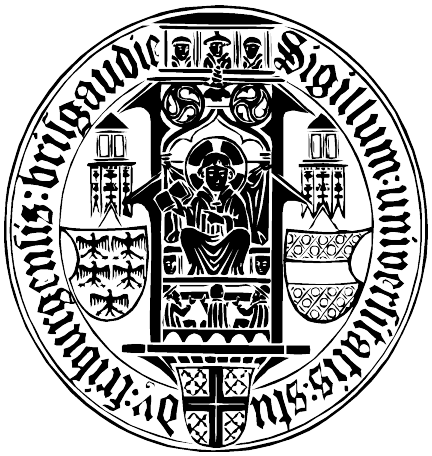


Search for lepton-flavour violation in $pp \to X \to \tau e(\tau\mu)$ decays with the ATLAS experiment

Master thesis

Emanuel Jona Dorbath

Supervisor: Prof. Dr. Markus Schumacher

Fakultät für Mathematik und Physik
Albert-Ludwigs-Universität Freiburg

May 2020

Erklärung

Hiermit erkläre ich, die vorliegende Masterarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt zu haben. Wörtlich oder inhaltlich verwendete Quellen wurden entsprechend den anerkannten Regeln wissenschaftlichen Arbeitens (lege artis) kenntlich gemacht. Ich erkläre weiterhin, dass die vorliegende Masterarbeit weder vollständig noch in wesentlichen Teilen Gegenstand eines anderen Prüfungsverfahrens war oder ist.

Freiburg, den _____ _____

Emanuel Dorbath

Abstract

The standard model of particle physics predicts lepton-flavour conversation in particle decays. An observation of a lepton-flavour violating decay would be a clear indication for physics beyond the standard model. In this thesis, the sensitivity of a search for the lepton-flavour violating decays of the Higgs boson and Z boson into a τe or a $\tau\mu$ pair is determined, considering fully leptonic decays of the τ-lepton. This analysis uses the full data set of proton-proton collisions recorded by the ATLAS detector at the LHC in Run 2 at a center-of-mass energy of $\sqrt{s} = 13\,\text{TeV}$, corresponding to an integrated luminosity of $139\,\text{fb}^{-1}$. Two methods are developed to improve the sensitivity. Firstly, the so-called fake lepton background is estimated using the fake factor method and a dedicated extraction region. Secondly, regression neural networks are trained to improve the mass reconstruction. The neural networks are trained on simulated $H \to \tau e$ events. A comparison of the neural networks to existing mass reconstruction methods is performed. For the neural networks, a better mass resolution than for the existing mass reconstruction methods is observed, but the sensitivity is not improved.

Zusammenfassung

Das Standardmodell der Teilchenphysik sagt eine Erhaltung der Lepton-Flavor-Zahl in Teilchenzerfällen vorher. Die Beobachtung eines Lepton-Flavor-verletzenden Zerfalls ist ein klarer Hinweis auf Physik jenseits des Standardmodells. In dieser Arbeit wird die Sensitivität für eine Suche nach Lepton-Flavor-verletzenden Zerfällen des Higgs-Bosons und des Z-Bosons in ein τe oder ein $\tau\mu$ Paar durchgeführt, wobei das τ-Lepton vollständig leptonisch zerfällt. Die Analyse verwendet den kompletten, vom ATLAS Detektor während Run 2 des LHCs aufgezeichneten Datensatz an Proton-Proton Kollisionen, mit einer Schwerpunktsenergie von $\sqrt{s} = 13\,\text{TeV}$ welcher einer integrierten Luminosität von $139\,\text{fb}^{-1}$ entspricht. Zwei Methoden werden entwickelt um die Sensitivität zu verbessern. Erstens, der Untergrund, verursacht durch so genannte „Fake“-Leptonen, wird mit Hilfe der „Fake Faktor“-Methode und einer speziellen Extraktionsregion bestimmt. Zweitens, eine verbesserte Massenrekonstruktion durch regressive neuronale Netze. Die neuronalen Netze wurden auf simulierte $H \to \tau e$ Ereignisse trainiert. Ein Vergleich der neuronalen Netze zu existierenden Massenrekonstruktionsmethoden wird durchgeführt. Für die neuronalen Netze wird, im Vergleich zu den existierenden Massenrekonstruktionsmethoden, eine verbesserte Massenauflösung beobachtet, jedoch keine Verbesserung der Sensitivität.

Contents

1 Introduction

The standard model of particle physics (SM) describes the properties and interactions of all known elementary particles and is the most consistent model with respect to the observations. This model is based on relativistic quantum field theories describing the electromagnetic, weak and strong interactions. Only the gravitational interaction is not described by the SM.
Until 2012, all elementary particles predicted by the SM were experimentally observed, except the scalar Higgs boson. The Higgs boson is predicted by the Englert-Brout-Higgs-Guralnik-Hagen-Kibble mechanism [1–6], denoted in the following simply as Higgs mechanism, which allows the description of massive elementary particles without violating the gauge symmetry of the SM. In 2012, the ATLAS (**A** **T**oroidal **L**HC **A**pparatu**S**) and CMS (**C**ompact **M**uon **S**olenoid) collaboration announced the observation of a Higgs boson [7, 8]. The Higgs boson mass was determined in a combined analysis as $m_H = 125.09 \pm 0.21(\text{stat.}) \pm 0.11(\text{sys.})\,\text{GeV}$ [9]. Subsequent measurements, like the production and decay rates [10] or the spin of the Higgs boson [11], were all in good agreement with the SM predictions.
Models beyond the SM (BSM) predict properties of the Higgs boson and Z boson, which differ from the ones in the SM. One of these is the lepton-flavour violating (LFV) decay of those bosons. Several BSM models predict LFV decays of the Higgs boson [12–27] and Z boson [28–30] and with the observation of neutrino oscillation [31] it is known that the lepton-flavour is not conserved. From indirect searches [32,33], the LFV decay $H \to \mu e$ is restricted to a branching ration $\lesssim 10^{-8}$ and is not likely to be observed. The upper limit for the branching ratios of $H \to \tau\mu$ and $H \to \tau e$ are set to 1.43 % and 1.04 % respectively by direct searches at the LHC [34]. For the LFV decays of the Z boson into a τe or $\tau\mu$ pair, upper limits of the order of 10^{-5} are set [34]. These are rates to which the ATLAS detector is sensitive and a search for BSM physics can be performed in a well motivated way.
In this thesis, data of proton-proton collisions recorded by the ATLAS detector in Run 2 of the LHC are used, which covers the data-taking years between 2015 and 2018 at a centre-of-mass energy of $\sqrt{s} = 13\,\text{TeV}$ corresponding to an integrated luminosity of $139\,\text{fb}^{-1}$. The increased luminosity after Run 1 results in more collisions and hence, enables more precise measurements of the SM as well as a more detailed search for physics beyond the SM. In this thesis, a sensitivity study to search for the LFV decays $H/Z \to \tau\mu$ and $H/Z \to \tau e$ is performed in a purely leptonic final state. This restricts the τ-lepton decay modes to $\tau \to e\bar{\nu}_e\nu_\tau$ and $\tau \to \mu\bar{\nu}_\mu\nu_\tau$. Furthermore, only different-flavour final states are considered. Hence, the final state is characterized by an electron, a muon and missing energy, arising from the neutrinos.
There are two main goals of this thesis. The first goal is the estimation of the so-called fake lepton background due to non-prompt leptons and particles misidentified as leptons. For this, the so-called fake factor method is used in a specifically defined extraction region. The second goal is the reconstruction of the Higgs boson and Z boson mass in the LFV decays, using a regression neural network and comparing its performance to existing methods.
In chapter 2, a theoretical description of the SM is presented, which includes a description of the individual Lagrangians, the absence of LFV decays in the SM as well as a model independent SM extension inducing LFV decays of the Higgs boson. Furthermore, the production and decay modes of the Higgs boson are presented. Chapter 3 provides an overview of the LHC and the

ATLAS detector as well as its individual detector parts. In chapter 4, the individual object reconstruction methods are presented. A fundamental overview of neural networks is presented in chapter 5, introducing the most important quantities, the training scheme as well as the optimization of neural networks. This is followed in chapter 6 by an overview of the used analysis strategy, a description of the so-called symmetry method as well as the importance of a precise mass reconstruction. In chapter 7, the individual simulated signal and background processes are described and the applied analysis selection is presented. In chapter 8 the estimation of the fake lepton background is described, which includes a description of the used extraction region, the kinematic dependencies and determination of the fake factor as well as a performed closure test. The mass reconstruction of the Higgs boson and Z boson via a regression neural network is presented in chapter 9. This covers the simulated events which are used to train the neural network, the performed optimization as well as a comparison to existing mass reconstruction. This thesis is summarised in chapter 10.

2 Theory

This chapter provides an overview of the theoretical concepts used in this thesis. Firstly, the Standard Model (SM) of particle physics is introduced, which describes the observed elementary particles and their interactions. The achievements and limitations are presented as well as a description of the SM as quantum field theory. In section 2.2, an extension of the SM is presented, inducing LFV Higgs boson decays. This is followed in section 2.3 with a more precise description of the Higgs boson, its production and decay modes in the SM as well as the induced LFV decays.

2.1 Standard Model of Particle Physics

The Standard Model of particle physics describes the observed elementary particles, their known interactions and properties. This includes all known fundamental interactions except gravitation. At the energy scales of present particle accelerators this is sufficient, since the gravitational force is negligible on the relevant distances of the interactions.

All elementary particles of the SM are classified either as *fermions* or as *bosons*, depending on their intrinsic angular momentum, called *spin*. The spin quantum number of fermions is a half integer, while for bosons it is an integer. All elementary fermions have a spin quantum number of 1/2.

Fermions are separated further into two different classes: *quarks* and *leptons*. Quarks are fermions with a colour charge, while leptons do not have any colour charge. Since the strong interaction couples to the colour charge, quarks take part in it while leptons do not. Both, quarks and leptons, are separated further into two types of particles. Quarks with an electric charge of $+2/3\,e$ are called up-type quarks. These are the *up* (u), *charm* (c) and *top* (t) quark. The other three quark types, with an electric charge of $-1/3\,e$, are called down-type quarks. These are the *down* (d), *strange* (s) and *bottom* (b) quark. In the SM, the three colour charges *red*, *blue* and *green* are present. Hence, each quark occurs three times in the SM, once for each colour charge.

Similarly to quarks, leptons are separated by their electric charge, resulting in an electrically charged lepton sector and an electrically neutral sector. The charged lepton sector contains the *electron* (e), *muon* (μ) and *tau* (τ) leptons, which have the electric charge $-e = -1.602 \cdot 10^{-19} C$. Each charged lepton has a neutral partner, called *neutrino*. They are denoted as ν_e, ν_μ and ν_τ respectively. In the SM, the neutrinos are assumed to be massless.

For every fermion in the SM, there exists a corresponding anti-fermion with the same mass, spin and absolute value of the electric charge. However, the sign of the electric charge is inverted.

All fermions of one certain type, i.e. up-type quarks, down-type quarks, charged leptons and neutrinos, differ only in their mass and flavour. This allows to split the SM fermions into three so-called *generations*. For each generation, the leptons and quarks have a separate flavour quantum number, e.g. u and d have a common flavour quantum number. The flavour quantum number is an additive quantum number. Anti-particles have the same flavour quantum number as the particle, but with an inverted sign. For the leptons, this results in the three flavour quantum numbers L_e, L_μ and L_τ, where $L_{\ell^-} = L_{\nu_\ell} = +1$ for the respective particles and $L_{\ell^+} = L_{\bar{\nu}_\ell} = -1$ for the anti-particles. In table 2.1, all SM fermions are summarized.

The SM conserves the lepton-flavour quantum number, in both the neutrino sector and the charged lepton sector. In the electromagnetic and strong interaction, the quark-flavour is conserved as well. However, the quark-flavour is violated by the charged weak current interactions, described with the Cabibbo-Kobayashi-Maskawa (CKM) matrix.
Quarks do not exist as free particle, but only in colour neutral bound states. The most common bound states are the baryons and mesons. Baryons consist of three quarks, whereas mesons are made up by a quark anti-quark pair. The 'regular' matter we know consists of the two baryons, the proton (uud) and the neutron (udd), and the electron. Besides the proton and the electron no other heavy elementary or composed stable particle is known yet.

Table 2.1: Mass and electric charge of all three generations of fermions in the SM [35]. The uncertainty on the charged lepton masses is at least four orders of magnitude smaller than the stated mass value.

	Generation			Electric
	1	2	3	charge [e]
Up-type Quarks	up (u)	charm (c)	top (t)	$+2/3$
Mass [MeV]	$2.2^{+0.5}_{-0.4}$	1275^{+25}_{-35}	$(173 \pm 0.4) \times 10^3$	
Down-type Quarks	down (d)	strange (s)	bottom (b)	$-1/3$
Mass [MeV]	$4.7^{+0.5}_{-0.3}$	95^{+9}_{-3}	$(4.18^{+0.04}_{-0.03}) \times 10^3$	
Charged Leptons	electron (e)	muon (μ)	tau (τ)	-1
Mass [MeV]	0.511	106	1.78×10^3	
Neutrinos	electron neutrino (ν_e)	muon neutrino (ν_μ)	tau neutrino (ν_τ)	0
Mass [eV]	< 2 (95% CL)	$< 0.19 \times 10^6$ (90% CL)	$< 18.2 \times 10^6$ (95% CL)	

The vector bosons of the SM have spin one and are the mediators of the elementary forces between the elementary particles. The Higgs boson is the only boson in the SM which does not have spin one, but spin zero. The massless *photon* (γ) mediates the electromagnetic force by coupling to the electric charge. Since it is not electrically charged, the photon does not couple to itself. It is furthermore neutral with respect to the colour and the weak charge. The strong interaction is mediated by eight massless *gluons* (g). These are all electrically neutral but carry colour charge themselves and hence, do interact with each other.
The massive spin-1 bosons of the SM are the Z, W^+ and W^- boson. These are the mediators of the weak interaction and couple to the weak isospin, which is in its absolute value for all left-handed fermions 1/2 and for all right-handed fermions zero. Both W bosons carry electric charge and the third component of the weak isospin is in its absolute value unity. The Z boson is electrically neutral and its third component of the weak isospin is zero. The weak hypercharge $\mathrm{Y_W}$ is connected, together with the third component of the weak isospin $\mathrm{I_{W,3}}$, to the electric charge Q by the Gell-Mann-Nishijima relation $Q = \mathrm{I_{W,3}} + \mathrm{Y_W}/2$. Hence, all three massive bosons have a weak hypercharge of zero. Furthermore, the massive bosons do interact with each other. All three massive bosons do not have any colour charge. In table 2.2, the three fundamental interactions are summarized.
The Higgs boson completes the SM of particle physics. It is neutral with respect to electric and colour charge, has a spin zero, a third component of the weak isospin of $-1/2$ and couples to

Table 2.2: Mediators of the three fundamental SM interactions and their mass and electric charge [35].

Interaction	Mediator	Mass [GeV]	Electric charge [e]
Strong	gluon (g)	0	0
Electromagnetic	photon (γ)	0	0
Weak	Z boson (Z)	91.2	0
	W boson ($W^{\pm}$)	80.4	± 1

all massive elementary particles. In 2012, a candidate for the SM Higgs boson with a mass of approximately 125 GeV was observed by ATLAS [7] and CMS [8] at the LHC. So far, all measured properties are consistent with the SM predictions, but more precise measurements are needed to study its nature in more detail.
The SM is a very good model to describe the physics of elementary particles at energies of present particle colliders, however there are still limitations. The model does neither include gravity nor any dark matter candidate. Furthermore, it cannot explain, for instance, the existence of exactly three generations of fermions. Finally, the individual particle masses and parameter constants in the model must be determined experimentally and are not predicted by the model itself.
In the following, the SM is described on basis of theoretical concepts, which is based on Ref. [36].

2.1.1 The Standard Model Gauge Structure

The SM is described in a theoretical way as a local gauge invariant quantum field theory (QFT). The gauge structure is given as $\mathrm{SU(3)_C} \times \mathrm{SU(2)_{I_W}} \times \mathrm{U(1)_{Y_W}}$, with the $\mathrm{U(1)_{Y_W}}$ group of weak hypercharge $\mathrm{Y_W}$, the $\mathrm{SU(2)_{I_W}}$ group of weak isospin $\mathrm{I_W}$ and the $\mathrm{SU(3)_C}$ group of colour charge. Particles in a quantum field theory are described as quantized fields, which obey the laws of both, special relativity and quantum mechanics. Any naive mass term in the SM Lagrangian would break the gauge invariance and is therefore not included. The so-called *Higgs mechanism* describes a spontaneous symmetry breaking via the introduction of a scalar field, the Higgs field, which leads to the observed massive gauge bosons and fermions.
The SM Lagrangian $\mathcal{L}_{\mathrm{SM}}$ is given by

$$\mathcal{L}_{\mathrm{SM}} = \mathcal{L}_{\mathrm{YM}} + \mathcal{L}_{\mathrm{ferm}} + \mathcal{L}_{\mathrm{H}} + \mathcal{L}_{\mathrm{Yuk}} \tag{2.1}$$

with the individual parts discussed in the following.
The Yang-Mills Lagrangian $\mathcal{L}_{\mathrm{YM}}$ describes the free propagation and self-interaction of the gauge fields. It is given by

$$\mathcal{L}_{\mathrm{YM}} = -\frac{1}{4} W^i_{\mu\nu} W^{i,\mu\nu} - \frac{1}{4} B_{\mu\nu} B^{\mu\nu} - \frac{1}{4} G^a_{\mu\nu} G^{a,\mu\nu} \tag{2.2}$$

with the field-strength tensors

$$\begin{aligned} W^i_{\mu\nu} &= \partial_\mu W^i_\nu - \partial_\nu W^i_\mu - g\epsilon^{ijk} W^j_\mu W^k_\nu, && i,j,k = 1,2,3 \\ B_{\mu\nu} &= \partial_\mu B_\nu - \partial_\nu B_\mu, \\ G^a_{\mu\nu} &= \partial_\mu G^a_\nu - \partial_\nu G^a_\mu - g_s f^{abc} G^b_\mu G^c_\nu, && a,b,c = 1,\ldots,8 \end{aligned} \tag{2.3}$$

of the respective gauge fields. The gauge field B_μ corresponds to the $\mathrm{U(1)_{Y_W}}$ group, W^i_μ to the $\mathrm{SU(2)_{I_W}}$ group and G^a_μ the $\mathrm{SU(3)_C}$ group. The gauge couplings of the individual groups are

denoted as g', g and g_s. For the non-abelian groups, $\mathrm{SU(2)_W}$ and $\mathrm{SU(3)_C}$, the structure constants defined by the commutation relations are noted as ϵ^{ijk} and f^{abc} respectively.
The second term in equation (2.1) describes the interaction between the gauge fields and the fermions, given by

$$\mathcal{L}_{\mathrm{ferm}} = i\bar{\Psi}_L \not{D} \Psi_L + i\bar{\psi}_{\ell_R} \not{D} \psi_{\ell_R} + i\bar{\Psi}_Q \not{D} \Psi_Q + i\bar{\psi}_{u_R} \not{D} \psi_{u_R} + i\bar{\psi}_{d_R} \not{D} \psi_{d_R} \tag{2.4}$$

where $\Psi_L = (\psi_{\nu_{\ell_L}}, \psi_{\ell_L})^\intercal$ are the left-handed $\mathrm{SU(2)_W}$ spinor doublets of the charged leptons $\ell = \mathrm{e}, \mu, \tau$ and the neutrinos $\nu_\ell = \nu_\mathrm{e}, \nu_\mu, \nu_\tau$. The left-handed quark doublet $\Psi_Q = (\psi_{u_L}, \psi_{d_L})^\intercal$ contains the spinor of one up-type quark $u = u, c, t$ and one down-type quark $d = d, s, b$ of the same generation. The respective right-handed $\mathrm{SU(2)_W}$ singlets are denoted as ℓ_R, u_R and d_R. Right-handed neutrinos are not present in the SM and thus no term with ν_{ℓ_R} is included.
The interaction between fermions is included in the covariant derivative

$$D_\mu = \partial_\mu + ig\mathrm{I}^i_\mathrm{W} W^i_\mu + ig' \frac{\mathrm{Y_W}}{2} B_\mu + ig_s T^a_\mathrm{C} G^a_\mu \tag{2.5}$$

which can be written with the Feynman slash notation as $\not{D} = D_\mu \gamma^\mu$. The generators of the gauge groups are denoted as $\mathrm{Y_W}$, I^i_W and T^a_C. With $\mathrm{I}^i_\mathrm{W} = \sigma^i/2$ (σ^i = Pauli matrices) for the left-handed $\mathrm{SU(2)_W}$ doublets and $\mathrm{I}^i_\mathrm{W} = 0$ for the right-handed singlets it follows that the W bosons only couple to left-handed fermions and are thus maximally parity violating. Finally, the generators for the $\mathrm{SU(3)_C}$ group are $T^a_\mathrm{C} = \lambda^a/2$ for the quark triplets, with the Gell-Mann matrices λ^a, and $T^a_\mathrm{C} = 0$ for leptons.
So far, the stated $\mathrm{SU(2)_W} \times \mathrm{SU(1)_{Y_W}}$ gauge fields are presented in their gauge basis (W^i_μ, B_μ) and not in their mass basis, i.e. $W^\pm_\mu$, Z_μ, A_μ which correspond to the particles $W^\pm$, Z and γ. With the requirement that the coupling structure of the photon is the same as in quantum electrodynamics, i.e. blind to parity and proportional to $Q\bar{\psi}\not{A}\psi$, the photon field A_μ and the Z boson field Z_μ are obtained as

$$\begin{pmatrix} Z_\mu \\ A_\mu \end{pmatrix} = \begin{pmatrix} \cos\theta_\mathrm{W} & -\sin\theta_\mathrm{W} \\ \sin\theta_\mathrm{W} & \cos\theta_\mathrm{W} \end{pmatrix} \begin{pmatrix} W^3_\mu \\ B_\mu \end{pmatrix} \tag{2.6}$$

which corresponds to a rotation of the basis. The weak mixing angle θ_W is given by

$$\cos\theta_\mathrm{W} = c_\mathrm{W} = \sqrt{1 - s^2_\mathrm{W}} = \frac{g}{\sqrt{g^2 + g'^2}} \tag{2.7}$$

and the electromagnetic coupling strength e by

$$e = \frac{gg'}{\sqrt{g^2 + g'^2}}. \tag{2.8}$$

With this, the remaining two fields are derived as

$$W^\pm_\mu = \left(W^1_\mu \mp iW^2_\mu\right)/\sqrt{2} \tag{2.9}$$

which correspond to the charged gauge bosons $W^\pm$.
The description of the SM up to here does not allow for mass terms for any of the so far mentioned particles. Any naive mass term, proportional to $W^i_\mu W^{i,\mu}$ for the W bosons or $\bar{\psi}_{f_L}\psi_{f_R} + \bar{\psi}_{f_R}\psi_{f_L}$ for fermions violates the gauge invariance. The introduction of the Higgs field allows particle masses while maintaining the gauge invariance.

2.1.2 Electroweak Symmetry Breaking

The Lagrangian of the Higgs sector in equation (2.1) is given by

$$\mathcal{L}_H = (D_\mu \Phi)^\dagger (D^\mu \Phi) - V(\Phi), \tag{2.10}$$

which introduces a new complex scalar $\mathrm{SU(2)_{I_W}}$ doublet $\Phi = (\phi^+, \phi^0)^\intercal$. This doublet has a weak hypercharge of $\mathrm{Y_{W,\Phi}} = 1$, with ϕ^+ carrying the charge $+e$ and ϕ^0 being neutral. The second term in equation (2.10) describes the self-interaction of Φ by the potential

$$V(\Phi) = -\mu^2(\Phi^\dagger\Phi) + \frac{\lambda}{4}(\Phi^\dagger\Phi)^2 \tag{2.11}$$

which is constrained such that the model remains gauge invariant and is renormalizable. This results in the two real free parameters μ^2 and λ in $V(\Phi)$. The latter is required to be greater zero to maintain vacuum stability. If μ^2 is positive as well, it gives rise to a non-vanishing vacuum expectation value (VEV) v. Finding the minimum of the potential $V(\Phi)$ results in

$$\Phi_0^\dagger \Phi_0 = v^2/2 \tag{2.12}$$

where Φ_0 is the vacuum state and $v^2 = 2\mu^2/\lambda = 1/\left(\sqrt{2}\mathrm{G_F}\right)$, with the Fermi constant $\mathrm{G_F}$. Since the vacuum is electrically neutral, the upper component of Φ_0 must vanish. Hence, Φ_0 is fixed up to a phase and can be chosen as $\Phi_0 = (0, v/\sqrt{2})^\intercal$.
The freedom to choose the vacuum state Φ_0 reflects the spontaneous breakdown of the initial $\mathrm{SU(2)_{I_W}} \times \mathrm{U(1)_{Y_W}}$ symmetry to the electromagnetic $\mathrm{U(1)_{EM}}$ symmetry. Expanding the Higgs field around the vacuum yields

$$\Phi = \begin{pmatrix} \phi^+ = \phi_1 + i\phi_2 \\ (v + H + i\chi)/\sqrt{2} \end{pmatrix} \tag{2.13}$$

where H is the real and scalar Higgs field. The three unphysical Goldstone boson fields are the complex field $\phi^+ = \phi_1 + i\phi_2$ and the real field χ. These are connected to the vacuum state by gauge transformations in such a way, that a gauge can be found, known as *unitary-gauge*, in which these three fields vanish.
The covariant derivative in equation (2.10) contains the interaction between the gauge boson fields and Φ. With $\mathrm{Y_{W,\Phi}} = 1$, $\mathrm{I^i_{W,\Phi}} = \sigma^i/2$ and $T^a_{\mathrm{C},\Phi} = 0$ the covariant derivative can be rewritten in the mass basis as

$$D_\mu = \partial_\mu + \frac{i}{2}\begin{pmatrix} \frac{g^2-g'^2}{\sqrt{g^2+g'^2}}Z_\mu + 2eA_\mu & \sqrt{2}gW_\mu^+ \\ \sqrt{2}gW_\mu^- & -\frac{g}{c_\mathrm{W}}Z_\mu \end{pmatrix} \tag{2.14}$$

where equations (2.6-2.9) are used. In the unitary-gauge, the Lagrangian $\mathcal{L}_H$ can be written as

$$\begin{aligned}\mathcal{L}_{H,U} =& \frac{1}{2}\partial_\mu H \partial^\mu H + \frac{g^2}{4}(v+H)^2 W_\mu^+ W^{-\mu} + \frac{g^2}{8c_\mathrm{W}^2}(v+H)^2 Z_\mu Z^\mu \\ &+ \frac{\mu^2}{2}(v+H)^2 - \frac{\lambda}{16}(v+H)^4\end{aligned} \tag{2.15}$$

where the identity $W_\mu^{+\dagger} = W_\mu^-$ is used in the second term. This Lagrangian contains bilinear terms in $W_\mu^\pm$, Z_μ and H. These represent the mass terms for the corresponding gauge bosons $W^\pm$, Z and the Higgs boson H. The respective masses are given by

$$M_W = \frac{gv}{2}, \qquad M_Z = \frac{M_W}{c_\mathrm{W}}, \qquad M_H = \sqrt{2\mu^2} = \sqrt{\lambda v^2} \tag{2.16}$$

where both W bosons share the same mass M_W. With the derived mass terms, μ^2, v and λ can be eliminated from equation (2.15), given as

$$\begin{aligned}\mathcal{L}_{H,U} =& \frac{1}{2}\partial_\mu H \partial^\mu H + \frac{1}{2}M_H^2 H^2 - \frac{g M_H^2}{4 M_W} H^3 - \frac{g^2 M_H^2}{32 M_W^2} H^4 \\ &+ M_W^2 W_\mu^+ W^{-\mu} + \frac{1}{2} M_Z^2 Z_\mu Z^\mu \\ &+ g M_W H W_\mu^+ W^{-\mu} + \frac{g^2}{4} H^2 W_\mu^+ W^{-\mu} + \frac{g M_Z}{2 c_\mathrm{W}} H Z_\mu Z^\mu + \frac{g^2}{4 c_\mathrm{W}^2} H^2 Z_\mu Z^\mu, \end{aligned} \tag{2.17}$$

where an irrelevant constant is neglected. The first line describes the kinematics and self-interaction of the Higgs field. For this, the electrically neutral, spinless Higgs boson is associated with a mass M_H. This reduces the four degrees of freedom that the doublet Φ initially has by one. The remaining degrees of freedom are used up by the now massive gauge fields $W_\mu^\pm$ and Z_μ, described in the second line of equation (2.17). In the last line, the interaction between the Higgs boson and the weak gauge bosons $W^\pm$ and Z is described.
The Lagrangian $\mathcal{L}_{H,U}$ predicts the massive, scalar Higgs boson, three massive gauge bosons and the massless electromagnetic photon field A_μ. The masses of the W bosons and the Z boson are connected by the factor $\cos\theta_\mathrm{W}$. Furthermore, interactions between one or two Higgs bosons with the massive gauge bosons are predicted, as well as self-interactions of up to four Higgs bosons. Despite the massive gauge bosons, the model remains gauge invariant.
Both parameters, μ^2 and λ, are fixed by the mass of the Higgs boson and the VEV of $v \approx 246\,\mathrm{GeV}$.

2.1.3 Yukawa couplings

Besides the interaction with the massive gauge bosons, the Higgs boson also couples to fermions called *Yukawa couplings*. In fact, these couplings give fermions their masses. The Lagrangian of the coupling of left-handed fermion doublets Ψ with right-handed fermion singlets ψ to the Higgs doublet Φ is given in the most general form as

$$\mathcal{L}_\mathrm{Yuk} = -\bar{\Psi}_L G_\ell \psi_{\ell_R} \Phi - \bar{\Psi}_Q G_u \psi_{u_R} \tilde{\Phi} - \bar{\Psi}_Q G_d \psi_{d_R} \Phi + \mathrm{h.c.} \tag{2.18}$$

where h.c. corresponds to the hermitian conjugate and $\tilde{\Phi} = i\sigma^2 \Phi^*$ is the charge-conjugate Higgs doublet. Here, Ψ and ψ are three-spinors represented in flavour space. The arbitrary 3×3 matrices G_f, with $f = \ell, u, d$, contain the coupling strengths between the different fermion-flavours. Hence, the off-diagonal elements of G_f mix left- and right-handed fermions of different generations.
By transforming the denoted *flavour basis* $(\psi_{f_{\tau,1}}, \psi_{f_{\tau,2}}, \psi_{f_{\tau,3}})$, with $\tau = L, R$, into a *mass basis* $(\hat{\psi}_{f_{\tau,1}}, \hat{\psi}_{f_{\tau,2}}, \hat{\psi}_{f_{\tau,3}})$ the off-diagonal elements in G_f become zero. This transformation is done by a unitary matrix U,

$$\hat{\psi}_{f_{\tau,i}} = U_{ij}^{f_\tau} \psi_{f_{\tau,j}}, \qquad i,j = 1,2,3 \tag{2.19}$$

which also transforms the matrices G_f into their diagonal form

$$U^{f_L} G_f U^{f_R \dagger} = \frac{\sqrt{2}}{v} \mathrm{diag}\,(m_{f_1}, m_{f_2}, m_{f_3}) \tag{2.20}$$

with the mass m_{f_i} of the fermion f_i. The unitary matrix for the charged leptons U^{ℓ_τ} can also be used to transform the left-handed neutrino fields as long as the neutrinos are considered massless.

With $\bar{\psi}_{f_\tau} = \bar{\hat{\psi}} U^{f_\tau}$, the Lagrangian in equation (2.18) becomes in the unitary-gauge

$$\mathcal{L}_{\text{Yuk},U} = -\sum_f \sum_i m_{f_i} \left(1 + \frac{H}{v}\right) \left(\bar{\psi}_{f_{i_L}} \psi_{f_{i_R}} + \bar{\psi}_{f_{i_R}} \psi_{f_{i_L}}\right) \tag{2.21}$$

where the first sum runs over the three fermion types ℓ, u, d and the second one over the three generations. This Lagrangian contains two important parts: the first one assigns a mass m_{f_i} to each fermion. The second one describes the coupling of the Higgs boson to all massive fermions with strength m_{f_i}/v. Hence, heavy fermions are more likely to interact with the Higgs boson. By construction, no mixing term remains.

2.1.4 Flavour Violation in the SM

The basis transformation of the fermions in the Yukawa Lagrangian $\mathcal{L}_{\text{Yuk}}$ must also be carried out in the fermion Lagrangian $\mathcal{L}_{\text{ferm}}$, given in equation (2.4). With $\Psi_L = (\psi_{\nu_{\ell_L}}, \psi_{\ell_L})^\intercal$ and the transformation stated in equation (2.19), the Lagrangian can be written as

$$\begin{aligned}
\mathcal{L}_{\text{ferm}} =& i \left(\bar{\hat{\psi}}_{\nu_{\ell_L}} U^{\ell_L}, \bar{\hat{\psi}}_{\ell_L} U^{\ell_L}\right) \not{D} \begin{pmatrix} (U^{\ell_L})^\dagger \hat{\psi}_{\nu_{\ell_L}} \\ (U^{\ell_L})^\dagger \hat{\psi}_{\ell_L} \end{pmatrix} + i \bar{\hat{\psi}}_{\ell_R} U^{\ell_R} \not{D} (U^{\ell_R})^\dagger \hat{\psi}_{\ell_R} \\
&+ i \left(\bar{\hat{\psi}}_{u_L} U^{u_L}, \bar{\hat{\psi}}_{d_L} U^{d_L}\right) \not{D} \begin{pmatrix} (U^{u_L})^\dagger \hat{\psi}_{u_L} \\ (U^{d_L})^\dagger \hat{\psi}_{d_L} \end{pmatrix} \\
&+ i \bar{\hat{\psi}}_{u_R} U^{u_R} \not{D} (U^{u_R})^\dagger \hat{\psi}_{u_R} + i \bar{\hat{\psi}}_{d_R} U^{d_R} \not{D} (U^{d_R})^\dagger \hat{\psi}_{d_R},
\end{aligned} \tag{2.22}$$

where the neutrinos are transformed with the same unitary matrix as the charged leptons, since the neutrinos are approximated as being massless in the SM. The up-type quarks and down-type quarks are transformed separately by the respective matrices.
Equation (2.22) can be simplified by commuting U^{f_τ} and $\not{D}$. This is possible, since U^{f_τ} is defined in the flavour space and $\not{D}$ is proportional to the identity matrix in the flavour space [37]. Furthermore, the identity of the unitary matrices is used, i.e. $U^{f_\tau}(U^{f_\tau})^\dagger = I$, where I is the identity matrix. With this, the left-handed lepton doublet and the right-handed fermion singlet terms are simplified. The Lagrangian is given by

$$\begin{aligned}
\mathcal{L}_{\text{ferm}} =& i \bar{\hat{\Psi}}_L \not{D} \hat{\Psi}_L + i \bar{\hat{\psi}}_{\ell_R} \not{D} \hat{\psi}_{\ell_R} \\
&+ i \left(\bar{\hat{\psi}}_{u_L} U^{u_L}, \bar{\hat{\psi}}_{d_L} U^{d_L}\right) \not{D} \begin{pmatrix} (U^{u_L})^\dagger \hat{\psi}_{u_L} \\ (U^{d_L})^\dagger \hat{\psi}_{d_L} \end{pmatrix} + i \bar{\hat{\psi}}_{u_R} \not{D} \hat{\psi}_{u_R} + i \bar{\hat{\psi}}_{d_R} \not{D} \hat{\psi}_{d_R}
\end{aligned} \tag{2.23}$$

with $\hat{\Psi}_L = (\hat{\psi}_{\nu_{\ell_L}}, \hat{\psi}_{\ell_L})^\intercal$. Only the term of left-handed quarks can not be simplified, since the up-type and down-type quarks are transformed with two distinct matrices. The off-diagonal elements of the covariant derivative $\not{D}$ result in a mixture of $\hat{\psi}_{u_L}$ and $\hat{\psi}_{d_L}$ which causes bilinear terms in U^{u_L} and U^{d_L}. These matrices are in general not identical and hence, the basis transformation has a non-vanishing effect in the quark sector. However, the diagonal elements of $\not{D}$ are not affected by this transformation, but only the off-diagonal elements. The latter elements originate due to the charged weak current, i.e. the coupling to the $W^\pm$ bosons.
Separating the interaction of the charged weak current with the left-handed quarks from the other interactions, the corresponding Lagrangian is given by

$$\mathcal{L}_{\text{ferm}|_{q \leftrightarrow W^\pm}} = -\frac{1}{2} \left(\bar{\hat{\psi}}_{u_L} U^{u_L}, \bar{\hat{\psi}}_{d_L} U^{d_L}\right) \begin{pmatrix} 0 & \sqrt{2} g W_\mu^+ \\ \sqrt{2} g W_\mu^- & 0 \end{pmatrix} \gamma^\mu \begin{pmatrix} (U^{u_L})^\dagger \hat{\psi}_{u_L} \\ (U^{d_L})^\dagger \hat{\psi}_{d_L} \end{pmatrix}$$

$$= -\frac{1}{\sqrt{2}} g \left(\bar{\hat{\psi}}_{u_L} V W_\mu^+ \gamma^\mu \hat{\psi}_{d_L} + \bar{\hat{\psi}}_{d_L} V^\dagger W_\mu^- \gamma^\mu \hat{\psi}_{u_L} \right)$$

where the matrix $V = U^{u_L}(U^{d_L})^\dagger$ is the Cabbibo-Kobayashi-Maskawa (CKM) [38, 39] matrix. In the entire SM, the CKM matrix is the only remaining effect from the fermion basis transformation. However, the elements of the CKM matrix are not predicted by the SM, but must be derived experimentally. The CKM matrix is found to be non-diagonal. This means that in the interaction with a charged weak boson, different quark-flavours mix and hence, the quark-flavour is not a conserved quantity in the SM.
In the lepton sector of the SM, no such non-diagonal matrix is present, since both, charged leptons and neutrinos are transformed by the same unitary matrix. Hence, no lepton-flavours are mixed in the weak interaction and therefore no lepton-flavour violation is present. To summarize, any lepton-flavour violating term in the Yukawa coupling is eliminated by the transformation of the fermions into the *mass basis*, which does not result in any new lepton-flavour violating process in the remaining parts of the SM. The lepton-flavour is a conserved quantity in the SM.

2.2 Charged Lepton-Flavour Violation

The SM does not predict any charged lepton-flavour violation (LFV). However, several models beyond the SM (BSM) predict a lepton-flavour violating Higgs boson decay. This includes models with more than one Higgs doublet [12–15], composite Higgs models [16, 17] and many others [18–27]. The individual models are not discussed, but an effective and model-independent LFV Lagrangian $\mathcal{L}_{\text{LFV}}$ is introduced as extension to the SM Lagrangian.
This effective Lagrangian contains non-diagonal couplings between the Higgs field and the charged leptons, inducing the LFV Higgs boson decays. It can be written as

$$\mathcal{L}_{\text{LFV}} = \frac{c_{ij}}{\sqrt{2}} H \bar{\hat{\psi}}_{\ell_{i_L}} \hat{\psi}_{\ell_{j_R}}, \quad i,j = 1,2,3, \quad c_{ii} = 0 \tag{2.24}$$

with the coupling strength c_{ij}. The spinors of the left-handed charge leptons of generation i are denoted as $\hat{\psi}_{\ell_{i_L}}$ and the right-handed ones as $\hat{\psi}_{\ell_{i_R}}$. Since no flavour diagonal coupling ($i = j$) is added to the SM Lagrangian by $\mathcal{L}_{\text{LFV}}$, the coupling of the Higgs field to same-flavour leptons is unaffected. However, the non-diagonal elements cause LFV Higgs boson decays. Similarly to the elements of the CKM matrix, the individual coupling strength must be derived experimentally.
In the SM, lepton-flavour violating terms were eliminated by performing a basis transformation, i.e. from the *flavour basis* into the *mass basis*. This was possible because the matrix describing the fermion couplings to the Higgs boson is proportional to the corresponding mass matrix and therefore, both are diagonal in the *mass basis*. The introduced Lagrangian $\mathcal{L}_{\text{LFV}}$ only adds couplings between the fermions and the Higgs boson, but no mass terms. This prevents the removal of $\mathcal{L}_{\text{LFV}}$ by changing the fermion basis again. Thus, $\mathcal{L}_{\text{LFV}}$ enables LFV Higgs boson decays in the charged lepton sector.

2.3 The Higgs Boson

The Higgs mechanism in the SM explicitly predicts the existence of an additional scalar and electrically neutral boson, the Higgs boson. In 2012, both the ATLAS and CMS collaboration announced the discovery of a SM Higgs boson candidate with a mass close to 125 GeV [7, 8]. For a given mass of the Higgs boson, the SM predicts all its properties. Investigating these properties for the 125 GeV Higgs boson candidate allows to check if it is indeed the SM Higgs boson or

one of an extended model. Since the SM does not allow LFV decays of the Higgs boson, the observation of such decays would indicate that the observed Higgs boson is one of an extended model and not of the SM.
In this thesis, the LFV Higgs boson is almost identically simulated as the SM Higgs boson. The production mechanisms, decay modes and rates as well as the kinematics of the LFV Higgs boson are the same as predicted by the SM, except that in addition LFV decays are allowed. This is reasonable, since so far all properties of the observed 125 GeV Higgs boson are consistent with the SM predictions.

2.3.1 Production and Decay of the Higgs Boson

The main SM processes that produce a Higgs boson are the gluon-gluon fusion (ggH), the vector boson fusion (VBF) and the vector boson associated WH and ZH production. The corresponding Feynman diagrams are presented in figure 2.1.

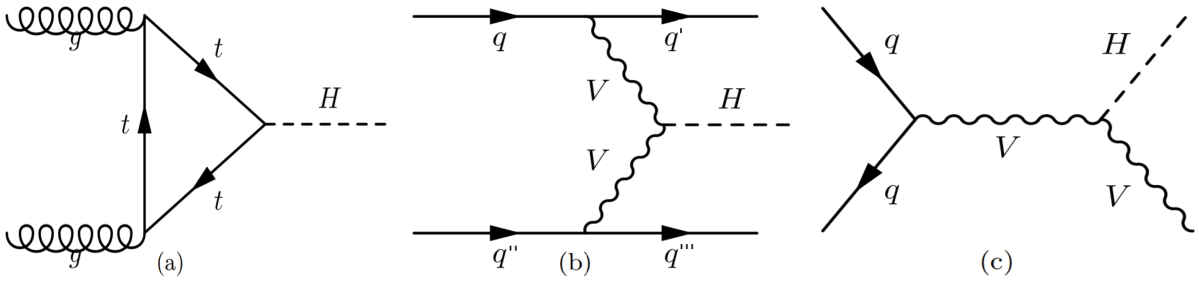


Figure 2.1: Leading order Feynman diagrams for gluon-gluon fusion (a), vector boson fusion (b) and vector boson associated (c) Higgs boson production [40].

Of the denoted production processes, ggH has the largest cross-section (88.9%). The VBF production process has the second largest (6.9%), followed by the associated WH production (2.5%) and finally the corresponding ZH process (1.6%). In table 2.3 the respective total cross-sections are listed.

Table 2.3: Total cross-section for different Higgs boson production processes at $\sqrt{s} = 13\,\text{TeV}$ [41].

Process	Cross-section [pb]
ggH	48.58
VBF	3.782
WH	1.373
ZH	0.884

The cross-section of the ggH process is by more than one order of magnitude larger than the second largest cross-section and hence, dominates the Higgs boson production. Both outgoing quarks in the VBF process result in jets. In most cases, these jets are separated by a large pseudorapidity difference and the jet system has typically a large invariant mass m_{jj}. In the vector boson associated Higgs boson production, the vector boson decays either leptonically or hadronically. In figure 2.2, the cross-section as function of the Higgs boson mass in the vicinity of 125 GeV is shown for different production processes.
Since the Higgs boson couples to all massive particles, it is able to decay at leading-order into massive particle and anti-particle pairs. In addition, higher-order contributions allow also decays

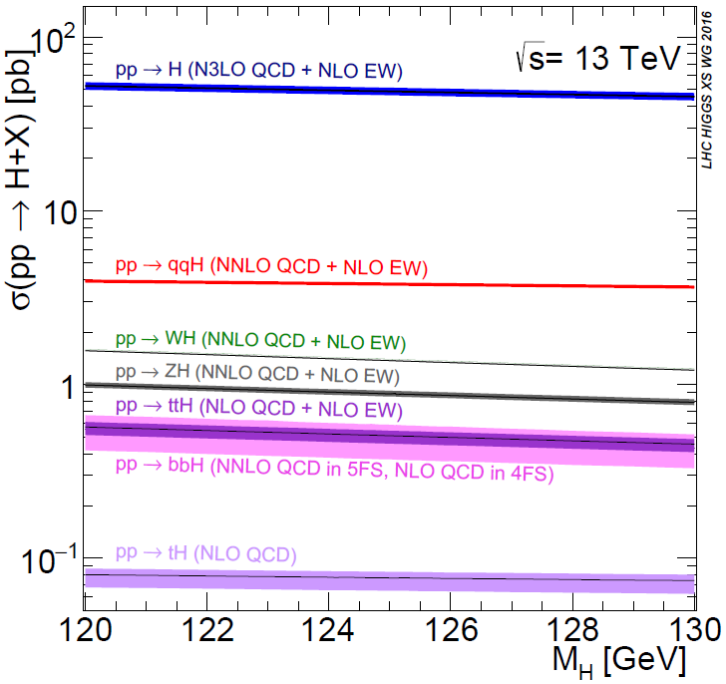


Figure 2.2: SM Higgs boson cross-section at $\sqrt{s} = 13\,\text{TeV}$ as function of the Higgs boson mass in the vicinity of 125 GeV [42].

into massless particle pairs or a massive and a massless particle. The respective branching ratios for the most important decay modes of the Higgs boson candidate with a mass of 125 GeV are presented in table 2.4.

Table 2.4: Branching ratios of the most important Higgs boson decay modes for a Higgs boson with a mass of 125 GeV [42].

$H \to$	$b\bar{b}$	W^+W^-	gg	$\tau^+\tau^-$	$c\bar{c}$	ZZ	$\gamma\gamma$	$Z\gamma$	$\mu^+\mu^-$
Branching ratio [%]	58.24	21.37	8.19	6.27	2.89	2.62	0.23	0.15	0.02

Most of these process are of no further interest in this analysis, since they are suppressed by the applied selection. However, $H \to \tau^+\tau^-$ has a significantly large branching ratio and a similar final state as the investigated LFV signal. Thus, this decay mode needs to be considered in the search for LFV decays of the Higgs boson. The decay process $H \to W^+W^- \to e\mu 2\nu$ has also a similar final state as the LFV signal. In figure 2.3, the SM Higgs boson branching ratio as function of the Higgs boson mass is presented.

LFV decays of the Higgs boson are possible in three decay modes: $H \to \tau e$, $H \to \tau\mu$ and $H \to \mu e$. However, the branching ratio of $H \to \mu e$ is already limited from indirect searches to $\lesssim 10^{-8}$ [32, 33]. The upper limit on the branching ratio of $H \to \tau e$ is 1.04% and the one of $H \to \tau\mu$ is 1.43% [34]. At the LHC, direct searches for these decays are performed. In the following, only LFV decays $H \to \tau e$ and $H \to \tau\mu$ are considered, where the τ-lepton decays leptonically resulting in a different flavour final state. In figure 2.4 the decay of a Higgs boson into a τ-lepton and either an electron or a muon (ℓ_0) is shown. The τ-lepton decays further into two neutrinos and a light lepton (ℓ_1). In this thesis, the τ-lepton is required to decay into the light lepton flavour, which is not the flavour of ℓ_0. Hence, an electron, a muon and two neutrinos are

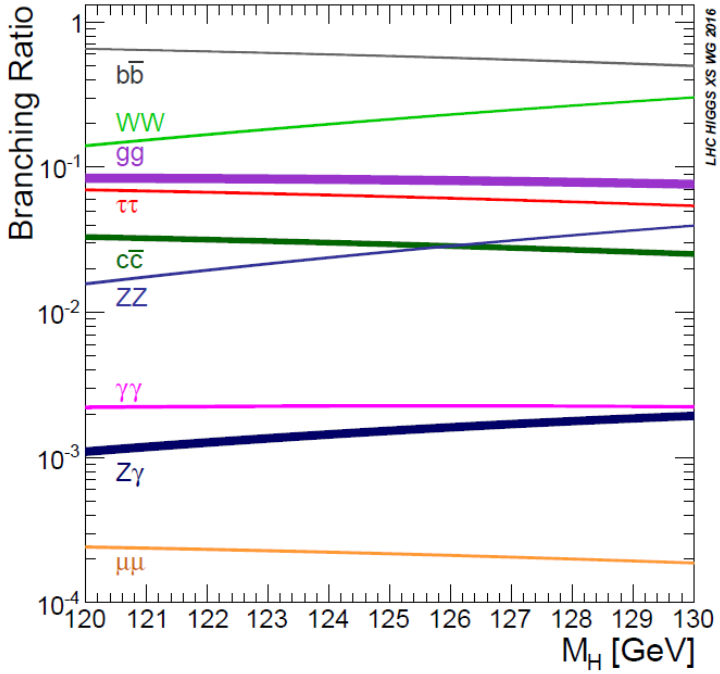


Figure 2.3: SM Higgs boson branching ratio at $\sqrt{s} = 13\,\text{TeV}$ as function of the Higgs boson mass in the vicinity of 125 GeV [42].

present in the final state. However, the neutrinos only interact weakly and thus are not detected directly by the detector, but cause missing transverse energy.

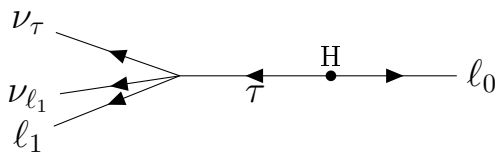


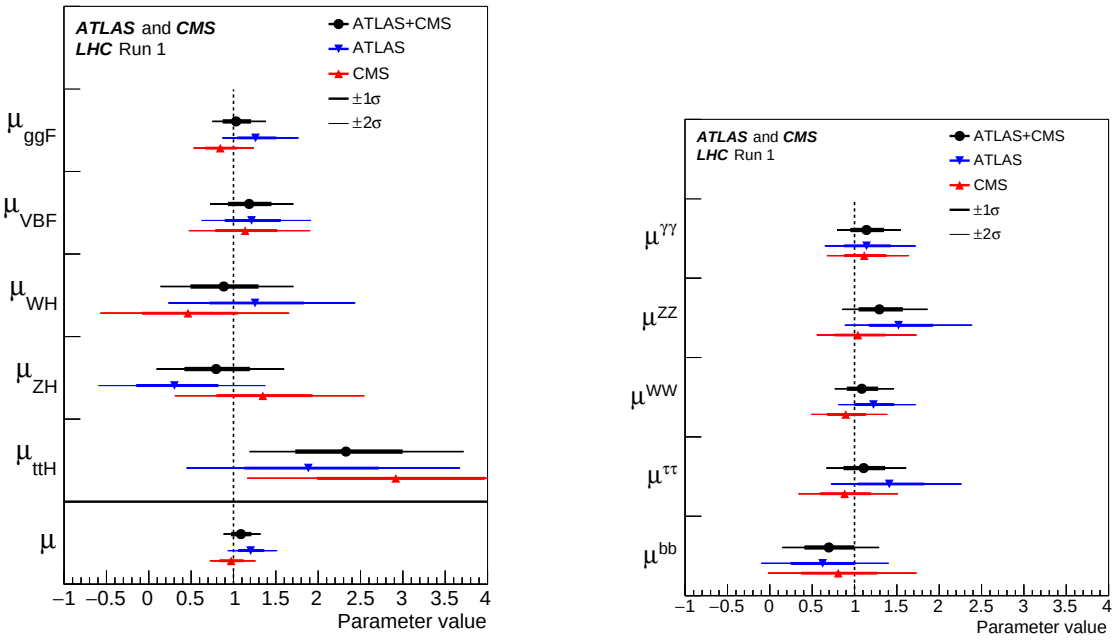
Figure 2.4: Feynman diagram of LFV Higgs boson decay into a τ lepton and a light charged lepton ℓ_0. The τ-lepton decays further into either an electron or a muon and two neutrinos.

Since ℓ_0 originates directly from the Higgs boson, it has in general a larger momentum than the lepton from the τ-decay. Furthermore, the τ-lepton and ℓ_0 have a high momentum due to the large mass difference to the Higgs boson. This results in approximately the same direction of flight of the decay products of the τ-lepton, i.e. ℓ_1 and the neutrinos.

2.3.2 The SM Higgs Boson Candidate

Since the discovery of the Higgs boson candidate, the properties of this new particle have been measured. So far, all measured properties do not show any significant deviation from the SM predictions.
As the SM predictions of the Higgs boson properties strongly depend on the Higgs boson mass, a precise measurement of its mass is crucial. Both collaborations presented a combined mea-

surement of the Higgs boson mass using the complete Run 1 data (2011-2012) of the LHC. A Higgs boson mass of $m_H = 125.09 \pm 0.21(\text{stat.}) \pm 0.11(\text{sys.})\,\text{GeV}$ was obtained [9]. For data of the 2016 data-taking year of Run 2 of the LHC, the CMS collaboration obtained a Higgs boson mass of $m_H = 125.26 \pm 0.20(\text{stat.}) \pm 0.08(\text{sys.})\,\text{GeV}$ [43]. The ATLAS collaboration derived for the data-taking years 2015 and 2016 a Higgs boson mass of $m_H = 124.86 \pm 0.27\,\text{GeV}$ [44].
Furthermore, a combined measurement of the Higgs boson production and decay rates was performed. The signal strength for the production process i and decay modes f is defined as $\mu_i^f = (\sigma_i \cdot \text{BR}^f)_{\text{obs.}}/(\sigma_i \cdot \text{BR}^f)_{\text{SM}}$, with the observed cross-section and branching ratio in the numerator and the ones predicted by the SM in the denominator. The derived best-fit signal strength values for the production and decay of the Higgs boson are shown in figure 2.5. Within the uncertainties, most of the values are in good agreement with the SM predictions. The global signal strength μ, which is the same for all production processes i and decay modes f ($\mu = \mu_i^f$) was derived as $\mu = 1.09^{+0.11}_{-0.10}$, being within 1σ of the SM prediction [10].



(a) Production

(b) Decay

Figure 2.5: Combined ATLAS and CMS measurements for Higgs boson production (a) and decay (b) signal strengths. A best fit value of $\mu = 1$ corresponds to the SM prediction [10].

Finally, the spin and parity of the SM Higgs boson candidate were measured. The SM predicts a spin 0 particle with positive parity. Alternative hypotheses were excluded above a 95% confidence level by the ATLAS and CMS collaboration [11, 45].
To summarize, all measurements up to now suggest that the observed Higgs boson at 125 GeV is indeed very SM-like. For the possible observation of a LFV decay of the Higgs boson, more precise measurements are necessary.

3 Experimental Setup at CERN

In this chapter, an overview of the experimental setup is presented. The LHC accelerator is described in section 3.1. In section 3.2, the general purpose detector ATLAS is presented in more detail. The structure of the detector and its most important detector parts as well as the used trigger system is described.

3.1 The Large Hadron Collider

The LHC (**L**arge **H**adron **C**ollider) [46] at CERN (**C**onseil **E**uropéen pour la **R**echerche **N**ucléaire) is presently the most powerful particle accelerator for proton-proton or heavy ion collisions. It consists of two 26.7 km long accelerator rings, which intersect each other at the experimental facilities, producing particle collisions at centre-of-mass energies $\sqrt{s}$ currently up to 13 TeV. The proton beams contain up to 2808 bunches separated by 25 ns and every bunch consists of about 10^{11} protons. Superconducting magnets with a dipole field up to 8.33 T force the beam on a circular path. Quadrupole magnets are used to focus the beams.
The number of collisions is proportional to the luminosity $\mathscr{L}$ of the accelerator, which is only depended on machine parameters. For two equally shaped Gaussian beam bunches, the instantaneous luminosity can be written as

$$\mathscr{L} = \frac{N_b^2 n_b f_{\text{rev}} \gamma_r}{4\pi \varepsilon_n \beta^*} F, \tag{3.1}$$

where N_b is the number of particles per bunch and n_b is the number of bunches per beam. The revolution frequency is noted as f_{rev}, the relativistic gamma factor as γ_r, the normalized transverse beam emittance as ε_n and the beta-function at the collision point as β^*. Finally, F denotes the geometric luminosity reduction factor due to the crossing angle of the beams at the interaction point [47].

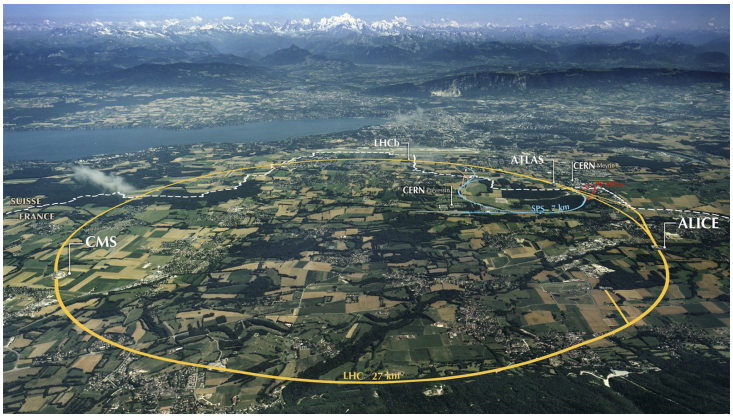


Figure 3.1: Position of the four detectors in the LHC ring [48].

At each of the four beam crossing points a detector is placed. These are the two general purpose detectors ATLAS and CMS as well as the ALICE and LHCb detectors. In figure 3.1 the position of these detectors in the LHC ring are shown.
In this thesis, data collected by the ATLAS detector from the so-called Run 2 of the LHC are used. This includes data-taking years from 2015 to 2018 at a center-of-mass energy of $\sqrt{s} = 13\,\text{TeV}$ and a corresponding integrated luminosity for proton-proton collisions of $\int \mathscr{L} = 139\,\text{fb}^{-1}$ [49]. In figure 3.2, the respective integrated luminosities for all data-taking years since 2011 are shown. In Run 2, much higher luminosities than in Run 1 are achieved.

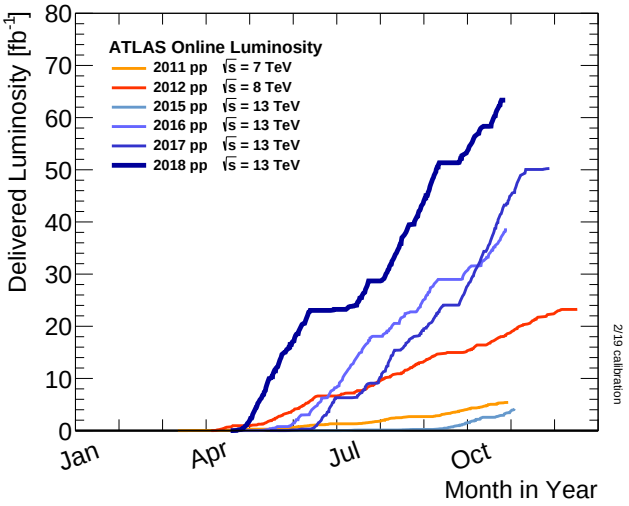


Figure 3.2: Integrated luminosity of all data-taking years since 2011 [49].

3.2 The ATLAS Detector

The ATLAS (**A** **T**oroidal **L**HC **A**pparatu**S**) detector is one of two general purpose detectors at the LHC. It is designed to perform precision measurements as well as to search for physics beyond the standard model. The detector is build up cylindrically and consists of several self-contained sub-detectors, shown in figure 3.3. The inner most detector part is the inner detector which is used to track charged particles. It is surrounded by the electromagnetic and hadronic calorimeters. The most outside parts of the ATLAS detector are the muon spectrometers. Two magnetic fields are present in the ATLAS detector, a solenoid field in the inner detector and a toroid field in the muon system.
The laboratory system of the ATLAS detector is called ATLAS coordinate system and has its origin in the nominal interaction point. It is defined as a right-handed coordinate system, where the z-axis is parallel to the beam line. Hence, the transverse plane is spanned by the xy-plane. The positive x-axis points from the origin of the coordinate system towards the centre of the LHC ring and the positive y-axis, is pointing upwards. With the right-handed coordinate system, the positive z-axis is called side-A and the negative one side-C.
The azimuthal angle ϕ is defined as angle around the z-axis in a right-handed way. Instead of the polar angle θ, defined as angle from the positive z-axis to the measured particle, the pseudorapidity

$$\eta = -\ln \tan \frac{\theta}{2} \tag{3.2}$$

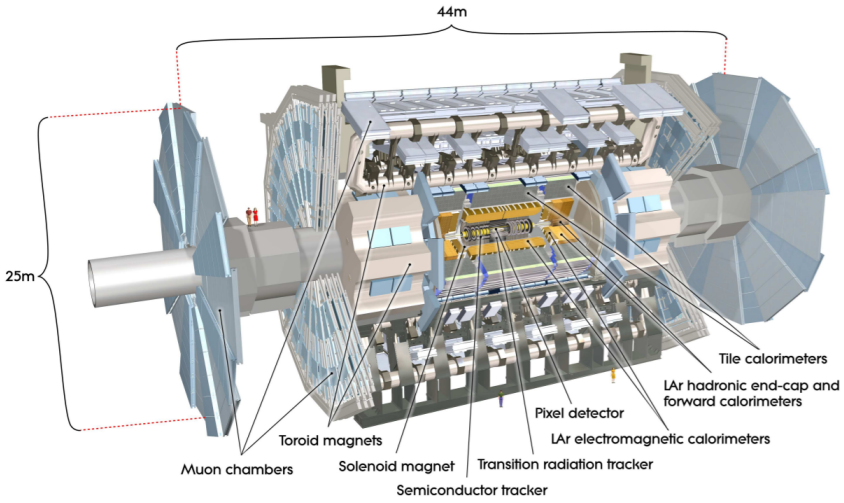


Figure 3.3: Overview of the ALTAS detector [50].

is used. With these angles, the distance in the azimuthal-pseudorapidity space is defined as

$$\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}, \tag{3.3}$$

where $\Delta\phi$ and $\Delta\eta$ are the respective angular differences.
The transverse momentum $\vec{p}_T$ and transverse Energy $\vec{E}_T$ are given as

$$\vec{p}_T = \begin{pmatrix} p_x \\ p_y \end{pmatrix}, \qquad \vec{E}_T = \begin{pmatrix} E_x \\ E_y \end{pmatrix} \tag{3.4}$$

and their magnitude, denoted as p_T and E_T respectively, is derived with the usual Euclidean norm.
In following, the individual sub-detectors are described in more detail. More information and details can be found in Ref. [50].

3.2.1 Inner Detector

The inner detector is located in the centre of the ATLAS detector. It is used to perform precise momentum measurements of charged particles, primary and secondary vertex measurements as well as to measure particle tracks. The central solenoid immerses the inner detector in a 2 T strong magnetic field. In figure 3.4, an overview of the inner detector is shown.
The inner detector is build up by the Pixel detector, the Semiconductor Tracker (SCT) and the Transition Radiation Tracker (TRT). Both, the Pixel detector and the SCT, cover the pseudorapidity of $|\eta| < 2.5$ and achieve the highest granularity around the vertex region. The TRT covers the region $|\eta| < 2.0$. All detector parts are split in a barrel region and two end-cap regions. In the barrel region, the Pixel detector and the SCT are arranged on concentric cylinders around the beam axis and the TRT straw tubes are ordered parallel to the beam axis, while in the end-cap regions all parts are ordered perpendicular to the beam axis.
The pixel layers are arranged in such a way, that a track usually crosses three of them. For the SCT, eight strip layers are typically crossed by one track, resulting in four measured space points due to the orthogonal orientation of the strips in each layer. The TRT provides about 36 hits per track, but in contrast to the other two detector parts, it does only provide $R - \phi$ position information, whereas the others also provide information on the z-position.

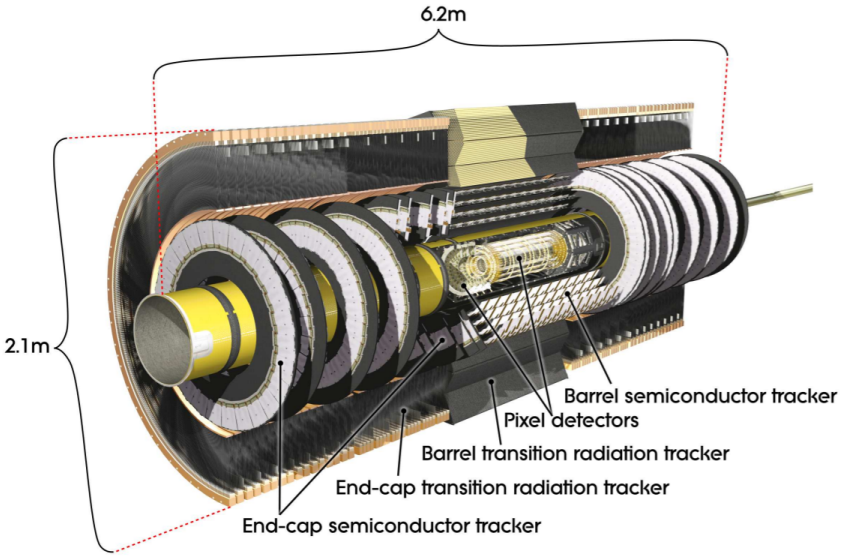


Figure 3.4: Overview of the inner detector [50].

The inner detector is designed to achieve the resolution of the transverse momentum

$$\sigma_{p_T}/p_T = 0.05\%\ p_T \oplus 1\%,$$

where σ_{p_T} is the uncertainty of the transverse momentum measurement.

3.2.2 Calorimeters

In figure 3.5, an overview of the electromagnetic and hadronic calorimeter is displayed. Both calorimeters are designed to provide a good containment of electromagnetic and hadronic showers. This is ensured by a total thickness of the calorimeter of > 22 radiation lengths in the barrel region and > 24 radiation lengths in the end-cap regions. The hadronic calorimeter has a radial depth of about 7.4 interaction lengths. Both, the electromagnetic and hadronic calorimeter are sampling calorimeters.
The electromagnetic calorimeter is designed to precisely measure the energy of electrons and photons. It has a very fine granularity using lead as absorber and liquid argon as active material. Similar to the inner detector parts, it is separated in a barrel region and two end-cap regions. The barrel region covers a pseudorapidity of $|\eta| < 1.475$ and is made up of two identical half-barrels. The end-cap consists of two coaxial wheels, where the outer wheel covers the region $1.375 < |\eta| < 2.5$ and the inner wheel $2.5 < |\eta| < 3.2$.
Over its full range, accordion-shaped electrodes and lead absorber plates build up the calorimeter, providing a complete symmetry in ϕ without any azimuthal cracks. In the central region $|\eta| < 1.8$, a presampler, consisting of a liquid argon layer, corrects for energy losses of electrons and photons prior to the calorimeter.
The electromagnetic calorimeter has a resolution of

$$\sigma_E/E = 10\%/\sqrt{E} \oplus 0.7\%$$

for the energy E of electrons and photons.
The hadronic calorimeter consists of a tile calorimeter in the central region, a hadronic liquid argon end-cap and a liquid argon forward calorimeter for large pseudorapidities. It is used to

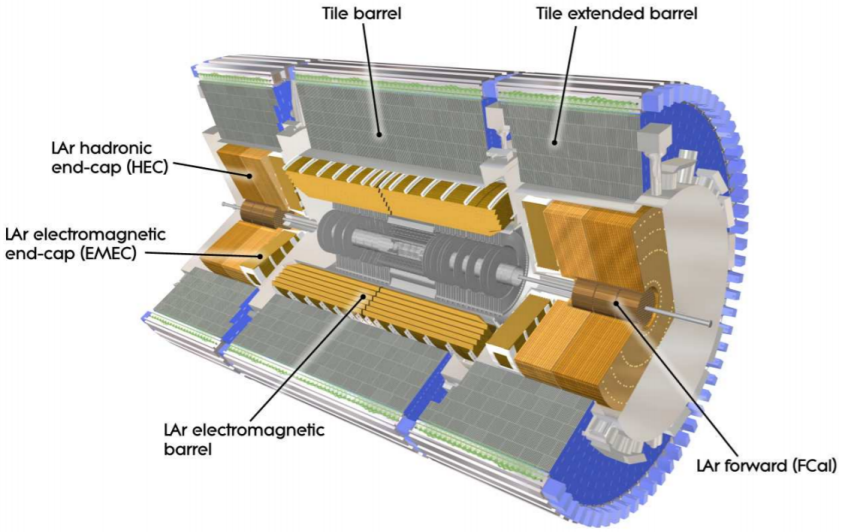


Figure 3.5: Overview of the electromagnetic and hadronic calorimeter [50].

reconstruct jets and the measurement of missing transverse energy. The central tile calorimeter consists of a barrel part, covering $|\eta| < 1.0$, and an extended barrel part for pseudorapidities between 0.8 and 1.7. Steel is used as absorber material and scintillating tiles as the active material. The hadronic end-cap calorimeter consists of two independent wheels for each end-cap. Each wheel is build up by wedge-shaped copper modules with liquid argon as active material. The end-cap calorimeter covers a region of $1.5 < |\eta| < 3.2$. In each end-cap, three modules are placed, building up the forward calorimeter, which covers a pseudorapidity region of $3.1 < |\eta| < 4.9$. The first one uses copper as absorber and is optimised for electromagnetic measurements. The other two measure primarily the energy of hadronic interactions and use tungsten as absorber. Liquid argon is used as active material for the forward calorimeter.
With

$$\sigma_E/E = 50\%/\sqrt{E} \oplus 3\%$$

for the tile and end-cap calorimeters, and

$$\sigma_E/E = 100\%/\sqrt{E} \oplus 10\%$$

for the forward calorimeter, the designed resolutions of jets are much larger than the one of the electromagnetic calorimeter. However, it is sufficient to satisfy the physics requirements.

3.2.3 Muon System

The muon system is designed to measure the momentum of charged particles which cross the calorimeters. It is located in the very outermost part of the ATLAS detector, as shown in figure 3.6.
Three different parts build up the muon system. Those parts are the barrel region for $|\eta| < 1.4$, the transition region with $1.4 < |\eta| < 1.6$ and the end-cap region for $1.6 < |\eta| < 2.7$. In the barrel region, the charged particles are deflected by the large barrel toroid, providing a bending power between 1.5 Tm and 5.5 Tm. Two magnets inserted in both ends of the barrel toroid provide the magnetic deflection in the end-cap region, with a bending power of 1 Tm to 7.5 Tm. In the

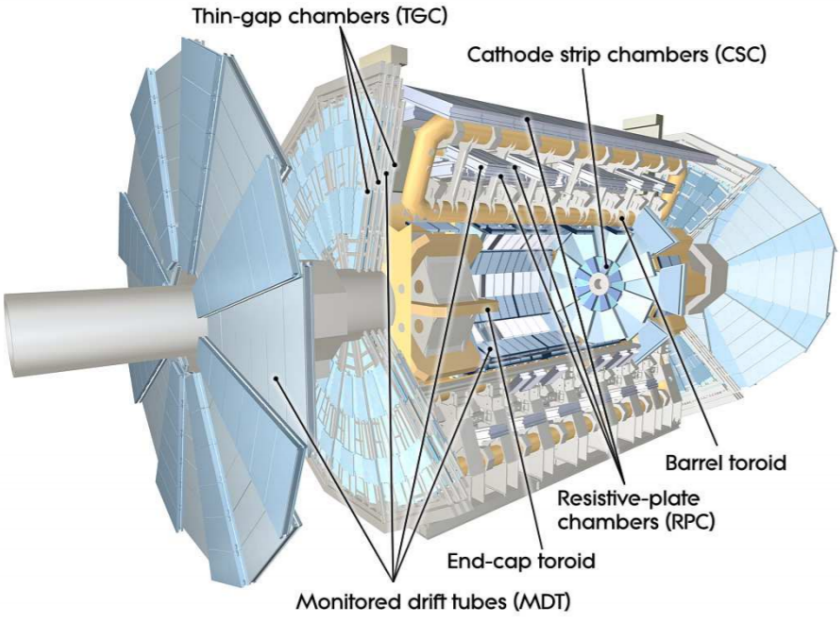


Figure 3.6: Overview of the muon system [50].

transition region, a combination of both magnets deflects the particles. This results in a magnetic field mostly orthogonal to the particle trajectories.
For the largest part of the muon system, monitored drift tubes (MDT) are used to measure the particle trajectories. At pseudorapidities of $2.0 < |\eta| < 2.7$, cathode strip chambers (CSC) are used in addition, which have a higher granularity.
In addition, the muon system is used to trigger particles with pseudorapidities up to 2.4. For this, resistive plate chambers are used in the barrel region and thin gap chambers in the end-cap region. The trigger chambers serve three purposes: providing well defined p_T thresholds, bunch-crossing identification and measuring the particle coordinate in the direction orthogonal to the one determined by the MDT's and CSC's.
At a transverse momentum of 1 TeV a resolution of

$$\sigma_{p_T}/p_T = 10\%$$

is achieved for muons.

3.2.4 Trigger System

The proton bunches in the LHC collide at a crossing rate of about 40 MHz. However, the ATLAS detector is only able to store the collision events at a rate of about 1 kHz. This requires a efficient trigger system which decides if a collision event is stored or not. The ATLAS trigger system [51] is based on two distinct systems. One system is the hardware-based level-1 trigger and the other the software-based high level trigger (HLT) [52].
The level-1 trigger uses information of the electromagnetic and hadronic calorimeter with a reduced granularity as well as the muon system. Events with either high transverse momentum muons, electrons, photons, taus or jets as well as high missing transverse energy are searched for. It also defines Regions-of-Interest (RoIs), which are further investigated by the HLT. The level-1 trigger reduces the initial bunch crossing rate of 40 MHz to 100 kHz, with a decision time of less than 2.5 μs.

Information from the inner detector hits, the full calorimeter information as well as data from the muon detectors is used by the HLT. The HLT is applied to the RoIs of the event. With the full event information available for the ROIs, the HLT can apply additional requirements, e.g. a given isolation requirement or have a transverse momentum above a given threshold. The HLT reduces the level-1 trigger rate to 1 kHz, with a processing time of 200 ms.
In figure 3.7, an overview of the ATLAS trigger system and data acquisition system is presented.

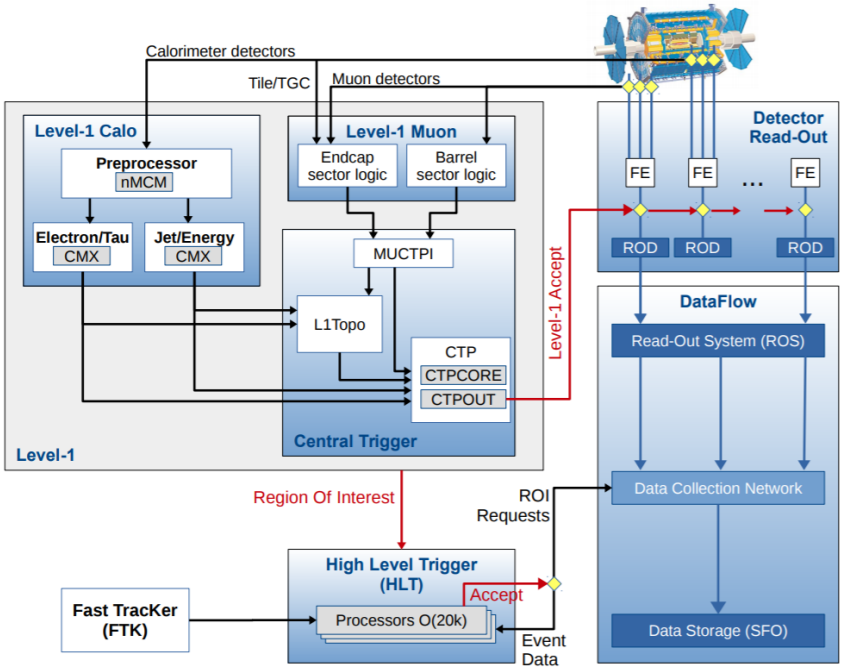


Figure 3.7: Overview of the ATLAS trigger system and data acquisition system [52].

4 Object Reconstruction and Triggers

The reconstruction of objects detected with the ATLAS detector is described in section 4.1. This includes electrons, muons, jets and the missing transverse energy. In section 4.2, the used single lepton and di-lepton triggers are described.

4.1 Object Reconstruction

The ATLAS detector is not able to detect a particle or jet itself, but only its energy deposition and the hits in the tracking detectors. Several algorithms use this information, to reconstruct the trajectories and properties of the respective objects. However, this reconstruction is not unique, but often several objects can be candidates for the reconstruction of the observed detector signals. Different requirements for the identification and isolation of these reconstructed objects is used, to select with a high probability the correct particle candidate.
A detailed description of the reconstruction of hadronically decaying τ-leptons can be found in Ref. [53].

4.1.1 Electrons

In this analysis, only electrons with a pseudorapidity $|\eta| < 2.47$ are considered. Because of a significant worse resolution and identification, electrons are excluded if they are detected in a transition area between the barrel and the end cap calorimeter, corresponding to a pseudorapidity of $1.37 \leq |\eta| \leq 1.52$, since .
The electrons are reconstructed using the information from the inner detector and the electromagnetic calorimeter. First, the clusters of energy deposits are identified in the calorimeter. Then, tracks of charged particles in the inner detector are matched to those clusters. If several tracks are matched, the primary electron track is chosen by using the distance in the azimuth angle ϕ and pseudorapidity η between the track and the centre of the cluster. Furthermore, tracks are required to hit the pixel detector and the silicon-strip detector. Finally, the energy of the electron candidate is calculated using the energy deposition in the calorimeter cluster. Energy loss effects like bremsstrahlung are taken into account.
For the identification of the electrons, a likelihood-based method is used. Three different working points with increasing likelihood thresholds are defined: *loose*, *medium* and *tight*. More background is rejected by using a higher threshold. However, with an increasing threshold, the identification efficiency decreases. The exact identification definitions and the used procedures can be found in Ref. [54].
Electron isolation criteria are applied, to reduce the background due to misidentified hadrons as well as electrons originating from heavy-flavour decays. These criteria use both the calorimeter-based isolation and track-based isolation. The isolation qualifies the detector activity in a cone around the electron candidate. Electron candidates from a heavy particle like the Higgs boson result only in a little additional detector response in the cone, whereas semileptonic decays of a heavy quark results in additional jets and thus in a larger detector activity [54].

The contribution of electrons resulting from secondary decay processes, so-called non-prompt electrons, can be reduced by using the longitudinal impact parameter $z_0 \sin\theta$ and the significance of the transverse impact parameter $d_{0,\text{signif.}}$. The definitions of d_0, z_0 and $\sin\theta$ are shown in figure 4.1.

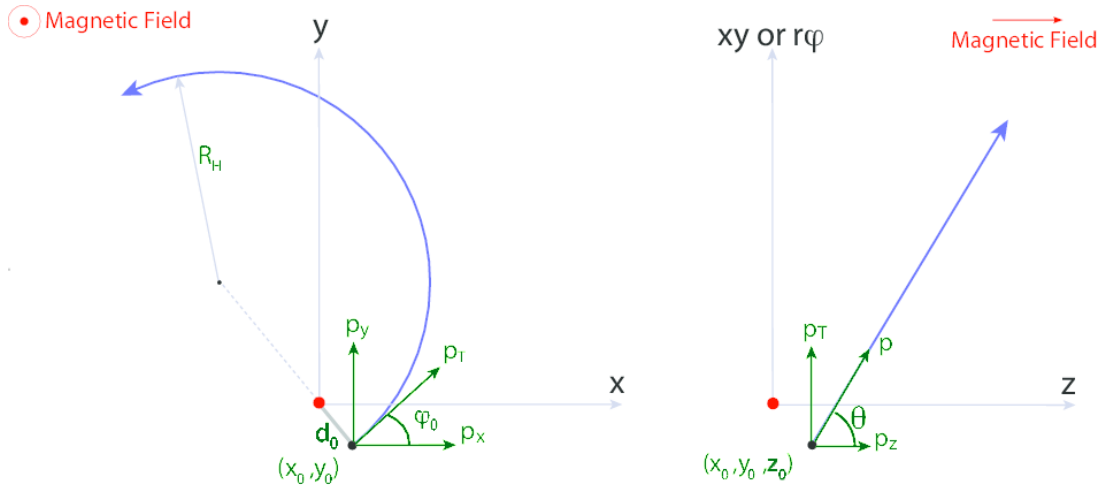


Figure 4.1: Definitions of d_0, z_0 and $\sin\theta$. The primary vertex is indicated as red dot [55].

The transverse impact parameter d_0 is defined as the shortest distance of a track to the primary vertex in the transverse plane, i.e. the x-y-plane. Its significance is derived as

$$d_{0,\text{signif.}} = d_0/\sigma_{d_0} \tag{4.1}$$

with the variance σ_{d_0} of the transverse impact parameter. Analogous, z_0 is defined as distance in z-direction between the primary vertex and the point on the track, used to evaluate d_0. This distance is multiplied with $\sin\theta$, where $\sin\theta = |\vec{p}_T|/|\vec{p}|$ with the (transverse) momentum $\vec{p}$ ($\vec{p}_T$) of the respective lepton. Events with large impact parameter values are more likely to include non-prompt electrons.
In this thesis, only electrons of the *medium* identification working point with a transverse energy greater 15 GeV are used, which corresponds to an identification efficiency $\varepsilon_{\text{id}} > 75\,\%$. In general, the electron isolation efficiency is dependent on different kinematic variables. These are mainly the transverse energy and the pseudorapidity. The used *Gradient* isolation is constructed in such a way, that the efficiency is only dependent on the transverse energy, eliminating the pseudorapidity dependency, given by $\varepsilon_{\text{iso}} = 0.1143 \cdot p_T + 92.14\,\%$. Electrons must have a longitudinal impact parameter $|z_0 \sin\theta| < 0.5$ and a significance of the transverse impact parameter $|d_{0,\text{signif.}}| < 10$.

4.1.2 Muons

Only muons with a pseudorapidity $|\eta| < 2.5$ are considered. The reconstruction and identification of muons is based on information of the inner detector and the muon spectrometer as well as the calorimeters.
Muons are at first reconstructed independently in the inner detector and the muon spectrometer. Inside each muon chamber of the spectrometer, hit patterns are searched to form segments. These segments are obtained by performing a straight-line fit to the hits which are aligned in the bending plane of the detector. A muon track candidate is constructed by fitting the hits from segments in different layers. Since this can result in several track candidates sharing the same segments, an

overlap removal between the different tracks is performed. This selects the best matching single track or allows that the segment is shared by two tracks.
The muon candidates in the muon spectrometer are combined with those in the inner detector. This is done in most cases by an outside-in pattern recognition, where the candidates in the spectrometer are extrapolated inwards to be matched with an inner detector track. This global fit results in a combined muon candidate. The inside-out recognition is also used. However, due to the muon identification discussed in the following, their contribution is below 1%.
Also for the muon identification, the three working points *loose*, *medium* and *tight* are used. The *loose* identification is designed to maximize the muon reconstruction efficiency, whereas the *tight* identification maximizes the purity of the reconstructed muons, by reducing the contamination due to pion and kaon decays. With the *medium* identification, the systematic uncertainties are minimized. The exact definitions are found in Ref. [56].
The impact parameters $d_{0,\text{signif.}}$ and $z_0 \sin\theta$ can be used to suppress the contribution of the cosmic muon background.
In this thesis, only *medium* identified muons with a transverse momentum greater 10 GeV are used, which corresponds to an identification efficiency $\varepsilon_{\text{id}} > 98\,\%$. For the muon isolation, the track-based criterion *FCTightTrackOnly_FixedRad* is applied. This isolation compares the transverse momentum of the muon candidate to the sum of transverse momenta of all tracks in a cone around it and rejects background processes with additional particles or jets. The cone size is fixed for a transverse momentum of the muon candidate larger than 50 GeV, which ensures a high isolation efficiency. Below this threshold, the cone size is inverse proportional to the transverse momentum, improving the isolation performance for muons produced in the decay of particles with a large transverse momentum. The isolation efficiency ε_{iso} is above 93 % [56]. The same requirements on $z_0 \sin\theta$ and $d_{0,\text{signif.}}$ as for the electrons are applied for the muons.

4.1.3 Jets

The reconstruction of jets is performed by the so-called *anti-k_t* algorithm using a radius parameter of $R = 0.4$ [57]. This algorithm uses topologically-grouped clusters of the calorimeter cells, where the individual calorimeter cells must exceed four times the expected average electronic and pile-up noise in the cell. In addition, the so-called particle flow algorithm is applied. This is a cell-based energy subtraction algorithm, which removes overlaps between the energy and momentum measurements in the inner detector and the calorimeters, respectively [58]. Jets are required to have a transverse momentum greater 20 GeV and to be within the η-range of $|\eta| < 4.5$.
The jets containing b-quarks are tagged using the *MV2c10* algorithm at the *FixedCutBEff_85* working point. This algorithm utilises a boosted decision tree with several training variables like the p_T and η of the jets. A more detailed description of the algorithm with all used training variables can be found in Ref. [59].
In the central pseudorapidity range ($|\eta| < 2.4$), another requirement is added, to reduce jets originating from additional beam interactions. For this, the jet-vertex-tagger (JVT) is introduced, which is a two dimensional likelihood discriminant. Both dimensions are based on variables, which are calculated in terms of the transverse momenta of tracks associated with the respective jet as well as tracks originating from additional beam interactions [60]. Jets in this pseudorapidity region which have a transverse momentum smaller than 60 GeV are required to have a JVT > 0.2.

4.1.4 Missing Transverse Energy

The ATLAS detector is not interacting directly with neutrinos. Thus, no individual reconstruction of those is possible, but some information can be obtained, using the information gained

from the measurement of all detectable particles. The accelerated protons in the LHC do not have any momentum perpendicular to the beam axis. Due to momentum conservation, this is also true for the final state of the produced particles. In case of an ideal detector, the negative of all measured transverse momenta corresponds to the combined transverse neutrino momentum, noted as missing transverse energy $\vec{E}_T^{\text{miss}}$. However, detector resolution effects are also contributing as well as further undetected particles.
The absolute value of the missing transverse energy is given as

$$|\vec{E}_T^{\text{miss}}| = E_T^{\text{miss}} = \sqrt{E_{T,x}^{\text{miss}\,2} + E_{T,y}^{\text{miss}\,2}} \tag{4.2}$$

with the components $E_{T,x}^{\text{miss}}$ along the x-axis and $E_{T,y}^{\text{miss}}$ along the y-axis. Each component of the missing transverse energy is reconstructed using *hard-event* and *soft-event* signals. The hard-event signals consist of all reconstructed jets and particles, i.e. electrons, muons, photons and hadronically decaying τ-leptons. Each of these objects must fulfil the respective reconstruction quality requirements as well as the kinematic thresholds. The soft-event signals are defined as all charged-particle detector signals of one event, which are not included in the hard-event signals. This includes, inter alia, independent pile-up interactions as well as particles and jets which do not fulfil the quality requirements.
The respective components are derived as

$$E_{T,x(y)}^{\text{miss}} = -p_{T,x(y)}^{e} - p_{T,x(y)}^{\mu} - p_{T,x(y)}^{\gamma} - p_{T,x(y)}^{\tau_{\text{had}}} - p_{T,x(y)}^{\text{jet}} - p_{T,x(y)}^{\text{soft}} \tag{4.3}$$

using the transverse momenta of the hard-event and soft-event contributions. Since all detector subsystems are used, the reconstruction of the missing transverse energy is very challenging.
A detailed description of the missing transverse momentum and its performance can be found in Ref. [61].

4.1.5 Overlap Removal

Two reconstructed objects are not allowed to be arbitrary close to each other, since this would cause ambiguity about the object identities. This is accomplished by an overlap removal. A scheme is used, eliminating one of two reconstructed objects by using in most cases the quantity $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$, with the angular differences $\Delta\phi$ and $\Delta\eta$ between the two objects.
In case of an electron and a jet with $\Delta R < 0.2$, the jet is removed. The remaining jets are categorized into pile-up and non-pile-up jets, where pile-up jets have a transverse momentum smaller 60 GeV, are within $|\eta| < 2.4$ and have a JVT < 0.59. The electron is removed, if the difference ΔR is smaller than 0.4 to a remaining non-pile-up jet.
For a muon and a jet, two criteria must be fulfilled by the jet to be removed. The first criterion is $\Delta R < 0.2$. Simultaneously, the jet must have less than three tracks or the transverse momentum requirements $p_T(\mu)/p_T(\text{jet}) > 0.5$ and $p_T(\mu)/\sum_k p_T^{\text{track}_k}(\text{jet}) > 0.7$ must be fulfilled. The $\sum_k p_T^{\text{track}_k}(\text{jet})$ is the scalar transverse momentum sum of tracks associated with the jet. The muon is removed, if $\Delta R < 0.4$ to a remaining non-pile-up jet.
Any calorimeter-tagged muon that shares an inner detector track with an electron is removed. Electrons are removed, if an inner detector track is shared with any remaining muon.

4.2 Triggers

Events are only used, if at least one lepton fires a single lepton trigger or a di-lepton trigger is fired. The trigger strategy is based on logical *OR* combinations of different single lepton and

di-lepton triggers. For every trigger, a p_T threshold requirement for the respective lepton is set. However, to avoid turn-on efficiency effects of the triggers in the low p_T-range, this threshold is increased by an additional offline p_T threshold. The used single lepton triggers as well as the corresponding offline p_T requirements are shown in table 4.1 for electrons and in table 4.2 for muons and the respective di-lepton triggers are presented in table 4.3.

Table 4.1: Electron trigger requirements. The electron transverse momentum is required to surpass both trigger and offline p_T requirements.

Year	Single lepton trigger	Trigger p_T/GeV	Offline p_T/GeV
2015	*HLT_e24_lhmedium_L1EM20VH*	24	25
	HLT_e60_lhmedium	60	
	HLT_e120_lhloose	120	
2016, 2017, 2018	*HLT_e26_lhtight_nod0_ivarloose*	26	27
	HLT_e60_lhmedium_nod0	60	
	HLT_e140_lhloose_nod0	140	

The data-taking years 2016, 2017 and 2018 have common single lepton and di-lepton triggers. For leptons passing the single lepton trigger, the offline p_T requirement is raised to 25 (27) GeV for electrons in 2015 (2016-2018) and 21 (27.3) GeV for muons. In case of the di-lepton triggers, the electron is required to have a higher transverse momentum than the muon. For the individual triggers for the 2015 data-taking year and the 2016-2018 data-taking years, the electron p_T is required to be larger than 18 GeV and the muon p_T larger than 14.7 GeV. In case of the trigger used in all data-taking years, these thresholds are modified to 27 GeV and 10 GeV.

Table 4.2: Muon trigger requirements. The muon transverse momentum is required to surpass both trigger and offline p_T requirements.

Year	Single lepton trigger	Trigger p_T/GeV	Offline p_T/GeV
2015	*HLT_mu20_iloose_L1MU15*	20	21
	HLT_mu50	50	
2016, 2017, 2018	*HLT_mu26_ivarmedium*	26	27.3
	HLT_mu50	50	

Table 4.3: Di-lepton trigger requirements. The electron and muon transverse momenta are required to surpass both trigger and offline p_T requirements.

Year	Di-lepton trigger	Trigger p_T/GeV	Offline p_T/GeV
2015	*HLT_e17_lhloose_mu14*	$p_T^e > 17$ $p_T^\mu > 14$	$p_T^e > 18$ $p_T^\mu > 14.7$
2016, 2017, 2018	*HLT_e17_lhloose_nod0_mu14*	$p_T^e > 17$ $p_T^\mu > 14$	$p_T^e > 18$ $p_T^\mu > 14.7$
All	*HLT_e26_lhmedium_nod0_* *L1EM22VHI_mu8noL1*	$p_T^e > 26$ $p_T^\mu > 8$	$p_T^e > 27$ $p_T^\mu > 10$

5 Neural Networks

Machine learning (ML) techniques are widely used in particle physics analyses. The different ML techniques are enabled to exploit the underlying scheme by the large amount of produced data. Thus, the experimental sensitivity can be improved by applying ML with respect to traditional cut-based concepts [62]. One of these techniques are Neural Networks (NN) or more modern Deep Neural Networks (DNN). Two techniques are typically used: supervised and unsupervised learning. In latter, the algorithm tries to find unknown patterns within data, whereas for supervised learning it learns to reproduce known pattern. Hence, the target value in the case of supervised learning is known, whereas in unsupervised learning it is not. Only supervised learning is considered in the following.
In section 5.1, an introduction to NNs is presented. This includes the most important quantities, the two DNN tasks regression and classification, the description of hyper-parameters as well as regularization of NNs and the important concept of error back propagation. In section 5.2, the training process of NNs is described. The challenges of optimizing NNs are discussed in section 5.3.

5.1 Introduction to neural networks

In high energy particle physics (HEP), the separation between background processes and the signal process is a main challenge to perform precision measurements or to find physics beyond the SM. A better separation or separating a resonant signal process can be achieved by an improved mass reconstruction. NNs and DNNs provide a new technique, which benefits from the large amount of produced data in HEP.

5.1.1 Important quantities of neural networks

The aim of any NN is to map some input features on one or more output values. Input features are not chosen by the NN itself, but are defined in advance. In case of HEP, this can be the energy of a particle or its direction of flight. At best, these are important quantities, which have a significant influence on the desired output. They are then propagated through the NN to derive the output(s), denoted as L.
The fundamental constituents of any NN are the so-called *nodes*. For each node, all input values are combined into one output value. In the following, each input value is given as x_i, with the total number of inputs N. Further, w_i are the corresponding weights. With this, the output y per node can be derived as

$$y = f\left(\vec{w} \cdot \vec{x} + b\right) \tag{5.1}$$

with the transposed vectors $\vec{x} = (x_1, \ldots, x_N)^\intercal$ and $\vec{w} = (w_1, \ldots, w_N)^\intercal$. The bias b is an individual constant for each node.
The function $f\left(\vec{z}\right)$, with $z = \vec{w} \cdot \vec{x} + b$, is called *activation function* or simply *activation*. The *rectified linear function* (ReLu) is one of the most used activation functions [63, 64]. It is defined

as

$$f(z) = \begin{cases} z & z \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

The function has two linear parts which enables easy computations of the derivatives. For supervised learning with large amounts of data, a superior performance to other activation functions was shown Ref. [64].
An important property of the output y in equation (5.1) is its smoothness. For any small change of the initial weights $\Delta\vec{w}$ and the bias Δb,

$$\Delta y \approx \sum_{i}^{N} \frac{\partial y}{\partial w_i} \Delta w_i + \frac{\partial y}{\partial b} \Delta b \tag{5.3}$$

is a good approximation of the resulting output change Δy. The change of the output Δy is linear in the weight change $\Delta\vec{w}$ and bias change Δb. Hence, a small $\Delta\vec{w}$ or Δb results in a similarly small Δy.
With this knowledge, one layer of a NN is defined as a set of M nodes. In general, each layer of a NN can have a different number of nodes and the bias and weights of each node are different with respect to the others. The output for every node is derived with equation (5.1).
The total NN can be split into three regions: *input layer*, *hidden layers* and *output layer* as displayed in figure 5.1. The input layer lists the N input features that were chosen. The last layer is the output layer, consisting of L nodes. The hidden layers include all layers, which are neither an input nor an output layer.
In general, DNNs have many hidden layers, whereas NNs only have a few or even only one. There is no sharp definition separating DNNs and NNs by their respective number of hidden layers. Thus, both terms are used to describe a NN with the general structure described previously.

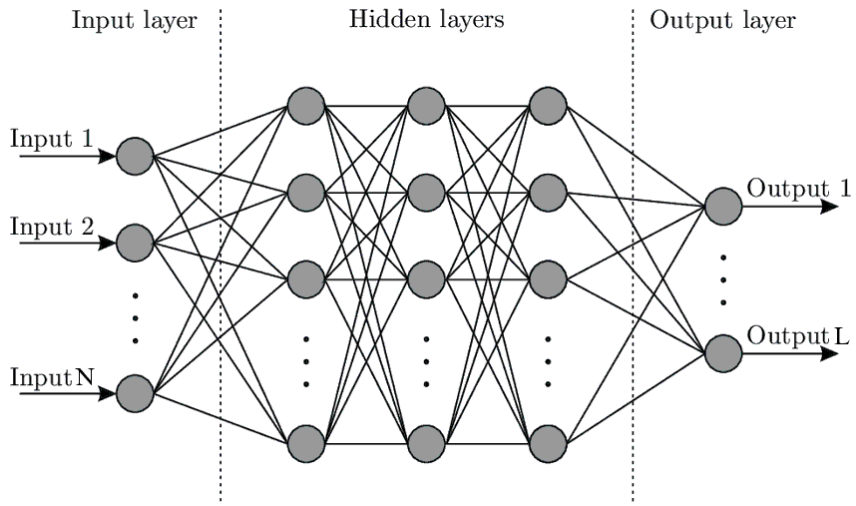


Figure 5.1: Architecture of a neural network. The network is typically split into an input layer, an output layer and hidden layers in between, indicated by the vertical dashed lines [65].

In feedforward NNs, the output of one layer is the input of the next layer. The information is only process into the direction of the output layer. In contrast to this, recurrent NNs allow loops within the NN. This means, that after a defined delay the output of one layer can be used as

input for a previous layer or the same layer [66]. In the following, only feedforward networks are considered.
To quantify how good the NN predicts the desired output, a so-called *loss function* or *cost function* is defined. Let the NN have L nodes in the output layer, than the respective outputs can be combined to $\vec{y}(\vec{x}, \mathbf{w}, \vec{b}) = (y_0, \ldots, y_L)^\intercal$ as NN output, where the weight matrix is given as $\mathbf{w} = (\vec{w}_0 \cdots \vec{w}_L)$. With K nodes in the last hidden layer, the dimensionality of the weight matrix is $K \cdot L$. The vector $\vec{d}(\vec{x})$ is the true known output corresponding to the input event $\vec{x}$. Then, the mean squared error loss function is defined as

$$C(\mathbf{w}, \vec{b}) = \frac{1}{\mathcal{N}} \sum_{i}^{\mathcal{N}} |\vec{y} - \vec{d}|^2 \tag{5.4}$$

where the sum runs over all present input events $\mathcal{N}$. The closer this function gets to zero, the better the NN predicts the output for all input events. Hence, a set of weights and biases is sought, which minimizes the loss function. For this purpose, a so called *optimizer* is introduced. The *gradient descent* optimizer is a widely used ML technique to minimize the loss function. This method calculates the gradient of the loss function with respect to all weights and biases. The weights of the output node n are given as $\vec{w}_n$ and the respective bias as b_n. To simplify, these are combined into the variable $\vec{u}_n = (\vec{w}_n, b_n)^\intercal$. By calculating

$$u_{n,i} \to u'_{n,i} = u_{n,i} - \alpha \frac{\partial C}{\partial u_{n,i}} \tag{5.5}$$

with a fixed constant $\alpha > 0$, a new set of weights and biases $\vec{u}'_n = (\vec{w}'_n, b'_n)^\intercal$ is obtained. The change of the loss function is then approximated as

$$\Delta C = C\left(\vec{u}'\right) - C\left(\vec{u}\right) \approx \vec{\nabla}_{\vec{u}} C \cdot \Delta \vec{u} = -\alpha \cdot (\vec{\nabla}_{\vec{u}} C)^2 \tag{5.6}$$

where $\vec{\nabla}_{\vec{u}}$ is the gradient with respect to $\vec{u} = (\vec{u}_0, \ldots, \vec{u}_L)^\intercal$. As α is per definition greater than zero, this method results by construction in a decreasing loss function in this approximation. For each weights and bias update, the average over all input events $\mathcal{N}$ is derived. Thus, this method becomes very time consuming for large data sets.
A more time efficient method with a better performance is the *stochastic gradient descent* (SGD). For each update, only a small randomly chosen subset of the input data is used. Hence, the average over this subset is much faster calculated than over the total data set. With this method, the NN can achieve a faster learning process, but the exact gradient of the loss function is lost. However, an approximation of the gradient is sufficient to obtain values close enough to the actual minimum of the loss function. However, SGD does need much time to escape local minima or saddle points [67].
A more modern optimizer is the *adaptive moment* (Adam) optimizer, which is based on the SGD. This method updates exponentially the mean $\vec{m}$ (1^{st} moment) and the uncentered variance $\vec{v}$ (2^{nd} moment) of the gradient of the loss function. Both moments have the same dimensionality as the variable $\vec{u}$ in equation 5.6. The exponential decay rate of the 1^{st} moment is defined by the parameter $\beta_1 = 0.9$ and the one of the 2^{nd} moment by the parameter $\beta_2 = 0.999$. Both moments are initialized as vectors of zeros, resulting in a bias in the beginning of the training. However, the Adam optimizer uses both decay parameters to correct this bias. Thus, for node n and the t^{th} parameter update, the 1^{st} and 2^{nd} moments are given by

$$m_{n,i}(t) = \left[\beta_1 \cdot m_{n,i}(t-1) + (1-\beta_1) \cdot g_{n,i}(t)\right] / \left(1 - \beta_1^t\right) \tag{5.7}$$

$$v_{n,i}(t) = \left[\beta_2 \cdot v_{n,i}(t-1) + (1-\beta_2) \cdot g^2_{n,i}(t)\right] / \left(1-\beta_2^t\right) \tag{5.8}$$

where $g_{n,i}(t) = \partial C / \partial u_{n,i}$ is the gradient of the loss function and β_i^t the respective decay parameter to the power of t. The denominator performs the described bias correction.
The t^{th} updated weights and biases are calculated as

$$u_{n,i}(t) = u_{n,i}(t-1) - \alpha \cdot m_{n,i}(t) / (\sqrt{v_{n,i}(t)} + \epsilon) \tag{5.9}$$

with a constant $\epsilon = 10^{-7}$ to avoid a zero division in case of a vanishing 2^{nd} moment. Thus, the actual updated stepsize of the weights and biases are limited by the fixed constant α. Furthermore, this method works well with only a small set of the total input data as well as with sparse gradients and is thus capable to escape much faster saddle points of the loss function than purely the SGD. With the performed bias-correction, more stability in the early training process is gained however, after sufficient many updates, this bias correction becomes close to unity [68].
To summarize, optimizers enable the NN to obtain a set of weights and biases which minimize the loss function. This corresponds to a NN, which predicts the output for all input data best. Thus, the loss function is used as the quantity separating good working NNs from bad ones.

5.1.2 Regression DNN in Contrast to Classification DNN

Regression and classification are two fundamental important tasks for supervised DNNs. In the case of classification, the network is meant to decide which category a respective input corresponds to. For instance, the input events are split into signal like events and background like events. To do so, each output node corresponds to exactly one category and the NN predicts a probability for every input event to correspond to the respective category. The target values are restricted to 0 and 1, where a 0 indicates that this event does not correspond to the respective category. Hence, the activation functions in the output layer are restricted to an interval of $[0, 1]$ and values close to 1 predict an agreement to the corresponding category. Widely used activation functions, fulfilling this requirement are the sigmoid and the softmax function.
In case of regression, one numerical or several values are predicted for each input event. For HEP, the predicted numerical value can be the invariant mass of a decayed particle. The target values are real numbers and so are the ones of the NN. Thus, a linear activation function is used for the nodes in the output layer.
The main difference between these two cases is the output. Whereas classification-based problems require probabilities and thus values between 0 and 1, regression needs real values. Thus, different activation functions are needed in the output layer [67].

5.1.3 Hyper-parameters of Neural Networks

Every neural network has specific parameters, which are not and can not be adapted by the network itself in the training process. They are predefined before any learning process is started. These parameters are called *hyper-parameters* and their setting is of high importance for good training results.
The *number of epochs* determines how many times the neural network trains on the total set of training examples. A too small number of epochs will prevent the neural network from learning all features correctly. After some number of epochs however, the NN will not improve further. Thus, a reasonably large value is the best choice for this hyper-parameter. A number of epochs greater than 100 is in many cases sufficient, but with different techniques the value can be raised much higher without worsening the NN performance [63].

Related to this is the so called *mini-batch* or *batch size*. Most NNs do not use all training events for each weight and bias update, but only a subset. The size of this subset is defined by the batch size. Large values estimate the gradient of the loss function more precisely, but also require more memory during the computation. Small batch sizes introduce stochastic noise, which potentially helps the NN to adapt better to the underlying scheme within the training events. In most cases, a small batch size is better than a large one [69]. In general, a value around 30 up to a few hundred is used [63].
The single most important hyper-parameter is the *learning rate*. It is the fixed constant α introduced in equation (5.5). Consequently, the learning rate defines how large the steps are, which the NN takes, when it updates its weights and biases. As the update of weights and biases is the main challenge of any NN, a well chosen learning rate is essential. Large learning rates will lead to a fast minimization of the loss function, but also to large jumps during the determination in the weights and biases. Thus, the actual minimum of the loss function is more likely to be missed. If the predefined value for α is too small, the training will take much longer, and the minimization can get stuck in a local minimum without a way to escape. Common values for the learning rate are between 10^{-6} and 1.
Strongly connected to this hyper-parameter is the *decay*. The decay parameter reduces the initially fixed learning rate. In each parameter update, the learning rate gets smaller, enabling the network to find the actual minimum of the loss function more easily. The updated learning rate after t iterations is given as $\alpha' = \alpha/(1 + \text{decay} \times t)$, where the decay is usually chosen to be smaller unity.
Finally, the number of nodes in each layer as well as the number of hidden layers must be defined in advance. In each layer, different number of nodes can be used. There is no general scheme what relation between the different number of nodes in each layer is to be preferred. However, it is important, that the first hidden layer has more nodes than there exist input features to ensure that no information of the input features is lost. The actual number of nodes depend on the number of input features however, more nodes are more preferential than less [63]. Similar to number of nodes, there is no general rule, describing how many hidden layers are performing best. It is expected, that a NN with more layers is able to solve more complex tasks, but only one hidden layer with enough nodes is sufficient for most tasks [67].

5.1.4 Regularization of a Neural Network

An important concern is the one of regularization. While training, NNs can adapt too strong to the used input data, which is called *overfitting*, resulting in a loss of generalization. The aim of regularization is to prevent the NN from overfitting. Thus, the NN is modified during its learning process in such a way that its generalization improves. This means, that a more general model is learned by the NN instead of the specific one that matches the training data best. Regularization results almost always in a worse training loss function than unregularized training, but this is intended to prevent the acquisition of intrinsic features of the particular training set resulting in better generalization properties of the NN [67].
One way of regularization is the so called *dropout* method. This method eliminates temporarily specific nodes with a given probability p. For those nodes, all connections to previous or following nodes are eliminated for one update of weights and biases through the optimizer. For each update of the weights and biases, a new set of nodes is chosen. After the completion of the training, all nodes are set as active, but their outgoing weights are scaled by $(1 - p)$. This approximates the average over all neural networks during training. A typical value of the dropout probability is 0.5 [70].

Another common regularization method is *early stopping*. For this, an additional hyper-parameter called *patience* is introduced. In general, the NN trains until it reaches the maximum number of epochs. With early stopping, the training can be stopped earlier. During training, the loss function is monitored. If its value does not shrink for a given number of epochs, the training is stopped. The patience fixes the number of epochs within the loss function must reduce [63].
Both methods are meant to prevent overfitting of the NN which deteriorates its performance during training, but improves its generalization.

5.1.5 Error Back Propagation

The method of back propagation is used to update the weights and biases in the NN. For this, the updated weights and biases in the output layer are calculated at first and these are than used, to update the previous layer. Thus, the weights and biases update is propagated backwards through the NN.
The gradient of the loss function in equation (5.6) is derived by the so-called *backpropagation algorithm*. For this, lets define w^l_{jk} as the weight, connecting the k^{th} neuron of layer $(l-1)$, with the j^{th} neuron of layer l. The bias of the j^{th} neuron in the l^{th} layer is denoted as b^l_j and the respective node output as y^l_j.
Similar to equation (5.1)

$$y^l_j = f\left(\sum_k w^l_{jk} y^{l-1}_k + b^l_j\right) \tag{5.10}$$

connects the output of the j^{th} neuron in layer l to the previous layer using the activation function f. The error of neuron j in layer l is defined as

$$\delta^l_j \equiv \frac{\partial C}{\partial z^l_j} \tag{5.11}$$

with $z^l_j = \sum_k w^l_{jk} y^{l-1}_k + b^l_j$. The error describes how good the weights and bias of the respective node are optimized. An error close to zero corresponds to a very good optimized node.
With these definitions, the error in the output layer M is given as

$$\delta^M_j = \frac{\partial C}{\partial y^M_j}\frac{\partial y^M_j}{\partial z^M_j} = \frac{\partial C}{\partial y^M_j}\frac{\partial f(z^M_j)}{\partial z^M_j} \tag{5.12}$$

Since both the loss function and the activation are known, their derivatives with respect to the given parameters can be calculated fast. The error of any node j in layer l is derived as

$$\delta^l_j = \sum_k w^{l+1}_{kj}\delta^{l+1}_k \frac{\partial f(z^l_j)}{\partial z^l_j} \tag{5.13}$$

by applying the chain rule to equation (5.11). The errors are now connected to the derivatives of the loss function with respect to the weights and biases. This is achieved by applying again the chain rule to equation (5.11). The derivatives are given as

$$\frac{\partial C}{\partial w^l_{jk}} = y^{l-1}_k \delta^l_j \tag{5.14}$$

$$\frac{\partial C}{\partial b} = \delta^l_j \tag{5.15}$$

and enable the calculation of the gradient in equation (5.6). Thus, by propagating the error δ_j^l backwards through the network, weights and biases can be updated and the NN is able to adjust to the given task [66].

5.2 Training of a Neural Network

A NN needs to be trained, validated and tested on data before it can be applied.

5.2.1 Importance of Train-, Validation- and Test-Sets

To obtain a well performing NN, the initial data set is split into three categories: a training, a validation and a test set. The training set is the largest one in order to provide the NN with as much statistics as possible during the training. It is used to let the NN learn a specific model. Thus, the loss function of the training set is optimized to obtain the best weights and biases.

The validation set is used to quantify the generalization of the trained NN. After each epoch, the NN is evaluated on the validation set by calculating the loss function. As this data set is completely independent of the training set, so is its loss function. Hence, a small validation loss function verifies that the NN is well generalized and is not overfitted. The early stopping regularization is only applied on the validation loss function and not on the training loss function. Furthermore, the validation loss function is used to define the optimal hyper-parameters, described in more detail in section 5.3. In addition, the validation set can be used to correct the trained NN in hindsight. If there is, for instance, a systematic offset, a constant can be added to the output to adjust the NN prediction.

At the end, the NN is tested on the test set. The performance of the trained NN is examined by comparing the true target values with the ones predicted by the NN. No more modifications or optimizations of the NN are carried out.

5.2.2 Preprocessing of the Data-Set

Preprocessing is done prior to the training of the NN and is applied to all of the data. The data set is processed in such a way, that the NN benefits from it. One useful procedure is the standardization of the input features. This means, that based on all events of the three sets, all input features are shifted by their respective mean towards zero and divided by their standard deviation. Thus, for all input features, the mean value is zero and the variance unity. This basically corresponds to a shifted and rescaled axis and helps the NN to learn faster [71].

Furthermore, the used coordinate system can be shifted or rotated in order to eliminate one or more input features. By rotating, for instance, the azimuth angle of a given detected particle always to $\phi = 0$, this input feature can be eliminated without loosing information about the event. This reduces the complexity of the dependencies the NN needs to learn and simplifies the learning process. The same purpose is fulfilled by a decorrelation of the input features, e.g. with a principle component analysis, as explained in more detail in section 5.3.3.

Finally, before the training starts, the data set events are shuffled randomly. This prevents that events which have very similar properties are combined in the same batch and helps the NN to adjust to a more general model [71].

5.3 Optimizing the Neural Network Architecture, Hyper-parameters and Input Features

The best configuration of a NN for a specific task is not obvious. For each architecture, the best hyper-parameters must be determined and the input features must be chosen. A decorrelation of the latter is possible and can be performed with the principle component analysis.

5.3.1 Determining the Best Hyper-parameter Configuration

The best trained NN for a particular task requires a specific value for each hyper-parameter. However, these values are not known in advance. Thus, some kind of search must be performed. As several hyper-parameters are tested, this can become very extensive. Many distinct configurations must be examined and compared to each other.
In general, the validation loss is used to determine how well a NN is trained. For each hyper-parameter configuration, the validation loss must be determined, requiring several trainings and thus a significant amount of computing memory and time.
So called hyper-parameter searches are a specific approach to find appropriate values. The considered searches are described in the following.

5.3.2 Grid-, Random- and Grid-Random-Hyper-parameter Search

One approach to find the optimal hyper-parameters is the so-called *grid-search*. For the hyper-parameters a set of predefined values is chosen and all possible combinations are used to train a NN. All obtained validation losses are compared after training. In figure 5.2, the resulting minimization of the validation loss is displayed. All hyper-parameters except the learning rate are fixed. The actual values are at this point of no interest as later on several hyper-parameters are optimized simultaneously. The optimal learning rate obtained with the grid-search for this generic configuration is located at the minimum of the validation loss function.

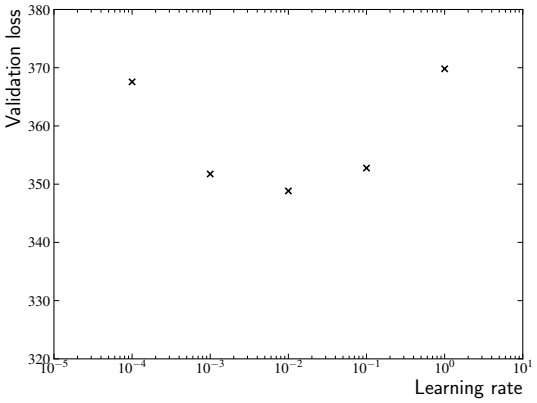


Figure 5.2: Concept of validation loss minimization using grid-search. For a fixed configuration, different learning rates are used to train the NN. As a result, the optimal value is located at the minimum of the validation loss, here 10^{-2}.

The general idea of the grid-search can be seen in figure 5.3. This method is very straight forward, but has also a few drawbacks. First, several combinations of hyper-parameters will not result in

a well trained NN, for instance, a high learning rate in combination with a low decay. Second, all hyper-parameters are on a predefined grid. Thus, all information in the vicinity of the grid points is unknown. This is especially problematic for important hyper-parameters as the optimal value is more likely missed which results in a significant worse performance of the NN. In addition, it is very time consuming for higher dimensional searches, since each hyper-parameter is tested several times with the exact same value [72].
An alternative to the grid-search is a *random-search*. Here, for each hyper-parameter, instead of specific values rather an allowed parameter range is specified. Then, for each iteration, random values are chosen within these ranges. This is also displayed in figure 5.3. As the values are chosen randomly in each training, the hyper-parameters are assigned different values each time. Thus, hyper-parameters are more likely to find their optimal value. However, there is no prevention against clustering of the hyper-parameter values in a region far from the optimal value. This may result in an incomplete coverage of the possible parameter space [72].

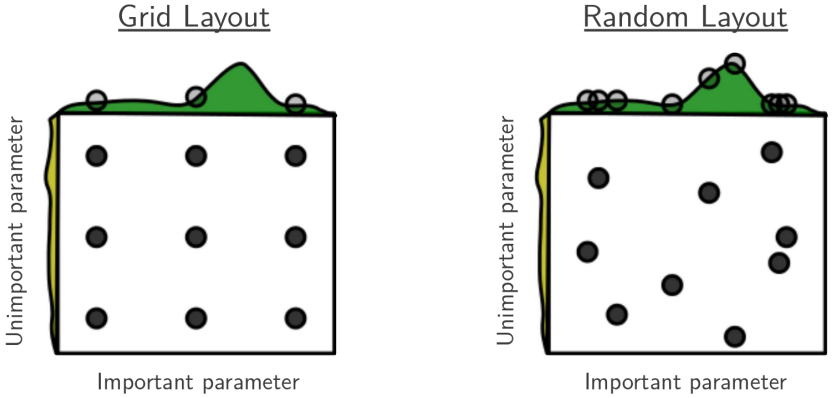


Figure 5.3: Schematics for grid and random hyper-parameter search. Each black dot represents a specific hyper-parameter configuration. The influence of an important hyper-parameter on the validation loss is shown on the top side fo the box, where a higher value corresponds to a better optimization. An unimportant parameter is displayed on the left side of the box. Random search is able to find a better configuration than grid search [72].

In this thesis, a combination of both, a grid-random-search, is tested which provides a much more reliable result. As in the grid-search optimization, specific values are predefined for all hyper-parameters. For each grid point however, all hyper-parameters are separately shifted by a random value. Thus, no value is used twice and a more precise optimization is performed. This method combines the advantages of both grid and random searches. Unreasonable configurations are still possible, but a reasonable coverage of the full parameter space is ensured.
The time consumption of the training can be shortened by means of parallelizations for all hyper-parameter searches.

5.3.3 Description of the Principle Component Analysis

The NN is expected to learn the correlation between different input features as well as their individual importance for the output. In case of many input features, this is not always the case. The NN can be helped by decorrelating the input features and eliminating the least important ones. A common way to do so is the principle component analysis (PCA).

The PCA can be used for two tasks: transform the input features into orthogonal variables and reduce the total number of input features. To do so, the initial input features $\vec{X} = (X_0, \ldots, X_N)^\intercal$ are standardized as described in section 5.2.2. A element of the Pearson correlation matrix $\rho_{i,j}$ is calculated as

$$\rho_{i,j} = \frac{E[X_i \cdot X_j] - E[X_i] \cdot E[X_j]}{\sigma_{X_i} \cdot \sigma_{X_j}}$$

for the input features X_i and X_j with the standard deviation denoted as σ_{X_i} and σ_{X_j}. $E[\ldots]$ indicates the expectation value of the respective input features or input feature product [73]. By definition, the matrix is symmetric and hence an orthonormal basis of eigenvectors of this matrix can be found. Transforming the initial inputs into this new basis, results in a new set of uncorrelated input features. The eigenvalues λ_i and eigenvectors $\vec{v}_i$ of the correlation matrix are calculated by solving the equation

$$\rho \cdot \vec{v}_i = \lambda_i \vec{v}_i$$

The transformation matrix is then constructed from the eigenvectors: Each column is one eigenvector. If the number of eigenvectors is smaller than the rank of the matrix ρ, orthonormal vectors are added to the transformation matrix with eigenvalues of zero. Thus, it is ensured that the transformation matrix has as many columns as initial input features exist. Every event is transformed by multiplying the transformation matrix from the right to the initial input feature vector

$$\vec{X} \to \vec{X}' = \vec{X}^\intercal \cdot (\vec{v}_0 \cdots \vec{v}_N)$$

Hence, the obtained uncorrelated input features are a linear combination of the initial input features [74].
In addition, a number of transformed input features can be eliminated. As the initial input features are standardized, the eigenvalue describes how many initial input features the corresponding eigenvector represents. Thus, an eigenvector with an eigenvalue greater than unity contains information of more than one initial input feature and the corresponding transformed input feature should be used in all cases. To eliminate an eigenvector and thus a transformed input feature, it has to be verified, that the eigenvalue is small compared to the sum over all eigenvalues [74].
Furthermore, the transformation matrix can be used to reduce the number of initial input features $\vec{X}$. As the transformation matrix is multiplied from the right to the initial input features, each initial input feature corresponds to exactly one row of the transformation matrix. By deriving some kind of norm of this row, an importance of the initial input feature is obtained. A reasonable threshold is defined eliminating all initial input features which have a norm smaller than this threshold. Thus, the complexity of the NN can be reduced without transforming the events. Furthermore, the meaning of the individual input features remains clear, whereas by performing the PCA transformation this clarity is lost.

6 Searches for Lepton-Flavour Violation in $H \to \tau\ell$ and $Z \to \tau\ell$ decays

The standard model of particle physics predicts the conservation of lepton-flavour number. In the non-charged lepton sector, however, this concept is questioned by the observation of neutrino oscillation. This means, that after some distance of flight, a neutrino of one flavour can be observed as a neutrino of another flavour, violating the predicted conservation [31]. Thus, the predicted lepton-flavour conservation does not hold in all cases, which is predicted by several extensions of the standard model [12–27]. Furthermore, these extensions predict also the violation of this conservation in the charged lepton sector. The charged lepton-flavour violation (LFV) is predicted to be induced by the Higgs boson and the Z boson, decaying into a lepton pair of two different flavours. However, the branching ratio for $H(Z) \to e^{\pm}\mu^{\mp}$ decays is limited indirectly to $\lesssim 10^{-8}$ [32,33] and is not likely to be observed. Hence, only $H(Z) \to \tau^{\pm}e^{\mp}$ and $H(Z) \to \tau^{\pm}\mu^{\mp}$ decays are studied. At a confidence level of 95%, an upper limit on the branching ratio for $H \to \tau e$ from direct searches for LFV in Run 1 of the LHC was set to 1.04% and for $H \to \tau\mu$ to 1.43% [34]. The upper limits on LFV decays of the Z boson into a τe or $\tau\mu$ pair are of the order of 10^{-5} [34].
To become sensitive to this signal, a good SM background estimation is crucial. However, a Monte-Carlo (MC) simulated background may result in large systematic uncertainties. This can be avoided by using the fully data driven *symmetry method* [75]. An overview of the general analysis strategy is given in section 6.1, followed by a description of the symmetry method in section 6.2. Further, the *maximum likelihood method* is presented, used to determine the significance of a signal excess over the SM background. Finally, existing mass reconstruction methods are described.

6.1 Analysis Strategy

The aim of this analysis is the observation of lepton-flavour violating decays of the Higgs boson and Z boson. For this decays into $\tau^{\pm}e^{\mp}$ and $\tau^{\pm}\mu^{\mp}$ lepton pairs are considered, with the τ lepton decaying further into a lepton and two neutrinos. However, the neutrinos escape the detector without any interaction. Hence, the signal consists of exactly two leptons of different flavours and opposite charge as well as missing energy.
A main assumption of this analysis is, that the SM processes are symmetric with respect to the exchange of electrons by muons and vice versa. The mass difference between the electron and the muon can be neglected at the energy scales of the LHC. This symmetry is used to split all data into two samples, one where the electron has a larger transverse momentum than the muon ($e\mu$ sample) and one where it has a smaller one (μe sample). Since all SM processes are expected to be equally distributed in both samples, the $e\mu$ sample is used as SM background estimation of the μe sample and vice versa. In case of a lepton-flavour violating $H \to \tau^{\pm}\mu^{\mp}$ decay, the event is more likely to end up in the μe sample than in the $e\mu$ one, resulting in a signal excess. Thus, a fully data-driven analysis can be carried out.
However, due to detector effects and reconstruction algorithms, this symmetry can be broken

without actual LFV processes. One effect are the detector efficiencies for electrons and muons which are unequal and must be corrected. Furthermore, electrons and muons are detected completely different, which produces different contributions to the fake background due to other objects being misidentified as leptons. Thus, the fake background estimation must be performed separately in a data driven way, described in chapter 8.
With a performed background estimation, the signal excess is looked for in the reconstructed mass distribution of the Higgs boson and the Z boson respectively. The signal significance is derived using the maximum-likelihood method. An enhanced mass reconstruction is important to separate signal and background processes. This can be achieved by using a regression neural network with various input features of the measured objects, described in chapter 9.

6.2 The Symmetry Method

Monte-Carlo simulations are used in many analyses for the SM background estimation. However, to do so several theory and simulation uncertainties are propagated, which can result in large systematic uncertainties in the analysis. This can be avoided by using the fully data driven symmetry method for the background estimation [75]. Two assumptions are used for this method:

1. Lepton universality for SM background processes, which means that electrons and muons have approximately the same kinematic distributions for sufficiently high energies as those present at the LHC. Thus, the prediction from SM processes are approximately the same when replacing an electron with a muon and vice versa.

2. The signal processes $H \to \tau e$ or $H \to \tau\mu$ break this symmetry, since in the $H \to \tau e$ process, the electron has in general a larger transverse momentum than the muon which originates from the τ-lepton decay. Analogous, in the $H \to \tau\mu$ process the muon has in general a larger transverse momentum than the electron.

This method is sensitive to the branching ratio of both symmetry breaking Higgs boson decays, but only if these are not of equal size. In the case of $BR(H \to \tau e) \approx BR(H \to \tau\mu)$, the method will provide no sensitivity to any lepton-flavour violating decay. However, previous limits indicate a non-equal branching ratio [34].
The leptonic τ-lepton decay induces a hierarchy of the electron and muon with respect to the transverse momentum. Since the τ-lepton decays in three particles, the resulting lepton has in general a softer transverse momentum than the one originating directly from the Higgs boson decay. This hierarchy is used to split the data into two distinct samples.
To do so, the leading and sub-leading lepton of each event are defined, as the lepton with the highest and second highest transverse momentum p_T in the final state, respectively. The data is then split into two exclusive samples:

- μe sample: $p_T^\mu \geq p_T^e$
- $e\mu$ sample: $p_T^e > p_T^\mu$

In case of LFV decays, the leading lepton is assumed to originate directly from the Higgs boson decay and the sub-leading lepton from the τ-lepton decay. With the first assumption, the SM background of a $H \to \tau\mu$ decay can therefore be estimated from the inverted lepton sample $e\mu$ and for a $H \to \tau e$ decay vice versa, i.e. the μe sample.
In a study performed by Bressler et al. (2014) the symmetry method was tested the first time, displayed in figure 6.1 [75]. The collinear mass distributions of MC simulated events at a centre of

mass energy of 8 TeV are shown for the $e\mu$ and μe sample. A branching ratio of 2% is assumed for the $H \to \tau\mu$ decay and zero for the $H \to \tau e$ decay. In the mass range between 100 and 150 GeV, a signal peak is visible, inducing an asymmetry as assumed. Furthermore, no additional deviation between the two samples in all other mass ranges is visible, supporting the first assumption.

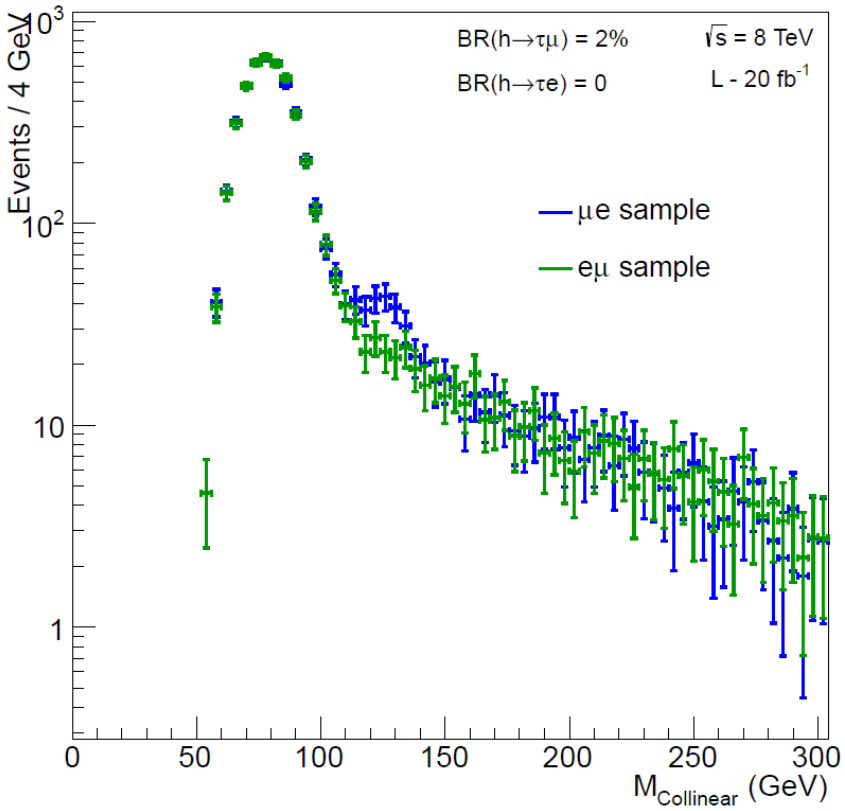


Figure 6.1: Collinear mass for the $e\mu$ and μe sample. A branching ratio of 2% is for $H \to \tau\mu$ assumed and zero for $H \to \tau e$. Both distributions are determined from Monte Carlo based predictions at a centre of mass energy of 8 TeV [75].

Since electrons and muons are detected by different parts of the ATLAS detector, there are several detector effects resulting in a broken symmetry without any lepton-flavour violating process being present. One effect are the different trigger, reconstruction, isolation and identification efficiencies for electrons and muons. All these efficiencies are combined to one common efficiency for each sample, noted as $\epsilon^{e\mu}$ and $\epsilon^{\mu e}$ respectively. The combined efficiency is derived as

$$\epsilon^{e\mu(\mu e)} = \epsilon^{\text{trigger}} \cdot \epsilon^{e} \cdot \epsilon^{\mu}, \tag{6.1}$$

where the efficiency of the electron ϵ^e and muon ϵ^μ contain the reconstruction, isolation and identification efficiency. Both efficiencies ϵ^e and ϵ^μ are dependent on the transverse momentum and pseudorapidity of the respective lepton. For a $e\mu$ event, $\epsilon^{e\mu}$ is derived by equation (6.1), whereas the efficiency $\epsilon^{\mu e}$ is obtained from a 'virtual' event, where the kinematic properties of the electron and muon are exchanged. The detected data yield $N^{e\mu}$ in the $e\mu$ sample is scaled by

$$\tilde{N}^{\mu e} = \frac{\epsilon^{\mu e}}{\epsilon^{e\mu}} \cdot N^{e\mu}, \tag{6.2}$$

where $\tilde{N}^{\mu e}$ is than used as the SM background estimate for the μe sample. For the estimation of $\tilde{N}^{e\mu}$, all indices are switched, i.e. $e\mu \leftrightarrow \mu e$.

Electrons and muons are differently misidentified, resulting in different fake backgrounds. This is taken into account, by performing separate fake background estimations, with a dependency of the fake background contribution on the transverse momentum and pseudorapidity of the misidentified lepton. The contributions are respectively subtracted from the data to restore the symmetry between the electron and the muon. With this, equation (6.2) gets modified to

$$\tilde{N}^{\mu e} = \frac{\epsilon^{\mu e}}{\epsilon^{e\mu}} \cdot (N^{e\mu} - f^{e\mu}) + f^{\mu e} \tag{6.3}$$

where $f^{e\mu}$ and $f^{\mu e}$ are the respective fake background contributions. The estimation of the fake background is described in more detail in chapter 8.
The symmetry method was used by ATLAS in Run 1 of the LHC [34].

6.3 Extracting the Signal with the Maximum Likelihood Method

The observation of a signal in the data for physics beyond the standard model requires some kind of statistical significance for the signal above the SM background.
For this, two hypotheses are formulated. One is the null-hypothesis, which states, that only the SM processes are present and the data is fully described by those. The other one is the signal plus background hypothesis, which assumes a signal above the SM processes and states, that the data is described by the sum of both [76].
The observed data is tested against a stated hypothesis by calculating the p-value. This is the probability, that the observed data, under assumption of the hypothesis, is equal or more incompatible with the actual data observed. Thus, a sufficient small p-value excludes the stated hypothesis. An equivalent value is the Gaussian significance Z. It is connected to the p-value in such a way, that a Gaussian distributed variable located Z standard deviations above its mean, has an upper-tail probability half as large as the p-value. Is a significance of $Z \geq 3$ ($p \leq 2.7 \cdot 10^{-3}$) observed in the data with respect to the null hypothesis, it is called evidence and for $Z \geq 5$ ($p \leq 5.7 \cdot 10^{-7}$) an observation.
The p-value and the significance Z can be estimated by the maximum-likelihood method (ML-method) described in the following.
The measurements for observable O per event are filled into a histogram with N bins, with $\vec{n} = (n_1, \ldots, n_N)$ being the number of events per bin observed in the data. The expectation value from signal and background in bin i is given as

$$E[n_i] = \mu s_i + b_i \tag{6.4}$$

with the number of signal events s_i and background events b_i. The parameter μ describes the signal strength, where $\mu = 1$ corresponds to the signal plus background hypothesis and $\mu = 0$ to the null-hypothesis. Since all bins are assumed to be independent of each other and each event is randomly distributed, the number of events in each bin is Poisson distributed, given as

$$\mathcal{P}_{E[n_i]}(k) = \frac{E[n_i]^k}{k!} e^{-E[n_i]} \tag{6.5}$$

with the mean and variance $E[n_i]$.
The likelihood function is defined as function of a parameter of $\mathcal{P}$ for a given k, i.e. $L(\kappa|k) = L_k(\kappa)$, where κ is a parameter of $\mathcal{P}$. In one bin, the likelihood function is obtained as

$$L_{n_i}(\mu) = \frac{E[n_i]^{n_i}}{n_i!} e^{-E[n_i]} \tag{6.6}$$

with a dependency on the strength parameter μ and a fixed parameter $k = n_i$. Thus, the likelihood function specifies the probability of n_i events for a given μ. Taking the product of the probabilities of all bins, the likelihood function can be written as

$$L(\mu, \theta) = \prod_{j=1}^{N} \frac{(E[n_j])^{n_j}}{n_j!} e^{-E[n_j]} \cdot g(\theta) \tag{6.7}$$

where any dependency on nuisance parameters is accounted for in the function $g(\theta)$, which is in many cases a Gaussian distribution. These parameters are additional dependencies which are not of direct interest, but their influence on the likelihood function must be taken into account. In general, there are several nuisance parameters and thus θ is a vector.
Maximizing this function provides the respective unconditional estimators $\hat{\mu}$ and $\hat{\theta}$, which describe the observed data best. In addition, the likelihood function is maximized for one specific signal strength, which is given by the stated hypothesis. This results in the conditional estimator $\hat{\hat{\theta}}$, which describes the observed data best for the given signal strength.
Minimizing the negative logarithm of equation (6.7)

$$\ln L(\mu, \theta) = \sum_{j=1}^{N} n_j \ln E[n_j] - E[n_j] - n_j + \ln g(\theta) \tag{6.8}$$

is equivalent to the maximization of the likelihood function. However, minimizing sums is computational much more efficient than products. Furthermore, any term without a dependency on μ or θ can be neglected in the minimization process.
The Neyman-Pearson lemma states that in the absence of unknown parameters the optimal test statistic for the test of the background-only hypothesis against the signal plus background hypothesis is the likelihood ratio $L(\mu = 0)/L(\mu = 1)$ [77]. In the presence of unknown parameters, the profile likelihood ratio is empirically a very good test statistic, given as

$$\lambda(\mu) = \frac{L(\mu, \hat{\hat{\theta}})}{L(\hat{\mu}, \hat{\theta})} \tag{6.9}$$

with the conditional likelihood function in the numerator and the unconditional one in the denominator. The unconditional likelihood function is always greater equal than the conditional likelihood function, since both μ and θ are optimized simultaneously, resulting the maximal likelihood value. Furthermore, all expectation values in equation (6.4) are non-negative and thus the profile likelihood ratio is restricted to $\lambda \in [0, 1]$. The proposed hypothesis has a good agreement with the data if λ is close to unity. Thus, assuming the background-only hypothesis, $\lambda(\mu = 0) \to 0$ suggests processes not covered by the background-only hypothesis.
The test statistic can be defined as

$$t_\mu = -2 \ln \lambda(\mu) \tag{6.10}$$

where incompatibility between the hypothesis and the data is described by higher values of t_μ. In the case of $\mu = 0$

$$Z_0 = \sqrt{t_0} \tag{6.11}$$

provides the searched Gaussian significance assuming that the Wilks' approximation is valid [76]. A larger Gaussian significance indicates the existence of a process not covered in the background-only hypothesis.

The so-called *Asimov dataset* is defined in such a way, that when it is used to evaluate the estimators μ and θ of the likelihood function, the true parameter values are obtained. This means, that assuming a signal strength μ' for simulated events, the maximization of the likelihood function for an Asimov dataset will return $\hat{\mu} = \mu'$. Thus, this dataset can only be constructed by simulated signal and background events, but not by actually measured data. Neglecting all possible nuisance parameters, the profile likelihood ratio assuming the background-only hypothesis is given by

$$\lambda(\mu = 0) = \frac{L(0)}{L(\hat{\mu})} \tag{6.12}$$

with a signal strength of $\hat{\mu} = 1$ in case of an Asimov dataset. This dataset allows a fast estimation of the expected significance of a signal process and thus of the expected sensitivity.
Evaluating equation 6.11 for the background-only hypothesis without any nuisance parameters and a assumed signal strength of $\hat{\mu} = 1$ results in

$$Z = \sqrt{-2 \sum_{j=1}^{N} n_j \cdot \ln\left(\frac{b_j}{s_j + b_j}\right) - 2 \cdot s_j}. \tag{6.13}$$

In this analysis, a signal excess is searched in the mass distribution and the ML-method is applied to this distribution. Thus, a higher signal significance is derived with a better mass reconstruction.

6.4 Mass Reconstruction of the Higgs Boson and Z Boson Candidate

The LFV signal processes consist of a Higgs or Z boson, which decays into a τ-lepton and either an electron or a muon, with the τ-lepton decaying further into a light lepton and two neutrinos. Thus, an electron, a muon and two neutrinos are present in the final state. However, the neutrinos escape the detector undetected, resulting in missing transverse energy. The individual momenta of the neutrinos are unknown. In general, the mass of the Higgs boson is reconstructed by using the detected leptons and the missing transverse energy.
One way to reconstruct the mass of the Higgs boson relies on the collinear mass approximation [78]. The large mass difference between the Higgs-boson and the final leptons results in a large Lorentz boost of the decay leptons. This allows to neglect the respective lepton masses for the Higgs-mass reconstruction and leads to a collinear flight direction of the decay products of the τ-lepton.
In the transverse plane of the detector, the momentum of the τ-lepton ($\vec{p}_{T,\tau}$) can be written as

$$\vec{p}_{T,\tau} = \frac{\vec{p}_{T,\ell_1}}{x} \approx \vec{p}_{T,\ell_1} + \vec{E}_T^{\text{miss}} \tag{6.14}$$

with the momentum fraction x of the sub-leading lepton ($\vec{p}_{T,\ell_1}$), stemming from the τ-decay, with respect to the initial τ-lepton. It is assumed that the missing transverse energy $\vec{E}_T^{\text{miss}}$ consist only of neutrino contributions. However, in general additional detector effects result in a missing transverse energy as well.
The invariant mass of the Higgs boson is given as

$$m_H^2 = (p_\tau^\mu + p_{\ell_0}^\mu)^2 \tag{6.15}$$

where p_i^μ denotes the respective four-momentum. Neglecting the lepton masses and approximating $p_\tau^\mu \approx p_{\ell_1}^\mu / x$, the mass of the Higgs boson is approximated as

$$m_{\mathrm{H}}^2 \approx 2 \frac{p_{\ell_1}^\mu \cdot p_{\ell_0}^\mu}{x} \tag{6.16}$$

Multiplying equation (6.14) with $\vec{p}_{T,\ell_1}$, the equation can be solved for the momentum fraction

$$x = \frac{\vec{p}_{T,\ell_1}^{\,2}}{\vec{p}_{T,\ell_1}^{\,2} + \vec{p}_{T,\ell_1} \cdot \vec{E}_T^{\mathrm{miss}}} \tag{6.17}$$

With the collinear approximation of the τ-lepton decay products, equation (6.17) is simplified to $x = |\vec{p}_{T,\ell_1}|/(|\vec{p}_{T,\ell_1}| + |\vec{E}_T^{\mathrm{miss}}|)$.
The pseudorapidity is given as $\eta = -\ln\tan(\theta/2)$ with the polar angle θ. Using this and neglecting the masses of the leptons, the collinear mass can be derived from equation (6.16)to

$$m_{\mathrm{coll}}^2 = 2|\vec{p}_{T,\ell_0}| \left(|\vec{p}_{T,\ell_1}| + |\vec{E}_T^{\mathrm{miss}}| \right) (\cosh \Delta\eta - \cos \Delta\phi) \tag{6.18}$$

with $\Delta\phi$ being the azimuth difference between leading and sub-leading lepton and $\Delta\eta$ the corresponding pseudorapidity difference. However, the missing transverse energy and the sub-leading lepton are not aligned in general. In case of an $H \to \tau e$ decay, the azimuth difference is shown in figure 6.2. The majority of the events are not perfectly aligned impairing the approximation.

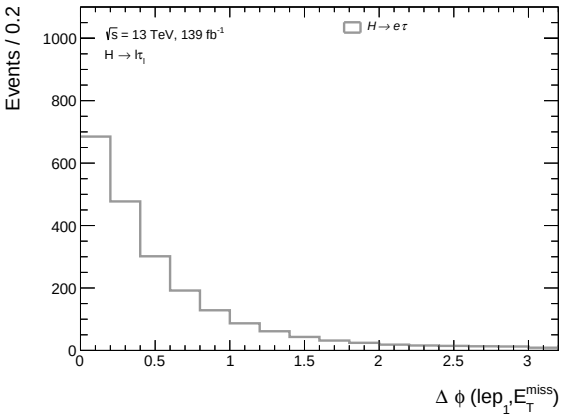


Figure 6.2: Azimuth angle difference between the missing transverse energy and the sub-leading lepton in the $H \to \tau e$ decay process. In many cases, they are not perfectly aligned to each other which impairs the collinear approximation.

Furthermore, this method is very sensitive to the resolution of the missing transverse energy, which results in most cases in an overestimation of the reconstructed mass [79]. This broadened resolution results in a worse LFV signal sensitivity Z, as discussed in equation (6.13).

An alternative mass reconstruction method is the missing mass calculator (MMC) [79]. The invariant τ mass is derived as

$$m_\tau^2 = \left(p_{2\nu}^\mu + p_{\ell_1}^\mu \right)^2 \tag{6.19}$$

with the combined neutrino four-momentum $p^{\mu}_{2\nu}$. The mass of the neutrino system is derived as

$$m^2_{2\nu} = E^2_{2\nu} - \vec{p}^{\,2}_{2\nu} = E^2_{2\nu} - p^2_{2\nu,z} - {E_T^{\text{miss}}}^2 \tag{6.20}$$

where the relation ${E_T^{\text{miss}}}^2 = p^2_{2\nu,x} + p^2_{2\nu,y}$ is used, stating that all missing energy is due to the neutrinos. Thus, two equations are obtained

$$\vec{p}_{2\nu} \cdot \vec{p}_{\ell_1} = \left(m^2_{2\nu} + m^2_{\ell_1} + 2E_{2\nu}E_{\ell_1} - m^2_{\tau}\right)/2 \tag{6.21}$$

$$p^2_{2\nu,z} = E^2_{2\nu} - m^2_{2\nu} - {E_T^{\text{miss}}}^2 \tag{6.22}$$

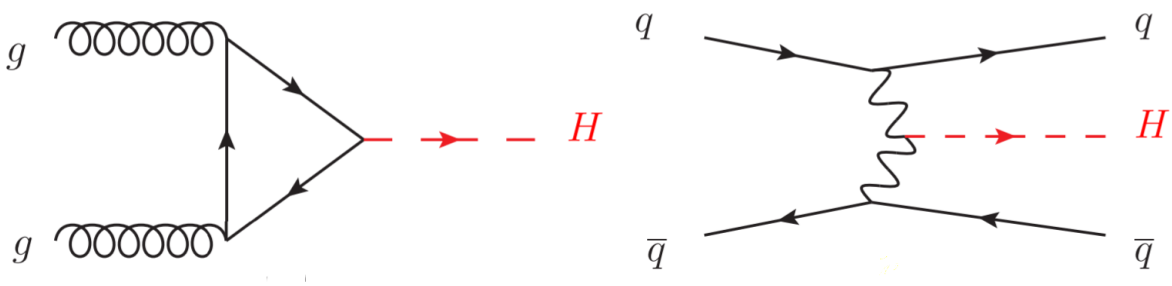
where $p_{2\nu,z}$, $E_{2\nu}$ and $m_{2\nu}$ are unknown parameters. This is an under-determined system of equations, which allows one unknown parameter to be freely chosen.
The idea of the MMC is to perform a scan, for which the system of equations can be solved. This is done, by a scan over the x- and y-component of the neutrino momentum and eliminating $E_{2\nu}$ in the equations. For both components, 40 equidistant scan points are used. However, taking detector effects into account, the values are smeared out corresponding to the respective resolutions. The scan allows to calculate for each point a new mass for the neutrino system by setting the mass of the individual neutrinos to zero. Thus, $p_{2\nu,z}$ and $E_{2\nu}$ can be derived by solving the equations (6.21) and (6.22). This completes the reconstruction of the neutrino system and thus of the τ-lepton. For each scan point, the invariant mass of the $\tau\ell_0$-lepton system $m_{\tau\ell_0}$ is calculated by equation (6.15). The weight of each scan point is calculated by probability distribution functions obtained with simulated events. For all three objects, separate probabilities are derived and multiplied to obtain the combined weight.
Over all scan points, a reconstructed mass distribution $m_{\tau\ell_0}$ is produced, where every point is weighted by its probability. Now the MMC mass value can be extracted in two different ways. The mass value, which is the most probable of the distribution, i.e. the bin containing the maximum, is chosen as final estimator and is noted as MMC_{mlm}. Another version of the MMC, noted as MMC_{maxw}, takes as reconstructed mass the scan point with the largest individual weight.
In the process of the parameter scan, it is possible that no parameter value provides a solution for the system of equations, since only a specific parameter space is scanned. Events with no solution of the system of equations, lacking a mass reconstruction of the Higgs or Z boson, must be eliminated from the data set, if the MMC is used.
This method improves significantly the mass resolution with respect to the collinear mass approximation. Fewer assumptions and approximations are necessary, allowing a more reliable mass reconstruction. However, a major drawback is a more complicated and time intensive calculation. Thus, new mass reconstruction methods are of great interest.

7 Signal and Background Processes and Analysis Selection

This chapter presents the signal and background processes, considered for the LFV analysis and the applied analysis selection. The individual signal and background contributions are described and their impact is discussed. All signal and background events are simulated with Monte-Carlo (MC) methods. For the particle propagation through the ATLAS detector and simulation of its response, GEANT4 [80] is used. The influence of several proton-proton collisions in the same bunch crossing is simulated by overlaying the produced hits by several types of events and weighting the simulated events with a so-called pile-up weight, such that the average number of interactions match the data best. Furthermore, the MC simulation is modified to fit the data best by various efficiency correction factors. For instance, events with sufficient low properties like the energy or transverse momentum are eliminated on generator level by the generator filter. Hence the monitored cross-section is lower than the initially generated one and must be adjusted by the filter efficiency. Further efficiency factors are the trigger, isolation and identification efficiencies [81].
A summary of the respective generators and cross-sections for the individual processes are presented in this chapter, followed by the applied analysis selection.

7.1 Signal Processes

All considered LFV signal processes contain a Higgs boson decaying into a $\tau\ell$ pair, where the lepton ℓ is either an electron or a muon. This results in the four major signal production processes: gluon-gluon fusion (ggH), vector boson fusion (VBF) and the associated WH and ZH production. For all of the Higgs LFV decay processes, a branching ratio of 1% is assumed. The τ-lepton decays are simulated with `PYTHIA8` and leptonic decays to $e\nu_e\nu_\tau$ and $\mu\nu_\mu\nu_\tau$ as well as hadronic decays are considered in the production.
The ggH and VBF processes are simulated at next-to-leading-order (NLO) in perturbative quantum chromodynamics (pQCD) using the `POWHEG` [82] event generator in combination with `PYTHIA8` [83] for the hadronisation and parton showering. For ggH processes the parton distribution function set PDF4LHC15 NNLO [84] is used, whereas for the VBF production PDF4LHC15 NLO is used. The underlying event is tuned with `AZNLO_CTEQ6L1` [85,86].
The LFV decay of the Higgs boson is modelled with `PYTHIA8`. A precision up to next-to-next-to-leading-order (NNLO) in pQCD is achieved by a reweighting of the NLO cross-section with a k-factor [87]. The NLO cross-section is derived with `MiNLO` [88]. A cross-section times branching ratio of 485.8 fb is used for the ggH signal process. In case of $H \to \tau e$ a generator filter efficiency of 0.640 is used and for $H \to \tau\mu$ of 0.638. The VBF events are generated with a cross-section times branching ratio of 37.8 fb and for both decay modes the filter efficiency is 0.672 [41].
Feynman diagrams for ggH and VBF Higgs boson production are shown in figure 7.1. In the VBF production mode, two jets are produced in addition to the Higgs boson.
For the signal processes WH and ZH, `POWHEG` in combination with `PYTHIA8` is used in the same way as before, with the parton distribution function set PDF4LHC15 NLO. A matrix element



(a) ggH (b) VBF

Figure 7.1: Feynman diagrams for ggH (a) and VBF (b) Higgs boson production [35].

rescaling is performed with `MiNLO` to account for next-to-leading order effects [88]. All decay modes of the W and Z bosons are considered. For the $WH \to \tau\ell$ process, a cross-section times branching ratio of 8.5 fb is assumed and for $ZH \to \tau\ell$ 8.4 fb respectively [41]. A filter efficiency of unity is used for both signal processes WH and ZH.
A Feynman diagram for the WH and ZH Higgs boson production is displayed in figure 7.2.

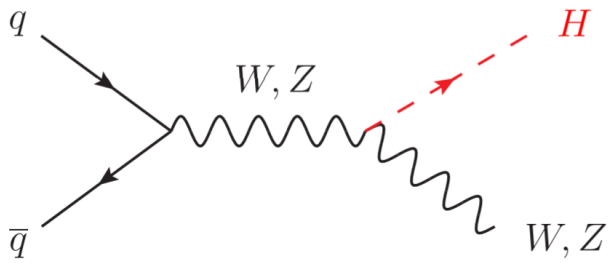


Figure 7.2: Feynman diagram for W/Z associated Higgs boson production [35].

Any heavy flavour decay is simulated with `EvtGen` [89].

7.2 Background Processes

All background processes are simulated with MC methods, except the contribution due to fake leptons. The latter is described in detail in chapter 8.
The leading SM background process is the production of a Z boson or a virtual photon γ^* decaying into two τ-leptons, which decay further as $Z/\gamma^* \to \tau^\pm\tau^\mp \to e^\pm\mu^\mp + \nu_e + \nu_\mu + 2\nu_\tau$. The kinematics of the final state are very similar to the LFV signal process, so that Z/γ^* production is an irreducible background. The cross-section of $Z/\gamma^* \to \tau\tau$ production is by four orders of magnitude larger than the cross-section of the signal process [90]. Hence, a good separation between this SM background and the LFV signal is crucial for the observation of a signal excess. The $Z/\gamma^* \to \tau\tau$ production is simulated using `Sherpa`(v2.2.1) [91] and the parton distribution function NNPDF30NNLO [92]. Processes with less than three jets are simulated up to NLO in pQCD and processes with more than three jets at LO precision. `Sherpa` uses its own model for hadronisation and parton showering.
The second largest contribution results from diboson processes, i.e. the production of WW, ZZ and WZ. The diboson systems decaying into $\ell\ell + 2\nu$ have the largest contribution of all diboson processes to the measurement backgrounds. Processes, with one boson decaying into quarks are

simulated with Sherpa (2.2.1). The parton distribution function NNPDF30NNLO is used and a precision up to NLO in pQCD is achieved for processes with zero or one jet and LO in pQCD if more than one jet is present. For leptonic decays of both bosons Sherpa (2.2.2) is used, with the parton distribution function NNPDF30NNLO. This production is split in events with at least one same flavour opposite charge lepton pair with an invariant mass below four GeV and events where the invariant mass of all lepton pairs is above four GeV. The first one is called in the following 'lowMll' and the latter 'highMll'. In case of the lowMll production, events without jets are simulated at NLO in pQCD and otherwise at LO. The highMll production has a similar precision, but in addition events with one jet are simulated up to NLO in pQCD as well.
The SM top quark background contribution is about the same size as the diboson contribution, displaying a similar invariant dilepton mass distribution. The $t\bar{t}$ production is considered as well as the single top quark production Wt. All decay modes of the $t\bar{t}$ and Wt production are considered in the simulation. The $t\bar{t}$ contribution is after the event selection about four times larger than the Wt contribution. Both, $t\bar{t}$ and single top productions are simulated with POWHEG in combination with PYTHIA8 for the parton showering and hadronisation. For both top quark productions the parton distribution function NNPDF30NLO is used. Heavy flavour decays are simulated using EvtGen and the underlying event is tuned with A14_NNPDF23LO.
The SM background contributions due to $Z/\gamma^* \to e^{\pm}e^{\mp}$ and $Z/\gamma^* \to \mu^{\pm}\mu^{\mp}$ are also produced and are simulated in the same way as $Z/\gamma^* \to \tau\tau$.
Contributions through gluon-gluon fusion and vector boson fusion $H \to \tau\tau$ productions are also considered. Leptonic or semi-leptonic decays of the τ-leptons are taken into account. The Higgs boson in association with a $t\bar{t}$-pair results mostly in events with many jets. Inclusive ZH and WH processes are also considered. Events are simulated using the event generator POWHEG and PYTHIA8 for the particle showering. For all SM Higgs boson processes, the parton distribution function NNPDF30NLO is used. The EvtGen is used for heavy flavour decays and MiNLO to derive the NLO cross-section precision in pQCD. The underlying event is tuned with AZNLO_CTEQ6L1.
In table 7.1 the generators of the signal and background processes are summarized, together with the corresponding cross-section times branching ratio.

7.3 Analysis Selection

The analysis selection is optimized to reduce backgrounds to the LFV signal process and to select candidate events from LFV $H \to e\tau \to e\mu 2\nu$ or $H \to \mu\tau \to \mu e 2\nu$ decays. A summary of the selection criteria is given in table 7.2.
Events are selected in shich at least one reconstructed vertex is present, containing at least two tracks [93]. Exactly one electron and one muon are required for each event, including a veto against hadronically decaying τ-leptons.
The leading lepton, with the highest p_T, and sub-leading lepton are required to be of opposite sign charge and must fulfil the identification and isolation criteria specified in section 4.1.
The events are assigned to the $e\mu$ channel if the electron has the higher transverse momentum and to the μe channel otherwise. It is ensured that at least on trigger is fired as described in section 4.2. Events are eliminated, if a lepton is outside the pseudorapidity range which is covered by the inner detector or within the transition area of the barrel and the end caps of the detector. This is applied to both electrons and muons, to ensure a symmetry between both.
The minimal transverse momentum threshold of the leading and sub-leading lepton are used to reduce mainly the $Z \to \tau\tau$ background and to avoid unreliable measurements since the detector is not sensitive to arbitrary low momenta. These two requirements reduce the signal by about 13 %, whereas $Z \to \tau\tau$ contributions are reduced by over 50 %.

Table 7.1: Cross-section times branching ratio and generators of simulated signal and background processes. For the cross-section calculation, generator filter efficiencies are included. The processes are ordered by their contribution in the signal region after the analysis selection.

Process	Generator	Cross-section × branching ratio [pb]
Signal		
$ggH \to \tau e(\tau\mu)$	POWHEG+PYTHIA8	0.311 (0.310)
VBF$\to \tau e(\tau\mu)$	POWHEG+PYTHIA8	0.025 (0.025)
WH$\to \tau e(\tau\mu)$	POWHEG+PYTHIA8	0.019 (0.019)
ZH$\to \tau e(\tau\mu)$	POWHEG+PYTHIA8	0.009 (0.009)
Background		
$Z/\gamma^* \to \tau\tau$	Sherpa	2721.0
WW	Sherpa	12.7
WZ	Sherpa	11.0
ZZ	Sherpa	4.9
$t\bar{t}$	POWHEG+PYTHIA8	87.7
single top	POWHEG+PYTHIA8	71.7
$Z/\gamma^* \to ee/\mu\mu$	Sherpa	9038.0
$ggH\ H \to \tau\tau$	POWHEG+PYTHIA8	0.6
VBF $H \to \tau\tau$	POWHEG+PYTHIA8	0.4
$t\bar{t}H$	POWHEG+PYTHIA8	0.5
WH	POWHEG+PYTHIA8	0.09
ZH	POWHEG+PYTHIA8	0.06

Table 7.2: Summary of the analysis selection criteria. Only events fulfilling all requirements are kept.

Selection number	Cut name	Selection criteria
Selection 1	reconstructed vertex	$N_{pvx} \geq 1$
Selection 2	two leptons	$N_e + N_\mu = 2$
Selection 3	tau veto	no hadronic τ decays
Selection 4	opposite charge	$q_{\ell_0} \cdot q_{\ell_1} \leq 0$
Selection 5	identification & isolation	flag (see chapter 4)
Selection 6	channel	$e\mu$ if $p_T^e > p_T^\mu$, μe if $p_T^\mu \geq p_T^e$
Selection 7	trigger	flag (see section 4.2)
Selection 8	pseudorapidity	$\lvert\eta\rvert < 2.47$ & ($\lvert\eta\rvert < 1.37$ or $1.52 > \lvert\eta\rvert$)
Selection 9	leading lepton	$p_T(\ell_0) \geq 35\,\text{GeV}$
Selection 10	sub-leading lepton	$p_T(\ell_1) \geq 15\,\text{GeV}$
Selection 11	visible mass	$30\,\text{GeV} < m_{\text{vis}} < 150\,\text{GeV}$
Selection 12	b-jets veto	no b-jets
Selection 13	impact parameters	$\lvert d_{0,\text{signif.}}\rvert < 10$ & $\lvert z_0 \sin\theta\rvert < 0.5\,\text{mm}$

In figure 7.3 both momenta are shown before a tighter requirement is applied. The transverse momentum distribution of the leading lepton for the signal process has its peak at around 60 GeV, which allows to raise the threshold and eliminate further background contributions.
The visible mass of the dilepton system is required to be within the range of 30 GeV to 150 GeV. As shown in figure 7.4, this mainly reduces top quark background contributions. This background

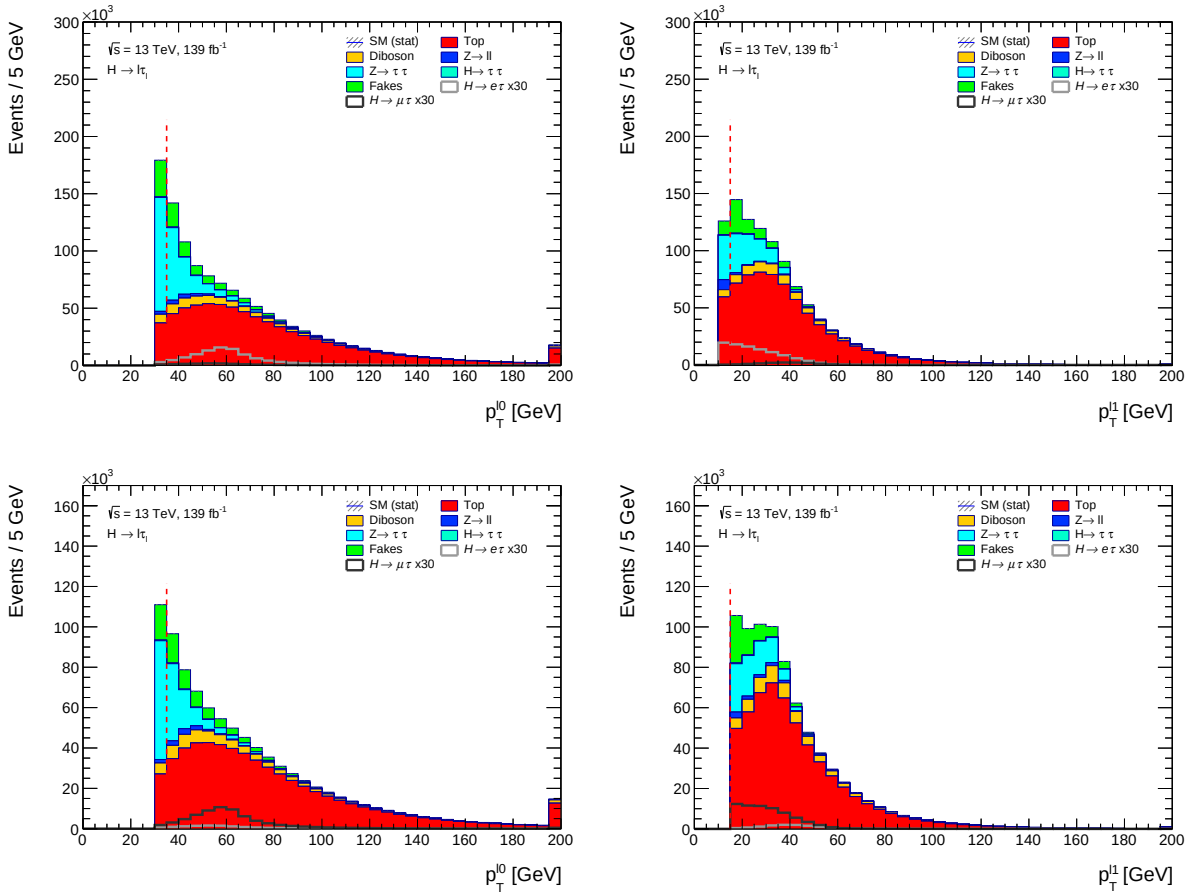


Figure 7.3: Transverse momentum distribution of the leading (left) and sub-leading lepton (right) in the $e\mu$ (top) and μe-channel (bottom), before a tighter selection requirement is applied. Only events with $p_T(\ell_0) \geq 35\,\text{GeV}$ and $p_T(\ell_1) \geq 15\,\text{GeV}$ are used, indicated by the dashed line. The $p_T(\ell_1)$ requirement has no influence in the μe channel, since in the production electrons are required to have a transverse momentum greater 15 GeV. The LFV signal is scaled in the figure by a factor of 30.

contribution is reduced further by applying a veto on jets containing b-quarks, shown in figure 7.4. These jets are tagged by *MV2c10* at the *FixedCutBEff_85* working point. With these two requirements, the top quark background is reduced by about 93 %, whereas the signal is only reduced by about 8 %.

Finally, the longitudinal impact parameter $z_0 \sin\theta$ and the significance of the transverse impact parameter $d_{0,\text{signif.}}$ are restricted for both leptons to $|z_0 \sin\theta| < 0.5\,\text{mm}$ and $|d_{0,\text{signif.}}| < 10$. The applied requirements reduce additional fake background contributions by about 28 % and the signal by 4 %. In figure 7.5, both distributions in the $e\mu$-channel are displayed before the respective selections are applied.

In table 7.3 and 7.4, the expected event yield in the $e\mu$ and μe-channel for the individual signal and background processes are summarized for specific selections. The respective ratio of signal over background is displayed as well. All data-taking years are combined.

With the applied selection, the signal over background ratio is about 1.3% in both channels. However, to enhance this ratio machine learning techniques are used to separate the signal

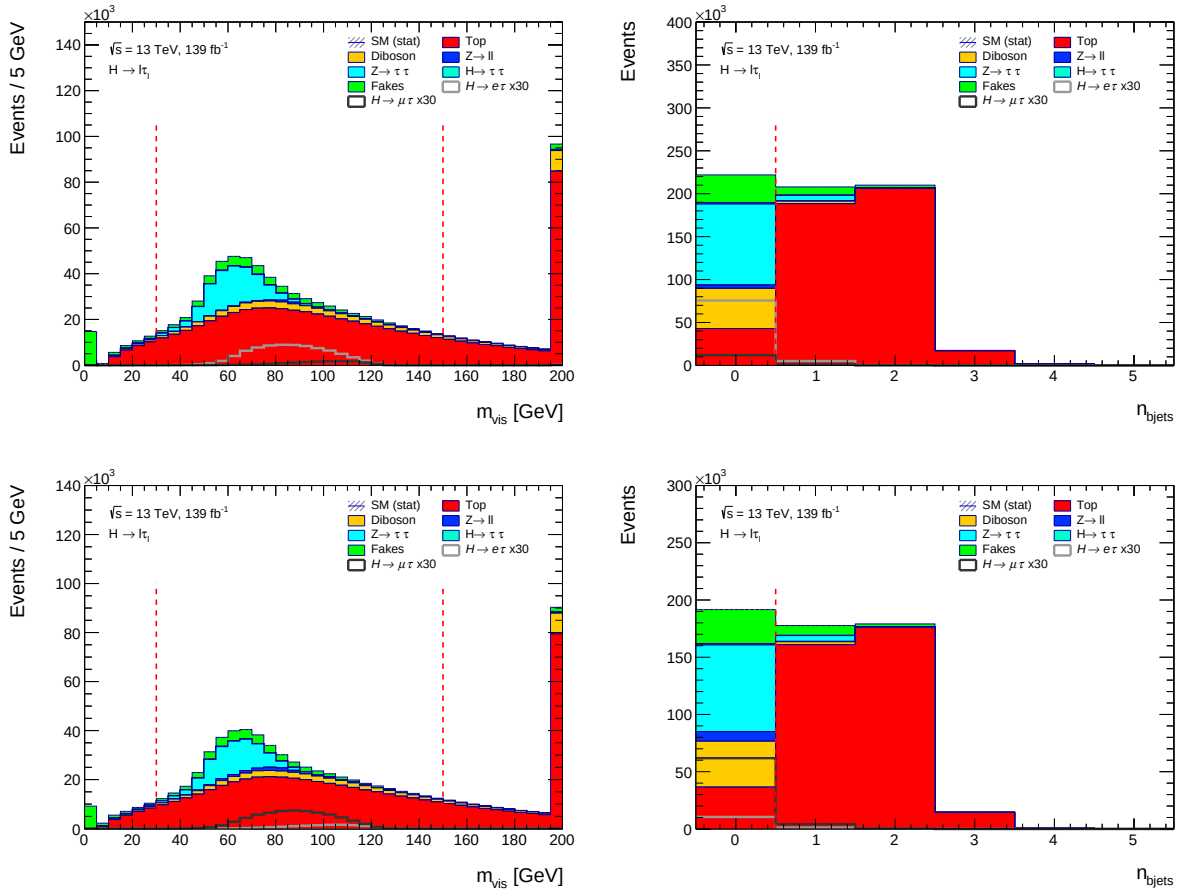


Figure 7.4: Visible mass m_{vis} of the dilepton system (left) and number of b-quark jets (right) in the $e\mu$ (top) and μe-channel (bottom), before the respective selection is applied. Events with a visible mass outside of 30 GeV and 150 GeV or any jet containing b-quarks are eliminated, indicated by the vertical dashed lines. Mainly background contributions from top quarks are reduced. The LFV signal is scaled in the figure by a factor of 30.

processes better from the background processes.

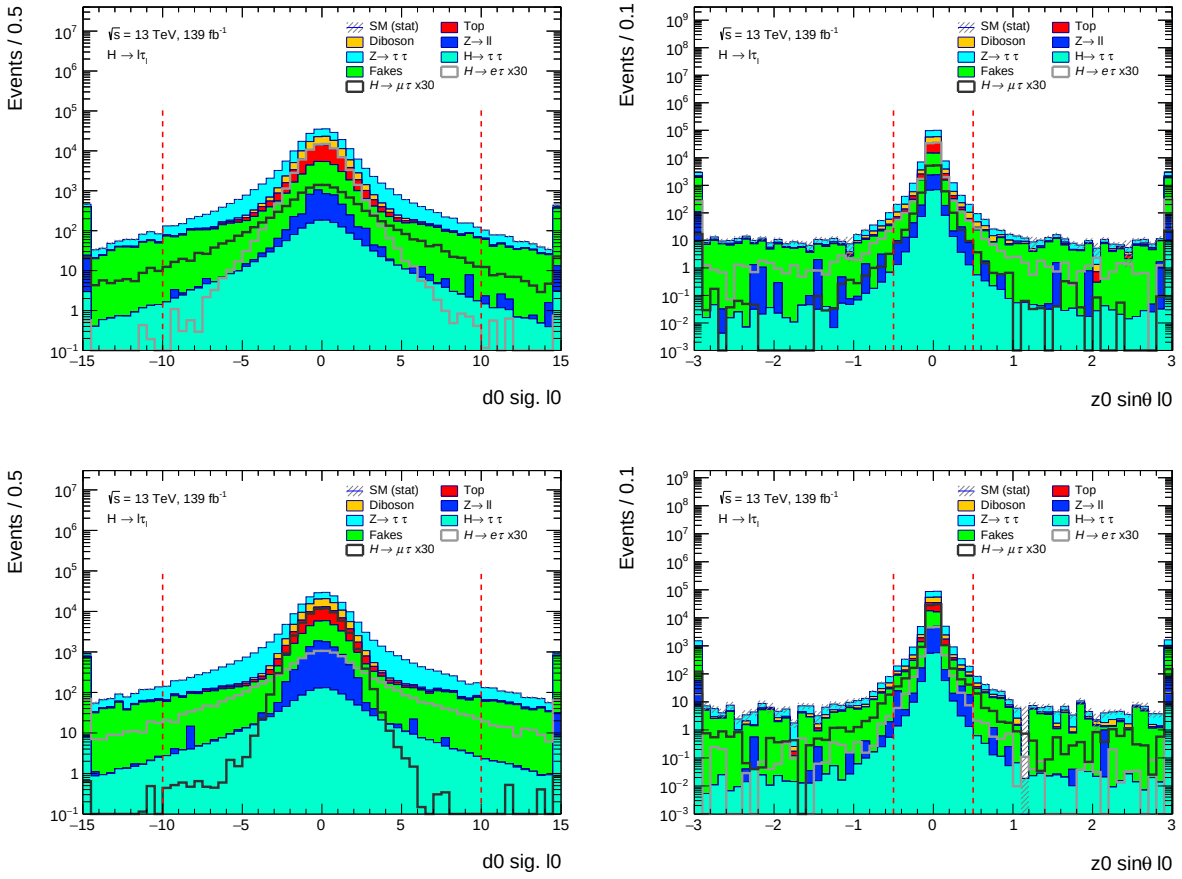


Figure 7.5: The significance of the transverse impact parameter $d_{0,\text{signif.}}$ (left) and the longitudinal impact parameter $z_0 \sin\theta$ (right) for the leading lepton in the $e\mu$ (top) and μe-channel (bottom), before the respective selection is applied. Events with $|d_{0,\text{signif.}}| \geq 10$ and $z_0 \sin\theta \geq 0.5\,\text{mm}$ are eliminated, indicated by the dashed lines. The LFV signal is scaled in the figure by a factor of 30.

Table 7.3: Expected event yield of signal and background processes in the $e\mu$-channel for specific selections. The background $H \to \tau\tau$ contains all Higgs boson processes discussed in 7.2.

Process	Selection 9	Selection 10	Selection 11	Selection 12	Selection 13
Signal					
$ggH \to \tau e$	2983.3 ± 5.2	2403.5 ± 4.7	2403.3 ± 4.7	2257.2 ± 4.5	2119.5 ± 4.4
$ggH \to \tau\mu$	359.9 ± 1.7	351.8 ± 1.7	351.6 ± 1.7	314.7 ± 1.6	309.2 ± 1.6
$VBF \to \tau e$	231.5 ± 0.5	185.0 ± 0.5	184.9 ± 0.5	163.3 ± 0.4	150.8 ± 0.4
$VBF \to \tau\mu$	61.5 ± 0.3	59.3 ± 0.3	59.3 ± 0.3	52.0 ± 0.2	50.4 ± 0.2
$WH \to \tau e$	108.0 ± 0.7	90.7 ± 0.6	81.4 ± 0.6	67.0 ± 0.5	63.4 ± 0.5
$WH \to \tau\mu$	32.7 ± 0.4	31.6 ± 0.4	27.5 ± 0.3	22.5 ± 0.3	22.0 ± 0.3
$ZH \to \tau e$	44.5 ± 0.3	35.9 ± 0.3	35.2 ± 0.3	25.8 ± 0.2	24.1 ± 0.2
$ZH \to \tau\mu$	12.5 ± 0.2	12.0 ± 0.2	11.5 ± 0.1	8.5 ± 0.1	8.3 ± 0.1
Background					
$Z \to \tau\tau$	142316 ± 213	103505 ± 169	101286 ± 159	93750 ± 152	88391 ± 149
diboson	75571 ± 97	69227 ± 93	50036 ± 80	46512 ± 78	45751 ± 77
top	706355 ± 242	646816 ± 232	455819 ± 194	44444 ± 70	43675 ± 69
Fakes	81402 ± 518	69479 ± 512	45432 ± 223	32365 ± 167	21995 ± 148
$Z \to \ell\ell$	13461 ± 392	4711 ± 238	4265 ± 226	3888 ± 221	2550 ± 175
$H \to \tau\tau$	2814 ± 4	2219 ± 3	2061 ± 3	1570 ± 3	1439 ± 3
Sig/Bkg	$3.75 \cdot 10^{-3}$	$3.54 \cdot 10^{-3}$	$4.79 \cdot 10^{-3}$	$1.31 \cdot 10^{-2}$	$1.35 \cdot 10^{-2}$

Table 7.4: Expected event yield of signal and background processes in the μe-channel for specific selections. Selection 10 does not change the event yield, since in the production electrons are required to have a transverse momentum greater than 15 GeV. The background $H \to \tau\tau$ contains all Higgs boson processes discussed in 7.2.

Process	Selection 9	Selection 11	Selection 12	Selection 13
Signal				
$ggH \to \tau e$	307.6 ± 1.6	307.6 ± 1.6	275.1 ± 1.5	265.3 ± 1.5
$ggH \to \tau\mu$	1965.8 ± 4.2	1965.4 ± 4.2	1847.1 ± 4.0	1798.9 ± 4.0
$VBF \to \tau e$	51.6 ± 0.2	51.5 ± 0.2	45.1 ± 0.2	42.5 ± 0.2
$VBF \to \tau\mu$	147.5 ± 0.4	147.4 ± 0.4	130.3 ± 0.4	125.4 ± 0.4
$WH \to \tau e$	28.5 ± 0.3	24.3 ± 0.3	19.5 ± 0.3	18.9 ± 0.3
$WH \to \tau\mu$	76.6 ± 0.6	67.6 ± 0.5	55.5 ± 0.5	54.3 ± 0.5
$ZH \to \tau e$	10.4 ± 0.1	10.1 ± 0.1	7.3 ± 0.1	7.0 ± 0.1
$ZH \to \tau\mu$	29.8 ± 0.2	29.1 ± 0.2	21.3 ± 0.2	20.8 ± 0.2
Background				
$Z \to \tau\tau$	83227 ± 154	81231 ± 144	75268 ± 139	72080 ± 137
diboson	60489 ± 92	42829 ± 75	39856 ± 74	39359 ± 73
top	563840 ± 214	389054 ± 177	37792 ± 63	37439 ± 63
Fakes	585701 ± 281	41137 ± 230	29649 ± 188	22727 ± 173
$Z \to \ell\ell$	9440 ± 354	8778 ± 342	8359 ± 338	7559 ± 320
$H \to \tau\tau$	1812 ± 3	1673 ± 3	1265 ± 3	1179 ± 3
Sig/Bkg	$3.37 \cdot 10^{-3}$	$4.61 \cdot 10^{-3}$	$1.25 \cdot 10^{-2}$	$1.29 \cdot 10^{-2}$

8 Estimation of Fake Lepton Background

The majority of the background to the LFV τe and $\tau\mu$ final states results from physical processes with a similar signature as final state, such as $Z \to \tau\tau \to e\mu 2\nu$ and $t\bar{t} \to e\mu 2\nu + 2b$. A smaller, yet significant part results from so called *fake leptons*. These are objects, for instance jets, which are misidentified in the reconstruction as leptons. Another contribution comes from non-prompt leptons. These are correctly reconstructed leptons, but do not originate from the original decaying particle, i.e. in case of the LFV search the Higgs boson. Leptons from the τ-lepton decay are also not counted as fake leptons. Non-prompt leptons mostly result from the decay of hadrons with a bottom or charm quark content. The sources and contributions of the fake leptons are different for electrons and muons. Thus, the estimation is performed separately for electrons and muons [94]. The notation *fake leptons* is used in the following for both fake leptons and non-prompt leptons.
Those fake leptons have a significant background contribution in the signal region (SR). Thus, it is important to estimate this background contribution well. Monte Carlo simulation does not model the fake background in a reliable way, hence a data driven estimate is necessary.
The method which will be used is the so-called *fake factor method*, which is described in section 8.1. Section 8.2 presents the Z+jets extraction region, which is used to determine the fake factor. This is followed by the estimation of a normalisation factor for the prompt WZ background in section 8.3. The determination of the fake factor with its kinematic dependencies is described in section 8.4, followed by a closure test with the resulting fake lepton estimation in section 8.5.

8.1 The Fake Factor Method

In the fake factor method, a transfer factor, also called *fake factor*, between two lepton identification requirements is calculated for the fake leptons. At first the lepton identification definitions are presented, followed by the fundamental assumptions of the fake factor method. This section closes with the scheme to calculate the fake factor.

8.1.1 ID and anti-ID Definitions

The fake factor method uses two orthogonal lepton identification requirements, called *ID* and *anti-ID*. The anti-ID region is enriched with fake leptons with only a small contribution of prompt leptons, whereas the ID region has a larger contribution of prompt leptons, but through the choice of phase space, is as well enriched in fake leptons compared to the SR.
The definition of the ID criteria is the same as in the SR. ID electrons require *Gradient* isolation and *medium* identification, whereas for the ID muons *FCTightTrackOnly_FixedRad* isolation and *medium* identification is demanded. For the anti-ID definition, one or both of these criteria are inverted, resulting in an orthogonal set of events. Since all muons are required to be *medium* identified on a more fundamental level, only the isolation requirement is inverted. For electrons, either the isolation or the identification requirements must fail, while still passing loose identification. Figure 8.1, shows the ID and anti-ID definitions for electrons and muons, respectively.

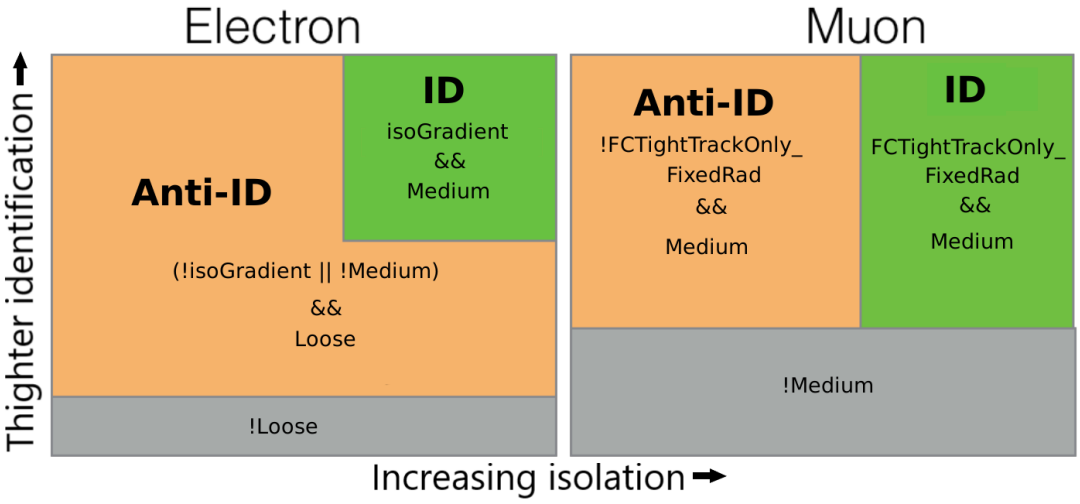


Figure 8.1: Schematic of the ID and anti-ID requirements used in the fake factor method for the estimation of the fake background contribution. The isolation criteria given as x-axis, the identification requirement as y-axis. Both are tightened for increasing axis values. The definitions are chosen such that there is no overlap between both regions. The grey area is displayed for completeness, but is not used in the fake factor method.

8.1.2 Description of the Fake Factor Method

For a given lepton flavour, it is assumed that the number of leptons satisfying the ID requirements is proportional to the number of leptons satisfying the anti-ID requirements, independent of other selection criteria. The fake factor f connects the two lepton identification definitions. With this, the leptonic fake background in the signal ID region can be estimated as

$$N_{\text{fake}}^{\text{SR,ID}} = f \cdot N_{\text{fake}}^{\text{SR,anti-ID}} \tag{8.1}$$

where the $N_{\text{fake}}^{\text{SR,anti-ID}}$ are only those events, which fulfil all SR requirements except the identification criteria and $N_{\text{fake}}^{\text{SR,ID}}$ which fulfil all SR requirements. Rewriting equation 8.1, the fake factor

$$f = N_{\text{fake}}^{\text{ID}} / N_{\text{fake}}^{\text{anti-ID}} \tag{8.2}$$

can be written as the ratio of fake events in the ID and the anti-ID region.
To calculate the fake factor, an extraction region (ER) is defined, which is kinematically close enough to the SR, but tailored to select the fake factor. It is assumed, that the ratio in equation 8.2 is the same in both the ER and the SR. With this assumption, the fake factor calculated in the extraction region can be applied to estimate the fake leptons in the signal ID region [95].

8.1.3 Fake Factor Calculation

The fake factor is obtained in the extraction region as

$$f = N_{\text{fake}}^{\text{ER,ID}} / N_{\text{fake}}^{\text{ER,anti-ID}} \tag{8.3}$$

This factor can then be used to estimate the leptonic fake background in the SR by using equation 8.1. However, the fake factor is not well described as a constant for all kinematic values. Leptons

in the ID and anti-ID region are distributed differently in specific kinematic observables. This means, the fake contribution in the ID definition can not be simply obtained by scaling the fake contribution in the anti-ID definition, but kinematic dependencies of these parameters must be taken into account as $f \to f(x, y, \dots)$. Equation 8.1 then becomes

$$N_{\text{fake}}^{\text{SR,ID}} = f(x, y, \dots) \cdot N_{\text{fake}}^{\text{SR,anti-ID}}(x, y, \dots) \tag{8.4}$$

where x and y are examples of kinematic dependencies.

8.2 Z+jets Extraction Region

To obtain the fake factor, a phase space region of high purity of fakes is desired, called *Z+jets extraction region*. The Z boson decays to two leptons which represent a very clear signature in the final state. The leptons are oppositely charged and of same-flavour and their invariant mass is close to the mass peak of the Z boson. Consequently, any additional reconstructed lepton candidate is likely to be a fake lepton. An additional reason is the large cross-section of the Z boson reducing the statistical uncertainty.

8.2.1 Selection

For the Z+jets extraction region, exactly three leptons are required, either electrons or muons. For all electrons, loose ID is required and for all muons medium ID. Two same-flavour (SF) leptons with opposite charge (OS) are Z-tagged, if their invariant mass is between 80 GeV and 100 GeV. If more than one pair of leptons fulfils this requirement, the pair closer to the Z peak of $m = 91.188\,\text{GeV}$ is chosen [35]. For the Z-tagged leptons, medium ID is required in addition. The missing transverse energy $\vec{E}_T^{miss}$, defined in equation 4.2, is required to be below 60 GeV. This reduces the contamination due to $WZ \to \ell\ell\ell\nu$ events. For all leptons, a requirement on the transverse momentum of $p_T \geq 15\,\text{GeV}$ is used. The third lepton, which is not Z-tagged, is called ℓ_{probe} in the following. The transverse mass

$$m_T(\ell_i) = \sqrt{2 \cdot |\vec{p}_T(\ell_i)| \cdot |\vec{E}_T^{miss}| \cdot \left(1 - \cos\Delta\phi\left(\vec{p}_T(\ell_i), \vec{E}_T^{miss}\right)\right)} \tag{8.5}$$

of ℓ_{probe} must satisfy $m_T(\ell_{\text{probe}}) < 40\,\text{GeV}$, which also rejects WZ events. For all three leptons, the same impact parameter requirements as in the SR are used as well as the same crack veto and b-jets veto, presented in section 7.3
For the Z+jets extraction region, an overview of the criteria is also given in table 8.1. In addition, events are then split depending on whether the third non Z-tagged lepton passes the ID or anti-ID requirement.
Events are only used, if at least one of the two Z-tagged leptons fired the trigger. This is done by an logical OR combination of single lepton triggers, as described in section 4.2.

8.2.2 Simulation of Prompt Background

Since this is a data driven estimation, contaminations by real prompt lepton processes in both the ID and anti-ID region must be estimated and subtracted before calculating and applying the fake factor. This is done by simulated events for the respective prompt background processes. Thus, the yields become

$$N_{\text{fake}}(x, y) = N_{\text{data}}(x, y) - N_{\text{MC,prompt}}(x, y) \tag{8.6}$$

Table 8.1: Selection for the Z+jets extraction region. The ID and anti-ID requirements are applied in addition to the probe lepton ℓ_{probe}.

Quantity	Requirement
Number of leptons (e, μ)	3
Z tagging of leptons	2 SF, OS leptons with 80 GeV$< m_{\ell\ell} <$ 100 GeV and medium ID
Missing transverse energy E_T^{miss}	< 60 GeV
Transverse momentum p_T	≥ 15 GeV
Transverse mass m_T (ℓ_{probe})	< 40 GeV

for both ID and anti-ID as well as extraction and signal region.

Only prompt background contributions with at least three real leptons in the Z+jets ER are considered. The background processes are $WZ \to \ell\ell\ell\nu$, $ZZ \to \ell\ell\ell\ell$ and inclusive WH decays where the Higgs boson decays into two tau leptons. The tau leptons decay further into electrons or muons and two neutrinos, hence resulting in a three lepton signature. These processes are estimated with MC simulation. The used simulated samples are listed in table 8.2.

Table 8.2: Simulated event samples for background processes in the Z+jets ER. Each sample is shown together with its respective cross-section.

Process	Sample name	Cross-section/pb
WZ	Sherpa_222_NNPDF30NNLO_lllv	4.5765
	Sherpa_222_NNPDF30NNLO_lllv_lowMllPtComplement	2.9708
ZZ	Sherpa_222_NNPDF30NNLO_llll	1.252
	Sherpa_222_NNPDF30NNLO_llll_lowMllPtComplement	1.4484
WH	PowhegPy8EG_NNPDF30_AZNLO_ WmH125J_Winc_MINLO_tautau	0.033417
	PowhegPy8EG_NNPDF30_AZNLO_ WpH125J_Winc_MINLO_tautau	0.052685

In table 8.3 the event yield for the prompt background processes as well as the data in the Z+jets extraction region is shown in case ℓ_{probe} is an electron or a muon. From the considered prompt background processes, WZ has the largest contribution.

Table 8.3: Event yield for prompt background processes and data in Z+jets ER in case ℓ_{probe} is an electron or a muon.

	Electron		Muon	
	ID	anti-ID	ID	anti-ID
data	7000 ± 84	17696 ± 133	3987 ± 63	5130 ± 72
WZ	1361 ± 9	167 ± 3	1523 ± 8	43 ± 1
ZZ	871 ± 6	192 ± 5	697 ± 8	138 ± 7
WH	0.35 ± 0.05	0.04 ± 0.02	0.40 ± 0.06	0.01 ± 0.01

8.3 Normalisation of WZ Background

The WZ process has the highest cross-section among the considered background processes and has also the largest contribution of all considered background processes in the Z+jets region. It was shown in previous measurements, that the NLO cross-section for WZ production did not predict correctly the observed event yield [96]. Hence, an additional normalisation of the WZ cross-section to data is performed.
For this, a specific region is used, which is orthogonal to the Z+jets extraction region.

8.3.1 WZ Normalisation Region

The WZ normalisation region must be orthogonal to the Z+jets ID region and enriched in WZ events. For this purpose, the two requirements E_T^{miss} and $m_T\,(\ell_{\text{probe}})$ are inverted. The kinematics of ℓ_{probe} ought to be similar of those in the Z+jets ER. This is accomplished by comparing the shape of the $p_T\,(\ell_{\text{probe}})$ distribution for several possible values for the E_T^{miss} and $m_T\,(\ell_{\text{probe}})$ requirements with the one of the Z+jets extraction region. The requirements, resulting in the most similar shape are chosen to be used for determining the WZ normalisation region.

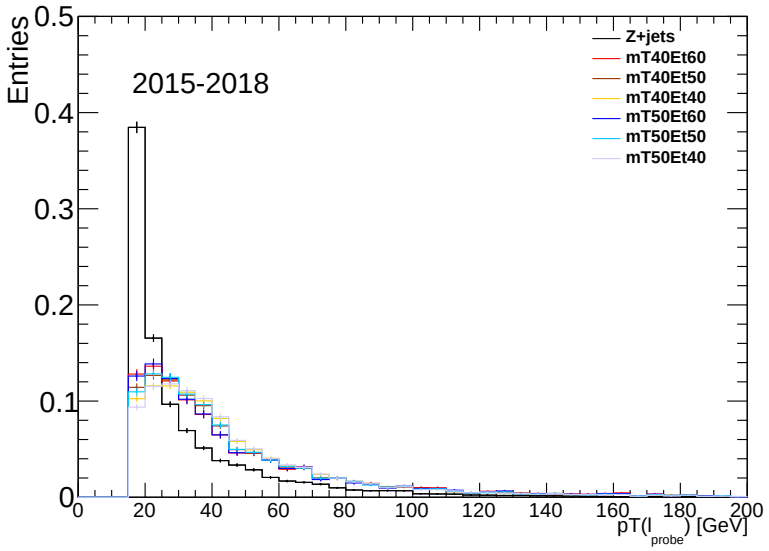


Figure 8.2: Normalised transverse momentum distributions of the probe lepton for different selection criteria for a possible WZ normalisation region as well as the Z+jets region. The label mTXEtY corresponds to the requirements $m_T\,(\ell_{\text{probe}}) \geq X$ GeV and $E_T^{\text{miss}} \geq Y$ GeV.

The different shapes are shown in figure 8.2. With either $m_T\,(\ell_{\text{probe}}) \geq 40\,\text{GeV}$ or $m_T\,(\ell_{\text{probe}}) \geq 50\,\text{GeV}$, the orthogonality with the Z+jets region is always ensured. Thus, the E_T^{miss} selection has no restriction due to the orthogonality requirement and can be also lowered below the SR value after inversion. Selections of $E_T^{miss} \geq 40\,\text{GeV}$, $E_T^{miss} \geq 50\,\text{GeV}$ and $E_T^{miss} \geq 60\,\text{GeV}$ have been investigated.
The $E_T^{miss} \geq 60\,\text{GeV}$ requirement is found to result in the closest approximation of the Z+jets ER shape for $p_T\,(\ell_{\text{probe}})$ compared to the others. Even higher E_T^{miss} requirements result in an even better agreement, but this would raise the statistical uncertainty further and is therefore not

considered. The requirement on the transverse mass $m_T(\ell_{\text{probe}})$ does not show any significant influence. A requirement of $m_T(\ell_{\text{probe}}) \geq 50\,\text{GeV}$ is used.
The WZ normalisation region is thus defined by the same selection requirements as in the Z+jets region, except $E_T^{miss} \geq 60\,\text{GeV}$ and $m_T(\ell_{\text{probe}}) \geq 50\,\text{GeV}$. The $p_T(\ell_{\text{probe}})$ distribution for predictions compared to data in the WZ normalisation region is shown in figure 8.3. The contamination from other process is below 6 %, demonstrating a very good purity of the WZ normalisation region.

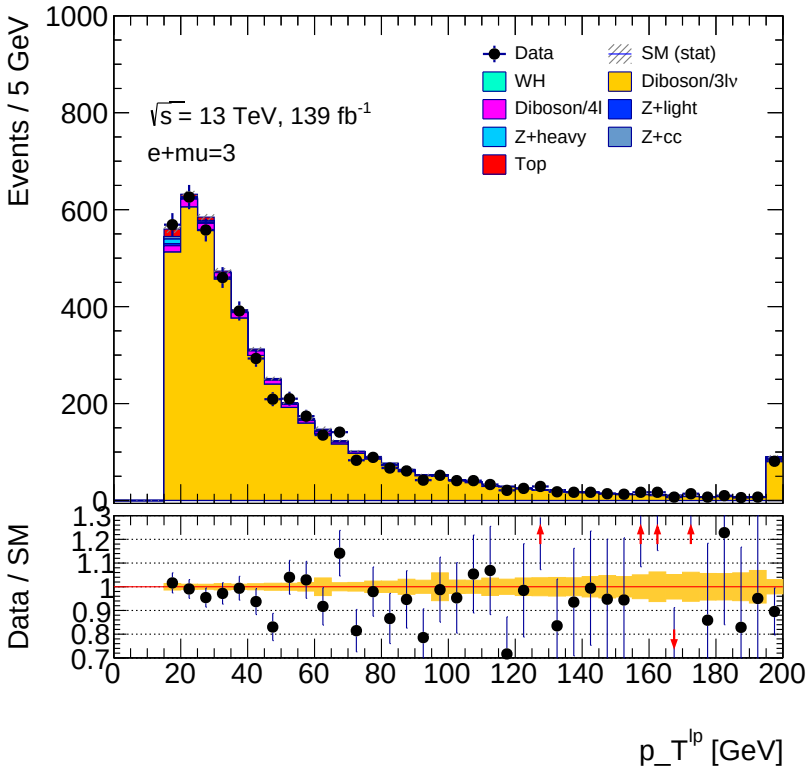


Figure 8.3: Distribution of $p_T(\ell_{\text{probe}})$ in the WZ normalisation region, for predictions of SM contributions compared to data. The lower panel shows the ratio of data over predictions. Processes with a Z boson and jets induced by a b-quark are combined in Z+heavy, whereas jets induced by a c-quark are noted as Z+cc. All other Z+jets processes are combined in Z+light. Diboson processes with three leptons and a neutrino in the final state are noted as Diboson/3lν, and the ones with four leptons in the final state as Diboson/4l.

The event yields of the simulated events are presented table 8.4 for different selection steps.

8.3.2 Estimation of the WZ Normalisation

The estimation of the WZ normalisation factor is done by comparing the transverse mass distribution $m_T(\ell_{\text{probe}})$ in simulation with data. The distribution is shown in figure 8.4. All processes except WZ are fixed as constant background in the following.
The simplest way to estimate the WZ normalisation factor is by calculating the total number of events in data minus the other background processes as well as the one predicted from WZ

Table 8.4: Event yield of the simulated processes in the WZ normalisation region for different selection steps. The Z-tagging and ID selections are the same as for the Z+jets extraction region. For E_T^{miss} a value greater 60 GeV is required and for the transverse mass a value greater than 50 GeV.

	No Cuts	Z-tagging	E_T^{miss}	$m_T(\ell_{\text{probe}})$	ID
Diboson/3lν	43419 ± 107	32416 ± 49	9281 ± 22	6429 ± 19	4476 ± 16
Diboson/4l	17633 ± 112	6646 ± 28	447 ± 6	211 ± 5	134 ± 2
Top	91408 ± 66	11244 ± 23	5079 ± 15	3022 ± 12	69 ± 2
Z+light	165227 ± 1655	102301 ± 1293	3437 ± 219	1032 ± 127	19 ± 4
Z+heavy	88215 ± 494	71128 ± 434	2336 ± 60	400 ± 23	17 ± 5
Z+cc	42438 ± 480	31074 ± 406	1112 ± 63	234 ± 31	12 ± 5
WH	33.8 ± 0.6	3.7 ± 0.2	1.0 ± 0.1	0.5 ± 0.1	0.3 ± 0.1

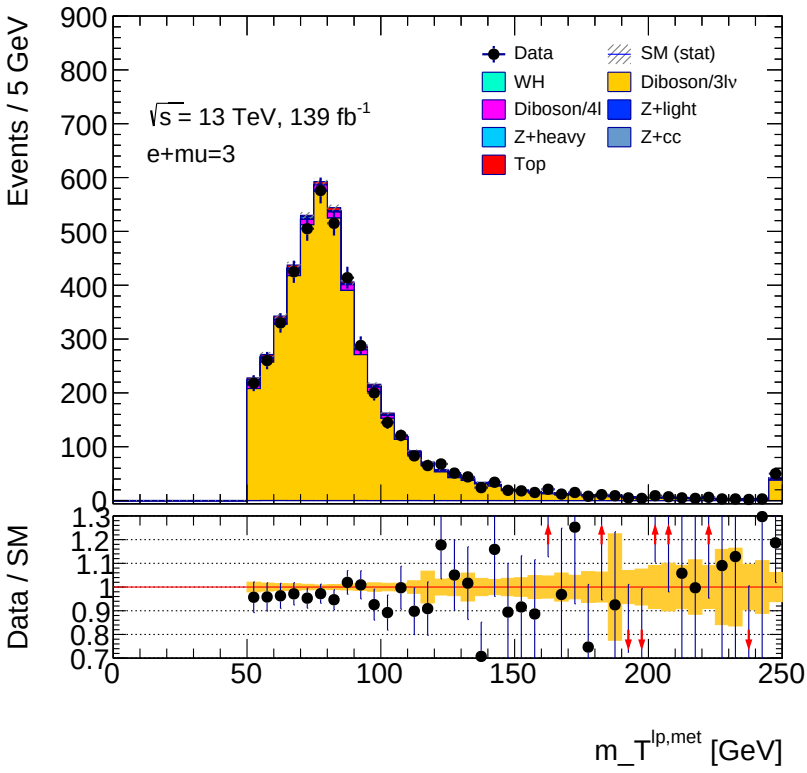


Figure 8.4: Distribution of $m_T\,(\ell_{\text{probe}})$ in the WZ normalisation region for data compared to the sum of predictions from simulations. The lower panel shows the ratio of data over predictions. The legend is described in figure 8.3.

simulations. The normalisation factor is then derived as

$$\alpha_{\text{yield}} = \frac{N_{\text{data}} - N_{\text{bkg}}}{N_{\text{WZ}}} \tag{8.7}$$

with N_{data} and N_{WZ} being the respective yields of data and WZ and N_{bkg} the yield of the other background processes.
A more sophisticated way for the estimation of the WZ normalisation factor is a χ^2-fit of the

$m_T(\ell_{\text{probe}})$ distribution to data. For this purpose, the area around the peak in the $m_T(\ell_{\text{probe}})$ distribution is considered, namely $50\,\text{GeV} \leq m_T(\ell_{\text{probe}}) \leq 150\,\text{GeV}$. The chosen fit range mitigates the influence of the high $m_T(\ell_{\text{probe}})$-tail on the fit, which has minimal impact on the total number of events.
A binned extended maximum likelihood fit is performed. The fit returns an adjusted WZ yield with its respective uncertainty. By dividing the new yield $N_{WZ}^{\text{post-fit}}$ with the one prior to the fit, $N_{WZ}^{\text{pre-fit}}$, the normalisation factor

$$\alpha_{\text{fit}} = \frac{N_{WZ}^{\text{post-fit}}}{N_{WZ}^{\text{pre-fit}}} \tag{8.8}$$

is obtained. The statistical uncertainty on the normalisation factor is given as (fit error) $/N_{WZ}^{\text{pre-fit}}$. The WZ normalisation factors, investigated separately for 2015+2016 (c16a), 2017 (c16d) and 2018 (c16e) are listed in table 8.5. The results for the different campaigns are consistent within the statistical uncertainties, justifying the combination of all.

Table 8.5: Obtained WZ normalisation factors separately for the three data-taking campaigns: 2015+2016 (c16a), 2017 (c16d) and 2018 (c16e). Both, the yield and fit estimation are shown.

Campaign	α_{fit}	α_{yield}
c16a	0.968 ± 0.029	1.008 ± 0.032
c16d	0.909 ± 0.026	0.948 ± 0.029
c16e	0.926 ± 0.023	0.957 ± 0.026

For the 2015-2018 combined estimation, the compatibility of the four lepton channels is investigated. The channels are defined by the Z-tagged lepton pair and ℓ_{probe}. Since only electrons and muons are present in the final state, the four channels are eee, $ee\mu$, $\mu\mu e$ and $\mu\mu\mu$. The results for the different channels are shown in table 8.6. All channels except the $\mu\mu\mu$ channel agree within one standard deviation and in case of the $\mu\mu\mu$ channel an agreement within two standard deviations is observed.

Table 8.6: Obtained WZ normalisation factors for the four lepton channels eee, $ee\mu$, $\mu\mu e$ and $\mu\mu\mu$ for data from 2015-2018. Both, the yield and fit estimation are shown.

Channel	α_{fit}	α_{yield}
eee	0.968 ± 0.032	1.009 ± 0.035
$\mu\mu e$	0.919 ± 0.030	0.954 ± 0.034
$\mu\mu\mu$	0.873 ± 0.028	0.919 ± 0.031
$ee\mu$	0.975 ± 0.030	0.997 ± 0.032

The results from table 8.5 and 8.6 justify a combination of all three data-taking campaigns and four lepton channels to one common normalisation factor. A systematic uncertainty due to the fit range of α_{fit} is considered, by deriving a normalisation factor $\alpha_{\text{fit,full}}$ in the fit range $50-250\,\text{GeV}$. This results in a larger normalisation factor and the difference $\alpha_{\text{fit,full}} - \alpha_{\text{fit}}$ is then assigned as systematic uncertainty on α_{fit}.
The final result for the yield method is given by

$$\alpha_{\text{yield}} = 0.968 \pm 0.016$$

and

$$\alpha_{\text{fit}} = 0.932 \pm 0.015(\text{stat}) \pm 0.036(\text{syst}) \tag{8.9}$$

for the χ^2-fit. Both scaling factors agree within the uncertainties. For the estimation of the fake factor, α_{fit} is used. The statistical and systematic uncertainty are treated as uncorrelated, which results in a combined uncertainty of the normalisation factor of 0.039.

8.4 Fake Factor Estimation

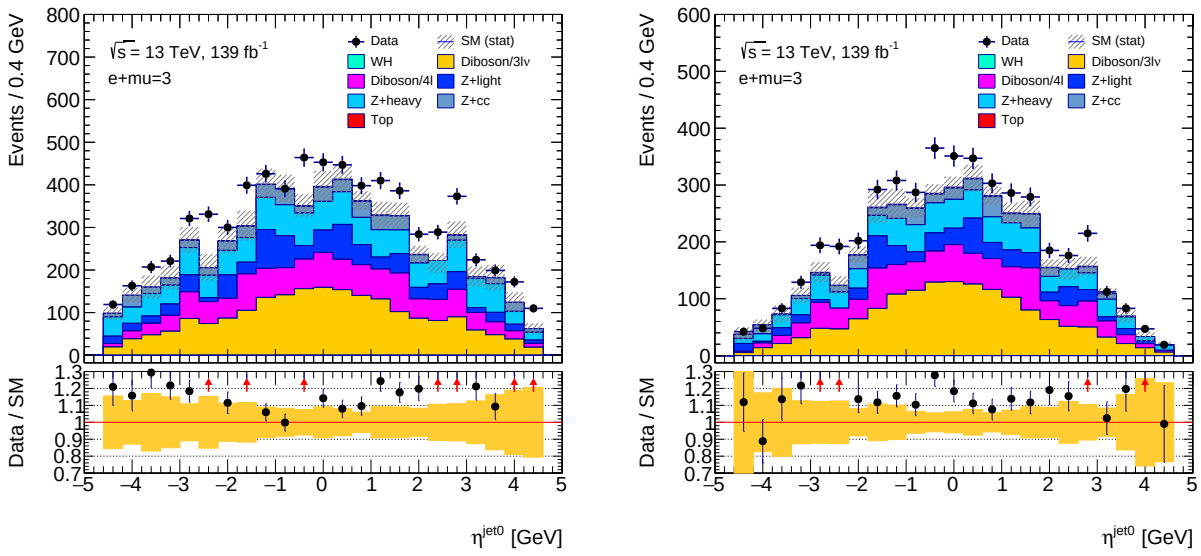
The fake factor is derived using equation 8.3 in the Z+jets extraction region. The contamination from real prompt lepton processes is subtracted from the data using MC simulations. These processes are WZ, ZZ and WH inclusive decays. The WZ contribution is scaled by a normalisation factor α_{fit}, stated in equation 8.9. In addition, truth matching is applied to the MC samples. This means, that events are only used, if no mismatching in the reconstruction algorithm is present. Hence, only reconstructed electrons and muons are used, which are no fake leptons, but in fact electrons and muons. With this, only simulated events are subtracted, when the reconstructed leptons are real prompt leptons, originating directly from the respective signal particle.
The obtained data set is split into an electron and a muon set, depending on the flavour of ℓ_{probe}. Both sets are treated separately resulting in two fake factor estimations. Further, kinematic dependencies of the fake factor are taken into account as stated in equation 8.4.

8.4.1 Kinematic Dependencies of the Fake Factor

Kinematic dependencies of the fake factor are taken into account as a function of p_T and η of the probe lepton. A binning of [15, 20, 25, 35, 1000] GeV in p_T is chosen. For the pseudorapidity, only the absolute value is considered. As a start, for both electrons and muons, two bins [0, 1.37] and [1.52, 2.47] are used. The region between 1.37 and 1.52, covers the crack region and is thus empty.
An additional observable that is investigated is the ΔR between the nearest jet and ℓ_{probe}. The nearest jet is defined as the jet with the smallest $\Delta R = \sqrt{\Delta\phi^2 + \Delta\eta^2}$ to the probe lepton. Due to the overlap removal, no events with $\Delta R < 0.4$ are present.
Figure 8.5 shows the η distribution of the leading jet with the lepton ID selection in the Z+jets extraction region, for (a) jets with $p_T \geq 20\,\text{GeV}$ and (b) jets with $p_T \geq 30\,\text{GeV}$. The modelling in the $p_T \geq 30\,\text{GeV}$ shows improvements for high pseudorapidities, which are difficult to simulate with MC. Thus, only jets are considered which lie above this threshold.
For all events without any jet fulfilling the requirement of $p_T^{\text{jet}} \geq 30\,\text{GeV}$, a separate fake factor is estimated. This corresponds to an implicit binning in $N_{\text{jets}} = 0$ and $N_{\text{jets}} \geq 1$.

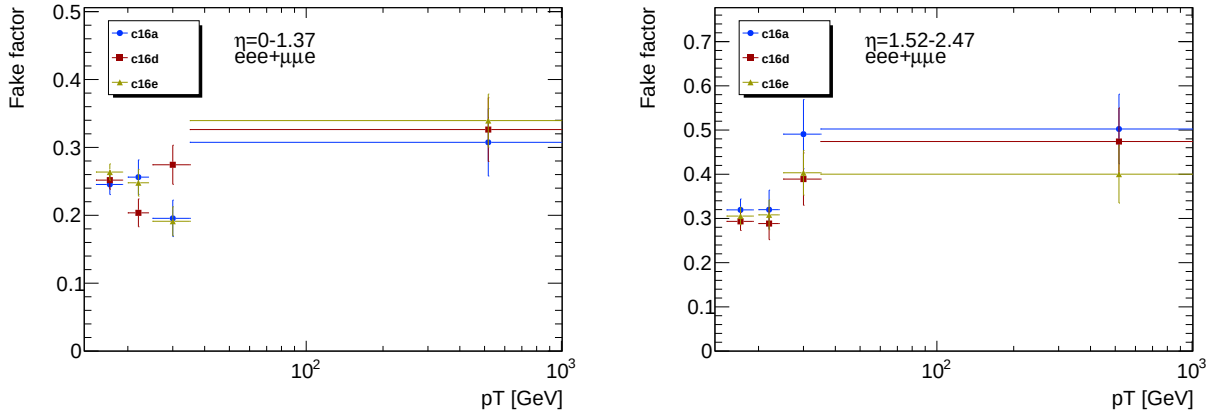
8.4.2 Determination of the Fake Factor

The fake factor is calculated according to equation 8.2, binned in the probe lepton p_T, η and $\Delta R(\text{nearest jet}, \ell_{\text{probe}})$. Figures 8.6 and 8.7 show the results for all campaigns separately, both for the electron and the muon. The obtained values agree reasonably well among the three campaigns with each other, only in the central p_T range for the first electron η bin small deviations are observed. Justified by the reasonable agreement, a common fake factor estimation is performed in the following for the three campaigns.
The resulting fake factors for both electrons and muons are shown in figure 8.8. For the electron, clear differences in the fake factor for the two η bins are visible. This difference is not visible



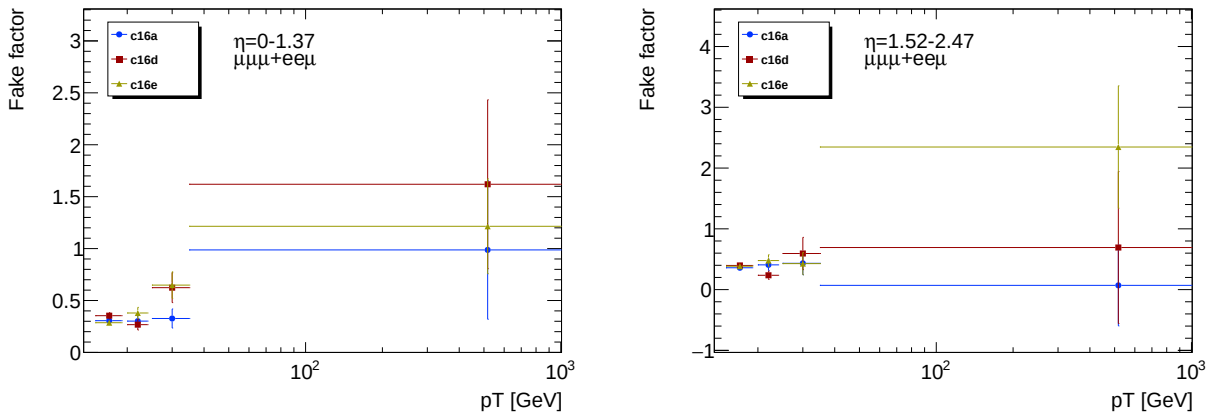
(a) $p_T^{\text{jet}} \geq 20\,\text{GeV}$ (b) $p_T^{\text{jet}} \geq 30\,\text{GeV}$

Figure 8.5: Leading jet η distribution for different jet p_T thresholds of 20 GeV (a) and 30 GeV (b). No truth matching and no WZ scaling is applied here. The legend is described in figure 8.3.



(a) $|\eta| < 1.37$ (b) $|\eta| > 1.52$

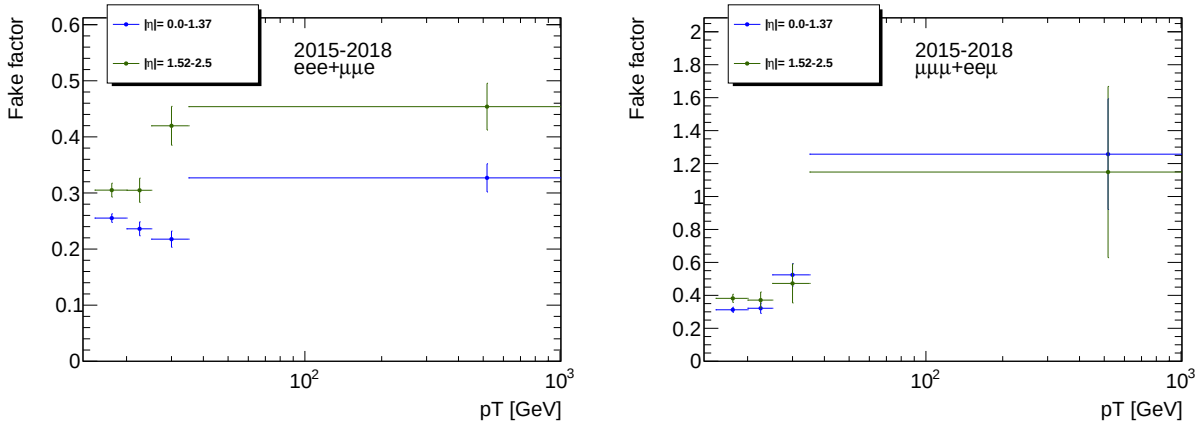
Figure 8.6: Electron fake factors for three all data-taking campaigns as a function of p_T for $|\eta| < 1.37$ (a) and $|\eta| > 1.52$ (b).



(a) $|\eta| < 1.37$ (b) $|\eta| > 1.52$

Figure 8.7: Muon fake factors for all three data-taking campaigns as a function of p_T for $|\eta| < 1.37$ (a) and $|\eta| > 1.52$ (b).

for the muon. Therefore, the two $|\eta|$ bins are merged for muons. The combined fake factors for muons are shown in figure 8.9.



(a) Electron (b) Muon

Figure 8.8: Fake factor binned in $p_T\,(\ell_{\text{probe}})$ and $\eta\,(\ell_{\text{probe}})$ for the electron (a) and for the muon (b).

The extracted fake factors for electrons and muons are presented in table 8.8 and 8.9, with statistical and systematic uncertainties included.
The systematic uncertainties on the fake factors result from the background subtraction and the following sources are considered: theoretical uncertainties on ZZ and WH as well as the normalisation uncertainty on WZ. In table 8.7, the theoretical uncertainties on ZZ and WH are presented. For both, scale uncertainties as well as PDF uncertainties are taken into account. In case of the ZZ production, the uncertainty is calculated at next-to-leading order in perturbative QCD. The uncertainty from the choice of PDF is derived as 2 % and the scale uncertainty as $4 - 5$ %. This results in a theoretical uncertainty of 6 % [97].
The WH uncertainty is also calculated at next-to-leading order in perturbative QCD. The W^+H

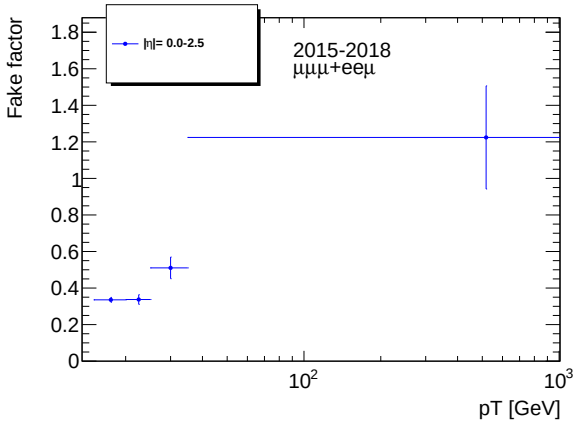


Figure 8.9: Muon fake factor combined for the entire $|\eta|$ range as function of p_T.

Table 8.7: Theoretical uncertainties of ZZ and WH [42, 97].

	ZZ	W^+H	W^-H	$W^\pm H$
Theoretical uncertainty	$+6\,\%$	$+1.9\,\%$	$+2.0\,\%$	$+2.2\,\%$
	$-6\,\%$	$-1.9\,\%$	$-2.1\,\%$	$-2.2\,\%$

scale uncertainty at $\sqrt{s} = 13\,\text{TeV}$ is given as $^{+0.5}_{-0.7}\,\%$. With the sum in quadrature of PDF and α_s uncertainties of $\pm 1.8\,\%$, this results in a combined theory uncertainty of $\pm 1.9\,\%$. For the W^-H process, the scale uncertainty is estimated as $^{+0.4}_{-0.7}\,\%$ and results in a combined uncertainty of $\pm 2.0\,\%$. A combined theoretical uncertainty on W^+H and W^-H is used as $\pm 2.2\,\%$ which covers both individual uncertainties [42].
Due to the WZ normalisation, discussed in section 8.3, no theoretical uncertainties are applied on WZ. Instead, the uncertainty of α_{fit} is propagated through to the fake factor estimation.
For all considered systematic sources, the uncertainty is propagated to the fake factor uncertainty by varying the corresponding process normalisation by the respective uncertainty up and down. This results in a new fake factor, denoted as $f_{i,\text{syst}}$. The fake factor without variation is denoted as f_{nom}.
The difference $\Delta f_i = f_{i,\text{syst}} - f_{\text{nom}}$ indicates the systematic uncertainty in the fake factor. The differences Δf_i are very similar in their absolute value for both up and down variation. Hence, the uncertainty is symmetrized by averaging the absolute values of the respective up and down variations. The resulting systematic uncertainty on the fake factor due to WH is suppressed by the ones of ZZ and WZ. In the low p_T range, both dominant systematic uncertainties are of equal size, whereas for the highest p_T range, Δf_{WZ} becomes about twice as large as Δf_{ZZ}.

Table 8.8: Muon fake factor and systematic uncertainties.

$p_T\,(\ell_{\text{probe}})$ [GeV]	Fake factor	Stat. unc.	Sys. unc.	Combined
15-20	0.335	0.013	0.004	0.014
20-25	0.337	0.026	0.009	0.028
25-35	0.510	0.059	0.028	0.065
35-1000	1.224	0.282	0.258	0.382

The dependence of the fake factor on $\Delta R\,(\text{nearest jet}, \ell_{\text{probe}})$ is shown in figure 8.10. For this, the fake factor is integrated over $p_T\,(\ell_{\text{probe}})$ and $\eta\,(\ell_{\text{probe}})$. The events with $N_{\text{jets}} = 0$ are shown

Table 8.9: Electron fake factor for $|\eta| < 1.37$ (a) and $|\eta| > 1.52$ (b) and systematic uncertainties.

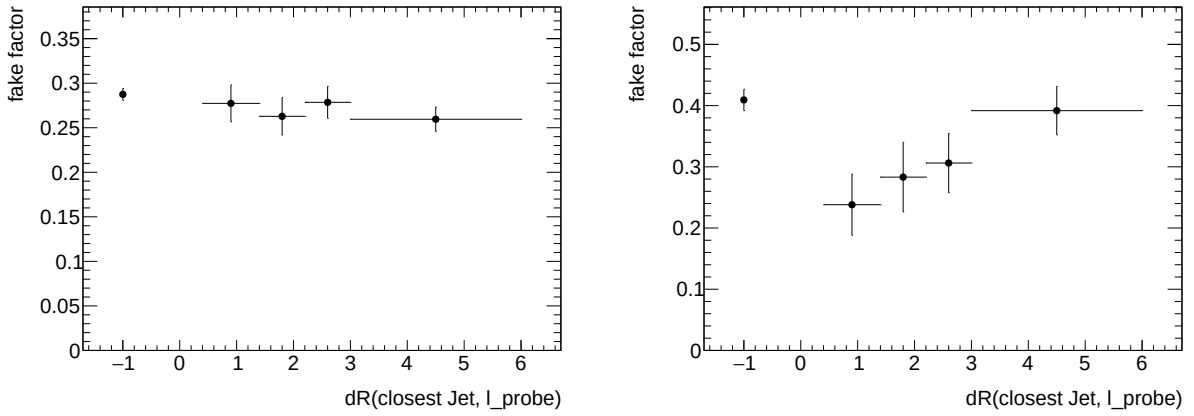
(a) $|\eta| < 1.37$

$p_T\,(\ell_{\text{probe}})$ [GeV]	Fake factor	Stat. unc.	Sys. unc.	Combined
15-20	0.255	0.008	0.001	0.008
20-25	0.236	0.012	0.003	0.012
25-35	0.218	0.014	0.005	0.015
35-1000	0.327	0.025	0.017	0.030

(b) $|\eta| > 1.52$

$p_T\,(\ell_{\text{probe}})$ [GeV]	Fake factor	Stat. unc.	Sys. unc.	Combined
15-20	0.305	0.012	0.001	0.012
20-25	0.305	0.022	0.003	0.022
25-35	0.420	0.035	0.007	0.036
35-1000	0.454	0.042	0.017	0.045

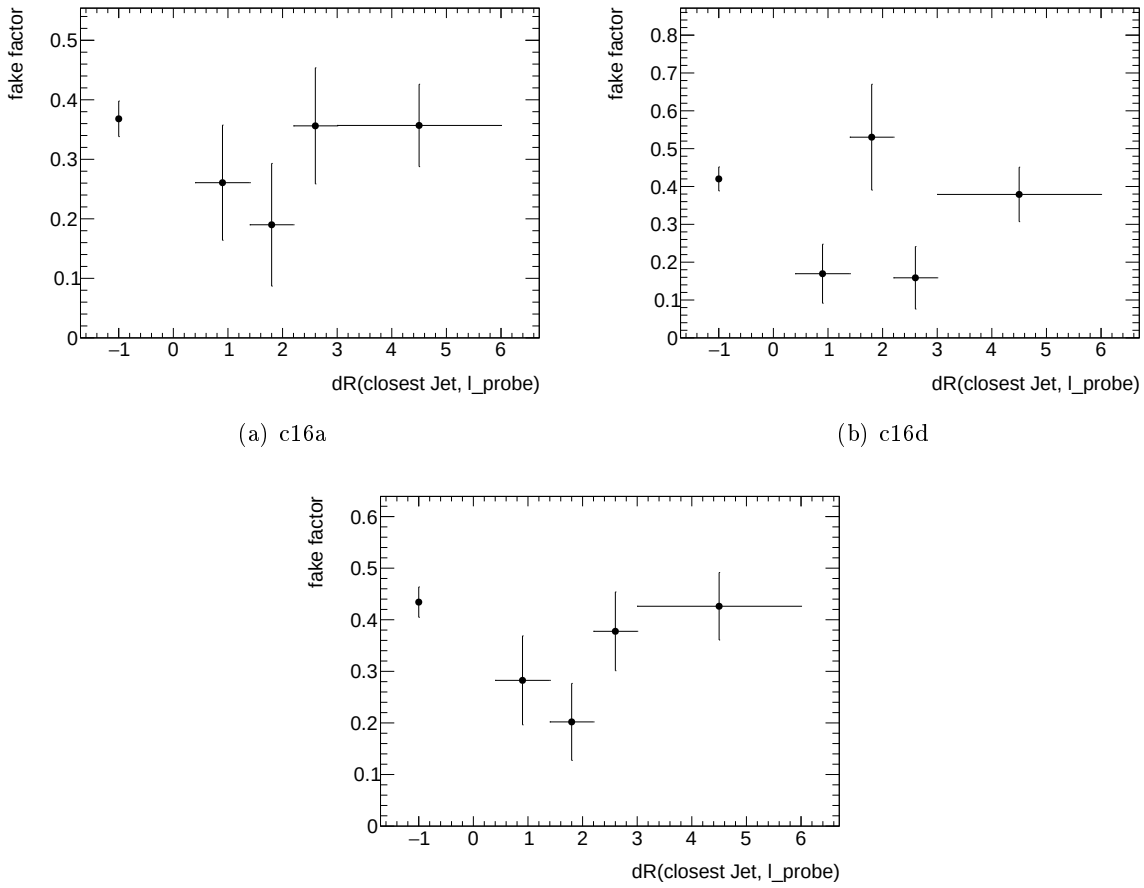
in the bin below 0.



(a) Electron (b) Muon

Figure 8.10: The fake factor binned in $\Delta R\,(\text{nearest jet}, \ell_{\text{probe}})$. The values obtained for the electron are shown in figure a) and those for the muon in b).

In the case of the electron, no significant dependence on $\Delta R\,(\text{nearest jet}, \ell_{\text{probe}})$ is visible. For the muon a difference between $N_{\text{jets}} = 0$ and the first $\Delta R\,(\text{nearest jet}, \ell_{\text{probe}})$ bin is observed. To investigate this further all three data-taking campaigns are split up again. This is shown in figure 8.11.
In the campaign c16a, no dependence of the muon fake factor on ΔR is observed within the uncertainties and in c16d, strong statistical fluctuations are visible without a clear trend. For the campaign c16e, the trend similar to the combined fake factor for the three data-taking campaigns is observed. Since two out of the three campaigns show no significant dependency of the fake factor on ΔR and the campaign c16d is behaving very different compared to the others, no additional ΔR binning is applied.
Additional fake factor dependencies are investigated. The number of jets N_{jets} has a possible influence. An implicit binning in the context of the ΔR dependency was already performed. The



(a) c16a

(b) c16d

(c) c16e

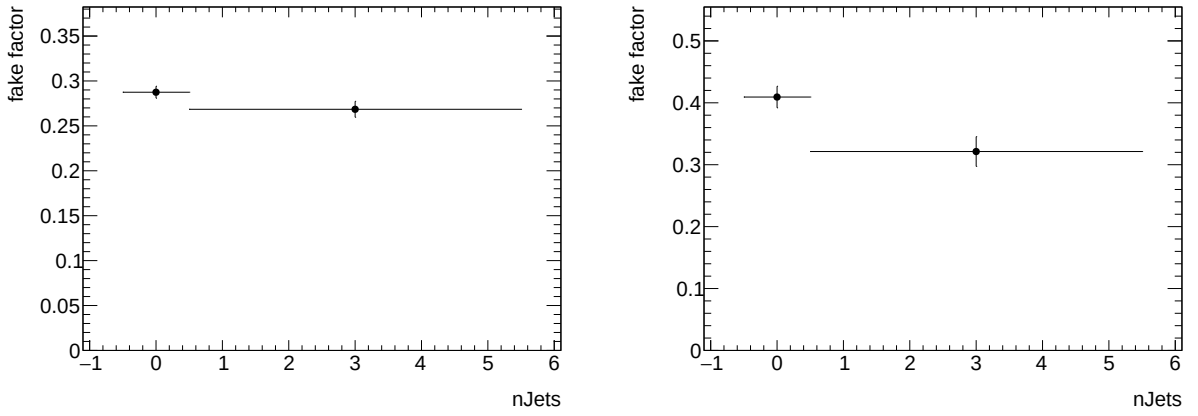
Figure 8.11: Muon fake factor binned in $\Delta R\,(\text{nearest jet}, \ell_{\text{probe}})$, separately for the three data-taking campaigns c16a (a), c16d (b) and c16e (c).

explicit dependency of the fake factor on N_{jets} is shown in figure 8.12. It is divided in $N_{\text{jets}} = 0$ and $N_{\text{jets}} > 0$. Jets are required to have a transverse momentum of $p_T \geq 30\,\text{GeV}$. The muon fake factor shows a slight dependency on N_{jets}. Thus, taking this additional dependency into account, improvements especially for the muon could be achieved.

Further, a dependency of the fake factor on the significance of the transversal impact parameter $|\text{d}_{0,\text{signif.}}|$ is investigated. A binning of $[0, 0.5, 1, 2, 3, 5, 10]$ is chosen integrating over the complete p_T and η range. The resulting fake factor dependency is displayed in figure 8.13. In the case of an electron fake, a dependency is visible. A $|\text{d}_{0,\text{signif.}}|$ binning with $[0, 3, 10]$ could be considered.

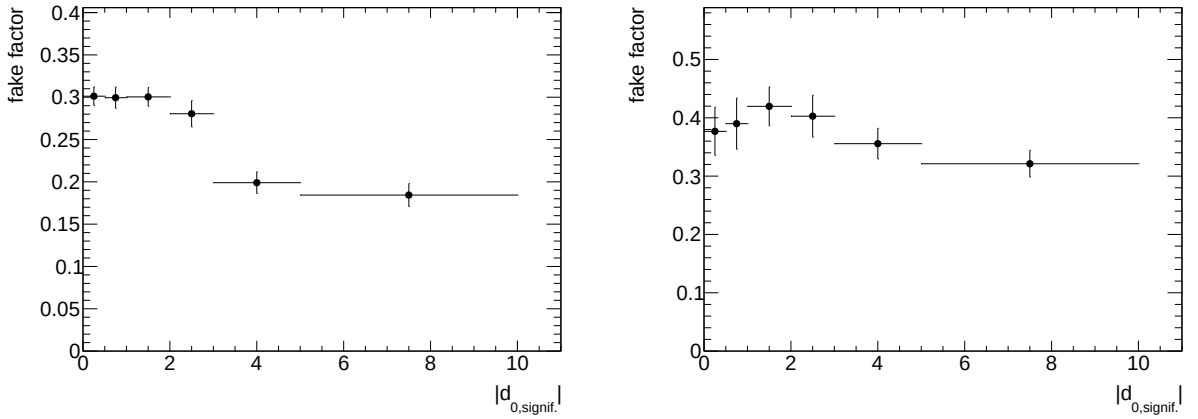
Lastly, the dependency of the longitudinal impact parameter $|\text{z}_0 \sin\theta|$ on the fake factor is investigated, binned in $[0, 0.025, 0.05, 0.075, 0.1, 0.2, 0.5]$ mm. The dependency is very similar to the one of the transversal impact parameter and a binning of the fake factor in this parameter may improve the fake estimation.

However, dependencies of the fake factor on $|\text{d}_{0,\text{signif.}}|$, $|\text{z}_0 \sin\theta|$ or N_{jets} are not take into account in the following closure test.



(a) Electron (b) Muon

Figure 8.12: Fake factor binned in $N_{\text{jets}} = 0$ and $N_{\text{jets}} > 0$ for the electron (a) and the muon (b). Jets are required to have an $p_T \geq 30\,\text{GeV}$.



(a) Electron (b) Muon

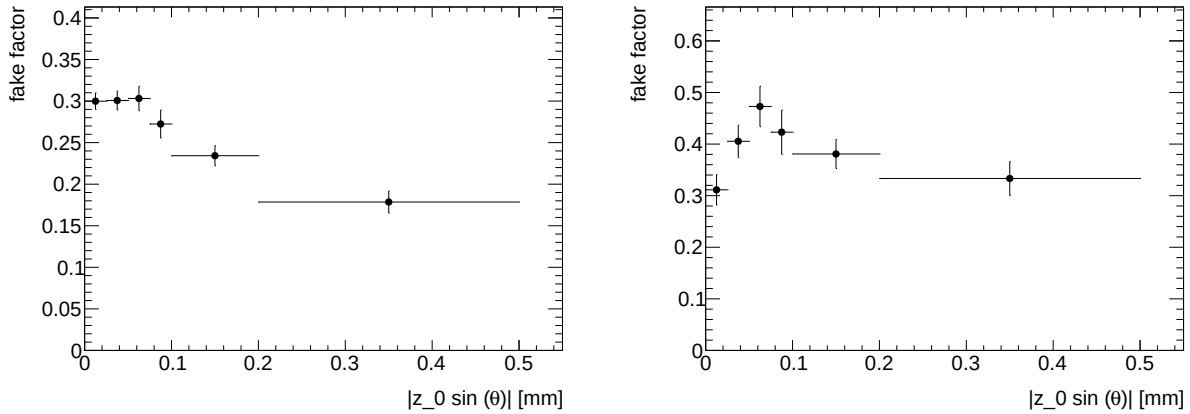
Figure 8.13: Fake factor binned in the significance of the transverse impact parameter $|d_{0,\text{signif.}}|$ for the electron (a) and the muon (b).

8.5 Closure Test for the Fake Facotrs

The fake factors obtained in the previous section are applied as closure test to the anti-ID selection in the Z+jets extraction region and compared to the ID selection.

8.5.1 Procedure for the Closure Test

The closure test is performed to verify a correct estimation and implementation of the fake factors and search for possible mismodelling or further kinematic dependencies of the fake factor.
Fakes in the Z+jets ID region are estimated with equation 8.4. Prompt background processes are subtracted in the anti-ID region, with the WZ background being scaled by α_{fit}. As before, truth matching for prompt leptons is applied to simulated events. In the ID region, the total yields are given as the sum of the estimated fakes and the prompt background. The WZ background is



(a) Electron (b) Muon

Figure 8.14: Fake factor binned in the longitudinal impact parameter $|z_0 \sin\theta|$ for the electron (a) and the muon (b).

again scaled by α_{fit}. The prediction in the ID Z+jets extraction region is then compared to data. For the variables p_T and η, which are explicit dependencies of the used fake factor, a ratio between data and fake estimation close to unity is expected. This is shown and validated in figure 8.15. In both, the muon p_T and the electron η, slight deviations are visible, however within statistical uncertainties the background estimation is in good agreement with the data.
The uncertainty on the data is given as $\Delta N_{\text{data}} = \sqrt{N_{\text{data}}}$. For the background, statistical and systematic uncertainties are considered. The statistical uncertainty in the background $\Delta N^2_{\text{stat,total bkg}}$ is given by the sum of squares of event weights in the simulation.
Each systematic uncertainty of the fake factor Δf_i is used for an up and down variation of the nominal fake factor. This results in a new fake estimation. The difference between the varied yields and the nominal yield $\Delta N_{i,\text{syst}} = N_{i,\text{syst}} - N_{\text{nom}}$ is the corresponding systematic uncertainty. These are then combined to one systematic uncertainty

$$\Delta N_{\text{syst}} = \sqrt{\Delta N^2_{\text{WZ,syst}} + \Delta N^2_{\text{ZZ,syst}} + \Delta N^2_{\text{WH,syst}}}$$

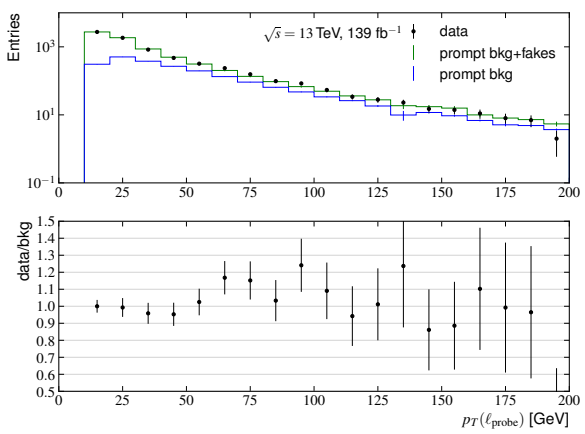
by the sum in quadratures. By analogy with this variation, ΔN_{FF} includes the uncertainty induced by the statistical uncertainties of the fake factor and $\Delta N_{\text{syst,prompt}}$ the combined systematic uncertainty on the simulated prompt processes by deriving the sum of quadratures of the individual uncertainties. With this, the total background uncertainty is given as

$$\Delta N_{\text{total bkg}} = \sqrt{\Delta N^2_{\text{stat, total bkg}} + \Delta N^2_{\text{syst,prompt}} + \Delta N^2_{\text{FF}} + \Delta N^2_{\text{syst}}}$$

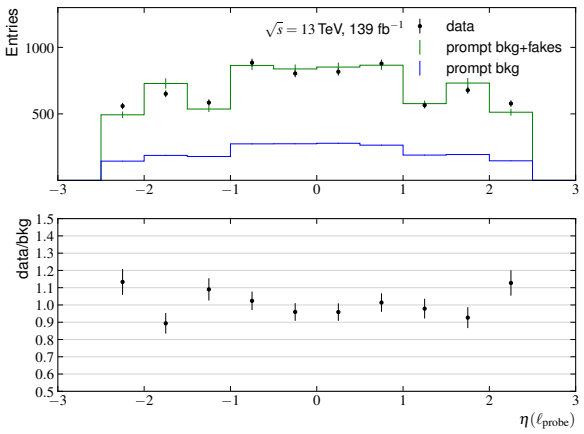
The uncertainty on the ratio of data and estimated background is obtained by the sum in quadrature of relative uncertainties

$$\Delta\left(\frac{N_{\text{data}}}{N_{\text{total bkg}}}\right) = \frac{N_{\text{data}}}{N_{\text{total bkg}}} \times \sqrt{\left(\frac{\Delta N_{\text{data}}}{N_{\text{data}}}\right)^2 + \left(\frac{\Delta N_{\text{total bkg}}}{N_{\text{total bkg}}}\right)^2}$$

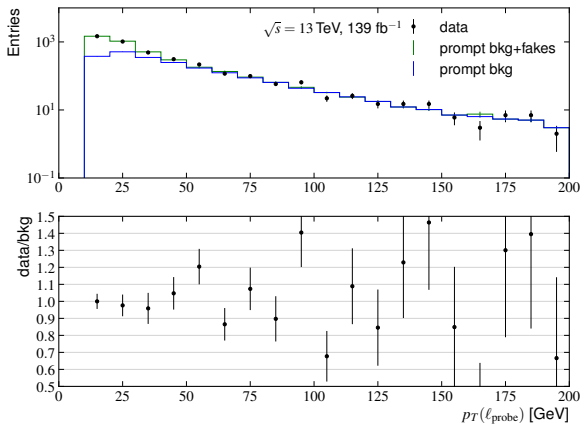
displayed as error bars in the ratio plots.
In table 8.10, the observed data yield is compared to the estimated background yield. The latter is split in simulated prompt processes and the estimated fake contribution.



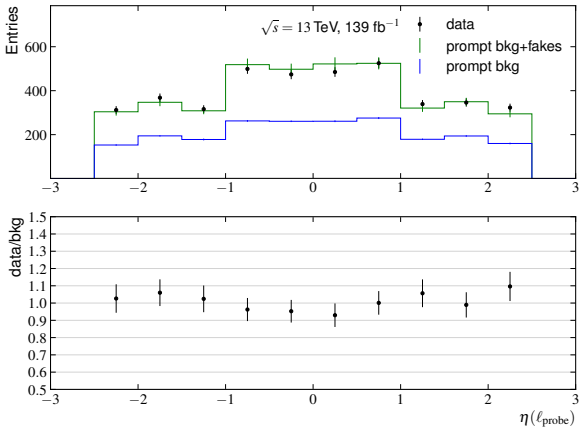
(a) Electron p_T



(b) Electron η



(c) Muon p_T



(d) Muon η

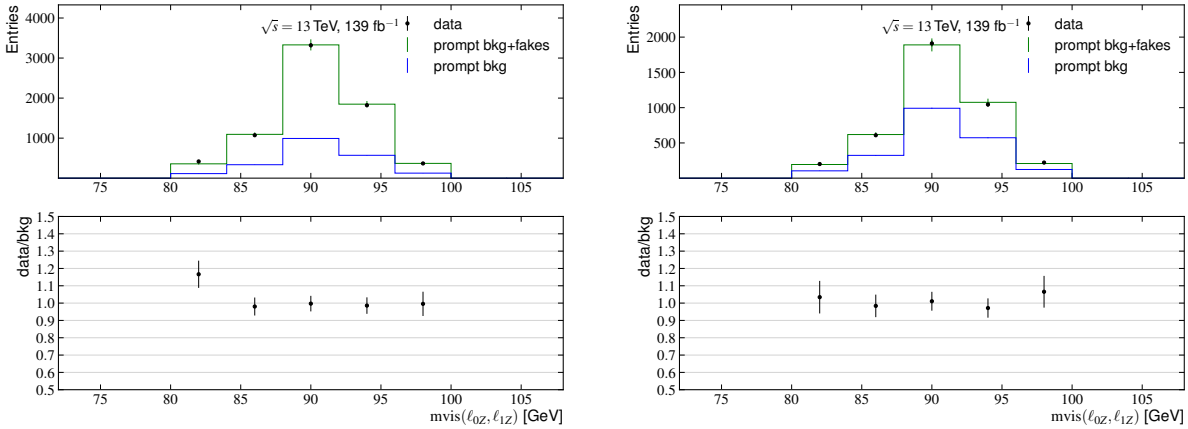
Figure 8.15: Closure test for the fake factor determination in the Z+jets ID region as function of p_T (left) and η (right) for electrons (top) and muons (bottom). The data is compared to the combined estimate of the prompt lepton backgrounds and the fake estimate through the fake factor method. Contributions from the prompt lepton backgrounds and the fake estimate are separately shown. The lower panel shows the ratio between data and the estimated background. For a discussion of the displayed uncertainties see the text.

Table 8.10: Data and estimated background yield for electron and muon fakes. The statistical and systematic uncertainties of the prompt background and the fake estimation are included. The statistical uncertainty of the fake factor estimation dominates all others.

	Data	Background	Prompt background	Fake background
Electron fake	7000 ± 83.7	7000 ± 287.3	2139.8 ± 105.8	4860.2 ± 267.1
Muon fake	3987 ± 63.1	3987 ± 188.4	2117.2 ± 101.9	1869.8 ± 158.5

8.5.2 Closure Test for Other Observables

A qualitative evaluation of the fake factor estimation is performed by comparing data to the estimated background in variables which are not used for the fake factor estimation.
Figure 8.16 shows the visible mass of the two Z tagged leptons. In general, good agreement is observed, only in the lowest bin a underestimation of data by approximately 20 % is visible for electron fakes.



(a) Electron mvis($\ell_{0,Z}, \ell_{1,Z}$)

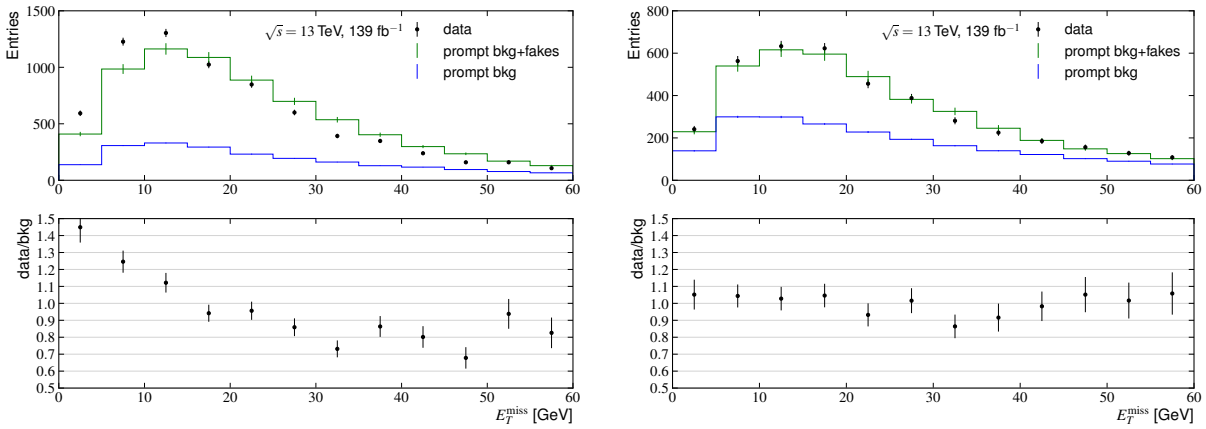
(b) Muon mvis($\ell_{0,Z}, \ell_{1,Z}$)

Figure 8.16: Visible mass of both Z-tagged leptons for an electron probe (a) and muon probe (b). The plot style follows the one in figure 8.15.

The missing transverse energy is shown in figure 8.17. For muon fakes, data and estimation are found to be in a good agreement. However, for electron fakes, the data over background ratio shows an underestimation up to 45 % and an overestimation up to 33 %. A negative slope in the ratio is visible as well. A reason for this could be fake factor dependencies which are not considered yet.
In figure 8.18, the transverse mass of the probe lepton is displayed. In case of muon fakes, a good estimation is observed. Similar to the missing transverse energy, a trend in the ratio of data over prediction is visible for electron fakes, but in contrast to the missing transverse energy, this slope is positive.
The lepton impact parameters $d_{0,\text{signif.}}$ and $z_0 \sin\theta$ are displayed in figure 8.19. For both, a good estimation around the peak value is visible. However, for larger $|d_{0,\text{signif.}}|$ and $|z_0 \sin\theta|$ values, an overestimation is observed for electron probe leptons.

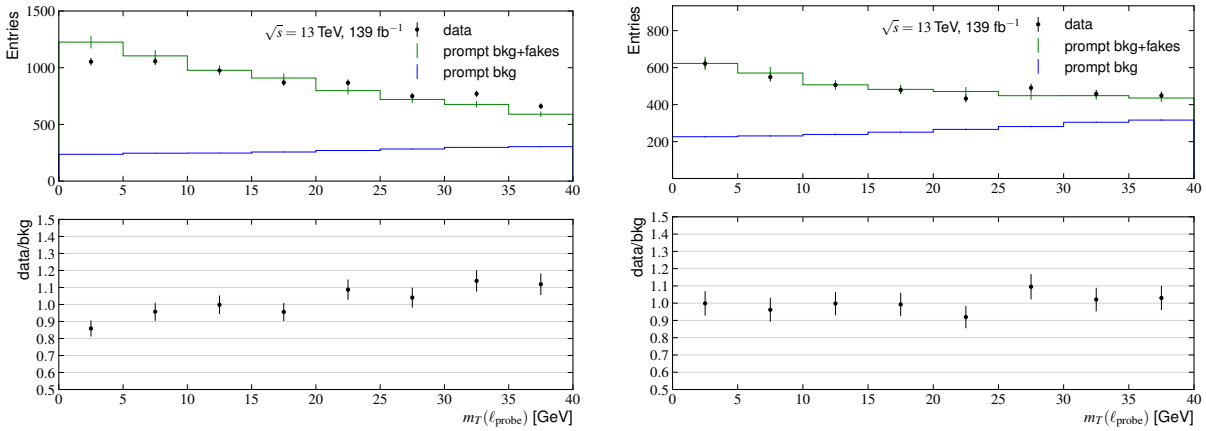
8.5.3 Discussion

In general, the closure test verifies the determination and application of the fake factor as well as the fake factor method for the fake background estimate. Most investigated quantities show good agreement between data and background in the Z+jets ID region. Issues in the modelling are seen for electrons in the missing transverse energy E_T^{miss}, the transverse mass $m_T(\ell_{\text{probe}})$ and the impact parameters.
These deviations could be solved by taking an additional dependency into account. As presented in section 8.4, both $d_{0,\text{signif.}}$ and $z_0 \sin\theta$ have influence on the fake factor in case of an electron probe lepton. However, their actual influence for the fake estimation in the Z+jets region must



(a) Electron (b) Muon

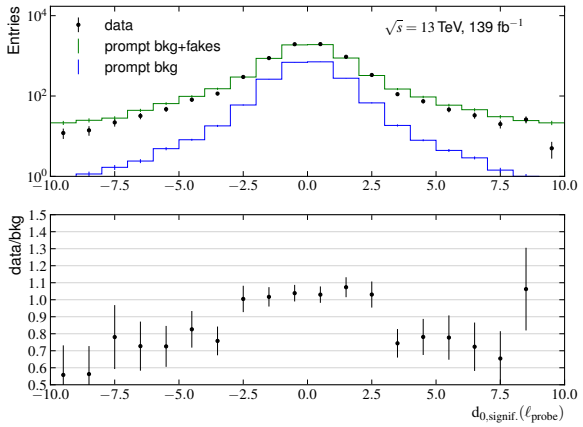
Figure 8.17: Missing transverse energy for an electron as probe lepton (a) and a muon (b). The plot style follows the one in figure 8.15.



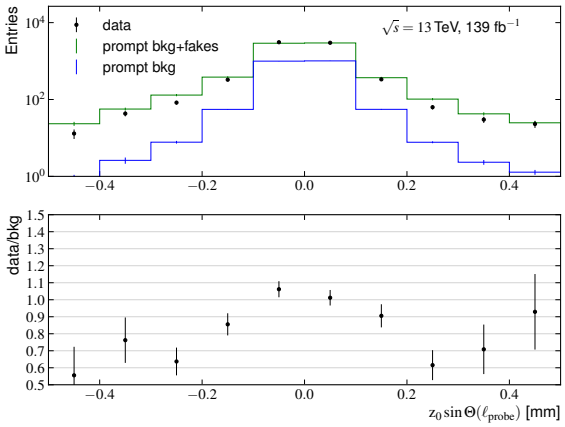
(a) Electron (b) Muon

Figure 8.18: Transverse mass of ℓ_{probe} for an electron as probe lepton (a) and a muon (b). The plot style follows the one in figure 8.15.

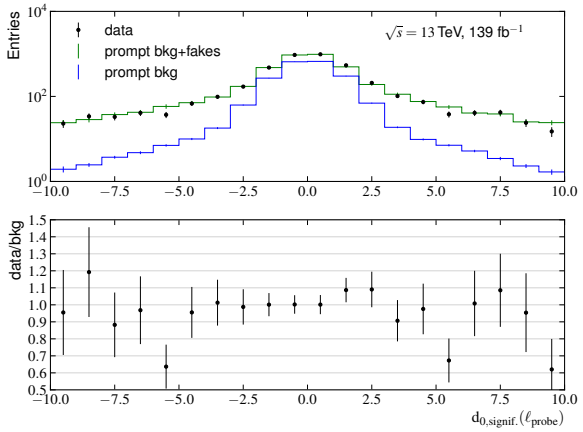
be verified. A binning in $\Delta R\,(\text{nearest jet}, \ell_{\text{probe}})$ or N_{jets} of the fake factor will not resolve the mismodelled E_T^{miss} and $m_T(\ell_{\text{probe}})$, since only a fake factor dependency in case of a muon probe lepton was observed.



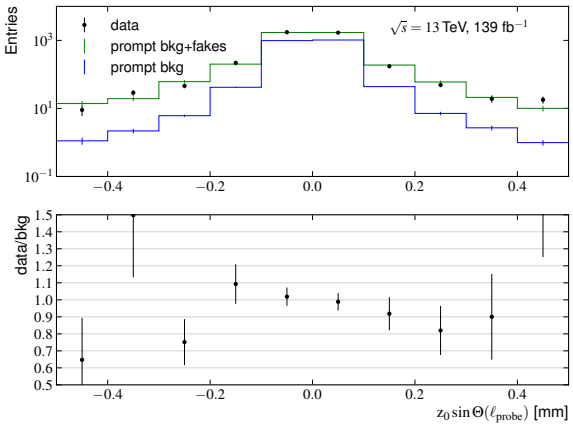
(a) Electron $d_{0,\text{signif.}}$



(b) Electron $z_0 \sin\theta$



(c) Muon $d_{0,\text{signif.}}$



(d) Muon $z_0 \sin\theta$

Figure 8.19: Impact parameters $d_{0,\text{signif.}}$ (left) and $z_0 \sin\theta$ (right) for an electron as probe lepton (top) and a muon (bottom). The plot style follows the one in figure 8.15.

9 Higgs Boson Mass Reconstruction via a Regression Neural Network

An integral part for the observation of final states originating from LFV decays is the separation of different processes that produce two different flavour leptons in the final state. To increase the sensitivity to such events originating from LFV Higgs boson decays, a precise reconstruction of the Higgs boson mass might be a very valuable separation parameter and enable a better signal-to-background separation. This may be achieved by using a regression neural network (NN) to reconstruct the Higgs boson mass.
The NN is competing with existing mass reconstruction techniques, namely the collinear mass approximation and the missing mass calculator (MMC), described in section 6.4. The MMC is a more advanced method and it is expected to outperform the collinear mass approximation. However, in prior studies at the CMS experiment, an improved mass reconstruction of the NN compared to existing methods was shown [98]. The signal-to-background significance for a $H \to \tau\tau$ signal and a Drell-Yan background was improved to 16.5 ± 0.2 compared to the significance of a existing reconstruction algorithm of 11.2 ± 0.1. Hence, especially the comparison between the NN and the MMC is of interest. In addition to the signal-to-background separation, the NN might gain in terms of computing time compared to the relatively resource-intensive MMC calculation.
Section 9.1 describes the selection of simulated events used in the training of the NN. The used hyper-parameters, architectures and input features for the various NNs are presented in section 9.2, as well as the used optimization procedures. The trained NNs are compared with existing mass reconstruction methods in section 9.3. This includes comparisons on the basis of the test set as well as with different Monte Carlo simulated signal and background processes.

9.1 Simulation and Event Selection

The NN is trained on a large range of Higgs boson masses to obtain a more generalized and reliable mass reconstruction of the signal processes as well as the background processes. For the simulated training events, only Higgs boson production processes via gluon-gluon fusion are considered, since these have the highest contribution among all Higgs production processes. These processes are simulated in the same way as described in section 7.1. The LFV Higgs boson is simulated to result in a τe pair. No decay into the $\tau\mu$ lepton pair is considered. However, due to lepton universality, it is expected that a NN trained on one decay mode reconstructs the Higgs mass similarly well for both decay modes. The lowest simulated mass point is at 60 GeV and the highest at 270 GeV, with steps of 5 GeV in between. This results in 43 different mass points. Furthermore, events are only simulated for the 2018 data-taking year with the respective pile-up profile.
The general selection is the same as described in section 7.3, however some modifications are performed to gain more statistics. No hadronic τ veto, no trigger requirements and no pseudo-rapidity selection on the leptons is applied. Furthermore, the threshold of $p_T(\ell_0)$ is reduced to 15 GeV. Since a large mass range is investigated, no selection on the visible mass is performed.

Finally, the requirements on the impact parameters are dropped as well.
For each mass point, between 10 000 to 15 000 events are generated. However, low mass points are more likely to fail the event selection resulting in less statistics. The efficiency of the event selection is shown for each mass point in figure 9.1. It ranges from 45 % to 78 %. This effect can be compensated by weighting each mass point such that the sum over all event weights after the selection is the same for every mass point. Hence, in the training process of the NN, individual events in the lower mass range become more important than events of the higher mass points.

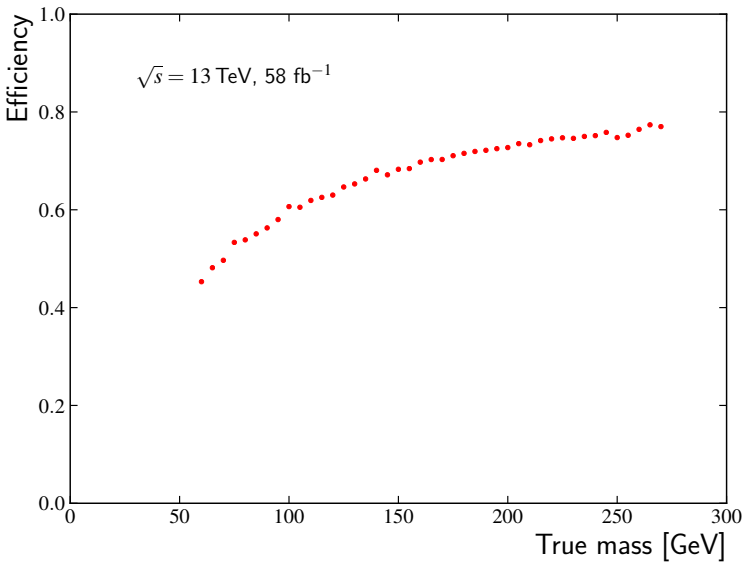


Figure 9.1: Efficiency of the event selection for all simulated $H \to \tau e$ mass points.

9.2 Optimization of Neural Network Hyper-parameters and Input Features

The simulated events are divided into three sets: 80% training set, 10% validation set and 10% test set. The individual purpose of each set is described in section 5.2.1. Several feedforward NNs are optimized via a grid-random hyper-parameter search. All NNs are written with the Python high-level library Keras [99] in combination with TensorFlow [100] as backend, which executes operations such as tensor products.

9.2.1 Utilized Hyper-parameters

All nodes in the hidden layers of the NN use the ReLu activation. As needed for the mass reconstruction, the NN has only one output node, providing the predicted mass. This node uses a linear activation, allowing also negative values of the predicted mass. The used loss function is the mean squared error function, given in equation (5.4).
To prevent the NN from overfitting, early stopping with a patience of five epochs is used. In general, no other regularization methods are applied. However, one NN is trained by using dropout as additional regularization. For the minimization of the loss function, the Adam optimizer is used. The maximum number of epochs is set for all NNs to 200.

9.2.2 Neural Network Architectures

In principle, three different architectures are used. For the most simple one, the flat layer architecture, every layer of the NN has exactly the same number of nodes (N-N-N). Two, three and four layer NNs are trained with this architecture. Studies in Ref. [101] showed that this architecture generally performs equally well or better than others. However, this can depend on the data and the task. Thus, two additional architectures are tested: a pyramid-like structure and an onion-like one. In the latter case, the first and last hidden layer have the least nodes. The number of nodes in each layer increases towards the middle of the NN and decreases afterwards in such a way, that a symmetry with respect to the central layer is obtained (N-2N-N). The pyramid structure describes a NN, where the number of nodes decreases with each hidden layer (3N-2N-N). For the pyramid-like and onion-like layer architectures, only three layer NNs are trained. The number of nodes N for the NNs is between 200 and 400.

9.2.3 Sets of Input Features

Two different sets of input features are used. Both sets are based on the leading lepton, the sub-leading lepton and the missing transverse energy. More advanced input features are included in the first set, whereas only the fundamental quantities of the four-momenta are used in the second set. The first set of input features is noted as IF1 and the second one as IF2.
The input features in both sets are standardized as described in section 5.2.2. Furthermore, the coordinate system is rotated in both cases such that the azimuth angle $\phi(\text{MET}) = 0$. The azimuth angle of both leptons is restricted to $(-\pi, +\pi]$. This restriction is also applied to differences of azimuth angles.
The first set of input features is presented in table 9.1.

Table 9.1: First set of input features for the regression NN, determining the Higgs boson mass.

Parameter	Description
$p_T(\ell_0)$	Transverse momentum of leading lepton
$p_T(\ell_1)$	Transverse momentum of sub-leading lepton
E_T^{miss}	Missing transverse energy
$\Delta\eta(\ell_0, \ell_1)$	Pseudorapidity difference between leading and sub-leading lepton
$m_T(\ell_0)$	Transverse mass of leading lepton and missing transverse energy
$m_T(\ell_1)$	Transverse mass of sub-leading lepton and missing transverse energy
m_{coll}	Collinear mass
m_{vis}	Visible mass
S	Significance of missing transverse energy
$\Delta\phi(\ell_0, \ell_1)$	Azimuth angle difference between leading and sub-leading lepton
$\Delta\phi(\ell_0, E_T^{\text{miss}})$	Azimuth angle difference between leading lepton and missing transverse energy
$\Delta\phi(\ell_1, E_T^{\text{miss}})$	Azimuth angle difference between sub-leading lepton and missing transverse energy

The transverse momenta of the leading and sub-leading lepton as well as the missing transverse energy are used as well as the respective azimuth angle differences. Leptons originating from heavy particles like the top quark have in general a large transverse momenta which can be used to separate these processes from the signal process. Further, the pseudorapidity difference between the leading and the sub-leading lepton is used. For low mass points, the leptons are more likely to have the same pseudorapidity than for high mass points. Important input features are the collinear mass and the visible mass of the dilepton system.
The transverse mass, defined in equation (8.5), is also used as input feature in this set. Especially the transverse mass of the leading lepton shows a dependency on the mass of the Higgs boson.

The last one in this set of input features is the significance of the missing transverse energy. This significance is calculated as log-likelihood ration of the hypothesis that the total transverse momentum due to undetected particles is equal to zero against the hypothesis that it is unequal zero. It can be written as

$$S^2 = 2 \ln \left(\frac{\max_{\vec{p}_T^{\,\text{und.}}=\vec{0}} L(\vec{E}_T^{\text{miss}})}{\max_{\vec{p}_T^{\,\text{und.}}\neq\vec{0}} L(\vec{E}_T^{\text{miss}})} \right),$$

where the numerator and denominator are maximised likelihood functions for a given E_T^{miss} under the constraints of no undetected transverse momentum ($\vec{p}_T^{\,\text{und.}} = \vec{0}$) and non-zero undetected transverse momentum ($\vec{p}_T^{\,\text{und.}} \neq \vec{0}$) [102].
For five mass points, the m_{coll} and $p_T(\ell_0)$ distributions are displayed in figure 9.2. As expected, the collinear mass is already decently predicting the actual mass. The mean value of all five distributions is within $\pm 8\,\text{GeV}$ of the actual mass. However, the standard deviation increases with each mass point up to $\approx 27\,\text{GeV}$ for the 255 GeV mass point. The leading p_T distribution becomes broader for higher masses.

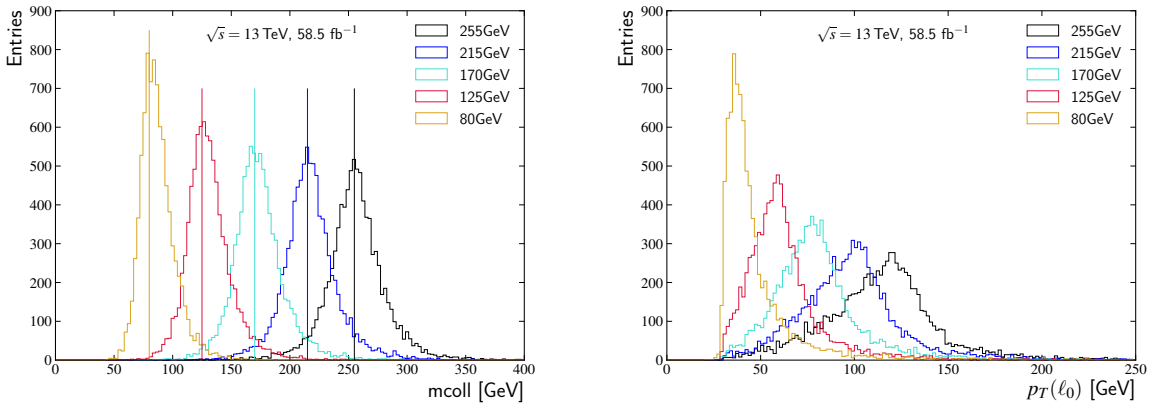


Figure 9.2: Kinematic distributions as a function of m_{coll} (left) and $p_T(\ell_0)$ (right) for different mass points of LFV decaying Higgs bosons, which are used as input features in the training of the NN. The actual mass values are indicated by the vertical lines.

In figure 9.3, the m_{vis} and $p_T(\ell_1)$ distributions are shown. Both distributions are much broader for high masses than for low masses.
The distributions of E_T^{miss} and S are shown in figure 9.4. For small masses, both distributions are shifted to smaller values and the E_T^{miss} distribution becomes sharper.
In figure 9.5, the distributions of $m_T(\ell_0)$ and $m_T(\ell_1)$ are displayed as well as the ones of the angular differences $\Delta\phi$ and pseudorapidity difference $\Delta\eta(\ell_0, \ell_1)$. For higher masses, the $m_T(\ell_0)$ distribution is located around larger values and is broader distributed, whereas $m_T(\ell_1)$ is shifted slightly to smaller values. The distributions of $\Delta\phi(\ell_0, \ell_1)$ and $\Delta\phi(\ell_0, E_T^{\text{miss}})$ are shifted for higher masses more to $\pm\pi$, whereas the $\Delta\phi(\ell_1, E_T^{\text{miss}})$ distribution has a sharper peak around 0. This indicates an alignment of the sub-leading lepton and the missing transverse energy. For smaller masses, the $\Delta\eta(\ell_0, \ell_1)$ distribution is increasingly located at smaller values.
During the training phase, the NN tries to learn how the respective input features influence the reconstructed mass and adjust its weights and biases to predict the mass points more precisely.

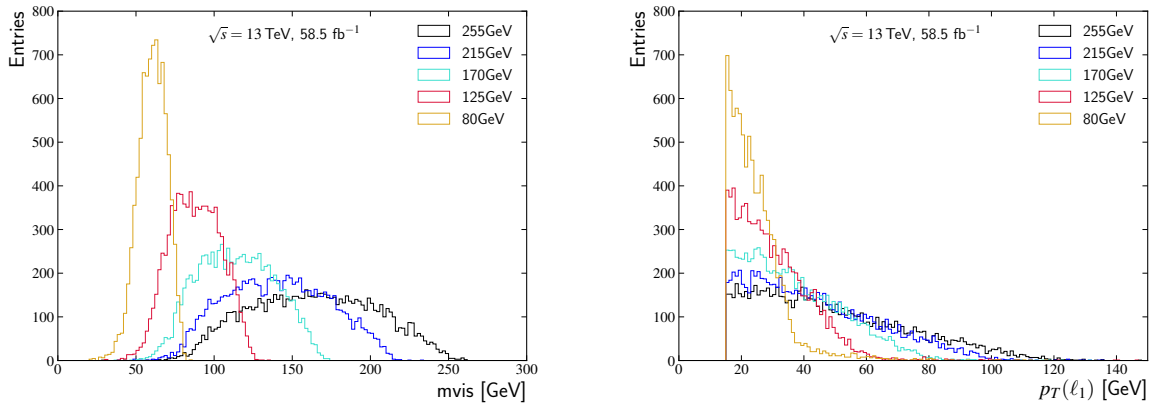


Figure 9.3: Kinematic distributions as a function of m_{vis} (left) and $p_T(\ell_1)$ (right) for different mass points of LFV decaying Higgs bosons, which are used as input features in the training of the NN.

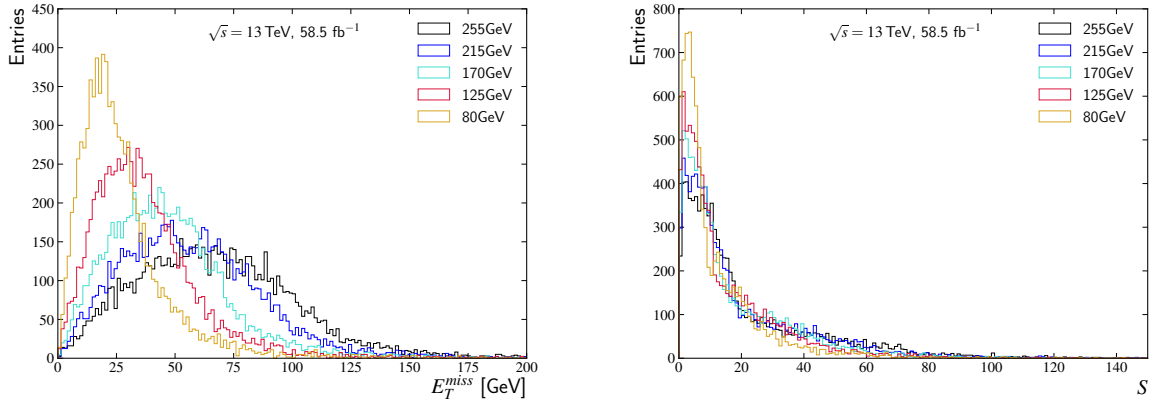


Figure 9.4: Kinematic distributions as a function of E_T^{miss} (left) and S (right) for different mass points of LFV decaying Higgs bosons, which are used as input features in the training of the NN.

IF1 includes measured quantities as well as calculated quantities like the collinear mass. On the one hand, this allows more advanced input features which might help the NN to reconstruct the mass more precisely. On the other hand, this can also lead to a bias in the training, as the NN could learn that only the collinear mass is important for the mass reconstruction since it is already a decently working reconstruction method.

For comparison, an additional NN is trained only using the fundamental four-momenta of the final state particles. For the azimuthal angles, the 2π symmetry of the detector is factored in by taking the sine and cosine, so that the NN does not have to extract it from the input features itself. In contrast to IF1, no input feature in IF2 is already a mass reconstruction method. Due to the rotation of the coordinate system, $\phi(E_T^{\text{miss}})$ is not present in this set. The input features of the second set are listed in table 9.2.

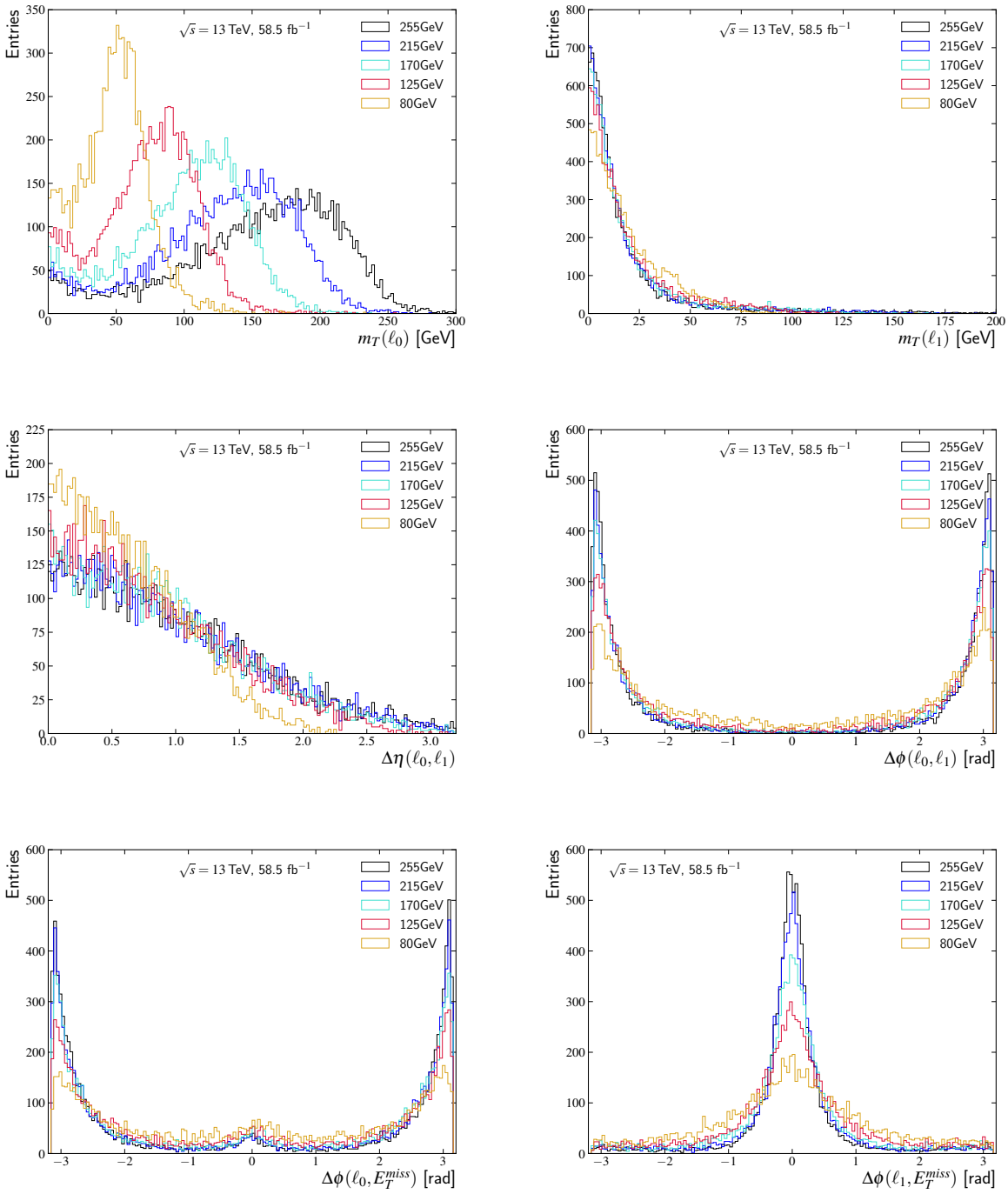


Figure 9.5: Kinematic distributions as a function of $m_T(\ell_0)$ (top left), $m_T(\ell_1)$ (top right), $\Delta\eta(\ell_0, \ell_1)$ (middle left), $\Delta\phi(\ell_0, \ell_1)$ (middle right), $\Delta\phi(\ell_0, E_T^{\text{miss}})$ (bottom left) and $\Delta\phi(\ell_1, E_T^{\text{miss}})$ (bottom right) for different mass points of LFV decaying Higgs bosons, which are used as input features in the training of the NN.

In figure 9.6, the distributions of $E(\ell_0)$ and $E(\ell_1)$ are shown. For higher masses, both distributions are located at higher values and become broader.

Table 9.2: Second set of input features for the training of a NN to reconstruct the LFV Higgs boson mass. Only the fundamental four-momenta are used.

Parameter	Description
$p_T(\ell_0)$	Transverse momentum of leading lepton
$p_T(\ell_1)$	Transverse momentum of sub-leading lepton
E_T^{miss}	Missing transverse energy
$E(\ell_0)$	Energy of leading lepton
$E(\ell_1)$	Energy of sub-leading lepton
$\eta(\ell_0)$	Pseudorapidity of leading lepton
$\eta(\ell_1)$	Pseudorapidity of sub-leading lepton
$\sin\phi(\ell_0)$, $\cos\phi(\ell_0)$	Azimuth angle of leading lepton
$\sin\phi(\ell_1)$, $\cos\phi(\ell_1)$	Azimuth angle of sub-leading lepton

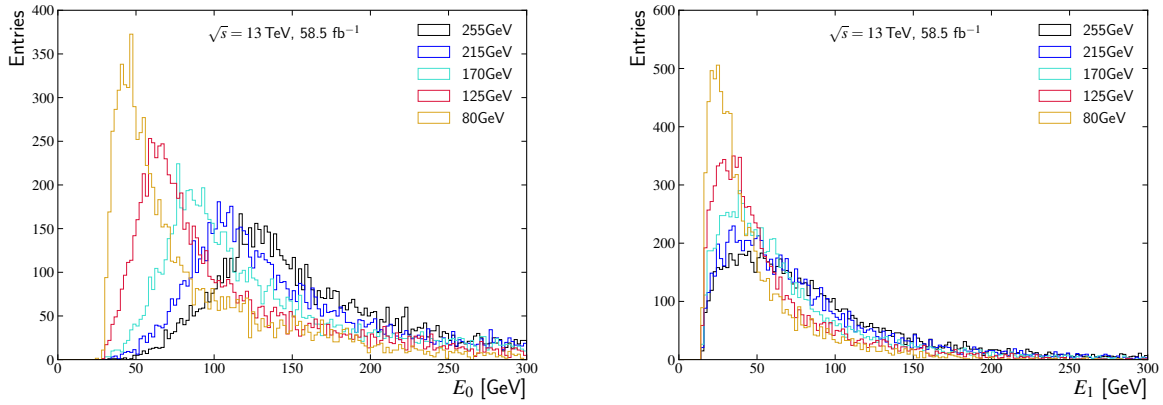


Figure 9.6: Kinematic distributions as a function of $E(\ell_0)$ (left) and $E(\ell_1)$ (right) for different mass points of LFV decaying Higgs bosons, which are used as input features in the training of the NN.

The η, $\sin\phi$ and $\cos\phi$ distributions of the leading and sub-leading lepton are displayed in figure 9.7. The pseudorapidity distributions are for all masses approximately the same. Both $\sin\phi$ distributions have a sharper peak around 0 for higher masses and the $\cos\phi$ distributions are shifted stronger to -1 and $+1$ for the leading and sub-leading lepton, respectively. This is due to the rotation of the coordinate system as well as the enhance collinearity of the sub-leading lepton and the missing transverse energy for higher masses.

9.2.4 Hyper-parameter Optimization with Grid-Random-Search

The free hyper-parameters of the NN training, i.e. the learning rate, the decay, the nodes in each layer and the batch size are optimized with a grid-random search as described in section 5.3.2. If dropout is used, the dropout rate is considered as optimized hyper-parameter as well.
For the optimization, NNs are trained for each combination of values tested for the free hyper-parameters within a given range, along randomly varied grid points. NNs without dropout are trained with 36 different hyper-parameter configurations and if dropout is used, 216 differenct configurations are tested.

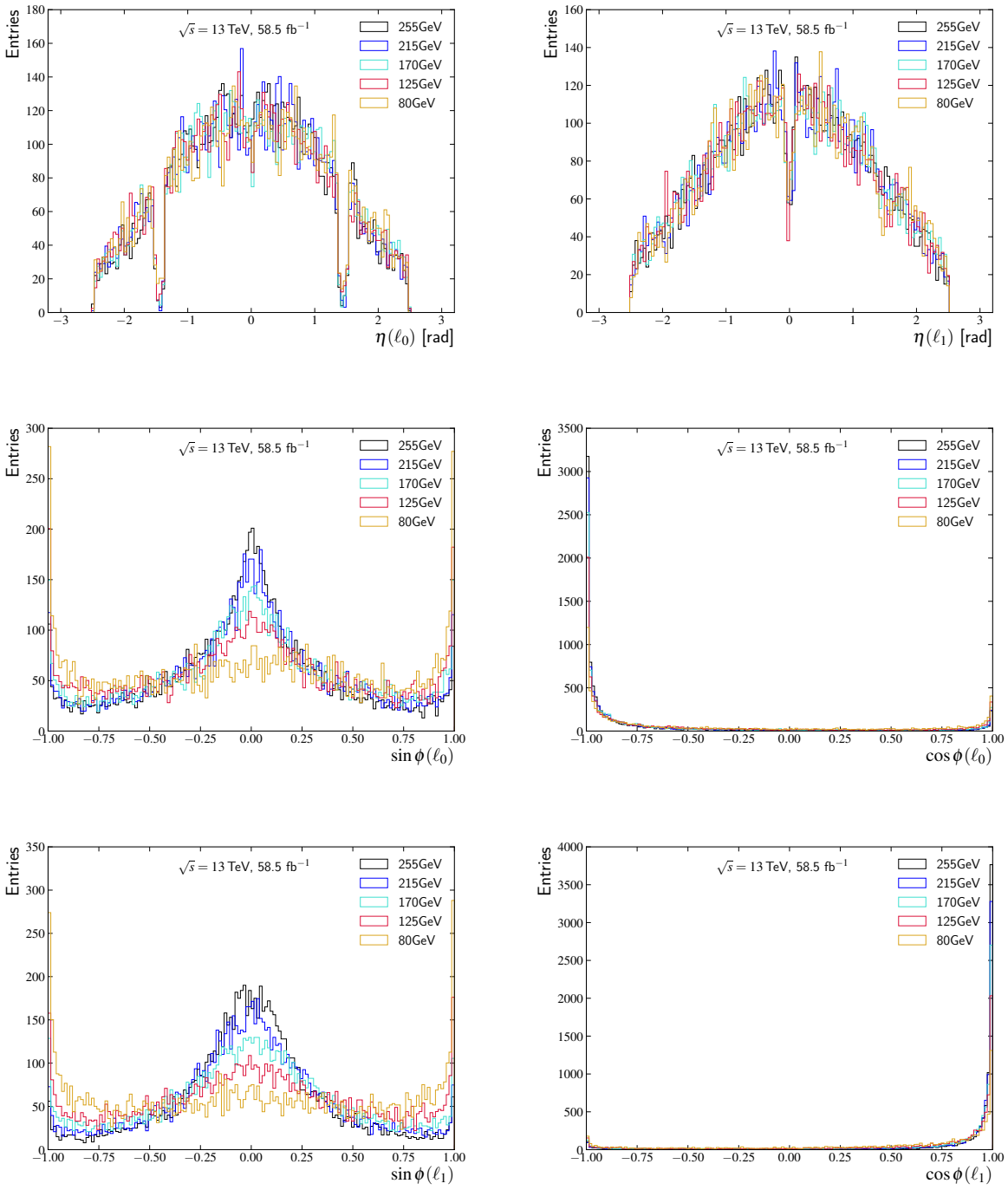


Figure 9.7: Kinematic distributions as a function of $\eta(\ell_0)$ (top left), $\eta(\ell_1)$ (top right), $\sin\phi(\ell_0)$ (middle left), $\cos\phi(\ell_0)$ (middle right), $\sin\phi(\ell_1)$ (bottom left) and $\cos\phi(\ell_1)$ (bottom right) for different mass points of LFV decaying Higgs bosons, which are used as input features in the training of the NN.

The random shift of one hyper-parameter i at the grid point g_i is given by

$$\Delta g_i = \frac{g_i}{10} \cdot R_i$$

with the random value R_i restricted to the interval $(-1, 1]$. Noting $\vec{g}$ as the vector build from all free hyper-parameters, the new grid point is derived as

$$\vec{g}' = \vec{g} + \Delta\vec{g}.$$

All NNs except the one with dropout regularization use the predefined grid for the grid-random search shown in table 9.3.

Table 9.3: Predefined grid for the grid-random search for NNs without dropout regularization.

Free hyper-parameter	Number of grid points	Grid values
Learning rate	2	0.001, 0.01
Decay	2	0.001, 0.01
Nodes	3	200, 300, 400
Batch size	3	30, 50, 100

The case of an onion architecture, the parameter 'nodes' refers to the number of nodes in the first and last hidden layer. For every other hidden layer, the position P of the layer in the layer architecture, or to be precise its difference in position with respect to the first or last hidden layer, is used to calculate the number of nodes. The number of nodes is calculated as

$$\text{nodes for position } P = \text{nodes} \times (P + 1) \tag{9.1}$$

resulting in integer multiples of the value 'nodes'.
For the pyramid architecture the optimization procedure has one modification. The parameter 'nodes' refers only to the number of nodes in the last hidden layer and the position P is the distance to this last hidden layer. Hence, the respective nodes in each layer are calculated as in equation (9.1).
If dropout regularization used, a different grid is defined as the additional regularization is expected to influence the optimal values for most of the hyper-parameters. The predefined grid is shown in table 9.4.

Table 9.4: Predefined grid for the grid-random search for NNs with dropout regularization.

Free hyper-parameter	Number of grid points	Grid values
Learning rate	3	0.01, 0.1, 0.5
Decay	2	0.01, 0.1
Nodes	3	200, 300, 400
Batch size	3	50, 100, 200
Dropout rate	4	0.05, 0.1, 0.2, 0.4

This additional regularization is only applied to flat layer architectures where all hidden layers have the same number of nodes.

9.2.5 Application of the Principle Component Analysis

For the NN training based on the IF1 input features, a principle component analysis (PCA), as discussed in section 5.3.3, is applied. As a result of the PCA, the 12 input features are combined as linear combinations among which the two least important ones are dropped from the training of the NN. The transformation matrix, constructed by the eigenvectors of the input feature

correlation matrix, is used to derive the linear combinations. With this, 10 uncorrelated input features are obtained which still cover over 99% of the initial input feature information.
Furthermore, the transformation matrix is used to train one additional NN with a reduced number of input features of the IF1 set and a flat layer architecture. As discussed in section 5.3.3, each row of the transformation matrix corresponds exactly to one initial input feature, here IF1. Deriving the absolute value norm $\sum_i |V_{ji}|$ for row j of the transformation matrix V, the corresponding input feature is eliminated from the set, if the norm does not exceed a given limit. This limit is set to be 0.7. As a result, all azimuth differences as well as the significance of the missing transverse energy are eliminated as input features for this specific NN.

9.2.6 Resulting Hyper-parameters and Input Features

In total, 10 different optimized NNs are trained, tested and compared. A summary of the NNs is presented in table 9.5. Two (IF1L2F), three (IF1L3F) and four layer (IF1L4F) NNs with a flat layer architecture are compared as well as three layer NNs with a pyramid-like (IF1L3P) and onion-like (IF1L3O) architecture.
A neural network (IF1L3F-no m_{coll}) with the IF1 input features, but without the collinear mass is trained as well. It is used to verify, that the NN is able to predict the Higgs boson mass correctly, missing an actual mass reconstruction as input feature. One NN (IF1L2F-dropout) is also trained using dropout regularization in each hidden layer. Using the transformation matrix of the PCA method, one NN (IF1L3F-redPCA) with reduced IF1 input features is trained. Another one (IF1L3F-uncPCA) uses the full PCA method and has only 10 linear uncorrelated input features. Finally, one NN (IF2L4F) with the IF2 input features is used to investigate the capability of the NN to reconstruct the mass with only the most fundamental input features.

Table 9.5: Architecture and hyper-parameters of the 10 NNs, optimized with the grid-random hyper-parameter search. The names of the NNs are structured as: set of input features, number of layers, architecture. The latter F describes a NN with a flat layer architecture, whereas P and O represent the pyramid-like and onion-like architecture respectively. For the NN IF1L3F-no m_{coll}, the collinear mass is not included as input feature. IF1L2F-dropout uses a dropout regularization with a dropout rate of 4.6% in each hidden layer. The last two noted NNs use the PCA, where IF1L3F-uncPCA has uncorrelated and reduced input features and IF1L3F-redPCA reduced input features.

NN name	Input features	Layers	Nodes	Learning rate	Decay	Batch size
IF1L2F	IF1	2	304-304	0.0109	0.0009	51
IF1L3F	IF1	3	365-365-365	0.0105	0.0010	28
IF1L4F	IF1	4	275-275-275-275	0.0095	0.0011	45
IF2L4F	IF2	4	279-279-279-279	0.0093	0.0010	31
IF1L3F-no m_{coll}	IF1	3	191-191-191	0.0109	0.0001	54
IF1L2F-dropout	IF1	2	204-204	0.0905	0.0095	198
IF1L3P	IF1	3	1281-854-427	0.0103	0.0107	96
IF1L3O	IF1	3	309-618-309	0.0108	0.0001	31
IF1L3F-uncPCA	IF1	3	218-218-218	0.0094	0.0011	51
IF1L3F-redPCA	IF1	3	194-194-194	0.0103	0.0010	51

9.2.7 Bias in Mass Prediction and its Correction

After the training the performance of each NN is evaluated based on simulated events in the validation set. The Higgs boson mass predicted by the NN is compared to the true value.

The relative deviation

$$\text{bias} = \frac{\text{prediction} - \text{target}}{\text{target}} \tag{9.2}$$

between the predicted mass and the true mass value, denoted as target, is called *bias*. It describes how precise the NN predicts the actual mass, and is evaluated for each target mass separately. The arithmetic mean of the bias is derived over all events in one mass point of the validation set. Figure 9.8 displays the bias as a function of the true Higgs boson masses. For each mass point, the x-value is derived as arithmetic mean of the derived mass values of the NN. The uncertainties are calculated as standard error of the mean, i.e. the standard deviation divided by the square root of the number of events.
A mean value that deviates from zero indicates that there is an offset in the prediction of the NN for this mass point. A systematic trend can be observed and is in general well described by a polynomial fit of 3$^{\text{rd}}$ degree. However, in the low mass range, the fit is less steep than the observed bias dependency.

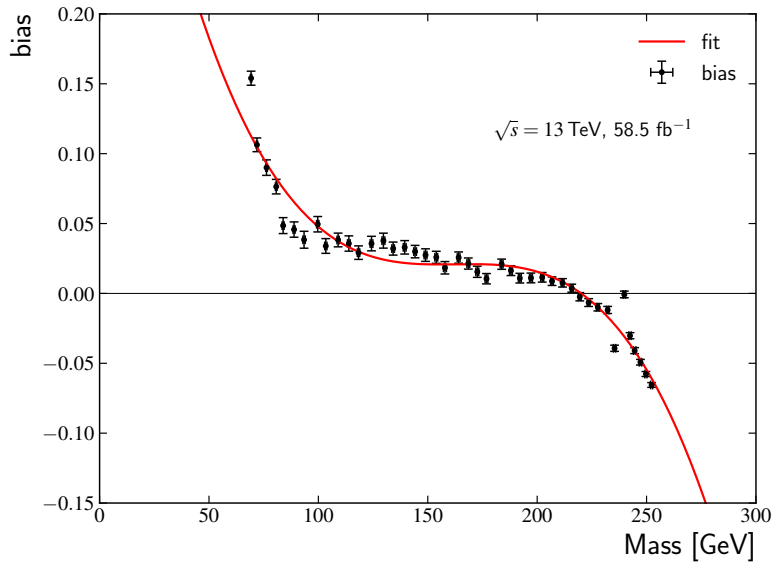
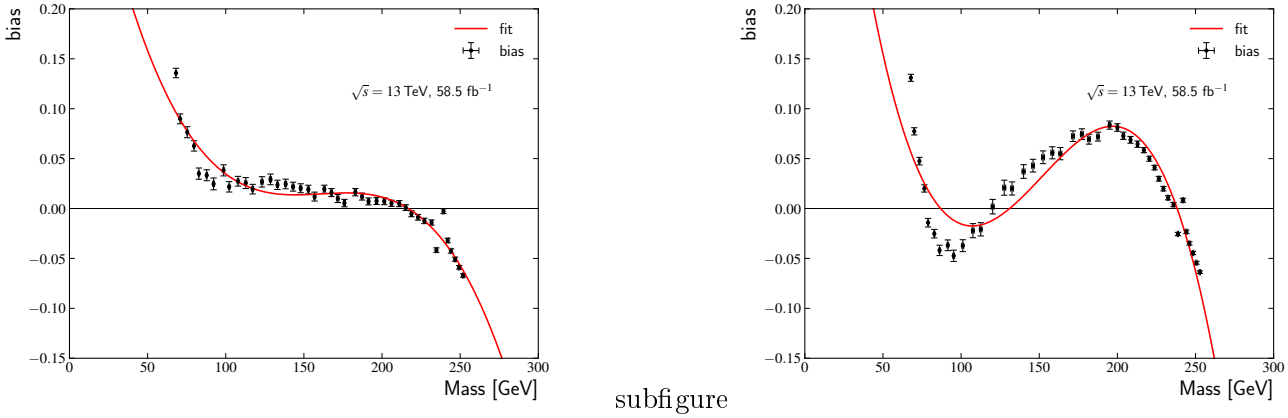


Figure 9.8: Bias in the derived mass from the NN IF1L4F after training for each mass point. The points from the MC simulation (black markers) are fitted with a polynomial fit of 3$^{\text{rd}}$ degree (red line).

In the central mass region between 100 GeV and 200 GeV the bias is rather small and flat at approximately 2 %. At low and high masses, below 100 GeV and above 260 GeV, the bias becomes larger than 5 % and increases rapidly. This bias is tested in two ways. In the first one, a NN is trained with many more nodes per layer, such that more free parameters are available to adapt to the mass reconstruction task. For this, all hyper-parameters are the same as for the NN IF1L3F, but with 1000 nodes in each layer. In the second test, the low and high masses during the training process of the NN IF1L3F where weighted by an order of magnitude larger than in the standard training. However, the first NN has almost the exact same dependency and the latter NN shows a much stronger dependency, displayed in figure 9.9.
The polynomial fit can be used to correct the systematic trend. Using equation (9.2), the corrected mass reconstruction of the NN can be derived on an event-by-event basis from the initial mass



subfigure

Figure 9.9: Bias in the derived mass of two NNs with alternative training processes. Both have the IF1L3F hyper-parameters, but one has 1000 nodes per layer (left) and for the other, mass points below 90 GeV and above 210 GeV are weighted by an order of magnitude larger than in the standard training (right). The points from the MC simulation (black markers) are fitted with a polynomial fit of 3$^{\text{rd}}$ degree (red line).

reconstruction of the NN and the performed polynomial fit, given by

$$\text{corrected mass} = \frac{\text{initial mass}}{\text{correction fit(initial mass)}+1} \tag{9.3}$$

where the polynomial fit depends only on the initial mass reconstruction, before the performed correction is applied. In table 9.6, the obtained fit parameters for all NNs are presented.

Table 9.6: Obtained parameters of the performed polynomial fit $a + b \cdot \text{initial mass} + c \cdot (\text{initial mass})^2 + d \cdot (\text{initial mass})^3$.

NN	$a/10^{-1}$	$b/10^{-3}$	$c/10^{-5}$	$d/10^{-7}$
IF1L2F	4.93	-9.04	5.65	-1.17
IF1L3F	4.94	-9.09	5.67	-1.16
IF1L4F	5.13	-9.12	5.62	-1.15
IF2L4F	5.43	-9.84	6.10	-1.25
IF1L3F-no m_{coll}	4.71	-8.65	5.39	-1.11
IF1L2F-dropout	5.05	-9.07	5.59	-1.15
IF1L3P	5.16	-9.47	5.89	-1.21
IF1L3O	6.14	-10.74	6.48	-1.30
IF1L3F-uncPCA	5.24	-9.74	6.09	-1.25
IF1L3F-redPCA	4.62	-8.57	5.41	-1.12

The distribution of the biases over all selected events in the validation set for the mass point of 75 GeV is shown in figure 9.10, with and without the bias correction fit. With the correction applied, the mean of the distribution is clearly shifted towards zero, hence the overall bias is reduced. This observed for all NNs.
Connected to the bias is the resolution. The resolution of the mass reconstruction is defined as the standard deviation of the difference between the NN mass reconstruction and the actual mass

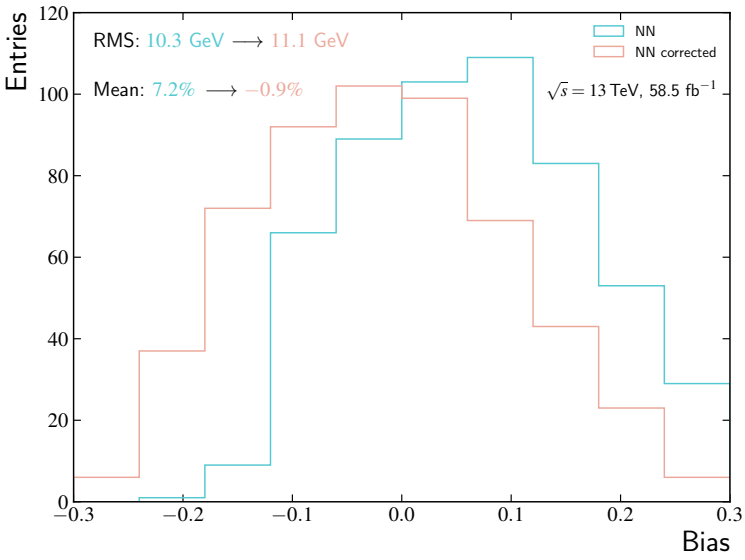


Figure 9.10: Bias in the mass reconstruction from the IF1L4F NN for the 75 GeV mass point, evaluated for the validation dataset. Both, the bias distribution without bias correction (blue) and with bias correction applied (orange) are shown. The resolution and mean of the bias are indicated in the top left.

for a given mass point. This quantity describes how peaked the reconstructed mass distribution is. Thus, a small resolution provides a better separation between the individual mass points. However, the application of the bias correction in general worsens the resolution. Since the main goal of this work is to improve the resolution to gain separation power between signal and background, this correction fit is not applied in the following.

9.3 Comparison of NNs to existing Mass Reconstruction Methods

The trained NNs are compared with each other and to existing mass reconstruction methods, namely the collinear mass and missing mass calculator described in section 6.4. The evaluation of the bias and resolution is performed using data from the test set and for Higgs LFV and Z LFV simulated events. The separation power between signal and background is compared for both LFV signal processes.

9.3.1 Comparison of various Mass Reconstruction Techniques with Simulated Events from the Test Set

The bias and resolution are evaluated for four different true mass ranges, which ensure enough statistics. The lowest mass range contains the Z boson mass peak with a range of $[60, 100)$ GeV. The second mass range includes the Higgs boson mass peak and is set to $[100, 150]$ GeV. The mass range above is split in $(150, 200]$ GeV and $(200, 270]$ GeV. To calculate the resolution in the respective mass range, the difference between the reconstructed mass and the true mass is filled in a histogram. Each event is weighted equally and the resolution is derived as standard deviation of the histogram. Similar to this, a histogram is filled with the bias of each of those

events and the noted bias is given as mean value of the histogram.
An improvement of the resolution in the first two mass ranges is of high interest to enable a better separation between the signal-like processes and the irreducible $Z \to \tau\tau$ background. However, to improve the separation of signal-like processes and the background due to top and diboson contributions, higher masses must also be predicted reasonably well. The bias is summarized in table 9.7 and the resolution in table 9.8. The uncertainty on the bias is derived as $\sigma/\sqrt{N}$ and on the resolution as $\sigma/\sqrt{2N-2}$, where σ is the standard deviation and N the number of events.

Table 9.7: Bias of the trained NNs and existing mass reconstruction methods. All values are stated as percentage difference with respect to the true masses. The abbreviations are described in table 9.5.

Mass reconstruction	$[60, 100)$ GeV	$[100, 150]$ GeV	$(150, 200]$ GeV	$(200, 270]$ GeV
IF1L2F	6.4 ± 0.2	2.1 ± 0.1	1.2 ± 0.1	-2.3 ± 0.1
IF1L3F	6.2 ± 0.2	2.0 ± 0.1	1.2 ± 0.1	-2.2 ± 0.1
IF1L4F	7.4 ± 0.2	2.8 ± 0.1	1.6 ± 0.1	-2.1 ± 0.1
IF2L4F	7.2 ± 0.2	2.5 ± 0.1	1.5 ± 0.1	-2.1 ± 0.1
IF1L3F-no m_{coll}	6.1 ± 0.2	1.9 ± 0.1	1.1 ± 0.1	-2.2 ± 0.1
IF1L2F-dropout	7.0 ± 0.2	2.4 ± 0.1	1.1 ± 0.1	-2.6 ± 0.1
IF1L3P	6.5 ± 0.2	1.9 ± 0.1	0.8 ± 0.1	-2.7 ± 0.1
IF1L3O	8.8 ± 0.2	3.0 ± 0.1	1.4 ± 0.1	-2.4 ± 0.1
IF1L3F-uncPCA	6.2 ± 0.2	1.7 ± 0.1	0.8 ± 0.1	-2.7 ± 0.1
IF1L3F-redPCA	5.9 ± 0.2	2.1 ± 0.1	1.4 ± 0.1	-2.0 ± 0.1
m_{coll}	10.3 ± 0.3	4.0 ± 0.2	2.0 ± 0.1	0.9 ± 0.1
MMC_{maxw}	-1.5 ± 0.2	-2.9 ± 0.2	-2.5 ± 0.1	-2.0 ± 0.1
MMC_{mlm}	4.2 ± 0.2	0.4 ± 0.2	-0.6 ± 0.1	-0.8 ± 0.1

Table 9.8: Resolution of the trained NNs and existing mass reconstruction methods. All stated values are in units of GeV. The abbreviations are described in table 9.5.

Mass reconstruction	$[60, 100)$ GeV	$[100, 150]$ GeV	$(150, 200]$ GeV	$(200, 270]$ GeV
IF1L2F	10.6 ± 0.1	16.5 ± 0.1	19.0 ± 0.1	18.3 ± 0.1
IF1L3F	10.6 ± 0.1	16.6 ± 0.1	19.0 ± 0.1	18.3 ± 0.1
IF1L4F	10.7 ± 0.1	16.6 ± 0.1	18.9 ± 0.1	18.0 ± 0.1
IF2L4F	10.9 ± 0.1	16.7 ± 0.1	19.2 ± 0.2	18.3 ± 0.1
IF1L3F-no m_{coll}	10.7 ± 0.1	16.6 ± 0.1	19.0 ± 0.1	18.4 ± 0.1
IF1L2F-dropout	10.7 ± 0.1	16.5 ± 0.1	18.8 ± 0.1	18.3 ± 0.1
IF1L3P	10.5 ± 0.1	16.5 ± 0.1	18.9 ± 0.1	18.2 ± 0.1
IF1L3O	10.5 ± 0.1	16.3 ± 0.1	18.7 ± 0.1	18.0 ± 0.1
IF1L3F-uncPCA	10.6 ± 0.1	16.5 ± 0.1	18.9 ± 0.1	18.1 ± 0.1
IF1L3F-redPCA	10.7 ± 0.1	16.8 ± 0.1	19.2 ± 0.2	18.3 ± 0.1
m_{coll}	14.8 ± 0.2	19.5 ± 0.2	22.8 ± 0.2	25.7 ± 0.2
MMC_{maxw}	12.1 ± 0.1	17.4 ± 0.1	20.9 ± 0.2	24.0 ± 0.2
MMC_{mlm}	11.6 ± 0.1	16.6 ± 0.1	20.2 ± 0.2	23.5 ± 0.2

In general, all NNs perform similarly well with respect to bias and resolution. The NN trained without the collinear mass (IF1L3F-no m_{coll}) performs slightly better in the bias, but marginally worse in terms of resolution compared to the NN with the same architecture, but including m_{coll}

as input feature. The NN trained only with the four momenta (IF2L4F) accomplishes a similar resolution and bias as the other NNs, but is in most cases slightly below the average of all NNs. The observations verify:

- A NN does not need an actual mass reconstruction as input feature for a good performance of the mass reconstruction.
- The compared NN layer architectures perform equally well. The NN with a pyramid-like architecture (IF1L3P) performs very similar to the same NN with a flat layer architecture. In case of the NN with an onion-like architecture (IF1L3O), a slightly better resolution with a worse bias is observed.
- Input features can be eliminated from the input feature set without impairing the NN mass reconstruction performance. This is shown by the NN IF1L3F-uncPCA and NN IF1L3F-redPCA. In case of the former NN, decorrelated input features are used and two of those are eliminated from the set. The NN IF1L3F-redPCA uses only eight input features of the first input feature set. In all mass ranges, both NNs perform very similar as the NN IF1L3F, which has the same layer architecture, but all input features of the first set.

The trained NNs are compared to the collinear mass approximation m_{coll} and the two variants of the missing mass calculator MMC_{maxw} and MMC_{mlm}. Compared with the collinear mass reconstruction, the NNs perform better with respect to the resolution and almost always better with respect to the bias. This is also the case for the NN without the collinear mass as input feature (IF1L3F-no m_{coll}) as well as the NN using only the four momenta (IF2L4F). In general, the MMC_{maxw} mass reconstruction performs worse than the NNs, but better than the collinear mass. Compared to the NNs, only the MMC_{mlm} mass reconstruction has a very similar mass resolution in the range between 100 and 150 Gev, but still a slightly worse one in the other mass ranges. However, its bias is in general much closer to zero than the ones of the NNs. The resolution for each mass point of the NN IF1L3F-uncPCA and the three existing mass reconstruction techniques is shown in figure 9.11.
Both MMC variants take advantage of an additional event selection due to the parameter scan over the x- and y-component of the di-neutrino momentum, discussed in section 6.4.
To summarize, the trained NNs outperform in almost all cases the existing mass reconstruction methods in terms of the mass resolution. Compared to the existing mass reconstruction methods, the resolution can be improved significantly in the higher mass ranges by at least 1 GeV and up to 7.7 GeV. This corresponds to an improved resolution between 5 % and 30 %. Improvements for lower mass ranges are of the order of 1 GeV, which is equivalent to an improvement of 10 %. However, the MMC_{mlm} has the smallest bias, except in the lowest mass range, where MMC_{maxw} has the smallest bias.

9.3.2 Investigation of the Bias at Low Masses

To investigate a possible improvement of the NN bias, two additional mass points are added to each event sample at 50 GeV and 55 GeV. The IF1L3F configuration is used with a new hyper-parameter search resulting in a learning rate of 0.0099, a decay of 0.0010, a batch size of 28 and 280 nodes in each layer. This NN is called IF1L3F-low in the following. The obtained bias in the validation set as well as the performed polynomial fit is displayed in figure 9.12 next to the IF1L3F NN without the two lower mass points. In table 9.9, the polynomial fit parameters for both NNs are presented.

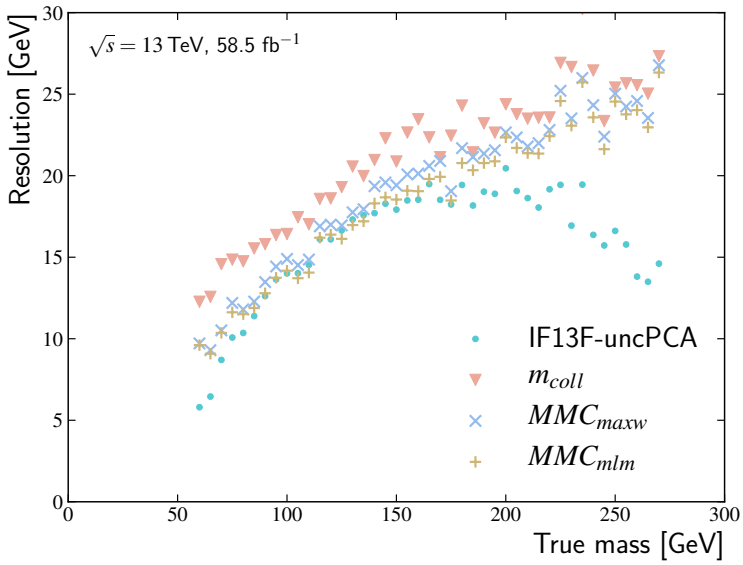


Figure 9.11: Resolution for the NN IF1L3F-uncPCA and the three existing mass reconstruction techniques for all mass points. The NN shows a better resolution for masses below 90 GeV and above 175 GeV.

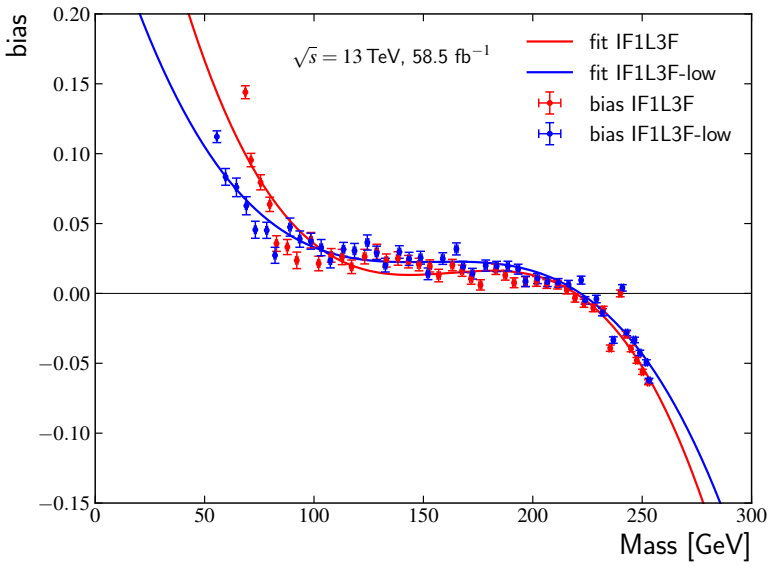


Figure 9.12: Bias of the NNs IF1L3F-low and IF1L3F in the validation set as function of the simulated mass points. The simulated data points are fitted with a polynomial of 3$^{\text{rd}}$ degree.

With the additional mass points, IF1L3F-low displays an improved bias in the low mass range. The bias at the 60 GeV mass point in the test set is improved from 13.8% to 7.4%. In addition, the polynomial fit describes the lower tail better with the added mass points than without, which

Table 9.9: Obtained parameters of the performed polynomial fit $a + b \cdot$ initial mass $+ c \cdot$ (initial mass)$^2 + d \cdot$ (initial mass)^3 for the NNs IF1L3F-low and IF1L3F.

NN	$a/10^{-1}$	$b/10^{-3}$	$c/10^{-5}$	$d/10^{-7}$
IF1L3F	4.94	-9.09	5.67	-1.16
IF1L3F-low	2.96	-5.38	3.52	-0.76

is expressed by a reduced χ^2/ndf from $1.6 \cdot 10^{-4}$ compared to $8.1 \cdot 10^{-5}$.

Table 9.10: Bias (a) and resolution (b) for the NNs IF1L3F-low and IF1L3F using data of the test set. The bias is given in % and the resolution in GeV.

(a) Bias

Mass reconstruction	[60, 100) GeV	[100, 150] GeV	(150, 200] GeV	(200, 270] GeV
IF1L3F	6.2 ± 0.2	2.0 ± 0.1	1.2 ± 0.1	-2.2 ± 0.1
IF1L3F-low	4.5 ± 0.2	3.0 ± 0.1	1.7 ± 0.1	-1.6 ± 0.1

(b) Resolution

Mass reconstruction	[60, 100) GeV	[100, 150] GeV	(150, 200] GeV	(200, 270] GeV
IF1L3F	10.6 ± 0.1	16.6 ± 0.1	19.0 ± 0.1	18.3 ± 0.1
IF1L3F-low	11.4 ± 0.1	16.7 ± 0.1	19.2 ± 0.2	18.4 ± 0.1

The obtained resolution and bias in the test set for the respective mass ranges are presented in table 9.10. For the lowest mass range an improvement in the bias by about 1.7 % is observed, however the bias deteriorates for two out of three higher mass points. The resolution worsens for all four mass ranges, in the lowest mass range by about 0.8 GeV. Including the two additional mass points to the lowest mass range, i.e. $[50, 100)$, the resolution improves to 10.8 GeV, but the bias does also increase to 5.2%.
To summarize, the additional low mass points improve the bias of the NN in the low mass range enabling predictions closer to the true mass values. However, a worse resolution is observed which denies an improved mass reconstruction. It is still unclear, why all NNs show this increasing bias effect at both edges of the mass point range and further studies are necessary to understand and reduce the bias.

9.3.3 Mass Reconstruction for Higgs Boson LFV and Z Boson LFV decays

In the LFV analysis $H(Z) \to \tau e$ and $H(Z) \to \tau \mu$ decays are studied. A good mass reconstruction of these signal processes and the SM background processes is crucial to become sensitive to a signal excess. The performance of the trained NNs and the exisiting mass reconstruction methods is evaluated for simulated Higgs LFV and Z LFV processes as well as the most dominant and irreducible $Z \to \tau\tau$ background. For the Higgs LFV processes, dependencies of the mass reconstruction methods on different production processes, decay modes and channel assignments are investigated. The performed event selection is the same as in the signal region (SR), described in chapter 7.
In case of the simulated Higgs LFV processes, the resolution and the arithmetic mean of the mass reconstruction is compared for the two dominant Higgs production processes, gluon-gluon fusion (ggH) and vector boson fusion (VBF). The results are listed in table 9.11.
As previously, all NNs perform similarly well. For all mass reconstruction methods, the Higgs mass is slightly better reconstructed for the ggH process than for the VBF process, though the

difference is mainly around 1 GeV. For the mass resolution as well, the performance is better for the ggH production than the VBF production. This difference is most striking in case of the collinear mass, where the resolution is better by 3.5 GeV in ggH than in VBF. For all other mass reconstruction methods, the resolution in ggH is better by about 1.5 GeV.

The NNs outperform both the collinear mass and the MMC_{maxw} mass reconstruction as they did on the test set, however the best performance is obtained by the MMC_{mlm} mass reconstruction. In comparison to the MMC_{mlm}, the resolution of the NNs is approximately equal. However, the means of the NN predictions are further away of the Higgs mass peak than for the MMC_{mlm}, indicating a smaller bias.

Table 9.11: Arithmetic mean and resolution of the reconstructed Higgs mass for gluon-gluon fusion and vector boson fusion. The uncertainties on the mean and resolution of all mass reconstruction techniques are below 0.05 GeV. The abbreviations are described in table 9.5.

Mass reconstruction	Gluon-gluon fusion		Vector boson fusion	
	Mean [GeV]	Resolution [GeV]	Mean [GeV]	Resolution [GeV]
IF1L2F	128.7	16.0	129.9	17.5
IF1L3F	128.5	16.0	129.8	17.6
IF1L4F	129.6	16.0	130.8	17.6
IF2L4F	129.1	16.1	130.1	17.7
IF1L3F-no m_{coll}	128.5	16.0	130.3	17.5
IF1L2F-dropout	129.1	15.9	130.3	17.4
IF1L3P	128.5	15.9	129.4	17.4
IF1L3O	129.8	15.7	130.9	17.2
IF1L3F-uncPCA	128.2	16.0	129.4	17.6
IF1L3F-redPCA	128.7	16.2	129.7	17.7
m_{coll}	130.8	19.0	134.8	22.5
MMC_{maxw}	122.2	16.7	122.4	18.1
MMC_{mlm}	126.3	15.9	126.5	17.5

The NNs are trained on $H \to \tau e$ decays. Thus, a possible difference in the mass reconstruction between τe and $\tau\mu$ decays is investigated for the ggH process. The arithmetic mean and resolution values of the mass reconstruction methods are shown in table 9.12. In general, no significant dependency of the mass reconstruction methods on the decay into a τe or a $\tau\mu$ pair is observed. Only the collinear mass shows a worse resolution in case of the $H \to \tau\mu$ decay. Neither the NNs nor the MMC variants show any significant dependency.

Furthermore, the $H \to \tau e$ is split into the $e\mu$ and μe channels, as described in section 6.2, to investigate a possible dependency of the mass reconstruction methods on the assigned channel. About 11% of all $H \to \tau e$ events end up in the 'mismatched' μe channel. In table 9.13, the arithmetic mean and resolution are displayed. The NNs and the MMC_{maxw} reconstruct the Higgs mass better for the 'mismatched' μe channel, whereas for the collinear mass and the MMC_{mlm} it is the other way around. For all reconstruction methods, except the MMC_{mlm}, the bias in the μe channel is worse than in the $e\mu$ channel. These schemes are also seen in the $H \to \tau\mu$ signal process. A NN trained on both $H \to \tau e$ and $H \to \tau\mu$ may benefit from the additional events in the μe channel.

Since the main background in the SR results from the Z boson decaying into two τ-leptons, it is also important that this peak is well resolved. The results are presented in table 9.14.

Table 9.12: Arithmetic mean and resolution of the via ggH produced Higgs boson, split into τe decays and $\tau\mu$ decays. The uncertainties on the mean and resolution of all mass reconstruction techniques are below 0.05 GeV. The abbreviations are described in table 9.5.

Mass reconstruction	$H \to \tau e$		$H \to \tau\mu$	
	Mean [GeV]	Resolution [GeV]	Mean [GeV]	Resolution [GeV]
IF1L2F	128.4	15.9	129.1	16.1
IF1L3F	128.3	16.0	128.9	16.1
IF1L4F	129.4	15.9	129.9	16.1
IF2L4F	128.9	16.0	129.4	16.1
IF1L3F-no m_{coll}	128.2	16.0	128.9	16.1
IF1L2F-dropout	128.8	15.8	129.5	16.0
IF1L3P	128.2	15.8	128.8	15.9
IF1L3O	129.5	15.7	130.1	15.8
IF1L3F-uncPCA	127.9	15.9	128.5	16.0
IF1L3F-redPCA	128.5	16.1	129.0	16.2
m_{coll}	130.4	18.7	131.3	19.3
MMC_{maxw}	122.0	16.6	122.5	16.7
MMC_{mlm}	126.0	15.9	126.7	16.0

Table 9.13: Arithmetic mean and resolution split into of the $e\mu$ and μe channel for a via ggH produced Higgs boson, decaying into a τe pair. The uncertainties on the mean and resolution in the $e\mu$ channel are for all mass reconstruction techniques below 0.05 GeV. In the μe channel, all uncertainties are 0.1 GeV. The abbreviations are described in table 9.5.

Mass reconstruction	$e\mu$ Channel (correctly matched)		μe Channel (mismatched)	
	Mean [GeV]	Resolution [GeV]	Mean [GeV]	Resolution [GeV]
IF1L2F	128.1	16.1	130.9	14.5
IF1L3F	128.0	16.1	130.1	14.5
IF1L4F	129.0	16.1	131.7	14.5
IF2L4F	128.6	16.2	130.9	14.8
IF1L3F-no m_{coll}	127.9	16.1	130.5	14.5
IF1L2F-dropout	128.5	16.0	131.5	14.4
IF1L3P	127.9	16.0	130.4	14.4
IF1L3O	129.2	15.9	131.7	14.1
IF1L3F-uncPCA	127.6	16.1	129.9	14.6
IF1L3F-redPCA	128.2	16.3	130.2	14.5
m_{coll}	129.4	17.9	137.2	22.3
MMC_{maxw}	122.2	16.6	119.8	15.9
MMC_{mlm}	126.2	15.8	124.1	16.8

The NNs have the best resolution of the Z boson mass peak, outperforming even the MMC_{mlm}. The collinear mass has the smallest bias in the mass reconstruction and the MMC_{maxw} the largest, whereas the NNs and the MMC_{mlm} have a similar bias. All but the collinear mass do underestimate the Z boson mass of 91.188 GeV. However, this is expected since all techniques but the collinear mass approximation assume a LFV decay, whereas here a decay into a τ-lepton pair is present. Furthermore, this underestimation is expected to improve the separation between the Higgs LFV signal and the Z boson background.

Table 9.14: Arithmetic mean and resolution of the mass reconstruction for the $Z \to \tau\tau$ background. The uncertainties on the mean and resolution are for all mass reconstruction techniques below 0.02 GeV. The abbreviations are described in table 9.5.

Mass reconstruction	Mean [GeV]	Resolution [GeV]
IF1L2F	87.1	19.0
IF1L3F	86.7	19.1
IF1L4F	87.8	19.3
IF2L4F	87.6	19.2
IF1L3F-no m_{coll}	86.8	19.3
IF1L2F-dropout	87.5	19.2
IF1L3P	87.0	18.9
IF1L3O	88.8	18.8
IF1L3F-decPCA	86.6	19.0
IF1L3F-redPCA	86.6	19.2
m_{coll}	93.2	22.8
MMC_{maxw}	82.2	20.1
MMC_{mlm}	86.5	19.6

In figure 9.13, the $Z \to \ell\ell$ background contributions are shown for the μe channel.

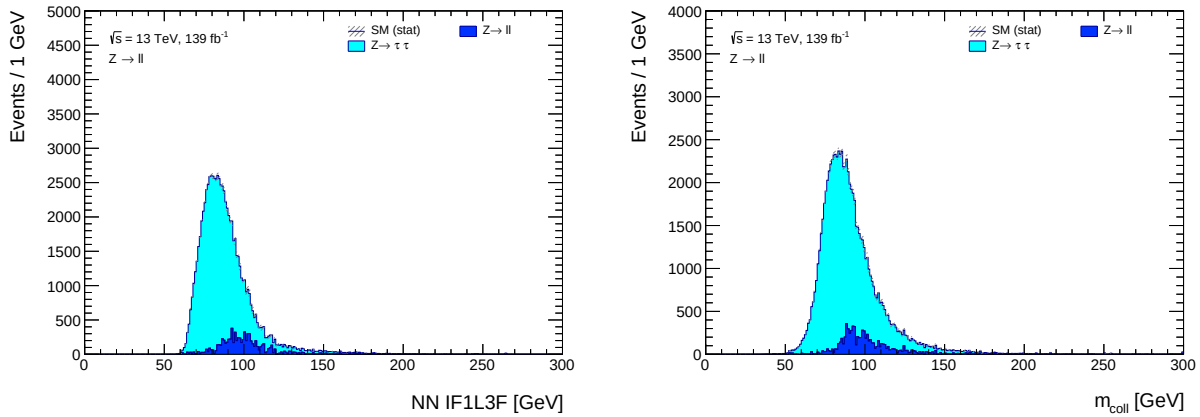


Figure 9.13: Mass reconstruction of the Z boson background in the μe channel for the NN IF1L3F (left) and the collinear mass approximation (right). The $Z \to \tau\tau$ distribution is shown in light blue and the $Z \to ee/\mu\mu$ distribution in dark blue.

The mass reconstruction methods are also applied to simulated Z boson events, where the Z boson decays as $\text{Z} \to \tau\ell$. The respective mean and resolution of all NNs is presented in table 9.15 next

to the other mass reconstruction methods. The resolution of the NNs and the MMC_{mlm} differs at most 0.3 GeV and is superior to both other mass reconstructions. Both, m_{coll} and MMC_{mlm} have a similar arithmetic mean as the NNs, however the resolution of the collinear mass approximation is by about 2.4 GeV worse. In comparison with the SM $Z \to \tau\tau$ background, the mean of the NN and the MMC_{mlm} mass reconstruction deviates slightly more from the actual Z boson mass. The MMC_{maxw} predicts the actual mass very well.

Table 9.15: Arithmetic mean and resolution of the reconstructed LFV Z boson, considering $Z \to \tau e$ and $Z \to \tau\mu$ decays. The uncertainties on the mean and resolution of all mass reconstruction techniques are 0.1 GeV. The abbreviations are described in table 9.5.

Mass reconstruction	Mean [GeV]	Resolution [GeV]
IF1L2F	96.7	13.5
IF1L3F	96.5	13.5
IF1L4F	97.8	13.5
IF2L4F	97.2	13.7
IF1L3F-no m_{coll}	96.4	13.6
IF1L2F-dropout	97.3	13.5
IF1L3P	96.8	13.5
IF1L3O	98.6	13.3
IF1L3F-decPCA	96.5	13.5
IF1L3F-redPCA	96.4	13.7
m_{coll}	98.5	16.1
MMC_{maxw}	91.7	14.4
MMC_{mlm}	96.4	13.6

9.3.4 Separation Power of Signal and Background in Higgs Boson LFV and Z Boson LFV decays

The maximum likelihood method, described in section 6.3, is used to estimate the separation power of signal and background for the different mass reconstruction methods. For this, all considered Higgs LFV signal and SM background processes are the same as in the SR as well as the performed event selection, described in chapter 7. The background includes $Z \to \ell\ell$, diboson and top quark background contributions as well as SM $H \to \tau\tau$ decays. The fake background contribution is estimated with the results of chapter 8. For $H \to \tau e$ and $H \to \tau\mu$ a branching ratio of 1% is assumed.
To evaluate the significance of a particular signal, the background-only hypothesis is tested, without considering systematic uncertainties. The expected Gaussian significance Z is calculated with equation (6.11), separately for the $e\mu$ and μe channel. The uncertainties are derived from the 95 % confidence interval of the estimated signal strength $\hat{\mu}$. This interval is given by

$$\hat{\mu} \pm 1.96 \cdot \frac{1}{\sqrt{-\frac{\partial^2 \ln L(\mu)}{\partial \mu^2}|_{\mu=\hat{\mu}}}}$$

with the logarithm of the likelihood function $\ln L$. For both, the up and down variation of the signal strength a corresponding expected significance $Z^{\pm}$ is derived. The difference between $Z^{\pm}$ and the obtained expected significance $Z(\hat{\mu})$ is indicated as uncertainty [103].
The expected significance of all trained NNs deviates at most by 0.1, shown in table 9.16 next to the other mass reconstruction methods. The found expected significances are in the range of 10 to

14 and correspond to a p-value below 10^{-22}, which strongly rejects the stated background-only hypothesis. For all mass reconstruction methods, the $H \to \tau e$ signal in the $e\mu$ channel has a higher significance than the $H \to \tau\mu$ signal in the μe channel. However, this difference is within the 95% confidence interval for all methods except the MMC_{mlm}. In addition, more τe than $\tau\mu$ events pass the event selection, resulting in higher signal statistics in the $e\mu$ channel. Within the 95% confidence interval, all NNs have the same expected significance as the collinear mass approximation, but are outperformed by the MMC variants. The expected significance of the NNs is about 25 % smaller than the expected significances of the two MMC variants.

Table 9.16: Expected significance Z for different mass reconstruction methods in the $e\mu$ and μe channels, considering signals from $H \to \tau e$ and $H \to \tau\mu$ respectively. A branching ratio of 1% for $H \to \tau e$ and $H \to \tau\mu$ is assumed. The uncertainties are obtained from the 95 % confidence interval of the estimated signal strength. The abbreviations are described in table 9.5.

Mass reconstruction	$e\mu$ channel Expected signif. Z	μe channel Expected signif. Z
IF1L2F	$10.5^{+0.7}_{-0.8}$	9.3 ± 0.8
IF1L3F	$10.5^{+0.7}_{-0.8}$	9.4 ± 0.8
IF1L4F	$10.5^{+0.7}_{-0.8}$	9.3 ± 0.8
IF2L4F	$10.5^{+0.7}_{-0.8}$	9.3 ± 0.8
IF1L3F-no m_{coll}	$10.4^{+0.7}_{-0.8}$	9.3 ± 0.8
IF1L2F-dropout	$10.5^{+0.7}_{-0.8}$	9.4 ± 0.8
IF1L3P	$10.5^{+0.7}_{-0.8}$	9.4 ± 0.8
IF1L3O	$10.5^{+0.7}_{-0.8}$	9.3 ± 0.8
IF1L3F-uncPCA	$10.5^{+0.7}_{-0.8}$	9.3 ± 0.8
IF1L3F-redPCA	$10.5^{+0.7}_{-0.8}$	9.3 ± 0.8
m_{coll}	9.6 ± 0.7	8.5 ± 0.7
MMC_{maxw}	13.5 ± 0.7	12.1 ± 0.7
MMC_{mlm}	13.8 ± 0.7	12.3 ± 0.7

To investigate the difference between the NNs and the MMC variants, the background is split into three contributions: (i) fake contributions, (ii) $Z \to \ell\ell$ and $H \to \tau\tau$ processes and (iii) diboson and top quark contributions. The expected significance is calculated considering only one background contribution in the $e\mu$ channel separately by neglecting the respective other background contributions. In table 9.17 the expected significances are presented for the individual background contributions. Since the three background contributions are individually smaller than the total background and the signal remains the same, the expected significance becomes larger. In case of the $Z \to \ell\ell$ and $H \to \tau\tau$ background contribution, the NNs performs as well as both MMC variants, whereas for the fake background it performs slightly worse. However, a large difference is visible for the diboson and top background. The MMC variants have an expected significance which is about 45 % larger than the one of the NNs.

The difference can be explained by the additional event selection for the MMC method. As described in section 6.4, using the MMC requires a selection of those events, where a solution for the system of equations is found. This results in a large suppression of the diboson and top background by over 60%, displayed in figure 9.14.

This additional selection is also applied to the events which are fed in the NN, to test its influence. The expected significances in the $e\mu$ and μe channels are presented in table 9.18. The values for

Table 9.17: Expected significance Z for all mass reconstruction methods of the $H \to \tau e$ signal decay in the $e\mu$ channel for three different background contributions. The uncertainties are obtained from the 95 % confidence interval of the estimated signal strength. The abbreviations are described in table 9.5.

Mass reconstruction	$Z \to \ell\ell$ & $H \to \tau\tau$ Expected signif. Z	Fakes Expected signif. Z	Diboson & top Expected signif. Z
IF1L2F	$28.2^{+0.7}_{-0.8}$	25.1 ± 0.2	12.6 ± 0.4
IF1L3F	$28.2^{+0.7}_{-0.8}$	25.1 ± 0.2	12.7 ± 0.4
IF1L4F	$28.1^{+0.7}_{-0.8}$	25.1 ± 0.2	12.7 ± 0.4
IF2L4F	$27.9^{+0.7}_{-0.8}$	25.0 ± 0.2	12.7 ± 0.4
IF1L3F-no m_{coll}	$28.1^{+0.7}_{-0.8}$	25.0 ± 0.2	12.6 ± 0.4
IF1L2F-dropout	$28.0^{+0.7}_{-0.8}$	25.0 ± 0.2	12.7 ± 0.4
IF1L3P	$28.2^{+0.7}_{-0.8}$	25.1 ± 0.2	12.7 ± 0.4
IF1L3O	$28.2^{+0.7}_{-0.8}$	25.0 ± 0.2	12.7 ± 0.4
IF1L3F-uncPCA	$28.1^{+0.7}_{-0.8}$	25.0 ± 0.2	12.6 ± 0.4
IF1L3F-redPCA	$28.1^{+0.7}_{-0.8}$	25.1 ± 0.2	12.6 ± 0.4
m_{coll}	22.0 ± 0.6	23.1 ± 0.2	12.3 ± 0.4
MMC_{maxw}	27.2 ± 0.7	27.7 ± 0.2	18.3 ± 0.2
MMC_{mlm}	27.9 ± 0.7	27.9 ± 0.2	18.8 ± 0.2

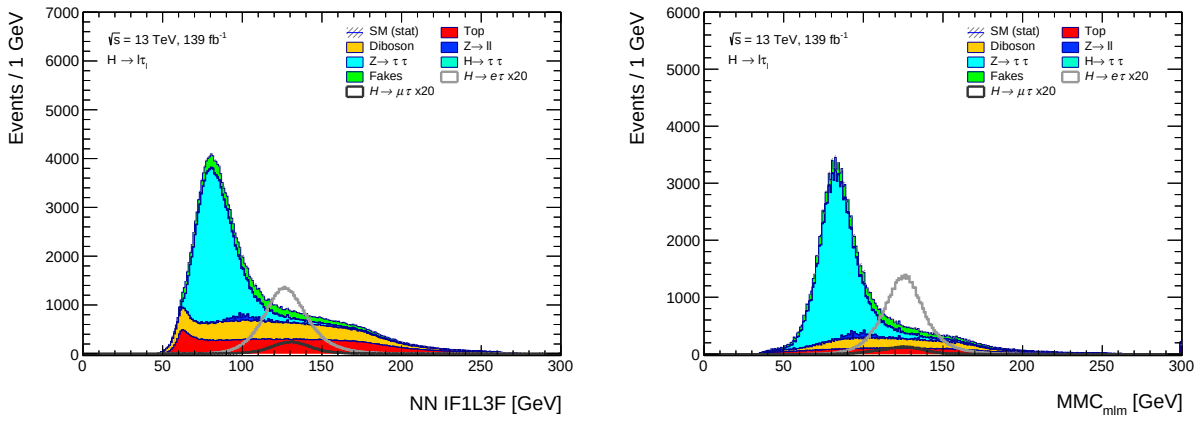


Figure 9.14: Mass reconstruction of Higgs boson LFV decays in the $e\mu$ channel for the NN IF1L3F (left) and MMC_{mlm} (right). In case of the MMC algorithm, a much smaller diboson (orange) and top (red) background contribution is visible. The LFV signal is scaled in the figure by a factor of 20.

the MMC variants are the same as before. However a clear improvement for both the NN and the collinear mass approximation is visible. With this additional event selection, the expected significance of the NNs is as well as the MMC methods.

Next to the Higgs LFV decay, the LFV decay of the Z boson is of great interest as well. The Z boson has a much larger production cross-section than the Higgs boson. However, the existing limits on the branching ratio of the LFV decay are by three orders of magnitude smaller. For the lepton-flavour violating Z boson, a branching ratio of $5 \cdot 10^{-5}$ is assumed. For the NN IF1L3F and the MMC_{maxw}, the reconstructed mass in the $e\mu$ channel is shown in figure 9.15.

Table 9.18: Expected significance Z in the $e\mu$ and μe channels for LFV $H \to \tau e$ and $H \to \tau \mu$ decays respectively. The additional event selection of the MMC algorithm is applied to all mass reconstruction methods. A branching ratio of 1% for $H \to \tau e$ and $H \to \tau \mu$ is assumed. The uncertainties are obtained from the 95 % confidence interval of the estimated signal strength. The abbreviations are described in table 9.5.

Mass reconstruction	$e\mu$ channel Expected signif. Z	μe channel Expected signif. Z
IF1L2F	13.8 ± 0.7	12.4 ± 0.7
IF1L3F	13.8 ± 0.7	12.5 ± 0.7
IF1L4F	13.8 ± 0.7	12.5 ± 0.7
IF2L4F	13.8 ± 0.7	12.6 ± 0.7
IF1L3F-no m_{coll}	13.8 ± 0.7	12.5 ± 0.7
IF1L2F-dropout	13.8 ± 0.7	12.5 ± 0.7
IF1L3P	13.8 ± 0.7	12.6 ± 0.7
IF1L3O	13.7 ± 0.7	12.6 ± 0.7
IF1L3F-uncPCA	13.8 ± 0.7	12.4 ± 0.7
IF1L3F-redPCA	13.8 ± 0.7	12.4 ± 0.7
m_{coll}	13.0 ± 0.7	11.2 ± 0.7
MMC_{maxw}	13.5 ± 0.7	12.1 ± 0.7
MMC_{mlm}	13.8 ± 0.7	12.3 ± 0.7

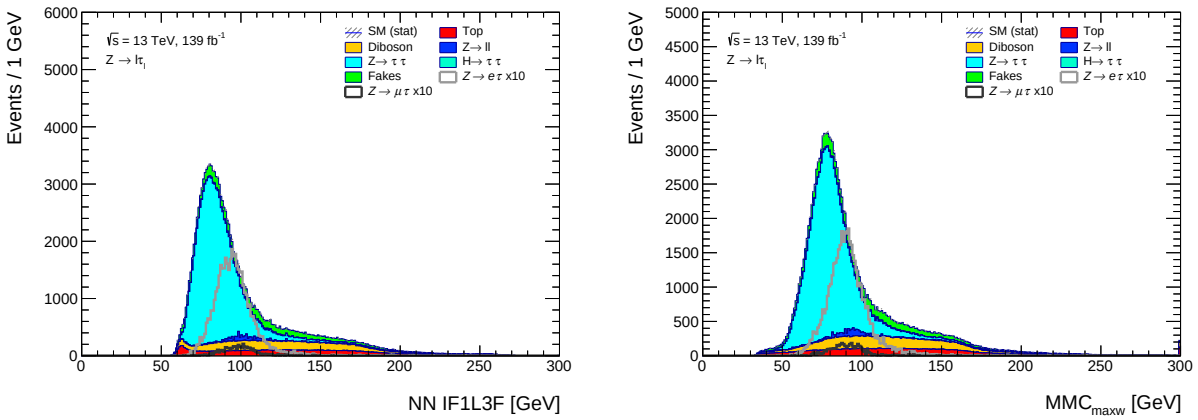


Figure 9.15: Mass reconstruction of Z boson LFV decays in the $e\mu$ channel for the NN IF1L3F (left) and MMC_{maxw} (right). The background processes are shown and the MMC event selection is applied. A branching ratio of $5 \cdot 10^{-5}$ is assumed. The LFV signal is scaled in the figure by a factor of 10.

The two signal decays $Z \to \tau e$ and $Z \to \tau \mu$ are considered. As before, the expected significance is split into the $e\mu$ and μe channels, presented in table 9.19. The additional MMC event selection is applied. In the $e\mu$ channel, the NNs have a similar expected significance as the MMC_{mlm} method, which is better than the collinear mass approximation and the MMC_{maxw} method. The NN IF2L4F has the highest expected significance in the μe channel. However, since this NN performs similar as the other NNs in all other investigations, it is assumed that the high significance is a result of an unexpected behaviour of the likelihood function. Further studies

are necessary to verify if this assumption is correct or if indeed this NN has the best separation power in the μe channel. The MMC_{maxw} method has the second highest significance in the μe channel and the other NNs perform similar to the MMC_{mlm} method.

Table 9.19: Expected significance Z for different mass reconstruction methods in the $e\mu$ and μe channels, considering $Z \to \tau e$ and $Z \to \tau \mu$ decays respectively. A branching ratio of $5 \cdot 10^{-5}$ for $Z \to \tau e$ and $Z \to \tau \mu$ is assumed. The uncertainties are obtained from the 95 % confidence interval of the estimated signal strength. The abbreviations are described in table 9.5.

Mass reconstruction	$e\mu$ channel Expected signif. Z	μe channel Expected signif. Z
IF1L2F	16.9 ± 0.2	15.4 ± 0.2
IF1L3F	17.0 ± 0.2	15.5 ± 0.2
IF1L4F	17.0 ± 0.2	15.5 ± 0.2
IF2L4F	17.0 ± 0.2	17.7 ± 0.2
IF1L3F-no m_{coll}	17.0 ± 0.2	15.5 ± 0.2
IF1L2F-dropout	17.0 ± 0.2	15.5 ± 0.2
IF1L3P	17.0 ± 0.2	15.5 ± 0.2
IF1L3O	17.0 ± 0.2	15.5 ± 0.2
IF1L3F-uncPCA	17.1 ± 0.2	15.6 ± 0.2
IF1L3F-redPCA	17.1 ± 0.2	15.5 ± 0.2
m_{coll}	16.1 ± 0.2	14.8 ± 0.2
MMC_{maxw}	16.8 ± 0.2	16.4 ± 0.2
MMC_{mlm}	17.1 ± 0.2	15.6 ± 0.2

9.3.5 Discussion

Several regression NNs were trained to reconstruct the mass of the Higgs boson in LFV decays. A comparison of the NNs to existing mass reconstruction methods was performed using data from the test set as well as for MC simulated Higgs LFV and Z LFV decays. The separation power of the signal and background processes was evaluated using the maximum likelihood method.

On the test set, all NNs outperform existing mass reconstruction methods in most cases. The resolution with respect to the MMC_{mlm} is improved by about 1 GeV in the lower mass range and up to 5.5 GeV in the high mass range. Only between 100 GeV and 150 GeV no improvement is observed. However, the bias of the NNs is large in the vicinity of the lower training mass limit, influencing the Z boson mass reconstruction. By adding more mass points below 60 GeV, the bias in the low mass range is reduced. However, this results in a worse resolution and a larger bias in all other mass ranges. Further studies are necessary to understand and reduce the bias.

All NNs are able to predict the mass of MC simulated Higgs boson as well as the MMC_{mlm} does and thus, better than the collinear mass approximation and the MMC_{maxw}. Despite being only trained on $H \to \tau e$ events, the NNs performed equally well in case of $H \to \tau \mu$ events. A neural network trained on both, $H \to \tau e$ and $H \to \tau \mu$ events, needs to be investigated and may outperform the MMC_{mlm}, due to the additional events in the μe channel. Furthermore, the NNs outperformed all other mass reconstruction methods with respect to the 'mismatched' channels, i.e. the μe channel of the $H \to \tau e$ decay and the $e\mu$ channel of the $H \to \tau \mu$ decay. In the LFV analysis, the 'mismatched' channel is used to estimate the SM background of the 'matched'

channel. Hence, an improved mass reconstruction of the 'mismatched' channel allows a better background estimation.
Both MMC variants have a higher expected significance than all NNs and the collinear mass approximation, separating the LFV signal from the SM background. However, this results mainly from the suppressed diboson and top background due to the additional event selection required for the MMCs. When applying this additional requirement also to the events reconstructed by the NNs, a similar expected significance as the MMC methods is obtained, with the difference being covered by the 95% confidence interval. Since the NNs are supposed to replace the MMCs, this event selection is not wanted. However, an additional classification NN would have to be trained to imitate this suppression of the diboson and top background.
Similar results are obtained for a lepton-flavour violating decay of a Z boson.
The improved resolution of the reconstructed mass in the mass range below 100 GeV and above 150 GeV is a promising result to use NNs. However, the polynomial dependency of the NN bias worsens the mass reconstruction. This dependency is still not fully understood and needs to be investigated further. Due to the additional event selection, which is mandatory if the MMC is used, the NNs have a worse separation power than both MMC variants. It is pointless to substitute the MMC by a NN, while using this additional selection, since this undoes the benefit of the much faster computation time the NNs have. To summarize, in the current state the MMC_{mlm} method shows the overall best performance and is recommended to used for the mass reconstruction in the LFV analysis.

10 Conclusion

In this thesis a sensitivity study to search for the LFV decays of the Higgs boson and the Z boson in the $H/Z \to \mu\tau \to \mu e 2\nu$ and $H/Z \to e\tau \to e\mu 2\nu$ decay modes has been presented. Simulated events of the ATLAS detector at a centre-of-mass energy of $\sqrt{s} = 13\,\text{TeV}$ with an integrated luminosity of $139\,\text{fb}^{-1}$ were used.
The events were split into two distinct sets, called $e\mu$ and μe, where the $e\mu$ (μe) set was enriched by events of the $H/Z \to e\tau$ ($H/Z \to \mu\tau$) decay. The dominant irreducible background process to the LFV signal is $Z \to \tau\tau$, followed by the diboson and top background processes and the fake lepton background. Small contributions result from $Z \to \ell\ell$ and $H \to \tau\tau$ decays.

The background contribution due to non-prompt leptons and particles misidentified as leptons, so-called fake leptons, was estimated with the data driven fake factor method. A dedicated Z+jets extraction region was defined by two orthogonal lepton identification requirements. The derived fake factor connects these two orthogonal regions in a linear way. Contaminations due to background processes with prompt leptons were estimated with Monte Carlo simulations and subtracted from the data.
A dependency of the fake factor on different kinematic observables was investigated. In case of a fake electron, dependencies on the transverse momentum and the pseudorapidity of the fake lepton were identified, whereas in case of a fake muon no dependency on the pseudorapidity was observed within statistical uncertainties. The estimated fake factor with these dependencies was applied as closure test in the Z+jets extraction region. A good agreement of the estimated fake background with the data could be verified for most cases. However, deviations for the distributions of the missing transverse energy as well as the impact parameters were visible. These deviations may be solved by taking into account additional dependencies of the fake factor, like the impact parameters for which a fake factor dependency was observed.

Existing mass reconstruction methods, i.e. the collinear mass approximation as well as the two variants of the missing mass calculator MMC_{maxw} and MMC_{mlm}, were compared to a newly method exploiting regression neural networks (NN). The NNs were trained on a via gluon-gluon fusion produced Higgs boson in a mass range between 60 GeV and 270 GeV with steps of 5 GeV in between. Higgs boson decays into a τe pair were considered. The architecture and hyperparameters of the NNs were optimized. For the mass reconstruction bias of the predicted mass by the NNs, a polynomial dependency on the true mass point values was observed. In the lower mass range, this dependency could be reduced by extending the mass range in the training to 50 GeV. However, this resulted in a worse resolution and bias in all other mass ranges. Further studies are necessary to understand and reduce this dependency. The NNs have a better resolution than the collinear mass approximation and the MMC_{maxw} mass reconstruction. The MMC_{mlm} mass reconstruction is able to compete with the NNs, however only in the mass range between 100 GeV and 150 GeV, where the resolutions are approximately the same. With respect to the MMC_{mlm}, the resolution was improved by the NNs in the mass range below 100 GeV by about 1 GeV. In the mass range above 150 GeV an improvement between 1 GeV and 5 GeV was achieved with respect to the existing mass reconstruction methods.

For simulated LFV decays of the Higgs boson, the mass reconstruction methods were compared for the two most important production processes, the gluon-gluon fusion (ggH) and the vector boson fusion (VBF). The reconstructed Higgs boson mass is more precisely reconstructed in case of the ggH production by at least 1.5 GeV for all reconstruction methods. For both production processes, the NNs perform similar to the MMC_{mlm} method, despite being trained only on ggH processes, which verifies a good generalisation of the training. The mass resolution of the NNs is the same for ggH produced $H \to \tau e$ and $H \to \tau\mu$ decays. This once more verifies the generalisation of the training, despite only $H \to \tau e$ decays have been used in the training processes. The $H \to \tau e$ decay mode is split further into the $e\mu$ and μe channel. Both, the NNs and the MMC_{maxw} provide a better resolution for the 'mismatched' μe channel, whereas the collinear mass and the MMC_{mlm} have a better resolution for the $e\mu$ channel. This improved resolution of the 'mismatched' channel by the NNs was also seen in the $H \to \tau\mu$ decay mode. A NN trained on both decay modes may benefit from the additional events in the μe channel.
The separation power of signal and background processes for the existing mass reconstruction methods was compared to the one of the NNs by performing a maximum likelihood fit to the reconstructed mass distribution without considering systematic uncertainties. The separation power was investigated in both channels separately for Higgs LFV decays and Z LFV decays. Taking into account all background processes, the NNs have a higher separation power than the collinear mass approximation, but are not able to obtain the same separation power as both MMC variants. However, performing the maximum likelihood fit for the individual background processes separately, by neglecting classes of background processes, showed that this is due to the contributions from the diboson and top quark background. Using the MMC requires an additional mandatory event selection which heavily reduces the diboson and top background. By applying this selection also to the events when the mass is reconstructed by the NNs, the separation power is improved significantly and the expected significance in both channels is almost the same as the one when using the MMC_{mlm}.
In general, all trained NNs are able to reconstruct the Higgs boson and Z boson mass as well as the MMC_{mlm} method and better than the MMC_{maxw} method and the collinear mass approximation. In the lower mass range, the bias of the mass reconstruction of the NNs was improved by adding additional mass points below 60 GeV. However, this resulted in a worse resolution and bias in all other mass ranges. Further studies are necessary to understand the behaviour of the NNs better. Both variants of the MMC are able to separate the LFV signal processes better from the SM background processes than the NNs. However, this is due to the additional mandatory event selection of the MMC which heavily reduces the contributions from the diboson and top background. Replacing the MMC by a NN while using this additional event selection to exploit the background suppression is pointless, since it would undo the benefit of the much faster computation time of the NN. However, a classification NN trained on this additional event selection may be used to take advantage of the background suppression, while maintaining the much faster computation time. Since the NNs were only trained on $H \to \tau e$ decays, expanding the training data by $H \to \tau\mu$ decays may improve the NNs even further.
In the current state the MMC_{mlm} mass reconstruction has the overall best performance and is recommended to be used as mass reconstruction method in the LFV analysis.

List of Figures

List of Tables

Bibliography

[1] F. Englert and R. Brout, "Broken Symmetry and the Mass of Gauge Vector Mesons," *Physical Review Letters*, vol. 13, no. 9, pp. 321–323, 1964, 10.1103/PhysRevLett.13.321.

[2] P. W. Higgs, "Broken symmetries, massless particles and gauge fields," *Physics Letters*, vol. 12, no. 2, pp. 132–133, 1964, 10.1016/0031-9163(64)91136-9.

[3] P. W. Higgs, "Broken Symmetries and the Masses of Gauge Bosons," *Physical Review Letters*, vol. 13, no. 16, pp. 508–509, 1964, 10.1103/PhysRevLett.13.508.

[4] G. S. Guralnik, C. R. Hagen, and T. W. B. Kibble, "Global Conservation Laws and Massless Particles," *Physical Review Letters*, vol. 13, no. 20, pp. 585–587, 1964, 10.1103/PhysRevLett.13.585.

[5] P. W. Higgs, "Spontaneous Symmetry Breakdown without Massless Bosons," *Physical Review*, vol. 145, no. 4, pp. 1156–1163, 1966, 10.1103/PhysRev.145.1156.

[6] T. W. B. Kibble, "Symmetry Breaking in Non-Abelian Gauge Theories," *Physical Review*, vol. 155, no. 5, pp. 1554–1561, 1967, 10.1103/PhysRev.155.1554.

[7] ATLAS Collaboration, "Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC," *Phys. Lett. B*, vol. 716, pp. 1–29, 2012, 10.1016/j.physletb.2012.08.020.

[8] CMS Collaboration, "Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC," *Physics Letters B*, vol. 716, no. 1, pp. 30–61, 2012, 10.1016/j.physletb.2012.08.021.

[9] ATLAS Collaboration and CMS Collaboration, "Combined Measurement of the Higgs Boson Mass in *pp* Collisions at $\sqrt{s} = 7$ and 8 TeV with the ATLAS and CMS Experiments," *Physical Review Letters*, vol. 114, no. 19, 2015, 10.1103/PhysRevLett.114.191803.

[10] ATLAS Collaboration and CMS Collaboration, "Measurements of the Higgs boson production and decay rates and constraints on its couplings from a combined ATLAS and CMS analysis of the LHC *pp* collision data at $\sqrt{s} = 7$ and 8 TeV," *Journal of High Energy Physics*, vol. 2016, no. 8, p. 579, 2016, 10.1007/JHEP08(2016)045.

[11] ATLAS Collaboration, "Evidence for the spin-0 nature of the Higgs boson using ATLAS data," *Phys. Lett. B*, vol. 726, pp. 120–144, 2013, 10.1016/j.physletb.2013.08.026.

[12] J. D. Bjorken and S. Weinberg, "Mechanism for Nonconservation of Muon Number," *Physical Review Letters*, vol. 38, no. 12, pp. 622–625, 1977, 10.1103/PhysRevLett.38.622.

[13] J. L. Diaz-Cruz and J. J. Toscano, "Probing lepton flavour violation with Higgs boson decays H-> li+lj," *Physical Review D*, vol. 62, no. 11, p. 1, 2000, 10.1103/PhysRevD.62.116005.

[14] M. Arana-Catania, E. Arganda, and M. J. Herrero, "Non-decoupling SUSY in LFV Higgs decays: A window to new physics at the LHC," *Journal of High Energy Physics*, vol. 2013, no. 9, p. 036, 2013, 10.1007/JHEP09(2013)160.

[15] A. Arhrib, Y. Cheng, and O. C. Kong, "Comprehensive analysis on lepton flavor violating Higgs boson to $\mu^{\mp}\tau^{\pm}$ decay in supersymmetry without R parity," *Physical Review D*, vol. 87, no. 1, 2013, 10.1103/PhysRevD.87.015025.

[16] K. Agashe and R. Contino, "Composite Higgs-Mediated FCNC," *Physical Review D*, vol. 80, no. 7, p. 231, 2009, 10.1103/PhysRevD.80.075016.

[17] A. Azatov, M. Toharia, and L. Zhu, "Higgs Mediated FCNC's in Warped Extra Dimensions," *Physical Review D*, vol. 80, no. 3, p. 1, 2009, 10.1103/PhysRevD.80.035016.

[18] H. Ishimori *et al.*, "Non-Abelian Discrete Symmetries in Particle Physics," *Progress of Theoretical Physics Supplement*, vol. 183, pp. 1–163, 2010, 10.1143/PTPS.183.1.

[19] G. Perez and L. Randall, "Natural Neutrino Masses and Mixings from Warped Geometry," *Journal of High Energy Physics*, vol. 2009, no. 01, p. 077, 2009, 10.1088/1126-6708/2009/01/077.

[20] M. Blanke *et al.*, "Delta F=2 Observables and Fine-Tuning in a Warped Extra Dimension with Custodial Protection," *Journal of High Energy Physics*, vol. 2009, no. 03, p. 001, 2009, 10.1088/1126-6708/2009/03/001.

[21] G. F. Giudice and O. Lebedev, "Higgs-dependent Yukawa couplings," *Physics Letters B*, vol. 665, no. 2-3, pp. 79–85, 2008, 10.1016/j.physletb.2008.05.062.

[22] J. A. Aguilar-Saavedra, "A minimal set of top-Higgs anomalous couplings," *Nuclear Physics B*, vol. 821, no. 1-2, pp. 215–227, 2009, 10.1016/j.nuclphysb.2009.06.022.

[23] M. E. Albrecht *et al.*, "Electroweak and Flavour Structure of a Warped Extra Dimension with Custodial Protection," *Journal of High Energy Physics*, vol. 2009, no. 09, p. 064, 2009, 10.1088/1126-6708/2009/09/064.

[24] A. Goudelis, O. Lebedev, and J.-h. Park, "Higgs-induced lepton flavor violation," *Physics Letters B*, vol. 707, no. 3-4, pp. 369–374, 2012, 10.1016/j.physletb.2011.12.059.

[25] D. McKeen, M. Pospelov, and A. Ritz, "Modified Higgs branching ratios versus CP and lepton flavor violation," *Physical Review D*, vol. 86, no. 11, 2012, 10.1103/PhysRevD.86.113004.

[26] A. Crivellin, G. D'Ambrosio, and J. Heeck, "Addressing the LHC flavour anomalies with horizontal gauge symmetries," *Physical Review D*, vol. 91, no. 7, 2015, 10.1103/PhysRevD.91.075006.

[27] A. Crivellin, G. D'Ambrosio and J. Heeck, "Explaining $h \to \mu^{\pm}\tau^{\mp}$, $B \to K^*\mu^+\mu^-$ and $B \to K\mu^+\mu^-/B \to Ke^+e^-$ in a two-Higgs-doublet model with gauged $L_\mu - L_\tau$," *Physical Review Letters*, vol. 114, no. 15, p. 361, 2015, 10.1103/PhysRevLett.114.151801.

[28] F. Gabbiani, J. H. Kim, and A. Masiero, "$Z^0 \to b\bar{s}$ and $Z^0 \to \tau\bar{\mu}$ in SUSY: Are They Observable?" *Phys. Lett. B*, vol. 214, pp. 398–402, 1988, 10.1016/0370-2693(88)91384-6.

[29] T.-K. Kuo and N. Nakagawa, "Lepton Flavor Violating Decays of Z^0 and τ," *Physical Review D*, vol. 32, p. 306, 1985, 10.1103/PhysRevD.32.306.

[30] J. I. Illana and T. Riemann, "Charged lepton flavor violation from massive neutrinos in Z decays," *Physical Review D*, vol. 63, p. 053004, 2001, 10.1103/PhysRevD.63.053004.

[31] The Super-Kamiokande Collaboration, "Evidence for oscillation of atmospheric neutrinos," *Physical Review Letters*, vol. 81, no. 8, pp. 1562–1567, 1998, 10.1103/PhysRevLett.81.1562.

[32] R. Harnik, J. Kopp, and J. Zupan, "Flavor Violating Higgs Decays," *JHEP*, vol. 03, p. 026, 2013, 10.1007/JHEP03(2013)026.

[33] G. Blankenburg, J. Ellis, and G. Isidori, "Flavour-Changing Decays of a 125 GeV Higgs-like Particle," *Phys. Lett. B*, vol. 712, pp. 386–390, 2012, 10.1016/j.physletb.2012.05.007.

[34] ATLAS Collaboration, "Search for lepton-flavour-violating decays of the Higgs and Z bosons with the ATLAS detector," *The European physical journal. C, Particles and fields*, vol. 77, no. 2, p. 70, 2017, 10.1140/epjc/s10052-017-4624-0.

[35] M. Tanabashi *et al.*, "Review of Particle Physics," *Physical Review D*, vol. 98, no. 3, p. 030001, 2018, 10.1103/PhysRevD.98.030001.

[36] S. Dittmaier and M. Schumacher, "The Higgs Boson in the Standard Model - From LEP to LHC: Expectations, Searches, and Discovery of a Candidate," *Progress in Particle and Nuclear Physics*, vol. 70, pp. 1–54, 2013, 10.1016/j.ppnp.2013.02.001.

[37] P. B. Pal, "An Introductory Course of Particle Physics," Hoboken: CRC Press, 2015.

[38] N. Cabibbo, "Unitary Symmetry and Leptonic Decays," *Physical Review Letters*, vol. 10, no. 12, pp. 531–533, 1963, 10.1103/PhysRevLett.10.531.

[39] M. Kobayashi and T. Maskawa, "CP -Violation in the Renormalizable Theory of Weak Interaction," *Progress of Theoretical Physics*, vol. 49, no. 2, pp. 652–657, 1973, 10.1143/PTP.49.652.

[40] C. Schillo, "Search for the Standard Model Higgs Boson and Test of CP Invariance in Vector-Boson Fusion Production of the Higgs Boson in the Fully Leptonic $H \to \tau^+\tau^- \to \ell^+\ell^- 4\nu$ Final State in Proton-Proton Collisions with the ATLAS Detector at the LHC," Dissertation, Universität, Freiburg, 2016, 10.6094/UNIFR/11333.

[41] ATLAS Collaboration, "Measurements of the Higgs boson production, fiducial and differential cross sections in the 4ℓ decay channel at $\sqrt{s} = 13$ TeV with the ATLAS detector," 2018, ATLAS-CONF-2018-018.

[42] D. d. Florian *et al.*, "Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector", 2016, 10.23731/CYRM-2017-002.

[43] CMS Collaboration, "Measurements of properties of the Higgs boson decaying into the four-lepton final state in pp collisions at $\sqrt{s} = 13$ TeV," *JHEP*, vol. 11, p. 047, 2017, 10.1007/JHEP11(2017)047.

[44] ATLAS Collaboration, "Measurement of the Higgs boson mass in the $H \to ZZ^* \to 4\ell$ and $H \to \gamma\gamma$ channels with $\sqrt{s} = 13$ TeV pp collisions using the ATLAS detector," *Phys. Lett. B*, vol. 784, pp. 345–366, 2018, 10.1016/j.physletb.2018.07.050.

[45] CMS Collaboration, "Study of the Mass and Spin-Parity of the Higgs Boson Candidate Via Its Decays to Z Boson Pairs," *Physical Review Letters*, vol. 110, no. 8, 2013, 10.1103/PhysRevLett.110.081803.

[46] L. Evans and P. Bryant, "LHC Machine," *Journal of Instrumentation*, vol. 3, no. 08, pp. S08 001–S08 001, 2008, 10.1088/1748-0221/3/08/S08001.

[47] M. Hostettler and G. Papotti, "Luminosity Lifetime at the LHC in 2012 Proton Physics Operation," in *4th International Particle Accelerator Conference*, 2013, p. TUPFI029, s3.cern.ch/inspire-prod-files-b/.

[48] Sarah Charley, "The LHC does a dry run," 2015, symmetrymagazine.org.

[49] ATLAS Collaboration, "Luminosity determination in pp collisions at $\sqrt{s} = 13$ TeV using the ATLAS detector at the LHC," 2019, ATLAS-CONF-2019-021.

[50] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider," *JINST*, vol. 3, p. S08003, 2008, 10.1088/1748-0221/3/08/S08003.

[51] ATLAS Collaboration, "Performance of the ATLAS Trigger System in 2010," *The European physical journal. C, Particles and fields*, vol. 72, no. 1, p. S08003, 2012, 10.1140/epjc/s10052-011-1849-1.

[52] A. R. Martínez, "The Run-2 ATLAS Trigger System," *Journal of Physics: Conference Series*, vol. 762, p. 012003, 2016, 10.1088/1742-6596/762/1/012003.

[53] ATLAS Collaboration, "Measurement of the tau lepton reconstruction and identification performance in the ATLAS experiment using *pp* collisions at $\sqrt{s} = 13$ TeV," 2017, ATLAS-CONF-2017-029.

[54] ATLAS Collaboration, "Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton-proton collision data at $\sqrt{s} = 13$ TeV," *The European physical journal. C, Particles and fields*, vol. 79, no. 8, p. T05008, 2019, 10.1140/epjc/s10052-019-7140-6.

[55] "Tracking-CP analysis guidelines for DC14," twiki.cern.ch/twiki/InDetTrackingDC14, 07.05.2020.

[56] ATLAS Collaboration, "Muon reconstruction performance of the ATLAS detector in proton-proton collision data at $\sqrt{s}$=13 TeV," *The European physical journal. C, Particles and fields*, vol. 76, no. 5, p. 043, 2016, 10.1140/epjc/s10052-016-4120-y.

[57] M. Cacciari, G. P. Salam, and G. Soyez, "The anti-k_t jet clustering algorithm," *Journal of High Energy Physics*, vol. 2008, no. 04, p. 063, 2008, 10.1088/1126-6708/2008/04/063.

[58] ATLAS Collaboration, "Jet reconstruction and performance using particle flow with the ATLAS Detector," *Eur. Phys. J. C*, vol. 77, no. 7, p. 466, 2017, 10.1140/epjc/s10052-017-5031-2.

[59] ATLAS Collaboration, "Measurements of *b*-jet tagging efficiency with the ATLAS detector using $t\bar{t}$ events at $\sqrt{s} = 13$ TeV," *JHEP*, vol. 08, p. 089, 2018, 10.1007/JHEP08(2018)089.

[60] ATLAS Collaboration, "Tagging and suppression of pileup jets," 2014, ATLAS-CONF-2014-018.

[61] ATLAS Collaboration, "Performance of missing transverse momentum reconstruction with the ATLAS detector using proton-proton collisions at $\sqrt{s} = 13$ TeV," *The European physical journal. C, Particles and fields*, vol. 78, no. 11, p. 063, 2018, 10.1140/epjc/s10052-018-6288-9.

[62] S. Schramm, "Machine learning at CERN: ATLAS, LHCb, and more," in *Proceedings of The 39th International Conference on High Energy Physics — PoS(ICHEP2018)*, U.-k. Yang, J. E. Kim, and Y. Kim, Eds. Trieste, Italy: Sissa Medialab, 2019, p. 158, 10.22323/1.340.0158.

[63] Y. Bengio, "Practical Recommendations for Gradient-Based Training of Deep Architectures," 2012, arXiv:1206.5533v2.

[64] X. Glorot, A. Bordes, and Y. Bengio, "Deep Sparse Rectifier Neural Networks," in *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, G. Gordon, D. Dunson, and M. Dudík, Eds., vol. 15. Fort Lauderdale, FL, USA: PMLR, 2011, pp. 315–323, proceedings.mlr.press.

[65] F. Bre, J. M. Gimenez, and V. D. Fachinotti, "Prediction of wind pressure coefficients on building surfaces using artificial neural networks," *Energy and Buildings*, vol. 158, pp. 1429–1441, 2018, 10.1016/j.enbuild.2017.11.045.

[66] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Representations by Back-propagating Errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986, 10.1038/323533a0.

[67] I. Goodfellow, Y. Bengio, and A. Courville, "Deep Learning," Cambridge, Massachusetts and London, England: MIT Press, 2016, deeplearningbook.org.

[68] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2014, arXiv:1412.6980v9.

[69] D. R. Wilson and T. R. Martinez, "The General Inefficiency of Batch Training for Gradient Descent Learning," *Neural Netw*, vol. 16, no. 10, pp. 1429–1451, 2003, 10.1016/S0893-6080(03)00138-2.

[70] N. Srivastava *et al.*, "Dropout: A Simple Way to Prevent Neural Networks from Overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014, 10.5555/2627435.2670313.

[71] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient BackProp," in *Neural Networks: Tricks of the Trade, This Book is an Outgrowth of a 1996 NIPS Workshop*. Berlin, Heidelberg: Springer-Verlag, 1998, pp. 9–50, yann.lecun.com.

[72] J. Bergstra and Y. Bengio, "Random Search for Hyper-Parameter Optimization," *Journal of machine learning research*, vol. 13, no. Feb, pp. 281–305, 2012, jmlr.org/papers/volume13/bergstra12a.

[73] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 2010, 10.1080/14786440109462720.

[74] S. Wold, K. Esbensen, and P. Geladi, "Principal Component Analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37–52, 1987, 10.1016/0169-7439(87)80084-9.

[75] S. Bressler, A. Dery, and A. Efrati, "Asymmetric lepton-flavor violating Higgs decays," *Physical Review D*, vol. 90, no. 1, 2014, 10.1103/PhysRevD.90.015025.

[76] G. Cowan, K. Cranmer, E. Gross, and O. Vitells, "Asymptotic formulae for likelihood-based tests of new physics," *The European physical journal. C, Particles and fields*, vol. 71, no. 2, p. 60, 2011, 10.1140/epjc/s10052-011-1554-0.

[77] J. Neyman and E. S. Pearson, "On the Problem of the Most Efficient Tests of Statistical Hypotheses," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, vol. 231, pp. 289–337, 1933, jstor.org/stable/91247.

[78] R. Ellis, I. Hinchliffe, M. Soldate, and J. J. van der Bij, "Higgs Decay to $\tau^+\tau^-$: A Possible Signature of Intermediate Mass Higgs Bosons at the SSC," *Nucl.\ Phys.\ B*, vol. 297, pp. 221–243, 1988, 10.1016/0550-3213(88)90019-3.

[79] A. Elagin, P. Murat, A. Pranko, and A. Safonov, "A New Mass Reconstruction Technique for Resonances Decaying to di-tau," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 654, no. 1, pp. 481–489, 2011, 10.1016/j.nima.2011.07.009.

[80] S. Agostinelli *et al.*, "Geant4—a simulation toolkit," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 506, no. 3, pp. 250–303, 2003, 10.1016/S0168-9002(03)01368-8.

[81] ATLAS Collaboration, "The ATLAS Simulation Infrastructure," *Eur. Phys. J. C*, vol. 70, pp. 823–874, 2010, 10.1140/epjc/s10052-010-1429-9.

[82] S. Frixione, P. Nason, and C. Oleari, "Matching NLO QCD computations with Parton Shower simulations: The POWHEG method," *Journal of High Energy Physics*, vol. 2007, no. 11, p. 070, 2007, 10.1088/1126-6708/2007/11/070.

[83] T. Sjöstrand *et al.*, "An Introduction to PYTHIA 8.2," *Computer Physics Communications*, vol. 191, pp. 159–177, 2015, 10.1016/j.cpc.2015.01.024.

[84] J. Butterworth *et al.*, "PDF4LHC recommendations for LHC Run II," *Journal of Physics G: Nuclear and Particle Physics*, vol. 43, no. 2, p. 023001, 2016, 10.1088/0954-3899/43/2/023001.

[85] J. Botts *et al.*, "CTEQ Parton Distributions and Flavor Dependence of Sea Quarks," *Physics Letters B*, vol. 304, no. 1-2, pp. 159–166, 1993, 10.1016/0370-2693(93)91416-K.

[86] J. Pumplin *et al.*, "New Generation of Parton Distributions with Uncertainties from Global QCD Analysis," *Journal of High Energy Physics*, vol. 2002, no. 07, p. 012, 2002,10.1088/1126-6708/2002/07/012.

[87] K. Hamilton, P. Nason, E. Re, and G. Zanderighi, "NNLOPS simulation of Higgs boson production," *Journal of High Energy Physics*, vol. 2013, no. 10, p. 1, 2013, 10.1007/JHEP10(2013)222.

[88] K. Hamilton, P. Nason, and G. Zanderighi, "MINLO: Multi-scale improved NLO," *Journal of High Energy Physics*, vol. 2012, no. 10, p. 2916, 2012, 10.1007/JHEP10(2012)155.

[89] D. Lange, "The EvtGen particle decay simulation package," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 462, pp. 152–155, 2001, 10.1016/S0168-9002(01)00089-4.

[90] CMS Collaboration, "Measurement of the $Z\gamma^* \to \tau\tau$ cross section in pp collisions at $\sqrt{s} =$ 13 TeV and validation of τ lepton analysis techniques," *The European physical journal. C, Particles and fields*, vol. 78, no. 9, p. 1, 2018, 10.1140/epjc/s10052-018-6146-9.

[91] T. Gleisberg *et al.*, "Event generation with SHERPA 1.1," *Journal of High Energy Physics*, vol. 2009, no. 02, p. 007, 2009, 10.1088/1126-6708/2009/02/007.

[92] The NNPDF Collaboration, "Parton distributions for the LHC Run II," *Journal of High Energy Physics*, vol. 2015, no. 4, p. 21, 2015, 10.1007/JHEP04(2015)040.

[93] ATLAS Collaboration, "Performance of the ATLAS Inner Detector Track and Vertex Reconstruction in the High Pile-Up LHC Environment," 2012, ATLAS-CONF-2012-042.

[94] ATLAS Collaboration, "Search for anomalous production of prompt same-sign lepton pairs and pair-produced doubly charged Higgs bosons with $\sqrt{s} = 8$ TeV pp collisions using the ATLAS detector," *Journal of High Energy Physics*, vol. 2015, no. 3, p. 349, 2015, 10.1007/JHEP03(2015)041.

[95] J. Alison, "The Road to Discovery: Detector Alignment, Electron Identification, Particle Misidentification, WW Physics, and the Discovery of the Higgs Boson," Dissertation, 2015, 10.1007/978-3-319-10344-0.

[96] ATLAS Collaboration, "Measurement of the $W^{\pm}Z$ boson pair-production cross section in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS Detector," *Physics Letters B*, vol. 762, pp. 1–22, 2016, 10.1016/j.physletb.2016.08.052.

[97] ATLAS Collaboration, "Multi-Boson Simulation for 13 TeV ATLAS Analyses," 2017, ATL-PHYS-PUB-2017-005.

[98] P. Bärtschi, C. Galloni, C. Lange, and B. Kilminster, "Reconstruction of τ lepton pair invariant mass using an artificial neural network," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 929, pp. 29–33, 2019., 10.1016/j.nima.2019.03.029

[99] F. Chollet *et al.*, "Keras," 2015, keras.io.

[100] Martín Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems," 2015, tensorflow.org.

[101] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring Strategies for Training Deep Neural Networks," *J. Mach. Learn. Res.*, vol. 10, pp. 1–40, 2009, jmlr.org/papers/volume10/larochelle09a.

[102] ATLAS Collaboration, "Object-based missing transverse momentum significance in the ATLAS detector," 2018, ATLAS-CONF-2018-038.

[103] G. Cowan, "Statistical data analysis," ser. Oxford science publications. Oxford: Clarendon Press, 1998, physics.mff.cuni.cz.