Production of Hybrid Data Samples for Data Driven Background Determination in the $H \rightarrow \tau\tau$ Channel
Produktion von hybriden Datensätzen für die
datengetriebene Bestimmung von
Untergründen im $H \rightarrow \tau\tau$ Kanal

Zur Erlangung des akademischen Grades

MASTER OF SCIENCE

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

MASTERARBEIT

von

B. Sc. Sebastian Brommer

geb. 26.05.1993
in Stuttgart

Abgabedatum: 12. März 2019

Referent: Priv.-Doz. Dr. Roger Wolf
Korreferent: Prof. Dr. Günter Quast

Institut für Experimentelle Teilchenphysik
Abstract

The embedding technique is a data driven method, where $\mu\mu$ events are selected from data, and the muons are replaced by simulated $\tau$-lepton decays. In this way, a hybrid event is created, which only relies on the simulation for the well-understood $\tau$-lepton decay. The remainder of the event, by construction, provides a better description of the data than full simulation, especially for challenging simulation tasks, such as the underlying event or multijet production. Within the scope of this thesis, the embedding technique was applied to the data measured by CMS in 2017. The performance of the method concerning the lepton isolation, the electron identification, the calorimeter cleaning, and the HLT reconstruction of hadronically decaying $\tau$-leptons was studied and improved. In addition, the embedding method was used in a $H \rightarrow \tau\tau$ analysis and the determination of all necessary correction factors has been performed. Distributions using embedded events are statistically compatible with observed data, as demonstrated using goodness-of-fit tests while resulting in reduced uncertainties compared to simulation. The performed validations and improvements of the technique demonstrate that the method provides a robust estimation of Standard Model processes with a $\tau\tau$ final state.
Contents

1 Introduction 3
  1.1 The Standard Model of Particle Physics ......................... 4
    1.1.1 The Higgs Mechanism ..................................... 4
    1.1.2 The $H \to \tau\tau$ Decay Channel .......................... 6

2 The CMS Experiment 9
  2.1 The CMS Detector ........................................... 9
  2.2 The CMS Trigger System ..................................... 12
    2.2.1 High-Level Trigger Hadronic Tau Reconstruction ............. 13
  2.3 Reconstruction ............................................. 15
    2.3.1 Track and Vertex Reconstruction ............................ 16
    2.3.2 Particle Flow ........................................... 18
    2.3.3 Lepton Reconstruction and Identification ................... 19

3 The Embedding Technique 23
  3.1 Embedding Strategy ........................................ 24
    3.1.1 Selection Step .......................................... 24
    3.1.2 Cleaning Step .......................................... 26
    3.1.3 Generator Step .......................................... 27
    3.1.4 Merging Step .......................................... 30
  3.2 Production of Embedded Events using 2017 Data ................. 32
    3.2.1 Workflow ............................................. 32
    3.2.2 Performance of the Embedded Event Production ............. 34

4 Validation and Improvement of the Embedding Technique 39
  4.1 Validation .................................................. 39
    4.1.1 Applying $\mu \to \mu$ Embedding on Simulated Events ........... 39
    4.1.2 Applying $\mu \to e$ Embedding on Simulated Events .......... 47
    4.1.3 Applying $\mu \to \tau$ Embedding on Simulated Events ........ 54
  4.2 Improvements of the Embedding Technique ....................... 56
    4.2.1 Simulation of HLT $\tau_h$ Reconstruction in Embedding ....... 56

5 Application of Embedded Events in an Analysis 63
  5.1 Corrections and Uncertainties for Embedded Events .......... 63
CHAPTER 1

Introduction

The Standard Model of particle physics (SM) provides a very good and precise description of elementary particles and their interactions. Since its current formulation in 1970, the SM correctly predicted multiple new particles like the top quark, the W, or the Z boson. In 2012 the CMS and ATLAS experiments at the Large Hadron Collider (LHC) discovered a new boson with a mass of $m_H = 125.09$ GeV and properties consistent with a Higgs boson [1–3]. This was considered to be the last missing piece of the SM. Since the discovery, many measurements have been made, not only to study the observed boson in more detail, but also to search for additional Higgs bosons. Deviations in the behavior of the observed Higgs boson may point to physics beyond the Standard Model (BSM).

The branching fraction of the observed Higgs boson decaying into a pair of $\tau$-leptons is the second largest of all fermions. The decay of a Higgs boson into a pair of $\tau$-leptons was first observed with a significance of 5$\sigma$ in a combination of CMS and ATLAS results [4]. With the LHC upgrade to a center of mass energy of $\sqrt{s} = 13$ TeV, both experiments established the decay channel independently [5, 6]. In addition, the decay $H \rightarrow \tau\tau$ is of particular interest in BSM searches [7, 8], since many beyond the Standard Model theories with more than one Higgs boson predict a modified coupling of additional Higgs bosons to $\tau$-leptons.

The cross section of a Higgs boson decaying into $\tau$-leptons is small compared to other SM processes with a $\tau\tau$ final state. The most important background process is the decay of the Z boson into $\tau$-leptons. As a good understanding of background processes is crucial, the embedding technique [9] was developed as a method to describe SM processes with a $\tau\tau$ final state from data. For this method, $Z \rightarrow \mu\mu$ events are selected from data, and the selected muons are replaced by simulated $\tau$-lepton decays. This results in a hybrid event, that only relies on simulation for the well described $\tau$-lepton decay. The remainder of the event provides a better description of the data than simulation, without any need of additional tuning. This is especially true for challenging simulation tasks, such as the underlying event or multijet production.

The method was successfully used by the CMS collaboration in Higgs Boson analyses from 2009 to 2013 [1, 3]. The ATLAS collaboration also used a similar technique [10] in several analyses. Together with the upgrade of the CMS detector, the method was re-
implemented and first used as crosscheck in a search for additional neutral Higgs Bosons in the context of the MSSM [11] by the CMS collaboration.

In chapters 1 and 2 of this thesis, the standard model of particle physics and the CMS detector will be introduced. Chapter 3 contains a detailed description of the embedding technique and the production of embedded events using data measured by the CMS experiment in 2017. In chapter 4, multiple validation studies and improvements of the technique are presented. The application of embedded events in a typical $H \rightarrow \tau\tau$ analysis is illustrated in chapter 5. A short summary is given in chapter 6.

1.1 The Standard Model of Particle Physics

The SM is a quantum field theory that describes elementary particles and their interactions. The three fundamental forces, the electromagnetic, the weak, and the strong force are mediated by gauge bosons, particles with integer spin. Half integer spin particles, the fermions, form matter.

The electromagnetic force is mediated by the massless photon. It describes interactions between particles with an electric charge.

The mediators of the weak force are the $W^\pm$ and $Z^0$ bosons. These are massive particles with a mass of 80.39 GeV and 91.19 GeV respectively. Due to the high mass of the $W^\pm$ and $Z^0$ bosons, the range of the weak force is limited to subatomic distances. The electromagnetic force and the weak force can be described in a unified electroweak theory [12].

The strong force is described by quantum chromodynamics (QCD) [13]. It is mediated by eight gluons. Gluons are massless gauge bosons that carry color charge, which has three different states referred to as red, green, and blue, plus their corresponding anti colors. At short distances, the coupling strength of the strong force decreases, and strongly interacting particles can be considered quasi-free. Due to the nature of the potential of the strong force, the creation of a new quark-antiquark pair gets energetically favorable with increasing distance. The result of this effect, called Confinement is that a color neutral quark antiquark pair is created, if the energy stored in the potential is large enough. Therefore, all stable particles are color neutral.

1.1.1 The Higgs Mechanism

In the following, a brief introduction to the Higgs Mechanism will be given, a detailed description can be found in [14]. In the standard model, the different fields are described by a Lagrangian density $\mathcal{L}$, using the Lagrange formalism. From a postulated Lagrangian density, an equation describing the field dynamics can be derived. By enforcing the Lagrangian to be invariant under local gauge transformations, the gauge bosons and their interactions mentioned above, enter the theory. The corresponding symmetry groups are a $SU(3)$ group for the QCD and a $SU(2) \times U(1)$ group for the electroweak theory.

However, the resulting gauge bosons are required to be massless, which was not observed by the experiment for the $W^\pm$ and $Z^0$ bosons. Therefore, an extension of the SM is required. The Higgs mechanism [15–17], proposed in the 1960s uses spontaneous symmetry breaking
of the SU(2) symmetry in order to generate mass terms for the electroweak gauge bosons and fermions. If a system is invariant under a transformation given by a symmetry group, but the symmetry is broken in its ground state, the symmetry is spontaneously broken. The Higgs Mechanism is included in the SM by adding an additional term to the SM Lagrangian density

$$L_{\text{Higgs}} = \partial_\mu \phi^\dagger \partial^\mu \phi - V(\phi),$$

(1.1)

where

$$V(\phi) = -\mu^2 \phi^\dagger \phi + \lambda \left( \phi^\dagger \phi \right)^2$$

(1.2)

is the Higgs potential and $\phi$ is a new SU(2) doublet field

$$\phi = \begin{pmatrix} \phi_+ \\ \phi_0 \end{pmatrix}$$

(1.3)

where both $\phi_+$ and $\phi_0$ are complex fields. The local gauge invariance can be enforced by replacing the derivative $\partial_\mu$ by the covariant derivative $D_\mu$

$$\partial_\mu \rightarrow D_\mu = \partial_\mu + ig Y_\phi \frac{g'}{2} B_\mu +igt^a W^a_\mu$$

(1.4)

where $B_\mu$ is the field belonging to the U(1) symmetry and $g'$ is the corresponding coupling constant. The $W^a_\mu$ fields belong to the SU(2) symmetry together with the coupling constant $g$. The field $\phi$ can be expanded around its ground state in the minimum of the Higgs potential

$$v = \sqrt{\frac{\mu^2}{2\lambda}}$$

(1.5)

In principle, any form of expansion of the field $\phi$ could be chosen. The common choice in the physical gauge is

$$\phi = \begin{pmatrix} 0 \\ v + \frac{H}{\sqrt{2}} \end{pmatrix},$$

(1.6)

where $H$ is a new scalar field. The kinematic term of the Higgs Lagrangian density can be evaluated using the expansion of $\phi$ in equation 1.6 and the covariant derivative from equation 1.4. This leads to

$$D_\mu \phi^\dagger D^\mu \phi = \frac{1}{2} \partial_\mu H \partial^\mu H + \frac{g^2 + g'^2}{4} \left( v + \frac{H}{\sqrt{2}} \right)^2 Z_\mu Z^\mu + \frac{g^2}{4} \left( v + \frac{H}{\sqrt{2}} \right)^2 W^+_\mu W^-_\mu$$

(1.7)

In this result, mass terms for the $W^\pm$ and $Z^0$ bosons along with their couplings to the scalar field $H$ can be identified. Their masses are

$$m_{W^\pm} = \frac{1}{2} g v \quad \text{and} \quad m_Z = \frac{1}{2} \sqrt{g^2 + g'^2} v$$

(1.8)

The electroweak gauge bosons obtain their masses through their coupling to the non-vanishing vacuum expectation value $v$. Additionally, a new field $H$, the Higgs field can
be identified in the Lagrangian density. The $H$ field obtains its mass via self coupling. The particle belonging to the $H$ field can be identified with the new particle that was discovered by the ATLAS and the CMS Collaborations in 2012, since it has the same properties.

By introducing a coupling between the lepton doublets and the field $\phi$, mass terms for the leptons can be introduced into the theory. For the electron and the electron neutrino, the Lagrangian density for such a Yukawa coupling is

$$\mathcal{L}^{e}_{\text{Yuk}} = -\lambda e \bar{\psi}_L \phi e_R - \lambda e \bar{e}_R \phi^\dagger \psi_L$$

With $\psi_L = (\nu_e, e_L)^T$ and the chosen expansion of $\phi$ (equation 1.6) in the vicinity of the ground state results in

$$\mathcal{L}^{e}_{\text{Yuk}} = -\lambda e \bar{\nu} e - \frac{\lambda e}{\sqrt{2}} H \bar{e} e$$

The mass of the electron is then given by $m_e = \lambda_e v$ and the coupling of the Higgs field to the electron is proportional to the mass of the fermion. For vector bosons, the coupling is proportional to the squared mass of the vector boson.

### 1.1.2 The $H \rightarrow \tau \tau$ Decay Channel

At the LHC [18], the most common Higgs boson production channels are gluon-gluon-fusion and vector-boson-fusion. Their feynman diagrams are shown in Figure 1.1 (left). Depending on the mass of the Higgs boson, the branching fractions of the possible decay products vary, as shown in Figure 1.1 (right). As the $\tau$-lepton and the $b$ quark are the two heaviest fermions that can be directly produced in a Higgs boson decay, the $H \rightarrow \tau \tau$ and $H \rightarrow b\bar{b}$ decay are the most promising channels for measuring and probing the Higgs boson Yukawa coupling to fermions.

As shown in Figure 1.1, the branching ratio of a Higgs boson decaying into a $b$ quark pair is about one order of magnitude higher than the branching ratio of a decay into a $\tau$-lepton pair. However, distinguishing between jets produced through the strong interaction, referred to as QCD multijets and $b$ jets that are produced by decaying $b$ quarks originating from a Higgs boson decay, is difficult. The cross section of QCD jet production is several orders of magnitude higher than the combined cross section for all Higgs production mechanisms resulting $H \rightarrow b\bar{b}$ decays.

Hadronically decaying $\tau$-leptons, denoted as $\tau_h$ in the following, have a different signature than QCD jets. In addition, a $\tau$-lepton can decay into electrons or muons, which both can be reconstructed with high precision. The decay of a Higgs boson into $\tau$-leptons is easier to identify than the decay of a Higgs boson into $b$ quarks. In the following, the different decay modes of $\tau$-leptons will be discussed in detail.

The $\tau$-lepton is the heaviest known lepton, with a mass of 1.777 GeV [20]. Apart from its leptonic decay modes, the $\tau$-lepton can also decay hadronically. In the leptonic case, the $\tau$-lepton decays into two neutrinos and a lighter lepton, an electron or a muon:

$$\text{leptonic} \quad \tau^- \rightarrow \ell^- \bar{\nu}_\ell \nu_\ell$$
1.1 The Standard Model of Particle Physics

Figure 1.1: On the left, the two leading production mechanisms for Higgs bosons at the LHC are displayed. On the right, the branching ratios of the different standard model Higgs boson decay modes, depending on the Higgs boson mass, are shown [19].

Out of all $\tau$-leptons 35.2% decay leptonically. The remaining 64.8% of $\tau$-leptons decay hadronically, typically forming a narrow $\tau$-jet. The hadronic decays can be further categorized, depending on the number of charged and neutral hadrons in the final state. The $\tau$-lepton decay into one charged hadron is referred to as "one-prong decay". The decay into three charged particles is referred to as "three-prong decay". In either case, additional $\pi^0$ candidates, which further decay into pairs of photons, may be present in the decay:

\begin{align*}
\text{One prong} & \quad \tau^- \rightarrow \pi^- \nu_\tau \\
\text{One prong + } \pi^0 & \quad \tau^- \rightarrow \pi^- \nu_\tau \pi^0 \\
\text{One prong + 2 } \pi^0 & \quad \tau^- \rightarrow \pi^- \nu_\tau \pi^0 \pi^0 \\
\text{Three prong} & \quad \tau^- \rightarrow h^+ h^- h^- \nu_\tau
\end{align*}

where $h$ refers to lighter hadron that can be produced like a pion or a kaon.

In a $H \rightarrow \tau\tau$ decay, the $\tau$-leptons appear in pairs. The different decay modes of a di-$\tau$ pair and the corresponding branching fractions are listed in Table 1.1. Usually, only the top four final states in Table 1.1 $e\mu$, $e\tau_h$, $\mu\tau_h$, and $\tau_h\tau_h$ are used in physics analyses. The remaining final states $ee$ and $\mu\mu$ only have a very small branching fraction. In addition, the decay of a Z boson into two leptons ($Z \rightarrow ll$) is an irreducible background with a much higher yield than the Higgs signal in these two channels.

At the CMS experiment, the $e\tau_h$ and $\mu\tau_h$ channels are the most accessible ones, because one of the decay products is an isolated electron or muon. This requirement suppresses a lot of background from QCD multijet events. The full hadronic channel $\tau_h\tau_h$ is more challenging, but also has the highest branching fraction.
Table 1.1: Different decay modes of $\tau$-lepton pairs, where one $\tau$-lepton decays into the first particle, and the other $\tau$-lepton into the second particle [20].

<table>
<thead>
<tr>
<th>Decay Mode</th>
<th>Branching Fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_h\tau_h$</td>
<td>42 %</td>
</tr>
<tr>
<td>$\mu\tau_h$</td>
<td>23.1 %</td>
</tr>
<tr>
<td>$e\tau_h$</td>
<td>22.5 %</td>
</tr>
<tr>
<td>$e\mu$</td>
<td>6.2 %</td>
</tr>
<tr>
<td>$ee$</td>
<td>3 %</td>
</tr>
<tr>
<td>$\mu\mu$</td>
<td>3 %</td>
</tr>
</tbody>
</table>

Processes, where two genuine $\tau$-lepton decays are present, and processes, where either a lepton or a jet are misidentified as a $\tau$-lepton decay are background processes for a H $\rightarrow$ $\tau\tau$ analysis. The largest background, coming from Z $\rightarrow$ $\tau\tau$ decays, can be estimated by using the embedding technique, which will be introduced in chapter 3. Additionally, top quark pair production ($t\bar{t}$) and vector boson pair production (diboson) can also result in a $\tau\tau$ final state. QCD multijets, the production of W bosons with associated jets ($W+$jets), and Z $\rightarrow$ $ll$ events are processes that can be misidentified as a $\tau\tau$ final state, if one or more objects in the detector are misidentified as a $\tau_h$. 


The LHC is a particle accelerator at the European Organization for Nuclear Research (CERN) close to Geneva, Switzerland. With a total length of nearly 27 km and a design center-of-mass energy of 14 TeV it is the biggest particle accelerator ever build. Between 2010 and 2018 the accelerator was operated at 7, 8 and 13 TeV center-of-mass energy, and was used to accelerate both protons and ion nuclei. Four major experiments located around the ring are able to study the collisions generated by the LHC. The four experiments are A Large Ion Collider Experiment (ALICE), LHC-beauty (LHCb), A Toroidal LHC ApparatuS (ATLAS), and the Compact Muon Solenoid (CMS) experiment.

In this chapter, the CMS detector and its different components will be discussed. In addition, the triggering system of CMS and the reconstruction of measured collisions with a focus on leptons will be discussed.

2.1 The CMS Detector

The CMS detector was built to study a wide range of different processes, like searches for the SM Higgs boson, precision measurements of SM parameters, dark matter searches or searches for the evidence of physics beyond the standard model. This is reflected in the design and structure of the detector.

The detector consists of multiple layers around one of the interaction points of the LHC. At the interaction point, the two beam lines of the accelerator cross each other to cause particle collisions. A sketch of the CMS detector design is shown in Figure 2.1. A silicon tracker, an electromagnetic (ECAL), and the majority of the hadronic (HCAL) calorimeter are located inside the super conduction solenoid of the detector. Outside of the magnet, muon chambers are included within the return yoke of the magnet system. The whole detector is split up into multiple regions, the cylindrical barrel region, which is located around the interaction point, and the endcap regions, which are two disk shaped regions to cover the forward region. As shown in Figure 2.2, the barrel components are horizontal in $z$ direction while the endcap components are vertical in $z$ direction.

The inner tracker [21] is the component closest to the interaction point. It is wrapped around the interaction point like a barrel. The inner tracker is used to detect the starting point and trajectory of charged particles originating from the particle collision. Since the
inner tracker is placed inside a magnetic solenoid, charged particles are forced on a helix shaped trajectory allowing the measurement of the transverse particle momentum ($p_T$). The inner tracker consists of semi conducting particle detectors made from silicon. The innermost part is made up of three layers of pixel detectors, followed by 10 layers of strip detectors. In total, the inner tracker covers a pseudorapidity region of $|\eta| \leq 2.5$ with a regional resolution of up to $10 \mu m$. The total sensor surface is about $205 m^2$ and has more than 76 million readout channels.

The electromagnetic calorimeter (ECAL) [23] is located around the inner tracker. Its purpose is to measure the energy of electrons and photons. It is built from lead tungstate (PbWO$_4$) crystals, a transparent crystal with a density of $\rho = 8.28 g cm^{-1}$. The crystal scintillates when excited by traversing charged particles, which allows a precise measurement of electromagnetic showers originating from electrons. In both, the barrel and the endcap region, the total length of the calorimeter corresponds to $\approx 25$ radiation lengths ($X_0$), which is enough to fully absorb the energy of electrons and photons. The ECAL barrel (EB) covers a region up to $|\eta| < 1.479$ and the ECAL endcap (EE) a region of $1.479 < |\eta| < 3.0$. In addition to EB and EE, a preshower detector, which is located in the endcap region, is also part of the ECAL. The preshower detector is made up of silicon strips and used to distinguish between single high energy photons and pairs of low energy photons.
2.1 The CMS Detector

Figure 2.2: A cross section of one quadrant of the CMS detector in the $R - z$ plane \[26\]. A focus is set on the muon system of the detector. In the barrel region, drift tubes (shown in brown) are used to detect muons, while in the endcap region cathode strip chamber modules (shown in green) are applied. Resistive plate chambers (shown in blue) are installed in both regions.

The hadronic calorimeter (HCAL) \[24\] is located outside of the ECAL. It is used to measure the energy of strongly interacting particles, mostly protons, neutrons, or pions. The hadronic calorimeter is a sampling calorimeter made from brass plates as absorber material and a plastic scintillator as sensitive material. This type of calorimeter is used in both the barrel region, which covers a region of $|\eta| < 1.3$ and the endcap region, covering a region between $1.3 < |\eta| < 3.0$. Additionally, a forward detector is installed further down the beam line to cover the region of $3.0 < |\eta| < 5.0$. The forward detector is made of glass fibers integrated into steel to account for the increased radiation exposure in this region. Finally, the hadron outer detector (HO) is located outside the solenoid, to increase the energy measurement precision of the HCAL, as high energy hadrons are able to pierce the HCAL and the solenoid. These particles are then detected in the hadron outer detector.

The superconducting solenoid of the CMS detector \[25\] provides a homogenous magnetic field of 3.8 T. The direction of the magnetic field is parallel to the beam-pipe. The return yoke of the solenoid is located outside the magnet and houses the muon system.

The most outer part of the CMS detector is the muon system \[27\]. It is made up from different types of independent gaseous detectors, referred to as muon chambers, in order to measure and reconstruct tracks of muons. Drift tube (DT) chambers are used in the...
barrel region up to $|\eta| > 1.2$, resistive plate chambers (RPC) are applied in the barrel and the endcap region up to $|\eta| > 1.8$. The endcap region from $0.9 < |\eta| < 2.4$ is also instrumented using cathode strip chambers (CSC). The muon system is shown in Figure 2.2. Since it is located in the return yoke of the magnetic field of the inner detector, a magnetic field is also present in the muon system, which helps to measure the muon momentum more precisely.

### 2.2 The CMS Trigger System

The LHC was designed to operate at a luminosity of $1 \times 10^{34}$ cm$^{-2}$ s$^{-1}$ and a beam crossing frequency of 40 MHz. During the second measurement period of the LHC (Run 2), from 2015 to 2018, the luminosity exceeded the design luminosity. The number of additional proton-proton collisions (pp collisions) besides the hard process in one bunch crossing, (pileup), exceeded 50 collisions. Each collision, also called event, uses about 1 MB of uncompressed disk space. The immense amount of events needs to be greatly reduced in order to be able to handle the storage and processing of the measured data. In addition, most of the collisions do not contain hard processes of interest for the CMS collaboration. Therefore, the trigger system of CMS \[28–31\] is implemented to achieve data reduction without losing relevant collisions.

The CMS trigger system consists of two levels as shown in Figure 2.3. The first level, L1, is made up of field programmable gate arrays, custom hardware within the detector itself. They are used to perform a basic analysis of the event in order to determine, whether it is of potential interest. This interest can arise, if typical signatures for muons, electrons, photons, jets, hadronically decaying $\tau$-leptons, or large amounts of missing energy are detected. The L1 system uses inputs from the electromagnetic calorimeter, the hadronic calorimeter, and the muon system. Within a few microseconds, the information from the different sub detectors is combined. Then, a global decision whether an event is of possible interest, based on the measured energy deposits, is made. Meanwhile, the data associated to the event is stored in a buffer. In order to monitor the overall performance of the detector and the trigger system, a small amount of rejected events is also passed on. The L1 trigger reduces the event rate from 40 MHz to about 100 kHz. Potential particle candidates identified in the L1 step, referred to as L1 seeds, are passed to the next step as starting points for further analysis.

The second level, the High-Level Trigger (HLT), is a software based trigger solution. Its purpose is to further reduce the event rate from 100 kHz to about 1 kHz. During the HLT step, a down-sized version of the full reconstruction software of CMS is used, in order to ensure a faster computation time in comparison to the full event reconstruction. The main differences are a simplified track reconstruction, and the limitation on a regional particle reconstruction, where only the area around a L1 seed is analyzed. The average run time of the HLT reconstruction lies in the order of 100 ms. The HLT step is calculated on a cluster of commercial computers, with a total of 26000 CPU cores. After a successful HLT reconstruction, a trigger bit to flag the event is set, and the event is stored at the computing site at CERN for the full reconstruction.
2.2 The CMS Trigger System

Figure 2.3: The trigger setup of the CMS experiment [32]. It consists of two levels, the L1 level implemented in custom hardware, and the HLT that is running on a computer farm near the detector. The buffers in between hold the collision data long enough to perform the needed reconstruction steps.

2.2.1 High-Level Trigger Hadronic Tau Reconstruction

The CMS trigger configuration includes many different configurations in order to identify different particles. These configurations are referred to as trigger paths. Several trigger paths are sensitive to \( \tau_h \). The reconstruction and identification of \( \tau_h \)-candidates in the trigger system [33] is split up into four distinct steps. First, the L1 \( \tau_h \) seed is calculated, then, during the HLT reconstruction, three additional levels of reconstruction are performed. The algorithms chosen for the L1 and HLT reconstruction are focused on fast and efficient processing, rather than on precision. The detailed requirements of a trigger path may vary and depend on the chosen configuration, however the reconstruction procedure is identical for all trigger paths sensitive to \( \tau_h \). This procedure will be described in the following and is illustrated in Figure 2.4.

A \( \tau_h \)-candidate produces a \( \tau_h \)-jet in the calorimeter. A jet consists of multiple particles that can be found within a narrow cone. Typically, they originate from hadronizing quarks or gluons, containing many different particles. A \( \tau_h \)-jet only contains few charged and neutral hadrons. Additionally, if the initial \( \tau \)-lepton has a high \( p_T \), the resulting \( \tau_h \)-jet will be very narrow; about 90% of the \( \tau_h \)-jet transverse energy \( E_T \) will be located in a small cone of \( \Delta R < 0.2 \) around the \( \tau_h \)-jet axis. These features are used to identify a \( \tau_h \)-candidate and distinguish it from quark or gluon jets.

In the L1 step, energy deposits in the electromagnetic and hadronic calorimeter are organized into trigger towers (TT). The TT form a grid in the \( \eta - \phi \) plane splitting the calorimeters into small rectangular sections. If a trigger tower exceeds a transverse energy of \( E_T > 2 \text{ GeV} \), it will be used as a seed for finding an L1 \( \tau_h \)-candidate. Around a seeding trigger tower, additional TTs surrounding the seed are added to form a cluster,
The CMS Experiment

if they also contain a sufficient amount of energy. Several clusters may be merged into one seed, as the $\tau_h$ decay products may be more spread out, due to the fact that the trajectories of charged particles are bent in the presence of a magnetic field, while the trajectories of neutral decay products are unaffected. The energy of the L1 $\tau_h$ seed is defined as the sum of the energies in all clustered trigger towers. Finally, the isolation of the candidates is checked, by comparing the energy of the cluster with the energy in a $(6 \times 9)$ grid in the $\eta - \phi$ plane around the seeding TT as sketched in Figure 2.4 (top left). A more detailed description of the L1 $\tau_h$-candidate reconstruction algorithm can be found in [34]. Clusters, which pass this procedure, are used as L1 $\tau_h$ seeds in the HLT sequence.

The HLT sequence for identifying $\tau_h$-candidates consists of multiple steps, in order to successively reduce the amount of data that has to be processed. The next step in the sequence, L2, is the formation of calorimeter jets using the L1 seeds from the previous step. The calorimeter jets are constructed using the anti-$k_t$ algorithm [35] with a radius parameter of $R = 0.2$ around the L1 $\tau_h$ seed. Selection requirements on the $p_T$ and $\eta$ of the calorimeter jet are applied.

![Diagram of CMS trigger sequence](image)

**Figure 2.4:** Illustration the trigger reconstruction and identification of $\tau_h$-candidates. The different steps are explained in the text. In the top left, a sketch of the clustering in the L1 step is shown. In the top right, a sketch of a signal cone around a leading charged hadron candidate, and an isolation cone around the signal cone are shown. This procedure is used in the L2.5 and the L3 step to identify $\tau_h$-jets.

14
In the second HLT step L2.5, tracks in a $\eta - \phi$ region of $\Delta \eta \times \Delta \phi = 0.5 \times 0.5$ around the L2 calorimeter jets are reconstructed. The process of track reconstruction is explained in section 2.3.1. Based on the reconstructed tracks, a primary vertex (PV) is reconstructed, using an algorithm to determine the origin of multiple tracks [36]. The reconstructed tracks are matched to the L2 $\tau_h$-candidate if they are located in the vicinity of the L2 jet and can be associated to the PV. The isolation of the $\tau_h$-candidate is checked, using the tracks originating from the PV. Since $\tau_h$-jets are narrow jets, there should be little to no energy found around a signal cone. The $\tau_h$-candidate is considered isolated, if the sum of track $p_T$ around the candidate in a isolation cone of $0.15 < \Delta R < 0.4$ around the L2 seed is less than a certain threshold. Only tracks, originating from the PV are considered. This process is illustrated in Figure 2.4 (top right).

In the L3 step, the final HLT step, a regional reconstruction based on the particle flow algorithm (PF) is performed. The region is set around the L2.5 $\tau_h$-jet from the previous step. Like during the L2.5 step, a signal cone with a size between 0.08 and 0.12 in $\Delta R$, depending on the $p_T$ of the jet is formed around the leading charged hadron. In addition an isolation cone of $\Delta R < 0.4$ around this signal cone is defined. The $\tau_h$ candidate is build by considering all neutral hadrons in the signal cone as $\pi^0$ candidates, the leading charged hadron, and up to three additional charged hadrons. This way, all important hadronic decay modes of the $\tau$-lepton are considered. The vertex closest to the track of the leading charged hadron in z-direction is considered the PV. Depending on the $p_T$ of the additional charged particles in the isolation cone, three different isolation working points are defined. If a $\tau_h$ candidate passes the quality requirements imposed in this step, a flag, signaling that the trigger path for this particle was passed, is set. This information can be used in a later analysis.

2.3 Reconstruction

In order to reconstruct the hard scattering process, all detector information available should be used as efficiently as possible. Depending on the type of stable particle, different signatures can be found in the detector. This is illustrated in Figure 2.5.

- **Muons** are minimum ionizing particles. They deposit only small amounts of energy when traversing the different sub detectors. Due to their charge, the measured trajectories of muons in the inner tracker and the muon system are bent. In addition, small amounts of energy in both the ECAL and HCAL are deposited.

- **Electrons** are stopped in the ECAL, causing an electromagnetic shower. In addition, electrons are identified by their curved track in the inner tracker. Due to the low mass, a significant amount of the energy loss of electrons is due to radiation of bremsstrahlung photons rather than ionization effects. As a result, the trajectory of electrons in the inner tracker can change significantly.

- **Charged hadrons** are stopped in the HCAL. They are identified using their energy deposits both in the ECAL and HCAL. In addition, their trajectory is measured in the inner tracker.
Neutral hadrons are also stopped in the HCAL. Energy deposits of neutral hadrons can be found in the ECAL and HCAL, however, since their electric charge is zero, no energy deposits and therefore no trajectory is present in the inner tracker.

Photons result in an electromagnetic shower in the ECAL. No trajectory can be measured in the inner tracker. However, the conversion of a photon into an $e^+e^-$ pair may occur in the inner tracker.

In the reconstruction and identification step, the measured data are processed using numerous algorithms, utilizing the specific behaviors of the different particles in order to gain knowledge of the collision.

2.3.1 Track and Vertex Reconstruction

Track reconstruction is a very important part of the reconstruction, since multiple pp collisions take place during a single bunch crossing. More than 1000 particles may traverse the silicon tracker during a single bunch crossing. Therefore, a high track-finding efficiency whilst having a low number of wrongly reconstructed tracks (ghost tracks) is a challenging task. In addition, the primary vertex, the origin of the signal process, is be determined from tracks.
The basic idea of track reconstruction is to combine individual measurements of energy deposits in the different tracker cells, referred to as hits, into tracks. Tracks describe the trajectory of a charged particle. The location, where the two proton beams collide, the beamspot, is used as an reference point during the reconstruction. In the following, the track and vertex reconstruction of the CMS experiment are discussed in more detail. A complete description of the procedure can be found in [38], and a schematic cross section of the tracker is shown in Figure 2.6.

The CMS collaboration uses an iterative algorithm called iterative tracking. During each iteration, tracks are reconstructed using all tracker hits available to the algorithm. Reconstructed tracks are then saved and removed from the collection of remaining tracker hits. This reduces the complexity of the possible hit combinations and allows the reconstruction of less obvious tracks in the following iteration. During each iteration, four different steps are processed:

1. Track seeds are build by combining two to three tracker hits. The rough estimate of the track parameters provided by the track seeds is used in the second step. Track parameters are the location, the momentum, the charge and the direction of a potential particle trajectory.

2. A Kalman filter algorithm [39] is used to find additional hits. Based on the track seed, the particle trajectory is extrapolated. The predicted trajectory follows a perfect helix within a uniform magnetic field. If, within the prediction, a suitable tracker hit is found in a successive detector layer, the found tracker hit is added to the emerging track collection. The track parameters are then updated, taking the location of the additional hit into account. This process is repeated multiple times.

3. The found tracks are refitted and smoothed by a dedicated algorithm.

4. A set of selection requirements that determine the quality of a track like the number of tracker hits or the track fit quality are tested for a reconstructed track. The track is discarded if the requirements are not met. If the track passes the requirements, all hits assigned to the track are removed from the collection of hits. After all potential tracks are tested, the four steps are repeated.

After all tracks in the event are found, the PV of the event is reconstructed. The PV is essential to distinguish between the collision of interest and pileup. Tracks are extrapolated back to the pp interaction, in order to locate their origin. This is done in three steps:

1. Tracks with an extrapolated origin in the vicinity of the beamspot are selected.

2. The selected tracks are clustered based on their distance to the beamspot in $z$ direction. For each cluster, a rough estimate of a vertex position is calculated.

3. The exact vertex position is determined using a dedicated fitting algorithm. All tracks within a cluster are taken into account.
For each vertex, a probability can be calculated whether the origin of a reconstructed track is actually that vertex. During Run II the geometrical resolution of vertices with a sum of track $p_T$ higher than 100 GeV was $14 \mu m$ in $x$, $y$ and $19 \mu m$ in $z$ direction [40].

2.3.2 Particle Flow

In order to reconstruct and identify all particles, like electrons, muons, photons, charged, and neutral hadrons in an event, the CMS collaboration uses an Particle Flow algorithm [37, 41] that combines the information of the different CMS subdetectors listed in section 2.2 to achieve the best possible particle identification and energy measurement. This results in a list of all particles in an event, which are then analyzed further to identify higher level objects that contain multiple particles like jets or $\tau$-leptons. In order to perform a global event description, a finely granulated segmentation in each subdetector is necessary. The PF algorithm will be explained in the following.

In the first step of the PF reconstruction, the energy deposits in each subdetector are reconstructed individually. This includes the reconstruction of tracks as described in section 2.3.1, reconstruction of tracks in the muon system, and reconstruction of the calorimeter clusters in the ECAL and HCAL. The resulting objects are the basic building blocks of the PF reconstruction and are referred to as PF elements.

Since most particles deposit energy in multiple subdetectors, multiple PF elements can be associated to a single particle. Therefore, PF elements are linked using a linking algorithm. A link between two PF elements is established, if certain requirements are met, depending on the type of PF elements to be linked. As an example a track from [Image]
the inner tracker and a calorimeter cluster are linked, if the track position extrapolated to the calorimeter surface lies within the area of the calorimeter cluster.

Multiple linked PF elements form an entity called PF blocks. In order to reduce the computational complexity of linking all PF elements, particles are identified in a given order. All PF elements belonging to a found PF block are removed from the event to speed up the linking of the remaining PF elements. At first, muons, electrons, and isolated photons are reconstructed. The remaining PF elements are identified as hadrons. ECAL cluster without an associated track are considered a photon and HCAL cluster without an associated track are considered neutral hadrons. Finally, a post-processing step is applied, to reduce the misidentification which can occur due to the removal of PF elements.

2.3.3 Lepton Reconstruction and Identification

Since leptons are of special interest in this thesis, a more detailed description of muon, electron and \( \tau \)-lepton reconstruction and their identification is given in this section.

**Muon Reconstruction and Identification**

A detailed description of muon reconstruction and identification is given in [42]. Muon tracks are reconstructed independently in the inner tracker and in the muon system, using the track reconstruction described in section 2.3.1. By combining the track information of the two subdetectors, three types of muon tracks can be identified:

- **Standalone-muon** tracks are tracks found in the muon system. Without a matching track in the inner tracker, standalone-muon tracks have a high probability to originate from external sources, like cosmic rays. They are discarded in most analyses.

- **Tracker muon** tracks are constructed by extrapolating tracks from the inner tracker to the muon system. If at least one matching muon segment is found, the track is considered a tracker muon track. These tracks have a high efficiency in regions with less active detector material and in the low-\( p_T \) region, as only one matching hit in the muon system is required. However, tracker muons can originate from misidentified remnants of hadronic showers, as these remnants can reach the first layer of muon chambers outside the solenoid.

- **Global muon** tracks are constructed by matching standalone-muon tracks from the muon system with tracks found in the inner tracker. This approach is called outside-in strategy. If a match is found, the combined track is refitted to optimize the track parameters using the information of both subdetectors.

Tracker muon tracks and global muon tracks are combined into a single track candidate, if they share the same inner track. At the CMS experiment, the reconstruction efficiency of muon tracks is about 99%. The resulting tracks are passed to the PF algorithm.

In order to identify a reconstructed muon, an algorithm combining various variables from the reconstruction step, like the \( \chi^2 \) value of the track fit, is applied. In an analysis,
an appropriate working point can be chosen, which corresponds to a defined muon identification efficiency.

**Electron Reconstruction and Identification**

A detailed description of the electron reconstruction and identification is given in [43]. Most of the electron energy is deposited in the ECAL. The resulting energy deposits can be concentrated on a small geometric area, if little bremsstrahlung occurs in the tracker. However, the energy deposits can also be spread out in a larger region, if most of the electron energy is converted into bremsstrahlung photons before the electron reaches the ECAL. These radiated photons mainly spread in the \( \phi \) direction, as they are emitted tangential to the electron track. The photons convert back into a pair of electrons, resulting in additional electromagnetic showers in the ECAL. A dedicated algorithm is used to match multiple ECAL clusters together, forming a Supercluster (SC).

Bremsstrahlung also affects the reconstruction of electron tracks. The high radiative energy loss causes big changes in the trajectory curvature and therefore, the default Kalman filter approach has a reduced hit finding efficiency. A special Gaussian-sum filter (GSF) algorithm, an extension of the Kalman filter, is applied for the electron track reconstruction [44]. The algorithm is based on the assumption that the bremsstrahlung energy loss of electrons can be described as a Gaussian mixture. Therefore, all tracker hits used for the electron track reconstruction are a weighted mixture of multiple tracker hits rather than just a tracker hit per layer.

The GSF algorithm is more computationally intensive than the regular tracking, so track seeds must fulfill special selection criteria for the GSF algorithm to be applied. There are two types of GSF seeds:

- **Tracker-driven** seeds are tracks, that were reconstructed using the Kalman filter. If the track extrapolation onto the ECAL surface matches the location and energy of a SC within certain selection criteria, the track is used as a seed for the GSF algorithm.

- **ECAL-driven** seeds are identified by extrapolating a trajectory from a reconstructed SC to the interaction point using an outside-in approach. If matching trajectory seed is found, it is used as a GSF algorithm seed.

Often, electrons have both tracker-driven and ECAL-driven seeds.

Electrons are reconstructed by associating superclusters with GSF or regular tracks. The extrapolation of a track onto the ECAL surface has to match the position of a SC. Based on the assumption, that a bremsstrahlungs photon is radiated tangential to the electron track at every tracker layer, additional PF calorimeter cluster are associated to the electron track.

A multi variant (MVA) based algorithm, combining several quantities sensitive to the signature of electrons is used to identify the reconstructed electrons. The input variables of the discriminator can be grouped into three different categories.
• Track related variables – These are variables related to the reconstruction of the electron track, e.g. the number of tracker hits or the track fit quality for both the regular kalman filter track and the specialized GSF track.

• Calorimeter related variables – These variables quantify the shape of the electromagnetic shower in the ECAL. Variables used are for example the size of the SC in $\eta$ and $\phi$ or the ratio of energy deposits in the HCAL over the ECAL. Here, the fact that electromagnetic showers are narrower than hadronic showers, and that only little electron energy is deposited in the HCAL are used to distinguish between electrons and misidentified electrons, like jets with large electromagnetic components.

• Matching related variables – These are variables that determine the quality of the track and calorimeter combination. These include variables sensitive to the geometric positioning of the electron track and the SC, of the compatibility between the momentum measurement in the tracker and the energy measurement in the calorimeter.

The resulting discriminator has several working points, which quantify the efficiency and the purity of the electrons in a sample, when applying the discriminator.

**Tau Reconstruction and Identification**

As explained in section 1.1.2, the $\tau$-lepton can decay both hadronically and leptonically. Leptonically decaying $\tau$-leptons are identified by reconstructing the resulting electrons or muons. Hadronically decaying $\tau$-leptons are reconstructed by using a hadron-plus-strip algorithm. A detailed description can be found in [33].

The reconstruction of $\tau_h$-jets formed by the decay products of a $\tau_h$-candidate is seeded by reconstructed jets. The relevant $\tau_h$ decay modes are listed in section 1.1.2. Photons originating from the $\pi^0$ decay have a high probability of converting into $e^+e^-$ pairs while traversing the inner tracker. To reconstruct the $\pi^0$, the resulting photons and electrons are clustered into strips in the $\eta$-$\phi$ plane. The actual size of the strip varies, as the decay products of a $\tau_h$-candidate with a high $p_T$ will be boosted in the direction of the $\tau_h$-candidate, resulting in a smaller strip. The strip size is limited to 0.3 in $\phi$ and 0.15 in $\eta$ while having a minimal size of 0.05 in both directions. By combining the resulting strips with reconstructed charged hadrons in the seeded jet, the decay mode of the $\tau_h$ can be determined.

After the reconstruction, multiple discriminators are used to separate $\tau_h$-candidates from quark and gluon jets, electrons, and muons.

The discrimination between quark or gluon jets and $\tau_h$-jets is done using an MVA-based algorithm [45]. As mentioned in section 2.2.1, jets from $\tau_h$ and quark- or gluon jets have several differences:

• A $\tau_h$-jet contains fewer hadrons.
Due to the $\tau$-lepton lifetime of $290.31 \times 10^{-15}$ s [20] the $\tau$-lepton can propagate a measurable distance before decaying. As a result, the constituents of a $\tau_h$ jet originate from a secondary vertex slightly displaced from the PV of the event, which is not the case for quark or gluon jets.

In the vicinity of the $\tau_h$-jet less energy can be found, as $\tau_h$-jets are collimated and isolated.

The discriminating algorithm utilizes variables sensitive to these differences. As electrons emit bremsstrahlung, their signature can be identified as a $\pi^0$ decay, and isolated electron may be misidentified as a $\tau^- \to \pi^- \nu_\tau \pi^0$ decay. A discriminator, taking into account information from the ECAL, the HCAL and the tracker, is used to reduce the $e \to \tau_h$ misidentification. To discriminate against $\mu \to \tau_h$ misidentification, a $\tau_h$-candidate is vetoed, if a signal in the muon system is found in the direction of the $\tau_h$-candidate.
The Embedding Technique

A common way of estimating background processes for an analysis is using the Monte Carlo (MC) simulation. The simulation has to provide an accurate description of the observed data, in order to allow for a meaningful comparison. However, the simulation of challenging processes is not trivial and, in general, simulated events need to be carefully tuned to the data in control regions before providing a good estimate. In addition, the simulation process is computationally expensive, especially when simulating the complete detector response in the presence of up to 50 or more pileup events. Current prognosis for the High Luminosity LHC (HL-LHC), an upgraded version of the current LHC, predict an average of 120-200 pileup collisions per bunch crossing, which will further increase the computational effort needed to produce appropriate pileup modeling.

The embedding technique [9] is a method that helps to avoid such problems. By using a data driven approach, its major purpose is to provide a better description of $\tau\tau$ events resulting from the Drell-Yan process [46] than Monte Carlo simulation. The idea is to use a recorded event and embed the decay of two $\tau$-leptons into the event, without modifying the rest of the event. In the following $\mu \rightarrow \tau$ embedding, refers to the selection of events with two muons recorded pp collisions and replacing them with simulated $\tau$-lepton decays. The natural target of this selection are $Z \rightarrow \mu\mu$ events. Because of lepton universality, the probability of a $Z$ boson to decay into a $\mu\mu$ pair is equal to the probability of a decay into a $\tau\tau$ pair. The muon pair is required to pass a set of selection criteria aimed at selecting $Z \rightarrow \mu\mu$ events. All energy deposits of the two selected muons are removed from the event, resulting in a cleaned event. Based on the kinematic properties of the two muons, the decay of two $\tau$-leptons is simulated. In the last step, the simulated $\tau$-lepton decays are merged back into the cleaned event, creating a hybrid event, where everything but the $\tau$-lepton decays is taken directly from data. This ensures less need for tuning and a better description of pileup events, the underlying event, and production of additional jets. In addition, many corrections and corresponding uncertainties necessary for the Monte Carlo simulation are obsolete for embedded event samples. In this chapter, the embedding technique and the production of embedded event samples will be discussed in detail.
The Embedding Technique

3.1 Embedding Strategy

The embedding technique is split into four distinct steps that are processed one by one on an event-by-event basis. A sketch of these steps is shown in Figure 3.1. In this chapter the different steps will be explained in detail. In the following, embedded events will refer to events which were produced by applying the embedding technique on either a simulated sample or measured data. Within this thesis, the data sample of pp collisions measured in 2017 by the CMS experiment will be used to demonstrate the technique, if not noted otherwise.

3.1.1 Selection Step

The selection step is the first step in the embedding procedure. The goal of the chosen selection, listed in Table 3.1, is to select as many \( \mu\mu \) events as possible, whilst keeping a high purity of \( Z \to \mu\mu \) events. If more than one suitable \( \mu\mu \) pair can be found in the event, the pair with dimuon pass \( m_{\mu\mu} \) closest to the \( Z \) boson mass is chosen. Due to the
Table 3.1: List of the selection criteria applied to pp collision data in 2017 in order to select $\mu\mu$ events suitable for the embedding technique.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Selection Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trigger</td>
<td>Trigger path sensitive to two muons with $p_{T}^{\mu,1} &gt; 17$ GeV, $p_{T}^{\mu,2} &gt; 8$ GeV</td>
</tr>
<tr>
<td>Leading Muon $p_{T}$</td>
<td>&gt; 17 GeV</td>
</tr>
<tr>
<td>2nd leading Muon $p_{T}$</td>
<td>&gt; 8 GeV</td>
</tr>
<tr>
<td>Muon $</td>
<td>\eta</td>
</tr>
<tr>
<td>Muon Track Type</td>
<td>Global Muon track</td>
</tr>
<tr>
<td>Di-muon Mass $m_{\mu\mu}$</td>
<td>&gt; 20 GeV</td>
</tr>
<tr>
<td>Di-muon Charge</td>
<td>0</td>
</tr>
</tbody>
</table>

loose selection requirements, additional processes besides $Z \rightarrow \mu\mu$ have, although small, contributions to the resulting event sample.

The expected contributions of different processes to this sample were checked by applying the selection on simulated events. The result is shown in Figure 3.2. As expected, the resulting event sample is dominated by $Z \rightarrow \mu\mu$ events, making up about 97% of the entire selected event sample. In the lower tail of the $m_{\mu\mu}$ spectrum, additional contributions originate from QCD multijet (0.84%) and $Z \rightarrow \tau\tau$ events (0.74%). However, because of the low $p_{T}$ of the two muons in this region, these events only have a limited probability of passing the selection criteria applied during the generator step. This will be explained in more detail in section 3.1.3. Therefore, the contributions from QCD multijet and $Z \rightarrow \tau\tau$ events will contribute only marginally to the final $\mu \rightarrow \tau$ embedded event samples.

A contribution of 0.78% comes from $t\bar{t}$ production, being most apparent in the higher mass region of $m_{\mu\mu}$. A top quark pair decays into a $\mu\mu$ final state if both W bosons originating from the t quarks decay into a muon plus a muon neutrino. Due to lepton universality, the decay probability of $t\bar{t} \rightarrow \mu\mu X$ is identical to the decay probability of $t\bar{t} \rightarrow \tau\tau X$. Therefore, as studied in [47], instead of further suppressing the background from $t\bar{t}$ production, $t\bar{t} \rightarrow \mu\mu X$ events are subsumed into the embedded event samples. In a later analysis, that uses embedded events, referred to as target analysis in the following, the fact that $t\bar{t} \rightarrow \tau\tau X$ events are included in a $\mu \rightarrow \tau$ embedded event sample has to be considered. To prevent double counting, $t\bar{t} \rightarrow \tau\tau X$ events have to be removed from the remaining part of simulated $t\bar{t}$ events in the target analysis.

The application of the selection criteria listed in Table 3.1 introduces a selection bias on the resulting event sample. This bias can be corrected for by applying correction factors binned in the $p_{T}$ and $\eta$ of the selected muons. The correction factors are applied for the trigger selection, the reconstruction, and the identification efficiency of the $\mu\mu$ pair. After applying these correction factors, the number of events in the embedded event sample is an estimate of the number of expected $\tau\tau$ events in data, without the need for further corrections.
The Embedding Technique

Figure 3.2: Distribution of $m_{\mu\mu}$ after applying the selection criteria listed in table 3.1 on simulated events. The plots show how different processes are expected to contribute to the event selection suitable for embedding.

3.1.2 Cleaning Step

During the cleaning step, all energy deposits of the two selected muons are removed from the event. As mentioned in section 2.3, muons may deposit energy in any subdetector of the CMS experiment.

- Hits in the inner tracker associated to the global muon track of the selected muons are removed from the collections of tracker hits.

- Calorimeter entries associated with the muons are identified by checking all calorimeter cells crossed by the extrapolated trajectory of the muon tracks. If a calorimeter cell is crossed, the energy in this cell is set to zero. This is done in both the ECAL and HCAL.

- Hits in the muon chambers associated to the global muon track of the selected muons are removed from the collections of muon chamber hits.

An example of the cleaning step is shown in Figure 3.3. On the left, the energy deposits in the calorimeters are shown prior to the cleaning step. On the right, the muon track and the calorimeter energy deposits crossed by the extrapolated muon trajectory are removed from the event.

The cleaning of both the ECAL and HCAL is not always optimal and energy remnants of the muon may remain in the detector. As discussed in section 2.3.2, during the particle flow reconstruction, additional ECAL energy deposits without an associated particle flow candidate are reconstructed as photon candidates, while additional energy deposits in the HCAL are reconstructed as neutral hadron candidates. In addition, the complete
3.1 Embedding Strategy

Figure 3.3: An example of the cleaning step. A subset of the ECAL (red boxes) and HCAL (blue boxes) energy entries in the $\eta - \phi$ plane before (on the left) and after (on the right) the cleaning step are shown. All energy deposits crossed by the extrapolated trajectory of the muon track, as well as the muon track itself (red line) are removed. All energies are given in GeV.

removal of all energy in a calorimeter cell crossed by the muon track may influence the reconstruction of other particles close to the selected muons. After a detailed study presented in section 4.1.1, the approach to set the cell energy to zero was chosen as being most appropriate.

3.1.3 Generator Step

During the generator step, the decay of two $\tau$-leptons is simulated, using the kinematic properties of the selected muons. The hadronization is simulated using PYTHIA 8.2 [48]. Embedded events are produced in a total of six different final states. Four final states cover the four main decay channels of $\tau$-lepton pairs listed in table 1.1: $e\mu$, $e\tau_h$, $\mu\tau_h$, and $\tau_h\tau_h$. Two additional samples are produced by replacing the selected muons with electrons, referred to as $\mu \rightarrow e$ embedding and replacing the selected muons with muons, referred to as $\mu \rightarrow \mu$ embedding. The latter two samples are used for validation purposes, as described in section 4.1, and to derive the necessary correction factors needed for electrons and muons in embedded events. This is described in section 5.1.1. All six samples are produced using the same selected $\mu\mu$ events.

In the case of $\mu \rightarrow \tau$ embedding, at least two neutrinos are included in all final states. Neutrinos can only be detected by missing transverse momentum in the event. As the amount of energy transferred to the decay products is not fixed, the energy of the visible decay products of the decay can be significantly lower than the energy of the initially selected muons, if the neutrinos emerge from the decay with higher energy. Events with low visible energy are of little interest in a target analysis, as they are removed by the kinematic requirements of the analysis, which are typically set by the loosest trigger path sensitive to the given final state. For example, in the $\mu\tau_h$ channel, the lowest trigger path sensitive to a muon and a $\tau_h$-candidate in the event requires a muon with $p_T$ larger than 20 GeV, and a $\tau_h$-candidate with $p_T$ larger than 27 GeV. Therefore, it is not beneficial to produce embedded events with kinematic properties much lower than the requirements
The Embedding Technique

Table 3.2: The selection requirements applied to a simulated lepton pair in the generator step. They are determined based on the requirements of later analyses. The first decay product refers to the first particle in the name of the final state, if both particles are identical, the particles are ordered by $p_T$. In the $e\mu$ final state, either the electron or the muon can be the particle with higher $p_T$. The branching ratio is taken from [20].

<table>
<thead>
<tr>
<th>Final State</th>
<th>1st Decay Product</th>
<th>2nd Decay Product</th>
<th>Branching Ratio (BR)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\tau_\ell \tau_\ell$</td>
<td>$p_T^\tau &gt; 33\text{ GeV}$ $\eta^\tau &lt; 2.2$</td>
<td>$p_T^\ell &gt; 33\text{ GeV}$ $\eta^\ell &lt; 2.2$</td>
<td>42%</td>
</tr>
<tr>
<td>$\mu\tau_\ell$</td>
<td>$p_T^\mu &gt; 18\text{ GeV}$ $\eta^\mu &lt; 2.2$</td>
<td>$p_T^\tau &gt; 18\text{ GeV}$ $\eta^\tau &lt; 2.4$</td>
<td>23.1%</td>
</tr>
<tr>
<td>$e\tau_\ell$</td>
<td>$p_T^e &gt; 22\text{ GeV}$ $\eta^e &lt; 2.2$</td>
<td>$p_T^\tau &gt; 18\text{ GeV}$ $\eta^\tau &lt; 2.4$</td>
<td>22.5%</td>
</tr>
<tr>
<td>$e\mu$</td>
<td>$p_T^e &gt; 10\text{ GeV}$</td>
<td>$p_T^\mu &gt; 21\text{ GeV}$</td>
<td>6.2%</td>
</tr>
<tr>
<td>$p_T^e &gt; 21\text{ GeV}$</td>
<td>$p_T^\mu &gt; 10\text{ GeV}$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mu \to e$</td>
<td>$p_T^\mu &gt; 22\text{ GeV}$</td>
<td>$p_T^e &gt; 10\text{ GeV}$</td>
<td>not needed</td>
</tr>
<tr>
<td>$\mu \to \mu$</td>
<td>$p_T^\mu &gt; 17\text{ GeV}$</td>
<td>$p_T^\mu &gt; 8\text{ GeV}$</td>
<td>not needed</td>
</tr>
</tbody>
</table>

applied in a target analysis, since those events would not be used in the analysis anyway. To reduce the amount of CPU time spent on the simulation of energy deposits in the detector, a filter with requirements listed in table 3.2 is applied prior the simulation of the detector response.

In order to increase the number of decays passing these selection requirements, the simulation of the embedded lepton decay is repeated 1000 times for each event. This increases the number of events, that fulfilled the selection criteria for the visible decay products. If at least one simulated decay passes the selection criteria, the last passing simulated decay is stored and passed on to the detector simulation. A weight defined by

$$w_{gen} = \frac{N_{pass}}{N_{tot}} \cdot BR$$

(3.1)

is calculated for every passing event, where $N_{pass}$ is the number of decays passing the selection requirements, and $N_{tot}$ is the total number of trials, in this case 1000. The weight describes the probability of the event to pass the kinematic requirements and ensures, that the bias towards decays with kinematic properties close to the kinematic requirements is corrected correspondingly. This way, events, for which the simulated decay passes the selection requirements almost every time, receive a higher weight than events with only a few passing trials. With the help of this procedure, depending on the final state, more than 50% of the initially selected $\mu\mu$ events can be used to create embedded events. The number of events in an embedded sample is up to 50 times higher, than the number of expected $Z \to \tau\tau$ events in data, depending on the final state.

The total number of trials $N_{tot}$ was optimized to maximize the number of events passing the selection requirements. In Figure 3.4, the acceptance rate of a small $Z \to \mu\mu$ sample for different number of total decay simulation trials is shown. After $\approx 500 - 1000$ events, the number of events passing the selection does not increase anymore, which is why more than 1000 trials would only take up more computation time without any
3.1 Embedding Strategy

Figure 3.4: The rate of events passing the selection requirements in the $e\tau_h$ final state. Only a small set of randomly selected $Z \rightarrow \mu\mu$ events is shown, which is why the acceptance rate is higher than the total acceptance rate in the $e\tau_h$ final state. The error bars represent the binomial uncertainty for each value of $N_{\text{tot}}$ and the red line represents the chosen value of $N_{\text{tot}}$. The average time needed for one simulation trial is in the order of $O(\text{ms})$.

increase in the rate accepted events. The uncertainty for the weight can be calculated using the binomial uncertainty

$$
\sigma(\epsilon) = \sqrt{\frac{\epsilon(1-\epsilon)}{N_{\text{tot}}}}
$$

(3.2)

The chosen value of $N_{\text{tot}} = 1000$ corresponds to an uncertainty of $\approx 1.5\%$ on the $w_{\text{gen}}$ weight. For $\mu \rightarrow e$ and $\mu \rightarrow \mu$ embedded events, the simulation of the respective lepton is only performed once, because no neutrinos are produced for these processes. Furthermore, the $\mu \rightarrow e$ and $\mu \rightarrow \mu$ embedded event samples are not intended for application in any analysis.

After the simulation of the $\tau$-lepton decay, the response of the detector is simulated. During this simulation, only the $\tau$-lepton decays are present in the detector so the decays are simulated in an otherwise empty detector. The simulation of a $\tau_h\tau_h$ decay is shown in Figure 3.5. After the detector simulation, the simulation of the HLT is applied. Because only the two $\tau$-leptons are present during the HLT simulation, only trigger paths sensitive to $\tau$-leptons and their decay products can be used for embedded events in the later analysis.

In all other cases, corrections determined from data have to be applied. As an example a trigger path sensitive to the production of a Higgs boson via vector-boson-fusion requires additional event information other than the $\tau$-lepton decay. Furthermore, the geometry of the detector during the generator step and the geometry of the detector during all other steps differ. E.g. it is necessary to apply a correction on the beamspot and the
The Embedding Technique

Figure 3.5: Simulation of a $\tau^- \tau^-$ decay in the generator step. The detector is empty, besides the decay products of the two $\tau$-leptons.

primary vertex of the generated event to compensate for geometrical differences. Further details are given in section 4.2.1.

3.1.4 Merging Step

The final step of the embedding technique is to merge the simulated lepton decays back into the cleaned event. After merging the two events, the full reconstruction of the event is performed, starting from the lowest level possible.

Ideally, the merging would be done at the level of individual tracker hits and calorimeter cell entries. However, problems arise due to the geometrical shift between the geometry in the generator step, where a simulated detector geometry is used, and the actual geometry during the data taking. The location of all subdetector components relative to each other has to be measured and well-defined, in order to achieve a high geometrical resolution, especially in the inner tracker. This calibration process is called alignment.

The term different geometries refers to different alignment settings. The statistical accuracy of the alignment should remain significantly below the typical intrinsic silicon hit resolution of between 10 and 30 $\mu$m [49]. The alignment is different for the simulated and the actual detector geometry. The geometric shifts between those two geometries can be in the order of mm, and there is no straightforward way of converting one alignment setting into another.

These $O(\text{mm})$ shifts are enough to completely displace a track in the inner tracker. Additionally, the geometric differences are not uniform, they may differ from layer to layer, including shifts and rotations. As a result, a track of the simulated decay, that
3.1 Embedding Strategy

was well reconstructed in the generator step, may not be identified as one single track after merging on the level of tracker hits, because of displaced hits. As the particle flow reconstruction relies on a good track finding, this is a problem that can not be easily overcome with such an approach.

On the other hand the merging has to be done before particle identification is applied, as particle identification, among other things, relies on the isolation of a particle. If the particles were identified directly after the generator step, the detector would be empty apart from the two $\tau$-lepton decays. As a result, the particles would be more isolated during their identification, than they actually are in the merged event.

Keeping these two effects in mind, the merging is performed on the level of the inputs needed to reconstruct the particles. The merging strategy is optimized to minimize the impact of geometric effects [50], whilst being at the earliest possible reconstruction step. The chosen process is sketched in Figure 3.6. Tracks and calorimeter clusters of the simulated decays are reconstructed during the generator step. The reconstructed tracks of the decay products and calorimeter entries on the level of reconstructed calorimeter clusters are merged into the cleaned event. This ensures that the combination of different subdetector measurements in order to reconstruct and identify a particle is based on the full event information available, while at the same time effects due to the differences in the geometries are minimized. Since the merging of two different events is a complicated procedure, checks for unwanted effects or problems have to be made. Several studies concerning this are presented in section 4.1.

![Event Merging Strategy](image)

**Figure 3.6:** Description of the merging strategy chosen for embedding. The signals in the subdetectors are reconstructed individually for the simulated decay and the cleaned event. They are merged before the information of the different subdetectors is combined using the PF algorithm.
3.2 Production of Embedded Events using 2017 Data

In the following, the production of embedded events using pp collisions measured by the CMS experiment in 2017 are presented. First, the computational setup will be presented, followed by a more detailed analysis of the performance of the 2017 embedded event production.

3.2.1 Workflow

The CMS experiment provides a dataset with $\mu\mu$ events that were measured by the CMS detector. This preselected dataset (referred to as DoubleMuon) is used as input for the embedding technique. The cleaning, the generator, and the merging step require the full detector information, as those steps access and modify low level detector information like tracker hits or calorimeter entries. This information is only available, when using the earliest uncompressed data tier of the CMS experiment, the RAW event information [51]. The most important data tiers defined by the CMS experiment and the required disk space per event are listed in table 3.3.

The embedding technique is applied on an event-by-event basis, meaning each event is processed separately and all four steps have to be applied to a single event individually. This gives the possibility to easily parallelize the embedded event production by splitting the production into many small pieces (jobs). Each job processes only a small fraction of the whole dataset. In the case of embedded event production, each job has been configured to process 1000 events. This number of events per job was chosen to match the specifications set by the computing resources and confine the average runtime of a single job. The processing is done using multiple high performance computing (HPC) resources available to the high energy physics community. Jobs are submitted via the job submission tool grid-control [52], together with the high-throughput computing software HTCondor which is responsible for the handling of the individual jobs and the allocation of slots on the different computing resources available [53].

The six different embedding final states listed in table 3.2 are processed individually. This implies that the input dataset has to be read a total of six times. On average, about 30% of all events in the DoubleMuon dataset pass the selection requirements listed in table 3.1, reducing the amount of events suitable for the embedding technique by a

<table>
<thead>
<tr>
<th>Data Tier</th>
<th>Description</th>
<th>Event Size in kB</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAW</td>
<td>raw detector information</td>
<td>$\approx$ 700</td>
</tr>
<tr>
<td>RECO</td>
<td>all reconstructed physics objects</td>
<td>$\approx$ 1300</td>
</tr>
<tr>
<td>RAWRECO</td>
<td>combination of RAW and RECO</td>
<td>$\approx$ 2000</td>
</tr>
<tr>
<td>AOD</td>
<td>version of RECO, optimized for analyses</td>
<td>$\approx$ 450</td>
</tr>
<tr>
<td>MINIAOD</td>
<td>version of AOD, optimized for analyses</td>
<td>$\approx$ 45</td>
</tr>
<tr>
<td>NANOAOD</td>
<td>version of MINIAOD, optimized for analyses</td>
<td>target: $&lt;1$</td>
</tr>
</tbody>
</table>
3.2 Production of Embedded Events using 2017 Data

factor of three. Since jobs are split by events in the input sample, it is efficient to apply a pre-selection prior to the embedded event production. During the pre-selection, the requirements of the selection step (table 3.1) are applied. This ensures that the number of events that really enter the process per embedded event production job is maximized.

The DoubleMuon dataset, used as input for the pre-selection, is stored at multiple storage sites distributed around the world, and can be accessed via the Worldwide LHC Computing Grid (WLCG) [55, 56]. The output of the pre-selection is stored at the storage facility of the GridKa computation center in Karlsruhe. Due to a misconfiguration in the already available configuration of the pre-selection, the output was stored in the RAWRECO data format, instead of RAW, which would have been sufficient. Therefore the size of the output dataset is bigger than the input dataset despite of the reduction of the number of events. The majority of the processing was done, using the bwForCluster NEMO provided by the bwHPC project [57]. The pre-selection process is illustrated in Figure 3.7.

The pre-selected dataset is used as an input for the embedded event production. The processing steps are similar to the pre-selection described in the previous paragraph. During the embedded event production, all four steps of the procedure, including the selection step, are applied to the pre-selected dataset. As steps are processed sequentially, a temporary local RAWRECO file is created after each step and used as the input of the following step. The temporary files are deleted after the job is completed and only the output of the merging step is saved to the storage facility provided by the DESY. The output is saved in the MINIAOD data format [58], listed in table 3.3, which is optimized

---

**Figure 3.7:** A flowchart of the event pre-selection for the embedded event production. The selection step is applied to the DoubleMuon dataset and the resulting dataset is stored at the storage facility of the GridKa computation center.
The Embedding Technique

Embedded event production

Figure 3.8: A flowchart of the production of embedded events. The embedding technique is applied to the pre-selected dataset. Each final state is processed individually. The output is saved to a storage facility provided by DESY.

to provide all information necessary for an analysis, while requiring as little disk space as possible. The computation workflow for the embedded event production is illustrated in Figure 3.8.

3.2.2 Performance of the Embedded Event Production

The number of collisions measured by the CMS experiment in 2017 corresponds to an integrated luminosity of $L = 41.5\text{fb}^{-1}$. The DoubleMuon dataset used as input for the pre-selection consists of 219,230,732 events with a total size of 166.3 TB. The pre-selected dataset contains all events suitable for the embedding technique. It consists of 65,170,166 events with a size of 201.22 TB. The total number of events was reduced by an overall factor of three, when comparing the pre-selected dataset to the DoubleMuon dataset, whereas the filesize of the pre-selected dataset increased by about 21% due to the reason given above.

The pre-selected dataset is used as an input for the embedded event production campaign. The details of the $\mu \rightarrow \tau$ embedded event production are listed in table 3.4. The details for the production of the $\mu \rightarrow e$ and $\mu \rightarrow \mu$ embedded event sample can be found in Appendix B.1.1.

As expected, the embedded event sample in the $\tau_h\tau_h$ final state is the smallest in terms of number of events. The selection criteria applied during the generator step are tighter in the $\tau_h\tau_h$ channel compared to the other final states. On the other hand, the $e\mu$ channel has the highest acceptance rate due to its loose selection criteria. For each channel, 65,173
3.2 Production of Embedded Events using 2017 Data

Table 3.4: Details of the $\mu \rightarrow \tau$ embedded event production split by final state. The acceptance rate shows how many events from the input dataset ended up in the respective embedded event sample. The pre-selected dataset is used as the input for every final state. Therefore the number of jobs needed to process the one final state is always the same.

<table>
<thead>
<tr>
<th>Input</th>
<th>Events</th>
<th>Size</th>
<th>No. Jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>65,170,166</td>
<td>201.22 TB</td>
<td>65.173</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Final State</th>
<th>$\mu\tau_h$</th>
<th>$e\tau_h$</th>
<th>$\tau_h\tau_h$</th>
<th>$e\mu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Events</td>
<td>34,683,979</td>
<td>32,788,982</td>
<td>17,338,560</td>
<td>37,713,920</td>
</tr>
<tr>
<td>Size</td>
<td>1567.3 TB</td>
<td>1.564 TB</td>
<td>0.907 TB</td>
<td>1.751 TB</td>
</tr>
<tr>
<td>Acceptance Rate</td>
<td>53.22 %</td>
<td>50.31 %</td>
<td>26.61 %</td>
<td>57.87 %</td>
</tr>
</tbody>
</table>

Table 3.5: The average runtime of the four steps of the embedding technique.

<table>
<thead>
<tr>
<th>Step</th>
<th>Avg. Runtime $T_{step}$</th>
<th>$T_{step}/T_{tot}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Selection</td>
<td>5.0 s</td>
<td>32.5 %</td>
</tr>
<tr>
<td>Cleaning</td>
<td>4.4 s</td>
<td>28.6 %</td>
</tr>
<tr>
<td>Generator</td>
<td>4.9 s</td>
<td>31.8 %</td>
</tr>
<tr>
<td>Merging</td>
<td>1.1 s</td>
<td>7.1 %</td>
</tr>
<tr>
<td>Total $T_{tot}$</td>
<td>15.4 s</td>
<td>100 %</td>
</tr>
</tbody>
</table>

jobs were processed, resulting in a total of 391,038 jobs. The average runtime of all jobs during the production was 5.0 hours.

For $\mu \rightarrow \tau$ embedded event production the average runtime per step for a single event is listed in table 3.5. In addition, every job requires some time for initialization and the transfer of data.

In order to measure the performance of an embedded event production job, 200 production jobs per final state were monitored and analyzed in detail. All jobs use the same input events and all jobs were processed using the bwForCluster NEMO with an identical configuration to ensure comparability.

In Figure 3.9, all jobs show a CPU efficiency above 94%, irrespective of the final state. The CPU is mostly idle during data transfer. On average, the CPU efficiency is 97.9%. No correlation between the CPU efficiency and the runtime of a job is visible, as shown in Figure B.1c.

The runtime of a job is shown in Figure 3.10. The runtime, the disk space required, and the output size of a job are all related to the acceptance rate during the generator step of the final state. If more events pass the selection during the generator step, more events have to be processed, which takes more time and results in a larger output file. As the $\tau_h\tau_h$ final state has the lowest acceptance rate of all the $\mu \rightarrow \tau$ embedded final states, the average runtime of $\tau_h\tau_h$ jobs is the shortest. The runtime of $\mu \rightarrow e$ and $\mu \rightarrow \mu$ embedded production jobs is shorter in comparison to $\mu \rightarrow \tau$ embedded production jobs,
Figure 3.9: The CPU efficiency of a job embedded event production job. The colored box represents the 50% quantile, and the vertical line inside the box represents the median CPU efficiency. The whiskers represent the 5 and 95% quantile and each small dot represents one single production job.

because the simulation of the leptons is only performed once instead of 1000 times. On the other hand, the acceptance rates of $\mu \rightarrow e$ and $\mu \rightarrow \mu$ embedding are higher than the acceptance rate of any $\mu \rightarrow \tau$ embedded final state. However, the time saving due to the faster generator step is dominant, which is why the average runtime of the $\mu \rightarrow e$ and $\mu \rightarrow \mu$ embedding is shorter than for $\mu \rightarrow \tau$ embedding.

The correlation between the runtime and the acceptance rate can be seen when comparing the runtime of a job versus the disk space used, displayed in Figure 3.11. The $\mu \rightarrow e$ and $\mu \rightarrow \mu$ jobs have a higher disk usage whilst having a lower runtime than the other final states. The average input size is 1.51 GB/job. Because each step produces a temporary local output for the next step, the average disk usage is about 6.95 GB/job. The disk usage is capped at about 9 GB/job, due to limitations of the bwForCluster NEMO.

Overall, the embedded event production jobs show a very efficient usage of the provided resources. Additional performance variables are shown in Appendix B.1.2. A few optimizations are still possible for future productions e.g. the pre-selection process to store the output in the RAW data format. In addition, in order to reduce the disk usage of the jobs, a cleanup of the temporary local RAWRECO files could be applied after each embedding step.

More than 2.4 Million CPU hours were invested to produce the embedded event samples using the 2017 data. The whole production took about five weeks. The total size of the output files is 12.2 TB for all embedded events that were produced during the production campaign, about 1-3 TB per final state. This corresponds to a data reduction by a factor
3.2 Production of Embedded Events using 2017 Data

Figure 3.10: Average runtime of an embedded event production job.

of 100 compared to the input dataset, due to the selection during the generator step and the utilization of the MINIAOD format.
Figure 3.11: Runtime plotted against the disk usage of a job for the six different final states. The marginal distributions show the distribution of the runtime and disk usage of all final states combined. The black curves show the marginal distributions of all jobs regardless of their final state.
CHAPTER 4

Validation and Improvement of the Embedding Technique

4.1 Validation

The method can be validated by applying the embedding technique on a selection of simulated $Z \rightarrow \mu\mu$ events. Muons, electrons and $\tau$-leptons are used as embedded leptons to perform various validation studies. By applying the embedding technique on a simulated event sample, differences introduced by the method can be studied. In the following, the findings of these studies will be presented.

4.1.1 Applying $\mu \rightarrow \mu$ Embedding on Simulated Events

Selecting $Z \rightarrow \mu\mu$ events in simulation and replacing the selected muons with simulated muons is the sanity check of the technique. Ideally, there should be no difference between a selected $Z \rightarrow \mu\mu$ event and the resulting event after applying $\mu \rightarrow \mu$ embedding. All differences between the $\mu \rightarrow \mu$ embedded and the input $Z \rightarrow \mu\mu$ event sample must be understood and controlled. The study is performed on an event-by-event basis, since all events are included in both datasets. Final state radiation (FSR) of muons results in additional photons in the detector, which may not be reconstructed as a part of a muon, causing a smearing of the muon energy and momentum. The simulation of FSR can be turned off in the generator step of the embedding procedure. No FSR of the simulated muons helps to identify differences that are caused by the embedding technique and not by the FSR. The details of the event sample used for this study are given in Appendix B.2.1. In the following, various conclusions of this study are discussed.

Overall, good agreement between the $Z \rightarrow \mu\mu$ simulation and the $\mu \rightarrow \mu$ embedded event sample is observed. When looking at geometric variables like $\eta_\mu$ or $\phi_\mu$, as shown in Figure 4.1 (top row), both the simulation and the $\mu \rightarrow \mu$ embedded event sample follow the geometric acceptance of the detector. For example, the gaps in the muon system at $|\eta| = 0.3$ and $|\eta| = 0.9$ (see Figure 2.2) can be found in both the simulation and the $\mu \rightarrow \mu$ embedded event sample.
Figure 4.1: Comparison of distributions of simulated $Z \rightarrow \mu \mu$ events (red line) and $\mu \rightarrow \mu$ embedded events (blue dots). In the relative isolation plot, the vertical red line represents a selection requirement of $I_{\text{rel}} > 0.15$, typically used in target analyses. The error bands do not correspond to the uncertainty on the difference of the two samples, because the two samples are fully correlated. However, if the differences are within the statistical uncertainty of the simulated $Z \rightarrow \mu \mu$ sample, which contains about ten times more events than expected embedded events in target analyses, the differences are considered acceptable.
Mass of the Di–Muon System

When looking at the distribution of $m_{\mu\mu}$, it is noticeable that the Z boson mass peak in the simulated event sample has a better resolution than in the $\mu \rightarrow \mu$ embedded event sample. This is shown in Figure 4.1 (bottom left). In $\mu \rightarrow \mu$ embedded events, the muon pair is reconstructed twice, once during the selection step, and a second time after merging the two simulated muons back into the rest of the event. Each reconstruction causes a smearing effect on the muon energy and momentum distribution due to the finite detector resolution of the CMS detector. As mentioned in [47], a correction for this effect was calculated and applied. However, the correction was calculated based on a MC sample with conditions corresponding to the 2016 data taking. The correction does not completely mediate the effect in the simulation that is based on the 2017 conditions. However, as will be shown in section 4.1.3, in the actual use case of $\mu \rightarrow \tau$ embedding, the smearing effect caused by the neutrinos from the $\tau$-lepton decay is much larger than the smearing caused by the additional muon reconstruction. The correction derived based on 2016 conditions is still applied.

Relative Muon Isolation

The relative isolation of a muon $I_{\text{rel}}^\mu$ is defined as the ratio of the $p_T$ of all PF candidates from the PV around the muon in a cone with $\Delta R = 0.4$, over the $p_T$ of the muon. The exact definition of $I_{\text{rel}}^\mu$ is

$$I_{\text{rel}}^\mu = \frac{\sum_n p_T^{PV} + \max(0, \sum_n p_T^\gamma + \sum_n p_T^{NH} - 0.5 \cdot \sum_n p_T^{PU})}{p_T^{\mu}}$$

(4.1)

with $p_T^{PV}$ being the $p_T$ of charged PF candidates associated to the primary vertex of the event, $p_T^\gamma$ the energy of particle flow photons, $p_T^{NH}$ the energy of neutral hadrons, and $p_T^{PU}$ the $p_T$ of charged PF candidates not associated to the primary vertex. The relative isolation of $\mu \rightarrow \mu$ embedded muons shows a shift towards less isolated muons, which is evident from Figure 4.1 (bottom right). There are multiple explanations for this effect, like insufficient calorimeter cleaning and FSR of the initially selected muons. These reasons will be discussed in the following.

Most of the time, muons are very well isolated, having a relative isolation of zero. However, due to the fact that the relative isolation can not be lower than zero, and that in $\mu \rightarrow \mu$ embedding, muons are reconstructed a second time, the isolation can only fluctuate towards values larger than zero. More isolated muons after the second reconstruction are unlikely since most of the muons initially already have an isolation of zero. This asymmetrical effect causes a shift towards less isolated muons.

The muon isolation can also fluctuate towards higher values if additional neutral hadron PF candidates are present in the event. They may appear as remnants of the ECAL and HCAL calorimeter cleaning process, as described in section 3.1.2. During the PF reconstruction, explained in section 2.3.2, small energy deposits in the ECAL and HCAL without an associated particle or track are reconstructed as additional photons.
and neutral hadrons respectively. If some calorimeter energy deposits of the initially selected muon are not removed during the cleaning step, they will not be associated to the simulated muon in the generator step and instead, will be reconstructed as an additional neutral hadron candidate.

This effect is visible, when comparing the mean transverse momentum flux per muon in simulated and $\mu \rightarrow \mu$ embedded events. A $\Delta R \leq 0.4$ cone in the $\eta$-$\phi$ plane is constructed around the muon. The cone is split into 100 small rings around the muon candidate, each ring corresponding to an annular area of $[\Delta R_i, \Delta R_i + 0.004]$. In each ring, the $p_T$ of all PF candidates other than the muon is summed. The resulting $p_T$-sums are filled into a histogram binned in $\Delta R$.

For a uniformly distributed $p_T$ flux in the vicinity of the muon, a linear increase of the $p_T$-flux is expected. The $p_T$-flux of neutral hadrons around one muon in events with $m_{\mu \mu}$ between 80 and 100 GeV is shown in Figure 4.2. The embedded sample has a higher $p_T$-flux of neutral hadrons, due to the insufficient hadronic calorimeter cleaning mentioned above. On average, 181 MeV of additional energy coming from neutral hadron candidates can be found in the cone around the muons in $\mu \rightarrow \mu$ embedded events. The difference is calculated by subtracting the total flux of the two samples.

When comparing the number of reconstructed PF candidates in the event, the number of reconstructed neutral hadron candidates in the embedded event sample is also shifted towards higher values. This trend is shown in Figure 4.3 (left). Again, this can be explained with the insufficient HCAL cleaning. When applying a selection requirement
of \( p_T > 5 \text{ GeV} \), no shift between the simulated and the embedded event sample is visible, which implies that the energy of the additional hadron candidates, responsible for the shift, is small.

A second effect that can affect the relative muon isolation is the contribution of FSR of the initially selected muons. The selected muons can emit additional FSR photons in any part of the detector. If these photons are missed during the reconstruction in the selection step, they are either reconstructed as individual particles or completely missed. This leads to a reduction of the muon energy and momentum, and less isolated muons in the embedded event sample.

In order to study the effect of FSR, an additional selection requirement is applied to the simulated \( Z \rightarrow \mu\mu \) event sample to only select \( \mu\mu \) pairs without FSR from the two muons. Using the information available from the simulation of the \( Z \rightarrow \mu\mu \) event sample, all events with at least one photon radiated by a muon, are rejected. Due to this requirement, 62% of all \( Z \rightarrow \mu\mu \) events in the validation sample are rejected. The resulting event sample consists only of \( Z \rightarrow \mu\mu \) events without any FSR remnants of the initially selected muons, and is used to produce an FSR free \( \mu \rightarrow \mu \) event sample.

The contribution of FSR photons can be now checked by studying the \( p_T \)-flux of photons in the vicinity of the muon. Figure 4.4a shows the \( p_T \)-flux of the full simulated \( Z \rightarrow \mu\mu \) event sample and the corresponding \( \mu \rightarrow \mu \) embedded event sample. Figure 4.4b shows the same quantity for the subset of the initial samples, where only events without any FSR from the initially selected \( \mu\mu \) pair, as described above, are included.

In the subset without FSR, the additional photon energy in a cone of \( \Delta R \leq 0.4 \) around the muon is reduced from 74 MeV to 34 MeV. This means that half of the shift in the
muon isolation can be addressed to the final state radiation of the initially selected muons, which is not completely removed from the event during the cleaning step.

**Removal of Muon Energy Deposits in the Calorimeters**

As explained in section 3.1.2, in the current detector cleaning approach, all energy deposits in calorimeter cells crossed by an extrapolated trajectory of the muon track are set to zero. This results in an average energy deposit of 0.7 GeV removed in the ECAL, and an average of 7.0 GeV of energy removed from the HCAL. If the muon deposits energy in a cell not directly crossed by the extrapolated muon trajectory, this energy is not removed in the current setup. This can occur, if the trajectory is very close to the edge of a calorimeter cell, or if the muon loses energy by radiating photons. In addition, the extrapolated trajectory is only a rough estimate of the actual trajectory of the muon, as it is only based on the information of the inner tracker and the muon system available. On the other hand, removing all energy in a calorimeter cell could actually lead to an overcorrection concerning the surrounding of the muon, as the total energy deposit in a calorimeter cell can come from more than one single particle.

An alternative approach of the cleaning procedure is to calculate the average energy loss of a muon in the calorimeter material and to subtract this energy loss predicted by the theory from the energy deposited in each corresponding calorimeter cell. The average energy loss of a muon is described by

\[-\frac{dE}{dx} = a(E) + b(E)E \]  

(4.2)
4.1 Validation

Figure 4.5: Illustration of the calculation of the distance used in the alternative cleaning approach. The distance between the entry and exit point of the extrapolated muon trajectory is used to calculate the theoretical energy deposit of the muon in each calorimeter cell. The procedure is identical for the ECAL and the HCAL.

where \( a(E) \) is the energy loss by ionization described by the Bethe equation [20] and \( b(E) \) is the summed energy loss due to \( e^+ e^- \) pair production, bremsstrahlung, and photonuclear contributions. The contribution of the different effects depend on the energy of the muon and the traversed material. For energies below \( \approx 100 \text{ GeV} \), depending on the material, ionization is the dominant source of energy loss. In [59], values for \( a(E) \) and \( b(E) \) for different materials depending on the energy of the muon are calculated and listed. As explained in section 2.1, the ECAL of the CMS detector is made of lead tungstate crystals and the HCAL is made of brass as absorber material, an alloy composed of 70% copper and 30% zinc. The energy loss of a mixed material can be calculated using the weighted sum of the energy loss in the individual materials

\[
\frac{-dE}{dx} = \sum_i w_i a_i(E) + b_i(E)E
\]  

Instead of setting the energy deposit in crossed calorimeter cells to zero, the mean energy loss of the muon in a calorimeter cell is calculated. Depending on the calorimeter type and the \( p_T \) of the muon, an average muon energy loss per cm is calculated, using the tables listed in [59]. This energy loss is multiplied by the distance from the entry point to the exit point of the muon trajectory of each calorimeter cell, to get an estimate of the energy deposited in this calorimeter cell. The calculation of the expected distance is illustrated in Figure 4.5. This approach will result in removing less energy compared to the approach of setting the energy in all crossed cells to zero.

The alternative cleaning approach was applied to a simulated \( Z \rightarrow \mu\mu \) event sample and compared to the standard approach. Figure 4.6 shows the relative isolation of a muon in an \( \mu \rightarrow \mu \) embedded event sample using the standard cleaning approach, a \( \mu \rightarrow \mu \) embedded event sample using the alternative cleaning approach and the simulated \( Z \rightarrow \mu\mu \) event sample. Apart from the relative isolation of the muon, no difference between the two cleaning methods is observed. The alternative cleaning leads to less isolated muons with
Validation and Improvement of the Embedding Technique

Figure 4.6: The relative isolation of one muon in a simulated $Z \rightarrow \mu\mu$ event sample (red line), a $\mu \rightarrow \mu$ embedded event sample (blue dots) and an $\mu \rightarrow \mu$ embedded event sample using an alternative cleaning approach (turquoise dots).

respect to the simulated $Z \rightarrow \mu\mu$ event sample. Therefore, the approach of setting the energy of all calorimeter cells crossed by the extrapolated trajectory of the muon to zero is more compatible with the simulated $Z \rightarrow \mu\mu$ sample, than subtracting the average energy loss of a muon.

All effects described in the previous paragraphs lead to the conclusion, that the shift in the relative isolation of embedded muons in a $\mu \rightarrow \mu$ embedded events sample is a result of calorimeter cleaning remnants and FSR of the initially selected muons. Energy deposits outside the extrapolated muon track are not taken into account by the chosen cleaning approach. The trend towards less isolated embedded particles is compensated by correction factors as explained in section 5.1.1.
4.1 Validation

4.1.2 Applying $\mu \rightarrow e$ Embedding on Simulated Events

Electrons are identified using a multivariate discrimination method described in Section 2.3.3. The identification relies on the full event information and utilizes multiple variables closely related to the geometric position and the shape of the electron tracks and associated superclusters. Therefore, the distribution of the input variables in the embedding technique must be well described, in order to ensure a good modeling of electron identification. The optimization of the merging step described in section 3.1.4 was aimed at a more realistic electron discrimination in embedded events.

The performance of the electron identification in embedded events can be studied by comparing the efficiency of the electron identification in $\mu \rightarrow e$ embedded event samples using data as input with the efficiency measured in data or in simulated $Z \rightarrow ee$ event samples. The efficiency of the electron identification is measured using the "tag-and-probe" method which is explained in section 5.1.1.

The electron identification efficiency of a simulated $Z \rightarrow ee$ event sample, a $\mu \rightarrow e$ embedded event sample, and measured $ee$ events is shown in Figure 4.7. On the left, the efficiency of the WP90 and on the right, the efficiency of the WP80 is shown. The working points correspond to two configurations of the electron discriminator and are designed to result in an electron identification efficiency of 90% and 80% respectively. The electron identification efficiency of the embedded event sample is about 30% (10%) lower for the WP80 (WP90) working point, compared to the efficiency in simulation and data.

![Efficiency of the electron identification in data (black dots), a simulated $Z \rightarrow ee$ event sample (red dots) and a $\mu \rightarrow e$ embedded event sample (blue dots). In the lower panel of the figure, the ratio of the efficiency in data over the efficiency in simulation and the $\mu \rightarrow e$ embedding are shown. The efficiency of the electron identification in the $\mu \rightarrow e$ embedded event sample is $\approx 10\%$ lower for the WP90 and $\approx 30\%$ lower for the WP80 working point, compared to the efficiency in simulation and data.](image-url)
This discrepancy is mitigated by correction factors, which reweight the embedded event sample to have an identification efficiency identical to the efficiency in data. However, large correction factors may point to problems intrinsic to the embedding technique and especially the merging strategy and have to be followed up.

In 2017, 25 different input variables are used for the electron discrimination. As mentioned in Section 2.3.3, the different input variables of the electron identification can be grouped into three main categories: tracking observables, calorimeter observables, and observables determining the quality of the matching between the calorimeter information and inner tracks. A list of all input variables can be found in Table B.3.

In Figure 4.8, a selection of four different input variables is shown. In the top right, $f_{\text{brem}}$, the fraction of energy lost by the electron due to bremsstrahlung

\[ f_{\text{brem}} = \frac{p_{\text{in}} - p_{\text{out}}}{p_{\text{in}}} \]  

is shown, where $p_{\text{in}}$ is the electron track momentum closest to the beam spot, and $p_{\text{out}}$ the momentum of the electron when exiting the tracker, extrapolated to the ECAL surface.

On the top left, the ratio of energy deposited in the hadronic calorimeter $E_{\text{HCAL}}$ over the energy deposited in the electromagnetic calorimeter $E_{\text{ECAL}}$ is shown. An electron should deposit only very little to no energy in the HCAL.

In the lower row, the distributions of

\[ |\Delta \eta_{\text{in}}| = |\eta_{\text{sc}} - \eta_{\text{extrap}}| \]  

\[ |\Delta \phi_{\text{in}}| = |\phi_{\text{sc}} - \phi_{\text{extrap}}| \]

are shown, where $\eta_{\text{sc}}$ ($\phi_{\text{sc}}$) is the $\eta$ ($\phi$) position of the electron supercluster, and $\eta_{\text{extrap}}$ ($\phi_{\text{extrap}}$) is the $\eta$ ($\phi$) position of an extrapolation of the electron track to the ECAL surface. This extrapolation is based on the innermost tracker hit.

When comparing the distributions of all input variables in a $\mu \to e$ embedded event sample using data as input with a simulated $Z \to ee$ event sample, especially variables related to the matching between the subdetectors show larger differences. As explained in section 3.1.4, reconstructed tracks and calorimeter clusters of the simulated decay and the cleaned event are combined during the merging step. After that, the full event information is available, and the information from the whole detector can be combined using the PF algorithm, as explained in section 2.3.2. The combination is performed in the detector geometry used in data. However, the tracks and calorimeter clusters of the simulated lepton decay are reconstructed during the generator step, hence using a simulated detector geometry. The resulting differences can be related to the differences of the detector geometry used during the generator step and the detector geometry used in data. In the following, the reason for the discrepancy in the distribution of $\Delta \eta_{\text{in}}$ will be demonstrated.

The two positions needed for the calculation of $\Delta \eta_{\text{in}}$, the position of the supercluster and the innermost tracker hit, are determined during the generator step. A single tracker hit contains several quantities:
Figure 4.8: The distribution of four input variables to the electron identification discriminator in a $\mu \rightarrow e$ embedded event sample using 2017 data as input (blue dots) and a simulated $Z \rightarrow ee$ event sample (red line). The different variables are explained in the text.
• a unique number (DetID) which is used to globally identify the tracker cell that contains the hit,
• the relative position of the tracker hit in the $x$–$y$ plane within the inner tracker,
• the momentum of the estimated trajectory,
• the charge of the estimated trajectory,
• the direction of the estimated trajectory.

The problem, which arises due to the change of geometry is illustrated in Figure 4.9. The global location of the tracker hit is not stored explicitly. Instead, it is determined by the position of the tracker cell associated with the hit. The tracker cell is identified using the DetID and its global position is determined from the corresponding alignment. When changing the geometry, the alignment changes, however the DetID associated to the tracker hit does not. As a result, the location of the tracker cell is moved to a different location and the global position of the innermost tracker hit changes as well.

The extrapolation of the innermost tracker hit, that is used for the calculation of $\Delta \eta_{\text{in}}$ is performed during the merging step with the geometry used in data. But now, the extrapolation starts from a shifted position, which results in a different extrapolation position on the ECAL surface. This results in a larger distance between the position of the supercluster and the extrapolation, and as a consequence results in higher values for $\Delta \eta_{\text{in}}$.

**Figure 4.9:** Illustration of the extrapolation of the innermost track hit during the generator and the merging step. A sector of the CMS detector in the $x$–$y$ plane is shown. In this example, the innermost track hit is shifted in the $x$–$y$ plane due to a different alignment configuration, resulting in a shifted extrapolation of the innermost tracker hit that does not match anymore with the associated supercluster.
4.1 Validation

All detector geometries are based on an ideal model of the CMS detector. For a simulated detector geometry, the ideal model is slightly modified to mimic the detector conditions during data taking, so all simulated geometries are similar to each other. The detector geometry used in data is determined by various alignment measurements. The measurements determine, how the ideal model of the CMS detector has to be altered, in order to represent the real world detector. As a result, the detector geometry used in data can be different from the simulated detector geometries.

If the higher $\Delta \eta_{in}$ values are related to geometric differences, the agreement between a simulated $Z \to ee$ sample and a $\mu \to e$ embedded event sample using simulated $Z \to \mu\mu$ events as input is expected to be better. In this case, both the detector geometry of the input sample and the detector geometry used during the generator step are simulated geometries. The resulting distribution of $\Delta \eta_{in}$ and $\Delta \phi_{in}$ is shown in Figure 4.10. The differences are smaller compared to the differences shown in Figure 4.8 (lower row) since both geometries are nearly identical. Small deviations are still visible, as the detector geometry used during the generator step and the detector geometry used for the simulated $Z \to \mu\mu$ events are different due to technical limitations.

It is possible to quantify the shift between the detector geometry used during the generator and the merging step, by comparing the global position of the innermost track hit in the generator and in the merging step. In Figure 4.11, the shift of the innermost tracker hit position is shown. The blue histogram corresponds to the shift observed in a $\mu \to e$ embedded event sample using simulated $Z \to \mu\mu$ events as input. Although the position of the pixel hits fluctuates, the median shift is zero in $x$, $y$, and $z$ direction. When using data as input for $\mu \to e$ embedding, the median shift is 1.12 mm, −1.08 mm, and −3.17 mm in $x$, $y$, and $z$ direction.
Figure 4.11: The shift of the position of the innermost track hit, due to the use of different geometries in the generator and the merging step. Shown are the shift in the x, y, and z direction. The orange histogram corresponds to the shift observed when using simulated $Z \rightarrow \mu\mu$ events as input. The blue histogram corresponds the shift observed when performing $\mu \rightarrow e$ embedding using data as input. While the average shift position when switching from a simulated to a simulated geometry stays at zero, the average shift when switching from a simulated to data geometry is in $O(\text{mm})$. 
One possible explanation for the shifts in $O(\text{mm})$ is the upgrade of the pixel detector. Between the 2016 and 2017 data taking, the pixel detector of CMS has been upgraded and repositioned. The results in a different alignment when comparing the 2016 and 2017 detector geometry used in data. A study described in [60] states, that the differences of the alignment in 2016 and 2017 are in $O(\text{mm})$, which is compatible with the observed shifts in Figure 4.11.

Figure 4.12, shows the distributions of $\Delta\eta_{\text{in}}$ and $\Delta\phi_{\text{in}}$ for a $\mu \to e$ embedded event sample using 2016 data as input and a corresponding simulated $Z \to ee$ event sample. The two distributions show better agreement than the distributions shown in Figure 4.8 (lower row). This means that the discrepancy in $\Delta\eta_{\text{in}}$ using 2017 data as input for embedded event samples may be directly related to the repositioning of the pixel detector between the 2016 and 2017 data taking.

In principle, this alignment shift could be corrected for. In the concrete configuration, this is not possible, as this correction would be different for every geometry, and would differ depending on the actual hit position. Instead, the correction factors for the electron identification correct for the lower efficiency observed in the embedded event samples.
Validation and Improvement of the Embedding Technique

Figure 4.13: The relative isolation $I_{e,\mu}^{\text{rel}}$ of the electron (muon) in the $e\tau_h (\mu\tau_h)$ final state. The embedded event sample is represented by blue dots. The simulated $Z \rightarrow \tau\tau$ event sample is shown as a red line. The vertical red line represents a selection value on the relative isolation, that is typically used in target analyses. The red error band corresponds to the statistical uncertainty of the simulation.

4.1.3 Applying $\mu \rightarrow \tau$ Embedding on Simulated Events

For the validation of the use case of the embedding technique, simulated $Z \rightarrow \mu\mu$ events are used as input for the production of $\mu \rightarrow \tau$ embedded events in the $e\mu, e\tau_h, \mu\tau_h$, and $\tau_h\tau_h$ final states. The resulting event samples can be compared with a simulated $Z \rightarrow \tau\tau$ event sample, which uses the identical simulation configuration as the simulated $Z \rightarrow \mu\mu$ event sample that is used as input. Therefore, the $\mu \rightarrow \tau$ embedded event sample and the simulated $Z \rightarrow \tau\tau$ event sample can be directly compared.

In Figure 4.13, the relative isolation of the electron and the muon in the $e\tau_h$ and $\mu\tau_h$ final states are shown. In this comparison, the events are independent of each other, and the error band can be interpreted as the statistical uncertainty of the simulated event sample. A shift in the embedded event samples towards less isolated particles is not visible, compared to the shift observed in the $\mu \rightarrow \mu$ embedded event sample shown in Figure 4.1 (bottom right).

In Figure 4.14, the distributions of $m_{\text{vis}}$ in the four different final states are shown. In all four final states, the embedded event samples are compatible with the corresponding simulated $Z \rightarrow \tau\tau$ event sample. It should be noted, that especially in the $\tau_h\tau_h$ final state, the simulated event sample consists of only about 4000 events, while the embedded event sample contains about ten times more events. The smearing of the mass peak as observed in the case of $\mu \rightarrow \mu$ embedded events shown in Figure 4.1 (bottom left) is not visible in any of the four final states. The smearing caused by the neutrinos produced during the $\tau$-lepton decay is larger than the effect caused by the muon reconstruction during the selection. As expected, $m_{\text{vis}}$ takes lower than $m_{\mu\mu}$ of the initially selected...
muons due to the non observable neutrinos that emerge from the decay. In the $e\mu$ final state $m_{\text{vis}}$ is the smallest because both $\tau$-leptons decay leptonically, resulting in four neutrinos in the final state.

Overall, the various validation studies show that the embedding technique is able to achieve an event description comparable to simulated events. Correction factors on the identification and isolation are used to correct for the differences between embedded event samples and data that have been described in this section.

**Figure 4.14**: The visible mass $m_{\text{vis}}$ in the $e\mu$, $e\tau$, $\mu\tau$, and $\tau\tau$ final state. The embedded event sample is represented by blue dots, and the simulated $Z \rightarrow \tau\tau$ event sample is shown as a red line.
4.2 Improvements of the Embedding Technique

4.2.1 Simulation of HLT $\tau_h$ Reconstruction in Embedding

Because new particles are simulated during the generator step (see section 3.1.3), the High-Level Trigger sequence has to be simulated as well, which is also done during the generator step of the procedure. In addition to the HLT sequence, all L1 triggers are simulated as well. The CMS trigger system is explained in chapter 2.2.

It was observed that trigger paths sensitive to hadronically decaying $\tau$-leptons show different results comparing $\mu \to \tau$ embedded events in the $\tau_h\tau_h$ final state, denoted as $\mu \to \tau_h$ in the following, with simulated $Z \to \tau_h\tau_h$ events. As an example, the di-$\tau$ mass of the two $\tau$-leptons in the $\tau_h\tau_h$ channel is shown in Figure 4.15. Whilst the event yield of the embedded and the simulated event sample differs by about 4% without any HLT information, the yield of the embedded event sample is about 35% lower when requiring a trigger path sensitive to a successful HLT reconstruction of two hadronically decaying $\tau$-leptons. This difference is shown in Figure 4.15 (right). In the semileptonic $e\tau_h$ and $\mu\tau_h$ final states, a similar effect can be observed, when requiring a trigger path sensitive to one $\tau_h$ and one leptonically decaying $\tau$-lepton. Figure 4.16 shows the discrepancy in the $\mu\tau_h$ final state.

Since this effect appears across all triggers containing hadronically decaying $\tau$-leptons the issue is most likely located in the HLT simulation during the generator step. The HLT reconstruction of $\tau_h$-candidates is described in section 2.2.1. During the generator step, the $\tau$-lepton decays are simulated in an otherwise empty detector. As an example, the trigger sequence of the $\text{DoubleTightChargedIsoPFTau35}$ trigger path, a trigger path sensitive to the presence of two $\tau_h$-candidates with a $p_T$ larger than 35 GeV, was studied in detail. In total, the trigger path consists of 312 different reconstruction steps.

The HLT simulation can be performed using a $Z \to \tau\tau$ simulated event sample. During the generator step, the simulated $\tau$-lepton decays have to meet selection criteria (listed in Table 3.1). The identical selection criteria are applied to a simulated $Z \to \tau_h\tau_h$ event sample, which allows a detailed comparison between the HLT simulation in the embedding and in the simulation case.

As explained in section 2.2.1, during the HLT reconstruction multiple selection requirements are applied to the $\tau_h$-candidates. In Figure 4.17, the rejection rate of these different selection requirements, denoted as filter steps in the following, are shown for a $Z \to \tau_h\tau_h$ simulated event sample, and a $\mu \to \tau$ embedded event sample in the $\tau_h\tau_h$ final state. The filter names corresponding to the filter steps are listed in Table 4.1.
4.2 Improvements of the Embedding Technique

Figure 4.15: Distribution of $m_{\text{vis}}$ before (left) and after (right) requiring a trigger path sensitive to two $\tau_h$-candidates. A $\mu \to \tau_h$ embedded event sample is represented by blue dots and a simulated $Z \to \tau_h \tau_h$ event sample is represented by a red line. The event yield for embedded event sample is 35% lower than the simulated event sample yield, when requiring the trigger.

Figure 4.16: Distribution of $m_{\text{vis}}$ before (left) and after (right) requiring a trigger path sensitive to one $\tau_h$-candidate and one muon. A $\mu \to \tau$ embedded event sample in the $\mu \tau_h$ final state is represented by blue dots, and a simulated $Z \to \mu \tau_h$ event sample is represented by a red line. The event yield for the embedded event sample is 30% lower than the simulated event sample yield, when requiring the trigger.
Validation and Improvement of the Embedding Technique

![Comparison of the filter rejection rate for $\mu \rightarrow \tau_h$ embedded (blue) and $Z \rightarrow \tau \tau$ simulated events (red). The relative rejection of each filter step with respect to the preceding steps are shown. The names corresponding to the numbered filter steps are listed in Table 4.1.](image)

**Figure 4.17:** Comparison of the filter rejection rate for $\mu \rightarrow \tau_h$ embedded (blue) and $Z \rightarrow \tau \tau$ simulated events (red). The relative rejection of each filter step with respect to the preceding steps are shown. The names corresponding to the numbered filter steps are listed in Table 4.1.

### Table 4.1: Filter names corresponding to the steps used in Figure 4.17.

<table>
<thead>
<tr>
<th>Step</th>
<th>Filter Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>hltL1sDoubleTauBigOR</td>
</tr>
<tr>
<td>2</td>
<td>hltDoubleL2Tau26eta2p2</td>
</tr>
<tr>
<td>3</td>
<td>hltDoubleL2IsoTau26eta2p2</td>
</tr>
<tr>
<td>4</td>
<td>hltPFTauTrackReg</td>
</tr>
<tr>
<td>5</td>
<td>hltDoublePFTau35Reg</td>
</tr>
<tr>
<td>6</td>
<td>hltDoublePFTau35TrackPt1TightChargedIsolationReg</td>
</tr>
<tr>
<td>7</td>
<td>hltDoublePFTau35TrackPt1TightChargedIsolationDz02Reg</td>
</tr>
</tbody>
</table>

As apparent from the rejection rates, the simulation and the embedding technique differ in most of the filter steps. Most notable are step 1, 3, 4, 6, and 7, where the difference in rejection rates is 20% or higher. During the first filter step the existence of two suitable L1 $\tau_h$ candidates is checked. In step 3, 6 and 7, the isolation of the HLT $\tau_h$-candidate is probed. Since the detector is empty during the generator step, it is expected, that only very few embedded events are rejected by this selection criteria. The only step, where significantly more embedded events than simulated events are rejected is step four, the hltPFTauTrackReg filter. In the simulated event sample, only $\approx 2\%$ of the events are rejected, while more than $40\%$ of embedded events are rejected during this filter step.

In the hltPFTauTrackReg filter step, a check whether there is at least one HLT $\tau_h$-candidate with $\eta \leq 2.5$ and $p_T \geq 0$ can be found in the event, is performed. For the
reasons given below, these selection criteria are often not met in the embedded event sample, which is one of the reasons, why the embedded event sample has a lower event yield when requiring a trigger path sensitive to a $\tau_h$-candidate.

The HLT $\tau_h$-candidates are reconstructed in a reconstruction step called \texttt{RecoTauProducer}. During the L3 step of the HLT $\tau_h$ reconstruction, tracks are reconstructed in a small region around the HLT $\tau_h$-candidate. Afterwards, those tracks are used to determine the primary vertex of the event. The vertex reconstruction in HLT is done by searching for at least two matching tracks in $z$ direction \cite{36}. The distance between the tracks and the potential vertex $z_{\text{offset}}$ must be smaller than 0.5 cm, or the track is excluded from the vertex reconstruction. Finally, at least two tracks must originate from the potential vertex. If a vertex is found, it is associated to the $\tau_h$ candidate.

During the HLT simulation of the embedding technique, the detector is empty apart from the two $\tau$-lepton decays. As shown in \cite{38}, the PV resolution gets better, the more tracks are used to reconstruct the vertex. Since the two decays produce only a few tracks, the resolution of the reconstructed PV is poor. As a fallback solution, the beamspot of the event is used as the reconstructed primary vertex, if the vertex reconstruction requirements listed above are not met. This leads to the fact, that the beamspot is used as the PV during the HLT simulation of the embedding procedure.

One selection requirement of the HLT $\tau_h$ reconstruction is the distance in $z$ direction between the reconstructed PV and the extrapolation of the track of the leading charged hadron in the $\tau$-jet cone to the beamline. This distance may not be larger than 0.2 cm. However, the beamspot is not the origin of the simulated $\tau$-lepton decays and therefore, if the beamspot is the only reconstructed vertex in the event, this requirement is often not met, causing the L3 step of the HLT reconstruction to fail.

As a suitable solution for this problem, a new reconstruction step to bypass the poor vertex reconstruction and therefore the possible fallback of using the beamspot, was implemented. The vertex used in the simulation of the two $\tau$-lepton decays is determined by the position of the originally selected muon pair. This location is used by the new reconstruction step and converted into a primary vertex. Additional information stored in association with the vertex, like fit quality or resolution uncertainty, are adjusted such that the new vertex will pass all selection criteria of the HLT sequence. The vertex reconstruction step in the HLT simulation of the embedded technique is then being replaced by this new reconstruction step. This mitigates the problem of poor vertex reconstruction if only a few tracks are present in the detector. The new HLT sequence is illustrated in Figure 4.18.

With this modification, an acceptance increase of 20 % in the total event yield in the $\tau_h\tau_h$ channel can be achieved. A similar solution can be applied in the $e\tau_h$ and $\mu\tau_h$ final state. In these final states, the trigger yields when requiring trigger paths sensitive to a $\tau_h$-candidate, are increased by $\approx 30\%$. A comparison of the event yield before and after the modification is shown in Figure 4.19.

The improved vertex finding makes the HLT simulation of embedded events more robust and provides a better description of trigger paths sensitive to a hadronically decaying $\tau$-lepton in embedding. The remaining differences of 20% in the $\tau_h\tau_h$, and 10% in the $e\tau_h$ and $\mu\tau_h$ final state between the simulated event sample and the embedding
Validation and Improvement of the Embedding Technique

Figure 4.18: Modified version of the HLT $\tau_h$ reconstruction sequence. With respect to the original sequence shown in Figure 2.4, the vertex reconstruction is replaced by a new reconstruction step (blue). The vertex used during the simulation of the $\tau$-lepton decay is picked up by the new reconstruction step and converted it into the only primary vertex used within the HLT $\tau_h$ reconstruction sequence.

Figure 4.19: Comparison of simulation (red) and embedding using the old HLT simulation method (blue) and the new one (green). The distribution of $m_{\text{vis}}$ in the $\tau_h\tau_h$ final state when requiring a trigger path sensitive to two $\tau_h$-candidates is shown on the left. The distribution of $m_{\text{vis}}$ in the $\mu\tau_h$ final state when requiring a trigger path sensitive to one $\tau_h$-candidate and one muon is shown on the right.
event sample yield can be compensated by applying correction factors derived using the 'tag-and-probe' method, that will be explained in section 5.1.1. With the improved vertex finding, the correction factors are significantly reduced.
CHAPTER 5

Application of Embedded Events in an Analysis

When applying $\mu \to \tau$ embedded events in a target analysis, simulated $Z \to \tau\tau$ events as well as $t\bar{t}$ and diboson events with two genuine $\tau$-leptons in the final state ($t\bar{t} \to \tau\tau X$ and $WW, ZZ, WZ \to \tau\tau X$) are replaced by the embedded event sample. In this chapter, the corrections and uncertainties needed for the application of embedded events in a target analysis are presented. In addition, the application of $\mu \to \tau$ embedded events in the $e\tau_h$ and $\mu\tau_h$ final state as well as a comparison with simulated $\tau\tau$ events will be discussed.

5.1 Corrections and Uncertainties for Embedded Events

Several corrections and uncertainties are applied to the embedded event sample:

- The correction factors for the $\mu\mu$ selection during the selection step (as explained in section 3.1.1) are derived using the "tag-and-probe" method described in the next section. The correction factors are binned in $p_T$ and $\eta$ of the selected muon and compensate any biases introduced by the modeling of the $\mu\mu$ selection as well as the finite detector resolution. Due to this correction, all information and respective uncertainties of the integrated luminosity and the cross section of $\tau\tau$ event production can be omitted in the target analysis. A global uncertainty of 2% per selected muon is applied, resulting in a total uncertainty of 4%.

- The weight $w_{\text{gen}}$ (explained in section 3.1.3) is introduced due to the repeated simulation of the $\tau$-lepton decay during the generator step. The weight is derived individually for every event and is applied on an event by event basis.

- Residual differences between the simulated $\tau$-lepton decay in a $\mu \to \tau$ embedded event and genuine $\tau$-leptons are discussed in chapter 4. In order to compensate these differences, correction factors for electrons and muons are applied. The correction factors are binned in $p_T$- and $\eta$. They correct the efficiency of the lepton isolation, identification, and trigger path requirements in the target analysis.
• For \( t\bar{t} \rightarrow \tau \tau X \) events, an uncertainty of 10% on the expected shape of the distribution of such events is introduced due to insufficient knowledge of the exact \( t\bar{t} \) contamination in \( \mu \rightarrow \tau \) embedded event samples.

• An uncertainty of 1.2% on the \( \tau \) energy scale is applied. On top of that, a 2% yield uncertainty per \( \tau \) is applied to compensate the larger tracking efficiency, due to the detector being empty during the generator step (explained in section 3.1.3).

Furthermore, the uncertainty model of the target analysis has to be adopted accordingly.

5.1.1 The Tag and Probe Method

The "tag-and-probe" (TP) method \([61, 62]\) is a method to determine the efficiency of different selection criteria. Those efficiencies can then be used correct the efficiency in a sample that is not data. Here, the efficiency of electron (muon) identification, the isolation requirement and the trigger paths sensitive to an electron (muon) are derived. This is done for both electrons and muons separately. The efficiencies are calculated for a \( \mu \rightarrow e (\mu \rightarrow \mu) \) embedded event sample, a simulated \( Z \rightarrow ee (Z \rightarrow \mu \mu) \) event sample, and data. The efficiencies are binned in \( p_T \) and \( \eta \) of the lepton. In the following, the process of calculating the efficiencies will be explained in more detail.

For the TP method, a sample consisting of suitable lepton pairs is needed. Therefore, the decay products of a known mass resonance, in this case the \( Z \) boson, are selected in order to probe the efficiency of different selection criteria. A suitable pair consists of two leptons of the same flavor: one "tag" lepton, which is required to pass tight selection criteria and is therefore very likely to be a genuine muon or electron, and one "probe" lepton which is loosely selected. Since both leptons are expected to be of the same flavor due to being a decay product of the \( Z \) boson, any "probe" lepton not passing a selection requirement lowers the efficiency of that criteria. By checking the number of "probe" leptons passing the selected criteria, an efficiency can be determined.

Because the two leptons are expected to originate from the decay the \( Z \) boson resonance, the visible mass of the pair is required to be between 65 GeV and 115 GeV. The requirements for the "tag" and the "probe" lepton for muons and electrons are listed in table 5.1. The \( p_T \) of the "tag" lepton is required to be at least one GeV larger than the minimal requirement of the "tag" trigger path, so the kinematic region where the "tag" trigger path is not fully efficient yet (the turn-on region) is excluded. On top of the requirements listed in the table, the two leptons are required to be well separated in the \( \eta-\phi \) plane \((\Delta R > 0.5)\) and of opposite charge. A single pair of leptons can result in two independent TP pairs if both leptons pass the "tag" requirements. In this case, each lepton is considered as "tag" lepton and "probe" lepton once.

Since the trigger paths use isolation and identification requirements, the different efficiencies are calculated sequentially, meaning that only pairs where the "probe" lepton satisfies the identification requirements are used for the isolation efficiency calculation, and only pairs passing the isolation requirements are used for the trigger path efficiency calculation. The total efficiency is then given by the product of all efficiencies

\[
\epsilon(ID,Iso,Trigger) = \epsilon(Trigger|ID,Iso) \cdot \epsilon(Iso|ID) \cdot \epsilon(ID)
\] (5.1)
Table 5.1: Selection requirements for a TP lepton pair. A pair is found if one lepton passes the 'tag' requirements and the other lepton passes the 'probe' requirements. The impact parameters $|d_{xy}|$ and $|d_z|$ denote the distance between the PV and the extrapolated track of the lepton. These requirements ensure that the selected leptons originate from the PV of the event.

| Electron | $p_T$ | $|\eta|$ | ID     | $|d_{xy}|$ | $|d_z|$ | $I_{rel}^\ell$ | Trigger |
|---------|------|--------|--------|---------|-------|--------------|---------|
| Tag     | > 36 GeV | < 2.5  | WP90   | < 0.045 | < 0.2 | < 0.10 | ✓ |
| Probe   | > 10 GeV | < 2.5  | -      | -       | -     | -          | -       |

| Muon    | $p_T$ | $|\eta|$ | ID     | $|d_{xy}|$ | $|d_z|$ | $I_{rel}^\ell$ | Trigger |
|---------|------|--------|--------|---------|-------|--------------|---------|
| Tag     | > 28 GeV | < 2.4  | IDMedium | < 0.045 | < 0.2 | < 0.15 | ✓ |
| Probe   | > 10 GeV | < 2.4  | TrackerMuon | -       | -     | -          | -       |

In $|\eta|$, the binning

- electron: [0.0, 1.0, 1.4442, 1.56, 2.1, 2.5]
- muon: [0.0, 0.9, 1.2, 2.1, 2.4]

is chosen, in order to incorporate the geometric properties of the ECA and the muon system. The $\eta$ bin [1.4442, 1.56] corresponds to the transition region between the barrel and the endcap region of the ECAL. In $p_T$, the binning is chosen to be as fine as possible, while keeping the number of events in each bin large enough.

The efficiency is calculated by checking if the 'probe' lepton passes the criteria for which the efficiency is measured. Based on this requirement, the TP pairs are sorted into passing and failing pairs. Both subsets are tested against a background plus Z boson as signal hypothesis, using the invariant mass distribution of the dilepton system. The efficiency is the ratio of the signal yield in the distribution of passing pairs over the signal yield for failing plus passing pairs.

The correction factors, which are applied to the embedded event samples are the ratio of the efficiency found in data over the efficiency found in $\mu \rightarrow e$ ($\mu \rightarrow \mu$) embedding:

\[
\text{Correction Factor} = \frac{\epsilon(\text{data})}{\epsilon(\text{embedding})} \quad (5.2)
\]

Figure 5.1 (left column) shows the efficiency of the electron isolation $I_{rel}^e < 0.15$, the efficiency of the WP90 electron identification, and the efficiency of a trigger path sensitive to an electron above 35 GeV in an $\eta$ region between 0 and 1.0. The isolation and trigger path correction factors are below 5% apart from the trigger turn-on region, where larger

\[^1\text{HLT_Ele35_WPTight_Gsf_v}\]
\[^2\text{HLT_IsoMu27_v}\]
correction factors are expected. The reason for the larger correction factors of up to 30\% for the electron identification is discussed in section 4.1.2.

The efficiencies for muons in the $0.0 > \eta > 0.9$ region are shown in Figure 5.1 (right column). The efficiency was measured for $I_{\text{rel}}^\mu < 0.15$, the IDMedium muon identification and a trigger path sensitive to one muon with $p_T$ larger than 27 GeV. The correction factors are mostly smaller than 5\%, except for the trigger turn-on region.
5.1 Corrections and Uncertainties for Embedded Events

Figure 5.1: The isolation, trigger efficiency for electrons and muons. The electron efficiencies for the first \( \eta \) bin \([0, 1.0]\) are shown on the left and the muon efficiencies for the first \( \eta \) bin \([0, 0.9]\) are shown on the right. The lower panel of each figure displays the correction factors, that are applied to the respective sample.
5.2 Application on Data

In this section, the application of $\mu \rightarrow \tau$ embedded events to model $\tau\tau$ backgrounds in a typical target analysis are presented. The selection and uncertainty model is based on a $H \rightarrow \tau\tau$ analysis in the $e\tau_h$ and $\mu\tau_h$ final states [9, 63], and is done inclusively.

5.2.1 Event Selection

The selection requirements applied are shown in Table 5.2. In the $e\tau_h$ ($\mu\tau_h$) final state, a trigger path sensitive to an electron (muon) with $p_T > 27$ (24) GeV is required. In addition, a trigger path sensitive to one $\tau_h$ and one electron (muon), where the electron (muon) is required to have at least 24 (19) GeV is included. Therefore, the $p_T$ threshold for the electron (muon) was chosen to be one GeV above the lowest trigger path $p_T$ threshold to ensure, that the trigger path is fully efficient. All leptons must pass identification requirements as described in section 2.3.3, and the $\tau_h$ must pass additional discrimination against being misidentified as an electron or muon. In both final states, the electron (muon) and the $\tau$-lepton are required to have opposite charge, and must be well separated in the $\eta$–$\phi$ plane ($\Delta R > 0.5$). Finally, the transverse mass

$$m_T^\ell = \sqrt{2p_T^\ell E_T^{miss}(1 - \cos(\Delta \phi))}$$

(5.3)

of the electron (muon) must be smaller than 50 GeV to suppress the contribution from $W$ with associated jets and $t\bar{t}$ processes. Here, $E_T^{miss}$ is the missing transverse energy and $\Delta \phi$ is the angle between the lepton and the missing transverse energy vector in the $\eta$–$\phi$ plane.

5.2.2 Additional Background Processes

As explained in section 1.1.2, the main background contributions in an $H \rightarrow \tau\tau$ analysis besides $Z \rightarrow \tau\tau$ are

- $W +$ jets: This process contributes e.g. if the W decays leptonically into a lepton and a neutrino, and the jet is misidentified as a $\tau_h$ (jet fakes). Processes, where a jet is misidentified as a $\tau_h$ are estimated using the fake factor method ($F_F$) [64], which is explained in the next paragraph.

- QCD multijet production: A contribution in the $e\tau_h$ and $\mu\tau_h$ final state is found, if one jet is misidentified as a $\tau_h$ and additionally, a lepton, which does not originate

| Table 5.2: Selection requirements for $e\tau_h$ and $\mu\tau_h$ final state. The index 1 refers to electron or the muon, and the index 2 refers to the $\tau_h$ candidate. |
|----------------|----------------|--------|--------|--------|
| $p_T,1$ | $p_T,2$ | $|\eta_1|$ | $|\eta_2|$ |
| $e\tau_h$ | $> 25$ GeV | $> 30$ GeV | $< 2.1$ | $< 2.3$ | $< 0.15$ |
| $\mu\tau_h$ | $> 20$ GeV | $> 30$ GeV | $< 2.1$ | $< 2.3$ | $< 0.15$ |

68
5.2 Application on Data

from a $\tau$-lepton decay is selected. This process is also estimated using the FF method.

- $Z \rightarrow \ell\ell$: This process contributes in the $e\tau_h$ and $\mu\tau_h$ final state, if one of the leptons is misidentified as a $\tau_h$. This process is estimated using simulation.

- $t\bar{t}$ events can contribute to a $\tau\tau$ final state, and are partially estimated by the embedded event samples due to the reasons explained in section 3.1.1. The contribution from $t\bar{t}$ events resulting in a prompt electron (muon) plus $\tau$-lepton final state is estimated using simulation.

- Diboson processes may also decay into two genuine $\tau$-leptons. They are estimated by the embedded event samples like $t\bar{t}$ events. Diboson events resulting in a prompt electron (muon) plus $\tau$-lepton final state are estimated using simulation.

The FF method is a data-driven method to estimate backgrounds in $\tau$-lepton analyses, that originate from quark or gluon jets being misidentified as a $\tau_h$ (called jet $\rightarrow \tau_h$ or jet fake). The main idea of the method is to derive correction factors (the fake factors $F_F$) and apply them on events with alternative $\tau_h$ ID requirements to get an estimate of the contributions from jets misidentified as $\tau_h$ in the analysis. For the alternative ID requirement, the $\tau_h$ isolation has to fulfill the very loose working point but not fulfill the medium and tight working point. All events can then be sorted into two regions, the signal region (SR) with the ID requirements of the analysis, were genuine $\tau_h$ and jet fakes are present, and the application region (AR) with the alternative ID requirement. The AR mostly contains jet fakes, with a minor contribution of a few percent by genuine $\tau_h$ decays.

The $F_F$ are measured as a function of the $p_T$ of the $\tau_h$-candidate in different determination regions (DR$_p$). In every DR$_p$, a certain background process is enriched. For example the number W+jets events is enriched by requiring $m_T > 70$ GeV. The $F_F$ are determined by

$$F_F = \frac{N_{\text{nominal}}}{N_{\text{altered}}}$$

Here, $N_{\text{nominal}}$ is the number of events that pass the ID requirements of the analysis, and $N_{\text{altered}}$ is the number of events passing the alternative ID requirements. The final contribution of all jet fake backgrounds is calculated by a multiplying the number of events in the AR with a weighted sum of the $F_F$ determined in different DR$_p$

$$N(\text{jet} \rightarrow \tau_h) = N(\text{AR}) \cdot \sum_p R_p F^p_F$$

where $N(\text{AR})$ is the number of events in the AR and $F^p_F$ if the $F_F$ determined in the corresponding DR$_p$. The contribution of each DR$_p$ to the AR is estimated using simulation and $R_p$ is the relative contribution factor. Using this method, event with jets misidentified as $\tau_h$ can be modeled without using simulated event samples.
5.2.3 Comparison of $\mu \rightarrow \tau$ Embedded Events with Simulated $Z \rightarrow \tau\tau$ Events

In Figure 5.2, the distribution of $m_{\text{vis}}$ in the $e\tau_h$ and $\mu\tau_h$ final state after the event selection described in section 5.2.1 and applying a likelihood fit are shown. The $\mu \rightarrow \tau$ embedded events are used to estimate $Z \rightarrow \tau\tau$ along with diboson and $t\bar{t}$ production decaying into to genuine $\tau$-leptons. All corrections specific to the embedding technique, listed in section 5.1.1, are applied to the $\mu \rightarrow \tau$ embedded event samples. $W$+jets, QCD multijet, and $t\bar{t}$ events, which contribute because a jet is misidentified as a $\tau_h$, are combined in the jet $\rightarrow \tau_h$ event class and are estimated using the $F_F$ method. This way, 76% of all events in the $e\tau_h$ and 90% of all events in the $\mu\tau_h$ final state are estimated using data-driven methods.

For the likelihood fit the $\mu \rightarrow \tau$ embedded event sample is treated as signal and the uncertainty model used in the SM $H \rightarrow \tau\tau$ analysis presented in [63] is used to describe all other uncertainties. All uncertainties are implemented as nuisance parameters in the likelihood model.

The fitting procedure can be repeated, using the simulation to estimate all processes with two genuine $\tau$-leptons, in order to compare the embedded event samples with the simulation. This includes $Z \rightarrow \tau\tau$, the $t\bar{t}$ and diboson production with two genuine $\tau$-leptons in the final state. For the simulated samples, corrections for lepton identification, isolation and corresponding trigger paths are applied. In addition, corrections on the jet energy and resolution, corrections on the $Z$ boson recoil, b jet identification and misidentification efficiency, and pileup distribution corrections are applied. The resulting distribution of $m_{\text{vis}}$ is also shown in Figure 5.2 as a red line in the top panel and red dots in the lower panel of the figure.

In addition, the distribution of $m_{\text{vis}}$ in the $e\tau_h$ and $\mu\tau_h$ final state before applying a likelihood fit are shown in Figure 5.3. Both methods show fluctuation within the background uncertainties. The simulation has a small tendency of underestimating the yield of observed events, while the embedded samples show a small overestimation of the observation.

The comparison of $\mu \rightarrow \tau$ embedded events and simulation can also be done in other variables. Figure 5.4 shows the number of jets and the leading jet $p_T$ in the $e\tau_h$ and $\mu\tau_h$ final states.

To further quantify the result of the likelihood fits, goodness-of-fit tests can be performed. In this case, a generalized version of the $\chi^2$ test, the saturated likelihood model (SAT) [65, 66] test was used. The SAT test results in a more accurate p-value, if the uncertainty model contains correlated uncertainties. In addition, the Kolmogorov-Smirnov test (KS), which is sensitive to differences between the cumulative distribution function and the empirical distribution function, and the Anderson-Darling test (AD), sensitive to the integral of the differences between the cumulative distribution function and the empirical distribution function are applied. The normalization, with respect to the expectation, resulting from the fit, and the p-values from the three different goodness-of-fit tests, are listed in Table 5.3. The p-values are obtained by performing pseudo experiments based
5.2 Application on Data

Figure 5.2: Postfit distribution of $m_{\text{vis}}$ in the $e\tau_h$ (left) and $\mu\tau_h$ (right) final state. The $F_F$ method is used to estimate jet $\to \tau_h$ events and $\mu \to \tau$ embedded events are used for the estimation of $\tau\tau$ backgrounds. The black dots in the lower panel of the figure represent the ratio between the estimation using embedded events (Bkg.) and data (Obs.). The red line in the upper panel represents the estimation using simulation to estimate $\tau\tau$ backgrounds and the red dots in the lower panel represent the ratio between simulation and data.

Figure 5.3: Prefit distribution of $m_{\text{vis}}$ in the $e\tau_h$ (left) and $\mu\tau_h$ (right) final state.
Figure 5.4: Postfit distribution of Number of jets and $p_T$ of the leading jet in the $e \tau_h$ (left) and $\mu \tau_h$ (right) final state.
5.2 Application on Data

Table 5.3: Fit results for the $e\tau_h$ and $\mu\tau_h$ final state and the p-value obtained from the SAT, the KS and the AD goodness-of-fit test.

<table>
<thead>
<tr>
<th>Final state</th>
<th>Fit</th>
<th>SAT</th>
<th>KS</th>
<th>AD</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e\tau_h$, sim.</td>
<td>1.096 $^{+0.071}_{-0.066}$</td>
<td>0.433</td>
<td>0.783</td>
<td>0.455</td>
</tr>
<tr>
<td>$e\tau_h$, emb.</td>
<td>0.970 $^{+0.060}_{-0.056}$</td>
<td>0.224</td>
<td>0.601</td>
<td>0.279</td>
</tr>
<tr>
<td>$\mu\tau_h$, sim.</td>
<td>1.010 $^{+0.064}_{-0.060}$</td>
<td>0.101</td>
<td>0.970</td>
<td>0.491</td>
</tr>
<tr>
<td>$\mu\tau_h$, emb.</td>
<td>0.987 $^{+0.060}_{-0.057}$</td>
<td>0.554</td>
<td>0.931</td>
<td>0.302</td>
</tr>
</tbody>
</table>

on the expectation, where all nuisances are varied within they allowed uncertainty. For each p-value, 1000 pseudo experiments were performed.

Both the embedded events and the simulated events show a good agreement with the data, so both ways of estimating genuine $\tau\tau$ backgrounds are compatible with the measurements. All p-values indicate statistical compatibility with the data. The normalization of the $\mu \rightarrow \tau$ embedded events samples also shows a good agreement with the data.
Conclusion and Outlook

The embedding technique is an established method that provides a data driven estimation of SM processes resulting in a $\tau\tau$ final state. Compared to simulation, embedded events require significantly less tuning to provide accurate modeling of the data. This applies to quantities related to the underlying event or pileup collisions. Furthermore, the computational effort for the production of embedded events is significantly lower than for the production of an equal number of simulated events.

An estimation of the total yield of expected $\tau\tau$ events in data can be obtained directly from the embedded samples itself, making a scaling by the luminosity and cross section of production processes obsolete. The corresponding uncertainties, as well as uncertainties related to the jet energy scale, the missing transverse momentum scale and resolution, and the identification and misidentification of b jets do not affect embedded events.

Regarding the upcoming upgrade to the HL-LHC with an even higher number of pileup collisions, the method will gain even more importance. Data driven estimation methods often provide computationally more efficient solutions compared to the full simulation.

Within the scope of this thesis, embedded event samples in the $e\mu$, $e\tau_h$, $\mu\tau_h$, and $\tau_h\tau_h$ final states were produced using data measured in 2017. The intensive validation studies of the produced events show that the method is very robust. Correction factors are used to treat all remaining differences to data. Further improvements of the method are closely related to the details of the CMS detector structure and reconstruction algorithms. Issues related to the lepton isolation, the electron identification, the calorimeter cleaning, and the HLT reconstruction of $\tau_h$-candidates were studied and improved. In addition, the embedding method was applied in a $H \to \tau\tau$ target analysis and the determination of all necessary correction factors has been performed.

Distributions using embedded events are statistically compatible with observed data, as demonstrated using goodness-of-fit tests. Compared to the estimation using simulated events, embedded events do not require additional tuning, which results in reduced uncertainties. The data was recorded in 2017 using the CMS detector and corresponds to an integrated luminosity of $41.5\text{ fb}^{-1}$. More than 2.4 million CPU hours were invested to produce the embedded event samples that are available to all analysts within the CMS collaboration. The 2017 embedded event samples are currently used in an SM $H \to \tau\tau$ analysis [63] for the estimation of $\tau\tau$ backgrounds.
There are still several possibilities for improvement of the technique. By further utilizing the information that is used by the CMS Particle Flow algorithm to reconstruct muons during the selection step, the current cleaning approach can be further improved. In addition, studies and potential corrections of the embedded events can be applied to provide a description of $\tau$-lepton polarization effects in embedded events. This would allow the usage of embedded events in analyses targeting CP related properties. Another improvement would be the possibility to convert the energy deposits from the simulated decays into the alignment used in data, prior to the full reconstruction and particle identification. This would allow reconstruction effects, like the tracking efficiency to be taken into account since the full detector information is available. The embedding technique could also be modified to provide a data driven estimate of $Z \rightarrow \nu\nu$. Even an expansion of the method to estimate $W \rightarrow \nu\ell\ell$ backgrounds, where only one lepton decay has to be simulated, would be possible.
Bibliography


Appendix

B.1 Performance of Embedded Event Production

B.1.1 Details of the Production of Validation Samples

<table>
<thead>
<tr>
<th></th>
<th>Events</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input</td>
<td>65,170,166</td>
<td>201.22 TB</td>
</tr>
<tr>
<td>No. Jobs</td>
<td>65,173</td>
<td></td>
</tr>
<tr>
<td>Final State</td>
<td>$\mu \to e$</td>
<td>$\mu \to \mu$</td>
</tr>
<tr>
<td>Output</td>
<td>45,398,261</td>
<td>2222 TB</td>
</tr>
<tr>
<td>Size</td>
<td>2222.8 TB</td>
<td>2703.8 TB</td>
</tr>
<tr>
<td>Acceptance Rate</td>
<td>69.99%</td>
<td>99.0 %</td>
</tr>
</tbody>
</table>

B.1.2 Performance Plots

Here, additional performance measures of the embedded event production are shown.
Appendix

Figure B.1: Correlation plots between multiple variables. Each dot represents one single job, and the colors determines the final state.
Figure B.2: Comparison of embedded event production jobs. The colored box represents the 50% quantile, and the vertical line inside the box represents the mean CPU efficiency. The whiskers represent the 5 and 95% quantile and each small dot represents one single production job.

B.2 Validation Studies

B.2.1 $\mu \rightarrow \mu$ Embedding Validation Dataset

Table B.2: Details of the simulation sample used for the validation via $\mu \rightarrow \mu$ embedding

<table>
<thead>
<tr>
<th>Type</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>/DYJetsToLL_M-50_TuneCP5_13TeV-madgraphMLM-pythia8</td>
</tr>
<tr>
<td></td>
<td>/RunIIWinter17DR-NZSPU40to70_94X_upgrade2018_realtistic_v8-v3/GEN-SIM-RAW</td>
</tr>
<tr>
<td>Energy</td>
<td>13 TeV</td>
</tr>
<tr>
<td>GlobalTag</td>
<td>94X_upgrade2018_realtistic_v8</td>
</tr>
<tr>
<td>Number of Events</td>
<td>9239494</td>
</tr>
<tr>
<td>Preselection</td>
<td>$Z \rightarrow \mu\mu$ process, $p_T(\mu) &lt; 7$, $</td>
</tr>
</tbody>
</table>
### B.2.2 Electron Identification Variables

**Table B.3:** List of all variables used in the 2017 electron identification discriminator (egmGsfElectronIDs:mvaEleID-Fall17-v2)

<table>
<thead>
<tr>
<th>Variable Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position of the supercluster in $\eta$</td>
<td>$\eta_{SC}$</td>
</tr>
<tr>
<td>$\sigma_{\eta\eta}$</td>
<td>$\sigma_{\eta\eta} = \frac{\sum (\eta_i - \eta)^2 w_i}{\sum w_i}$. The sum runs over all ECAL cells in a 5x5 cluster, $\eta_i$ is the $\eta$ position of a cell and $w_i$ depends on the energy contained in a cell</td>
</tr>
<tr>
<td>$\sigma_{\phi\phi}$</td>
<td>$\sigma_{\phi\phi} = \frac{\sum (\phi_i - \phi)^2 w_i}{\sum w_i}$. The sum runs over all ECAL cells in a 5x5 cluster, $\phi_i$ is the $\phi$ position of a cell and $w_i$ depends on the energy contained in a cell</td>
</tr>
<tr>
<td>One minus the highest energy of a 1x5 cluster around the electron seed over the energy in a 5x5 cluster around the electron seed</td>
<td>$1 - E(1x5) E(5x5)$</td>
</tr>
<tr>
<td>Ratio of the energy in a 3x3 cluster around the electron calorimeter seed over the energy in a 5x5 cluster around the electron seed</td>
<td>$\frac{E(3x3)}{E(5x5)}$</td>
</tr>
<tr>
<td>Size of the supercluster in $\eta$</td>
<td>$\Delta \eta_{SC}$</td>
</tr>
<tr>
<td>Size of the supercluster in $\phi$</td>
<td>$\Delta \phi_{SC}$</td>
</tr>
<tr>
<td>See 4.1.2</td>
<td>$E_{SC}$</td>
</tr>
<tr>
<td>$\frac{E_{HCAL}}{E_{ECAL}}$</td>
<td>$E_{PS}$</td>
</tr>
<tr>
<td>$\frac{E_{PS}}{E_{RAW}}$</td>
<td>psEOverEraw</td>
</tr>
<tr>
<td>Ratio of Supercluster energy in the preshower over the raw energy.</td>
<td></td>
</tr>
<tr>
<td>Number of hits of the Kalman filter track</td>
<td>$N_{Hits}$</td>
</tr>
<tr>
<td>Fit quality of the Kalman filter track</td>
<td>$\chi^2$</td>
</tr>
<tr>
<td>Fit quality of the GSF track</td>
<td>$\chi^2$</td>
</tr>
<tr>
<td>Fit Number of hits of the GSF track</td>
<td>$N_{Hits}$</td>
</tr>
<tr>
<td>Number of hits from the GSF track</td>
<td>$N_{Miss}$</td>
</tr>
<tr>
<td>Number of expected missing Hits of the GSF track</td>
<td>$N_{Miss}$</td>
</tr>
<tr>
<td>Probability of a Conversion vertex</td>
<td>$P_{conv}$</td>
</tr>
</tbody>
</table>
### B.2 Validation Studies

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\frac{E_{SC}}{p_e}$</td>
<td>ep</td>
</tr>
<tr>
<td>$\frac{E_{SC}}{p_{out}}$</td>
<td>eelepout</td>
</tr>
<tr>
<td>$\frac{1}{E_{SC}} - \frac{1}{p_e}$</td>
<td>IoEmIop</td>
</tr>
</tbody>
</table>

#### Matching

| $|\Delta \eta_{in}|$ | deltaetain |
| $|\Delta \phi_{in}|$ | deltaphiin |
| $|\Delta \eta_{seed}|$ | deltaetaseed |

#### Misc.

| $\rho$ | rho |
| $p_{t,e}$ | pt |

- Ratio of the Supercluster energy over the electron track momentum extrapolated to the beamspot.
- Ratio of the Supercluster energy over the electron track momentum extrapolated to the outermost tracker hit.
- One over the Supercluster energy minus one over the electron track momentum.
- See 4.1.2
- See 4.1.2
- Difference in $\eta$ between the seed cluster and the track extrapolated from the outermost tracker hit to the ECAL surface.

- Number of Pileup per area.
- $p_T$ of the electron.
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Production and decay modes of the Higgs boson</td>
<td>7</td>
</tr>
<tr>
<td>2.1</td>
<td>A sketch of the CMS detector showing the different sub detectors.</td>
<td>10</td>
</tr>
<tr>
<td>2.2</td>
<td>Cross section of the CMS detector</td>
<td>11</td>
</tr>
<tr>
<td>2.3</td>
<td>CMS trigger setup</td>
<td>13</td>
</tr>
<tr>
<td>2.4</td>
<td>HLT $\tau_h$ reconstruction</td>
<td>14</td>
</tr>
<tr>
<td>2.5</td>
<td>Slice of CMS detector with particle signatures</td>
<td>16</td>
</tr>
<tr>
<td>2.6</td>
<td>Slice of the CMS tracker</td>
<td>18</td>
</tr>
<tr>
<td>3.1</td>
<td>Sketch of the Embedding Technique</td>
<td>24</td>
</tr>
<tr>
<td>3.2</td>
<td>Distribution of $m_{\mu\mu}$ after applying selection criteria</td>
<td>26</td>
</tr>
<tr>
<td>3.3</td>
<td>Example of the Calorimeter Cleaning</td>
<td>27</td>
</tr>
<tr>
<td>3.4</td>
<td>Acceptance Rate depending on the Number of decay simulation trials</td>
<td>29</td>
</tr>
<tr>
<td>3.5</td>
<td>Simulation of a decay during the generator step</td>
<td>30</td>
</tr>
<tr>
<td>3.6</td>
<td>Flowchart of the merging strategy</td>
<td>31</td>
</tr>
<tr>
<td>3.7</td>
<td>Embedded pre-selection flowchart</td>
<td>33</td>
</tr>
<tr>
<td>3.8</td>
<td>Embedded event production procedure</td>
<td>34</td>
</tr>
<tr>
<td>3.9</td>
<td>Average CPU efficiency of an embedded event production job</td>
<td>36</td>
</tr>
<tr>
<td>3.10</td>
<td>Average runtime of an embedded event production job</td>
<td>37</td>
</tr>
<tr>
<td>3.11</td>
<td>Runtime over disk usage of an embedded event production job</td>
<td>38</td>
</tr>
<tr>
<td>4.1</td>
<td>Comparison of $Z \rightarrow \mu\mu$ simulation and $\mu \rightarrow \mu$ embedding</td>
<td>40</td>
</tr>
<tr>
<td>4.2</td>
<td>$p_T$-flux of neutral hadrons</td>
<td>42</td>
</tr>
<tr>
<td>4.3</td>
<td>PF Neutral Hadron candidates in a $Z \rightarrow \mu\mu$ simulated event sample and $\mu \rightarrow \mu$ embedded event sample</td>
<td>43</td>
</tr>
<tr>
<td>4.4</td>
<td>Effects of FSR in the initial $Z \rightarrow \mu\mu$ selection.</td>
<td>44</td>
</tr>
<tr>
<td>4.5</td>
<td>Illustration of the calculation of the distance used in the alternative cleaning approach</td>
<td>45</td>
</tr>
<tr>
<td>4.6</td>
<td>Impacts of an alternative cleaning approach</td>
<td>46</td>
</tr>
<tr>
<td>4.7</td>
<td>Efficiency of the electron identification</td>
<td>47</td>
</tr>
<tr>
<td>4.8</td>
<td>Distribution of four electron identification input variables</td>
<td>49</td>
</tr>
<tr>
<td>4.9</td>
<td>Illustration of the merging issue of the innermost state</td>
<td>50</td>
</tr>
</tbody>
</table>
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.10</td>
<td>Distribution of $\Delta \eta_{\text{in}}$ and $\Delta \phi_{\text{in}}$ in $\mu \rightarrow e$ embedded event sample</td>
<td>51</td>
</tr>
<tr>
<td>4.11</td>
<td>Shift of the innermost track hit when changing the geometry</td>
<td>52</td>
</tr>
<tr>
<td>4.12</td>
<td>Distribution of $\Delta \eta_{\text{in}}$ and $\Delta \phi_{\text{in}}$ in $\mu \rightarrow e$ embedded event sample using 2016 data</td>
<td>53</td>
</tr>
<tr>
<td>4.13</td>
<td>$I_{\text{rel}}$ in the $e\tau_{h}$, and $\mu\tau_{h}$ final state</td>
<td>54</td>
</tr>
<tr>
<td>4.14</td>
<td>$m_{\text{vis}}$ in the $e\tau_{h}$, $\mu\tau_{h}$, $\tau_{h}\tau_{h}$, and $e\mu$ final state</td>
<td>55</td>
</tr>
<tr>
<td>4.15</td>
<td>Difference in the $\tau_{h}\tau_{h}$ final state when requiring a sensitive trigger path</td>
<td>57</td>
</tr>
<tr>
<td>4.16</td>
<td>Difference in the $\mu\tau_{h}$ final state when requiring a sensitive trigger path</td>
<td>57</td>
</tr>
<tr>
<td>4.17</td>
<td>HLT $\tau_{h}$ reconstruction filter rejection rates</td>
<td>58</td>
</tr>
<tr>
<td>4.18</td>
<td>New version of HLT $\tau_{h}$ reconstruction used in the embedding technique</td>
<td>60</td>
</tr>
<tr>
<td>4.19</td>
<td>Comparison of old and new HLT $\tau_{h}$ reconstruction sequence in the $\tau_{h}\tau_{h}$ and $\mu\tau_{h}$ channel in $m_{\text{vis}}$</td>
<td>60</td>
</tr>
<tr>
<td>5.1</td>
<td>Correction factors for Electrons and Muons</td>
<td>67</td>
</tr>
<tr>
<td>5.2</td>
<td>Postfit $m_{\text{vis}}$ in $e\tau_{h}$ (left) and $\mu\tau_{h}$ (right)</td>
<td>71</td>
</tr>
<tr>
<td>5.3</td>
<td>Pretfit $m_{\text{vis}}$ in $e\tau_{h}$ (left) and $\mu\tau_{h}$ (right)</td>
<td>71</td>
</tr>
<tr>
<td>5.4</td>
<td>Postfit distribution of $N_{\text{jets}}$ and $p_{\text{T}}$ of the leading jet in $e\tau_{h}$ (left) and $\mu\tau_{h}$ (right) final state</td>
<td>72</td>
</tr>
<tr>
<td>B.1</td>
<td>2D performance comparison of embedded event production</td>
<td>84</td>
</tr>
<tr>
<td>B.2</td>
<td>Performance comparison of embedded event production</td>
<td>85</td>
</tr>
</tbody>
</table>
# APPENDIX D

## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1</td>
<td>Decay modes of $\tau \tau$ pairs</td>
<td>8</td>
</tr>
<tr>
<td>3.1</td>
<td>Selection requirements for suitable $\mu \mu$ events</td>
<td>25</td>
</tr>
<tr>
<td>3.2</td>
<td>Selection requirements in the generator step</td>
<td>28</td>
</tr>
<tr>
<td>3.3</td>
<td>List of the different data tiers defined by the CMS collaboration</td>
<td>32</td>
</tr>
<tr>
<td>3.4</td>
<td>Production statistics of $\mu \rightarrow \tau$ embedded event production</td>
<td>35</td>
</tr>
<tr>
<td>3.5</td>
<td>Average runtime of the four different steps</td>
<td>35</td>
</tr>
<tr>
<td>4.1</td>
<td>Filter names corresponding to the steps used in Figure 4.17</td>
<td>58</td>
</tr>
<tr>
<td>5.1</td>
<td>Selection requirements for electron and muon TP pairs</td>
<td>65</td>
</tr>
<tr>
<td>5.2</td>
<td>Analysis selection for $e\tau_h$ and $\mu\tau_h$</td>
<td>68</td>
</tr>
<tr>
<td>5.3</td>
<td>Fit results in the $e\tau_h$ and $\mu\tau_h$ final state</td>
<td>73</td>
</tr>
<tr>
<td>B.1</td>
<td>Production statistics for $\mu \rightarrow e$ and $\mu \rightarrow \mu$ embedding</td>
<td>83</td>
</tr>
<tr>
<td>B.2</td>
<td>Details of simulated validation sample</td>
<td>85</td>
</tr>
<tr>
<td>B.3</td>
<td>Variables used for electron identification</td>
<td>86</td>
</tr>
</tbody>
</table>
Erklärung der selbständigen Anfertigung der Masterarbeit

Hiermit erkläre ich, dass ich die Masterarbeit mit dem Titel

»Production of Hybrid Data Samples for Data Driven Background Determination in the \(H \rightarrow \tau\tau\) Channel«

selbständig und unter ausschließlicher Verwendung der angegebenen Hilfsmittel angefertigt habe.

Sebastian Brommer
Karlsruhe, den 12. März 2019