Higgs Boson Cross Section Measurement in the W-Associated Production and Fully Leptonic Final State at $\sqrt{s}$ = 13 TeV with the ATLAS Experiment using Deep Neural Networks

by
Thomas Honig

JOHANNES GUTENBERG UNIVERSITÄT MAINZ

Masterarbeit in Physik
vorgelegt dem Fachbereich Physik, Mathematik und Informatik (FB 08)
der Johannes Gutenberg-Universität Mainz
am 28.12.2020

1. Gutachter: Prof. Dr. Volker Büscher
2. Gutachter: Prof. Dr. Matthias Schott

Ich versichere, dass ich die Arbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt, sowie Zitate kenntlich gemacht habe.

Thomas Honig
Mainz, den 28.12.2020

Thomas Honig
ETAP - Experimentelle Teilchen- und Astroteilchen Physik
Institut für Physik
Staudingerweg 7
Johannes Gutenberg-Universität D-55099 Mainz
thonig@students.uni-mainz.de

Abstract

In this thesis, a neural network-based multivariate analysis is presented for the cross section measurement of the associated Higgs production channel

$$pp \to W \to WH \to WWW^* \to 3\ell\, 3\nu$$

with the Higgs boson decaying into two W bosons and a fully leptonic final state. The analysis focuses on the most sensitive signal region of a leptonic final state without lepton pairs of the same lepton flavour and different electric charge. The data set that is used comprises 139 fb^{-1} of proton-proton collisions collected at the ATLAS detector at a center of mass energy of $\sqrt{s}$ = 13 TeV. In the analysis a multiclass deep neural network is employed to suppress the most dominant backgrounds from top and WZ decays. The expected significance is measured to be Z_0 = 2.93.

Table of Contents

List of Abbreviations

ATLAS	A Toroidal LHC ApparatuS
ALICE	A Large Ion Collider Experiment
API	Application Programming Interface
ANN	Artificial Neural Network
AUC	Area Under the Curve
BDT	Boosted Decision Tree
CERN	Conseil Européen pour la Recherche Nucléaire
CKM-Matrix	Cabibbo-Kobayashi-Maskawa-Matrix
CMS	Compact Muon Solenoid
CR	Control Region
CSC	Cathode Strip Chambers
DNN	Deep Neural Network
ECal	Electromagnetic Calorimeter
FCAL	Forward CALorimeter
ggF	gluon-gluon Fusion
HCal	Hadronic Calorimeter
HEP	High Energy Physics
HLT	High Level Trigger
IBL	Insertable B-Layer
ID	Inner Detector
LEIR	Low Energy Ion Ring
LHC	Large Hadron Collider
LHCb	Large Hadron Collider beauty
LHCf	Large Hadron Collider forward
LINAC	Linear Particle Accelerator
MC	Monte Carlo
MDT	Monitored Drift Tubes
MoEDAL	Monopole and Exotics Detector At the Lhc
MS	Muon Spectrometer

MVA	Multivariate Analysis
PDF	Parton Distribution Function
PS	Proton Synchrotron
PSB	Proton Synchrotron Booster
QED	Quatum Electro Dynamics
QFT	Quantum Field Theory
ROC	Receiver Operating Characteristic
RPC	Radiative Plate Chamber
SCT	SemiConductor Tracker
SFOS	Same Flavour Opposite Sign
SPS	Super Proton Synchrotron
SR	Signal Region
STXS	Simplified Template Cross Section
TGC	Thing Gap Chamber
TRT	Transition Radiation Tracker
TOTEM	TOTal Elastic and diffractive cross section Measurement
VBF	Vector Boson Fusion

1 | Introduction and Motivation

Modern particle physics is an essential pillar of our understanding of the fundamental laws of nature. It deals with the building blocks of matter and radiation, so called elementary particles, and the interactions between them, also referred to as forces and further aims to understand, to explain and to predict nature on a fundamental level. Over the last decades the field developed a theory called the Standard Model of particle physics. This model provides our current understanding of the interaction of aforementioned elementary particles through mediators which are themselves particles. As a theory, the Standard Model is remarkably successful being compatible with a huge range of experimental observations.

The latest milestone and last remaining puzzle in the completion of this theory is the Higgs Boson which has been discovered by the ATLAS[1] and CMS[2] experiments at the Large Hadron Collider in 2012. Its mass has been measured to be 125.10 ± 0.14 $\mathbf{GeV}$[1] and the spin has been determined to be 0.[3] This was the result of a long and exhaustive experimental search that had its motivational origins in the 1960s. Along the development of a theoretical backbone of particle physics based on the Quantum Field Theory (QFT), in 1964 Higgs[4], Brout and Englert[5] theorized the existence of a massive scalar elementary particle. The so called Higgs boson was expected to arise from a mechanism by which all Standard Model particles acquire their masses.

Despite the huge predictive power of the Standard Model, it is not capable of explaining e.g. the nature of dark matter whose existence has been implied by gravitational effects observed in space. If deviations between Standard Model predictions and experimental results would be observed in e.g. analyses including the Higgs boson, this would open the door to physics beyond the Standard Model. Thus, a precise and systematic measurement of the coupling to other Standard Model particles seems a promising path towards a broader understanding of Standard Model particle physics and beyond. The masses of particles such as the weak gauge bosons $W^{\pm}$ and $Z^{(0)}$ are given by the coupling to the Higgs field. Hence, measuring the coupling of e.g. Higgs and $W^{\pm}$ is expected to gain further insight into the mechanism's nature. A suitable physical decay channel is the associated Higgs production with a W boson radiating a Higgs boson and a subsequent decay of the Higgs boson into a $W^{+}W^{-}$ pair since the Higgs boson couples only to W bosons in this case.

Chapter 2 and 3 summarize the theoretical foundations and the ATLAS experiment at the LHC followed by brief introduction to neural networks and related concepts in chapter 4. Chapter 5 presents the analysis in the $WH \to WWW^{*} \to 3\ell\, 3\nu$ channel and chapter 6 gives a summary with the main results as well as an outlook.

[1]This thesis follows the convention of natural units meaning that $\hbar = c = 1$.

2 | Theoretical Foundations

In the following, a brief introduction into the relevant theoretical foundation is given mainly following Thomson et al.[6], Povh et al.[7] and Meschede et al.[8].

2.1 The Standard Model of Particle Physics

The Standard Model of Particle Physics (SM) is the current most accurate theory for the description of experimental data from the field of High Energy Physics (HEP). The theoretical foundation as indicated in Figure 2.1 can be based on three theoretical pillars:
First, Quantum Field Theory (QFT) provides a mathematical representation of particles as well as their interactions.[1] Their dynamics can be expressed in QFT-compatible equations of motion like the Dirac, Klein-Gordon or Proca equation.
Second, the local gauge principle which uses a group symmetry approach to obtain the exact form of the particles' interaction term in this theory.
Third, the Higgs mechanism that explains the generation mechanism of particle masses.
The particles and their interactions, namely electromagnetic, weak and strong, as well as the Higgs mechanism are described in more detail in chapters 2.1.1 and 2.1.2 respectively.

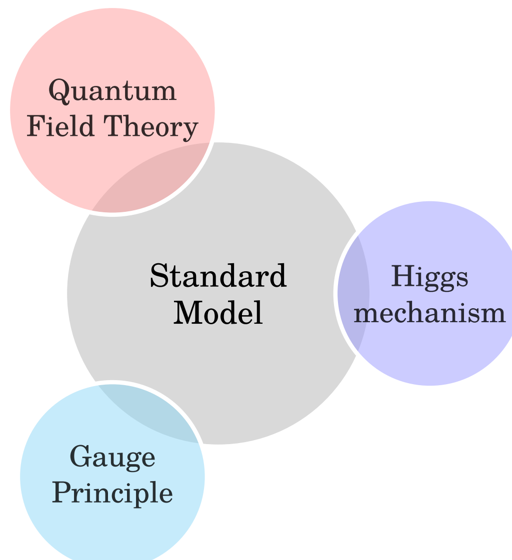


Figure 2.1: Theoretical fundamental ingredients of the Standard Model. The list of ingredients would be completed under consideration of the equally important experimental data and tests which are omitted for the sake of a theory chapter.

While the gravitational force is technically the fourth fundamental form of interactions, it is still a current topic of research to connect it to the Standard Model theory. Additionally, it can be omitted in a high energy physics study like this due to its low relative strength compared to the three aforementioned forces as illustrated in Table 2.1.

[1] From hereon the term "interaction" and the term "force" are used interchangeably as commonly done in particle physics

Interaction	Relative Strength	Couples to	Range [m]
strong	1	color	$\sim 10^{-15}$
electromagnetic	10^{-2}	electric charge	∞
weak	10^{-15}	weak charge	$< 10^{-15}$
gravitational	10^{-41}	mass	∞

Table 2.1: Fundamental Forces ordered by relative strength normalized to the strong force.[8]

2.1.1 The Standard Model Particles

The particles of the SM can be grouped into fermions which are particles of half odd integer spin and bosons whose spin is of integer nature. While the fermionic quarks and leptons are the building blocks of matter, bosons act as mediators in interactions between elementary particles. During the successive discovery of these particles a certain order has been established: three generations of fermions have been defined. The first generation consists of the lightest and stable fermions (u, d, e, ν_e), while generation two (c, s, μ, ν_μ) and three (t, b, τ, ν_τ) include the heavier, unstable fermions in ascending mass order. Additionally, every particle has its antiparticle that has the same mass but opposite quantum numbers, like charge and spin.
The upcoming section will deal with fermions and leptons in more detail. A compact representation of the particles is given in Figure 2.2.

Leptons and Quarks

On one hand, leptons can be distinguished between the charged species e^-, μ^-, τ^- and their electrically neutral partners ν_e, ν_μ, ν_τ. The antiparticles postulated by the Dirac equation have opposite quantum numbers and are denoted by their charge or by a bar in case of neutrinos ($e^- \rightarrow e^+$, $\nu_e \rightarrow \bar{\nu}_e$). The Standard Model assumes neutrinos to be massless. However, experiments have observed neutrino flavor oscillations which imply that neutrinos are massive particles while the exact physical mass values are still a current subject of research.
Quarks on the other hand can be distinguished between up-type quarks (u, c and t) of electrical charge $+\frac{2}{3}$ e and down-type quarks (d, s, b) of charge $-\frac{1}{3}$ e. Their corresponding antiparticles are denoted by a bar ($u \rightarrow \bar{u}$) with the distinction that anti-up-type quarks have charge $-\frac{2}{3}$ e and anti-down-type quarks have charge $+\frac{1}{3}$ e.

All 12 mentioned fermions are subject to the weak interaction. Furthermore, all 9 charged fermions participate in the electromagnetic interaction. Additionally, quarks carry a color charge. Hence, only quarks are subject to the strong interaction. These particles have never been observed as free particles but only in colorless bound states called hadrons.[6]

Bosons

The here discussed bosons act as the mediators of the different interactions. Each of these so called force carriers has spin 1. For the weak interaction there exist three massive mediators, the charged $W^{\pm}$ bosons and the neutral Z^0 boson. The short interaction range given in Table 2.1 can be explained by the high mass of $W$and Z (compare Figure 2.2). The photon γ is the mass- and chargeless equivalent for the electromagnetic interaction. It has an infinite interaction range. Eight massless and electrically neutral gluons g form the group of mediators of the strong interaction. Because they carry a color charge similar to quarks, they can interact with each other. While they should have an infinite interaction range just like the photon, it is this very self interaction exchanging colors that restricts the gluons to the low interaction range reported in Table 2.1.

The remaining boson from the SM given in Figure 2.2 is the Higgs boson H. The mechanism that generates the mass of every SM particle is going to be explained in more detail in the upcoming sections. In contrast to the other bosons, H has spin 0. Therefore one refers to it as a scalar boson while W, Z, γ and g are referred to as vector bosons of non-zero spin.

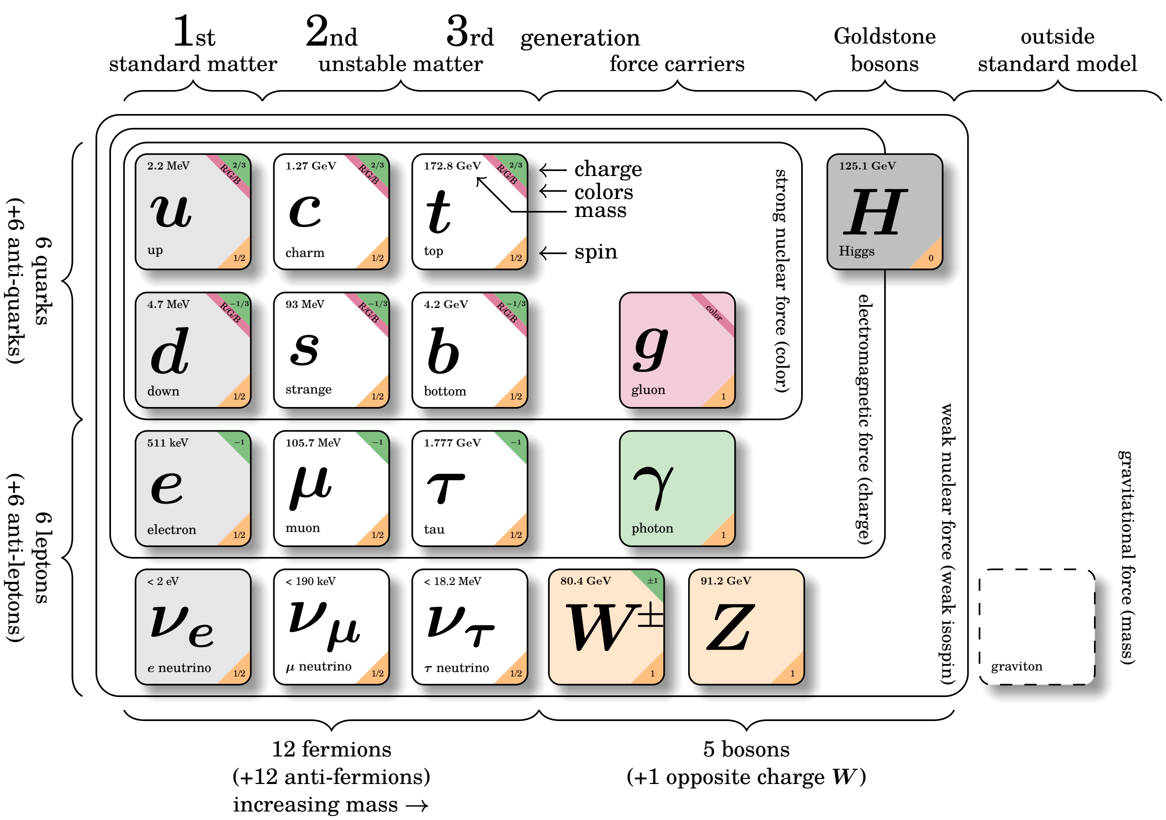


Figure 2.2: Summary diagram of the particles of the SM.[3,9]

2.1.2 The Standard Model Interactions

Electromagnetic Interaction

The electromagnetic interaction is the exchange of photons γ by charged particles. The interaction's theoretical foundation is incorporated into the SM by applying a Quantum Field Theory called Quantum Electro Dynamics (QED). A visual representation of a Quantum Field Theory can be obtained by using so called Feynman diagrams. Figure 2.3 illustrates such a diagram: a three point vertex of γ the gauge boson of QED and an incoming and outgoing electron. At the vertex (illustrated with a dot), it is common to indicate the coupling strength between particle and mediator. For QED the coupling strength is proportional to the electric charge. Feynman diagrams shown in this thesis have their positive time axis propagating from left to right following the usual convention.[7] In electromagnetism one essential requirement on the physical electric field $\vec{E}$ and the physical magnetic field $\vec{B}$ is their invariance under a gauge transformation[10]. This assures the fields to be independent from any initial choices on their reference frame. The interpretation of this requirement is, broadly speaking, that the laws of physics should not change through the frame of an observer. The QED/QFT equivalent to this, is the demand for a local $U(1)$ gauge symmetry.

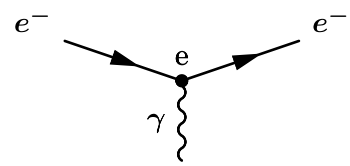


Figure 2.3: Fundamental interaction vertex of Quantum Electro Dynamics that all charged particles can undergo. Here, an interaction between the gauge boson γ and an incoming and outgoing electron is illustrated. Based on Ref.[6].

In order to satisfy this condition, a new field A_μ has to be introduced. This new field can be identified as the mediator particle γ while the fermion's charge q ($q = -\mathrm{e}$ in case of an electron) serves as coupling constant in QED. The terms in the resulting Lagrangian, that is a function of the fields, their derivatives, and their space time coordinates, describe the free fermion field, the coupling between the fermion field and the gauge field and the free gauge field itself.

Weak Interaction

As illustrated in Figure 2.4 the weak interaction can be split into two different types of interactions. While interactions including a W boson are referred to as charged current wherein there is always a change in flavor from ingoing to outgoing fermion, interactions including a Z boson are referred to as neutral current which never changes the flavor of participating fermions. These reactions are proportional to the respective coupling constants g_W and g_Z.

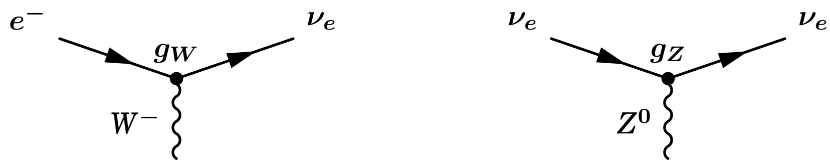


Figure 2.4: Fundamental interaction vertices of the weak interaction that all twelve fermions can undergo. In the left figure a weak flavor-changing charged current is illustrated and in the right figure a weak flavor-conserving neutral current is illustrated. Based on Ref.[6].

Since the W bosons carry a charge of ± 1 e and the electric charge has to be conserved in any particle exchange, only pairs of fermions that differ in charge corresponding to the charge of $W^\pm$ are coupled in the (weak) charged current interaction. In the case of leptons, the charged representative always couples with its uncharged neutrino partner. The interacting lepton pairs can be represented in dublets as follows:

$$\begin{pmatrix}\nu_e\\ e^-\end{pmatrix}, \begin{pmatrix}\nu_\mu\\ \mu^-\end{pmatrix}, \begin{pmatrix}\nu_\tau\\ \tau^-\end{pmatrix}.$$

The possible interaction dublets where the quarks' charge differs by ± 1 e are

$$\begin{pmatrix}u\\ d'\end{pmatrix}, \begin{pmatrix}u\\ s'\end{pmatrix}, \begin{pmatrix}u\\ b'\end{pmatrix}, \begin{pmatrix}c\\ d'\end{pmatrix}, \begin{pmatrix}c\\ s'\end{pmatrix}, \begin{pmatrix}c\\ b'\end{pmatrix}, \begin{pmatrix}t\\ d'\end{pmatrix}, \begin{pmatrix}t\\ s'\end{pmatrix}, \begin{pmatrix}t\\ b'\end{pmatrix}.$$

Here q' represents the weak eigenstate $|q'\rangle$ corresponding to a linear combination of the physical mass eigenstates. The transformation between weak and mass eigentstates is given by the *Cabibbo-Kobayashi-Maskawa*-Matrix (CKM-Matrix). In the following part additional characteristics of the weak interaction are going to be discussed intending to further indicate the special nature of this interaction and to lead to the QFT representation.

In 1956 the Wu-experiment[11] provided a clear demonstration of the weak interaction's parity violation in the exchange of a $W^\pm$ boson. The experiment found that parity is not conserved in the weak force, in contrast to the electromagnetic and strong interaction. The conclusion from this observation is that only left-handed fermions and right-handed antifermions participate in the weak interaction. Left- and right-handed corresponds here to the particles' *helicity*[2]

$$h = \frac{\vec{s}\cdot\vec{p}}{|\vec{p}|} \tag{2.1}$$

with the particle's spin $\vec{s}$ and momentum $\vec{p}$. In case spin and momentum are parallel, helicity is positive and one speaks of an right-handed particle. If however spin and momentum are anti-parallel, one speaks of an left-handed particle. Thus, all dublets given above should be considered left-handed in case they participate in the weak interaction in exchange with a W boson.

[2]Strictly speaking, the theory behind it describes how a particle transforms in a left- or right-handed representation of the Poincaré group. This property is called *chirality* χ. However, in the ultra-relativistic case where $E \gg m$ the helicity's quantum mechanical expectation value $\langle h\rangle$ is equal to the χ.

As it was discussed for QED, it is also possible to require a symmetry for the charged current weak interaction. The latter is the result of requiring an invariance under $SU(2)_L$ local phase transformation for the entire Lagrangian. The letter L accounts for the left-handed nature. The main consequences from this requirement is the introduction of three gauge fields W_μ^k from which the W boson can be identified from a linear combination of two out of three gauge fields. Furthermore two additional degrees of freedom, the weak isospin $\vec{T}$ with third component T_3, are introduced with which weak isospin dublets are formed, only containing left-handed/right-handed fermion/anti-fermion pairs from the dublets given above. Finally, the $SU(2)_L$ group does not commute which gives rise to the experimentally evident self-interaction of W and Z.

To identify the remaining $W_\mu^{(k)}$ and to include the Z boson into this theory, Glashow, Salam and Weinberg (GSW)[12,13] introduced a unification of the electromagnetic and weak interaction. In this *electroweak* model, the QED gauge symmetry $U(1)$ is replaced by an alternative $U(1)_Y$ gauge symmetry which in turn needs one further gauge field B_μ. The subscript Y stands for the *hypercharge* which is an additional degree of freedom that can be directly connected to the electromagnetic charge. The mediators γ and Z can be expressed as quantum mechanical superposition of B_μ and the remaining $W_\mu^{(k)}$.

Comparing the electromagnetic coupling constant with the weak coupling constant yields

$$\frac{e^2}{g_W^2} \sim 0.23^{[14]} \tag{2.2}$$

and might give reason to believe that the strength of QED interactions is less than the strength of the weak interaction. This is not true in reality and can be explained by the Feynman propagators. For QED interactions the propagator describing the exchange of a photon is

$$\frac{-ig_{\mu\nu}}{q^2} \tag{2.3}$$

where q is the momentum transfer defined as the difference in momentum before and after an interaction and $g^{\mu\nu}$ is the metric tensor. For the weak interaction, the propagator in lowest order, which is sufficient for this discussion, can be defined as

$$\frac{-ig_{\mu\nu}}{q^2 - m^2} \tag{2.4}$$

with an additional squared mass term of Wor Z boson in the denominator which is of the order $\sim (80\ \mathrm{GeV})^2$ for $W$and $\sim (91\ \mathrm{GeV})^2$ for Z. As the interaction probability is proportional to the square of the propagator term, it gets highly suppressed with the additional mass terms and it becomes clear why the weak interaction is termed "weak" and why it is much weaker than QED as shown in Table 2.1.

Strong Interaction

The QFT describing the interaction of quarks with gluons is termed Quantum Chromo Dynamics (QCD). The corresponding symmetry for QCD is an invariance under $SU(3)_C$ local phase transformation, where C accounts for a color. In similarity to the weak interaction, the requirement of a local gauge symmetry introduces eight new fields G^a_μ. Their non-commuting properties can be interpreted as possible gauge boson self-interactions which are illustrated in Figure 2.5 (middle and right vertex). In fact the fields can be identified as the experimentally observed gluons. Further, a threefold of degrees of freedom has been introduced and its values have been named the colors red, green and blue. The formalism of QCD has certain similarities to QED in terms of their structures: The interaction is mediated by one massless gauge boson in case of QED and eight massless gauge bosons in case of QCD. Instead of a single electric charge, QCD has three color charges r, g and b which have to be conserved as well as q in QED. While charged antiparticles in QED carry an opposite electric charge, QCD follows this pattern with antiquarks carrying an opposite color charge $\bar{r}$, $\bar{g}$ and $\bar{b}$.

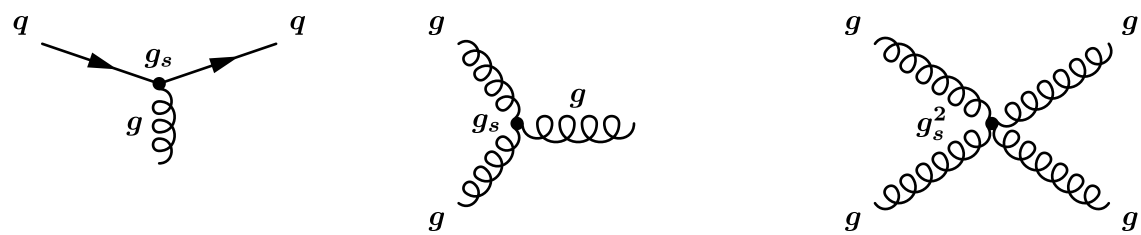


Figure 2.5: Fundamental interaction vertex and gluon self-interaction vertices of the strong interaction. The left vertex illustrates and incoming and outgoing quark interacting with a gluon. Middle and right vertices depict the triple and quartic gluon vertices that describe the gluon self-interaction. Based on Ref.[6].

As quarks carry color which has to be conserved at every particle vertex, it becomes clear from looking at a quark scatter process that gluons as mediators have to carry a color and different anticolor at the same time. From nine possible color combinations, group theory defines eight of them to have a physical interpretation. This set can be represented in the following octet

$$r\bar{g},\ r\bar{b},\ g\bar{b},\ g\bar{r},\ b\bar{r},\ b\bar{g},\ \frac{1}{\sqrt{2}}\left(r\bar{r}-g\bar{g}\right)\ \text{and}\ \frac{1}{\sqrt{6}}\left(r\bar{r}+g\bar{g}-2b\bar{b}\right).$$

The remaining unphysical singlet

$$\frac{1}{\sqrt{3}}\left(r\bar{r}+b\bar{b}+g\bar{g}\right)$$

is technically color-neutral. With the latter is not possible to interchange colors, it does therefore not participate in the strong interaction. In the following, effects originating from the color aspect are discussed. As this is still subject of current research like lattice QCD the explanations are going to be only qualitative. The intention here is to give a short overview of the phenomenology.

While many experiments have collected evidence for the existence of quarks, no experiment has ever detected a free quark. The *color confinement* hypothesis explains this as

followed: a colored object like a quark or a gluon can not propagate as a free particle but is confined into a color-neutral singlet state. The foundation of this hypothesis builds the gluon-gluon self-interaction argument. During an interaction of two quarks, it is possible for gluons to interact in between. This is believed to form color tubes of gluons. Eventually, quarks arrange in colorless bound states, so called hadrons. The two most primitive colorless bound states need to either comprise r, g and b or pairs like $r\bar{r}$, $g\bar{g}$ or $b\bar{b}$. The type of states with three quarks of color r, g and b are called *baryons* and the type of state with quark of color and corresponding anticolor are called *mesons*. The only quark that does not participate in the formation of hadrons is the top quark. Due to its lifetime being less than the time scale of the strong interaction, it may decay beforehand.
The color confinement has another very prominent impact on high energy physics collider experiments: For a produced $q\bar{q}$ pair progressively separating from each other back-to-back in the center of mass frame, the color confinment forces the enlarging color tube two separate into two $q\bar{q}$ pairs as this is energetically more favorable. This process goes on until it is not anymore favorable to create more $q\bar{q}$ pairs energy-wise. The newly formed quarks then arrange into hadrons which leads to two large sprays of hadrons called jets that propagate along the original paths of the $q\bar{q}$ pair produced in the initially so-called *hard scatter*.

Higgs mechanism

Besides of the Standard Model's inability to explain experimentally observed neutrino oscillations, the incorporated requirement for a $SU(3)_C \times SU(2)_L \times U(1)_Y$ symmetry prohibits the introduction of non-zero particle masses. In fact, explicit mass terms violate the required local gauge symmetry. While this is acceptable for the mediators of QED and QCD with their massless photon and gluon, it is in contradiction to the observed masses of W and Z as mediators of the weak interaction. Furthermore, this problem is not restricted to the gauge bosons as it can be shown that the requirement for fermion masses violates the mandatory gauge invariance, too.
This problem can be solved by introducing the Higgs mechanism into the theory of the Standard Model. As the Higgs boson is an integral component and subject of research of this thesis, the mechanism is explained in more detail. If not indicated differently, a summary of the works by Brout, Englert and Higgs as well as Glashow, Salam and Weinberg published by Thomson et al.[6] serves as a reference for the derivation of the mechanism.

Spontaneous Symmetry Breaking

The Lagrangian of a real scalar field ϕ is given by

$$\mathcal{L} = \frac{1}{2}\left(\partial_\mu \phi\right)\left(\partial^\mu \phi\right) - V(\phi) \tag{2.5}$$

$$= \frac{1}{2}\left(\partial_\mu \phi\right)\left(\partial^\mu \phi\right) - \frac{1}{2}\mu^2\phi^2 - \frac{1}{4}\lambda\phi^4. \tag{2.6}$$

Here $\frac{1}{2}\left(\partial_\mu \phi\right)\left(\partial^\mu \phi\right)$ can be interpreted as the kinetic energy of ϕ, $\frac{1}{2}\mu^2\phi^2$ can be associated with a mass of the scalar particle and the final $\frac{1}{4}\lambda\phi^4$ term can be associated with

possible self-interactions. This potential $V(\phi)$ has been initially chosen as it is the simplest potential possible for the discussion. The general procedure is to find the state of minimum energy of ϕ also referred to as the vacuum expectation value v. In case of the Lagrangian given above there are two possibilities. In any way λ is required to be non-negative to find a finite and therefore physical minimum. The first case where μ^2 is assumed to be positive, the potential $V(\phi)$ has its minimum at zero as depicted in the left graph of Figure 2.6. Thus having identified the vacuum state corresponding to $\phi = 0$, the Lagrangian describes a massive scalar particle field with real mass μ.
As there is no limitation on the choice of μ^2, it can also be chosen to be negative. In this case the minima are at

$$\phi = \pm v = \pm \left| \sqrt{\frac{-\mu^2}{\lambda}} \right| \tag{2.7}$$

as illustrated in the right graph of Figure 2.6. This has two important consequences: First, the minimum energy is not anymore at zero and is therefore termed to be a non-zero vacuum expectation value v. Furthermore, the graph in Figure 2.6 shows two minima between which the field ϕ can choose. Since the non-vanishing vacuum state breaks the Lagrangian's symmetry this behavior is referred to as *spontaneous symmetry breaking*. The second consequence is the fact that the choice of μ^2 makes it impossible to interpret the associated Lagrangian term as mass.
A physical mass of the field can be obtained from the Lagrangian by using a pertubative approach that varies the scalar field ϕ around zero by the amount $\eta(x)$ as $\phi(x) = v + \eta(x)$. Inserting this into the original Lagrangian at the beginning of the chapter yields

$$\mathcal{L}(\eta) = \frac{1}{2}\left(\partial_\eta \phi\right)\left(\partial^\eta \phi\right) - \frac{1}{2} m_\eta^2 \eta^2 - V(\eta) \tag{2.8}$$

with $m_\eta = \sqrt{-2\mu^2}$ and $V(\eta) = \lambda v \eta^3 + \frac{1}{4}\lambda\eta^4 + \text{const.}$ In that way it is possible to receive a term that can be interpreted as the mass of the scalar field induced by a process referred to as spontaneous symmetry breaking. Additionally, the terms in $V(\eta)$ proportional η^3 and η^4 can be interpreted as self-interaction terms as discussed in to previous paragraphs.

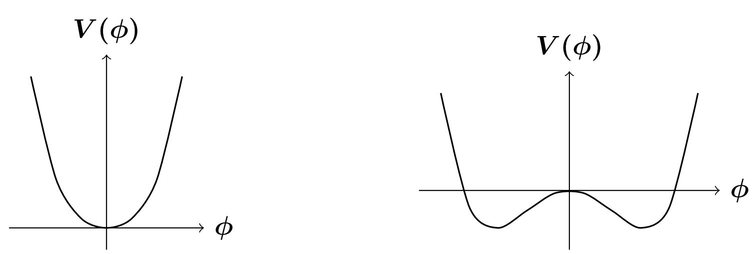


Figure 2.6: Visualization of the scalar field's potential $V(\phi) = \frac{1}{2}\mu^2\phi^2 - \frac{1}{4}\lambda\phi^4$. While λ has to be positive to get finite minimum, μ^2 can be either positive (left plot) or negative (right plot). In case $\mu^2 > 0$, the minimum is at $\phi = 0$, whereas for $\mu^2 < 0$, the minima can be calculated to be at $\phi = \pm\left|\sqrt{\frac{-\mu^2}{\lambda}}\right|$.

Generating Mass Terms for Gauge Bosons

As it can be shown that the spontaneous symmetry breaking generalizes to complex scalar fields

$$\phi = \frac{1}{\sqrt{2}} (\phi_1 + i\phi_2), \tag{2.9}$$

such a complex field is used in the upcoming discussion.
Applying the previously explained concept of requiring a $U(1)$ local gauge symmetry that provides in its consequence a new gauge field B_μ, a combined Lagrangian for the complex field ϕ and the gauge field B can be formed. Up to this point, the gauge field B is supposed to be massless since a mass term would violate gauge invariance as discussed. When following the procedure from the spontaneous symmetry breaking by expanding the complex scalar field around the vacuum state, a massive scalar field η and a massless Goldstone boson ξ are produced. The latter emerges from the additional degree of freedom of the complex scalar field. Furthermore, the originally massless gauge field B acquires mass. With an appropriate gauge transformation, which is termed *unitary gauge*, the degree of freedom corresponding to the Goldstone boson ξ can be replaced by the mass-induced longitudinal polarisation of the previously massless B. Now following the expansion around the vacuum state after breaking symmetry the Lagrangian can be expressed in parts describing the massive scaler field termed h, the massive gauge boson field B, the interaction between h and B and the self-interaction of h.

To stress that the previously named field η, that emerges from the pertubative approach, can in fact be identified as the physical Higgs field, it has been renamed to h in this scenario. Thus, the explained technique yields a $U(1)$ gauge-symmetric Lagrangian that is able to describe a massive scalar boson associated with the Higgs field h together with a massive gauge boson B. Furthermore, a mass of the gauge boson can be identified to be $m_B = gv$ which shows a direct relation to the gauge coupling strength g and the vacuum expectation value of h. Finally, the Higgs boson mass can be found to be $m_H = \sqrt{2\lambda}v$. While the vacuum expectation value can be theoretically calculated to $v = 246$ GeV, λ is a free parameter of the theory and therefore m_H is a priori unknown.

The Higgs Boson in the Standard Model

As a third step, it needs to be explained how the Higgs boson is eventually incorporated into the SM by discussing the relevant steps in order to embed the above explained Higgs mechanism into $U(1)_Y \times SU(2)_L$ local gauge symmetry from the electroweak interaction.
The complex scalar field arranged in a weak isospin doublet is

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}. \tag{2.10}$$

To account for the charges of the Standard Model particles, the first component of the field has charge. As explained for the weak interaction, the two components differ by unit of charge (+1 e for ϕ^+ and -1 e for its complex conjugated associate). The corresponding Lagrangian is

$$\mathcal{L} = (\partial_\mu \phi)^\dagger (\partial^\mu \phi) - V(\phi) \tag{2.11}$$
$$= (\partial_\mu \phi)^\dagger (\partial^\mu \phi) - \mu^2 \phi^\dagger \phi - \lambda (\phi^\dagger \phi)^2. \tag{2.12}$$

Carrying out the symmetry breaking formalism, the field can be expanded as followed

$$\phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}. \tag{2.13}$$

This representation is chosen from all degenerated states emerging from symmetry breaking to set the minimum of the potential to a non-vanishing expectation value at the neutral scalar field ϕ^0. This translates into the massless photon one observes in experiments. Replacing the derivatives in the Higgs-Lagrangian with covariant derivatives that satisfy the required $SU(2)_L \times U(1)_Y$ symmetry of the form

$$\partial_\mu \rightarrow D_\mu = \partial_\mu + i g_W \vec{T} \cdot \vec{W}_\mu + i g' \frac{Y}{2} B_\mu \tag{2.14}$$

brings back the hyper charge and weak isospin from the electroweak unification and sets everything into place to derive the masses of the mediator particles from the symmetry breaking principle. These can be found to be

$$m_\gamma = 0 \tag{2.15}$$

as mass of the photon,

$$m_W = \frac{1}{2} g_W v \tag{2.16}$$

as mass of the W boson and

$$m_Z = \frac{g_W}{2 \cos \theta_W} v \tag{2.17}$$

as mass of the Z boson. Rearranging the kinetic term of the Higgs-Lagrangian results in an expression whose terms directly relate to the coupling between Higgs and W or Z, respectively. Expressing it in terms of the physical $W^\pm$ fields, one finds the expression

$$\frac{1}{2} g_W^2 v W_\mu^- W^{+\mu} h \tag{2.18}$$

that describes the coupling hW^+W^-, which is the coupling of interest in this thesis, with a corresponding coupling strength of

$$g_{HWW} = \frac{1}{2} g_W^2 v \overset{(2.16)}{=} g_W m_W. \tag{2.19}$$

In a similar fashion it can be argued that the symmetry breaking induces masses for the Standard Model fermions.[15] Working through the formalism yields the coupling constant g_f, also referred to as Yukawa coupling, that determines the coupling strength between Higgs and fermions to be

$$g_f = \sqrt{2}\frac{m_f}{v} \overset{(2.16)}{=} \frac{m_f}{2m_W} g_W. \tag{2.20}$$

Thus, fermions not only receive their masses from the Higgs field, the latter also couples to fermions. The neutrino masses are not yet explained within this formalism and are still subject of current research.

2.1.3 Properties of the Higgs

The intention of the previous discussion was to conclude that the Higgs can decay into any massive Standard Model particle pair $x\bar{x}$ with $m_H > 2m_x$. On top of that, the coupling is proportional to the mass m_x of the particle as indicated in Equations 2.16, 2.17 and 2.20. This is in agreement with recent experimental observations[16] which also show that the decay $H \to b\bar{b}$ has the highest branching fraction. The observation of decays into massless gauge bosons like $H \to \gamma\gamma$ can be explained similarly as done in the upcoming chapter. The following section aims to give a short overview of the possible production and decay modes of the Higgs boson motivated by the corresponding parts of the theory that have been described so far.

Production Modes at the LHC

Based on the coupling rules in the previous chapter, theoretical production cross sections can be calculated in principle. In proton-proton collisions, these cross sections furthermore depend on the momentum fraction carried by the proton's constituents namely two up-quarks, a down-quark and gluons, generally termed *partons*. The probability to find a parton a from a hadron A with momentum fraction x_A is given by the *Parton Distribution Function* (PDF) $f_{a/A}(x_A)dx_A$.[17] The latter has to be determined experimentally.

Production cross sections of proton-proton collisions including the creation of a Higgs boson are illustrated in Figure 2.7. The left subfigure shows the production cross section for Higgs masses near the observed value $m_H = 125$ GeV at a center-of-mass energy of $\sqrt{s} = 13$ TeV. The process of gluon-gluon fusion ($pp \to H$) in blue has by far the highest production cross section of ~ 50 pb. It is followed by the process of vector-boson-fusion ($pp \to qqH$) in red with a production cross section of ~ 4 pb. The third most probable mode is the associated Higgs production with a W ($pp \to WH$) in green with ~ 1.5 pb followed by the associated Higgs production with a Z boson ($pp \to ZH$) in grey. The two rarest production modes considered still accessible at the LHC in descending cross section order are top-antitop-fusion ($pp \to t\bar{t}tH$), bottom-antibottom-fusion ($pp \to bbH$) and associated Higgs production with a top quark ($pp \to tH$).
The right subplot gives an estimate of production cross sections of the aforementioned production channels as a function of center-of-mass energy assuming a Higgs mass of 125 GeV. All channels show a monotonically increase in cross section with increasing center-of-mass energy.

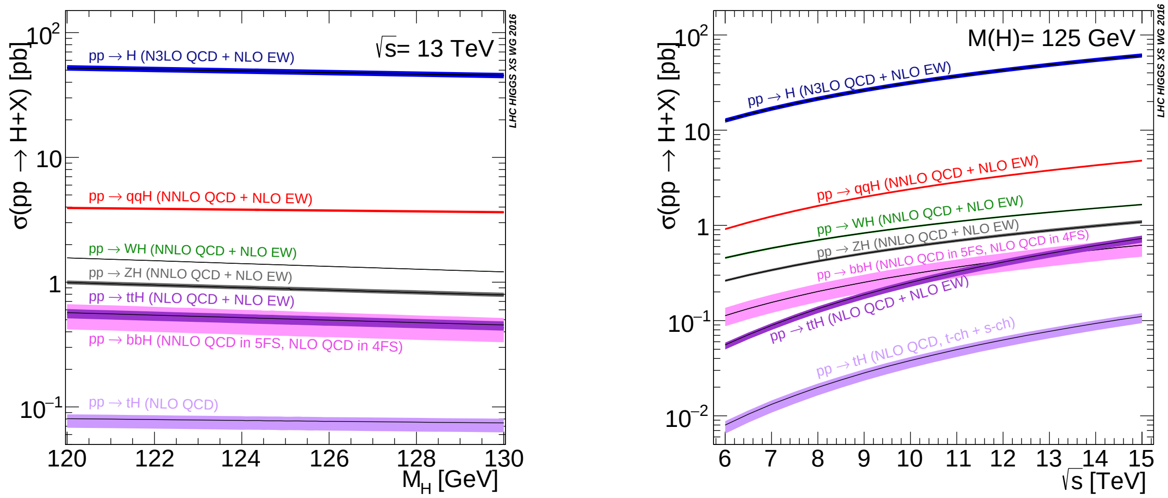


Figure 2.7: Left: the production cross section in pb for a narrow mass spectrum around the measured Higgs mass of 125 GeV at $\sqrt{s} = 13$ TeV. **Right:** the production cross sections as a function of center-of-mass energy assuming a Higgs mass of 125 GeV. The following physical processes are considered: gluon-gluon Fusion (blue), vector-boson-fusion (red) associated Higgs production with W (green) and Z (grey), top-antitop-fusion (dark purple), bottom-antibottom-fusion (light purple) and associated Higgs production with a top quark (lilac).[18]

Decay Modes

In Figure 2.8, the three lowest-order diagrams of the Higgs decay into SM particles are illustrated together with the couplings derived in Equations 2.16, 2.17 and 2.20. The vertex expressions indicate that the largest branching ratios are reserved for the most massive particles due to the dependence of the coupling to the mass of the particle which is itself proportional to the branching ratio.

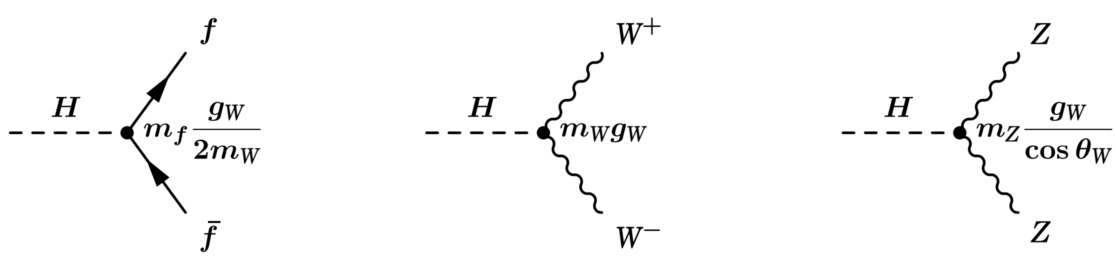


Figure 2.8: Fundamental Feynman diagrams for the Higgs decay. From left to right: Higgs decaying into two fermions $f\bar{f}$, Higgs decaying into WW and Higgs decaying into ZZ. Based on Ref.[6]

By taking into account the masses from Figure 2.2, the relation $m_H > 2m_x$ yields that from all possible SM particle pairs it is the b-quark pair with a mass of 2×4.8 GeV that is predicted to have the highest branching fraction from all possible Higgs decays. A broad summary on the predicted branching ratios of the Higgs boson for a narrow window around the experimentally extracted mass $m_H = 125$ GeV is given in Figure 2.9. The figure shows that $b\bar{b}$ has the highest theoretically predicted branching ratio with roughly 57%[6], followed by the decay $H \to WW^*$ although $m_H < 2m_W$. This is due to the fact

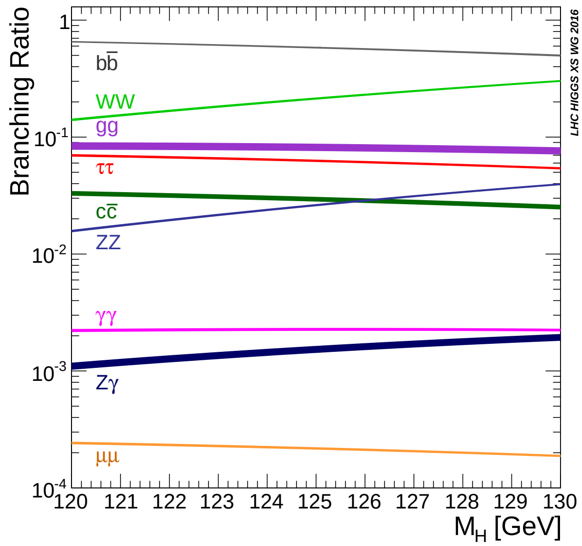


Figure 2.9: Branching ratio of possible Higgs decay modes. $H \to WW^*$ has the second highest branching ratio and is therefore a favorable mode for a production cross section measurement.[18]

that one of the W bosons is produced off-mass-shell with $q^2 < m_W^2$ also denoted as W^*. To describe such an off-(mass-)shell W boson, one additional term $(im_W\Gamma_W)$ has to be added in the denominator of the propagator term from Equation 2.4 which also accounts for the finite lifetime of the W boson. The expression

$$\frac{1}{q^2 - m_W^2 + im_W\Gamma_W} \tag{2.21}$$

indicates that the presence of an off-shell W, or Z, tends to suppress the matrix element that is proportional to the branching fraction. Nevertheless, the coupling term for H and W, $m_W g_W$, ensures by its large value the high branching fraction compared to the remaining decay modes depicted in Figure 2.9.

Finally, Figure 2.9 illustrates the fairly high branching ratios of the Higgs decay modes into massless particles like g or γ. In fact, $H \to \gamma\gamma$ was one of the first channels in which the Higgs boson was observed. Such a decay mode is possible via loops of virtual top quarks or W bosons. Due to the high top-quark mass and therefore high coupling with the Higgs, these modes can compete with the decays into massive particles.

Properties of the leptonic $H \to WW^*$ decay produced at the LHC

This chapter deals with the kinematic details of the decay

$$q\bar{q} \to WH \to WWW^* \to 3\ell\, 3\nu$$

originating from proton-proton collisions as performed at the LHC as illustrated in Figure 2.10. In this analyis, ℓ denotes charged leptons of flavor e, or μ. τ leptons are not considered as described in the chapter on the experimental setup.

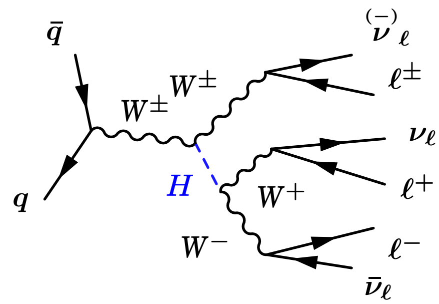


Figure 2.10: Feynman diagram of the signal process $q\bar{q} \to WH \to WWW^* \to 3\ell\,3\nu$.

The analysis requires three charged leptons in the final state which are supposed to origin from the associated Higgs production with a W boson, where the W boson decays into one ℓ-ν_ℓ pair and the Higgs decays into two W bosons that themselves decay into ℓ-ν_ℓ pairs. The spin-0 nature of the H boson has direct consequences on this decay. To conserve angular momentum, the spin components of the W bosons from the Higgs decay along the decay axis, m_s, have to sum up to zero. In the non-vanishing case of e.g. $m_{s_{W^\pm}}$ = ± 1 and $m_{s_{W^\mp}}$ = ∓ 1, the spins of the leptons have to sum up to $\pm$ 1 as well and consequently, the ℓ-ν_ℓ pair is required to have aligned spins. Here the helicity comes into play: as discussed, the SM requires left-handed particles and right-handed antiparticles only. Following this rule and considering momentum conservation, the only possible configuration comprises a scenario wherein the neutrino spin is anti-aligned with respect to the anti-neutrino and thus their momenta being aligned (compare Figure 2.11). Thus, ν_ℓ and $\bar{\nu}_\ell$ propagate in the same direction as well as ℓ^+ and ℓ^- in the restframe of the decaying particles. The angle between the charged leptons therefore tends to be small.

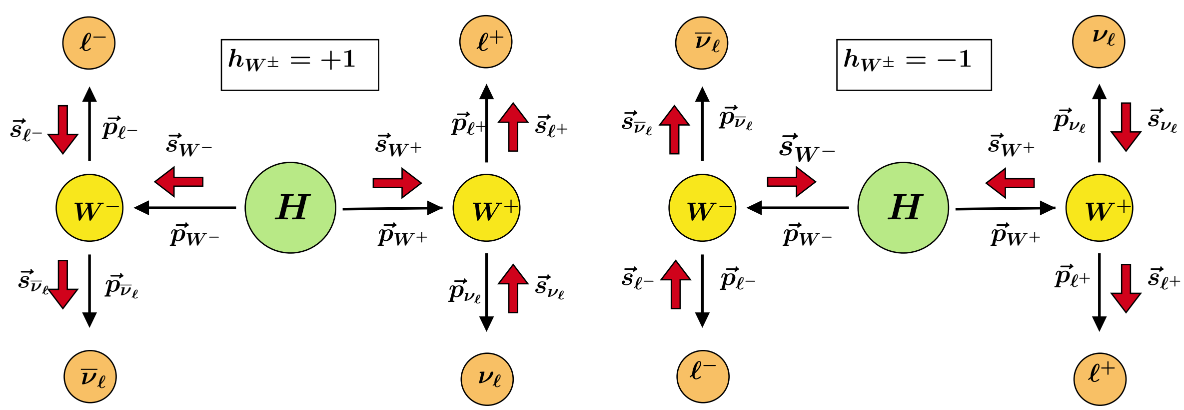


Figure 2.11: Schematic drawing of the decay $H \to WW^*$ with emphasis on the spin-momentum correlations of the decay products. While the black solid arrows illustrate the direction of the momentum vectors, the red solid arrows visualize the particles' spin direction. The decays of H and W are shown in their restframes.

3 | The ATLAS Experiment at the LHC

The *Large Hadron Collider* (LHC)[19] is a particle accelerator facility of about 26.7 km circumference located at CERN, the European Organization for Nuclear Research, on the Swiss-French border near Geneva, Switzerland. The accelerator ring of superconducting magnets, located in a tunnel 100 m underground, is capable of accelerating protons and heavier nuclei and has been equipped with seven detectors installed at and around the four beam interaction points.

The four big experiments with their detectors build around the collision points are *A Large Ion Collider Experiment* (ALICE)[20], *A Toroidal LHC ApparatuS* (ATLAS)[21], the *Compact Muon Solenoid* (CMS)[22] and the *Large Hadron Collider beauty* (LHCb) experiment[23]. The remaining three experiments are called *TOTal Elastic and diffractive cross section Measurement* (TOTEM)[24] located in the vicinity of the CMS interaction point, *Large Hadron Collider forward* (LHCf)[25] installed nearby ATLAS and the *Monopole and Exotics Detector at the LHC* (MoEDAL) experiment[26] close to the LHCb detector.

The main goal of the experiments at the LHC is to test the Standard Model's predictions and to answer questions about e.g. the origin of mass as it has been successfully done with the observation of the Higgs boson in 2012. Since the Standard Model is incomplete, the LHC also serves to give hints on not yet confirmed theories like supersymmetry and particles or phenomena that might explain dark matter or dark energy, respectively. The upcoming chapter deals with an overview of the key aspects of the LHC with emphasis on the ATLAS experiment followed by a dedicated chapter on the ATLAS detector.[27]

3.1 The Large Hadron Collider (LHC)

With the first beam at the LHC[28] in September 2008, many following milestones have been achieved. In December 2009 the so far highest center-of-mass energy of $\sqrt{s}$ = 2.36 TeV[29] of a particle physics collision experiment was achieved followed by a successive increase up to center-of-mass energies of $\sqrt{s}$ = 7-8 TeV for a two year data taking in 2011-2012 (*Run 1*) and a final center-of-mass energy of $\sqrt{s}$ = 13 TeV for a three-years data taking period from 2015 to the end of 2018 (*Run 2*). To reach such high energies a step-wise acceleration procedure is carried out:
Before the beam is injected into the LHC it is accelerated in four pre-accelerators with successively increasing energy. In case of protons, the first step is the LINAC 2 (*LINear ACcelerator* 2) , an accelerator providing protons of 50 MeV. These protons are fed forward into the *Proton Synchrotron Booster* (PSB) that accelerates the protons up to 1.4 GeV followed by the *Proton Synchrotron* (PS) where energies of 26 GeV are reached. In

the ultimate stage before the LHC, the *Super Proton Synchrotron* (SPS), the protons are further accelerated up to 450 GeV before being injected into main ring namely the LHC. A schematic drawing of the mentioned facilities' location at CERN is given in Figure 3.1. Lead ions are accelerated in LINAC 3, followed by the *Low Energy Ion Ring* (LEIR), PS and SPS before entering LHC.[30]

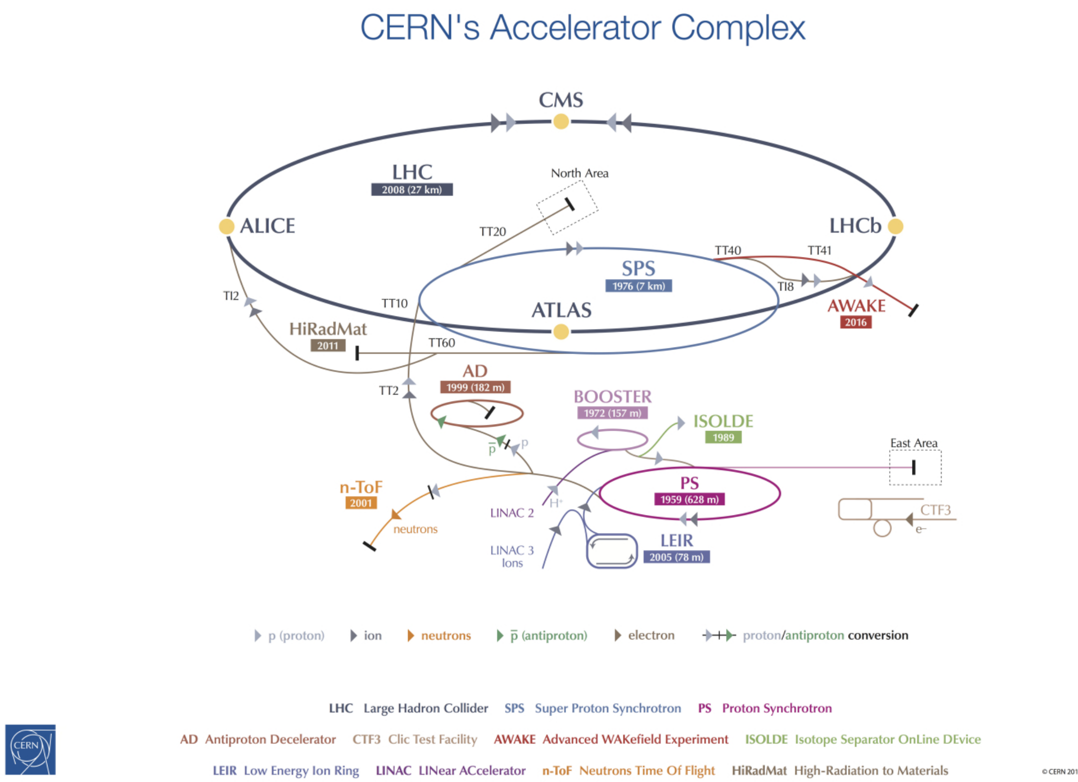


Figure 3.1: Schematic drawing of the LHC and the location of its corresponding pre-accelarators as well as the four main experiments ALICE, ATLAS, CMS and LHCb.[19]

The LHC consists of two neighbouring beam lines into which the hadrons are injected clock-wise and counter clock-wise in order to collide them. The particles are accelerated by 1232 superconducting dipole magnets operating at 1.9 K (-271.3°C). During Run 2 protons are accelerated to nominal energies of up to 6.5 TeV (equivalent to 99.999997% speed of light) corresponding to a center-of-mass energy of $\sqrt{s} = 13$ TeV whereas ions are accelerated to nominal energies of 2.56 TeV per nucleon. 2808 bunches with a spacing of 7.5 m per proton beam including 1.2×10^{11} protons each, take 11245 turns per second resulting in $\sim$ 1 billion collisions per second at ATLAS.

Besides the aforementioned center-of-mass energy $\sqrt{s}$, the time-integrated *luminosity* $\mathcal{L}$ is another central key quantity in particle collisions. In general it describes the number of particle interactions per area. It is defined as the integral over time of the instantaneous luminosity L_0:

$$\mathcal{L} = \int L_0 \, dt = \int \frac{\dot{N}}{\sigma} \, dt = \frac{N}{\sigma}. \tag{3.1}$$

Here $\dot{N}$ describes the interaction rate, N is the number of interactions per time interval dt and σ is the interaction cross section over all possible interactions. The latter depends on $\sqrt{s}$ and is expressed in units of $[\sigma] = 1\ \mathrm{b(arn)} = 10^{-28}\mathrm{m}^2$.

While the design instantaneous luminosity for protons at the LHC has been predicted to be $\sim 1\times 10^{34}\ \mathrm{cm}^{-2}\mathrm{s}^{-1}$, the LHC exceeded this target in 2017 with $2.06\times 10^{34}\ \mathrm{cm}^{-2}\mathrm{s}^{-1}$ at twice the nominal expected value.[31] Until the end of the second round of data taking Run 2 the time-integrated luminosity at ATLAS accumulated to 139 fb^{-1} of events that satisfy all quality requirements to use them for physics analyses as shown in the left plot of Figure 3.2.

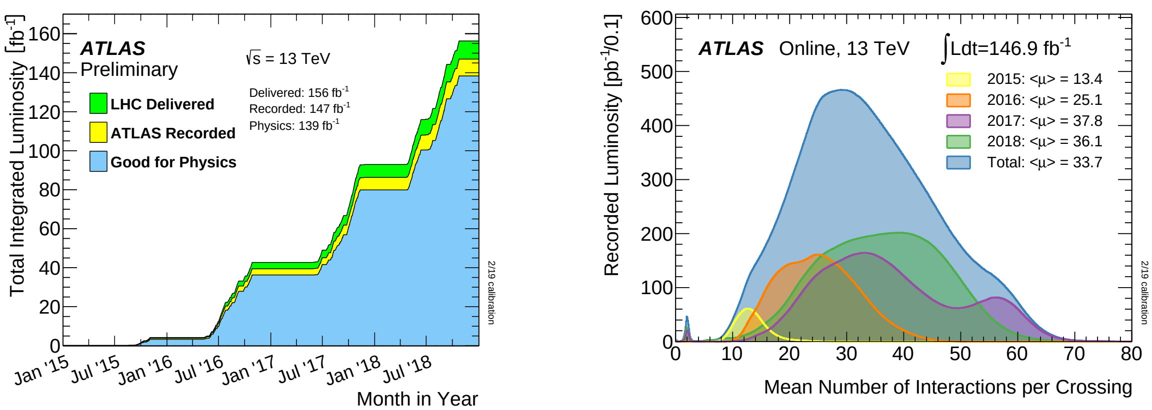


Figure 3.2: Left: Time-integrated luminosity for Run 2 from beginning of 2015 until the end of 2018, split into three categories. Good for physics analyses (139 fb^{-1} in light blue), by ATLAS recorded (147 fb^{-1} in yellow) and total by the LHC delivered (156 fb^{-1} in green). **Right:** Mean number of interactions per bunch crossing for the Run 2 period from 2015 to 2018. The online luminosity from the right figure corresponds to the time-integrated luminosity recorded by ATLAS in the left plot in yellow.[32,33]

As mentioned in the previous paragraph, each proton bunch includes roughly 1.2×10^{11} protons at the start. Thus multiple proton interactions can take place per bunch crossing. Typical numbers of this phenomenon termed $<\mu>$, the mean number of interactions per crossing, are given in the right plot of Figure 3.2. As this quantity is related to the luminosity, it is plotted against the latter for the full Run 2 data taking period and each year separately. A total mean number of interactions per bunch crossing of $<\mu_{Run2}> = 33.7$ has been observed for full Run 2 with a maximum of $<\mu_{2017}> = 37.8$ in 2017.

When operating at high luminosities resulting in equivalently high numbers of $<\mu>$, rare interesting physics processes might not be detected in favor of less interesting and less rare background processes. This effect is called *pile-up* and it is one of the main challenges the electronics and algorithms of the detector have to overcome.

3.2 The ATLAS Experiment

The technical core of the ATLAS experiment is its detector. The in 100 m below ground located cylindrical construction weighs 7000 t, is 44 m long and has a transverse diameter of 25 m. The detector is a multi-layer instrument and consists of several subsystems which are wrapped concentrically around the beam with the interaction point in the detector's middle. It is furthermore equipped with a strong magnet system employed to bend the trajectories of charged particles originating from the collisions in order to measure their momentum.

The four major subsystems inside the detector are the *Inner Detector* (ID) measuring the direction, momentum and charge of electrically charged particles emerging from proton-proton collisions, the *calorimeter* measuring the energy loss of a particle crossing the system, the *Muon Spectrometer* (MS) identifying and measuring the momentum of muons and the aforementioned magnet system.
A cylindrical detector like the ATLAS detector is often distinguished into two main regions as points of reference: the cylindrical part around the beam line is termed barrel while the regions at the two detector's ends perpendicular to the beam line are referred to as *end-caps*. A more detailed overview of the detector's components is depicted in Figure 3.3 which serves as a visual guide through the upcoming chapter dealing with the geometry and the components.[34]

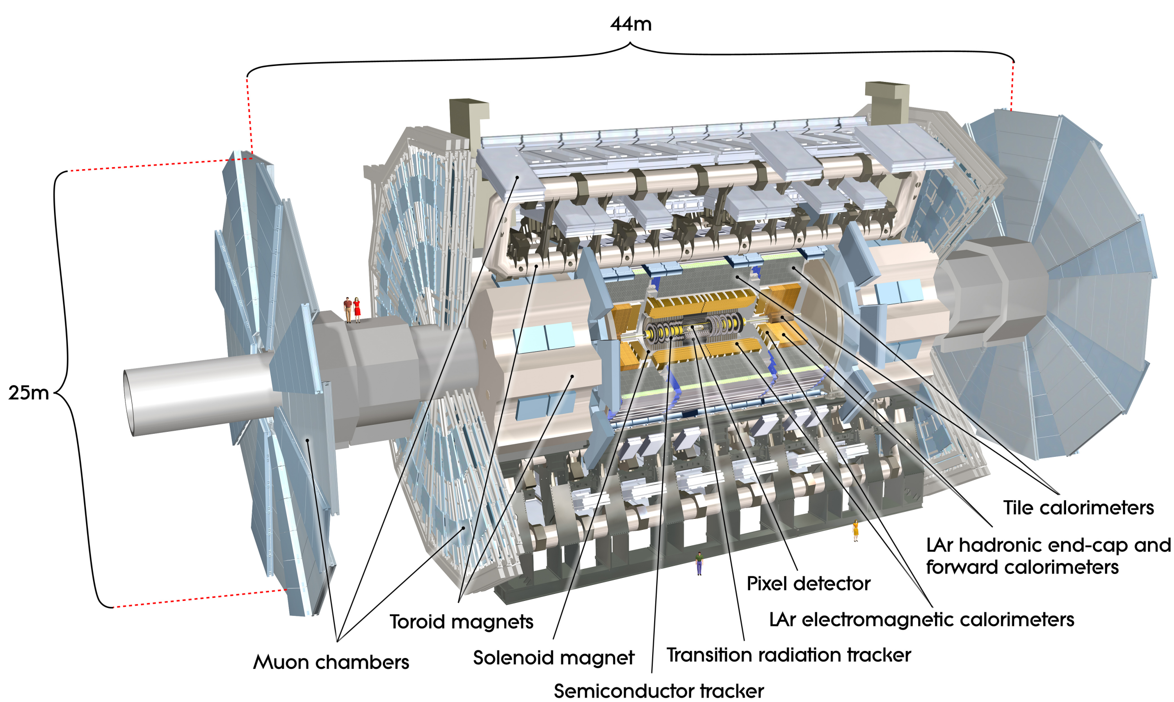


Figure 3.3: Schematic drawing of the ATLAS detector and its components.[35]

3.2.1 The ATLAS Event Kinematics

In this chapter the geometry of the ATLAS detector is explained. Additionally, physical quantities are introduced that are used in the analysis.
The established ATLAS coordinate system is a three-dimensional right-handed coordinate system. The z-axis is aligned with the beam axis such that the positive x-axis points away from the detector towards the center of the LHC and the positive y-axis points upwards. To exploit the shape of the detector, the position vector is usually expressed in cylindrical or spherical coordinates r, φ and θ.

Here φ is the azimuth angle defined as the angle between the positive x-axis and the position vector $\vec{r}_{xy}$ projected onto the x-y-plane. The polar angle θ is defined as the angle between the positive z-axis and position vector $\vec{r}$. An illustration of these two definitions is shown in Figure 3.4.

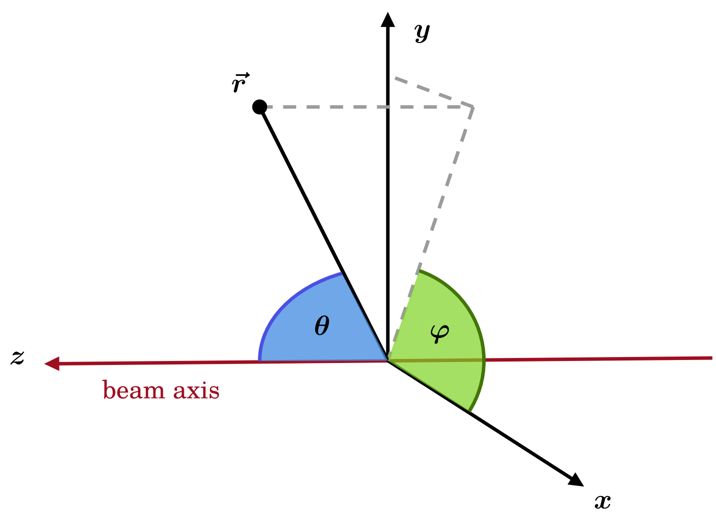


Figure 3.4: Schematic drawing of the right-handed coordinate system used in ATLAS. The z-axis is aligned with the beam, the positive x-axis points towards the center of the LHC and the y-axis points upwards completing the right-handed triad.

In pp-collisions as carried out at the LHC, the momentum fractions of the partons is a priori unknown as the center-of-mass energy can only be expressed in terms of the full protons. Thus it is reasonable to measure the momentum in the transverse plane to the beam axis as the vectorial sum of these momenta is negligible in the initial state. This quantity

$$|\vec{p}_T| = \sqrt{p_x^2 + p_y^2} \equiv p_T \tag{3.2}$$

is termed *transverse momentum*. In case of "invisible" particles that can not be detected by the detector as discussed in the upcoming chapter, the energy of the event might not sum up to the initial energy available.

This *missing energy* can be expressed in the transverse plane as

$$\sum_{\text{observable}} \vec{p}_T + \vec{E}_T^{miss} = 0 \tag{3.3}$$

$$\Leftrightarrow \vec{E}_T^{miss} = -\sum_{\text{observable}} \vec{p}_T. \tag{3.4}$$

With

$$\vec{E}_T^{miss} = \begin{pmatrix} E_x^{miss} \\ E_y^{miss} \end{pmatrix} \tag{3.5}$$

it is further possible to get an expression of the azimuth direction of the missing transverse energy as follows:

$$\varphi^{E_T^{miss}} = \tan^{-1}\left(E_y^{miss}/E_x^{miss}\right). \tag{3.6}$$

Here, an E_T^{miss} dependent observable termed *transverse mass* for the sum of four vectors[1] of two particles ℓ_0 and ℓ_1 is defined as

$$m_T^{\ell_0,\ell_1} = \sqrt{\left(E_T^{\ell_0,\ell_1} - E_T^{miss}\right)^2 - \left|\vec{p}_T^{\,\ell_0,\ell_1} - \vec{p}_T^{\,miss}\right|^2}. \tag{3.7}$$

Another typical mass-related quantity can be constructed by using the four vector notation: the *invariant mass* of n particles is given by

$$m_{inv} = \sqrt{\left(\sum_{k=1}^{n} p_k^{\mu}\right)^2} = \sqrt{\left(\sum_{k=1}^{n} E_k\right)^2 - \left(\sum_{k=1}^{n} \vec{p}_k\right)^2}. \tag{3.8}$$

Furthermore, the parton system may be boosted in one direction along the beam line since the momentum fraction of the colliding partons may not be equal in general. Therefore, high-energy physics takes advantage of quantities invariant under boosts. A quantity like this, that is also used at ATLAS, is the so called *rapidity*

$$y = \frac{1}{2}\ln\left(\frac{E + |p_z|}{E - |p_z|}\right). \tag{3.9}$$

However, it is not the rapidity itself but the differences in rapidity that are invariant under boosts and therefore the same in two different frames, $\Delta y' = \Delta y$. In the ultra-relativistic case $E \approx |\vec{p}|$, the rapidity y can be approximated by the so called *pseudorapidity*

$$\eta = \frac{1}{2}\ln\left(\frac{E + |p_z|}{E - |p_z|}\right) \overset{E\approx|\vec{p}|}{=} -\ln\left[\tan\left(\frac{\theta}{2}\right)\right]. \tag{3.10}$$

Hence, in the ultra-relativistic regime of high-energy physics experiments like ATLAS, the polar angle θ can be expressed in terms of η with differences in η being approximately invariant under boosts. $\eta = 0$ corresponds to the entire plane obtained by rotating the

[1]The four vector notation $p^{\mu} = (E, \vec{p})^T$ is used.

y-axis around the z-axis. Comparing two particles with indices 1 and 2 with respect to their spherical coordinates, one useful quantity is

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\varphi)^2} \tag{3.11}$$

where

$$\Delta\eta = |\eta_1 - \eta_2| \tag{3.12}$$

is their separation in η and

$$\Delta\varphi = \begin{cases} |\varphi_1 - \varphi_2|, & \text{if } |\varphi_1 - \varphi_2| \leq \pi \\ 2\pi - |\varphi_1 - \varphi_2|, & \text{if } |\varphi_1 - \varphi_2| > \pi \end{cases} \tag{3.13}$$

is the separation in φ as illustrated in Figure 3.5:
The left subfigure shows the spatial locations of two particles denoted by indices 1 and 2 at $\vec{r}_1$ and $\vec{r}_2$ in a coordinate system as it is used in the ATLAS experiment. The spatial information is further split up into $\eta_{1,2}$ and $\varphi_{1,2}$. The right subfigure illustrates the corresponding angular separation calculated by ΔR.

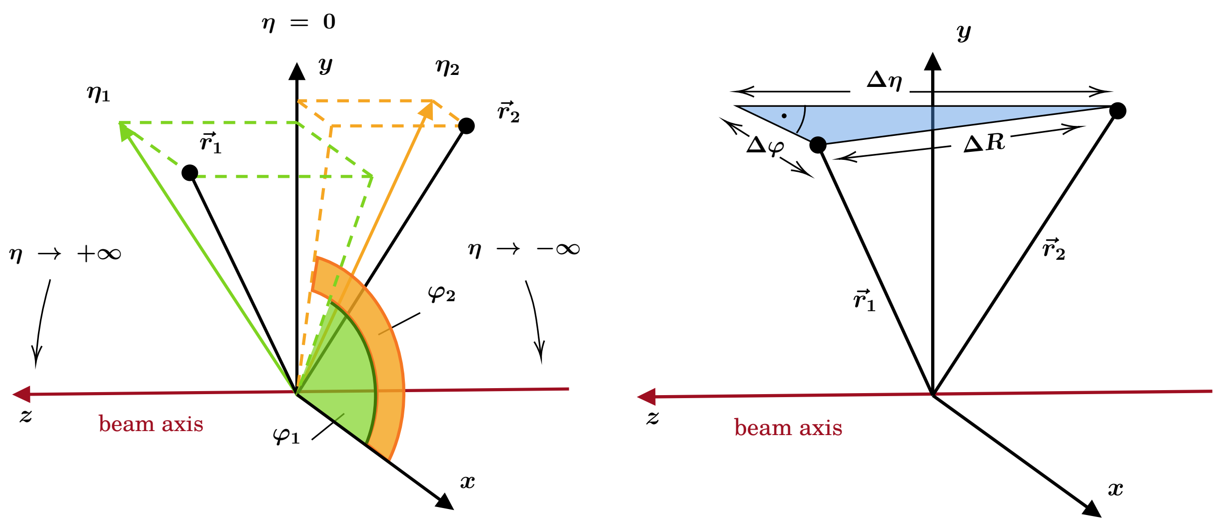


Figure 3.5: Left: Two spatial vectors $\vec{r}_1$ and $\vec{r}_2$ with corresponding φ and η information in a right-handed coordinate system like it is used for the ATLAS experiment, **Right:** The corresponding difference in η and φ of the vectors $\vec{r}_1$ and $\vec{r}_2$ from which the angular separation ΔR can be derived.

Another spatial quantity of interest is the *impact parameter* which describes the minimal distance between a particle's reconstructed trajectory and its corresponding reconstructed *primary vertex*. This is, broadly speaking, the closest approach to the interaction point from which the particle is believed to originate.
The quantity can be split up into a transversal component d_0 in the x-y-plane and a longitudinal component z_0 along the z-axis as illustrated in Figure 3.6.

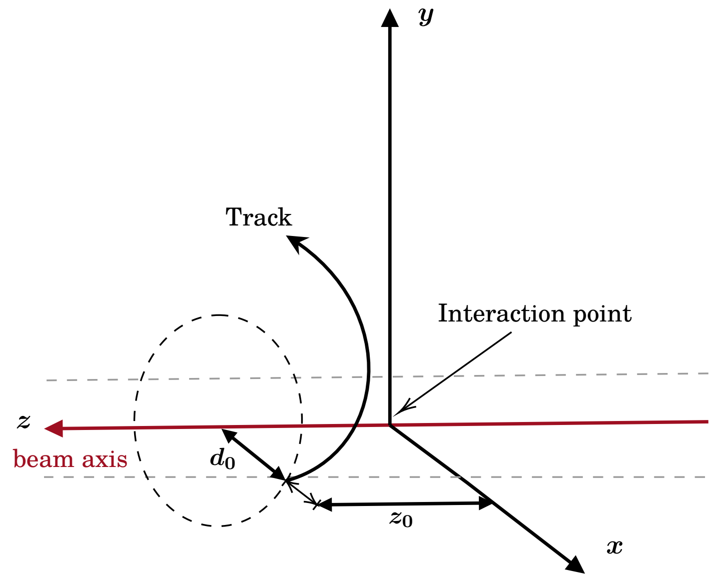


Figure 3.6: Illustration of the transverse impact parameter d_0 and the longitudinal impact parameter z_0 inspired by Ref.[36].

3.2.2 The Detector Components

The Magnet System

The ATLAS magnet system is employed to bend the charged particles' trajectories and consists of three main systems: The innermost part is called *central solenoid magnet* and is installed around the ID. With a total of 9 km of superconducting wires it induces a magnetic field of 2 T at a nominal current of 7.773 kA. It is furthermore surrounded by a toroidal magnetic system that is installed farther outwards in the barrel region (*barrel toroid*) as well as in the end-caps (*end-cap toroid*). This system consists of 8 separate coils each, inducing a 4 T magnetic field based on 100 km superconducting wires working at a nominal current of 20.5 kA.[37]

The Inner Detector

The ID illustrated in the left part of Figure 3.7 constitutes the first detector layers around the beam including beam pipe and interaction point. It is thus the first detector component, particles emerging from the collisions can interact with. It comprises three subdetector systems, namely the *pixel detector*[38], the *SemiConductor Tracker* (SCT)[39] both covering $|\eta| < 2.5$ and the *Transition Radiation Tracker* (TRT)[40] covering $|\eta| < 2.0$ as shown in more detail in the right part of Figure 3.7. With help of the surrounding solenoid magnet inducing the magnetic field which forces the charged particle onto bent trajectories, the three subsystems are employed to measure and identify the momentum and charge, respectively. Additionally, the tracked trajectory can be used to estimate the origin of the charged particle.

The pixel detector consists of 80 million pixels arranged in modules of which roughly 80% are used in three barrel layers and the remaining modules are used in 3 pixel disks in each end-cap. The whole system covers a full area of 1.7 $\mathbf{m^2}$ installed around the interaction point. Furthermore, each pixel has an effective resolution of $14{\times}115$ $\mu\mathbf{m^2}$. One additional layer has been installed as innermost part with only 3.3 cm distance to the beam line and has been used as of beginning of Run 2. This layer called *Insertable B-Layer* (IBL)[41] is supposed to maintain the tracking performance in the harsh radiation environment in immediate vicinity to the beam by adding a further active layer comprising another 13 million readout channels of size 50×250 $\mu\mathbf{m^2}$.
The SCT that is wrapped around the pixel detector is a silicon microstrip tracker of more than six million incorporated readout strips. It is made up of four cylindrical double layers and nine planar end-cap discs at each of the two detector ends. Having a readout strip every 80 μm on the silicon allows for a position resolution of charged particles of 17 μm and 580 μm per layer transverse and longitudinal to the strips, respectively.
The outermost subsystem of the ID is the TRT which is a gaseous detector. It consists of small drift tubes with a diameter of 4 mm that are made from a mixture of Kapton and carbon fibres. Inside each tube there is a gold-plated tungsten wire with a diameter of 31 μm. By applying a voltage of $\sim$ 1.5 kV on the outer wall and grounding the wire, each tube, filled with a gas mixture of 70 % Xe, 27 % CO_2 and 3 % O_2, serves as a proportional counter. The TRT's barrel region contains 52544 tubes, aligned with the beam axis over a length of 1.5 m and the end-caps contain 122880 tubes each. In between the tubes polymer fibres/foils are employed in the barrel/end-caps to induce transition radiation that may be emitted by charged particles propagating through the material. As this effect depends on the particle's E/m, it is strongest for electrons allowing to differentiate the latter from charged hadrons as part of the particle identification chain. While the spatial resolution of 120 μm is worse compared to the ability of the inner tracking detectors, this is kept in balance by the large number of hits per track of $\sim$ 30. Together with a high sampling frequency of the provided wire signals a time information on the nanosecond level can be achieved.

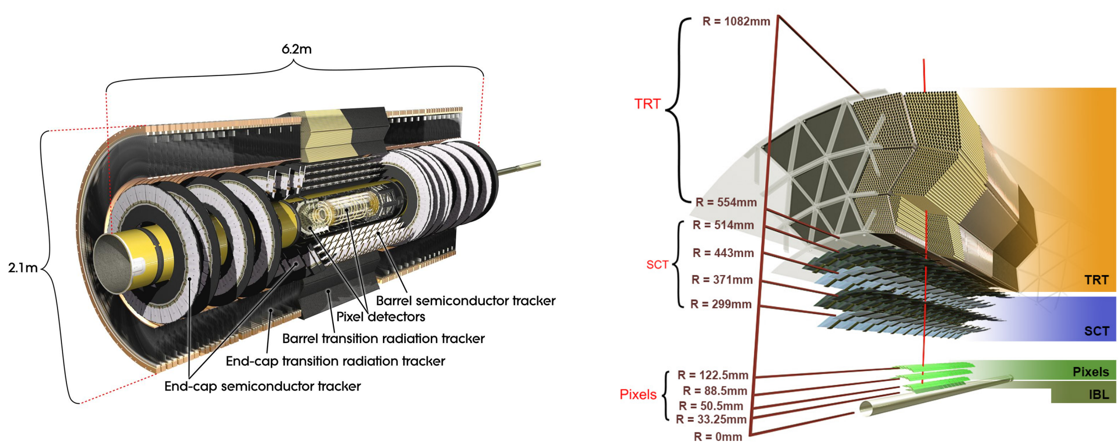


Figure 3.7: Left: Schematic drawing of the full Inner Detector comprising the pixel detector, the SRT and the TRT in the barrel and end-cap regions. **Right:** A close-up on the aforementioned subsystems in the barrel region.[42,43]

The Calorimeter

The calorimeter illustrated in Figure 3.8 is placed around the ID and the solenoid magnet system and is employed to measure a particle's full energy except for muons and neutrinos which hardly interact with the calorimeter material and detector, respectively. As the other detector components, the calorimeter consists of subsystems.[44]
The *Electromagnetic Calorimeter* (ECal) inside the ATLAS detector has two parts. One is covering $|\eta| < 2.5$ in the barrel and one is covering $2.5 < |\eta| < 3.2$ in the end-cap region. Inside an ECal the two effects *bremsstrahlung* and *pair production* compete. Qualitatively, a primary photon γ may enter the calorimeter material and converts into an e^+e^- pair in case of sufficient energy. The e^+e^- pair may interact with the material inducing bremsstrahlung which in turn produces photons again. With sufficient initial energy, this may lead to a cascade of γ and e^+e^--pairs, also referred to as electromagnetic *shower*. The length of the cascade can be expressed in units of X_0 termed *radiation length* which is a typical, material-dependent measure of this propagation. The interplay of the mentioned effects can produce up to 2^n particles after n radiation lengths X_0.
The ATLAS detector uses a calorimeter that comprises lead as high-density absorber material to ensure the particles deposit all their energy and liquid argon as active material in alternating layers. The charged particles crossing the active material may interact with it and produce light emissions that are proportional to the deposited energy. The calorimeter is segmented into small cells that are read out individually. This makes it further possible to associate the deposited energy with space points.

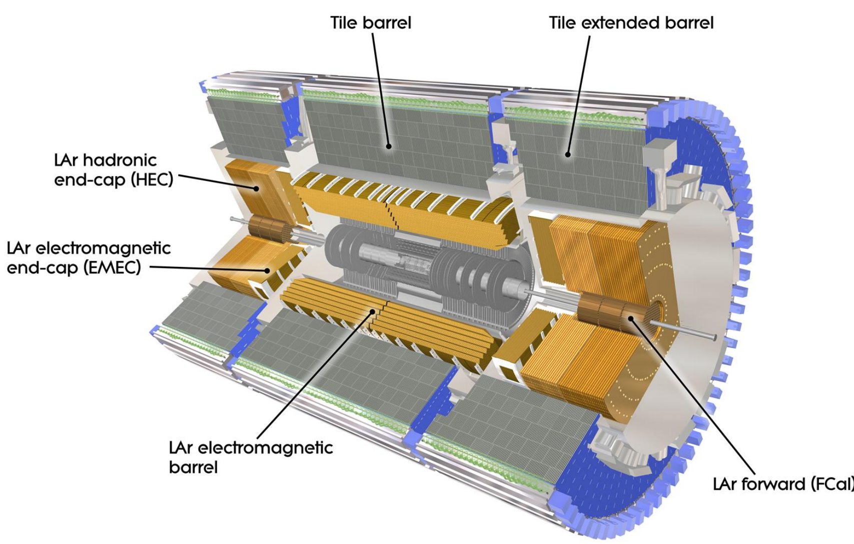


Figure 3.8: Schematic drawing of the electromagnetic and hadronic calorimeter components installed in the ATLAS detector.[45]

The *Hadronic Calorimeter* (HCal) at the ATLAS detector is wrapped around the ECal and is supposed to measure the energy of hadrons which deposit only parts of their energy in the ECal. Inelastic interactions between hadrons and the nuclei from the calorimeter material produce secondary hadrons and thereby a shower as discussed for the ECal.

The typical length scale for these interactions is given by the absorption length λ which is ~ 30 times larger than X_0 in case of lead. Thus the HCal is supposed to be larger than the ECal in order to avoid showers leaking out of the detector range.
The ATLAS HCal surrounding the ECal consists of steel as passive material and crystal scintillators as active component for the tile barrel region and for the tile extended barrel region. The HCal installed in the end-caps and the forward region (FCAL) comprises liquid argon as active material and copper and tungstem as passive material, respectively.

Muon Spectrometer

In contrast to electrons, muons are minimum ionizing particles[46] and deposit very little of their initial energy in the detector material due to their high mass. Therefore, a dedicated muon spectrometer is installed as outermost layer of the ATLAS detector as muons are the only interacting species that leave the calorimeter.

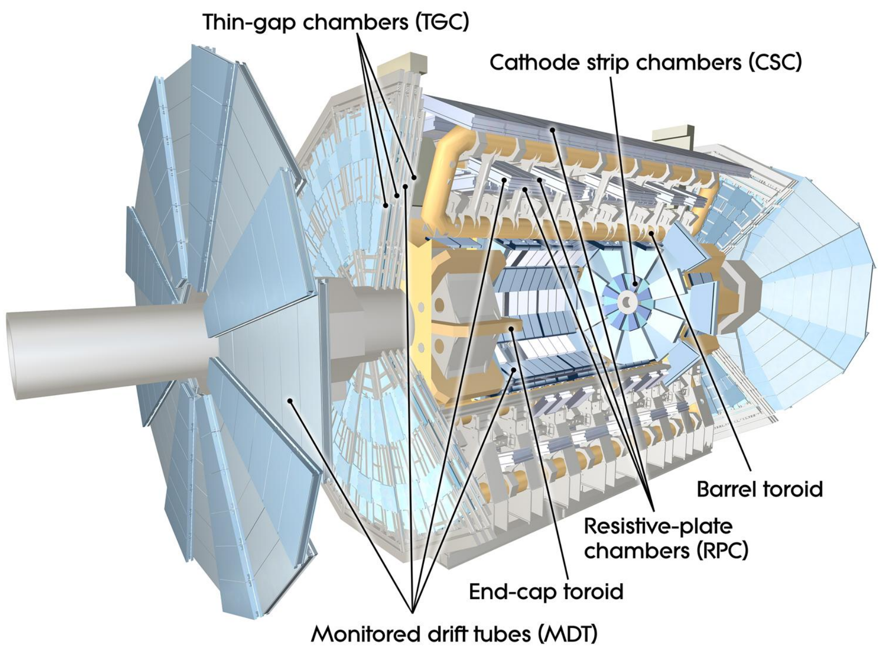


Figure 3.9: Schematic drawing of the muon spectrometer with its subcomponents, the RPC, the CSC, the TGC and the MDT.[47]

The in Figure 3.9 illustrated MS consists of multiple subcomponents: There are *Thin Gap Chambers* (TGC) that are employed for triggering interesting physics processes and for a second coordinate measurement at the ends of the detector. Furthermore, the *Resistive Plate Chambers* (RPC) equipped with an electric field of 5 $\mathbf{kVmm^{-1}}$ which are supposed to trigger events and to perform a second coordinate measurement in the central barrel region. Finally, three layers of *Monitored Drift Tubes* (MDT) measure the curves of the tracks with a resolution of 80 μm and the Cathode Strip Chambers (CSC) measure again the track's coordinates at the ends of the detector with a resolution of 60 μm.

Combining the Detector's Information

After having discussed the technical details of the detector and its submodules this section gives a brief overview on the importance of the linkage of these information in order to understand which kind of a particle has been observed as scatched in Figure 3.10. Most of the particles that are being produced in a particle collision are unstable i.e., have a finite lifetime. As the latter is often too short to reach the detector components, the collision products decay beforehand producing particles that eventually reach the detector and interact with it.
Hence, for most particles only the decay products are visible in ATLAS. A particle like the neutral pion π^0 for instance decays into two photons which would be detected in the ECal as discussed in the previous chapters while a particle decaying into an e^+e^- pair could be traced back to its position of decay with help of the information from the bent track seen by the Inner Detector and the deposited energy in the ECal.
Further, a proton emerging from the particle collision might have a bent trajectory, would interact with the ID and eventually be stopped in the HCal by depositing its energy. A muon can be tracked with the ID and the MS, whereas a neutrino can not be detected at all.
Thus combining the information from the multiple subdetectors is crucial to correctly identify and/or reconstruct a particle. Furher, combining all visible information makes it possible to infer about invisible parts of events such as non-interacting neutrinos that carry away energy.

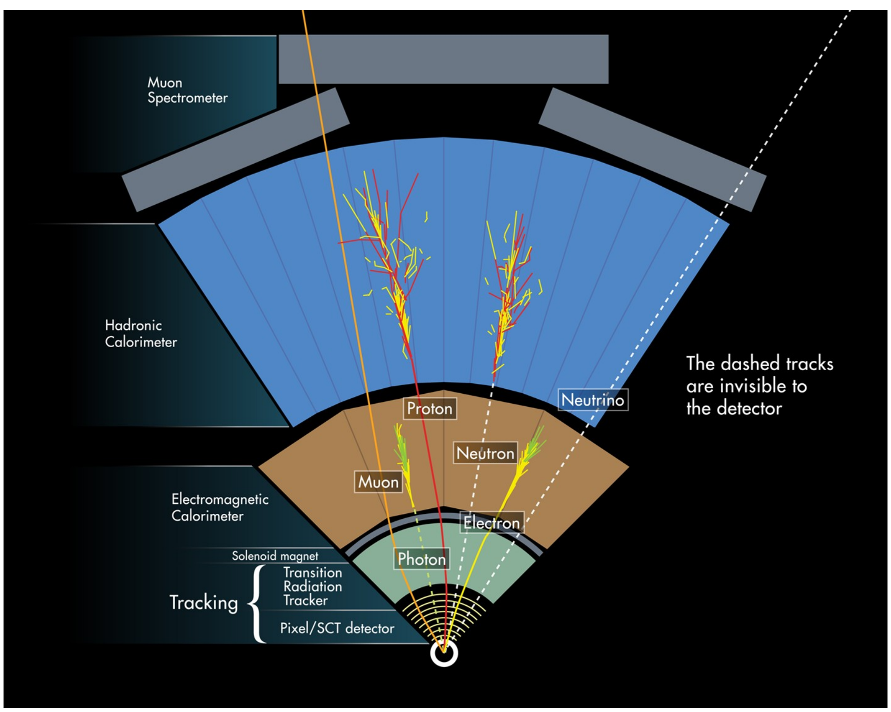


Figure 3.10: Schematic drawing of the cross section of the ATLAS detector summarizing the submodules and possible particle interactions with the material.[48]

The Trigger System

The high luminosity of the LHC as discussed in chapter 3.1, yields ~ 1 billion collisions per second. To the current day, there is no technology that is able to simultaneously process or to store the amount of data equivalent to recording every collision at such high rates. Thus, a system needs to be introduced that already picks only interesting events from a physics point of view in order to reduce the high event rates. The trigger system of ATLAS serves as such a decision maker.

It consists of a hardware-based first *level trigger* (Level-1) and a software based *High Level Trigger* (HLT). The Level-1 trigger reduces the event rate initially equivalent to the bunch crossing rate of ~ 30 MHz down to 100 kHz. The events accepted by the Level-1 trigger are read out and send to the HLT that further reduces the event rate from 100 kHz down to 1 kHz. The accepted events are written to disk to eventually use them in physics analyses.[49]

3.2.3 Event Reconstruction

If an event passes the above mentioned trigger system it is forwarded to reconstruction algorithms employed to reconstruct all possible physical details of the event including an identification of the participating particles. As discussed earlier it is reasonable to combine information from multiple subdetector systems, thus the algorithms usually rely on several inputs to reconstruct the particles as discussed in detail in the upcoming sections.

Reconstruction of Tracks and the Primary Vertex

The ID track reconstruction exploits local and global pattern recognition algorithms to connect measurements originating from the same charged particle. These measurements are fed into a track fitting tool to estimate the track's parameters. For charged particles the so called *perigee parameters* consisting of the transverse impact parameter d_0, the longitudinal impact parameter z_0, azimuth angle φ, polar angle θ and the ratio of charge and momentum q/p are extracted. From the information of the reconstructed tracks, their origin at beam line level can be determined. For a track to be considered in the determination of the primary vertex, certain requirements have to be fulfilled. Each candidate track needs to have $p_T > 400$ MeV and $|\eta| < 2.5$. Furthermore a sufficient number of pixel detector and SCT hits is required. The primary vertex search is carried out in the following steps:
A set of good tracks as defined above is chosen together with a seed position for the primary vertex. Vertex and tracks are used to determine a best fit, iteratively downweighting less compatible tracks. If the fit converges and at least two tracks are compatible with the found vertex, all tracks that are not compatible with the vertex are dropped to use them in the next vertex search.[50]

Electrons and Muons

In the analysis only leptons of flavor $\ell = e, \mu$ are considered. In contrast to e and μ, τ leptons can decay inside the beam pipe due to their short lifetime translating into a short mean free path of 87 μm and therefore introduce a deviating decay signature with a large variety of possible final states.[3]

Using the reconstructed track information from the ID and the shower profile from the ECal, electrons are identified in the central barrel region within $|\eta| < 2.47$ omitting $1.37 < |\eta| < 1.52$ corresponding to the transition region between the barrel and the end-caps. To take into account changes in the track curvature due to energy loss via bremsstrahlung, a *Gaussian Sum Filter* (GSF)[51] approach is used by ATLAS to improve the electron track parameter estimation. To ensure the electrons originate from the primary vertex it is required that every electron satisfies $|d_0/\sigma\,(d_0)\,| < 5.0$ and $|z_0 \sin\theta| < 0.5$ mm. Here $\sigma\,(d_0)$ is the uncertainty on the transverse impact parameter d_0 derived from the covariance matrix of the perigee parameters.

Muons are reconstructed based on information from the ID and the MS within $|\eta| < 2.5$. As for electrons, requirements on the impact parameters are set. For muons they are defined to be $|d_0/\sigma\,(d_0)\,| < 3.0$ and $|z_0 \sin\theta| < 0.5$ mm.

Lepton Isolation

The analysis described in chapter 5 deals with a purely leptonic final state with muons and electrons, where the leptons in the signal signature $WH \to WWW^* \to 3\ell\, 3\nu$ originate from three W bosons. These leptons tend to be well separated from other particles in terms of their angular distance ΔR. However, background leptons from e.g. in-flight decays of (heavy) hadrons that tend to have many close-by objects can also be reconstructed as lepton candidates for the analysis.
Hence, discriminants are constructed using additional tracks from the ID and energy deposits from the ECal in a certain cone ΔR around the lepton candidate. While leptons originating from W bosons are expected to have a rather clean environment and are therefore well *isolated*, leptons from heavy-flavour decays are expected to have more tracks or energy in their vicinity due to hadronization, i.e. are less isolated. So called isolation working points are defined to create measures on the purity requirement in the cones around the leptons. There are two types of isolation. First, the track isolation that is based on the sum of the transverse momenta of all tracks within a cone of radius ΔR around the lepton track, excluding the latter. The track isolation $p_{T,\,\text{varcone},\,\Delta R}$ is defined with a variable cone size of

$$\Delta R = \min\left(\frac{10\text{ GeV}}{p_T}, R_{max}\right) \tag{3.14}$$

with p_T of the main track and R_{max} is the maximum cone size of typically 0.2, 0.3 or 0.4.[52]
Second, the calorimeter isolation that follows an equivalent idea to the track-based isolation by summing over all energies that are deposited in clusters within a cone of radius ΔR around the cluster of the lepton candidate, excluding the latter. The energy-based isolation variable is referred to as $E^{\text{topo}}_{T,\,\text{cone},\Delta R}$.

Jets

The jet reconstruction is based on an algorithm that combines calorimeter clusters. To this end the *anti-k_T* algorithm[53] is used with a radius parameter of R = 0.4 to reconstruct jets from this region. The latter algorithm is further based on the distance measures d_{ij} describing the distance between particle denoted with index i and j and the distance d_{iB} that is the distance between particle with index i and the beam B:

$$d_{ij} = \min(k_{ti}^{-2}, k_{tj}^{-2})\frac{R_{ij}^2}{R^2}, \tag{3.15}$$

$$d_{iB} = k_{ti}^{-2}. \tag{3.16}$$

Here, $R_{ij}^2 = (y_i - y_j)^2 + (\varphi_i - \varphi_j)^2$, while R, k_{ti}, y_i and φ_i are the fixed radius parameter, the transverse momentum, the rapidity and the azimuth angle of the particle indexed by i. Jets are required to satisfy p_T > 30 GeV within $|\eta|$ < 2.5.

For jets originating from a b-quark there is a dedicated *multivariate analysis* (MVA) based identification technique. The MVA approach called DL1r[54] uses input variables related to the distinct mean life time of the b-quark that makes it possible for the latter to travel a significant amount of time away from the interaction point which would introduce a secondary vertex for the reconstructed particles from the heavy-flavor jet. B-jets are accepted beyond p_T > 20 GeV and within $|\eta|$ < 2.5.

Missing Transverse Energy

Due to the fact that neutrinos can not be detected by the detector they can carry away significant amounts of energy that is not deposited in the detector. The transverse component $\vec{E}_T^{miss}$ can be calculated on event-by-event basis after Equation 3.4. For the actual calculation, all fully reconstructed particles, all high p_T hadronic jets as well as all energies from non-associated clusters or tracks with p_T > 500 MeV which satisfy a set of quality requirements contribute.[55]

3.3 Monte-Carlo and Detector Simulation

In high-energy physics experiments simulations are often consulted as a tool to guide the analysis through decisions on e.g. how to enhance signal-to-background ratios in the real data or to give estimates on the rate and nature of events that can be expected to be obtained from the data.

From the theoretical knowledge of the SM predictions, cross sections as measures for the probability of certain physics reaction and event kinematics can be derived. Implementing these predictions into software is often done in the scope of Monte-Carlo-simulations (MC-simulations). The basic idea is to stochastically generated pseudo-random experiments that, if done often enough, eventually converge towards distributions that can be expected from theoretical predictions. Besides cross sections and initial and final state event kinematics, the generators also serve to model gluons radiating other gluons which eventually creates particle cascades termed *parton shower* and the process of hadronization described in the theory chapter.

The purely kinematic distributions (MC truth) are applied to the detector response by folding them with the machine's sensitivities and resolutions. This is based on software that specializes in the interaction between particles and anorganic materials like the ones mentioned in the detector chapter. A typical software that simulates the interaction of particles with matter is e.g. GEANT4.[56] MC truth events that are folded with the detector response are referred to as detector level MC events. These are then reconstructed using the same algorithms that are used for data obtained from the detector.

4 | Deep Neural Networks in Physics Analyses

4.1 Neural Networks

This chapter is dedicated to discuss the machine learning approach used in the analysis together with certain applied techniques. First, the theoretical backbone of neural networks termed *backpropagation* is explained followed by the discussion of a common issue in machine learning called *overtraining* and a common practice of monitoring this phenomenon referred to as *cross-validation*. Finally, an overview of used techniques in the scope of the framework Keras[57] is given.

4.1.1 Theoretical Introduction into Artificial Neural Networks

The structure of an artificial neural network (ANN)[1] is conceptually based on the makeup of a biological brain. It consists of a set of so called artificial neurons which are organized in layers. The artificial neurons model the propagation of a signal through the network. Except for the neurons in the first layer, generally each neuron is connected to all neurons in the previous layer via an activation function. The activation of a neuron is the sum over all activations of neurons in the previous layer, each multiplied by a respective weight. Additionally, a bias is added to the sum and the sum is encapsulated in a differentiable activation function σ_a that typically maps the input values onto an interval (0,1) in order to numerically smooth the training of the neural network. The activations for the neurons in the first layer are given by the inputs to the neural network, hence the first layer is called input layer. The final layer is called output layer and the layers in between are called hidden layers as illustrated in Figure 4.1. The activation can be calculated as

$$a_i^{(l)} = \sigma_a\left(z_i^{(l)}\right) = \sigma_a\left(\sum_j w_{ij}^{(l)} a_j^{(l-1)} + b_i^{(l)}\right), \tag{4.1}$$

where $a_i^{(l)}$ denotes the activation of the i^{th} neuron in the layer with the index (l) and $b_i^{(l)}$ labels the corresponding bias. $w_{ij}^{(l)}$ is the weight connecting the j^{th} neuron in layer $(l-1)$ to the i^{th} neuron in layer (l).

[1]The terms artificial neural network and deep neural network (DNN) are used interchangebly in this thesis.

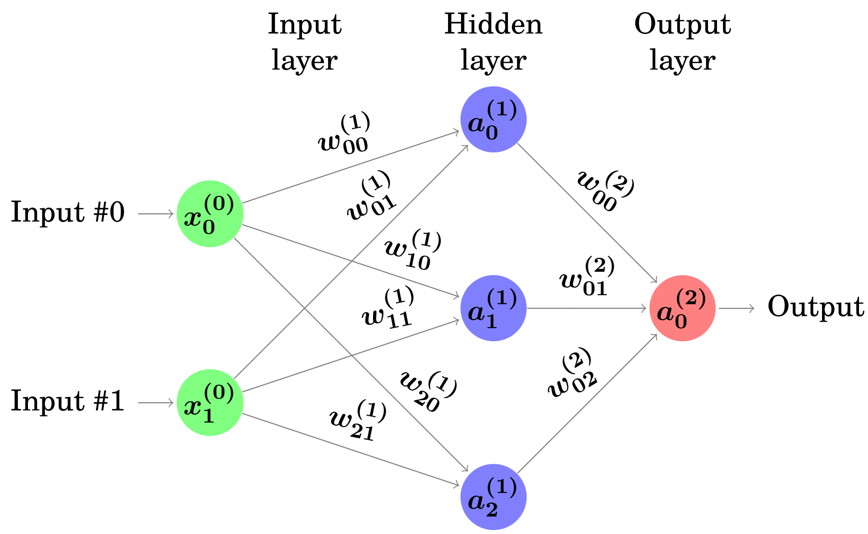


Figure 4.1: Exemplary sketch of the input and output layer of an ANN architecture with the indexing as described in the section on the theory of backpropagation.[58]

To evaluate the performance of an ANN the cost function is introduced. The *cost* or *loss function* measures the deviation of the desired output to the actual output of the network. In the first part of this work the binary cross entropy loss function is used:

$$C = -\sum_j \left[E_j \ln \left(a_j^{(n)} \right) + (1 - E_j) \ln \left(1 - a_j^{(n)} \right) \right] \tag{4.2}$$

where E_j denotes the desired result of the j^{th} neuron in the output layer (n).

Before the network is of any use it has to be trained. Training the network means tweaking its weights and biases so that its output matches the desired output on a given input. To achieve this the cost function has to be (locally) minimized. This is the case when the gradient of the cost function with respect to the weights and biases approaches zero. Therefore in every step of the training, the weights and biases are adjusted in opposite direction of the gradient of the cost function. This derivative is further multiplied by the learning rate η_0 to ensure the stability of this numerical minimization. Following the derivation in appendix A.1 the correction of the weights and biases per training step is:

$$\Delta w_{ij}^{(l)} = -\eta_0 \, \frac{\partial C}{\partial w_{ij}^{(l)}} = -\eta_0 \cdot \delta_i^{(l)} \cdot a_j^{(l-1)} \tag{4.3}$$

$$\Delta b_i^{(l)} = -\eta_0 \, \frac{\partial C}{\partial b_i^{(l)}} = -\eta_0 \cdot \delta_i^{(l)} \tag{4.4}$$

where $\delta_i^{(l)}$ is termed *output error* that comprises a deviation of the desired result E_j and the activation value in the output node. It should be noted that the structure of the result would not change in case of a different loss function. The only adaption has to be done within the output error $\delta_i^{(l)}$. Furthermore, the calculation is valid for any activation function that is applied or tested in this work as it has not been specified in the derivation in the appendix. Finally, the derivation is based on a gradient descent approach. As this method is computationally heavy, modern machine learning frameworks usually use optimization algorithms based on stochastic gradient descent methods[59] with optional adaptive estimations of first-order and second-order moments.

4.1.2 Overfitting and Underfitting

Overfitting or overtraining describes the sensitivity of e.g. a neural network with respect to the statistical variance in the data set. If it is not taken care of the effect, an ANN can tend to see statistical fluctuation as part of the underlying structure in the data and therefore models the training data too well. This leads to the adaption of arbitrary structures which in turn can cause failures when applying the thereby evaluated model on new test data. An illustration of the described behavior referred to as overfitting is given in Figure 4.2 (green line).

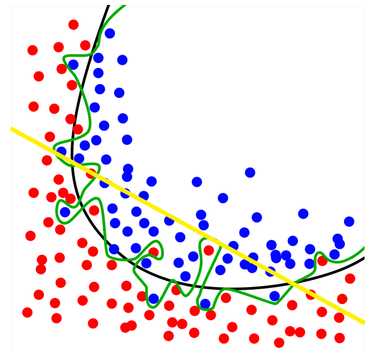


Figure 4.2: Example of overfitting represented by the green line. The model has adapted the variance of the data and might fail on a different validation sample.[60]

Underfitting or undertraining can be understood as, roughly speaking, the opposite of overfitting: in the case of underfitting, the ANN is not capable to accurately reproduce details in the distribution of the data. In contrast to overfitting, the ANN can neither find a good model while training nor generalize to new data which might lead to a low performance in the prediction. As an example one can think of trying to fit data following a non-linear trend with a linear model. An illustration of the latter example is given in Figure 4.2 (yellow line).

Overfitting can occur in the process of *hyperparameter optimization* which is the process of evaluating different settings for a neural network. The parameters can be set to optimally perform on one training data set available. However, this does not guarantee a good performance on unknown data sets in general as it might lead to implicitly introducing information about the training data in the model. One possible method to monitor this problem is a technique called cross-validation.

4.1.3 K-Fold Cross-Validation

Using k-fold cross-validation, the full data set is split into a training data set and a validation data set. The latter is never going to be used for training purposes. The training data set is further split into k *folds* as illustrated in Figure 4.3. Next, the neural network is trained on $k-1$ folds that make up the internal training set (all indicated in blue) and is evaluated with some performance measure on the remaining internal test set (indicated in green) for each of the k folds. As the internal test set changes each split (Split 1, 2 and 3 in the exemplary Figure 4.3 given) an average performance measure can be calculated over all splits.

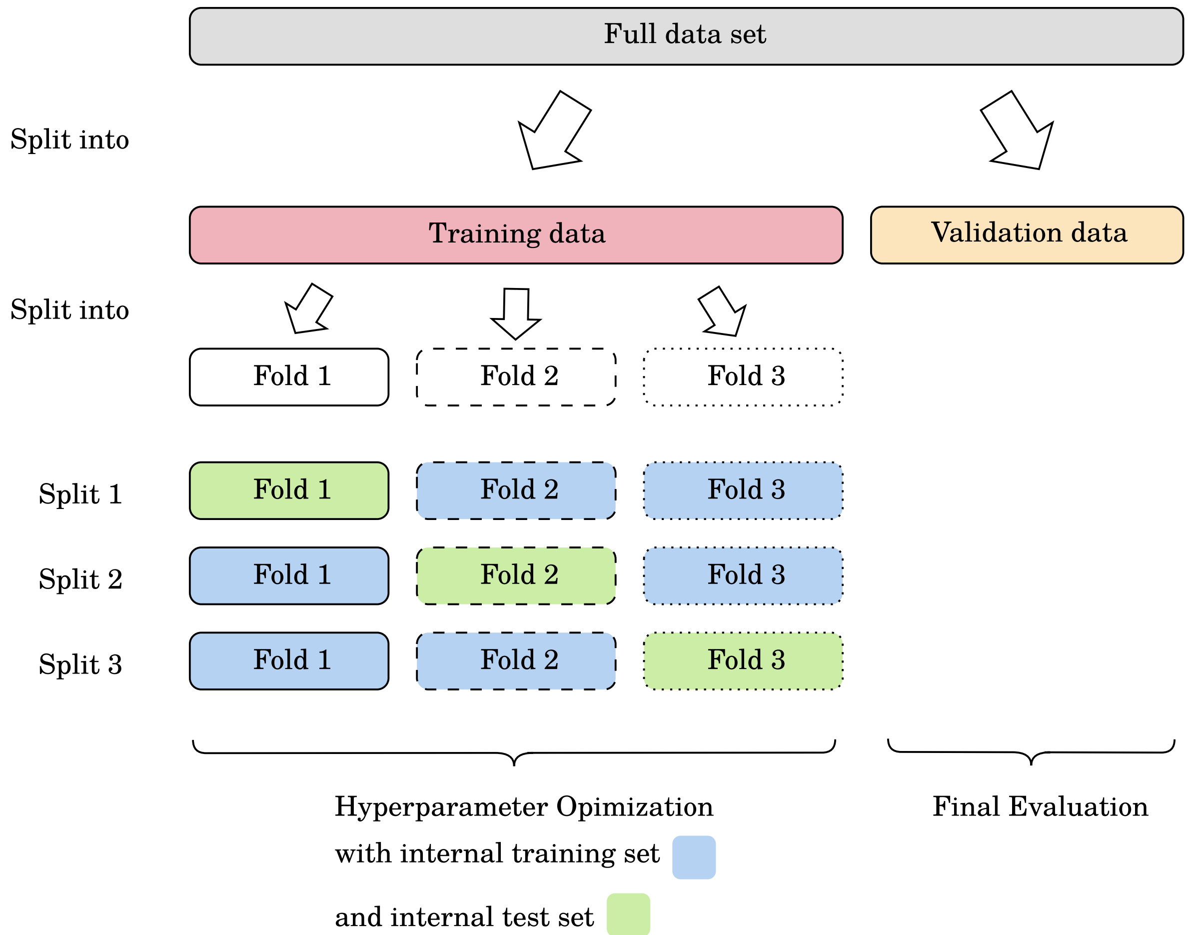


Figure 4.3: Schematic representation of k-fold cross-validation. The full data set is split up into training data and test data. The training data is then split up into three folds, two are used for training and the remaining one is used for an internal test.

4.2 Machine Learning Concepts and Keras

Keras[57] is a Python deep learning library. As an application programming interface (API) of the machine learning framework TensorFlow[61], it is used to implement neural networks. In the following all structures and functions of Keras used in this work are briefly explained. The specific settings chosen for each of the function is finally discussed in the analysis section. The core concepts of every machine learning approaches using neural networks can be split up into three parts:

Creating a Model

An ANN model consists of an input layer, hidden layers and an output layer which define the model's architecture. The size of the input layer is defined by the number of input variables fed into the network.

While the hidden part of the network can consist of an arbitrary number of layers with an arbitrary number of neurons per layer, the output layer's dimension is defined by the problem to solve (e.g. binary, multiclass or multilabel problems, etc.). As discussed in the previous section, every neuron has an activation function, the latter is defined layerwise in Keras. The API provides a wide range of commonly used activation functions by the domain, further details are given in the analysis chapter.

Additionaly, Keras provides so called regularization layers that are employed to prevent the network from overtraining. Besides layers that fold their inputs with e.g. a normal distribution, another prominent method is *dropout*: since the network adjusts all weights per training circle in order to minimize the loss/cost function, weights of neurons have to adapt for features in the training data. This adaption can influence neighboring neurons in their learning process such that they rely on the neuron capable of learning the mentioned feature of the data set. This behavior is called complex co-adaption and can lead to overfitting. To prevent from this, dropout is employed to set a choosable fraction of randomly chosen neurons and their corresponding weights to 0 (compare Figure 4.4) during the back propagation process in each training step. Doing this, the remaining neurons have to react on the missing neurons which decreases the complex co-adaption and thereby prevents from overfitting. It should be noted that dropout is only active at training time.

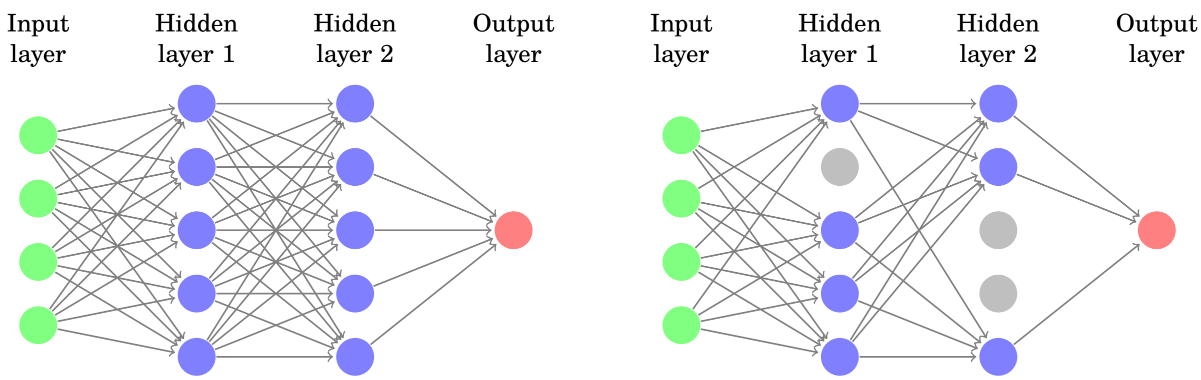


Figure 4.4: Left: Ordinary ANN with four input neurons, two hidden layers and an output neuron. **Right**: ANN with dropout applied. The neurons in gray are dropped and therefore not considered during training.

Preprocessing

Fully connected neural networks usually have a more stable training process and therefore perform better when the components of the input data are mapped onto the same range or order of magnitude, preferably $\mathcal{O}(1)$. This can be achieved by removing the components mean and scale. If the component suffers from outliers, more robust transformation algorithms can be applied as shown in the analysis chapter.

Training the Neural Network

During the training procedure of a neural network the weights and biases are tweaked in each training step. A training step can be defined in terms of the so-called *batch size*. The latter is a hyperparameter defining the number of samples propagating through the network before the weights and biases are updated accordingly. A common approach is to set the batch size to 1 < batch size < size of training set.
A second hyperparameter is a so-called *epoch*. It is defined as one full iteration through the whole training set. Thus, one epoch usually consists of multiple batches. After each epoch the network is evaluated on the held out test data by predicting on the samples and comparing with the truth information. The average deviation can be calculated based on a provided loss function. The corresponding loss can further be monitored to check for overtraining. If the loss of the test set exceeds the loss of the training set in any epoch, the network might have been adapting too much on the training data and lost the ability of generalization. Hence, the comparison of training and test loss over all epochs can be a good first indication of a network's capability of prediction on unknown data sets. The monitoring process is not restricted on the loss function. In principle an arbitrary number of performance measures can be calculated after each epoch.
Furthermore, it is possible to prune the training time by requiring an improvement of a certain performance measure with e.g. every epoch. As soon as this does not improve anymore, the training would be stopped. This technique is called *early stopping*. If the true result is known and can be compared with the network's output, as it is the case in this work, the procedure is referred to as *supervised* training.

4.3 Performance Measures

In physics analyses neural networks are usually employed to classify an event as signal or background. The overall performance can e.g. be measured as the fraction of correctly classified events. In a two-class problem e.g. signal (positive class) against the most prominent background (negative class) there are four possible results for every event. If the signal is classified as positive class, the result is termed *true positive* while a misclassified signal event is termed *false negative*. On the other hand, correctly classified background events are referred to as *true negative* and misclassified background events are called *false positive*. These four quantities can be combined in numerous ways to construct figures of merit. A figure of merit that is used in this analysis is the area under the *Receiver Operating Characteristics Curve* (ROC) curve.

Receiver Operating Characteristics Curve (ROC Curve)

The ROC curve relies on the the *background efficiency* which is the fraction of correctly classified background events as a function of the *signal efficiency* which is the fraction of correctly classified signal events. In the ideal case of a perfect classifier the ROC's *area under the curve* (AUC) is supposed to be 1. This area gives a first qualitative measure of the performance and is especially helpful to compare different classifiers on the same data sample.

Significances

The experiment which the analysis is based on, is a counting experiment. With a generally high number of events and a low probability of the signal to occur, the Poisson approximation can be applied to describe e.g. the estimator of the significance[62]

$$Z_p = \sqrt{2\left(n \ln\left[\frac{n(b+\sigma_b^2)}{b^2+n\sigma_b^2}\right] - \frac{b^2}{\sigma_b^2}\ln\left[1+\frac{\sigma_b^2(n-b)}{b(b+\sigma_b^2)}\right]\right)} \tag{4.5}$$

with $n = s + b$ where s is the number of signal events and b is the number of background events. σ_b is the MC uncertainty on the background. In case of a vanishing uncertainty, the expression simplifies to

$$Z_p = \sqrt{2\left((s+b)\ln\left[1+\frac{s}{b}\right] - s\right)} \tag{4.6}$$

5 | Analysis in the Channel $WH \to WWW^* \to \ell\nu\ell\nu\ell\nu$

The following analysis that aims to extract a significance with respect to the WH process based on MC data at $\sqrt{s} = 13$ TeV with an integrated luminosity of 139 fb^{-1} is divided into several steps. After the definition of the preselection criteria with associated signal and control regions, two neural networks are developed to suppress the most dominant background processes $t\bar{t}$ and WZ in the region of interest. The information that is gained from the two individual studies is combined in a subsequent section that deals with the further development of a multiclass neural network (*multiclassifier*) that is employed to suppress the most dominant background process simultaneously. Based on the multiclassifier output, a dedicated signal region is defined. The events that pass this step are considered in the statistical interpretation of calculating the significance.

The obtained results are compared to the previous WH analysis[63] based on data collected during 2015 and 2016 at $\sqrt{s} = 13$ TeV with an integrated luminosity of 36.1 fb^{-1}. The analysis employed Boosted Decision Trees (BDTs) as multivariate analysis technique. BDTs consist of decision trees that split input data accordingly to the single values of the input variables at so called nodes to maximize the information gain. A tree terminates in so called leaves whose output can be interpreted as class probability. A more detailed introduction on BDTs and the procedure of boosting can be found in e.g. Ref.[64] or Ref.[65]. A detailed discussion on the hyperparameter optimization for the previous analysis can be found in Ref.[66]. To put it in a nutshell, two individual Boosted Decision Trees were used to suppress the most dominant backgrounds $t\bar{t}$ and WZ.

5.1 Background Processes

The analysis aims to extract the signature $WH \to WWW^* \to 3\ell\ 3\nu$. Besides Higgs-including decay channels like ggF, VBF and $t\bar{t}H$, there are multiple diboson, triboson, and top-quark as well as Z+jets decay processes that can form a trilepton final state. An overview of the considered processes can be found in Table 5.1. From the diboson processes, $WZ/W\gamma^*$ is the most dominant background processes due to exactly three leptons in the final state. In the category of triboson processes (VVV), WWW has the most similar signature due to three leptonically decaying W bosons, therefore significantly contributing to the irreducible background. Regarding processes including top-quarks, the $t\bar{t}$ process is the most dominant one. Here, so called *non-prompt* or *fake* leptons that originate from jets can be reconstructed as third lepton. Same applies for processes like Z+jets, where two leptons originate from the Z decay and a third lepton is faked by a jet. Diboson processes including more than and triboson processes including at least one Z boson such as ZZ^* or WWZ etc. can get accepted in the analysis in case one lepton is not detected if it e.g. propagates very close to the beam line.

Process	MC Generator
Higgs	
VH ($H \to WW^*$)	POWHEG-BOX[67,68,69]+Pythia8[70]
VH ($H \to \tau\tau$)	POWHEG-BOX+Pythia8
$gg \to H$ ($H \to WW^*$)	POWHEG-BOX+Pythia8
VBF ($H \to WW^*$)	POWHEG-BOX+Pythia8
$t\bar{t}H$ ($H \to WW^*$)	MadGraph5[71]+Pythia8
Diboson	
$WZ/W\gamma^*$, $ZZ/Z\gamma^*$, WW^*	Sherpa 2.2.2[72]
Triboson (VVV)	
WWW, WWZ, WZZ, ZZZ	Sherpa 2.2.2
Top	
$t\bar{t}$	POWHEG-BOX+Pythia8
$t\bar{t}V$	MadGraph5+Pythia8
Single Top:	
Wt	POWHEG-BOX+Pythia8
tZ	MadGraph5+Pythia8
V+jets	
Z+jets	MadGraph5+Pythia8

Table 5.1: MC simulated processes that are taken into account in the analysis. The left column names the process while the right column states the MC generator. The VH ($H \to WW^*$) process where $V = W$ is treated as signal. Any other appearing process is treated as background.

5.2 Preselection and Signal Region Criteria

The preselection scheme includes a set of criteria to identify event candidates that enter the final analysis. Every event needs to fulfill the following requirements:

- **Number of leptons: nLep = 3**,
  since the WH final state of interest consists of three reconstructed leptons.
- **Sum of lepton charges:** $\sum_{i=0}^{2} q^{\ell_i} = \pm 1$ e,
  to account for the charge of the associated $W^{\pm}$ boson.
- **Transverse lepton momentum:** $p_T^{\ell_i} > 15$ GeV
  Threshold on the lepton transverse momentum to reduce low momentum background contamination by fake leptons.
- **Isolation Working Point: FCTight**[73]
  Events with three reconstructed leptons that satisfy the following optimized calorimeter and track isolation criteria are accepted:

	$\frac{E_{T,\,\text{cone},\,20}^{\text{topo},\ell}}{p_T^{\ell}}$	$\frac{p_{T,\,\text{varcone},\,30}^{\ell}}{p_T^{\ell}}$
Electrons	< 0.06	< 0.06
Muons	< 0.15	< 0.04

  The isolation cut is applied to suppress fake leptons from jets that have more particles in their vicinity compared to leptons from W bosons. Electrons (< 0.06) are required to be more isolated in the calorimeter than muons (0.15) due to their high energy loss. The track isolation requirement is similar for muons (< 0.04) and electrons (< 0.06).

A preselection-based signal region (SR) in principle can be defined by requiring a final state including three charged leptons of flavor e or μ and a total charge of ± 1 e to account for the charge of the W boson. The so defined signal region can be divided into three distinct regions of 0, 1 and 2 *same flavor opposite sign* (SFOS) lepton pairs as summarized in Table 5.2. The flavor corresponds to the lepton type while the sign is the sign of the lepton's charge.

SR1 (2 SFOS)	**SR2 (1 SFOS)**	**SR3 (0 SFOS)**
$e^+ e^- e^-$	$e^+ e^- \mu^-$	$e^+ e^+ \mu^-$
$e^- e^+ e^+$	$e^- e^+ \mu^+$	$e^- e^- \mu^+$
$\mu^+ \mu^- \mu^-$	$\mu^+ \mu^- e^+$	$\mu^+ \mu^+ e^-$
$\mu^- \mu^+ \mu^+$	$\mu^+ \mu^- e^-$	$\mu^- \mu^- e^+$

Table 5.2: Possible flavor-charge-combinations for the three signal regions. Any cyclic permutation is allowed as well.

This analysis deals with region SR3 including 0 SFOS that is also referred to as *Z-depleted* signal region. It is expected to have the highest signal to background ratio (S/B) due to the additional suppression of lepton pairs originating from Z decays as included in SR1 and SR2 which are also referred to as *Z-dominated* signal regions.

To avoid any naming ambiguity of the three reconstructed final state leptons the following enumeration is applied:
The lepton of unique charge that originates from one of the W bosons from the Higgs decay is termed ℓ_0. As explained in chapter 2.1.3, 2 out of 3 spin correlations states force the second lepton that originates from the Higgs decay to propagate into the same direction as ℓ_0. Thus, the reconstructed lepton of opposite charge that has a minimum angular distance ΔR to ℓ_0 is termed ℓ_1 while the remaining reconstructed lepton that fulfills the total charge and isolation requirements is termed ℓ_2. The enumeration is illustrated in Figure 5.1.

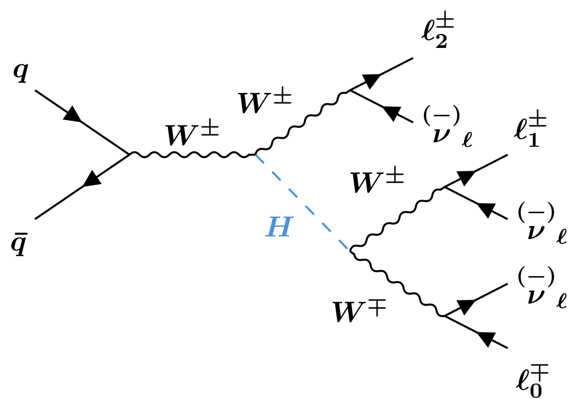


Figure 5.1: Feynman diagram including the enumeration scheme used in the analysis of the leptons from the WH process.

Figure 5.2 summarizes the MC truth origin information of the three lepton candidates in the WH MC in SR3. In more than 99% of the weighted events, ℓ_0 stems from the Higgs decay, while ℓ_1 originates almost 86% from the Higgs and in more than 13% of the weighted events from the associated W boson. For ℓ_3 it is the opposite case.

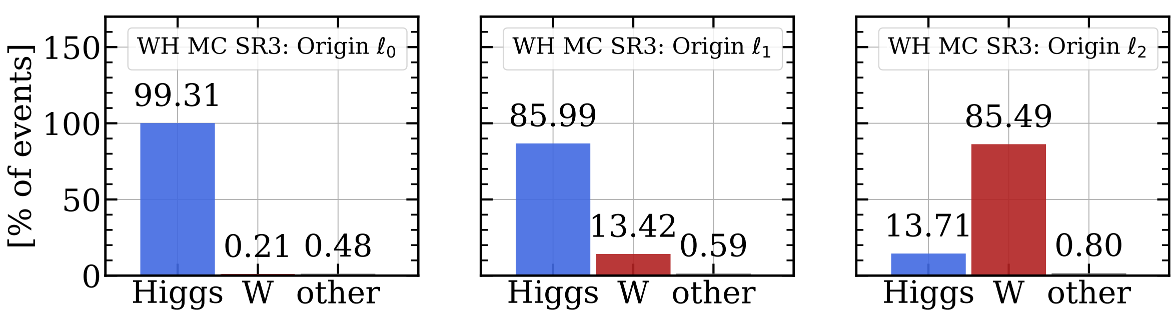


Figure 5.2: Origin information of three lepton candidates in the WH MC in SR3 distinguished between Higgs, W boson and other. The latter category results if the origin of the reconstructed particle that matches with the truth particle is determined to be neither Higgs nor W boson.

Subsequently applying the preselection and SR3 cut scheme to the MC events listed in Table 5.1 yields the number of MC events summarized in the upper Table 5.3 (referred to as *cutflow* table). With a time-integrated luminosity of 139 fb^{-1}, the simulations predict roughly 17 WH events against 529 background events in the Z-depleted signal region. The column "Other Higgs" includes any process that involves a Higgs boson namely ZH, VBF, ggF, $t\bar{t}H$ and $H\rightarrow \tau^+\tau^-$ which sum up to almost 25 MC events. With a total of about 379 MC events, top quark related decay channels dominate the background composition with close to 72%. From all top backgrounds it is the $t\bar{t}$ process with 279 MC events that induces the majority of events. The Diboson process $WZ/W\gamma^*$ contributes with roughly 70 events and the triboson processes VVV amount to roughly 24 MC events. All top backgrounds together with $WZ/W\gamma^*$ and VVV make up approximately 90% of the background.

$\sqrt{s} = 13$ TeV $\mathcal{L} = 139\ \mathrm{fb}^{-1}$ (Full Run 2)	WH	Other Higgs	$WZ/W\gamma^*$	VVV	$t\bar{t}$	$t\bar{t}V$	Single Top	Total Bkg.	Sign.
Preselection	69.82 ± 0.11	194.82 ± 0.68	25021.03 ± 36.14	139.92 ± 0.72	1118.54 ± 6.60	1202.14 ± 2.67	422.73 ± 3.39	44810.84 ± 442.40	0.14 ± 0.00
SR3: 0 SFOS	17.01 ± 0.06	24.91 ± 0.21	69.70 ± 1.76	24.35 ± 0.34	279.17 ± 3.30	77.89 ± 0.76	21.92 ± 1.71	529.05 ± 9.88	0.68 ± 0.01

$\sqrt{s} = 13$ TeV $\mathcal{L} = 139\ \mathrm{fb}^{-1}$ (Full Run 2)	WH	Other Higgs	$WZ/W\gamma^*$	VVV	$t\bar{t}$	$t\bar{t}V$	Single Top	Total Bkg.	Data/MC
Preselection	69.82 ± 0.11	194.82 ± 0.68	25021.03 ± 36.14	139.92 ± 0.72	1118.54 ± 6.60	1202.14 ± 2.67	422.73 ± 3.39	44810.84 ± 442.40	0.91 ± 0.01
$t\bar{t}$ CR	0.94 ± 0.01	10.23 ± 0.08	57.60 ± 0.97	2.75 ± 0.11	167.82 ± 2.53	63.44 ± 0.66	12.71 ± 1.19	338.36 ± 6.01	1.08 ± 0.06
WZ CR	6.60 ± 0.03	27.68 ± 0.31	6460.54 ± 21.92	27.30 ± 0.27	37.13 ± 1.21	6.48 ± 0.18	23.15 ± 0.92	7142.28 ± 67.41	0.98 ± 0.01

Table 5.3: Top: Event yields with statistical uncertainties at preselection level and in the Z-depleted signal region. The row named "Preselection" includes events that pass the preselection cuts, the row below includes the subset of events that further satisfy the 0 SFOS requirement. These MC events are used for the nominal analysis. The column "Other Higgs" includes any other process involving a Higgs boson namely ZH, VBF, ggF, $t\bar{t}H$ and $H \rightarrow \tau^+\tau^-$ events. These are treated as background as well. The last column shows the Poisson significance taking into account the total background. **Bottom:** Event yields with statistical uncertainties at preselection level as well as in the $t\bar{t}$ and WZ control region. The last column shows the data to MC yield ratio.

5.3 Control Regions

Control Regions (CR) are phase space regions in which one of the background processes usually dominates any other process with a huge excess of events. They can be employed to perform data-MC shape comparisons in order to monitor possible MC-mismodelling and to extract normalization factors to account for discrepancies between data and MC yields.

For this analysis one control region for WZ and one control region for $t\bar{t}$ is implemented. To gain higher event yields, the regions are defined by requiring 1 or 2 SFOS lepton pairs. The scheme starts at the preselection level for both WZ and $t\bar{t}$. To receive an as pure as possible WZ background with three prompt leptons one requires at most 1 jet and no b-jets both within a narrow window of 25 GeV around the difference between the Z boson mass m^Z and the invariant mass of same-sign candidate leptons m_{inv}^{SFOS} to include the majority of Z decays from the WZ process. The $t\bar{t}$ region is defined outside of this invariant mass window to ensure orthogonality to the WZ region. At least one jet is required to account for initial and final state radiation and exactly 1 b-jet is required to account for the expected jets from heavy-flavor decays. For events from both regions it is further required to pass a low invariant mass and missing transverse energy threshold.

The MC event yields for the two control regions are given in bottom Table 5.3. In the $t\bar{t}$ control region the latter process dominates the event yields with roughly 168 MC events corresponding to almost 50% of the 338 remaining MC events. The second most dominant background is the very similar process $t\bar{t}$V with 63 MC events. Furthermore, a non-negligible amount of 58 WZ MC events enter the $t\bar{t}$ control region. The data to MC ratio of 1.08 ± 0.06 indicates a good agreement between data and MC yield with a slight tendency towards a higher amount of data.

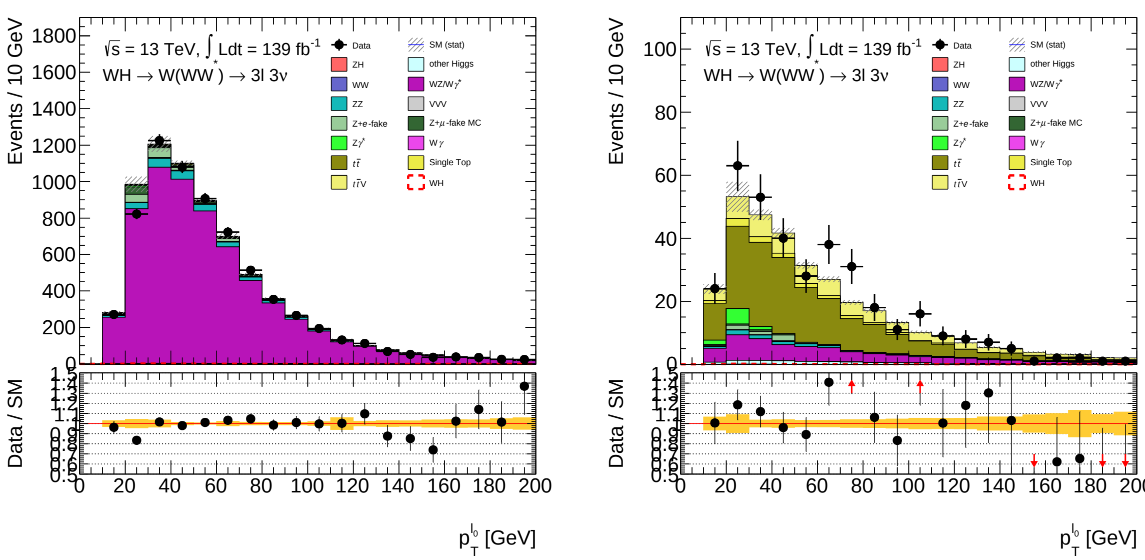


Figure 5.3: Left: Distribution of the transverse momentum of ℓ_0 in the WZ CR. **Right:** Distribution of the transverse momentum of ℓ_0 in the $t\bar{t}$ CR

In contrast, the WZ control region shows a very good enhancement of WZ MC events with a purity of more than 90% corresponding to 6461 MC events. Furthermore, the data to MC ratio of 0.98 ± 0.01 indicates a very good agreement between data and MC yield slightly overestimating the number of MC events.
Figure 5.3 exemplary illustrates the transverse momentum of lepton ℓ_0 in the WZ control region (left) and the $t\bar{t}$ control region (right). A good data to MC shape agreement can be observed in the case of the WZ control region. Even though not at comparably high rates as WZ, it becomes further visible that processes such as ZZ^* and Z+jets are likely to enter the WZ CR. This can be explained by the requirements of 1 or 2 SFOS pairs and the cut on the narrow window around the Z peak with an additional fake lepton from a jet or a missing lepton that is not detected in case of ZZ^*. The transverse momentum distribution of ℓ_0 in the $t\bar{t}$ control region shows fluctuations yet an overall good agreement between the distribution of data and MC. A full cut scheme is depicted in Figure 5.4. Additionally the the Z-depleted signal region SR3 is shown for completeness.

Preselection		
nSFOS $\geq$ 1		nSFOS = 0
$n_{\text{Jet}} \leq 1$	$n_{\text{Jet}} > 0$	—
$n_{\text{b-Jet}} = 0$	$n_{\text{b-Jet}} = 1$	—
$\left\lvert m_{inv}^{\text{SFOS}} - m_{inv}^{Z} \right\rvert < 25$ GeV	$\left\lvert m_{inv}^{\text{SFOS}} - m_{inv}^{Z} \right\rvert \geq 25$ GeV	—
$m_{inv}^{\ell\ell} > 12$ GeV		—
$E_T^{miss} > 50$ GeV		—
WZ CR	$t\bar{t}$ CR	SR 3

Figure 5.4: Cutflow scheme starting at the preselection level and summarizing the three distinct regions: the nominal signal region SR3, the WZ control region and the $t\bar{t}$ control region.

The upcoming sections 5.4 and 5.5 deal with the training and optimization of binary neural networks. Since $t\bar{t}$ and WZ are two most prominent backgrounds, a network is trained to separate WH from $t\bar{t}$ followed by the section on the network training to separate WH from WZ. For both background processes Boosted Decision Trees serve as classifier baseline using the configurations from the previous WH analysis. In section 5.6 the knowledge on the backgrounds gained in the former sections 5.4 and 5.5 is combined to construct, train and optimize a multiclass neural network capable of simultaneously separating WH from the mentioned backgrounds.

5.4 ANN$_{t\bar{t}}$: WH against $t\bar{t}$

In the Z-depleted signal region the most dominant background to the WH decay with three charged leptons at the LHC is the gluon induced top quark pair decay into a W-b pair where the $W^{\pm}$ bosons further decay leptonically and the $b/\bar{b}$-quark form a particle jet from which one charged lepton can be reconstructed as third candidate lepton as schematically depicted in Figure 5.5. Following the WH lepton enumeration, Figure 5.6 illustrates the truth origin distribution of the $t\bar{t}$ MC sample. More than 99% of ℓ_0 originate from a top quark decay into a W boson ($t \to W \to \ell\nu$) while in one third of all events ℓ_1 or ℓ_2 are fake leptons from a heavy flavor hadron. In order to properly distinguish the signal from this background it is essential to identify the fake lepton.

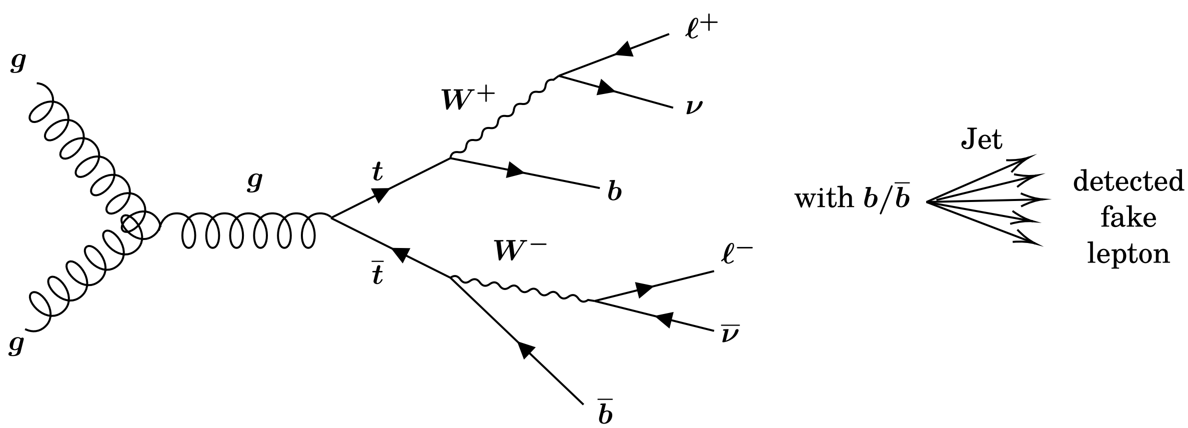


Figure 5.5: Feynman diagram of the dominant $t\bar{t}$ background process. The b-quarks can form jets from which a lepton can be detected as lepton completing the trilepton criterion.

A study on the MC reconstructed kinematic variables is carried out using the MC truth origin information. Comparing the transverse momentum distribution of prompt leptons from the $t \to W \to \ell\nu$ decay chain with the equivalent distribution from leptons originating from a b-quark associated decay shows that the latter tends towards small p_T values of less than 30 GeV including almost 60% of all fake leptons. Thus the kinematic variables of the leptons with minimal p_T is one reasonable approach for the choice of potential variables for the neural network.

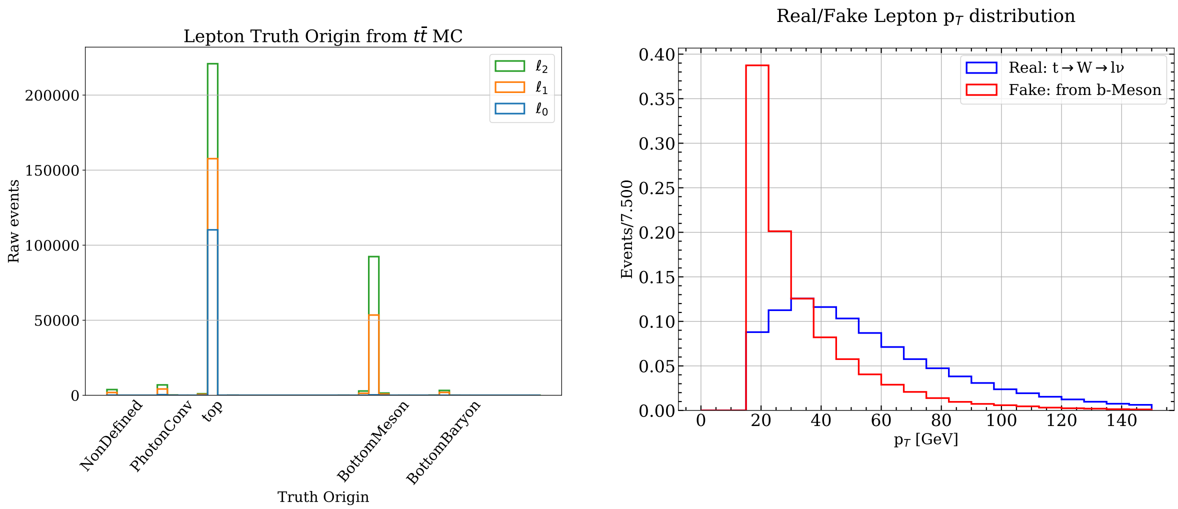
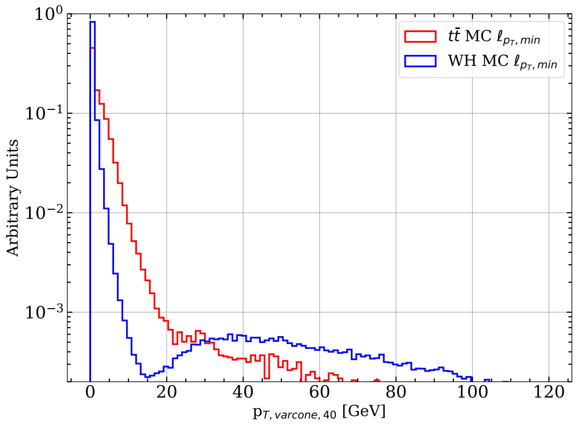


Figure 5.6: Left: Truth origin distribution of the three lepton candidates over all events that passed the SR3 cut stage in case of the $t\bar{t}$ MC. **Right:** p_T distribution for leptons originating from a b-jet (red) and leptons originating from the top associated W decay (blue) in the $t\bar{t}$ MC.

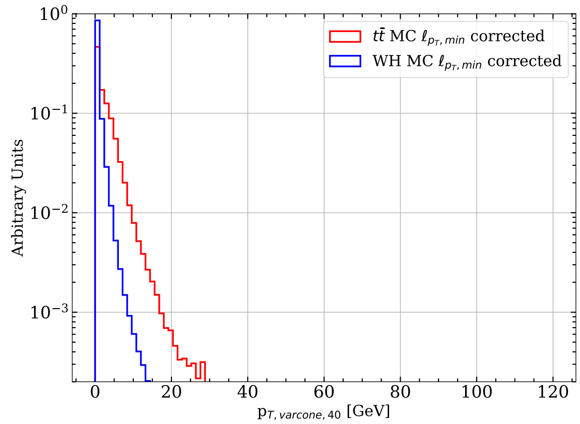
Since fake leptons are products of the b-jet decay and propagate within its particle environment emerging from hadronization they tend to be less isolated. This translates into broader energy cone and momentum cone distributions compared to equivalent distributions of prompt leptons. Therefore, isolation variables are expected to be a crucial part of the set of variables fed to the neural network. Before using the isolation variables for the training, a correction has to be applied that accounts for the overlap of cones of close prompt leptons in the calculation of isolation variables. The effect that occurs especially for high cone radii of e.g. $\Delta R = 0.4$ can be seen in Figures 5.7 (a) and 5.7 (c). Here, the calorimeter isolation variable $E_{\mathrm{T,\,cone,\,40}}^{\mathrm{topo}}$ and the track isolation variable $p_{\mathrm{T,\,varcone,40}}$ of the lepton with minimum transverse momentum ℓ_{minp_T} is depicted. Following a steep negative slope, the number of WH MC events rises again above values of $p_{\mathrm{T,\,varcone,40}} > 15$ GeV and $E_{\mathrm{T,\,cone,\,40}}^{\mathrm{topo}} > 25$ GeV while the number of $t\bar{t}$ MC events steadily decreases. Since the prompt leptons from the WH decay are well isolated, this behavior is not expected and therefore needs to be removed.

The correction algorithm from the previous analysis is reused with a change in the case of muons. While for electrons the reconstructed energy is subtracted to correct the energies $E_{\mathrm{T,\,cone}}^{\mathrm{topo}}$, for muons a constant subtraction of 3 GeV is applied approximately following Ref.[74]. Regarding the $p_{\mathrm{T,\,varcone}}$ variables the transverse momentum that is measured in the ID is subtracted from the nominal value.

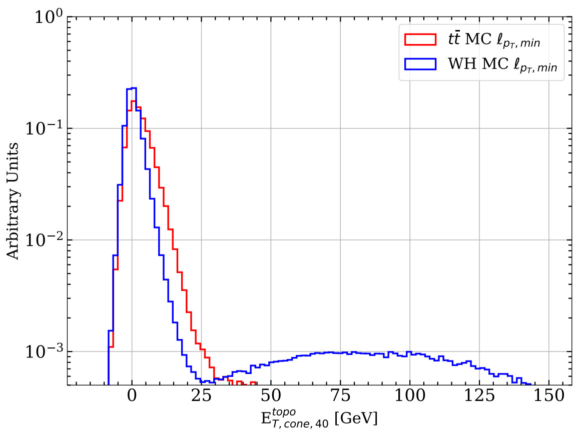
The correction is applied to all three lepton candidates, results are shown in Figures 5.7 (b) and 5.7 (d) for ℓ_{minp_T}. The previous rise of WH events towards higher energy and momentum values is removed while the correction barely changes the $t\bar{t}$ distributions.



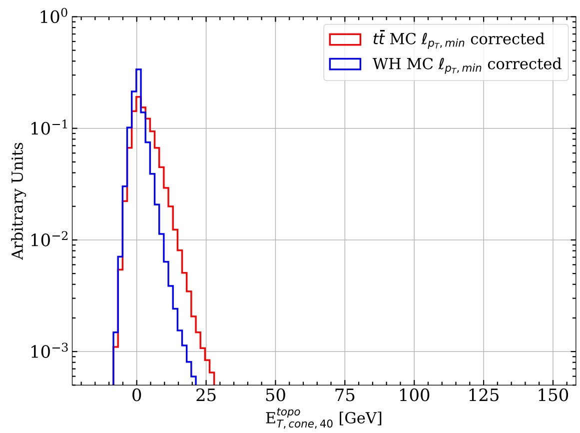
(a) Uncorrected $p_{\mathrm{T,\,varcone,\,40}}^{\ell_{minp_T}}$



(b) Corrected $p_{\mathrm{T,\,varcone,\,40}}^{\ell_{minp_T}}$



(c) Uncorrected $E_{\mathrm{T,\,cone,\,40}}^{\mathrm{topo},\ell_{minp_T}}$



(d) Corrected $E_{\mathrm{T,\,cone,\,40}}^{\mathrm{topo},\ell_{minp_T}}$

Figure 5.7: Left Column: Uncorrected isolation variables on the example of the large radius cones. **Right Column:** Corrected isolation variables.

5.4.1 Training and Validation MC Data Sets

For the training and validation step different MC data sets are used. For training and hyper parameter optimization WH and $t\bar{t}$ events from SR12 are used to guarantee orthogonality between the events used in the training and the events worked with in the following nominal analysis. The choice of SR12 is possible without any restrictions since neither of the two process depends on the SFOS criteria as no Z boson is involved. A loose isolation working point[1] is chosen to enhance the size of the training sample. With these criteria 1.001.300 WH raw events and 110.987 $t\bar{t}$ raw events passed the preselection into SR12. During the training the events are reweighted to have an equal class balance in order not to bias the network in favor of one class. After training and hyper parameter optimization the network is validated on the nominal WH and $t\bar{t}$ SR3 MC which are used for the analysis.

[1]This isolation working point is called FCLoose and is defined Ref.[73].

5.4.2 Variable Selection

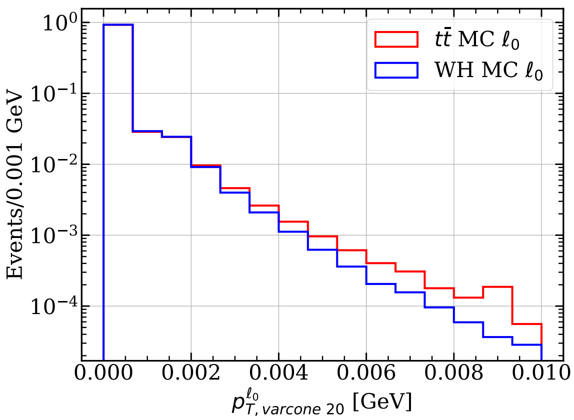
To find the most powerful set of variables different combinations are systematically tested. The optimal set is decided on the highest area under the ROC curve that gives a first estimate on the separation power of the network.
Additionally, the sets of variables are tested with the WH enumeration scheme as well as with the lepton of minimum p_T approach. While the right plot in Figure 5.6 shows a good separation between signal and background it turns out that the missing information from the two remaining leptons is essential for the network to perform optimally. Thus the identified set of most powerful variables uses information from all three leptons resulting in 44 input variables. This set can be represented in different kinematic categories explaining their corresponding impacts.

Fake Lepton Identification Variables

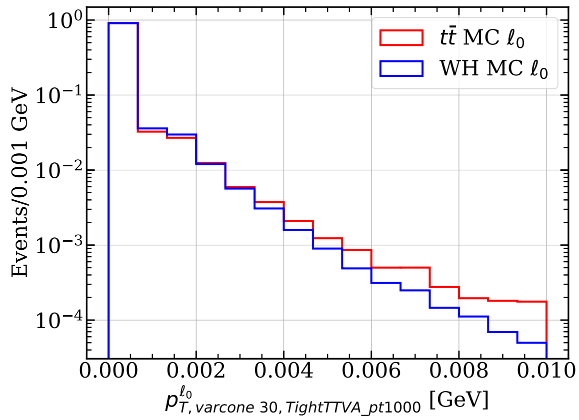
The momentum based isolation variables $p_{\text{T, varcone,20}}$, $p_{\text{T, varcone, TightTTVA\_pt500,30}}$, $p_{\text{T, varcone, TightTTVA\_pt1000,30}}$ and $p_{\text{T, varcone,40}}$ as well as the calorimeter energy based variables $E^{\text{topo}}_{\text{T, cone, 20}}$ and $E^{\text{topo}}_{\text{T, cone, 40}}$ are used for all three lepton candidates ℓ_0 (subplots in Figure 5.8), ℓ_1 (subplots in Figure 5.9) and ℓ_2 (subplots in Figure 5.10). With increasing cone size and lepton number the deviation of the distributions of WH and $t\bar{t}$ become more visible. This can be explained on one hand by the fact that the increase in cone size in principle allows for more particles to participate in the cone variable calculation. Therefore the distributions of $t\bar{t}$ will be broader due to the possible jet environment. On the other hand the deviation with respect to the lepton number can also be explained with the left plot of Figure 5.6. In contrast to ℓ_0, the probability of ℓ_1 or ℓ_2 to be the less isolated fake lepton is non-negligible.

Furthermore, the significance of the transverse impact parameter $d_0/\sigma\,(d_0)$ and the longitudinal impact parameter $z_0 \times \sin\theta$ are chosen. Their distributions are expected to tend towards greater values in absolute for $t\bar{t}$ due to the distinctive long life time of the b-quark as depicted in the subplots of Figure 5.11 for the significance of the transverse impact parameter and the longitudinal parameter.

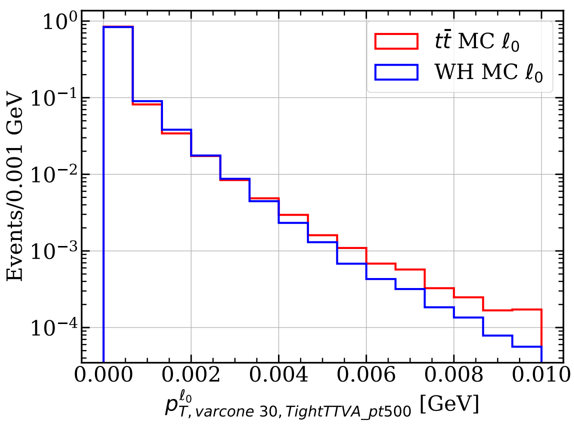
While the ℓ_0 distributions do not show a significant deviation between WH and $t\bar{t}$, the training shows that these variables contribute to the improvement of the network's performance. Thus the network benefits from the different shapes with respect to the three leptons or even from correlations. Additionally, the lepton p_T distributions as illustrated in Figure 5.12 are provided to the network to exploit the tendency of low p_T values in the case of fake leptons.



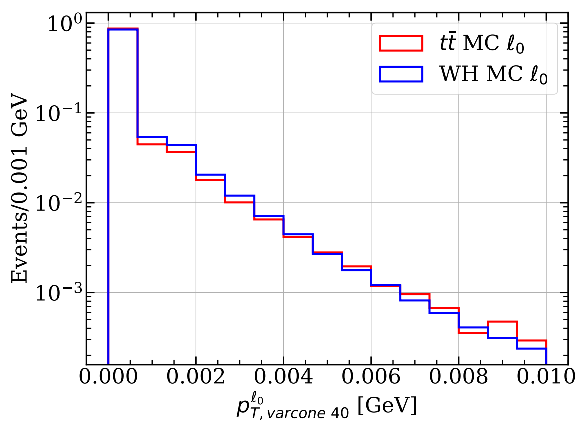
(a) $p^{\ell_0}_{\mathrm{T,\ varcone,\ 20}}$



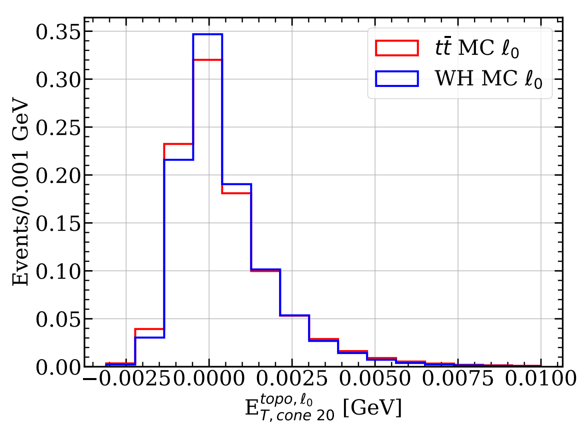
(b) $p^{\ell_0}_{\mathrm{T,\ varcone,\ TightTTVA\_pt1000,\ 30}}$



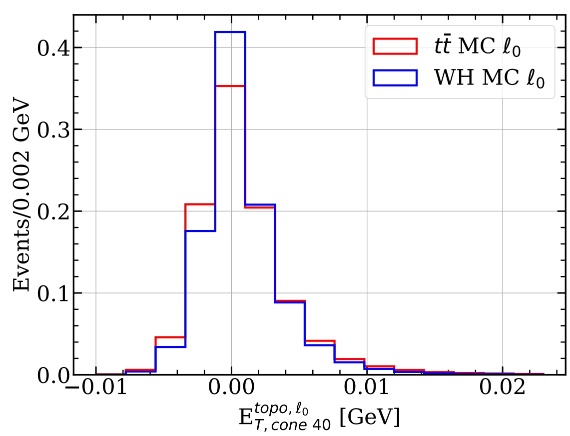
(c) $p^{\ell_0}_{\mathrm{T,\ varcone,\ TightTTVA\_pt500,\ 30}}$



(d) $p^{\ell_0}_{\mathrm{T,\ varcone,\ 40}}$

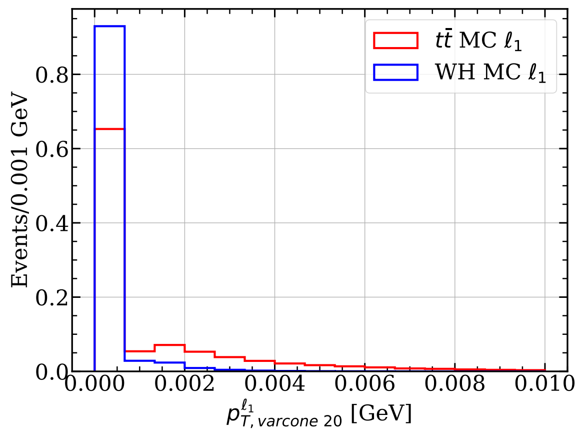


(e) $E^{\mathrm{topo},\ell_0}_{\mathrm{T,\ cone,\ 20}}$

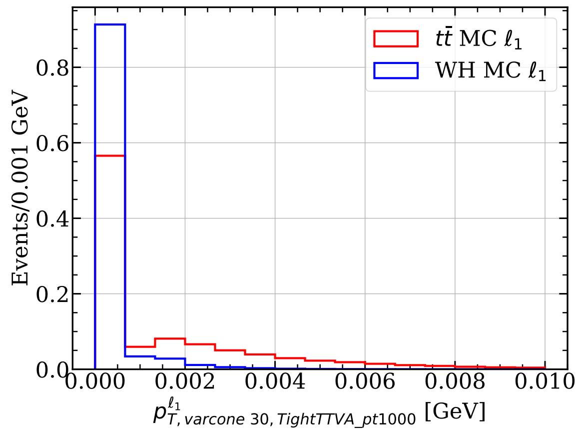


(f) $E^{\mathrm{topo},\ell_0}_{\mathrm{T,\ cone,\ 40}}$

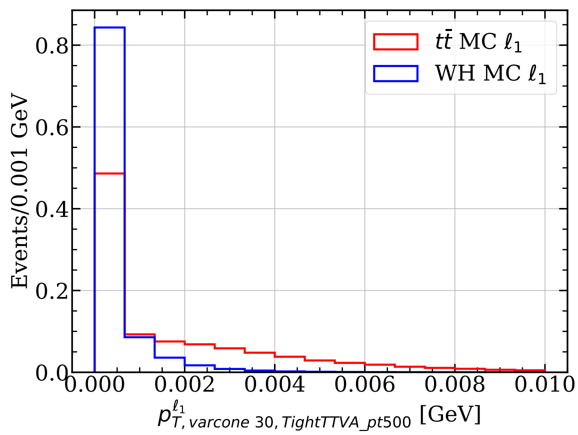
Figure 5.8: Normalized isolation variables of ℓ_0.



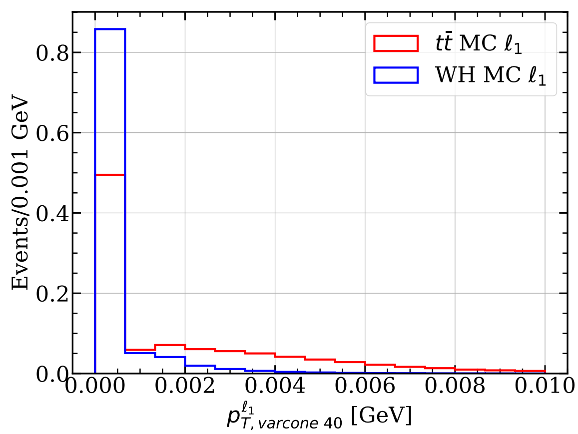
(a) $p_{\mathrm{T,\,varcone,\,20}}^{\ell_1}$



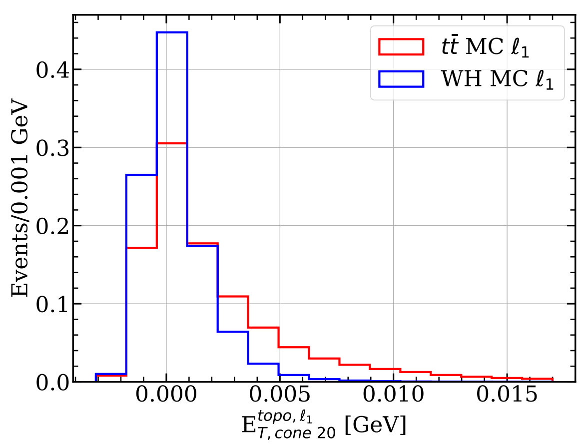
(b) $p_{\mathrm{T,\,varcone,\,TightTTVA\_pt1000,\,30}}^{\ell_1}$



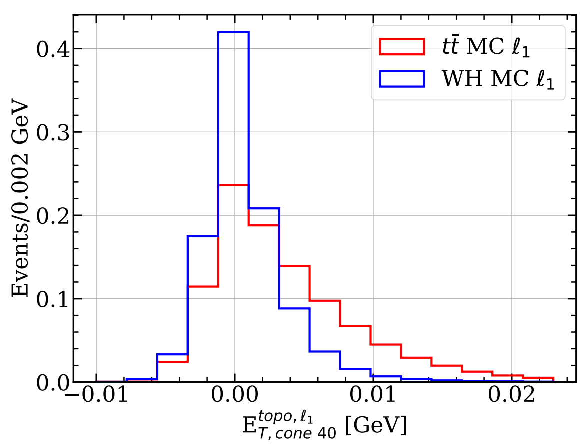
(c) $p_{\mathrm{T,\,varcone,\,TightTTVA\_pt500,\,30}}^{\ell_1}$



(d) $p_{\mathrm{T,\,varcone,\,40}}^{\ell_1}$

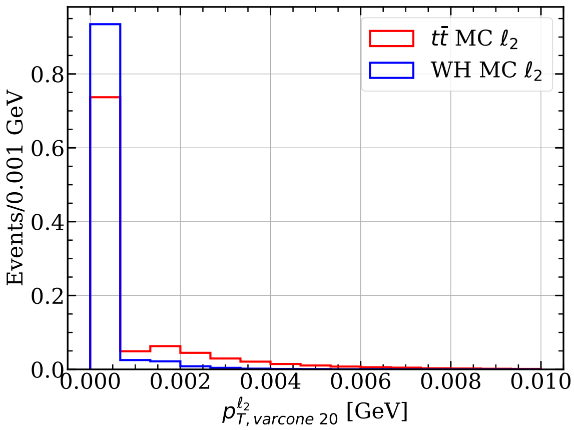


(e) $E_{\mathrm{T,\,cone,\,20}}^{\mathrm{topo},\ell_1}$

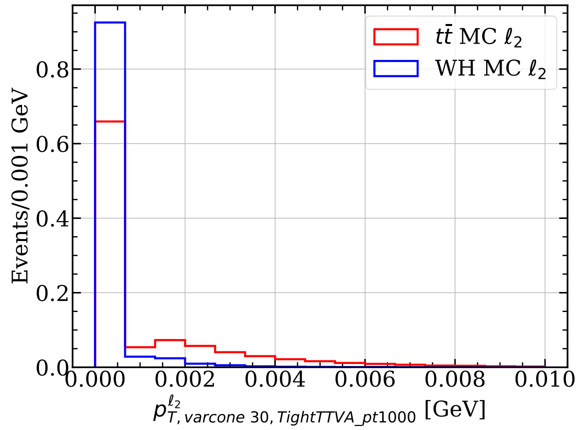


(f) $E_{\mathrm{T,\,cone,\,40}}^{\mathrm{topo},\ell_1}$

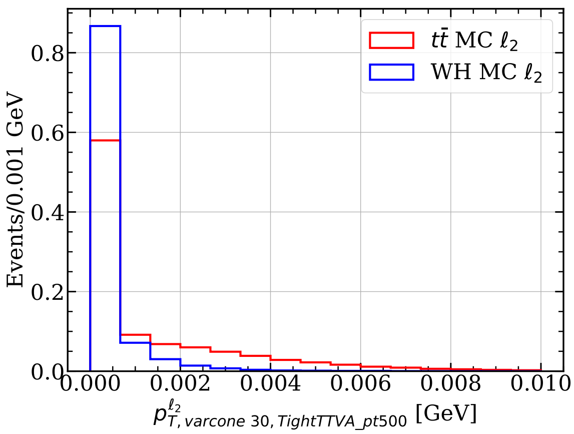
Figure 5.9: Normalized isolation variables of ℓ_1.



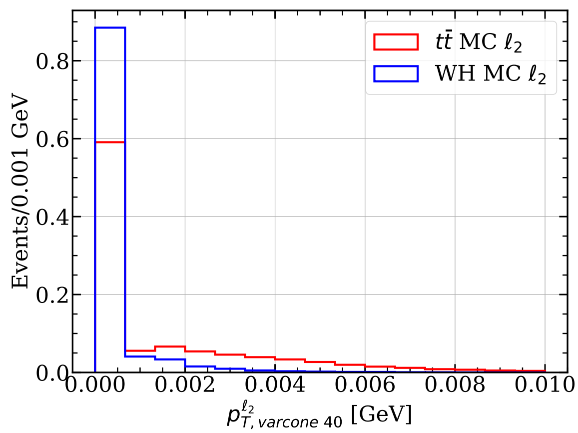
(a) $p^{\ell_2}_{\mathrm{T,\,varcone,\,20}}$



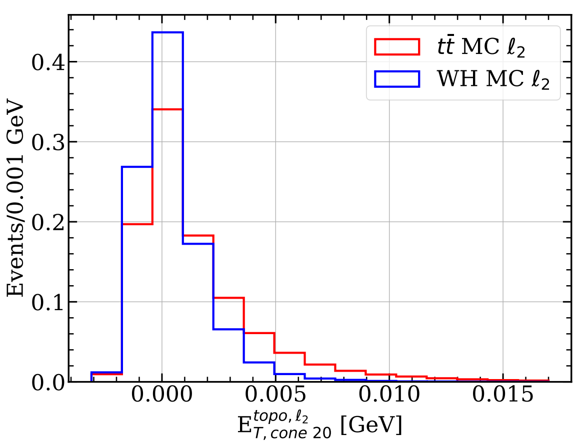
(b) $p^{\ell_2}_{\mathrm{T,\,varcone,\,TightTTVA\_pt1000,\,30}}$



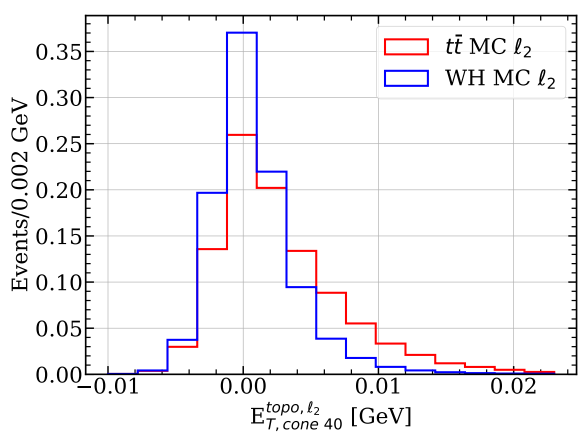
(c) $p^{\ell_2}_{\mathrm{T,\,varcone,\,TightTTVA\_pt500,\,30}}$



(d) $p^{\ell_2}_{\mathrm{T,\,varcone,\,40}}$

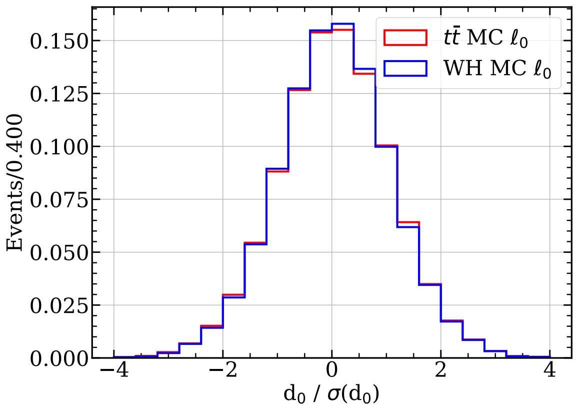


(e) $E^{\mathrm{topo},\ell_2}_{\mathrm{T,\,cone,\,20}}$

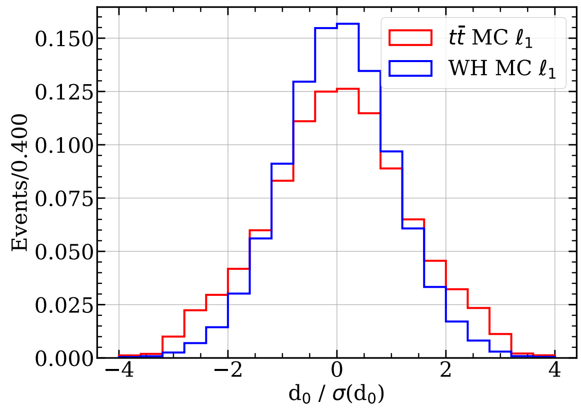


(f) $E^{\mathrm{topo},\ell_2}_{\mathrm{T,\,cone,\,40}}$

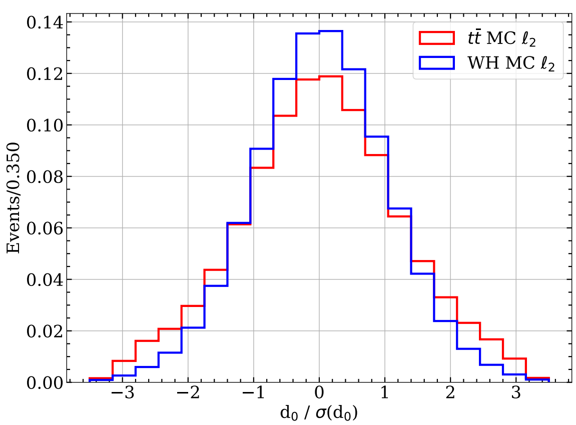
Figure 5.10: Normalized isolation variables of ℓ_2.



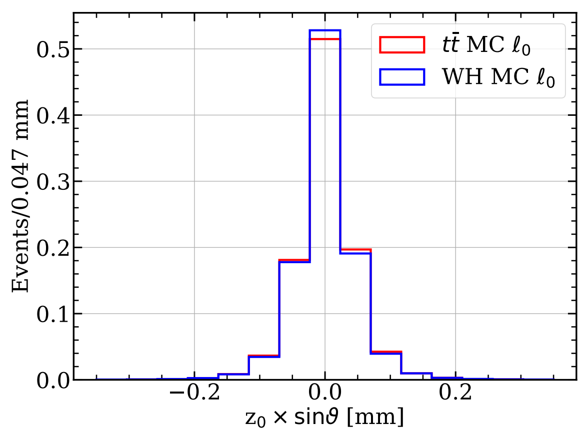
(a) $d_0/\sigma(d_0)^{\ell_0}$



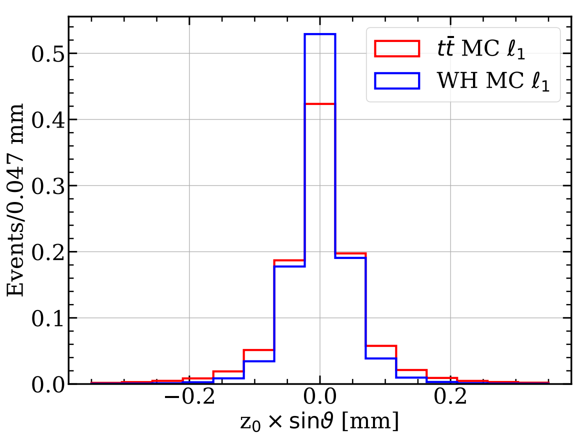
(b) $d_0/\sigma(d_0)^{\ell_1}$



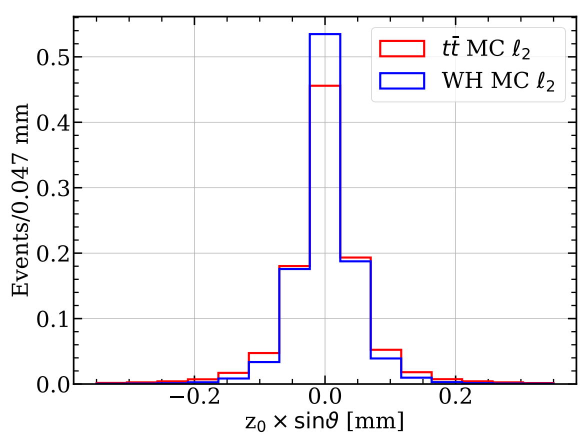
(c) $d_0/\sigma(d_0)^{\ell_2}$



(d) $z_0 \times \sin(\theta)^{\ell_0}$



(e) $z_0 \times \sin(\theta)^{\ell_1}$



(f) $z_0 \times \sin(\theta)^{\ell_2}$

Figure 5.11: Normalized transverse impact parameter significance distributions and normalized longitudinal impact parameter distributions of all three lepton candidates.

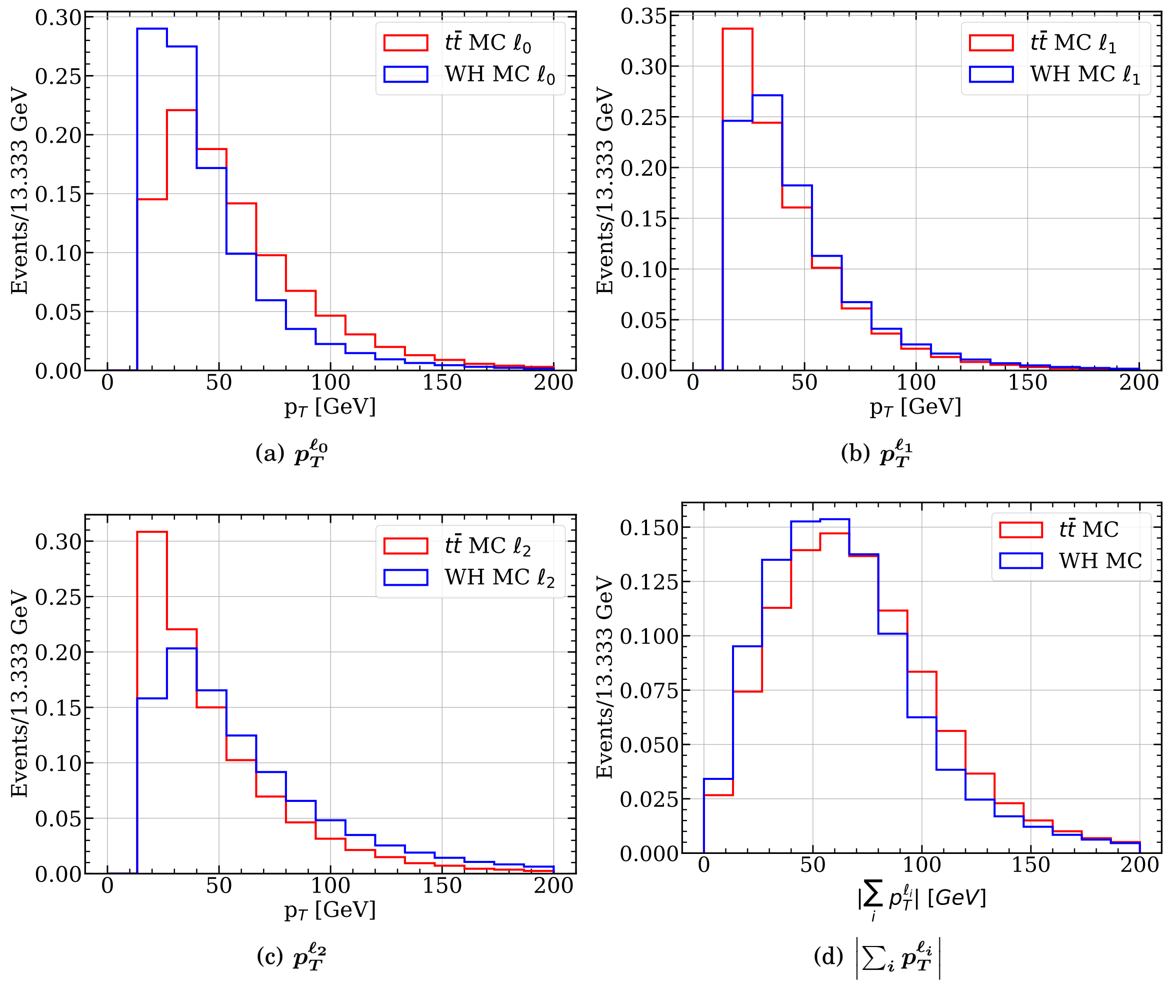
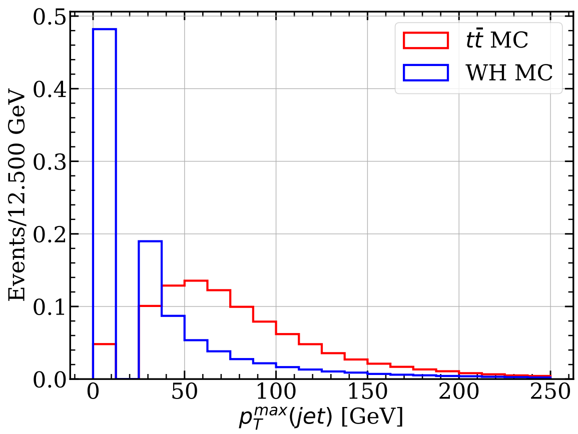


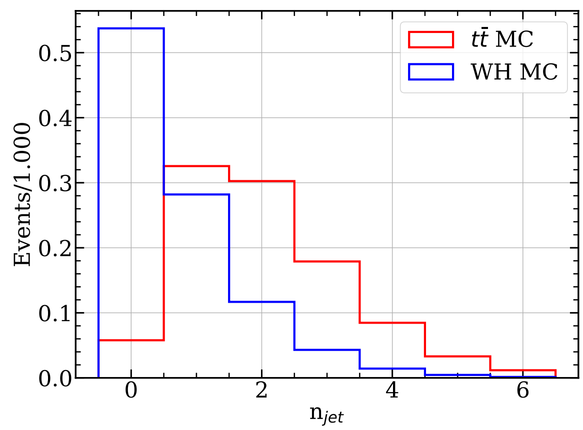
Figure 5.12: Normalized p_T distributions of all three lepton candidates as well as the normalized sum of p_T distribution.

Jet Related Variables

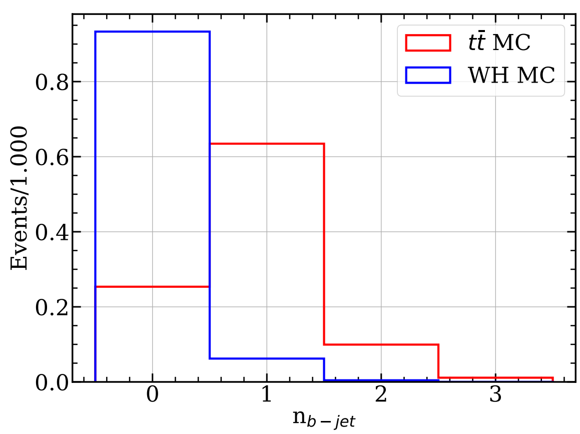
An important difference between WH and $t\bar{t}$ is the presence of two b-jets due to the top decays. Correspondingly sensitive variables are the number of (b-) jets and the transverse momentum of the leading jet (p_T^{max}(jet)) which is the jet with maximum p_T. In the absence of jets this value is set to zero and explains the peak in the corresponding WH distribution of Figure 5.13 (a) that also illustrates the jet p_T threshold. The Figures 5.13 (b) and 5.13 (c) furthermore show the expected higher jet activity in the $t\bar{t}$ process.



(a) p_T^{max}(jet)



(b) n_{jet}



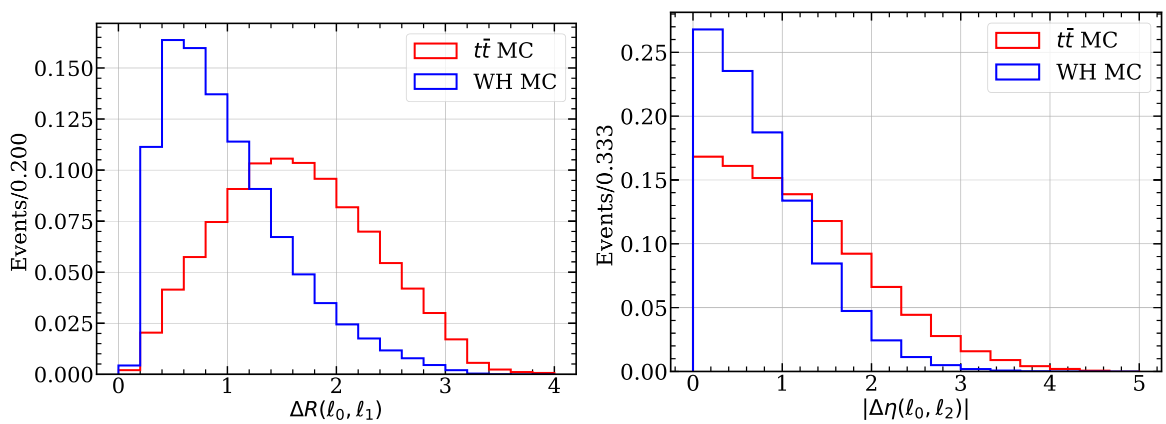
(c) $n_{b-\text{jet}}$

Figure 5.13: Normalized p_T^{max}(jet), n_{jet} and $n_{b-\text{jet}}$ distributions from left to right.

Spin Correlation and Lepton Systems Variables

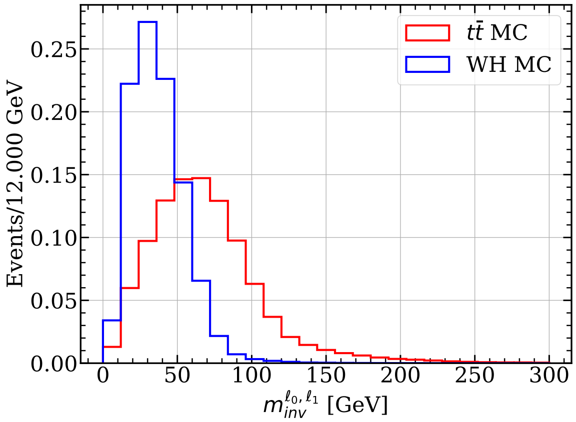
The possible spin correlation between the leptons from the Higgs decay can be exploited with help of variables like the angular distance ΔR, or the invariant mass. While the leptons from $t\bar{t}$ are not necessarily correlated, their angular distance and invariant mass are expected to tend towards higher values for the ℓ_0-ℓ_1-system as depicted in Figure 5.14 (a) and Figure 5.15 (a). The transverse mass $m_T^{\ell_0,\ell_1}$ distributions as depicted in Figure 5.15 (d) show a separation as the WH distributions is supposed to peak close to the Higgs mass.

Information on the ℓ_1-ℓ_2-system is provided by using the absolute difference of $\Delta\eta^{\ell_1,\ell_2}$ and the invariant/transverse mass of this lepton combination. For WH one lepton is expected to originate from the $H \to WW^*$ decay and the other is expected from the associated Wdecay while in the case of $t\bar{t}$ the leptons are expected to originate from different top quarks to conserve the total charge. Further systematic studies show that the network can benefit from the complementary invariant/transverse mass information of the ℓ_0-ℓ_2-system variables with expected contribution of a Higgs decay lepton.

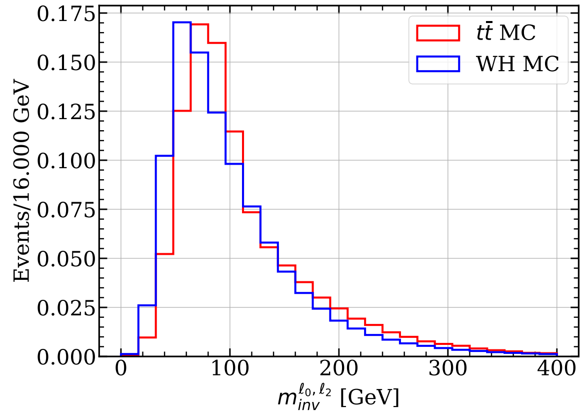


(a) $\Delta R^{\ell_0,\ell_1}$ (b) $\Delta\eta^{\ell_1,\ell_2}$

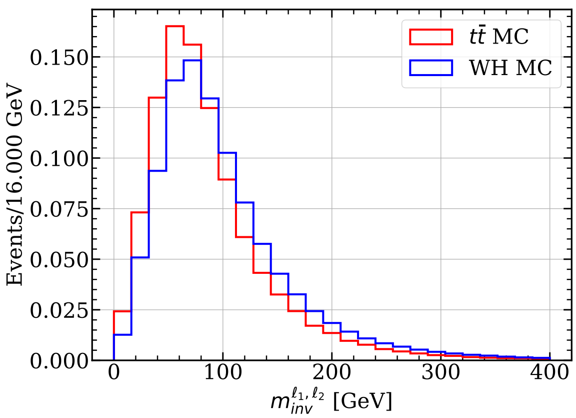
Figure 5.14: Normalized distribution of the angular separation in ΔR of ℓ_0 and ℓ_1.



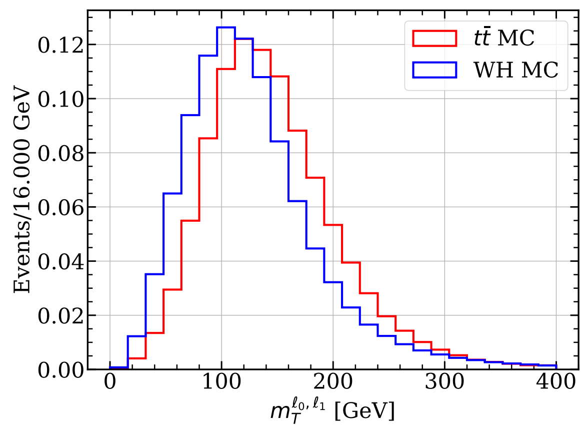
(a) $m_{inv}^{\ell_0,\ell_1}$



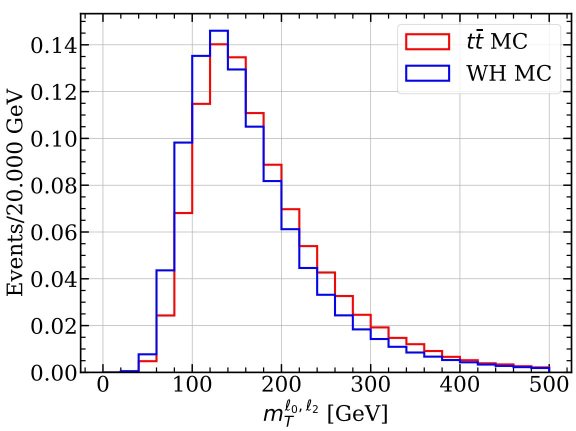
(b) $m_{inv}^{\ell_0,\ell_2}$



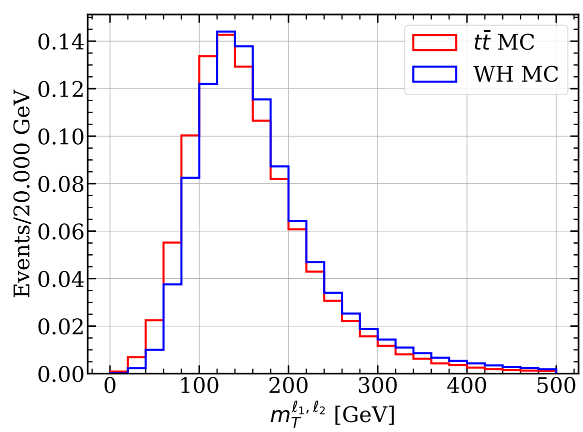
(c) $m_{inv}^{\ell_1,\ell_2}$



(d) $m_T^{\ell_0,\ell_1}$



(e) $m_T^{\ell_0,\ell_2}$

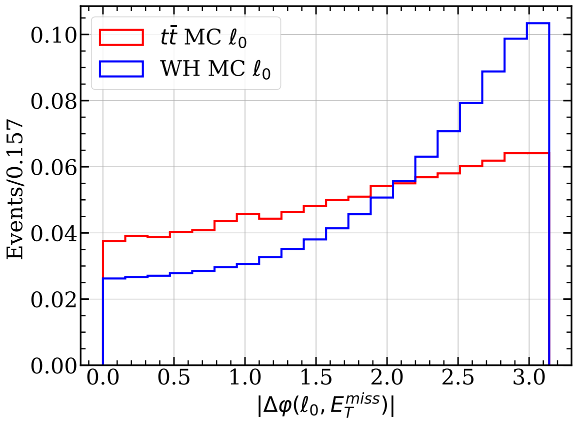


(f) $m_T^{\ell_1,\ell_2}$

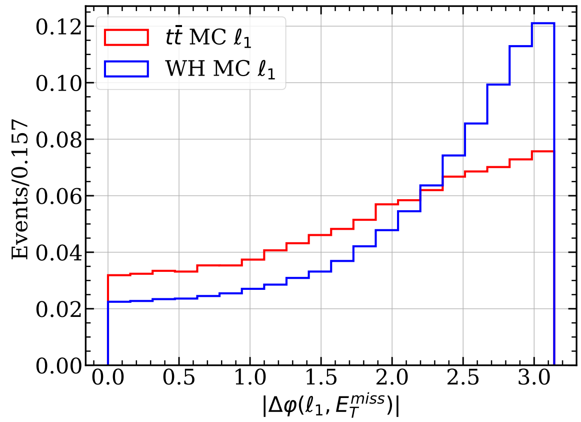
Figure 5.15: Normalized invariant mass distributions and normalized transverse mass distributions for all three dilepton combinations.

Additional Event Variables

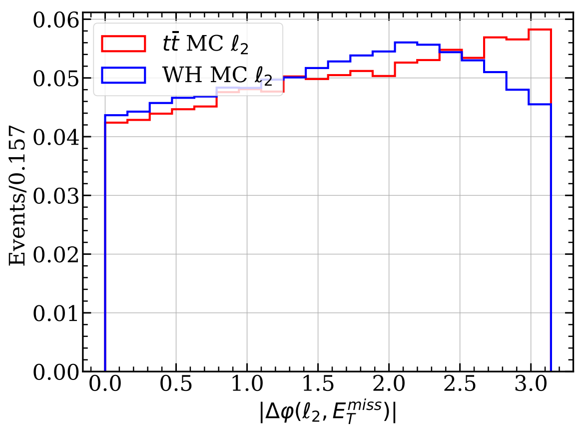
Another difference between WH and $t\bar{t}$ is the number of neutrinos of 3 and 2 respectively. The neutrino energy and the angles between them introduce a topological difference that can be observed in quantities like the missing transverse energy E_T^{miss}. Additionally, its corresponding $\varphi^{E_T^{miss}}$ direction in the transverse plane and the difference in φ with respect to the three candidate leptons $\Delta\varphi^{\ell_i, E_T^{miss}}$ is used as input variables and is shown in the subplots of Figure 5.16 and Figure 5.17. A summary of all input variables for $\mathrm{ANN}_{t\bar{t}}$ is given in Table 5.4.



(a) $\Delta\varphi^{\ell_0, E_T^{miss}}$

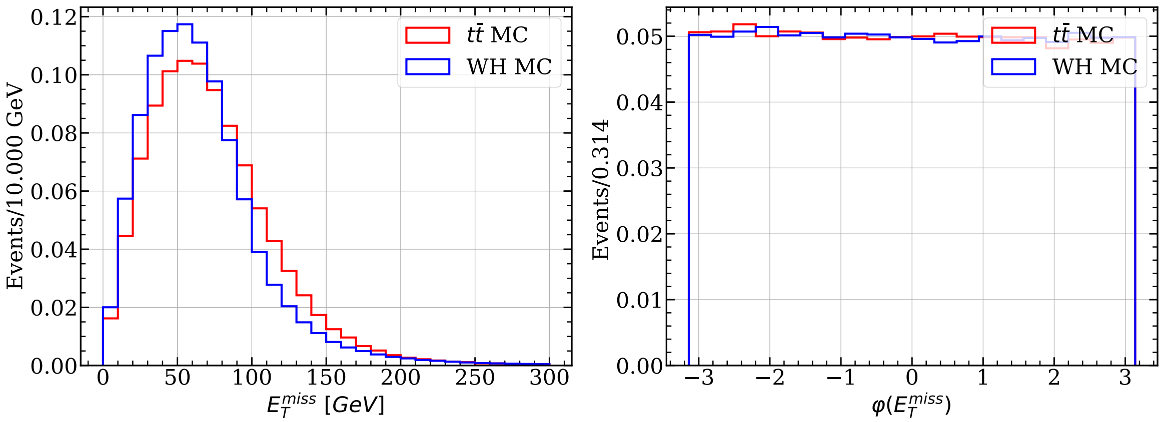


(b) $\Delta\varphi^{\ell_1, E_T^{miss}}$



(c) $\Delta\varphi^{\ell_2, E_T^{miss}}$

Figure 5.16: Normalized distributions of the difference in φ of E_T^{miss} and all three lepton candidates.



(a) E_T^{miss}

(b) $\varphi^{E_T^{miss}}$

Figure 5.17: Normalized distribution of the missing transverse energy and corresponding φ direction.

ANN$_{t\bar{t}}$ input variables

$d_0/\sigma\ (d_0)^{\ell_0}$	$d_0/\sigma\ (d_0)^{\ell_1}$	$d_0/\sigma\ (d_0)^{\ell_2}$	E_T^{miss}
$z_0 \times \sin\ (\theta)^{\ell_0}$	$z_0 \times \sin\ (\theta)^{\ell_1}$	$z_0 \times \sin\ (\theta)^{\ell_2}$	$\varphi^{E_T^{miss}}$
$p_T^{\ell_0}$	$p_T^{\ell_1}$	$p_T^{\ell_2}$	$\left\|\sum_i p_T^{\ell_i}\right\|$
p_T^{max}(jet)	n_{jet}	$n_{b-\text{jet}}$	$\Delta\varphi^{\ell_0,E_T^{miss}}$
$m_{inv}^{\ell_0,\ell_1}$	$m_{inv}^{\ell_0,\ell_2}$	$m_{inv}^{\ell_1,\ell_2}$	$\Delta\varphi^{\ell_1,E_T^{miss}}$
$m_T^{\ell_0,\ell_1}$	$m_T^{\ell_0,\ell_2}$	$m_T^{\ell_1,\ell_2}$	$\Delta\varphi^{\ell_2,E_T^{miss}}$
$p_{\text{T, varcone, 20}}^{\ell_0}$	$p_{\text{T, varcone, TightTTVA\_pt1000, 30}}^{\ell_0}$	$p_{\text{T, varcone, TightTTVA\_pt500, 30}}^{\ell_0}$	$p_{\text{T, varcone, 40}}^{\ell_0}$
$p_{\text{T, varcone, 20}}^{\ell_1}$	$p_{\text{T, varcone, TightTTVA\_pt1000, 30}}^{\ell_1}$	$p_{\text{T, varcone, TightTTVA\_pt500, 30}}^{\ell_1}$	$p_{\text{T, varcone, 40}}^{\ell_1}$
$p_{\text{T, varcone, 20}}^{\ell_2}$	$p_{\text{T, varcone, TightTTVA\_pt1000, 30}}^{\ell_2}$	$p_{\text{T, varcone, TightTTVA\_pt500, 30}}^{\ell_2}$	$p_{\text{T, varcone, 40}}^{\ell_2}$
$E_{\text{T, cone, 20}}^{\text{topo},\ell_0}$	$E_{\text{T, cone, 40}}^{\text{topo},\ell_0}$	$E_{\text{T, cone, 40}}^{\text{topo},\ell_1}$	$E_{\text{T, cone, 40}}^{\text{topo},\ell_1}$
$E_{\text{T, cone, 20}}^{\text{topo},\ell_2}$	$E_{\text{T, cone, 40}}^{\text{topo},\ell_2}$	$\Delta R^{\ell_0,\ell_1}$	$\Delta\eta^{\ell_1,\ell_2}$

Table 5.4: Summary of input variables for ANN$_{t\bar{t}}$

5.4.3 Hyperparameter Optimization

The considered hyperparameters that are optimized to find the best performing neural network based on the highest area under the ROC curve are summarized in Table 5.5. Simultaneously, a grid search is carried out to find the optimal network architecture. In the final setup the ANN is trained on a batch size of 4096 samples over 80 epochs with an early stopping after 10 epochs in case the area under the ROC curve has not increased. The training is carried out in a 5-fold cross-validation with a training size of 80% of the full data set and a test size of 20% respectively. Adam[75] is used as stochastic optimization method. The learning rate is set to 0.001 while the loss function is the binary crossentropy from Equation 4.2. The swish activation function

$$\mathrm{swish}(x) = \frac{x}{1 + e^{-x}}$$

performs best together with the dropout method applied for regularization purposes.

Hyperparameter	Tested [within]	Final Value
Batch Size	[1024, 12.288]	4096
Epochs	[30, 100]	80
k folds	5, 10	5
Training Sample Size	80%	80%
Testing Sample Size	20%	20%
Optimizer	Adam	Adam
Learning Rate	[0.01, 0.0001]	0.001
Loss	Binary Crossentropy	Binary Crossentropy
Scaler	Robust/MinMax/Standard	RobustScaler
Activation	ReLu/SeLu/Swish	Swish
Regularisation	AlphaDropout/Dropout	Dropout

Table 5.5: Summary of hyperparameters that are tested and the set of final hyperparameters used in the $\mathrm{ANN}_{t\bar{t}}$ network configuration.

The data that is fed into the neural network is scaled using Scikit Learn's[76] "RobustScaler" in order to ensure a stable learning procedure. This scaler can be applied to remove a distribution's median and to scale it accordingly to the range between first quartile and the third quartile which makes the procedure robust against outliers. The effect is studied and is visualized in a boxplot format for a selected set of representative input variables in Figure 5.18. The top subplot shows the unscaled input variables ranging over multiple orders of magnitude from $\mathcal{O}(10^0)$ to $\mathcal{O}(10^3)$. The bottom plot shows the scaled input variable ranges now being much more narrow within the absolute of $\mathcal{O}(10^0)$ to $\mathcal{O}(10^1)$.

The final ANN architecture consists of 44 input variables, three hidden layers of 64, 64 and 32 neurons with corresponding dropout rates of 0.35, 0.35 and 0.3 respectively and a single output neuron with a sigmoid activation function

$$\mathrm{sigmoid}(x) = \frac{1}{1 + e^{-x}}$$

applied to it.

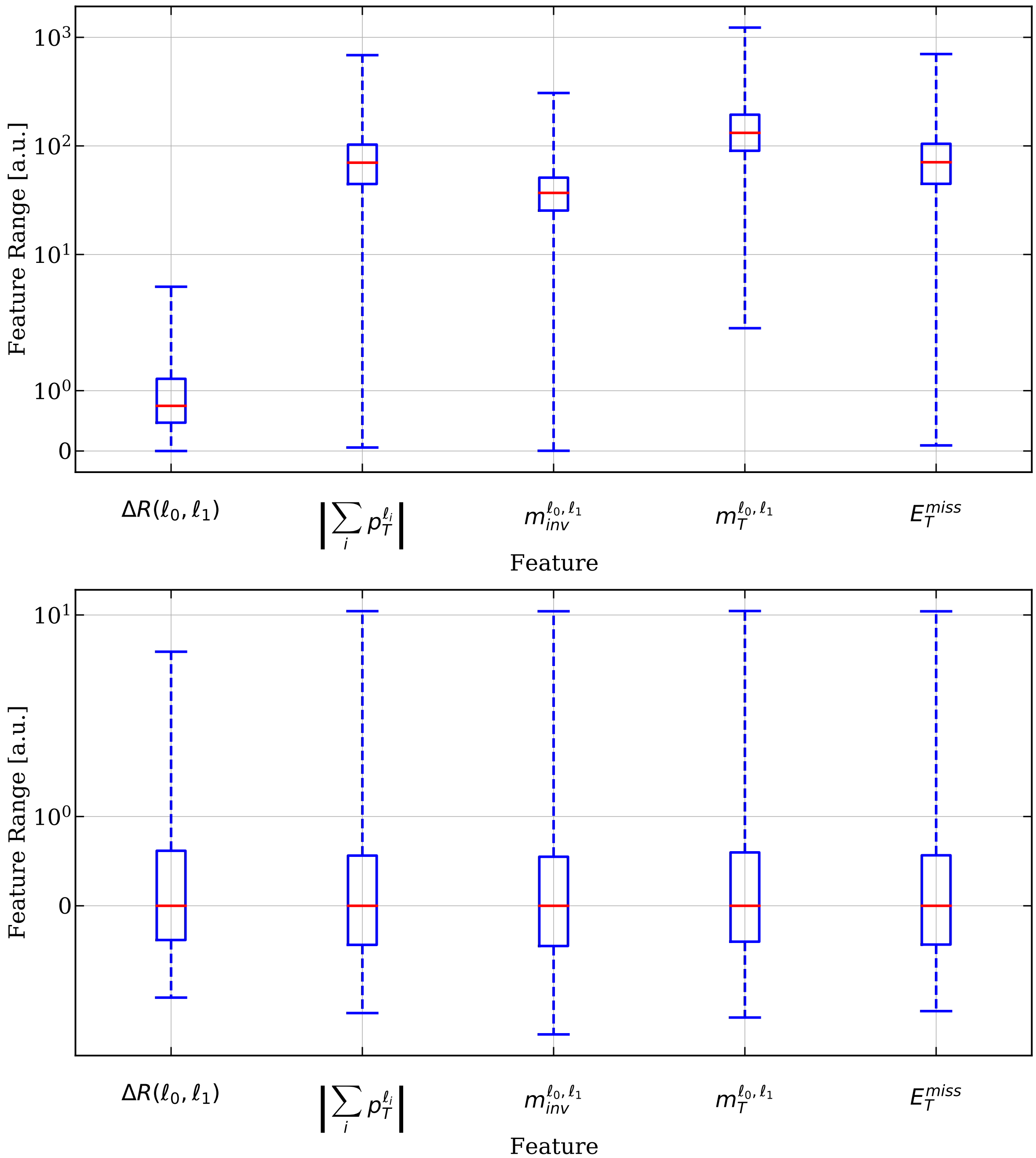


Figure 5.18: Visualization of the effect of the scaler "RobustScaler" on the range of selected input variables $\Delta R^{\ell_0,\ell_1}$, $\left|\sum_i p_T^{\ell_i}\right|$, $m_{inv}^{\ell_0,\ell_1}$, $m_T^{\ell_0,\ell_1}$, E_{miss}^T. **Top:** Unscaled input variables ranging over multiple orders of magnitude from $\mathcal{O}(10^0)$ to $\mathcal{O}(10^3)$. **Bottom:** Scaler applied, the median is removed and the variable range is within the absolute of $\mathcal{O}(10^0)$ to $\mathcal{O}(10^1)$.

A retrained BDT with the hyperparameter configuration from the previous WH analysis is used as classifier baseline. Due to the lack of variables from a dedicated heavy-flavor $\mathrm{BDT_{HFL}}$, the input variables of the latter are used in addition to the input variables of the $t\bar{t}$ discriminating classifier $\mathrm{BDT}_{t\bar{t}}$. A summary of the hyperparameters and the input variables is given in Table 5.6.

$\mathrm{BDT}_{t\bar{t}}$ Configuration	
NTrees	500
MinNodeSize	1%
maxDepth	5
BoostType	AdaBoost

$\mathrm{BDT}_{t\bar{t}}$ Input Variables
$(p_T, \eta, d_0/\sigma(d_0))^{\ell_{minp_T}}$
$m_{inv}^{\ell_0\ell_1}$ (OC)
$m_{inv}^{\ell_1\ell_2}$ (SC)
$\Delta R^{\ell_0\ell_1}$
$\Delta\eta^{\ell_1\ell_2}$
p_T^{max}(jet)
$p_{T,\mathrm{varcone},20,30,40}^{\ell_{minp_T}}$
$E_{\mathrm{T,\ cone,\ 20,40}}^{\mathrm{topo},\ell_{minp_T}}$

Table 5.6: Left: Hyperparameter configuration of the $\mathrm{BDT}_{t\bar{t}}$ used to define the classifier baseline. **Right:** Corresponding input variables to the $\mathrm{BDT}_{t\bar{t}}$ used in this analysis.

5.4.4 Training Results

This chapter gives a summary of the training results from the best performing set of hyperparameters. The first figures of merit are the development of the loss, the area under the ROC curve and the Poisson significance. The latter is calculated considering the most significant bin of the ANN prediction on the test data of WH and $t\bar{t}$ MC used in the training as a function of epoch as depicted in Figure 5.19. The mean and standard deviation over all 5 folds is displayed. The loss curves in the first panel of Figure 5.19 indicate a stable training process. The network seems to have found a distinct local minimum between epoch 20 and 30 indicated by a sudden loss drop. The mean loss curve corresponding to the prediction on the test data set (green line) stays well below the mean loss curve of the respective predictions on the train data set (blue line) which indicate a good generalization behavior of the network. While the mean loss is decreasing, the mean area under the mean ROC curve and the mean Poisson significance both show a steady increase which can be interpreted as an increasing separation power as a function of epoch.

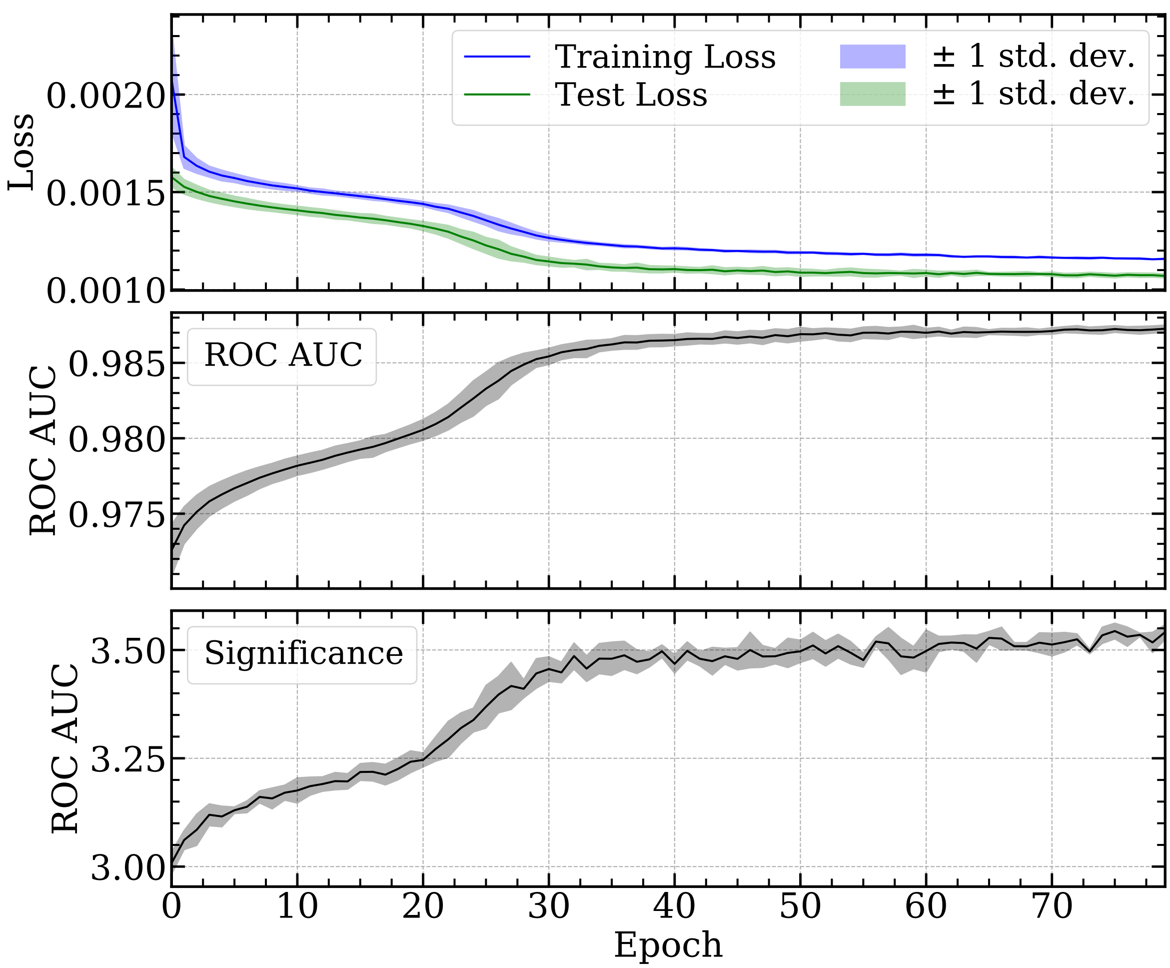


Figure 5.19: Evolution of the loss for train (blue) and test (green) data fold in the first panel, the area under the ROC curve in the middle panel and the Poisson significance in the bottom panel for the test data.

A second measure of the training's goodness is the shape comparison between the ANN's prediction on the train data set (histograms) and the prediction on the test data set (dots) over 2 exemplary folds for better readability as illustrated in the top plot in Figure 5.20, where the signal in blue is supposed to peak at values close to 0 and the background in red is supposed to peak at values towards 1. While the prediction ratio in the bottom plot slightly fluctuates around a value of 1, there is no critical deviation within the MC uncertainties. Especially the ratio of the most significant bin is in good agreement for both WH and $t\bar{t}$ over the folds and allows for the conclusion that the network gained the ability to generalize. The same behavior is observed for folds 2, 4 and 5 as illustrated in appendix B.1.

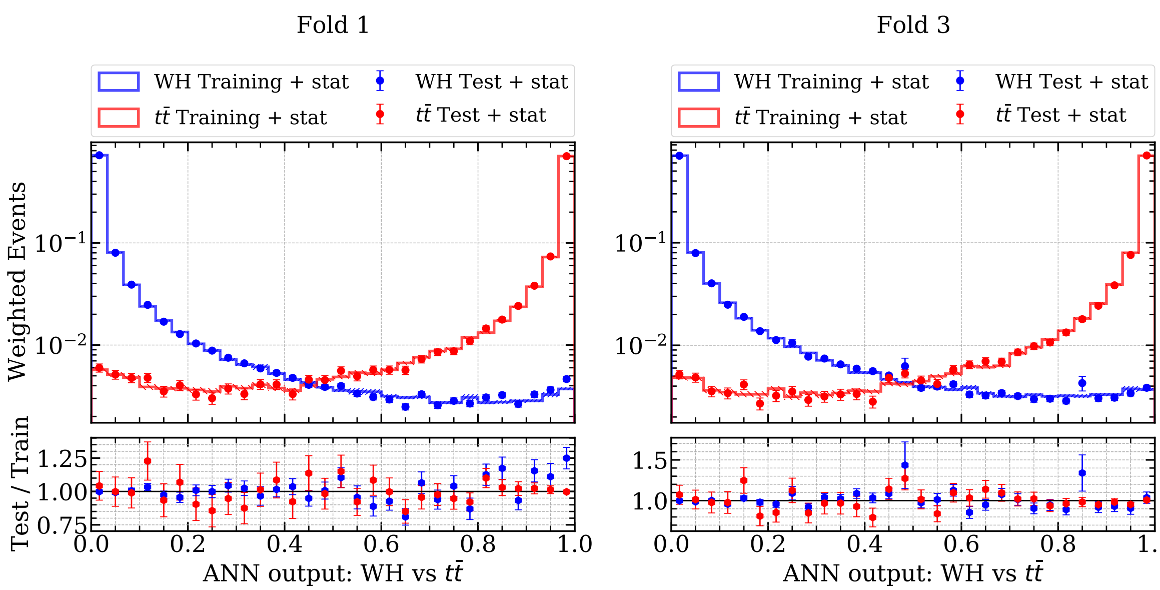


Figure 5.20: The ANN$_{t\bar{t}}$'s prediction on train (dots) and test (histograms) data folds over 2 exemplary folds in the upper plots with the corresponding ratios in the bottom plots.

The third figure of merit in this evaluation on the network's performance is the area under the ROC curve. Figures 5.21, 5.22 and 5.23 illustrate the ROC curves for the predictions on the train data set of 80% of the WH and $t\bar{t}$ SR12 MC ("Training"), the predictions on the test data set of the respective remaining 20% of the WH and $t\bar{t}$ SR12 MC ("Testing") and the ROC curves of the prediction on the nominal WH and $t\bar{t}$ SR3 MC ("Validation") respectively. The subplots are divided as followed:
The upper left plot shows the ANN ROC curve with background rejection measure $1/\epsilon_b$ against the signal efficiency for all 5 folds individually where ϵ_b is the background efficiency. The upper right plot shows the mean ANN ROC curve with the same background rejection measure and standard deviation in blue and the mean BDT ROC curve with standard deviation in green. The bottom left plot shows the ANN ROC curve with background rejection measure $1-\epsilon_b$ against the signal efficiency again for all 5 folds individually while the bottom right plot shows the mean ANN ROC curve with the same background rejection measure and standard deviation in blue and the mean BDT ROC curve with standard deviation in green.

For the prediction on the train data set the following conclusions can be drawn from the four plots: The neural network beats the retrained BDT in terms of their signal-to-background separation power with a mean area under the ROC curve of 0.99 to 0.96. Furthermore, the ANN has an over all much stronger background rejection than the BDT as suggested by the upper row plots. Finally, for both ANN and BDT the performance between the 5 folds is in very good agreement.

Equivalent conclusions can be drawn for the test data set and the validation data set. Additionally, the ANN also shows comparable performance throughout train, test and validation step as summarized with the highest areas under the ROC for both classifiers in Table 5.7.

	Training	**Testing**	**Validation**
$\text{BDT}_{t\bar{t}}$	0.96	0.96	0.96
$\text{ANN}_{t\bar{t}}$	0.99	0.99	0.99

Table 5.7: Areas of the best ROC curves in separating WH from $t\bar{t}$ of $\text{BDT}_{t\bar{t}}$ and $\text{ANN}_{t\bar{t}}$.

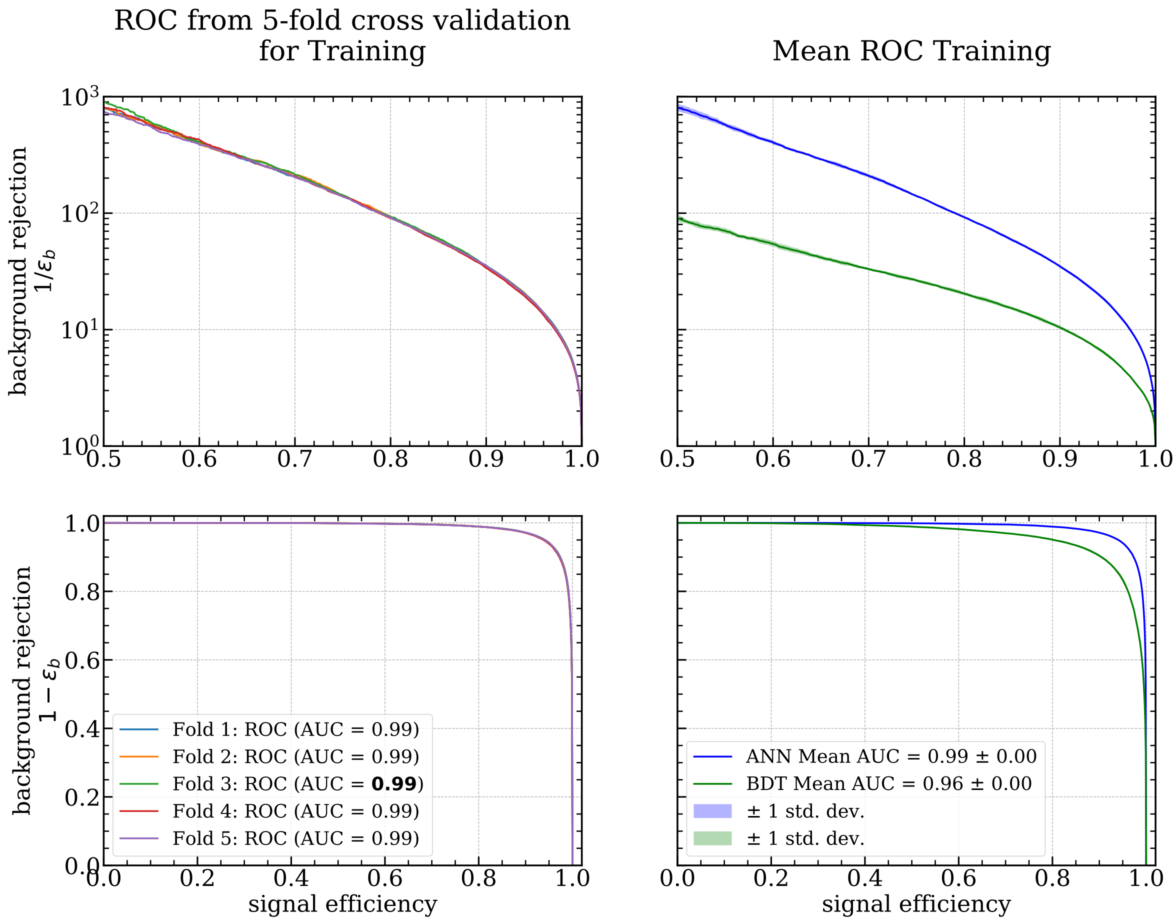


Figure 5.21: Plots based on the prediction on the train data folds for **Top Row:** ANN's ROC curve with background rejection $1/\epsilon_b$ for the individual folds (left) and the mean ROC curve of ANN (blue) and BDT (green) with background rejection $1/\epsilon_b$ and standard deviation. **Bottom Row:** ANN's ROC curve with background rejection $1 - \epsilon_b$ for the individual folds (left) and the mean ROC curve of ANN (blue) and BDT (green) with background rejection $1 - \epsilon_b$ and standard deviation.

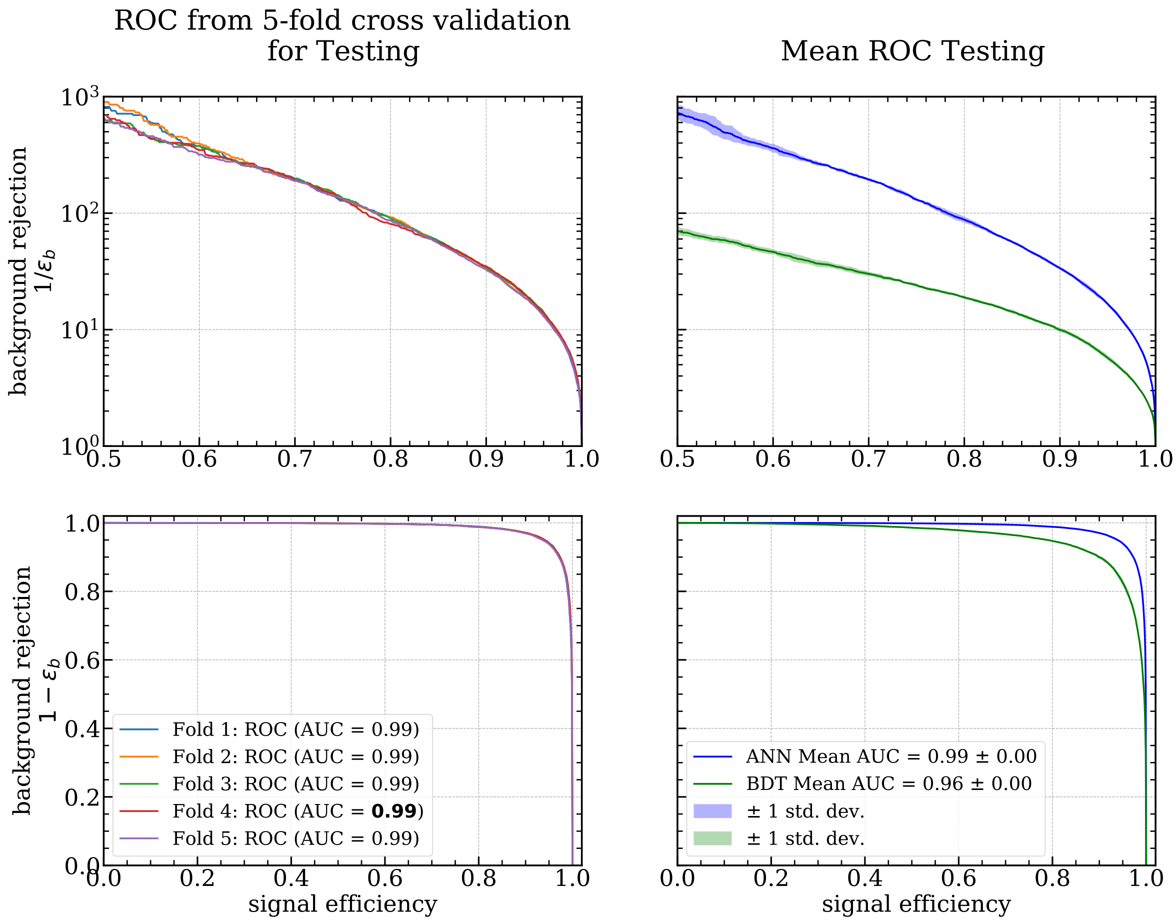


Figure 5.22: Plots based on the prediction on the test data folds for **Top Row:** ANN's ROC curve with background rejection $1/\epsilon_b$ for the individual folds (left) and the mean ROC curve of ANN (blue) and BDT (green) with background rejection $1/\epsilon_b$ and standard deviation. **Bottom Row:** ANN's ROC curve with background rejection $1-\epsilon_b$ for the individual folds (left) and the mean ROC curve of ANN (blue) and BDT (green) with background rejection $1-\epsilon_b$ and standard deviation.

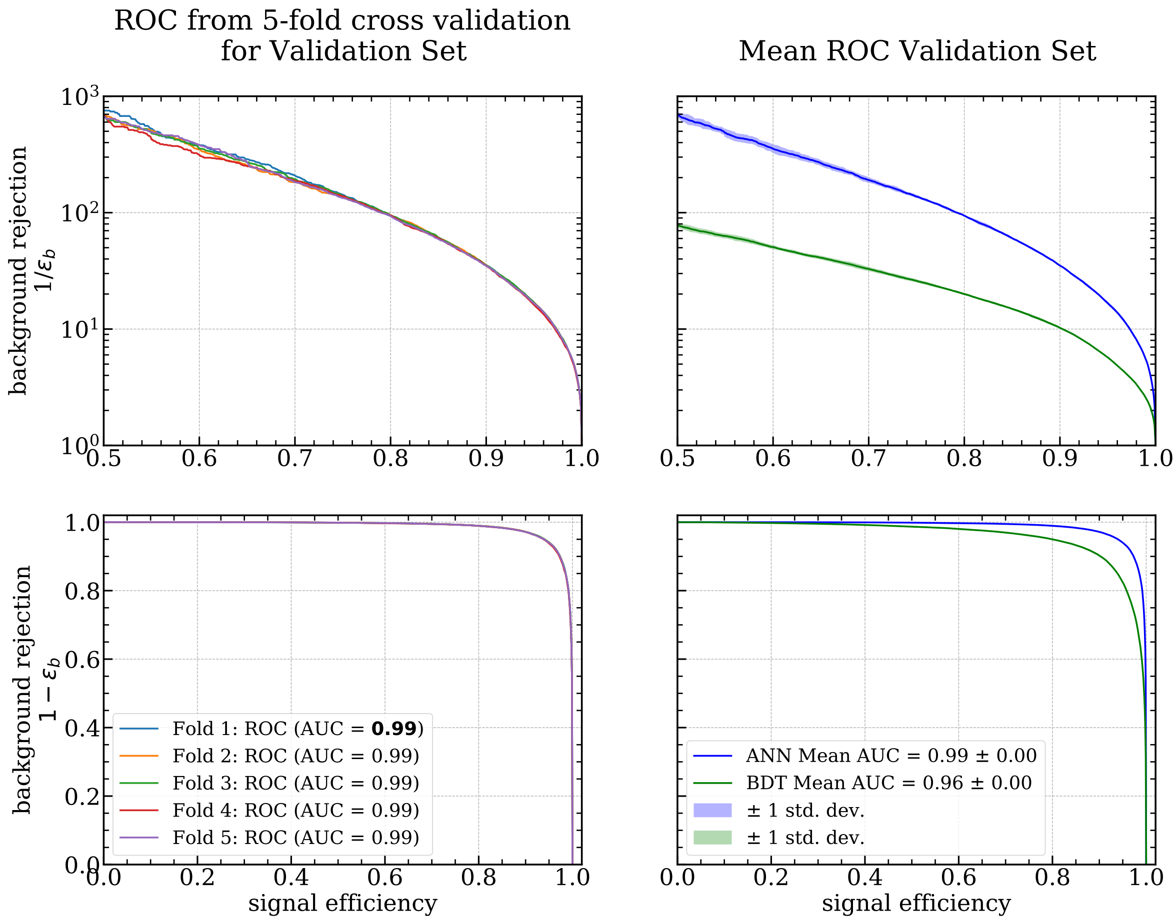


Figure 5.23: Plots based on the prediction on the validation data for **Top Row:** ANN's ROC curve with background rejection $1/\epsilon_b$ for the individual folds (left) and the mean ROC curve of ANN (blue) and BDT (green) with background rejection $1/\epsilon_b$ and standard deviation. **Bottom Row:** ANN's ROC curve with background rejection $1-\epsilon_b$ for the individual folds (left) and the mean ROC curve of ANN (blue) and BDT (green) with background rejection $1-\epsilon_b$ and standard deviation.

A simple signal region can be defined from the $\mathrm{ANN}_{t\bar{t}}$ output by scanning along the output node to find the most significant bin with respect to the Poisson significance. This technique only serves for preliminary purposes to receive a first estimate of the network's power in separating WH from the dominant $t\bar{t}$ background. The most sensitive region is found for events with output node values below 0.006.
Hereby a Poisson significance of 5.14 is gained only considering $t\bar{t}$ as background and a Poisson significance of 0.94 is gained under consideration of all backgrounds with a total of 9.62 ± 0.04 remaining WH MC events and 1.14 ± 0.21 remaining $t\bar{t}$ MC events. Thus, only taking into account a classifier for the $t\bar{t}$ background, the overall significance can be increased by almost 30% from 0.68 to 0.94 as summarized in Table 5.8. Additionally, the impact of requiring 0 b-jets is tested after the cut on the neural network output. The effect turns out to be not significant with respect to the number of $t\bar{t}$ MC events and even worsens the Poisson significance within it's statistical uncertainty due to a decrease in WH MC events. The absent improvement can be explained by the fact the number of b-jets is an input variable to the network. Thus, it learns the b-veto during the training. Finally, a systematic study of (input) variables has been carried out in order to search for any $t\bar{t}$ event excess not being caught by the network in order to further enhance the significance. This approach did not yield in a significant improvement.

$\sqrt{s} = 13$ TeV $\mathcal{L} = 139$ fb^{-1} (Full Run 2)	WH	$t\bar{t}$	Total Bkg.	Sig. ($t\bar{t}$)	Sig. (Total Bkg.)
SR3: 0 SFOS	17.01 ± 0.06	279.17 ± 3.30	529.05 ± 9.88	0.99 ± 0.01	0.68 ± 0.01
$\mathrm{ANN}_{t\bar{t}}$ < 0.006	9.62 ± 0.04	1.14 ± 0.21	29.37 ± 7.75	5.14 ± 0.15	0.94 ± 0.02
b-Veto	9.58 ± 0.04	1.13 ± 0.21	29.23 ± 7.75	5.14 ± 0.17	0.93 ± 0.02

Table 5.8: Event yields with statistical uncertainty of the given processes at cut stage SR3: 0 SFOS, a cut on the $\mathrm{ANN}_{t\bar{t}}$ applied and successive b-veto.

5.4.5 ANN$_{t\bar{t}}$ in the Control Regions

The $t\bar{t}$ control region is consulted to compare the shape of the input variables in MC to data. In case of severe mis-modelling, the input variables can introduce high systematic uncertainties and the consideration of entirely removing this variable from the training can often be seen as trade-off decision between gain in performance to decrease in uncertainty. Figure 5.24 depicts the transverse impact parameter significance $d_0/\sigma(d_0)$ of ℓ_2 in the WZ and $t\bar{t}$ control region as one of the important variables in the fake lepton identification together with the isolation variables. Data and MC show a reasonable agreement within their statistical uncertainty especially in the signal enriched bins. However, there is a certain fluctuation in the data to MC ratio of the $t\bar{t}$ control region that is also visible in the $t\bar{t}$ control region plots of Figures 5.25 and 5.26 showing $E_{\mathrm{T,\,cone,\,20,\,40}}^{\mathrm{topo}}$ of ℓ_1 and ℓ_0 respectively. In case of the WZ control region data and MC are in goof agreement. The fluctuations in the $t\bar{t}$ case can be explained by the low yields in the control region as well as the presence of Z+jets MC events which are observed to introduce large statistical uncertainties in both cases WZ and $t\bar{t}$ respectively.

An alternative data-driven event yield estimation for fake lepton production channels like Z+jets is developed during the time of writing of this thesis which could also replace the purely MC based $t\bar{t}$ control region in case of higher event yields and less statistical uncertainty. More detailed information on this method can be found e.g. in Ref.[77]. All remaining input variables in the WZ and $t\bar{t}$ control regions show a similar modelling and are illustrated in appendix B.3 and B.4.

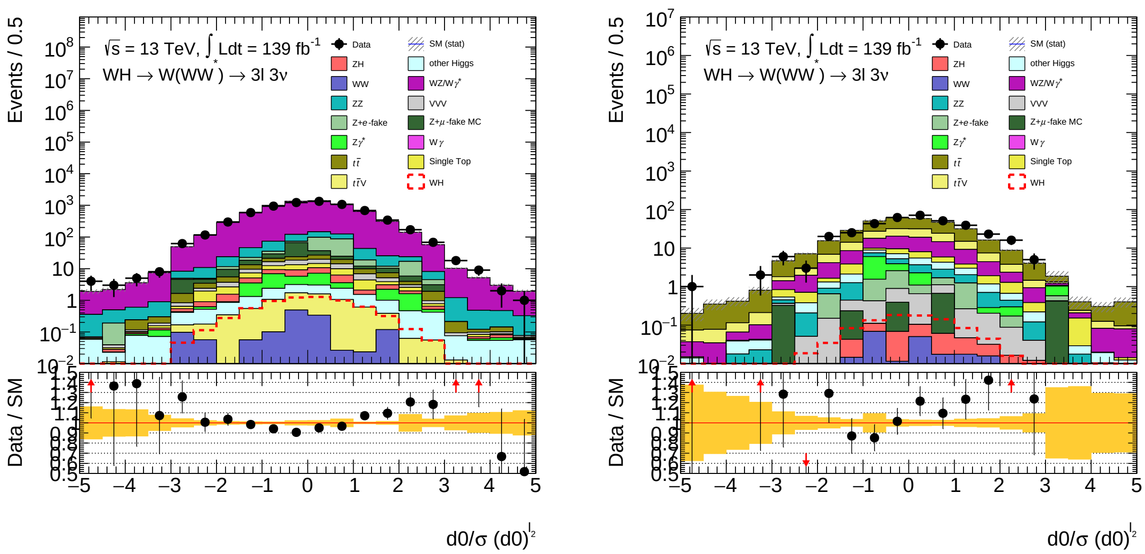


Figure 5.24: Transverse impact parameter significance of ℓ_2 considering the full background contribution in the WZ control region (left) and the $t\bar{t}$ control region (right) for data and MC with statistical uncertainties.

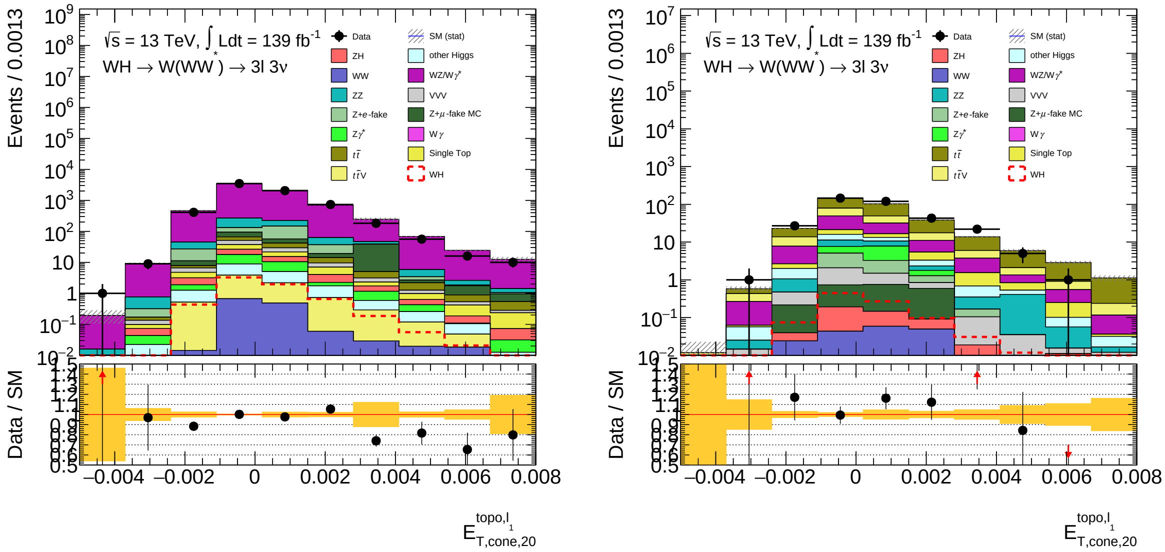


Figure 5.25: $E_{\mathrm{T,\,cone,\,20}}^{\mathrm{topo}}$ of ℓ_1 considering the full background contribution in the WZ control region (left) and the $t\bar{t}$ control region (right) for data and MC with statistical uncertainty.

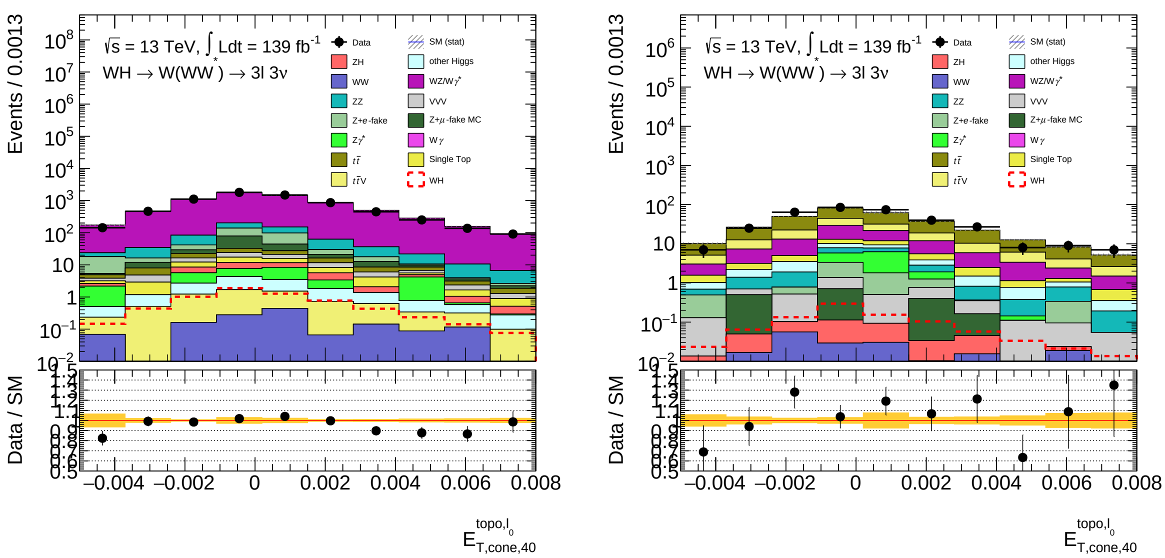


Figure 5.26: $E_{\mathrm{T,\,cone,\,40}}^{\mathrm{topo}}$ of ℓ_0 considering the full background contribution in the WZ control region (left) and the $t\bar{t}$ control region (right) for data and MC with statistical uncertainty.

Figure 5.27 shows the ANN output distribution in the two control regions. While for the $t\bar{t}$ control region a certain excess of MC events is visible in the most significant output bin, it is still within statistical uncertainty of MC and data. A possible explanation could again be the presence of Z+μ-fake events in the region of interest. The WZ control region seems to confirm the argument of statistical limitation in the $t\bar{t}$ case. The data to MC ratio tends to fluctuate much less. While the most significant bin also shows a slight excess of MC events, the ratio is much closer to the expected value of 1 and therefore in good agreement with the expectations.

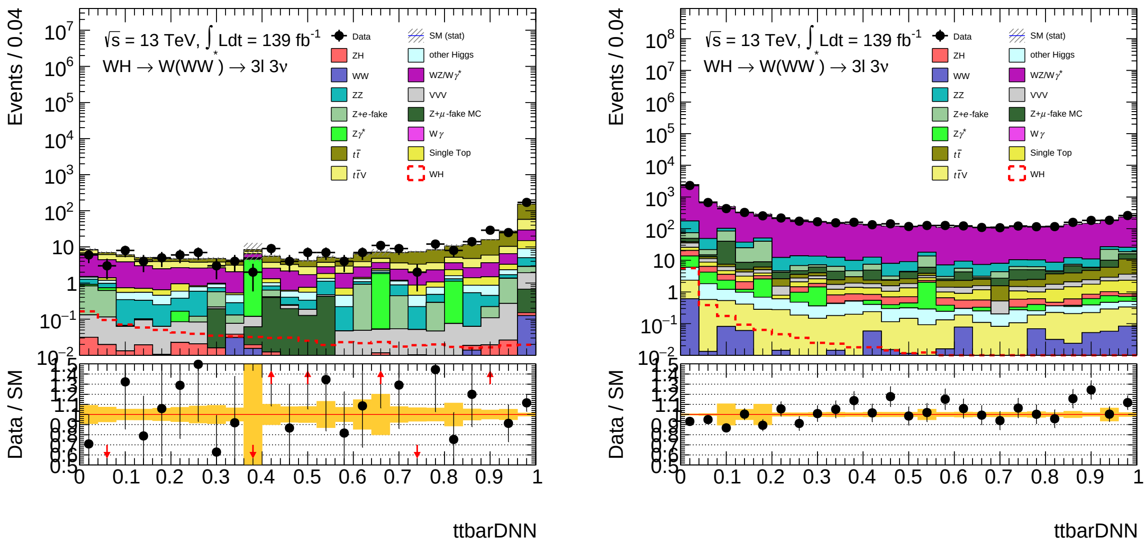


Figure 5.27: ANN$_{t\bar{t}}$ output considering the full background contribution in the $t\bar{t}$ control region (left) and the WZ control region (right) for data and MC with statistical uncertainty.

5.5 ANN_{WZ}: WH against WZ

The WZ decay into three charged leptons constitutes the second most dominant background after $t\bar{t}$ in the Z-depleted signal region. As depicted in Figure 5.28, it either contributes via a leptonic decay of the W with a decay of the Z into $\tau^+\tau^-$ which further decay into leptons of different flavor or via a leptonic decay of the W together with the Z decaying into an e^+e^- pair. While the latter process should not enter SR3 due to the requirement of 0 SFOS lepton pairs, it can in fact pass the cut in case one of the electrons is assigned the wrong charge due to a mismeasurement of the track in the magnetic field. Furthermore, one of these high energetic electrons can radiate a hard bremsstrahlung photon that can convert into an e^+e^- pair. If the electron of opposite charge is reconstructed as a candidate lepton due to the induced track ambiguity because of additional hits in the detector, this can lead to a charge misidentification which is referred to as *charge flip*. A charge flip can in principle happen to electrons from $Z \to \tau\tau$ as well, but since they split their energy with four neutrinos, the $Z \to ee$ process contributes the dominant part of charge flip in the relevant energy range. More details regarding this effect can be found in e.g. Ref.[78].

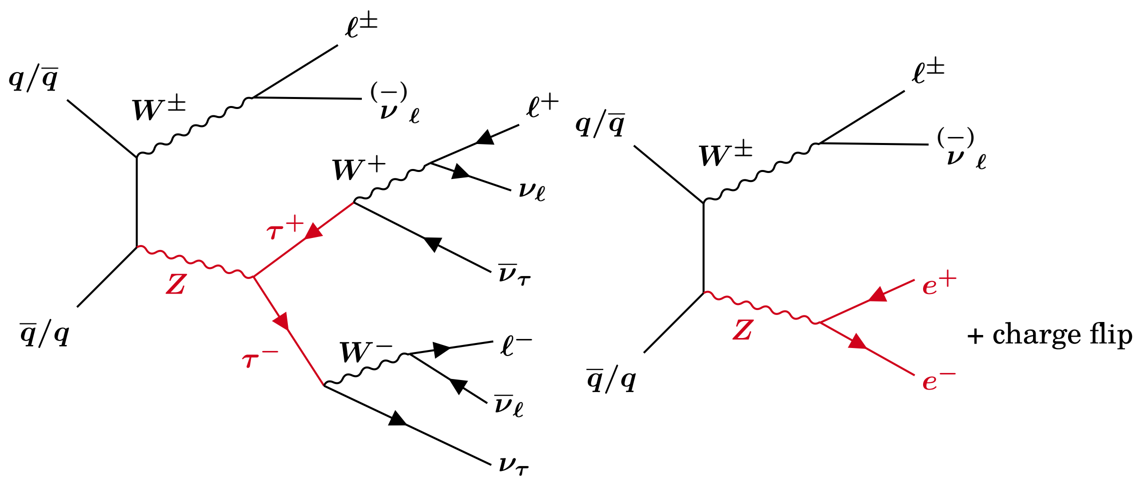


Figure 5.28: Feynman diagrams of WZ production in pp collisions. **Left:** Decay of the Z boson into a pair of τ leptons decaying into W boson that themselves decay into leptons. **Right:** Decay of the Z boson into an electron-positron pair. The latter decay passes the 0 SFOS requirement and enters SR3 when one electron or positron undergoes a charge flip.

The left plot in Figure 5.29 depicts the invariant mass distribution of same sign electrons $m_{inv}^{e^\pm e^\pm}$ in SR3. As discussed, $Z \to e^+e^-$ events where an electron undergoes a charge flip can pass the requirements. Hence the distribution shows a maximum at the rest mass of the Z boson. Additionally, events with e.g. $W^+ \to e^+\nu$ and $Z \to \tau^+\tau^- \to e^+\mu^-\ 4\nu$[2] (or charge conjugated) may contribute to the invariant mass distribution of same-sign electrons.

[2]Neglecting a detailed neutrino description for better readability.

The right plot in Figure 5.29 indicates the number of MC events where the product of the sum of reconstructed charges and the sum of truth charges is positive, referred to as q ID (in green), and the number of MC events where the product of the sum of reconstructed charges and the sum of truth charges is negative, referred to as q MID (in red).

$$\text{q ID} \equiv \left(\sum_{i=0}^{2} q_{reco}^{\ell_i}\right) \times \left(\sum_{i=0}^{2} q_{truth}^{\ell_i}\right) > 0, \tag{5.1}$$

$$\text{q MID} \equiv \left(\sum_{i=0}^{2} q_{reco}^{\ell_i}\right) \times \left(\sum_{i=0}^{2} q_{truth}^{\ell_i}\right) < 0. \tag{5.2}$$

For events where the product is positive the charge identification is considered to be successful, neglecting double charge flip events, while events with a negative product are assumed to include a lepton with misidentified charge. The ratio of q ID to q MID in SR3 is 38.14 : 29.05 = 1.32 and can, together with $m_{inv}^{e^\pm e^\pm}$, be used as a good figure of merit in the conclusion of what the network has learned.

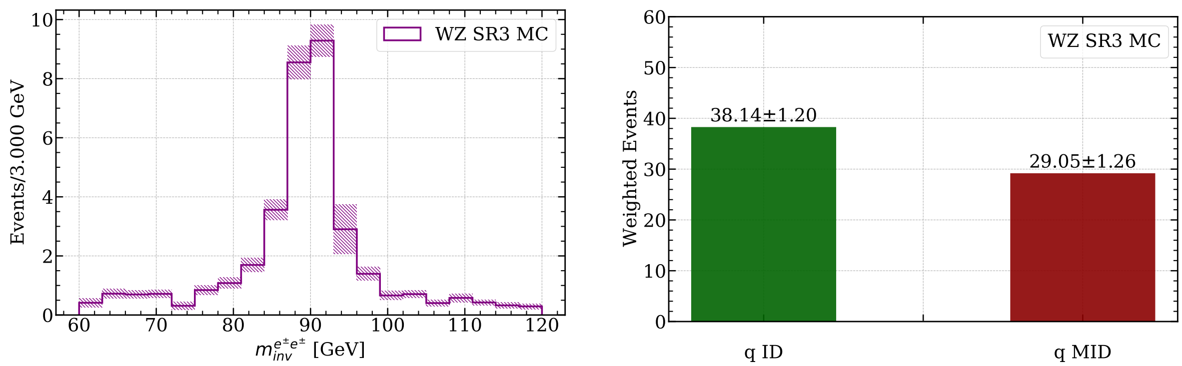


Figure 5.29: Left: Invariant mass distribution of same sign same sign electron pairs $m_{inv}^{e^\pm e^\pm}$ around the rest mass of Z in the WZ SR3 MC. **Right:** Number of events where product of the sum of reconstructed charges and the sum of truth charges is positive, referred to as q ID (green) and the number of events where the product of the sum of reconstructed charges and the sum of truth charges is negative, referred to as q MID

5.5.1 Training and Validation MC Data Sets

Like in the previous chapter the WH SR12 sample is used for training and hyperparameter optimization. Since lepton flavor related variables in the WZ process are sensitive to the zero and non-zero SFOS signal regions, a dedicated training sample is produced in order to have sufficient number of events for a WZ SR3 training sample. Even though only well isolated leptons are expected in both processes, the same loose isolation criteria as in chapter 5.4.1 are applied in order to enhance the size of the training samples as much as possible. This yields again in 1.001.300 WH SR12 raw events and furthermore 78.330 WZ SR3 raw events. Again, the events are reweighted to have an equal class balance in order not to bias the network in favor of one class. As in the case of $t\bar{t}$ the network is finally validated on the nominal WH and WZ SR3 MC which are used for the analysis.

Additionally, a dedicated study with fast simulation WH and WZ SR3 MC samples both including more than a million raw events is carried out. The samples include variables based on smeared particle four-vectors instead of variables derived from the full ATLAS detector reconstruction. Despite much larger training samples, the approach is restricted to the kinematic variables and therefore the network's performance is observed to be worse than by using the variables described in subsection 5.5.2.

5.5.2 Variable Selection

$W(Z\rightarrow \tau^+\tau^-)$ Identification Variables

In the WZ process with the Z decaying into a τ lepton pair the final state includes 4 neutrinos from two τ decays and one neutrino from $W \rightarrow \ell\nu$. The number of degrees of freedom due to the non-detectable neutrinos can be reduced from $5 \times 4 = 20$ down to 7 by neglecting the neutrino rest mass due to $E_\nu \approx |\vec{p}_\nu|$ and assuming collinearity between the decay products of τ leptons originating from the Z decay. When further combining the momentum vectors of the neutrinos from the same τ decay, the number of degrees of freedom reduces down to 5. With only four constraints however (m_W, m_Z and momentum conservation in the transverse plane), the system is underdetermined.

Therefore a previous analysis[79] developed a numerical routine that fixes one of the five unknowns by scanning along the range of possible values of the latter, trying to calculate the remaining four unknowns with this choice. If all unknowns are calculated and the values are physically reasonable, the result is considered a successful reconstruction. Since in the procedure not all results have a physical solution for kinematic reasons, the fraction of successful reconstructions of the $Z \rightarrow \tau\tau$ decay, termed F_{α_1}, is calculated and used as input instead of trying to find an approximation for the direct reconstruction of the Z boson rest mass. Due to the absence of a Z boson in the WH process, the distribution of the latter tends towards lower values as illustrated in the according normalized distributions for WH in blue and WZ in red in the left plot of Figure 5.30. While WH even peaks at a value of $\mathrm{F}_{\alpha_1} = 0.5$, both distributions drop beyond this value. It corresponds to a corner case where only one out of two possible solutions is calculable as input for subsequent calculations within the routine.

In addition, the distribution of the transverse impact parameter significance is expected to differ in WZ from the corresponding distribution in WH. It can be explained by the fact that leptons from the $Z \rightarrow \tau\tau$ decay tend to show a greater distance to the primary vertex due to the distinct lifetime of the τ lepton which translates into larger parameter values. The right plot in Figure 5.30 and the plots in Figure 5.31 depict the normalized $d_0/\sigma(d_0)$ distributions of all three lepton candidates and reflect this behavior.

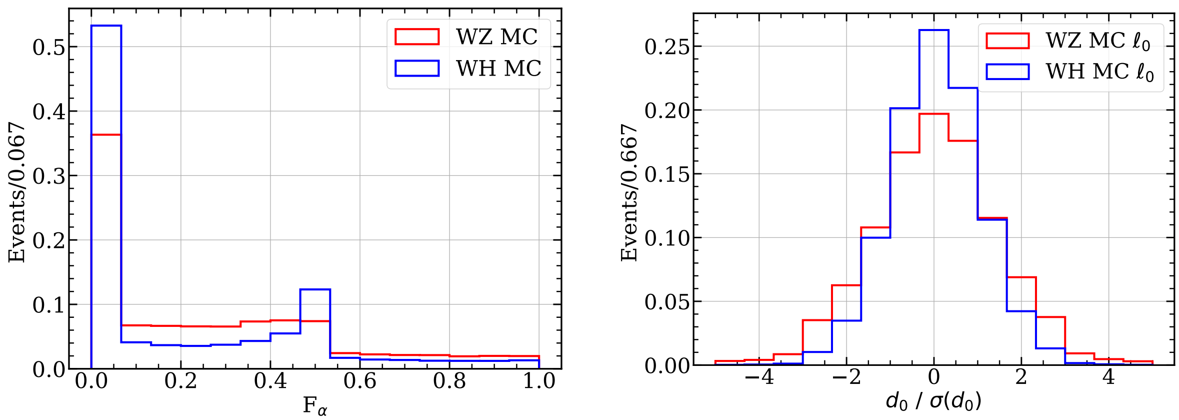


Figure 5.30: Left: Normalized distribution of F_{α_1}. **Right:** Normalized distribution of the transverse impact parameter significance of ℓ_0

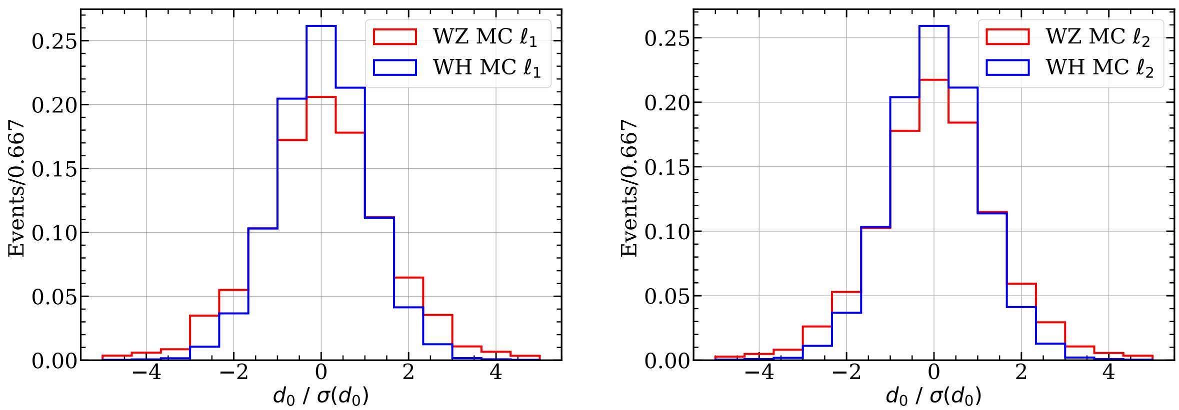


Figure 5.31: Left: Normalized distribution of the transverse impact parameter significance of ℓ_1. **Right:** Normalized distribution of the transverse impact parameter significance of ℓ_2

$W(Z\rightarrow e^+e^-$ + Charge Flip) Identification Variables

The process $W(Z\rightarrow e^+e^-)$ can enter SR3 in case of a misidentified electron that originates from a bremstrahlung-induced photon conversion. This way a high p_T electron from the Z decay can be misidentified as a result of the ambiguity of multiple electron tracks in the reconstruction step caused by the cascade $e^-_{\text{hard}} \rightarrow e^-_{\text{soft}} + \gamma_{\text{hard}} \rightarrow e^-_{\text{soft}} + e^-_{\text{soft}} + e^+_{\text{hard}}$. Bremsstrahlung can affect the track reconstruction leading to a mismeasured p_T as well as it can have an impact on the calorimeter-based measurement due to leakage of energy outside of the reconstructed calorimeter cluster. A way of identifying electrons that were produced in photon conversion is by considering the ratio of the calorimeter energy over the track momentum E/p. Ideally, the ratio would be 1 since the energy momentum relation for high relativistic particles simplifies to $E \approx p$. However, energy losses of the electron translate into a higher track curvature from which a lower momentum is calculated in the reconstruction. Thus, charge flip electrons tend to have $E/p > 1$.

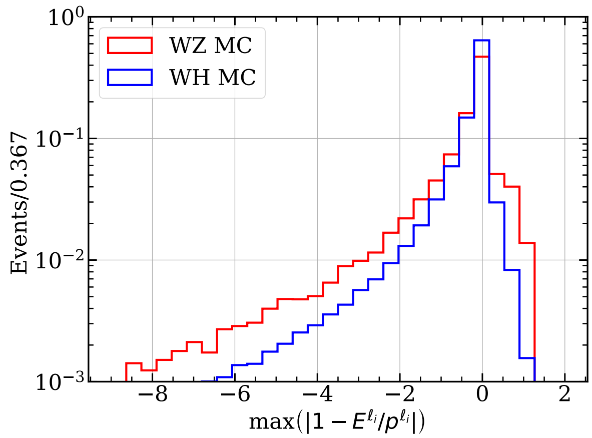
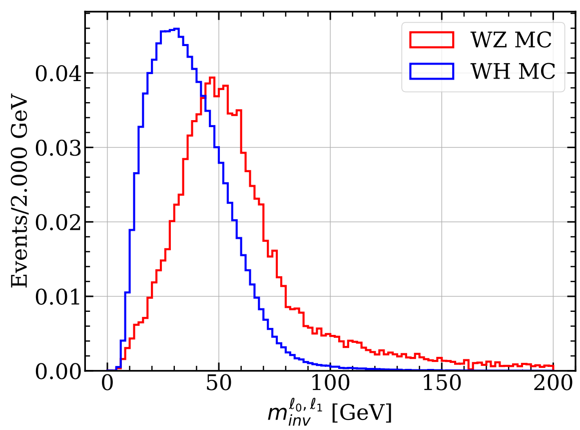


Figure 5.32: Normalized distribution of the greatest deviation of E^{ℓ_i}/p^{ℓ_i} compared between all three candidate leptons.

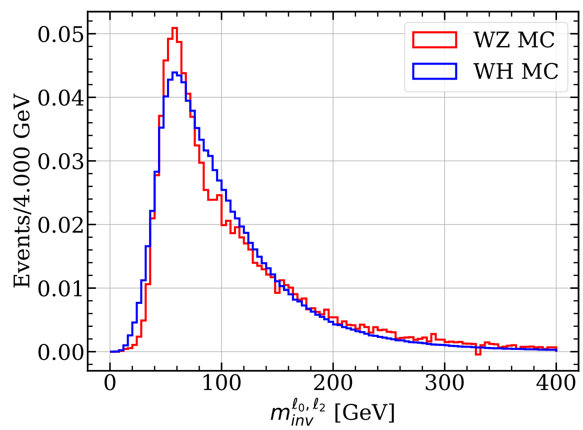
A variable is constructed that measures the greatest deviation of all three E^{ℓ_i}/p^{ℓ_i} from 1, $\max\left(\mathrm{abs}\left(1 - E^{\ell_i}/p^{\ell_i}\right)\right)$, per event. Here E^{ℓ_i} is the reconstructed calorimeter-based energy and p^{ℓ_i} is the track-based momentum of lepton candidate ℓ_i. Since muons are minimum ionizing particles and the ANN is supposed to focus on electrons originating from photon conversion, the muons' E/p can be set to 1 as a good approximation. In Figure 5.32 the resulting normalized distributions are shown. It should be noted that the algorithm picks the greatest deviation in absolute, but the sign is kept in the input variable which is why negative values are visible. As expected the WZ distribution in red shows more deviation from 1 than the corresponding WH distribution does. Additionally, different combinations such as $\min\left(1 - E^{\ell_i}/p^{\ell_i}\right)$ and $\max\left(1 - E^{\ell_i}/p^{\ell_i}\right)$ are tested as input variables. These do not improve the performance as much as $\max\left(\mathrm{abs}\left(1 - E^{\ell_i}/p^{\ell_i}\right)\right)$ and are therefore not considered as input variables.

Spin Correlation and Lepton Systems Variables

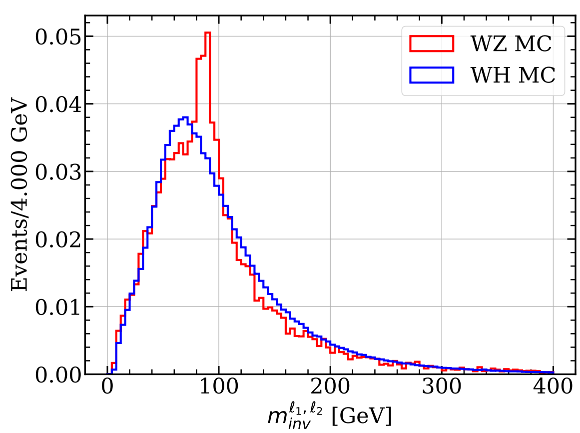
As explained in the chapter on $t\bar{t}$, the spin correlation in the Higgs decay forces the invariant mass distributions of $m_{inv}^{\ell_0,\ell_1}$ towards smaller values compared to the corresponding background distributions as illustrated in Figure 5.33 (a). Further the normalized invariant mass distribution of the same sign lepton candidates $m_{inv}^{\ell_1,\ell_2}$ in Figure 5.33 (c) shows a distinct resonance around the Z rest mass in the background distribution and can therefore be also considered an important variable in the charge flip detection. Systematic studies show that the invariant mass distributions $m_{inv}^{\ell_0,\ell_2}$ and $m_{inv}^{\ell_0,\ell_1,\ell_2}$ (Figures 5.33 (b,d)) improve the performance by adding complementary information on the ℓ_0-ℓ_2 system.



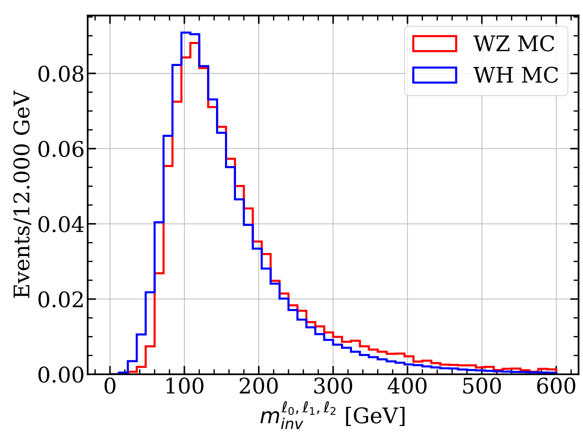
(a) $m_{inv}^{\ell_0,\ell_1}$



(b) $m_{inv}^{\ell_0,\ell_2}$



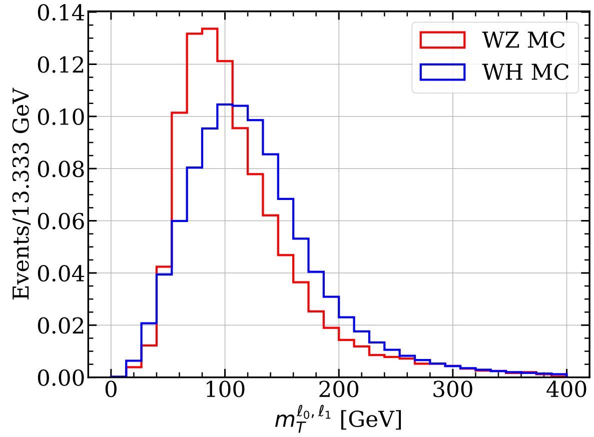
(c) $m_{inv}^{\ell_1,\ell_2}$



(d) $m_{inv}^{\ell_0,\ell_1,\ell_2}$

Figure 5.33: Normalized invariant mass distributions of all possible lepton candidate combinations.

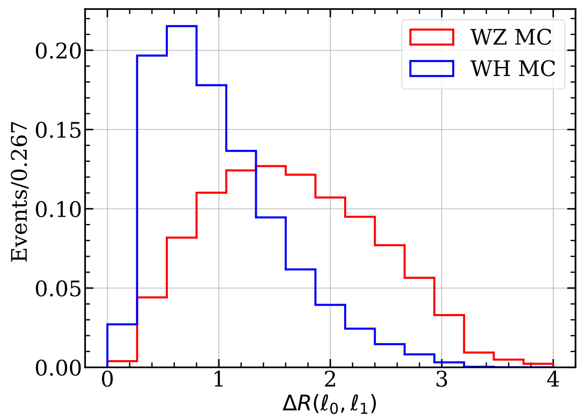
The impact on the performance of the transverse mass distributions is studied with the result that $m_T^{\ell_0,\ell_1}$ is most powerful in terms of increasing the area under the ROC curve and the Poisson significance respectively. Figure 5.34 illustrates that the transverse mass of the ℓ_0-ℓ_1 system shows a tendency to peak around the Z rest mass for the background distribution and to peak towards the Higgs rest mass for the signal distribution.



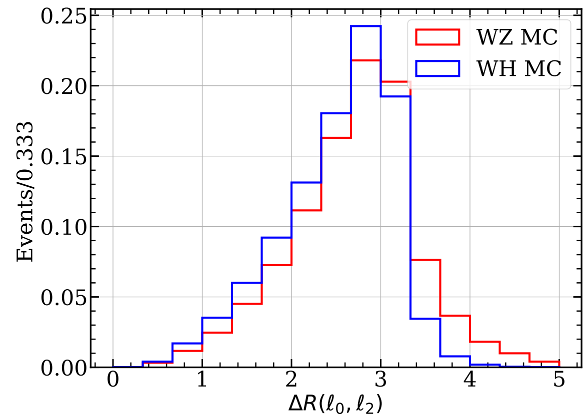
(a) $m_T^{\ell_0,\ell_1}$

Figure 5.34: Normalized transverse mass distributions of the ℓ_0-ℓ_1 system.

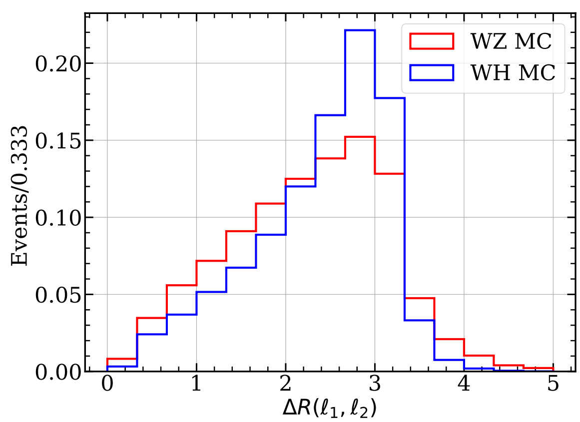
Angular distributions are tested with the conclusion that the network can again benefit from the spin correlations between ℓ_0 and ℓ_1 as well as from the difference in the same sign lepton system due to the fact ℓ_1 likely stems from the Higgs decay while ℓ_2 is more probable originating from the associated W decay. Therefore $\Delta R^{\ell_0,\ell_1}$ and $\Delta\eta^{\ell_0,\ell_1}$ tend towards smaller values in the case of WH compared to WZ as can be observed in the normalized distributions in Figures 5.35 (a) and 5.35 (d). In Figures 5.35 (c) and 5.35 (f), depicting the ℓ_1-ℓ_2 system, the normalized distributions differ due to the fact that these leptons are likely to originate from a massive spin 0 (H) and a massive spin 1 particle (W) for WH in contrast to WZ where the leptons are decay products of to massive spin 1 particles. In a similar way but for oppositely charged leptons however, the network gains performance when taking into account $\Delta R^{\ell_0,\ell_2}$ and $\Delta\eta^{\ell_0,\ell_2}$ that are illustrated in Figures 5.35 (b) and 5.35 (e).



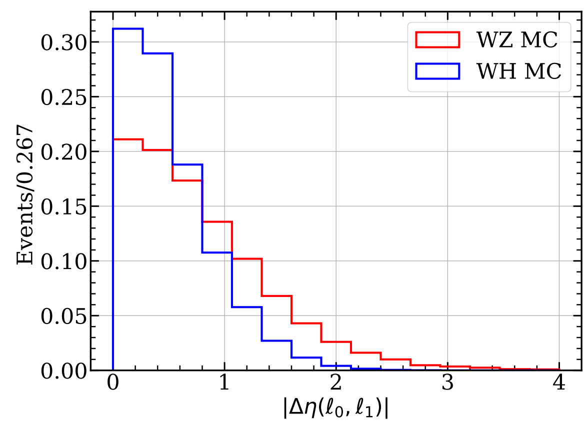
(a) $\boldsymbol{\Delta R^{\ell_0,\ell_1}}$



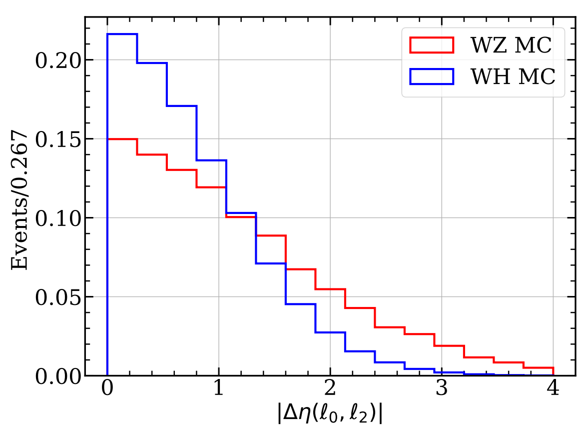
(b) $\boldsymbol{\Delta R^{\ell_0,\ell_2}}$



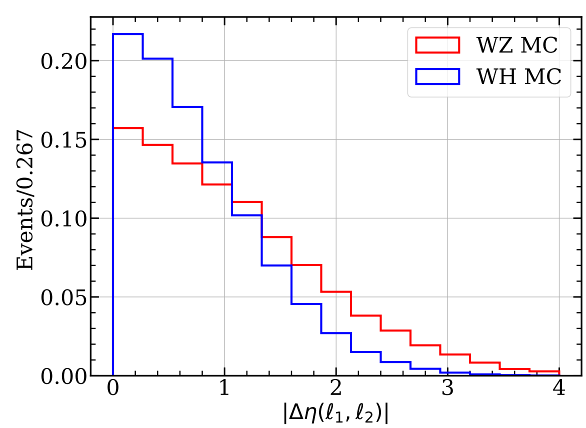
(c) $\boldsymbol{\Delta R^{\ell_1,\ell_2}}$



(d) $\boldsymbol{\Delta \eta^{\ell_0,\ell_1}}$



(e) $\boldsymbol{\Delta \eta^{\ell_0,\ell_2}}$

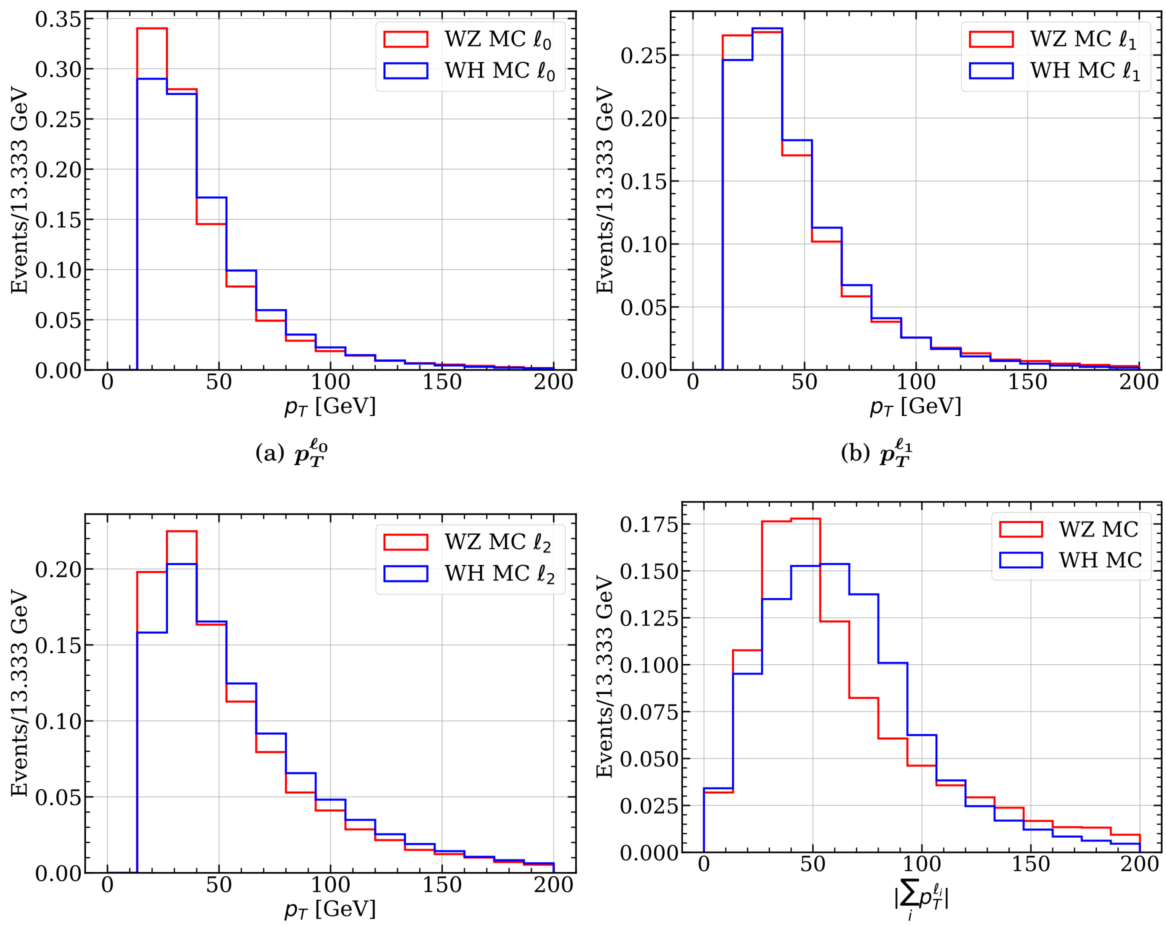


(f) $\boldsymbol{\Delta \eta^{\ell_1,\ell_2}}$

Figure 5.35: Normalized distributions of the angular distance $\boldsymbol{\Delta R}$ for all lepton candidate combinations and normalized distributions of the difference in η for all lepton candidate combinations

Additional Variables

Additional input variables that are used in the training are the lepton candidate's p_T as well as their sum where the difference in the distributions due to the additional neutrinos from the $Z \to \tau\tau$ decay becomes most visible (compare Figure 5.36). The excess of neutrinos in the $W(Z \to \tau\tau)$ process further affects E_T^{miss} related variables. The missing transverse energy E_T^{miss}, its $\varphi^{E_T^{miss}}$-direction in the transverse plane and the difference between the latter and φ^{ℓ_i} of the reconstructed candidate leptons are shown in Figure 5.37. With the total number of jets, n_{jet}, another event topological variable is used. A summary of all input variables to ANN_{WZ} is given in Table 5.9.



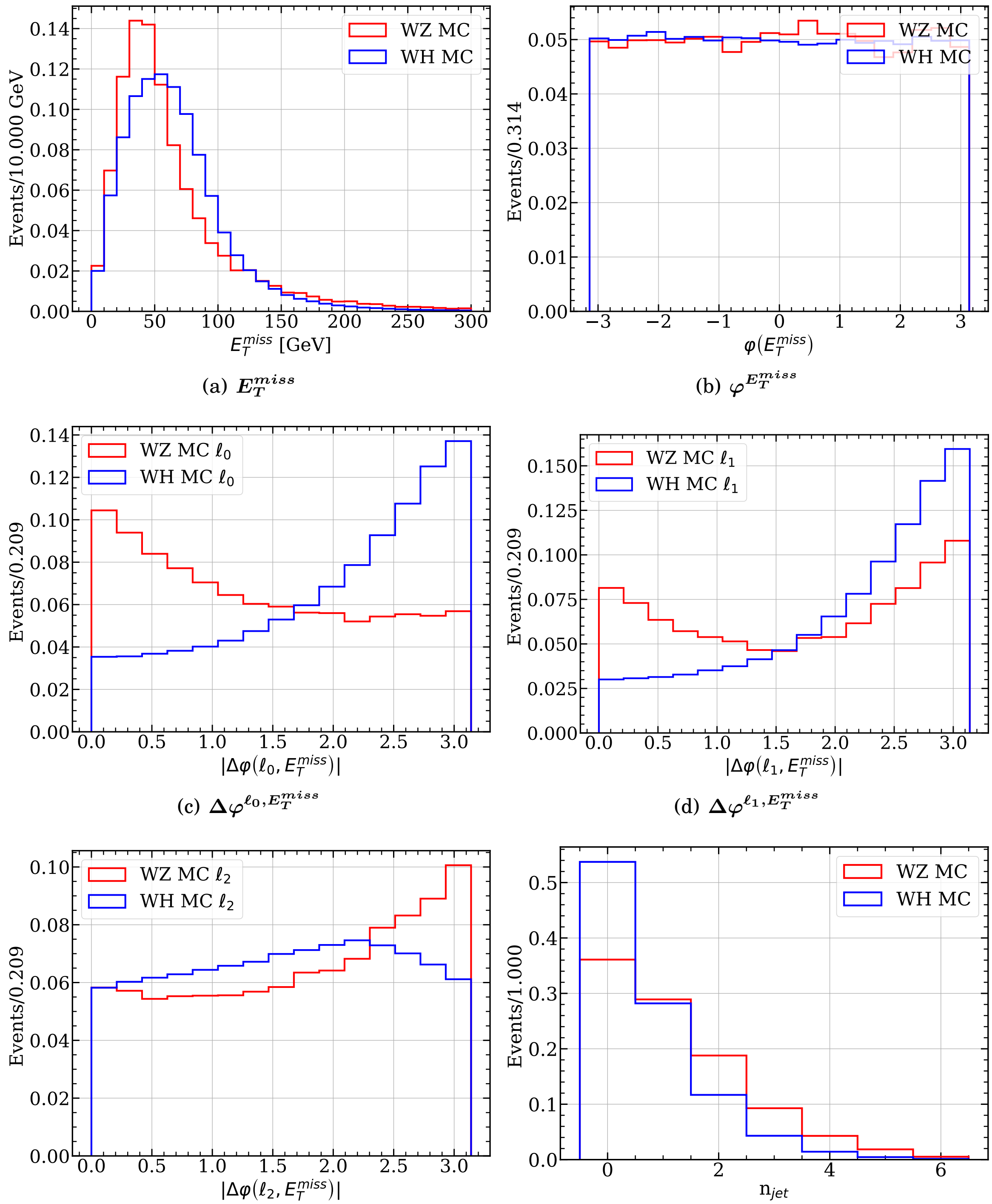
(a) $p_T^{\ell_0}$

(b) $p_T^{\ell_1}$

(c) $p_T^{\ell_2}$

(d) $\left|\sum_i p_T^{\ell_i}\right|$

Figure 5.36: Normalized p_T distributions of all three lepton candidates as well as the normalized sum of p_T distribution.



(a) E_T^{miss} (b) $\varphi^{E_T^{miss}}$

(c) $\Delta\varphi^{\ell_0, E_T^{miss}}$ (d) $\Delta\varphi^{\ell_1, E_T^{miss}}$

(e) $\Delta\varphi^{\ell_2, E_T^{miss}}$ (f) n_{jet}

Figure 5.37: Normalized distributions of E_T^{miss} and corresponding direction in the transverse plane $\varphi^{E_T^{miss}}$. Normalized distributions of the difference in φ of E_T^{miss} and all lepton candidates and n_{jet}.

ANN$_{WZ}$ input variables

F_{α_1}	$d_0/\sigma(d_0)^{\ell_0}$	$d_0/\sigma(d_0)^{\ell_1}$	$d_0/\sigma(d_0)^{\ell_2}$
$\max\left(\left\|1 - E^{\ell_i}/p^{\ell_i}\right\|\right)$	$m_T^{\ell_0,\ell_1}$	E_T^{miss}	$\varphi^{E_T^{miss}}$
$m_{inv}^{\ell_0,\ell_1}$	$m_{inv}^{\ell_0,\ell_2}$	$m_{inv}^{\ell_1,\ell_2}$	$m_{inv}^{\ell_0,\ell_1,\ell_2}$
$\Delta R^{\ell_0,\ell_1}$	$\Delta R^{\ell_0,\ell_2}$	$\Delta R^{\ell_1,\ell_2}$	n_{jet}
$p_T^{\ell_0}$	$p_T^{\ell_1}$	$p_T^{\ell_2}$	$\left\|\sum_i p_T^{\ell_i}\right\|$
$\Delta\eta^{\ell_0,\ell_1}$	$\Delta\eta^{\ell_0,\ell_2}$	$\Delta\eta^{\ell_1,\ell_2}$	
$\Delta\varphi^{\ell_0,E_T^{miss}}$	$\Delta\varphi^{\ell_1,E_T^{miss}}$	$\Delta\varphi^{\ell_2,E_T^{miss}}$	

Table 5.9: Summary of input variables for ANN$_{WZ}$

5.5.3 Hyperparameter Optimization

Analogously to subsection 5.4.3 a hyperparameter optimization procedure is carried out simultaneously with a network architecture grid search. In the final setup the ANN is trained on a batch size of 4096 samples over 40 epochs with an early stopping after 10 epochs in case the area under the ROC curve has not increased. Training is done in a 5-fold cross-validation with 80% training data size and 20% test data size respectively. Again, Adam is used as optimization algorithm with a learning rate of 0.001 and binary crossentropy loss and swish activation function with dropout applied. Furthermore the network again performed best with the input variables scaled by the "RobustScaler". The most crucial part is the network architecture. Here, a slightly larger configuration compared to the $t\bar{t}$ case of three layers with 80, 64 and 16 neurons with corresponding dropout rates of 0.33, 0.30 and 0.26 is found to perform best. All hyperparameters are summarized in Table 5.10.

Hyperparameter	Tested [within]	Final Value
Batch Size	[1024, 12.288]	4096
Epochs	[30, 100]	40
k-Folds	5, 10	5
Training Sample Size	80%	80%
Testing Sample Size	20%	20%
Optimizer	Adam	Adam
Learning Rate	[0.01, 0.0001]	0.001
Loss	Binary Crossentropy	Binary Crossentropy
Scaler	Robust/MinMax/Standard	RobustScaler
Activation	ReLu/SeLu/Swish	Swish
Regularisation	AlphaDropout/Dropout	Dropout

Table 5.10: Summary of hyperparameters that are tested and the set of final hyperparameters used in the ANN_{WZ} network configuration.

The hyperparameters and input variables of the retrained BDT_{WZ} are summarized in Table 5.11 where $d_0/\sigma(d_0)^{\ell_{mid p_T}}$ and $d_0/\sigma(d_0)^{\ell_{min p_T}}$ are the transverse impact parameters of the candidate lepton with lowest p_T and the candidate lepton of opposite charge to the latter with medium p_T.

BDT_{WZ} Configuration	
NTrees	500
MinNodeSize	4%
maxDepth	3
BoostType	AdaBoost
Shrinkage	0.5
nCuts	20

BDT_{WZ} Input Variables
$p_T^{\ell_0}, p_T^{\ell_1}, p_T^{\ell_2}$
E_T^{miss}
$m_{inv}^{\ell_0 \ell_1}$
$m_{inv}^{e^\pm e^\pm}$
$m_T^{\ell_0 \ell_1}$
$d_0/\sigma(d_0)^{\ell_{mid p_T}}$
$d_0/\sigma(d_0)^{\ell_{min p_T}}$

Table 5.11: Left: Hyperparameter configuration of the BDT_{WZ} used to define the classifier baseline. **Right:** Corresponding input variables to the BDT_{WZ} used in this analysis.

5.5.4 Training Results

Similar to the training results in the section on $t\bar{t}$, in this section a summary is given of the training results with help of the introduced figures of merit from chapter 5.4.4. In Figure 5.38 the mean loss is drawn together with the mean area under the ROC curve and the mean Poisson significance determined by a scan on the output distributions for a cut to maximize it, each as a function of epoch taking all 5 folds into account. The loss curve corresponding to the predictions on the test data (first panel green line) stay well below the respective curve corresponding the the loss of the predictions on the train data (blue line) which allows for the conclusion that no complex co-adaption is visible from the loss development while smoothly approaching a local minimum. Mean area and the ROC curve and Poisson significance of the prediction on the test data are well increasing with increasing epoch with no peculiarities in their development.

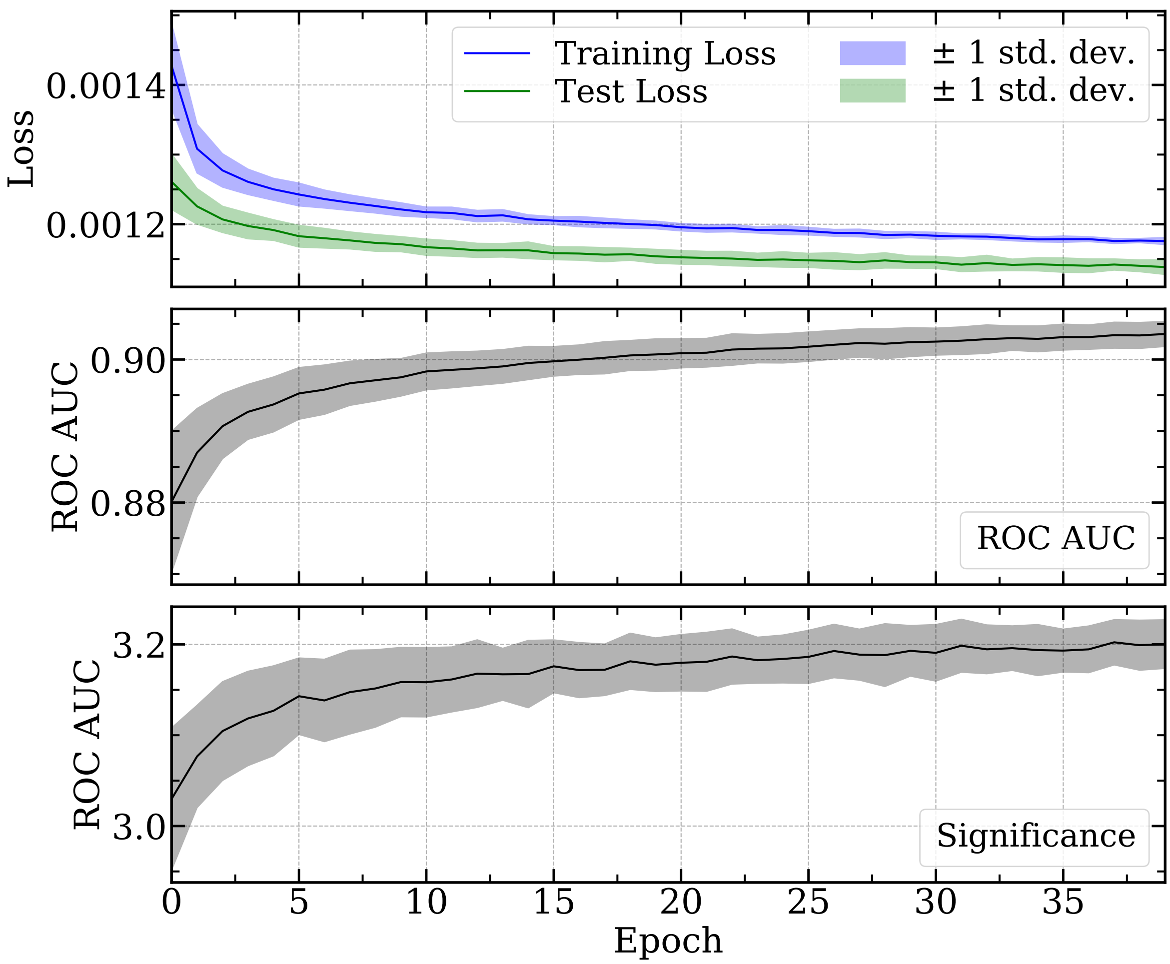


Figure 5.38: Evolution of the loss for train (blue) and test (green) data fold in the first panel, the area under the ROC curve in the middle panel and the Poisson significance in the bottom panel for the test data.

Second, the shape comparison between the predictions on the train data set (histograms) and the test data set (dots) for signal (blue) peaking towards 0 and background (red) peaking towards 1 in Figure 5.39 does not show any severe discrepancies throughout the visible folds 1 and 3. The same behavior can be observed for the remaining folds 2, 4 and 5 in appendix B.2

Due to the low number of samples in the WZ training data set, the ratio slightly fluctuates but especially in the most significant bin not more than $\pm 1\sigma$ around the nominal value which is in good agreement with the hypothesis that the network learned to generalize in this scope.

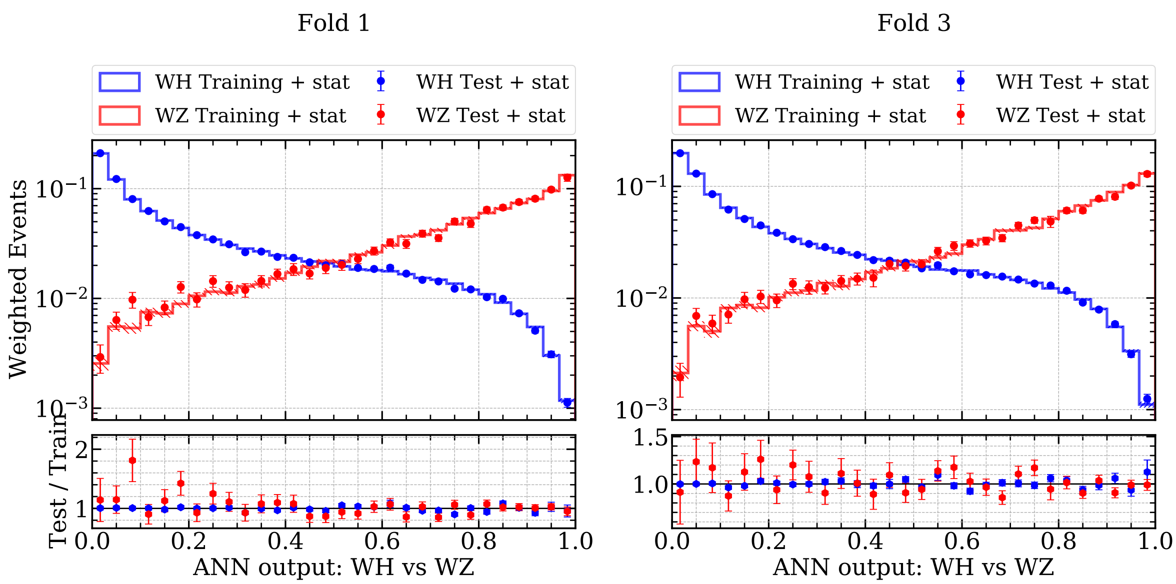


Figure 5.39: The ANN$_{WZ}$'s prediction on train (dots) and test (lines) data folds over the exemplary folds 1 and 3 in the upper plots with the corresponding ratios in the bottom plots.

Finally, the area under the ROC curve is compared for train data (Figure 5.40), test data (Figure 5.41) and validation data (Figure 5.42). In all three steps a very good agreement in performance is observed over the 5 folds. Further, the neural network beats the retrained BDT$_{WZ}$ on average with 0.91/0.91/0.91 to 0.88/0.87/0.88 in case of predictions on train/test/validation data which also indicates a stable ANN performance throughout known and unknown used samples. The highest area under the ROC curve are summarized and compared to the retrained BDT$_{WZ}$ in Table 5.12.

	Training	**Testing**	**Validation**
BDT$_{WZ}$	0.88	0.87	0.88
ANN$_{WZ}$	0.91	0.91	0.91

Table 5.12: Best ROC curves in separating WH from WZ of BDT$_{WZ}$ and ANN$_{WZ}$.

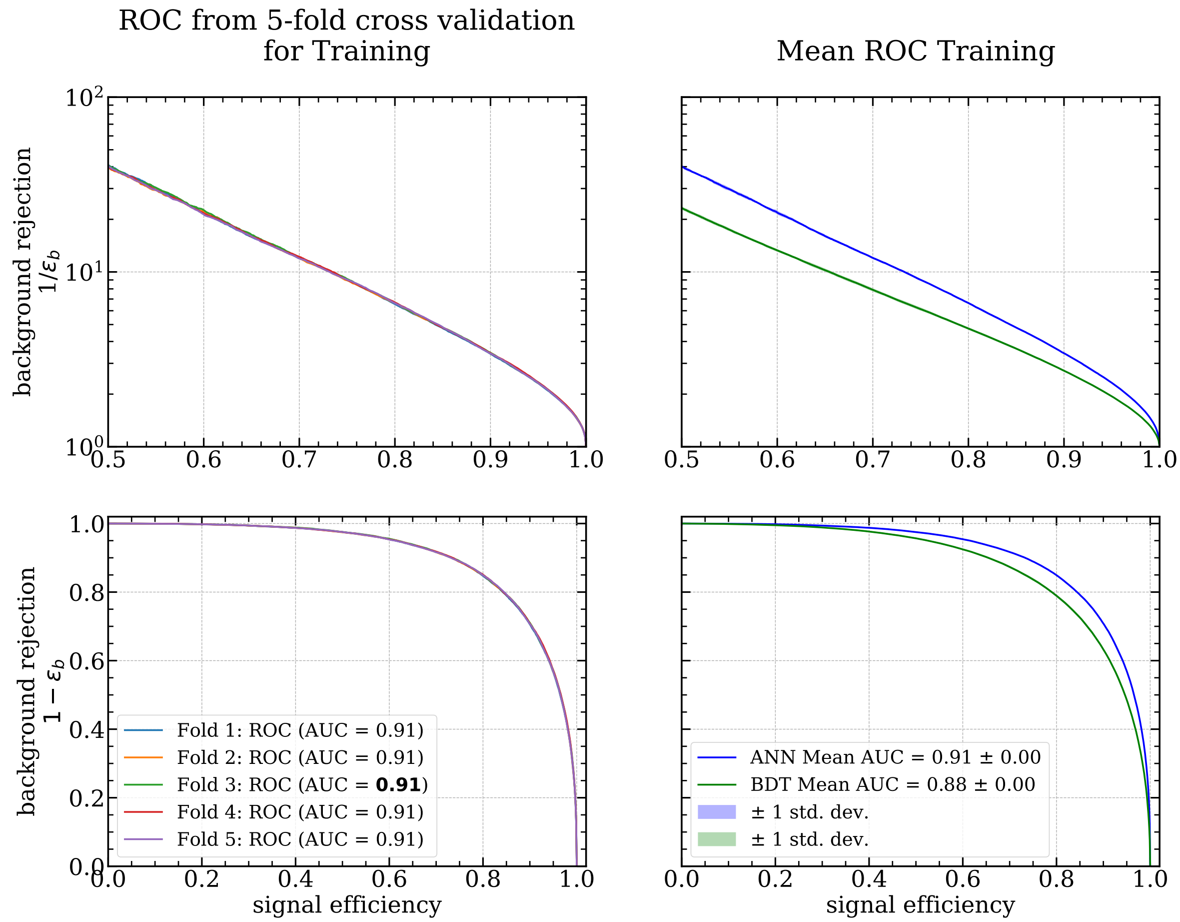


Figure 5.40: Plots based on the prediction on the train data folds for **Top Row:** ANN's ROC curve with background rejection $1/\epsilon_b$ for the individual folds (left) and the mean ROC curve of ANN (blue) and BDT (green) with background rejection $1/\epsilon_b$ and standard deviation. **Bottom Row:** ANN's ROC curve with background rejection $1-\epsilon_b$ for the individual folds (left) and the mean ROC curve of ANN (blue) and BDT (green) with background rejection $1-\epsilon_b$ and standard deviation.

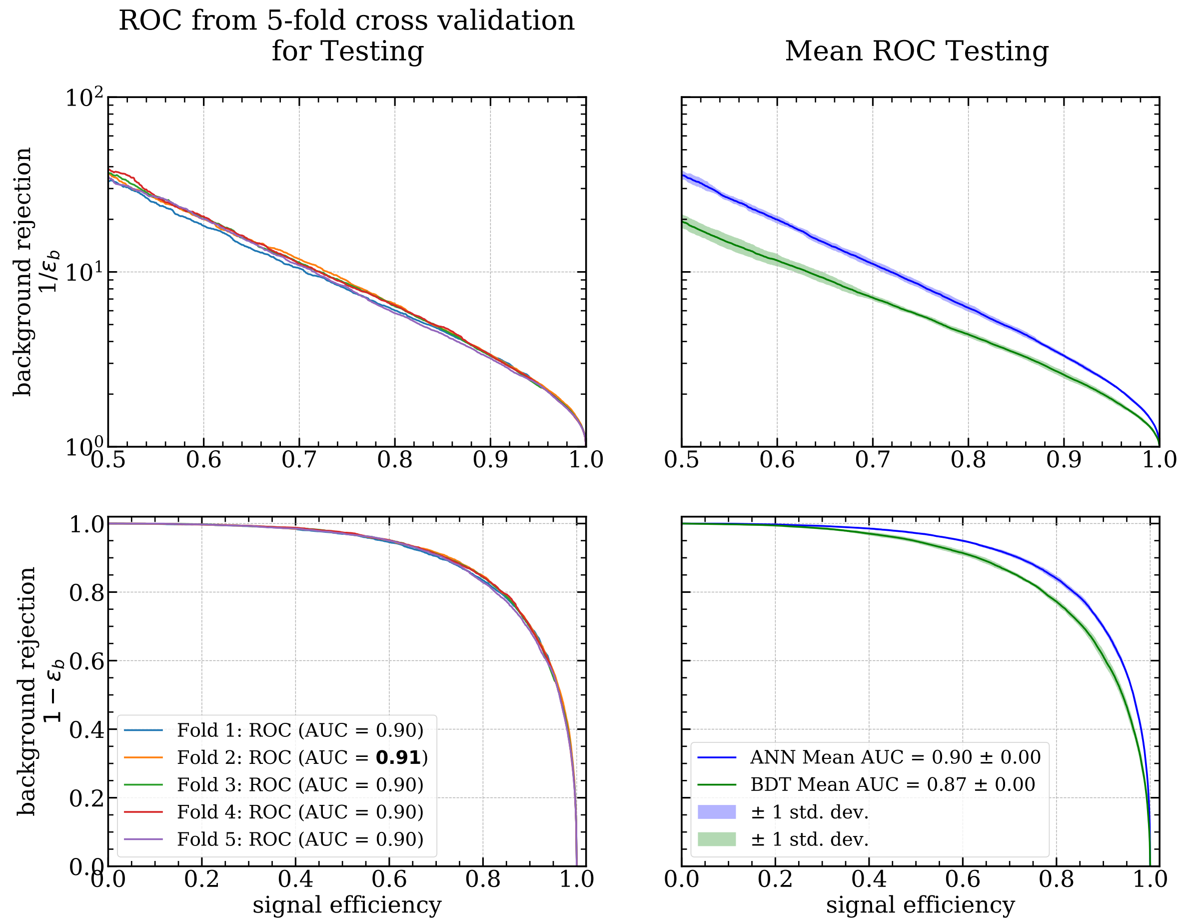


Figure 5.41: Plots based on the prediction on the test data folds for **Top Row:** ANN's ROC curve with background rejection $1/\epsilon_b$ for the individual folds (left) and the mean ROC curve of ANN (blue) and BDT (green) with background rejection $1/\epsilon_b$ and standard deviation. **Bottom Row:** ANN's ROC curve with background rejection $1-\epsilon_b$ for the individual folds (left) and the mean ROC curve of ANN (blue) and BDT (green) with background rejection $1-\epsilon_b$ and standard deviation.

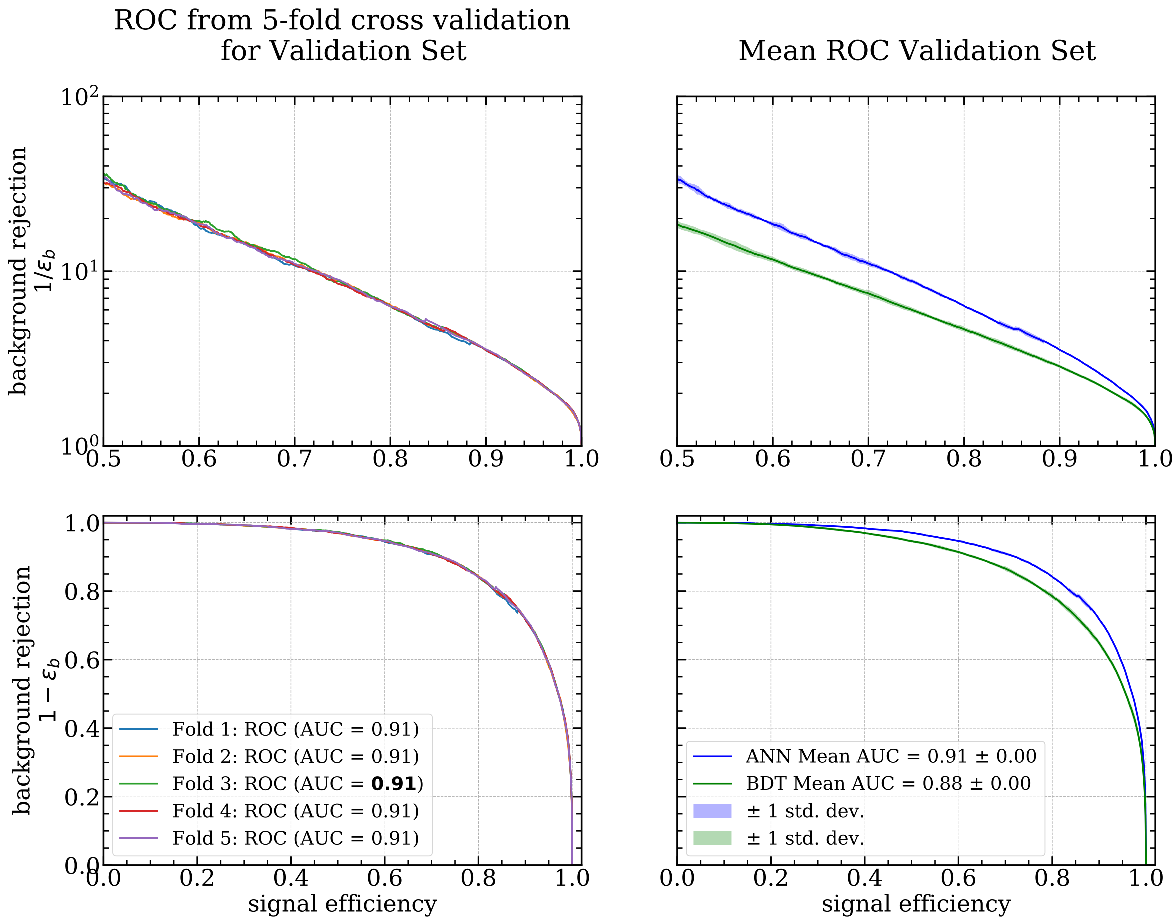


Figure 5.42: Plots based on the prediction on the validation data for **Top Row:** ANN's ROC curve with background rejection $1/\epsilon_b$ for the individual folds (left) and the mean ROC curve of ANN (blue) and BDT (green) with background rejection $1/\epsilon_b$ and standard deviation. **Bottom Row:** ANN's ROC curve with background rejection $1-\epsilon_b$ for the individual folds (left) and the mean ROC curve of ANN (blue) and BDT (green) with background rejection $1-\epsilon_b$ and standard deviation.

Scanning along the ANN_{WZ} output node to find the most significant bin yields a cut at $\text{ANN}_{WZ} < 0.145$. The corresponding yields are summarized in Table 5.13. With 8.65 $\pm$ 0.04 remaining WH MC events and 1.95 $\pm$ 0.28 WZ MC events in SR3 this equals to a significance of 4.15 only considering these two processes. In the scope of the full background contribution that totals 37.59 $\pm$ 1.15 the cut on the network output improves the Poisson significance from 0.68 to 1.34 which is almost a factor of 2. Additionally requiring a b-veto to suppress $t\bar{t}$ induced backgrounds the total background efficiently reduces to 13.24 $\pm$ 0.71 which corresponds to a significance of 2.05 for full Run 2 MC in the most significant bin which is an improvement by more than factor 3 with respect to the 0 SFOS cut level.

$\sqrt{s} = 13$ TeV $\mathcal{L} = 139$ fb^{-1} (Full Run 2)	***WH***	***WZ***	**Total Bkg.**	**Sig. (*WZ*)**	**Sig. (Total Bkg.)**
SR3: 0 SFOS	17.01 ± 0.06	69.70 ± 1.76	529.05 ± 9.88	1.92 ± 0.02	0.68 ± 0.01
$\text{ANN}_{WZ} < 0.145$	8.65 ± 0.04	1.95 ± 0.28	37.59 ± 1.15	4.15 ± 0.14	1.34 ± 0.02
b-Veto	8.33 ± 0.04	1.90 ± 0.28	13.24 ± 0.71	4.05 ± 0.14	2.05 ± 0.04

Table 5.13: Event yields with statistical uncertainty of the given processes at cut stage SR3: 0 SFOS, a cut on the ANN_{WZ} applied and subsequent b-veto. Last column summarizes the Poisson significance at each cut level.

5.5.5 ANN_{WZ} in the Control Regions

Similar to the previous chapter, distributions of the input variables in both control regions are investigated. As one of the most important variables, Figure 5.43 depicts the invariant mass distribution of the ℓ_0-ℓ_1 system $m_{inv}^{\ell_0,\ell_1}$ in the $t\bar{t}$ control region (left) and WZ control region (right). The distributions show good agreement in both regions especially in bins with high fractions of signal contribution. Again, the statistical limitations in the $t\bar{t}$ control region become visible leading to higher statistical uncertainties and slight fluctuations. Going towards higher values of the $m_{inv}^{\ell_0,\ell_1}$, the distribution decreases in the number of events and statistical precision.
Figure 5.44 shows the important E/p related variable that is employed to gain information on WZ events with charge flip. A similar conclusion on the shapes can be drawn here, the Z+jets contribution further introduces higher MC uncertainties. The data to MC ratio in the most significant bins is in good agreement with 1. All remaining input variables in the WZ and $t\bar{t}$ control regions show a similar modelling and are illustrated in appendix B.3 and B.4.

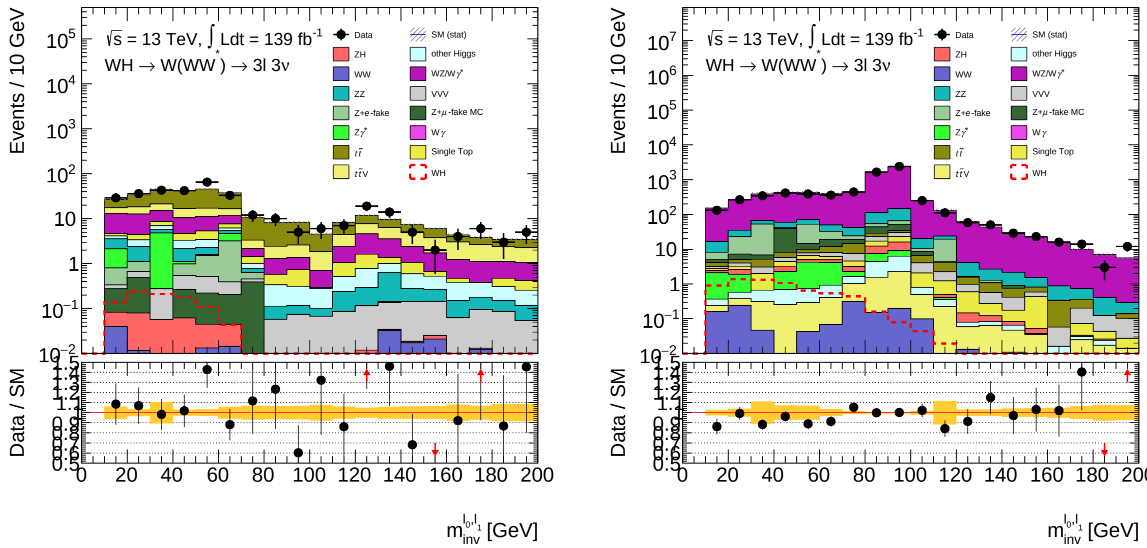


Figure 5.43: Invariant mass distribution $m_{inv}^{\ell_0,\ell_1}$ in the $t\bar{t}$ control region (left) and the WZ control region (right) under consideration of the full background contribution for data and MC with statistical uncertainties.

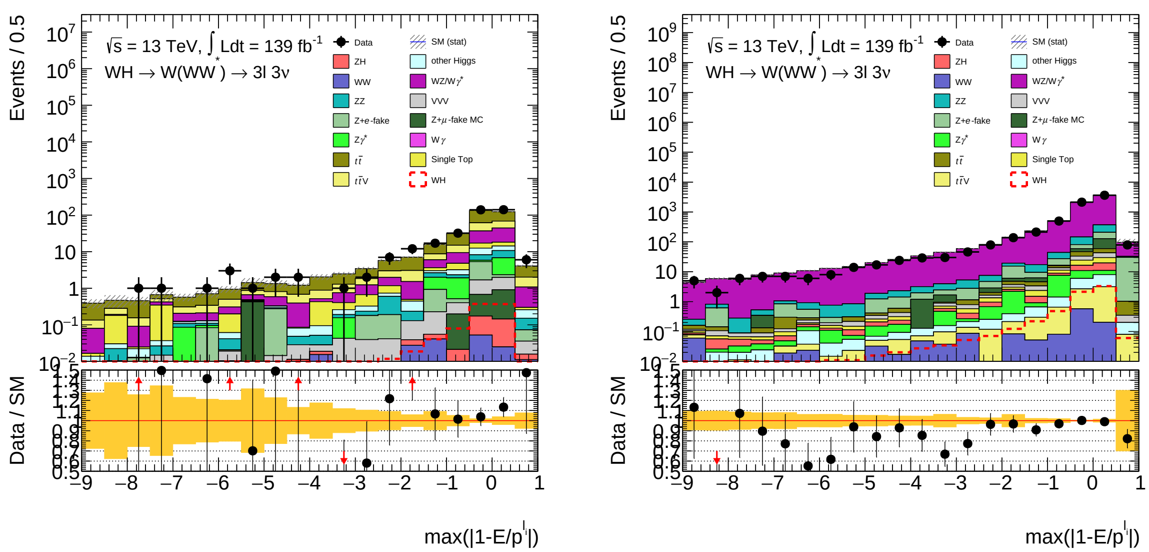


Figure 5.44: Greatest deviation of E^{ℓ_i}/p^{ℓ_i} from 1 in the $t\bar{t}$ control region (left) and the WZ control region (right) under consideration of the full background contribution for data and MC with statistical uncertainties.

Regarding the ANN_{WZ} output distributions for the two control regions given in Figure 5.45 the data to MC ratio tends to fluctuate again for the $t\bar{t}$ CR within the statistical uncertainties. For the WZ CR a slight excess of MC events is visible in the most significant signal region bin while a slight excess of data is visible at the other end of the distribution. Both might again be explained by the Z+jets MC modelling which is under investigation at the time of writing in order to be replaced by a data-driven approach that could account for this.

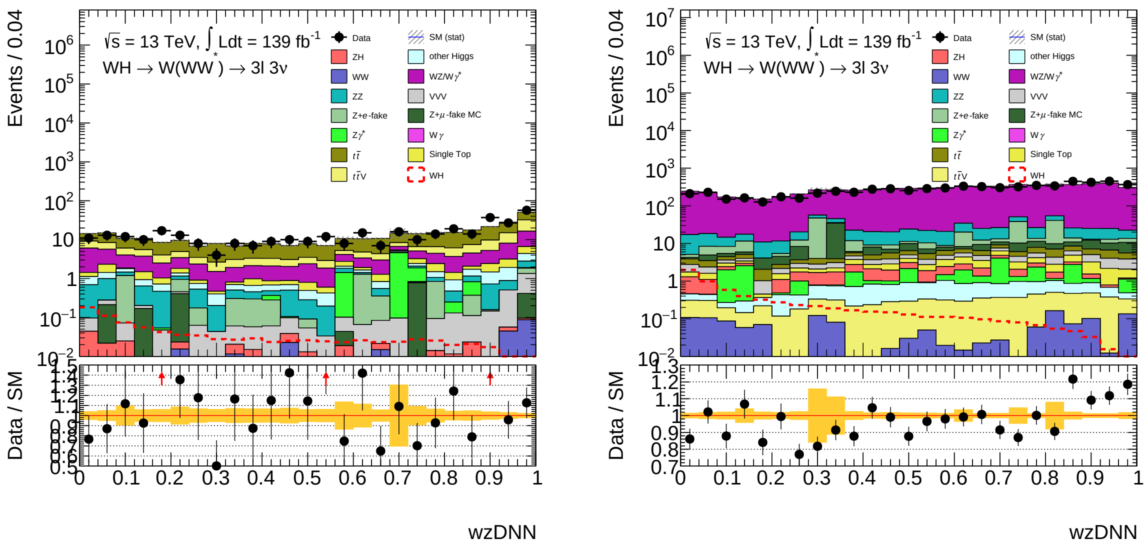


Figure 5.45: ANN_{WZ} output considering the full background contribution in the $t\bar{t}$ control region (left) and the WZ control region (right) for data and MC with statistical uncertainties.

5.6 Towards a Multiclassifier

The previous sections demonstrated that the trained and optimized neural networks can outperform the retrained BDTs in terms of the area under the ROC curve. This section deals with the development of a multiclassifier employed to suppress the most dominant backgrounds $t\bar{t}$ and WZ simultaneously. Its performance is compared with the results obtained from the combination of $\mathrm{BDT}_{t\bar{t}}$ and BDT_{WZ} that have been used in the previous iteration of the analysis. As additional baseline serve the developed binary neural networks $\mathrm{ANN}_{t\bar{t}}$ and ANN_{WZ} which are combined to a multiclassifier as well (from hereon denoted as *unified binary ANN*). As illustrated in appendix B.3, the unification is achieved by replacing the output layer of the binary ANNs with a single concatenation layer that is followed by an output layer with three outputs to account for the multiclass case. The weights are loaded from ANN_{WZ} and $\mathrm{ANN}_{t\bar{t}}$ and only the new last layer is set to be trainable. The same loss function and activation function for the output are used as in the case of the multiclassifier which is discussed next (compare section 5.6.3).

5.6.1 Training and Validation MC Data Sets

The multiclassifier is trained on MC data sets of all three processes WH, WZ and $t\bar{t}$ at once. The same training samples are used as in the previous chapters including 1.001.300 WH SR12 MC events, 78.330 WZ SR3 MC events and 110.987 $t\bar{t}$ SR12 MC events. During training the weights of the three classes are scaled to achieve an equal class balance.

5.6.2 Variable Selection

The simplest approach is to use the set of all variables that are used as inputs for the binary networks ANN_{WZ} and $\text{ANN}_{t\bar{t}}$. Systematically dropping variables from the training routine during the hyperparameter optimization further shows that the longitudinal impact parameter $z_0 \times \sin\theta$ does not improve the performance anymore and is therefore removed from the training. The observation that the multiclassifier needs less inputs than both binary classifiers in total can be explained by the fact that a multiclass representation can exploit inter-correlations from the different classes within the same network. All input variables for the multiclassifier are summarized in Table 5.14.

Multiclassifier input variables

$d_0/\sigma\left(d_0\right)^{\ell_0}$	$d_0/\sigma\left(d_0\right)^{\ell_1}$	$d_0/\sigma\left(d_0\right)^{\ell_2}$	E_T^{miss}
$\Delta R^{\ell_0,\ell_1}$	$\Delta R^{\ell_0,\ell_2}$	$\Delta R^{\ell_1,\ell_2}$	$\varphi\left(E_T^{miss}\right)$
$\Delta\eta^{\ell_0,\ell_1}$	$\Delta\eta^{\ell_0,\ell_2}$	$\Delta\eta^{\ell_1,\ell_2}$	$m_{inv}^{\ell_0,\ell_1,\ell_2}$
$p_T^{\ell_0}$	$p_T^{\ell_1}$	$p_T^{\ell_2}$	$\left\lvert\sum_i p_T^{\ell_i}\right\rvert$
$p_T^{max}(\text{jet})$	n_{jet}	$n_{b-\text{jet}}$	$\Delta\varphi^{\ell_0,E_T^{miss}}$
$m_{inv}^{\ell_0,\ell_1}$	$m_{inv}^{\ell_0,\ell_2}$	$m_{inv}^{\ell_1,\ell_2}$	$\Delta\varphi^{\ell 1,E_T^{miss}}$
$m_T^{\ell_0,\ell_1}$	$m_T^{\ell_0,\ell_2}$	$m_T^{\ell_1,\ell_2}$	$\Delta\varphi^{\ell_2,E_T^{miss}}$
$p^{\ell_0}_{\text{T, varcone, 20}}$	$p^{\ell_0}_{\text{T, varcone, TightTTVA\_pt1000, 30}}$	$p^{\ell_0}_{\text{T, varcone, TightTTVA\_pt500, 30}}$	$p^{\ell_0}_{\text{T, varcone, 40}}$
$p^{\ell_1}_{\text{T, varcone, 20}}$	$p^{\ell_1}_{\text{T, varcone, TightTTVA\_pt1000, 30}}$	$p^{\ell_1}_{\text{T, varcone, TightTTVA\_pt500, 30}}$	$p^{\ell_1}_{\text{T, varcone, 40}}$
$p^{\ell_2}_{\text{T, varcone, 20}}$	$p^{\ell_2}_{\text{T, varcone, TightTTVA\_pt1000, 30}}$	$p^{\ell_2}_{\text{T, varcone, TightTTVA\_pt500, 30}}$	$p^{\ell_2}_{\text{T, varcone, 40}}$
$E^{\text{topo},\ell_0}_{\text{T, cone, 20}}$	$E^{\text{topo},\ell_0}_{\text{T, cone, 40}}$	$E^{\text{topo},\ell_1}_{\text{T, cone, 40}}$	$E^{\text{topo},\ell_1}_{\text{T, cone, 40}}$
$E^{\text{topo},\ell_2}_{\text{T, cone, 20}}$	$E^{\text{topo},\ell_2}_{\text{T, cone, 40}}$	F_{α_1}	$\max\left(\left\lvert 1 - E^{\ell_i}/p^{\ell_i}\right\rvert\right)$

Table 5.14: Summary of input variables for the multiclassifier

5.6.3 Hyperparameter Optimization

The hyperparameter space is scanned in order to find the best performing setup. The crucial parameter is again the network architecture. Its final configuration includes an input layer with 48 input variables that is followed by three layers of 80, 32 and 16 neurons with swish activation function and corresponding dropout rates of 0.37, 0.3 and 0.26, respectively. To account for a multiclass problem, the loss function is changed to the categorical crossentropy

$$C = -\sum_{j=0}^{2} E_j \cdot \log\left(a_j^n\right) \tag{5.3}$$

that runs over all considered classes. E_j is the desired result and a_j^n the activation as introduced in chapter 4.1.1 where j is the index of the component in the output vector whose dimension equals the number of classes in the output layer (n). The activation function of the output layer is changed to the softmax function

$$\text{softmax}(x)_j = \frac{e^{x_j}}{\sum_{k=0}^{2} e^{x_k}} \tag{5.4}$$

that constrains the sum over the activation of all output nodes to be equal to 1. The final set of parameters further includes a batch size of 4096 samples, trained over up to 100 epochs with early stopping applied, 5-fold cross-validation with 80% of the training sample used as train data and 20% of the training sample used as test data. Adam is employed as stochastic optimization method along with a learning rate of 0.001. The input data is preprocessed using the "RobustScaler". All hyperparameter settings are summarized in Table 5.15.

Hyperparameter	Tested [within]	Final Value
Batch Size	[1024, 12.288]	4096
Epochs	[30, 100]	100
k-Folds	5, 10	5
Training Sample Size	80%	80%
Testing Sample Size	20%	20%
Optimizer	Adam	Adam
Learning Rate	[0.01, 0.0001]	0.001
Loss	Categorical Crossentropy	Categorical Crossentropy
Scaler	Robust/MinMax/Standard	RobustScaler
Activation	ReLu/SeLu/Swish	Swish
Regularisation	AlphaDropout/Dropout	Dropout

Table 5.15: Summary of hyperparameters that are tested and the set of final hyperparameters used in the multiclassifier network configuration.

5.6.4 Training Results

Figure 5.46 depicts the mean categorical crossentropy loss and the mean area under the ROC curve with respective standard deviation against the number of training epochs over all 5 folds. The mean test loss in green stays well below the training loss in blue under consideration of the standard deviation per epoch. The most important figure of merit for the multiclassifier is the performance of the separation of WH against all other background as illustrated in the bottom panel of Figure 5.46. The area under the ROC curve with respect to WH against WZ and $t\bar{t}$ is increasing with decreasing loss and reaches a plateau towards epoch > $\sim$ 70 from whereon the early stopping applies.

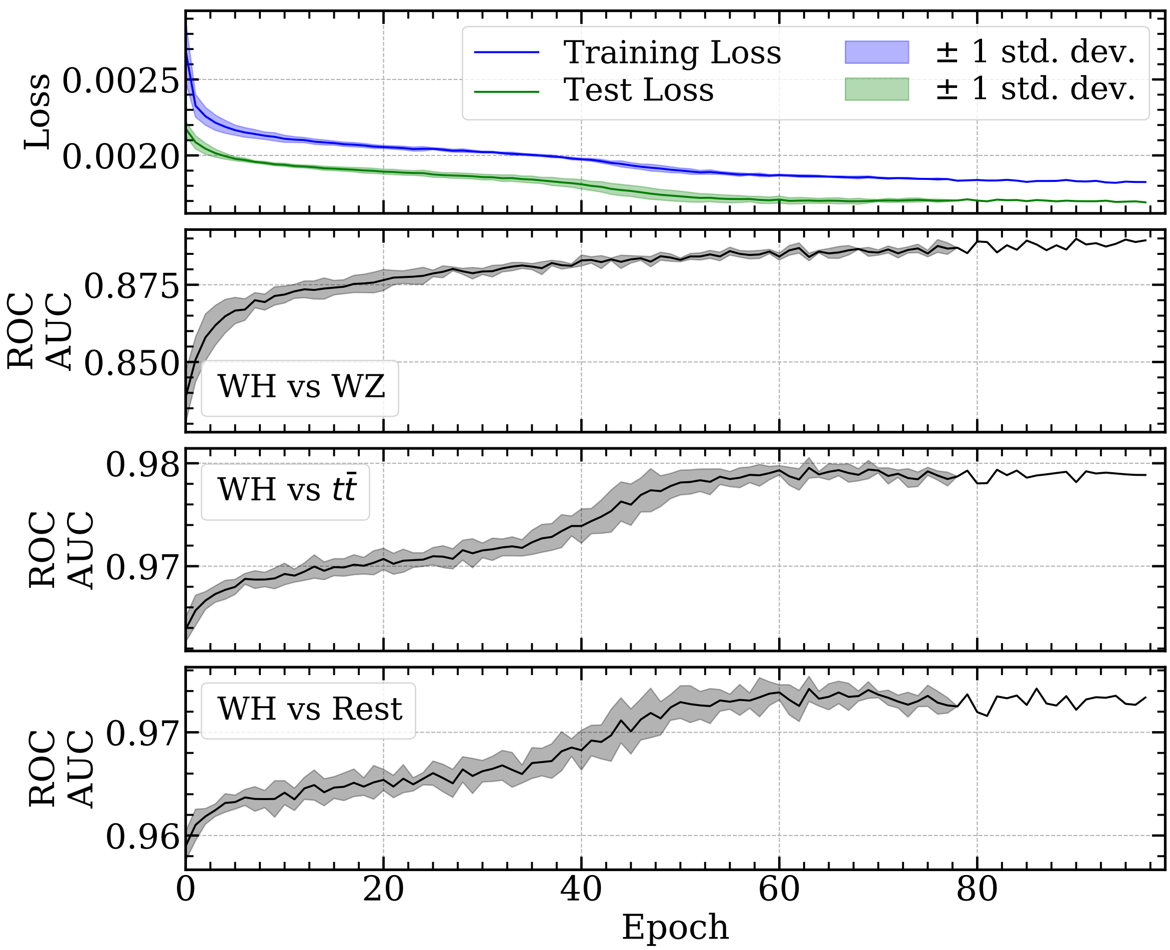


Figure 5.46: Evolution of the mean categorical crossentropy loss with standard deviation for training (blue) and test (green) data fold in the first panel, followed by the mean area under the ROC curve with standard deviation for WH vs $WZ/t\bar{t}$ (second/third panel) and for WH vs both classes in the bottom panel for the test data.

Furthermore, the train to test ratios of the prediction on WH in blue, WZ in red and $t\bar{t}$ in yellow show a good agreement with 1, especially in the most significant signal bin close to 0. Figure 5.47 illustrates this on the example of only fold 1 and 3 for better readability. The remaining folds behave similarly and are given in the appendix B.4.

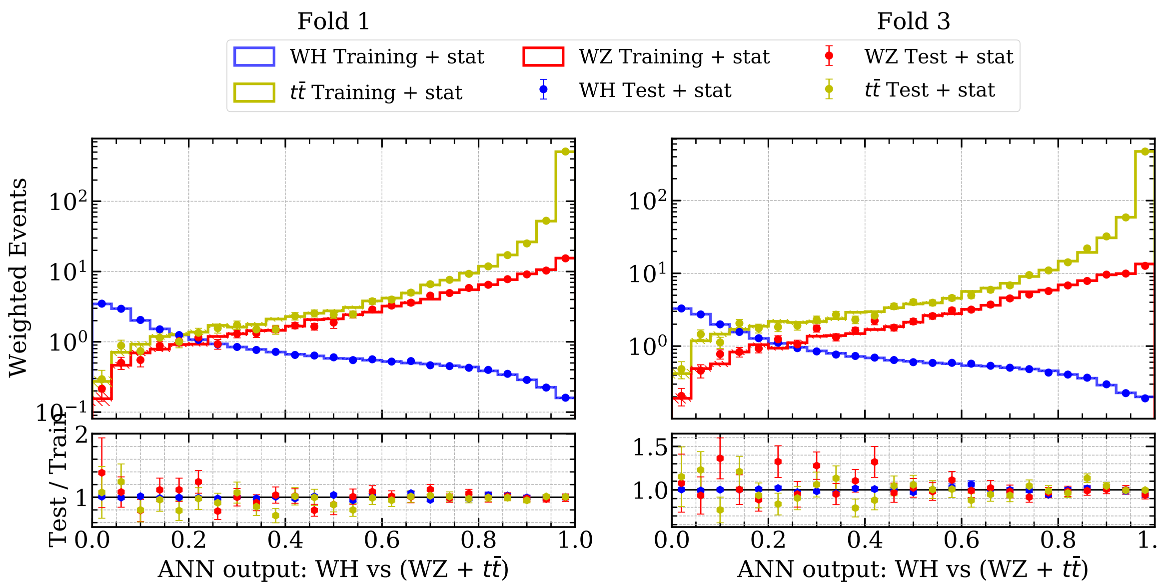


Figure 5.47: The multiclassifier's prediction on train (dots) and test (lines) data folds in the upper plots with the corresponding ratios in the bottom plots on the example of fold 1 and 3.

The area under the ROC curve is studied and the performance is observed to be in good agreement over all five folds. The corresponding plots can be found in appendix B.6. In the following the performance of the multiclassifier is further studied in terms of the significance as the quantity of interest in the final statistical interpretation. The three dimensional output is reparameterized by taking the difference of the WZ node and the $t\bar{t}$ node

$$\Delta \text{ node} \equiv WZ \text{ node} - t\bar{t} \text{ node}$$

without any loss of generality for a simplified visualization and optimization of the analysis. The two dimensional mapping of the multiclassifier prediction on the 139 fb^{-1} nominal MC samples from SR3 that are used in the statistical analysis is visualized in Figure 5.48 considering WH (red) with the most dominant backgrounds WZ (purple) and $t\bar{t}$ (dark gold) as well as the third most dominant background VVV (grey) and the additional top-quark related background contributions from $t\bar{t}W$ and $t\bar{t}Z$ referred to as $t\bar{t}V$ (yellow) and Wt and tZ referred to as single top (sand).

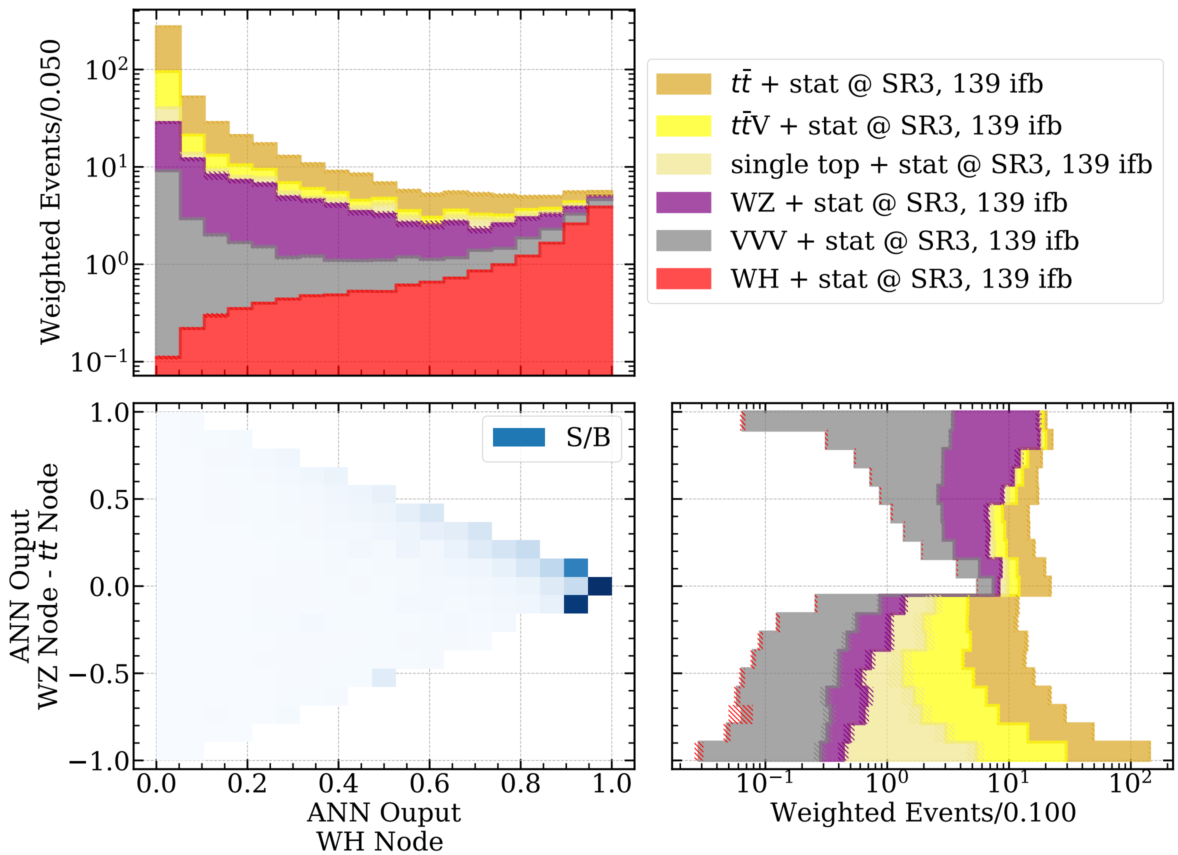


Figure 5.48: Predictions of the multiclassifier on WH and the most dominant background processes. **Upper left:** WH node with signal peaking towards values of 1. **Bottom right:** $\boldsymbol{\Delta}$ node with WZ peaking towards values of +1 and $t\bar{t}$ peaking at values towards -1. **Bottom left:** S/B distribution where only bins with non-zero signal contribution are filled. Bins of dark color indicate high values of S/B.

The upper left plot shows the network output in the WH node. The WH MC peaks towards values of 1 in the WH node while WZ and $t\bar{t}$ are highly suppressed in this region. Despite the fact that the multiclassifier has not been trained on VVV, $t\bar{t}V$ or single top, these processes show a clear tendency towards background-like values of 0, too.

The bottom right plot shows the predictions in the $\boldsymbol{\Delta}$ node output with the same processes as above. Here, WZ peaks at +1, while $t\bar{t}$ peaks at -1. As expected from the construction of the $\boldsymbol{\Delta}$ node, the signal distribution is located at values around 0 and reflects the very good separation from $t\bar{t}$ with a sharp decrease towards negative values. Slightly more signal events are categorized to be WZ-like. Furthermore, the majority of VVV MC events is located at values greater than 0 reflecting the similarity to WZ in the form of well isolated leptons from the boson decays. The corresponding effect is seen for the top-quark related backgrounds which are pushed by the multiclassifier into the $t\bar{t}$ region due to the presence of b-jets and the therefore less isolated leptons. Additionally, the plot indicates that a much cleaner $t\bar{t}$ control region could be constructed with a similar amount of top events but an almost negligible number WZ events compared to the ~ 50 events in the current cut-based $t\bar{t}$ control region. A multiclassifier-based $t\bar{t}$ control region could e.g. be defined within WH node < 0.5 & $\mathbf{\Delta}$ node < -0.1. This could give complementary Z-depleted control region information in the modelling of the observables which could be compared with the current cut-based Z-enriched $t\bar{t}$ control region and they could be combined in the final step of the statistical interpretation. A further modelling check of $W(Z\rightarrow\ \tau\tau)$ seems possible with help of an additional WZ control region within e.g. WH node < 0.5 & $\mathbf{\Delta}$ node > 0.1.

The bottom left plot illustrates the signal to background ratio (S/B) in the two dimensional plane of the output nodes. It shares its x-axis with the WH node plot and its y-axis with the $\mathbf{\Delta}$ node plot. The triangular shape is a residual effect of the constraint introduced by the softmax activation function normalizing the sum over all output nodes to be equal to 1. Bins of darker color indicate higher values of S/B than bins of light color.

The most significant bin with respect to the Poisson significance is extracted in this parameterization for the unified binary classifiers ANN_{WZ} and $\text{ANN}_{t\bar{t}}$ and the multiclassifier for 139 fb^{-1} MC data as summarized in Table 5.16. The multiclassifier outperforms the unified ANN with a Poisson significance of 2.64 to 2.56 due to a better background rejection of 5.29 ± 0.41 to 6.35 ± 0.44 MC events at 7.14 ± 0.03 to 7.47 ± 0.03 signal MC events which translates into S/B values of 1.35 and 1.18, respectively.

$\sqrt{s} = 13$ TeV $\mathcal{L} = 139$ fb^{-1} (Full Run 2)	**Unified Binary ANNs**	**Multiclassifier**
Significance	**2.56**	**2.64**
S/B	**1.18**	**1.35**
S	$\mathbf{7.47 \pm 0.03}$	$\mathbf{7.14 \pm 0.03}$
B	$\mathbf{6.35 \pm 0.44}$	$\mathbf{5.29 \pm 0.41}$
Cut on x-axis	WH node $\geq \mathbf{0.884}$	WH node $\geq \mathbf{0.870}$
Cut on y-axis	$\mathbf{-0.001} \leq \mathbf{\Delta}$ node $< \mathbf{0.200}$	$\mathbf{-0.012} \leq \mathbf{\Delta}$ node $< \mathbf{0.130}$

Table 5.16: Results of the unfied binary ANNs and the multiclassifier in the most significant bin. The significance corresponds to the Poisson signifiance. The x-axis corresponds to the WH node while the y-axis corresponds to the $\mathbf{\Delta}$ node.

A comparison with the BDT combination from the previous analysis can be done in the following ways: The multiclassifier-based most significant signal bin is calculated on MC normalized to 36.1 fb^{-1} or by scaling up the original BDT yields for signal and background by a factor of $^{139}/_{36.1}$ as a rough estimate.

In the case of the normalization to 36.1 fb^{-1}, the multiclassifier improves the significance compared to the BDT from 1.15 to 1.28. A $\sim 10\%$ enhancement of the signal yield from 2.17 ± 0.02 to 2.40 ± 0.02 and a reduction of WZ MC events by 34% from 1.06 ± 0.19 to 0.70 ± 0.19 can be seen in Table 5.17.

With 0.61 ± 0.16 remaining weighted events in the case of the multiclassifier, it achieves a comparable $t\bar{t}$ suppression under consideration of the statistical MC uncertainties with respect to the BDT that recorded 0.54 ± 0.24. It is remarkable that this performance is achieved despite the lack of powerful heavy-flavor variables that have been used in the previous analysis which demonstrates the ability of the network to correctly identify fake leptons with help of more fundamental information like the isolation variables. Furthermore, an improvement in the reduction of $t\bar{t}V$ by more than 40% from 0.25 ± 0.03 down to 0.14 ± 0.03 is observed. In the case of single top, the very low statistical precision does not allow for a meaningful conclusion. While the BDT cut is set to include no events from the latter process, the multiclassifier bin includes 4 raw events (0.26 ± 0.18 in the weighted case) translating into a relative MC statistical uncertainty of 70%. Finally, a better ZZ^* suppression is achieved with 0.05 ± 0.02 MC events accepted by the multiclassifier to 0.21 ± 0.10 MC events accepted by the BDTs.

When scaling the BDT event yields of the previous analysis up to 139 fb^{-1} and applying the multiclassifier to the full Run 2 MC data set, the picture is very similar. The background contribution to the most significant multiclassifier bin with a total of 6.09 ± 0.41 MC events is significantly smaller with respect to 11.17 ± 0.66 MC events in the BDT case yielding in a higher Poisson significance of 2.45 for the multiclassifier and 2.26 for the BDT, respectively. To get a first idea of the effect of systematic uncertainties on the significance, corresponding theory uncertainties from the previous analysis[66] summarized in Table 5.17 are applied as an approximation increasing the performance differences between the multiclassifier and the BDT combination with significances of 2.33 and 2.06, respectively. A further improvement of the significance is expected when taking into account the shape of the network's output distribution for signal and background.

	WH	Other Higgs	$WZ/W\gamma^*$	ZZ^*	VVV	$t\bar{t}$	$t\bar{t}$V	Single Top	Total Bkg.	Sign.
Multiclasifiers most significant Bin (36.1 fb^{-1})	2.40 ± 0.02	0.27 ± 0.03	0.70 ± 0.19	0.05 ± 0.02	0.60 ± 0.05	0.61 ± 0.16	0.14 ± 0.03	0.26 ± 0.18	2.64 ± 0.32	1.28 ± 0.06
$BDT_{WZ} \geq 0.15$ & $BDT_{t\bar{t}} \geq 0.3$ (36.1 fb^{-1})	2.17 ± 0.02	0.24 ± 0.03	1.06 ± 0.19	0.21 ± 0.10	0.60 ± 0.05	0.54 ± 0.24	0.25 ± 0.03	0.00 ± 0.00	2.90 ± 0.33	1.15 ± 0.06
Multiclassifier most significant Bin (139 fb^{-1})	7.14 ± 0.03	0.60 ± 0.05	1.45 ± 0.24	0.11 ± 0.04	1.64 ± 0.08	1.55 ± 0.25	0.38 ± 0.05	0.26 ± 0.18	6.09 ± 0.41	2.45 ± 0.06
$BDT_{WZ} \geq 0.15$ & $BDT_{t\bar{t}} \geq 0.3$ (scaled to 139 fb^{-1})	8.35 ± 0.04	0.92 ± 0.06	4.08 ± 0.40	0.81 ± 0.20	2.23 ± 0.10	2.08 ± 0.47	0.96 ± 0.06	0.00 ± 0.00	11.17 ± 0.66	2.26 ± 0.13
Systematic uncertainty[66]	5%	–	16%	33%	9%	35%	13%	–		
Multiclassifier most significant Bin (139 fb^{-1})										2.33 ± 0.06
$BDT_{WZ} \geq 0.15$ & $BDT_{t\bar{t}} \geq 0.3$ (scaled to 139 fb^{-1})										2.06 ± 0.12

Table 5.17: Event yields with statistical uncertainties and Poisson significance in the most significant bin of the multiclassifier for 36.1 fb^{-1} (first row), the BDT combination of the previous analysis for 36.1 fb^{-1} (second row), the multiclassifier for 139 fb^{-1} (third row) and the BDT combination scaled up to 139 fb^{-1} (forth row). The fifth row indicates the relative systematic uncertainties obtained in the previous analysis. The sixth and seventh row show only the Poisson significance for 139 fb^{-1} under consideration of the systematic uncertainties from the row above.

5.6.5 The Multiclassifier in the Control Regions

The distributions of the multiclassifer output nodes displayed in Figures 5.49, 5.50 and 5.51 show a similar behavior to the output nodes of the binary classifiers in the control regions as defined in Figure 5.4. Due to the low yields in the majority of bins of the output nodes in the $t\bar{t}$ control region the data to MC ratio fluctuates around 1 like in the previous chapters 5.4.5 and 5.5.5. Yet the ratio is in good agreement especially in regions with a higher signal contribution. The output nodes in the WZ control region show a better agreement except for the last bin in the WZ node in Figure 5.50 where a 20% excess of data is visible. Again the Z+jets MC contributes to this bin and might be the reason for the underestimation of MC. A data-driven estimate may improve this.

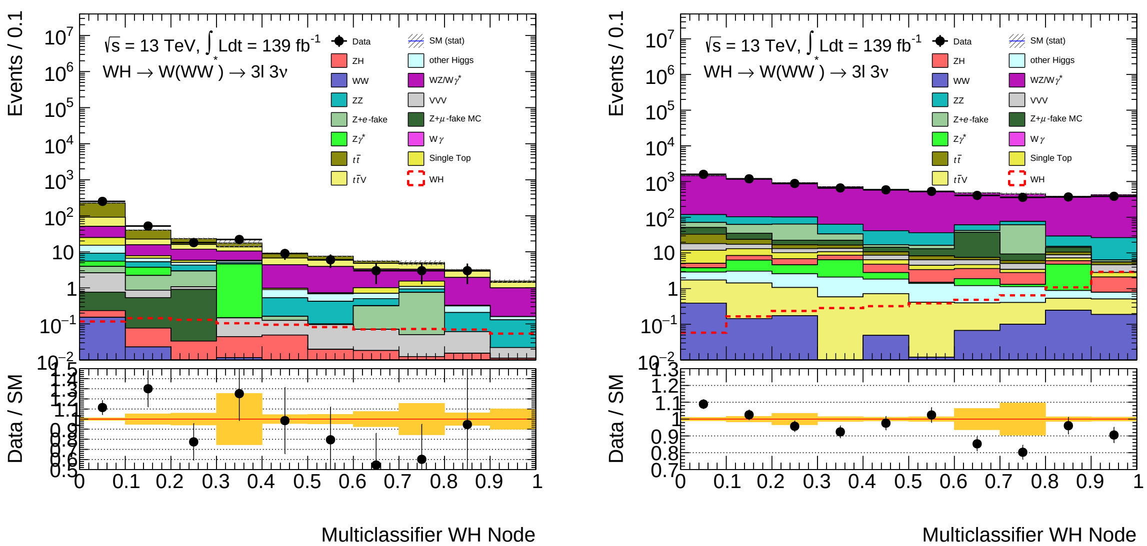


Figure 5.49: Multiclassifier output of the WH node considering all relevant background contributions in the $t\bar{t}$ control region (left) and the WZ control region (right) for data and MC with statistical uncertainties.

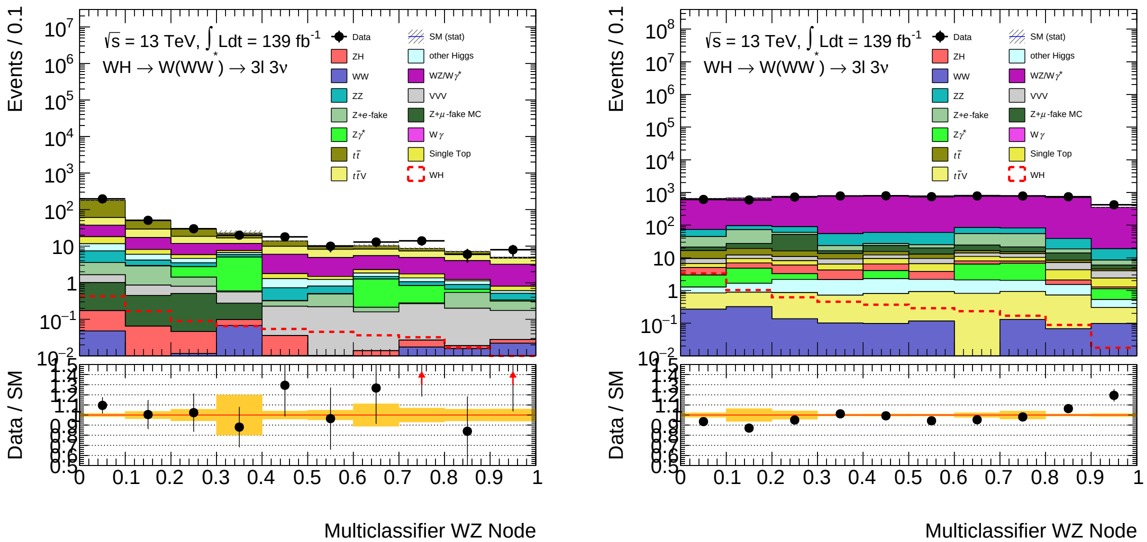


Figure 5.50: Multiclassifier output of the WZ node considering all relevant background contributions in the $t\bar{t}$ control region (left) and the WZ control region (right) for data and MC with statistical uncertainties.

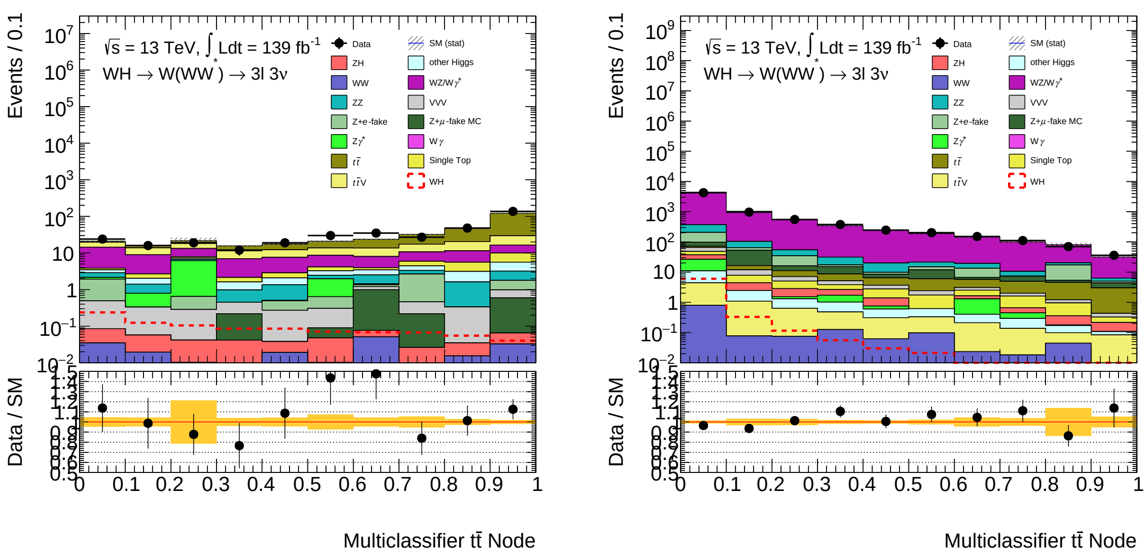


Figure 5.51: Multiclassifier output of the $t\bar{t}$ node considering all relevant background contributions in the $t\bar{t}$ control region (left) and the WZ control region (right) for data and MC with statistical uncertainties.

5.7 Z-depleted Signal Region

The final Z-depleted signal region is defined under consideration of the shape from signal and background in the multiclassifier output using multiple bins. The bin edges of each axis in the two dimensional output plane are chosen such that each bin includes the same amount of signal events. In that way, a flat signal distribution is obtained on the individual axis which grants access to a set of bins with high S/B as illustrated in Figure 5.52. The applied signal flattening yields bins with small background contribution in regions of high signal density which are studied in detail in the upcoming section.

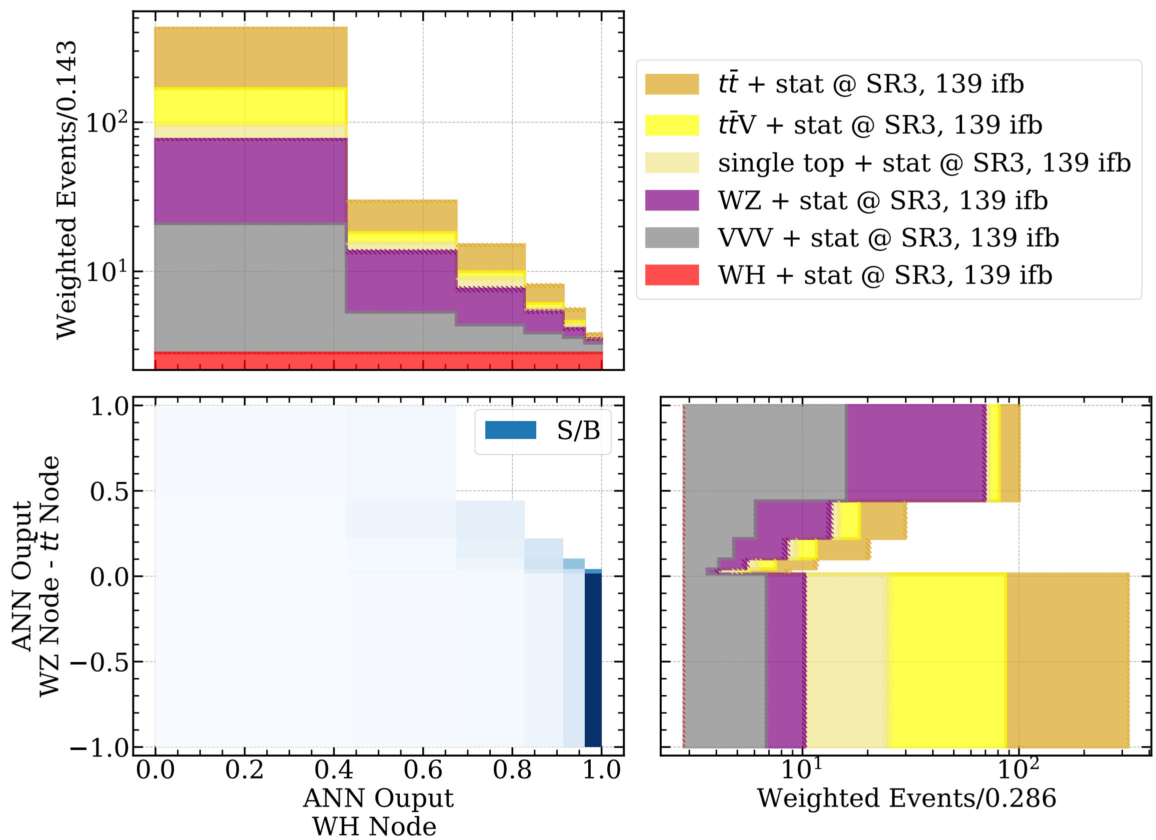


Figure 5.52: Multiclassifier output in the two dimensional plane spanned by the WH node and the Δ node under consideration of the most dominant background processes. The WH distribution is flattened.

An optimization is performed using the most dominant background processes top, WZ and VVV to find the optimal number of bins per axis along with the following requirements: the optimal number of bins in the WH node and the Δ node, which is not necessarily the same, is found when the sum in quadrature of the Poisson significance from the considered bins is maximized with the requirement that every bin includes $t\bar{t}$ and WZ events with a relative MC uncertainty of less then 100% in order to ensure a stable fitting routine in the statistical interpretation step. Any combination of numbers between 2 and 10 bins per axis is tested.

The best number of bins is found to be 6 for both axis. Figure 5.53 shows the final bin selection with Poisson significance and signal to background ratio per bin while the corresponding event yields for WH, top and $WZ+VVV$ are shown in Figure 5.54. Every bin taken into account for the signal region is marked with a red rectangle. Adjacent bins of similar S/B are merged in order to reduce fluctuations and relative statistical MC uncertainties. In this way, 6 bins are considered to define the multiclassifier-based Z-depleted signal region. The bins are summarized in Table 5.18 and the yields are summarized in Table 5.19 at the end of the chapter.

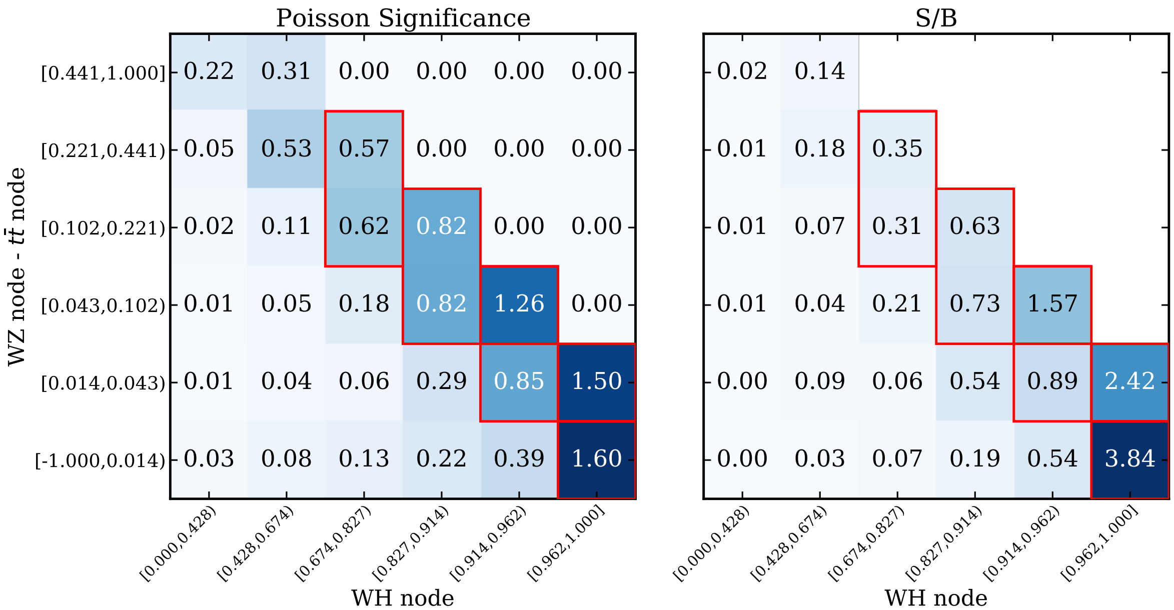


Figure 5.53: Poisson significance (left) and signal to background ratio (right) per bin using the optimized binning. The bins marked with a red rectangle form the Z-depleted signal region.

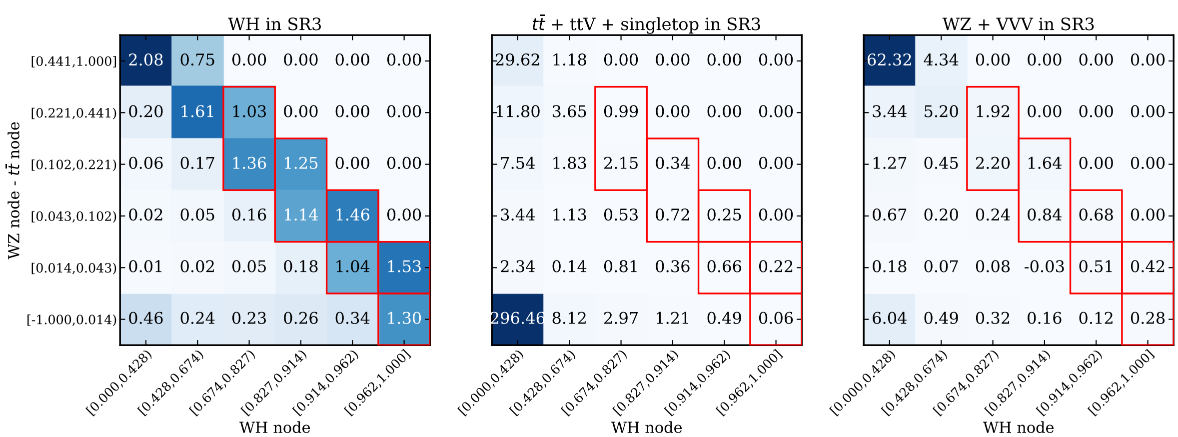


Figure 5.54: Event yields of WH (left), top (middle) and $WZ+VVV$ (right) per bin using the optimized binning. The bins marked with a red rectangle form the Z-depleted signal region.

A systematic study of variables is carried out in order to find phase space regions that still suffer from background excess. In this way, the distribution of the invariant mass of two same-sign electrons $m_{inv}^{e^\pm e^\pm}$ is identified to be a lucrative candidate since WZ is still observed to peak around the rest mass of the Z. As depicted in Figure 5.55, a cut-and-count scan is employed to find the edges of a Z veto window that improves the Poisson significance the most. This window is found to be $m_{inv}^{e^\pm e^\pm} \neq$ [85;95] **GeV**.

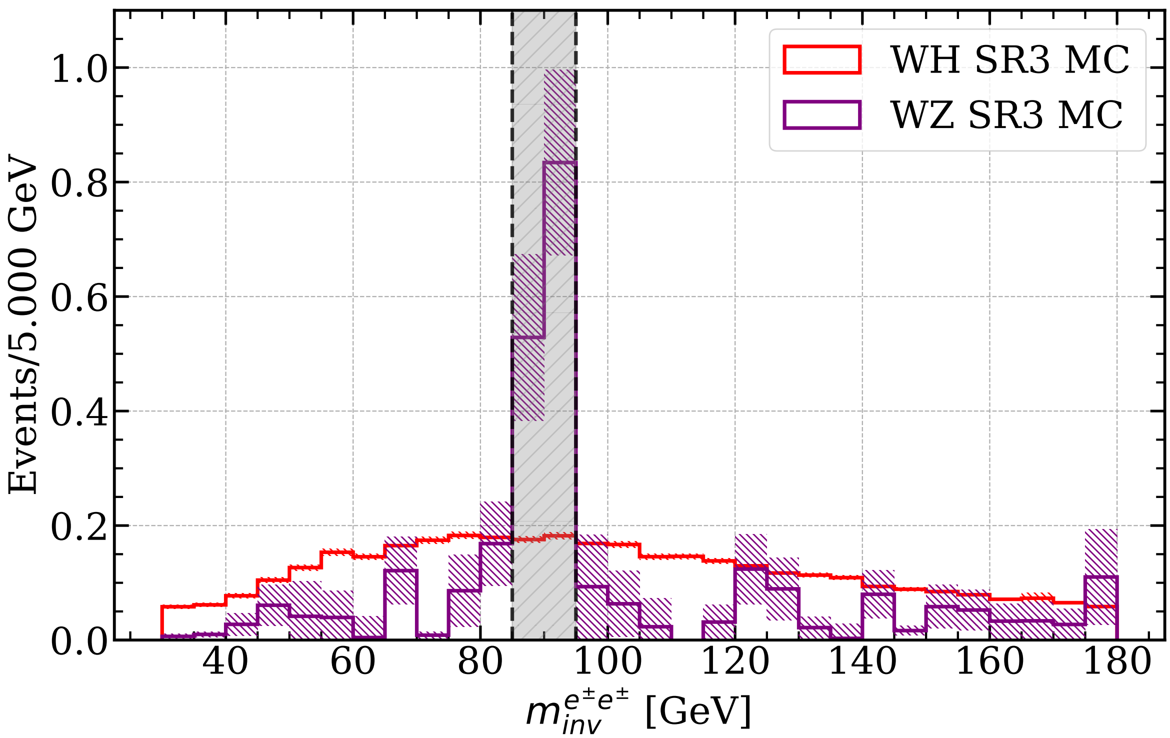


Figure 5.55: Invariant mass distribution of same-sign electron pairs in the Z-depleted signal region of WH (red) and WZ (purple). A cut is applied around to veto events originating from the remaining Z peak.

The final signal region with cut on $m_{inv}^{e^\pm e^\pm}$ is shown in Figure 5.56 together with all relevant background estimations from MC in descending S/B order from left to right. Since the analysis is still blinded, no data is shown. The most significant bin shows a signal to background ratio of ~ 3.5. The suppression of $t\bar{t}$ and WZ is clearly visible. Additionally, the network is able to suppress the background processes $t\bar{t}V$ and single top although it is not explicitly trained on. Furthermore, the contribution of the second VH process, ZH, is well suppressed in the first four bins and shows only small yields in the least significant bins 5 and 6.

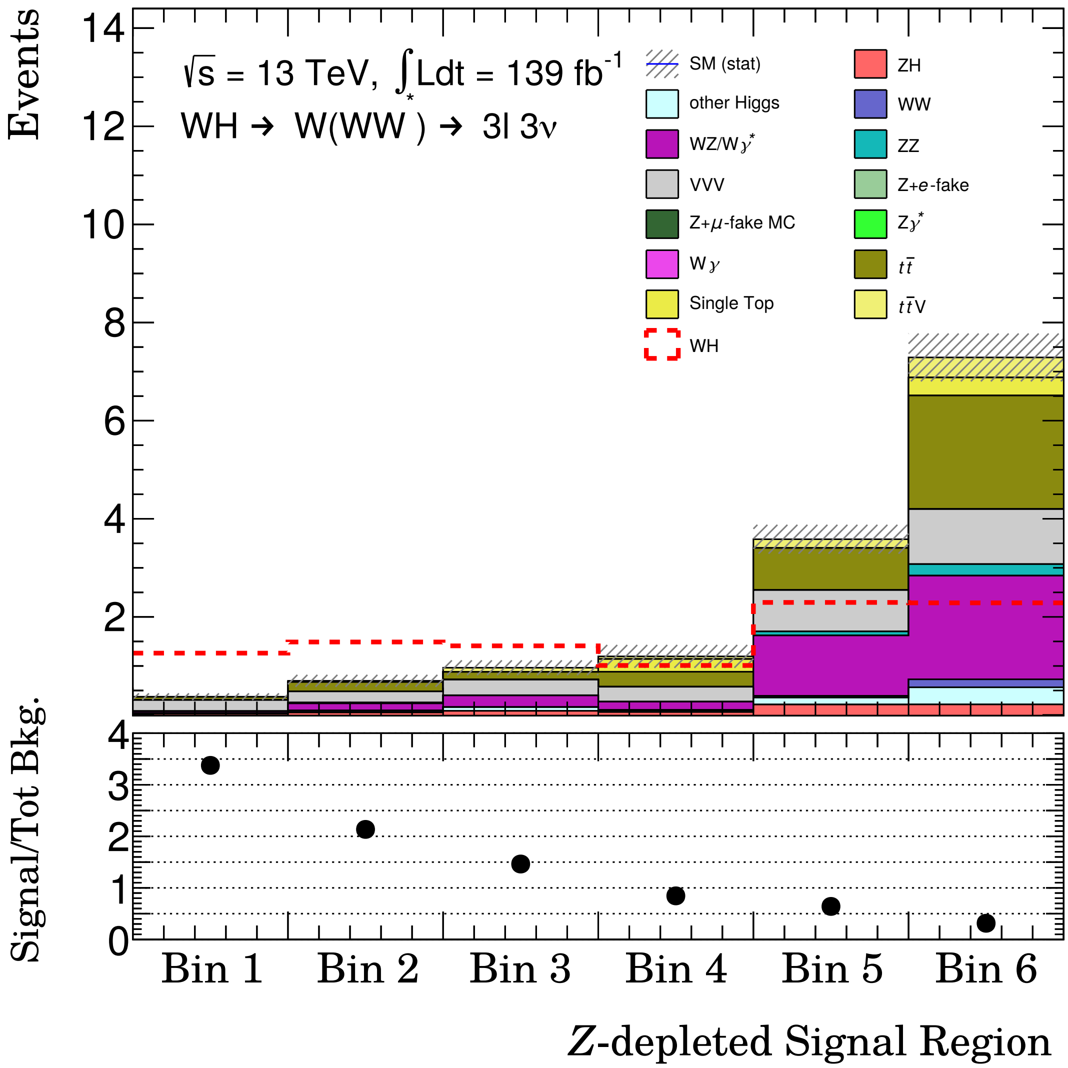


Figure 5.56: The Z-depleted signal region after the cut on $m_{inv}^{e^{\pm}e^{\pm}}$ with the full MC background contribution. **Top:** Event yields in the six signal region bins. The analysis is still blinded, hence no data is shown. **Bottom:** Signal to background ratio per signal region bin.

The largest contribution to the three most significant bins is given by VVV. With over 90% of all VVV it is the WWW process with 2.87 $\pm$ 0.12 events integrated over all 6 signal region bins that has the largest impact from the four considered VVV processes since it is the only one without a Z boson and therefore an irreducible background. A systematic variable study does not show any further improvement by applying additional cuts, since sensitive variables like quantities exploiting the $H \rightarrow WW^* \rightarrow \ell\nu\ell\nu$ spin correlation or the total transverse energy are part or constructed from variables that are part of the training. A training sample with roughly 20.000 WWW MC raw events is derived from SR12 to include the process in the training procedure with an additional node for VVV. However, due to the low number of training events the signal efficiency suffers significantly yielding in a smaller Poisson significance. In a second more subtle approach, the $\boldsymbol{WWW}$ training sample is merged with the WZ training sample such that the network already sees the process during training. Again, the attempt yields a smaller Poisson significance compared to the one reported in Table 5.17 since the trade-off in a decreased WZ rejection to the increase in the rejection of $\boldsymbol{VVV}$ is observed to be not large enough.

Finally, the performance of the multiclassifier in the identification of q (M)ID events as defined in Equation 5.2 in the crucial WZ process is studied. As shown in the upper panel of Figure 5.57 the q ID (green) to q MID (red) ratio of WZ events in the full SR3 is 38.14 : 29.05 = 1.31.

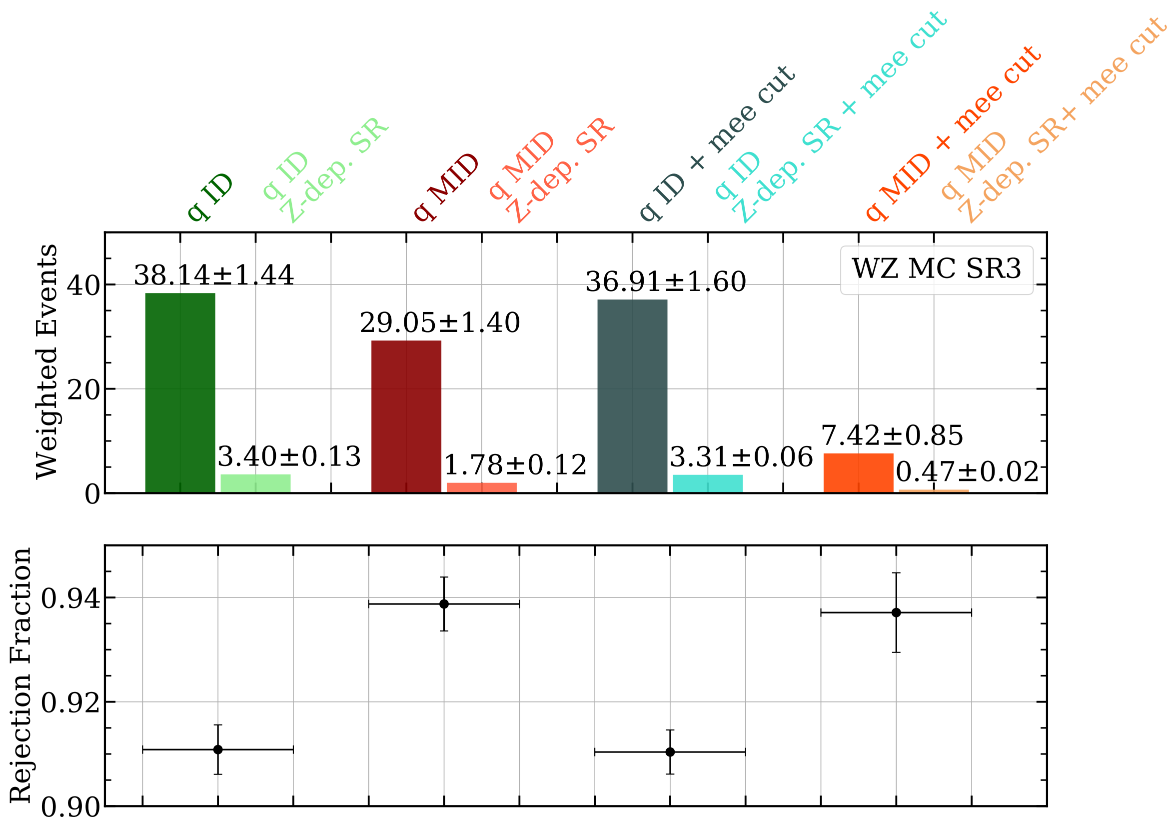


Figure 5.57: Top: Event yields before and after entering the multiclassifier-based Z-depleted signal region for q ID (green), q MID (red), q ID with additional cut on $m_{inv}^{e^\pm e^\pm}$ (turquoise) and q MID with additional cut on $m_{inv}^{e^\pm e^\pm}$ (orange). **Bottom:** Rejection fraction of WZ events from full SR3 (dark colors) compared to the Z-depleted signal region (light colors).

When applying the cut on $m_{inv}^{e^\pm e^\pm}$, the q ID (turquoise) to q MID (orange) ratio increases to 36.91 : 7.42 = 4.97 by effectively removing q MID events from the Z peak. This procedure allows for four different categories that can be compared:
The performance of the network on q ID events, on q MID events, on q ID events with additional cut on $m_{inv}^{e^\pm e^\pm}$ and finally on q MID with cut on $m_{inv}^{e^\pm e^\pm}$ by comparing the numbers from full SR 3 (dark colors) to the Z-depleted signal region (light colors). To do so, a

possible figure of merit is the rejection fraction as the percentage of correctly identified WZ events per category by the neural network when keeping only the 6 most sensitive bins as illustrated in the bottom panel of Figure 5.57. In the Z-depleted signal region this fraction of q ID WZ events amounts to $(91.1 \pm 0.5)\%$ while the network performs even better in classifying q MID WZ events with a rejection fraction of $(93.9 \pm 0.5)\%$. Further applying the cut on $m_{inv}^{e^{\pm}e^{\pm}}$ in the Z-depleted signal region yields a fraction of $(91.4 \pm 0.5)\%$ and of $(93.7 \pm 0.7)\%$ correctly classified q ID and q MID WZ events, respectively.

It can be concluded, that the network learns to effectively suppresses q MID WZ events associated to e.g. charge flip of electrons by almost 94% slightly stronger than q ID WZ events associated to e.g. the $Z \rightarrow \tau\tau$ decay that are suppressed by more than 90% independent from $m_{inv}^{e^{\pm}e^{\pm}}$. These figures of merit could also potentially be used for future analyses on the topic for quantitative benchmark comparisons.

Bin 1	**Bin 2**	**Bin 3**	**Bin 4**	**Bin 5**	**Bin 6**
WH node ≥ 0.962	WH node ≥ 0.962	$0.914 \leq WH$ node < 0.962	$0.914 \leq WH$ node < 0.962	$0.827 \leq WH$ node < 0.914	$0.674 \leq WH$ node < 0.827
Δ node < 0.014	$0.014 \leq \Delta$ node < 0.043	$0.043 \leq \Delta$ node < 0.102	$0.014 \leq \Delta$ node < 0.043	$0.043 \leq \Delta$ node < 0.221	$0.102 \leq \Delta$ node < 0.441

Table 5.18: Summary of the Z-depleted signal region bins.

$\sqrt{s} = 13TeV$ $\mathcal{L} = 139fb^{-1}$ (Full Run 2)	WH	Other Higgs	$WZ/W\gamma^*$	VVV	$t\bar{t}$	$t\bar{t}$V	Single Top	Total Bkg.	Sign.
Multiclassifier Z-dep. SR: Sum of 6 Bins	10.12 ± 0.04	1.34 ± 0.08	5.36 ± 0.44	3.13 ± 0.12	3.98 ± 0.40	0.78 ± 0.07	0.63 ± 0.28	16.02 ± 0.69	2.27 ± 0.04
$m_{inv}^{e^\pm e^\pm} \neq [85; 95]$ GeV	9.75 ± 0.04	1.27 ± 0.07	3.94 ± 0.38	3.04 ± 0.11	3.89 ± 0.39	0.75 ± 0.07	0.63 ± 0.28	14.11 ± 0.65	2.32 ± 0.04
Bin 1	1.26 ± 0.01	0.03 ± 0.01	0.04 ± 0.03	0.23 ± 0.03	0.06 ± 0.05	0.00 ± 0.01	0.00	0.37 ± 0.07	1.50 ± 0.08
Bin 2	1.49 ± 0.01	0.06 ± 0.01	0.14 ± 0.07	0.23 ± 0.03	0.19 ± 0.09	0.02 ± 0.01	0.00	0.70 ± 0.12	1.40 ± 0.08
Bin 3	1.41 ± 0.01	0.16 ± 0.03	0.24 ± 0.12	0.32 ± 0.04	0.16 ± 0.08	0.08 ± 0.02	0.00	0.96 ± 0.15	1.19 ± 0.07
Bin 4	1.01 ± 0.01	0.10 ± 0.02	0.17 ± 0.07	0.31 ± 0.04	0.31 ± 0.12	0.05 ± 0.02	0.26 ± 0.18	1.20 ± 0.23	0.80 ± 0.06
Bin 5	2.30 ± 0.02	0.36 ± 0.04	1.23 ± 0.21	0.84 ± 0.06	0.86 ± 0.18	0.18 ± 0.03	0.00 ± 0.00	3.59 ± 0.29	1.09 ± 0.04
Bin 6	2.29 ± 0.02	0.56 ± 0.05	2.11 ± 0.28	1.12 ± 0.07	2.31 ± 0.30	0.41 ± 0.05	0.37 ± 0.21	7.29 ± 0.49	0.79 ± 0.02

Table 5.19: Total event yields with statistical uncertainties for the multiclassifier-based signal region in the first row, followed by a cut on $m_{inv}^{e^\pm e^\pm}$. Bin 1 to 6 indicate to event yields in the different Z-depleted signal reigon bins. The last row further states the corresponding Poisson significance.

5.8 Statistical Interpretation

The statistical interpretation is carried out in the scope of *hypothesis testing*. In this way the compatibility of a hypothesis H^1 including the Higgs Boson in the $WH \rightarrow WWW^* \rightarrow 3\ell\ 3\nu$ channel is tested against an alternative hypothesis H^0 including SM background only. The section briefly summarizes the maximum likelihood method, while details can be found in e.g. Ref.[80].
For each hypothesis a *likelihood function*

$$L(\mu, \theta) = \prod_{i=1}^{n} f(x_i|\mu, \theta) \tag{5.5}$$

can be constructed where f is a probability density function corresponding to a fixed data point x_i with free parameters μ and θ in the case of H^1, while H^0 corresponds to $\mu = 0$. The free parameters μ and θ are the parameter of interest to be calculated and a set of nuisance parameters, respectively. In this analysis the nuisance parameters so far only take into account statistical uncertainties from MC backgrounds and scale the expected yields mainly constrained from the control region bins. The parameter of interest in this case is defined as the ratio of the observed cross section to the predicted cross section from theory.

The goal of the maximum likelihood method is to estimate the free parameters so that under the assumed hypothesis the actual data is most probable. Thus, the best estimates of the paramater values are those maximizing the likelihood function. In practice one often minimizes the negative log-likelihood for computational reasons. This hypothesis test is carried out using a likelihood ratio test as a test statistic

$$t_\mu = -2 \ln \frac{L\left(\mu, \hat{\hat{\theta}}(\mu)\right)}{L(\hat{\mu}, \hat{\theta})}. \tag{5.6}$$

Here, μ is the parameter of interest, $\hat{\hat{\theta}}(\mu)$ corresponds to the value of θ at μ while $\hat{\mu}$ and $\hat{\theta}$ maximize the likelihood.
Furthermore, the *p-value*

$$p_\mu = \int_{t_{\mu,obs}}^{\infty} \tilde{f}(t_\mu|\mu) dt_\mu \tag{5.7}$$

can be introduced which corresponds to the probability to find a value of t_μ corresponding to a worse agreement with data under the assumption that H^0 is true. Here, $t_{\mu,obs}$ is the value of the test statistic t_μ observed from the data and $\tilde{f}(t_\mu|\mu)$ corresponds to probability density of the test statistic under the assumption of μ. Eventually, the p-value can be converted into a Gaussian variable's significance Z_0 using the corresponding quantile

$$Z_0 = \Phi^{-1}(1 - p_\mu). \tag{5.8}$$

Since the real data is still blinded at the time of writing, pseudo-data based on MC is used that corresponds to the expectation for $\mu = 1$. In this way the expected significance can still be measured which serves as the figure of merit and gives a first estimate on the result to expect from real data.

The inputs to the estimation are the 6 multiclassifer-based signal region bins with subsequent cut on the $m_{inv}^{e^{\pm}e^{\pm}}$ to veto the remaining Z peak and one bin per control region of WZ and $t\bar{t}$. The fit yields an expected significance of 2.93 as shown in Table 5.20 which is close to the commonly used threshold of 3σ that corresponds to an *evidence*. For the full picture, the systematic uncertainties need to be evaluated and implemented into the fit routine which is beyond the scope of this thesis. However, due to the small background contributions, especially in the most significant bins as stated in Table 5.19, the bin depended systematic uncertainties are expected to be small as well. In this way it is likely that the analysis is still statistically limited and thus a significance that is derived on statistical uncertainties only yields already a good first estimation of the final results. A summary of the fit results is given in Table 5.20.

Parameter	**Expected**
Z_0	**2.930**
p_0	**0.002**
μ	$\mathbf{1.000^{+0.495}_{-0.417}}$

Table 5.20: Fit Results obtained from the Z-depleted signal region using 6 signal region bins and one bin per control region.

6 | Summary and Outlook

This thesis deals with the analysis of the Higgs boson production in association with a W boson and the decay of the Higgs boson into two W bosons with a fully leptonic final state in the Z-depleted signal region as most sensitive region to the $W(H\to WW^*)$ process. The data set used in the analysis corresponds to events from proton-proton collision at a center of mass energy of 13 TeV over a time-integrated luminosity of 139 fb^{-1} collected at the ATLAS experiment.

For the first time in this decay channel at ATLAS a neural network is employed to suppress the background processes. In a first step two binary neural networks are developed, trained and optimized to separate WH events from WZ and $t\bar{t}$ events individually. The knowledge gained from these extensive studies is used to build a multiclass neural network capable of predicting on the three classes simultaneously. A most powerful set of variables is identified partially including variables from the previous analysis as well as new variables such as an E/p related quantity giving information on charge flip events. The multiclassifier outperforms the unified binary neural networks as well as a BDT combination developed in a previous analysis with respect to the Poisson significance.

With a comparable rejection performance on $t\bar{t}$ despite the lack of powerful variables from the previous analysis, an improvement in $t\bar{t}V$ rejection by more than 40% and an improvement in the reduction of the crucial background WZ by 34% with respect to the BDTs, the analysis clearly demonstrates the power and advantage of a multiclass neural network.

Another advantage is the possible construction of complementary control regions with help of the multiclassifier output. These regions could be consulted for additional modelling checks for variables of interest in regions that are orthogonal to the cut-based control regions.

A dedicated multiclassifier-based Z-depleted signal region is developed from which an expected significance on the full Run 2 MC data set of

$$Z_0 = 2.93$$

is determined which is close to the evidence threshold of 3σ.

With an additional dedicated training sample and training on VVV as a remaining background in the most significant signal region bins, this goal would be in reach. As the multiclassifier outperforms the BDT combination from the previous iteration, the new classifier will serve as foundation of the next ATLAS publication on the measurement of the VH production in the WH trilepton decay channel. Before unblinding the real data, next steps are to determine systematic uncertainties and to establish a data-driven fake estimate. In the upcoming years the high luminosity LHC is going to be built.[81] With a predicted time-integrated luminosity of 3000 fb^{-1} the significance is expected to be even higher and a statistical limitation of the results is expected to become less relevant.

A further prospect is to carry out a simplified template cross section (STXS) measurement[82] that aims to introduce a granularity within a single production mode by splitting up into kinematic regions. Matching the STXS granularity with experimentally observed sensitive regions in a staged approach could grant access to dedicated beyond Standard Model bins sensitive to new physics. A multiclassifer as developed in this thesis could help to purify these STXS bins.

Using machine learning algorithms like (deep) neural networks comes with the difficulty of understanding the reason behind their predictions. Therefore it seems promising to study the network output as a function of its input to check to plausibility of a network's decision or to get insights that help to improve the performance. Possible approaches are e.g. to develop local models around the decision threshold that is set to define regions of interest in order to better understand the decision of a network or to cluster events in order to find out a network's reason behind the specific prediction.[83] At the time of writing, studies based on the classifiers developed in thesis are carried out to investigate on this promising topic.

Acknowledgements

Not provided here for privacy reasons.

Appendices

A | Deep Neural Networks

A.1 Backpropagation in a Nuthsell

Derivation of $\frac{\partial C}{\partial w^l_{ij}}$ and $\frac{\partial C}{\partial b^l_i}$ for $\Delta w^{(l)}_{ij}$ and $\Delta b^{(l)}_i$, respectively.

We can apply the chain rule:

$$\frac{\partial C}{\partial w^{(l)}_{ij}} = \frac{\partial C}{\partial a^{(l)}_i}\frac{\partial a^{(l)}_i}{\partial z^{(l)}_i}\frac{\partial z^{(l)}_i}{\partial w^{(l)}_{ij}} \tag{A.1}$$

$$\frac{\partial C}{\partial b^{(l)}_i} = \frac{\partial C}{\partial a^{(l)}_i}\frac{\partial a^{(l)}_i}{\partial z^{(l)}_i}\frac{\partial z^{(l)}_i}{\partial b^{(l)}_i} \tag{A.2}$$

We further define the output error:

$$\delta^{(l)}_i = \frac{\partial C}{\partial a^{(l)}_i}\frac{\partial a^{(l)}_i}{\partial z^{(l)}_i} \tag{A.3}$$

For the weights and biases in the output layer inserting the definitions in equations 4.1 and 4.2 yields:

$$\delta^{(n)}_i = \left[\frac{1-E_i}{1-a^{(n)}_i} - \frac{E_i}{a^{(n)}_i}\right]\cdot\sigma'_a\left(z^{(n)}_i\right) \tag{A.4}$$

$$\frac{\partial C}{\partial w^{(n)}_{ij}} = \delta^{(n)}_i \cdot a^{(n-1)}_j \tag{A.5}$$

$$\frac{\partial C}{\partial b^{(n)}_i} = \delta^{(n)}_i \tag{A.6}$$

To find the derivative of the cost function with respect to a weight or bias in a layer that is not the output layer it is necessary to find the expression:

$$\delta^{(l-1)}_i = \frac{\partial C}{\partial a^{(l-1)}_i}\frac{\partial a^{(l-1)}_i}{\partial z^{(l-1)}_i} = \left(\sum_j \frac{\partial C}{\partial a^{(l)}_j}\frac{\partial a^{(l)}_j}{\partial z^{(l)}_j}\frac{\partial z^{(l)}_j}{\partial a^{(l-1)}_i}\right)\frac{\partial a^{(l-1)}_i}{\partial z^{(l-1)}_i} = \sum_j \delta^{(l)}_j w^{(l)}_{ij}\cdot\sigma'_a\left(z^{(l-1)}_i\right). \tag{A.7}$$

Consequently $\delta^{(l-1)}_i$ can be calculated from the $\delta^{(l)}_j$ in the preceding layer starting from the last layer ($l = n$). This process is called back propagation.

B | Backup Plots

B.1 ANN$_{t\bar{t}}$

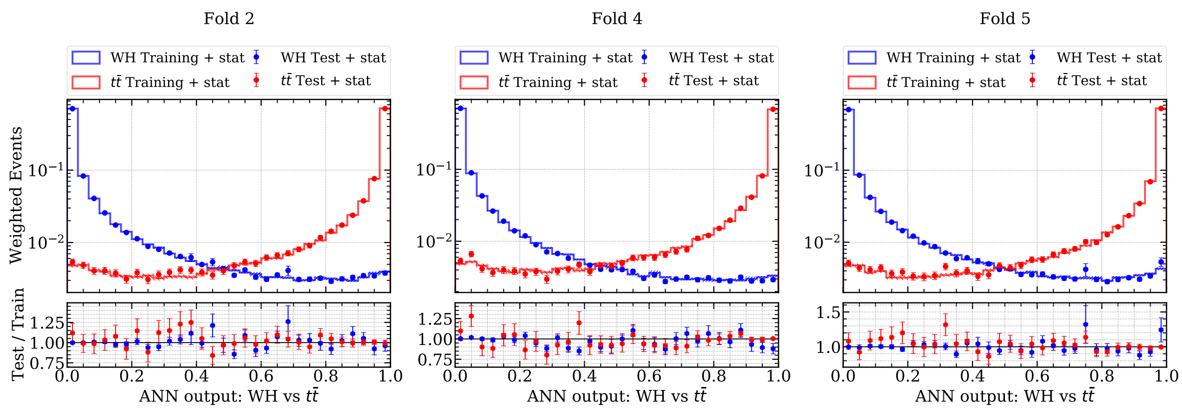


Figure B.1: The ANN$_{t\bar{t}}$'s prediction on train (dots) and test (lines) data folds in the upper plots with the corresponding ratios in the bottom plots on the example of fold 2, 4 and 5.

B.2 ANN$_{WZ}$

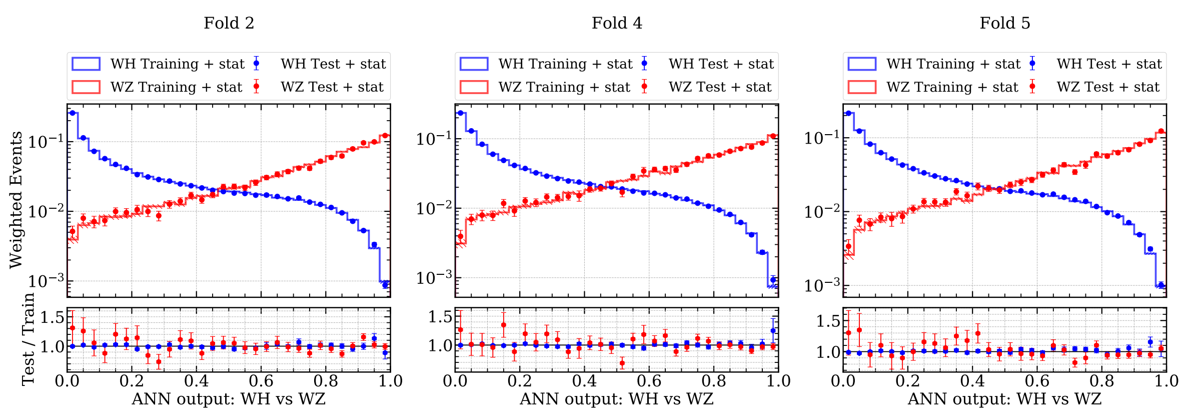
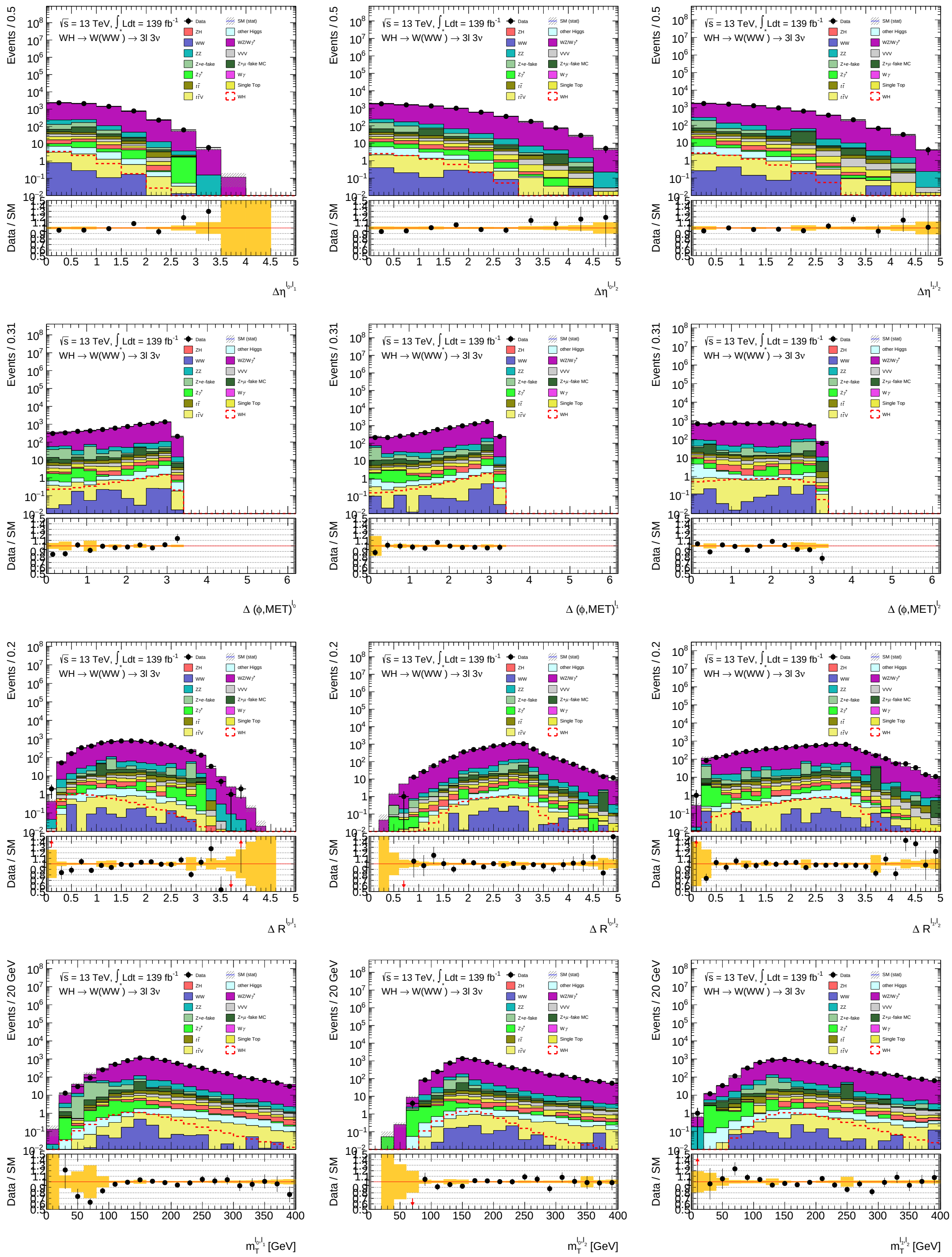
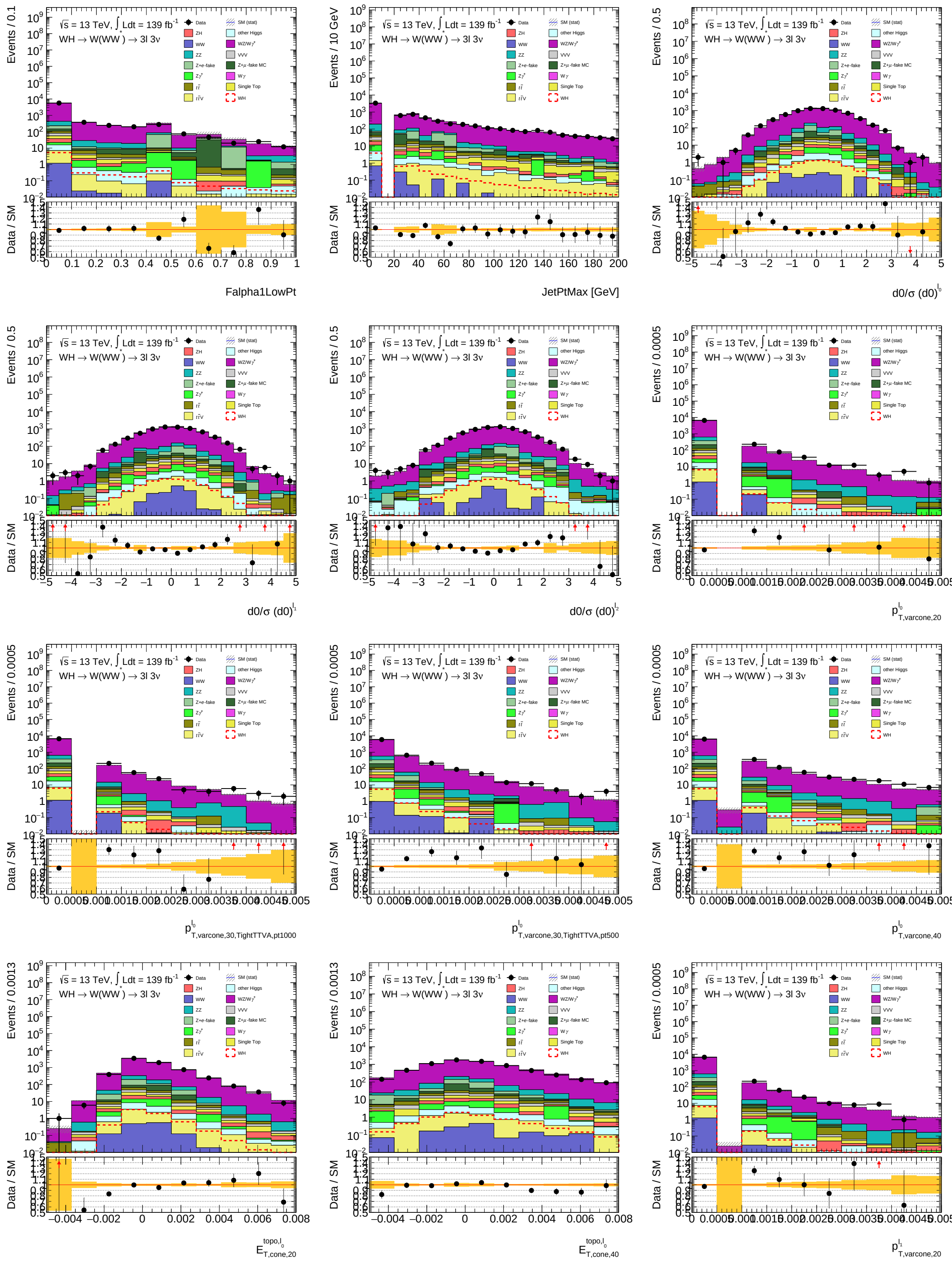
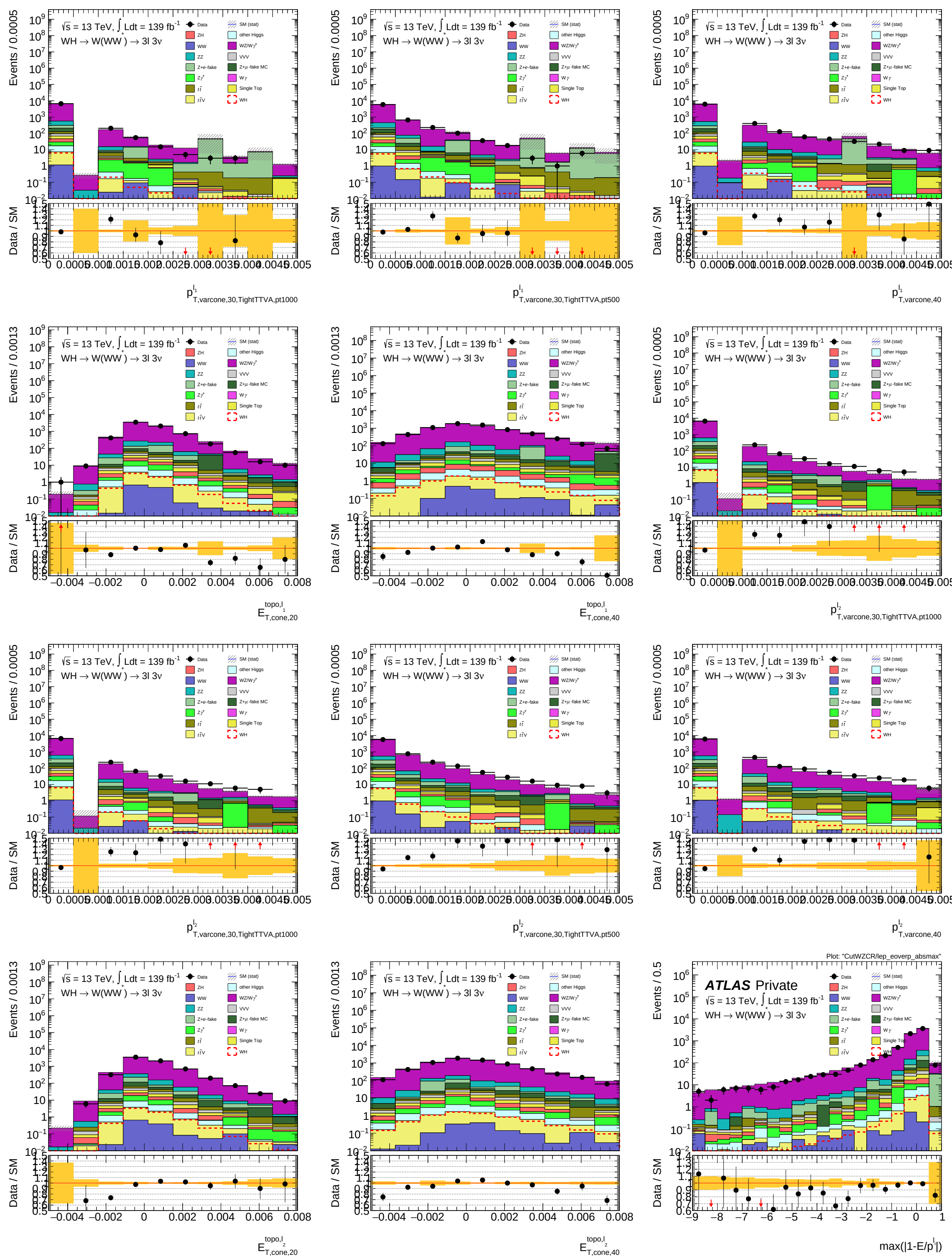


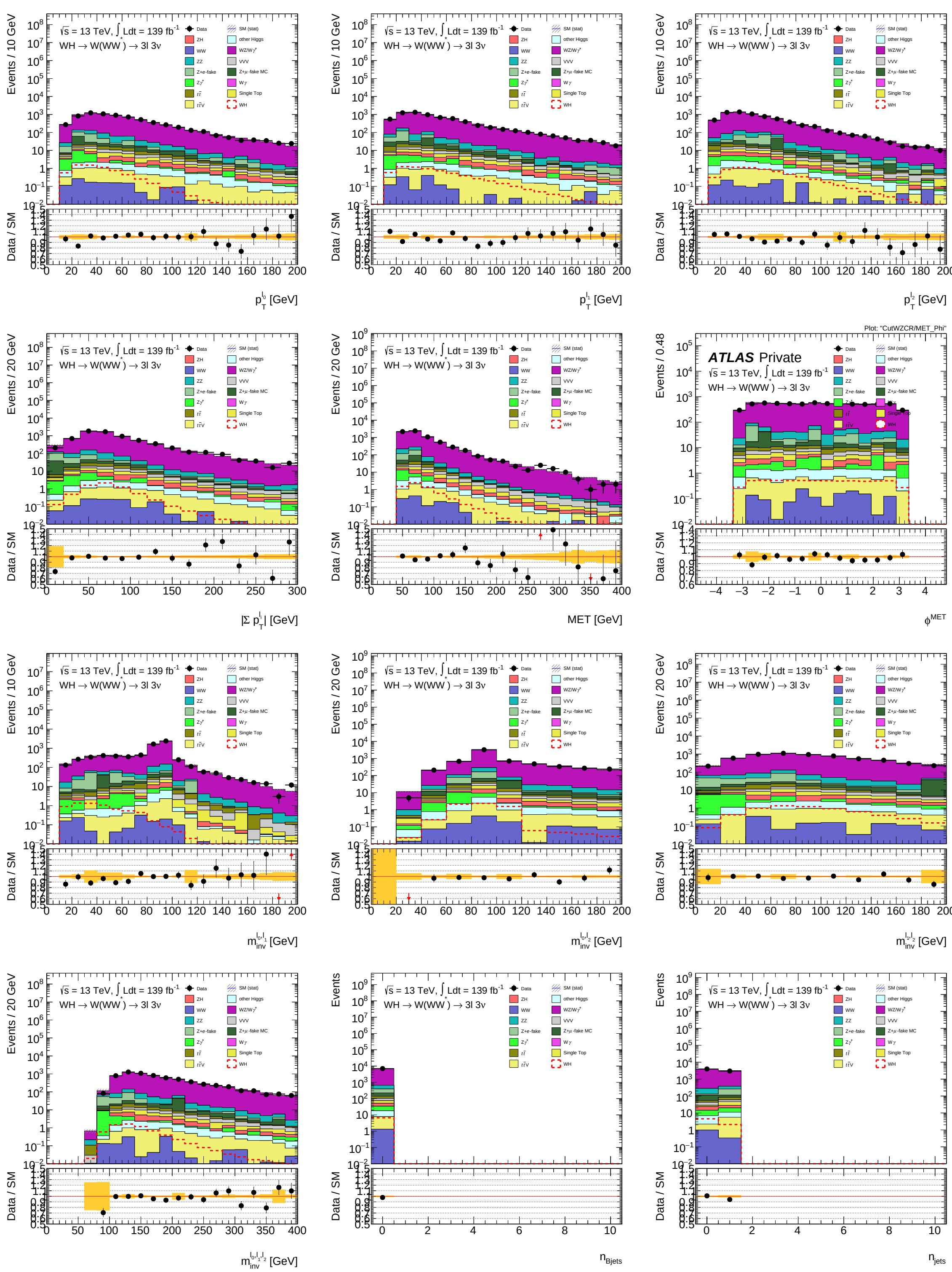
Figure B.2: The ANN$_{WZ}$'s prediction on train (dots) and test (lines) data folds in the upper plots with the corresponding ratios in the bottom plots on the example of fold 2 , 4 and 5.

B.3 *WZ* Control Region: Input Variables



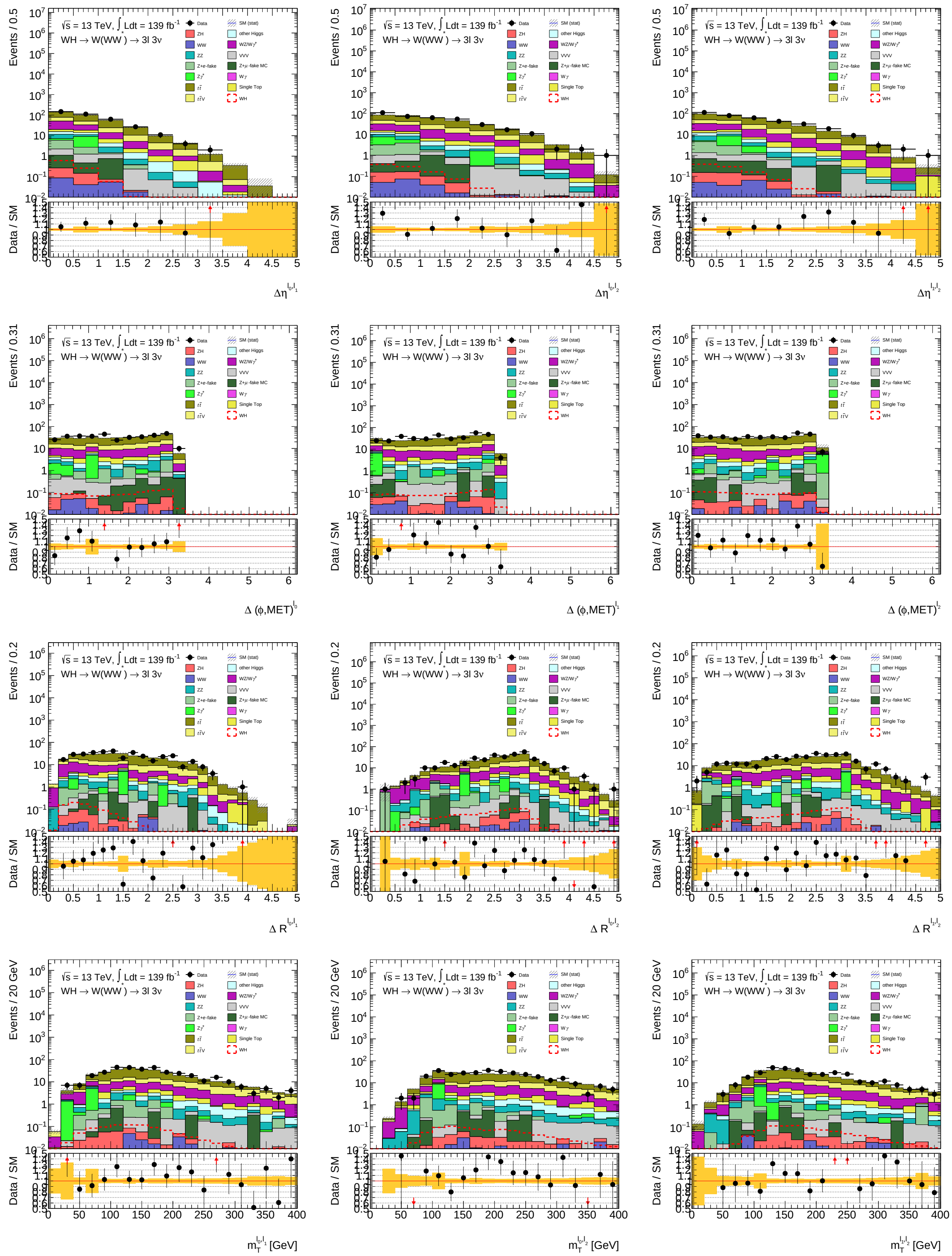

√s = 13 TeV, ∫ Ldt = 139 fb⁻¹
WH → W(WW) → 3l 3ν
Data
SM (stat)
ZH
other Higgs
WW
WZ/Wγ*
ZZ
VVV
Z+e-fake
Z+μ-fake MC
Zγ*
Wγ
tt̄
Single Top
tt̄V
WH
Data / SM
Falpha1LowPt
JetPtMax [GeV]
Events / 0.1
Events / 10 GeV
Events / 0.5
Events / 0.0005
Events / 0.0013

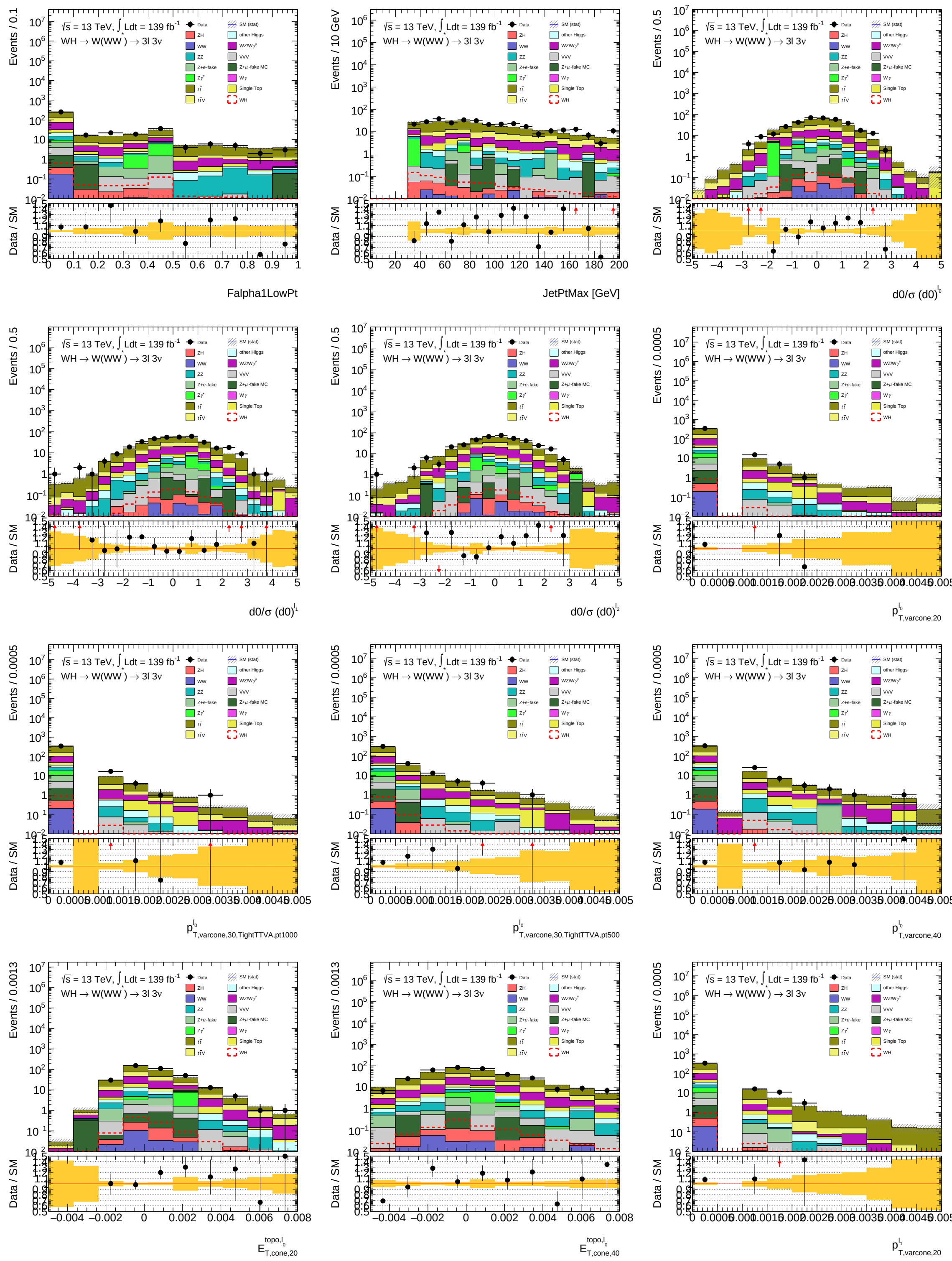

Plot: "CutWZCR/lep_eoverp_absmax"
ATLAS Private
√s = 13 TeV, ∫Ldt = 139 fb-1
WH → W(WW*) → 3l 3ν
Data
SM (stat)
ZH
other Higgs
WW
WZ/Wγ*
ZZ
VVV
Z+e-fake
Z+μ-fake MC
Zγ*
Wγ
tt̄
Single Top
tt̄V
WH
Data / SM
Events / 0.0005
Events / 0.0013
Events / 0.5
max(|1-E/p|)

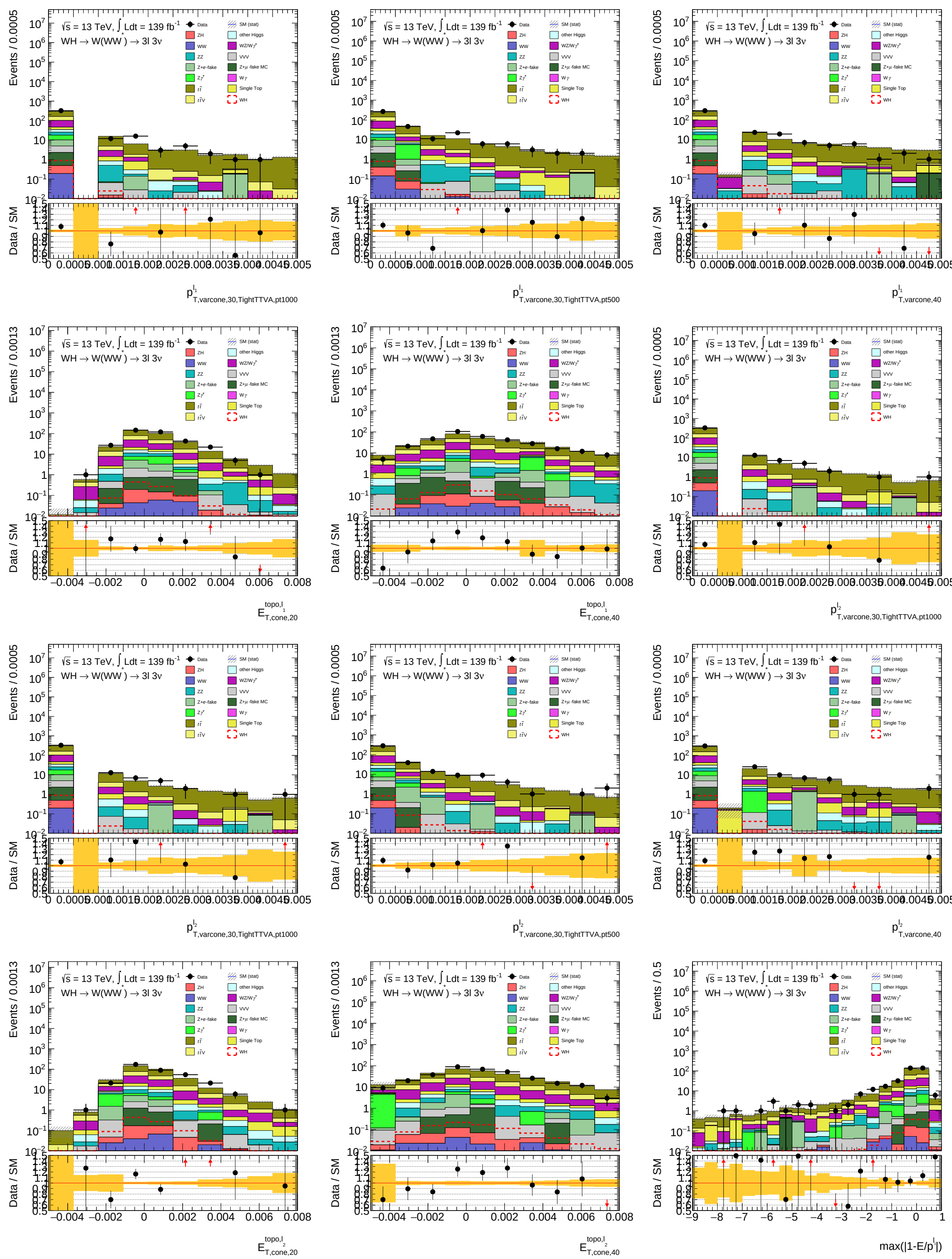


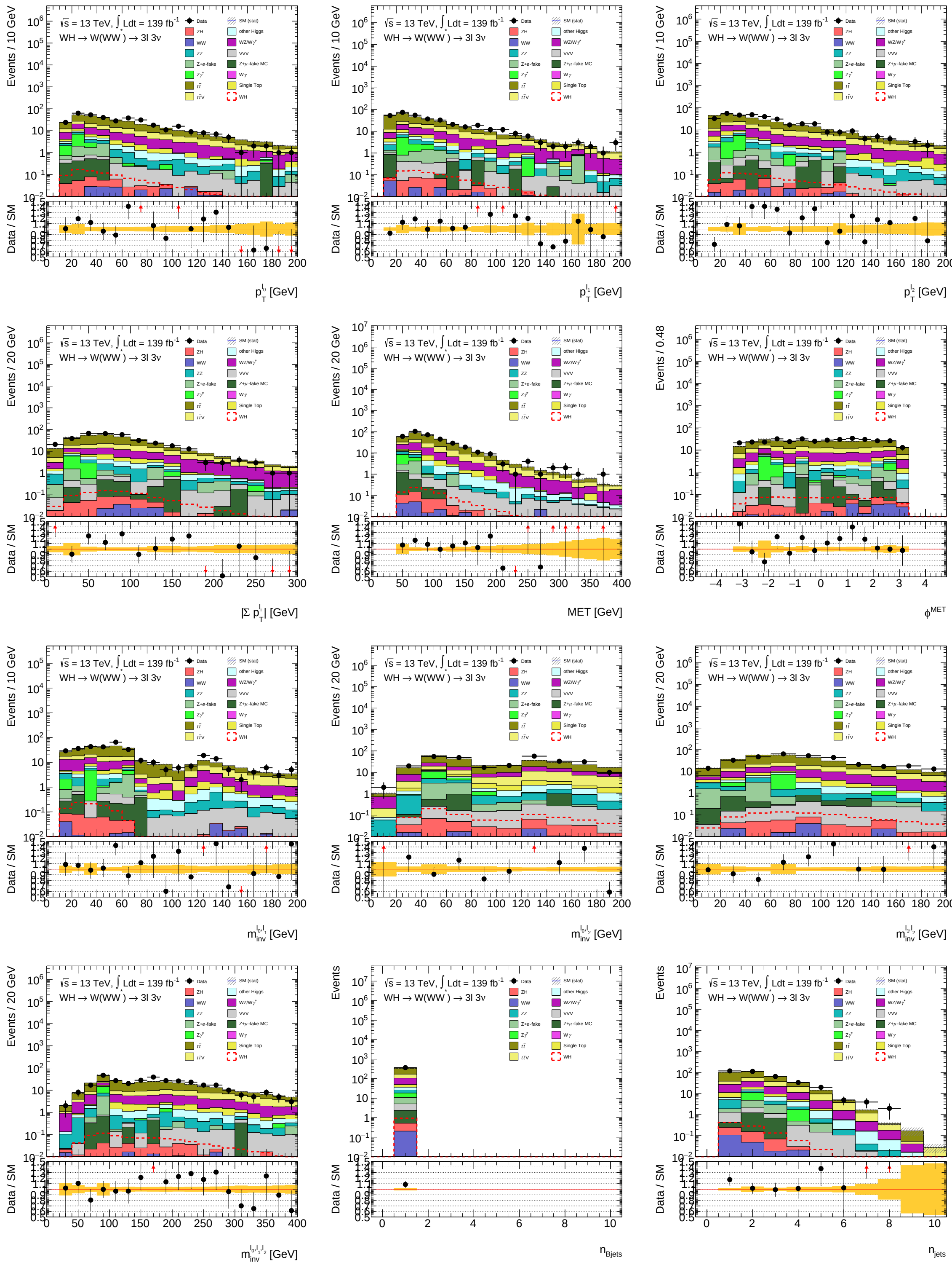
√s = 13 TeV, ∫ Ldt = 139 fb⁻¹
WH → W(WW*) → 3l 3ν
ATLAS Private
Plot: "CutWZCR/MET_Phi"
Data
SM (stat)
ZH
other Higgs
WW
WZ/Wγ*
ZZ
VVV
Z+e-fake
Z+μ-fake MC
Zγ*
Wγ
tt̄
Single Top
tt̄V
WH
Events / 10 GeV
Events / 20 GeV
Events / 0.48
Events
Data / SM
p_T^{l_0} [GeV]
p_T^{l_1} [GeV]
p_T^{l_2} [GeV]
|Σ p_T^l| [GeV]
MET [GeV]
φ^MET
m_inv^{l_0 l_1} [GeV]
m_inv^{l_0 l_2} [GeV]
m_inv^{l_1 l_2} [GeV]
m_inv^{l_0 l_1 l_2} [GeV]
n_Bjets
n_jets

B.4 $t\bar{t}$ Control Region: Input Variables




√s = 13 TeV, ∫ Ldt = 139 fb⁻¹
WH → W(WW*) → 3l 3ν
Data
SM (stat)
ZH
other Higgs
WW
WZ/Wγ*
ZZ
VVV
Z+e-fake
Z+μ-fake MC
Zγ*
Wγ
tt̄
Single Top
tt̄V
WH
Data / SM
Falpha1LowPt
JetPtMax [GeV]
d0/σ (d0)
p_T,varcone,20
p_T,varcone,30,TightTTVA,pt1000
p_T,varcone,30,TightTTVA,pt500
p_T,varcone,40
E_T,cone,20 topo
E_T,cone,40 topo


√s = 13 TeV, ∫ Ldt = 139 fb^{-1}
WH → W(WW) → 3l 3ν
Data
SM (stat)
ZH
other Higgs
WW
WZ/Wγ*
ZZ
VVV
Z+e-fake
Z+μ-fake MC
Zγ*
Wγ
$t\bar{t}$
Single Top
$t\bar{t}V$
WH
Events / 0.0005
Events / 0.0013
Events / 0.5
Data / SM
$p_{T,varcone,30,TightTTVA,pt1000}^{l_1}$
$p_{T,varcone,30,TightTTVA,pt500}^{l_1}$
$p_{T,varcone,40}^{l_1}$
$E_{T,cone,20}^{topo,l_1}$
$E_{T,cone,40}^{topo,l_1}$
$p_{T,varcone,30,TightTTVA,pt1000}^{l_2}$
$p_{T,varcone,30,TightTTVA,pt500}^{l_2}$
$p_{T,varcone,40}^{l_2}$
$E_{T,cone,20}^{topo,l_2}$
$E_{T,cone,40}^{topo,l_2}$
max(|1-E/p^{l}|)



√s = 13 TeV, ∫Ldt = 139 fb⁻¹
WH → W(WW*) → 3l 3ν
Data
SM (stat)
ZH
other Higgs
WW
WZ/Wγ*
ZZ
VVV
Z+e-fake
Z+μ-fake MC
Zγ*
Wγ
tt̄
Single Top
tt̄V
WH
Events / 10 GeV
Events / 20 GeV
Events / 0.48
Events
Data / SM
p_T^{l0} [GeV]
p_T^{l1} [GeV]
p_T^{l2} [GeV]
|Σ p_T^l| [GeV]
MET [GeV]
φ^MET
m_inv^{l0l1} [GeV]
m_inv^{l0l2} [GeV]
m_inv^{l1l2} [GeV]
m_inv^{l0l1l2} [GeV]
n_Bjets
n_jets

B.5 Unified Binary ANN

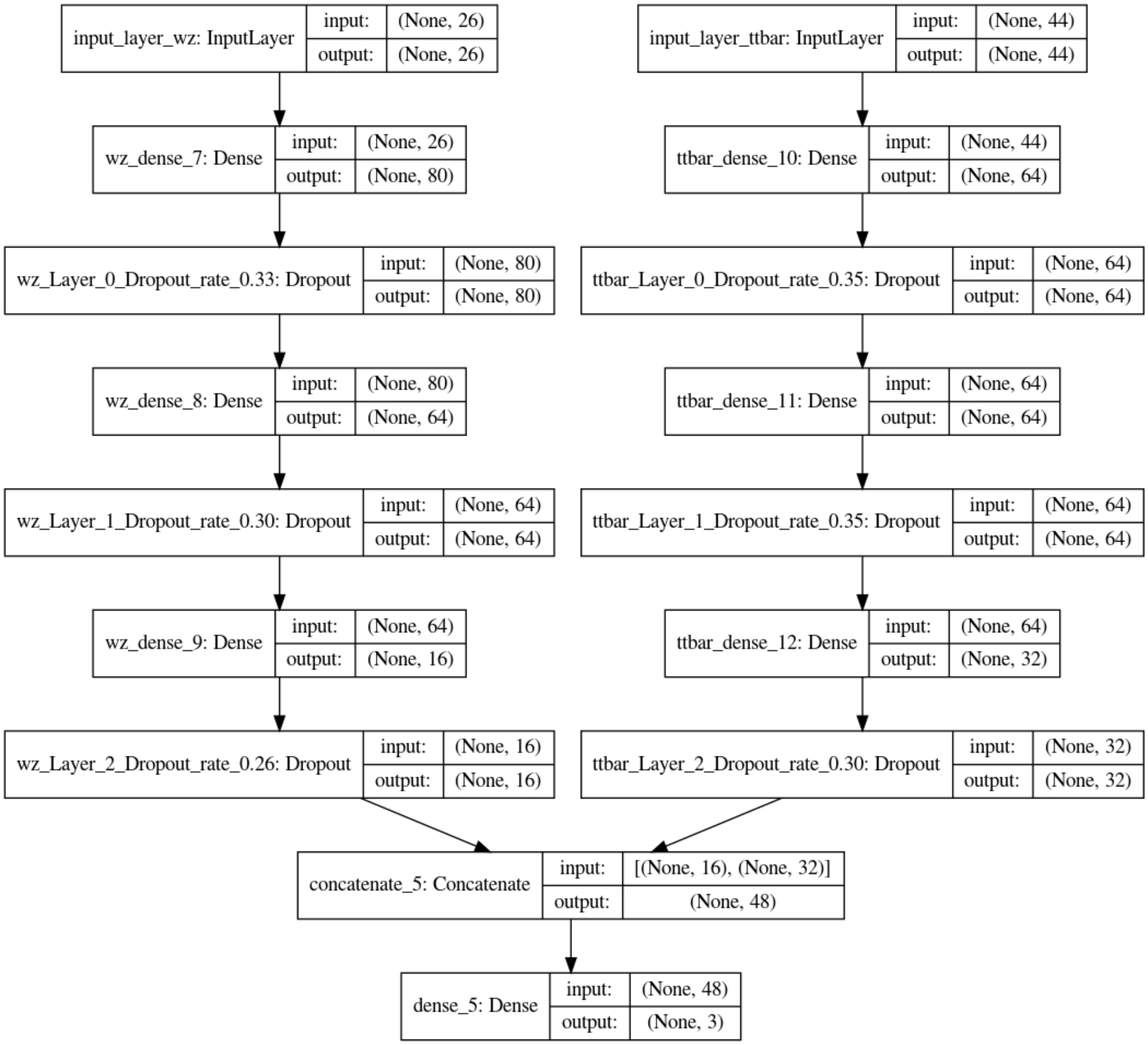


Figure B.3: Schematic drawing of the network architecture of the unified binary ANN.

B.6 Multiclassifier

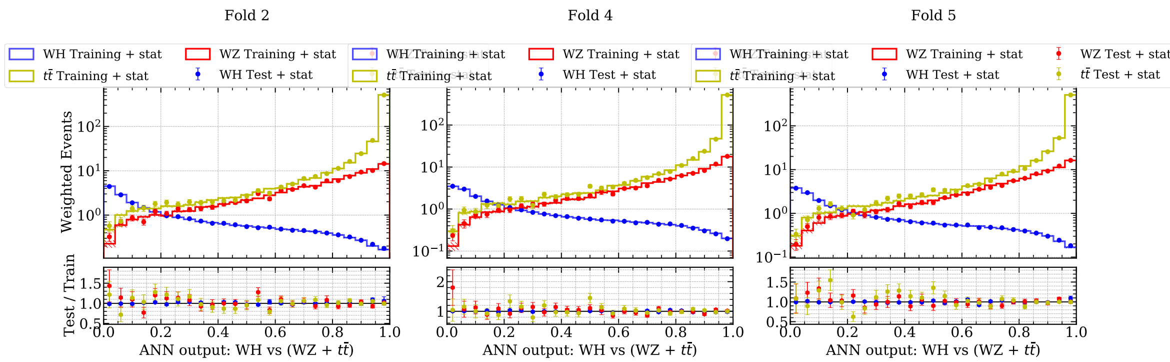


Figure B.4: The multiclassifier's prediction on train (dots) and test (lines) data folds in the upper plots with the corresponding ratios in the bottom plots on the example of fold 2, 4 and 5.

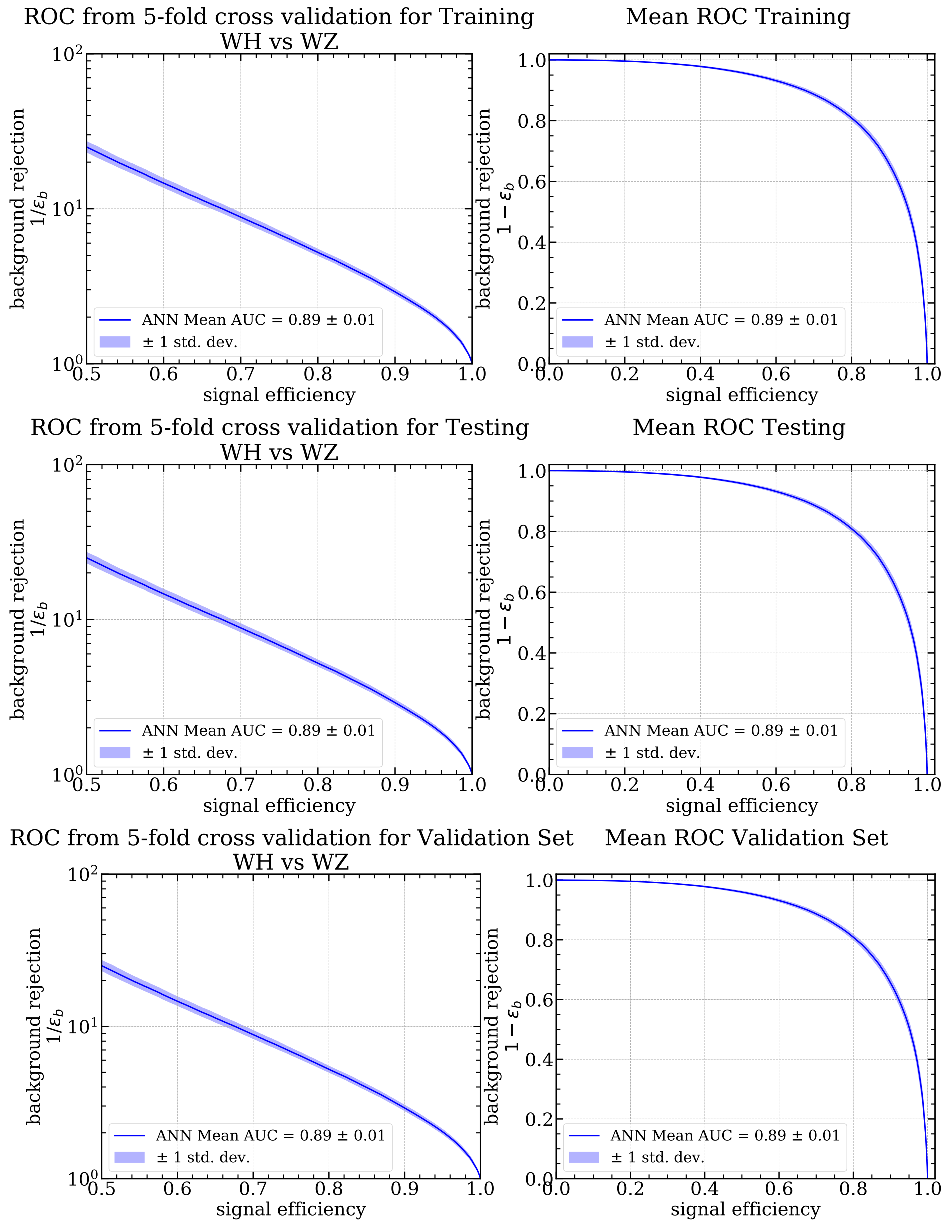


Figure B.5: Mean ROC of multiclassifier for training, test and validation.

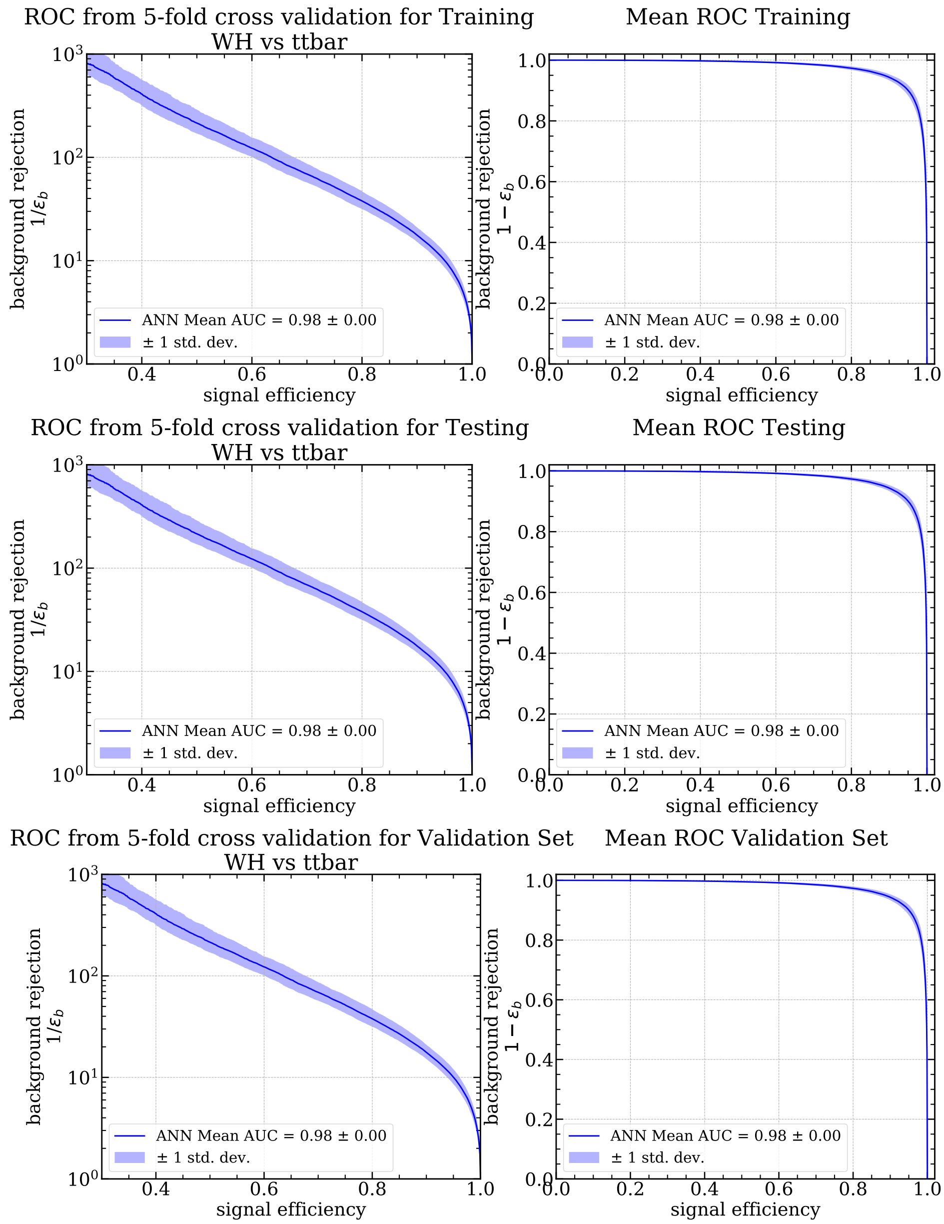


Figure B.6: Mean ROC of multiclassifier for training, test and validation.

References

[1] The ATLAS Collaboration. A Particle Consistent with the Higgs Boson Observed with the ATLAS Detector at the Large Hadron Collider. *Science*, 338(6114):1576–1582, 2012.

[2] CMS Collaboration. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716(1):30 – 61, 2012.

[3] Particle Data Group. Review of Particle Physics. *Progress of Theoretical and Experimental Physics*, 2020(8), 08 2020. 083C01.

[4] Peter W. Higgs. Broken Symmetries and the Masses of Gauge Bosons. *Phys. Rev. Lett.*, 13:508–509, Oct 1964.

[5] F. Englert and R. Brout. Broken Symmetry and the Mass of Gauge Vector Mesons. *Phys. Rev. Lett.*, 13:321–323, Aug 1964.

[6] Mark Thomson. *Modern Particle Physics*. Cambridge University Press, 2013.

[7] B. Povh, K. Rith, C. Scholz, and F. Zetsche. *Povh, B: Teilchen und Kerne.: Eine Einführung in die physikalischen Konzepte.*, volume 9 of *Springer-Lehrbuch*. Springer, 2014.

[8] D. Meschede and C. Gerthsen. *Gerthsen Physik*, volume 25 of *Springer-Lehrbuch*. Springer, 2015.

[9] Carsten Burgard. A Standard Diagram of the Current Standard Model of Physics. `https://texample.net/tikz/examples/model-physics/`, 12 2016.

[10] F. Scheck. *Theoretische Physik 3*, volume 4 of *Springer-Lehrbuch*. Springer, 2017.

[11] C. S. Wu, E. Ambler, R. W. Hayward, D. D. Hoppes, and R. P. Hudson. Experimental Test of Parity Conservation in Beta Decay. *Phys. Rev.*, 105:1413–1415, Feb 1957.

[12] Sheldon L. Glashow. Partial-Symmetries of Weak Interactions. *Nuclear Physics*, 22(4):579 – 588, 1961.

[13] Steven Weinberg. A Model of Leptons. *Phys. Rev. Lett.*, 19:1264–1266, Nov 1967.

[14] A. M. Sirunyan, A. Tumasyan, W. Adam, F Ambrogi, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, and et al. Measurement of the Weak Mixing Angle using the Forward–Backward Asymmetry of Drell–Yan Events in pp Collisions at 8 TeV. *The European Physical Journal C*, 78(9), Sep 2018.

[15] Hideki YUKAWA. On the Interaction of Elementary Particles. I. *Proceedings of the Physico-Mathematical Society of Japan. 3rd Series*, 17:48–57, 1935.

[16] The ATLAS Collaboration. Combined Measurements of Higgs Boson Production and Decay using up to $80\ \text{fb}^{-1}$ of proton-proton Collision Data at $\sqrt{s} = 13$ TeV Collected with the ATLAS Experiment. *Phys. Rev. D*, 101:012002, Jan 2020.

[17] Davison E. Soper. Parton Distribution Functions. *Nuclear Physics B - Proceedings Supplements*, 53(1-3):69–80, Feb 1997.

[18] LHC Higgs Cross Section Working Group (LHCXSWG). Higgs cross sections and decay branching ratios. `https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWG#Documents_and_Practical_Guide`.

[19] CERN. Facts and Figures about the LHC. `https://home.cern/resources/faqs/facts-and-figures-about-lhc`.

[20] The ALICE Collaboration. The ALICE Experiment at the CERN LHC. *Journal of Instrumentation*, 3, 2008.

[21] The ATLAS Collaboration. The ATLAS Experiment at the CERN Large Hadron Collider. *Journal of Instrumentation*, 3, 2008.

[22] The CMS Collaboration. The CMS experiment at the CERN LHC. *Journal of Instrumentation*, 3(08):S08004–S08004, aug 2008.

[23] The LHCB Collaboration. The LHCB detector at the LHC. *Journal of Instrumentation*, 3, 2008.

[24] The TOTEM Collaboration. The TOTEM Experiment at the CERN Large Hadron Collider. *Journal of Instrumentation*, 3(08):S08007–S08007, aug 2008.

[25] The LHCf Collaboration. The LHCf Detector at the CERN Large Hadron Collider. *Journal of Instrumentation*, 3(08):S08006–S08006, aug 2008.

[26] Vasiliki A Mitsou and. The MoEDAL Experiment at the LHC: Status and Results. *Journal of Physics: Conference Series*, 873:012010, jul 2017.

[27] CERN. LHC FAQ - The Guide. `https://home.cern/sites/home.web.cern.ch/files/2018-07/CERN-Brochure-2017-002-Eng_0.pdf`.

[28] Lyndon Evans and Philip Bryant. LHC Machine. *Journal of Instrumentation*, 3(08):S08001–S08001, aug 2008.

[29] Leonardo Rossi. More collisions at 2.36 tev. `https://atlas.cern/updates/atlas-news/more-collisions-236-tev`, 12 2009.

[30] CERN. CERN's Accelerator Complex. `https://home.cern/science/accelerators/accelerator-complex`.

[31] Corinne Pralavorio. Record Luminosity: well done LHC. `https://home.cern/news/news/accelerators/record-luminosity-well-done-lhc`.

[32] ATLAS experiment. Total Integrated Luminosity at Run 2 (13 TeV pp Data only). `https://atlas.web.cern.ch/Atlas/GROUPS/DATAPREPARATION/PublicPlots/2018/DataSummary/figs/intlumivstimeRun2DQall.pdf`.

[33] ATLAS experiment. Number of Interactions per Crossing. https://atlas.web.cern.ch/Atlas/GROUPS/DATAPREPARATION/PublicPlots/2018/DataSummary/figs/mu_2015_2018.pdf.

[34] ATLAS Experiment. Detector & Technology. https://atlas.cern/discover/detector.

[35] Joao Pequenao. Computer generated image of the whole atlas detector. https://cds.cern.ch/record/1095924.

[36] A Salzburger. The ATLAS Track Extrapolation Package. Technical Report ATL-SOFT-PUB-2007-005. ATL-COM-SOFT-2007-010, CERN, Geneva, Jun 2007.

[37] The ATLAS Collaboration. Magnet System. https://atlas.cern/discover/detector/magnet-system, 8 2020.

[38] G Aad et al. ATLAS Pixel Detector Electronics and Sensors. *Journal of Instrumentation*, 3(07):P07007–P07007, jul 2008.

[39] J R Pater. The ATLAS Semi Conductor Tracker Operation and Performance. *Journal of Instrumentation*, 7(04):C04001–C04001, apr 2012.

[40] A Vogel. ATLAS Transition Radiation Tracker (TRT): Straw Tube Gaseous Detectors at High Rates. Technical Report ATL-INDET-PROC-2013-005, CERN, Geneva, Apr 2013.

[41] M Capeans, G Darbo, K Einsweiller, M Elsing, T Flick, M Garcia-Sciveres, C Gemme, H Pernegger, O Rohne, and R Vuillermet. ATLAS Insertable B-Layer Technical Design Report. Technical Report CERN-LHCC-2010-013. ATLAS-TDR-19, Sep 2010.

[42] Joao Pequenao. Computer Generated Image of the ATLAS Inner Detector. Mar 2008.

[43] Georges Aad et al. Alignment of the ATLAS Inner Detector in Run-2. 7 2020.

[44] *ATLAS liquid-argon calorimeter: Technical Design Report*. Technical Design Report ATLAS. CERN, Geneva, 1996.

[45] Joao Pequenao. Computer Generated Image of the ATLAS Calorimeter. Mar 2008.

[46] PDG. Passage of Particles Through Matter. 2018.

[47] Joao Pequenao. Computer Generated Image of the ATLAS Muons Subsystem. Mar 2008.

[48] Joao Pequenao. Event Cross Section in a Computer Generated Image of the ATLAS Detector. Mar 2008.

[49] Aranzazu Ruiz-Martinez and ATLAS Collaboration. The Run-2 ATLAS Trigger System. Technical Report ATL-DAQ-PROC-2016-003, CERN, Geneva, Feb 2016.

[50] Federico Meloni. Primary Vertex Reconstruction with the ATLAS Detector. Technical Report ATL-PHYS-PROC-2016-163. 12, CERN, Geneva, Oct 2016.

[51] Improved Electron Reconstruction in ATLAS using the Gaussian Sum Filter-Based Model for Bremsstrahlung. Technical Report ATLAS-CONF-2012-047, CERN, Geneva, May 2012.

[52] ATLAS Collaboration. Electron and Photon Performance Measurements with the ATLAS Detector using the 2015–2017 LHC proton-proton Collision Data. *Journal of Instrumentation*, 14(12):P12006–P12006, dec 2019.

[53] Matteo Cacciari, Gavin P Salam, and Gregory Soyez. The anti-ktjet Clustering Algorithm. *Journal of High Energy Physics*, 2008(04):063–063, apr 2008.

[54] Philipp Jonas Windischhofer. Heavy-Flavour Jet Tagging in ATLAS. Jan 2020.

[55] Performance of Missing Transverse Momentum Reconstruction for the ATLAS Detector in the first proton-proton Collisions at at $\sqrt{s}$= 13 TeV. Technical Report ATL-PHYS-PUB-2015-027, CERN, Geneva, Jul 2015.

[56] S. Agostinelli et al. Geant4—A Simulation Toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250 – 303, 2003.

[57] François Chollet et al. Keras. `https://keras.io`, 2015.

[58] Kjell Magne Fauske. Neural Network. `https://texample.net/tikz/examples/neural-network/`, 12 2006.

[59] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, 2017.

[60] Wikimedia Commons. File:overfitting.svg — wikimedia commons, the free media repository, 2020. [Online; accessed 19-October-2020].

[61] Martín Abadi et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems, 2015. Software available from tensorflow.org.

[62] William Buttinger and Michel Lefebvre. Formulae for Estimating Significance. Technical Report ATL-COM-GEN-2018-026, CERN, Geneva, Oct 2018.

[63] The ATLAS Collaboration. Measurement of the Production Cross Section for a Higgs Boson in Association with a Vector Boson in the H→WW* $\rightarrow \ell\nu\ell\nu$ Channel in pp Collisions at $\sqrt{s} = 13$TeV with the ATLAS Detector. *Physics Letters B*, 798:134949, 2019.

[64] Trevor Hastie et al. *The Elements of Statistical Learning*. 2017.

[65] Y. Freund and R. E. Schapire. A Short Introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, 14, 1999.

[66] Natalie Wieseotte. "Higgs Boson Cross Section Measurement in the W-Associated Production and Fully Leptonic Final State at $\sqrt{s}$ = 13 TeV with the ATLAS Experiment, PhD Thesis".

[67] Paolo Nason. A New Method for Combining NLO QCD with Shower Monte Carlo Algorithms. *JHEP*, 11:040, 2004.

[68] Stefano Frixione, Paolo Nason, and Carlo Oleari. Matching NLO QCD Computations with Parton Shower Simulations: the POWHEG Method. *JHEP*, 11:070, 2007.

[69] Simone Alioli, Paolo Nason, Carlo Oleari, and Emanuele Re. A General Framework for Implementing NLO Calculations in Shower Monte Carlo Programs: the POWHEG BOX. *JHEP*, 06:043, 2010.

[70] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. A Brief Introduction to PYTHIA 8.1. *Comput. Phys. Commun.*, 178:852–867, 2008.

[71] J. Alwall, R. Frederix, S. Frixione, V. Hirschi, F. Maltoni, O. Mattelaer, H.-S. Shao, T. Stelzer, P. Torrielli, and M. Zaro. The Automated Computation of Tree-Level and Next-to-Leading Order Differential Cross Sections, and their Matching to Parton Shower Simulations. *Journal of High Energy Physics*, 2014(7), Jul 2014.

[72] Enrico Bothmann, Gurpreet Singh Chahal, Stefan Höche, Johannes Krause, Frank Krauss, Silvan Kuttimalai, Sebastian Liebschner, Davide Napoletano, Marek Schönherr, Holger Schulz, and et al. Event Generation with Sherpa 2.2. *SciPost Physics*, 7(3), Sep 2019.

[73] Yohei Yamaguchi. Recommended Isolation Working Points, 2020. [Online; accessed 24-November-2020].

[74] Konstantinos Nikolopoulos, Dimitrios Fassouliotis, Christine Kourkoumelis, and Alan Poppleton. Event-by-Event Estimate of Muon Energy Loss in ATLAS. *IEEE Trans. Nucl. Sci.*, 54:1792–1796, 2007.

[75] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, 2017.

[76] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[77] Ralf Gugel. *Measurement of Higgs Boson Production via Gluon Fusion and Vector-Boson Fusion in the H->WW* Decay Mode with the ATLAS Experiment at the LHC at $\sqrt{s}$ = 13 TeV*. PhD thesis, 01 2019.

[78] Aad G. Abbott B. et al. Aaboud, M. 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*, 79, 2019.

[79] Marcel Weirich. “Suche nach dem Higgs-Boson im Kanal V H $\rightarrow$ V WW$^{(*)}$ am LHC unpublished, N.D.”.

[80] Glen Cowan, Kyle Cranmer, Eilam Gross, and Ofer Vitells. Asymptotic Formulae for Likelihood-based Tests of New Physics. *The European Physical Journal C*, 71(2), Feb 2011.

[81] Apollinari G., Béjar Alonso I., Brüning O., Fessia P., Lamont M., Rossi L., and Tavian L. *High-Luminosity Large Hadron Collider (HL-LHC): Technical Design Report V. 0.1*. CERN Yellow Reports: Monographs. CERN, Geneva, 2017.

[82] D. et al. de Florian. *Handbook of LHC Higgs Cross Sections: 4. Deciphering the Nature of the Higgs Sector*. CERN Yellow Reports: Monographs. Oct 2016. 869 pages, 295 figures, 248 tables and 1645 citations. Working Group web page: https://twiki.cern.ch/twiki/bin/view/LHCPhysics/LHCHXSWG.

[83] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *CoRR*, abs/1602.04938, 2016.