechanism [1–3] proposed by Robert Brout, Francois Englert and Peter Higgs in 1964.

**Strong Interaction** The strong interaction is described by quantum chromodynamics (QCD). The charge of the QCD is the colour charge; three different colour charges ( usually referred to as red, green, and blue), and their corresponding anti-colour charges exist. The strong force is mediated by gluons, which are massless bosons with a spin of one. For large distances or small energies, the potential of the strong force becomes large enough that more colour-charged particles are created. Due to this phenomenon called confinement, all particles with a colour charge form colour-neutral objects, called hadrons. The coupling strength is minimal for high energies or small distances, resulting in asymptotic freedom, where colour-charged particles are quasi-free.

**Electroweak Interaction** The description of electroweak (EWK) interactions, the unification of the electromagnetic and the weak force, is based on the works of Shelden Glasgow, Abdus Salam and Steven Weinberg, who developed the *Glashow-Salam-Weinberg* theory between 1960 and 1970 [18–20]. The charge corresponding to the $\mathrm{U}(1)_Y$ symmetry in the electroweak theory is the weak hypercharge. The charge corresponding to the $\mathrm{SU}(2)_L$ is the weak isospin. Within the theory, three gauge bosons $W^i_\mu$ for the $\mathrm{SU}(2)_L$ with a coupling constant $g$ and one gauge boson $B_\mu$ for the $\mathrm{U}(1)_Y$ group with the coupling constant $g'$ exist. The observable bosons of the EWK interaction, the two charged massive W bosons, a neutral massive Z boson, and a neural massless photon are superpositions of the fundamental gauge bosons. Due to the large masses of the W and Z bosons, the weak interaction only has a short range. Within the EWK interaction, a distinction is made between left-handed and right-handed particles as the $\mathrm{SU}(2)_L$ is a chiral symmetry. Regarding the weak isospin left-handed particles are described as doublets, while right-handed particles are described as singlets

$$\Psi_L = \begin{pmatrix} \nu_L \\ l_L \end{pmatrix} \text{ (leptons)}, \begin{pmatrix} u_i \\ d'_i \end{pmatrix} \text{ (quarks)}, \qquad \Psi_R = l_R, \tag{2.1}$$

where $u_i$ denotes an up-type and $d'_i$ a down-type flavour eigenstate of the $i$-th generation. The flavour eigenstates are defined as superpositions of the mass eigenstates

$$d'_i = \sum_{i,j} V_{ij} d_j, \tag{2.2}$$

where $V_{ij}$ is the Cabibbo-Kobayashi-Maskawa (CKM) matrix [21, 22].

## The Brout-Englert-Higgs Mechanism

The Brout-Englert-Higgs Mechanism was the last missing piece of the SM, which was required to explain the masses of the elementary particles. Due to the required SM symmetries, a naive addition of mass terms for bosons and fermions is not possible. Mass terms for the bosons break the gauge symmetry of the SM Lagrangian and mass terms for the fermions break the chiral symmetry of the $\mathrm{SU}(2)_L$. Instead, a complex doublet field $\Phi$ is introduced

$$\Phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}. \tag{2.3}$$

This self-interacting field spontaneously breaks the $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ symmetry to generate mass terms through the phenomenon of spontaneous symmetry breaking. The Lagrangian of the new doublet is

$$\mathcal{L}_{\mathrm{Higgs}} = \left(D_\mu \Phi\right)^\dagger \left(D^\mu \Phi\right) \; - V(\Phi), \tag{2.4}$$

where $V(\Phi)$ is the potential

$$V(\Phi) = \mu^2 \Phi^\dagger \Phi + \lambda (\Phi^\dagger \Phi)^2. \tag{2.5}$$

In the case of $\mu^2 < 0$, the potential has an infinite amount of minima. As a result, the vacuum expectation value (VEV) $v$ of the Higgs field $\Phi$ does not vanish. One can freely choose the ground state of the doublet to be:

$$\langle \Phi \rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v \end{pmatrix} \quad \text{with } v = \sqrt{-\frac{\mu^2}{2\lambda}}. \tag{2.6}$$

To describe the system after the symmetry breaking, a Taylor expansion around the VEV is made. Thereby, the neutral component of the Higgs field $\phi^0$ is chosen such that the $\mathrm{U}(1)_Y$ symmetry remains unbroken to ensure that the photon remains massless. The field is then denoted as

$$\Phi = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + \mathrm{H} \end{pmatrix}. \tag{2.7}$$

Together with Equation (2.4) and under consideration of the $\mathrm{SU}(2)_L \times \mathrm{U}(1)_Y$ gauge covariant derivative, one obtains the mass terms for the W and Z bosons:

$$m_W^2 = \frac{1}{4} g^2 v^2 \tag{2.8}$$

$$m_Z^2 = \frac{1}{2} (g^2 + g'^2) v^2. \tag{2.9}$$

The remaining field H corresponds to the SM Higgs boson. Its mass is given by

$$m_{\mathrm{H}} = \sqrt{2\lambda} v. \tag{2.10}$$

Since the self-coupling parameter of the potential $\lambda$ is unknown, the mass of this neutral Higgs boson is the last free parameter of the SM.

**Table 2.1.:** The SM Higgs production cross sections at a center-of-mass energy of $\sqrt{s}$ = 13 TeV in pp collisions. The mass of the SM higgs boson is set to $m_{\mathrm{H}} = 125$ GeV. The values are taken from [23].

| Production Mode | Cross Section [pb] |
|---|---|
| Gluon fusion (ggH) | $48.61^{+5.6\%}_{-7.4\%}$ |
| Vector boson fusion (VBF) | $3.77^{+2.1\%}_{-2.1\%}$ |
| Associated with W boson (WH) | $1.36^{+2.0\%}_{-2.0\%}$ |
| Associated with Z boson (ZH) | $0.88^{+4.1\%}_{-3.5\%}$ |
| Associated with $\mathrm{t\bar{t}}$ pair (ttH) | $0.51^{+6.8\%}_{-9.9\%}$ |

The masses of the fermions are generated by the Yukawa coupling term, which, e.g. for the electron is given by

$$\mathcal{L}^{e}_{\text{Yukawa}} = -f_e \left( \bar{e_R} \Phi^{\dagger} \Psi_{e,L} \right) + \text{h.c.}. \tag{2.11}$$

After an expansion in the ground state of the Higgs field, the Lagrangian becomes

$$\mathcal{L}^{e}_{\text{Yukawa}} = -f_e \left( \frac{v}{\sqrt{2}} + \frac{\mathrm{H}}{2} \right) \bar{e}e. \tag{2.12}$$

The resulting mass of the electron is then given by

$$m_e = \frac{f_e v}{\sqrt{2}}. \tag{2.13}$$

The same mechanism generates the masses of the other lepton generations and the quarks. For all fermions, the masses are proportional to the VEV of the Higgs field. The value of $v$ can be determined from the Fermi coupling $G_F$ to be

$$v = \sqrt{2} G_F^{-1/2} = 246.22\,\text{GeV}. \tag{2.14}$$

## The Higgs Boson

One consequence of the Brout-Englert-Higgs mechanism is the existence of a neutral Higgs boson, which was discovered by the ATLAS [6] and CMS [7] Collaborations in 2012. The dominant mechanisms for the production of the SM Higgs boson in proton-proton (pp) collisions are listed in Table 2.1, and a selection of leading order Feynman diagrams is shown in Figure 2.2. The most dominant production mechanism at the LHC is gluon fusion (ggH), followed by vector-boson fusion (VBF) and Higgs-strahlung (VH). In most analyses, the contribution vom VBF and V(qq)H where the vector boson decays hadronically are combined and denoted as qqH, since the two production modes are indistinguishable.

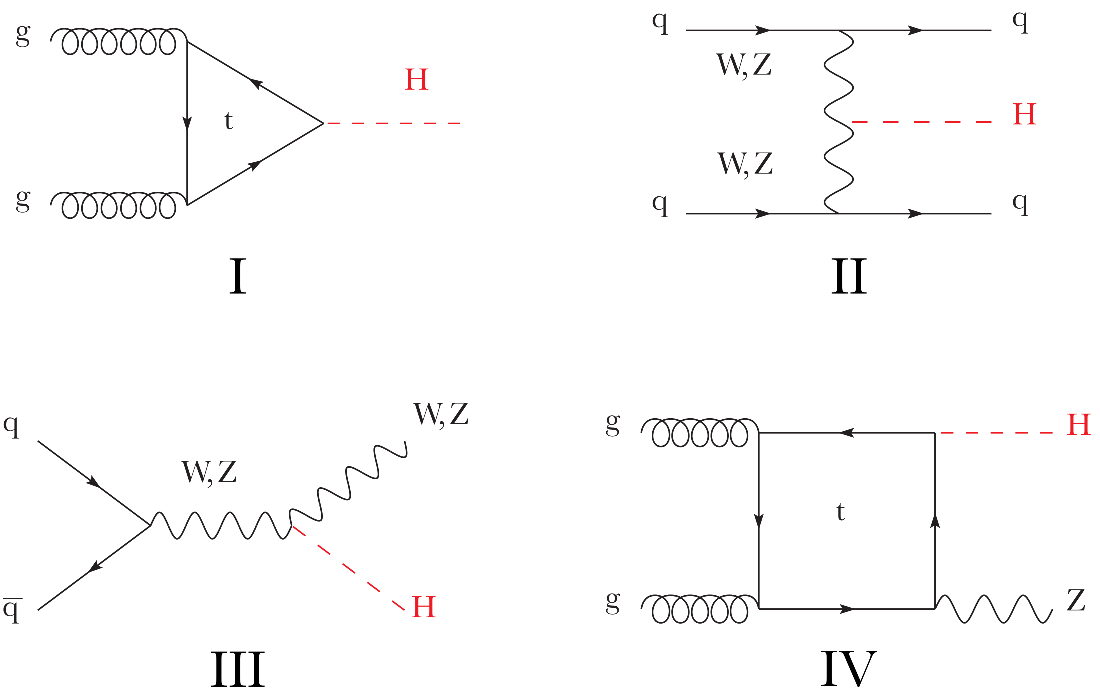


**Figure 2.2.:** An overview of the dominant Higgs boson production modes at the LHC, which are gluon fusion (ggH, I), vector boson fusion (VBF, II), Higgs-strahlung and associated production with a gauge boson (WH, ZW, III & IV). Taken from [24].

The SM Higgs boson can decay into any massive particle of the SM; however, the coupling to fermions is proportional to the fermion mass, while the coupling to vector bosons is proportional to the mass squared

$$g_{\mathrm{H}ff} = \frac{m_f}{v}, \qquad g_{\mathrm{H}VV} = \frac{2m_V^2}{v}. \tag{2.15}$$

Due to the observed mass of $m_{\mathrm{H}} = (125.38 \pm 0.14)\,\mathrm{GeV}$ [25], the decay to two vector bosons is only possible with one off-shell boson, reducing the branching fraction of the WW and ZZ decay. Therefore, the decay to a b quark pair is the dominant SM Higgs boson decay at the LHC. The branching fractions for all final states are depicted in Figure 2.3. For the discovery in 2012, the contributing channels were $\mathrm{H} \rightarrow \gamma\gamma$, $\mathrm{H} \rightarrow \mathrm{ZZ}^* \rightarrow 4\mathrm{l}$ and later also $\mathrm{H} \rightarrow \mathrm{WW}^* \rightarrow \mathrm{l}\nu\mathrm{l}\nu$. Despite the small branching fraction of less than 1%, the $\mathrm{H} \rightarrow \gamma\gamma$ channel [26], provides a very clean experimental signature, forming a resonance in the di-$\gamma$ spectrum. The same is true for the $\mathrm{H} \rightarrow \mathrm{ZZ}^* \rightarrow 4\mathrm{l}$ channel [27], where the four leptons in the final state allow for full reconstruction of the resonance. The $\mathrm{H} \rightarrow \mathrm{WW}^* \rightarrow \mathrm{l}\nu\mathrm{l}\nu$ channel [28] has a larger branching fraction but suffers from larger background contributions from $\mathrm{t\bar{t}}$ production and nonresonant production of two W bosons.

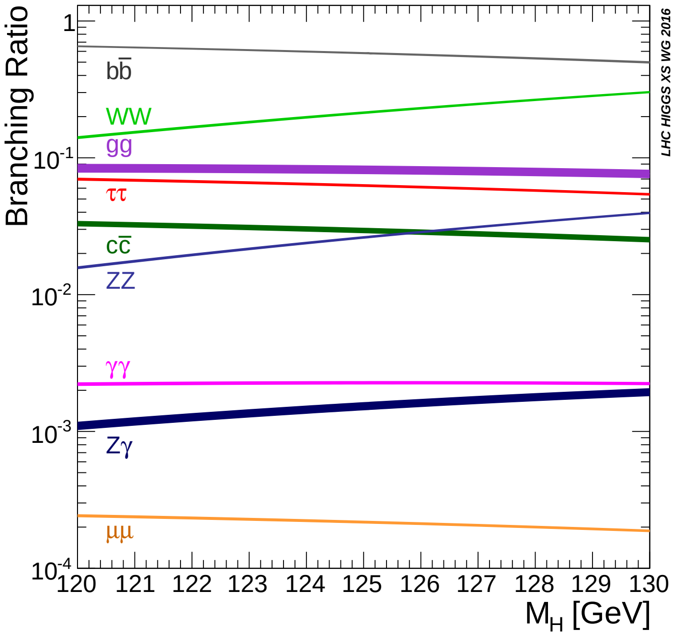


**Figure 2.3.:** The branching fractions of the SM Higgs boson final states as a function of the Higgs boson mass. Taken from [35].

During LHC Run II, the SM Higgs boson couplings to the 3rd generation fermions, to b quarks [29], to top quarks via the $t\bar{t}H$ production [30] and τ leptons [31, 32] were observed. In the $H \rightarrow b\bar{b}$ channel, a challenge is posed by discriminating the hadronized b quark decays from the SM Higgs decay and background processes. The $H \rightarrow \tau\tau$ channel has a smaller branching fraction but is experimentally more accessible, due to the leptonic decays of τ leptons. The leptonic decays represent a cleaner experimental signature. The b quarks and the τ leptons final state provide direct access to the Yukawa coupling of the SM Higgs boson. Measurements in the $H \rightarrow \tau\tau$ channel are the main motivation for this thesis. In Section 4.1, a selection of recent Run II Higgs boson analyses in di-τ final states from the CMS collaboration are outlined. In 2021, the evidence of the SM Higgs boson decay to a pair of muons was reported by the CMS collaboration [33].

Since the discovery, numerous measurements in more differential phase spaces have been performed. In addition, the measurements from multiple channels are combined to provide a more precise measurement of the SM Higgs boson properties. The most recent measurements by the CMS collaboration are summarized in [34].

## The Tau Lepton

With a mass $m_\tau = 1.776$ GeV and a lifetime of $\tau_\tau = 290.3$ fs [24] the τ lepton is the heaviest of the three known leptons. The τ lepton always decays via the exchange of a W boson and has both leptonic and hadronic decay modes. The two possible Feynman graphs for the τ lepton decays are shown in Figure 2.4. The branching fractions for the most important decay modes are listed in Table 2.2. With a probability of 35.2%, a τ lepton

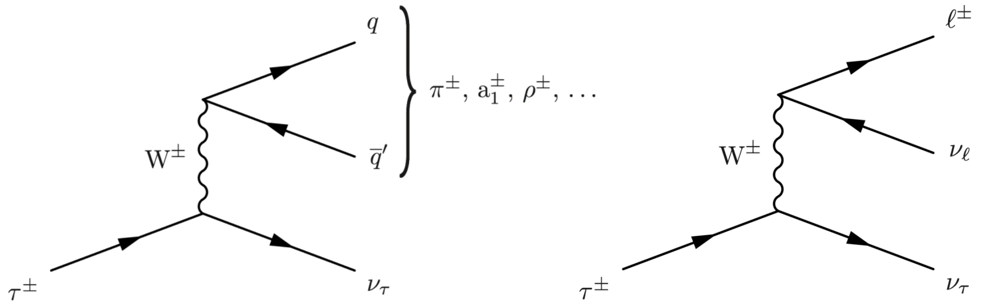


**Figure 2.4.:** Feynman graphs for the τ lepton decay modes. The hadronic decay mode is shown on the left, the leptonic decay mode on the right. Taken from [36]

**Table 2.2.:** Branching fractions for the most common τ lepton decay modes. Only the decays for a negatively charged τ lepton are shown. Charged hadrons are denoted as $h^{\pm}$. Values are taken from [24]. The $\tau_h$ decay mode obtained via Equation (2.16) is shown in the last column.

| Name | decay mode | branching fraction [%] | Decay Mode (DM) |
|---|---|---|---|
| One prong | $\tau \rightarrow h^- \nu_\tau$ | 10.82 | 0 |
| One prong + $\pi^0$ | $\tau \rightarrow h^- \nu_\tau \pi^0$ | 25.49 | 1 |
| One prong + 2 $\pi^0$ | $\tau \rightarrow h^- \nu_\tau \pi^0 \pi^0$ | 9.26 | 2 |
| One prong + 3 $\pi^0$ | $\tau \rightarrow h^- \nu_\tau \pi^0 \pi^0 \pi^0$ | 1.04 | 3 |
| Three prong | $\tau \rightarrow h^+ h^- h^- \nu_\tau$ | 8.99 | 10 |
| Three prong + $\pi^0$ | $\tau \rightarrow h^+ h^- h^- \nu_\tau \pi^0$ | 2.74 | 11 |
| Electron | $\tau \rightarrow \nu_\tau \nu_e e$ | 17.82 | - |
| Muon | $\tau \rightarrow \nu_\tau \nu_\mu \mu$ | 17.39 | - |

decays leptonically, while the rest of the possible decays are hadronic, mainly into charged Pions and Kaons and potentially additional neutral Pions. The hadronic decay modes of the τ lepton are denoted as $\tau_h$. The decay mode (DM) (see Table 2.2) of a $\tau_h$ will further on be associated with an integer number according to

$$\mathrm{DM} = 5 \cdot (N_{\mathrm{chr}} - 1) + N_{\pi^0} \tag{2.16}$$

where $N_{\mathrm{chr}}$ is the number of charged hadrons (prongs) in the hadronic decay and $N_{\pi^0}$ is the number of $\pi^0$ in the hadronic decay.

For Higgs analyses, final states with two τ leptons are of interest. Combining the τ lepton decay modes from Table 2.2, a total of six di-τ final states emerge, as shown in Figure 2.5. The two fully leptonic final states ττ → ee and ττ → μμ are neglected in most analyses since a large irreducible background from Z → ee and Z → μμ decays is present in these final states.

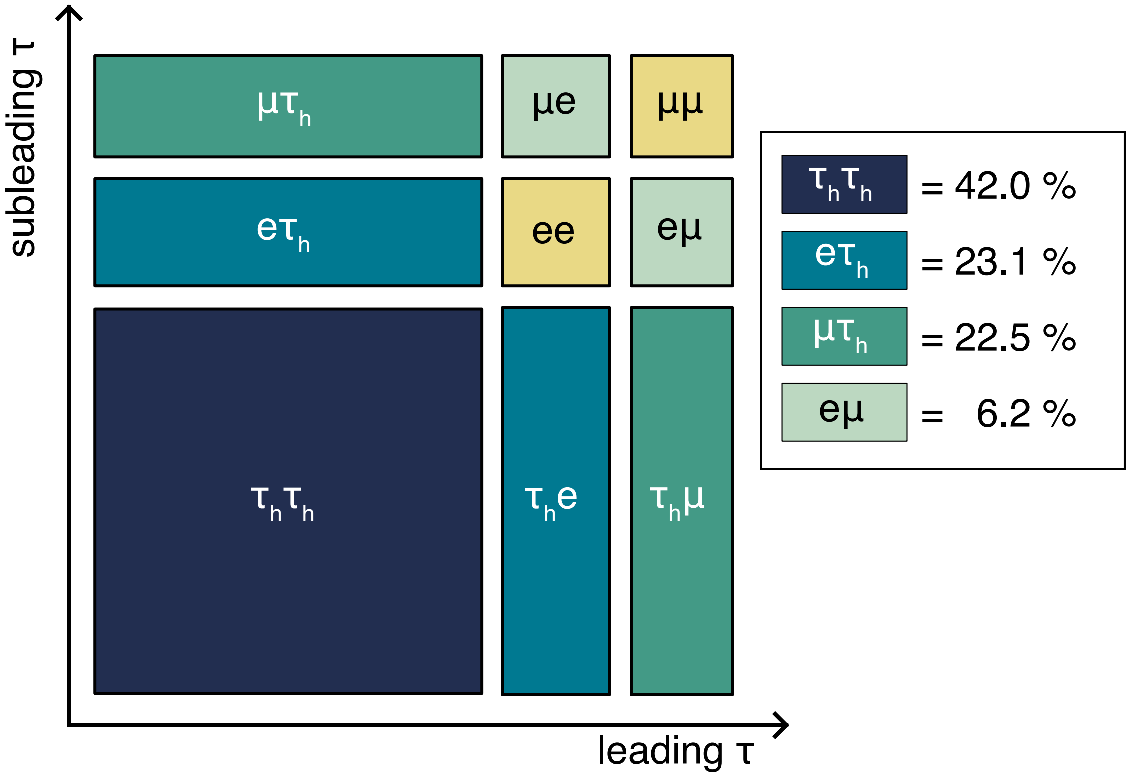


**Figure 2.5.:** Sketch of the possible di-τ final states and their branching fractions. The first letter corresponds to the decay mode of the first τ lepton, the second letter to the decay mode of the second τ lepton.

# 3 | The CMS Experiment

The CMS detector is a multi-purpose particle detector located at the European Organization for Nuclear Research (CERN) site in Geneva, Switzerland. The CMS detector was built and is operated by the CMS Collaboration, a multi-national organization with more than 3000 members. A set of accelerators is operated at CERN, of which the LHC is the largest and most powerful. In this chapter, the accelerator complex at CERN, as well as the CMS detector, are described. Furthermore, the reconstruction algorithms used for event reconstruction the CMS detector are introduced.

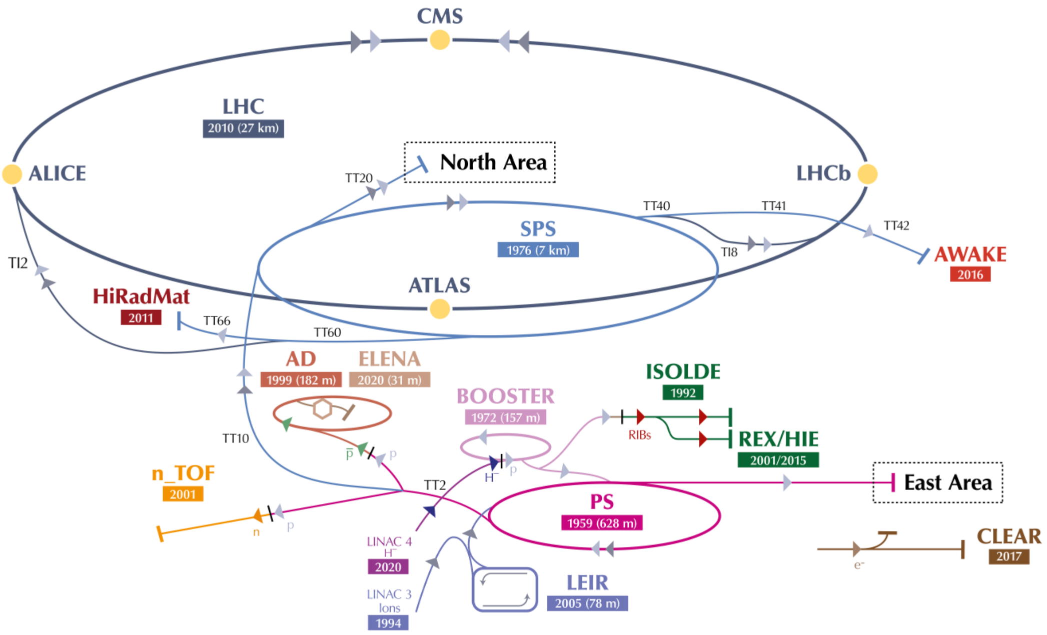


**Figure 3.1.:** A sketch of the accelerator complex at CERN, adapted from [37].

## 3.1. The Large Hadron Collider

A sketch of the complete CERN accelerator complex is shown in Figure 3.1. The LHC is a hadron collider with a total length of 27 km and has been in operation since 2010. The primary purpose of the LHC is to accelerate proton beams, but other hadrons, such as ion nuclei, can also be accelerated. The ring consists of two beam pipes with hadrons accelerated in opposite directions. The hadrons are grouped into bunches, small packets of hadrons. During the Run II measurement period between 2016 and 2018, a single bunch consisted of $\sim 1.15 \times 10^{11}$ protons. In 2017, the number of bunches increased from 2244 to 2556 bunches per beam. During Run II, the LHC was operated at a center-of-mass energy of 13 TeV, 6.5 TeV per proton beam. The bunch spacing was set to 25 ns, resulting in a collision rate of 40 MHz.

Crossing points, where the two particle beams cross each other, are located at four locations along the two beam lines of the LHC. These crossing points are surrounded by four large experiments, the CMS experiment [5], the A Toroidal LHC Apparatus (ATLAS) experiment [4], the A Large Ion Collider Experiment (ALICE) [38] experiment, and the LHC-beauty (LHCb) [39] experiment. The CMS and ATLAS experiments are two multi-purpose detectors designed to study a wide range of different physics topics, while the latter two have a more specialized physics program. The LHCb detector is used to study processes involving b quarks with great precision, while the ALICE Collaboration focuses on studies of heavy ion collisions.

To measure the cross section $\sigma$ of a physics process, the *integrated luminosity* of the accelerator has to be known. The number of observed occurrences of a physics process, also called events $N$, is connected to the cross section via

$$N = \sigma \cdot L_{\text{int}}, \tag{3.1}$$

where the integrated luminosity $L_{\text{int}}$ is defined as the integral of the *instantaneous luminosity*

$$L_{\text{int}} = \int \mathcal{L}(t) dt. \tag{3.2}$$

The integrated luminosity is a measure for the size of the collected data of the experiment and is usually reported in units of inverse femtobarn ($1\text{fb}^{-1} = 10 \times 10^{-43}\,\text{m}^{-2}$). The instantaneous luminosity of a collider with bunches of particles is defined as

$$\mathcal{L}(t) = f_{\text{coll}} \frac{n_1 n_2}{4\pi\sigma_x\sigma_y}, \tag{3.3}$$

where $f_{\text{coll}}$ is the average collision frequency, $n_1$ and $n_2$ are the number of particles in the two bunches, and $\sigma_x$ and $\sigma_y$ are the spread of the bunches in the x- and y-direction respectively.

Since the instantaneous luminosity is directly connected to the cross section of a process, the value has to be measured with great precision [40–42]. In the CMS experiment, several different measurements are combined to achieve a luminosity uncertainty of 1.6% for the combined Run II data set. The integrated luminosities of the different eras of the Run II data set are listed in Table 3.1. In Figure 3.2, collection of data over time is visualized.

At the beginning of 2022, the Run III data taking of the LHC was started, using similar collider conditions as the Run II data taking. The centre-of-mass energy was increased to 13.6 TeV. Run III is expected to last till 2025 to collect a total of $L_{\text{int}} = 250\,\text{fb}^{-1}$ of proton-proton collisions [44]. After another long shutdown phase of three years, during which the collider and the four experiments will undergo significant upgrades, the High Luminosity LHC (HL-LHC) will be operational, which is expected to provide a total of $L_{\text{int}} = 3000\,\text{fb}^{-1}$ of proton-proton collisions till 2037.

**Table 3.1.:** Integrated luminosity of proton-proton collisions of the Run II data set, split by measurement era.

| Era | Integrated Luminosity $L_{\mathrm{int}}$ [fb$^{-1}$] |
|---|---|
| 2016 | 36.33 |
| 2017 | 41.48 |
| 2018 | 59.83 |
| Total Run II | 137.65 |

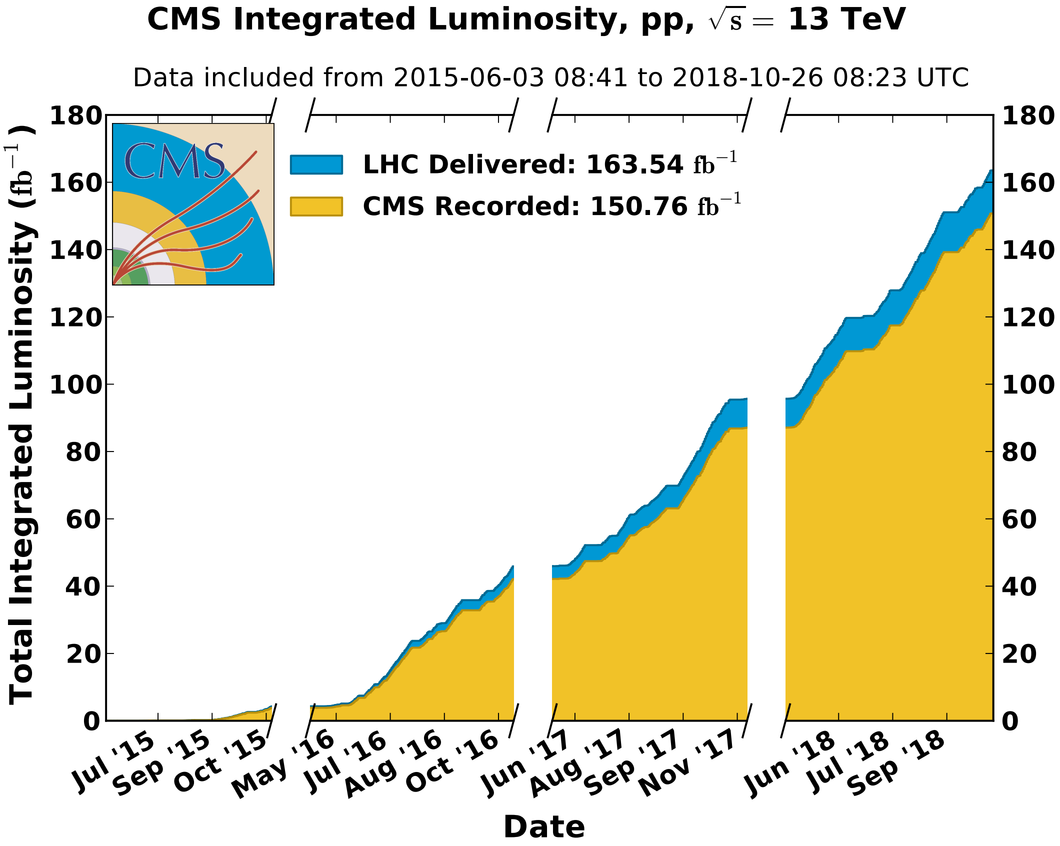


**Figure 3.2.:** Integrated luminosity collected by the CMS experiment between 2015 and 2018. In the Run II data set used within this thesis, the data collected in 2015 is omitted since the magnet of the CMS detector was turned off. Taken from [43].

## 3.2. The CMS Detector

The CMS detector is located at one of the crossing points of the LHC. The detector is designed as a multi-purpose detector to study different physics topics, ranging from precision measurements of the SM to searches for Dark Matter particles and studies of heavy ion collisions. The detector has a total length of 21 m and a diameter of 15 m. Right in the centre of the detector, the crossing point of the two LHC beam pipes is located, also called the interaction point. With its weight of about 14.000 t, the CMS detector is densely packed which is the reason for ”Compact” being in the name of the detector. A sketch of the whole detector is shown in Figure 3.3.

The detector consists of several subdetectors located around the interaction point in a cylindrical shape. To increase the instrumented area of the detector, two endcaps are added to the cylinders, forming the densely packed detector.

The innermost subdetector is the silicon tracker, which measures the trajectories of charged particles traversing outwards from the interaction point. Around the silicon tracker, the Electromagnetic Calorimeter (ECAL) is located. The purpose of the ECAL is to measure the energy of electrons, positrons and photons while stopping them completely. The next subdetector outwards of the ECAL is the Hadronic Calorimeter (HCAL). The HCAL is targeted towards measuring the energy of hadrons and stopping them completely. All those subdetectors are located in a homogeneous magnetic field of 3.8 T, which is generated by a superconducting solenoid which has an inner diameter of 6 m. Within the return yoke of the solenoid, the muon system is located. It measures the trajectory of muons, the only particles that can traverse the other subdetectors without being stopped. A full description of the detector can be found in [5]. The following will describe the different subdetectors in more detail.

### 3.2.1. CMS Coordinate System

By convention, the right-handed coordinate system used for the CMS detector originates at the interaction point. The x-axis is pointed at the centre of the LHC, while the z-axis is pointed along the beam line. The y-axis is perpendicular to the x- and z-axis. A visualization of the coordinate system is shown in Figure 3.4.

Since the protons used in the collisions are made up of quarks and gluons, the composite partons’ momenta are unknown. For this reason, only the *transverse momentum* of decaying particles is measured, which is defined as

$$p_{\mathrm{T}} = \sqrt{p_{\mathrm{x}}^2 + p_{\mathrm{y}}^2}. \tag{3.4}$$

By definition, the $p_{\mathrm{T}}$ of a particle is independent of the unmeasurable $p_{\mathrm{z}}$ component of the total particle momentum. To describe particle trajectories within the CMS detector more easily, a cylindrical coordinate system with the polar angle $\theta$, the azimuth angle $\phi$, and the axial coordinate $z$ is used, where $z$ is identical to the z-axis of the Cartesian coordinate system. The polar angle $\theta$ can be used to calculate the *pseudorapidity*

$$\eta = -\ln\left(\tan\frac{\theta}{2}\right), \tag{3.5}$$

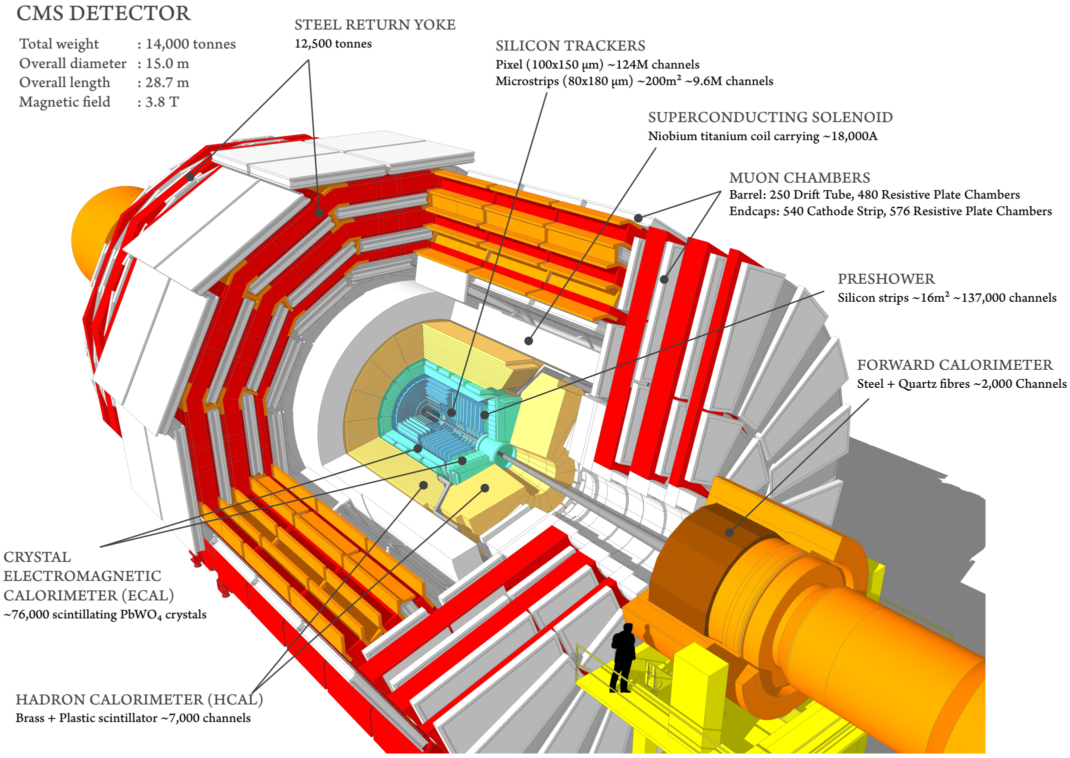

CMS DETECTOR
Total weight : 14,000 tonnes
Overall diameter : 15.0 m
Overall length : 28.7 m
Magnetic field : 3.8 T
STEEL RETURN YOKE
12,500 tonnes
SILICON TRACKERS
Pixel (100x150 μm) ~124M channels
Microstrips (80x180 μm) ~200m² ~9.6M channels
SUPERCONDUCTING SOLENOID
Niobium titanium coil carrying ~18,000A
MUON CHAMBERS
Barrel: 250 Drift Tube, 480 Resistive Plate Chambers
Endcaps: 540 Cathode Strip, 576 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 sketch of the CMS detector and its different subdetectors. The Sketch is adapted from [45].

which is an approximation of the *rapidity*

$$y = \ln\left(\frac{E + p_z c}{\sqrt{m^2 c^4 + p_\mathrm{T}^2 c^2}}\right) \tag{3.6}$$

of a particle, where E is the particle's energy, $m$ is the particle's mass, and $p_\mathrm{T}$ is the particle's transverse momentum. For massless particles, the rapidity is equal to the pseudorapidity. The difference in rapidity between two particles is invariant under Lorentz boosts along the z-axis. At the LHC, all particles are considered highly relativistic ($m \ll p$), which means Equation (3.5) is a valid approximation of the true rapidity, again eliminating the dependence on $p_\mathrm{z}$.

To quantify the spatial distance between two particles, their distance in the η-ϕ plane is used. It is defined as

$$\Delta\mathrm{R} = \sqrt{(\eta_1 - \eta_2)^2 + (\phi_1 - \phi_2)^2}, \tag{3.7}$$

where $\eta_1$ and $\phi_1$ are the coordinates of the first particle and $\eta_2$ and $\phi_2$ are the coordinates of the second particle.

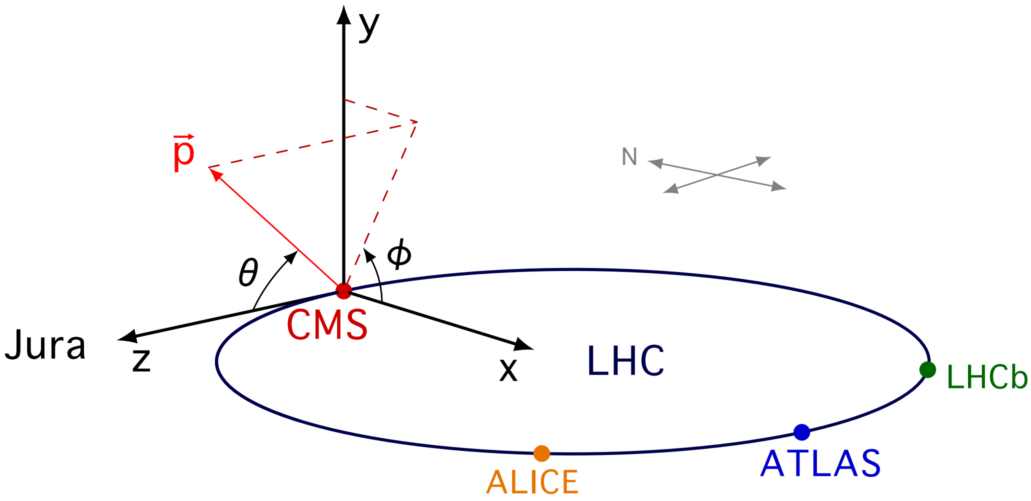


**Figure 3.4.:** An illustration of the CMS coordinate system, adapted from [46].

### 3.2.2. Inner Tracker

The inner tracker is located directly around the interaction point. A full description of its design is given in [47, 48]. The inner tracker itself is split into multiple smaller sections. Two silicon detectors are used within the inner tracker: silicon pixel detectors and silicon strip detectors. The CMS inner tracker is sketched in Figure 3.5.

The silicon pixel detector is located closest to the interaction point. At the beginning of the Run II measurement period, the silicon pixel detector consisted of three barrel layers and two endcap disk layers. The new beam pipe was installed during a technical stop at the end of 2016 and the beginning of 2017. Since the new beam pipe had a smaller diameter, an upgrade of the pixel detector was possible and allowed for a new design with four barrel layers and three endcap disk layers [49]. Within the four barrel layers and the three endcap disks, a total of 1856 sensor modules, each consisting of a silicon pixel sensor with 160 x 416 pixels, are installed. A single pixel has a size of 100 x 150 $\mu m^2$. In total, the pixel detector has 124 million readout channels. If a charged particle passes through a pixel sensor, a charge is deposited in the semiconducting material, which is then registered by a readout chip. The pixel detector has a full 4-layer coverage up to a pseudorapidity of $|\eta| > 2.5$ and allows for a spatial resolution of up to 15 μm.

The silicon strip detector is located outside of the silicon pixel detector. The innermost strip detector is the Tracker Inner Barrel (TIB), consisting of four layers of silicon strip detectors. Two sets of Tracker Inner Disk (TID)s, one for each direction along the beam pipe, consisting of three layers, are installed to cover the endcap regions. Those two strip detectors are surrounded by the Tracker Outer Barrel (TOB), which consists of 6 layers of silicon strip sensors. For the increased coverage in the z-direction, two Tracker Endcap (TEC) are installed, each consisting of a total of 9 disks. In total, the silicon strip detector has 9.3 million readout channels.

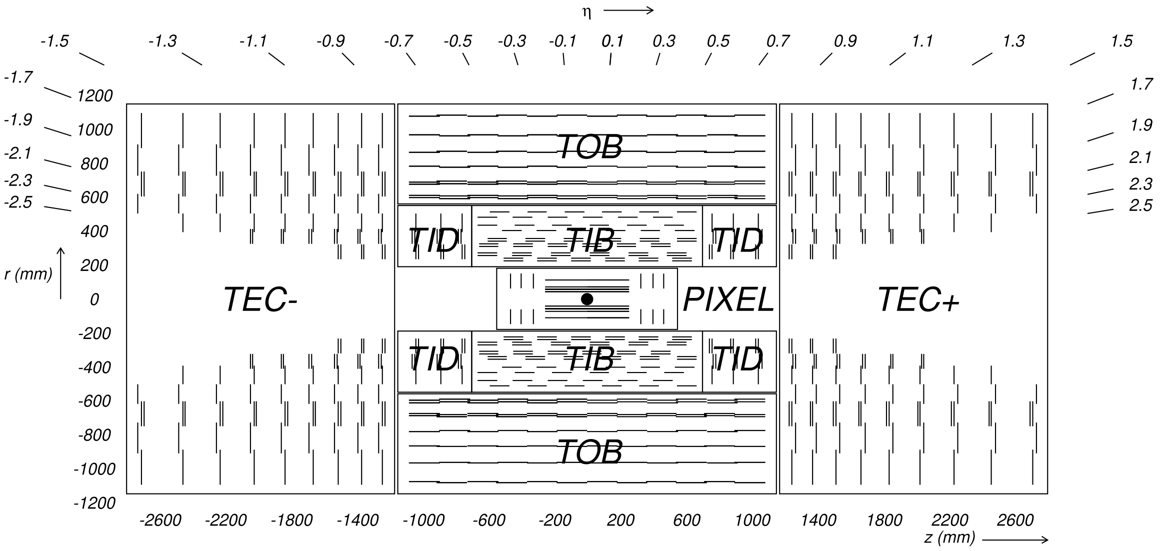


**Figure 3.5.:** A 2D sketch of the inner tracker of the CMS Experiment in the $x$-$z$ plane, taken from [5].

### 3.2.3. Electromagnetic Calorimeter

The Electromagnetic Calorimeter is used to measure the energy of electrons, positrons and photons. A full description of the ECAL is given in [50]. The ECAL is the subdetector surrounding the inner tracker and is made of 61200 lead tungstate ($PbWO_4$) crystals in the Barrel ECAL (EB) and 14648 lead tungstate crystals distributed over the two Endcap ECAL (EE) disks. Lead tungstate is a dense material with a *radiation length* $X_0$ of 7.39 g cm$^{-2}$ [51]. The radiation length $X_0$ is the mean distance a high-energy electron can propagate through a material while emitting $1/e$ of its initial energy via bremsstrahlung. The radiation length of $PbWO_4$ corresponds to a distance of 0.89 cm.

The region up to a pseudorapidity of up to $|\eta| < 1.479$ is covered by the EB of the ECAL, while the region $1.479 < |\eta| < 3.0$ is covered by the two EE disks. The Preshower (ES) detector ranges from $1.653 < |\eta| < 2.6$. Due to the placement of readout electronics, a small area between the barrel and the endcap $1.479 < |\eta| < 1.56$, has a suboptimal coverage compared to the rest of the detector. A sketch of the ECAL cross section is shown in Figure 3.6.

Since lead tungstate crystals can be used as a scintillator, the ECAL is a homogeneous calorimeter without any additional absorber material. Each crystal in the barrel region has a tapered shape, with a size of 22 x 22 $\mu m^2$ pointing towards the beam pipe and a size of 26 x 26 $\mu m^2$ on the rear end. The total length of a single crystal is 23 cm, which corresponds to a size of 25.8 $X_0$ and has a weight of around 1.5 kg. The use of the lead tungstate crystals allows for a very dense and compact but also weighty design of the ECAL. The crystals used on the endcap regions have a slightly larger front side while being shorter. In front of the endcap disks, a small preshower detector is located. It consists of two layers of absorber material (lead) with a total thickness of 3 $X_0$, each followed by a single layer of silicon strip sensors to increase the spatial granularity in this region. The

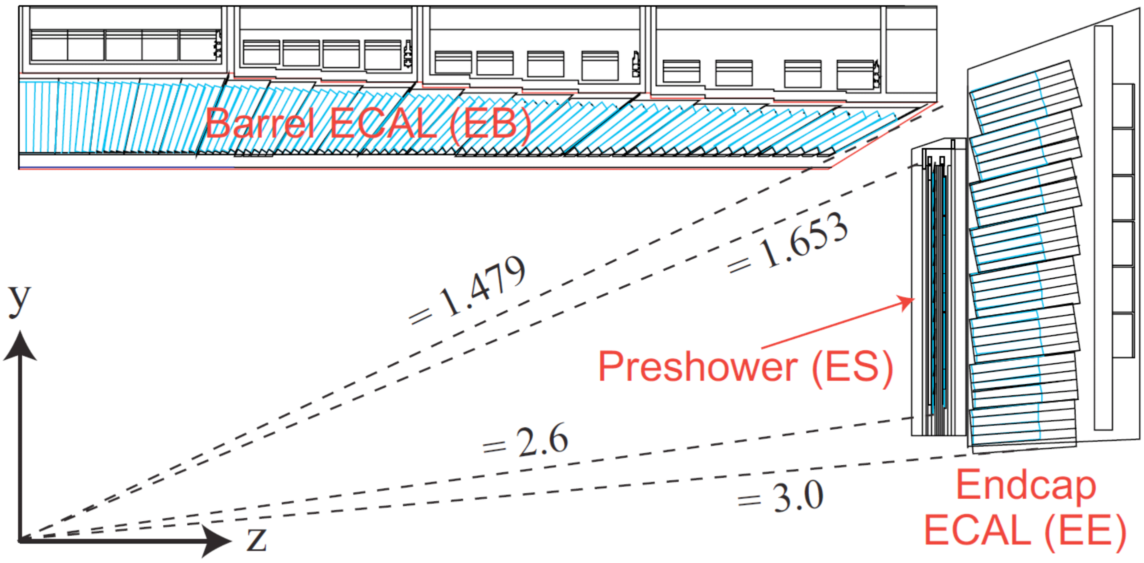


**Figure 3.6.:** A sketch of the ECAL cross section, showing the regional distribution of the EB, EE, and ES subdetectors. The sketch is taken from [52].

preshower detector makes it possible to distinguish between the decay of a low energy $\pi^0$ and a single high energy $\gamma$ in the forward region.

### 3.2.4. Hadronic Calorimeter

The HCAL is a sampling calorimeter consisting of alternating layers of absorber and scintillator material. A full description of the HCAL is given in [53]. In the CMS detector, brass is used as absorber material, while the scintillator is made from plastic. The HCAL is again split into four smaller subdetectors, the Hadron Barrel (HB) located in the barrel region, the Hadron Endcap (HE) located in the endcaps, the Hadron Forward (HF) located in the forward region, and the Hadron Outer (HO) calorimeters located just outside the solenoid. A schematic sketch of the HCAL is shown in Figure 3.7.

The size of the HCAL is defined by its length in nuclear interaction lengths $\lambda_I$. The nuclear interaction length is defined as the average distance a hadron can traverse before undergoing an interaction with the material. Typically, $\lambda_I$ is much larger than $X_0$. For the HCAL configuration of the CMS detector, $\lambda_I$ corresponds to a distance of 16.42 cm [54].

The HB consists of about 40000 plastic scintillator tiles, with a thickness of 9 mm, wedged between a steel plate at the beginning of the HCAL followed by brass absorber plates with a thickness of 50.5 mm in the first eight proceeding layers and 56.5 mm brass absorber plates in the remaining six layers. This corresponds to a size of 5.82 $\lambda_I$ for $|\eta| = 0.0$ and a size of 10.6 $\lambda_I$ for $|\eta| = 1.3$. Along the z-Axis, the HB is split into 32 sections, each corresponding to a width of $\Delta\eta = 0.087$. One slice of the HCAL section is also called an HCAL tower. The HE in the two endcap regions has a similar setup, though slightly different margins on the scintillator and absorber plates. Like the ECAL, the barrel and the endcap region cover a pseudorapidity up to $|\eta| < 3.00$. Since the HB is only $\sim 6$ $\lambda_I$ thick in the central region of the detector, an additional energy measurement outside the solenoid is beneficial. The HO

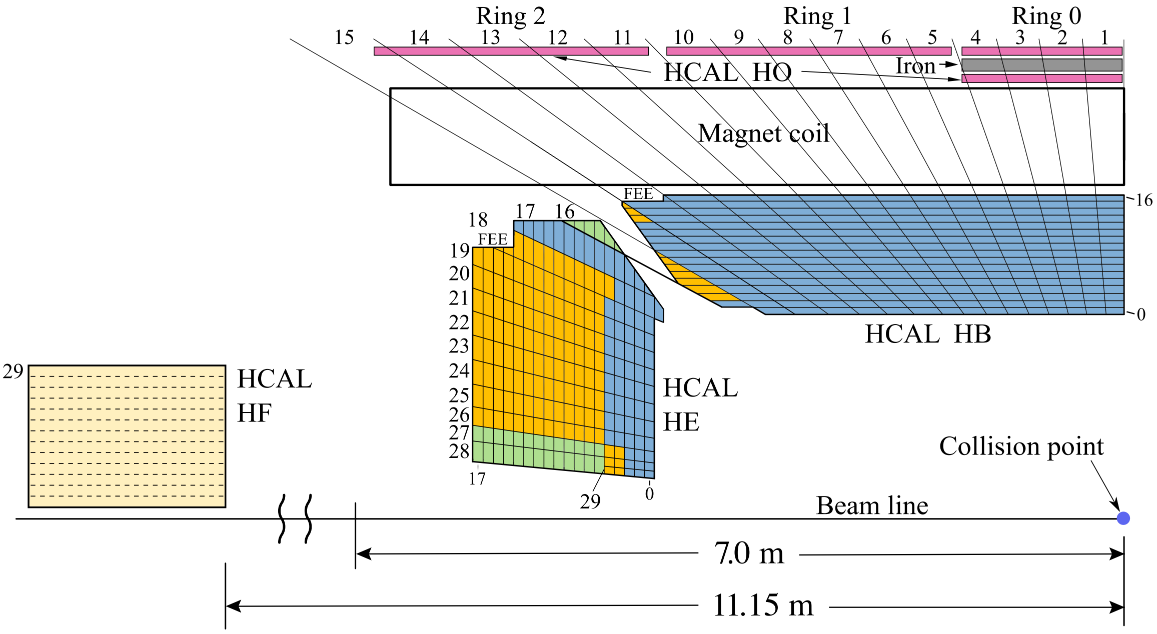


**Figure 3.7.:** A schematic sketch of the HCAL, showing the regional distribution of the HB, HE, HF and HO subdetectors. The figure is taken from [55].

serves this purpose and is used to detect and measure the energies of particles with large $p_T$ that were not entirely stopped by the calorimeters. The solenoid serves as an additional layer of absorber material in front of the HO. The HO consists of up to two layers of brass absorber covers up to a pseudorapidity of $|\eta| < 1.26$. Finally, the HF calorimeter is located 11.15 m in both directions from the interaction point, along the beam pipe. It covers an extended pseudorapidity range of $2.85 < |\eta| < 5.19$ and is made from steel embedded with quartz fibres.

### 3.2.5. Magnet

The superconducting solenoid embedded in the CMS detector is used to generate a homogeneous magnetic field of 3.8 T. The magnetic field is required for measuring particle trajectories since, due to the Lorentz force, charged particles have a bent trajectory in a magnetic field. The direction of the curvature of the particle trajectory is dependent on the particle's charge, and the bending is proportional to the particle's momentum. Therefore, the inner tracker, the ECAL and most parts of the HCAL are located within the solenoid. The solenoid has a diameter of 6.3 m, a length of 12.5 m, weights 220 t, and is operated at a temperature of 4.65 K. The return yoke used to guide the magnetic field outside of the magnet is made from iron and houses the muon system. A magnetic field with a strength of $\sim 2$ T is generated within the return yoke and is used for the track reconstruction in the muon system.

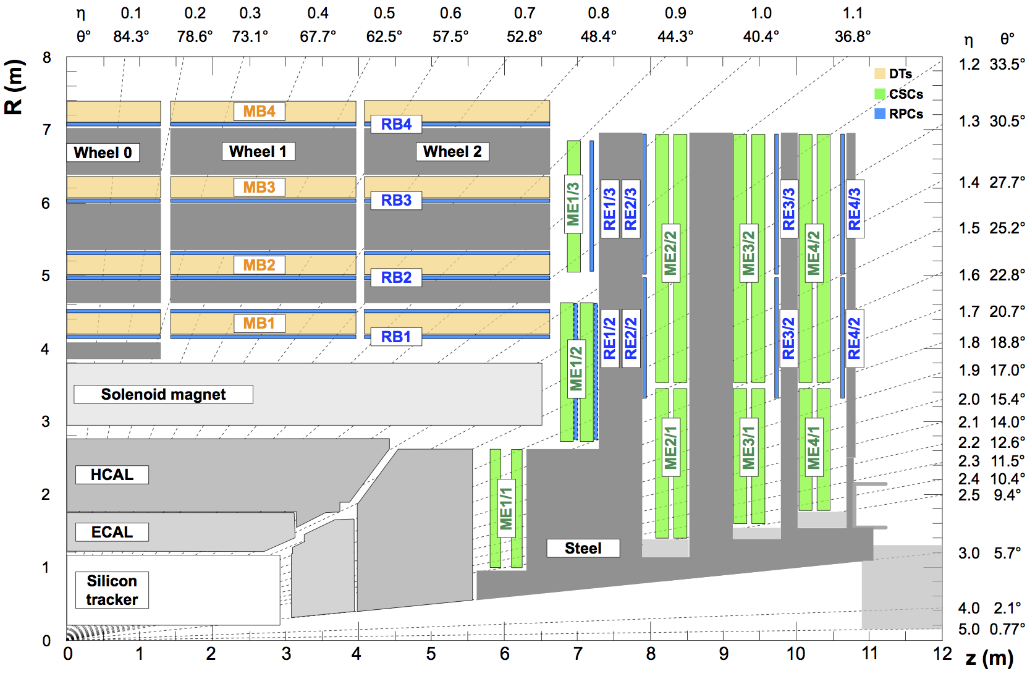


**Figure 3.8.:** A schematic sketch of the CMS muon system. The DTs located in the central region are shown in orange, the CSC of the endcap region is shown in green, and the RPCs located in the gap region are shown in blue. The figure is taken from [56].

### 3.2.6. Muon System

The muon chambers are located outside the solenoid and are partly housed within the return yoke of the magnet. The muon system is made up of three different types of gaseous particle detectors. The purpose of the muon system is to identify muons, measure muons' momentum, and provide a robust solution for triggering events with muons. A full description of the muon system is given in [57]. A schematic sketch of the muon system is shown in Figure 3.8.

The drift tube chambers (DT)s are located in the barrel region of the muon system and cover a pseudorapidity up to $|\eta| < 1.2$. Within a single DT, an anode wire is used to measure the drift time of electrons. Those electrons are generated when a muon traverses the chamber's gas, ionising it. Combining the information of when the muon initially entered the tube with the time when a signal was registered at the cathode wire and the position makes a precise 3D measurement of the position possible. The DT system has 172000 readout channels.

The cathode strip chambers (CSC) located in the two endcaps over a pseudorapidity range of $0.9 < |\eta| < 2.4$ are filled with many anode wires to measure the position of muons. Electrons from the ionization of the gas are measured at the anode wires, but instead of using the timing information, multiple wires are used to perform the 3D measurement. The CSC system has a total of $\sim 470000$ readout channels.

The resistive plate chambers (RPC) located at both the barrel and the endcaps cover a pseudorapidity range of up to $|\eta| < 1.9$. Each RPC consist of a gas-filled chamber with two oppositely charged plates. Their main purpose is to provide a fast measurement of a muon and good timing information, which is required for an accurate triggering of muon events. The RPC system has a total of $\sim 123000$ readout channels.

## 3.3. Reconstruction Algorithms

To convert the electric signals from the different parts of the detector into physics objects, different reconstruction algorithms, depending on the type of particle, are used. In the following sections, the reconstruction of particle tracks (Section 3.3.1) and calorimeter clusters (Section 3.3.3) is described. The information of the different subdetectors is then combined using the Particle Flow (PF) algorithm (Section 3.3.4), and based on the PF-candidates, physics objects like electrons, muons, jets, and $\tau$ leptons are reconstructed.

### 3.3.1. Track and Vertex Reconstruction

Since the inner tracker is located within a magnetic field, charged particles are bent on a helix-shaped trajectory. This trajectory of charged particles propagating from the interaction point through the detector is described by a so-called track. A track is reconstructed from a set of multiple tracker energy deposits from charged particles, called hits. A single hit is registered if, within the corresponding pixel, a minimal charge of 3000 - 4000 electrons is registered by the readout chip.

The CMS Collaboration uses the Combinatorial Track Finder (CTF) algorithm described in [58, 59] for the track reconstruction, which is an iterative algorithm based on the combinatorial Kalman Filter [60]. Since a single collision can contain up to 1000 particles, the track reconstruction is challenging due to the high combinatorial complexity. By using the iterative ansatz, tracks that are easy to identify are reconstructed first, therefore reducing the combinatorial complexity in subsequent iterations. If a hit is associated with a track, the hit is removed from the set of available hits for the next iteration.

Within a single iteration, a set of steps is performed by the algorithm:

- A track seed is generated. A seed consists of two to three hits and is used to set coarse starting parameters of trajectory.
- Using a Kalman Filter approach, an additional hit is added to the track seed, and the track parameters are updated if the additional hit can be found within the boundaries set by the track seed. This step is repeated multiple times unit no more matching hits are found or until the furthest layer of the tracker is reached.
- After all feasible hits corresponding to a track seed were found, the trajectory is refitted, using the complete information of all selected hits. In addition, smoothing using a Runge–Kutta propagator is performed. During the smoothing, effects from material interactions and possible inhomogeneities of the magnetic field, which can result in a deviation from the helix trajectory, are considered.

- The quality of the fit is checked against a set of defined quality thresholds. If these quality criteria are not fulfilled, the track is discarded. Typical criteria are the number of hits and layers used for the reconstruction or the $\chi^2$/dof value of the track fit.

The version of the tracking algorithm deployed during Run II uses ten iterations. The initial seed setting and the quality criteria applied in the different iterations vary.

In addition to the trajectories of the particles, the origin of the particles has to be determined. During a single bunch crossing, interactions of multiple protons are very likely to happen. Collisions of other protons apart from the hard process are called Pileup Collisions (PU) collisions. To isolate the hard process from PU collisions, it is necessary to determine the origin of every single particle in the event with high precision. The origin of the particles from the hard process is called primary vertex (PV).

The reconstruction of all vertices is possible once all tracks in the event have been reconstructed. The vertex algorithm consists of three steps:

- Tracks that originate close to the interaction point are selected by requiring a minimal amount of two pixel layers traversed and a reasonable quality of the track fit.
- All considered tracks are clustered based on their distance to the interaction point. The challenge of this clustering is to accurately determine all vertices in the event whilst not splitting tracks from the same vertex into different clusters. The clustering is performed using an annealing algorithm.
- For each cluster containing at least two tracks, a vertex fit is performed to determine the vertex position. In addition, each track is assigned a weight, corresponding to the probability that the track originates from this vertex.

In the end, the primary vertex (PV) is the vertex that corresponds to the hardest scattering in the event based on the tracking information as described in [61].

As shown in Figure 3.9, the resolution of the PV position improves with the number of tracks considered for the vertex fit. On average, the resolution is smaller than 20 µm if at least 50 tracks are associated with a vertex.

The *impact parameters* in the transverse and the longitudinal plane can be used to quantify the compatibility of a particle with the PV. The impact parameter $d_{\mathrm{XY}}$ is defined as the distance between the PV and the closest point of the particle track in the transverse plane, whereas the impact parameter $d_{\mathrm{Z}}$ is defined as the same distance in the longitudinal plane. For a particle to be compatible with the PV, the impact parameters are typically in the order of $\mathcal{O}(1\,\mathrm{mm})$, where $d_{\mathrm{Z}}$ can be a bit larger than $d_{\mathrm{XY}}$.

**Muon Track Reconstruction** For the reconstruction of muon tracks, the track information from the muon system is combined with the tracks found in the inner tracker. A full description of the method can be found in [56, 62]. Before a matching between tracks in the inner tracker and the muon system is possible, the individual hits in the muon system have to be reconstructed. The basic principles of muon hit reconstruction are outlined in Section 3.2.6. In the second step, individual tracks are reconstructed in the muon system, using the same iterative approach as in the inner tracker. The tracks reconstructed from the muon system are called *standalone muon tracks*.

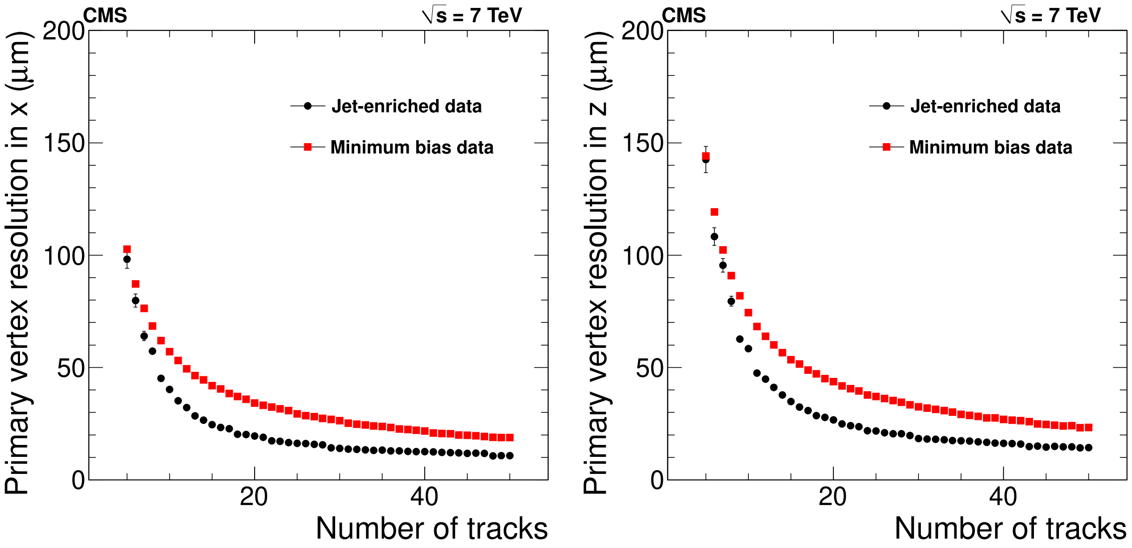


**Figure 3.9.:** The resolution of the PV in $x$- and $z$-direction depending on the number of tracks associated with the vertex. More tracks associated allow for a stronger constraint on the vertex position, resulting in an improved resolution. The figure is taken from [58].

The combination of inner tracker tracks and muon system tracks is performed using two approaches:

- **Inside-Out**: In this approach, tracks from the inner tracker are extrapolated "outwards" to the muon system. If at least one matching hit in a DT or CSC is found, the resulting track is considered a *tracker muon track*. This extrapolation is performed for all inner tracker tracks with $p_{\mathrm{T}} > 0.5\,\mathrm{GeV}$ and $p > 0.5\,\mathrm{GeV}$, where $p$ is the total momentum of the track.

- **Outside-In**: This approach is performed contrary to the Inside-Out approach by matching a track from the muon system with a track from the inner tracker. If both tracks can be propagated onto a joint surface and have a matching trajectory, the two tracks are combined and refitted. A muon with such a track is considered a *global muon track*.

Most analyses rely on global muon tracks since those have the lowest misidentification rate. Tracker muon tracks can also originate from secondary particles or decays that were not fully contained within the HCAL and therefore leak into the muon system. Standalone muon tracks can originate from decays outside of the detector like cosmic muons since they are not required to originate from any vertex of the recorded event.

**Electron Track Reconstruction** An extended version of the track reconstruction is run on top of the reconstructed tracks to identify electron tracks. A detailed description of the procedure can be found in [63, 64].

Contrary to all other charged particles, there is a high probability that photons are radiated from the electron via bremsstrahlung. This radiation can happen within the inner tracker, and since the photons from the bremsstrahlung can carry a significant portion of the electron energy, a considerable change of the trajectory curvature is possible. Bremsstrahlung photons are radiated tangentially to the electron trajectory and are more spread in the $\phi$ direction than in the $\eta$ direction.

The refitting of potential electron tracks is performed using a Gaussian Sum Filter (GSF) algorithm, as described in [65]. The GSF algorithm is advantageous compared to the Kalman Filter, as it allows for large changes in the trajectory due to a large energy loss of the electron. The GSF is based on the assumption that a gaussian mixture can model electron bremsstrahlung. Multiple hits within a layer can be used to reconstruct a single electron track. Since the GSF tracking is more computationally expensive, only a subset of potential tracks is refitted. Two types of seeds are used for the GSF track finding:

- **Tracker-driven** seeds are tracks that can be matched with a SuperCluster (SC) in the ECAL. The extrapolated ECAL impact point of the track has to lie within a small distance of the SC. The reconstruction of SCs is explained in Section 3.3.3.
- **ECAL-driven** seeds are constructed using an outside-in approach similar to global muon tracks; a trajectory originating from a reconstructed SC is extrapolated back to the interaction point. If a matching trajectory is found, it is used as a seed.

### 3.3.2. Tracker Alignment

For the track reconstruction, the actual position, orientation and potential twists with all tracker modules have to be known with a precision of 10 µm. Since it is impossible, to construct the detector with such precision, and movement of the modules can happen external effects such as temperate, positions and orientations of the modules have to be constantly measured. This process is called tracker alignment [66].

Tracking information from multiple sources, such as cosmic muons or tracks from pp collisions is used as input for the alignment procedure. The whole silicon tracker and all its components are parameterized, and then the best fit component positions are determined using a combined minimization of these $\mathcal{O}(100000)$ parameters. This procedure is repeated multiple times during the data-taking to account for long-term effects such as radiation damage or temperature. After the alignment is performed, the best estimate of all module positions is known for each measured collision.

### 3.3.3. Calorimeter Cluster Reconstruction

Calorimeter clusters are the other building blocks needed for the PF algorithm. A detailed description of the cluster reconstruction can be found in [63]. The calorimeter clusters are created by combining multiple calorimeter cells. The clustering is performed independently in the EB, the ES, the two EEs, the HB, and the two HEs. The algorithm is structured as follows:

- At first, a calorimeter cluster seed is generated. A cell is considered a cluster seed if the cell energy is larger than a given threshold and if no other cell in its surroundings contains more cell energy. The energy thresholds for a cluster seed are 230 (600) MeV in the EB (ES) and 800 (1100) MeV in the HB (HE).

- Cells adjacent to the seeds are added to the cluster if the energy measured in the cell is larger than a minimum threshold. These clusters are called *topological clusters*. The energy thresholds for a cell to be added to a cluster are 80 (300, 0.06) MeV in the EB (EE, ES) and 800 MeV in the HCAL.

- Within each topological cluster, a reconstruction is performed to identify the actual number of clusters since a topological cluster typically contains more than a single cluster, especially if the decay products are in close vicinity to each other.

To account for the bremsstrahlung from electrons, an additional algorithm combining multiple ECAL clusters to a SC is applied. To form a SC, multiple ECAL clusters spread in the $\phi$ direction can be combined to reconstruct the total energy of the electron and not lose the energy deposits of the emitted bremsstrahlung photons.

### 3.3.4. The Particle Flow Algorithm

The PF algorithm is the reconstruction algorithm used by the CMS Collaboration for the global reconstruction of all particles in an event. The algorithm is built upon combining the measurements from the different subdetectors and linking them together. A full description of the algorithm can be found in [63]. To utilise such a holistic reconstruction approach, fine granularity within all subdetectors is needed. This granularity is given for the inner tracker, the muon system and the ECAL. The HCAL has a more coarse granularity but is still sufficiently granular.

Objects, like tracks or calorimeter clusters originating from a subdetector reconstruction, are all considered as *PF elements*. A linking algorithm combines multiple PF elements into *PF blocks*. For the linking, only elements within a reasonable $\eta - \phi$ distance are considered to keep the algorithm's runtime low. Multiple types of links are possible:

- **Track - Cluster** The track is extrapolated from its last hit onto the calorimeter surface and linked to a calorimeter cluster if the extrapolated hit lies within the cluster surface.

- **Cluster - Cluster** This linking can be used to create a link between an ES Cluster and an ECAL cluster or between an ECAL cluster and an HCAL cluster. A link is established if the cluster position of the more granular calorimeter is compatible with the cluster position of the more coarse calorimeter. If they overlap, a link between an ECAL cluster and an SC is made.

- **Track - Track** A link between two tracks can be created if one track originates from a secondary displaced vertex, and the other track is connected to the secondary vertex and the PV of the event.

- **Track - Muon Track** A link between a track from the inner tracker and a track in the muon system can be established if a single track fit containing both can be performed.

After the linking, each PF block typically contains multiple PF elements originating from a single or a few particles. For every single PF block, the same particle reconstruction steps are performed. If one or multiple PF elements match the requirements for a given particle, the elements are removed from the PF block. The different types of particles are expected to leave different signatures in the detector, as illustrated in Figure 3.10. The different steps of the PF algorithm are designed to exploit these differences:

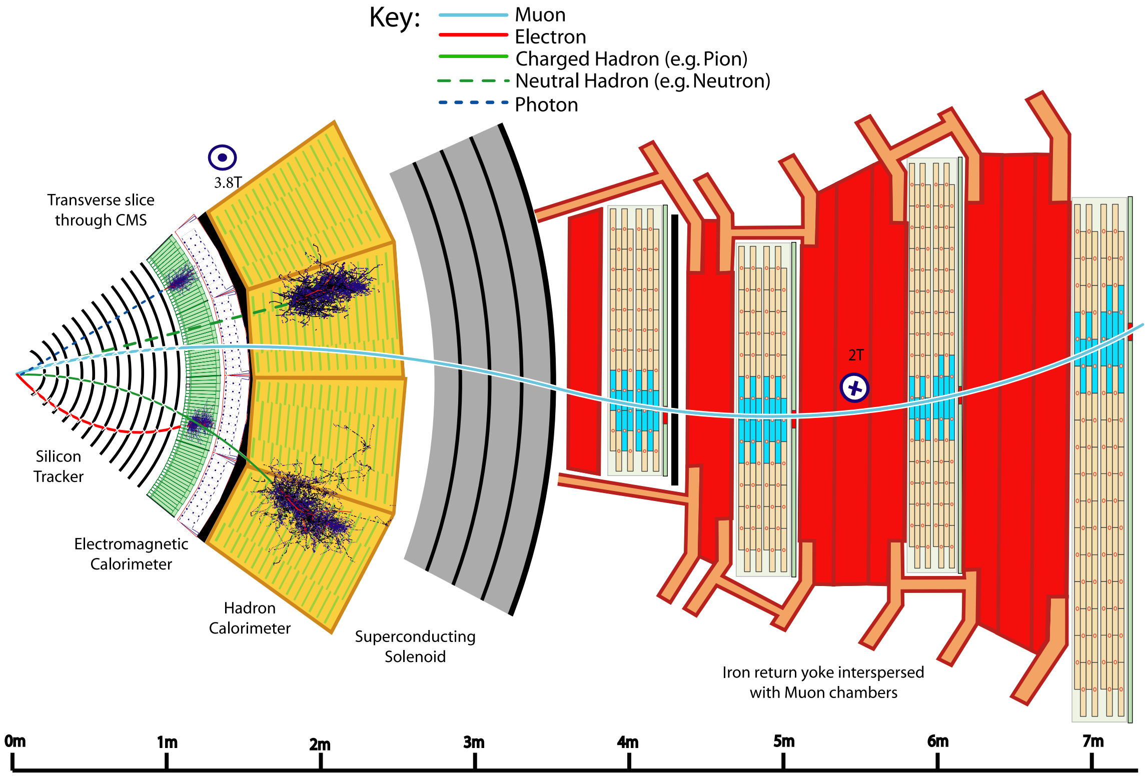

Key:
Muon
Electron
Charged Hadron (e.g. Pion)
Neutral Hadron (e.g. Neutron)
Photon
Transverse slice through CMS
3.8T
2T
Silicon Tracker
Electromagnetic Calorimeter
Hadron Calorimeter
Superconducting Solenoid
Iron return yoke interspersed with Muon chambers
0m
1m
2m
3m
4m
5m
6m
7m

**Figure 3.10.:** A visualization of particle interactions with the different subdetectors. Electrons leave a track in the tracker and energy cluster in the ECAL, while photons only result in an ECAL cluster. Charged hadrons deposit energy in the tracker, the ECAL, and the HCAL. Muons deposit only minimal energy in the ECAL and HCAL but can be identified via their track in the inner tracker and the muon system. The visualization is taken from [63].

1. **Muons**: Selection criteria for muons are based on the muon tracks. All energy deposits in the calorimeters with a distance of less than $\Delta R = 0.3$ from the muon trajectory are associated with the muon.
2. **Electrons**: For electrons, GSF tracks from the inner tracker have to be associated with a reconstructed SC from the ECAL. In addition, the energy in a linked HCAL

cluster must not be larger than 10% of the ECAL cluster energy since electrons will deposit most of their energy in the ECAL.

3. **Photons**: A SC that cannot be linked with a track from the inner tracker, and has an energy $E_{\mathrm{T}} > 10\,\mathrm{GeV}$ is considered an isolated photon. All remaining ECAL clusters that are not linked with a track or an HCAL cluster are considered non-isolated photons.

4. **Hadrons**: All remaining F elements are used for the reconstruction of Hadrons. Typical examples include a track linked to an ECAL cluster with lower and an HCAL cluster with larger energy. HCAL clusters without any additional link are considered neutral hadrons.

After the identification steps outlined above, all reconstructed signals are assigned to a reconstructed object, and the event is fully and globally reconstructed.

After the PF reconstruction, high-level objects like jets or event quantities like the missing transverse momentum can be calculated. Dedicated identification algorithms, often based on multivariate methods, are used to improve the misidentification versus reconstruction efficiency. By defining efficiency-purity conditions, called working points (WP), it is possible to obtain a standard set of identification criteria shared among all analyses. The algorithms used to identify jets, muons, electrons, and $\tau$ leptons are outlined in the following sections.

### 3.3.5. Jet Reconstruction and Identification

Quarks and gluons originating from a particle collision cannot be observed as free particles. Both quarks and gluons carry colour charges and are therefore affected by strong interactions. Due to confinement, single quarks cannot be observed. Instead, quarks and gluons always form colour-neutral hadrons. Typically, these colour-neutral hadrons will decay further until only stable particles remain. This process is called hadronization. Within the detector, the hadronized decay products are collimated in a narrow cone called a jet. Since the LHC is a hadron collider, the accurate reconstruction of jets is crucial for reconstructing the entire event. The jet reconstruction is complicated because the hard process is accompanied by PU collisions, resulting in numerous jets in the whole event.

To reconstruct the kinematic properties of the particle from which the jet originated, the momenta of all decay products must be combined using clustering algorithms. A detailed discussion of different jet algorithms can be found in [67]. All jet algorithms aim to identify all decay products of the same jet. The algorithms have to be collinear-safe and infrared-safe, meaning the addition of a soft or a collinear particle to the jet should not alter the jet axis and the number of jets.

At CMS, the anti-$k_t$ algorithm described in [68] is used for the jet reconstruction. It is a sequential clustering algorithm, which means that rather than trying to combine particles within a given cone, particles are sequentially recombined into a jet:

1. The distance between all objects $d_{ij}$ and the beam distances $d_{iB}$ are calculated:

$$d_{ij} = \min\left(p_{\mathrm{T},i}^{2k}\,,\ p_{\mathrm{T},j}^{2k}\right)\frac{\Delta \mathrm{R}_{i,j}^{2}}{R^{2}} \tag{3.8}$$

$$d_{iB} = p_{\mathrm{T},i}^{2k}. \tag{3.9}$$

   For the anti-$k_t$ algorithm, the value of $k$ is set to -1. In the beginning, the objects represent all particles available. The value of $R$ defines the radius in which particles are considered to be in the same jet. This value is set to 0.4 for most analyses.

2. The objects with the smallest distance $d_{ij}$ are combined to form a new object. If the distance $d_{iB}$ is the smallest of all distances, the algorithm is stopped, and the object is considered a jet. In the case of the anti-$k_t$ algorithm, the hard objects will be clustered first, and then soft objects will be added subsequently. As a result, the resulting jets often have a cone structure, which is not the case for the other algorithms, such as the $k_t$ algorithm, where $k$ is set to 1, and therefore soft objects are clustered first.

3. Repeat the procedure with the newly merged particles until all particles are combined into jets.

**Pileup Mitigation** It is important to know, if a jet or a particle belongs to the hard process or if it was produced by PU. The distribution of the number of PU collisions over the years and the average number of PU collisions is shown in Figure 3.11. To separate particles from the hard process and PU collisions, several algorithms were developed. In CMS the PileUp Per Particle Interaction (PUPPI) algorithm [69, 70], and the Charged-Hadron Subtraction (CHS) algorithm [63] are commonly used to identify particles from PU collisions and reduce their effect on the jet clustering.

The CHS algorithm relies on the information from the PV finding algorithm described in Section 3.3.1. All charged particles associated with PU vertices are removed. This removal is done before the application of the jet clustering. As a result, all charged particles associated with PU vertices are removed from the jet reconstruction. However, neutral and charged particles without vertex association are not removed and can appear in reconstructed jets.

The PUPPI algorithm is based on a more general approach, where each reconstructed particle is assigned a weight $w_{\mathrm{p}}$. This weight represents a probability that the particle belongs to the PV. Similar to the CHS algorithm, the PUPPI algorithm is applied before the jet clustering using all particles reconstructed by the PF algorithm. For the jet clustering, the $p_{\mathrm{T}}$ of all particles is scaled according to their assigned weight $w_{\mathrm{p}}$.

All charged particles assigned to the PV receive a weight of 1, while all charged particles assigned to PU vertices receive a weight of 0. A charged particle not assigned to any vertex is given a weight of one if its distance to the PV is smaller than 0.3 cm in the z-direction; otherwise, the weight is set to zero. For neutral particles, the weight is calculated using a discrimination variable $\alpha$

$$\alpha_i = \sum_{j\neq i}\left(\frac{p_{\mathrm{T},j}}{\Delta \mathrm{R}_{ij}}\right)^{2} \tag{3.10}$$

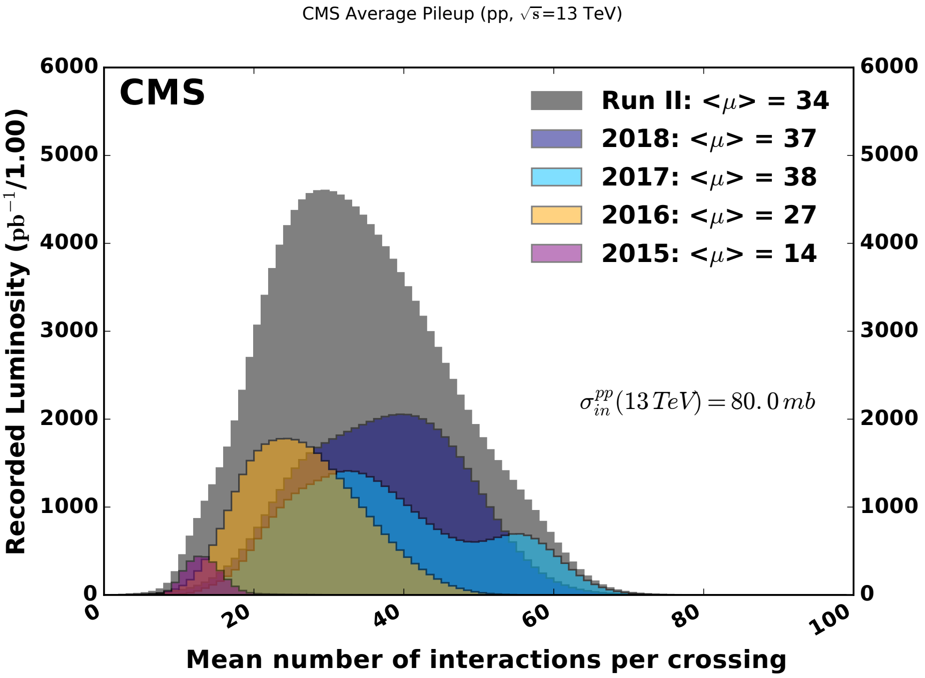


**Figure 3.11.:** Distribution of the average number of PU for pp collisions in 2015 (purple), 2016 (orange), 2017 (light blue), and 2018 (blue). The mean number of PU is shown in the legend.

where $\Delta R_{ij}$ is the distance between two particles. The particles $j$ consist of all charged particles from the PV if $|\eta_i| > 2.5$, meaning that tracking information is available and consists of all reconstructed particles otherwise. If $\Delta R_{ij}$ is larger than 0.4, the particle $j$ is not considered. If a neutral particle is close to a charged particle from the PV with large $p_T$, it will have a large value of $\alpha_i$, while a neutral particle far away from the high-$p_T$ particles of the PV will have a small value of $\alpha_i$.

To calculate the weight $w_p$, the value of $\alpha_i$ has to be converted into a probability. Using all charged particles from PU, the expected PU distribution of $\alpha_{PU}$ is calculated. Now the compatibility of the individual $\alpha_i$ values is calculated using a signed $\chi^2$ value

$$\chi_i^2 = \frac{(\alpha_i - \overline{\alpha_{PU}}) \cdot |\alpha_i - \overline{\alpha_{PU}}|}{\left(\alpha_{PU}^{RMS}\right)^2}, \tag{3.11}$$

where $\overline{\alpha_{PU}}$ is the median and $\alpha_{PU}^{RMS}$ is the standard deviation of the $\alpha_{PU}$ distribution. The $\chi_i^2$ values are expected to be distributed according to a $\chi^2$ distribution with one degree of freedom. The weight $w_i$ can then be calculated using the cumulative distribution function (CDF) of this $\chi^2$ distribution. As a result, particles with a large $\chi_i^2$ value are assigned large weights and are likely to belong to the PV. In the region $|\eta| > 2.5$, where no tracking information is available, $\overline{\alpha_{PU}}$ and $\alpha_{PU}^{RMS}$ are calculated using an extrapolation from the region $|\eta| < 2.5$. The extrapolation factors are determined using simulation.

**Jet Identification** The *DeepJet* algorithm [71–73] is used to identify if a jet originated from a gluon or a heavier quark like a b, or c quark. Due to the suppression of the $V_{ub}$ and $V_{cb}$ elements in the CKM matrix, hadrons containing a b quark have a longer lifetime than other hadrons. The same effect, although less strong, is present for hadrons containing c quarks. The lifetime of up to 2 ps allows b, and c hadrons to travel a short distance within the detector before decaying. In comparison to jets induced by gluons or light quarks, heavy-flavour jets can originate from a secondary vertex (SV), which is displaced from the PV by a few mm or cm. The distance between the PV and the SV and the existence of a SV can be used to differentiate between light- and heavy-flavour jets. To reconstruct a SV, a specialized version of the standard vertex finding algorithm is used.

The initial quark in a heavy-flavour jet has a larger mass, because the decay products of the quarks have, on average, a larger $p_{\mathrm{T}}$ compared to the rest of the jet constituents. Within jets, leptons are mainly produced in b and c quark decays. As a result, a high $p_{\mathrm{T}}$ lepton within the jet indicates a heavy-flavour jet.

The *DeepJet* algorithm is based on a deep neural network and yields six different probabilities for each jet:

- $P_{\mathrm{bb}}$: the probability that the jet contains at least two b hadrons
- $P_{\mathrm{b}}$: the probability that the jet contains at least one b hadrons decaying hadronically
- $P_{\mathrm{b,\,lep}}$: the probability that the jet contains at least one b hadrons decaying leptonically
- $P_{\mathrm{c}}$: the probability that the jet contains at least one c hadrons
- $P_{\mathrm{l}}$: the probability that the jet originated from a light quark
- $P_{\mathrm{g}}$: the probability that the jet originated from a gluon

The input variables are related to the quantities of the reconstructed particles used to cluster the initial jet, like momentum, track fitting information, and compatibility with the PV. This information is included for both neutral and charged particles within the jet. In addition, the number of reconstructed vertices and quantities related to the jet clustering are added. Finally, information targeting the SV fit and its impact parameters are included. In total, 650 input variables are used.

### 3.3.6. Muon Identification

The muon track reconstruction was explained in Section 3.3.1. Additional variables sensitive to muon misidentification are used to refine the identification of muons [56]. These variables include the number of hits in the inner tracker, the goodness-of-fit of the global muon track, the matching of the muon track in the inner and the muon system, and the muon segment compatibility. The muon segment compatibility is a score between zero and one, representing the compatibility of an extrapolated tracker muon trajectory with registered hits in the muon system. If the hits in the muon system are close to the extrapolated track, a high score is given, while a low score represents poor compatibility.

The main contributions for misidentified muons result from cosmic muons, or jets, where the shower is not entirely contained within the HCAL; therefore, some particles can leak into the muon system. The most common muon identification working points are

- **Loose**: A loose muon was reconstructed by the PF algorithm. It can be either a global muon or a tracker muon. The loose working point is designed to have high efficiency.

- **Medium**: A medium muon is always also a loose muon. In addition, the muon track has to traverse at least 80% of all silicon tracker layers. If the muon is global, the goodness-of-fit of the global track fit has to be $\chi^2/N_{\mathrm{dof}} < 3$, and the track from the muon system and the silicon tracker have to be consistent (the position match has to be $\chi^2 < 12$). Additionally, the $\chi^2$ of a kink finding algorithm must be smaller than 20 to reject muon tracks with a prominent kink. If the muon is only a tracker muon, it has to have a muon segment compatibility score larger than 0.451, whereas, for a global muon, the value must be larger than 0.303. The medium working point is targeted towards prompt muons. An efficiency of over 99% for Z boson decays is achieved using this working point.

- **Tight**: A tight muon is again always a loose muon and is also required to be a global muon. At least one hit in the pixel detector is required to suppress muons that were not produced instantly. The goodness-of-fit of the track fit must be $\chi^2/N_{\mathrm{dof}} < 10$ and include at least one hit in the muon system. In addition, the muon must originate from the primary vertex and have a good matching between the track in the silicon tracker and the muon system. The muon must originate from the PV, so the impact parameters must be $|d_{\mathrm{XY}}| < 0.2\,\mathrm{cm}$ and $|d_{\mathrm{Z}}| < 0.5\,\mathrm{cm}$. The tight working point is designed to have high purity.

For the muon momentum measurement, refitting of the final muon track is performed, taking into account that depending on the $p_{\mathrm{T}}$ of the muon, different fit settings result in different $p_{\mathrm{T}}$ resolutions. The refitting algorithm is described in [62]. For muons below 200 GeV, it is shown that a fit using only silicon tracker information results in the best momentum measurement. For larger muon momenta, the importance of information from the muon system increases and is therefore preferred for muons with a momentum above 200 GeV.

The *isolation* of a muon $I^{\mu}_{\mathrm{rel}}$ is defined as the ratio of the sum of the $p_{\mathrm{T}}$ of all particles other than the muon in the vicinity of the muon over the $p_{\mathrm{T}}$ of the muon itself:

$$I^{\mu}_{\mathrm{rel}} = \frac{\sum p_{\mathrm{T}}^{PV} + \max\left(0\,,\ \sum p_{\mathrm{T}}^{\gamma} + \sum p_{\mathrm{T}}^{NH} - \Delta\beta \cdot \sum p_{\mathrm{T}}^{PU}\right)}{p_{\mathrm{T}}^{\mu}}, \tag{3.12}$$

with $p_{\mathrm{T}}^{PV}$ being the $p_{\mathrm{T}}$ of charged particles associated with the primary vertex of the event, $p_{\mathrm{T}}^{\gamma}$ the energy of photons, $p_{\mathrm{T}}^{NH}$ the energy of neutral hadrons, and $p_{\mathrm{T}}^{PU}$ the $p_{\mathrm{T}}$ of charged particles not associated to the primary vertex. All particles in a cone of $\Delta\mathrm{R} = 0.4$ around the muon are considered. The factor $\Delta\beta$ is chosen to be 0.5, which is the estimated contribution of charged hadrons from PU collisions. Typically, a muon is considered well isolated, if $I^{\mu}_{\mathrm{rel}} < 0.15$.

### 3.3.7. Electron Identification

Electron identification is necessary to discriminate between genuine and falsely identified electrons, which mainly come from converted photons, misidentified jets with a significant ECAL energy deposition, or non-prompt electrons from b quark or c quark decays. An identification algorithm based on a boosted decision tree (BDT) is used. The algorithm is described in [64, 74]. The variables used within the BDT can be grouped into multiple categories:

- **ECAL - Tracker matching variables**: These variables are related to the matching between the track in the silicon tracker and the ECAL cluster. Some variables are sensitive to geometric differences between the position of reconstructed SC and the extrapolated electron track, whereas other variables are sensitive to the momentum determined using the track and the ECAL cluster.

- **Track variables**: These variables are related to the track parameters and the quality of both the initial track fit and the GSF refit of the track as described in Section 3.3.1.

- **Calorimeter cluster variables**: These variables are related to the energy deposition in the ECAL and HCAL. They are sensitive to the position and the shape of the electromagnetic shower within the calorimeter. Also, the energy ratio between energy deposited in the ECAL and the HCAL is considered. A prompt electron will have a significant portion of its energy deposited in the ECAL, whereas a misidentified jet will have more energy deposited in the HCAL.

The BDT was trained using electron candidates from Monte Carlo (MC) simulation. Multiple working points are provided; the most common ones are the ones with an efficiency of 80% and 90%.

The isolation of the electron can be considered to suppress the misidentification of electrons greatly. For electrons, isolation is defined as

$$I_{\mathrm{rel}}^{\mathrm{e}} = \frac{\sum p_{\mathrm{T}}^{PV} + \max\left(0\,,\ \sum p_{\mathrm{T}}^{\gamma} + \sum p_{\mathrm{T}}^{NH} - \rho \cdot A_{\mathrm{eff}}\right)}{p_{\mathrm{T}}^{e}}, \tag{3.13}$$

which is equivalent to the muon isolation in Equation (3.12) apart from the PU mitigation. The PU contribution is determined as the product of the medium transverse energy density per unit era $\rho$ and $A_{\mathrm{eff}}$, the area of the isolation region around the electron, depending on the electron η. The η dependence considers how large the PU density in each detector region is. The $A_{\mathrm{eff}}$ values are listed in Table 3.2. The signal cone, in which the electron isolation is calculated, is set to $\Delta R = 0.3$.

The measurement of the electron energy can be improved by combining the momentum determined from the electron track with the energy of the associated SC. The combined energy is defined as

$$E_{\mathrm{comb}} = \frac{E_{\mathrm{SC}}/\sigma_{\mathrm{E}}^2 + p_{\mathrm{track}}/\sigma_p^2}{1/\sigma_{\mathrm{E}}^2 + 1/\sigma_p^2}, \tag{3.14}$$

**Table 3.2.:** The effective area $A_{\text{eff}}$ that is used for the isolation calculation of the electron, depending on the η of the electron.

| $\|\eta\|$ | $A_{\text{eff}}$ |
|---|---|
| $[0.000, 1.000]$ | 0.1440 |
| $[1.000, 1.479]$ | 0.1562 |
| $[1.479, 2.000]$ | 0.1032 |
| $[2.000, 2.200]$ | 0.0859 |
| $[2.200, 2.300]$ | 0.1116 |
| $[2.300, 2.400]$ | 0.1321 |
| $[2.400, 2.500]$ | 0.1654 |

where $E_{\text{SC}}$ is the energy deposited in the SC, $p_{\text{track}}$ is the momentum determined from the electron track, and $\sigma_{\text{E}}^2$ and $\sigma_p^2$ are the resolution of the SC energy measurement and the momentum measurement, respectively. The resolution for the whole electron energy range is improved by using this combined measurement. The momentum measurement from the tracker has a better resolution for low-energy electrons. In contrast, the energy measurement from the SC has a superior resolution for high-energy electrons.

### 3.3.8. Missing Transverse Energy

The *missing transverse energy* (MET) is an important measure to get an experimental handle on the energy of neutrinos. A detailed description of the MET reconstruction can be found in [75]. The neutrinos can not be reconstructed within the detector as they do not interact with any subdetectors. However, their presence can be seen in the imbalance of the transverse momentum sum. The MET and the missing transverse momentum are defined as

$$E_{\text{T}}^{\text{miss}} = -\sum_{i=1}^{N} E_{\text{T},i} \tag{3.15}$$

$$p_{\text{T}}^{\text{miss}} = -\sum_{i=1}^{N} p_{\text{T},i} \tag{3.16}$$

To improve the MET reconstruction, the PU mitigation algorithm PUPPI described in Section 3.3.5 can be used [69], since the calcuation of the MET is highly dependent on the particles $i$ that are considered. By using the weights $w_i$ derived by the PUPPI algorithm, the $p_{\text{T}}^{\text{miss}}$ can be calculated as

$$p_{\text{T}}^{\text{miss}} = -\sum_{i=1}^{N} w_i \cdot p_{\text{T},i} \tag{3.17}$$

By incorporating the weights from the PUPPI algorithm, the overall MET reconstruction can be improved, resulting in an improved MET resolution and MET values closer to the true MET.

### 3.3.9. Tau Reconstruction and Identification

As described in Chapter 2, the $\tau$ lepton has both leptonic and hadronic decay modes. In addition, at least one neutrino is present in every $\tau$ lepton decay, making the $\tau$ lepton reconstruction a challenging task which requires all parts of the detector. For the leptonic decay modes of the $\tau$ lepton, only the electron or the muon of the decay can be reconstructed. For $\tau_\mathrm{h}$ decay modes, the Hadron-Plus-Strip reconstruction algorithm (HPS algorithm) is used [76–78] to reconstruct the visible part of the $\tau_\mathrm{h}$ decay.

**HPS Tau Reconstruction** All reconstructed jets are taken as input for the HPS algorithm. Since many of the hadronic decay modes include $\pi^0$ as decay products, which promptly decay further into two photons, the energy deposits in the ECAL are of special interest for the $\tau_\mathrm{h}$ reconstruction. In a given $\Delta\eta \times \Delta\phi$ region, all the energy deposits in the ECAL are combined into a *strip*. Similar to the electron reconstruction, the strips have a larger spread in the $\phi$ direction than in the $\eta$ direction. The spread is uneven due to the bremsstrahlung, which is radiated perpendicular to the track, and the orientation of the magnetic field. A new strip is generated via the following steps:

1. The electron or photon with the largest $p_\mathrm{T}$ not contained in any strip is used as a seed for the new strip.
2. The search window is defined as

   $$\Delta\eta = f\left(p_\mathrm{T}^{e/\gamma}\right) + f\left(p_\mathrm{T}^{\text{strip}}\right) \tag{3.18}$$

   $$\Delta\phi = g\left(p_\mathrm{T}^{e/\gamma}\right) + g\left(p_\mathrm{T}^{\text{strip}}\right) \tag{3.19}$$

   with $f$ and $g$ being defined as

   $$f(p_\mathrm{T}) = 0.20 p_\mathrm{T}^{-0.66} \tag{3.20}$$

   $$g(p_\mathrm{T}) = 0.35 p_\mathrm{T}^{-0.71}. \tag{3.21}$$

   The functions $f$ and $g$ were determined using simulated $\tau_\mathrm{h}$ decays. If the electron or photon with the second largest $p_\mathrm{T}$ is contained within this search window, it is merged with the seed strip, and the size of the strip is increased accordingly. The maximum size of the strip is set to

   $$\Delta\eta \times \Delta\phi = 0.15 \times 0.3. \tag{3.22}$$

3. If a new electron or photon was added, the position of the strip is updated via a $p_\mathrm{T}$ weighted average of all strip constituents.
4. The reconstruction of a strip is complete if no additional electron or photon was found in the search window. The reconstruction of a new strip is started.

After the strips are reconstructed, they are combined into $\tau_h$ candidates. To determine the correct $\tau_h$ decay mode, the mass of all visible decay products is used. The HPS algorithm can differentiate between the $\tau_h$ decay modes listed in Table 2.2

$$\mathrm{DM} \in \{0, 1, 2, 10, 11\}$$

by using the mass of the hadronic component of the $\tau_h$ decay and the mass of strips. No strip is expected for the one prong decay, so only the mass of the hadronic component is used to determine the decay mode. The additional $\pi^0$ in the decay is associated with the strip component of the $\tau_h$ candidate.

**Identification** After the reconstruction of $\tau$ lepton candidates using the HPS algorithm, an identification algorithm based on a deep neural network is used to discriminate between true $\tau$ leptons and electrons, true $\tau$ leptons and muons, and jets misreconstructed as $\tau$ leptons. A detailed description of the *DeepTau* algorithm used for this task is given in [79]. Compared to jets from quarks or gluons, $\tau_h$ jets are more narrow and consist only of a few decay products. Additionally, $\tau$ leptons are mainly produced as isolated leptons, so the region surrounding the $\tau_h$ jet is expected to be empty.

The *DeepTau* algorithm utilized two types of input features

- **High-level features** These features were used in previous identification approaches and are known to have a strong discrimination power. In total, 47 high-level features are used, including the Lorentz vector and charge of the $\tau_h$ candidate, the number and type of particles used during the HPS reconstruction, the isolation of the $\tau_h$ candidate, the compatibility of the $\tau_h$ candidate with the PV, variables related to the energy and spatial distribution of the strips.

- **Low-level features** These are features on the level of reconstructed PF candidates. Two grids in the $\eta - \phi$ plane surrounding the $\tau_h$ candidate are defined: an inner grid representing the signal cone of the $\tau_h$ jet and an outer grid representing the isolation cone of the $\tau_h$ jet. The two grids are visualized in Figure 3.12. The inner grid has a size of

  $$\eta \times \phi = 0.22 \times 0.22 \;\; \text{(split into 11x11 cells)}$$

  and the outer grid has a size of

  $$\eta \times \phi = 1.05 \times 1.05 \;\; \text{(split into 21x21 cells).}$$

  The two grids are filled with the information of all reconstructed PF candidates. The candidates are placed in the grid depending on their $\eta$ and $\phi$ coordinates. The information used varies depending on the type of PF candidate. The $p_T$, the charge, and information on the associated track and the PV are always used. For hadrons, information from the HCAL is added; for electrons, ECAL information is added; and for muons, variables related to ECAL energy deposits and the muon system are added.

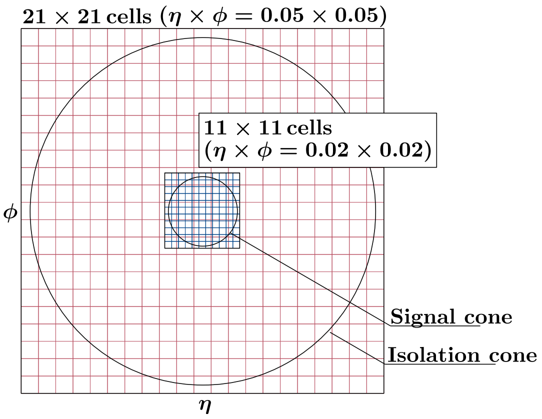


**Figure 3.12.:** Visualization of the inner and the outer grid used to identify the $\tau_\mathrm{h}$ jets with the *DeepTau* algorithm. Taken from [79].

The neural network of the classifier is created by connecting three separate subnetworks, one processing the high-level features and two processing the information in the inner and the outer grid, respectively. For processing the grid information, a set of convolutional layers is used, similar to how image processing is done in the computer vision field. The high-level feature network consists of multiple fully connected layers. In total, 105703 inputs are used.

The three subnetworks are combined using multiple fully connected layers, connecting to a final output layer with four nodes. Using a softmax activation function, the values of the output layer $y_i$ can be interpreted as probabilities of the $\tau_\mathrm{h}$ candidate belonging to the different classes $\tau_\mathrm{h}$, jet, electron or muon. The discriminator is then defined as

$$D_i(y) = \frac{y_\tau}{y_\tau + y_i} \qquad i \in \{\text{jet, electron, muon}\}, \tag{3.23}$$

where $i$ is the type of particle and $y_i$ is the probability of the $\tau_\mathrm{h}$ candidate belonging to the class $i$. In the following, the discriminators are denoted as *vsJet* ($D_\mathrm{jet}$), *vsEle* ($D_\mathrm{electron}$), and *vsMu* ($D_\mathrm{muon}$). Several working points are defined, ranging from signal efficiency of 40% for the tightest *vsJet* working point to signal efficiency of 99.5% for the loosest *vsEle* working point. The full set of working points is given in Table 3.3.

### 3.3.10. The CMS Trigger System

As mentioned before, LHC was operated at a collision rate of 40 MHz during Run II. A complete detector readout stored in a file takes about 2 MB of disk space, so an operation of the CMS detector without any trigger would result in an output rate of 80 TB $\mathrm{s}^{-1}$ which is far more than any modern storage system can handle.

Therefore, a trigger system consisting of two steps is used at the CMS detector to reduce the output rate to a few kHz. The two steps are the Level-1 trigger (L1), described in Section 3.3.10.1, and the High Level Trigger (HLT), described in Section 3.3.10.2. The trigger setup is illustrated in Figure 3.13.

**Table 3.3.:** The *DeepTau* $\tau_\mathrm{h}$ identification efficiency for the different working points and discriminators. The working points are defined using a simulated $\mathrm{H} \rightarrow \tau\tau$ sample with $\tau_\mathrm{h}$ $p_\mathrm{T} \in [30,70]$ GeV. Only four different working points are defined for the *vsMu* discriminator.

| **Working point** | *vsJet* | *vsEle* | *vsMu* |
|---:|---|---|---|
| VVVLoose | 98% | 99.5% | - |
| VVLoose | 95% | 99% | - |
| VLoose | 90% | 98% | 99.95% |
| Loose | 80% | 95% | 99.9% |
| Medium | 70% | 90% | 99.8% |
| Tight | 60% | 80% | 99.5% |
| VTight | 50% | 70% | - |
| VVTight | 40% | 60% | - |

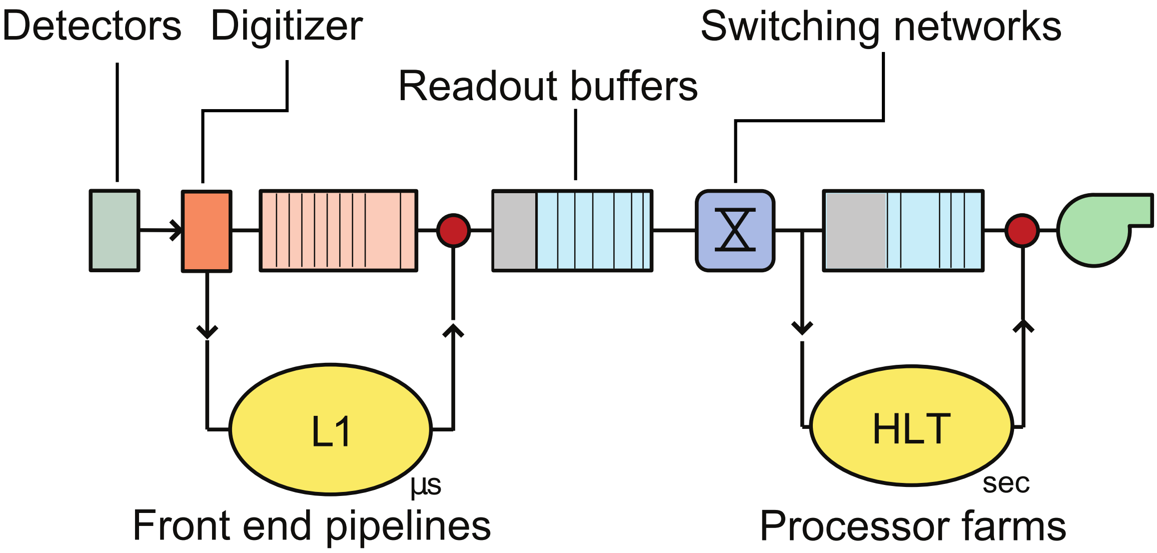


**Figure 3.13.:** The trigger setup of the CMS experiment, adapted from [80]. It consists of two levels, the L1 trigger and the HLT.

#### 3.3.10.1. L1 Trigger

The Level-1 trigger is the first trigger level of the CMS trigger setup. A detailed description of the L1 trigger system can be found in [48, 55]. The L1 trigger system is designed to have an output rate of 100 kHz, so, on average, every 400th event passes the L1 trigger. The time it takes until a decision if a collision should be kept or thrown away, is called the trigger latency. For the L1 trigger, the latency is 4 μs. To achieve this performance, the L1 trigger system is implemented using Field Programmable Gate Array (FPGA) chips. These chips are located inside the detector.

Within the L1 trigger system, several different *seeds* are defined. If at least one seed is accepted, the event is kept. As a result, the event has passed the trigger and will be processed by the HLT outside the detector. A seed can target different objects, namely the object can be a muon, a hadronic jet, a $\tau$ lepton, an electron or photon, the scalar sum of transverse energy $H_{\mathrm{T}}$, or the MET. For those objects, selection criteria can be made on the $\eta$, energy, or $p_{\mathrm{T}}$. Different types of seeds are possible:

- **Single seeds**: These are seeds targeting one single object, examples are a *SingleMuon* or a *SingleElectron* L1 seed.

- **Double seeds**: Similar to the **Single seeds**, but requiring two objects of the same type to be present, examples are a *DoubleMuon* or a *DoubleElectron* L1 seed.

- **Cross seeds**: Similar to the **Single seeds**, but requiring two objects of the different types to be present. Examples are a *Muon+Tau* or a *Electron+Tau* L1 seed. For some analysis, a dedicated L1 trigger seed is defined, which consists of even more than two objects. One example is a vector boson fusion trigger, targeting events with at least two jets and a large invariant mass.

To stay within the hard limit of a rate of 100 kHz, the selection thresholds have to be set accordingly. Single and double seeds take up about 75% of the total trigger budget. For muons and electrons, the thresholds are chosen to deliver an efficient selection of W- and Z-Boson decays, whereas for $\tau$ leptons, the thresholds target Higgs boson production. In total, about 150 of these physics L1 seeds are defined.

In addition to the seeds used for analysis, additional seeds used for calibration and monitoring are defined. These seeds can be *prescaled* to fit into the budget, which means that only every $N$-th event is accepted and processed further. Prescaled triggers are typically not used for analysis but are sufficient for the abovementioned tasks. During Run II, about 250 prescaled seeds were used, resulting in 400 seeds in the L1 trigger system.

The L1 trigger consists of two subsystems, the muon system trigger and the calorimeter trigger. On the level of the L1 trigger, no tracking information is available since it is impossible to perform a sufficiently good track reconstruction during the short latency of the L1 trigger. The two subsystems are combined into a single global trigger, which is used to check all possible trigger seeds in parallel.

The L1 muon system utilises signals from all three subdetectors available. Three separate track-finding algorithms are deployed, one for the barrel region, one for the overlap region and one for the endcap region. The results of these track-finding algorithms are combined and collected by the global trigger decision. For muon L1 seeds, only the muon tracks found this way are used without any calorimeter information.

The energy deposits from the ECAL and HCAL are read out and combined into calorimeter trigger towers (TTs) for the calorimeter trigger. A single TT consists of a 5x5 grid of ECAL crystals and the HCAL tower directly behind the crystal grid. This results in a TT size of $\Delta\eta \times \Delta\phi = 0.087 \times 0.087$ in the barrel region. In the endcap region, the TTs are larger, resulting in a size of $\Delta\eta \times \Delta\phi = 0.17 \times 0.17$.

Only the calorimeter information is used to identify electrons or photons on the L1 trigger level, making it impossible to differentiate between the two. For an $e/\gamma$ seed, the energy of a single TT has to be larger than 2 GeV. If this is the case, all surrounding TTs are added to the seed if their energy is larger than 1 GeV. A maximum of 8 TTs are added this way. To suppress misidentification, the energy from the ECAL crystals must be much larger than that from the HCAL tower. The seed can be required to be isolated by checking the energy of the surrounding TTs.

For $\tau_h$, a similar strategy is used; however, the total size of the seed is allowed to be greater since the decay products of the $\tau_h$ decay can be more spread out. Here, isolation can be used to suppress background from QCD-induced jets. For regular jets, a sliding window of $9 \times 9$ TTs is used, which corresponds to a jet cone of $\Delta R = 0.4$. The energy of a jet seed is defined as the sum of the energy of all 81 TTs in the window.

#### 3.3.10.2. High Level Trigger

The HLT is the second trigger level of the CMS trigger setup. A detailed description of the HLT system can be found in [81, 82]. During the HLT reconstruction, a simplified set of reconstruction algorithms described in the previous sections is used. It is run in parallel on a large computing cluster of more than 32000 CPU cores. Its purpose is to further reduce the rate from 100 kHz to a rate that can be transferred and stored on permanent storage. The maximum output rate is a few kHz, or about 5 Gbit $s^{-1}$ which is the maximum bandwidth available for the CMS detector. The maximum processing time per event is limited to 320 ms.

Similar to the L1 trigger system, different trigger paths are defined, each sensitive to a different type of event signature. If the selection criteria imposed by at trigger path are passed, a trigger flag is set. If an event has at least one trigger flag set, it will be accepted by the HLT and stored on permanent storage.

Compared to the full track reconstruction, the algorithm parameters are modified during the HLT track reconstruction to ensure a shorter runtime. In contrast to the algorithm outlined in Section 3.3.1, only triplets of pixel tracker hits are considered as track seeds. In addition, the tracker is split into multiple $\eta - \phi$ regions, where each region is processed independently. This splitting reduces the combinatorics of the track reconstruction. Also, only three (four during the 2018 measurement) iterations of track reconstruction are performed. This way, the runtime of the HLT track reconstruction is reduced by a factor of ten compared to the offline track reconstruction. To improve the runtime, object reconstruction at the HLT level is not done using the holistic PF approach but instead using several, more specialized algorithms.

For electrons, SC and pixel hits are matched. This input is then used as a seed for an electron track reconstruction using the GSF algorithm. L1 $e/\gamma$ seeds are then used to seed the SC reconstruction. In 2016, the lowest unprescaled single electron trigger had a $p_{\mathrm{T}}$ threshold of $p_{\mathrm{T}} > 25\,\mathrm{GeV}$. In 2017 and 2018, this threshold was increased to 32 GeV.

Muons are reconstructed in a two-step process, again seeded by the L1 trigger results. In the L2 step, a local reconstruction only using the muon system is performed. In the final L3 step, these local muons are combined with hits from the inner tracker. The combination is done using three subsequent algorithms, where the next algorithm is only used if the previous fails. Again, a set of isolation and identification criteria can be applied. For muons the $p_{\mathrm{T}}$ thresholds are $p_{\mathrm{T}} > 22(24)\,\mathrm{GeV}$ for 2016 (2017/2018). Single electron and single muon trigger paths are responsible for about 65% of the total rate.

To reconstruct $\tau_{\mathrm{h}}$ at the HLT level, the PF algorithm is used. There is no low-energy single $\tau_{\mathrm{h}}$ trigger path defined; instead, three cross trigger paths target the three main di-$\tau$ final states $\mathrm{e}\tau_{\mathrm{h}}$, $\mu\tau_{\mathrm{h}}$ and $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$. For the first one, a global PF reconstruction is used, whereas, for $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$, only a regional reconstruction is seeded using L1 $\tau_{\mathrm{h}}$ candidates. During the 2018 measurements, the HPS algorithm described in Section 3.3.9 was included in the HLT $\tau_{\mathrm{h}}$ reconstruction, improving its resolution.

For jets, a reconstruction via calorimeter towers and PF is possible. The usage of the PF algorithm results in a better resolution for low-energy jets, as the accuracy of the tracker can improve the energy resolution. At higher energies, die calorimeter information is sufficient. The same is true for the MET reconstruction.

# 4 The H → ττ Analyses Program during Run II

The analysis selection and the composition of background processes are essential for the design of data-driven estimation methods, especially of the τ-embedding method. While the SM H → ττ measurement was one of the primary motivations for the development of the method, its application is not limited to just this analysis. It can be applied to a wide range of τ lepton analyses. The method was used in several analyses of the Higgs physics program with di-τ final states, of the CMS Collaboration during Run II.

A short overview of the analyses that made use of the τ-embedding method during Run II is given in the first half of this chapter. In the second half, the selections used for the different di-τ final states, as well as the main background processes contributing to the SM H → ττ measurement, are described. Since this is one of the most inclusive analyses within the CMS Collaboration that uses τ lepton pairs, it was one of the main inputs used for the design of the τ-embedding method. Due to its high complexity, inclusiveness, and ambitious goals, the SM H → ττ analysis is a very well-suited proxy for di-τ analyses that the τ-embedding method is targeted at.

## 4.1. Run II Measurements

All the measurements presented here use the full Run II data set. In the following, two Higgs boson analyses in di-τ final states in the SM context and two BSM searches using τ leptons in the final states are outlined. The two SM analyses are the differential measurements of gluon-induced (ggH) and electroweak (qqH) Higgs boson production in the di-τ final states [9–11], and the analysis of the CP structure of the τ lepton Yukawa coupling [83]. The two BSM analyses are searches for additional Higgs bosons and vector leptoquarks in the di-τ final state in a mass range from 60 GeV to 3500 GeV [12–14], and a search for the decay of a heavy Higgs boson (H) into two lighter Higgs bosons h and $h_S$, of which h is the observed Higgs boson with a mass of 125 GeV in the ττbb final state [15, 16]. In the scope of this thesis, significant contributions were made to all these analyses, apart from the CP measurement.

### SM Higgs boson measurement

The analysis has three measurement targets:

- The inclusive signal strength relative to the SM prediction of Higgs boson production is measured in the di-τ final state.

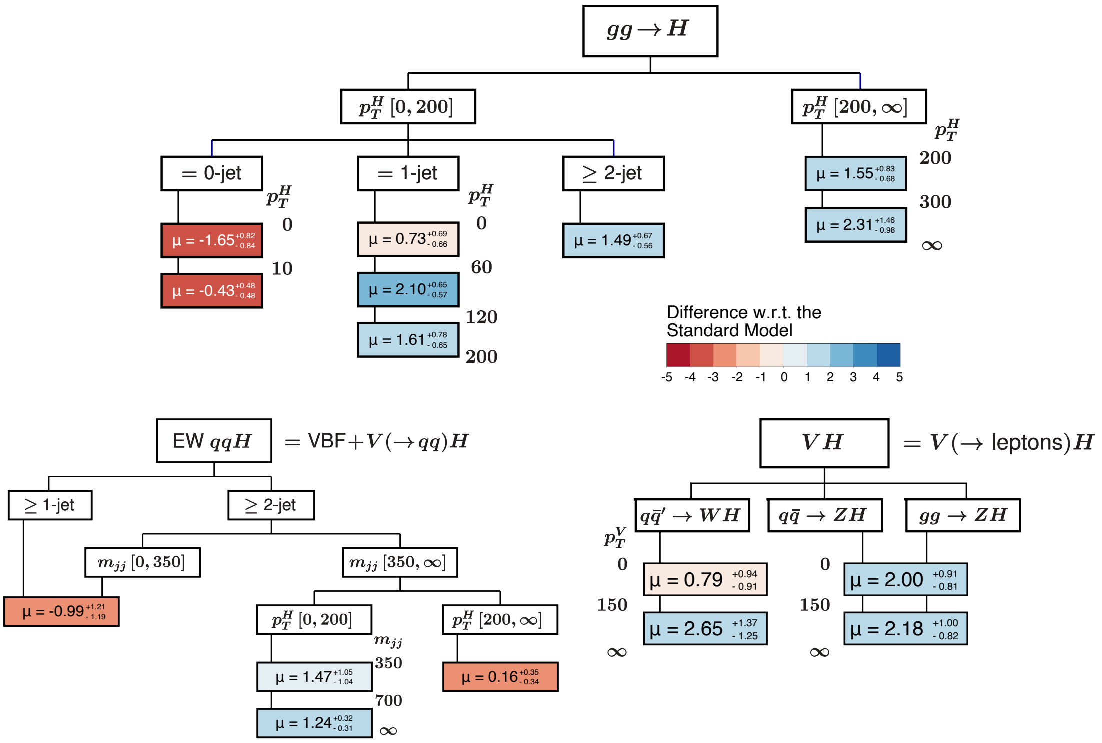

$gg \rightarrow H$
$p_T^H$ [0, 200]
$p_T^H$ [200, ∞]
= 0-jet
= 1-jet
≥ 2-jet
$p_T^H$
0
10
μ = -1.65$^{+0.82}_{-0.84}$
μ = -0.43$^{+0.48}_{-0.48}$
$p_T^H$
0
60
120
200
μ = 0.73$^{+0.69}_{-0.66}$
μ = 2.10$^{+0.65}_{-0.57}$
μ = 1.61$^{+0.78}_{-0.65}$
μ = 1.49$^{+0.67}_{-0.56}$
$p_T^H$
200
300
∞
μ = 1.55$^{+0.83}_{-0.68}$
μ = 2.31$^{+1.46}_{-0.98}$
Difference w.r.t. the Standard Model
-5 -4 -3 -2 -1 0 1 2 3 4 5
EW $qqH$ = VBF+$V(\rightarrow qq)H$
≥ 1-jet
≥ 2-jet
$m_{jj}$ [0, 350]
$m_{jj}$ [350, ∞]
μ = -0.99$^{+1.21}_{-1.19}$
$p_T^H$ [0, 200]
$p_T^H$ [200, ∞]
$m_{jj}$
350
700
∞
μ = 1.47$^{+1.05}_{-1.04}$
μ = 1.24$^{+0.32}_{-0.31}$
μ = 0.16$^{+0.35}_{-0.34}$
$VH$ = $V(\rightarrow$ leptons$)H$
$q\bar{q}' \rightarrow WH$
$q\bar{q} \rightarrow ZH$
$gg \rightarrow ZH$
$p_T^V$
0
150
∞
μ = 0.79$^{+0.94}_{-0.91}$
μ = 2.65$^{+1.37}_{-1.25}$
0
150
∞
μ = 2.00$^{+0.91}_{-0.81}$
μ = 2.18$^{+1.00}_{-0.82}$

**Figure 4.1.:** The results of the stage-1.2 STXS measurement. The measured signal strengths are colour coded according to their difference w.r.t. the SM prediction and given in units of $\sigma$. The values are taken from [9]. The visualization follows the visualization of the stage-1.2 STXS binning from [84].

- The signal strength is measured split by production mode. The considered production modes are gluon fusion (ggH), electroweak (qqH) and associated production with a vector boson (VH). The definitions of the production modes are chosen to correspond to the *stage-0* definitions in the simplified template cross section (STXS) scheme [35, 84]. Experimentalists and theorists developed this scheme as a common ground for differential Higgs boson measurements. In the STXS scheme, different phase space bins are defined depending on the transverse momentum of the Higgs boson and the number and mass of jets in the event. The bins are chosen such that migration effects between bins are minimal and that the bins are accessible for experimental measurements.
- The measurements of the observed SM Higgs boson production signal strengths are based on a total of 16 different STXS bins. The bins are chosen based on the *stage-1.2* STXS definitions.

The measurements of the ggH and qqH signal strengths are performed using a multi-classification neural network trained to identify Higgs boson decays. Two neural network setups are used, one for the inclusive and stage-0, and another one for the stage-1.2 measurements. The measurement of VH production is performed using a more traditional selection-based approach.

The SM H → ττ analysis aims for a model, that can describe the data within ± 5% uncertainty. The investigated phase space is defined by the HLT paths with the lowest available $p_{\mathrm{T}}$ thresholds. The analysis uses two data-driven methods to estimate a significant fraction of the contributing background processes. The τ-embedding method is an estimate of genuine di-τ events, and the $\mathrm{F_F}$ method is used to estimate the background originating from jets misidentified as $\tau_{\mathrm{h}}$ (jet $\rightarrow \tau_{\mathrm{h}}$). The $\mathrm{F_F}$ method is described in Section 4.2.4.

The inclusive signal strength was measured to be $0.82^{+0.10}_{-0.11}$, the signal strength for ggH was measured to be $0.67^{+0.20}_{-0.18}$, the signal strength for qqH was measured to be $0.81^{+0.17}_{-0.16}$, and the signal strength for VH was measured to be $1.79^{+0.47}_{-0.42}$. The results of the stage-1.2 measurements and their differences w.r.t. the SM prediction are shown in Figure 4.1. Apart from the more significant differences in the ggH 0-jet category, a good agreement with the SM prediction is observed.

### Analysis of the CP structure of the τ lepton Yukawa coupling

By parameterizing the Lagrangian of the τ lepton Yukawa coupling introduced in Equation (2.11) as

$$\mathcal{L}^{\tau}_{\mathrm{Yukawa}} = -\frac{m_{\tau}}{v}\mathrm{H}(\kappa_{\tau}\bar{\tau}\tau + \tilde{\kappa}_{\tau}\bar{\tau}i\gamma_5\tau), \tag{4.1}$$

the effective mixing angle

$$\alpha^{\mathrm{H}\tau\tau} = \arctan\left(\frac{\tilde{\kappa}_{\tau}}{\kappa_{\tau}}\right) \tag{4.2}$$

can be defined. This effective angle can be accessed experimentally via the angle $\phi_{CP}$, defined as the angle between the decay planes of the two τ leptons from the Higgs boson decay. In the CP-even scenario, an angle of 180° is favoured, while a value of 0° is favoured in the CP-odd scenario.

This analysis also uses the τ-embedding and $\mathrm{F_F}$ methods to estimate the major background processes. A multivariate (MVA) discriminant is used to separate between signal and background processes. This MVA score is then combined with the reconstructed value of $\phi_{CP}$ to extract the value of $\alpha^{\mathrm{H}\tau\tau}$ using a maximum likelihood fit. The observed value of

$$\alpha^{\mathrm{H}\tau\tau} = -1 \pm 19^{\circ}$$

excludes the CP-odd scenario at the $3\sigma$ level. The statistical uncertainty still dominates the uncertainty of the measurement.

### Search for additional Higgs bosons in the di-τ final state

The minimal supersymmetric SM (MSSM) is a possible minimal extension of the SM, where a more complex Higgs sector is proposed [85]. The MSSM predicts three neutral and two charged Higgs bosons. In this analysis, a model-independent search is performed. In addition, exclusion contours for several MSSM scenarios are set. For the model-independent search a mass range from 60 GeV up to 3.5 TeV is investigated. Exclusion limits on the product of the production cross section and the branching fraction for the decay into τ leptons are set for the production via gluon fusion and b quark-associated production.

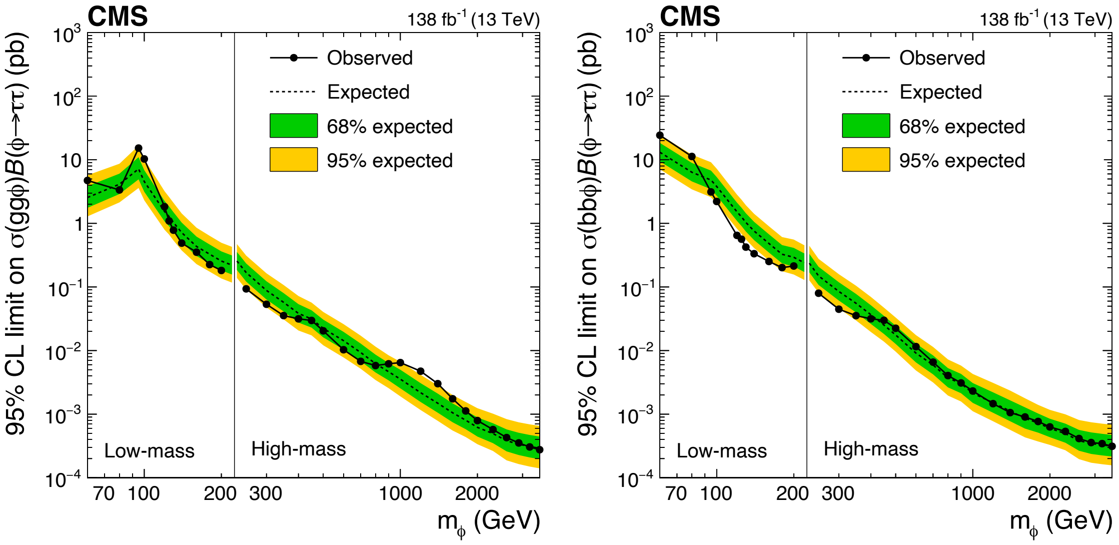


**Figure 4.2.:** The results of the model-independent search for additional Higgs bosons in gluon fusion (left) and b quark-associated production (left). The expected and observed 95% confidence level (CL) upper exclusion limits on the product of the production cross section and the branching fraction are shown. In gluon fusion, two excesses, one at 100 GeV and one at 1.2 TeV, are observed. Taken from [12].

The analysis uses the τ-embedding method and the $F_F$ method to estimate the major background processes. To increase the sensitivity, the search phase space is further split into a high-mass and a low-mass region. The split is performed at a hypothetical boson mass of 250 GeV. Each region is further subdivided into multiple categories. In the low-mass region, 26 categories and the mass of the di-τ system as discriminant are used. In the high-mass region, the total transverse mass [86]

$$m_{\mathrm{T}}^{\mathrm{tot}} = \sqrt{m_{\mathrm{T}}^2(\vec{p}_{\mathrm{T}}^{\,\tau_1}, \vec{p}_{\mathrm{T}}^{\,\tau_2}) + m_{\mathrm{T}}^2(\vec{p}_{\mathrm{T}}^{\,\tau_1}, \vec{p}_{\mathrm{T}}^{\,\mathrm{miss}}) + m_{\mathrm{T}}^2(\vec{p}_{\mathrm{t}}^{\,\tau_2}, \vec{p}_{\mathrm{T}}^{\,\mathrm{miss}})} \tag{4.3}$$

is used as discriminant, and the phase space is subdivided into 17 categories. The definition of the transverse mass $m_{\mathrm{T}}$ can be found in Equation (4.5). The resulting 95% CL exclusion limits for gluon fusion and b quark-associated production are shown in Figure 4.2. Two excesses at the $3\sigma$ level are observed for gluon fusion, one at 100 GeV and one at 1.2 TeV. In the b quark-associated production channel, no excess is observed.

### Search for the decay of a heavy Higgs boson into two lighter Higgs bosons in the ττbb final state

The basis for this search is the next-to-MSSM (NMSSM) [87], a possible extension of the MSSM. This extension has a total of 7 Higgs bosons; two charged and five neutral. This search focuses on the case, where a neutral heavy Higgs boson H decays into the SM Higgs boson denoted as h(125) and another lighter neutral Higgs boson $\mathrm{h_s}$ with $m_{\mathrm{h_s}} < m_{\mathrm{H}} - m_{\mathrm{h}}$. While the h(125) decays into a pair of τ leptons, the $h_s$ boson decays into a pair of b quarks.

This final state is fully included in the SM H → ττ analysis selection and can be realized by requiring two b-tagged jets in addition to the di-τ system. The flipped final state, where the SM Higgs boson decays into a pair of b quarks, is not considered in this analysis.

Again, the $F_F$ method and the τ-embedding method are applied in the search. In contrast to the previously discussed analyses, this analysis is performed in the $e\tau_h$, $\mu\tau_h$ and $\tau_h\tau_h$ final states. It is performed using a similar multi-classification approach as chosen for the SM H → ττ analysis. The main difference is that many different signal hypotheses are tested by training not one but a group of kinematically similar neural networks. Per final state, 68 neural networks were trained using different signal hypotheses, where the mass of the heavy Higgs boson H and the mass of the light Higgs boson $h_s$ is varied between

$$240 \leq m_{\mathrm{H}} \leq 3000\mathrm{GeV} \quad \text{and} \quad 60 \leq m_{\mathrm{h_s}} \leq 2800\mathrm{GeV} \tag{4.4}$$

While no excess is observed in the analysis, the results are used to set 95% CL upper exclusion limits on the product of the branching fraction and cross section of the process.

## 4.2. The Standard Model H → ττ Analysis

### 4.2.1. Event Selection

In the SM H → ττ measurement outlined in Section 4.1, the four major di-τ final states $e\tau_h$, $\mu\tau_h$, $\tau_h\tau_h$, and eμ are considered. The di-e and di-μ final states are omitted due to their small branching fractions, and large irreducible backgrounds from Z bosons decaying into pairs of electrons or muons. In each final state, the main step is to construct a τ pair based on a pair selection algorithm. Since τ → e (τ → μ) decays cannot be distinguished from prompt electrons (muons), all reconstructed electrons (muons) are used by the pair selection algorithm. Before the pair selection algorithm, quality criteria are applied to the reconstructed objects to determine whether they are good candidates for the given final state. The quality criteria coincide with the most inclusive HLT path thresholds for the required leptons; the $p_T$ thresholds of the offline event selection are set 1 GeV above the HLT path thresholds to avoid the turn-on region of the trigger.

Electrons and muons are required to be well-identified, well-isolated, and fully contained in the fiducial volume of the detector. In addition, it is required that the event was selected by an HLT path sensitive to the corresponding lepton. The detailed requirements for electrons and muons in the $e\tau_h$ and $\mu\tau_h$ final states are listed in Table 4.1. In the eμ final state, two HLT paths sensitive to one electron and one muon are used, where the leading electron (muon) is required to have a $p_T$ greater than 23 GeV. The subleading electron (muon) is required to have a $p_T$ larger than 15 GeV to ensure the full efficiency of the HLT path.

In the semileptonic final states, the $p_T$ of a $\tau_h$ candidate is required to be larger than 30 GeV. In the $\tau_h\tau_h$ final state, the $p_T$ of the two $\tau_h$ candidates is required to be larger than 40 GeV. All $\tau_h$ candidates are required to be well contained in the detector with $|\eta| < 2.1$. Their impact parameter has to be $d_z < 0.2$ cm. In addition, different WPs of the DeepTau algorithm described in Section 3.3.9 are used to suppress contributions from misidentified

**Table 4.1.:** Selection requirements for electrons and muons in the $e\tau_h$ and $\mu\tau_h$ final states.

| Criteria | Electron | Muon |
|---|---|---|
| transverse momentum | $p_T > 33\,\text{GeV}$ | $p_T > 25\,\text{GeV}$ |
| Isolation | $I^e_{rel} < 0.15$ | $I^\mu_{rel} < 0.15$ |
| pseudorapidity | $\lvert\eta\rvert < 2.5$ | $\lvert\eta\rvert < 2.4$ |
| HLT path | Single Electron Trigger | Single Muon Trigger |
| Identification | Wp90 | IsMedium |
| Impact parameter in cm | $d_z < 0.2\,, d_{xy} < 0.045$ | $d_z < 0.2\,, d_{xy} < 0.045$ |

**Table 4.2.:** Working points used for the reconstruction of the $\tau_h$ candidate in the $e\tau_h$, $\mu\tau_h$, and $\tau_h\tau_h$ final states.

| Final State | *vsJet* | *vsEle* | *vsMu* |
|---|---|---|---|
| $e\tau_h$ | Tight | Tight | VLoose |
| $\mu\tau_h$ | Tight | VVLoose | Tight |
| $\tau_h\tau_h$ | Tight | VVLoose | VLoose |

jets and leptons. In the $\mu\tau_h$ channel, a tighter *vsMuon* WP is used, while in the $e\tau_h$ channel, a tighter *vsEle* WP is used. The WP settings for all final states are listed in Table 4.2.

The pair selection algorithm to select the two objects of the $\tau$ pair is split into several steps:

- In all final states, a list of possible pair candidates is built from all permutations of the required objects. The requirements are:
  - $e\tau_h$ final state: one electron and one $\tau_h$ candidate
  - $\mu\tau_h$ final state: one muon and one $\tau_h$ candidate
  - $\tau_h\tau_h$ final state: two $\tau_h$ candidates
  - $e\mu$ final state: one electron and one muon candidate
- The two particles have to be well separated ($\Delta R > 0.5$) in the detector.
- If more than one possible pair is found, the quality and $p_T$ of the candidates are compared to determine the best-suited pair. The pair candidate is chosen based on the highest discriminator score of the DeepTau *vsJet* classifier for a $\tau_h$ candidate, and the lowest relative isolation $I^e_{rel}$ ($I^\mu_{rel}$) for an electron (muon) candidate. The proceeding particle property is only checked if the previous one is the same for more than one pair, for example, if the first particle is the same for all pairs. The order in which particle properties are checked is
  1. Quality of the first particle; In the $\mu\tau_h$ final state, the muon is the first particle, in the $e\tau_h$ final state, the electron is the first particle, the in $e\mu$ final state, the electron is the first particle, and in the $\tau_h\tau_h$ final state, the $\tau_h$ with the highest $p_T$ is used.

2. $p_T$ of the first particle
3. Quality of the second particle
4. $p_T$ of the second particle

In addition to the requirements listed above, the pair is required to be of opposite charge. To ensure that the analysis is orthogonal to other analyses with more leptons in the final states, a veto of additional electrons (muons) is applied in the $e\tau_h$ ($\mu\tau_h$) final state.

A selection cut on the transverse mass $m_T < 70\,\text{GeV}$ of the electron (muon) is applied in the $e\tau_h$ and $\mu\tau_h$ final states, where the transverse mass is defined as

$$m_T = \sqrt{2 p_T E_T^{\text{miss}} (1 - \cos \Delta\phi)}. \tag{4.5}$$

In Equation (4.5) $p_T$ corresponds the transverse momentum of the lepton, $E_T^{\text{miss}}$ to the MET of the lepton and $\Delta\phi$ to the angle between the lepton and the MET. This variable can be used to reject background from W bosons. For a lepton originating from a W boson decay, the angle between the lepton and the MET is large since the lepton, and the neutrino from the W boson decay are produced back to back in the centre-of-mass frame of the decay. Since the W boson is heavy, the angle between the lepton and the neutrino remains big in the lab frame. The resulting value of $m_T$ is large compared to, e.g. $Z \to \tau\tau$ events, where the MET does not have a distinguished direction and is composed of the contribution from multiple neutrinos.

Jets are reconstructed as described in Section 3.3.5. The jets are required to have a $p_T$ larger than 30 GeV and $|\eta| < 4.7$. In addition, each jet has to fulfil the jet identification requirements for the LOOSE working point described in [13]. For b-jets, the $p_T$ requirement is 20 GeV and $|\eta| < 2.4$. The b-jet has to pass the MEDIUM working point of the *DeepJet* algorithm described in Section 3.3.5. Both regular and b-jets are vetoed if they overlap with the two selected $\tau$ lepton candidates.

While the di-e and di-$\mu$ final states are not used in the analysis, these final states can be utilized to measure several correction factors and serve as additional control regions. A combination of single and cross HLT paths is used in both final states. The selection requirements are listed in Table 4.3. If more than one possible pair can be constructed, the pair with the largest mass of the dilepton system is selected. The two electrons (muons) are required to be well separated ($\Delta R > 0.5$) and of opposite charge.

### 4.2.2. Background Processes

The SM $H \to \tau\tau$ measurement is dominated by background processes. The contribution from signal events after event selection is less than 1% in all final states. Therefore, an accurate description of all background processes in the analysis is required. In the following, the different contributing background processes are described.

**Production of a Z boson** A Z boson can decay into a pair of $\tau$ leptons. The $Z \to \tau\tau$ and the $H \to \tau\tau$ decay have identical final state particles, with the primary difference being the mass of the Higgs and the Z boson. A leading order Feynman diagram for this

**Table 4.3.:** Selection Requirements for electrons and muons in the di-e and di-μ final states

| Criteria | Electron | Muon |
|---:|---|---|
| transverse momentum | $p_{\mathrm{T}} > 25\,\mathrm{GeV}$ | $p_{\mathrm{T}} > 18\,\mathrm{GeV}$ |
| Isolation | $I^{\mathrm{e}}_{\mathrm{rel}} < 0.15$ | $I^{\mu}_{\mathrm{rel}} < 0.15$ |
| pseudorapidity | $\lvert\eta\rvert < 2.5$ | $\lvert\eta\rvert < 2.4$ |
| HLT path | Single & Double Electron | Single & Double Muon |
| Identification | Wp90 | IsMedium |
| Impact parameter in cm | $d_{\mathrm{z}} < 0.2\,, d_{\mathrm{xy}} < 0.045$ | $d_{\mathrm{z}} < 0.2\,, d_{\mathrm{xy}} < 0.045$ |

process is shown in Figure 4.3 (top left). It is the dominant background process for all di-τ final states with at least one leptonic τ lepton decay. Due to the neutrinos produced during the τ lepton decay, no sharp resonance as the Z boson mass is visible in the invariant mass spectrum of the visible decay products ($m_{\mathrm{vis}}$). Instead, the resonance is smeared out and shifted to lower energies; for example, in the $\mu\tau_{\mathrm{h}}$ channel, the peak in the $m_{\mathrm{vis}}$ distribution (shown in Figure A.10 for the $\mu\tau_{\mathrm{h}}$ final state) is located at $\sim 65\,\mathrm{GeV}$.

The decay of a Z boson into two τ leptons is estimated via the τ-embedding method. In the $\mathrm{e}\tau_{\mathrm{h}}$ ($\mu\tau_{\mathrm{h}}$) channel, the Z boson decay to a pair of electrons (muons) can contribute if one of the electrons (muons) is misidentified as an $\tau_{\mathrm{h}}$ candidate. This contribution is denoted as Z → ll and is estimated using simulation. A small contribution comes from hadronic Z boson decays, estimated using the $\mathrm{F_F}$ method.

**QCD Multijet Production** Jets produced by strong interactions referred to as QCD multijet production can also end up in the event selection. In the full hadronic final state, two jets are misidentified as $\tau_{\mathrm{h}}$ candidates. Electrons can arise in the electromagnetic component of a jet, while muons can be produced through the decay of mesons, such as pions or kaons. In addition, semileptonic decays of b or c quarks can result in electrons or muons produced in a jet. In rare cases, a light quark or gluon-induced jet can be misidentified as an electron or muon. While leptons created in jets are generally not well isolated, some contribution is still expected due to the large branching fraction of the QCD multijet production. The contribution from this process is estimated using the $\mathrm{F_F}$ method. A sideband region estimation is used in the eμ channel since the $\mathrm{F_F}$ method is not applicable in the eμ final state. The sideband region is defined by inverting the opposite charge requirement of the analysis selection. This region, denoted as same-sign, is a pure QCD multijet control region since the jets of QCD multijet production are independent of each other.

**Production of Top Quark Pairs** At the LHC, top quark pairs ($\mathrm{t\bar{t}}$) are mainly produced via gluon-gluon fusion. Each of the two top quarks will immediately decay into a b quark and a W boson. With a branching fraction of more than 99%, this is the dominant decay mode. While b quarks can be observed as b-jets in the detector, the W boson can further decay to a lepton and the corresponding neutrino. One of the leading Feynman diagrams for this mechanism is shown in Figure 4.3(top right). Since the W boson can

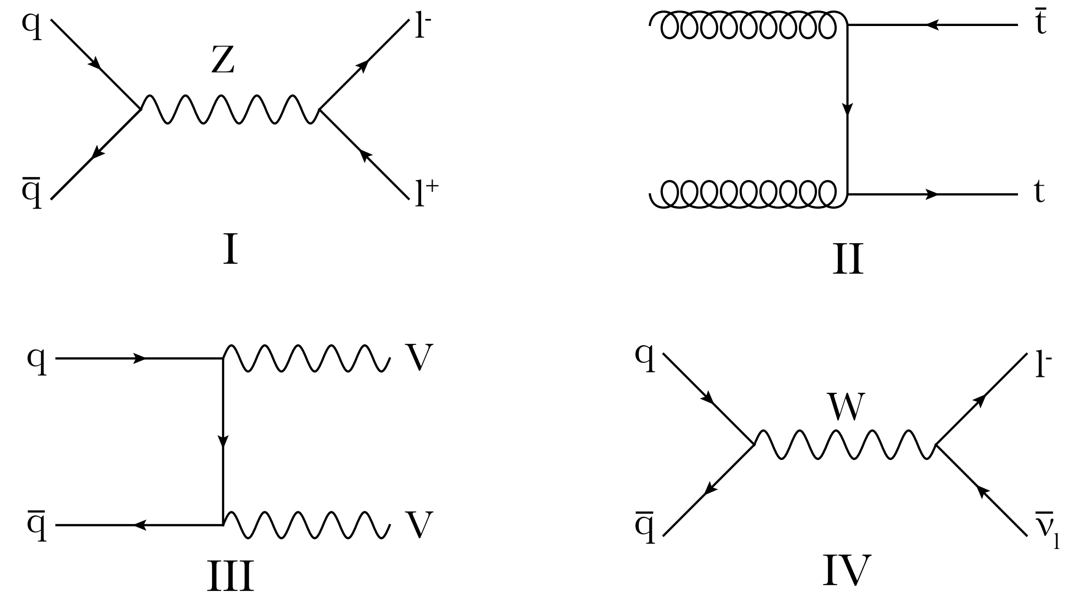


**Figure 4.3.:** A selection of possible leading order Feynman diagrams for the leading background processes. Shown are Z boson production (I), $t\bar{t}$ production (II), diboson production (III), and W boson production (IV).

decay leptonically (32.6%) and hadronically (67.4%), and the decay modes of the two W bosons are independent of each other, $t\bar{t}$ production is a background in all di-$\tau$ final states. If at least one W boson decays hadronically, which happens in 88.6% of the cases, the contribution is estimated using the $F_F$ method. If both W bosons decay into a $\tau$ lepton (1.2%), denoted as $t\bar{t}(\tau\tau)$, the contribution is estimated using the $\tau$-embedding method. The remaining leptonic decay modes are estimated using simulation.

**Diboson Production** Two vector bosons can be directly produced via quark-antiquark annihilation. The signature of this process is similar to the $t\bar{t}$ production, apart from the missing b-jets. A leading order Feynman diagram is shown in Figure 4.3 (bottom left). The method used for the estimation depends on the vector boson decay mode: hadronic decays are estimated using the $F_F$ method, the $\tau$-embedding method is used to estimate the case, where both vector bosons decay into a $\tau$ lepton ($VV(\tau\tau)$), and the rest is estimated using simulation.

**Production of a W boson associated with Jets** Events with a single W boson decay can end up in the event selection if the W boson decay products coincide with an additional jet, misidentified as a $\tau_h$ candidate. This process is denoted as W+jets production and is most relevant in the $e\tau_h$ and $\mu\tau_h$ final states. The contribution from this process is estimated using the $F_F$ method.

In Figure 4.4, the composition of events, split by the method used for the estimation, is shown. In all final states, the majority of processes are estimated using data-driven methods. In the $e\tau_h$, $\mu\tau_h$, and $e\mu$ final state, the $\tau$-embedding method is the most important method, while in the $\tau_h\tau_h$ final state, the $F_F$ method has the largest contribution.

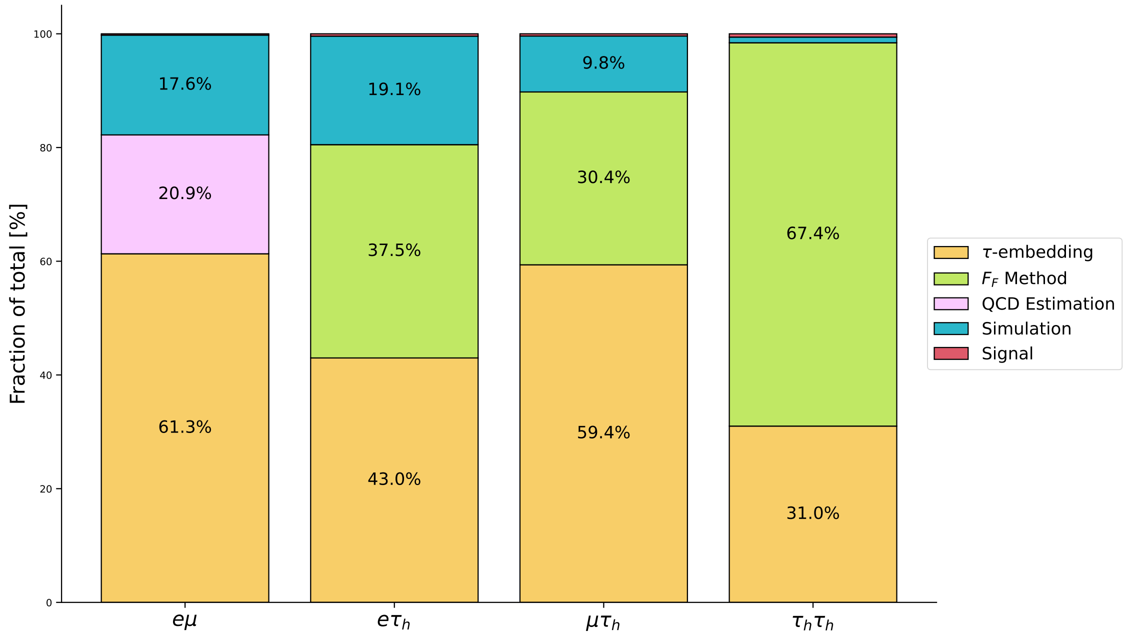


**Figure 4.4.:** Composition of background processes comprised by the methods that are used to estimate these processes in the four main di-τ final states. The signal contribution is always much smaller than 1% and is also estimated using simulation.

### 4.2.3. Simulated background processes

The simulated background processes are generated using several different event generators. The simulation for Z → ττ, Z → ll and W+jets is performed with leading order (LO) precision of the strong coupling constant $\alpha_s$ using MadGraph5_aMC@NLO (version 2.6.5) [88, 89]. For the simulation of the diboson process, the same event generator is used in next-to-leading order (NLO) precision [90]. The $t\bar{t}$ production is simulated using the POWHEG event generator with NLO precision [91–94]. The signal process for ggH [95, 96] and qqH [97] are also generated using the POWHEG event generator with NLO precision. The ggH sample is reweighted to match the next-to-NLO (NNLO) precision using the Nnlops event generator [98, 99]. The reweighing is performed using the $p_T$ of the Higgs boson and the jet multiplicity of the event.

For all event generators, the Nnpdf3.1 [100] set is used for the parton distribution functions. The hadronization, parton showering and τ lepton decays are simulated using Pythia (version 8.2) [101] interfaced with the event generators mentioned above. For the simulation of the underlying event, the CP5 tune [102] is used.

### 4.2.4. The $F_F$ Method

The $F_F$ method is an extrapolation method used to estimate the contribution of jets misidentified as $\tau_h$ candidates, denoted as jet → $\tau_h$. The method is illustrated in Figure 4.5. Within the CMS collaboration, the method was first used in a search for additional MSSM Higgs bosons in the di-τ final state [103] and a $Z/\gamma^* \longrightarrow \tau\tau$ cross section measurement [104]. Since then, the method has been the preferred method for estimating jet ⟶ $\tau_h$ contributions in di-τ analyses. While the general idea is the same, the method must be adapted for each analysis since it depends on the analysis selection, and the bias corrections need to be carefully tuned. A detailed description of the method can be found in [105].

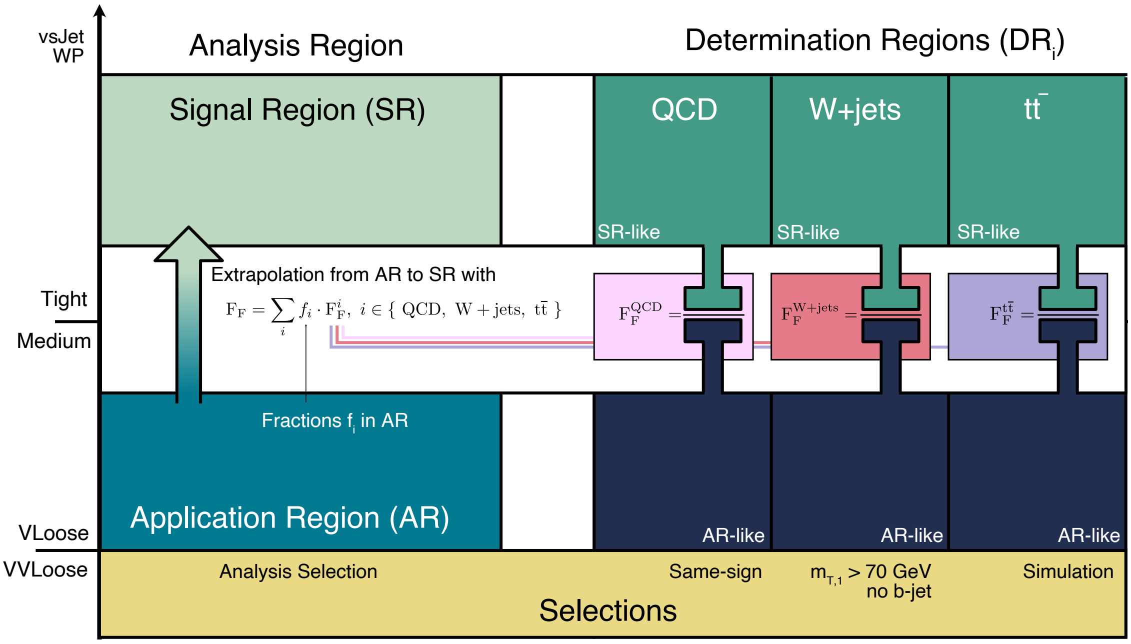


**Figure 4.5.:** Illustration of the $F_F$ method used in the semileptonic channels.

As mentioned in Section 4.2.2, jet ⟶ $\tau_h$ events originating from three different processes are estimated using the $F_F$ method:

1. QCD multijet production in the $e\tau_h$, $\mu\tau_h$, and $\tau_h\tau_h$ final states,
2. $t\bar{t}$ production with at least one hadronically decaying W boson in the $e\tau_h$ and $\mu\tau_h$ final states,
3. W+jets production in the $e\tau_h$ and $\mu\tau_h$ final states.

The basic idea of the method is to determine the contribution of jet ⟶ $\tau_h$ events in an application region (AR) and then extrapolate this contribution to the signal region (SR). The two regions are defined based on the *vsJet* τ lepton identification algorithm, as described in Section 3.3.9. While for the SR, the Tight working point is required, events in the AR must pass the VLoose, but not the Tight working point. As a result, the contribution from jet ⟶ $\tau_h$ events is enhanced in the AR. The contribution from jet ⟶ $\tau_h$ events to the signal region $N_{SR}$ can then be calculated via

$$N_{SR} = N_{AR} \cdot F_F, \tag{4.6}$$

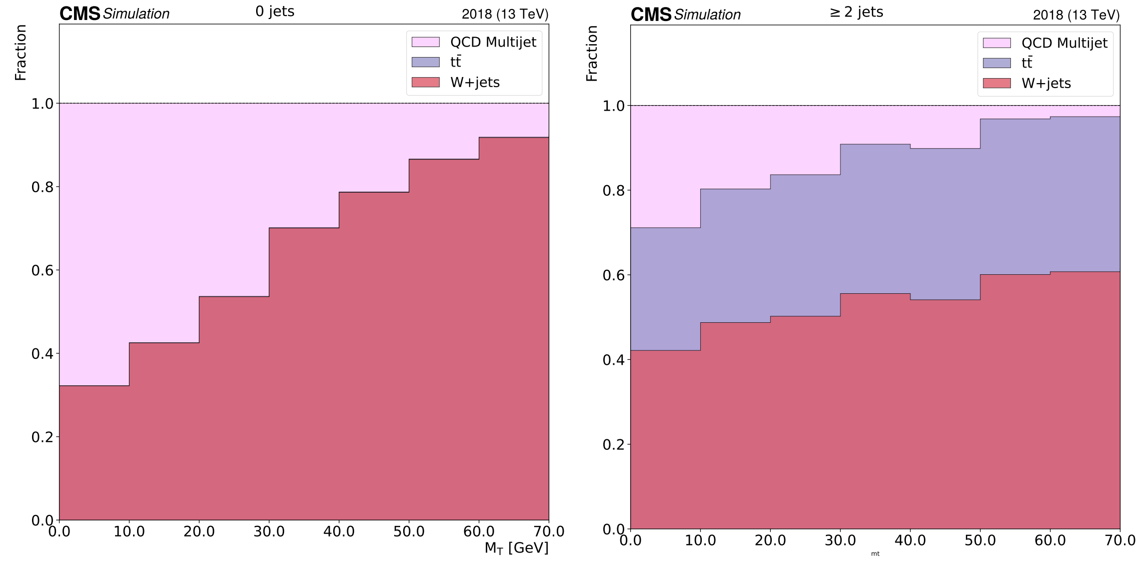


**Figure 4.6.:** Relative contributions (fractions) of the different processes to the signal region in the $e\tau_h$ and $\mu\tau_h$ final state. On the left, the fractions for the 0 Jet case are shown. The fractions for the ≥2 Jet case are shown on the right.

where $N_{\mathrm{AR}}$ is the number of events in the AR after the subtraction of an estimate of genuine $\tau$ lepton contributions, and $\mathrm{F_F}$ is an extrapolation function. The $\mathrm{F_F}$ is determined as a weighted sum

$$\mathrm{F_F} = \sum_i f_i \cdot \mathrm{F_F^{\mathit{i}}},\ i \in \{\ \mathrm{QCD},\ \mathrm{W{+}jets},\ \mathrm{t\bar{t}}\ \}, \tag{4.7}$$

where the fraction $f_i$ represents the probability that an event is of process $i$, and $\mathrm{F_F^{\mathit{i}}}$ is the extrapolation factor for the given process.

The fractions in the AR are estimated using the simulation. Since no simulation is available for QCD multijet production, this is estimated by subtracting the estimations from genuine $\tau$ leptons, W+jets production and $\mathrm{t\bar{t}}$ production from the data in the AR. Afterwards, the obtained fractions are normalized to 1. In the $e\tau_h$ and $\mu\tau_h$ final state, the fractions are measured as a function of $m_{\mathrm{T}}$ of the lepton and the number of jets ([0,1,≥ 2]). In Figure 4.6, the fractions for no jets and ≥ 2 jets are shown. While the $\mathrm{t\bar{t}}$ contribution is negligible in the zero jet case, it rises to ~ 20% in the ≥ 2 jet case. In the $\tau_h\tau_h$ final state, only a QCD multijet contribution is used, calculated as a function of $m_{\mathrm{vis}}$ and the number of jets.

To calculate the $\mathrm{F_F^{\mathit{i}}}$ for the three contributions, determination regions ($\mathrm{DR_i}$) that are orthogonal to the SR and AR are defined. Every $\mathrm{DR_i}$ is designed to obtain the extrapolation factor for that particular process. The extrapolation factors are determined by the ratio of events in the signal-like $\mathrm{DR_i}$ (passing the Tight *vsJet* working point) and the AR-like $\mathrm{DR_i}$ (passing the VLoose but not the Tight *vsJet* working point) region

$$\mathrm{F_F^i} = \frac{\mathrm{DR}^{i}_{\text{SR-like}}}{\mathrm{DR}^{i}_{\text{AR-like}}}. \tag{4.8}$$

The extrapolation factors are measured as a function of several variables, depending on the di-$\tau$ final state and the targeted process. A common dependency on $p_T$ of the $\tau_h$ candidate since the *vsJet* $\tau$ lepton identification efficiency has a strong $p_T$ dependence. A second common dependency is on the number of jets $N_{\text{jets}}$ in the event.

1. For QCD multijet production (used in the $e\tau_h$, $\mu\tau_h$, and $\tau_h\tau_h$ final states), the two selected particles of the pair are required to be of the same charge.The *same-sign* region serves as a QCD multijet control region with a purity of > 99% in the $\tau_h\tau_h$, and ~ 75% in the $e\tau_h$ and $\mu\tau_h$ final states. A smaller contamination from W+jets production is observed in the latter two. The extrapolation factors are measured as a function of the $p_T$ of the $\tau_h$ candidate and the number of jets. Since two $\tau_h$ candidates are present in the $\tau_h\tau_h$ channel, an event can be used twice, once for each $\tau_h$ candidate. Three jet multiplicity regions are used [0,1,$\geq$ 2].

2. For W+jets production (used in the $e\tau_h$ and $\mu\tau_h$ final states), the selection cut on $m_T$ of the lepton is inverted, and the presence of b-jets is vetoed. With this selection, a purity of more than 80% can be achieved. Extrapolation factors are measured as a function of $p_T$ of the $\tau_h$, number of jets, and the angular distance between the $\tau_h$ and the lepton. Three jet multiplicity regions [0,1,$\geq$ 2] , and two angular distance regions [$\Delta R(l, \tau_h < 3), \Delta R(l, \tau_h > 3)$] are used.

3. For $t\bar{t}$ production (used in the $e\tau_h$ and $\mu\tau_h$ final states), the $F_F^{t\bar{t}}$ is taken from simulation, with corrections applied. The extrapolation factors are measured as a function of $p_T$ of the $\tau_h$ and the number of jets. For $t\bar{t}$ production, only two jet bins [$\leq$ 1,$\geq$ 2] are used due to the lack of events in the 0 jet bin.

A closure correction is applied to account for a dependency on the lepton $p_T$ (on the second $\tau_h$ candidate, in the $\tau_h\tau_h$ final state). A second correction is used to remove the biases introduced by the definition of the $DR_i$.

The closure correction accounts for the dependence of the $F_F$ on the $p_T$ of the other $\tau$ pair particle. To not subdivide the $DR_i$ into too many regions, which would result in very few events for the $F_F$ determination and thus increased statistical uncertainties, this closure correction is applied after the determination of the $F_F$. In the $e\tau_h$ and $\mu\tau_h$ final states, the correction is applied to the $F_{F,QCD}$ and the $F_{F,W+jets}$. The closure correction is determined as a function of the electron (muon) $p_T$. In the $\tau_h\tau_h$ final state, the correction is applied to the $F_{F,QCD}$. In this case, the correction is determined as a function of the $p_T$ of the second $\tau_h$ candidate.

The bias correction is determined to account for the transfer from the SR to the $DR_i$. For the $F_{F,W+jets}$; this correction removes the bias introduced by using a different $m_T$ region. It is calculated using simulation. It is determined as a function of $m_{\text{vis}}$.

For the $F_{F,QCD}$, the correction addresses the bias introduced by the same-sign requirement. Since no suitable QCD multijet simulation is available, the correction is determined using an additional control region. Within this control region, the lepton is required to be non-isolated ($I_{\text{rel}}^{e,\mu} \in [0.15, 0.25]$). This control region is still orthogonal to the SR, where only well-isolated leptons are used. Depending on the lepton isolation, a second bias correction to the $F_{F,QCD}$ is used to address the bias for this control region.

In the $\tau_h\tau_h$ final state, where no lepton isolation can be used, the second $\tau_h$ candidate is used instead to obtain the control region. For the $F_{F,QCD}$, the bias correction is determined as a function of $m_{vis}$ and the lepton isolation (second $\tau_h$ candidate $p_T$) in the $e\tau_h$ and $\mu\tau_h$ ($\tau_h\tau_h$) final state(s).

In summary, the $F_F$ functions, including these corrections, are given as functions of the following variables:

| | |
|---|---|
| $\mathrm{F}_\mathrm{F}^\mathrm{QCD}\left(p_\mathrm{T}^{\tau_\mathrm{h}},\ \mathrm{N}_\mathrm{jets},\ I_\mathrm{rel}^{\mathrm{e},\mu},\ p_\mathrm{T}^{l},\ m_\mathrm{vis}\right)$ | $e\tau_h$ and $\mu\tau_h$ final states |
| $\mathrm{F}_\mathrm{F}^\mathrm{QCD}\left(p_\mathrm{T}^{\tau_\mathrm{h,1}},\ \mathrm{N}_\mathrm{jets},\ p_\mathrm{T}^{\tau_\mathrm{h,1}},\ m_\mathrm{vis}\right)$ | $\tau_h\tau_h$ final state |
| $\mathrm{F}_\mathrm{F}^\mathrm{W+jets}\left(p_\mathrm{T}^{\tau_\mathrm{h}},\ \mathrm{N}_\mathrm{jets},\ \Delta\mathrm{R}(l,\tau_\mathrm{h}),\ p_\mathrm{T}^{l},\ m_\mathrm{vis}\right)$ | $e\tau_h$ and $\mu\tau_h$ final states |
| $\mathrm{F}_\mathrm{F}^{\mathrm{t\bar{t}}}\left(p_\mathrm{T}^{\tau_\mathrm{h}},\mathrm{N}_\mathrm{jets},m_\mathrm{vis}\right)$ | $e\tau_h$ and $\mu\tau_h$ final states |

# 5 The τ-embedding method

The τ-*embedding method* is used to create a data-driven estimate of all processes with two genuine τ leptons in the final state. The resulting samples can be used as a replacement for simulated samples. On an event-by-event basis, events with two muons are selected from a recorded event in data, and the muons are replaced with simulated τ lepton decays. This way, the method mainly relies on measured collision events and only the decays of the τ leptons and their energy deposits in the CMS detector have to be simulated. The method results in more accurate modelling of several event properties, such as pileup or additional jets in the event, which requires a significant amount of tuning for regular fully simulated samples.

The process of the τ-embedding method can be split into four steps: the selection of two muons, the removal of the selected muons resulting in a cleaned event, the simulation of two τ lepton decays, and the combination of the simulation with the cleaned event. The workflow of the τ-embedding method is visualized in Figure 5.1.

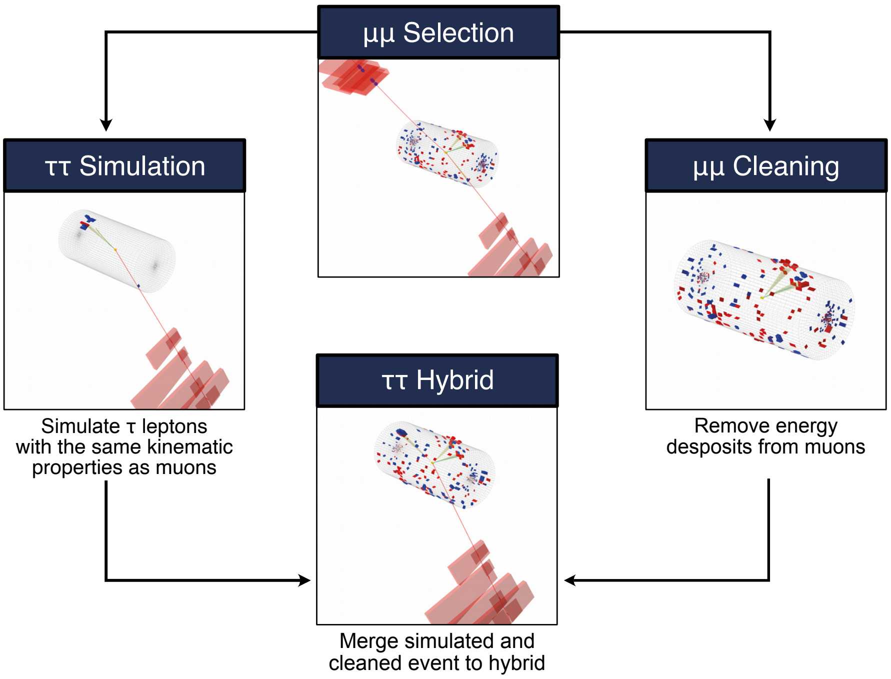

μμ Selection
ττ Simulation
μμ Cleaning
ττ Hybrid
Simulate τ leptons with the same kinematic properties as muons
Remove energy desposits from muons
Merge simulated and cleaned event to hybrid

**Figure 5.1.:** A visualization of the τ-embedding workflow. The first step is the selection of two muons. In the second and third steps, the removal of the selected muons from the event and the simulation of two τ lepton decays can, in principle, be executed at the same time. The final step is the combination of the simulation with the cleaned event. Adapted from [8].

The method was originally introduced by the CMS collaboration during Run I [106–108] and then completely reworked at the beginning of Run II. The ATLAS collaboration used a similar method during Run I [109], and a modified version which relies on distribution reweighing rather than event-by-event processing during Run II [110]. A description of the implementation used in CMS during Run II is given in [8]. It was developed, maintained and improved in the scope of this thesis and multiple master theses [111–114]. An in-depth discussion of the latest version of the method will be given in this chapter. Some of the analyses that used τ-embedded samples were outlined in Section 4.1.

The τ-embedding method is used to create six different event samples. An event sample produced using the τ-embedding method will be called a *τ-embedded sample* in the following. The four major τ-embedded samples corresponding to the four main di-τ final states $e\tau_h$, $\mu\tau_h$, $\tau_h\tau_h$, and $e\mu$, are created by enforcing the given final state during the τ lepton decay simulation. In addition, two τ-embedded samples used for the calculation of correction factors are created:

- a sample where the selected muons are replaced with simulated muons referred to as μ-Embedding ($\mu \rightarrow \mu$),
- a sample where the selected muons are replaced with simulated electrons referred to as e-Embedding ($\mu \rightarrow e$).

The $\mu \rightarrow \mu$ embedded sample has an additional benefit: it can be used to validate any biases introduced by the method since, in the ideal case, the muons in the $\mu \rightarrow \mu$ embedded event should have the same properties as the initial event. The selection and the cleaning step are identical for all six samples. The only difference between the sample processing occurs during the simulation step.

The version of the τ-embedding method that is described in this chapter is based on the latest reprocessing of the Run II data set of the CMS Collaboration. This data set is denoted as UL and was introduced to have a consistent version of the Run II data, that can be used for combinations with subsequent data sets. The UL data set should contain the best knowledge of all detector inefficiencies or defects. Due to this reprocessing, new τ-embedded samples, new simulated samples, and corrections for both are required. If not explicitly stated otherwise, all studies presented in this thesis are based on the UL data set.

## 5.1. Muon Selection

All events containing at least two muons reconstructed in the CMS detector are considered as input for τ-embedding. A list of the used data sets can be found in Table A.1 of the Appendix and are from now on denoted as DoubleMuon data sets. A set of selection criteria listed in Table 5.1 is applied, to select as many di-μ pairs as possible. The selection is chosen to be as inclusive as possible to minimize the need for additional corrections due to any selection biases.

The chosen HLT path is the lowest, unprescaled trigger path available. Other than the muon isolation requirement of the HLT path of $I^{\mu}_{\text{rel}} < 0.4$, no isolation requirement is set. Selecting only muons of opposite charge ensures that both muons originate from the

**Table 5.1.:** A list of the selection criteria applied for producing τ-embedded samples for Run II.

| Description | Selection Criteria |
|---|---|
| $p_\mathrm{T}$ of leading muon | $\geq 17\,\mathrm{GeV}$ |
| muon reconstruction quality | isGlobal and isLoose |
| mass of dimuon pair | $\geq 20\,\mathrm{GeV}$ |
| charge of two muons | opposite sign |
| muon multiplicity | $\geq 2$ |
| HLT path | Double Muon Trigger [1] |

same decay chain. The requirement of a globally reconstructed muon track minimises the selection of particles misreconstructed as muons. The reconstructed $p_\mathrm{T}$ of the leading muon and mass of the di-μ pair is chosen in line with the HLT path requirements.

After the selection, roughly 25% of all events in the DoubleMuon data set are selected, with corresponds to a total of 78.8 million events in the 2018 data set. To check the composition of different processes contributing to the event selection, the same selection criteria are applied to the simulation. To estimate the contribution from QCD multijet production, a simple extrapolation of di-μ events from a same-sign control region is used. In Figure 5.2, the composition for the 2018 data set is shown as a function of the di-μ mass ($m_{\mu\mu}$).

The vast majority of the selected events originate from Z bosons decaying into two muons. Since the muon and τ lepton couplings in the SM are identical, apart from the mass of the lepton, the cross section of Z → ττ and Z → μμ is identical. Since the mass of the Z boson is much larger than the mass of a τ lepton or muon, the difference between the cross section of the two decays is minimal.

The same is valid for contributions from leptonic $\mathrm{t\bar{t}}$ production ($\mathrm{t\bar{t}}(\mu\mu)$) and leptonic diboson production (VV(μμ)). In both cases, the two selected muons are produced via weak interactions from a W or Z boson decay, so the cross section is again identical if a

**Table 5.2.:** Event composition of the different processes contributing to the selected events in the 2018 data set. The composition is listed for events with a di-μ mass larger and smaller than 250 GeV.

| Process | Composition $m_{\mu\mu} < 250\,\mathrm{GeV}$ | Composition $m_{\mu\mu} > 250\,\mathrm{GeV}$ |
|---|---|---|
| VV(μμ) | 0.27 % | 6.13 % |
| W+jets | 0.19 % | 0.56 % |
| $\mathrm{t\bar{t}}(\mu\mu)$ | 0.89 % | 26.26 % |
| QCD multijet | 5.67 % | 2.49 % |
| $\mathrm{Z} \rightarrow \mu\mu$ | 92.28 % | 64.46 % |
| $\mathrm{Z} \rightarrow \tau(\mu)\tau(\mu)$ | 0.70 % | 0.10 % |

[1] HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_v* OR HLT_Mu17_TrkIsoVVL_Mu8_TrkIsoVVL_DZ_Mass8_v*

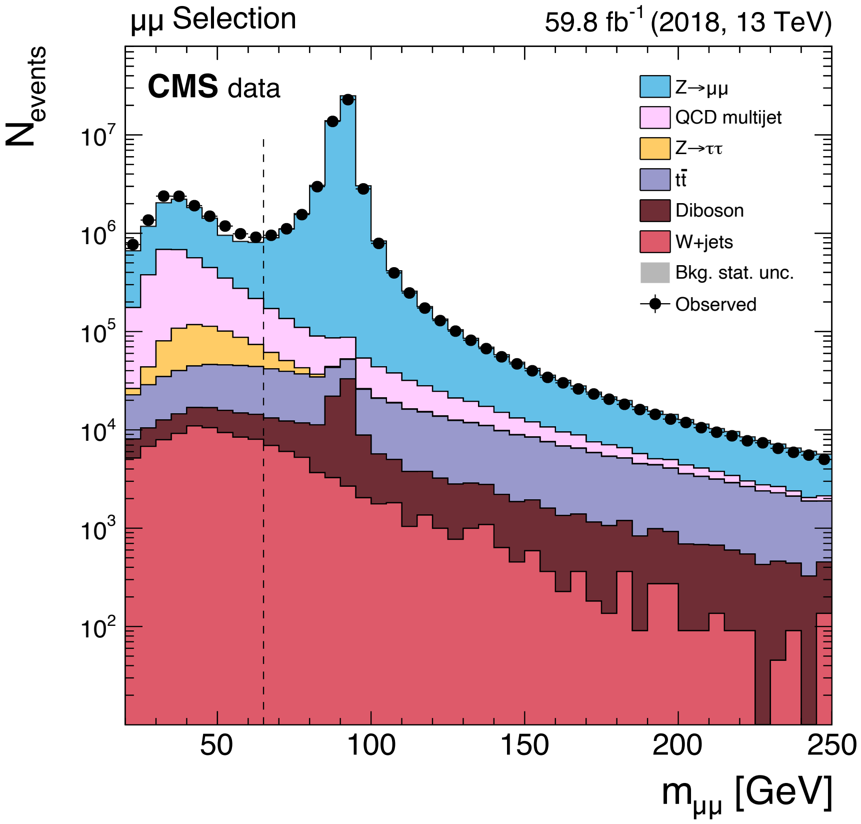


**Figure 5.2.:** Event composition after applying the muon selection of the τ-embedding method for the 2018 data set. The contributions from the different physics processes are estimated using simulation. The vertical line represents 65 GeV. At low mass, a small discrepancy between the data and simulation is visible, due to the QCD multijet extrapolation.

τ lepton replaces the muon. For all mentioned processes, the interaction is mediated by a W or Z boson. Therefore, no modification of the normalization of the individual process is needed, when replacing the muons with τ leptons.

With a proportion of 5.6%, the second-largest contribution to the selection comes from QCD multijet production, where muons can be produced via leptonic decays in the jet, or light quark- and gluon-induced jets may be misidentified as muons. The largest contribution from QCD multijet production is located in the lower tail of the di-μ mass spectrum, as shown in Table 5.2. This effect also holds for dimuon pairs originating from W+jets production. Such events are selected if the W boson decays into a muon and another jet in the event is misidentified as a muon.

QCD multijet and W+jets production do not have the same cross section when replacing a muon by a τ lepton. The probability of producing a muon in a jet is much higher than the probability of producing a τ lepton. In the jet, leptons can be produced via the decay of intermediate hadrons, mediated by a virtual W boson. However, the masses of these hadrons are much smaller than the W and Z boson masses; therefore, the mass difference

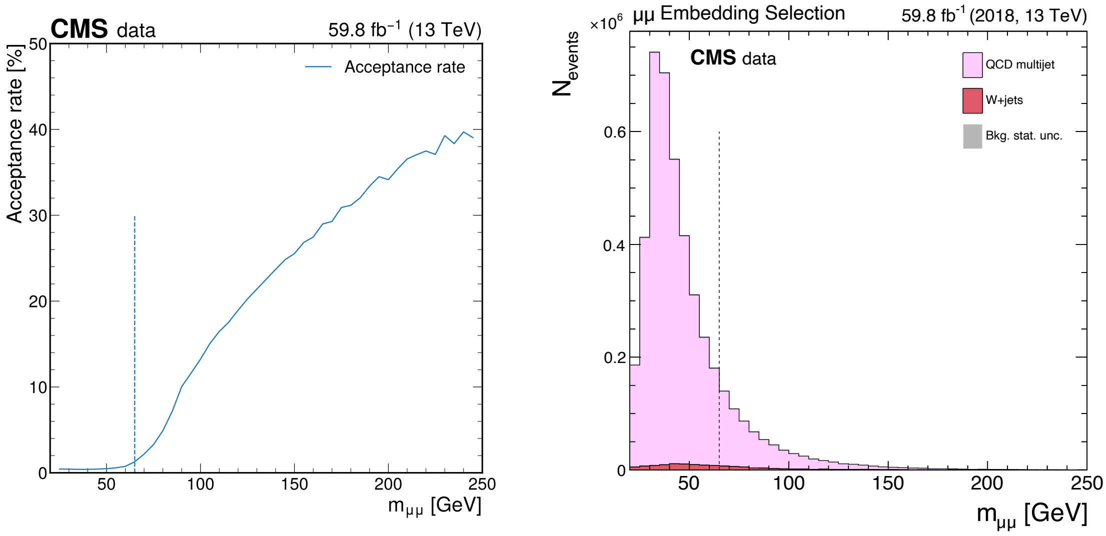


**Figure 5.3.:** The left plot shows the acceptance rate for events in the $\mu\tau_h$ final state using the analysis selection requirements described in Section 4.2.1. The right Figure shows the distribution of events from QCD multijet, W+jets and $Z \longrightarrow \tau\tau$ after the selection. In both figures, the vertical line represents the threshold of $m_{\mu\mu}$ = 65 GeV. Below this threshold, the acceptance rate shown on the left is well below 1% while 85% of the QCD multijet, 64% of the W+jets, and 83% of the $Z \longrightarrow \tau\tau$ contributions are also below.

between the τ lepton and the muon affects the available phase space and the branching ratio. Since the τ lepton is heavier than the muon, the cross section is smaller.

The smallest contribution comes from genuine $Z \longrightarrow \tau\tau$ events, where both τ leptons subsequently decay into a muon ($Z \longrightarrow \tau(\mu)\tau(\mu)$). These events generally have a smaller di-μ mass since the muons are accompanied by four neutrinos, which carry a significant amount of the available energy. The contributions from QCD multijet, W+jets, and $Z \longrightarrow \tau\tau$ events result in an overestimation of genuine di-μ events and, therefore, a potential overestimation of di-τ events.

While the initial selection of di-μ events is chosen as loose as possible, only a small subset of these events end up in a typical target analysis. If a muon isolation $I^{\mu}_{\text{rel}} < 0.15$ is required, the contribution from QCD multijet production to the selection drops to less than 1%. The acceptance rate of events from an τ-embedded sample in the $\mu\tau_h$ channel in a typical taget analysis over the di-μ mass is shown in Figure 5.3(left). Below a di-μ mass of 65 GeV, less than 1% of the events in the τ-embedded sample are accepted. This is due to the $p_T$ and isolation selection of the typical target analysis. In addition 85% of QCD multijet, 64% of W+jets, and 83 % of $Z \longrightarrow \tau\tau$ events have a di-μ mass of less than 65 GeV. The distributions are shown in Figure 5.3(right). As a result, the contributions from these three have a high probability of not being accepted in a typical target analysis. While a non-negligible portion of di-μ events selected originates from QCD multijet production, these events will not end up in a typical target analysis and no additional measures have to be taken.

The vast majority of selected events contain exactly two muons. Only 0.92% of all events contain more than two muons. However, if more than two muons can be found in an event, a decision, on which muons will be selected, must be made. One approach is to select the two muons, with $m_{\mu\mu}$ closest to the Z boson mass. This approach is correct for events where the muons originate from a Z boson decay. However, for $t\bar{t}$ and diboson production, it is not and no clear hypothesis on the di-μ mass can be made since the muons originate from separate W boson or top quark decays. A selection based on this approach introduces a bias towards mass pairs with a combined mass close to the Z boson mass and will be called Z boson hypothesis in the following.

An alternative approach to avoiding this bias is to select the muons with the largest $m_{\mu\mu}$ in the event. This approach will be called the largest mass hypothesis and was used in the most recent iteration of the τ-embedding. To validate, that using the largest mass hypothesis is the less biased approach, a comparison using simulation was performed. By using the generator information, it is possible to check, that the selected muons truly originate from the W bosons or top quarks. No change was observed for $VV(\mu\mu)$ since the number of events with more than two muons is very close to zero. For $t\bar{t}(\mu\mu)$ however, about 10% of all events contain more than two muons, which means a change in the selection algorithm will only affect top quark pair decays.

The results of the approaches are shown in Figure 5.4. For the Z boson hypothesis (top left), a clear bias towards a mass of 90 GeV is visible, whereas the largest mass hypothesis results in an unbiased distribution (top right). The main difference is the selection of the subleading muon as shown in Figure 5.4 (bottom row). For the Z boson hypothesis, a leading muon with large $p_T$ has to be combined with a low energy muon, to obtain a di-μ pair with a mass close to the Z boson mass. This low-energy muon does not originate from a leptonic top quark decay but most probably originates from one of the b-jets in the event and therefore has a smaller $p_T$. For the largest mass approach, the subleading muon can also have a large $p_T$. The distributions for the $p_T$ of the leading muon and η for the leading and subleading muon are shown in the Appendix in Figures A.2 to A.4.

The rate of correctly selected muons from $t\bar{t}(\mu\mu)$ increases from 90.36% to 95.71%. Together with the removed bias towards the Z boson mass, this leads to a more accurately modelled region above $m_{\mu\mu}$ >250 GeV. As listed in Table 5.2, 26.26% of all selected events in this region originate from $t\bar{t}(\mu\mu)$. Nearly 90% of these $t\bar{t}(\mu\mu)$ events contain well-isolated muons. Since this region is of special interest for BSM physics searches and is only very sparsely populated, accurate modelling of the high $m_{\mu\mu}$ region is important.

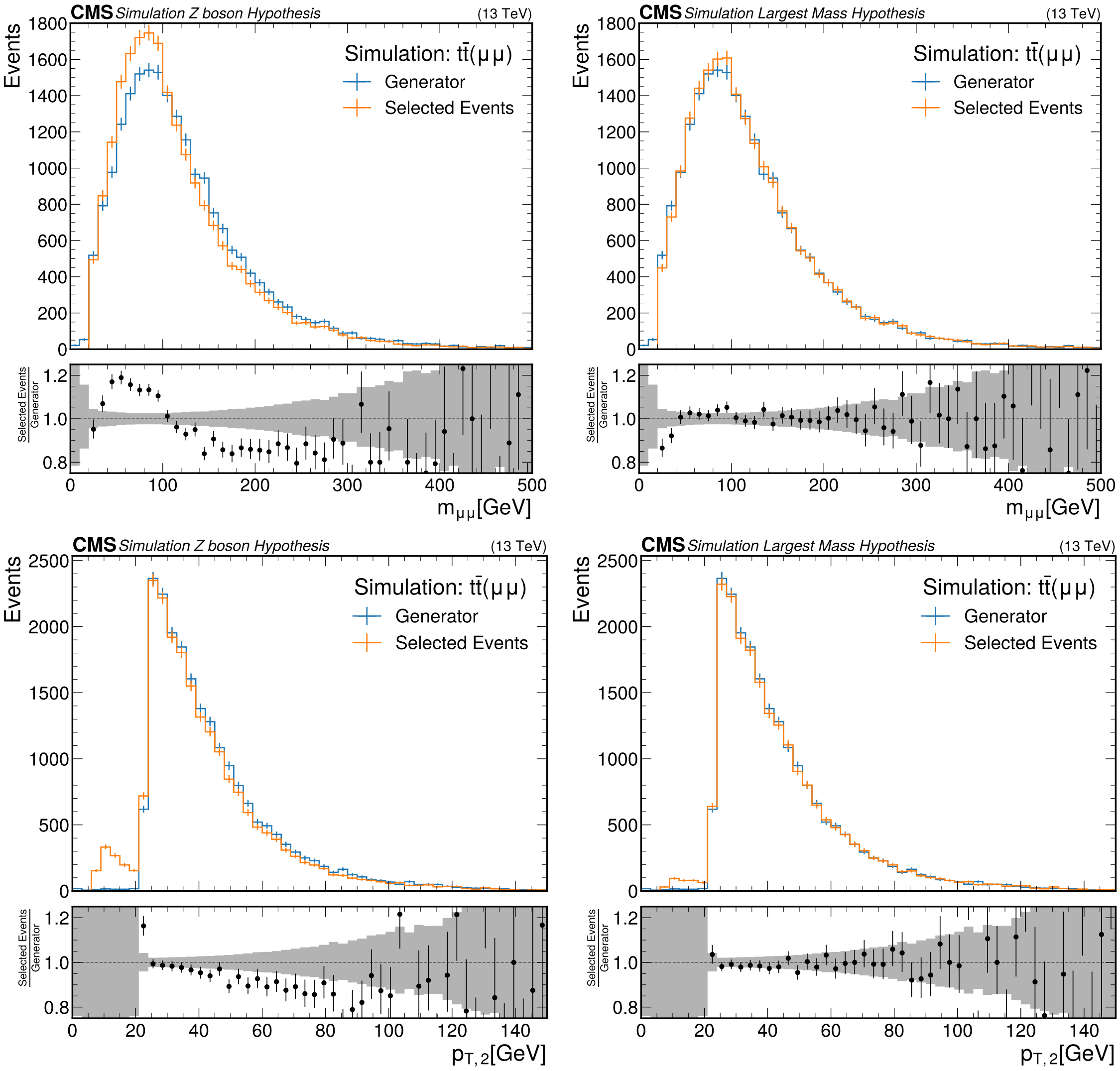


**Figure 5.4.:** Comparison between the Z boson mass (left row) and largest mass (right row) hypothesis during the selection step. The generator distribution represents the information of the true di-μ system used for the simulation. In the top row, $m_{\mu\mu}$ obtained from the selected dimuon system is shown, while in the bottom two, the $p_T$ of the subleading muon is shown. While a mismodelling is visible in the left plots, the agreement between the selected system and the generator is improved in the right plots. The distribution for additional variables can be found in Appendix A.

## 5.2. Cleaning of Muon Energy Deposits

After the selection, all energy deposits from the selected muons are removed from the event record by removing all detector signals associated with the muons. After the modification of the detector signals, the complete event reconstruction is rerun. This procedure will be referred to as *cleaning* in the following, and the reconstructed event with the selected muons removed is referred to as *Cleaned event*. A visualization of an event before and after the cleaning is shown in Figure 5.5.

All hits in the pixel and the strip detectors associated with the muons are removed. The associated hits are directly connected to the fitted tracks related to the muons during the particle flow algorithm. The same is done for hits in the DT, the CSC, and the RPC.

Since there is no clear association between calorimeter towers and muons, removing energy deposits in both the ECAL and the HCAL is only done implicitely. All ECAL clusters and HCAL towers crossed by the muon can be identified by using the trajectory of the global muon track. In the ECAL, all cell energies in the EB, EE and ES are set to zero if they are crossed by a muon trajectory. The same is done for the HB, HE, and HF towers of the HCAL. More sophisticated approaches such as relative energy removal based on the estimated energy loss due to the distance travelled in a tower were tested but did not yield any improved description. Removing all energy in the calorimeter cell ensures that no other particles can be reconstructed in the same place after the cleaning, which would affect the isolation of the embedded τ leptons. The chosen approach does not account for energy contributions from other particles in the affected calorimeter cells, and also does not take into account energy deposits from the muons themselves in the surrounding cells. However, since the calorimeters are sufficiently granular, the effect on the isolation of the embedded τ leptons is small. Nevertheless, as shown in Section 5.6, a shift in the isolation is visible in the μ → μ embedded validation.

Per muon an average energy of 0.65 GeV (8.25 GeV) is removed from the ECAL (HCAL). In addition, this procedure removes up to 30 hits per muon from the silicon tracker. The distribution of the number of removed hits and the energy of the removed ECAL clusters and HCAL towers is shown in Figure A.1 in the Appendix.

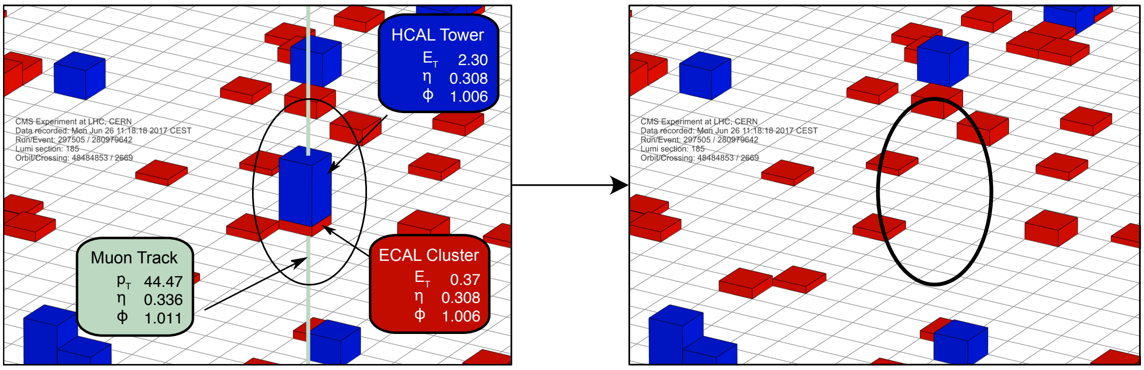

HCAL Tower
$E_T$ 2.30
η 0.308
φ 1.006
ECAL Cluster
$E_T$ 0.37
η 0.308
φ 1.006
Muon Track
$p_T$ 44.47
η 0.336
φ 1.011
CMS Experiment at LHC, CERN
Data recorded: Mon Jun 26 11:18:18 2017 CEST
Run/Event: 297505 / 280979642
Lumi section: 185
Orbit/Crossing: 48484853 / 2669

**Figure 5.5.:** A visualization of an event before and after the cleaning. The selected muon is removed from the event by removing all energy deposits associated with the muon and then rerunning the full event reconstruction.

## 5.3. Simulation of τ Lepton Decays

In this step, the decay of the two τ leptons is performed. A sketch of the workflow is shown in Figure 5.6 and will be explained in the following.

The kinematic properties of the selected muons are used to set up the simulation. The muons are assumed to originate from a two-body decay and to be perfectly reconstructed. For technical reasons, the simulated τ lepton decays have to originate from a common mother particle. For this, a Z boson with a Lorentz vector equal to the Lorentz vector of the dimuon system is chosen. As a result, the decay

$$\mathrm{Z} \longrightarrow \tau_1 \tau_2 \tag{5.1}$$

is defined, where the Lorentz vectors of $\tau_1$ and $\tau_2$ are calculated based on the Lorentz vectors of the selected muons. To obtain the Lorentz vectors of $\tau_1$ and $\tau_2$, a boost into the dimuon rest frame is performed. Then, a correction due to the mass difference between $m_\mu$ and $m_\tau$ is applied to the momentum components of the τ lepton Lorentz vector:

$$c_{\text{mass}} = \sqrt{\frac{(0.5 \cdot \mathrm{m}_{\mu\mu})^2 - m_\tau^2}{|\vec{p}_\mu|^2}} \tag{5.2}$$

where $0.5{\cdot}\mathrm{m}_{\mu\mu}$ is the energy of each muon in the dimuon rest frame, $m_\tau$ is the mass of the τ lepton, and $|\vec{p}_\mu|$ is the magnitude of the momentum of the muon in the dimuon rest frame. The energy of each τ lepton is identical to the energy of the corresponding muon; only the momentum is slightly adapted to account for the mass difference. Since both the τ lepton and the muon mass are much smaller than the particle momenta, the value of $c_{\text{mass}}$ is close to one, which corresponds to a momentum correction of $\mathcal{O}(100\,\text{MeV})$. This correction is negligible compared to the energy carried by the $\nu_\tau$ in the τ lepton decay, which is of $\mathcal{O}(10\,\text{GeV})$ but applied for completeness.

After the Lorentz vectors for the τ leptons are computed, the simulation of the decays is performed using PYTHIA 8.2 [101]. The simulation is performed inclusively without restrictions. Spin correlations and helicity effects, as well as all possible τ lepton decay modes with branching fractions $\mathcal{B} > 0.04$ are included in the τ lepton decay model used in PYTHIA 8.2. The most common decay modes are listed in Table 2.2.

In principle, a single simulation trial would be sufficient. To increase the number of di-τ decays ending up in a typical target analysis, two aspects have to be addressed:

1. The simulation is performed inclusively for all possible decay modes. Using only one simulation trial implies, that only a small fraction of decays end up e.g. in the eμ final state, whereas most decays would end up in the $\tau_h\tau_h$ final state.

2. During every τ lepton decay, at least one neutrino is produced, which leaves the detector undetected and can only be identified as MET. As a result, only a small fraction of the initial τ lepton energy may be visible in the detector. Since the selection of events for analyses mainly relies on the visible energy in the detector, events with large neutrino momenta have a high probability of being rejected by a typical analysis selection.

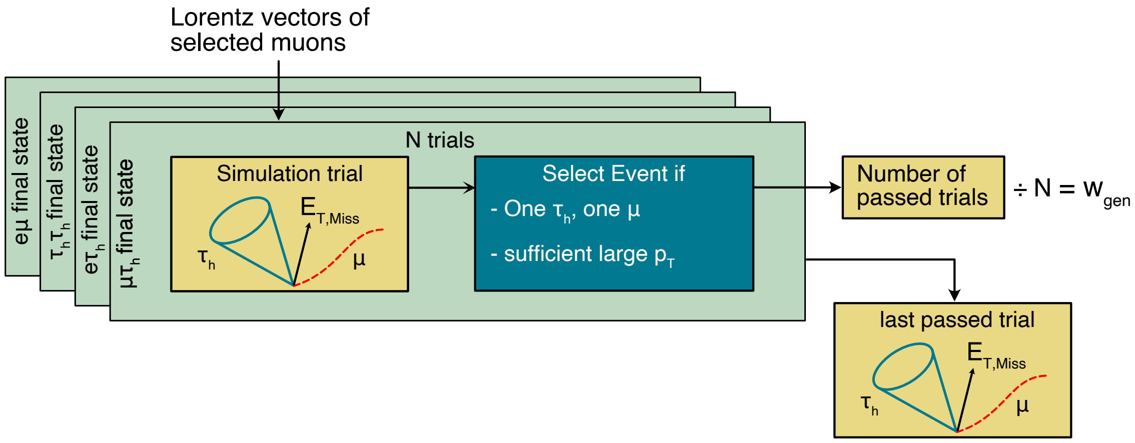


**Figure 5.6.:** A sketch of the simulation step of the τ-embedding. For 1000 simulation trials, a set of generator cuts is applied while counting the number of passing trials. After that, the fraction $w_{gen}$ is calculated, and the last successful simulation trial is used.

Applying a filter called *generator cuts* on the simulation outcome makes it possible to mitigate the first disadvantage. The filtering allows splitting the data into multiple τ-embedded samples, depending on the desired final state. For example, to enforce the $\mu\tau_h$ final state, one τ lepton has to decay into a muon, whereas the other τ lepton must decay into hadrons. Since the neutrinos produced during the τ lepton decays result in a shuffle of the τ lepton kinematics, every selected event can be reused multiple times, once per final state. The number of simulation trials can be increased to ensure that every final state is simulated at least once.

Using multiple simulation trials also helps to tackle the second aspect mentioned above. When enforcing the final state via generator cuts, it is possible to accept only those trials, where the visible decay products obtain a large portion of the initial τ lepton energy. These cuts increase the chance of simulating a constellation with a large amount of visible energy in the detector. In Figure 5.7 the visible $p_T$ of 10 million simulation trials while enforcing the $\mu\tau_h$ final state is shown. For the leptonic decay, the two neutrinos carry a significant fraction of the initial τ lepton energy, whereas, for the hadronic decay, the average visible $p_T$ is larger since only one neutrino is produced.

To summarise, the entire data set of selected events is reused six times, once for each of the four di-τ final states, as well as for the $\mu \rightarrow e$ and $\mu \rightarrow \mu$ embedded samples used for calibration. For every type of τ-embedded sample, a specific set of *generator cuts* is applied. The generator cuts on the $p_T$ and η of the visible τ lepton decay products are listed in Table 5.3. They align with the lowest available HLT path thresholds in each corresponding final state.

The main benefit of using multiple simulation trials is that the acceptance rate of τ embedded events is significantly increased in the phase space of a typical target analysis. It results in a large oversampling compared to the number of data events selected in the same phase space between 3 and 40, depending on the di-τ final state.

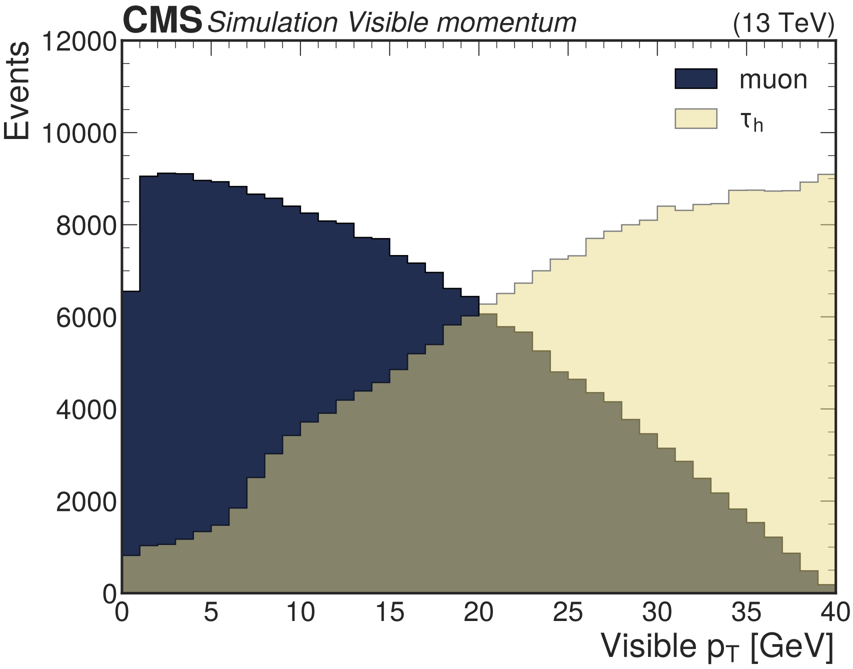


**Figure 5.7.:** Distribution of the visible $p_T$ of the τ lepton decay products using 10 million simulation trials where the $\mu\tau_h$ final state was enforced. The $p_T$ of both τ leptons is set to $p_{T,1} = 40$ GeV.

Setting the number of simulation trials to 1000 is sufficient to obtain a high acceptance rate for each desired final state while sustaining an acceptable processing time. The last simulation trial, which fulfils all *generator cuts*, is used and propagated to the detector simulation and reconstruction.

Since the repetition of the simulation combined with the generator cuts introduces a bias towards events with low neutrino energies, an additional generator weight $w_{\text{gen}}$ is calculated to reweight each event that passes the generator selection successfully. The weight is calculated as:

$$w_{\text{gen}} = \frac{N_{\text{passed}}}{N_{\text{trials}}} = \frac{N_{\text{passed}}}{1000}, \tag{5.3}$$

where $N_{\text{passed}}$ is the number of simulation trials that passed the generator cuts, and $N_{\text{trials}}$ is the number of simulation trials. Since the τ lepton simulation is performed inclusively, branching fractions for the different final states are automatically included in the generator weight and correspond to the maximum value that $w_{\text{gen}}$ can reach, if an infinite amount of trials was performed. The distribution of $w_{\text{gen}}$ is visualized in Figure 5.8. Here, in some rare cases, the value of $w_{\text{gen}}$ can be above to branching fraction, since only 1000 simulation trials are performed.

At first glance, using the same original di-μ events in more than one final state might introduce a statistical dependence of τ-embedded samples from different final states. However, a simulation trial in the $\mu\tau_h$ and the $e\tau_h$ final state coming from the same

**Table 5.3.:** Table of the selection criteria applied on the $p_T$ and $\eta$ of the visible decay products for the di-τ final states and the $\mu \rightarrow e$ and $\mu \rightarrow \mu$ embedded samples. Index 1 corresponds to the first particle in the final state name, and index 2 to the second particle. For $\tau_h\tau_h$, $\mu \rightarrow e$, and $\mu \rightarrow \mu$ embedded samples, index 1 corresponds to the lepton with the larger, index 2 to the lepton with the smaller $p_T$.

| Final State | $p_{T,1}$ [GeV] | $p_{T,2}$ [GeV] | $\lvert\eta_1\rvert$ | $\lvert\eta_2\rvert$ |
|---|---|---|---|---|
| $e\mu$ | > 9 | > 19 | < 2.5 | < 2.5 |
| $\mu\tau_h$ | > 18 | > 18 | < 2.2 | < 2.4 |
| $e\tau_h$ | > 18 | > 18 | < 2.2 | < 2.4 |
| $\tau_h\tau_h$ | > 20 | > 20 | < 2.4 | < 2.4 |
| $\mu \rightarrow e$ | > 22 | > 10 | < 2.5 | < 2.5 |
| $\mu \rightarrow \mu$ | > 17 | > 8 | < 2.5 | < 2.5 |

selected event will result in other kinematic properties of the τ leptons. In Figure 5.9 (top row), the comparison of the $p_T$ of the $\tau_h$ in the $e\tau_h$ and $\mu\tau_h$ final states, as well as the $p_T$ of the leading jet, are shown. Only events that are included in both τ-embedded samples are included. The jets, that are untouched by the method, result in two compatible distributions for the $e\tau_h$ and the $\mu\tau_h$ sample. On the histogram level, the shifts in the $\tau_h$ momentum due to the kinematic shuffling are also small. However, when comparing the $\tau_h$ $p_T$ on an event-by-event basis, as shown in Figure 5.9 (bottom right), the kinematic shuffling is visible. On average, a difference of 9.39 GeV and a correlation of 32% is observed. Although the final distributions are similar, individual events can have a completely different value for the $p_T$ of the $\tau_h$. In applications where not only a single variable but the whole event's content is of interest, e.g. neural network training, these event-by-event differences are relevant.

In the case of the $e\tau_h$ and $\mu\tau_h$ final state, the overlap of events common in both samples after applying an analysis selected is 38% of all $\mu\tau_h$ events (52% of all $e\tau_h$ events), as shown in Figure 5.9 (bottom left). While this common part of events has similar kinematic properties, the additional events in each sample will reshape the distributions to be different from each other. The individual events in different final states are distinct enough on an event-by-event and whole-sample basis that the different τ-embedded samples can be considered statistically independent.

After a successful trial was selected, the simulated di-τ decay is propagated through the regular CMS detector simulation, the simulation of the HLT response, and the event reconstruction. The processing is done the same way for simulated samples, with some necessary modifications applied to the reconstruction sequence.

The reconstruction of the di-τ decay is performed in an otherwise empty detector. As a result, some additional effects have to be taken into account. One example is the reconstruction of the PV. As shown in Figure 3.9, few tracks will result in a poor PV determination. Since the di-τ decay products only result in a handful of tracks, the reconstructed PV is not very accurate. Instead of rerunning the PV reconstruction algorithm on the simulated τ lepton decay, the PV is set to the location determined from the input event during the

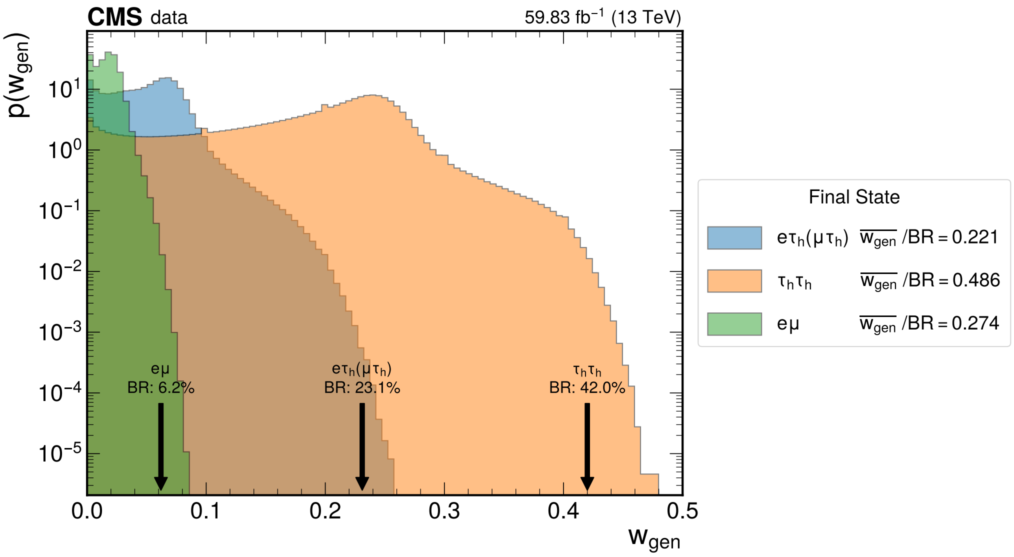


**Figure 5.8.:** The distribution of the generator weights $w_{\text{gen}}$ for the τ-embedded samples in the four primary di-τ final states. Since the branching fraction of each di-τ final state is directly included in $w_{\text{gen}}$, the maximal possible value corresponds to the branching fraction. In rare cases, higher $w_{\text{gen}}$ values can be obtained, since only 1000 trial simulations are performed.

selection step. The same replacement is done for the simulation of the HLT response, where a less complex version of the PV reconstruction is used.

In the simulation, the interaction point is set to the origin of the detector coordinate system. However, in data, the position the interaction point is measured and can change depending on the conditions of the LHC beam. To make the combination of the simulation with the cleaned event easier, the interaction point of the simulation is set to the same location as determined during the selection step.

Another implication of the chosen simulation setup is that only HLT paths sensitive to the two τ lepton decays will lead to meaningful results. More complicated HLT paths, such as a di-τ pair together with additional MET in the event, cannot be simulated correctly since the rest of the event content is not accessible during the HLT simulation. Apart from that, the HLT response for HLT paths only sensitive to the τ lepton decays may still look different from the HLT response in data. Typically, the efficiency of HLT paths in τ-embedded samples is higher than in data since the reconstruction and identification of particles are less challenging in an otherwise empty detector. It is necessary to derive a set of τ-embedding specific corrections to account for these differences. These corrections are described in Chapter 6. A second observed effect is that the HLT efficiency is much lower when the reconstruction of $\tau_h$ on the HLT level is performed. While some attempts to mitigate this effect have been made [114], the efficiency drop is still visible. As a solution, not the full HLT path is used, but only an intermediate result of the HLT filter sequence. The bias introduced by this intermediate filter usage is mitigated, by using τ-embedding specific HLT corrections.

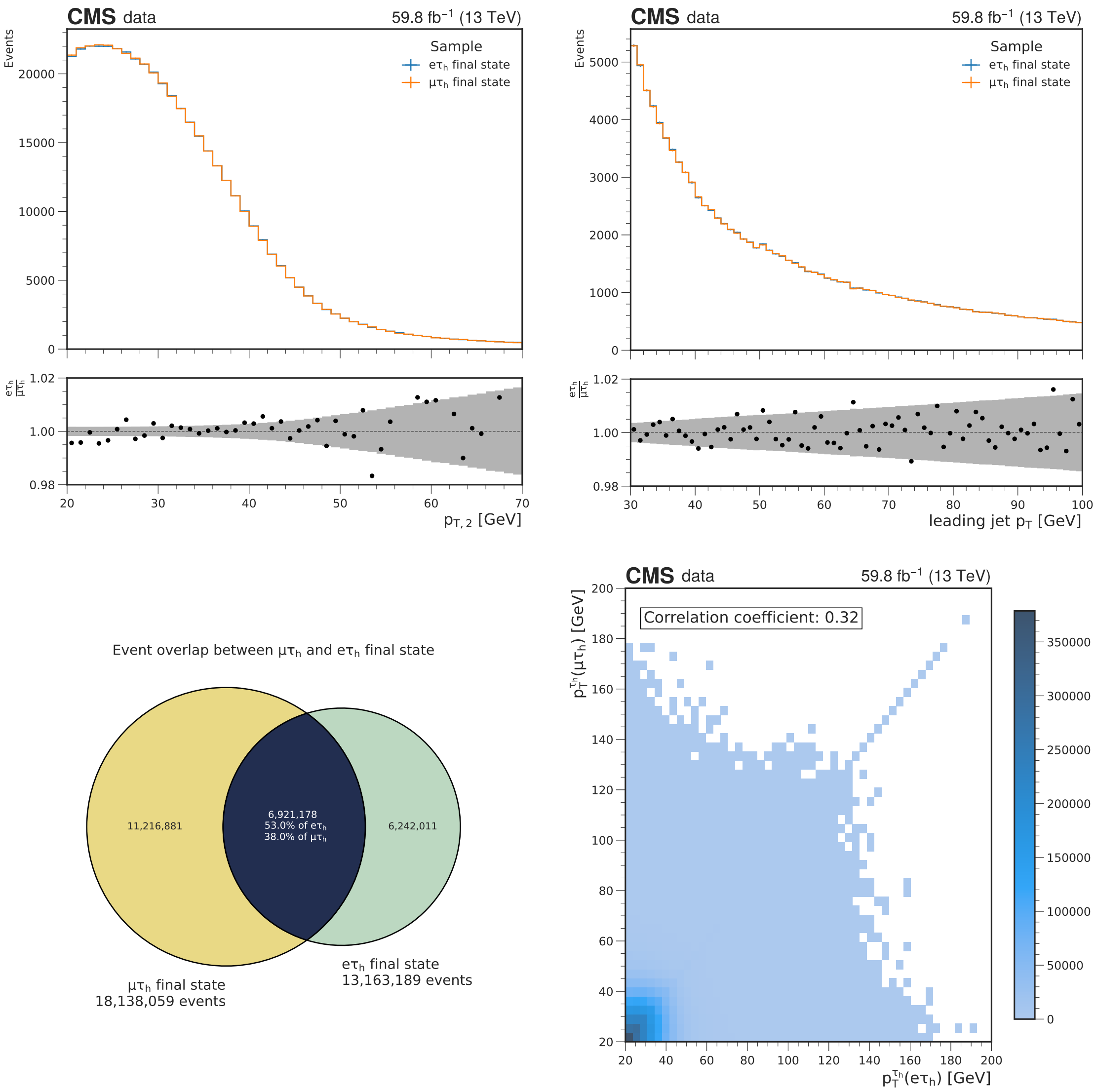


**Figure 5.9.:** For all comparisons, only events common between the $e\tau_h$ and $\mu\tau_h$ τ-embedded samples are shown. The grey band represents the statistical uncertainty of the $e\tau_h$ sample, however, since only common events are shown, this is only added as a representation of the expted statistical uncertainty. In the top row, the $p_T$ of the $\tau_h$ and the $p_T$ of the leading jet are shown. In the bottom left, the event overlap between the two final states is shown, and on the bottom right, the correlation between the $p_T$ of the electron and the muon in the $e\tau_h$ and $\mu\tau_h$ final states is shown.

## 5.4. Merging of Simulation and Data

After the simulation of the di-τ decay and the cleaning, the results of both steps have to be merged. After the combination, the CMS reconstruction is applied to end up with a single hybrid event. The chosen approach is visualized in Figure 5.10 and explained in the following.

To obtain a well-modelled simulation of physics processes, it is necessary to simulate the interaction of particles with the matter in the detector, including the active detector material, support structures and electronics. In the simulation tool GEANT4 [115–117], a model of the CMS detector and its components is implemented. In the model, each detector module is assigned a position in a 3D space, along with a rotation angle. The detector geometry is a set of parameters that describe the locations, rotations and orientations of all detector components relative to each other, as well as their global position. This information is most important for the inner tracker to reconstruct tracks from the measurements.

Ideally, the merging would take place on the level of measured energy deposits, ensuring that the reconstruction can be applied the same way it is done for the data. This strategy would imply, that the energy deposits obtained from the simulation of the τ lepton decays are added to the measured deposits of the cleaned event. Such a merging strategy is only meaningful if the position of all detector components is the same for the simulation and the data. Otherwise, a simulated track would be reconstructed in the wrong location or not even reconstructed at all. An illustration is given in Figure 5.11 where a slight shift in the tracker cell positions results in a failed track fit.

For the detector used during data-taking, the exact position of each detector module must be known. This knowledge is of particular interest for the silicon tracker since it is impossible to mount all tracker modules with the precision of 10 µm; however, this precision is needed to successfully reconstruct all tracks and resolve all vertices in an event. The actual position, orientation and potential twists with all tracker modules are determined using measurements. This process called tracker alignment is explained in Section 3.3.2.

However, due to technical limitations, the geometry used during the simulation is an idealized model of the detector. The information from the tracker alignment cannot be included, but instead, the idealized model is only slightly adapted to match the real detector geometry. It is not accurate enough for a merging of the τ lepton decay simulation and the cleaned event at the level of measured energy deposits.

A comparison between the tracker geometry used during the simulation and the measured detector geometry is shown in Figure 5.12. The shifts are in $\mathcal{O}(10\,\mathrm{cm})$, which is several orders of magnitude larger than the resolution needed for successful track reconstruction. It is visible that shifts include a sizeable global component of the complete detector as well as twists of the individual modules with respect to each other. While a correction of the global shift could be applied, correcting for the modules' twist is non-trivial.

Instead, the merging is performed on a level where the relative positions of different detector components only have a limited impact on the reconstruction outcome. The merging must occur before the application of PF, but otherwise, as late as possible. Therefore, the merging is performed on the level of PF inputs. A set of additional steps are

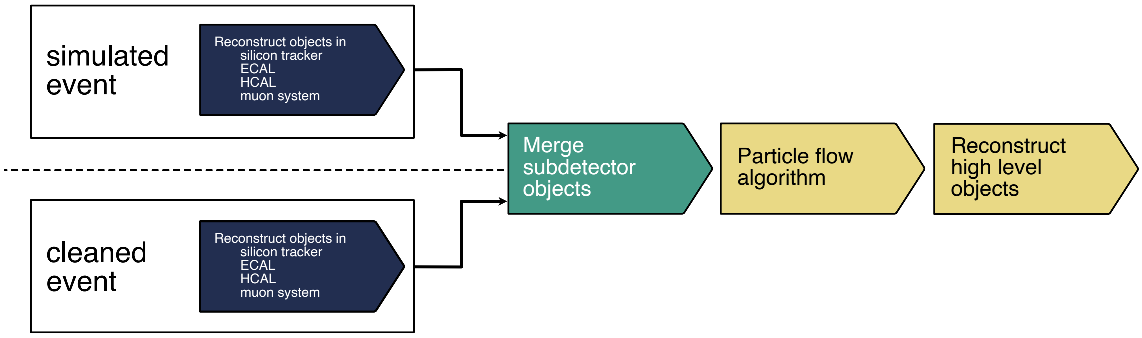


**Figure 5.10.:** A visualization of the merging strategy chosen for the τ-embedding. The impact of the different geometries can be minimized by merging the simulated di-τ decay with the cleaned event after the reconstruction of subdetector objects.

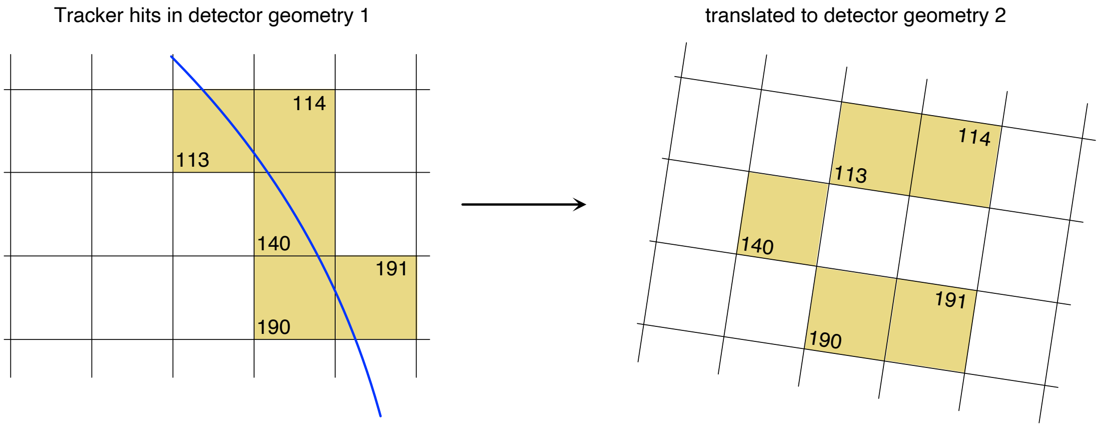

Tracker hits in detector geometry 1
114
113
140
191
190
translated to detector geometry 2
114
113
140
191
190

**Figure 5.11.:** An illustration of an issue during track reconstruction when using different geometries for simulation and data. In this example, a track results in energy deposits in the tracker cells 113, 114, 140, 190, and 191. In detector geometry 2, the position of the tracker cells is shifted relative to each other. As a result, the track from detector geometry 1 cannot be reconstructed in detector geometry 2.

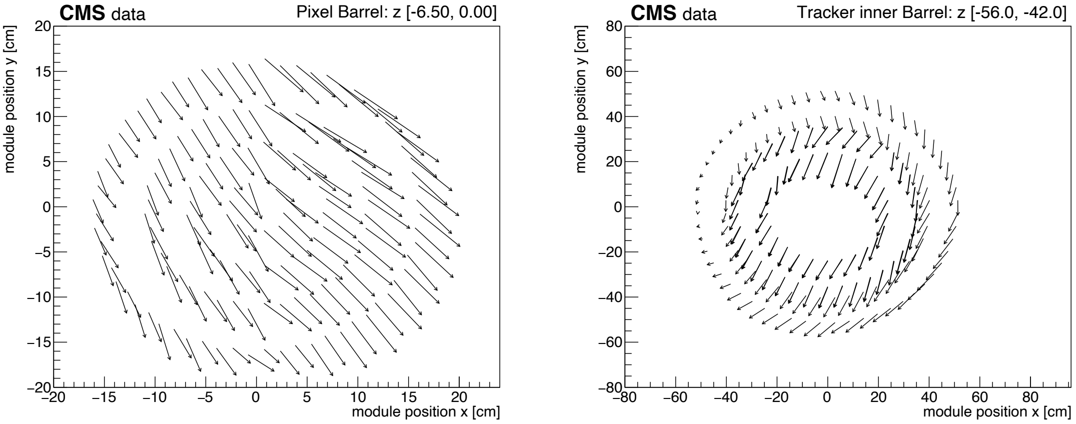


**Figure 5.12.:** Shown are the differences between the detector geometry used for the simulation and the detector geometry used during the 2018 data-taking. The arrows indicate the distance between the module position in data and simulation. The left plot shows the shifting for a set of inner pixel modules located in the negative z direction in the barrel. The right plot shows the shifts for a set of inner barrel modules in the negative z direction. In both cases, the shifts are in $\mathcal{O}(10\,\mathrm{cm})$ which is several magnitudes larger than the resolution required for successful track reconstruction. Additionally, shifts in the relative module positions can be seen since the arrows are twisted relative to each other.

injected into the reconstruction sequence, combining intermediate results of the simulated τ lepton decays and the cleaned event. These extra steps are the most delicate part of the τ-embedding procedure, as reconstruction objects that are not merged are not considered during the PF algorithm and therefore lost.

Some high-level objects still rely on the geometric compatibility of objects from different subdetectors, such as the Electron ID described in Section 3.3.7. Variables related to the geometrical compatibility of the electron track and the ECAL SC are not well modelled in τ-embedded samples. However, their impact can be mitigated by using dedicated lepton correction factors, as described in Section 6.2.

## 5.5. Production of τ-embedded samples

The production of τ-embedded samples is a computationally expensive task. The workflow that has to be applied slightly differs from the workflow described in the previous sections due to technical constraints in CMSSW [118], the software framework used for data processing in the CMS collaboration. Instead of performing the simulation and the cleaning steps in parallel, they have to be executed sequentially, and some additional steps are required. The actual workflow is shown in Figure 5.13. Both steps are performed using an HTCondor batch system, where the entire processing task is split into multiple jobs, and each job is responsible for processing a fraction of events. The workflow was applied to the whole Run II UL data set collected by CMS. In the following, only the workflow for the 2018 data set will be discussed.

The workflow is divided into two steps:

1. *Preselection*: This step is used to perform a preselection of events suitable for the τ-embedding. Essentially, the selection step described in Section 5.1 is performed, and afterwards, a filter is applied to remove events that do not pass the selection. The preselection is performed once per input data set.

2. Production of τ-embedded events: This step contains the steps described in the previous sections. As input, the preselection data set is used. This step is performed once per input data set and final state, i.e. a total of six times.

For the preselection, every single job is configured to process 3000 events. The filter during the preselection has an average reduction rate of 75%. Therefore each job results in an output file containing about 750 events. For the 2018 data-taking, the initial DoubleMuon data set contained $315.8 \times 10^6$ events, the preselection data set contained $78.8 \times 10^6$ events. The advantage to running this step before the τ-embedding production is that the amount of data that has to be processed is reduced.

In the CMS collaboration, data and simulation samples are centrally provided and available in multiple data tiers. A list of the most important data tiers is shown in Table 5.4. Most analyses are performed starting from MiniAod or NanoAod data tier since the information available at this data tier is sufficient. However, since the full detector information

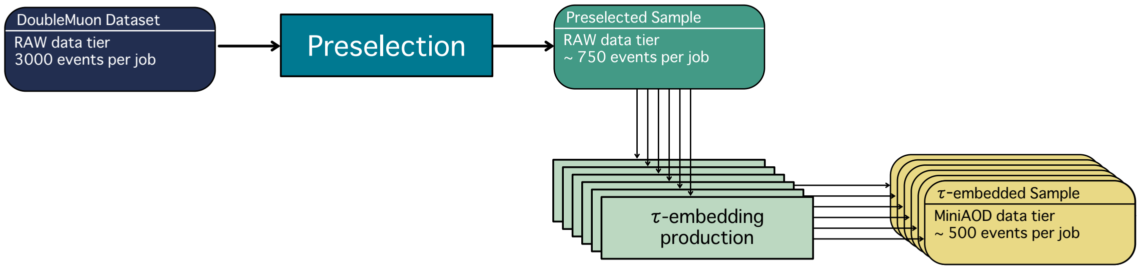


**Figure 5.13.:** Visualization of the workflow applied for the production of τ-embedded samples. At first, the preselection is performed to reduce the amount of data that has to be processed. Then the τ-embedding production is performed once per desired final state.

**Table 5.4.:** The most important data tiers used by the CMS Collaboration.

| Data Tier | Average event size [kB] | Description |
|---|---|---|
| Raw | ≈1500 | raw data, full detector readout |
| Reco | ≈3000 | reconstructed event |
| RawReco | ≈4500 | combination of Raw and Reco |
| Aod | ≈500 | reduced detector information |
| MiniAod [119] | ≈35 to 60 | further reduced detector information |
| NanoAod [120] | ≈1 to 2 | further reduced detector information |

is needed for the cleaning step, the data sets used for the τ-embedding method have to be available in the Raw data tier. The 2018 preselection data has a size of 122.9 TB, with an average size of 1.5 MB/event. The input and the output of the preselection step are stored in the Raw data tier.

The separate processing tasks of the τ-embedding production are visualized in Figure 5.14. A total of six tasks are performed in sequence in each job. The intermediate output files are only stored in the job, to be picked up by the next task. Saving all intermediate output files would require too much disk space; instead, only the output file of the last task is stored in the MiniAOD data tier.

1. **Selection**: This task is identical to the one performed during the preselection step. The RAW event is reconstructed, and two muons are selected. Strictly speaking, this would not be necessary; however, the output of the preselection step no longer contains any reconstruction information. The information could be kept by using the RawReco data tier as the output data tier of the preselection step. However, this would take up to three times more disk space for the preselection data set and would also increase the amount of data that each τ-embedding production job has to read. The input is a Raw event, the output is a RawReco event, and the average runtime is 16.74 s/event.

2. **Cleaning & LHE**: During this task, the cleaning of the di-μ signature as described in Section 5.2 is performed. After the cleaning, the full event reconstruction is performed resulting in the cleaned event. In addition to the cleaning, the simulation of the two τ leptons has to be prepared. Due to technical limitations, the information needed for the decay simulation must be set up before the decay simulation itself. Therefore the setup of the τ lepton decays described in Section 5.3 is performed, and the required information is stored in a Les Houches Event (LHE) file [121], which is a common interface for particle simulation tools such as Pythia. The input is taken from the Selection task, and the output is a custom RawReco event with the additional LHE information. The average runtime is 14.28 s/event.

3. **Simulation**: The simulation described in Section 5.3 is performed. The output is a custom data tier containing the Lorentz vectors of all simulated τ lepton decay products. After the simulation trials, a filter is applied to only continue the processing of events, where at least one decay has passed the acceptance cuts. The efficiencies for

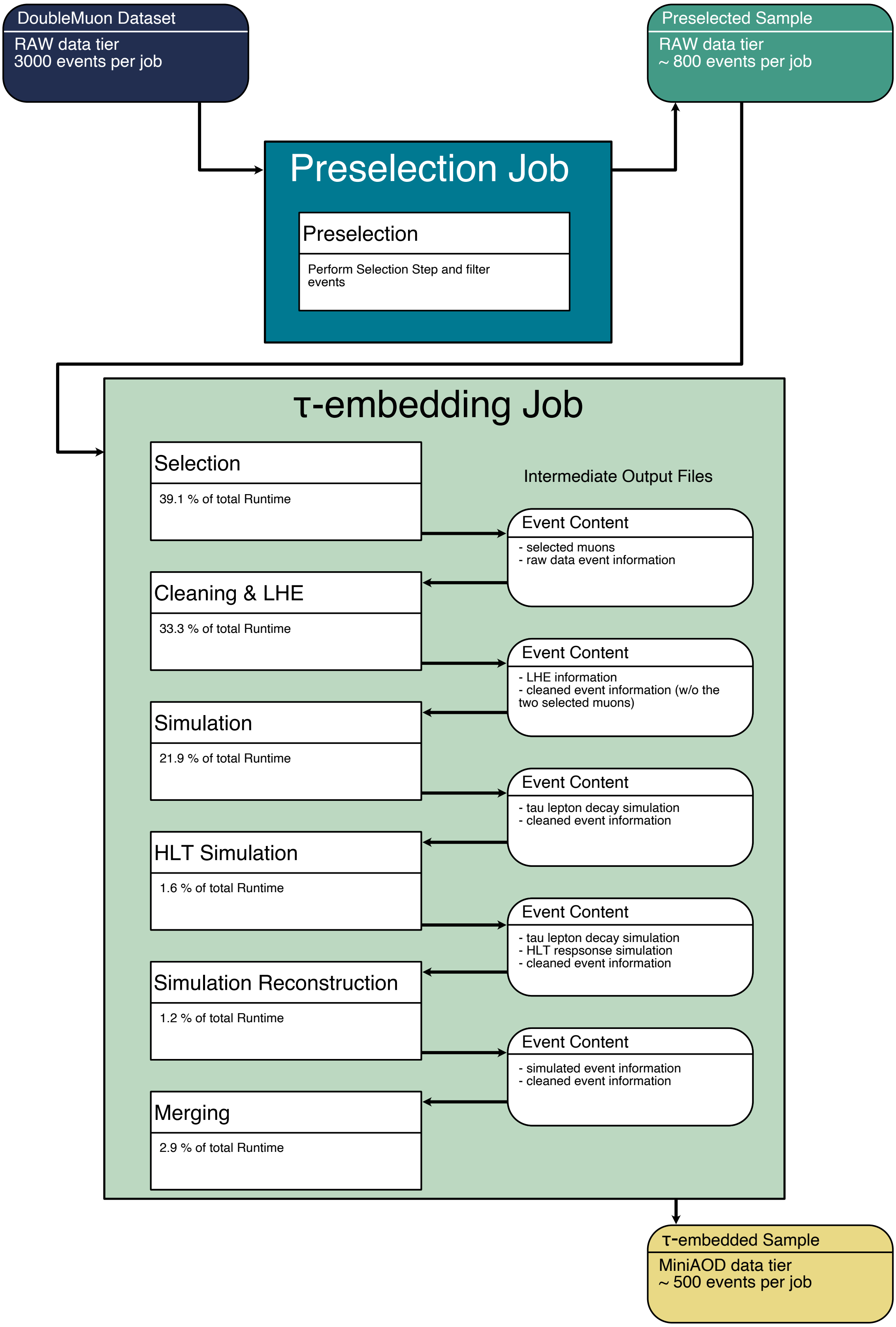


**Figure 5.14.:** Visualization of the tasks in the preselection and the τ-embedding production jobs.

the different final steps can be found in Table 5.5. Since the processing is performed in sequence, the information of the cleaned event is also contained in the output file. The average runtime is 9.37 s/event.

4. **HLT Simulation**: The response of the HLT is simulated during this task. This simulation must be performed with the identical software setup used during the data-taking. Therefore, a switch of the CMSSW version is needed. The usage of the same version ensures that the identical configuration and implementation of all reconstruction algorithms used during the data-taking are also used for the HLT response simulation. The output is again a custom data tier containing all information of the previous step, plus the simulated HLT response. The average runtime is 0.67 s/event.

5. **Simulation Reconstruction**: During this task, the reconstruction of the simulated τ lepton decays is performed. Since the detector is empty besides the decay products, the reconstruction is very fast. The output of this task contains the information of both the cleaned event and the simulated τ lepton decays after performing the reconstruction. The average runtime is 0.53 s/event.

6. **Merging**: The purpose of the last task is to combine the cleaned event with the simulated τ lepton decays. As described in Section 5.4, the merging is performed based on intermediate reconstruction results, which is why it is necessary to keep this information till the last task. After the merging, the Pat step is performed, which is used to calculate additional event information such as particle identification variables. The output of the final step is a MiniAod event. The average runtime is 1.23 s/event.

The distribution of the runtime per event of the different tasks is shown in Figure 5.15. The numbers also include the initial overhead of the task startup. The Selection and Cleaning & LHE tasks make up roughly 75% of the total runtime per event. This has several reasons: During both the Selection and Cleaning & LHE tasks, the full event reconstruction starting from the RAW data tier is performed. Full event reconstruction is by far the most computationally intensive task. In addition, the Selection task has a longer runtime than the Cleaning & LHE task since the input data for the first task has to be streamed from the grid, resulting in a higher I/O load. After the first task, the input information is available on the local disk within the job. Full event reconstruction is also performed during the Simulation Reconstruction task; however, in this case, the event contains only the products of the τ lepton decays, which is far less complex than the reconstruction of a full data event. The combined average runtime of the τ-embedding method is about 40 s per event, depending on the final state and the number of simulation trials.

In total, the τ-embedding production was split into 105 281 jobs per final state. The total runtime of the production was 3 547 298 h. A more detailed overview split by final state can be found in Table 5.5. On average, the production runtime was very similar across the different final states as shown in Figure 5.16. Apart from the $\tau_h\tau_h$ final state, which was produced using the lxplus batch system provided by CERN, the other five final states

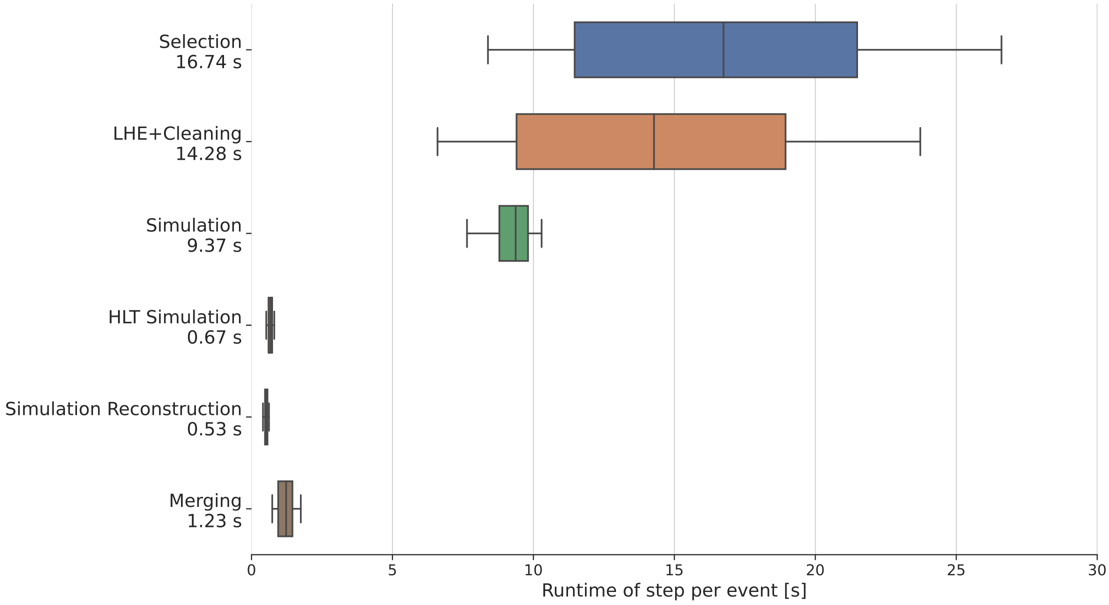


**Figure 5.15.:** Average runtime of the individual tasks of the τ-embedding production. The box represents the 50% quantile of the runtime; the whiskers represent the 95% quantile. The median value is represented by the vertical line; outliers are not shown.

were produced using the batch system of the ETP. In the ETP batch system, local resources and dynamically integrated opportunistic resources, such as the *bwForCluster NEMO* [122] were used.

The jobs themselves showed an excellent performance in terms of CPU utilization; the average utilization was 97% in single-threaded mode with an average runtime of 5.6 h. The memory usage of the jobs was defined beforehand and set to 3500 MB per job. The size of the output files and the runtime both depend on the efficiency of the generator filter, which is applied during the simulation step. All final states have an efficiency of more than 50%. In the $\tau_h\tau_h$ final state, the filter efficiency is the lowest, which is why the input files are the smallest, and the runtime is the lowest. For $\mu \rightarrow \mu$ embedded samples, no filter is applied, which is why the input files are the largest and the runtime is the highest. In addition, only a single simulation trial is performed for the $\mu \rightarrow \mu$ and $\mu \rightarrow e$ embedded samples, further reducing the runtime of the simulation step. A comparison of the average job runtime and the output file size is visualized in Figure 5.16.

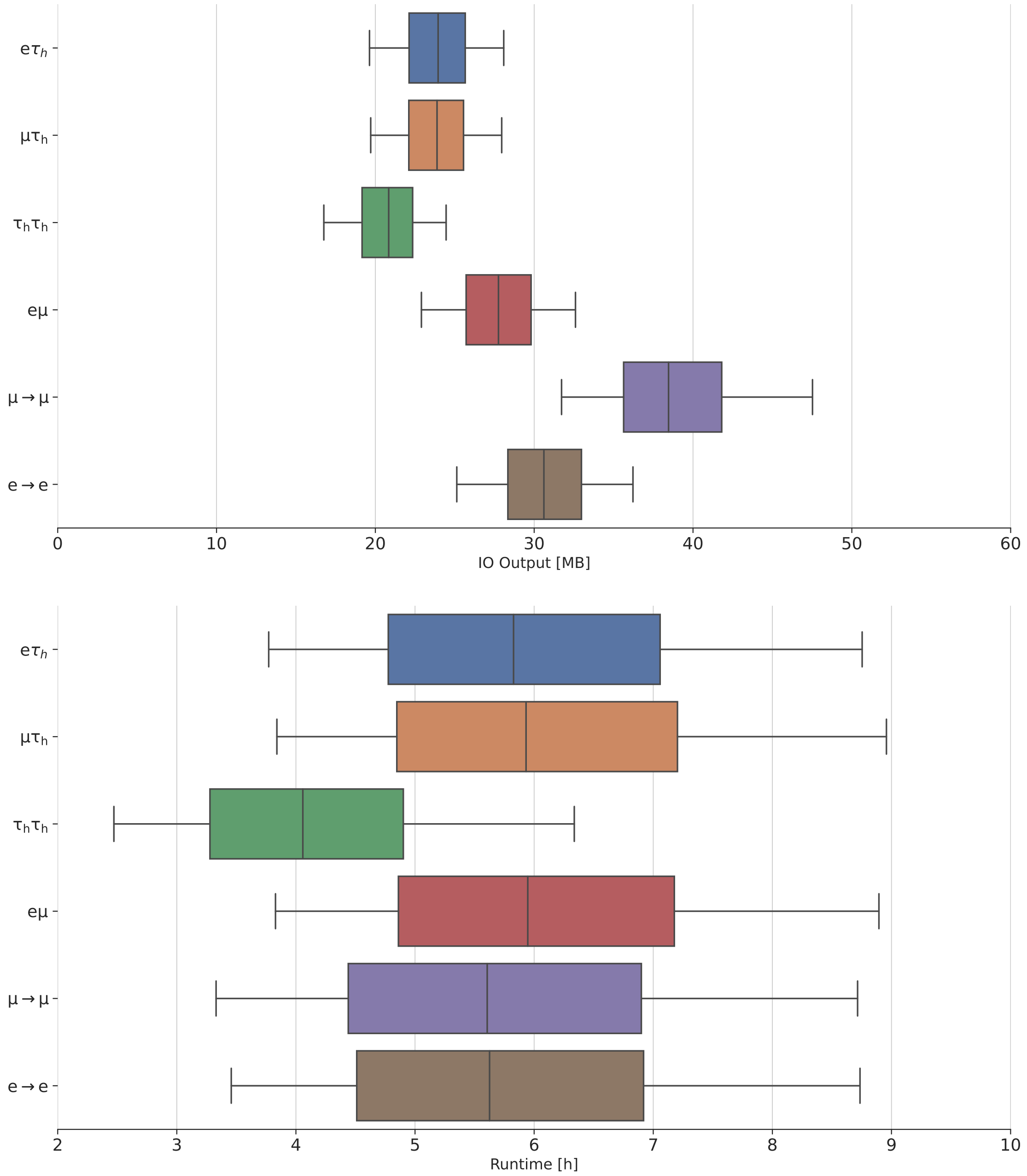


**Figure 5.16.:** The job runtime and the output file size for the τ-embedding production. On the top plot, the distribution of the output file sizes is shown, which directly corresponds to the number of events produced by the job. In the bottom plot, the job's runtime distribution is shown. The box corresponds to the 50% quantile of the respective distribution, whereas the whiskers correspond to the 95% quantile. The median value is given by the vertical in the middle of the box; outliers are not shown.

**Table 5.5.:** Summary of the runtime of the τ-embedding production jobs for the 2018 data set, split by final state.

| Final State | Runtime [h] | Number of Events | Data set Size [GB] | Filter Efficiency |
|---|---|---|---|---|
| $\mu\tau_h$ | 637 193 | 48 419 520 | 1593 | 61.4% |
| $e\tau_h$ | 624 542 | 48 421 858 | 1633 | 61.4% |
| $\tau_h\tau_h$ | 443 946 | 42 968 873 | 1385 | 54.5% |
| $e\mu$ | 636 199 | 55 484 486 | 1897 | 70.4% |
| $\mu \rightarrow e$ | 605 097 | 59 253 609 | 2133 | 75.2% |
| $\mu \rightarrow \mu$ | 600 320 | 78 801 838 | 2659 | 100.0% |
| Total | 3 547 298 | 333 350 184 | 11 300 | - |

## 5.6. Validation using μ → μ embedded events

The τ-embedding method can be validated by comparing the μ → μ embedded sample with the initial DoubleMuon data set. Any differences point to potential biases of the method. For this comparison, only events present in the DoubleMuon sample and the μ → μ embedded sample were used. In all comparison plots, the same events are used; therefore, no differences between the samples due to statistical uncertainty are expected. Instead, all differences are related to the τ-embedding method. The uncertainty band represents the statistical uncertainty of the μ → μ embedded sample and is only shown as a reference. Since the μ → μ embedded sample used for the validation contains much more events than expected in typical target analysis, an agreement within the shown statistical uncertainties is considered acceptable. For the comparison, both muons must be global and have a $p_T$ larger than 20 GeV. About 47 million events are considered for this comparison.

The direction of the muons is very well preserved by the method, as evident from the comparison of η and ϕ of the leading muon shown in Figure 5.17. The differences between the DoubleMuon and the μ → μ embedded sample are on the sub-percent level.

A difference can be seen in $I^{\mu}_{\mathrm{rel}}$ of the muons, as shown in Figure 5.18 (left). A trend towards less isolated muons in the τ-embedded sample is observed. This difference points to incomplete cleaning of the muon energy from the calorimeters. Any energy deposits not completely removed from the detector will result in a less isolated muon. In Figure 5.18 (right), the shift in the amount of energy that can be found in the isolation cone of $\Delta R = 0.3$ around muon before and after the μ → μ embedding is shown. For most events, no or a small change in energy is observed. On average, 126.6 MeV of additional energy can be found after performing the τ-embedding method. This energy roughly corresponds to the mass of a single muon. This finding is in line with the observation that the shift in $I^{\mu}_{\mathrm{rel}}$ is rather small. In a typical target analysis, a requirement of $I^{\mu}_{\mathrm{rel}} < 0.15$ is imposed. The observed median isolation shift is $0.0005$. In most cases, the isolation shifts from no surrounding energy deposits to a small number of energy deposits. Such a small shift does not impact the overall quality of the muon.

In the distribution of the di-μ mass in Figure 5.19 (left), a broader Z boson resonance is observed for the μ → μ embedded sample. The resolution of the Z boson resonance

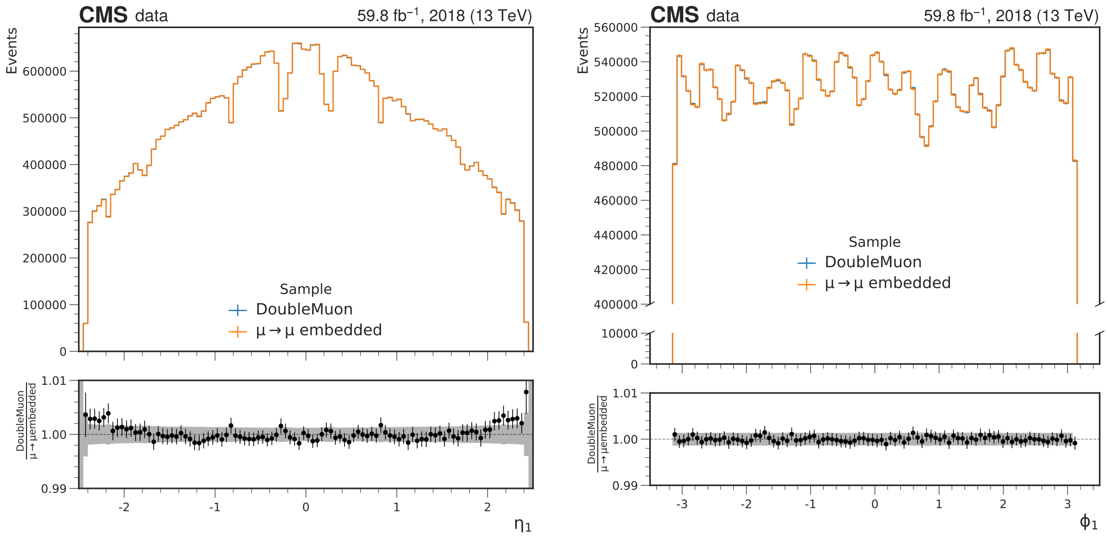


**Figure 5.17.:** The distribution of η and ϕ for the leading muon. Small differences are observed below the % level between the input sample and the μ ⟶ μ embedded sample. The statistical uncertainty on the μ ⟶ μ embedded sample is only shown as a reference, as only events common to both samples are shown.

is washed out since the CMS reconstruction sequence is applied twice for those muons, once during the selection step and then a second time during the simulation step. Each reconstruction results in a smearing of the muon energy measurement due to the finite resolution of the detector. As a result, the resonance width is increased, resulting in the distinctive double-peak structure in the ratio of the two samples. While this effect is visible in the μ ⟶ μ embedded validation, in the τ lepton decay simulation, the resolution of the Z boson resonance is much worse and shifted to lower energy due to the neutrinos produced in the decays. The difference in the distribution of the di-μ mass is therefore not of concern for the usage of τ-embedded samples.

In Figure 5.19 (right) the distribution of $E_T^{miss}$ calculated using the PUPPI algorithm is shown. A small trend towards less MET in the μ ⟶ μ embedded sample is visible. On average, a shift of 2 GeV per event is observed. In both cases, no MET is expected from the Z ⟶ μμ decay, so all MET contribution originates from reconstruction effects. Since the μ ⟶ μ embedded sample was reconstructed twice, small differences in the MET reconstruction are expected.

In Figure 5.20, the distribution of the number of jets (left) reveals a slightly harder jet spectrum in data compared to the μ ⟶ μ embedded sample. In the μ ⟶ μ embedded sample, ~ 100,000 events (0.03%) are found with 0 additional jets, compared to data. The distributions of the di-jet mass $m_{jj}$ (right) are in good agreement with each other.

Additional variables, as well as the distributions for the subleading muon, can be found in Figures A.5 to A.8. In general, a good agreement between the input sample and the μ ⟶ μ embedded sample is observed.

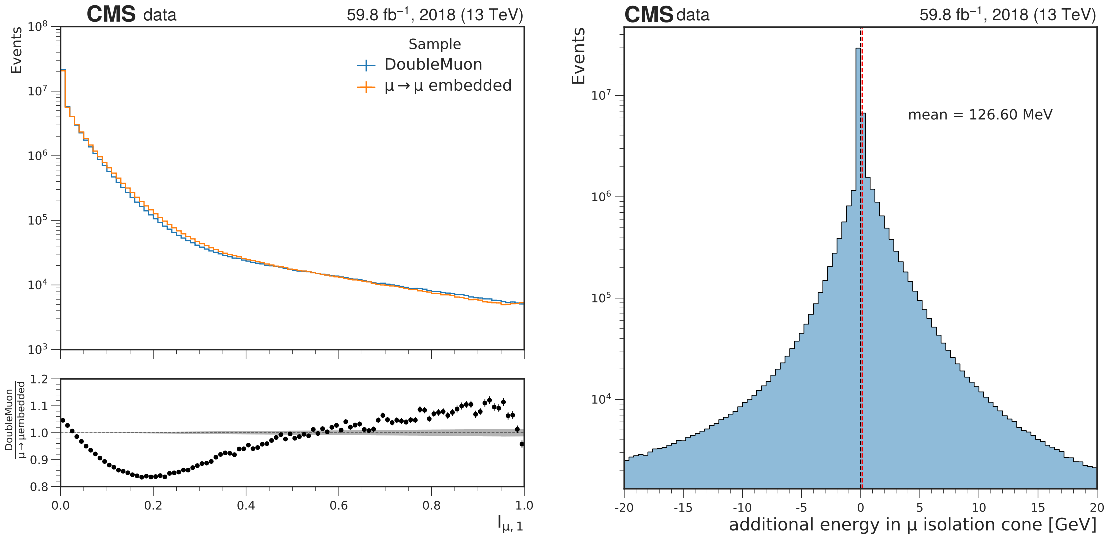


**Figure 5.18.:** The distribution of $I^{\mu}_{\text{rel}}$ of the leading muon and the difference of the amount of energy around the muon before and after the application of $\mu \to \mu$ embedding. The statistical uncertainty on the $\mu \to \mu$ embedded sample is only shown as a reference, as only events common to both samples are shown.

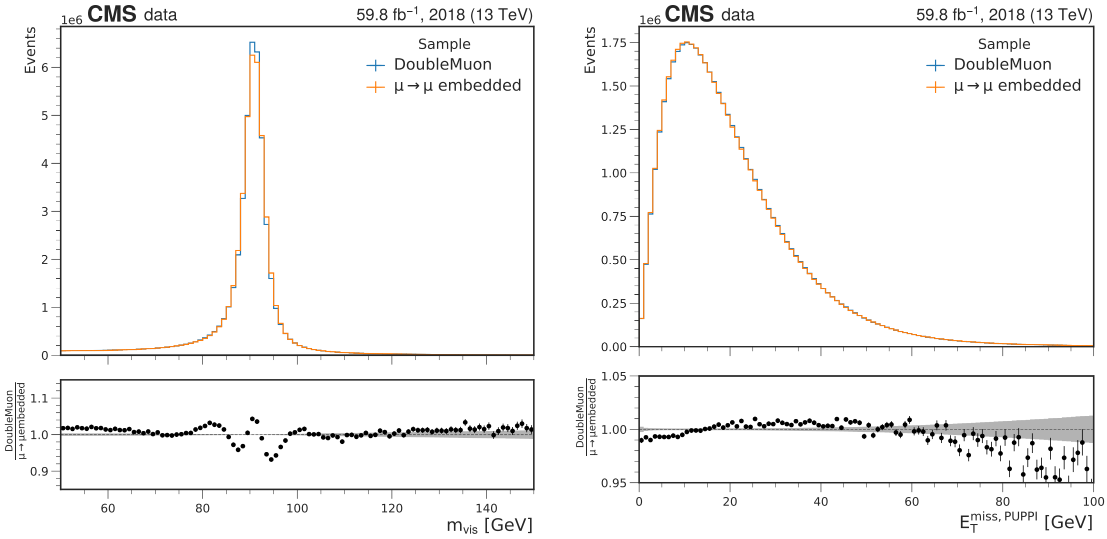


**Figure 5.19.:** The distribution of $m_{\text{vis}}$ of the di-μ system is shown on the left, while the distribution of $\mathrm{E}_{\mathrm{T}}^{\text{miss}}$ is shown on the right. The statistical uncertainty on the $\mu \to \mu$ embedded sample is only shown as a reference, as only events common to both samples are shown.

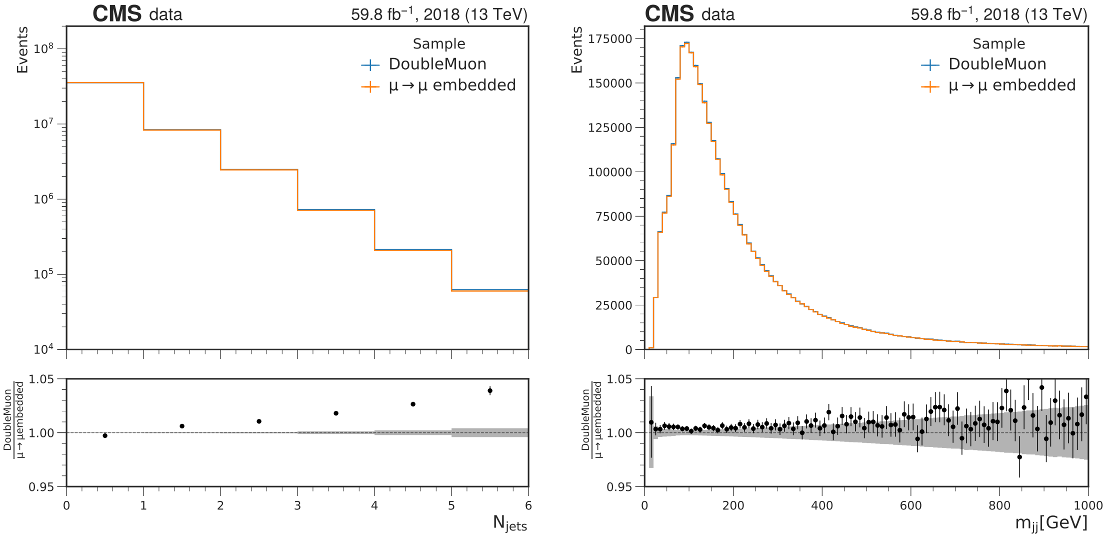


**Figure 5.20.:** The distribution of the number of jets (left) and the mass of the dijet system. The statistical uncertainty on the μ ⟶ μ embedded sample is only shown as a reference, as only events common to both samples are shown.

# 6 Towards the Application of $\tau$-embedding in Analyses

Several correction factors must be derived and applied to use $\tau$-embedded samples in a typical target analysis. Since the $\tau$ lepton decays in $\tau$-embedded samples are simulated in an otherwise empty detector, it is impossible to reuse the corrections derived for simulated samples. Instead, a new set of corrections has to be derived. In addition, some corrections are exclusively needed for $\tau$-embedded samples. In this chapter, the purpose and the measurement of all required corrections are described.

The necessary corrections are:

- The $w_{\mathrm{gen}}$ weight that is calculated during the simulation step as described in Section 5.3. This weight is used to correct the bias introduced by repeating the simulation of the $\tau$ lepton decays 1000 times, which greatly increases the number of events in the higher energy regions. This value is derived per event.

- To obtain the correct normalization of the $\tau$-embedded samples, it is possible to measure and unfold the efficiency of the criteria used during the selection step. Applying this efficiency correction eliminates the need for any luminosity scaling as it is required for simulated samples. The normalization is derived directly from the data by measuring the efficiency of the selection criteria and reverting it. This procedure corresponds to unfolding the detector effects responsible for any potential inefficiencies. The measurement is described in Section 6.1. These corrections are denoted as unfolding corrections.

- For electrons, corrections targeting the identification, the isolation and the trigger efficiency are needed. The corrections are derived using the $\mu \rightarrow \mathrm{e}$ embedded samples. They are required in the $\mathrm{e}\tau_{\mathrm{h}}$ and the $\mathrm{e}\mu$ final state. The procedure is described in Section 6.2.

- For muons, the corrections targeting the identification, isolation, and trigger efficiency are needed. These corrections are derived using the $\mu \rightarrow \mu$ embedded samples. They are required in the $\mu\tau_{\mathrm{h}}$ and the $\mathrm{e}\mu$ final state. The procedure is described in Section 6.2.

- For $\tau_{\mathrm{h}}$ decays, corrections for the vsJet identification efficiency have to be derived. The corrections are determined with a dedicated measurement in the $\mu\tau_{\mathrm{h}}$ final state described in Section 6.3. The corrections are applied in the $\mathrm{e}\tau_{\mathrm{h}}$, $\mu\tau_{\mathrm{h}}$ and $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ final states.

- For electrons and $\tau_h$, corrections for the measurement of the particle energy have to be derived. The measurements are described in Section 6.4. The energy measurement of the muon is considered to be sufficiently accurate so that no correction is applied.

## 6.1. Unfolding Corrections

The efficiency of the DoubleMuon trigger and the isLoose WP of the muon identification used during the selection described in Section 5.1 must be derived to correct the normalization of the τ-embedded samples. These types of efficiency measurements can be performed using the Tag and Probe method [123].

### 6.1.1. Tag and Probe Method

While one can directly obtain the efficiency for simulated samples using the information from the particle simulation, this is impossible for data. The Tag and Probe method does not rely on generator information and can be used to derive efficiencies $\epsilon_i$ for data, τ-embedded samples, and simulated samples.

The method assumes that the $Z \rightarrow \mu\mu$ and the $Z \rightarrow ee$ process can be measured with high precision and efficiency. In the following, the efficiency measurement using the Tag and Probe method is explained using the isLoose muon identification working point.

First, a very loose selection of events is performed. Two muons without any isolation, identification or trigger requirements are selected. As a baseline selection, the two muons are required to be separated by a minimum distance of $\Delta R = 0.5$ and have a $p_T > 7\,\text{GeV}$. To achieve an improved modelling of the Z boson resonance, a veto for final state radiation photons is applied, to remove all events, where a photon with a $p_T > 10\,\text{GeV}$ is reconstructed within $\Delta R = 0.4$ of the two muons.

After the event selection, each event can be used to build Tag and Probe pairs. The tag muon is required to be well-identified. This is ensured, by requiring the selection criteria listed in Table 6.1. A muon of this quality is assumed to originate from a Z-boson decay and therefore, the probe muon must also originate from a Z boson. Since the probe muon does not have to fulfil any requirements, other than the selection criteria described in the previous paragraph, it can be used to check, if the moun identification algorithm is able to correctly identify it. If all probe muons are correctly identified, the efficiency of the muon identification algorithm would be 100%. In most cases, both muons in the event can be used as a *tag* and *probe* muon, effectively doubling the number of Tag and Probe pairs available.

Assuming that both muons are genuine muons from a Z-boson decay, the pairs can now be sorted into a *pass* and *fail* region. The pair is put in the fail region if the probe muon does not pass the isLoose identification requirement. In this case, the muon identification algorithm could not correctly identify the probe muon, reducing the algorithm's efficiency. The di-μ mass is used as the discriminating variable in each region. One example of the pass and fail regions is shown in Figure 6.1. The efficiency measurement is performed in several phase space areas to include kinematic and regional effects within the efficiency measurements. Typically, the probe particle's $p_T$ and $\eta$ are used.

**Table 6.1.:** Selection criteria for the *tag* muon.

| Description | Selection Criteria |
|---|---|
| HLT path | Single Muon Trigger |
| muon $p_T$ | $> 25\,\text{GeV}$ |
| muon ID working point | isMedium |
| muon isolation | $< 0.15$ |

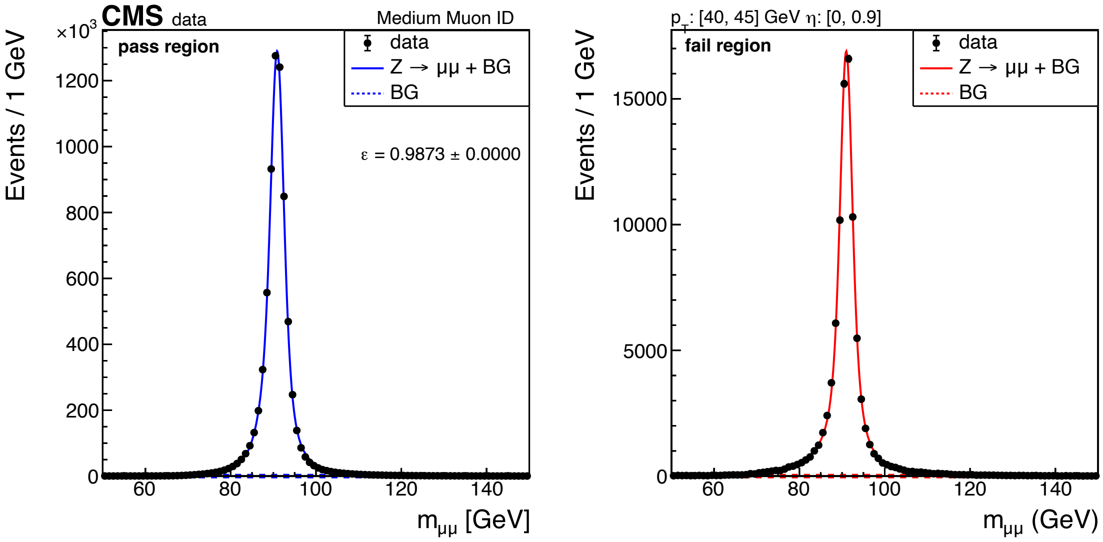


**Figure 6.1.:** Example for the histograms of the di-μ mass for the pass (left) and fail (right) regions used for the Tag and Probe measurement. The *probe* muon is required to have a $p_T$ between 40 and 45 GeV and both muons must have $|\eta| = [0.0, 0.9]$. The data is shown in black, the background model (BG) is visualized as a dashed line, and the solid line visualizes the fit result of the signal + background model.

A combined fit of the pass and fail regions is performed in every phase space bin *i*. In each region, the Z-boson peak is modelled via one Voigt function per region, a convolution of a gaussian and a Breit-Wigner function. For the background, an exponential function is used. The parameter of interest of the fit is the efficiency $\epsilon_\text{i}$ defined as

$$\epsilon_\text{i} = \frac{N_\text{pass,i}}{N_\text{pass,i} + N_\text{fail,i}} \tag{6.1}$$

where $N_{pass,i}$ and $N_{fail,i}$ are the number of signal events in the pass and fail regions for bin *i*. The numbers of events are obtained from the normalization of the Voigt functions.

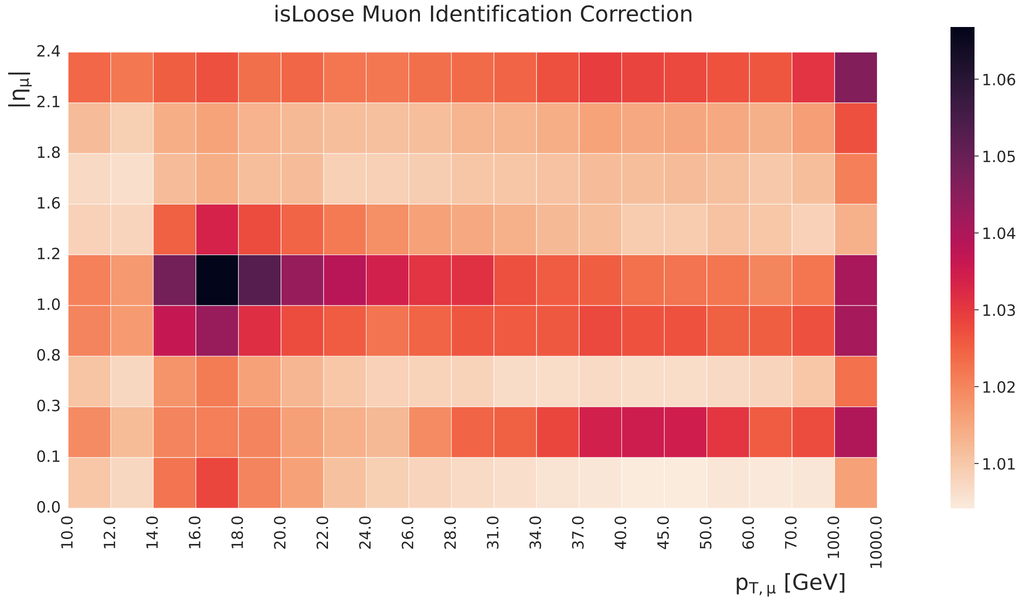


**Figure 6.2.:** Unfolding correction factor for the isLoose muon identification used during the selection step of the τ-embedding. Only for the region $|\eta| \in [0.8, 1.2]$ larger corrections of more than 4% are required. Due to the large number of events available for the measurement, the uncertainties on the measured efficiencies are on the sub-percent level.

### 6.1.2. Correction Factors

Only the efficiency in the data is derived for the selection efficiency corrections. The inverse of the efficiency in the data is applied to correct the efficiency for the τ-embedded samples

$$CF_{\mathrm{EMB}} = \frac{1}{\epsilon_{\mathrm{data}}}. \tag{6.2}$$

These unfolding correction factors are then applied per event to the τ-embedded samples based on the kinematics of the selected muons.

The resulting $CF_{\mathrm{EMB}}$ of the isLoose muon identification is binned in $p_{\mathrm{T}}$ and η of the selected muon and shown in Figure 6.2. Apart from a small drop in efficiency for $p_{\mathrm{T}}$ = [14,22] GeV the correction factors are close to one.

To obtain the efficiency of the DoubleMuon trigger used during the muon selection of the τ-embedding, a second Tag and Probe measurement is performed. Since the DoubleMuon trigger is a combination of two independent muon trigger legs, the total efficiency of the trigger path is given by:

$$\epsilon_{\mathrm{HLT}} = \epsilon_{8,1}\epsilon_{17,2} + \epsilon_{17,1}\epsilon_{8,2} - \epsilon_{17,1}\epsilon_{17,2} \tag{6.3}$$

where $\epsilon_{8,1}$ is the efficiency for the first muon to pass the 8 GeV HLT path, $\epsilon_{17,1}$ is the efficiency for the first muon to pass the 17 GeV HLT path, $\epsilon_{8,2}$ is the efficiency for the second muon to pass the 8 GeV HLT path, and $\epsilon_{17,2}$ is the efficiency for the second muon to

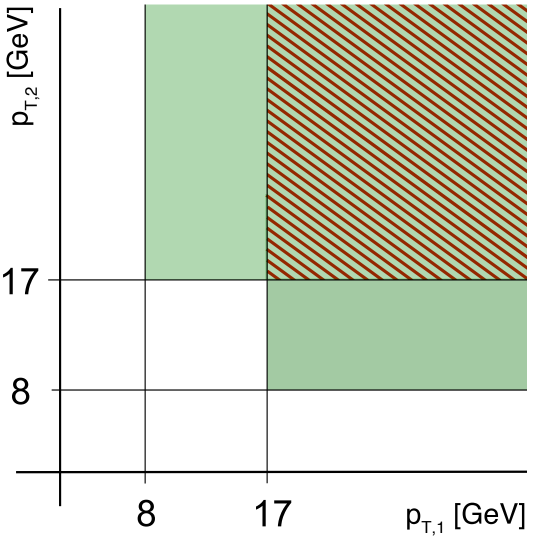


**Figure 6.3.:** Visualization of the calculation of the efficiency of the HLT path used for the selection step of the τ-embedding method. The green areas represent the first and second parts of Equation (6.3). Since the red-shaded area is added twice, the third part of the formula is a subtraction of this area.

pass the 17 GeV HLT path. The choice which muon is first and second is made randomly. A graphical representation of the formula is given in Figure 6.3. The resulting correction factor is binned in $p_T$ and η for the first and second muon, resulting in a four-dimensional parameter space

$$\epsilon_{\mathrm{HLT}}\left(p_{\mathrm{T},1}, p_{\mathrm{T},2}, \eta_1, \eta_2\right). \tag{6.4}$$

For the isLoose muon identification and the DoubleMuon trigger efficiency measurement, the $p_T$ and η bins were chosen as

$$\begin{aligned} p_t &\in [10, 12, 14, 16, 18, 20, 22, 24, 26, 28, 31, 34, 37, 40, 45, 50, 60, 70, 100, 1000] \\ |\eta| &\in [0, 0.1, 0.3, 0.8, 1.0, 1.2, 1.6, 1.8, 2.1, 2.4]. \end{aligned}$$

To obtain the normalization of the τ-embedded samples, the following unfolding corrections are applied

- $w_{\mathrm{gen}}$ from the simulation step,
- isLoose muon identification correction for the first muon,
- isLoose muon identification correction for the second muon,
- DoubleMuon trigger efficiency correction for the di-μ pair.

After that, the normalization of the τ-embedded samples is correctly scaled to the luminosity of the data used for the τ-embedding production. No additional scaling has to be used.

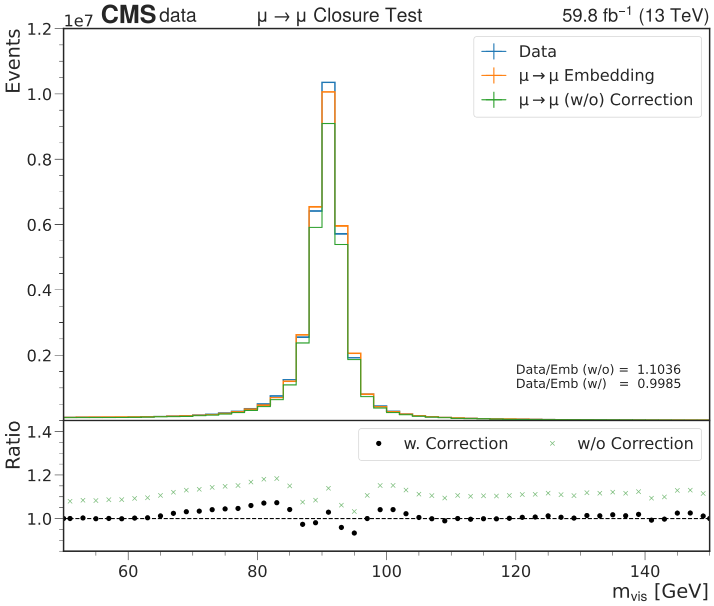


**Figure 6.4.:** A closure test of the unfolding correction for τ-embedded samples can be performed by comparing μ ⟶ μ embedded events with the DoubleMuon data set. After the application of the unfolding corrections, a difference of less than 0.15% in the total event yield is observed.

The closure of the derived unfolding corrections is tested by applying the correction factors to the μ ⟶ μ embedded samples and comparing the result with the DoubleMuon data set. Both muons are required to be global and must have a $p_T$ larger than 20 GeV. This can be used as a closure test of the method under the assumption that the efficiency for a Z ⟶ μμ event in the selected phase space region to be included in the DoubleMuon data set is 100%. Since the μ ⟶ μ embedded sample is based on the DoubleMuon data set, the unfolded μ ⟶ μ embedded distribution should match the DoubleMuon data set. The resulting $m_{\text{vis}}$ distribution is shown in Figure 6.4. Before applying the correction factors, the total yield of events in the data was 10.36% larger than the yield of the τ-embedding method samples. After the correction, an agreement within a difference of 0.15% is observed. The reconstruction effect discussed in Section 5.6 is still visible since the selection corrections only correct for the selection efficiencies.

In Figure 6.5, the $p_T$ (top row) and η (bottom row) distributions of the leading and subleading muons are shown before and after the application of the unfolding corrections. Due to the limited number of η bins, the efficiency deficits at the edges of the individual muon wheels are still visible, though much less prominent.

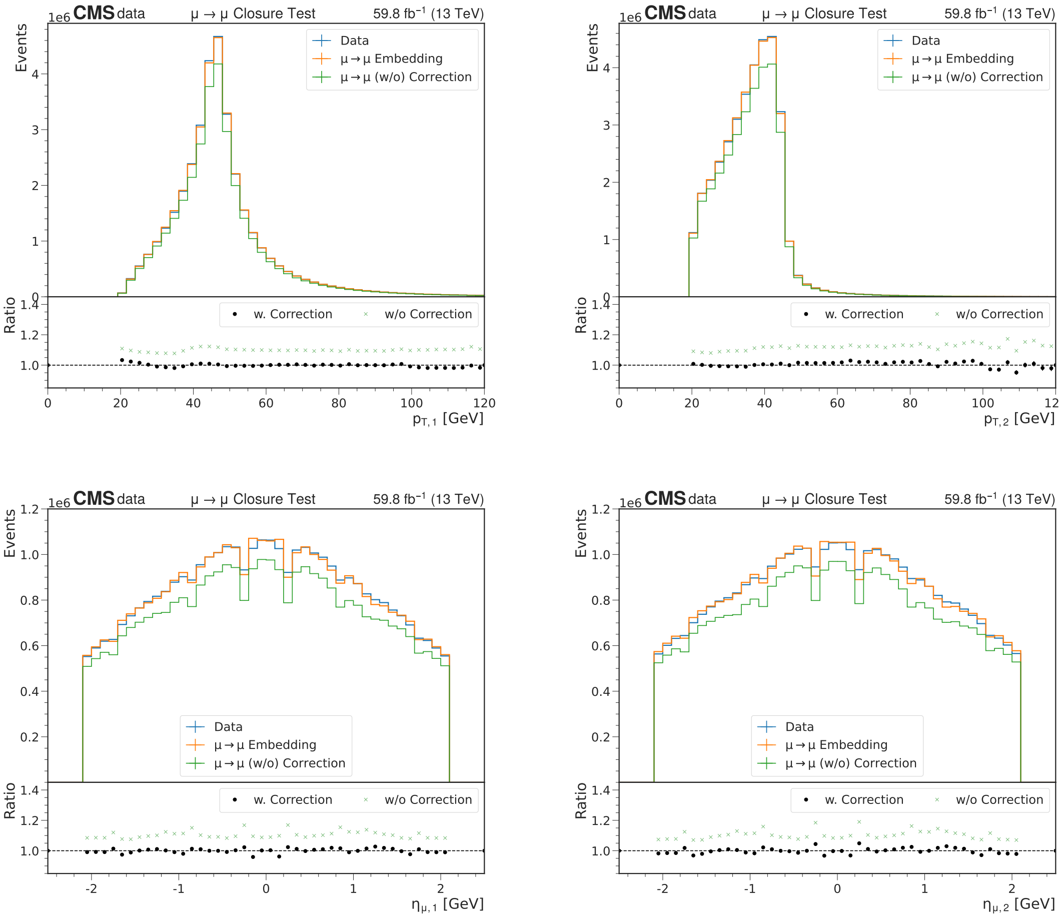


**Figure 6.5.:** Closure test of the unfolding correction for the τ-embedded samples using the four variables that are used to determine the efficiencies. The test is performed by comparing μ → μ embedded events with the DoubleMuon data set before and after the application of the unfolding corrections. In the top row, the $p_{\mathrm{T}}$ of the leading and subleading muon are shown. In the bottom row, the η of the leading and subleading muon are shown.

## 6.2. Electron and Muon Corrections

For the electron (muon) correction factors, the efficiency is measured for the data, τ-embedded samples, and simulated samples. The correction factors for the τ-embedded ($SF_{\text{EMB}}$) and simulated samples ($SF_{\text{MC}}$) can then be derived by calculating the ratios of the efficiencies:

$$SF_{\text{EMB,MC}} = \frac{\epsilon_{\text{data}}}{\epsilon_{\text{EMB,MC}}}. \tag{6.5}$$

These correction factors are applied as weights to every event in simulation or τ-embedded sample. The muon correction factors must be applied for all $\tau \rightarrow \mu$ decays in the $\mu\tau_{\text{h}}$ and eμ final states. The electron correction factors have to be applied for all $\tau \rightarrow \text{e}$ decays in the $\text{e}\tau_{\text{h}}$ and eμ final states. The main reason for different efficiencies is the simulation and reconstruction in the otherwise empty detector. During the τ-embedding, the simulation of the HLT response also takes place in the empty detector. The reconstruction algorithms are not subject to other particles and PU in the event. As a result, different *SF* for τ-embedding and regular simulation may occur.

An example for such a difference is the efficiency of a single isolated muon HLT path. Assume a non-isolated muon with a significant amount of hadronic activity in its vicinity. In the τ-embedded samples, the isolated muon HLT path for this muon still has high efficiency, because the muon will always be isolated during the HLT simulation. However, after merging, the simulated muon is placed into a region with high hadronic activity. However, the trigger response cannot be adapted to this situation.

For muons, correction factors for the isMedium muon identification, the muon isolation and the two single muon HLT paths[1] are derived. For the electron, correction factors for the 90WP electron identification working point, the electron isolation and two single electron HLT paths[2] are derived.

The efficiency measurements are performed using the Tag and Probe method described in the previous section. The different efficiencies are measured as conditional propabilities and have to be applied in combination:

$$\begin{aligned} \epsilon(\text{ID}) &= \epsilon(\text{ID}) \\ \epsilon(\text{Iso,ID}) &= \epsilon(\text{Iso}|\text{ID}) \cdot \epsilon(\text{ID}) \\ \epsilon(\text{HLT,Iso,ID}) &= \epsilon(\text{HLT}|\text{Iso, ID}) \cdot \epsilon(\text{Iso}|\text{ID}) \cdot \epsilon(\text{ID}). \end{aligned}$$

As a result, the combination of all corrections has to be applied during analysis to obtain the correct result. The efficiencies are measured as a function of $p_{\text{T}}$ and η. In Figure 6.6 (left column), the muon efficiency measurements in the region $|\eta| = [0.0, 0.9]$ are shown. The electron correction factors are measured separately for positive and negative η regions to better account for the differences between these two regions. In Figure 6.6 (right column), the efficiency measurements for $\eta = [0.0, 1.0]$ are shown.

For the isMedium muon identification (Figure 6.6, top left), the efficiency in the $\mu \rightarrow \mu$ embedded sample is very similar to simulation, while a small difference compared to data is visible. For electron identification (Figure 6.6, top right), a larger difference between the

[1] HLT_IsoMu24 and HLT_IsoMu27

[2] HLT_Ele32_WPTight_Gsf and HLT_Ele35_WPTight_Gsf

μ → e embedded samples and the simulation is visible. Here, the μ → e embedded sample efficiency is 80% compared to the expected efficiency of 90% for the 90WP. As mentioned in Section 3.3.7, the electron identification is based on a multivariate classifier that utilizes variables sensitive to the matching of the ECAL cluster and the electron track. In μ → e embedded samples, these variables have a different distribution than the simulated samples used for the classifier's training. Since the classifier was trained using simulation, it does not identify electrons in the μ → e embedded samples with the same efficiency.

In the middle row, the efficiencies and correction factors for a relative isolation of $I_{\mathrm{rel}} < 0.15$ are shown. Here, no large difference between the efficiencies is visible, resulting in small correction factors.

In the bottom row, the efficiencies for two trigger paths are shown. As expected, the efficiencies are higher in the μ → e embedded samples due to the empty detector during the HLT simulation. Nevertheless, the correction factors are in the order of 10% across the whole $p_{\mathrm{T}}$ range.

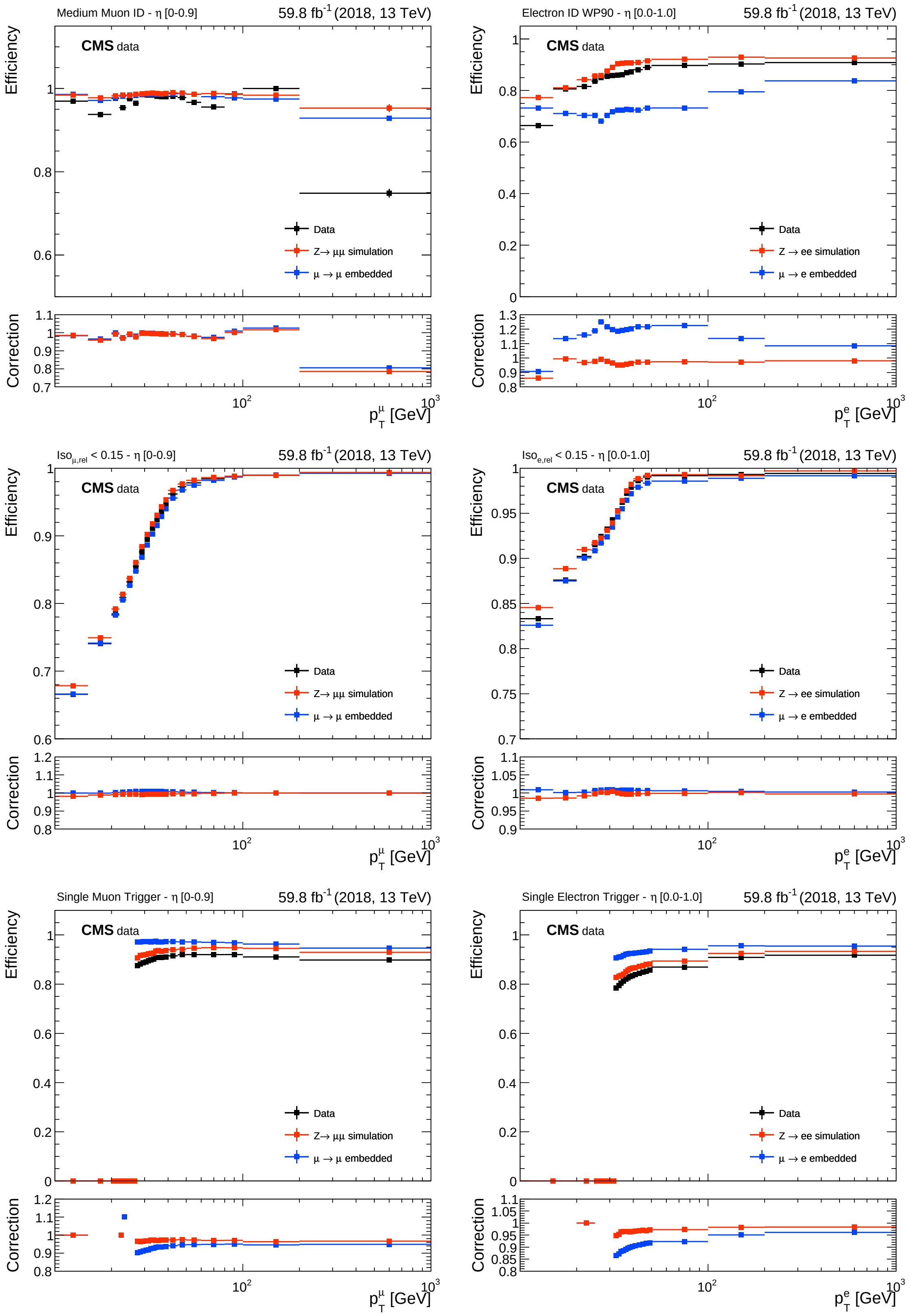


**Figure 6.6.:** A selection of the correction factors derived for muons and electrons. In the top row, the measurements for the electron and muon identification are shown. In the middle row, the measurements for isolation of $I_{\text{rel}} < 0.15$ are shown. The measurements for single electron and single muon HLT paths are shown in the bottom row.

## 6.3. Identification Correction of the $\tau_\mathrm{h}$

For the correction of the *vsJet* discriminant of the DeepTau $\tau_\mathrm{h}$ identification, a measurement in the $\mu\tau_\mathrm{h}$ final state is performed. The selection of events in the final state is based on the selection criteria described in Section 4.2.1. To increase the purity of genuine di-$\tau$ events, a selection cut on $D_\zeta$ >-25 [124] is applied, where the variable $D_\zeta$ is defined as

$$D_\zeta = p_\zeta^\mathrm{miss} - 0.85 p_\zeta^\mathrm{vis} \tag{6.6}$$

$$p_\zeta^\mathrm{miss} = \vec{p}_\mathrm{T}^\mathrm{\,miss} \cdot \vec{\zeta} \tag{6.7}$$

$$p_\zeta^\mathrm{vis} = (\vec{p}_\mathrm{T}^{\,\tau} + \vec{p}_\mathrm{T}^{\,\mu}) \cdot \vec{\zeta}, \tag{6.8}$$

and $\vec{\zeta}$ is a unit vector along the bisectional direction of the muon and the $\tau_\mathrm{h}$. The variable $D_\zeta$ can be used to differentiate between resonant di-$\tau$ decays, W+jets, and $\mathrm{t\bar{t}}$ production. For the latter two, no peak in the $D_\zeta$ distribution is expected. For resonant di-$\tau$ decays on the other hand, the values of $p_\zeta^\mathrm{miss}$ and $p_\zeta^\mathrm{vis}$ are expected to be similar, resulting in a peaking distribution around $D_\zeta \approx 0$. With a selection criterion of $D_\zeta$ > -25, events from W+jets and $\mathrm{t\bar{t}}$ production can be rejected. Additionally, the selection criterion on $m_{\mathrm{T},\mu}$ is lowered from 70 GeV to 60 GeV further increasing the purity of resonant di-$\tau$ decays.

The contributions of other processes are estimated using simulation. This includes W+jets, $\mathrm{t\bar{t}}$, diboson decays, and $\mathrm{Z} \to \mu\mu$ production. For QCD multijet production, an extrapolation from the same-sign region is used. After this selection, the $m_\mathrm{vis}$ distribution of the $\mu\tau_\mathrm{h}$ system is used as the discriminating observable. This distribution is shown in Figure 6.7 on the left.

In addition, a control region in the di-$\mu$ final state is used to constrain uncertainties of the $\tau$-embedded sample. For the control region, the $\mu \to \mu$ embedded sample is used to estimate the di-$\mu$ contribution from Z boson decays, which correspond to 99.9% of events in this region. Since the same selection, trigger, and identification requirements are used for the $\mu \to \mu$ and the $\tau$-embedded samples in the signal region, the same selection efficiency correction described in Section 6.1 is applied for the $\tau$-embedded sample in the $\mu\tau_\mathrm{h}$ final state and the $\mu \to \mu$ embedded sample. As a result, the uncertainties related to this correction can be constrained. As the di-$\mu$ control region contains about 100 times more events than the $\mu\tau_\mathrm{h}$ signal region, the selection efficiency correction is entirely determined by the measurements in the control region. Consequently, the measurement of the $\tau_\mathrm{h}$ identification efficiency in the $\mu\tau_\mathrm{h}$ final state is independent of the $\tau$-embedding selection efficiency correction. A counting experiment is performed using a single $m_\mathrm{vis}$ bin in the control region. The resulting yields are shown in Figure 6.7 on the right.

For the correction factor determination, an extended binned likelihood is constructed

$$\mathcal{L}(n_i|\mu, \theta_j) = \prod_{i\in\mathrm{bins}} P(n_i|\mu \cdot S(\theta_j) + B(\theta_j)) \cdot \prod_{j\in\mathrm{syst}} G(\theta_j^0|\theta_j) \tag{6.9}$$

where $n_i$ are the number of observed events in each bin, and $P(n_i|\mu \cdot S + B)$ represents the Poisson probability to observe $n_i$ events given the signal $S$ and background $B$ predictions. The Parameter Of Interest (POI) $\mu$ is used to scale the normalization of the contributing signal. Systematic uncertainties are incorporated by the terms $G(\theta_j^0|\theta_j)$, which serve as

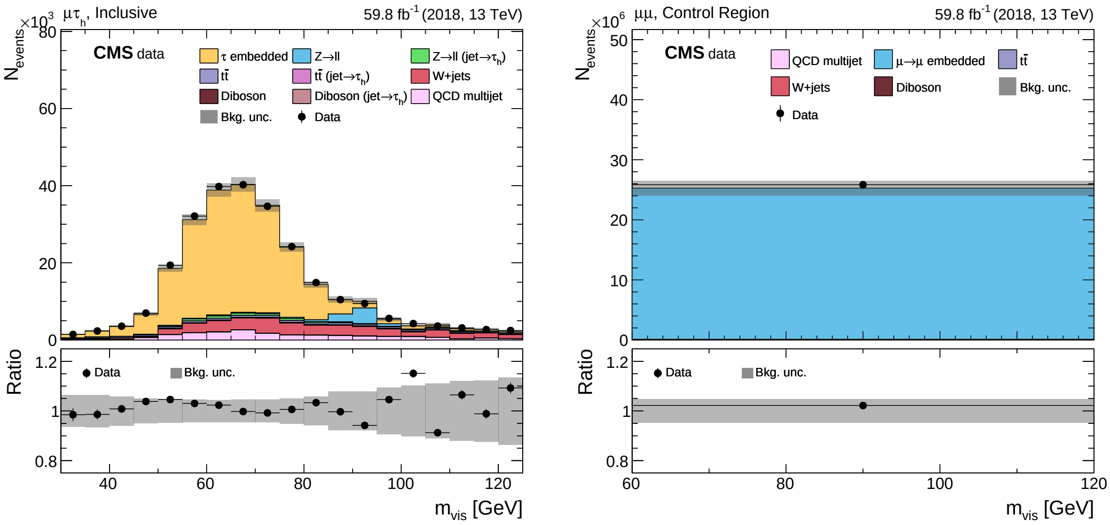


**Figure 6.7.:** On the left, the distribution of $m_{\text{vis}}$ in the $\mu\tau_h$ channel after the application of additional selection criteria for a purer selection of τ-embedded events is shown. The di-μ control region is shown on the right. The uncertainty band corresponds to the prefit uncertainty.

penalty terms for the fit. The parameter $\theta_j^0$ represents the nominal value of the nuisance parameter $j$. This likelihood function can also include statistical uncertainties by following the approach described in [125]. In this approach, instead of adding one nuisance parameter per bin and process, as suggested in [126], a single nuisance parameter scaling the yield of the sum of all processes in a bin is used. This approach reduces the number of parameters in the likelihood fit and has the same statistical representation since the statistical uncertainties of the different processes are independent and can thus be combined. These uncertainties are denoted as bin-by-bin uncertainties. After constructing the likelihood, a maximum likelihood fit is performed to determine the result.

For the maximum likelihood estimation, the $m_{\text{vis}}$ distribution and the di-μ control region are used while the yield of the τ-embedded sample is scaled by the POI $\mu$. The uncertainty model used for the measurement is nearly identical to the one described in Section 7.1.3, with a few differences:

- For the QCD multijet production, a 30% uncertainty on the extrapolation factor from the same-sign region is used.
- For the τ-embedded sample, no correction of the $\tau_h$ energy is applied and an uncertainty of 1.2% is used. Since the measurement of the $\tau_h$ identification correction and the $\tau_h$ energy correction depend on each other, an assumption has to be made here.

This measurement approach is delicate, as the signal strength $\mu$ can be influenced by the imperfect modelling of other background processes. Therefore it is essential that the uncertainty model of the background processes is close to the uncertainty model of a typical target analysis and includes the knowledge of all uncertainty sources coming from

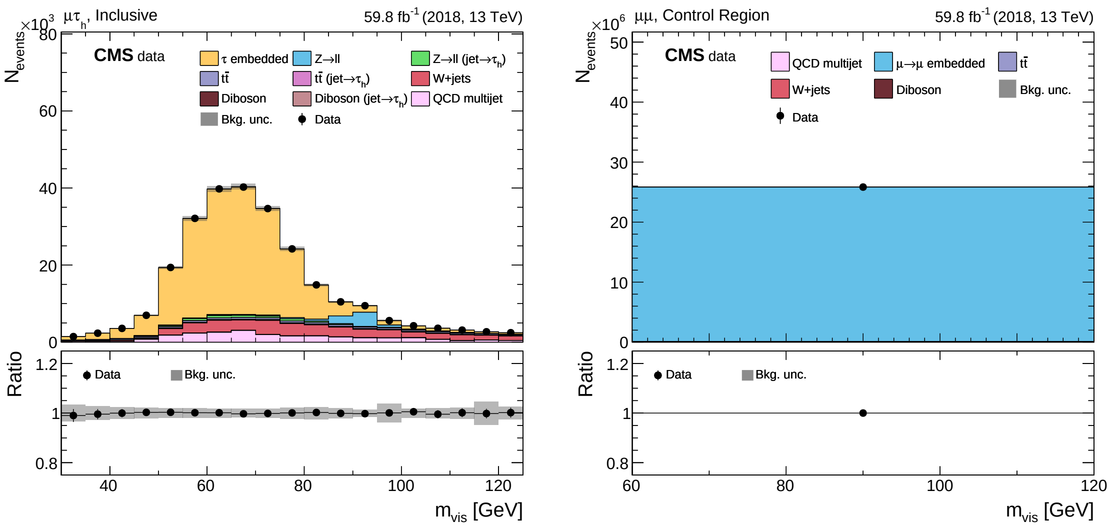


**Figure 6.8.:** On the left, the distribution of $m_{\text{vis}}$ in the $\mu\tau_\text{h}$ channel after the maximum likelihood fit is shown. The di-μ control region is shown on the right. The uncertainties shown are minimized, based on the provided uncertainty model.

other background processes. Otherwise, any miss-modelling could be attributed to the correction factor of the τ-embedded samples, and the result would be biased.

The correction factors are measured for all *vsJet* working points listed in Table 3.3. In addition, three separate categorizations are made:

- *Inclusive*, one single category. The resulting distributions after the fit are shown in Figure 6.8.
- *pt-binned*, by using multiple categories based on the $p_\text{T}$ of the $\tau_\text{h}$. The binning is

  $$p_\text{T}^\tau \in [20, 25, 30, 35, 40, \text{inf}].$$

  These correction factors are used in the $\text{e}\tau_\text{h}$ and $\mu\tau_\text{h}$ final states.
- *dm-binned*, by using multiple categories based on the $\tau_\text{h}$ DM. These correction factors are split into a One prong, One prong + $\pi^0$, and three prong category. They are used in the $\tau_\text{h}\tau_\text{h}$ final state.

The resulting correction factors for the Tight working point are shown in Figure 6.9. On the left, the *pt-binned* corrections are shown, on the right the *dm-binned* corrections. The most significant uncertainties of the measurement are the uncertainty on the QCD multijet estimate, which is constrained to ≈ 20% by the fit. Other impactful uncertainties are the bin-by-bin uncertainties of several less populated bins. The nuisance parameter related to the $\tau_\text{h}$ energy scale is pulled to a value slightly below one, indicating, that the energy of $\tau_\text{h}$ is a bit too large. The correction factors have an uncertainty of less than 3%, with a correction value close to one, indicating good modelling of the $\tau_\text{h}$ in τ-embedded samples.

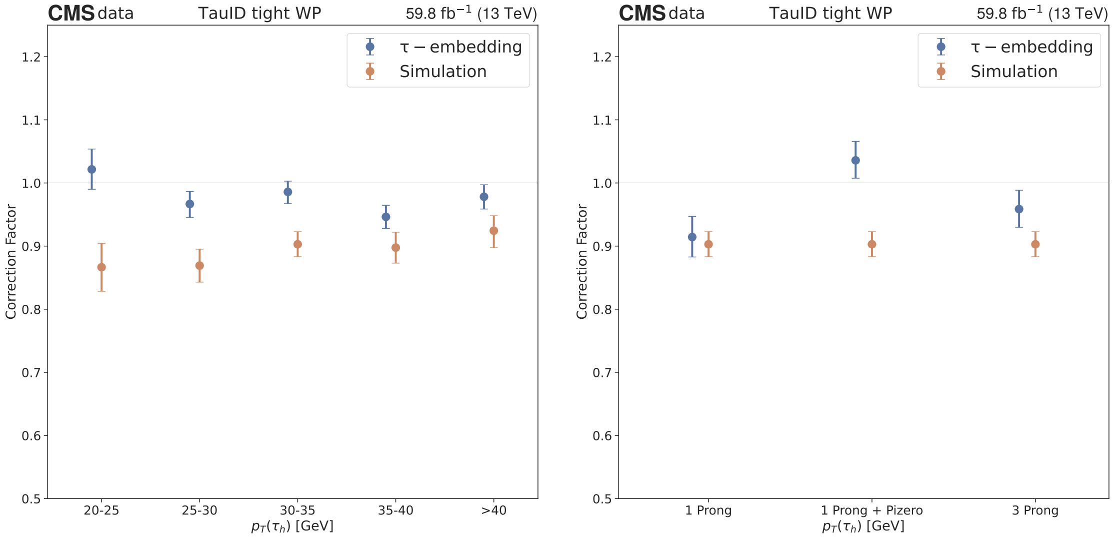


**Figure 6.9.:** On the left, the correction factors binned in $p_\mathrm{T}$ of the $\tau_\mathrm{h}$ for the TIGHT *vsJet* working point are shown. The correction factors binned in the decay mode of the $\tau_\mathrm{h}$ for the TIGHT *vsJet* working point are shown on the right. The correction factors for τ-embedding are shown in blue, and the correction factors that were derived for simulation are shown in red. The measurement procedure for simulation corrections is similar however, the resulting corrections are larger.

## 6.4. Energy Corrections for $\tau_\mathrm{h}$ and Electron

After the *vsJet* correction factors are determined, a measurement of the $\tau_\mathrm{h}$ energy scale is performed. This measurement is also performed in the $\mu\tau_\mathrm{h}$ channel; however, the $m_\mathrm{vis}$ distribution range is reduced to not include a reflection of the Z boson resonance from $\mathrm{Z} \to \mu\mu$, where one muon is misidentified as a $\tau_\mathrm{h}$. In addition, the $\mathrm{F_F}$ method described in Section 4.2.4 is used for the estimation of jet $\to \tau_\mathrm{h}$ events. In Figure 6.10, the input distribution is shown on the left.

To determine the correct energy scale, the energy of $\tau_\mathrm{h}$ in the τ-embedded sample is varied in steps of 0.1% between -2% and 2%. Since the $\mathrm{F_F}$ estimation of jet $\to \tau_\mathrm{h}$ depends on the number of events in the anti-isolated region, it also depends on the $\tau_\mathrm{h}$ energy. As a result, both the jet $\to \tau_\mathrm{h}$ and the τ-embedding contributions are varied.

Intermediate energy scales that were not explicitly calculated, e.g. -0.05%, can be obtained by performing an interpolation called template morphing. The yields from τ-embedded and jet $\to \tau_\mathrm{h}$ events are estimated based on the contributions from the available neighbouring $\tau_\mathrm{h}$ energy scale histograms. The morphed histograms are obtained by interpolating between the cumulative distribution functions of the two neighbouring histograms. If changes in the shape of the input templates are correlated with the change in $\tau_\mathrm{h}$ energy, an accurate interpolation can be performed.

The uncertainty model described in Section 7.1.3 is also used for this measurement. The corrections for the *vsJet* identification and their uncertainties as described in the previous

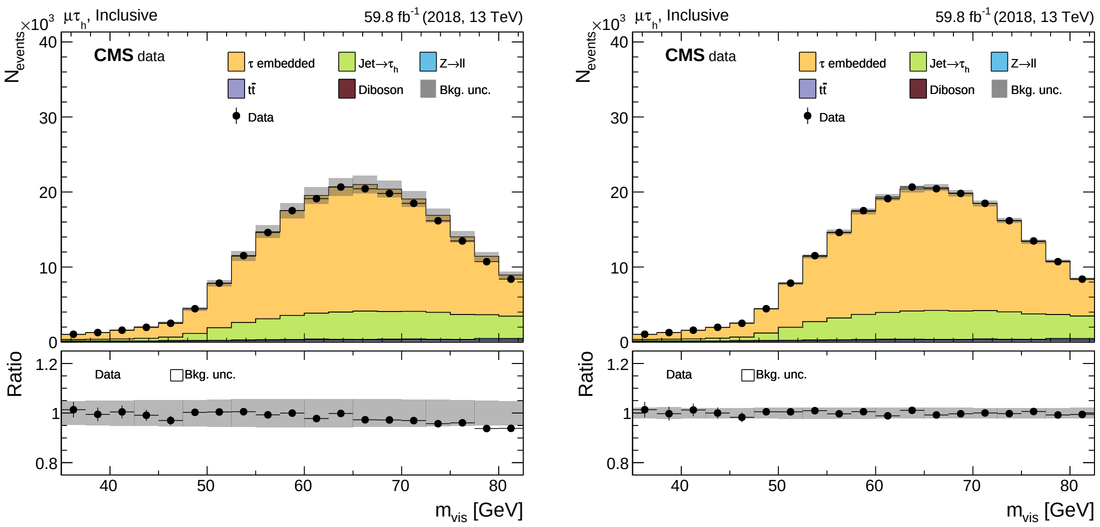


**Figure 6.10.:** On the left, the distribution of $m_{\text{vis}}$ used for the $\tau_{\text{h}}$ energy scale measurement before the fit is shown. On the right, the distribution after the fit is shown.

section are included. The $\tau_{\text{h}}$ energy scale correction factor can be determined using a likelihood scan. The resulting negative Log-likelihood $-2\ln\Lambda(r)$ of the scan is visualized in Figure 6.11 (left), whereas the resulting distribution is shown in Figure 6.10 on the right. The uncertainty of the correction factor is determined using the interval, where $-2\ln\Lambda = 1$, which corresponds to the 68% confidence level interval. A $\tau_{\text{h}}$ energy scale correction factor of

$$SF(\tau_{\text{h}}) = -\left(1.35^{+0.39}_{-0.39}\right)\ \%$$

is measured for the 2018 τ-embedded sample. This result is in line with the expectation from the *vsJet* correction measurement described in the previous selection. The measured correction factor is in good agreement with previous measurements. The scan shows that the morphing procedure does not reproduce the intermediate energy scale values perfectly.

The measurement of the electron energy scale is performed in the di-e final state. By using the $\mu \longrightarrow$ e embedded samples, the effect of the τ-embedding method on the measurement of the electron energy can be determined, and a correction factor derived. Contributions from other processes are modelled using MC samples; however, they only have a limited impact on the result since the contribution from $\mu \longrightarrow$ e embedded events dominates the measurement. More than 99% of events in this final state are modelled using $\mu \longrightarrow$ e embedded events. The selection of events is based on the di-e selection described in Section 4.2.1. Different electron energy scale corrections are applied to the τ-embedded sample to obtain different templates for a likelihood scan. Here, the energy is varied in steps of $0.05\%$ between $-1.5\%$ and $1.0\%$. Intermediate variations of the energy scale are again obtained using the morphing procedure outlined above.

The electron energy scale is measured separately for the barrel ($|\eta| \leq 1.479$) and endcap ($|\eta| > 1.479$) regions of the ECAL, due to the different subdetectors used as described in Section 3.2.3. The distribution of the di-e mass is used in the likelihood fit. In Figure 6.12,

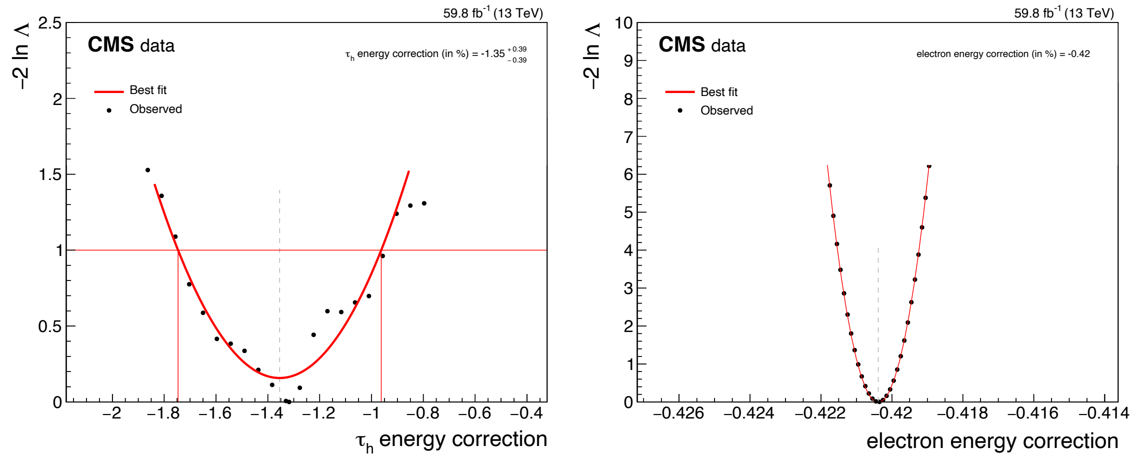


**Figure 6.11.:** The negative Log-likelihood of the $\tau_\mathrm{h}$ energy scale measurement is shown on the left. The negative Log-likelihood of the electron energy scale measurement is shown on the right.

the distribution of the di-e mass in the barrel region (left) without any correction is shown. The resulting negative log-likelihood scan is shown in Figure 6.11 (right).

Since the measurement is performed in a channel dominated by $\mu \rightarrow \mathrm{e}$ embedded events, a small change in the energy correction has a large influence on the fit. Instead of using the uncertainty obtained from the scan, a conservative uncertainty of $0.5\%$ ($1.25\%$) in the barrel (endcap) is used. The resulting corrections

$$SF(e, \text{barrel}) = -\left(0.42^{+0.5}_{-0.5}\right)\ \%$$
$$SF(e, \text{endcap}) = -\left(0.69^{+1.25}_{-1.25}\right)\ \%$$

are in good agreement with previous measurements. As apparent from the distribution after the fit shown in Figure 6.12 (right), no perfect modelling of the Z boson resonance as observed in the data is achieved. The Z boson resonance resolution is better in the $\mu \rightarrow \mathrm{e}$ embedded sample compared to the resolution in data. Additional smearing of the electron energy would be required to obtain an improved modelling. However, this smearing correction can be safely neglected since the correction will be applied for electrons coming from τ lepton decays, where much larger smearing comes from the two neutrinos produced during the decay.

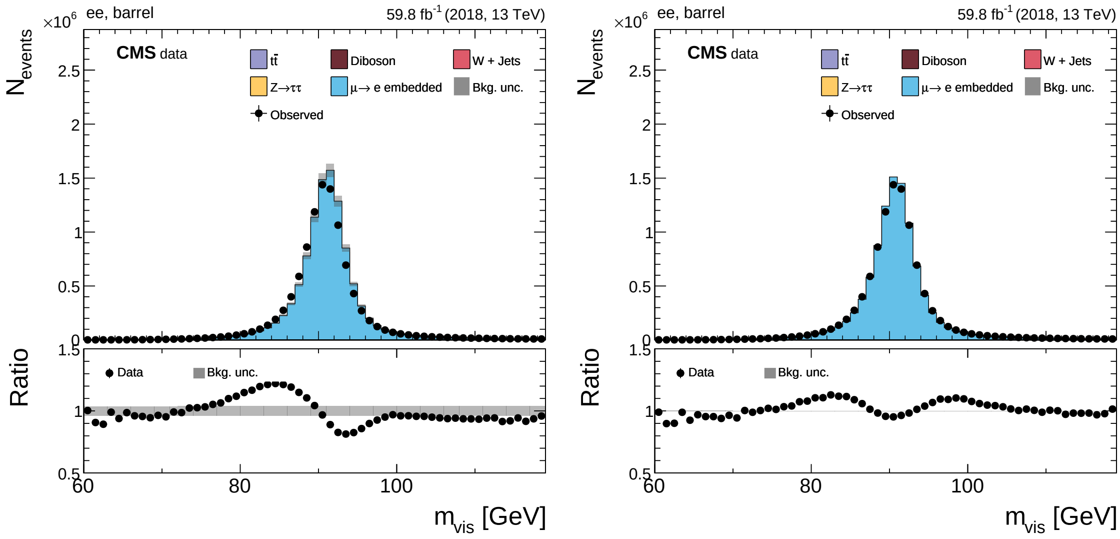


**Figure 6.12.:** On the left, the distribution of the di-e mass before the fit is shown. On the right, the distribution after the fit is shown.

## 6.5. Uncertainty Model

All corrections described in this chapter must be applied to τ-embedded samples when using them in a typical target analysis. In addition, the following uncertainty model has to be assigned.

- A flat 4% uncertainty is assigned to all τ-embedded events. This general uncertainty accounts for the uncertainty on the unfolding efficiencies and the general understanding of the τ-embedding method. This uncertainty is chosen conservatively.

- For the lepton corrections, a 2% uncertainty on the ID and Trigger corrections is assigned. Since the HLT response is different in τ-embedding and simulation, the efficiency of the HLT has to be treated as uncorrelated between τ-embedded and simulated events. However, the data are used for the calculation of $SF_{\mathrm{MC}}$ and $SF_{\mathrm{EMB}}$ in Equation (6.5). As a result, the correlation between the correction factors for simulation and τ-embedding are chosen to be 50%. This correlation can be implemented by using two nuisance parameters. The first nuisance parameter acts on simulation and the τ-embedded sample. For simulation, the full amplitude of the nuisance parameter is used; on the τ-embedded sample, 50% of the nuisance parameter amplitude is applied. The second nuisance parameter is only assigned to the τ-embedded sample with a strength of $\sqrt{1-0.5^2}$ [10]. The same is true for the electron and muon ID correction uncertainties. Since the isolation correction factors are small for both electrons and muons, no additional uncertainty is assigned.

- For the $\tau_{\mathrm{h}}$ identification corrections, the uncertainties coming from the measurement are used. In the $e\tau_{\mathrm{h}}$ and $\mu\tau_{\mathrm{h}}$ final state, the pt-binned corrections should be applied, whereas, in the $\tau_{\mathrm{h}}\tau_{\mathrm{h}}$ final state, the dm-binned correction should be applied. Since the data and some other background processes are shared between the measurements for τ-embedded samples and simulation, a correlation of 50% should be assumed.

- The uncertainty in the $\tau_\mathrm{h}$ energy scale correction is taken directly from the measurement. A correlation of 50% is assumed between τ-embedding and simulation.
- For the uncertainty in the electron energy scale, 0.5% in the barrel and 1.25% in the endcap are used. Since the measurement is performed differently for τ-embedding and simulation, no correlation is assumed.
- For the contribution from $\mathrm{t\bar{t}}(\tau\tau)$ included in the τ-embedded sample, an additional 10% uncertainty is added. The variations are calculated by using $\mathrm{t\bar{t}}(\tau\tau)$ simulation and then adding and subtracting 10% of the $\mathrm{t\bar{t}}(\tau\tau)$ template from the τ-embedded sample template. When applying τ-embedded events in the target analysis, the contribution from $\mathrm{t\bar{t}}(\tau\tau)$ has to be removed from the $\mathrm{t\bar{t}}$ simulation to avoid double counting this process. The same procedure is performed for $\mathrm{VV}(\tau\tau)$; however, since the contribution from this process is small, no additional uncertainty is assigned. The $\mathrm{VV}(\tau\tau)$ process is still to be removed from the diboson simulation to avoid double counting.

# 7 Application of τ-embedding in a H → ττ measurement

The τ-embedded samples can be utilized in several analysis scenarios. In this chapter, the application of the UL τ-embedded samples to an H → ττ measurement is presented. The analysis approach is based on the previous CMS measurement [9] and represents the foundation work of a Run II + Run III measurement. More than 300 $\mathrm{fb}^{-1}$ of measured collision data are expected for this analysis. The results presented in this Chapter are not targeted to repeat the existing measurement but to showcase the potential and the applicability of the τ-embedded samples in a real analysis scenario. A comparison of the results with a simulation-based approach and the existing measurement from [9] will be presented. While the τ-embedded samples are available for the full Run II data set, the comparisons will be limited to the $\mu\tau_\mathrm{h}$ channel and the era 2018 to highlight the differences between the approaches. The target of the analysis is to measure the signal strength for inclusive Higgs boson production and Higgs boson production split by the production modes ggH and qqH.

A new analysis framework has been developed in the scope of this thesis to cope with the enormous amount of data that has to be processed for the anticipated future analysis. A conceptual description of this new framework can be found in Appendix A.1.

## 7.1. Analysis Strategy

The event selection for the analysis was already described in Section 4.2.1. To measure the signal strength $\mu$ of the different production modes, multiple categories targeting different processes are defined. Past iterations of this analysis [9, 127] have shown that this categorization can best be performed using a neural network.

### 7.1.1. Neural Network Classification

A neural network with multiple output nodes is used for the classification task. Compared to binary classification, in which a neural network is trained to distinguish between a signal and a background event, a multi-class neural network is trained to distinguish between different signal and background processes. An event is represented by the neural network input vector $\vec{x}$ containing $i$ input features. The neural network's output is an output vector $\vec{y}$ with length $l$ corresponding to the number of categories in the analysis. The values in this output vector are called scores. By using a softmax activation function for the final layer of the network, the resulting $y_l$ values can be interpreted as probabilities for the event to belong to category $l$ since the sum of all scores is normalized to one.

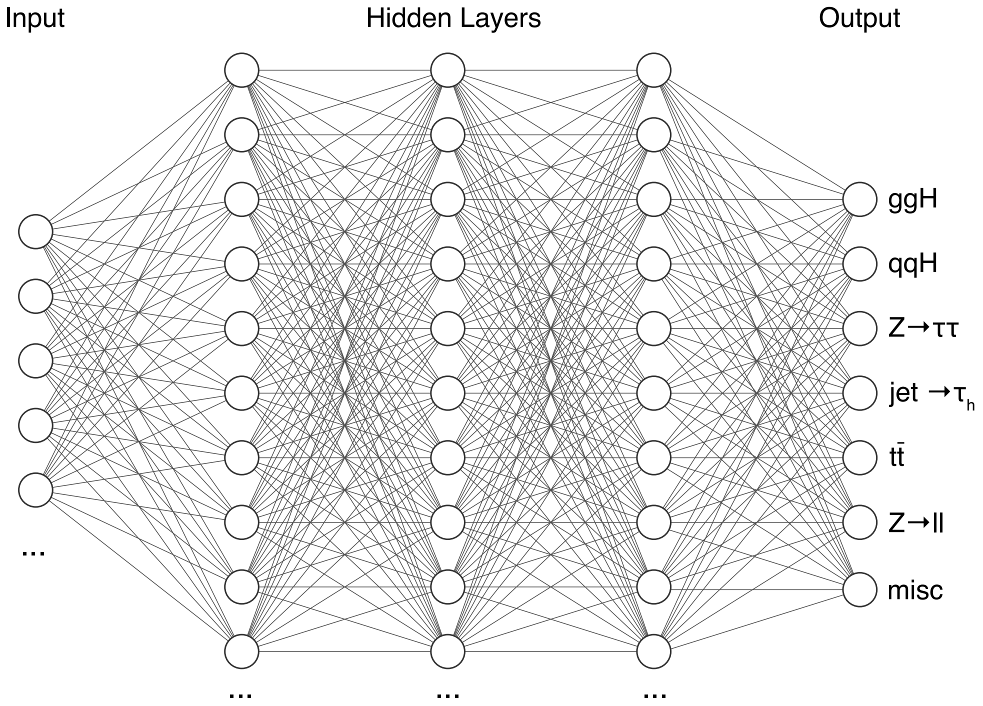


**Figure 7.1.:** Visualization of the neural network architecture.

**Neural Network Architecture** This analysis uses a feed-forward neural network with three hidden layers. Each hidden layer has 500 nodes and uses the hyperbolic tangent activation function. During the training, categorical cross entropy is used as loss function. The network is trained using the Adam optimiser [128] with a learning rate of 0.001. Before the training, the weights are initialized using the Glorot uniform initialization [129]. A dropout of 30% is applied after each hidden layer for regularization and to improve the neural network's generalization properties. In addition, an L2 regularization is applied to avoid weights that grow too large. The training is performed using Keras [130] with TensorFlow [131] as backend. A visualization of the network architecture is depicted in Figure 7.1.

Each output category is designed to target specific physics processes. Two signal categories for the two Higgs production modes are defined. In addition, four categories dedicated to different background processes are used. One category is dedicated to Z ⟶ ττ events, one to jet ⟶ $\tau_h$, one to tt̄ production, and one to Z ⟶ ll. The *misc* category serves as a category for all processes not covered by the other categories.

In Table 7.1 an overview of all categories for the analysis setup using τ-embedded samples and for the setup using simulated samples is given. When using τ-embedding during the training, one must remember that VV(ττ) and tt̄(ττ) are included in the τ-embedded samples. Since the two processes are included in the inclusive tt̄ and Diboson simulation, they must be removed when using τ-embedded samples. The two contributions can be assigned to their respective category when using simulation during the training.

In Figure 7.2, the number of events available is compared between the two training approaches. More than 40 times more τ-embedded events than simulated Z ⟶ ττ events

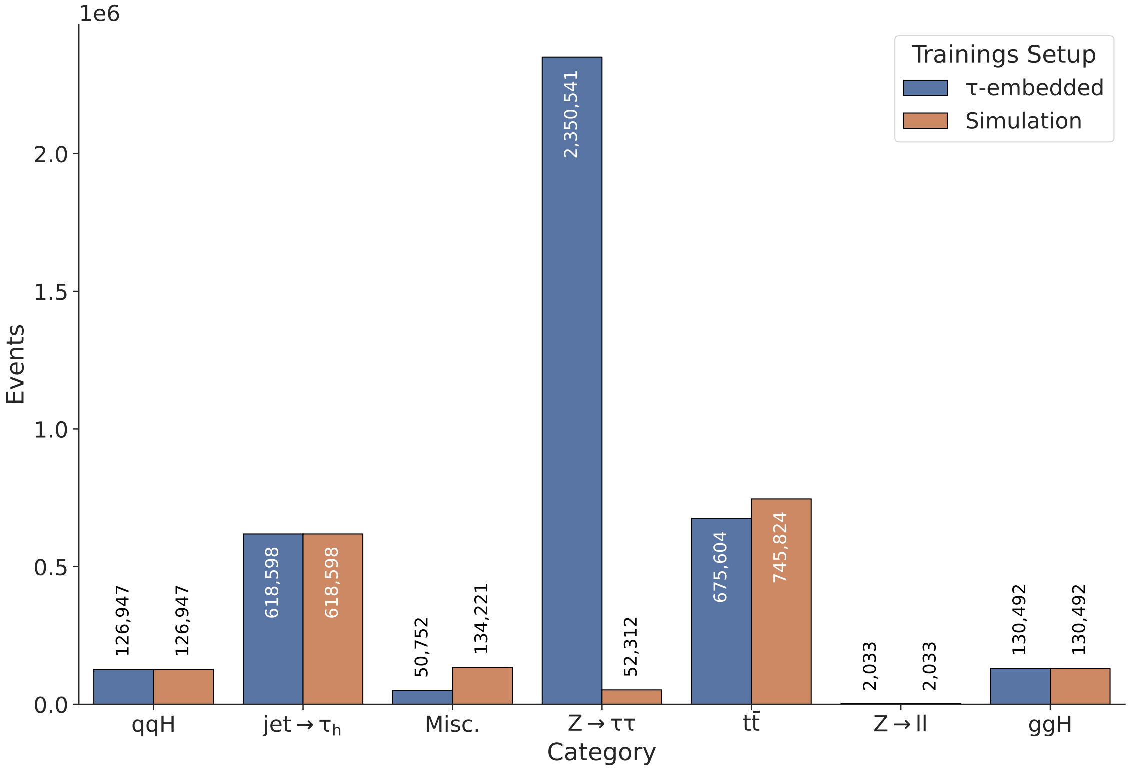


**Figure 7.2.:** Comparison of the number of events available for the training after the event selection. When using τ-embedded samples, the number of events available in the $Z \rightarrow \tau\tau$ category is much larger compared to regular simulation.

are available. A larger sample of events is always beneficial for the neural network training, as it prevents overfitting.

In each training iteration, a batch of 30 events per category is processed before the weights are updated. This approach ensures that τ-embedded events do not dominate the selection of events for a single batch. After one epoch, which consists of 1000 batches, the neural network is validated using the validation sample. The training is stopped after the validation loss does not improve for 50 epochs. The weights from the epoch with the lowest validation loss are used for the final neural network configuration.

The training is performed in a two-fold approach to utilise the entire data set. All data sets are split into two equally sized parts *A* and *B*. The training performed using data set *A* is used to evaluate data set *B* and vice versa. This approach ensures that one can use all available events in the final analysis without bias.

In Figure 7.3, the confusion matrix of the training using 2018 data and τ-embedded samples in the $\mu\tau_h$ channel is shown. The neural network performs well in identifying qqH production with an efficiency of 75%. Most qqH events have at least two jets, a signature the neural network can identify. The neural network can also efficiently identify $Z \rightarrow \tau\tau$, $t\bar{t}$ and $Z \rightarrow ll$ events. For jet → $\tau_h$, a more significant portion of events is misidentified as $Z \rightarrow \tau\tau$, which is the dominant background process in the $\mu\tau_h$ final state. The neural network has difficulties in correctly identifying ggH events. A more

**Table 7.1.:** Assignments of the different physics processes to the neural network categories. When using τ-embedding in the training, it is impossible to split VV(ττ) and $\mathrm{t\bar{t}}$(ττ) from the τ-embedded sample. The two processes are included in the Z ⟶ ττ category.

| Category | Training with τ-embedding | Training with Simulation |
|---:|---|---|
| ggH | H ⟶ ττ *(sim.)* | H ⟶ ττ *(sim.)* |
| qqH | H ⟶ ττ *(sim.)* | H ⟶ ττ *(sim.)* |
| Z ⟶ ττ | τ-embedding | *Z* ⟶ ττ *(sim.)* |
| jet ⟶ $\tau_h$ | $F_F$ method | $F_F$ method |
| $\mathrm{t\bar{t}}$ | $\mathrm{t\bar{t}}$(ll) *(sim.)* | $\mathrm{t\bar{t}}$(ll), $\mathrm{t\bar{t}}$(ττ) *(sim.)* |
| Z ⟶ ll | Z ⟶ ll *(sim.)* | Z ⟶ ll *(sim.)* |
| misc | VV(ll) *(sim.)* | VV(ll), VV(ττ) *(sim.)* |

considerable confusion with qqH, Z ⟶ ττ and Z ⟶ ll events can be observed. These four processes have a peaking structure in the di-τ mass distribution; ggH and qqH have the resonance at the Higgs boson mass, while Z ⟶ ττ and Z ⟶ ll have a resonance at the Z boson mass. Since the Z and the Higgs boson resonances with τ leptons in the final state overlap due to the neutrino contribution, it is hard to distinguish these four processes. The confusion matrix of the training using simulation yields only minor differences and is shown in the Appendix in Figure A.9.

### 7.1.2. Input Features

The input features for the training are selected based on two criteria:

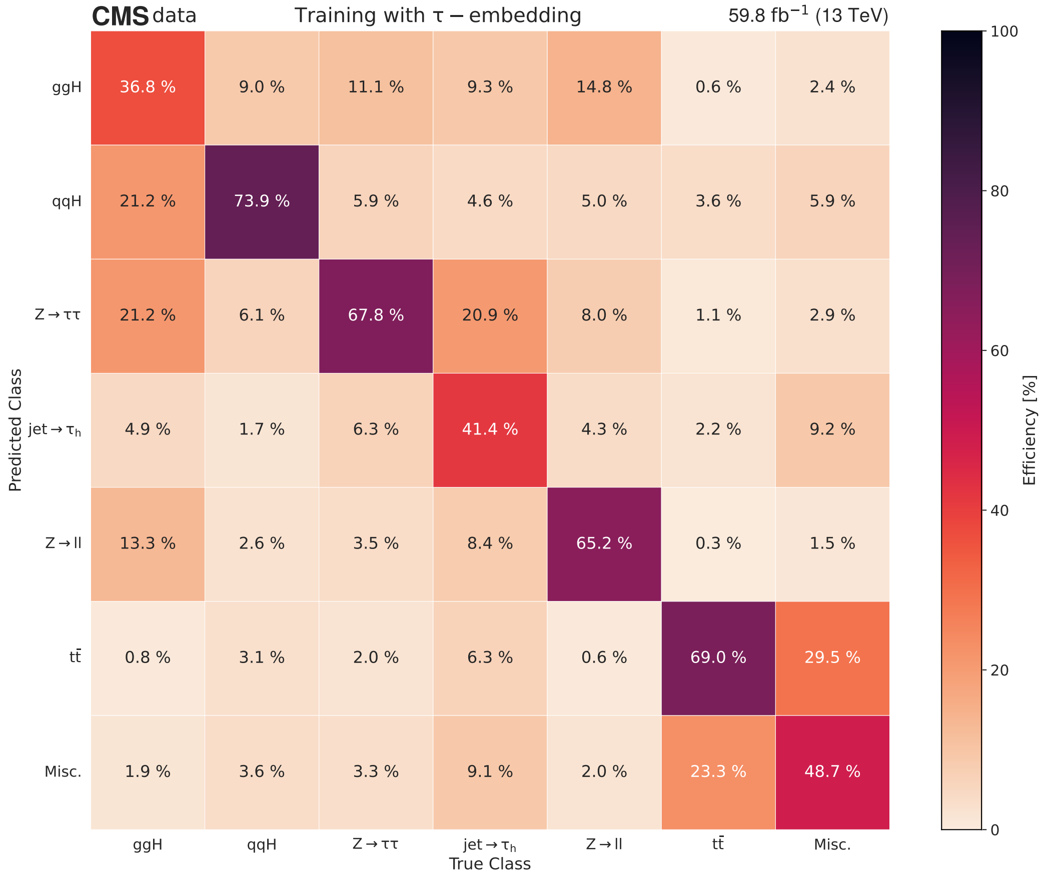
1. The feature must provide a good description of the data. Goodness-of-Fit (GoF) tests are performed to validate the modelling. A histogram with ten equally filled bins (or less) is created for each test. The bin edges are determined using the quantiles of the data, such that each bin contains 10% of the data. The saturated GoF [132, 133] test is the main figure of merit, but the Kolmogorov-Smirnov [134], and the Anderson-Darling [135] tests are also performed.

   The saturated test statistic is based on calculating the likelihood ratio of the full model likelihood (as described in Equation (6.9)) and the saturated likelihood $\mathcal{L}_{\text{saturated}}$. The latter represents the hypothesis that the observed data $x_i$ matches the hypothesis $\mu_i$ in every bin $i$. With a simplified example likelihood of the form

$$\mathcal{L} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_i^2}\right), \tag{7.1}$$

   where $\sigma_i$ represents the uncertainty on the $i$th bin value, the saturated likelihood is given by

$$\mathcal{L}_{\text{saturated}} = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma_i^2}}, \tag{7.2}$$



| Predicted Class \ True Class | ggH | qqH | Z→ττ | jet→τ$_h$ | Z→ll | tt̄ | Misc. |
|---|---|---|---|---|---|---|---|
| ggH | 36.8 % | 9.0 % | 11.1 % | 9.3 % | 14.8 % | 0.6 % | 2.4 % |
| qqH | 21.2 % | 73.9 % | 5.9 % | 4.6 % | 5.0 % | 3.6 % | 5.9 % |
| Z→ττ | 21.2 % | 6.1 % | 67.8 % | 20.9 % | 8.0 % | 1.1 % | 2.9 % |
| jet→τ$_h$ | 4.9 % | 1.7 % | 6.3 % | 41.4 % | 4.3 % | 2.2 % | 9.2 % |
| Z→ll | 13.3 % | 2.6 % | 3.5 % | 8.4 % | 65.2 % | 0.3 % | 1.5 % |
| tt̄ | 0.8 % | 3.1 % | 2.0 % | 6.3 % | 0.6 % | 69.0 % | 29.5 % |
| Misc. | 1.9 % | 3.6 % | 3.3 % | 9.1 % | 2.0 % | 23.3 % | 48.7 % |

**Figure 7.3.:** Confusion matrix of the training using 2018 data and τ-embedded samples in the $\mu\tau_h$ channel. Here, the confusion for fold *A* of the two-fold training is shown; however, the differences compared to fold *B* are on the sub-percent level.

since the exponential function of $\mathcal{L}$ is one when using $x_i = \mu_i$. The saturated test statistic $t_{\text{saturated}}$ can then be calculated using the ratio of the two likelihoods

$$\lambda_{\text{saturated}} = \frac{\mathcal{L}}{\mathcal{L}_{\text{saturated}}} = \prod_{i=1}^{N} \exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_i^2}\right), \tag{7.3}$$

and then calculating the logarithm of the ratio

$$t_{\text{saturated}} = -2\ln\lambda_{\text{saturated}}. \tag{7.4}$$

This test statistic $t_{\text{saturated}}$ is always positive, and the smaller the value, the better the agreement between the observation and the model. The saturated model approach can be generalized for arbitrarily complex likelihoods and thus represents a rather general approach to quantifying the goodness of a model. It does not consider all potential differences between observation and model, e.g. it does not consider the sign of differences between the data and the model.

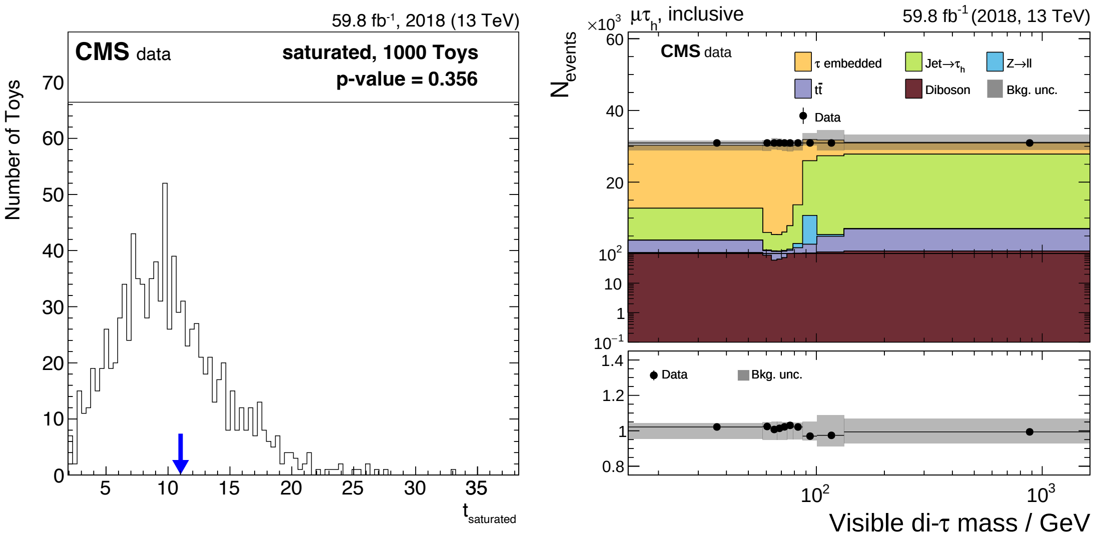


**Figure 7.4.:** On the left, the resulting test statistic $t_{\text{saturated}}$ for the saturated Goodness-of-Fit test applied to $m_{\text{vis}}$ in the $\mu\tau_{\text{h}}$ final state is shown. The blue arrow indicates the value of the observed test statistic $t_{\text{obs}}$. On the right, the input distribution of $m_{\text{vis}}$ is shown. The binning is chosen so that each bin contains $10\%$ of the observed data.

For the GoF test, 1000 toy distributions are generated by varying the model parameters within the given uncertainties. After the test statistic is calculated for each toy, a p-value can be obtained by comparing the observed test statistic value $t_{\text{obs}}$ to the distribution of the sample distribution of $t$. The p-value $p$ is calculated using the cumulative distribution function of the toy test statistics $f(t)$

$$p = \int_{t=t_{\text{obs}}}^{\infty} f(t)\mathrm{d}t. \tag{7.5}$$

In cases with a p-value below 5%, one decides to investigate further. One example distribution of the test statistic and the corresponding input distribution is shown in Figure 7.4 for $m_{\text{vis}}$ in the $\mu\tau_{\text{h}}$ final state.

2. The feature must have discriminating power to differentiate between processes. One can do this in two ways. The first way is to rank the discriminative power of the features used in the neural network. Such a ranking can be calculated using a Taylor expansion of the neural network output function as described in more detail in [136]. The first- and second-order Taylor coefficients can be used to rank features by their importance and disregard features with a negligible impact on the neural network scores. The second way is to use a toy data set, in which the data is replaced by the model expectation and compare the analysis results when using different input features in the neural network. A feature can be disregarded if it does not improve the sensitivity of the analysis.

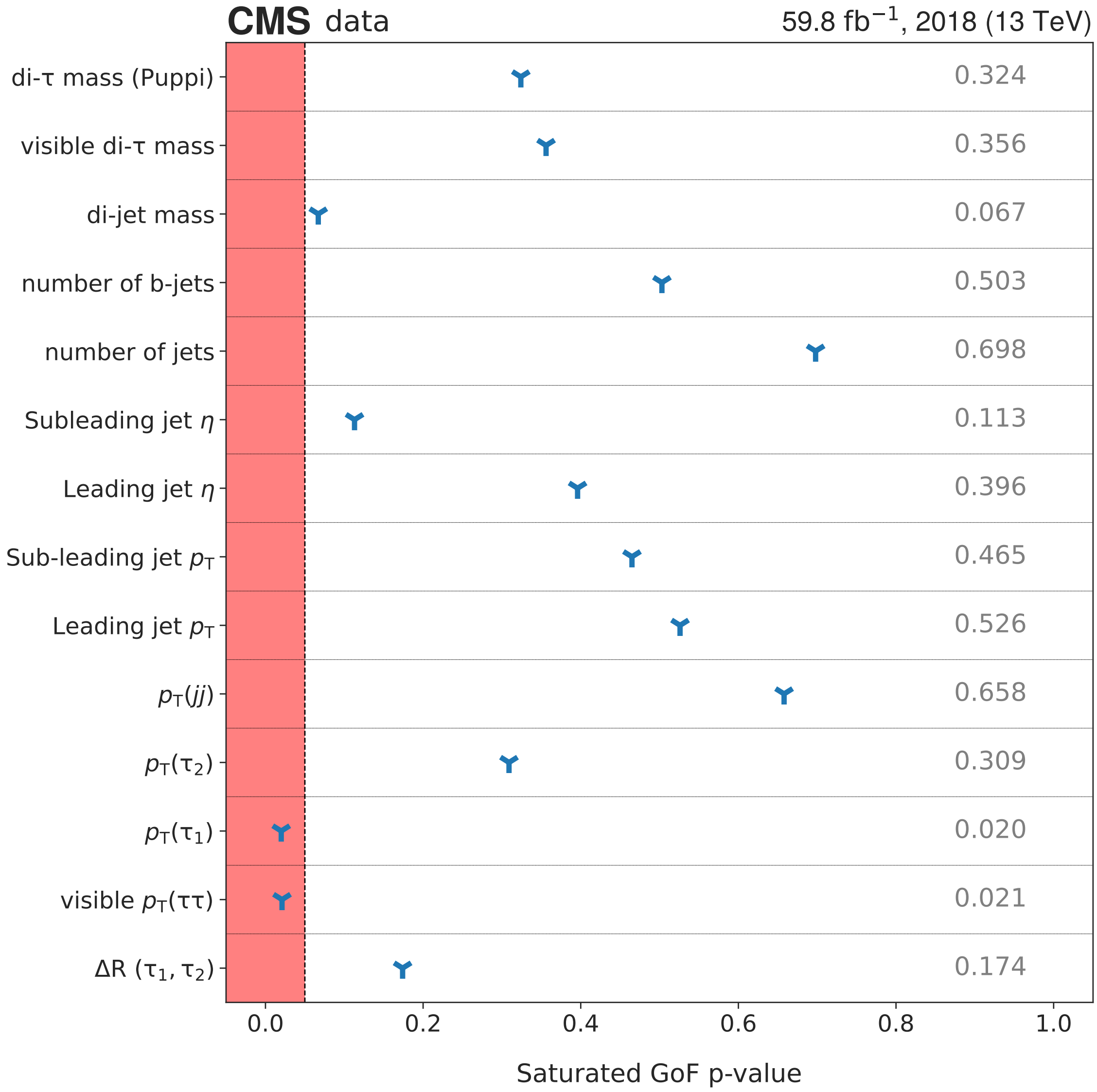


**Figure 7.5.:** GoF p-values for the $\mu\tau_h$ final state when using τ-embedding. The red area represents the region below a 5% significance level. The resulting p-values are listed on the right.

All input features undergo preprocessing before being used in the neural network training by setting their distributions to a mean of zero and a standard deviation of one. The preprocessing ensures that all features are treated equally in the training process. If a feature is not available for a given event (e.g. there is no second jet in the event), the value of the feature is set to -10.

The p-values of the saturated GoF test for 14 selected input features when using τ-embedded samples are shown in Figure 7.5. In the Appendix in Figures A.10 to A.16, the distributions of all selected input features can be found. They can be categorized into the following groups.

- **Mass Variables**: The most important features are $m_{\text{vis}}$ and the di-τ mass. While the first one is constructed from the visible decay products for the two τ leptons, the second one aims to reconstruct the true mass of the di-τ pair, taking the neutrinos into account. The algorithm is based on the SVFit algorithm [137] but uses a more simplified likelihood function during the evaluation, which improves the runtime of the algorithm by up to two orders of magnitude. The mass of the di-jet system $m_{jj}$ is the third mass variable used for the training.

- **Jet Variables**: Several variables related to the jets of the events are included in the training. The number of jets, and the $p_{\text{T}}$ and η of the two leading jets help to identify qqH production since two jets are expected for this production mode. The number of b-jets in the event can be used to tell apart $\text{t}\bar{\text{t}}$ production from other processes. The GoF tests show that all jet-related variables are well-modelled when using τ-embedded samples. This is expected, and one of the main benefits of using τ-embedded samples, since the jets in the event are untouched by the τ-embedding method and therefore identical to the jets observed in data.

- **Di-τ Variables**: The last set of variables is related to the two τ lepton decay products, namely the $p_{\text{T}}$, ΔR between the τ lepton candidates, and the visible $p_{\text{T}}$ of the τ lepton pair. For these input features, two p-values fall below the 5% level. The low p-value in the $p_{\text{T}}$ of the muon is related to an underestimation for muons below 35 GeV. The distribution is shown in Figure A.15. The same effect also results in the lower p-value of the visible $p_{\text{T}}$ of the τ lepton pair, where the deficit is also found in the region with smaller $p_{\text{T}}$. The corresponding distribution is shown in Figure A.16. Due to statistical fluctuations, some p-values below 5% are expected.

By using the coefficients from the Taylor expansion of the neural network output function, one can assign an importance score to each variable. The Taylor expansion is performed separately for each category. The score for the $j$th variable in a given category is calculated by summing up all first and second-order Taylor coefficients containing the variable $j$

$$S_{\text{imp},j} = \sum_{i} \frac{f_i}{f_{\text{tot}}} \quad j \in i, \tag{7.6}$$

where $f_i$ is the first or second-order Taylor coefficient of the variables $i$, and $f_{\text{tot}}$ is the sum of all Taylor coefficients. While the total value of the Taylor coefficients cannot be used to compare the importance of input features between different neural networks, one can use the normalized coefficients. The resulting importance scores when using $\tau$-embedded samples are shown in Figure 7.6.

In all categories, $m_{\mathrm{vis}}$ and the di-$\tau$ mass are the most important input features. These two features are more important for ggH, qqH, Z $\rightarrow$ $\tau\tau$, and Z $\rightarrow$ ll than for the other backgrounds, as expected due to their peaking structure. The $p_{\mathrm{T}}$ of the muon and the $\tau_{\mathrm{h}}$ are of equal importance across all categories. By exploiting the jet structure in the event, the neural network can identify qqH production, where more jets are expected. The same holds for $\mathrm{t\bar{t}}$ and diboson production, where at least two jets are expected for most events. As a result, the jet-related input features are of more importance for the qqH, the $\mathrm{t\bar{t}}$ and the misc category.

### 7.1.3. Corrections and Uncertainty Model

For the $\tau$-embedded samples, the corrections described in the previous Chapter and the uncertainty model described in Section 6.5 are applied. The following summarises the corrections and related uncertainties for signal, jet $\rightarrow$ $\tau_{\mathrm{h}}$, and simulated background processes. Uncertainties are modelled using nuisance parameters that are added to the likelihood used for the signal extraction.

- Electron and muon ID, isolation, and trigger corrections are derived as explained in Section 6.2. An uncertainty of 2% is applied for both the ID and the trigger corrections, 50% correlated with $\tau$-embedded samples.

- Corrections for the $\tau_{\mathrm{h}}$ and electron energy are derived and applied. For the uncertainty, a 50% correlation with $\tau$-embedded samples is applied for the $\tau_{\mathrm{h}}$ energy correction, while no correlation is applied for the electron energy.

- The simulation has to be scaled to the data luminosity. The combination of the luminosity measurements from Run II [40–42] result in a 1.6% uncertainty. This uncertainty does not have to be applied to the $\tau$-embedded samples since they are already corrected using the efficiencies described in Section 6.1.

- The energy resolution and scale of jets are corrected in simulation. The jet energy resolution amounts typically to 15% at 30 GeV and decreases to about 5% at 1 TeV [138]. The scale uncertainties are modelled using a split into 11 different sources. One additional nuisance parameter is added targeting the jet energy resolution. For $\tau$-embedding, these corrections and the corresponding uncertainties are not applied.

- During the simulation, the number of PU interactions in a simulated event is drawn from a Poisson distribution that models the PU profile of the data as shown in Figure 3.11. Since this profile is not known when the simulation is performed, a reweighting based on the measured PU profile is performed to obtain the same shape for simulation and data. For $\tau$-embedding, this correction is not needed.

- The $p_{\mathrm{T}}$ of top-quarks in $\mathrm{t\bar{t}}$ simulation is shifted towards larger values compared to the data. A reweighting depending on the $p_{\mathrm{T}}$ of the two simulated top quarks is

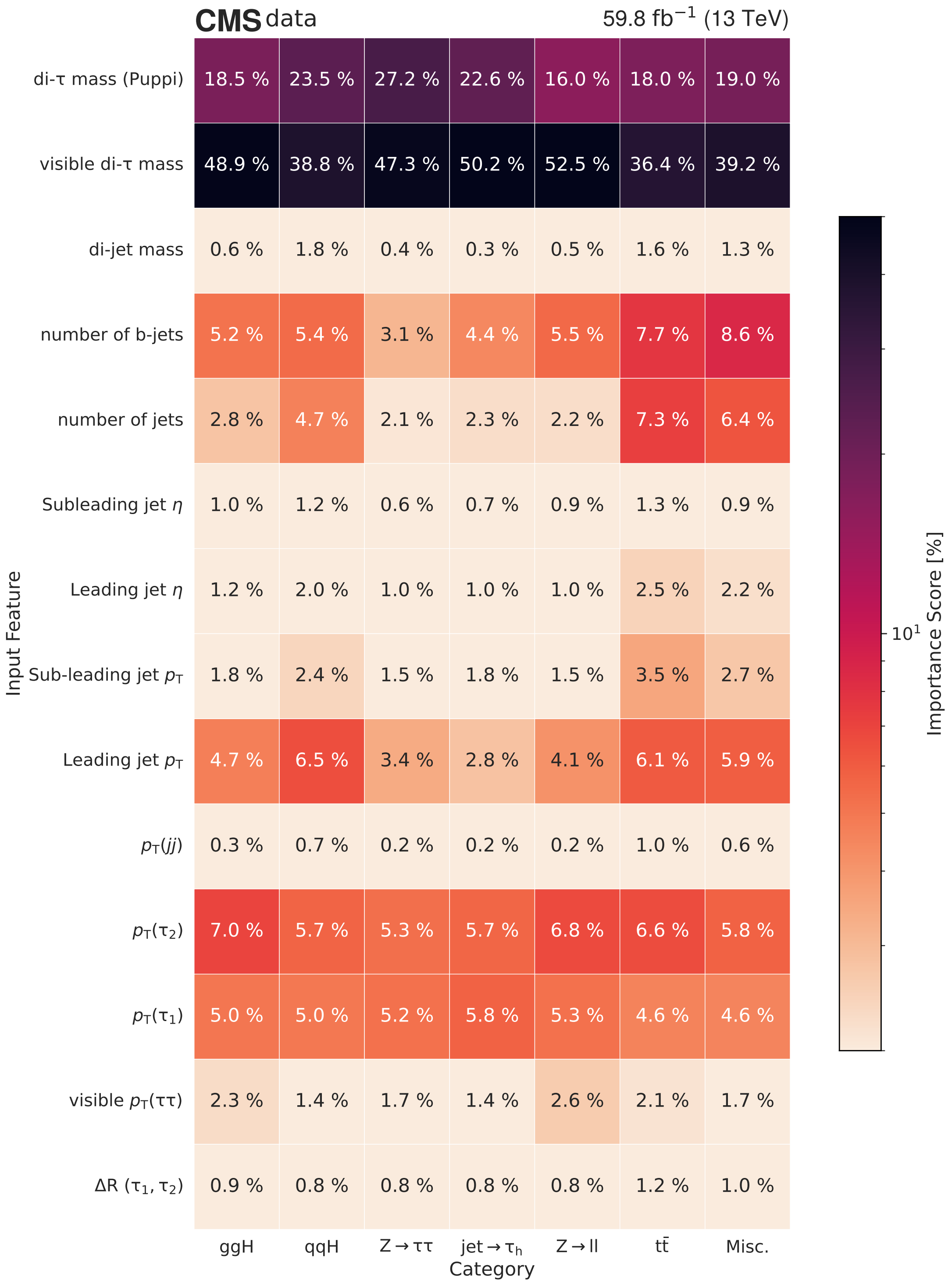


**Figure 7.6.:** Importance scores for the neural network training in the $\mu\tau_h$ final state when using τ-embedding.

applied to account for this effect. The uncertainty on this reweighting is modelled by applying it twice and not at all.

- Corrections for the identification of b-jets are applied for simulated samples.

- If the energy of a particle or jet is changed due to a correction, the MET has to be recalculated. In the case of a nonresonant simulation like $\mathrm{t\bar{t}}$ and diboson, an additional uncertainty in the unclustered energy, energy deposits not used for lepton or jet reconstruction, is applied.

- For resonant processes like $\mathrm{Z} \rightarrow \tau\tau$, $\mathrm{Z} \rightarrow \mathrm{ll}$, and Higgs boson production, a correction of the MET based on the recoil against the resonance particle is applied. This correction is calculated using $\mathrm{Z} \rightarrow \mu\mu$ events, where muons are reconstructed with high precision, and the true MET is zero since no neutrinos are present. The $p_{\mathrm{T}}^{\mathrm{miss}}$ is split into a parallel and a perpendicular component:

$$p_{\mathrm{T}}^{\mathrm{miss}} = p_{\parallel}^{\mathrm{miss}} e_{\parallel}^{\mathrm{Z}} + p_{\perp}^{\mathrm{miss}} e_{\perp}^{\mathrm{Z}}, \tag{7.7}$$

  where $e_{\parallel}^{\mathrm{Z}}$ and $e_{\perp}^{\mathrm{Z}}$ are the unit vectors parallel and perpendicular to the di-$\mu$ system which represents the Z boson. The corrections are then derived as a function of the $p_{\mathrm{T}}$ of the di-$\mu$ system and the number of jets by performing a quantile mapping of the $p_{\parallel}^{\mathrm{miss}}$ and $p_{\perp}^{\mathrm{miss}}$ distributions in data and simulation. A detailed description of this procedure can be found in [13]. Two nuisances, targeting the scale and the resolution from this correction, are added to the uncertainty model.

- For $\mathrm{Z} \rightarrow \mathrm{ll}$ events, an additional correction is applied to account for electrons (muons) misidentified as $\tau_{\mathrm{h}}$. In addition, a correction to improve the modelling of $\mathrm{Z} \rightarrow \mathrm{ll}$ is applied. This correction is derived based on the mass and $p_{\mathrm{T}}$ of the di-$\mu$ system in the data.

- For the $\mathrm{F_F}$ method, a set of nuisance parameters is added to the uncertainty model. These nuisances are split into normalization uncertainties and uncertainties that can alter the shape of the distribution. They are connected to the separate DRs and the additional closure corrections described in Section 4.2.4. Since the $\mathrm{F_F}$ extrapolation depends on the energies of the $\tau$ lepton decay products, the corrections on the $\tau_{\mathrm{h}}$ energy scale are propagated to the $\mathrm{F_F}$. For electrons and muons, this propagating is omitted due to the small impact on the shape of the jet $\rightarrow \tau_{\mathrm{h}}$ distribution. In total, 35 nuisance parameters are added to the uncertainty model to account for the uncertainties of the $\mathrm{F_F}$ method. A detailed description of the jet $\rightarrow \tau_{\mathrm{h}}$ method uncertainty model can be found in [105].

- Several signal estimation uncertainties are also included in the model. The signal uncertainties are split into PDF and $\alpha_s$, renormalization and factorization, and the $H \rightarrow \tau\tau$ branching fraction components. The uncertainties are chosen as described in [9].

- For the different simulated background processes, uncertainties on the cross sections are applied. The uncertainty values are listed in Table 7.2.

**Table 7.2.:** Uncertainties in the cross sections for the simulated background processes.

| Process | Uncertainty [%] |
|---|---|
| $t\bar{t}$ | 6.0 |
| diboson | 5.0 |
| W+jets | 4.0 |
| Z → ττ | 2.0 |
| Z → ll | 2.0 |

- Bin-by-bin uncertainties are included using the approach described in Section 6.3, where one nuisance parameter per bin is added to the model. The scaling of this nuisance parameter is determined by the contribution of the background estimates in the bin and calculated via

  $$s = n_{\mathrm{tot}} + x \cdot \epsilon_{\mathrm{tot}}, \tag{7.8}$$

  where $n_{\mathrm{tot}}$ is the sum of all background estimates

  $$n_{\mathrm{tot}} = \sum_{i=1}^{n} n_i, \tag{7.9}$$

  $\epsilon_{\mathrm{tot}}$ is the quadratic sum of all background uncertainties

  $$\epsilon_{\mathrm{tot}} = \sqrt{\sum_{i=1}^{n} \epsilon_i^2}, \tag{7.10}$$

  and $x$ is the parameter varied by the fit. The uncertainties $\epsilon_i$ are calculated using the weighted events via

  $$\epsilon_i = \sqrt{\sum_{j=1}^{m_i} w_{i,j}^2}, \tag{7.11}$$

  where $m_i$ is the number of events from process $i$ in the bin, and $w_{i,j}$ is the weight associated to event $j$.

### 7.1.4. Signal Extraction

Multiple histograms based on the neural network scores $y_l$ are constructed for signal extraction. Each event is assigned to the category that obtains the largest score from the neural network. Within the category, the value of $y_l$ is used to construct the histogram. Since a large score indicates a high probability for an event to belong to the given category, very pure control regions for the background classes can be obtained. These regions help to reduce the impact of background-related uncertainties in the signal regions since the control region distributions can be used to constrain these uncertainties. In Figure 7.7 the resulting histograms for the Z → ττ (left) and the jet → $\tau_{\mathrm{h}}$ (right) categories from the training in the $\mu\tau_{\mathrm{h}}$ final state using τ-embedded samples are shown. Both categories

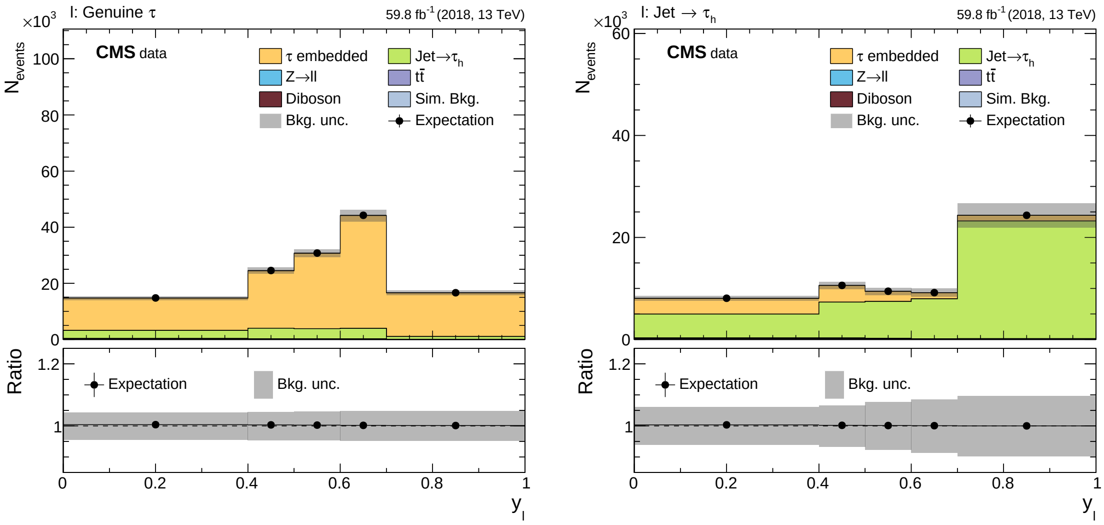


**Figure 7.7.:** Histograms of the neural network score for two background categories when training with τ-embedded samples. The score is used to sort the events within each category. On the left, the Z → ττ category is shown, and on the right, the jet → $\tau_h$ category. Both categories serve as a pure control region for the corresponding process. The signal plus background model expectation replaces the data.

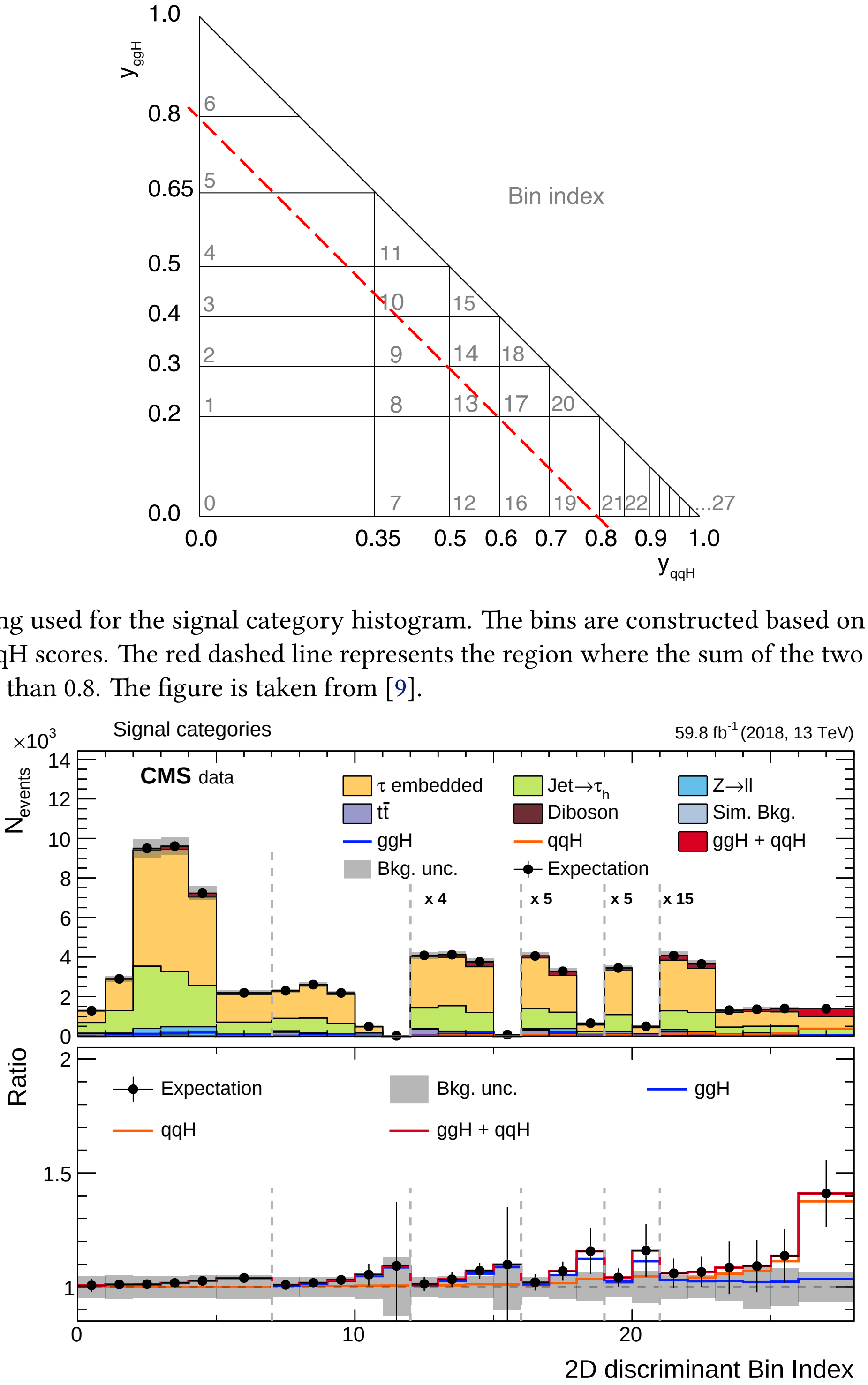
contain the corresponding process, with high purity when going to larger scores. The remaining background categories are shown in the Appendix in Figure A.17.

Due to the confusion of ggH and qqH, a sizable number of ggH events do not obtain a large ggH score. While the neural network cannot distinguish those events from qqH events, it can distinguish them from the background processes. As a result, the sum of the ggH and qqH scores could be used to recover this information. Since the measurement is simultaneously targeted towards the measuring ggH and qqH production, the two output scores are combined into a single histogram using a two-dimensional binning illustrated in Figure 7.8a. For the binning, the ggH and qqH scores are split into

$$y_{\mathrm{ggH}} \in [0.0, 0.2, 0.3, 0.4, 0.5, 0.65, 0.8, 1.0]$$
$$y_{\mathrm{qqH}} \in [0.0, 0.35, 0.5, 0.6, 0.7, 0.8, 0.85, 0.9, 0.92, 0.94, 0.96, 0.98, 1.0]$$

and then combined into 28 different bins. This way, a large signal over background ratio for qqH is created in the largest bins of the category, while several sensitive ggH bins are created, where the sum of the two scores is larger than 0.8, represented by the red dashed line. The resulting histogram is shown in Figure 7.8b.

The signal strength is then obtained using an extended binned likelihood as described in Equation (6.9). Two measurements are performed, the measurement of the inclusive H → ττ production cross section, as well as the measurement of qqH and ggH production.



**(a)** Binning used for the signal category histogram. The bins are constructed based on the ggH and qqH scores. The red dashed line represents the region where the sum of the two scores is larger than 0.8. The figure is taken from [9].

**(b)** The distribution of the signal category in the $\mu\tau_h$ channel. The data was replaced by the background plus signal expectation.

**Figure 7.8.:** Visualization of the signal category histogram. The binning is visualized at the top; the resulting signal category histogram is shown at the bottom.

## 7.2. Results

In the following, the results from the analysis in the $\mu\tau_h$ channel are discussed and compared within for different analysis setups:

1. $\tau$-embedding analysis: This is the analysis using the latest CMS data set and the latest iteration of the $\tau$-embedding method as described in the previous chapters.
2. Simulation analysis: In this analysis, the latest CMS data set is also used, but instead of $\tau$-embedding, regular simulation is used to estimate the genuine di-$\tau$ background.
3. Legacy analysis: These are the analysis results presented in [9]. This analysis uses a previous version of the Run II data and the $\tau$-embedded samples. For the training of the legacy analysis, data from 2016 to 2018 was used, and the selection of input features used for the training was slightly different. Instead of using the $\eta$ of the leading and subleading jets, only the $\Delta\eta$ between the two was used. In addition, two input features representing an estimate of the Higgs VBF hypothesis were included [139].

In the following comparisons, whenever $Z \to \tau\tau$ simulation is mentioned, the result is taken from the Simulation analysis, while the results labelled as $\tau$-embedding are taken from the $\tau$-embedding analysis.

The expected inclusive signal strength measurement results for the three analysis setups are shown in Table 7.3. The expected results are calculated by replacing the data with the signal plus background model. Using only the expectation allows for comparing the expected impact of different estimation methods and uncertainty models on the final result. The results for the measurement of the ggH and qqH signal strengths are shown in Table 7.4. Both measurements were performed using the same input but different POIs.

The $\tau$-embedding and the legacy analysis have a similar performance in the inclusive, and the ggH+qqH measurement, the differences between the expected constraints on the signal strengths are within 10%. The analysis using simulated $Z \to \tau\tau$ events is 40% less sensitive. The ability to constrain the ggH signal strengths is also about 47% weaker than the analyses using $\tau$-embedded samples. For the qqH signal strength, a 20% weaker constraint is obtained for the simulation analysis.

The reason for the large differences can be further investigated by combining uncertainties into different groups. The effect of one group on the signal strength can be obtained by performing a likelihood scan as shown in Figure 7.9. The uncertainties are split into four different groups:

- Bin-by-bin uncertainties (bbb) calculated via Equation (7.8),

**Table 7.3.:** The expected signal strength for the inclusive measurement in the $\mu\tau_h$ channel using the data from 2018.

| Signal | $\tau$-embedding Analysis | Simulation Analysis | Legacy Analysis |
|---|---|---|---|
| Inclusive | $1.00^{+0.26}_{-0.26}$ | $1.00^{+0.32}_{-0.33}$ | $1.00^{+0.24}_{-0.22}$ |

**Table 7.4.:** The expected signal strength for the ggH and qqH measurement in the $\mu\tau_h$ channel using the data from 2018.

| Signal | τ-embedding Analysis | Simulation Analysis | Legacy Analysis |
|---|---|---|---|
| ggH | $1.00^{+0.48}_{-0.46}$ | $1.00^{+0.78}_{-0.71}$ | $1.00^{+0.52}_{-0.49}$ |
| qqH | $1.00^{+0.38}_{-0.38}$ | $1.00^{+0.42}_{-0.42}$ | $1.00^{+0.35}_{-0.35}$ |

- Theory uncertainties (theory) representing all nuisance parameters associated with the signal model,
- Systematic uncertainties (syst) representing all remaining nuisance parameters,
- Statistical uncertainties (stat).

Starting from a scan, where all nuisance parameters are considered, the second scan is performed by freezing all bin-by-bin parameters. In every subsequent scan, all nuisance parameters of another group are frozen. Using this method, the combination of the individual uncertainties will result in the correct combined uncertainty. The freezing order can influence individual uncertainties as the correlations between nuisance parameters of different groups are not considered.

The resulting uncertainty splits for the ggH and qqH measurements are shown in Figure 7.10a and Figure 7.10b. The legacy and τ-embedding analyses have similarly sized uncertainties across all groups. For the simulation analysis, the biggest difference is the size of the bin-by-bin uncertainties. For the ggH and qqH measurements, the bin-by-bin uncertainty is more than two times larger for the simulation analysis, compared to the other two setups. Specifically, the bin-by-bin nuisance parameters in the signal category have much larger uncertainties when using Z → ττ simulation.

In Figure 7.11 the associated errors $\epsilon_i$ for τ-embedding and Z → ττ simulation are shown. These errors are used in Equations (7.8) and (7.10) to calculate the scaling of the corresponding bin-by-bin nuisance parameters. Although the two analyses use different neural network trainings, the contribution of genuine di-τ backgrounds to the signal region is very similar. In the τ-embedding analysis, 380,326 events are used to construct the genuine di-τ background in the signal region, while only 8,170 simulated events are used. The average event weight for a simulated Z → ττ event is 3.576 while the average weight for a τ-embedded event is 0.079. As a result, the associated uncertainty is much larger for the simulation, resulting in a larger bin-by-bin uncertainty.

This fact represents one of the largest benefits of using τ-embedded samples. As long as the phase space of the analysis is covered by the τ-embedded samples, the analysis will benefit from the huge number of available events. The production of a similar amount of simulated Z → ττ events would require a much larger CPU time. On average, the generation of a single Z → ττ simulated event takes about a minute[1], while, as presented in Section 5.5, the generation of a τ-embedded event takes about 40 s, depending on the final state. While the simulation remains irreplaceable in phase spaces where only a small

[1] Measured to be 62.3 s per event on a subset of 400,000 simulated Z → ττ events

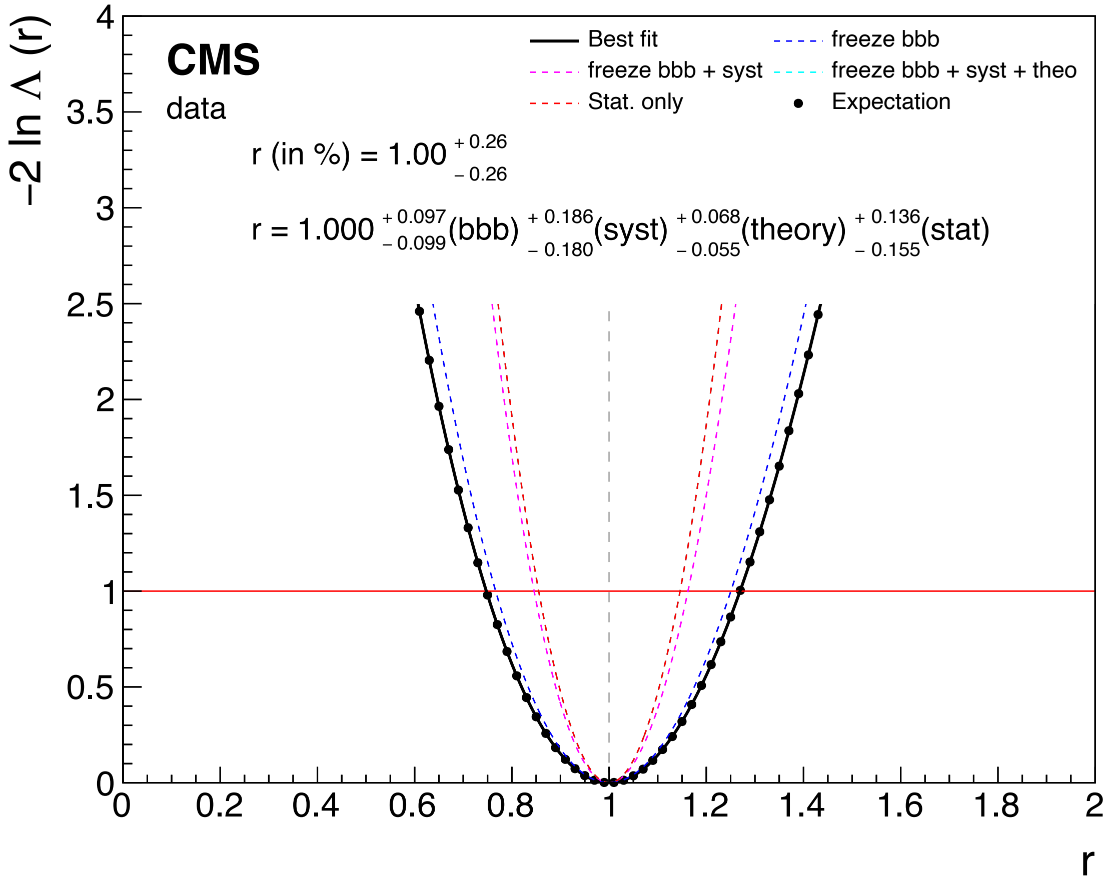
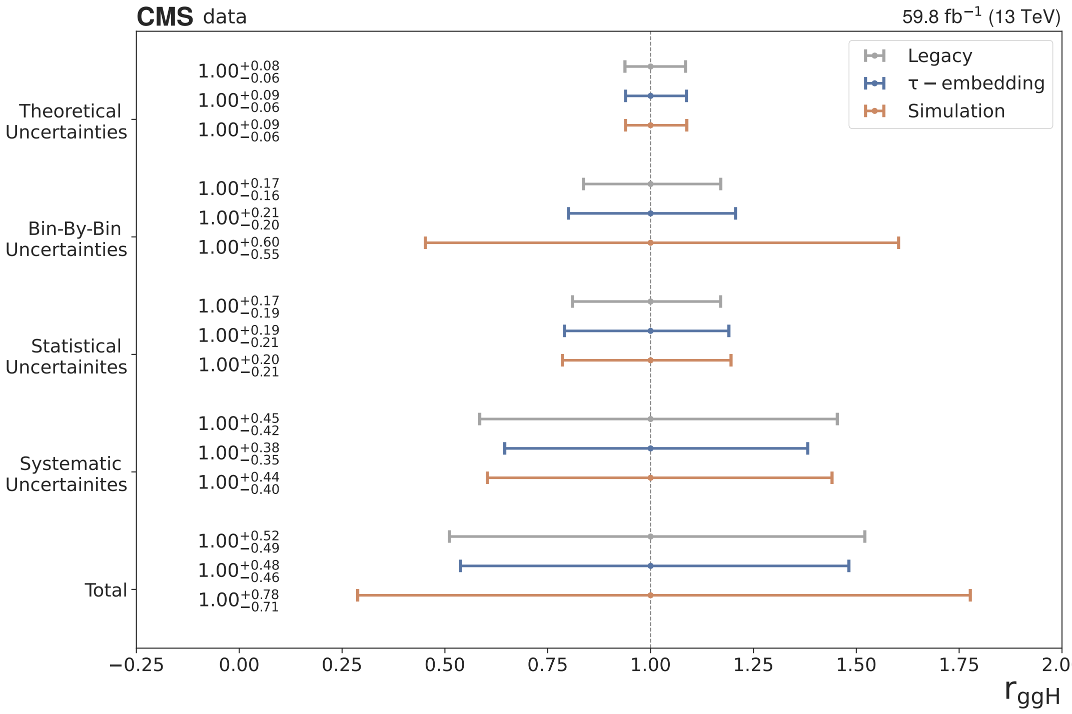
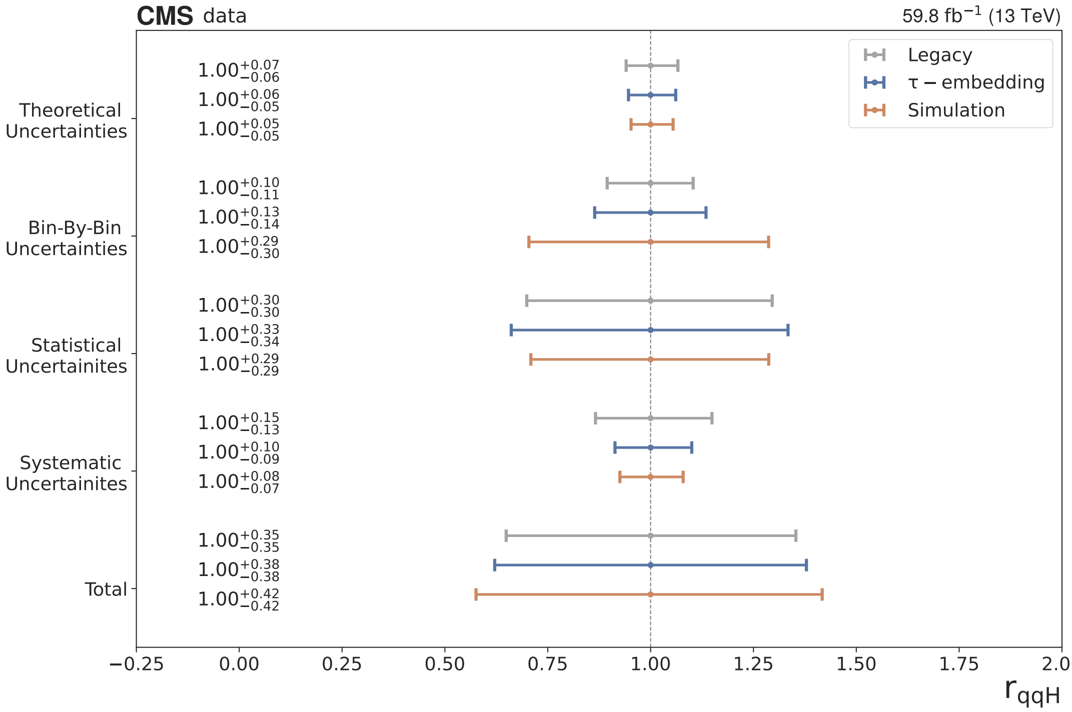


**Figure 7.9.:** The likelihood scan of the inclusive measurement using the τ-embedding analysis. The split of the uncertainties into different groups is calculated by consecutively freezing all nuisance parameters of the different groups. For the scans, the data was replaced by the signal plus background model.



**(a)** Uncertainty split for the measurement of the ggH signal strength.



**(b)** Uncertainty split for the measurement of the qqH signal strength.

**Figure 7.10.:** Uncertainty split for the measurement of the ggH and qqH signal strength, shown for the τ-embedding (blue), the simulation (orange) and the legacy (grey) analysis. The data was replaced by the signal plus background model.

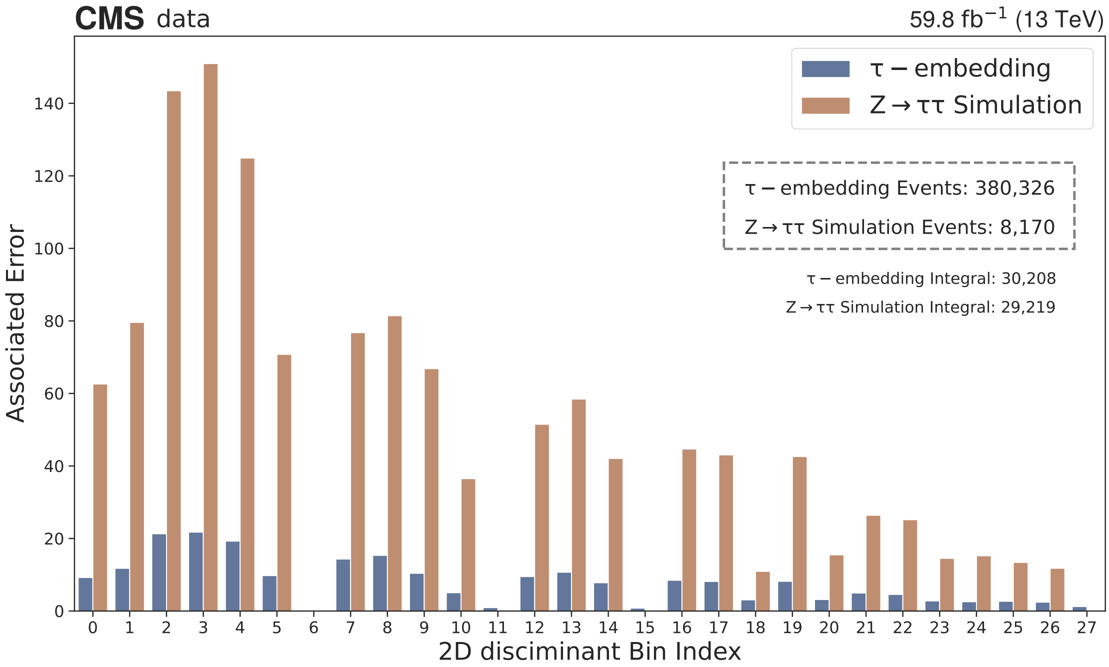


**Figure 7.11.:** Comparison of the errors $\epsilon_i$ of τ-embedding and $Z \longrightarrow \tau\tau$ simulation in the signal category used to calculate the bin-by-bin uncertainties. The bin numbers are depicted in Figure 7.8a. The associated error is zero if the bin is empty. For the τ-embedded sample, the average event weight is $0.079$ while it is $3.576$ for the $Z \longrightarrow \tau\tau$ simulation.

number of events are expected, τ-embedded samples pose an excellent estimation method in the phase space where lots of events are expected. This results in the comfortable position that two viable methods are available for estimating di-τ events, allowing for cross-checks between the two and an improved understanding of the underlying physics.

In Figure 7.12, the impact of different systematic uncertainties on the inclusive signal strength after the fit is shown. For this purpose, the systematic uncertainties mentioned in Sections 6.5 and 7.1.3 are combined into several groups:

- $\tau_h$ identification uncertainties,
- $\tau_h$ energy scale uncertainties,
- muon identification uncertainties,
- trigger uncertainties,
- process specific uncertainties. For τ-embedding, the 4% normalization uncertainty targeting the unfolding efficiency and the overall understanding of the method is listed. For $Z \longrightarrow \tau\tau$, the luminosity, the jet energy scale and resolution corrections, and the cross section uncertainty are listed.

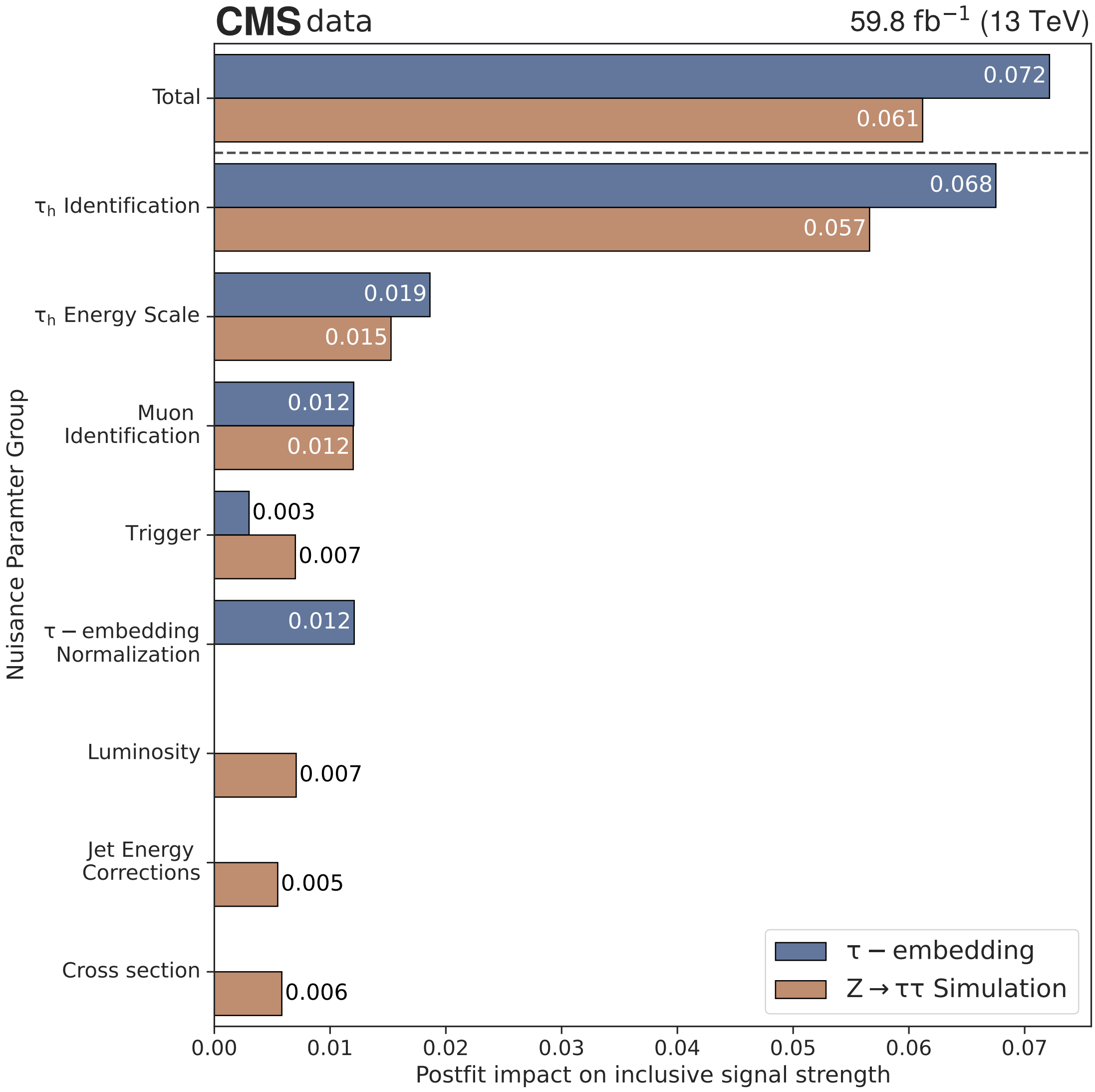


**Figure 7.12.:** Impact of the different systematic uncertainties assigned to $Z \longrightarrow \tau\tau$ and τ-embedded samples in the inclusive measurement. Only systematic nuisance parameters assigned to τ-embedded samples or $Z \longrightarrow \tau\tau$ simulation are shown. In the top row, the combined impact on the signal strength is shown; in the preceding rows, the impacts of individual groups of uncertainties are shown. For the comparison, correlations between nuisance parameters are not taken into account.

This estimated impact is calculated using the information from the covariance matrix. Only nuisance parameters assigned to τ-embedded samples or $\mathrm{Z} \rightarrow \tau\tau$ simulation are considered. In the comparison, correlations between nuisance parameters are not taken into account. In addition, many nuisance parameters are assigned to multiple processes, which makes it hard to completely isolate the effect of the $\mathrm{Z} \rightarrow \tau\tau$ simulation and τ-embedding on the signal strength. Nevertheless, this comparison gives a good indication of the primary uncertainty sources.

For the simulation analysis, the impact of uncertainties assigned to $\mathrm{Z} \rightarrow \tau\tau$ simulation is about 15% smaller than in the case of τ-embedded samples. The $\tau_\mathrm{h}$ identification uncertainties are the most impactful for both processes. The impact of the τ-embedding normalization uncertainty exclusively used for τ-embedded samples has roughly the same size as the combination of luminosity, jet correction and cross section uncertainties used for $\mathrm{Z} \rightarrow \tau\tau$ simulation. In the τ-embedding uncertainty model, fewer nuisance parameters are needed. By utilizing the $\mathrm{Z} \rightarrow \tau\tau$ category as the control region, the fit can constrain the τ-embedding normalization ( or in the case of using simulation the $\mathrm{Z} \rightarrow \tau\tau$ cross section uncertainty) by $\sim 20\%$.

While the impact of systematic uncertainties is smaller for simulation, the effect is outweighed by the increased bin-by-bin uncertainties. Both methods have a similar uncertainty model with similarly sized uncertainties, although fewer nuisance parameters are required for τ-embedded samples.

# 8 Conclusion

The τ-embedding method is a background estimation method, where events with two muons are selected in data. A simulated τ lepton decay replaces each muon. The rest of the event is left unchanged, resulting in a sample of hybrid events obtained mostly from data, where only the τ lepton decays are simulated.

In this thesis, a description based on the latest version of this method is given. The τ-embedding method has been and continues to be one of the cornerstones of several Higgs boson analyses in di-τ final states of the CMS Collaboration, performed using the Run II data set [9, 12, 15]. A reliable and accurate estimation of backgrounds with genuine τ leptons, mainly consisting of Z boson production in the subsequent decay into τ leptons ($Z \rightarrow \tau\tau$), is an essential ingredient for these analyses.

The method is validated using $\mu \rightarrow \mu$ embedded events, where simulated muons replace the selected muons. Studies show that the method can reproduce the kinematic properties of the original muon on the sub-percent level. On average, the difference in energy deposits in the vicinity of the replaced muons due to limitations in the removal of the original muons is below 200 MeV. Because of the undetectable neutrinos, residual differences of that size are far below the $p_T$ resolution of the involved τ lepton decays. The distributions of jets and PU collisions are not affected by the method.

While the production of two genuine τ leptons can also be estimated from the full simulation of all involved processes, the τ-embedding method offers some significant advantages. The production of a single τ-embedded event requires 35% less CPU time than the full simulation of a $Z \rightarrow \tau\tau$ event. The number of systematic variations required in the final analysis is reduced, jets are left unchanged, and no luminosity scaling is needed. The distributions of jets and PU collisions in τ-embedded events are correct without additional effort, resulting in good modelling of such quantities in events with two genuine τ leptons from the beginning.

While the generation of τ-embedded events is more CPU efficient than full process simulation, the significant number of events that have to be processed still turns the production of τ-embedded samples into a major enterprise. More than 8 million CPU hours were invested for the production of τ-embedded samples from the whole Run II data set, of which 3.5 million CPU hours were used for the production of the 2018 data set alone. The complete production was performed using opportunistic resources over a period of more than four months. Including all di-τ final states, the produced τ-embedded samples for 2018 contain about 195 million events. The simulated event sample of the $Z \rightarrow \tau\tau$ process produced for the 2018 data set and provided by the CMS Collaboration contains about 100 million events and took about 1.6 million CPU hours to produce.

While this might not appear like a gain at first glance the following points have to be taken into account for further assessment. Due to the generator level selection and multiple trials during the $\tau$ lepton decay simulation in the $\tau$-embedding method, the fraction of produced $\tau$-embedded events that enter the phase space of the typically targeted analyses is considerable. For example, about 2.3 million $\tau$-embedded events pass the selection criteria in the $\mu\tau_h$ final state described in Section 4.2.1. From the corresponding sample of simulated $Z \to \tau\tau$ events, only around 52,000 events pass the same selection criteria. With roughly double the CPU time invested, the $\tau$-embedding method results in about 40 times more events available for the analysis, an effect that can be observed across all di-$\tau$ final states. As demonstrated in Section 7.2, this significantly reduces uncertainties due to limited sample sizes of background estimates in the data model and can provide a more than 40% higher measurement accuracy in the final result.

In the scope of the HL-LHC, the $\tau$-embedding method will become even more important. Full event simulation will become significantly more complex in scenarios where an event pileup of 200 collisions per bunch crossing is expected. On the other hand, as long as it will be possible to select and identify muons, the $\tau$-embedding method will be applicable without any increase in complexity.

Limitations of the $\tau$-embedding method are given mainly through the technical setup. The method requires modifications of the standard event reconstruction used by the CMS Collaboration. The biggest issue is the simulation of the two $\tau$ lepton decays, which is performed using an otherwise empty detector. This setup results in reconstruction effects that require dedicated correction factors. The response of the HLT is also performed in the otherwise empty detector. As a result, no objects outside the di-$\tau$ decay can be obtained as part of the HLT simulation. Since the $\tau$-embedding method is based on selected di-$\mu$ events, the number of events in kinematic phase space regions, where few or no events of the modelled processes are expected, can become small. Examples are searches in mass ranges beyond the TeV scale or with high (b-)jet multiplicities. For such cases, where the statistical power of an inclusively produced, fully simulated $Z \to \tau\tau$ sample would also be insufficient, the coverage of $\tau$-embedded samples has to be checked. If the $\tau$-embedded event yield is insufficient, one should produce dedicated simulated event samples enriched in these phase space regions.

The principle of *particle embedding* can also be expanded to not only embed $\tau$ leptons but other particles. An effort to replace the selected muons with b quark decays is ongoing. A second effort is to apply the same principle to W boson decays, replacing the muon in $W \to \mu\nu$ events with a single $\tau$ lepton decay. Both applications are more complicated and still explorative. But they can build upon the existing and validated $\tau$-embedding procedure. The $\tau$-embedding method has also been proposed as a validation sample for $Z \to \nu\nu$ events. Such a sample can be created with the existing code setup by simply removing the selected muons without a replacement. Compared to full process simulation, such a sample suffers from reduced statistical power since the abundance of $Z \to \mu\mu$ events in data is only a third compared to $Z \to \nu\nu$.

An essential step towards the use of $\tau$-embedded samples for Run III is the integration of the technical setup into the official CMS reconstruction workflow management. Up to now, all $\tau$-embedded samples have been produced in private efforts by the ETP using opportunistic computing resources [140]. Integrating the $\tau$-embedding method into the reconstruction workflow management of CMS will allow the automatic production of τ-embedded samples and significantly improve the maintainability of the technique. This integration is anticipated by mid of 2023.

In summary, the τ-embedding method is a viable and robust method to provide an accurate and realistic model of all processes with two genuine τ lepton decays in the final state. For the CMS Collaboration, the τ-embedding method has been and foreseeably will remain the main method to estimate the background from processes with two genuine τ lepton decays in the final state. The method does not exclude the use of full simulation, e.g. in exotic regions of the kinematic phase space. Both background estimation methods have justified their use and can be used to improve our understanding of genuine τ lepton decays in the final state. The role of data-driven estimates like the τ-embedding method will become even more important in the scope of the HL-LHC, which poses new challenges in terms of cost, power, and processing time.

# List of Figures

# List of Tables

# Bibliography

[1] F. Englert and R. Brout. “Broken Symmetry and the Mass of Gauge Vector Mesons”. *Physical Review Letters* 13.9 (1964), pp. 321–323.
DOI: 10.1103/PhysRevLett.13.321.

[2] P. W. Higgs. “Broken Symmetries, Massless Particlees and Gauge Fields”. *Physics Letters* 12.2 (1964), pp. 132–133.
DOI: 10.1016/0031-9163(64)91136-9.

[3] Peter W. Higgs. “Broken Symmetries and the Masses of Gauge Bosons”. *Physical Review Letters* 13.16 (1964), pp. 508–509.
DOI: 10.1103/PhysRevLett.13.508.

[4] The ATLAS collaboration. “The ATLAS Experiment at the CERN Large Hadron Collider”. *Journal of Instrumentation* 3.8 (2008).
DOI: 10.1088/1748-0221/3/08/S08003.

[5] The CMS Collaboration. “The CMS Experiment at the CERN LHC”. *Journal of Instrumentation* 3.8 (2008).
DOI: 10.1088/1748-0221/3/08/S08004.

[6] The ATLAS collaboration. “Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC”. *Physics Letters, Section B: Nuclear, Elementary Particle and High-Energy Physics* 716.1 (2012), pp. 1–29.
DOI: 10.1016/j.physletb.2012.08.020. arXiv: 1207.7214.

[7] The CMS Collaboration. “Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC”. *Physics Letters, Section B: Nuclear, Elementary Particle and High-Energy Physics* 716.1 (2012), pp. 30–61.
DOI: 10.1016/j.physletb.2012.08.021. arXiv: 1207.7235.

[8] The CMS Collaboration. “An Embedding Technique to Determine $\tau\tau$ Backgrounds in Proton-Proton Collision Data”. *Journal of Instrumentation* 14.6 (2019).
DOI: 10.1088/1748-0221/14/06/P06032. arXiv: 1903.01216.

[9] The CMS Collaboration. “Measurements of Higgs Boson Production in the Decay Channel with a Pair of $\tau$ Leptons in Proton-Proton Collisions at $\sqrt{s} = 13$ TeV”. 2022.
DOI: 10.48550/arXiv.2204.12957.

[10] Sebastian Wozniewski. “Differential Cross Section Measurements in the $H \to \tau\tau$ Decay Channel with CMS Data of Proton-Proton Collisions at the Large Hadron Collider at CERN”. Karlsruhe Institute of Technology (KIT), 2021.
DOI: 10.5445/IR/1000128508.

[11] Stefan Wunsch. “Modern Machine Learning in the Presence of Systematic Uncertainties for Robust and Optimized Multivariate Data Analysis in High-Energy Particle Physics”. Karlsruher Institut für Technologie (KIT) / Karlsruher Institut für Technologie (KIT), 2021. 110 pp.
DOI: 10.5445/IR/1000129166.

[12] CMS Collaboration. “Searches for Additional Higgs Bosons and for Vector Leptoquarks in $\tau\tau$ Final States in Proton-Proton Collisions at $\sqrt{s} = 13$ TeV”. 2022.
DOI: 10.48550/arXiv.2208.02717. arXiv: 2208.02717.

[13] Artur Gottmann. “Global Interpretation of $\tau\tau$ Events in the Context of the Standard Model and Beyond”. Karlsruhe Institute of Technology (KIT), 2020.
DOI: 10.5445/IR/1000124886.

[14] Maximilian Burkart. “A Search for Additional Neutral Higgs Bosons in $\tau\tau$ Final States in pp Collisions at $\sqrt{s} = 13$ TeV (to Be Published)”. Karlsruhe Institute of Technology (KIT), 2022.

[15] The CMS Collaboration. “Search for a Heavy Higgs Boson Decaying into Two Lighter Higgs Bosons in the $\tau\tau$bb Final State at 13 TeV”. *Journal of High Energy Physics* 2021.11 (2021), pp. 1–54.
DOI: 10.1007/JHEP11(2021)057.

[16] Janek Bechtel. “A Novel Search for Di-Higgs Events in the $\tau\tau$+bb Final State in pp Collisions at 13 TeV at the LHC”. Karlsruhe Institute of Technology (KIT), 2021.
DOI: 10.5445/IR/1000130103.

[17] Andrew Purcell. “Go on a Particle Quest at the First CERN Webfest. Le Premier Webfest Du CERN Se Lance à La Conquête Des Particules”. BUL-NA-2012-269, 35/2012 (2012), p. 10.

[18] Sheldon L. Glashow. “Partial-Symmetries of Weak Interactions”. *Nuclear Physics* 22.4 (1961), pp. 579–588.
DOI: 10.1016/0029-5582(61)90469-2.

[19] Steven Weinberg. “A Model of Leptons”. *Physical Review Letters* 19.21 (1967), pp. 1264–1266.
DOI: 10.1103/PhysRevLett.19.1264.

[20] Abdus Salam. “Weak and Electromagnetic Interactions”. *Conf. Proc. C* 680519 (1968), p. 367.
DOI: 10.1142/9789812795915_0034.

[21] Nicola Cabibbo. “Unitary Symmetry and Leptonic Decays”. *Physical Review Letters* 10.12 (1963), pp. 531–533.
DOI: 10.1103/PhysRevLett.10.531.

[22] Makoto Kobayashi and Toshihide Maskawa. “CP-Violation in the Renormalizable Theory of Weak Interaction”. *Progress of Theoretical Physics* 49.2 (1973), pp. 652–657.
DOI: 10.1143/PTP.49.652. eprint: https://academic.oup.com/ptp/article-pdf/49/2/652/5257692/49-2-652.pdf.

[23] M. Cepeda et al. “Higgs Physics at the HL-LHC and HE-LHC”. 2019.
DOI: 10.48550/arXiv.1902.00134. arXiv: 1902.00134.

[24] Particle Data Group et al. “Review of Particle Physics”. *Progress of Theoretical and Experimental Physics* 2020.8 (2020), p. 083C01.
DOI: 10.1093/ptep/ptaa104.

[25] The CMS Collaboration. “A Measurement of the Higgs Boson Mass in the Diphoton Decay Channel”. *Physics Letters, Section B: Nuclear, Elementary Particle and High-Energy Physics* 805 (2020), p. 135425.
DOI: 10.1016/j.physletb.2020.135425. arXiv: 2002.06398.

[26] A. M. Sirunyan et al. “Measurements of Higgs Boson Production Cross Sections and Couplings in the Diphoton Decay Channel at $\sqrt{s} = 13$ TeV”. *Journal of High Energy Physics* 2021.7 (2021), p. 27.
DOI: 10.1007/JHEP07(2021)027.

[27] A. M. Sirunyan et al. “Measurements of Production Cross Sections of the Higgs Boson in the Four-Lepton Final State in Proton–Proton Collisions $\sqrt{s}$=13 TeV”. *The European Physical Journal C* 81.6 (2021), p. 488.
DOI: 10.1140/epjc/s10052-021-09200-x.

[28] The CMS Collaboration. “Measurements of the Higgs Boson Production Cross Section and Couplings in the W Boson Pair Decay Channel in Proton-Proton Collisions at $\sqrt{s} = 13$ TeV”. 2022.
DOI: 10.48550/arXiv.2206.09466. arXiv: 2206.09466.

[29] The CMS Collaboration. “Observation of Higgs Boson Decay to Bottom Quarks”. *Physical Review Letters* 121.12 (2018), p. 121801.
DOI: 10.1103/PhysRevLett.121.121801.

[30] CMS Collaboration. “Observation of $t\bar{t}$H Production”. *Physical Review Letters* 120.23 (2018), p. 231801.
DOI: 10.1103/PhysRevLett.120.231801.

[31] The CMS Collaboration. “Observation of the Higgs Boson Decay to a Pair of $\tau$ Leptons with the CMS Detector”. *Physics Letters, Section B: Nuclear, Elementary Particle and High-Energy Physics* 779 (2018), pp. 283–316.
DOI: 10.1016/j.physletb.2018.02.004. arXiv: 1708.00373.

[32] The ATLAS collaboration and The CMS Collaboration. “Measurements of the Higgs Boson Production and Decay Rates and Constraints on Its Couplings from a Combined ATLAS and CMS Analysis of the LHC Pp Collision Data at $\sqrt{s} = 7$ and 8 TeV”. *Journal of High Energy Physics* 2016.8 (2016), p. 45.
DOI: 10.1007/JHEP08(2016)045. arXiv: 1606.02266.

[33] The CMS Collaboration. “Evidence for Higgs Boson Decay to a Pair of Muons”. *Journal of High Energy Physics* 2021.1 (2021), pp. 1–68.
DOI: 10.1007/JHEP01(2021)148.

[34] The CMS Collaboration. “A Portrait of the Higgs Boson by the CMS Experiment Ten Years after the Discovery”. *Nature* 607.7917 (7917 2022), pp. 60–68. DOI: 10.1038/s41586-022-04892-x.

[35] D. de Florian et al. “Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector”. 2016. DOI: 10.23731/CYRM-2017-002. arXiv: 1610.07922.

[36] “Latex:Feynman [CMS Wiki Pages]”. URL: https://wiki.physik.uzh.ch/cms/latex:feynman#tau_decay (visited on 03/25/2022).

[37] Esma Mobs. “The CERN Accelerator Complex - 2019”. *The CERN accelerator complex - August 2018* (2019).

[38] The ALICE Collaboration. “The Alice Experiment at the CERN LHC”. *Journal of Instrumentation* 3.8 (2008). DOI: 10.1088/1748-0221/3/08/S08002.

[39] The LHCb Collaboration. “The LHCb Detector at the LHC”. *Journal of Instrumentation* 3.08 (2008), S08005–S08005. DOI: 10.1088/1748-0221/3/08/S08005.

[40] The CMS Collaboration. “CMS Luminosity Measurements for the 2016 Data Taking Period”. CERN Document Server. 2017. URL: https://cds.cern.ch/record/2257069 (visited on 03/11/2022).

[41] The CMS Collaboration. “CMS luminosity measurement for the 2017 data-taking period at $\sqrt{s} = 13$ TeV”. CERN Document Server. 2018. URL: https://cds.cern.ch/record/2621960 (visited on 10/02/2022).

[42] The CMS Collaboration. “CMS Luminosity Measurement for the 2018 Data-Taking Period at $\sqrt{s} = 13$ TeV”. CERN Document Server. 2019. URL: https://cds.cern.ch/record/2676164 (visited on 03/11/2022).

[43] The CMS Collaboration. “LumiPublicResults < CMSPublic < TWiki”. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/LumiPublicResults (visited on 03/11/2022).

[44] “LHC Long Term Schedule”. URL: https://lhc-commissioning.web.cern.ch/schedule/LHC-long-term.htm (visited on 03/11/2022).

[45] Tai Sakuma and Thomas McCauley. “Detector and Event Visualization with Sketchup at the CMS Experiment”. *Journal of Physics: Conference Series* 513 (TRACK 2 2014), p. 22032. DOI: 10.1088/1742-6596/513/2/022032.

[46] Izaak Neutelings. “CMS Coordinate System – TikZ.Net”. URL: https://tikz.net/axis3d_cms/ (visited on 02/03/2022).

[47] Rino Castaldi Patrice Siegrist Jean-Eudes Augustin, Michel Della Negra Ernst Radermacher CERN CERN MichelDellaNegra, and cernch ErnstRadermacher. “The CMS Tracker System Project : Technical Design Report” (1997).

[48] A Tapper and Darin Acosta. “CMS Technical Design Report for the Level-1 Trigger Upgrade”. CERN-LHCC-2013-011, CMS-TDR-12. 2013.

[49] The CMS Collaboration. “The CMS Phase-1 Pixel Detector Upgrade”. *Journal of Instrumentation* 16.2 (2021).
DOI: 10.1088/1748-0221/16/02/P02027. arXiv: 2012.14304.

[50] The CMS Collaboration. “The Electromagnetic Calorimeter Technical Design Report” (1997).

[51] Particle Data Group. “Atomic and Nuclear Properties of Lead Tungstate (PbWO4)”. URL: https://pdg.lbl.gov/2020/AtomicNuclearProperties/HTML/lead_tungstate.html (visited on 03/11/2022).

[52] A. Benaglia. “The CMS ECAL Performance with Examples”. *Journal of Instrumentation* 9.2 (2014), p. C02008.
DOI: 10.1088/1748-0221/9/02/C02008.

[53] The CMS Collaboration. “The Hadron Calorimeter Project - Technical Design Report” (June 1997 1997).

[54] S. Abdullin et al. “Design, Performance, and Calibration of CMS Hadron-Barrel Calorimeter Wedges”. *Eur. Phys. J. C* 55.1 (2008), pp. 159–171.
DOI: 10.1140/epjc/s10052-008-0573-y.

[55] The CMS Collaboration. “Performance of the CMS Level-1 Trigger in Proton-Proton Collisions at $\sqrt{s} = 13$ TeV”. *Journal of Instrumentation* 15.10 (2020).
DOI: 10.1088/1748-0221/15/10/P10017. arXiv: 2006.10165.

[56] The CMS Collaboration. “Performance of the CMS Muon Detector and Muon Reconstruction with Proton-Proton Collisions at $\sqrt{s} = 13$ TeV”. *Journal of Instrumentation* 13.6 (2018).
DOI: 10.1088/1748-0221/13/06/P06015. arXiv: 1804.04528.

[57] The CMS Collaboration. “The Muon Project Technical Design Report” (December 1997), p. 475.

[58] The CMS Collaboration. “Description and Performance of Track and Primary-Vertex Reconstruction with the CMS Tracker”. *Journal of Instrumentation* 9.10 (2014).
DOI: 10.1088/1748-0221/9/10/P10009. arXiv: 1405.6569.

[59] Walaa Elmetenawee. “CMS Track Reconstruction Performance during Run 2 and Developments for Run 3”. *Proceedings of Science* 390 (2021).
DOI: 10.22323/1.390.0733. arXiv: 2012.07035.

[60] R. Frühwirth. “Application of Kalman Filtering to Track and Vertex Fitting”. *Nuclear Inst. and Methods in Physics Research, A* 262.2-3 (1987), pp. 444–450.
DOI: 10.1016/0168-9002(87)90887-4.

[61] CMS Collaboration. “Technical Proposal for the Phase-II Upgrade of the Compact Muon Solenoid”. CMS technical proposal CERN-LHCC-2015-010, CMS-TDR-15-02. 2015.

[62] The CMS Collaboration. "Performance of CMS Muon Reconstruction in pp Collision Events at $\sqrt{s} = 7$TeV". *Journal of Instrumentation* 7.10 (2012).
DOI: 10.1088/1748-0221/7/10/P10002. arXiv: 1206.4071.

[63] The CMS Collaboration. "Particle-Flow Reconstruction and Global Event Description with the CMS Detector". *Journal of Instrumentation* 12.10 (2017).
DOI: 10.1088/1748-0221/12/10/P10003. arXiv: 1706.04965.

[64] The CMS Collaboration. "Electron and Photon Reconstruction and Identification with the CMS Experiment at the CERN LHC". *Journal of Instrumentation* 16.5 (2021), P05014.
DOI: 10.1088/1748-0221/16/05/P05014. arXiv: 2012.06888.

[65] W. Adam et al. "Reconstruction of Electrons with the Gaussian-sum Filter in the CMS Tracker at LHC". *Journal of Physics G: Nuclear and Particle Physics* 31.9 (2005), N9–N20.
DOI: 10.1088/0954-3899/31/9/N01. arXiv: physics/0306087.

[66] The CMS Collaboration. "Strategies and Performance of the CMS Silicon Tracker Alignment during LHC Run 2". *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 1037 (2022), p. 166795.
DOI: 10.1016/j.nima.2022.166795.

[67] Gavin P. Salam. "Towards Jetography". Version 2. *The European Physical Journal C* 67.3-4 (2010), pp. 637–686.
DOI: 10.1140/epjc/s10052-010-1314-6. arXiv: 0906.1833.

[68] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. "The anti-kt Jet Clustering Algorithm". *JHEP* 04 (2008), p. 63.
DOI: 10.1088/1126-6708/2008/04/063. arXiv: 0802.1189.

[69] The CMS Collaboration. "Pileup Mitigation at CMS in 13 TeV Data". *Journal of Instrumentation* 15.09 (2020), P09018–P09018.
DOI: 10.1088/1748-0221/15/09/P09018.

[70] Daniele Bertolini et al. "Pileup per Particle Identification". *Journal of High Energy Physics* 2014.10 (2014), p. 059.
DOI: 10.1007/JHEP10(2014)059. arXiv: 1407.6013.

[71] The CMS Collaboration. "Identification of B-Quark Jets with the CMS Experiment". *JINST* 8.CMS-BTV-12-001, CERN-PH-EP-2012-262 (2013), P04013.
DOI: 10.1088/1748-0221/8/04/P04013. arXiv: 1211.4462.

[72] The CMS Collaboration. "Performance of the DeepJet b Tagging Algorithm Using 41.9/Fb of Data from Proton-Proton Collisions at 13 TeV with Phase 1 CMS Detector". CERN, 2018.

[73] E. Bols et al. "Jet Flavour Classification Using DeepJet". *Journal of Instrumentation* 15.12 (2020).
DOI: 10.1088/1748-0221/15/12/P12012. arXiv: 2008.10519.

[74] The CMS Collaboration. “Performance of Electron Reconstruction and Selection with the CMS Detector in Proton-Proton Collisions at $\sqrt{s}$ = 8 TeV”. *Journal of Instrumentation* 10.6 (2015), P06005.
DOI: 10.1088/1748-0221/10/06/P06005. arXiv: 1502.02701.

[75] The CMS Collaboration. “Missing Transverse Energy Performance of the CMS Detector”. 2011.
DOI: 10.1088/1748-0221/6/09/P09001. arXiv: 1106.5048.

[76] The CMS Collaboration. “Performance of τ-Lepton Reconstruction and Identification in CMS”. *Journal of Instrumentation* 7.1 (2012).
DOI: 10.1088/1748-0221/7/01/P01001. arXiv: 1109.6034.

[77] The CMS Collaboration. “Reconstruction and Identification of τ Lepton Decays to Hadrons and Nτ at CMS”. *Journal of Instrumentation* 11.01 (2016), P01019–P01019.
DOI: 10.1088/1748-0221/11/01/P01019.

[78] The CMS Collaboration. “Performance of Reconstruction and Identification of τ Leptons Decaying to Hadrons and vτ in pp Collisions at $\sqrt{s}$ = 13 TeV”. *Journal of Instrumentation* 13.10 (2018), P10005.
DOI: 10.1088/1748-0221/13/10/P10005. arXiv: 1809.02816.

[79] The CMS Collaboration. “Identification of Hadronic Tau Lepton Decays Using a Deep Neural Network” (2022).
DOI: 10.1088/1748-0221/17/07/P07023.

[80] The CMS Collaboration. “Commissioning of the CMS High-Level Trigger with Cosmic Rays”. *Journal of Instrumentation* 5.03 (2010), T03005–T03005.
DOI: 10.1088/1748-0221/5/03/t03005.

[81] The CMS Collaboration. “The CMS Trigger System”. *Journal of Instrumentation* 12.1 (2017).
DOI: 10.1088/1748-0221/12/01/P01020. arXiv: 1609.02366.

[82] Laurent Thomas. “CMS High Level Trigger Performance at 13 TeV”. *PoS* ICHEP2018 (2019), p. 226.
DOI: 10.22323/1.340.0226.

[83] The CMS 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”. *Journal of High Energy Physics* 2022 (2022).
DOI: 10.1007/JHEP06(2022)012.

[84] “LHCHWGFiducialAndSTXS < LHCPhysics < TWiki”. URL: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHWGFiducialAndSTXS (visited on 10/18/2022).

[85] Pierre Fayet. “Supergauge Invariant Extension of the Higgs Mechanism and a Model for the Electron and Its Neutrino”. *Nuclear Physics B* 90 (1975), pp. 104–124.
DOI: 10.1016/0550-3213(75)90636-7.

[86] G. Aad et al. “Search for Neutral Higgs Bosons of the Minimal Supersymmetric Standard Model in pp Collisions at $\sqrt{s}$ = 8TeV with the ATLAS Detector”. *Journal of High Energy Physics* 2014.11 (2014), p. 56.
DOI: 10.1007/JHEP11(2014)056.

[87] Ulrich Ellwanger, Cyril Hugonie, and Ana M. Teixeira. “The Next-to-Minimal Supersymmetric Standard Model”. *Physics Reports* 496.1 (2010), pp. 1–77.
DOI: 10.1016/j.physrep.2010.07.001.

[88] J. Alwall et al. “The Automated Computation of Tree-Level and next-to-Leading Order Differential Cross Sections, and Their Matching to Parton Shower Simulations”. *Journal of High Energy Physics* 2014.7 (2014).
DOI: 10.1007/JHEP07(2014)079. arXiv: 1405.0301.

[89] Johan Alwall et al. “MadGraph 5: Going Beyond”. *Journal of High Energy Physics* 2011.6 (2011), p. 128.
DOI: 10.1007/JHEP06(2011)128. arXiv: 1106.0522.

[90] R. Frederix et al. “The Automation of Next-to-Leading Order Electroweak Calculations”. *Journal of High Energy Physics* 2018.7 (2018), p. 185.
DOI: 10.1007/JHEP07(2018)185. arXiv: 1804.10017.

[91] Paolo Nason. “A New Method for Combining NLO QCD with Shower Monte Carlo Algorithms”. *Journal of High Energy Physics* 8.11 (2004), pp. 1097–1124.
DOI: 10.1088/1126-6708/2004/11/040. arXiv: hep-ph/0409146.

[92] Stefano Frixione, Paolo Nason, and Carlo Oleari. “Matching NLO QCD Computations with Parton Shower Simulations: The POWHEG Method”. *Journal of High Energy Physics* 2007.11 (2007).
DOI: 10.1088/1126-6708/2007/11/070. arXiv: 0709.2092.

[93] Simone Alioli et al. “A General Framework for Implementing NLO Calculations in Shower Monte Carlo Programs: The POWHEG BOX”. *Journal of High Energy Physics* 2010.6 (2010), p. 43.
DOI: 10.1007/JHEP06(2010)043. arXiv: 1002.2581.

[94] Simone Alioli et al. “Jet Pair Production in POWHEG”. *Journal of High Energy Physics* 2011.4 (2011), p. 81.
DOI: 10.1007/JHEP04(2011)081. arXiv: 1012.3380.

[95] Simone Alioli et al. “NLO Higgs Boson Production via Gluon Fusion Matched with Shower in POWHEG”. *Journal of High Energy Physics* 2009.4 (2009), p. 2.
DOI: 10.1088/1126-6708/2009/04/002. arXiv: 0812.0578.

[96] E. Bagnaschi et al. “Higgs Production via Gluon Fusion in the POWHEG Approach in the SM and in the MSSM”. *Journal of High Energy Physics* 2012.2 (2012), p. 88.
DOI: 10.1007/JHEP02(2012)088. arXiv: 1111.2854.

[97] Paolo Nason and Carlo Oleari. “NLO Higgs Boson Production via Vector-Boson Fusion Matched with Shower in POWHEG”. *Journal of High Energy Physics* 2010.2 (2010), p. 37.
DOI: 10.1007/JHEP02(2010)037. arXiv: 0911.5299.

[98] Keith Hamilton et al. “NNLOPS Simulation of Higgs Boson Production”. *Journal of High Energy Physics* 2013.10 (2013), p. 222.
DOI: 10.1007/JHEP10(2013)222. arXiv: 1309.0017.

[99] Keith Hamilton, Paolo Nason, and Giulia Zanderighi. “Finite Quark-Mass Effects in the NNLOPS POWHEG+MiNLO Higgs Generator”. *Journal of High Energy Physics* 2015.5 (2015), p. 140.
DOI: 10.1007/JHEP05(2015)140. arXiv: 1501.04637.

[100] Richard D. Ball et al. “Parton Distributions from High-Precision Collider Data: NNPDF Collaboration”. *European Physical Journal C* 77.10 (2017), p. 663.
DOI: 10.1140/epjc/s10052-017-5199-5. arXiv: 1706.00428.

[101] Torbjörn Sjöstrand et al. “An Introduction to PYTHIA 8.2”. *Computer Physics Communications* 191.1 (2015), pp. 159–177.
DOI: 10.1016/j.cpc.2015.01.024. arXiv: 1410.3012.

[102] The CMS Collaboration. “Extraction and Validation of a New Set of CMS Pythia8 Tunes from Underlying-Event Measurements”. *European Physical Journal C* 80.1 (2020), p. 4.
DOI: 10.1140/epjc/s10052-019-7499-4. arXiv: 1903.12179.

[103] The CMS Collaboration. “Search for Additional Neutral MSSM Higgs Bosons in the $\tau\tau$ Final State in Proton-Proton Collisions at $\sqrt{s} = 13$ TeV”. *Journal of High Energy Physics* 2018.9 (2018).
DOI: 10.1007/JHEP09(2018)007. arXiv: 1803.06553.

[104] The CMS Collaboration. “Measurement of the $Z / \Gamma * \longrightarrow \tau\tau$ Cross Section in pp Collisions at $\sqrt{s} = 13$TeV and Validation of $\tau$ Lepton Analysis Techniques”. *European Physical Journal C* 78.9 (2018), p. 708.
DOI: 10.1140/epjc/s10052-018-6146-9. arXiv: 1801.03535.

[105] Janik Walter Andrejkovic. “Data-Driven Background Modeling for Precision Studies of the Higgs Boson and Searches for New Physics with the CMS Experiment”. 2022.

[106] Armin Burgmeier. “Data-Driven Estimation of Z0 Background Contributions to the Higgs Search in the H-$\tau\tau$ Channel with the CMS Experiment at the LHC”. Karlsruhe Institute of Technology (KIT), 2011.

[107] The CMS Collaboration. “Evidence for the 125 GeV Higgs Boson Decaying to a Pair of $\tau$ Leptons”. *Journal of High Energy Physics* 2014.5 (2014), p. 104.
DOI: 10.1007/JHEP05(2014)104. arXiv: 1401.5041.

[108] V. Khachatryan et al. “Search for Neutral MSSM Higgs Bosons Decaying to a Pair of Tau Leptons in pp Collisions”. *Journal of High Energy Physics* 2014.10 (2014), p. 160.
DOI: 10.1007/JHEP10(2014)160.

[109] The ATLAS collaboration. “Modelling $Z \longrightarrow \tau\tau$ Processes in ATLAS with $\tau$-Embedded $Z \longrightarrow \mu\mu$ Data”. *Journal of Instrumentation* 10.09 (2015), P09018–P09018.
DOI: 10.1088/1748-0221/10/09/P09018.

[110] Georges Aad et al. “Measurements of Higgs Boson Production Cross-Sections in the H⟶$\tau^+\tau^-$ Decay Channel in pp Collisions at $\sqrt{s}$=13 TeV with the ATLAS Detector”. 2022. arXiv: 2201.08269.

[111] Per Ahrens. “Implementation of the Electron Identification in μ ⟶ τ Embedded Hybrid Events”. Karlsruhe Institute of Technology (KIT), 2018.

[112] Artur Akhmetshin. “Embedding - a Data Driven Method to Estimate the Z⟶ττ Background in the H⟶ττ Searches”. Karlsruhe Institute of Technology (KIT), 2016.

[113] Janek Bechtel. “Cross-Check of the CMS Search for Additional MSSM Higgs Bosons in the Di-τ Final State Using μ ⟶ τ Embedded Events”. Karlsruhe Institute of Technology (KIT), 2017.

[114] Sebastian Brommer. “Production of Hybrid Data Samples for Data Driven Background Determination in the H ⟶ τ τ Channel”. Karlsruhe Institute of Technology (KIT), 2019.

[115] S. Agostinelli et al. “GEANT4 - A Simulation Toolkit”. *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 506.3 (2003), pp. 250–303.
DOI: 10.1016/S0168-9002(03)01368-8.

[116] J. Allison et al. “Geant4 Developments and Applications”. *IEEE Transactions on Nuclear Science* 53.1 (2006), pp. 270–278.
DOI: 10.1109/TNS.2006.869826.

[117] J. Allison et al. “Recent Developments in Geant4”. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 835 (2016), pp. 186–225.
DOI: 10.1016/j.nima.2016.06.125.

[118] “Cms-Sw/Cmssw”. 2022. URL: https://github.com/cms-sw/cmssw (visited on 07/22/2022).

[119] G Petrucciani, A Rizzi, and C Vuosalo. “Mini-AOD: A New Analysis Data Format for CMS”. *Journal of Physics: Conference Series* 664.7 (2015), p. 072052.
DOI: 10.1088/1742-6596/664/7/072052.

[120] Andrea Rizzi, Giovanni Petrucciani, and Marco Peruzzi. “A Further Reduction in CMS Event Data for Analysis: The NANOAOD Format”. *Epj Web of Conferences* 214.CMS-CR-2018-396 (2019). Ed. by A. Forti et al., p. 06021.
DOI: 10.1051/epjconf/201921406021.

[121] J. Alwall et al. “A Standard Format for Les Houches Event Files”. *Computer Physics Communications* 176.4 (2007), pp. 300–304.
DOI: 10.1016/j.cpc.2006.11.010. arXiv: hep-ph/0609017.

[122] 2021 HPC Team Freiburg University of. “bwForCluster NEMO”. bwForCluster NEMO. URL: https://www.nemo.uni-freiburg.de/ (visited on 07/24/2022).

[123] The CMS Collaboration. “Measurements of Inclusive W and Z Cross Sections in pp Collisions at $\sqrt{s}$ = 7 TeV”. *Journal of High Energy Physics* 2011.1 (2011), p. 80.
DOI: 10.1007/JHEP01(2011)080. arXiv: 1012.2466.

[124] David Alexander Mason. “Measurement of the Strange - Antistrange Asymmetry at NLO in QCD from NuTeV Dimuon Data”. Oregon U., 2006. DOI: 10.2172/879078.

[125] J. S. Conway. “Incorporating Nuisance Parameters in Likelihoods for Multisource Spectra”. 2011. arXiv: 1103.0354.

[126] Roger Barlow and Christine Beeston. “Fitting Using Finite Monte Carlo Samples”. *Computer Physics Communications* 77.2 (1993), pp. 219–228. DOI: 10.1016/0010-4655(93)90005-W.

[127] The CMS Collaboration. “Measurement of Higgs Boson Production and Decay to the $\tau\tau$ Final State”. CMS-PAS-HIG-18-032 (2019).

[128] Diederik P. Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. 2017. DOI: 10.48550/arXiv.1412.6980. arXiv: 1412.6980.

[129] Xavier Glorot and Yoshua Bengio. “Understanding the Difficulty of Training Deep Feedforward Neural Networks”. *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Yee Whye Teh and Mike Titterington. Vol. 9. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR, –May 15, 2010, pp. 249–256.

[130] François Chollet et al. “Keras”. 2015. URL: https://keras.io.

[131] TensorFlow Developers. *TensorFlow*. Version v2.8.2. Zenodo, 2022. DOI: 10.5281/zenodo.6574269.

[132] Robert D Cousins. “Generalization of Chisquare Goodness-of-Fit Test for Binned Data Using Saturated Models, with Application to Histograms”.

[133] Steve Baker and Robert D. Cousins. “Clarification of the Use of CHI-square and Likelihood Functions in Fits to Histograms”. *Nuclear Instruments and Methods in Physics Research* 221.2 (1984), pp. 437–442. DOI: 10.1016/0167-5087(84)90016-4.

[134] Frank J. Massey. “The Kolmogorov-Smirnov Test for Goodness of Fit”. *Journal of the American Statistical Association* 46.253 (1951), pp. 68–78. JSTOR: 2280095.

[135] T. W. Anderson and D. A. Darling. “Asymptotic Theory of Certain ”Goodness of Fit” Criteria Based on Stochastic Processes”. *The Annals of Mathematical Statistics* 23.2 (1952), pp. 193–212. DOI: 10.1214/aoms/1177729437.

[136] Stefan Wunsch et al. “Identifying the Relevant Dependencies of the Neural Network Response on Characteristics of the Input Space”. *Computing and Software for Big Science* 2.1 (2018), p. 5. DOI: 10.1007/s41781-018-0012-1.

[137] Lorenzo Bianchini et al. “Reconstruction of the Higgs Mass in $H \longrightarrow \tau\tau$ Events by Dynamical Likelihood Techniques”. *Journal of Physics: Conference Series* 513.2 (2014), p. 022035. DOI: 10.1088/1742-6596/513/2/022035.

[138] The CMS Collaboration. “Jet Energy Scale and Resolution in the CMS Experiment in pp Collisions at 8 TeV”. *Journal of Instrumentation* 12.2 (2017), P02014.
DOI: 10.1088/1748-0221/12/02/P02014. arXiv: 1607.03663.

[139] Andrei V. Gritsan et al. “Constraining Anomalous Higgs Boson Couplings to the Heavy-Flavor Fermions Using Matrix Element Techniques”. *Physical Review D* 94.5 (2016), p. 55023.
DOI: 10.1103/PhysRevD.94.055023. arXiv: 1606.03107.

[140] Janek Bechtel et al. “Performance of the bwHPC Cluster in the Production of μ → τ Embedded Events Used for the Prediction of Background for H →ττ Analyses” (2019).
DOI: 10.15496/publikation-29043.

[141] Sebastian Brommer et al. *KIT-CMS/CROWN: V.0.1*. Version v0.1. Zenodo, 2022.
DOI: 10.5281/zenodo.7181926.

[142] Fons Rademakers et al. *Root-Project/Root: V6.20/06*. Version v6-20-06. Zenodo, 2020.
DOI: 10.5281/zenodo.3895852.

[143] Danilo Piparo et al. “RDataFrame: Easy Parallel ROOT Analysis at 100 Threads”. *EPJ Web of Conferences* 214 (2019), p. 06029.
DOI: 10.1051/epjconf/201921406029.

[144] Vincenzo Eduardo Padulano et al. “Distributed Data Analysis with ROOT RDataFrame”. *EPJ Web of Conferences* 245 (2020), p. 03009.
DOI: 10.1051/epjconf/202024503009.

[145] Joram Berger et al. “ARTUS - A Framework for Event-based Data Analysis in High Energy Physics”. Version 1. 2015.
DOI: 10.48550/arXiv.1511.00852. arXiv: 1511.00852.

# A Appendix

## Samples used

| Sample | Dataset Name |
|---|---|
| **2016** | |
| Run2016B-ver1-HIPM | /DoubleMuon/Run2016B-v1/RAW |
| Run2016B-ver2-HIPM | /DoubleMuon/Run2016B-v2/RAW |
| Run2016C-HIPM | /DoubleMuon/Run2016C-v2/RAW |
| Run2016D-HIPM | /DoubleMuon/Run2016D-v2/RAW |
| Run2016E-HIPM | /DoubleMuon/Run2016E-v2/RAW |
| Run2016F-HIPM | /DoubleMuon/Run2016F-v1/RAW |
| Run2016F | /DoubleMuon/Run2016F-v1/RAW |
| Run2016G | /DoubleMuon/Run2016G-v1/RAW |
| Run2016H | /DoubleMuon/Run2016H-v1/RAW |
| **2017** | |
| Run2017B | /DoubleMuon/Run2017B-v1/RAW |
| Run2017C | /DoubleMuon/Run2017C-v1/RAW |
| Run2017D | /DoubleMuon/Run2017D-v1/RAW |
| Run2017E | /DoubleMuon/Run2017E-v1/RAW |
| Run2017F | /DoubleMuon/Run2017F-v1/RAW |
| **2018** | |
| Run2018A | /DoubleMuon/Run2018A-v1/RAW |
| Run2018B | /DoubleMuon/Run2018B-v1/RAW |
| Run2018C | /DoubleMuon/Run2018C-v1/RAW |
| Run2018D | /DoubleMuon/Run2018D-v1/RAW |

**Table A.1.:** Samples used as input for the tau embedding technique.

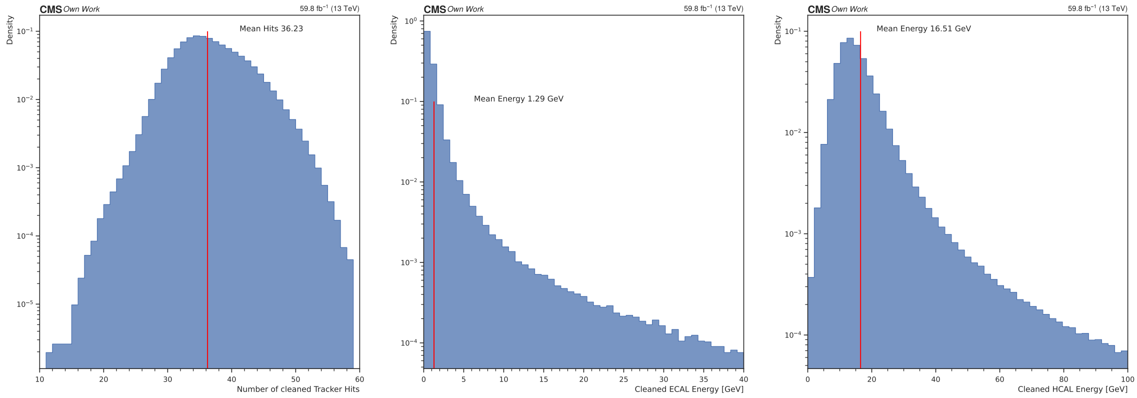


**Figure A.1.:** In the left plot, the number of removed hits is shown. The middle plot shows the energy distribution of the removed ECAL clusters, while the right plot shows the removed HCAL tower energy. For the HCAL energies, HB, HE, and HF are combined.

# Mass Selection Bias

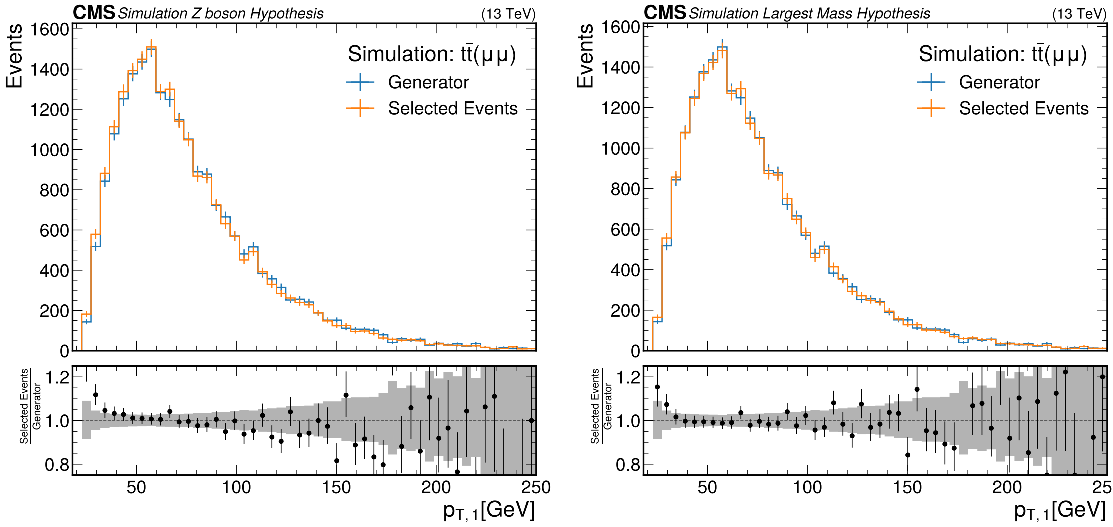


**Figure A.2.:** Comparison between the Z boson mass hypothesis(left) and the highest di-μ mass approach(right) for the distribution of the $p_T$ of the leading lepton.

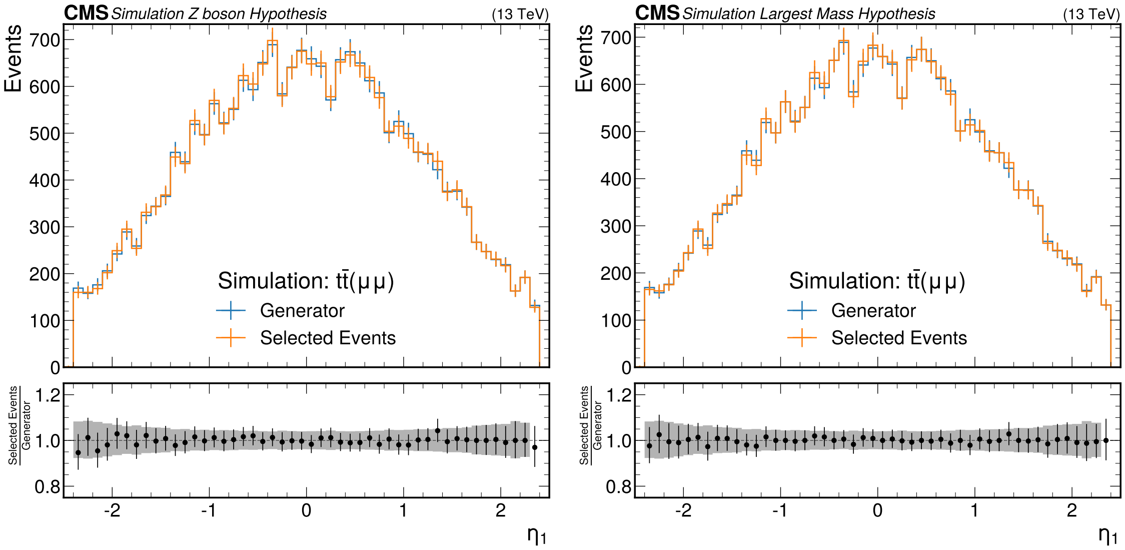


**Figure A.3.:** Comparison between the Z boson mass hypothesis(left) and the highest di-$\mu$ mass approach(right) for the distribution of the $\eta$ of the leading lepton.

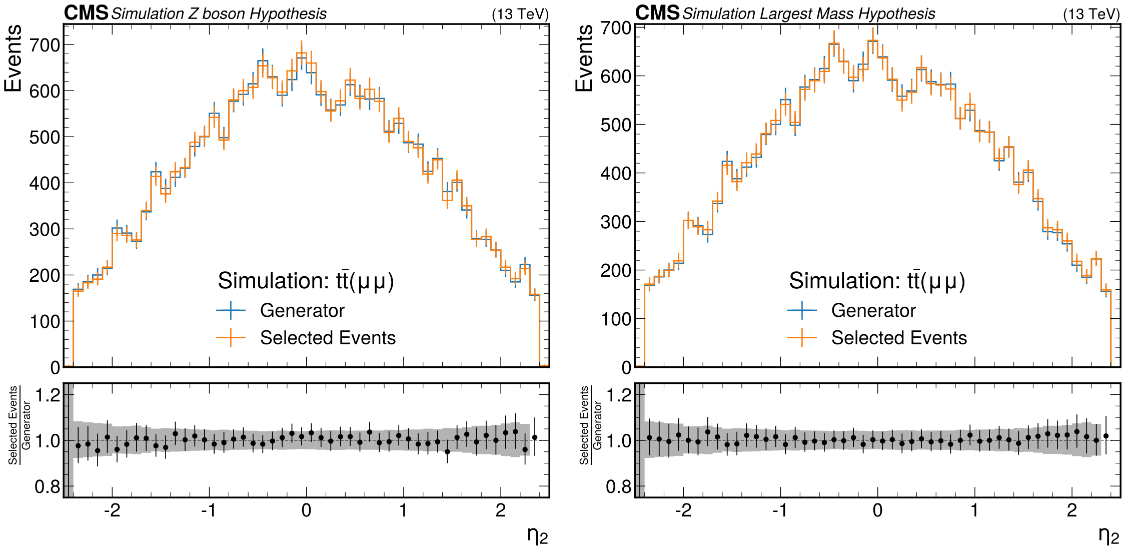


**Figure A.4.:** Comparison between the Z boson mass hypothesis(left) and the highest di-$\mu$ mass approach(right) for the distribution of $\eta$ the subleading lepton.

# μ → μ validation

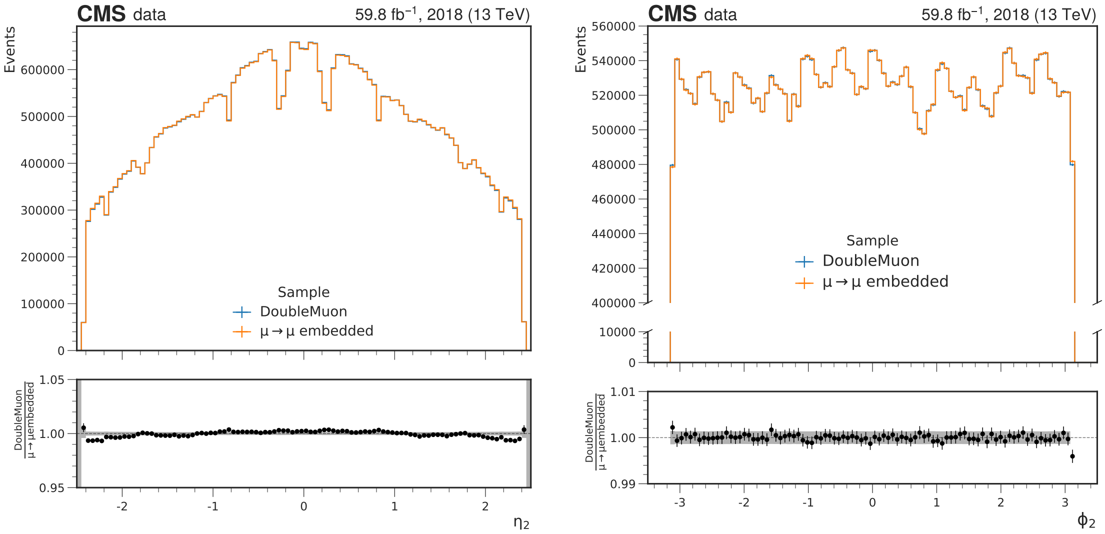


**Figure A.5.:** The distribution of η and ϕ for the subleading muon.

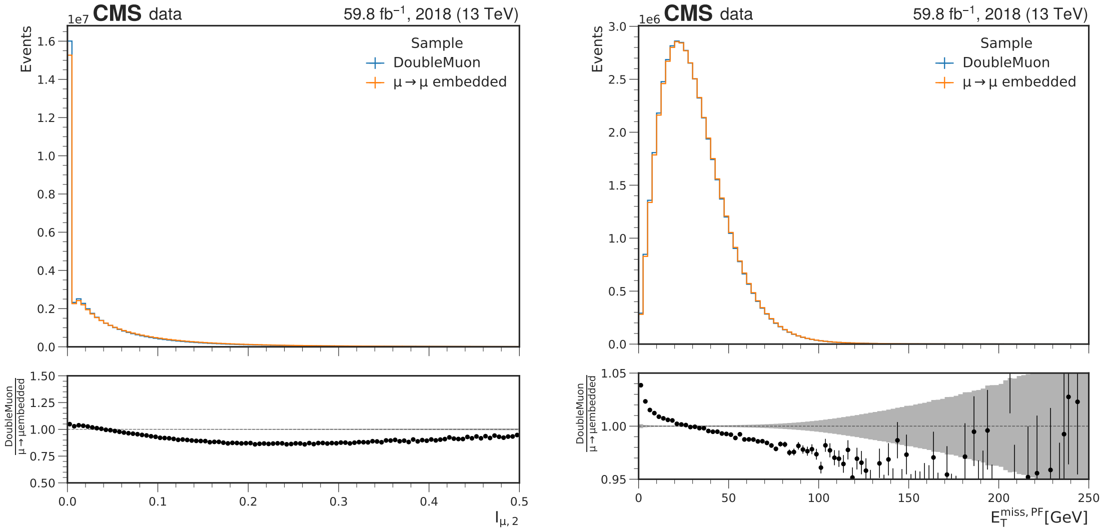


**Figure A.6.:** The distribution of $I^{\mu}_{\text{rel}}$ of the subleading muon and $E_T^{\text{miss}}$ using PF information.

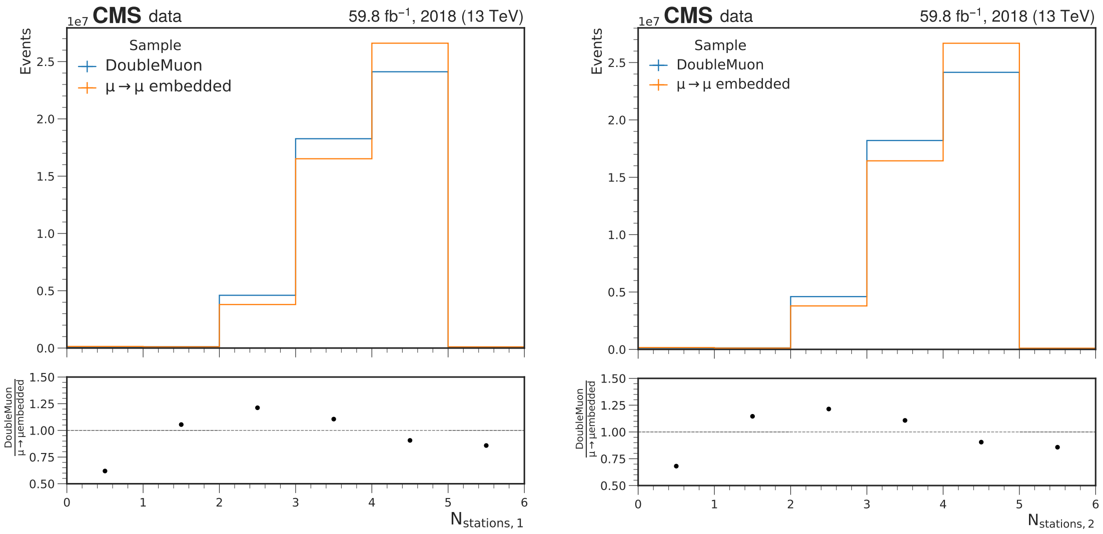


**Figure A.7.:** The distribution of the number of muon chamber hits for the leading and subleading muon.

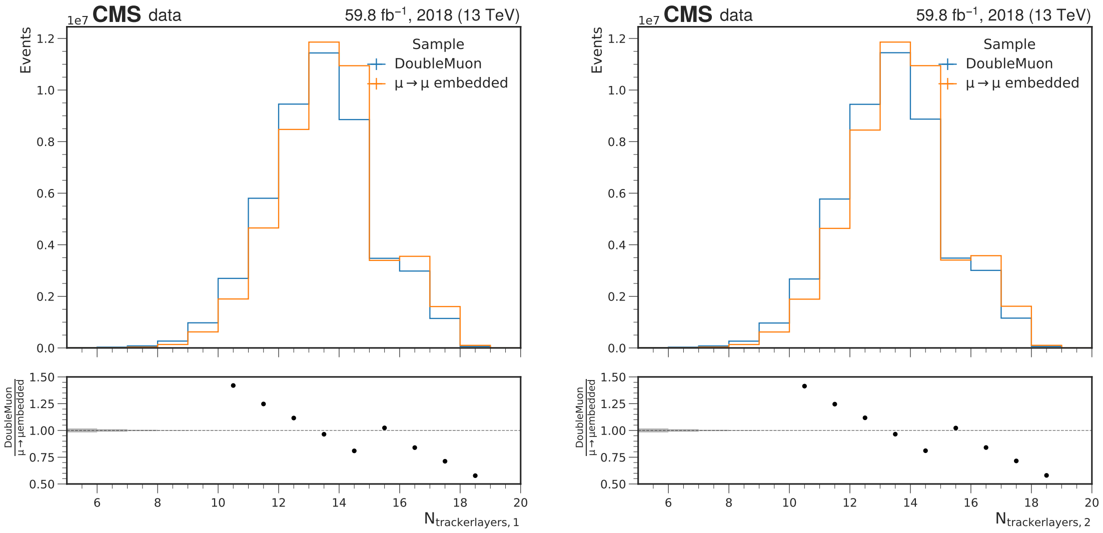


**Figure A.8.:** The distribution of the number of tracker layer hits for the leading and subleading muon.

## Simulation Training

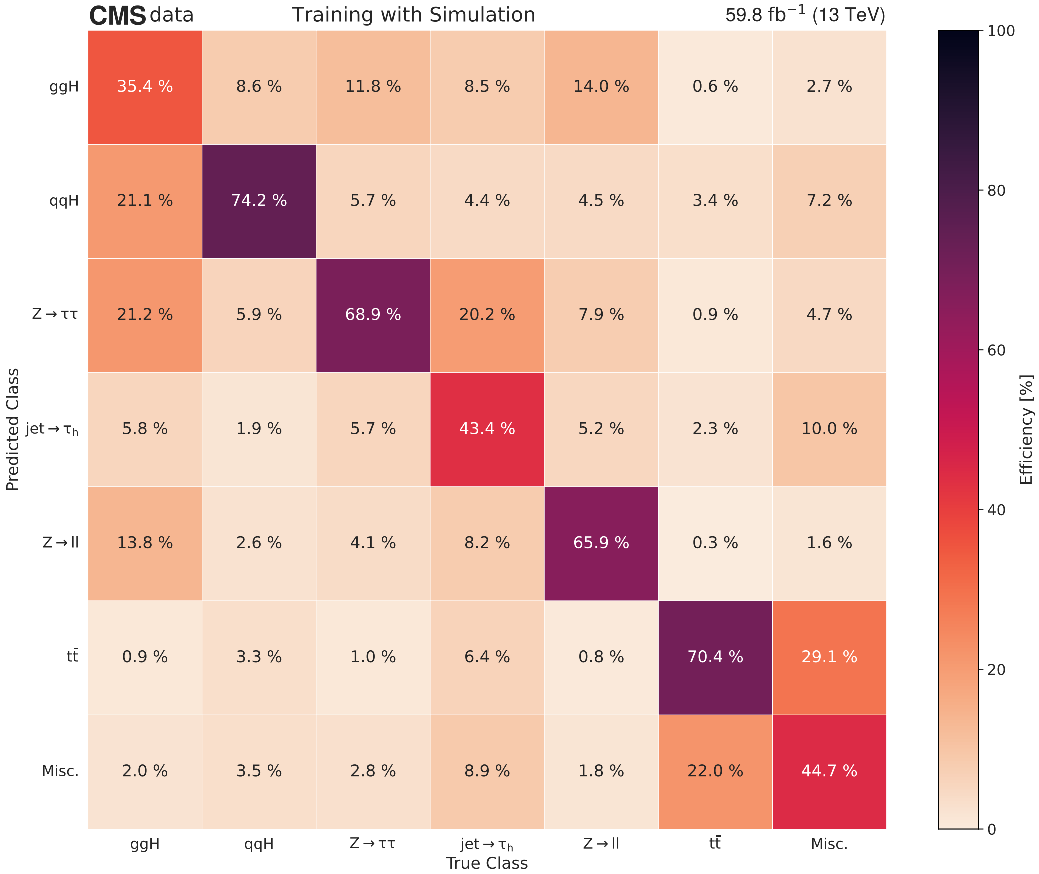


**Figure A.9.:** Confusion matrix of the training using 2018 data and simulation instead of $\tau$-embedded samples in the $\mu\tau_h$ channel.

## Control Plots for the $\mu\tau_h$ final state with 2018 data

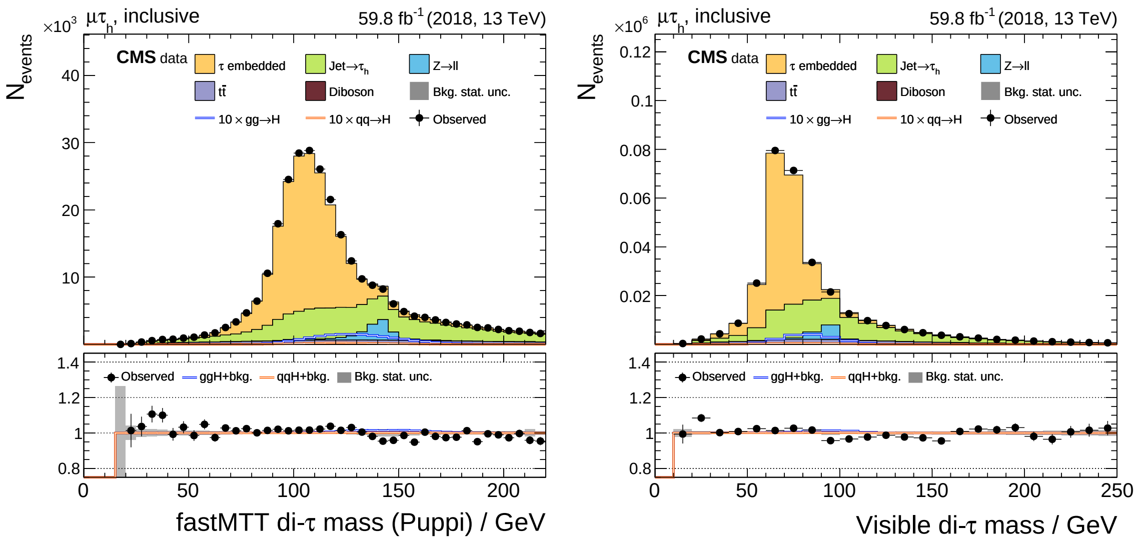


**Figure A.10.:** Control plot of di-τ mass (left) and the visible di-τ mass (right) in the $\mu\tau_h$ channel. Only statistical uncertainties are shown.

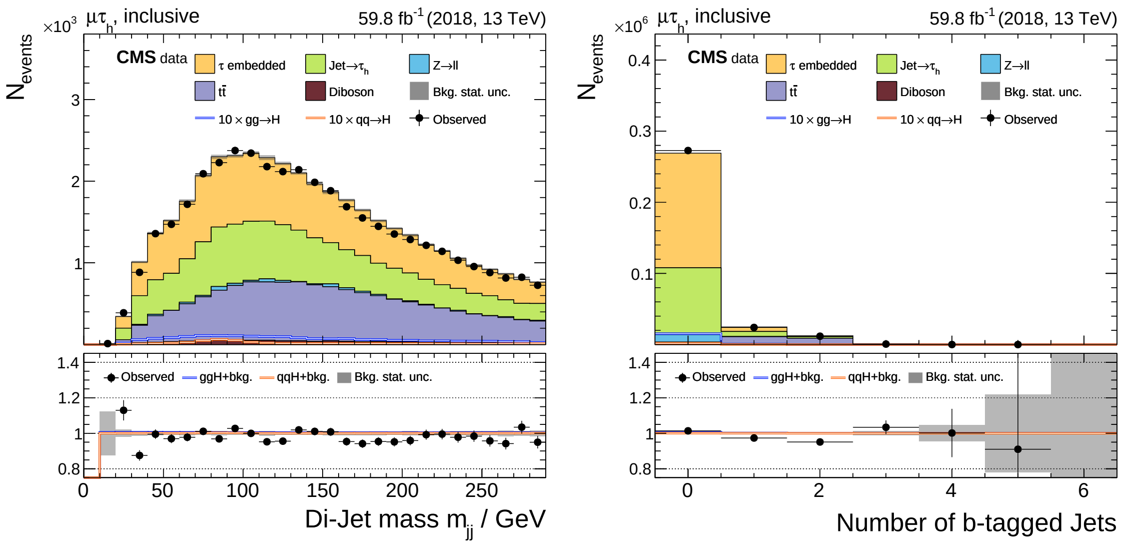


**Figure A.11.:** Control plot of di-jet mass (left) and the number of b-jets (right) in the $\mu\tau_h$ channel. Only statistical uncertainties are shown.

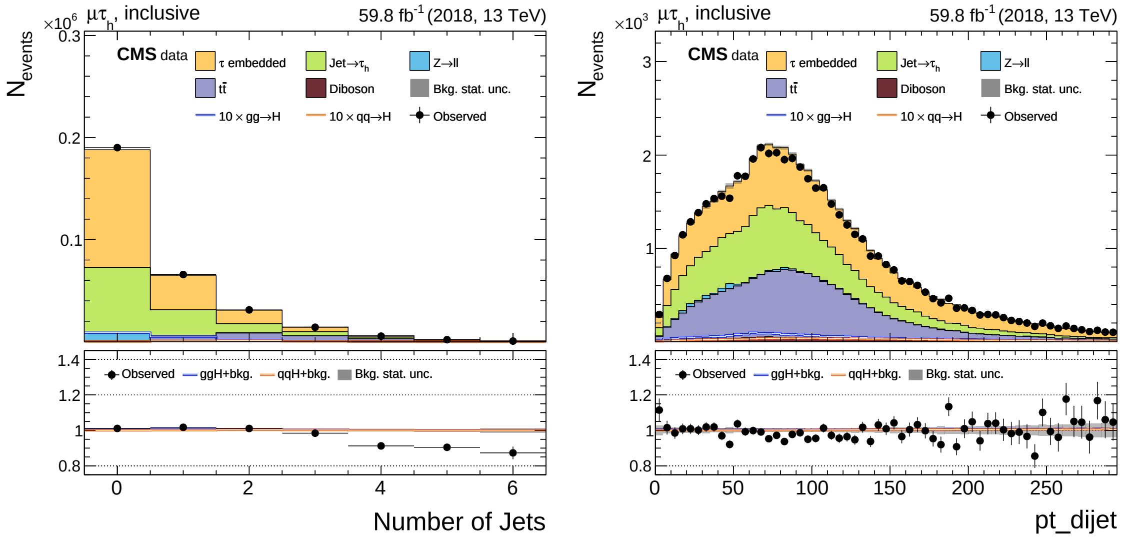


**Figure A.12.:** Control plot of the number of jets (left) and the $p_T$ of the di-jet system (right) in the $\mu\tau_h$ channel. Only statistical uncertainties are shown.

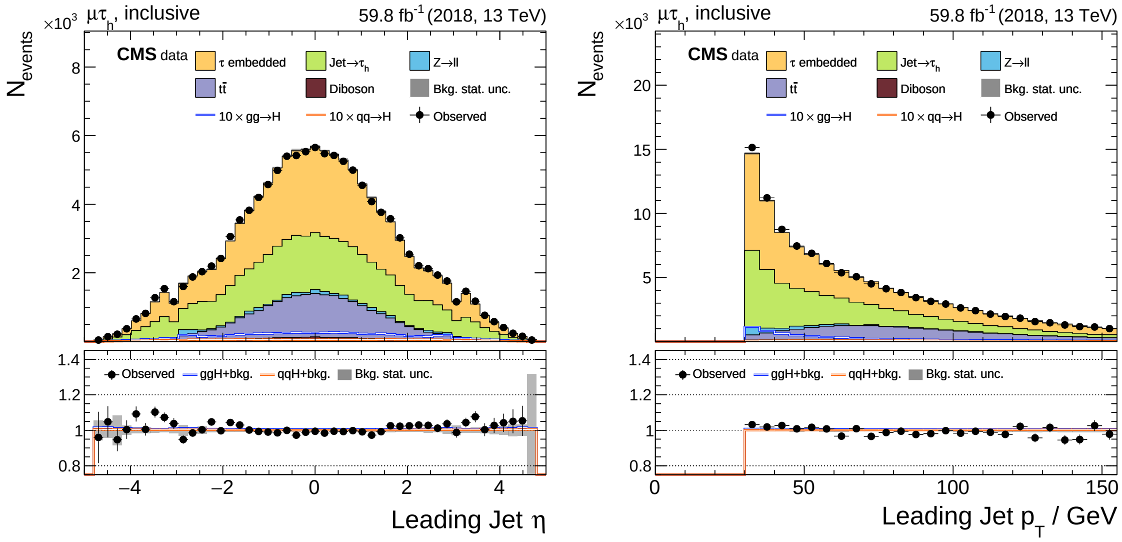


**Figure A.13.:** Control plot of $\eta$ (left) and $p_T$ (right) of the leading jet in the $\mu\tau_h$ channel. Only statistical uncertainties are shown.

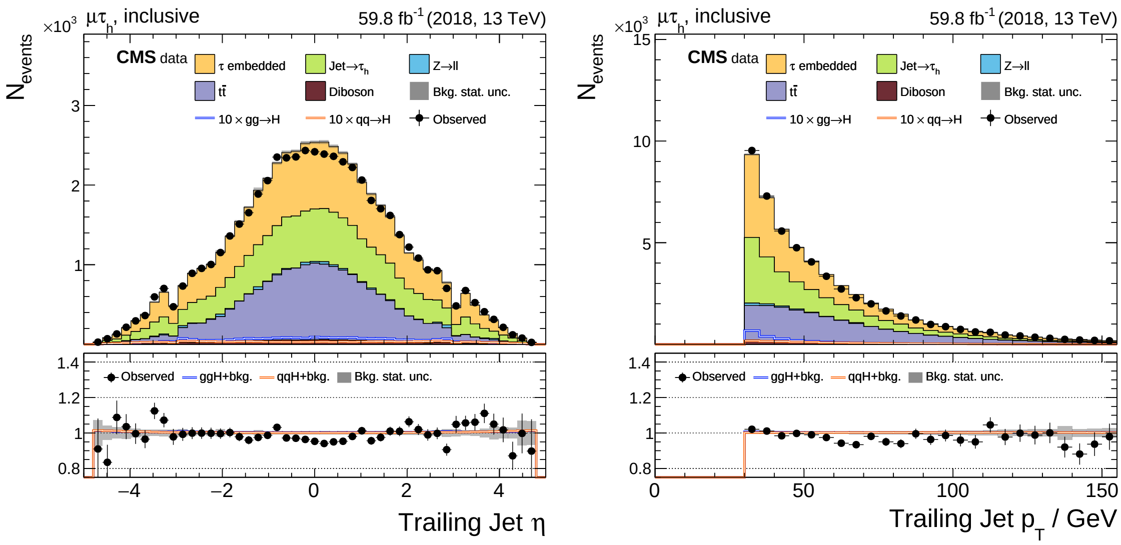


**Figure A.14.:** Control plot of η (left) and $p_T$ (right) of the subleading jet in the $\mu\tau_h$ channel. Only statistical uncertainties are shown.

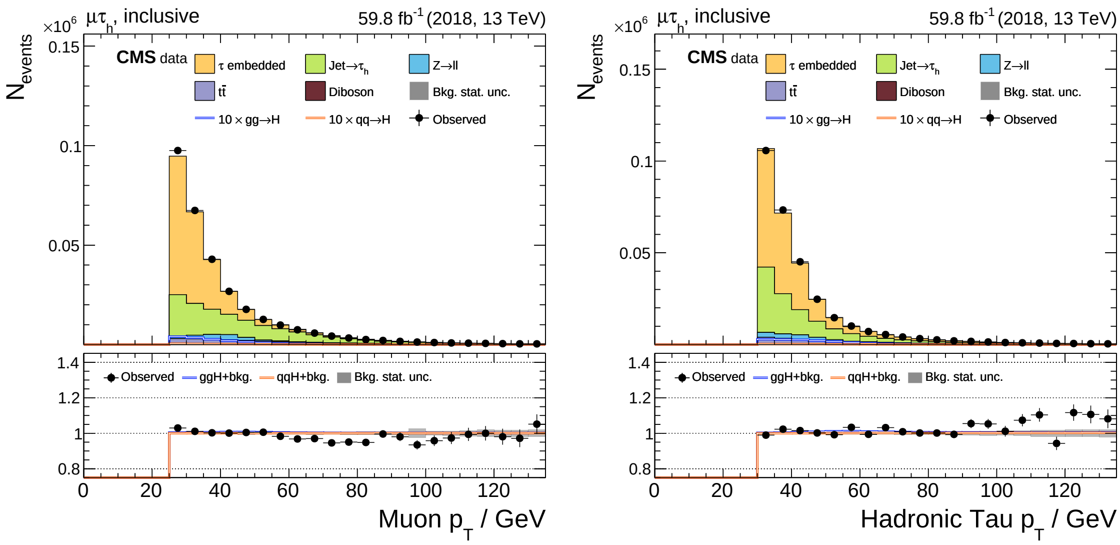


**Figure A.15.:** Control plot of $p_T$ of the muon (left) and the $p_T$ of the $\tau_h$ (right) in the $\mu\tau_h$ channel. Only statistical uncertainties are shown.

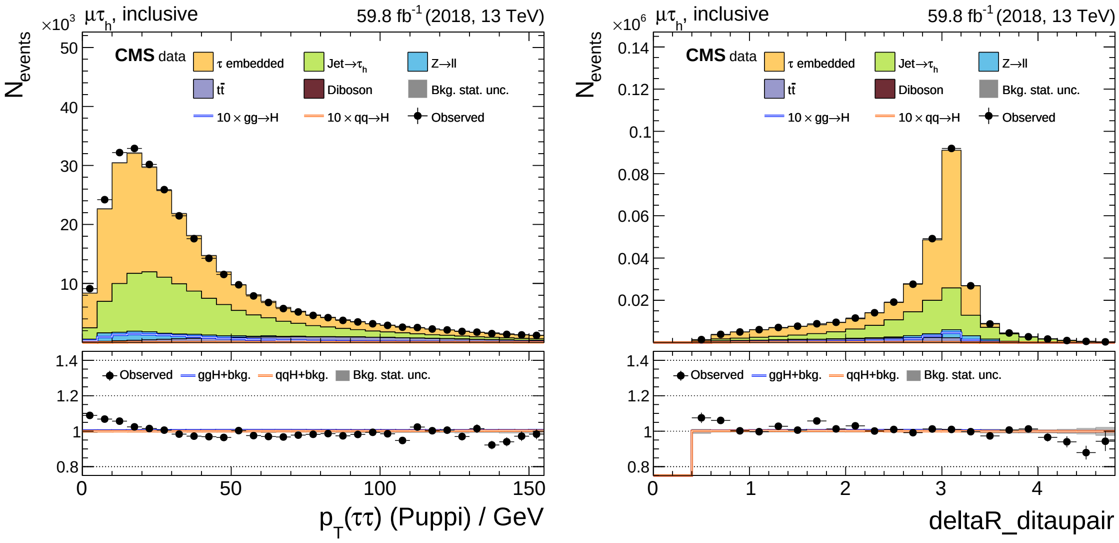


**Figure A.16.:** Control plot of the $p_T$ of the di-$\tau$ system (left) and the $\Delta$R between the muon and the $\tau_h$ (right) in the $\mu\tau_h$ channel. Only statistical uncertainties are shown.

# Anaylsis Background Categories

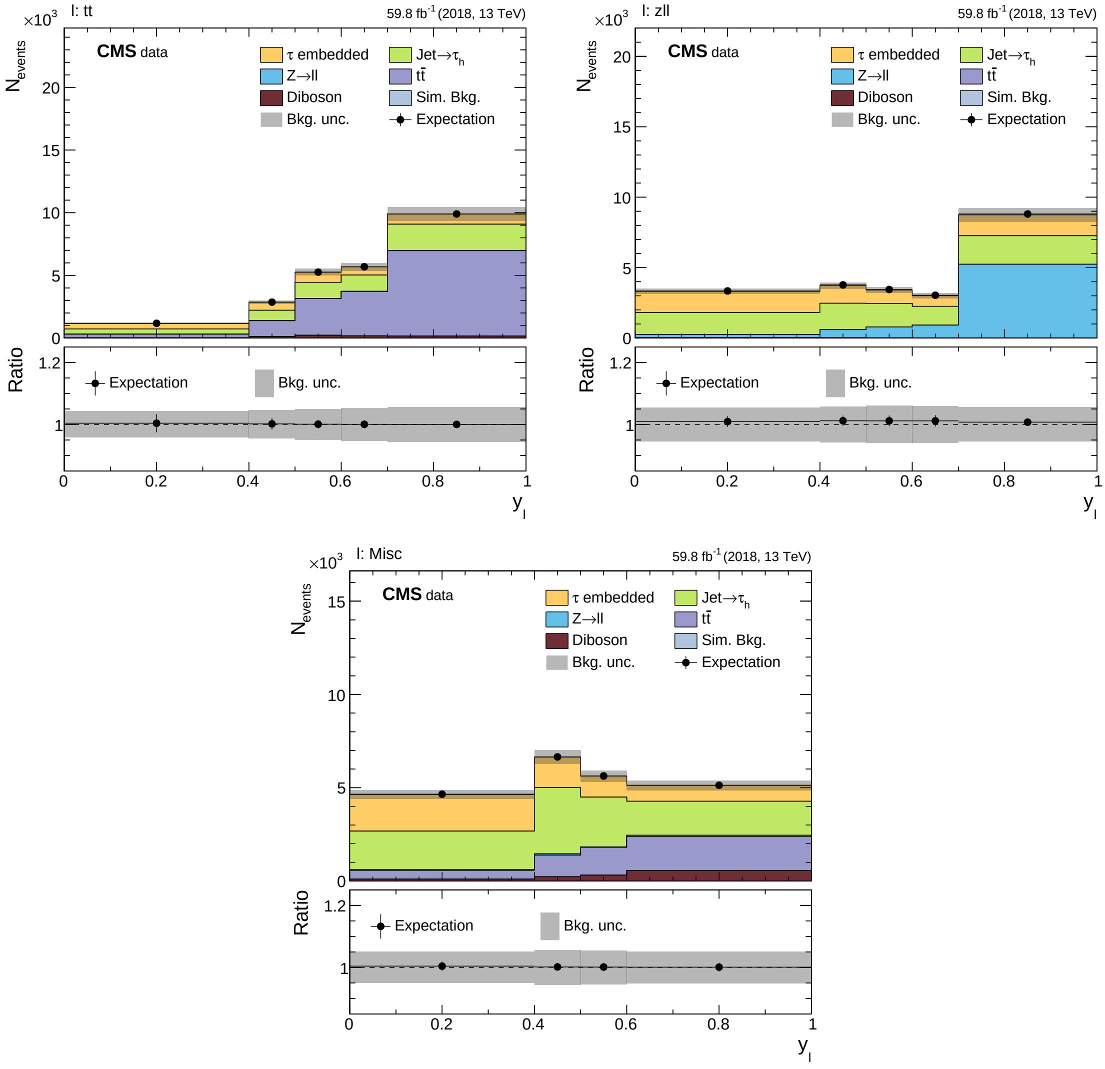


**Figure A.17.:** Histograms of the neural network score for the different categories. In the top row, the $t\bar{t}$ category is shown on the left, and the $Z \to ll$ categoy is shown on the right. In the bottom row, the Misc. categories is shown. For all distributions, the data was replaced by the background plus signal prediction.

## A.1. The CROWN Framework

In the scope of the upcoming Run III of the LHC and the increasing amount of available collision data, developments on the analysis software side are unavoidable in today's particle physics. Precision measurements, like the Run II Higgs physics program of the CMS Collaboration, are only possible by combining all available collision data and even more simulated events. Fast analysis turnaround cycles are also essential to test new ideas and push the boundaries of what is possible with the available data set. To ensure that these precision measurements continue, the CMS Collaboration developed the NanoAOD data tier [120]. In this data format, the relevant information of one event can be stored within 1 kB to 2 kB of disk storage.

To obtain physics results, several processing steps are required. In Figure A.18, a typical analysis workflow is depicted. Three different frameworks are needed to obtain the final results. At first, the NanoAOD data is converted into analysis tuples. These tuples contain only information relevant to the analysis. In addition, these tuples are flat, meaning that each variable is stored in a separate column. The tuples are then used to generate histograms, which are used for the subsequent statistical analysis. Whether this analysis tuple step is needed depends on the scope of the analysis. If the amount of data that has to be processed is too large, and if multiple people work together on a single analysis, it is often beneficial to have this intermediate step of analysis tuples. The analysis tuples represent a common ground to share among analysts. In addition, the creation of histograms directly from NanoAOD can be much slower if many events have to be processed.

The Crown framework [141] is an analysis tuple framework based on ROOT [142] RDataFrames (RDF) [143, 144]. RDF is a declarative interface that can be used to build a computation graph of different operations. The ROOT software handles details like I/O, parallelization and graph optimization in the background. The computation of the graph is also lazy-executed, meaning all computations are only performed when triggered by the user. The RDF ecosystem provides a potent tool that can be used for all types of analysis tasks. The RDF interface is implemented in C++, with an additional Python interface. In RDF, the two most important functions are `Define` and `Filter`. While `Define` allows the user to generate a new column in the data frame, `Filter` allows the user to select a subset

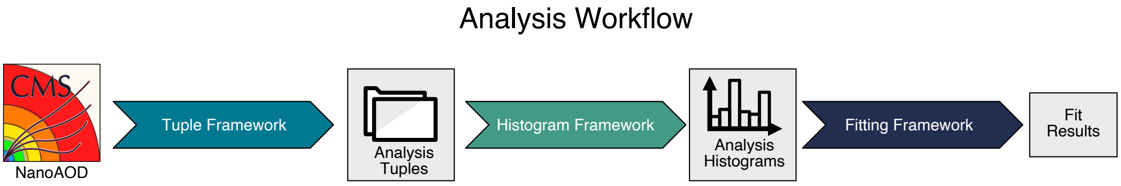


**Figure A.18.:** A sketch of a general analysis workflow. The NanoAOD data from CMS is converted into analysis tuples. These tuples are then used to generate histograms, which can then be used to perform the statistical analysis using a fitting framework.

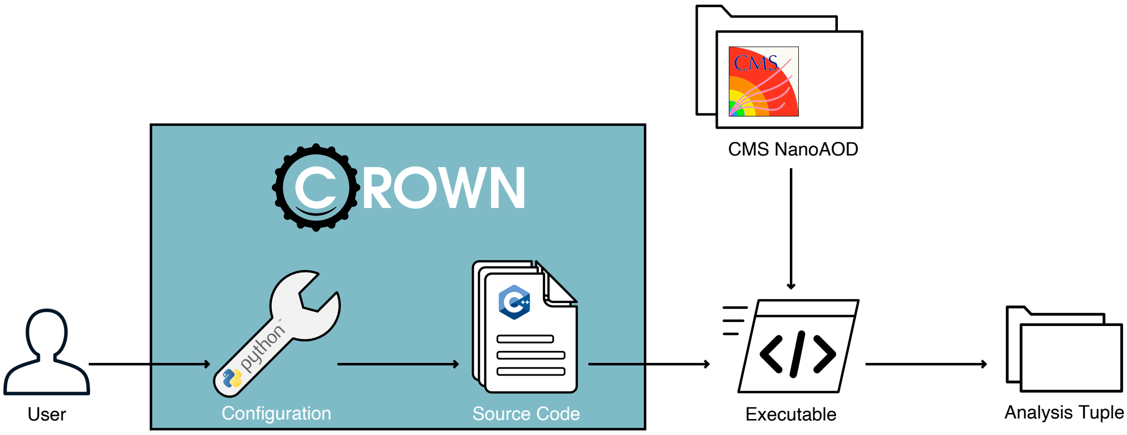


**Figure A.19.:** A sketch of the CROWN framework workflow. The user-provided configuration is used to generate C++ code, which contains the definition of an RDF computation graph. After compilation, the executable can be used to run this computation graph.

of the data frame. Performance studies have shown that the RDF interface provides one of the fastest implementations of a columnar data frame in the HEP ecosystem [144].

The CROWN framework was designed with two main ideas in mind: to provide a configuration-based interface to the user while performing as fast as possible. Rather than manually defining the RDF computation graph, the user only has to write a configuration. The framework handles the generation of the computation graph. PYTHON is used to keep the configuration's generation as flexible as possible. To obtain good performance, the PYTHON interface of RDF is omitted and instead, the PYTHON configuration is used to automatically generate C++ code. This auto-generated code contains the definition of an RDF computation graph. It is then compiled into an executable, which can be used to run the computation graph of the configuration provided by the user. The framework is not limited to using NANOAOD as input; in principle, any data set that can be processed with the RDF interface can serve as input data. This includes analysis tuples themselves, which allows the computation of high-level variables such as neural network scores afterwards. The overall workflow of the framework is depicted in Figure A.19.

The chosen approach yields some major benefits

- Since the computation graph used for the analysis tuple generation is compiled C++ code, it can benefit from compiler optimizations, which are not possible when using an interpreted language like PYTHON.

- The auto-generated code serves as an analysis preservation that can be used to understand and recreate every step of the analysis tuple generation.

- Once the executable is generated, it can be reused for any number of input files without any additional overhead.

- The user is still able to use PYTHON code to Define the configuration, which allows for a very flexible configuration.

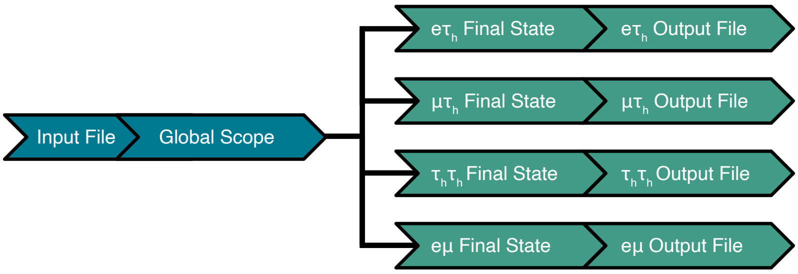


**Figure A.20.:** Sketch of the scopes concept in the CROWN framework as applied in the H → ττ measurement.

## Configuration

The configuration consists of building blocks used to build the auto-generated code. The dataflow is realized using quantities, producers, and parameters. Quantities represent columns in the data frame. They are either variables from the input data or newly derived variables. These new variables can be derived using a producer. The user has to specify, which quantities should be included in the output tuple, and which producers should be run. parameters are used to set user-Defined variables, e.g. selection criteria on the $p_T$ of an object. In the declaration of a producer, its input and output quantities are Defined, allowing the framework to automatically determine the correct processing order. In addition, filter functions, which reduce the size of the data set, are run as early as possible to avoid unneeded computations. The declaration of the producer also contains a mapping to a corresponding C++ function call. This mapping is used to generate the auto-generated code.

The computations are Defined in the CROWN C++ functions. While the basic structure of these functions is always identical and contains at least one RDF filter or Define command, arbitrary implementations using the RDF interface are possible. This abstraction layer allows performing more complicated tasks such as the di-τ pair selection algorithm described in Section 4.2.1 with a single function call.

In many analyses, a single input file is used to create analysis tuples for multiple final states. To ensure that every input file is only read once, the CROWN framework uses different scopes. The concept is visualized in Figure A.20. The global scope is always the top layer in the computation graph. All operations common to all events, regardless of the final state, can be performed in the global scope. After that, the computation graph is split based on the number of scopes Defined by the user. For example in the H → ττ measurement, the analysis tuples for the four final states $e\tau_h$, $\mu\tau_h$, $\tau_h\tau_h$, and eμ can be generated in a single run, without the need of reading the input data twice.

## Systematic Uncertainties

One major feature of the CROWN framework is the automatic calculation and tracking of systematic variations. The user can Define a new systematic variation in the configuration. A systematic variation can be a different configuration parameter, input quantity, or an entirely different producer. All output quantities, including quantities, that are calculated

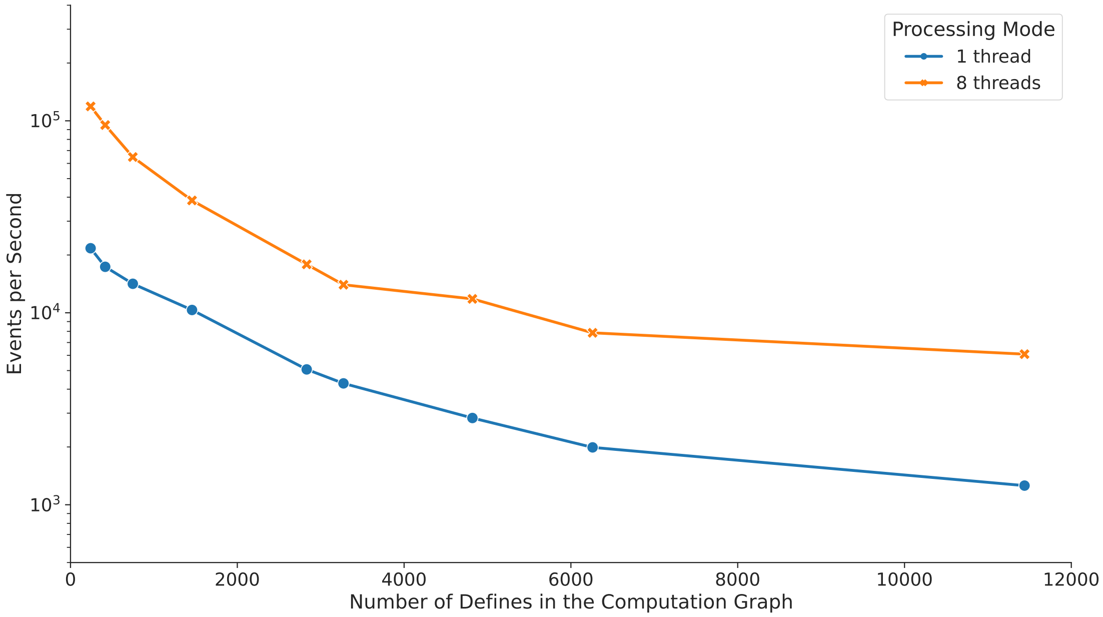


**Figure A.21.:** The scaling of the CROWN framework with multithreading. The larger the computation graph grows, the longer the setup time is. The setup time reduces the performance gain when using multiple threads since the setup is performed in single-thread mode.

based on `quantities` that are affected by this systematic variation are tracked. Multiple C++ function calls are generated during the code generation, one for each variation. This way, only the necessary variations are calculated, and the output is automatically tracked. `Quantities` unaffected by a systematic shift are not recalculated. The shifted `quantities` are stored in new columns using the `quantityname__shiftname` naming convention. This treatment of systematic variations results in a major improvement in performance and disk usage, as only quantities Defined by the user are calculated and stored.

## Performance

The CROWN framework is a significant step up in processing time. The analysis tuples of the Run II $H \rightarrow \tau\tau$ measurement [9] were generated using the ARTUS framework [145]. This framework was based on a C++ event loop and was able to process around 100 events per second and final state when including all required systematic variations. The CROWN framework achieves more than 4000 events per second for the same analysis tuple content. The size of output files is reduced by up to 70%, mainly due to the automatic tracking of systematic variations. Since RDF has built-in multithreading, the performance can be further improved by using multiple threads. In a realistic setup, when producing four final states in parallel with eight threads, the framework can process 6000 events per second, more than 200 times faster than its predecessor. Such a setup corresponds to more then 11,000 RDF `Define` calls in the computation graph.

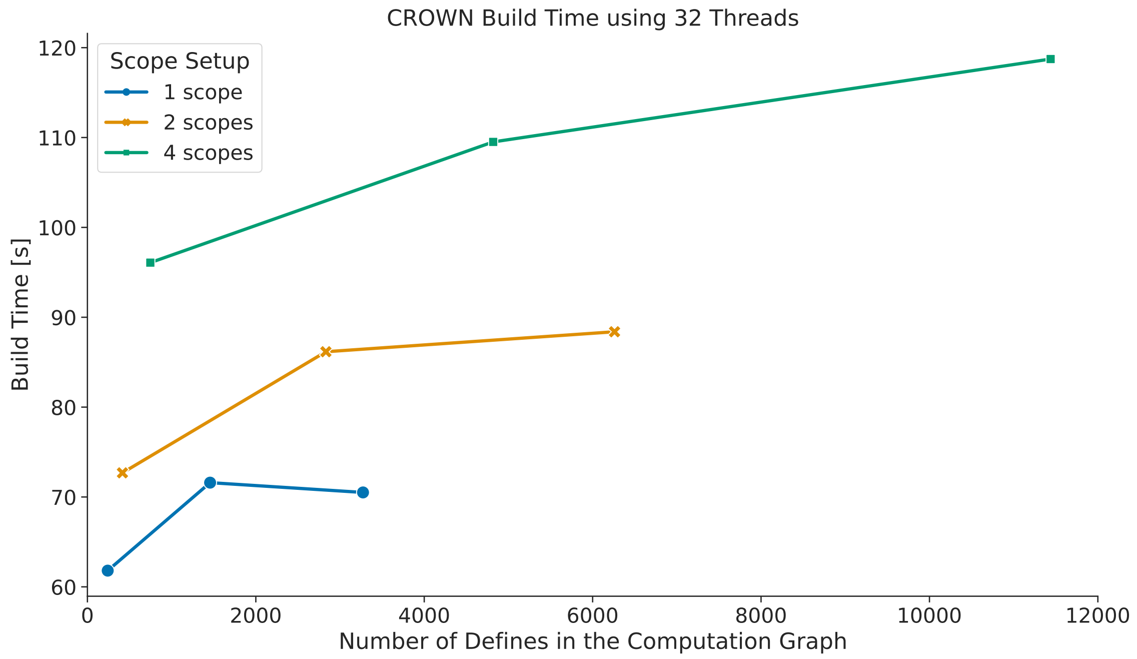


**Figure A.22.:** The scaling of the CROWN framework build time depends on the number of `scopes`. Additional `scope` increases the build time, while a larger computation graph has a limited impact on the build time.

In Figure A.21 the scaling of the CROWN framework with multithreading and increasing number of `Defines` is shown. All benchmarks are performed with local SSD access to the input file to exclude effects due to I/O limitations. The performance gain when using multithreading is not 100%, as the setup of the compilation graph is performed in single-threaded mode. This setup time depends on the size of the graph: for a small graph with 400 `Defines`, the setup time is 4 s, while for a big graph with about 10,000 `Defines`, the setup time increases to 240 s. For larger graphs, it is beneficial to reduce the impact of the setup time by increasing the number of events that are processed once the graph is loaded.

In Figure A.22, the build time of the executable, depending on the number of `Defines` and the number of `scopes`, are shown. All built benchmarks were performed using 32 threads. While adding new `scopes` results in a linear increase of the build time, an increase in `Define` calls only has a limited impact on the build time. Even large graphs with more than 10,000 `Define` calls can be compiled in less than 120 s. The build is only required once, and the executable can be reused for different input files.

The CROWN framework poses a significant step in analysis tuple generation. Compared to the previous framework, the processing is more than a magnitude faster while significantly reducing disk usage. The framework is also more flexible, as arbitrary computations can be performed. The RDF backend allows for built-in support of remote I/O, multithreading, and low-level optimizations. The framework is available as an open-source project on GitHub[1].

[1] https://github.com/KIT-CMS/CROWN/

# Danksagung

Ich möchte Professor Quast dafür danken, dass er mir vor mehr als drei Jahren die Möglichkeit geboten hat, an dem Thema meiner Masterarbeit weiterarbeiten zu können, und ich die $\tau$-embedding Methode weiterzuentwickeln und in Analysen zur Anwendung zu bringen. Über die Zeit meiner Promotion durfte ich viel über Analyse, Rekonstruktion und Datenprozessierung lernen, aber auch über gute, enge Zusammenarbeit an großen Themen. Er hat mir immer den Rücken freigehalten und sich um alle bürokratischen Hürden gekümmert.

Des Weiteren möchte ich Roger Wolf dafür danken, dass er meine Arbeit so exzellent betreut hat. Er hatte immer ein offenes Ohr und viele gute Ideen und Ansätze, wenn es mal nicht weiter ging. Auch bei schwierigeren Fragen hat er sich immer die Zeit genommen, um zusammen auf gute Lösungen zu kommen.

Außerdem möchte ich dem ETP, und im speziellen dem Admin Team und dem Computing Team danken. Meine Probleme, da $\tau$-embedding doch etwas aufwendiger ist, wurden stets erhört und gelöst. Es tut mir leid, dass ich manchmal wochenlang das Batch-System so belegt habe. Danke an Florian, Nils, Matthias und Robin. Dann möchte ich natürlich auch noch der $H \to \tau\tau$ Arbeitsgruppe danken, mit denen ich in den letzten Jahren so viel zusammengearbeitete habe: Maximilian, Ralf, Christian, Felix, Moritz, Sebastian, Artur, Stefan und Janek. Es hat mir viel Spaß gemacht, die Analysen mit euch zusammen voranzutreiben. Besonders hervorheben muss ich natürlich das Büro 8-16, mit den besten Schreibtisch Nachbarn, die man sich vorstellen kann! Grüße gehen raus an Christian, Ralf und Florian.

Als letztes möchte ich noch meiner Familie und meiner Freundin Luisa danken, dass sie mich die ganze Zeit, und speziell die anstregende Schreibphase über unterstützt haben und mir den Rücken frei gehalten haben. Ohne euch hätte ich es nicht so weit geschafft.

# Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die Dissertationsschrift mit dem Titel

*A data-driven method for Higgs boson analyses in di-$\tau$ final states for the LHC Run II and beyond*

selbständig angefertigt, alle benutzten Hilfsmittel vollständig und genau angegeben und alles kenntlich gemacht habe, was aus Arbeiten Anderer unverändert oder mit Abänderung entnommen wurde.

Ich versichere außerdem, dass ich die Dissertation nur in diesem und keinem anderen Promotionsverfahren eingereicht habe und dass diesem Promotionsverfahren keine endgültig gescheiterten Promotionsverfahren vorausgegangen sind.

Karlsruhe, den 31.10.2022
Sebastian Brommer