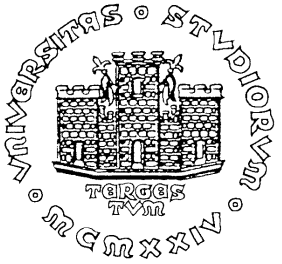


Università degli Studi di Trieste

Dipartimento di Fisica

Master Degree in Nuclear and Subnuclear Physics

Master Degree Thesis

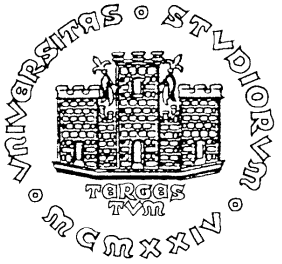
Classification Deep Neural Networks to optimize searches for Anomalous Charged Triple Gauge Couplings in $W^+ W^- \to l^+ l^- \nu_l \bar{\nu}_l$ events at the LHC with the ATLAS detector

Author:
Laura Pintucci

Supervisor:
Prof. Marina Cobal

Cosupervisor:
Dr. Giancarlo Panizzo

Academic Year 2020/2021



Università degli Studi di Trieste

Dipartimento di Fisica

Corso di Laurea Magistrale in Fisica Nucleare e Subnucleare

Tesi di Laurea Magistrale

Rete Neurale Profonda di Classificazione per ottimizzare la ricerca di Triple Gauge Coupling carichi anomali in eventi $W^+ W^- \to l^+ l^- \nu_l \bar{\nu}_l$ ad LHC con il detector ATLAS

Laureanda:
Laura Pintucci

Relatrice:
Prof. Marina Cobal

Correlatore:
Dr. Giancarlo Panizzo

Anno Accademico 2020/2021

Sommario

Negli esperimenti di fisica delle alte energie c'è una grande necessità di trovare metodi di analisi dati sempre più efficienti. In questa tesi studio la possibilità di utilizzare una rete neurale profonda per risolvere un problema di classificazone nell'ambito della fisica di LHC. In particolare applico questa tecnica ad un interessante caso di studio: l'estrazione dei limiti di esclusione per il coefficiente di Wilson c_{WWW} per i Triple Gauge Coupling carichi anomali attraverso lo studio della produzione dibosonica W^+W^-.

L'analisi è svolta utilizzando dei campioni di dati prodotti con il generatore di eventi MadGraph5_aMC@NLO, interfacciato con Pyhtia 8 e con il software Delphes. Inizialmente utilizzo un'analisi tradizionale con tagli sequenziali, definendo una regione di fiducia nello spazio delle fasi per il canale di decadimento dileptonico del bosone W. Quindi implemento con successo una rete neurale profonda, ottenendo una classificazione del segnale migliore con la variabile di output della rete neurale piuttosto che usando il tradizionale momento trasverso del leading lepton. Il valore della separazione $< S^2 >$ calcolato per le distribuzioni di queste due variabili è: $< S^2 >= 0.73$ per l'analisi tradizionale e $< S^2 >= 0.76$ per la rete neurale profonda. Infine ho calcolato i limiti di esclusione sia per la variabile tradizionale del momento trasverso del leading lepton, sia per la variabile di output della rete neurale profonda. Ho ottenuto i limiti con un livello di confidenza del 95 % (riportati nella tabella sottostante) con lo stimatore dei minimi quadrati utilizzando solo le incertezze statistiche, verificando che la rete neurale profonda permette di avere dei limiti più stringenti rispetto all'analisi con tagli sequenziali.

	c_{WWW} 95%CL [TeV^{-2}]
Lead lept p_T	[-3.2 , 3.2]
DNN output	[-2.3, 2.3]

Contents

Introduction

The sub-atomic world is effectively described by the Standard Model (SM) theory, and the Large Hadron Collider (LHC) program has been able until now to provide numerous experimental evidences of its validity, to be added to those from the many experiments in the past dedicated to its validation. Nonetheless there are different and substantial open questions, that can not be answered within the SM and whose solutions are being searched outside of this model, through the investigation of Beyond the Standard Model (BSM) physics.

In the field of High-Energy Physics (HEP) and at accelerators like LHC, BSM processes are studied via two main approaches: the direct searches for new physics, through the discovery of novel particles, and the indirect searches which tries to identify new physics phenomena as small deviations from the SM predictions. This second approach allows not only to study the possible presence of novel particles and interactions, but also to compute limits to new physics effects with respect to the SM. Triple Gauge Couplings (TGCs) are the self interactions of electroweak vector bosons and are an interesting and sensitive process to investigate BSM physics. Considering *anomalous* TGCs by adding to the SM Lagrangian new terms with additional parameters can be expressed within an Effective Field approach.

Data collected at LHC experiments, like ATLAS, need to be analysed in order to extract, out of their huge amount, information of interest. Traditionally, in HEP, data analysis uses a sequence of event selections based on variable distributions (referred to as sequential cut analysis) and then performs a statistical analysis of the selected data. Another technique, which is becoming more and more common, is Machine Learning (ML), a specific kind of algorithm able to learn how to predict a certain value directly from data, without being explicitly programmed to do so. In particular, Deep Learning (DL) is a sub-field of ML able to handle higher dimensional and more complex problems. Various studies have shown that shallow networks based on high-level features (manually-constructed non-linear variables) are outperformed by deep networks based on features with little or no pre-processing.

The purpose of this thesis work is to confirm the advantages of the implementation of DL techniques and in particular of the Deep Neural Networks (DNNs) in TGC data analysis. I take into consideration, as a case study, a currently interesting measurement: the computation of 95 % confidence level limit on one of the EFT parameters for anomalous charged Triple Gauge Couplings. In particular the aim of this thesis is to show that the DNN is able to better separate a BSM signal from the SM background with respect to the traditional sequential cut analysis. The analysis is performed with Monte Carlo simulated events of diboson W^+W^- production at the LHC with the ATLAS detector, at a centre of mass-energy of $\sqrt{s} = 13$ TeV and with different integrated luminosity assumptions: $\mathcal{L} = 36.1$ fb^{-1}, $\mathcal{L} = 139$ fb^{-1}, $\mathcal{L} = 300$ fb^{-1}, and $\mathcal{L} = 3000$ fb^{-1} . In this thesis, I study the diboson production in the dileptonic decay channel. The Monte Carlo generated data are also used to train the DNN, thanks to the fact that it is known a priori if they belong to BSM events or to the SM background.

The first chapter of this thesis is an introduction to the SM, its limits and how they can be overcome with BSM theories. There I will discuss also Triple Gauge Couplings, their definition, their possible BSM contributions in the EFT paradigm, and their study through diboson production. In the second chapter the main characteristics and features of LHC and ATLAS are illustrated. The third chapter introduces the theory of Deep Learning and in particular Deep Neural Networks. Particular attention is given to techniques used in this thesis work.

In the fourth chapter Monte Carlo generators used in this thesis are presented with their main characteristics. In chapter five the data samples produced for this work are presented. Then the sequential cut analysis implementation and the DNN analysis are illustrated. In particular the optimization of the DNN for the specific case of study is presented. The performances of the traditional analysis are compared with those of the DNN. Finally the conclusion are drawn, showing the results obtained.

Chapter 1

Standard Model and Triple Gauge Couplings

1.1 The Standard Model

The Standard Model (SM) is the theoretical model that describes the fundamentals of particle physics: the elementary particles and their interactions. It was originally developed by S. Glashow [1], S. Weinberg [2] and A. Salam [3] in the sixties and was improved during the second half of the 20th century. Particle physicists have verified SM predictions innumerable times, some of the most important evidences being: the discovery of the W and Z bosons [4], the discovery of the top quark [5], and the latest discovery of the Higgs boson in 2012 [6] at CERN.

The SM is a particular Quantum Field Theory (QFT), built using the Lagrangian formalism, in which particles are described as excited states of fields. These particles interact with each other through three main forces in the SM: the electromagnetic force, the weak force, and the strong force. The local gauge principle in the SM imposes invariance of physics under local phase transformations. The Higgs mechanism, included in the SM, explains mass generation for the particles in the model, driven by what can be seen as a kind of fourth fundamental interaction.

1.1.1 Particles and their Interactions

The elementary particles of the SM, shown in Fig. 1.1, have no internal structure and are divided in two groups based on their spin: *fermions* that have spin $1/2$ and obey the Fermi-Dirac statistics, and *bosons* that have integer spin and obey Bose-Einstein statistics.

Fermions are divided in two main categories based on their interactions: *quarks* that interact via the strong, the electromagnetic and the weak force, and *leptons* that do not interact via the strong force, while interacting via the electromagnetic and weak ones. *Quarks* are characterized by six flavours, one for each of them, and are divided in three generations with increasing masses (*(u,d)*, *(c,s)*, *(t,b)*). Each quark has its corresponding anti-quark, with same mass but opposite quantum numbers. Every generation is composed of two particles, one with charge $+2/3\ e$ and the other with charge $-1/3\ e$. Quarks also have a *colour* quantum number (*red*, *blue* or *green*) and are confined in bound states called *hadrons* in order to be in *colour* neutral states. *Hadrons* are divided in two groups depending on the number of quarks they are composed by: *mesons* are made of a quark and an anti-quark, whereas *barions* are made of three quarks. *Leptons* are divided in three generations with increasing mass, each formed by a negative charged particle (e, μ, τ) and the correspondent neutrino

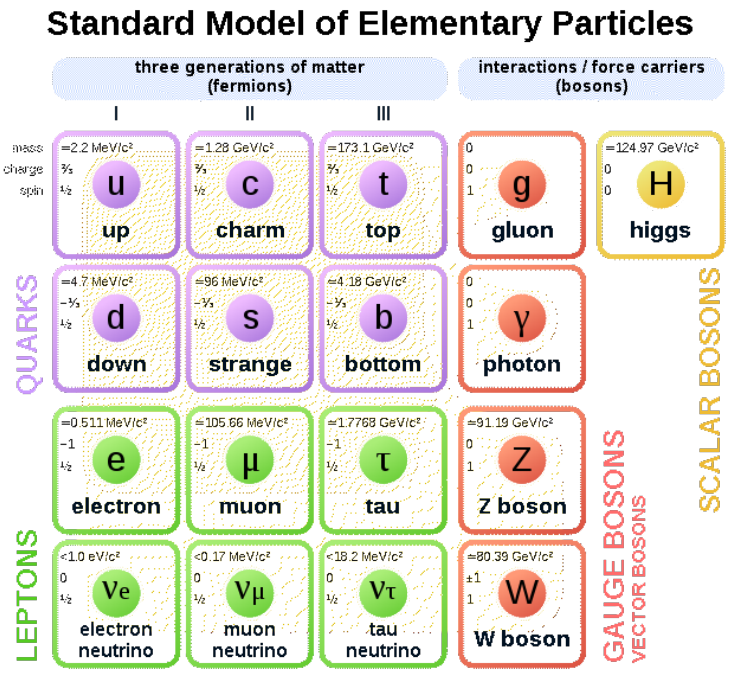


Figure 1.1: Diagram showing fundamental particles of the SM.

(ν_e, ν_μ, ν_τ), that is a light and neutral particle that interacts only via the weak force. Each lepton has its corresponding anti-particle with same mass but opposite quantum numbers.
Bosons are divided in *vector bosons* with spin 1 and *scalar bosons* with spin 0. The first ones are also called *gauge bosons* and are the force carriers of the three fundamental forces described by the SM. The strong force is carried by eight massless *gluons* g that carry colour charge, based on the $SU(3)_c$ gauge group. The weak force is carried by two charged *bosons* $W^\pm$ and a neutral one Z^0: all of them are massive particles. Finally, the electromagnetic force is carried by the massless photon γ. The SM Lagrangian in global obeys the $SU(3)_c \times SU(2)_L \times U(1)_Y$ symmetry. The only scalar *boson* in the SM is the *Higgs boson*, which arises from the spontaneous electroweak symmetry breaking, through the Higgs mechanism.

1.1.2 Quantum Electrodynamics and Electroweak Unification

The electromagnetic (EM) fields $\vec{E}$ and $\vec{B}$ can be described using the electromagnetic vector potential $\vec{A}$ and the scalar potential Φ. The Maxwell scalar and vector potentials can be written in four-vector notation as $A_\mu = (\Phi, -\vec{A})$. The Maxwell equations relate the electric field ($\vec{E}$) and the magnetic field ($\vec{B}$) to each other and can be written for free fields in a Lorentz covariant form as

$$\partial^\mu F_{\mu\nu} = 0 \tag{1.1}$$

where $F^{\mu\nu}$ is the EM field strength tensor defined as $F^{\mu\nu} = \partial_\mu A_\nu - \partial_\nu A_\mu$. Let's consider a local phase transformation:

$$\psi(x) \rightarrow \psi'(x) = e^{i\chi(x)}\psi(x) \tag{1.2}$$

where the function $q(x)$ is the phase. If this phase is constant in the whole space-time, equation 1.2 is a global phase transformation. The tensor $F_{\mu\nu}$ is manifestly

gauge invariant under a local phase transformation. This *invariance* is associated with the *gauge transformation* of the scalar and vector EM potentials that preserve the EM fields:

$$\Phi \rightarrow \Phi' = \Phi - \frac{d\chi}{dt} \quad \text{and} \quad \vec{A} \rightarrow \vec{A}' = \vec{A} - \vec{\nabla}\chi \tag{1.3}$$

where χ is an arbitrary scalar function. The Lagrangian density resulting in the equations of motion (eq: 1.1) for the free EM field is:

$$\mathcal{L}_{EM} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu}. \tag{1.4}$$

The Lagrangian describing the interaction of fermions and photons is obtained coupling the Dirac field to the Maxwell field:

$$\mathcal{L}_{D,EM} = \overline{\psi}(x)(i\gamma^{\mu}(\partial_{\mu} - ieA_{\mu}) - m)\psi(x) \tag{1.5}$$

where γ^{μ} are the Dirac matrices, $\psi(x) = (\psi_1(x), \psi_2(x), \psi_3(x), \psi_4(x))^T$ is the Dirac spinor, $\overline{\psi}(x) = \psi^{\dagger}(x)\gamma^0$, m is the fermion mass and A_{μ} can be interpreted as a massless gauge field. This Lagrangian is invariant under the local symmetry of the $U(1)$ group (see equation 1.2). If we define the gauge covariant derivative $D_{\mu} = \partial_{\mu} - ieA_{\mu}$, the complete gauge invariant QED Lagrangian is:

$$\mathcal{L}_{QED} = -\frac{1}{4} F_{\mu\nu} F^{\mu\nu} + \overline{\psi}(x)(i\gamma^{\mu} D_{\mu} - m)\psi(x). \tag{1.6}$$

The gauge local invariance principle is fundamental to introduce weak interactions too in the gauge invariant Lagrangian formalism. The Electroweak (EW) theory [7] combines together the EM theory with the weak interaction as two different manifestations of one single interaction. This interaction is based on the group $SU(2)_L \times U(1)_Y$, where $SU(2)_L$ is the weak isospin group and $U(1)_Y$ is the weak hypercharge group. The last one is an abelian group, with the corresponding gauge field B_{μ}, and it has a local gauge symmetry defined as:

$$\psi(x) \rightarrow \psi'(x) = e^{ig' \frac{Y}{2} \chi(x)} \psi(x) \tag{1.7}$$

where g' is the $U(1)_Y$ coupling constant and $\frac{Y}{2}$ is the weak hypercharge group generator. On the other hand, the $SU(2)_L$ is a non-abelian group, with the corresponding three gauge fields $W^{\mu}_{k=1,2,3}$, and it has a local gauge symmetry:

$$\psi(x) \rightarrow \psi'(x) = e^{ig_W \vec{\alpha}(x) \cdot \vec{T}} \ \psi(x) \tag{1.8}$$

where g_W is the $SU(2)_L$ coupling constant, $\vec{T} = \frac{1}{2}\vec{\sigma}$ are the three weak isospin group generators proportional to the Pauli matrices (σ_i), and $\vec{\alpha}(x)$ are three functions indicating the local phase at each point in space.
The fermion spinors ψ can be written as two parts: right-handed (RH) and left-handed (LH), defined as:

$$\psi(x)_R = \frac{1}{2}(1 + \gamma^5) \ \psi(x) \quad \text{and} \quad \psi(x)_L = \frac{1}{2}(1 - \gamma^5) \ \psi(x) \tag{1.9}$$

where γ^5 is proportional to the other four Dirac matrices. The transformation 1.8 affects only LH particles and RH antiparticles, which means that parity is not conserved in weak interactions. In this way fermions can be LH doublets, with weak isospin $I = \frac{1}{2}$, or RH singlets, with $I = 0$; for example the first leptonic generation family is

$$\begin{pmatrix}\nu_{e,L}\\ e_L\end{pmatrix}, \quad e_R \tag{1.10}$$

The $SU(2)_L \times U(1)_Y$ gauge group would imply the existence of four massless gauge bosons, of which the two charged physical gauge bosons are:

$$W^{\pm} = \frac{1}{2}(W^{\mu}_{(1)} \pm iW^{\mu}_{(2)}) \tag{1.11}$$

while for the following discussion it is useful to introduce two new physical gauge bosons A^{μ} and Z^{μ} as:

$$\begin{aligned} A^{\mu} &= B^{\mu}\cos\theta_W + W^{\mu}_{(3)}\sin\theta_W \\ Z^{\mu} &= -B^{\mu}\sin\theta_W + W^{\mu}_{(3)}\cos\theta_W \end{aligned} \tag{1.12}$$

where θ_W is the weak mixing angle. In the EW paradigm the EM interactions is a subgroup of the EW group and its generator Q is a linear combination of the EW generator:

$$Q = T_3 + \frac{Y}{2}. \tag{1.13}$$

This abstract EW theory would require, as stated before, the existence of four gauge bosons, which should be massless. On the contrary three of these gauge bosons ($W^{\pm}, Z$) are seen to be massive; to include these masses in the theory the Higgs mechanism is used, which triggers the EW spontaneous symmetry breaking.

1.1.3 The Higgs Mechanism and Particle Masses

The spontaneous symmetry breaking of the $SU(2)_L \times U(1)_Y$ local gauge symmetry, introduced by Higgs, Englert and Brout [8, 9], can be obtained considering the following Lagrangian

$$\mathcal{L} = (\partial_{\mu}\phi)^{\dagger}(\partial^{\mu}\phi) - V(\phi) \tag{1.14}$$

where ϕ is a weak isospin doublet of two complex scalar fields

$$\phi = \begin{pmatrix}\phi^{\dagger}\\ \phi^0\end{pmatrix} = \begin{pmatrix}\phi_1 + i\phi_2\\ \phi_3 + i\phi_4\end{pmatrix}. \tag{1.15}$$

In equation 1.14 the first part of the Lagrangian is the kinematic term and the second term is the potential, defined as $V(\phi) = \frac{1}{2}\mu^2\phi^{\dagger}\phi + \frac{1}{4}\lambda(\phi^{\dagger}\phi)^2$. The signs of the two real constants λ, μ^2 of the potential $V(\phi)$ determine the shape of the potential. The first one must be positive in order for the potential to be bounded from below, while the second can be both positive or negative. In the case of $\mu^2 < 0$ the potential has a set of degenerate minima (called vacuum states) identified by $\phi^{\dagger}\phi = {}^{v^2}/_2 = -{}^{\mu^2}/_{2\lambda}$. This way the local minimum is no longer unique causing the spontaneous symmetry breaking (SSB).

In order to maintain the photon mass equal to zero even after the symmetry breaking, the vacuum expectation value must be non-zero only for the first component of the field, taking the value:

$$\langle 0|\phi|0\rangle = \begin{pmatrix}0\\ \frac{v}{\sqrt{2}}\end{pmatrix}. \tag{1.16}$$

The field can be expanded around the minimum as:

$$\phi(x) = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1(x) + i\phi_2(x) \\ v + \eta(x) + i\phi_4(x) \end{pmatrix}. \tag{1.17}$$

Considering this expansion the field can be written in the unitary gauge as:

$$\phi(0) = \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ v + h(x) \end{pmatrix} \tag{1.18}$$

with $h(x)$ a scalar field. This way four fields are introduced, three of which give the transverse degree of freedoms for the $W^{\pm}$, Z bosons.
The Lagrangian in equation 1.14 is not *locally* gauge invariant under the $SU(2)_L \times U(1)_Y$ group. It becomes also locally gauge invariant by considering, instead of the partial derivative ∂_μ, the covariant derivative:

$$D_\mu = \partial_\mu + ig_W \vec{T} \cdot \vec{W}_\mu + ig' \frac{Y}{2} B_\mu. \tag{1.19}$$

The kinetic part of the Lagrangian 1.14, considering the covariant derivative in equation 1.19 and the unitary gauge, has terms quadratic in the bosons fields, which represents their masses. In particular, for the $W^{\pm}$ bosons the term is:

$$\begin{aligned} m_W^2 W_\mu^+ W_-^\mu &= \frac{1}{4} g_W^2 W_\mu^- W_+^\mu v^2 \\ m_W &= \frac{1}{2} g_W v \end{aligned} \tag{1.20}$$

for the Z boson the term of the Lagrangian is

$$\begin{aligned} \frac{1}{2} m_Z^2 Z_\mu Z^\mu &= \frac{1}{8} v^2 (g_W + g') \\ m_Z &= \frac{v}{2} \sqrt{g_W + g'} \end{aligned} \tag{1.21}$$

while for the γ boson there is no term proportional to $A_\mu A^\mu$, being it a massless boson.
The masses m_W and m_Z can be related through the mixing angle θ_W:

$$\cos\theta_W = \frac{g_w}{\sqrt{g_W^2 + g'^2}} \tag{1.22}$$

$$m_Z = \frac{m_W}{\cos\theta_W}. \tag{1.23}$$

Regarding fermions, their interaction with the Higgs boson takes place through the *Yukawa interaction*. The Yukawa Lagrangian is

$$-\mathcal{L}_{Yu} = \sum_{ij} \Gamma_u^{ij} \overline{Q}_L^i (i\sigma_2) \phi^* u_R^j + \Gamma_d^{ij} \overline{Q}_L^i \phi d_R^j + \Gamma_e^{ij} \overline{L}_L^i \phi e_R^j + [h.c.] \tag{1.24}$$

where $\Gamma_{u,d,e}^{ij}$ are the 3×3 complex Yukawa matrices, $\overline{Q}_L^i$ and $\overline{L}_L^i$ are the quarks and leptons left-handed (LH) doublets, u_R^j, d_R^j and e_R^j are the right-handed (RH) quarks and leptons singlets, ϕ is the Higgs doublet and $[h.c.]$ indicates the hermitian coniugate of all the previous terms. If we now consider small perturbations around

the vacuum state of the Higgs doublet, with the SSB, in the Lagrangian a term appears which is quadratic in the fermions fields, representing the mass term of the fermions with masses:

$$m_f = \frac{g_f v}{\sqrt{2}} \tag{1.25}$$

where g_f are the Yukawa couplings, that can be obtained diagonalizing the Γ matrices.
In the SM the neutrinos are assumed to be massless, anyway it is possible to give them masses with a mechanism similar to that of other fermions. Experiments shows that neutrinos should have little masses compared to other fermions, suggesting that a different mechanism might be responsible for their masses.

1.1.4 Quantum Chromodynamics

The SM describes the strong interactions between quarks and gluons through the gauge theory of the $SU(3)_c$ group, called Quantum Chromodynamics (QCD)[10]. The three colour charges of $SU(3)_c$ (r, g and b) are the non abelian analogue of correspond to the electric charge in QED. The quarks dynamic is defined by the colour symmetry in the $SU(3)_c$ group, which is non-Abelian and thus the interaction mediators, called gluons, can carry colour charge and interact with themselves. These eight gluons, related to the generators of the $SU(3)_c$ group, are the gauge bosons of the strong interaction. The QCD Lagrangian is:

$$\mathcal{L}_{QCD} = -\frac{1}{4} F_A^{\alpha\beta} F^A_{\alpha\beta} + \overline{\psi}_a (i\gamma_\mu D^\mu_{ab} - M_a \delta_{ab})\psi_b \tag{1.26}$$

where ψ_a are the quarks with spin $1/2$ in the fundamental representation ($a, b = 1, 2, 3$). The field strength tensor is defined as $F_A^{\alpha\beta} = \partial^\alpha A_A^\beta - \partial^\beta A_A^\beta + g_s f_{ABC} A_B^\alpha A_C^\beta$, where A_A^μ indicates the eight gluons ($A, B, C, ... = 1, 2, ..., 8$) in their adjoint representation and f_{ABC} is the structure constant of the $SU(3)_c$ group, which controls the interaction term, since the group is non-Abelian. The covariant derivative in the Lagrangian 1.26 is defined as $D^\mu_{ab} = \delta_{ab}\partial^\mu + i g_s (t^A)_{ab} A^\mu_a$ where $t^A = \frac{1}{2}\lambda^A$ are the group generators proportional to the Gell-Mann matrices, and the structure constant can be written as $[t_A, t_B] = i f_{ABC} t_C$. The QCD Lagrangian can be divided in three different parts

$$\mathcal{L}_{QCD} = \mathcal{L}_G + \mathcal{L}_q + \mathcal{L}_{int} \tag{1.27}$$

where the first part is the kinetic term for the gluon massless field $\mathcal{L}_G = -\frac{1}{4} F_A^{\alpha\beta} F^A_{\alpha\beta}$, the second one is the kinetic term for the massive quark fields $\mathcal{L}_q = \overline{\psi}_a (i\gamma_\mu \delta_{ab}\partial^\mu - M_a \delta_{ab})\psi_b$ and the last part represents the interaction between quarks and gluons $\mathcal{L}_{int} = -\overline{\psi}_a (g_s \gamma^\mu (t^A)_{ab} A^\mu_a)\psi_b$.
The strong coupling constant g_s depends on the energy scale of the interaction (Q^2): it is small for large momentum transfers and it is big for small momentum transfers. The first phenomenon is known as *asymptotic freedom*, meaning that at high energy (small distances) quarks behave almost as free particles, while the second is known as *colour confinement* quarks are bound together at small energy (large distances). When trying to separate quarks, a new pair of quarks with opposite colour charge appears from the vacuum, and new bound states are then created, as shown in figure 1.2 .
It is in fact impossible to isolate a quark, while they exist in structures called hadrons. These can be grouped in two categories: mesons, composed of a quark and an anti-quark ($q\overline{q}$), and baryons, made of three quarks (qqq or $\bar{q}\bar{q}\bar{q}$).

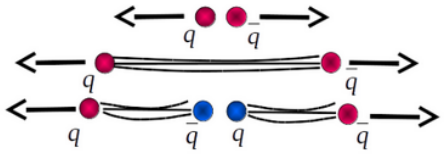


Figure 1.2: Schematic representation of colour confinement. When trying to distance a pair of quark and antiquark in a meson, a new couple of quark and antiquark is created and the system ends up in two bound state mesons.

The QCD Lagrangian has a built-in exact colour gauge symmetry and as a consequence the gluons are massless. This Lagrangian shows the existence of quarks, which carry a colour charge (r, g, b), and gluons, which carry a combination of colour charge, six of the gluons are colour charged $(r\bar{g}, r\bar{b}, ...)$ and two are neutral ($\frac{1}{\sqrt{2}}(r\bar{r} - g\bar{g})$ and $\frac{1}{\sqrt{6}}(r\bar{r} + g\bar{g} + 2b\bar{b})$).
Prontons and neutrons, the nuclei constituents, are made of partons, that is quarks and gluons. If one considers the collision between two protons, this can be described through these point-like partons, each one carrying a fraction of the total momentum. The production cross section of this event is obtained as the sum of two parts of the interaction: the first at large momentum transfer, which can be described in the perturbative QCD frame, while the second one at small momentum transfer, described with a parametrisation using the Parton Distribution Functions (PDFs). The PDF $f_{a/A}(x/Q^2)$ of parton a, in a proton A with a momentum p_A, is a function of the relative momentum of the parton $x = p_a/p_A$ and of the energy scale Q. These PDFs can not be calculated analytically but must be derived with a fit on various experimental data sets of different processes like deep inelastic scattering, Drell-Yann events and jet production.

1.2 Beyond the Standard Model

Nowadays the SM is the most accurate description of the experimentally observed phenomenology of particle physics processes. Its predictions have been proven with high precision during the last decades at the LHC accelerators and other experiments. After the discovery of the Higgs boson in 2012 all fundamentals particles predicted by the SM have been found. Despite this, there are some physics phenomena that can not be explained by the SM, and for this reason this last one is not considered a complete theory of particle physics. Researchers are then looking for evidences of physics Beyond the Standard Model (BSM). In the following, some puzzles unexplained within the SM are listed:

- **matter - antimatter asymmetry**: the SM predicts the existence of an almost equal percentage of matter and antimatter, while evidently the reality is far from this balance. Since free quarks are not observed (as explained earlier in 1.1.4), to study the baryon asymmetry, one must look back at the time when baryons were formed in the early universe, just after the Big Bang [11]. According to current cosmological models, in the early universe particles, antiparticles and radiation were in constant interaction through particle creation and annihilation, until, due to the Universe expansion, a critical temperature was reached and particle creation in thermal equilibrium stopped. If these particles were created in equal amounts of matter and antimatter, as the SM predicts, it would be expected that all matter particles would annihilate with

antimatter ones producing energy. This however is not observed.
The parameter which describes the matter - antimatter asymmetry is the *baryon asymmetry parameter*:

$$\eta = \frac{\tilde{n}_B}{n_\gamma} \tag{1.28}$$

where $\tilde{n}_B$ is the baryon number density and the bayon number is defined as the difference between the quarks and antiquarks number $n_B = n_q - n_{\bar{q}}$, and n_γ is the photon number density. The current experimental value for the baryon asymmetry parameter [12] is $\eta \sim (5.8 - 6.5) \cdot 10^{-10}$ (95% CL) from which can be derived the value for the baryon-antibaryon ratio of $\sim 10^{-4}$. There are various new theories, not yet verified, which try to justify this phenomenon, but the question still remains unsolved.

- **dark matter (DM)**: there are several suggestions of the existence of a type of matter that interacts only gravitationally but not electromagnetically, which is not described by the SM. Some of the more relevant ones are the galaxy rotation curve [13], gravitational lensing observations [14] and precision measurements of the cosmic microwave background (CMB) anisotropies [15]. In particular, the cosmological model used to explain cosmological observations is the ΛCDM model. This with the most recent measures of its parameters [15], gives indication of the Universe being composed by a $\sim 5\%$ of baryonic matter, $\sim 27\%$ of dark matter and $\sim 68\%$ of dark energy.
  Some of the dark matter candidates that have been proposed up to now are supersymmetric particles, primordial black holes and Weakly Interacting Massive Particles (WIMPs).

- **neutrino masses**: in the SM neutrinos are considered to be massless, but there are experimental evidences of the contrary, as already mentioned in section 1.1.3. In particular, experiments show a phenomenon called neutrino oscillations [16], which implies flavour mixing for neutrinos, possible only if neutrinos are massive.
  One can account for the small neutrino masses in the SM Lagrangian adding a new term, through the so called *Seesaw mechanism*. In the neutrino mass term, other than the usual Dirac term, the Majorana term is added too. The Lagrangian term would then be;

$$\mathcal{L}_M = -\frac{1}{2} \begin{pmatrix} \overline{\nu}_L & \overline{\nu}_R^c \end{pmatrix} \begin{pmatrix} 0 & m_D \\ m_D & M \end{pmatrix} \begin{pmatrix} \nu_L^c \\ \nu_R \end{pmatrix} + (h.c.) \tag{1.29}$$

  where ν_L^c and ν_R^c are respectively the left and right neutrino CP conjugate field (here CP is the Charge-Parity symmetry), m_D is the Dirac mass and M is the Majorana mass. Diagonalising the mass matrix and considering $m_D << M$, two different states, the light neutrino ν with mass $m_\nu \approx {}^{m_D^2}/_M$, and the heavy neutrino N with mass $m_N \approx M$, are obtained.

- **gravity**: the fourth fundamental interaction between particles is gravity, which is not included in the SM. In particular, gravity should have its own particle carrier, usually called *graviton*, as for the three other fundamental interactions. However the SM does not predict its existence.
  Nowadays there is a great effort of the scientific community to find a theory which unifies the general theory of relativity and the quantum theory. The two

main categories of solutions are: loop or non-local quantum gravity models that have a non-perturbative approach [17] [18]; and local and perturbative theories, such as supergravity [19]. The problem still remains open, since it has not been discovered a particle that could be the gravitation carrier and there is no testable model that can unifies quantum and gravitation theories.

To these it is worth to add an additional puzzle affecting every attempt of model building BSM: the **hierarchy problem**. Because of the instability of the Higgs boson mass at high energy scales, a lot of fine tuning is necessary to maintain the value of the Higgs mass that we observe at lower energies in the modern particle experiments. In the SM the particle masses depend on the energy scale that is being considered, and undergo loop corrections. For the Higgs boson this means that there are high mass corrections, quadratic in the energy scale, at high energy scales. To have a value of $m_H = 125$ GeV at the energy explored in modern particle experiments, a high precision fine tuning is needed, which is considered unnatural.

1.3 Triple Gauge Couplings

The subject of this thesis is the study of triple gauge couplings (TGCs), which in the SM are triggered by gauge bosons self-interaction. This phenomenon is particularly interesting since it gives a chance to study possible effects of BSM physics. These interactions are strictly related to the gauge group of the model which describes them, thus a deviation from the prediction of the SM in the measurements would offer important information on the underlying BSM physics.
Gauge boson self-interactions and possible anomalous couplings can be described within two conceptual frameworks: the anomalous coupling of the EW vector boson [20], and the Effective Field Theory (EFT) [21], that, as revived recently [22], can describe anomalous couplings too. In my study of the TGCs I use the EFT approach, which is largely used by the particle physics community to describe BSM physics effects.

1.3.1 Effective Field Theory

The EFT is a theoretical framework that does not rely on a specific model, and therefore can be used to study the effect of new particles or non-standard interactions of SM particles without introducing a specific extension of the SM. It is also important that EFT gives a way to compute the accuracy at which new physics can be excluded in a specific calculation based on certain data.
From the point of view of EFT the SM Lagrangian described in section 1.1 is a leading order approximation of a more general EFT, which is valid at the energy scale of modern particle physics experiments. The SM Lagrangian contains all terms of a QFT consistent with Lorentz symmetry, the field content introduced in sections 1.1.2, 1.1.3, and 1.1.4, and gauge invariance under the group $SU(3)_c \times SU(2)_L \times U(1)_Y$ up to mass-dimension four. However there are many other terms of higher dimension which respect the same invariance and which could be included in the EFT Lagrangian of the SM. These terms are coupled to a coefficient Λ raised to the appropriate power by dimensional analysis. In fact, the Lagrangian has dimension four, thus a term of dimension d has a coefficient of mass dimension $4-d$, which is written as the scale Λ . The general Lagrangian describing anomalous triple gauge couplings will then be:

$$\mathcal{L}_{EFT} = \mathcal{L}_{SM} + \sum_i \frac{c_i}{(\Lambda_i^{(6)})^2} O_i^{(d=6)} + ... \tag{1.30}$$

If the Λs in the Lagrangian are much higher than the energy scale being probed with a certain experiment, then it is expected that the theory predictions are unaffected by higher order terms, that could be considered $\mathcal{L}_{EFT} \simeq \mathcal{L}_{SM}$.
One possibility to discover and study new physics is to observe new particles production at high energy colliders. However there could be two problems: these particles can be too heavy to be produced at the energy presently available in modern colliders, or if these particles decay into final states that have a large SM background, which makes difficult their identification. Another way to study BSM physics is via precision measurements of SM processes: in fact new particles can contribute to these SM processes as intermediate states so that they can be searched as deviations of measured observables from SM predictions. For precision measurements the EFT provides a framework which gives the possibility to perform analysis whose results can be interpreted in a large class of BSM scenarios, without considering only one new physics model.

Gauge bosons can self-interact as a consequence of the non-abelian structure of the SM gauge groups, in particular the $SU(2)_L \times U(1)_Y$ gauge invariance and its spontaneous breaking determine the WWZ and $WW\gamma$ TGCs. New physics effects can be observed in the low-energy effective theory as anomalous TGCs. The effective Lagrangian describing a WWV vertex, where V can either be a Z or γ gauge boson, can be written as [23, 24, 25]:

$$\begin{aligned}\mathcal{L}_{WWV} = g_{WWV}(& g_1^V (W^+_{\mu\nu} W^{-\mu} - W^{+\mu} W^-_{\mu\nu}) V^\nu + \kappa_V W^+_\mu W^-_\nu V^{\mu\nu} + \\ & \frac{\lambda_V}{m_W^2} W^+_\mu W^-_\rho V^\mu_\rho) + i g_4^V W^+_\mu W^-_\nu (\partial^\mu V^\nu + \partial^\nu V^\mu) + \\ & i g_5^V \epsilon^{\mu\nu\rho\sigma} (W^+_\mu \partial_\rho W^-_\nu \partial_\rho - W^+_\mu W^-_\nu) V_\sigma + \\ & \tilde{\kappa}_V W^+_\mu W^-_\nu \tilde{V}^{\mu\nu} + \frac{\lambda_V}{m_W^2} W^{\nu+}_\mu W^{\rho-}_\nu \tilde{V}^\mu_\rho \;)\end{aligned} \tag{1.31}$$

where V_μ and $W^\pm_\mu$ are the field of the corresponding gauge bosons Z, γ, W^+, W^-, $W_{\mu\nu}$ is defined as $W_{\mu\nu} = \partial_\mu W_\nu - \partial_\nu W_\mu$ (similarly $V_{\mu\nu}$), and finally $\tilde{V}^{\mu\nu} = \frac{1}{2}\epsilon^{\mu\nu\rho\sigma} V_{\rho\sigma}$. The overall constant can be either $g_{WW\gamma} = -e$ or $g_{WWZ} = -e\ cot(\theta_W)$. In equation 1.31 the last four terms violate C and P conservation, moreover this Lagrangian is not gauge invariant for the $SU(2)_L \times U(1)_Y$ group.
The following five dimension-six operators, that are gauge-invariant, describe the electroweak effects of physics BSM:

$$\begin{aligned} O_{WWW} &= Tr[W_{\mu\nu} W^{\nu\rho} W^\mu_\rho] \\ O_W &= (D_\mu \phi)^\dagger W^{\mu\nu} (D_\nu \phi) \\ O_B &= (D_\mu \phi)^\dagger B^{\mu\nu} (D_\nu \phi) \\ O_{\tilde{W}WW} &= Tr[\tilde{W}_{\mu\nu} W^{\nu\rho} W^\mu_\rho] \\ O_{\tilde{W}} &= (D_\mu \phi)^\dagger \tilde{W}^{\mu\nu} (D_\nu \phi). \end{aligned} \tag{1.32}$$

The last two operators are those violating C and P conservation. Using these five operators it is possible to parameterize the leading effect of BSM physics on the electroweak vector boson self interaction. These operators have five corresponding

coefficients, called Wilson coefficients: $c_{WWW}, c_W, c_B, c_{\tilde{W}WW}, c_{\tilde{W}}$. The coefficients in the Lagrangian in equation 1.31 can be rewritten as a function of the Wilson coefficients, re-framing the anomalous couplings.

$$
\begin{aligned}
g_1^Z &= 1 + c_W \ \frac{m_Z^2}{2\Lambda^2} \\
\kappa_\gamma &= 1 + (c_W + c_B) \ \frac{m_W^2}{2\Lambda^2} \\
\kappa_Z &= 1 + (c_W - c_B tan^2(\theta_W)) \ \frac{m_Z^2}{2\Lambda^2} \\
\lambda_\gamma &= \lambda_Z = c_{WWW} \ \frac{3g^2 m_Z^2}{2\Lambda^2} \\
g_4^V &= g_5^V = 0 \\
\tilde{\kappa}_\gamma &= c_{\tilde{W}} \ \frac{m_Z^2}{2\Lambda^2} \\
\tilde{\kappa}_Z &= -c_{\tilde{W}} \ tan^2(\theta_W) \ \frac{m_Z^2}{2\Lambda^2} \\
\tilde{\Lambda}_\gamma &= \tilde{\Lambda}_Z = c_{\tilde{W}WW} \ \frac{3g^2 m_Z^2}{2\Lambda^2}.
\end{aligned}
\tag{1.33}
$$

It is then possible to invert these equations (1.33) to isolate the Wilson coefficient, and the expressions are valid considering anomalous couplings independent of energy. One of the conditions that an EFT should respect is the Unitarity Bound (UB) which is the request of the S-matrix to be unitary. The S-matrix relates the initial and final states of a scattering process. These UB enforce an upper bound on the partial wave amplitudes, and in the EFT the amplitude terms of dimension-six operators grow proportionally to s/Λ^2, which means that at high energy these terms can eventually violate the UB. It is possible to avoid this problem if one uses a vertex function approach instead of the Lagrangian method. The vertex approach considers the momentum-space analogue of the effective Lagrangian of the WWV vertex (1.31), obtaining a trilinear vector-boson vertex function [20, 25]:

$$
\begin{aligned}
\Gamma_V^{\alpha\beta\mu} = f_1^V (q-\bar{q})^\mu g^{\alpha\beta} - \frac{f_2^V}{m_W^2}(q-\bar{q})^\mu P^\alpha P^\beta + f_3^V (P^\alpha g^{\mu\beta} - P^\beta g^{\mu\alpha}) \\
+ i f_4^V (P^\alpha g^{\mu\beta} + P^\beta g^{\mu\alpha}) + i f_5^V \epsilon^{\mu\alpha\beta\rho}(q-\bar{q})_\rho \\
- f_6^V \epsilon^{\mu\alpha\beta\rho} P_\rho - \frac{f_7^V}{m_W^2}(q-\bar{q})^\mu \epsilon^{\alpha\beta\rho\sigma} P_\rho (q-\bar{q})_\sigma
\end{aligned}
\tag{1.34}
$$

where P, q and $\bar{q}$ are the four-momenta respectively of the bosons V, W^- and W^+, while the form factors f_i^V depend on P and are the analogues of the parameters in the Lagrangian 1.31. These form factors can be written, considering a series of requirements [26] that imply $f_1^\gamma(0) = 1$ and f_4^γ and f_5^γ proportional to P^2, as:

$$
\begin{aligned}
f_1^\gamma &= 1 + c_{WWW}\ \frac{3g^2P^2}{4\Lambda^2} \\
f_1^Z &= 1 + c_w\ \frac{m_Z^2}{\Lambda^2} - c_{WWW}\ \frac{3g^2P^2}{4\Lambda^2} \\
f_2^\gamma &= f_2^Z = c_{WWW}\ \frac{3g^2m_w^2}{2\Lambda^2} \\
f_3^\gamma &= 2 + (c_B + c_W)\ \frac{m_W^2}{2\Lambda^2} + c_{WWW}\ \frac{3g^2m_w^2}{2\Lambda^2} \\
f_3^Z &= 2 + [c_W(1 + cos^2(\theta_w)) - c_B sin^2(\theta_W)]\ \frac{m_Z^2}{2\Lambda^2} + c_{WWW}\ \frac{3g^2m_w^2}{2\Lambda^2} \\
f_4^\gamma &= f_5^Z = 0 \\
f_6^\gamma &= c_{\tilde{W}}\frac{m_W^2}{2\Lambda^2} - c_{\tilde{W}WW}\ \frac{3g^2m_w^2}{2\Lambda^2} \\
f_6^Z &= -c_{\tilde{W}}tan^2(\theta_W)\ \frac{m_W^2}{2\Lambda^2} - c_{\tilde{W}WW}\ \frac{3g^2m_w^2}{2\Lambda^2} \\
f_7^\gamma &= f_7^Z = -c_{\tilde{W}WW}\ \frac{3g^2m_w^2}{4\Lambda^2}.
\end{aligned}
\tag{1.35}
$$

Out of these form factors the first five ones, for both Z and γ bosons, are relative to terms which conserve C/P and depend on the three coefficients: c_{WWW}, c_W, c_B, while the last two coefficients (f_6^V and f_7^V) are relative to the terms that violate C/P and depend on the remaining two Wilson coefficients $c_{\tilde{W}WW}$ and $c_{\tilde{W}}$.
As mentioned previously, if one considers constant anomalous couplings, yields amplitudes grow proportional to $^s/_{m_W^2}$, meaning that for higher energies there will be, at a certain point, a UB violation. The form factors in 1.35 are regulated *ad hoc* in order not to exceed the UB condition. The condition to respect unitary bound can be restricted to the region where data are present, and since data respect this condition, an EFT that fits the data will consequently observe the UB.

1.3.2 Study of the Triple Gauge Couplings

Triple Gauge Couplings can be studied through diboson production, in particular in this thesis I consider W^+W^- production. The first measurements of W^+W^- production were carried at the LEP electron–positron collider [27] and at the Tetravon proton-antiproton collider [28, 29]. Multiboson production has been studied at LHC, both by the ATLAS [30] and the CMS [31] collaborations.
The SM production of W^+W^- happens mainly through three processes: the main one is the $q\bar{q}$ annihilation, the second one is the $gg \to W^+W^-$ process and the last one is the $H \to W^+W^-$ process. The second process occurs at higher order in perturbative QCD, and the last one is around ten times less probable than the other two. The first process mentioned, quark-antiquark annihilation, can generate as a result:

- W^+W^-, which is particularly important in testing EFT validity and to compute exclusion limits on WWV couplings. Studies of TGC via W^+W^- production have been done at the Tetravon [32] ad at LHC [33, 30]. Final states of this process can be either leptonic ($ll\nu\nu$) or semi-leptonic ($l\nu jj$);
- ZZ, whose production has been studied at LHC by the ATLAS [34] and CMS [35] collaborations. Final states of this process are four leptons or two leptons and two neutrinos. There are no TGC occurring with the production of ZZ in the SM;

- $W\gamma$ and $Z\gamma$, fundamental to find limits on TGC couplings, since they contribute to TGC generation. They have been studied in the leptonic final states, $W\gamma$ in the $l\nu\gamma$ and $Z\gamma$ in the $ll\gamma$ final state, at LHC by the ATLAS [36] and CMS [37] collaborations;
- WZ, that have been studied by ATLAS [38] and CMS [39] in the leptonic final state channel, and from which have been computed exclusion limits.

As stated above, in this thesis I study TGCs through the diboson production of W^+W^- generated by a quark-antiquark annihilation via the s-channel (thus not considering the t-channel nor the u-channel) and in particular the WWZ vertex. In the s-channel two particles interact with each other producing an intermediate particle that decays into two other new particles, as shown in figure 1.3.

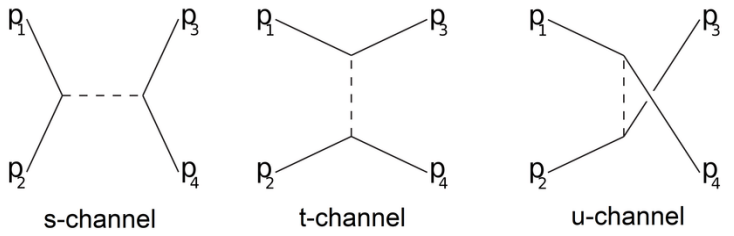


Figure 1.3: Representation of the three possible channels in a scattering process, the s-channel is the one used in this thesis [40, 41, 42].

The process $pp \to W^+W^-$ (shown in figure 1.4) is studied in the leptonic decay channel and I compute the confidence limits on the c_{WWW} Wilson coefficient. In this thesis analysis, the ATLAS study on WW production in [30] has been taken as a reference.

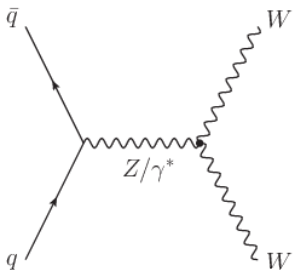


Figure 1.4: The s-channel of the $pp \to W^+W^-$ process [30].

Chapter 2

Physics at the LHC

The study of high energy physics has been requiring in the last decades experiments with an increasing energy and amount of data, in order to reach a better statistic precision. New developments in technologies and experimental tools helped to achieve these goals. In particular the Large Hadron Collider (LHC) was built as the latest part of the Cern complex accelerator. Colliders, as the LHC, accelerate particles beam to relativistic speed and make the beams collide to create particle interactions to study. In this chapter, an introduction to the LHC and to the ATLAS detector is given.

2.1 The Large Hadron Collider

The LHC is the largest and most powerful collider in the world [43], a two ring hadron accelerator, placed between Switzerland and France, near the city of Geneva. It has been built between 1998 and 2008 and is active since 2010. It is installed in an underground tunnel, that previously hosted the Large Electron-Positron Collider (LEP), with a radius of 27.6 km at a depth of around 100 m. The LHC can accelerate two counter-rotating proton beams up to an energy in the centre of mass frame ($\sqrt{s}$) of 14 TeV, which is now of 13 TeV, and an instantaneous luminosity that is currently of few 10^{34} cm^2 s^{-1}. The beams can also be made of heavy ions (Pb) that can be accelerated up to $\sqrt{s} = 5.5$ TeV. The whole tunnel is divided in eight straight sections and eight arches, and it has a complex system of superconducting magnets, refrigerated with super-fluid helium, which produce a field over 8 T, used to focus the beam.
The LHC is part of the Cern accelerator complex, an elaborate structure, as shown in figure 2.1. It enables to accelerate the particle beams and reach such a high energy in the centre of mass frame.
The protons are extracted from a hydrogen gas with an energy of 91 keV, and firstly injected in the LINAC2, which brings them to an energy of 50 MeV. After that the proton beam is brought to the Proton Synchrotron Booster (PSB) where the beam reaches the energy of 1.4 GeV. The beam arrives then in the Proton Synchrotron (PS) that make the protons reach $\sqrt{s} = 25$ GeV. Just before the LHC ring there is the Super Proton Synchrotron (SPS), a 6.9 km ring, that brings the beam energy to 450 GeV.

In figure 2.1 the four main experiments located in the LHC tunnel can be seen: ATLAS, CMS, ALICE and LHCb. These are found at the four points of collision of the proton beams. Nearby these interaction points there are also three other experiments: LHCf (Large Hadron Collider forward) [45], TOTEM (TOTal Elastic and diffractive cross section Measurement) [46], and MoEDAL (MOnopole and Exotical

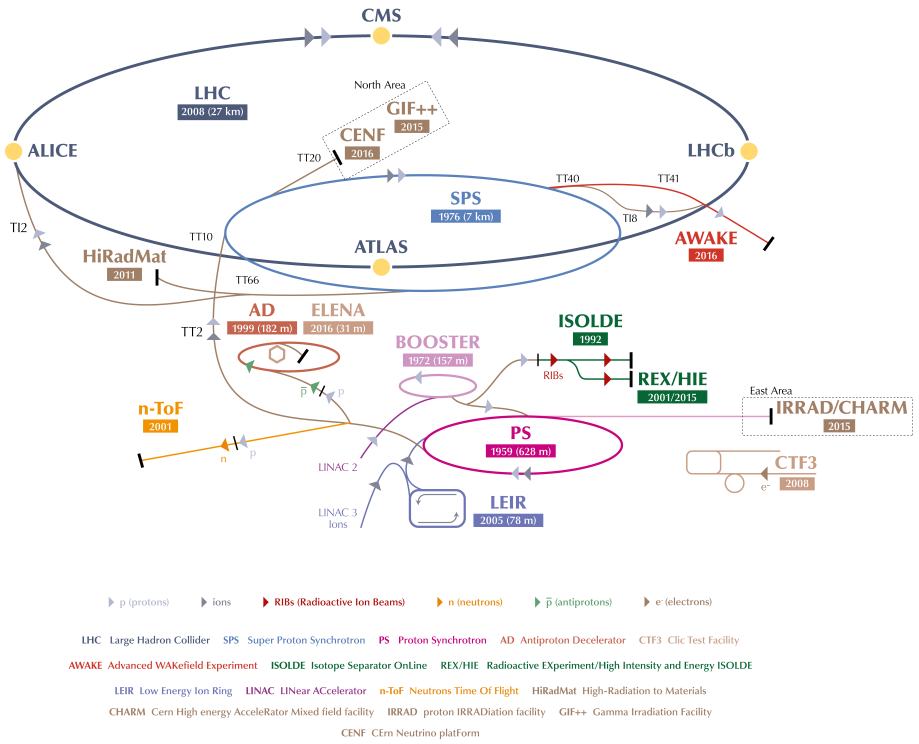


Figure 2.1: Sketch of the Cern accelerator complex[44].

Detector At the LHC) [47]. Here a brief description of the purpose of the four main LHC experiments is given:

- ATLAS (A Toroidal LHC ApparatuS) [48] is a general multi-purpose detector, designed to study as many possible events produced in pp collisions. Its main goal is to investigate SM physics and eventually also BSM physics;
- CMS (Compact Muon Solenoid) [49] as ATLAS is a multi-purpose detector and investigates the same phenomena. However, CMS is based on different technologies and materials and is also smaller than ATLAS;
- ALICE (A Large Ion Collider Experiment) [50] has the main purpose to study strong interactions between heavy ions in a high energy density regime. It also collects and studies data from pp collisions as a reference for the heavy-ion study;
- LHCb (Large Hadron Collider beauty) [51] studies heavy flavour physics, in particular hadrons containing bottom and charm quarks. It searches new possible sources of CP violation in the context of BSM physics.

At LHC each beam is split in bunches of protons with a time separation of 25 ns, which means a frequency of 40 MHz. The instantaneous luminosity is defined as in equation 2.1

$$L = {}^{dN}/_{dt}{}^{1}/_{\sigma} \tag{2.1}$$

where σ is the cross section of the process and ${}^{dN}/_{dt}$ is the number of events per second. This luminosity has reached a maximum of $L \sim 2 \times 10^{34}$ cm^2 s^{-1}. If we

consider two beams colliding, both of them with a Gaussian shape, the instantaneous luminosity can be written as

$$L = \frac{N_1 N_2 f N_b}{4\pi\sigma_x\sigma_y} \tag{2.2}$$

where N_1 and N_2 are the number of particles in each bunch, N_b is the number of bunches, f is the rate of collisions, and σ_x and σ_y are the bunch linear dimensions.

The ATLAS coordinate system is right-handed, with the z-axis parallel to the beam direction and the origin in the nominal interaction point. The x-axis is perpendicular to the beam direction pointing towards the centre of LHC, while the y-axis, also perpendicular to the beam, points upwards. The x-y plane, which is perpendicular to the beam direction, is called transverse plane. In order to describe events in the ATLAS detector, other variables can be introduced:

- $\phi = \cot\left(\frac{x}{y}\right)$ is the azimuthal angle measured around the z-axis;
- $\theta = \arctan\left(\frac{y}{z}\right)$ is the polar angle, measured from the z-axis;
- $y = \frac{1}{2}\ln\left(\frac{E+p_z}{E-p_Z}\right)$, where E is the energy and p_z the projection of the momentum on the z-axis, is the rapidity;
- $\eta = -\ln\ \tan\left(\frac{\theta}{2}\right)$ is the pseudo-rapidity, which can approximate the rapidity for highly relativistic particles;
- $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$ is the angular distance;
- $p_T = p\ \sin\theta$ is the transverse momentum, the projection of the particle momentum on the transverse plane;
- $E_T = E\ \sin\theta$ is the transverse energy, the projection of the particle energy on the transverse plane;
- E_T^{miss} is the missing transverse energy, which is the fraction of energy on the transverse plane that cannot be traced back to a specific reconstructed particle, but can be calculated thanks to the energy conservation on the transverse plane. The E_T^{miss} is used to reconstruct neutrinos at LHC, since this particles do not usually interact with the detector due to their particularly small cross section.

2.2 The ATLAS Detector

The ATLAS and CMS general purpose detectors investigate SM and BSM physics thanks to the events generated at LHC. In order to do so they need to have the following characteristics:

- fast components and electronics;
- highly efficient trigger that enable to downsize the amount of data stored, by rejecting background events;
- high detector granularity;
- large pseudo-rapidity acceptance and angular coverage;
- high level electromagnetic and hadronic calorimeters to have electron, photon, jets and E_T^{miss} measurements;

- good muon identification and measurements.

The ATLAS detector is the biggest detector at the LHC, 25m high and 44 m long, as shown in figure 2.2.

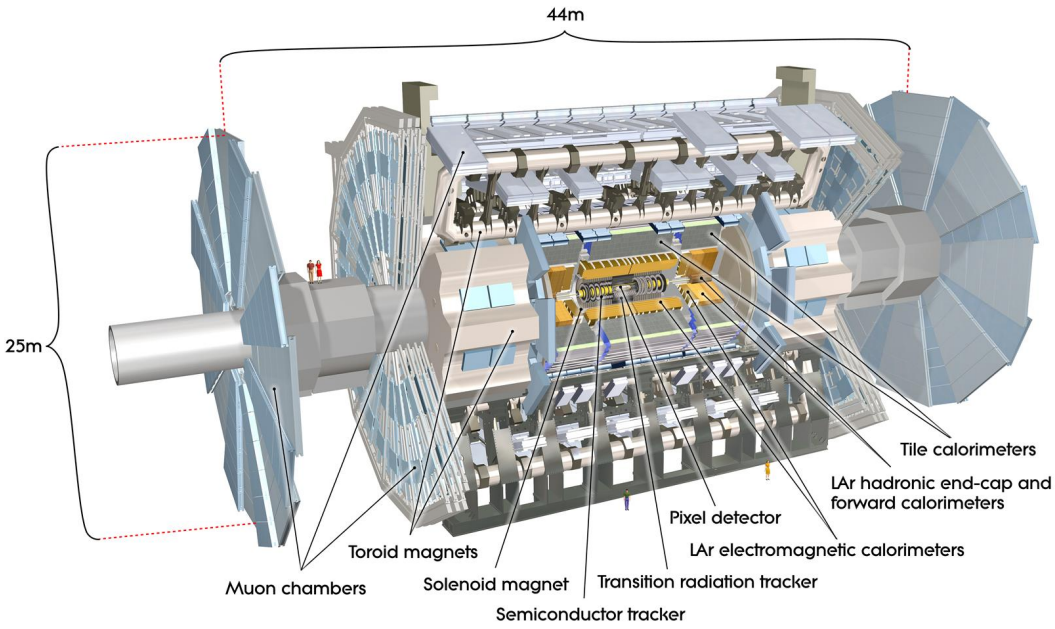


Figure 2.2: The ATLAS detector [48].

It has been built to measure the collision products across almost the full 4π solid angle, and thus it has a cylindrical structure along the beam line, centered at the interaction point. ATLAS is made of several sub-detector layers concentric along the beam axis with endcap regions perpendicular to the beam. The smallest and also closest to the interaction point is the Inner Detector (ID), which is surrounded by a solenoidal magnet that curve charged particle trajectories through a magnetic field. After that there is the Electromagnetic (EM) Calorimeter and then the Hadronic (Had) Calorimeter, both enclosed by the Muon Spectrometer (MS). Few additional and smaller forward detectors give information on the luminosity. In the next paragraphs a brief description of the sub-detectors is given.

2.2.1 The Inner Detector

The ID is 6.2 m long, has a diameter of 2.1 m and it is the innermost part of the ATLAS detector. Its main purpose is to reconstruct tracks of charged particles produced, measuring the p_T, the charge and the position of the particles and also the position of the primary and secondary vertex. This is possible also thanks to the 2 T axial magnetic field produced by the Central Solenoid (CS). The CS is made of a single-layer coil with a diameter of $\sim$ 2.5 m and 5.8 m long.

This sub-detector is made of three different sub-systems, shown in figure 2.3, and listed below:

- **the Pixel Detector** is a silicon detector built around the beryllium beam pipe. It is made of pixels, grouped in modules, each of 47232 pixels, arranged in three barrel layers and two end-caps, each made of three disks. The three layers are between 50 mm and 123 mm radius from the beam, while the end-cap

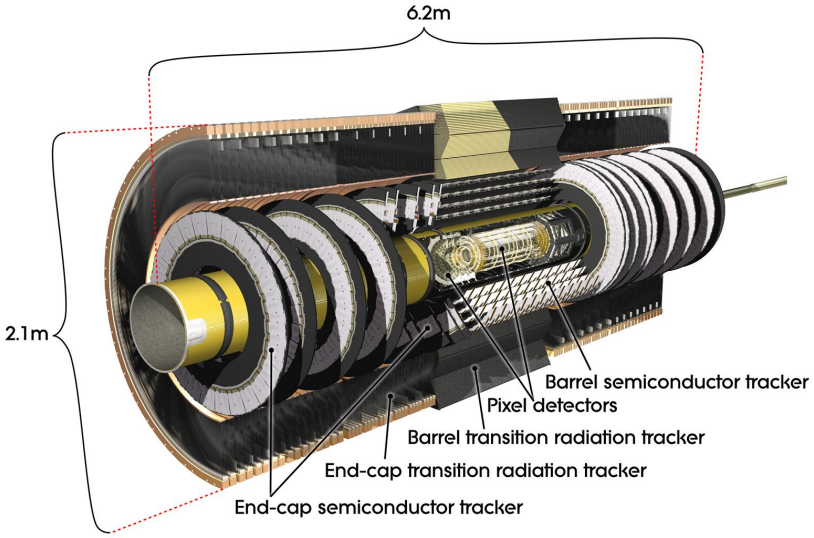


Figure 2.3: The Inner Detector [48].

disks are placed between $|z| = 495$ mm and 650 mm from the centre. Thus, the Pixel Detector gives as information three points for each track with $|\eta| < 2.5$;

- **the SemiConductor Tracker** (SCT) is made of silicon micro-strips with a pitch of 80 μm, grouped in 4088 modules. These are arranged in the barrel and two end-caps. The first one consists in four cylinders, almost 1.5 m long, with radius between $\sim$ 3 m and $\sim$ 5 m; the two end-caps are made of nine disks each at a distance from the center between $\sim$ 0.9 m and $\sim$ 2.7 m. Both the SCT and the Pixel Detector are refrigerated and maintained at a temperature of -7 $\hat{A}$°C. The SCT provides four space points for particles created by the beams interaction;

- **the Transition Radiation Tracker** (TRT), differently from the other two sub-systems, consists of gaseous drift tubes alternated with transition radiation material. The gas mixture is: 70% Xe, 27% CO_2 and 3%O_2 with an over-pressure between 5 and 10 mbar. It has an accuracy of 130 μm for particles with $|\eta| < 2$ and provides a continuous tracking with $\sim$36 hits per track. The TRT is also able to return information on Particle IDentification (PID) thanks to layers of straws interspersed with fibres or foils in the barrel and end-caps. This sub-detector operates at room temperature and is therefore insulated.

2.2.2 The Calorimeters

The calorimeter system, shown in figure 2.4, is designed to provide information on particles energy and position for events with $|\eta| < 4.9$ and to contain electromagnetic and hadronic showers. The system is composed of three different calorimeters: the Electromagnetic CALorimeter (ECAL), whose main purpose is precision measurements of electrons and photons, the Hadronic CALorimeter (HCAL), that, together with the Forward CALorimeter (FCAL), aims to reconstruct jets and measures the E_T^{miss}. This is possible thanks to the thickness of the active calorimeter $\sim 10\lambda$ (where λ is the interaction length) in the barrel and in the end-caps and around 11λ if the outer support is consider too.
The main characteristics of the three calorimeters are:

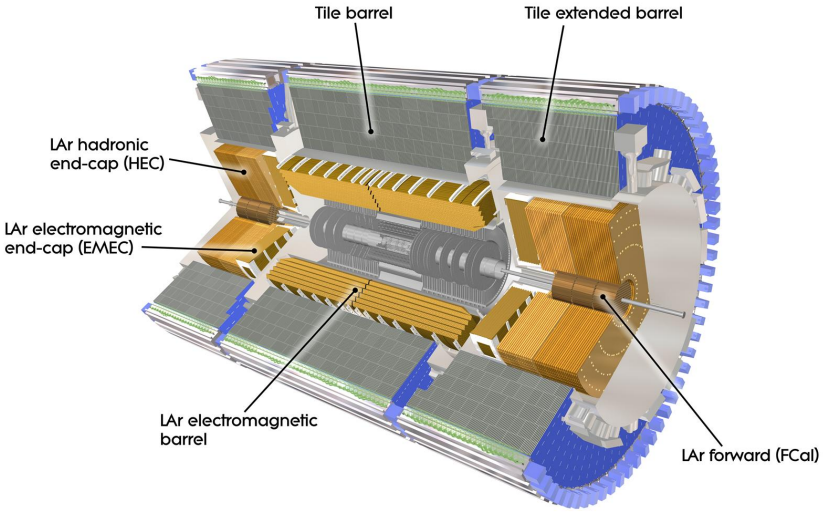


Figure 2.4: The Calorimeter System [48].

- the **ECAL** is composed by the barrel and the end-caps. The barrel calorimeter is made of two half-barrels that cover the range with $|\eta| < 1.475$ and are separated by 4 mm in the transverse plane. They have an inner diameter of ~ 2.8 m and the outer of ~ 4 m. Each half-barrel is divided in 16 modules, and the thickness of a single module varies from 22 X_0 at $\eta = 0$ to 33 X_0 at $|\eta| = 1.3$. Each end-cap is composed of two coaxial wheels, separated by 3 mm, and they cover respectively the regions $1.375 < |\eta| < 2.5$ and $2.5 < |\eta| < 3.2$. The outer wheel total thickness is between 24 X_0 and 38 X_0, while for the inner wheel it is between 26 X_0 and 36 X_0. The end-caps have an inner radius of ~ 0.3 m and an outer radius of ~ 2 m. Between the barrel and the end-caps there is the so called *crack region* ($1.37 < |\eta| < 1.52$), which has not good performance due to the presence of passive material in front of it;

- the **HCAL** is divided in two main parts: the barrel part is the TileCal (Tile Calorimeter) and the end-cap part is the HEC (Hadronic End-cap Calorimeter). The TileCal has steel absorber and scintillating tiles as active materials, it is composed of 64 modules and it is segmented in three layers. The TileCal is also composed of three sections, the central part and two extended barrels, which are linked to the central one with steel scintillator sandwiches. The central part cover the range $|\eta| < 1.0$, and the three layers are thichk ~ 1.5 λ, 4.1 λ and 1.8 λ, and it is 5.8 m long. The extended barrels cover the region $0.8 < |\eta| < 1.7$, they have three layers each with a thickness approximately of 1.5 λ, 2.6 λ and 3.3 λ and each extended barrel is 2.6 m long. Each HEC covers the region $1.5 < |\eta| < 3.2$ and is divided in two wheels that are found behind the ECAL end-cap. Each wheel is made of around 30 wedge-shaped modules, divided in two segments. The wheels and the modules are supported by a stainless-steel structure. The HEC is made of copper plated intervalled with 8.5 mm gaps filled with LAr;

- the **FCAL** covers the region $3.1 < |\eta| < 4.9$ and it is located in the end-caps cryostats. Each end-cap is composed of three modules: one made of copper and two made of tungsten. The copper one is optimised for electromagnetic measurements, while the tungsten ones are optimised for hadronic measure-

ments.

2.2.3 The Muon Spectrometer

The MS aims to reconstruct muons tracks and measure their p_T, and is also used to trigger on them. It is embedded in a magnetic field which curves the muon trajectory, provided, in the region with $|\eta| < 1.4$, by a large barrel toroid, in the region $1.6 < |\eta| < 2.7$, by two end-caps magnets and, in the region in between, by the combination of the fields. The barrel toroid is about 25 m long, it has a radius between 9.4 m and 20.1 m and it produces a field that varies from a minimum of 1.5 T×m to 5.5 T×m, while each end-cap toroid is around 5 m long, it has a radius between 1.65m and 10.7 m and it produces a field of about 1 to 7.5 T×m. Thanks to this magnetic field and to the knowledge of three different point of a muon track it is possible to determine the transverse momentum of the particle. Since the the toroidal magnet system in the MS is independent of the ID one, ATLAS can have two independent measurements of the muon p_T.
The MS is composed by different muon chambers, as shown in figure 2.5.

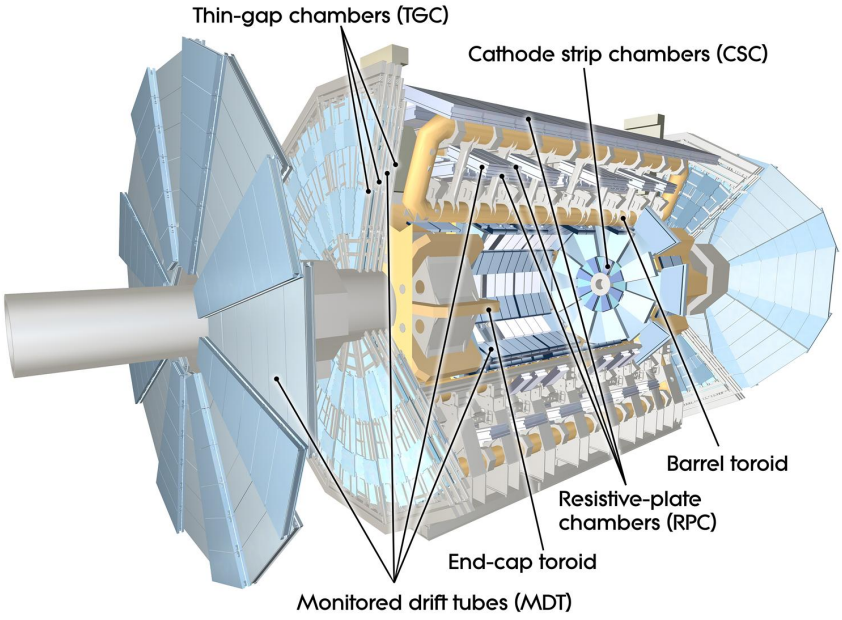


Figure 2.5: The Muon Spectrometer [48].

The Monitored Drift Tubes (MDT) measure the muon momentum in most of the η range, and is substituted by the Cathode Strip Chambers (CSC), that have higher granularity, only at large η and close to the interaction point. The trigger system has a range of $|\eta| < 2.4$ and is composed of Resistive Plate Chambers (RPC) in the barrel and Thin Gap Chambers (TGC) in the end-cap regions. The RPC cover the range $|\eta| < 1$ and are placed on three cylinders concentric to the beam, while the TGC cover the range $1 < |\eta| < 2.4$ and are positioned in four concentric disks. By the way, because of cable routing at $\eta = 0$, there is a gap. Overall this trigger system provides bunch-crossing identification, transverse momentum thresholds and measurements muon coordinates in the orthogonal direction with respect to that determined by the tracking chambers.

2.2.4 Luminosity Detectors

Three different detectors are used to measure the luminosity, a fundamental characteristic in every physics analysis. The first one measures the luminosity using Cerenkov Integrating Detector (**LUCID**), which detects inelastic *pp* scattering in the forward direction of the beam and it is located at ± 17 m from the interaction region. It produces luminosity information and is used as an online monitor by ATLAS, being also used to check beam losses before collisions happen. **ALFA** (Absolute Luminosity For ATLAS) is more distant from the interaction region, at around ± 240 m and detects elastic scattering at small angles. It is composed of scintillating fibre trackers that are close up to 1 mm to the beam. The third one is the **ZDC** (Zero-Degree Calorimeter), positioned at ± 140m measures neutral particles produced in the collisions in the region $|\eta| < 8.3$. It is made of layers of quartz rods alternated with tungsten plates. Finally there is also the Beam Conditions Monitor (BCM), which checks the stability of the particle beam.

2.2.5 Data Acquisition and Object Reconstruction

The amount of data produced by interactions at LHC is on average of 60 Terabytes per second, but only a fraction of these data are actually relevant for physics studies there is the need of a trigger system. ATLAS has a three level trigger [52] the **first level trigger** (L1) selects data based on energy depositions in the calorimeters and in the muon system, it searches for charged leptons and photons with high p_T and defines Regions of Interest (RoI) which are then passed to the next level trigger. The **second Level trigger** (L2) reduces the data flow to a few kHz in around few tens of ms. The L2 consider similar criteria to the L1 but with higher precision and granularity of the calorimeters and muon chambers considered. Finally, the **Events Filter** (EF), also called Third Level trigger ,selects the data in order to achieve an amount that can be handled offline, which for the Run II was of 1 kHz.

In order to have data that can be actually analysed to study SM and BSM processes, the various particles produced by the beam collision have to be reconstructed starting from the information collected by the detectors (object reconstruction). This is done by studying the type of interaction of a single event produced in the beam collision with the different detectors. In the case of an electron, for instance, it will leave a track passing through the ID and after that it will shower in the EM calorimeter. A neutral particle, instead, such as a neutron, will not interact with any detector until it will reach the hadronic calorimeter where it will create an hadronic shower. This way it is possible to identify almost uniquely every known particle. The object reconstruction is more difficult in case of particles with very small cross-section, as for the case of neutrinos, which can be partially reconstructed only via the transverse missing energy.
In the following paragraph, a brief description of the objects reconstruction strategy in ATLAS, for particles considered in this thesis work, is given:

- electrons identification is based on the shower shape, in particular on lateral and longitudinal variables from the ECAL layers, and on the energy leakage in the HCAL. In the region $|\eta| < 2.5$, where information on the track from the ID are available, electrons are selected based on how well the ECAL cell cluster matches with the ID track. Differently in the forward region with $2.5 < |\eta| < 4.9$, electrons are selected based on their shower shape against hadrons. ATLAS considers three main class of identification criteria: the "*Loose cuts*", the "*Medium cuts*", and the "*Tight cuts*". For instance, in the Run II for the *Tight* operating point the efficiency is minimum 55% at $E_T = 4.5$ GeV to a

maximum of 88% at $E_T = 100$ GeV [53]. Electrons reconstruction depends on their isolation, which gives a measure of how much activity is in the vicinity of the electron. ATLAS considers different kinds of isolation criteria: *Gradient* working points target a constant value of the isolation efficiency in η but dependent of E_T, *Fix* working points have fixed requirements and a changing isolation efficiency;

- muons reconstruction is based on information from the MS, the ECAL and the tracker. Two main strategies are defined: the "Stand Alone" which considers only the MS and the "Combined", that combines information from the MS and the ID. The identification of muons is achieved through four selection thresholds: *Loose*, *Medium*, *Tight* and *High-*p_T. Muon reconstruction efficiencies are measured close to 99% for Medium and Loose muons and in good agreement with MC predictions [54]. Muons have *Gradient* and *Fix* working point as for electrons;

- jets reconstruction is based on clusters, which are defined using calorimeter cells that are topologically connected. This is necessary to separate jet signals from multi-source background. The algorithm used by ATLAS to define clusters and reconstruct jets is provided by the FASTJET toolkit [55] and is called anti-k_T [56]. It relies on two variables, the first one being:

$$d_{i,j} = min\left(\frac{1}{p_{T_i}^2}, \frac{1}{p_{T_j}^2}\right) \times \frac{\Delta R_{ij}^2}{R} \tag{2.3}$$

  where i, j indicates two particles, p_T and ΔR have already been defined in 2.1, and R is the radius parameter, indicating the size of the jet, set for ATLAS to 0.4. The second variable $d_{i,B}$ represents the momentum space distance between the particle and the beam, $d_{i,B} = {}^1\!/_{p_{T_i}^2}$. The reconstructed jets are calibrated using the Jet Energy Scale. After that, jets undergo the Jet Vertex Tagging algorithm selection.

- missing transverse energy reconstruction is based on two separate contributions: one comes from the hard-event signals and the other from soft-events, that considers track not associated with any reconstructed object. It is defined as

$$\begin{aligned} E_T^{miss} &= \sqrt{(E_x^{miss})^2 + (E_y^{miss})^2} \\ E^{miss} &= E_e^{miss} + E_\mu^{miss} + E_{jet}^{miss} + E_{soft}^{miss} \end{aligned} \tag{2.4}$$

  where in the second equation each term is the opposite sum of the energies of all that type of object.

Chapter 3

Deep Neural Networks

When taking into consideration high-energy physics (HEP) problems and experiments, such as those studied at LHC and ATLAS, it is important to explore the nature of the data to be studied. LHC generates a vast number of collisions, each one with a high complexity: as a result, the data gathered are high dimensional and complex to be interpreted. Traditionally, they are analysed firstly with a selection of the data based on consecutive Boolean decisions, and in a second step through a statistical analysis. Most of the analyses considered, such as classification, regression problems, and hypothesis testing, are based on a statistical model $p(\vec{x}|\vec{\theta})$, which indicates the probability of observing $\vec{x}$ given the parameters $\vec{\theta}$. The problem is that the statistical model is not known explicitly in an analytical dorm, especially considering the high-dimensional space of the data collected by LHC. Even if one takes into account samples of simulated data that follow the physical laws of particles interactions, and determines the estimates with histograms or kernel-based density estimates, the solutions would demand almost impossible computational resources, because of the curse of dimensionality. It is then necessary to reduce the dimensionality of the data, which is traditionally made through different steps [57]. There is a big possibility to use machine learning (ML) techniques to improve the data reduction to select a lower dimensional space. ML have two different pros in this situation: not only there is a chance to obtain better selections and thus better final results, but it also reduces human intervention in the selection phase. In fact, traditionally the process of analyzing data requires a lot of efforts and time to develop algorithms, while when using ML techniques it is the computer that learns from the data and human intervention is smaller. In the next sections an introduction to ML techniques is discussed, together with deep neural networks and learning algorithms, posing particular emphasis to techniques and algorithms used in this thesis work.

3.1 Machine Learning

ML is a particular field of Artificial Intelligence. The name *machine learning* itself suggests what it is about: it is a class of algorithms that are able to make predictions and take decisions without knowing a priori the solution, but learning it from the data. There exist two main different classes of ML techniques: **supervised learning** algorithms, which are firstly trained on solved examples and after that are used to predict the outputs of similar unsolved data; and **unsupervised learning** techniques that group or select unsolved data based on similarities between features of the data. Unsupervised learning is often used to divide data or pre-process it, before other analysis. In this thesis work only supervised learning techniques are applied, thus only these will be further discussed.

Supervised learning algorithms are models that aim to find a relationship between a certain quantity to be predict, called *target*, and the features that are given as input. This way the algorithm can predict the output values for new data. Problems that can be solved with a supervised machine learning technique can be divided in two main groups: **classification** problems and **regression** problems. They differ on the type of output that the model wants to predict: in the first case the objective is to divide events in two (*binary classification*) or more (*multi-class classification*) classes, while for regression problems the aim is to approximate the real value of a certain quantity of interest. When considering binary classification the output variable is often defined as a real variable in the interval $[0, 1]$, and it indicates the probability of that event belonging to the class "0" or "1". In the following, particular attention is paid to the case of binary classification problems, being it the one considered in this thesis analysis.

In the **training phase** solved data, called *training set*, are presented to the algorithm, which tries to find the parameters values that best generalize the input data. Then the model should be able to predict the target even for *unseen data*. At this point the model generated is applied to other solved data without providing the algorithm the target values: this is called **testing phase** or validation phase. It is important in fact to evaluate the performance of the model, comparing the known target values to the ones predicted by the model. Finally, in the **application phase** the model is used to perform predictions on data whose target values are unknown. A set of input data can be represented as a matrix X, with N rows of observations $\vec{x}^{(i)}$. Each observation has m components or features that are given to solve the problem. Each observation has a corresponding label $\vec{y}^{(i)}$ which can be a scalar or have a certain dimension l, as shown in equation 3.1:

$$X = \begin{bmatrix} x_1^{(1)} & ... & x_m^1 \\ \vdots & \dots & \vdots \\ x_1^{(N)} & ... & x_m^{(N)} \end{bmatrix}, Y = \begin{bmatrix} y_1^{(1)} & ... & y_l^1 \\ \vdots & \dots & \vdots \\ y_1^{(N)} & ... & y_l^{(N)} \end{bmatrix} \tag{3.1}$$

The training set size can affect the performance of the algorithm: it highly depends on the complexity of the problem to be solved, on the specific algorithm used and on the noise of the data. Most tasks in HEP can be reformulated as machine learning problems: as a search for a certain function f that maps the input features to a low-dimensional space of a desired target

$$f : X \to Y. \tag{3.2}$$

The function f is chosen through the optimization of a certain metric $\mathcal{L}(f(X), Y)$ called *loss function*, which returns a measurement of how good $f(X)$ predicts the expected output Y. The loss function can be chosen by the user among different ones, and is calculated on all points of the training set. Some of the most commonly used functions for binary classification problems in machine learning are:

- the **cross-entropy**, often called **binary cross-entropy** in this field, that uses target values in the interval [0,1]. It calculates a score of the average difference between the actual and the predicted probability distributions for predicting the class "1" and it minimizes this score. The loss function is defined as:

$$l(q) = -\frac{1}{N} \sum_{i=1}^{N} y_i \ \log(p(y_i)) + (1 - y_i) \ \log(1 - p(y_i)) \tag{3.3}$$

  where N is the number of events, y_i is the label of the target value, and $p(y_i)$ is the predicted probability of the event target value being 1 for all the N events;

- the **squared hinge** loss, that computes the square value of the core of the hinge function. The hinge loss uses a target value in the range [-1,1] and it is defined as:

$$l = max(0,\ 1 - y_i(x_i - b)) \tag{3.4}$$

where y_i is the label of the target value, x_i is the output of the classifier decision, and b is a possible bias term;

- the **accuracy** that is the fraction of correctly labeled examples out of the total number of examples. This metric does not return a good evaluation of the classifier performances when the data are particularly unbalanced;

- the **Area Under the Curve** (AUC), that is the value of the area under the Receiver Operating Characteristic (ROC) curve. It is computed considering the True Positive Rate (TPR) versus the False Positive Rate (FPR) curve. These two indexes are defined considering two cases: *positives* that belong to the class with less elements, and *negatives* which are part of the more numerous class. The TPR and FPR are defined as:

$$\begin{aligned} TPR &= \frac{TP}{P} = \frac{P - FN}{P} = 1 - \frac{FN}{P} = 1 - FNR \\ FPR &= \frac{FP}{N} = \frac{N - TN}{N} = 1 - \frac{TN}{N} = 1 - TNR \end{aligned} \tag{3.5}$$

where P (N) indicates the number of elements in the *positive* (*negative*) class, TP is the number of true positives (positive cases correctly labeled), TN is the number of true negatives (negative cases correctly labeled), FN is the number of false negatives (positive cases incorrectly labeled as negative), and FP is the number of false positive (negative cases incorrectly labeled as positive).

3.2 Neural Networks

Neural Networks are an important class of machine learning techniques, widely used in many different fields, which are inspired by the human brain neurons. In a very simplified description of the biological network, each neuron receives electric signals from about a 1000 of cells, it elaborates the information and outputs a signal through synapses when the stimulus is above a certain threshold. McCulloch and Pitts in 1943 provided a mathematical description of this kind of network [58]: each neuron receives m inputs x_i with a corresponding weight w_i and the neuron provides an output signal different from zero only when the weighted sum of the inputs is above a threshold θ_0:

$$\hat{y} = \Theta\left(\sum_{i=1}^{m} x_i w_i + w_0\right) \tag{3.6}$$

where $\Theta(z)$ is the Heaviside step function that reproduces the two possible states of the neuron: *at rest* for $\Theta(z) = 0$ or *firing* when $\Theta(z) = 1$. The Heaviside function is the simplest possible choice and in fact it shows some limits when used in a model to learn complex tasks. In general in a learning algorithm small variations in the parameters should bring to small changes in the output. It is useful to replace the Heaviside function with a non-linear **activation function** $g(z)$. Equation 3.6 can be written as in 3.7 by introducing the activation function and the input and weight vectors (respectively $\vec{X}$ and $\vec{W}$):

$$\hat{y} = g\left(\vec{X}^T\vec{W} + w_0\right) = g(z) \tag{3.7}$$

where z from now on indicates the linear combination of the inputs. In practical tasks the target's domain can be bounded or unbounded and the activation function must be chosen consequently.

There exist many different activation functions: some of the most relevant ones are discussed in the following. They generally have a simple expression for the first derivative, since this is a useful property when applying an optimization algorithm, as explained below in 3.3. The **sigmoid** function

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3.8}$$

is bound between 0 and 1 and often used for classification purposes, in fact its finite output can be related to the probability of the input to belong to a certain class or not. The sigmoid function also has a simple first derivative expression that is a linear combination of the sigmoid itself:

$$\sigma'(z) = \sigma(z)[1 - \sigma(z)]. \tag{3.9}$$

On the other hand, this activation function suffers of a saturation issue: as $|z| \to \text{inf}$ then $\sigma'(z) \to 0$, which can lead to rapidly vanishing gradients. A different activation function, that does not suffer from the same issue, is the **Rectified Linear Unit** (**ReLU**) defined as:

$$g(z) = \begin{cases} z, & z > 0 \\ 0, & z \leq 0 \end{cases} \tag{3.10}$$

$$g'(z) = \begin{cases} 1, & z > 0 \\ 0, & z \leq 0 \end{cases} . \tag{3.11}$$

The first derivative of ReLU activation function (see 3.11) has an extremely basic expression and therefore is particularly quick to compute. A more complex activation function is the **Exponential Linear Unit** (**ELU**) defined as:

$$g(z) = \begin{cases} z, & z > 0 \\ \alpha(e^z - 1), & z \leq 0 \end{cases} \tag{3.12}$$

$$g'(z) = \begin{cases} 1, & z > 0 \\ \alpha e^z, & z \leq 0 \end{cases} . \tag{3.13}$$

A neural network is composed of multiple neurons, called nodes, that are simple functions. The number of neurons, the activation function and the kind of connections among the nodes are referred to as the network architecture. Neurons are organized into layers. The first one, that receive external data, is called *input layer*, while the last one, that produces the final output, is called *output layer*. All neurons in a layer receive the same inputs and apply the same activation function, but each of them uses a different set of weights w_i. Neural networks can be divided into two main groups, based on the kind of connections between layers [59]:

- *feed-forward* neural networks, where information passes from the input layer to internal neuron layers to the output layer;

- *recurrent* neural networks, where information is propagated forward and backwards and therefore the connection between neuron nodes creates a circle.

The simplest neural network is the single-layer perceptron, which is composed only by the input and the output layer. Neural networks with multiple layers are called **multi-layer perceptrons** and they tend to outperform the single layer ones [60]. In fact, single-layer perceptrons can combine input data only once and therefore cannot learn high level patterns. In a multi-layer perceptron the layers between the input and the output one are referred to as **hidden layers**. The most common kind of multi-layer perceptron is the one with *dense* layers, where all neurons of one layer are connected to all nodes of the previous and next layers.

A simple multi-layer perceptron structure is one with a single hidden layer as presented in picture 3.1.

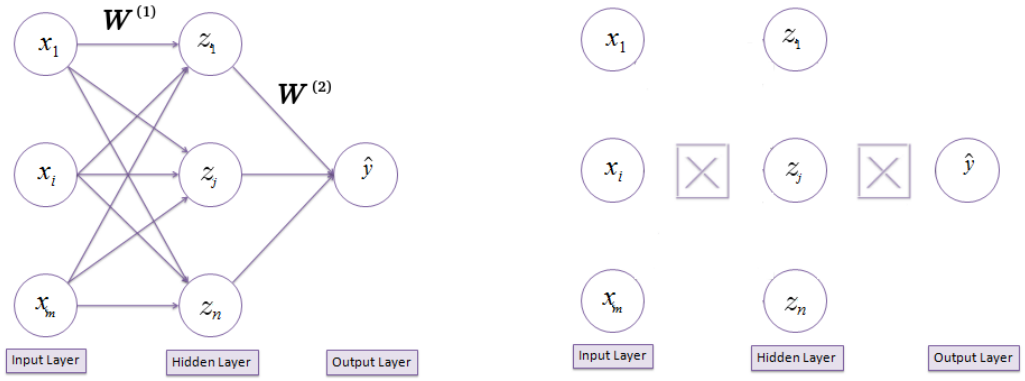


Figure 3.1: Sketch of a single-layer perceptron with a detailed notation in the left picture and the same neural network with a more compact notation in the right one.

This neural network has two weight matrices $W^{(1)}$ and $W^{(2)}$. The first one operates on the input features x_i and produces the z_i:

$$z_i = w_{o,i}^{(1)} + \sum_{j=1}^{m} x_j w_{j,i}^{(1)} = w_{o,i}^{(1)} + (\vec{X}^T \; W^{(1)})_i. \tag{3.14}$$

These values z_i are then passed to the application of the activation function of the hidden neurons $g(z_i)$, and the results obtained are used as input for the output layer. The second matrix $W^{(2)}$ is now applied to the $g(z_i)$ and its result is the input for the activation function $g^{(2)}$ of the output neuron namely:

$$\hat{y} = g^{(2)} \left(w_{0,i}^{(2)} + \sum_{j=1}^{n} g(z_j) w_{j,i}^{(2)} \right). \tag{3.15}$$

3.2.1 Deep Neural Networks

Deep Neural Networks (DNNs) are particularly large neural networks able to handle higher dimensional problems. They are arranged in many hidden layers that all contribute to elaborate data before the final prediction. These inspired the term *deep learning* (DL). Some DNN models have up to dozens of layers and hundreds of millions of parameters, and their training uses even more examples. A sketch of a DNN is shown in figure 3.2.

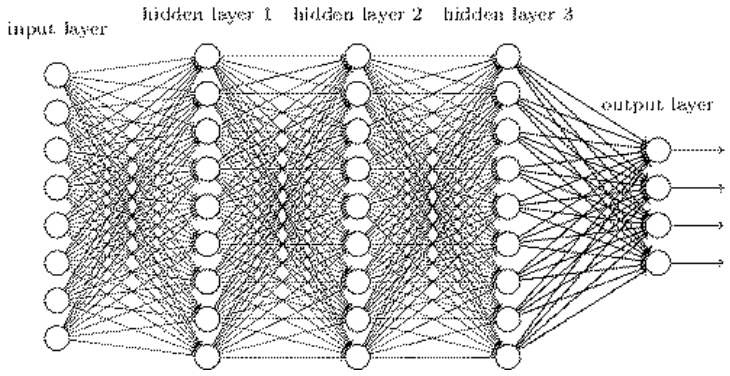


Figure 3.2: Sketch of a DNN with three dense hidden layers [61].

When using ML models it is often necessary to pre-process the data: the user devises variables that are sensible for the specific case of application combining the raw data features. On the other hand deep learning, thanks to its hierarchical structure, can learn such high level of abstraction on its own [62]. In the last decade DNNs have solved a wide variety of problems in different fields: an interesting example in high energy physics (HEP) is the application of DNNs to the classification problem of separating exotic particles from the background as e.g. in the paper [63], which compares the use of DNNs and raw data to more shallow techniques [63] with high-level features in the case of heavy Higgs bosons and SUSY particles.

If one considers a DNN with L hidden layers, each one with its activation function $g^{(l)}$, then the network algorithm can be written as a composite map:

$$\hat{Y}(X) = \left(g^{(L)} \circ ... \circ g^{(1)}\right)(X) \tag{3.16}$$

$$g^{(l)}(z) = g^{(l)}(W^{(l)}z + w_0^{(l)}) \tag{3.17}$$

where equation 3.17 shows the activation function for each layer l in $[0, L]$.

3.3 Algorithms

A deep neural network learning process is based on an optimization problem. The objective is to find the vector of parameters $\vec{\theta}$ that minimizes or maximizes the loss function $\mathcal{L}$. The problem being of minimization or depends on the nature of the loss function: they two cases are totally similar, thus in this section only minimization is treated. An **optimizer** is an optimization method that finds an approximate solution to the minimization problem, that in most cases is not analytically solvable. Optimizers can be grouped in three categories: *zero-order* methods use only information from the value of the loss function $\mathcal{L}(\theta)$, *first-order* methods use also the first derivative $\mathcal{L}'(\theta)$, and *second-order* methods require the Hessian of the loss function $\mathcal{L}''(\theta)$ too.

The **gradient descent** (GD) optimization is a first-order method based on the search of the optimal set of parameters θ following the opposite direction to that indicated by the gradient of the loss function $\mathcal{L}$. The gradient of a function in a certain point, in fact, shows the direction of fastest local ascent, and it can be written for the loss function as:

$$\nabla_\theta \mathcal{L} = \left(\frac{\partial \mathcal{L}}{\partial \theta_1}, \ldots, \frac{\partial \mathcal{L}}{\partial \theta_n}\right). \tag{3.18}$$

If one considers small θ variations, then the loss variation can be approximated as:

$$\Delta\mathcal{L}(\theta) \simeq \frac{\partial\mathcal{L}}{\partial\theta_1}\Delta\theta_1 + \ldots + \frac{\partial\mathcal{L}}{\partial\theta_n}\Delta\theta_n = \nabla_\theta\mathcal{L}\cdot\Delta\theta \tag{3.19}$$

where $\Delta\theta = (\Delta\theta_1, \ldots, \Delta\theta_n)$. In order for the loss function to decrease, the variation $\Delta\theta$ must be negative and it is convenient to define it as:

$$\Delta\theta = -\eta\nabla_\theta\mathcal{L}. \tag{3.20}$$

Here η is a positive parameter called **learning rate**. It can be proved that this definition of $\Delta\theta$ makes $|\Delta\mathcal{L}|$ maximum. The negative loss function variation can thus be written as:

$$\Delta\mathcal{L} = -\eta||\nabla_\theta\mathcal{L}||^2 < 0. \tag{3.21}$$

The solution for the optimization problem, at this point, is found iteratively: starting from a random value of θ_0 at first, then using equation 3.20 to compute the new value of the parameters for the next step. The loop stops when convergence is found over a certain number of iterations. Not all examples are used in one step: indeed often one or a small number of them are used for each iteration, as discussed in section 3.3.1. The DNN is trained over several **epochs**, that is when all the examples have been used once.

The learning rate (LR) determines how large the step in the gradient descent is. It can be fixed, but it is usually chosen so that it changes at each iteration. If one considers a fixed LR two opposite situations can happen (as shown in figure A.4): the LR is too small, therefore too many iterations are needed to converge and the optimization can be blocked in a local minimum; the LR is too big, which can prevent the algorithm to converge at all.

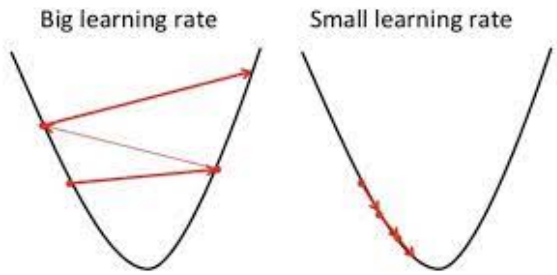


Figure 3.3: The left graph shows the case of a too large LR that does not make the loss function to converge to the minimum. On the right the case of a too small LR, which takes too much time to be trained [64].

There exist many different ways to consider an *adaptive* LR, that is a learning rate that varies at each iteration. In the following paragraph the Adam optimizer, that is used in the analysis of this thesis, is discussed.

The Adaptive Moment Estimation (**Adam**) [65] is an optimizer that updates the LR value at each step multiplying it by a certain factor, which depends on the running average of both the gradients (m) and the second moments (v) of the gradients. These two quantities are defined as:

$$m_\theta^{(t+1)} = \beta_1 m_\theta^{(t)} + (1-\beta_1)\nabla_\theta\mathcal{L}^{(t)} \tag{3.22}$$

$$v_\theta^{(t+1)} = \beta_2 v_\theta^{(t)} + (1-\beta_2)(\nabla_\theta\mathcal{L}^{(t)})^2 \tag{3.23}$$

where t is the index of the training iteration, β_1 and β_2 are the forgetting factors. The two moment estimates m and v are biased, therefore the corrected estimators $\hat{m}$ and $\hat{v}$ are computed:

$$\hat{m_\theta} = \frac{m_\theta^{(t+1)}}{1 - \beta_1^{(t+1)}} \tag{3.24}$$

$$\hat{v_\theta} = \frac{v_\theta^{(t+1)}}{1 - \beta_2^{(t+1)}} \ . \tag{3.25}$$

Finally the parameters can be updated as in equation 3.26.

$$\theta^{(t+1)} = \theta^{(t)} - \eta \frac{\hat{m}_\theta}{\sqrt{\hat{v}_\theta} + \epsilon} \tag{3.26}$$

where ϵ is a small scalar (ie. 10^{-8}) used to prevent division by 0. Even when using an adaptive LR as Adam, the optimizer requires an initial value for the LR, that must be decided by the user.

3.3.1 Stochastic Gradient Descent

The **Stochastic Gradient Descent** (SGD) is a variation of the GD algorithm. The GD computes the loss function and updates the model for each example in the training dataset, whereas the stochastic mechanism randomly chooses a single example in the training dataset and gets a rough estimation of the loss function gradient. At each iteration a different example is picked. The stochastic mechanism results in a very noisy approximation of the gradient with respect to the GD algorithm; on the other hand it is very fast and efficient. In fact, the gradient descent requires at each iteration to compute the partial derivative of the loss function for all the parameters of the model and for all the examples of the training datasets. Both the number of parameters and the number of examples can be huge in DNNs, of the order of hundreds of millions. Therefore at every update of the model the computer needs to keep saved all parameters values at runtime. In order to solve this problem the SGD calculates an approximation of the gradient which is less computationally expensive and more data efficient. The **mini-batch SGD** is in between the SGD and the GD methods. It splits the training dataset into small batches that are used to calculate the model loss function and to update the parameters. This compromise between the two extremes computes an unbiased, variable estimate of the full gradient, while reducing the variance compared to the single-sample approximation [66]. The **batch size** is a parameter that the user must choose and that defines the level of approximation in the gradient computation. The final gradient is obtained as the average of the gradient computed on the examples in the batch selected.

Chapter 4

Data Sample Simulation

Monte Carlo (MC) generators allow to simulate real collision data, given a certain model. They are fundamental tools for nowadays data analysis in HEP, useful to make predictions on complex final states from high energy particle collision events. These predictions can be used to test the goodness of a particular physics model (e.g. under a certain new physics hypothesis) by comparing simulated data with real data collected by experiments. Furthermore it is possible to study the performance of a detector through a simulation of the interaction of generated particles with different components of the detector itself.

The main idea of a MC generator is to produce a user-defined number N of points in the phase space of the considered process with a density based on the probability distribution f expected for that process. If we consider the trivial case of a single variable x in the interval $[x_{min}, x_{max}]$ with a probability distribution proportional to $f(x) \geqslant 0$, then it exists a pseudo-random number r in the interval [0,1], so that it is true:

$$\int_{x_{min}}^{x} f(x')dx' = r \int_{x_{min}}^{x_{max}} f(x')dx'. \tag{4.1}$$

By solving the expression, one finds that the analytical form of the indefinite integral $F(x)$ of the function $f(x)$ is:

$$F(x) = rF(x_{max}) + (1 - r)F(x_{min}) \tag{4.2}$$

If F^{-1} exists, it is possible to isolate x by inverting the equation 4.2 so that:

$$x = F^{-1}(rF(x_{max}) + (1 - r)F(x_{min})). \tag{4.3}$$

At this point one can generate a set R of pseudo-random uniformly distributed numbers in [0,1] and obtain a set of numbers distributed as $f(x)$ by using equation 4.3 on elements of R. If F^{-1} does not exist, a solution can be found numerically, since $f(x)$ is positive and thus $F(x)$ is monotone.

If the indefinite integral of $f(x)$ is undefined, the problem can be solved with a different method, called *Acceptance-Rejection* method. One considers a function $g(x)$ with a known indefinite integral $G(x)$ such that $g(x) \geqslant f(x)$ in the interval $[x_{min}, x_{max}]$. A set of random points distributed as $g(x)$ is generated, in the interval $[x_{min}, x_{max}]$ with the technique explained above for a density function whose indefinite integral is known. Each generated point is accepted if $f(x) \geqslant r'g(x)$, with r' being a pseudo-random number uniformly distributed in [0,1]. In this way, points are accepted with probability $f(x)/g(x)$. The set of points accepted is distributed as the desired probability density function f.

At the LHC the situation is more complicated: ones usually needs to find the expectation value of a certain observable O which is a function of the momenta of the n final-state particles. The expectation value can then be written as an integral in the phase space of the momenta, which has $3n$-4 dimensions (three components of momentum and four constraints for the energy-momentum conservation). In general the functions f considered cannot be analytically integrated. The MC method is then based on the idea of an integral as the average of the function f. Let's consider the integral:

$$I = \int_V d^m x f(\vec{x}) \tag{4.4}$$

where V is a region in the $\vec{x}$ space and $\vec{x}$ is a vector with m components. If one considers N pseudo-random uniformly distributed points in the region V, the mean value of f on those points, for the central limit theorem, is an unbiased estimator of the integral and the error can be evaluated as the variance of f. Therefore:

$$I[f] \simeq |V| < f >= |V| \frac{1}{M} \sum_{i=1}^{M} f(\vec{X}_i \tag{4.5}$$

$$E[f] = \sqrt{\frac{Var(f)}{M-1}}. \tag{4.6}$$

At this point one can compute the expectation value of a certain observable O, that can be related to measurable physical quantities in HEP (e.g. the cross section of a process), starting from a set of pseudo-random uniformly distributed points.

4.1 Monte Carlo Generators

In HEP experiments MC generators produce final states of particle collisions, following six main steps: the hard process, the parton shower, the hadronisation, the underlying event reconstruction, the decays of unstable particles, and the detector simulation. In figure 4.1 a sketch of a typical proton-proton collision is shown.

In this thesis proton-proton collisions are simulated with three Monte Carlo simulation tools, MadGraph5_aMC@NLO [68] interfaced with Pythia8 [69] and Delphes [70]. In the following a brief introduction of these three generators is given:

- **MadGraph5_aMC@NLO** is used here to generate the parton-level hard processes. The hard scattering is simulated with parton distribution functions (PDFs) to describe how partons join the process. MadGraph5_aMC@NLO is a framework able to perform LO and NLO computations. It is the result of the unification of MadGraph5 and aMC@NLO. Its structure has two main sections: the first one contains the theory model of the interaction Lagrangian and the related parameters (couplings and masses), the other section considers process-independent building blocks. MadGraph5_aMC@NLO can deal with user-defined Lagrangian at LO using the Mathematica package *FeynRules* [71] and then compute the matrix elements of the process and edit the process building blocks consequently. The framework uses tree-level techniques to construct LO and NLO matrix elements and is able to evaluate them quickly. In the special case of SM processes matrix elements can be computed at NLO for QCD and EW processes. In the output phase the information is transferred on physical memory. At this point the user can customize different characteristics

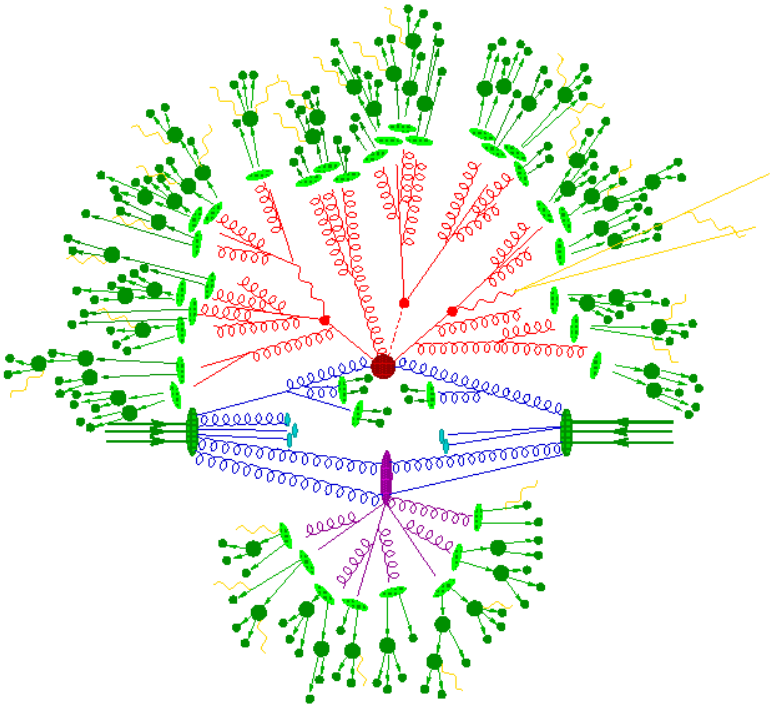


Figure 4.1: Schematic representation of a hard collision event as produced by a MC simulation [67].

of the process generation such as: choose the shower software, choose the detector simulator software, and use the MadSpin module to decay on-shell particles. Finally, in the running stage different tasks are performed such as the production of events;

- **Pythia 8** is a tool for the simulation of interactions between elementary particles at high centre of mass energy collisions. It is used in this work for the parton shower and the hadronization simulation. In the parton shower phase, coloured particles interact and decay via QCD radiative processes, until energies are so low that perturbation theory breaks. In the hadronization phase, clusters of partons are confined in hadrons. Additional simultaneous parton-parton collisions, beside the hard interaction, can happen because of the space-time extent of the proton bunches (causing the so-called pile-up effect) and, as a consequence, of multiple parton-parton interaction in the same protons, giving raise to the hard scattering (called underlying event); both effects are simulated by Pythia 8. Finally, since most particles are not stable, their decays are simulated: resonance decays are considered sequentially as part of the hard process, while hadron and lepton decays are performed isotropically after the hadronization. Pythia 8 code is written in C++ and is meant to be compatible with many other frameworks, for instance it can take inputs from matrix-element based tools, such as MadGraph5_aMC@NLO. The final state particle in Pythia are produced in vacuum, thus it does not consider interactions with the detector material. Pythia gives the possibility to change its parameters independently of each other and several Pythia tunes are available. The hadronization strategy is a linear confinement model: as the partons move apart they are considered as colour flux tubes stretched between q and $\bar{q}$.

- **Delphes** is a fast detector simulator which reproduces a cylindrical-symmetric

machine. Delphes converts input events in a set of universal objects with a module called *reader*, and then processes them with a sequence of modules, as shown in figure 4.2. These objects are then stored in Root TTree objects that can be directly analysed with the Root framework [72]. Delphes simulates three

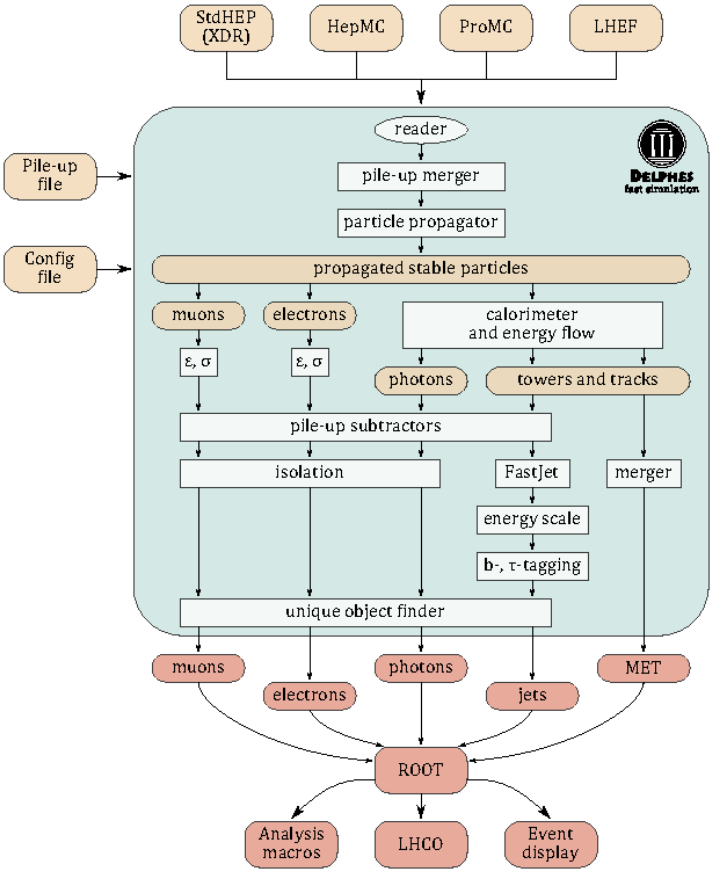


Figure 4.2: Typical work-flow chart of the Delphes fast simulation.

main components of a detector, the first one is the tracker with a magnetic field parallel to the beam direction. It is assumed that particles propagating in the tracker are reconstructed with a perfect angular resolution while smearing is used for the transverse momentum. This last one requires the information from the electromagnetic and hadron calorimeters, with finite pseudo-rapidity and azimuthal range, made by cell grid; finally the muon detection is performed through the simulation of a muon system.

Chapter 5

Data Analysis

In this chapter I investigate the potential improvements that could be obtained by implementing a DNN in a HEP specific classification problem. As an interesting case study, the search for anomalous TGCs in the WWZ vertex is considered. Here I present the strategy used to extract the corresponding exclusion limits on the c_{WWW} Wilson coefficients and I highlight the DNN analysis effectiveness by comparing the results with those derived by a traditional sequential cut analysis.

In section 5.1 I illustrate the tools and techniques which I used to generate the Monte Carlo signal and background samples. I assume here to perform a search for W^+W^- production in proton-proton collisions at a centre of mass energy of $\sqrt{s} = 13$ TeV with the ATLAS experiment, in two possible integrated luminosity scenarios: either $L = 36.1\text{fb}^{-1}$ or $L = 139\text{fb}^{-1}$. This allows me to compare the analysis results with two reference published results [30] ,[73]. The decay channel I studied is therefore the dileptonic one $W^+W^- \to e^{\pm}\nu_e\mu^{\mp}\nu_\mu$ where only events with one electron and one muon of opposite charges are considered. To compare the simulated data to real data from articles [30] and [73] more faithfully, I study and customize Delphes parameters in section 5.2.

A sequential cut analysis is performed in section 5.3. This analysis obtains a a signal region enriched in genuine TGC events out of the global phase space of all the final state particles [30].

In section 5.4, I implement a DNN analysis on events that passed the selections. I also study and optimize the structure and hyper-parameters of the DNN and choose the architecture that shows best performances. In section 5.5 I use the DNN selected to analyze data and obtain the DNN output distribution for signal and background events. This distribution is compared with the leading lepton p_T distribution, that is the one chosen in section 5.3 as the most sensitive to the c_{WWW} EFT coefficient. It is then possible to compare the separation between signal and background, obtained with the sequential cut analysis and the DNN. Finally, in section 5.6 the ${}^{c_{WWW}}/_{\Lambda^2}$ 95% confidence interval[1] is extracted by implementing a least squares estimator on two different distributions: the DNN output variable and the leading lepton p_T. It is then possible to quantify the level of improvement achieved by the DNN by comparing its performances to those of the traditional sequential cut analysis.

5.1 Simulated Data

Using the software tools presented in chapter 4 I generated the Monte Carlo signal sample of diboson production via EFT anomalous Triple Gauge Couplings with the

[1] From now on the parameter ${}^{c_{WWW}}/_{\Lambda^2}$ is referred to as c_{WWW} for notation compactness.

Wilson coefficient $c_{WWW} = 1$ TeV^{-2}. The decay channel used in this thesis is the dileptonic:

$$pp \to W^+W^- \to e^\pm \nu_e \mu^\mp \nu_\mu.$$

The SM contributions to the same leptonic final state are: SM diboson, SM $t\bar{t}$, Drell-Yan, top quark pair, Wt, W+jet and multi-boson production. Among these contributions, the main ones are the SM W^+W^- production with the same process as for the signal and the top quark pair production:

$$pp \to t\bar{t} \to W^+ b W^- \bar{b} \to e^\pm \nu_e \mu^\mp \nu_\mu b\bar{b}.$$

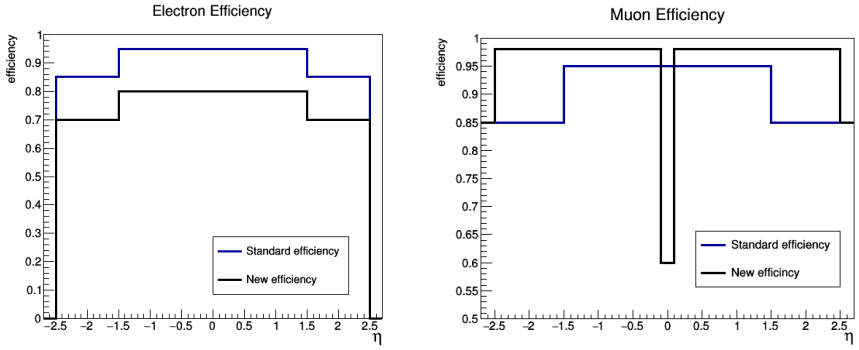
These two backgrounds are the ones considered in this work.

I used MadGraph5_aMC@NLO v 2.6.7[68] to generate simulated event samples at Leading Order (LO) both for the W^+W^- and the $t\bar{t}$ production, considering decay products with opposite charged leptons. The PDFs used with MadGraph5_aMC@NLO are the NNPDF2.3 [74]. Two samples, one for diboson and one for top pair production were created using the SM, while one sample of W^+W^- was generated with the EW_dim6 [75] model, considering only anomalous couplings. The Wilson coefficient c_{WWW} value used is 1 TeV^{-2}, while the other Wilson coefficients (discussed in 1.3.1) are set equal to zero. The cross sections obtained with MadGraph5_aMC@NLO for these three simulated samples are: $\sigma_{WW,SM} = 1.52$ pb for the W^+W^- SM sample, $\sigma_{t\bar{t},SM} = 11.37$ pb for the $t\bar{t}$ SM sample, and $\sigma_{WW,EFT} = 2.22 \cdot 10^{-4}$ pb for the W^+W^- EFT signal sample.

The parton shower and hadronization processes were simulated with Pythia v8.244 [69] using the NNPDF2.3 at LO [74], interfaced with the fast simulation software Delphes v3.3.0 [70] reproducing the ATLAS detector. Finally, events were reconstructed within the Root framework v6.2.4[72], as described in the next section.

5.2 Detector Simulation

When comparing simulated events with real data collected by experimental collaborations, and in particular here with [30] and [73], a more faithful fast simulation of the ATLAS detector is mandatory. In order to achieve this goal, I studied the detector simulation parameters, modifying accordingly those which turned out to mostly affect the present analysis. I pay particular attention to muons, electrons and jets reconstruction. As a first step the electron identification efficiency formula in the Delphes ATLAS card was modified in order to recreate the conditions obtained with the so called *tight* working point of 2.2.5. This efficiency formula, in the usual spirit of *fast* simulations, was expressed in the form of a step function in the electron p_T. I updated the formula through a fit to the identification efficiency plot for the *tight* working point taken from [53]. Similarly the muon identification efficiency is considered and modified with a fit on the identification efficiency plot for the *medium* working point taken from [54]. In figure 5.1 both the pre-existing and the modified electron and muon identification efficiency are shown. The effect of these updates in the identification efficiency formulas can be seen in figure 5.2, where, as an example, the dilepton invariant mass distribution for the SM background sample $W^+W^- \to e^\pm \nu_e \mu^\mp \nu_{mu}$ is shown both for events reconstructed with the modified identification efficiency functions and with the pre-existing ones.
Another important aspect when studying dileptonic final states is the isolation working point for electrons and muons. The one used in [30] are the so called *Gradient* working points, which can not be directly implemented in Delphes without changing the internal coding logic. As an alternative strategy, I studied which one of the *Fix*



(a) Electron identification efficiency (b) Muon identification efficiency

Figure 5.1: Electron and muon identification efficiency step functions. The black lines represent the modified efficiency, while the blue line the pre-existing efficiency in the Delphes ATLAS card.

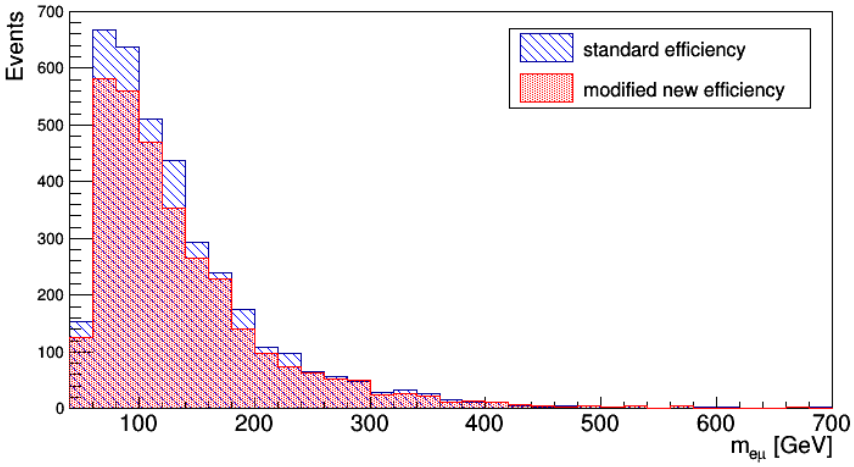


Figure 5.2: Dilepton invariant mass distribution for the SM W^+W^- sample. The red histogram considers events reconstructed with the modified identification efficiencies formulas for electrons and muons, while the blue histogram is made with the pre-existing identification efficiency formulas.

working points has the isolation efficiency most similar to the *Gradient* ones just mentioned. The isolation efficiencies for the various working points considered in this study were taken from [53] and [54]:the performances of the *Gradient* working point of [30] turned out to be the *Tight, Track only* working point for the electron isolation, and the *Fixed Cut Tight Track Only* working point for muon isolation. This was technically implemented by a change in the parameters ΔR_{max} and $p_{T_{ratio,max}}$ as shown in table 5.1.

Parameter	electron isolation	muon isolation
ΔR_{max}	0.2	0.3
$p_{T_{ratio,max}}$	0.06	0.06

Table 5.1: Isolation parameters modified in the Delphes card for electrons and muons.

Regarding the jet algorithm implementation I modified the radius parameter $R = 0.4$ (see 2.2.5), the minimum transverse momentum required for jets $p_T^{min} = 5$ GeV, and check that the jet finder algorithm chosen is the anti-k_T. Furthermore I modified the efficiency formulas for the b-jet tagging in order to recreate the condition described in [30] using the functions described and explained in [76].

5.3 Sequential Cut Analysis

Following the analysis in [30] I defined a region enriched in Triple Gauge Couplings W^+W^- events, called signal region, through a sequential cut analysis. As a basic requirement diboson event candidates must contain exactly one electron and exactly one muon of opposite charge, in particular events with a same-flavour lepton pair are discarded because they have a larger background from the Drell–Yan process. The selections for the final state leptons are:

- for **electrons** to have $p_T > 27$ GeV and a pseudorapidity $|\eta| < 2.47$, excluding the region $1.7 < |\eta| < 1.52$ corresponding to the transition region in the LAr calorimeter between the end-caps and the barrel (see 2.2.2);
- for **muons** to have $p_T > 27$ GeV and $|\eta| < 2.5$;
- for the **dilepton system** to have $p_T^{e\mu} > 30$ GeV and $m^{e\mu} > 55$ GeV

The requirements on the dilepton system, together with the one on **missing transverse energy** of $E_T^{miss} > 20$ GeV, target to minimize possible Drell-Yan background and to reduce below 1% the $H \to WW$ contribution [30].

Finally candidate events need to have no **jets** with $p_T > 35$ GeV and $|\eta| < 4.5$ and no b-jet with $p_T > 20$ GeV and $|\eta| < 2.5$. This requirement on b-jets strongly suppresses the background from top-quark production. In table 5.2 a summary of the above event selection requirements is given.

The selections implemented aim to suppress background events compared to signal ones: the selection efficiencies for the top quark pair production, the SM W^+W^- production and the simulated W^+W^- production via anomalous TGCs reported in table 5.3, show quantitatively how much this objective is reached.

5.4 DNN training phase

In order to prove the ability of a deep neural network to enhance the performances of a sequential cut analysis, I implemented a deep neural network to the classification

Selection requirement	Selection value
p_{T}^{ℓ}	> 27 GeV
η^{ℓ}	$\lvert\eta^{e}\rvert$ < 2.47 (excluding 1.37 < $\lvert\eta^{e}\rvert$ < 1.52), $\lvert\eta^{\mu}\rvert$ < 2.5
Number of additional leptons (p_{T} > 10 GeV)	0
Number of jets (p_{T} >35 GeV, $\lvert\eta\rvert$ <4.5)	0
Number of b-tagged jets (p_{T} > 20 GeV, $\lvert\eta\rvert$ <2.5)	0
$E_{\mathrm{T}}^{\mathrm{miss,track}}$	> 20 GeV
$p_{\mathrm{T}}^{e\mu}$	> 30 GeV
$m_{e\mu}$	> 55 GeV

Table 5.2: Summary of event selection criteria. Here l stands for electron or muon.

Sample	Selection efficiency
SM $t\bar{t}$	0.3%
SM W^+W^-	15%
BSM W^+W^- ($c_{WWW} = 1\mathrm{TeV}^{-2}$)	30%

Table 5.3: Selection efficiencies for the three simulated data samples.

problem of separating the diboson SM production from the W^+W^- produced via anomalous EFT effects.

Before the description of the DNN implementation in the next sections, In this paragraph the machine and software I used for the deep learning analysis are listed. The INFN[2] farm for scientific computing of Trieste provided the machine, based on Linux system, on which the code was run. The DNN analysis is performed with Python [77] programs. In particular for the DNN algorithm implementation I used Keras [78] together with TensorFlow (CPU-only version) [79].

5.4.1 Training Datasets

The samples generated and selected as described in 5.1 and 5.3 respectively were then used for the DNN training, testing and application (see 5.5). Specifically for the training and testing phase a leading background model was adopted: only the SM production W^+W^- sample was considered as background. The subset of signal and background simulated events samples used for the training and testing of the DNN are referred to as *training dataset* in this thesis. On the contrary, for the application of the DNN the $t\bar{t}$ sample is used together with the other two samples.

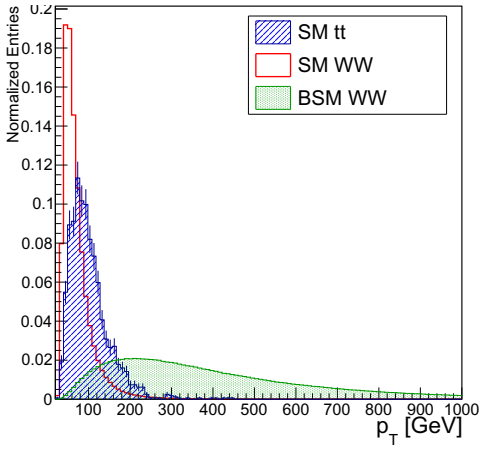
The signal and background subsamples considered as training dataset add up to 3200000 events, equally distributed among SM W^+W^- events and W^+W^- events produced via EFT TGCs. For the purpose of this work, from now on the first sample will be referred to as *background training dataset*, while the second one as *signal training dataset*.

As input for the DNN I chose low-level variables because of the deep learning technique used: in fact DL has the ability to automatically extract useful pieces of information from the input, without requiring to construct high level variables, based on more complex physics quantities. I took into consideration 14 observables describing the final state of the event: the leading lepton[3] transverse momentum $p_T^{lead,l}$, η and charge, the subleading lepton p_T, η and $\Delta\phi$ with respect to the leading

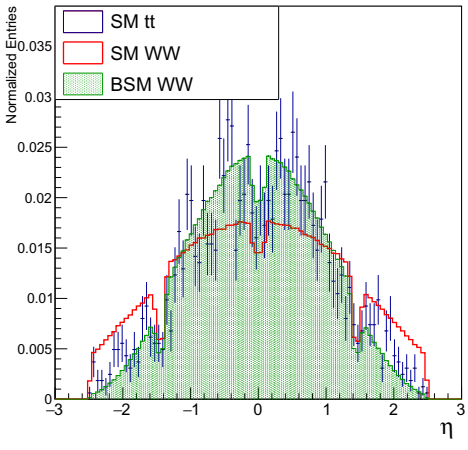
[2] the Italian National Institute for Nuclear Physics

[3] Here and in the following the leading lepton refers to the lepton with the higher transverse momentum p_T.

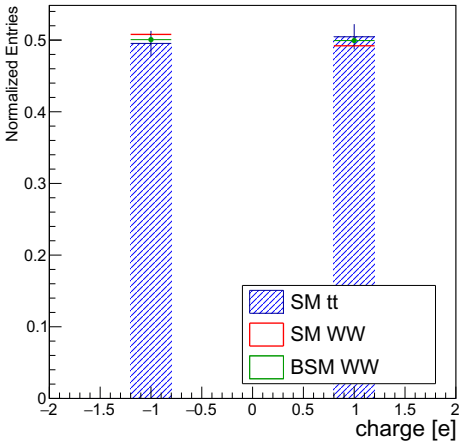
lepton direction, the missing transverse energy E_T^{miss} and its associated $\Delta\phi$, the jet p_T, η, charge, mass, and $\Delta\phi$, and the η of the opposite sum of all reconstructed objects, called *missing* η. The $\Delta\phi$ of an object is defined as the difference between the particle azimuthal angle ϕ and the leading lepton ϕ. In figures 5.3, 5.4, and 5.5 these 14 observables are shown for the signal training dataset, the background training dataset and for the $t\bar{t}$ sample.



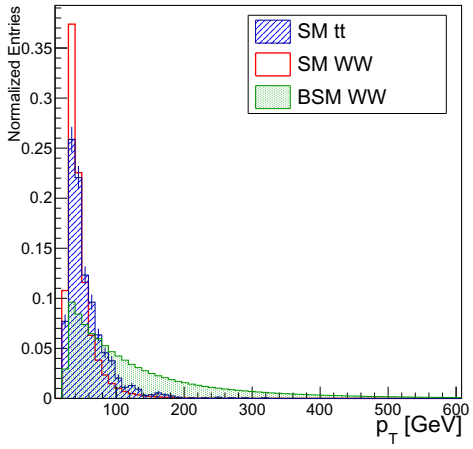
(a) Leading Lepton p_T



(b) Leading Lepton η



(c) Leading Lepton charge



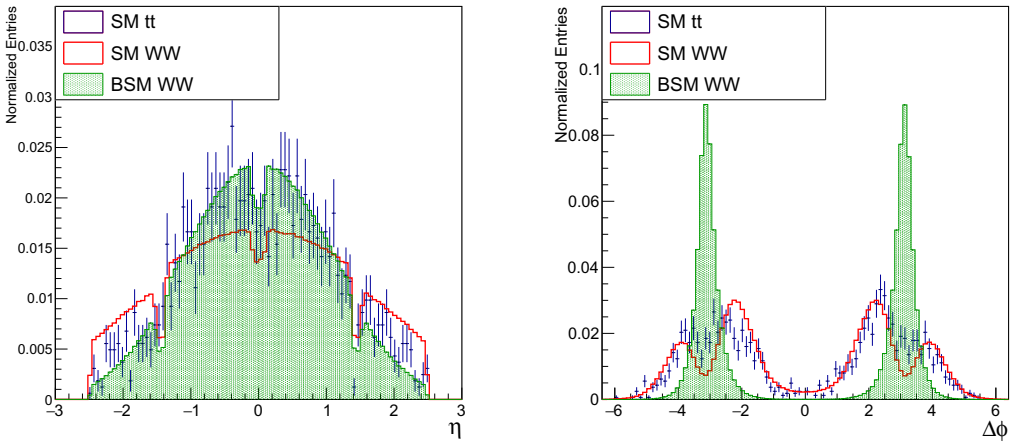
(d) Non leading Lepton p_T

Figure 5.3: Distributions of four different variables of the final event state for the $t\bar{t}$ sample in blue, the signal training dataset in green and the background training dataset in red.

Looking at figure 5.5(f) it can be seen that there is no significant difference between the jet charge distributions of the signal training dataset (in green) and of the background training datasets (in red). For this reason this variable was removed from the features input of the DNN. As a results the DNN takes as input 13 low level observables.

I designed the DNN so that it outputs a single variable, that is a combination of all the 13 input features, and that is related to the probability of an event to be part of the *signal training dataset* or of the *background training dataset*. This variable is called *DNN output* throughout this thesis. Figure 5.6 shows a sketch of the DNN process.

As explained in section 5.3 events with jets that have a transverse momentum $p_T > 35$ GeV are discarded. As a result some events have no jets, while others have one (or



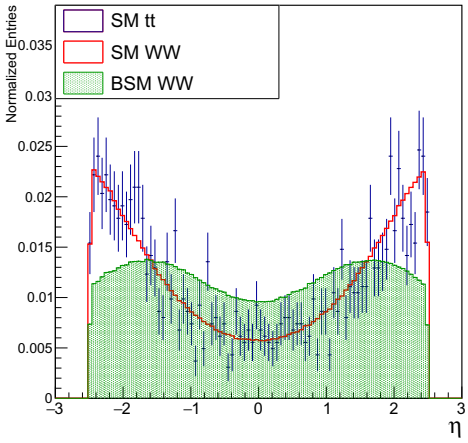
(a) Subleading Lepton η

(b) Subeading Lepton $\Delta\phi$

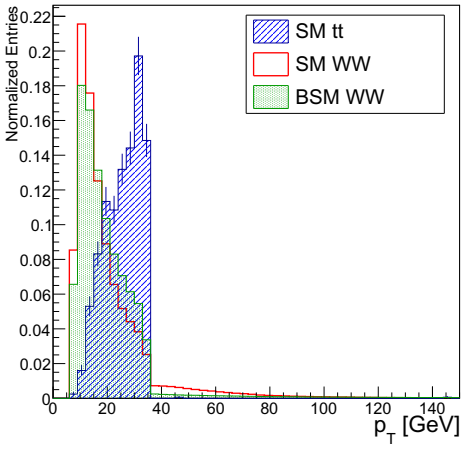
(c) Missing transverse energy

(d) Missing transverse energy $\Delta\phi$

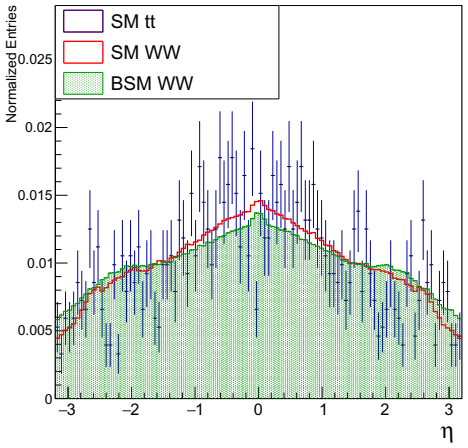
Figure 5.4: Distributions of four different variables of the final event state for the $t\bar{t}$ sample in blue, the signal training dataset in green and the background training dataset in red.



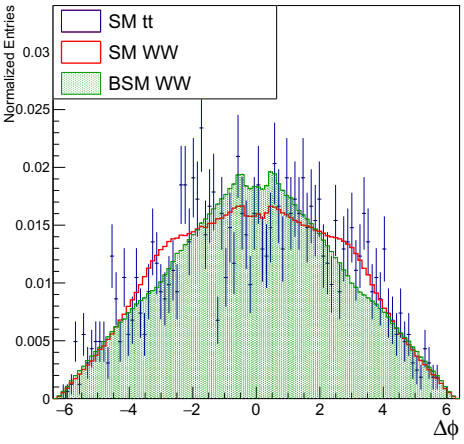
(a) Missing η



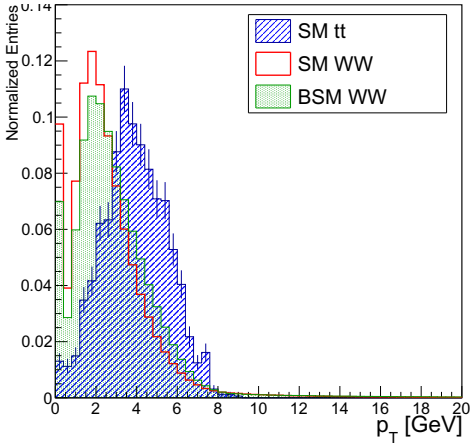
(b) Jet p_T



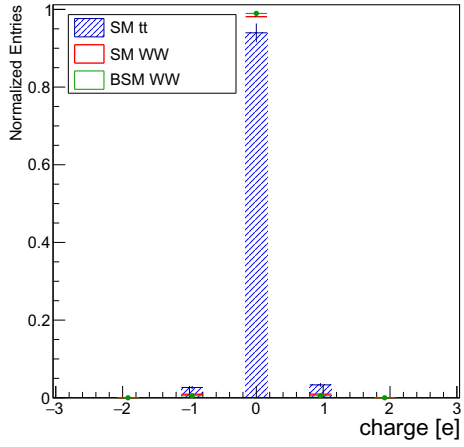
(c) Jet η



(d) Jet $\Delta\phi$



(e) Jet mass



(f) Jet charge

Figure 5.5: Distributions of six different variables of the final event state for the $t\bar{t}$ sample in blue, the signal training dataset in green and the background training dataset in red.

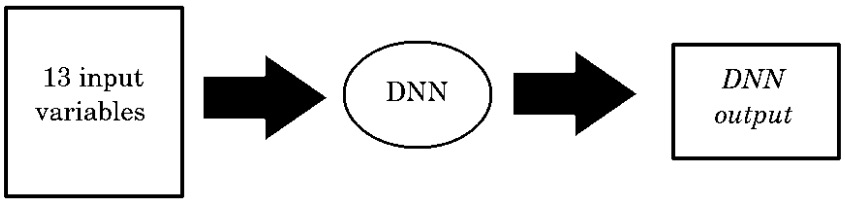


Figure 5.6: Sketch of the DNN working process: for each event the DNN takes as input 13 features and returns a single variable, the *DNN output*.

more) jet with $p_T < 35$ GeV. Anyway the DNN must take as input always the same number of features. Therefore, the four features relative to the jet are replaced by the value 0 for events with no jets. It could be interesting in future studies to consider two different dedicated DNNs, one for events with at least one jet and the other for events with no jets.

5.4.2 DNN model

The model used in this analysis is a deep feed-forward neural network, this means that all nodes in adjacent hidden layers are fully connected. The DNN uses the Adam optimizer and the gradient descent algorithm, in particular the mini-batch SGD one (see section 3.3).

The predictive power of a DNN tends to increase with the network complexity, which is given by the number of layers, the number of neurons per layers and in general by the number of parameters. Despite this growth in performances, DNNs that are too complex fail to generalize the relation between the input features and the output variable, and instead use their parameters to memorize characteristics of the specific events given in the training dataset. This is a common issue in ML, known as *overfitting*. It is expected, on general grounds, that the DNN performance reaches a plateau before starting to overfit. This plateau is the optimal point of work for the network. As a general rule, it is better to choose a simpler architecture for a certain DL algorithm, when it has performances similar to more complex ones. Once the structure of the DNN is fixed, the model has a certain number of hyper-parameters that must be chosen, which can heavily affect its performance. The number of epochs for a DNN is the number of times the algorithm uses the entire training dataset in the training.

The training of the DNN is guided by the loss function, as mentioned in 3.1, however the performance of the model can be evaluated with other metrics too. In this work, two alternative metrics are taken into consideration: the accuracy and the Area Under the ROC Curve (AUC), which are useful for binary classification problems, such as signal-background separation, as it is the case here. During the loss optimization the training dataset is randomly split in two groups, one is composed by 80% of the data and is used for training purposes, while the other 20% is used for the validation. The loss function is evaluated on the validation set (validation loss) without being optimized on it. In this way the performance of the DNN can be checked with data that the model has not used for the training. This validation procedure allows to estimate the presence and the amount of overfitting, which occurs when loss and validation loss start to diverge. The two alternative metrics are evaluated on both training and validation data without being optimized on those datasets.

5.4.3 DNN optimization

In order to obtain the best DNN performance I performed an optimization study on the architecture and hyper-parameters of the model. I investigated the structure of the DNN considering the number of hidden layers and the number of neurons per layer. At first the most basic model is used, with just one hidden layer and as many neurons as input features. In figure 5.7 the loss and alternative metrics curves are evaluated on the training set and on the validation one.

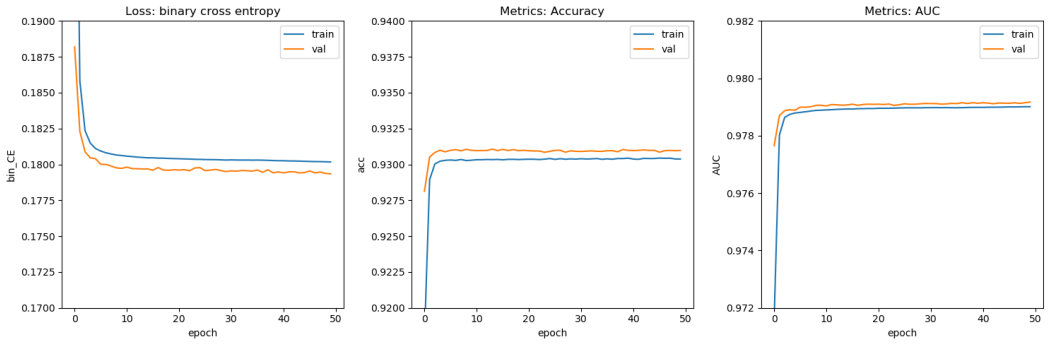


Figure 5.7: Loss and alternative metrics functions for a DNN with 1 hidden layer and 13 neurons for the curve evaluated on the training set (blue) and on the validation set (orange).

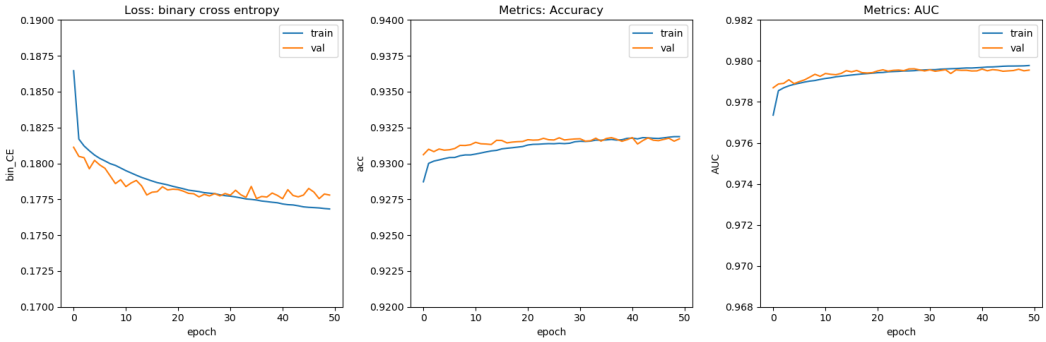
As a second step more complex architectures were studied, as summarized in the following:

- the **number of hidden layers** was varied in the set: 1, 2, 3, 4, 6 and 9. For the networks with three or more layers the performances were found to be similar but with increased time required. For this reason higher numbers of hidden layers were not considered. The loss and alternative metrics curves for 3 and 6 hidden layers are shown in figures 5.8(a), 5.8(b). The DNN with only one layer (as well as the basic model presented in figure 5.7) has worse values of the loss and alternative metrics than the DNN with 3 hidden layers. Concerning the model with 2 hidden layers I found that the model obtains similar performances to the 3 layers network after a number of epochs considerably greater, that implies more time needed. Therefore to optimize the computation time and the performances both for the training phase and for the application the number of hidden layers chosen is 3;

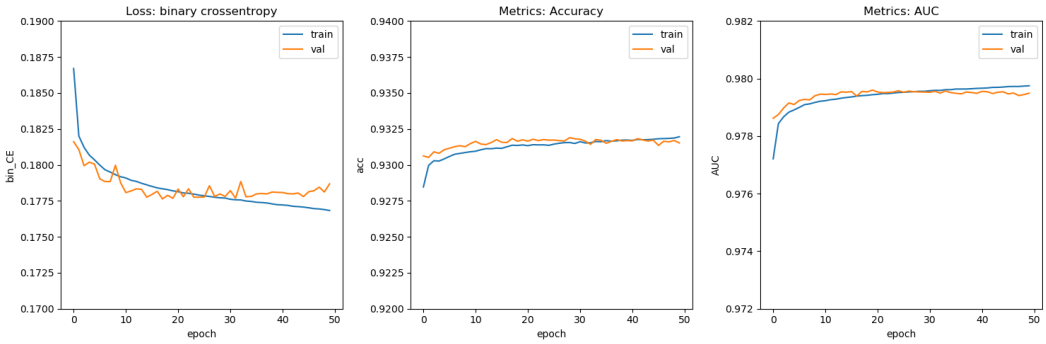
- the **number of neurons** per layer was varied in the set: 50, 100, 150, 200. The DNN with 50 and 100 neurons had similar performances but the first one requires more epochs than the second one, thus more time. In figure 5.9(a) the loss and alternative metrics curve are shown for the DNN with 50 neurons, while the above figure 5.8(a) considers the DNN with 100 neurons. The DNN architectures with 150 and 200 neurons per layer have similar performances to that with 100 neurons, but suffer from overfitting, as it can be seen in figure 5.9(b) for the DNN with 200 neurons per layer. Therefore I choose to use 100 neurons per layer in my DNN.

The loss and alternative metrics curves for the training and validation sets for the architectures not shown above are presented in appendix A for completeness.

Once the structure of the DNN is fixed to 100 neurons per layer with three hidden layers, I started studying the hyper-parameters of the DNN. As activation function

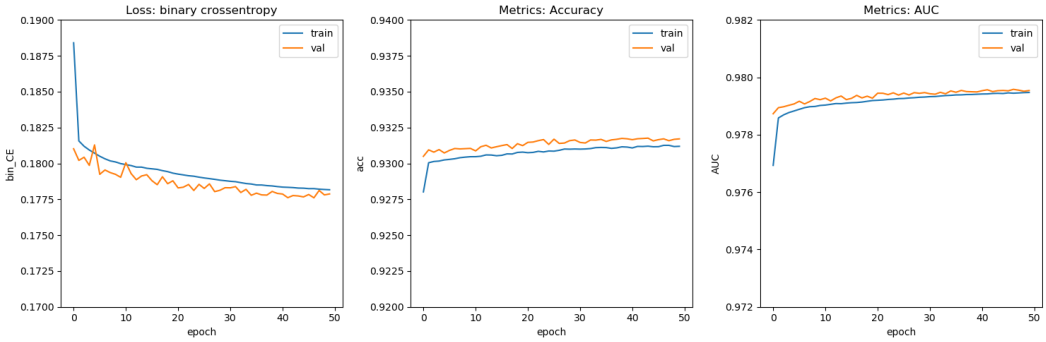


(a) 3 hidden layers.

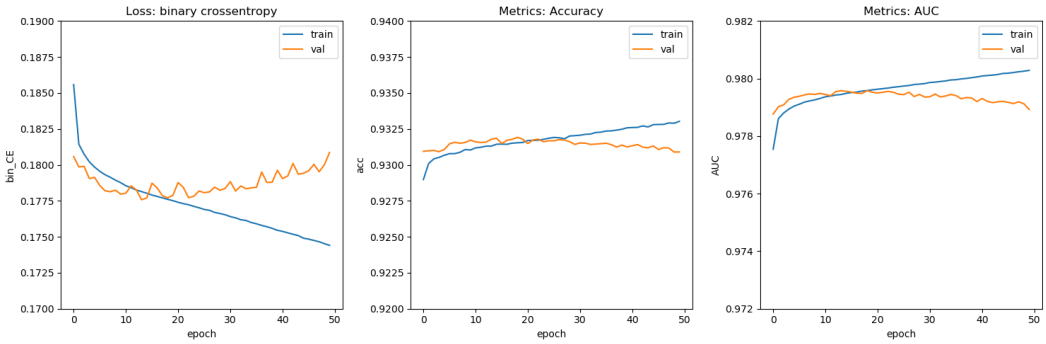


(b) 6 hidden layers.

Figure 5.8: Loss and alternative metrics functions for a DNN with 100 neurons per layer and three hidden layers 5.8(a) or with six hidden layers 5.8(b). In blue the curve evaluated on the training set and in orange on the validation set.



(a) 50 neurons per layer.



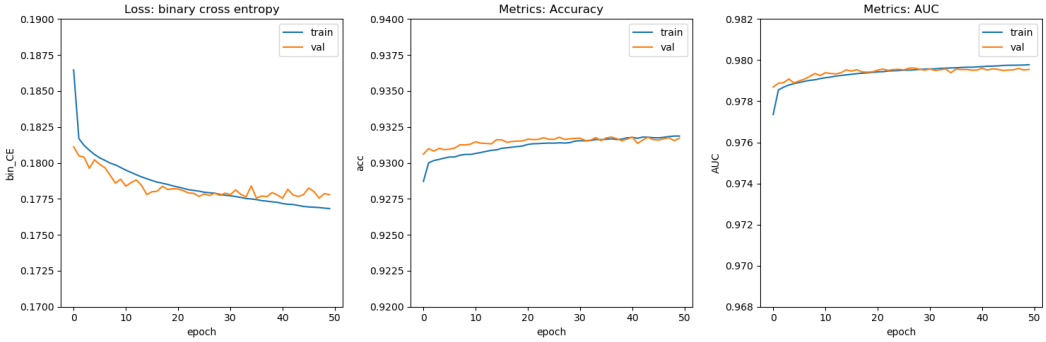
(b) 200 neurons per layer.

Figure 5.9: Loss and alternative metrics functions for a DNN with three hidden layers and 50 neurons per layer 5.9(a) and with 200 neurons per layer 5.9(b). In blue the curve evaluated on the training set and in orange on the validation set.

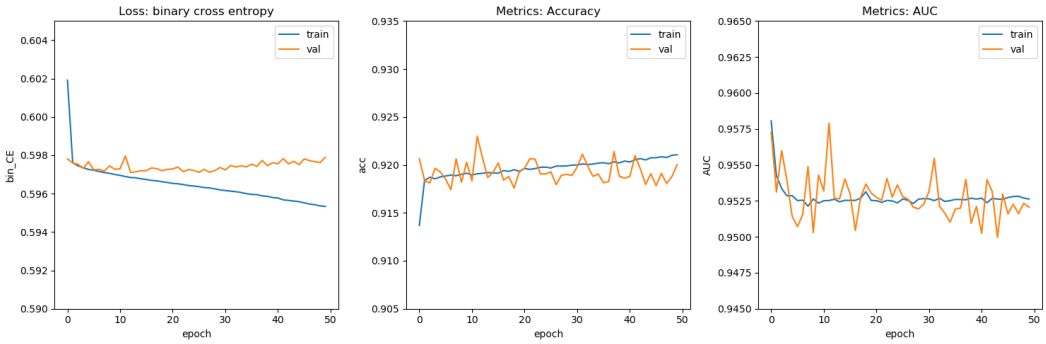
for the output layer of the network I choose the *sigmoid*, which is able to return a value in the range [0,1] and therefore useful for the binary classification problem studied in this thesis work. In particular the hyper-parameters scrutinized are:

- the **activation functions**: in particular I compared the *Relu*, and the *elu* (see section 3.2). The DNN with the *Relu* activation function has the best performances and its curves for the loss and alternative metrics are shown in figure 5.10(a). The same graphs for the DNN with the other activation function is shown in appendix A;

- the **loss function**: the binary cross entropy and the squared hinge. The network trained with the latter has considerably worse performances than the one with the binary cross entropy. In figure 5.10(a) and 5.10(b) the loss and alternative metrics curves for the DNN with binary cross entropy and squared hinge loss functions respectively are shown.



(a) *Relu* act. func. and binary cross entropy loss fun.



(b) *Relu* act. func. and squared hinge loss fun.

Figure 5.10: Loss and alternative metrics functions for a DNN with the structure described in the text, *binary cross entropy* 5.10(a) and *squared hinge* 5.10(b) loss function functions. In blue the curve evaluated on the training set and in orange on the validation set.

- the initial value of the **learning rate**, this was varied in the set: 10^{-2}, 10^{-3} and 10^{-4}. When the value used is 10^{-4} or 10^{-2} the DNN performances are worse than if the learning rate is set to 10^{-3}. The loss and alternative metrics for the different learning rate values are documented in appendix A;

- the **batch size**, varied in the set: 35, 50, 100 and 200. The batch size affects mainly the time required for the training and application of the DNN. I found that a batch size of 35 or 50 have similar performances and comparable time

needed for the training. If the batch size is 100 or 200 the time used for the training of one epoch decreases but the number of epochs necessary to have same performances as with a batch size of 50 increase. Nonetheless, I found that with a batch size value of 200 the time used to reach the best performances is the shortest;

In conclusion the DNN hyper-parameters I choose are the *Relu* activation function for all hidden layers, the *sigmoid* activation function for the output layer, the binary cross entropy as the loss function, a learning rate of 10^{-3} and a batch size of 200. With the latter parameters fixed, it was finally possible to start investigating the optimal number of epochs, both from the point of view of performances and of computational time. In order to do so, I considered the so called function *early stopping*[4] in TensorFlow, which computes the optimal epoch to stop the training by monitoring when the validation loss curve gets to a plateau. For the DNN architecture and the set of hyper-parameters chosen this value was found to be of 35 epochs, requiring about 22 minutes for the training. In tables 5.4 and 5.5 a summary of the chosen DNN parameters and structure is given.

DNN structure	optimal value
number of hidden layers	3
number of neurons per layer	100

Table 5.4: Summary of the structure chosen for the DNN.

hyper-parameter	optimal value
activation function output layer	sigmoid
activation function	Relu
loss function	binary cross entropy
learning rate	10^{-3}
batch size	200
number of epochs	35

Table 5.5: Summary of the hyper-parameters chosen for the DNN.

5.5 DNN application

The DNN model for binary classification, optimized and trained on the training dataset, is now applied to simulated data samples: $t\bar{t}$ events, W^+W^- SM production and W^+W^- EFT production with the Wilson coefficient $c_{WWW} = 1\text{TeV}^{-2}$. The distributions of the *DNN output* variable for the three samples are normalized to the conventional reference integrated luminosity of $L = 36.1$ fb^{-1} [30] and to the corresponding production cross sections (see 5.1). The scale factor used is defined as:

$$X = \frac{L \times \sigma_{MG5}}{N_{events}} \tag{5.1}$$

[4]The *early stopping* function in TensorFlow finds the optimal point to stop the training of a DNN. This function considers the values of the validation loss curve and choose the epoch corresponding to a stable value of the curve to avoid overfitting.

where σ_{MG5} is the process cross section computed by mg5_aMC@NLO and N_{events} is the number of simulated events for the sample considered. The same scaling operation is applied to all the following kinematical distributions, where not differently explicitly stated. The $p_T^{lead,l}$ distributions of the three samples too.

The $p_T^{lead,l}$ observable is chosen as the distribution most sensitive to the effect of the EFT operators among the ones obtained with the traditional sequential cut analysis [30]. The distributions both for the leading lepton p_T and for the *DNN output* are presented in figure 5.11.

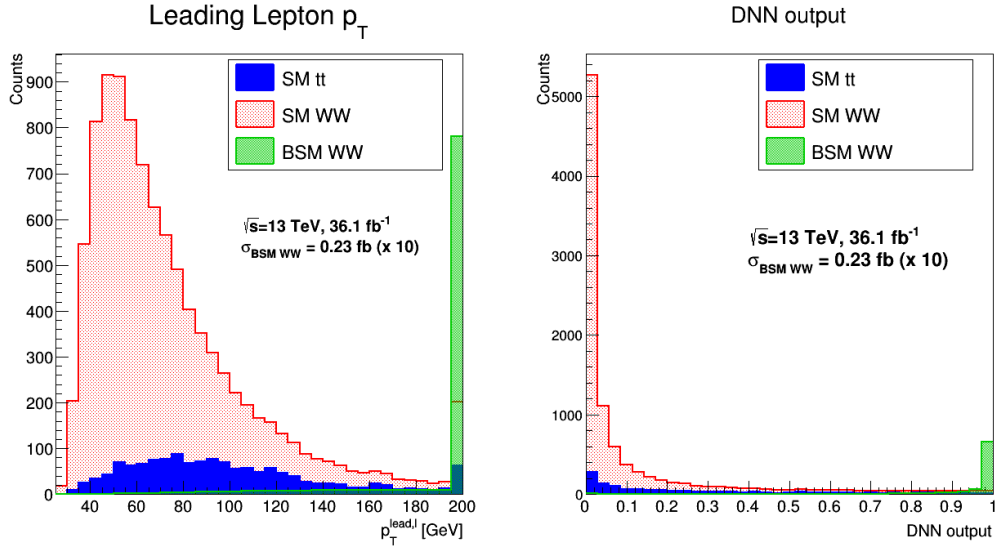


Figure 5.11: Distributions of the leading lepton p_T on the left and of the *DNN output* on the right. In blue the $t\bar{t}$ sample, in red the SM W^+W^- sample and in green the EFT W^+W^- sample with the coefficient $c_{WWW} = 1$ TeV^{-2}.

From the histograms in figure 5.11 it is clearly visible that the *DNN output* variable is able to better separate the signal component (the EFT W^+W^- sample) from the SM background rather than the traditional $p_T^{lead,l}$. In order to have a quantitative description of the power of the DNN, I computed the separation $< S^2 >$ for the two distributions, defined as:

$$< S^2 >= \frac{1}{2} \int \frac{(\hat{y_S} - \hat{y_B})^2}{\hat{y_S} + \hat{y_B}} dy \tag{5.2}$$

where y_S and y_B are the signal and background PDFs for the classifier y considered. The integral of equation 5.2 is understood as a sum over bins contents for binned distributions, as usual. The value obtained for the separation are collected in table 5.6: they represent a first quantitative proof that the separation is better for the *DNN output* variable rather than for the $p_T^{lead,l}$ distribution used with the sequential cut analysis only.

distribution	$< S^2 >$
$p_T^{lead,l}$	0.73
DNN output	0.76

Table 5.6: Separation values for the leading lepton p_T distribution and for the *DNN output* distribution. The statistical MC uncertainty on $< S^2 >$ was checked to be completely negligible in both cases.

5.6 Limits on TGCs

As a second quantitative way to show the improvement of the DNN analysis with respect to the sequential cut analysis, in this section I compute the 95% Confidence Level (CL) limit on the Wilson coefficient c_{WWW} with a least squares estimator, defined as:

$$\frac{\chi^2}{ndf} = \sum_{i=1}^{N} \frac{(n_i - y_i)^2}{\sigma_i^2 \; ndf} \tag{5.3}$$

where *ndf* is the number of bins of the histogram, i is the bin index, n_i is the expected yield in the EFT signal hypothesis in the $i-$th bin, y_i is the SM background in the $i-$th bin and σ_i^2 is the background uncertainty, which here and in the following is assumed to be of pure statistical nature.

In order to study the dependence of the estimator of equation 5.6 on the EFT coefficient c_{WWW}, I used the approximation that the cross-section $\sigma \propto c_{WWW}^2$, which holds when neglecting the subleading interference with the SM background. The 95 % CL for the c_{WWW} coefficient can be thus obtained by computing the p-values associated to these χ^2/ndf values. The limits I obtained when considering the leading lepton p_T and the *DNN output* as discriminating variables are, respectively:

$$\begin{aligned} c_{WWW} &\in [-3.2, 3.2] \text{ TeV}^{-2} \quad \text{for the } p_T^{lead,l} \\ c_{WWW} &\in [-2.3, 2.3] \text{ TeV}^{-2} \quad \text{for the } \textit{DNN output} \end{aligned} \tag{5.4}$$

The limits of equation 5.4 show again how the DNN analysis can improve the search for anomalous Triple Gauge Couplings with respect to a sequential cut analysis.

Finally it is interesting to derive the 95% CL limit on the c_{WWW} Wilson coefficient for different integrated luminosities, in particular the luminosity corresponding to the present full LHC Run-II collected data (139 fb^{-1}), the luminosity expected at the end of the LHC life (300 fb^{-1}) and the HL-LHC expected one (3000 fb^{-1}) [5]. These exclusion limits are computed taking into account only the difference in the integrated luminosity (L) with respect to the limits obtained in 5.4 as $\chi^2 \propto L$. The results obtained are collected in table 5.7.

	c_{WWW} 95%CL [TeV^{-2}]	
	Lead lept p_T	DNN output
2015/2016 Run-II	[-3.2 , 3.2]	[-2.3, 2.3]
full Run-II	[-2.3 , 2.3]	[-1.7, 1.7]
full LHC life	[-1.9 , 1.9]	[-1.4 , 1.4]
HL-LHC	[-1.1 , 1.1]	[-0.8 , 0.8]

Table 5.7: 95% CL for different integrated luminosity with statistic uncertainties only, both for the leading lepton p_T and for the *DNN output*.

It is interesting to notice how the exclusion limit here presented for the DNN analysis for the full LHC Run-II integrated luminosity is stringent than those obtained in [73] for the c_{WWW} Wilson coefficient. In this ATLAS paper the EFT operators are expressed in a different basis than the one used in this thesis (see 1.3.1). Therefore,

[5] The HL-LHC program [80] represents the performance evolution of the LHC. The term "high luminosity" explains the main purpose of this collider: in fact with a higher luminosity, HL-LHC will have a higher rate at which events can be observed.

to compare the exclusion limits in [73] with the one obtained in this thesis, I considered a multiplying factor that takes into consideration the basis transformation. The results in [73] constrained the c_{WWW} coefficient to a 95% confidence interval with a width of 9.4 TeV^{-2} while I obtained with the DNN analysis a width of 3.4TeV^{-2}. It is important to notice that tin this thesis analysis only statistic uncertainties are taken into consideration, while in [73] systematic ones are used too.

The exclusion limits in table 5.7 show that I improved the results obtained with the sequential cut analysis, through the implementation of a binary classification DNN. It is evident, in fact, that the deep learning technique constrained the c_{WWW} coefficient more than the traditional analysis.

Conclusions

At present collider experiments, as LHC, there is a continuous need to find more effective and efficient ways to analyse data, in order to extract from the raw data as much information as possible. In this thesis I performed a feasibility study of a deep learning data analysis tool: a deep neural network. I investigated the possibility to use a DNN for a classification problem using simulated data of diboson W^+W^- production at LHC with the ATLAS detector in the dilepton channel. I applied this analysis to an interesting case study: the extraction of exclusion limits on anomalous charged Triple Gauge Couplings, and I was able to show the power of deep learning techniques with respect to the traditional sequential cut analysis.

The extraction of confidence intervals associated to parameters describing BSM effects is a powerful mean to probe the predictive power of a BSM theory. In this thesis work, the computation of the 95 % CL on the c_{WWW} EFT coefficient was chosen as an example of a currently interesting exploration of physics beyond the SM and was therefore used as a case study for the application of the DNN. I generated simulated data samples through the hard-event generator mg5_aMC@NLO, interfaced with the shower/hadronization tool Phytia8 and with the fast-simulation software Delphes. At first I analysed the data with the traditional sequential cut analysis technique, by defining a fiducial phase space region for the leptonic decays of the W bosons through selection cuts on the kinematic distributions of the final state particles.

After the sequential cut analysis I successfully implemented a deep neural network in the feed-forward architecture and I was able to prove that it produces a better classification of the signal events, using the DNN output variable rather than the traditional leading lepton transverse momentum. The improved performances of the deep learning analysis with respect to the sequential cut analysis are shown by means of the obtained separation $< S^2 >$ computed between the background and signal distributions. For the sequential cut analysis the separation on the leading lepton transverse momentum distribution is $< S^2 >= 0.73$, while for the deep learning analysis the separation for the DNN output variable distribution was calculated to be $< S^2 >= 0.76$.

Finally, I used the events that passed the selections to compute the exclusion limits, both for the traditional leading lepton transverse momentum distribution and for the DNN output one. The 95% CL was obtained trough a least square estimator and deploying it with the computed distributions, considering statistic uncertainties only. The interval calculated considering the χ^2/ndf was rescaled to consider different integrated luminosity assumptions, corresponding to the 2015-2016 partial Run-II of LHC, the full LHC Run-II, the expected luminosity at the end of the ATLAS experiment at the LHC and finally the expected luminosity of the HL-LHC. In the following table I summarize the exclusion limits thus obtained:

	c_{WWW} 95%CL [TeV^{-2}]	
	Lead lept p_T	DNN output
2015/2016 Run-II	[-3.2 , 3.2]	[-2.3, 2.3]
full Run-II	[-2.3 , 2.3]	[-1.7, 1.7]

As it can be observed from the reported 95% CL values, it is evident that the DNN analysis significantly improves the results obtained by the traditional cut analysis, showing that the implementation of deep learning tools in HEP studies is a valid option to reach better results.

I found an optimal model and set of hyper-parameters for the DNN, through the optimization study conducted in this thesis work, and thus there should not be significant improvements considering different parameters. In future studies it could be interesting to evaluate the performances of the DNN including high-level features, such as the invariant mass and the transverse momentum of the dileptonic system, the b-tagging of the jets and others, as input variables. Furthermore it is expected that the results obtained with analysis conducted in this thesis could be enhanced by considering two different DNNs for events with one jet and with no jets.

Appendices

Appendix A

DNN Optimization Plots

In this appendix the plots relative to the DNN optimization (not already shown in section 5.4.3) are presented. Different structures and hyper-parameters are considered for the DNN. Each plot contains two curves: one (in blue) is the function evaluated on the training dataset, the other, called validation curve, (in orange) is the function evaluated on the validation dataset. The functions presented are the loss function, the accuracy and the AUC (defined in section 3.1).

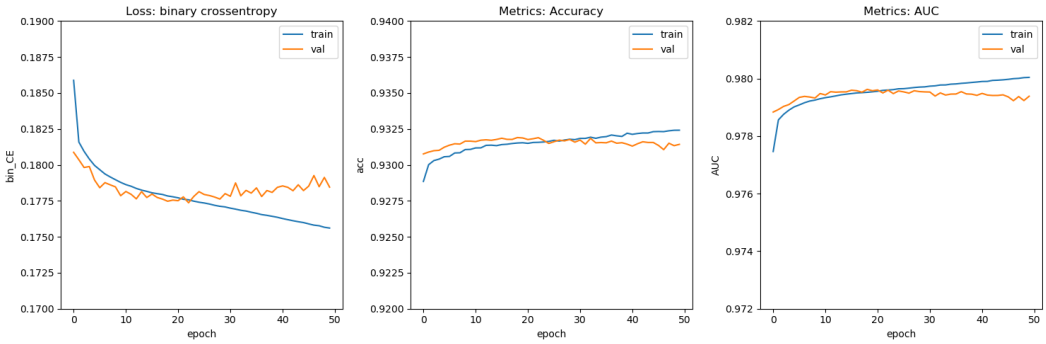


Figure A.1: Loss and alternative metrics functions for a DNN with 3 hidden layers and 150 neurons per layer. In blue the curve evaluated on the training set and in orange on the validation set.

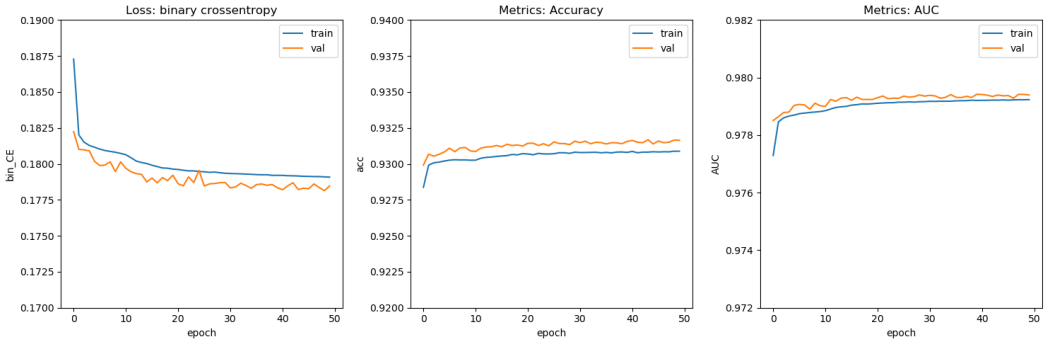
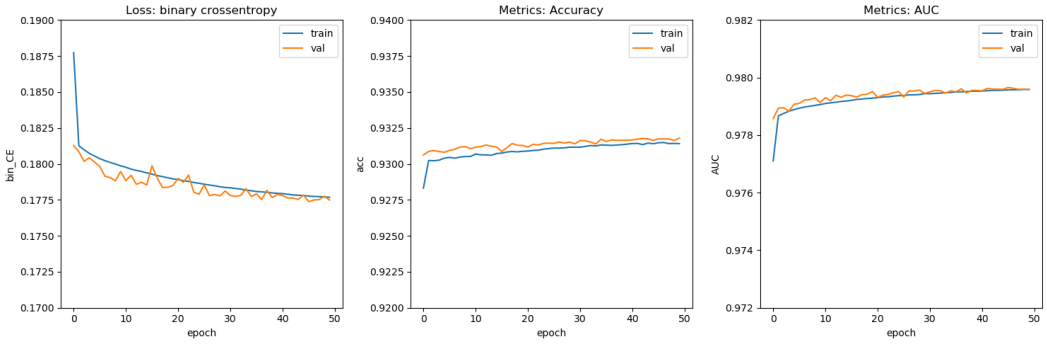
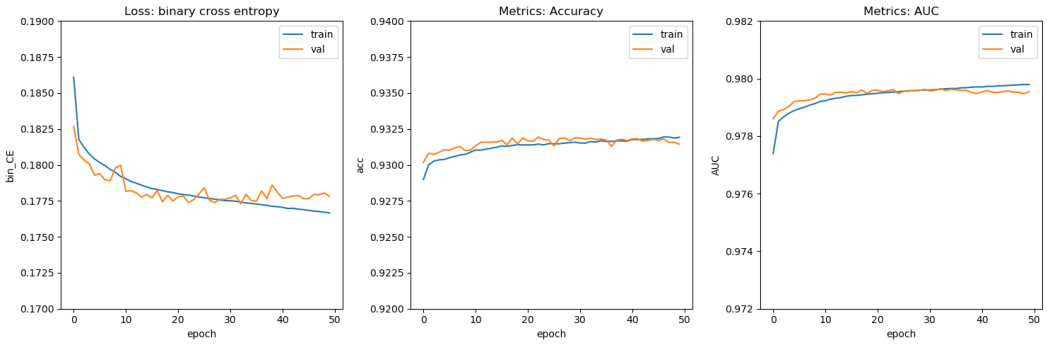


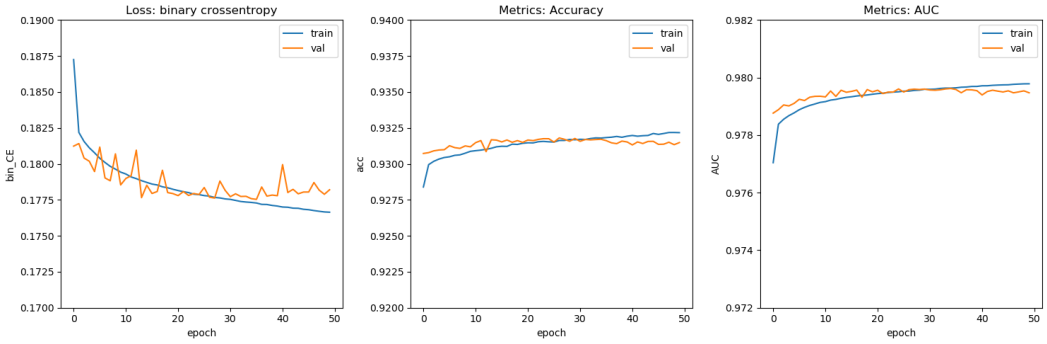
Figure A.2: Loss and alternative metrics functions for a DNN with the *Elu* activation function. In blue the curve evaluated on the training set and in orange on the validation set.



(a) 2 hidden layers

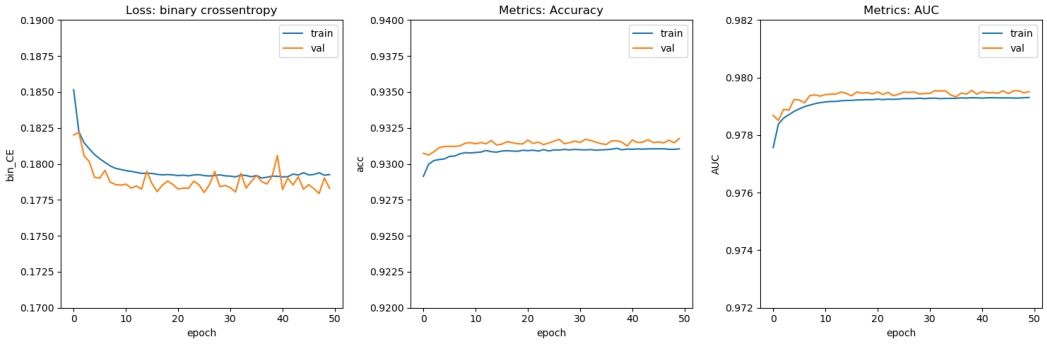


(b) 4 hidden layers

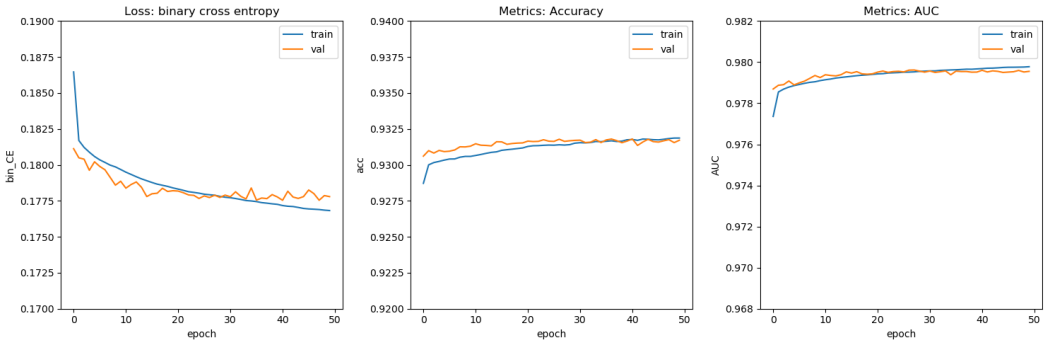


(c) 9 hidden layers

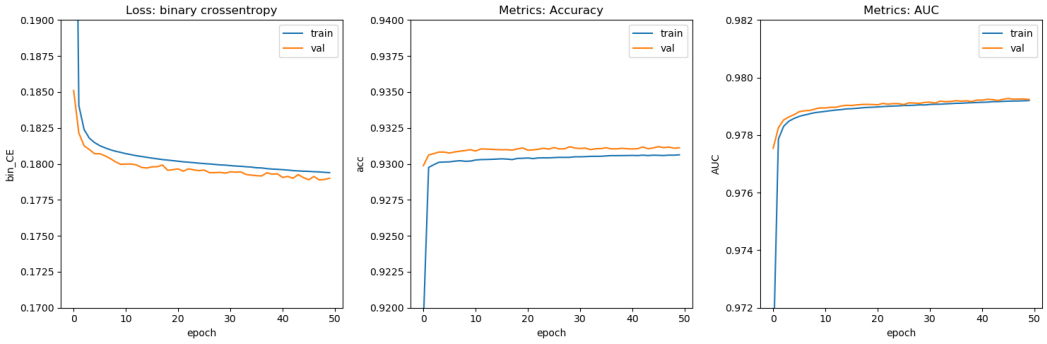
Figure A.3: Loss and alternative metrics functions for a DNN with 2 hidden layers A.3(a), 4 hidden layers A.3(b), and 9 hidden layers A.3(c). In blue the curve evaluated on the training set and in orange on the validation set.



(a) LR 10^{-2}



(b) LR 10^{-3}



(c) LR 10^{-4}

Figure A.4: Loss and alternative metrics functions for a DNN with initial value of the learning rate (LR) set to 10^{-2} A.4(a), to 10^{-3} A.4(b), and to 10^{-4} A.4(c). In blue the curve evaluated on the training set and in orange on the validation set.

Bibliography

[1] S. L. Glashow. Partial Symmetries of Weak Interactions. *Nucl. Phys.*, 22:579–588, 1961.

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

[3] A. Salam. Elementary Particle Physics: Relativistic Groups and Analyticity. *Eighth Nobel Symposium*, 367, 1968.

[4] G. Arnison et al. Experimental Observation of Lepton Pairs of Invariant Mass Around 95-GeV/c**2 at the CERN SPS Collider. *Phys. Lett. B*, 126:398–410, 1983.

[5] F. Abe et al. Observation of top quark production in $\bar{p}p$ collisions. *Phys. Rev. Lett.*, 74:2626–2631, 1995.

[6] Georges Aad et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Phys. Lett. B*, 716:1–29, 2012.

[7] Abdus Salam and John Clive Ward. Electromagnetic and weak interactions. *Phys. Lett.*, 13:168–171, 1964.

[8] Peter W. Higgs. Broken symmetries, massless particles and gauge fields. *Phys. Lett.*, 12:132–133, 1964.

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

[10] H. Fritzsch, Murray Gell-Mann, and H. Leutwyler. Advantages of the Color Octet Gluon Picture. *Phys. Lett. B*, 47:365–368, 1973.

[11] Gouranga C. Nayak. Matter-Antimatter Asymmetry of the Universe and Baryon Formation from Non-Equilibrium Quarks and Gluons. 9 2019.

[12] M. Tanabashi et al. Review of Particle Physics. *Phys. Rev. D*, 98(3):030001, 2018.

[13] Edvige Corbelli and Paolo Salucci. The Extended Rotation Curve and the Dark Matter Halo of M33. *Mon. Not. Roy. Astron. Soc.*, 311:441–447, 2000.

[14] Douglas Clowe, Marusa Bradac, Anthony H. Gonzalez, Maxim Markevitch, Scott W. Randall, Christine Jones, and Dennis Zaritsky. A direct empirical proof of the existence of dark matter. *Astrophys. J. Lett.*, 648:L109–L113, 2006.

[15] N. Aghanim et al. Planck 2018 results. VI. Cosmological parameters. *Astron. Astrophys.*, 641:A6, 2020. [Erratum: Astron.Astrophys. 652, C4 (2021)].

[16] Y. Fukuda et al. Evidence for oscillation of atmospheric neutrinos. *Phys. Rev. Lett.*, 81:1562–1567, 1998.

[17] Hanno Sahlmann. Loop Quantum Gravity - A Short Review. In *Foundations of Space and Time: Reflections on Quantum Gravity*, pages 185–210, 1 2010.

[18] Leonardo Modesto and Lesław Rachwał. Nonlocal quantum gravity: A review. *Int. J. Mod. Phys. D*, 26(11):1730020, 2017.

[19] P. Van Nieuwenhuizen. Supergravity. *Phys. Rept.*, 68:189–398, 1981.

[20] K. J. F. Gaemers and G. J. Gounaris. Polarization Amplitudes for e+ e- —> W+ W- and e+ e- —> Z Z. *Z. Phys. C*, 1:259, 1979.

[21] Steven Weinberg. Phenomenological Lagrangians. *Physica A*, 96(1-2):327–340, 1979.

[22] Celine Degrande, Nicolas Greiner, Wolfgang Kilian, Olivier Mattelaer, Harrison Mebane, Tim Stelzer, Scott Willenbrock, and Cen Zhang. Effective Field Theory: A Modern Approach to Anomalous Couplings. *Annals Phys.*, 335:21–32, 2013.

[23] Kaoru Hagiwara, S. Ishihara, R. Szalapski, and D. Zeppenfeld. Low-energy effects of new interactions in the electroweak boson sector. *Phys. Rev. D*, 48:2182–2203, 1993.

[24] Kaoru Hagiwara, J. Woodside, and D. Zeppenfeld. Measuring the WWZ Coupling at the Tevatron. *Phys. Rev. D*, 41:2113–2119, 1990.

[25] Kaoru Hagiwara, R. D. Peccei, D. Zeppenfeld, and K. Hikasa. Probing the Weak Boson Sector in e+ e- —> W+ W-. *Nucl. Phys. B*, 282:253–307, 1987.

[26] W. Buchmuller and D. Wyler. Effective Lagrangian Analysis of New Interactions and Flavor Conservation. *Nucl. Phys. B*, 268:621–653, 1986.

[27] S. Schael et al. Electroweak Measurements in Electron-Positron Collisions at W-Boson-Pair Energies at LEP. *Phys. Rept.*, 532:119–244, 2013.

[28] F. Abe et al. Observation of W^+W^- production in $\bar{p}p$ collisions at $\sqrt{s} = 1.8$ TeV. *Phys. Rev. Lett.*, 78:4536–4540, 1997.

[29] V. M. Abazov et al. Measurement of the WW production cross section with dilepton final states in p anti-p collisions at s**(1/2) = 1.96-TeV and limits on anomalous trilinear gauge couplings. *Phys. Rev. Lett.*, 103:191801, 2009.

[30] Morad Aaboud et al. Measurement of fiducial and differential W^+W^- production cross-sections at $\sqrt{s} = 13$ TeV with the ATLAS detector. *Eur. Phys. J. C*, 79(10):884, 2019.

[31] Albert M Sirunyan et al. W^+W^- boson pair production in proton-proton collisions at $\sqrt{s} = 13$ TeV. *Phys. Rev. D*, 102(9):092001, 2020.

[32] John Ellison and Jose Wudka. Study of trilinear gauge boson couplings at the Tevatron collider. *Ann. Rev. Nucl. Part. Sci.*, 48:33–80, 1998.

[33] Serguei Chatrchyan et al. Measurement of the W^+W^- Cross Section in pp Collisions at $\sqrt{s} = 7$ TeV and Limits on Anomalous $WW\gamma$ and WWZ Couplings. *Eur. Phys. J. C*, 73(10):2610, 2013.

[34] Morad Aaboud et al. Measurement of the ZZ production cross section in proton-proton collisions at $\sqrt{s} = 8$ TeV using the $ZZ \to \ell^-\ell^+\ell'^-\ell'^+$ and $ZZ \to \ell^-\ell^+\nu\bar{\nu}$ channels with the ATLAS detector. *JHEP*, 01:099, 2017.

[35] Vardan Khachatryan et al. Measurement of the W^+W^- cross section in pp collisions at $\sqrt{s} = 8$ TeV and limits on anomalous gauge couplings. *Eur. Phys. J. C*, 76(7):401, 2016.

[36] Georges Aad et al. Measurement of W^+W^- production in pp collisions at $\sqrt{s}$=7 TeV with the ATLAS detector and limits on anomalous WWZ and WWγ couplings. *Phys. Rev. D*, 87(11):112001, 2013. [Erratum: Phys.Rev.D 88, 079906 (2013)].

[37] Serguei Chatrchyan et al. Measurement of $W\gamma$ and $Z\gamma$ production in pp collisions at $\sqrt{s} = 7$ TeV. *Phys. Lett. B*, 701:535–555, 2011.

[38] Georges Aad et al. Measurement of the $W^{\pm}Z$ production cross section and limits on anomalous triple gauge couplings in proton-proton collisions at $\sqrt{s} = 7$ TeV with the ATLAS detector. *Phys. Lett. B*, 709:341–357, 2012.

[39] Vardan Khachatryan et al. Measurement of the WZ production cross section in pp collisions at $\sqrt{s} = 13$ TeV. *Phys. Lett. B*, 766:268–290, 2017.

[40] `https://en.wikipedia.org/wiki/File:S-channel.svg`. Accessed: 21/09/2021.

[41] `https://en.wikipedia.org/wiki/File:T-channel.svg`. Accessed: 21/09/2021.

[42] `https://en.wikipedia.org/wiki/File:U-channel.svg`. Accessed: 21/09/2021.

[43] LHC Machine. *JINST*, 3:S08001, 2008.

[44] `https://cds.cern.ch/record/2197559`. Accessed: 21/09/2021.

[45] O. Adriani et al. The LHCf detector at the CERN Large Hadron Collider. *JINST*, 3:S08006, 2008.

[46] G. Anelli et al. The TOTEM experiment at the CERN Large Hadron Collider. *JINST*, 3:S08007, 2008.

[47] James Pinfold et al. Technical Design Report of the MoEDAL Experiment. 6 2009.

[48] G. Aad et al. The ATLAS Experiment at the CERN Large Hadron Collider. *JINST*, 3:S08003, 2008.

[49] S. Chatrchyan et al. The CMS Experiment at the CERN LHC. *JINST*, 3:S08004, 2008.

[50] K. Aamodt et al. The ALICE experiment at the CERN LHC. *JINST*, 3:S08002, 2008.

[51] A. Augusto Alves, Jr. et al. The LHCb Detector at the LHC. *JINST*, 3:S08005, 2008.

[52] M. Abolins et al. The ATLAS Data Acquisition and High Level Trigger system. *JINST*, 11(06):P06008, 2016.

[53] Morad Aaboud et al. Electron reconstruction and identification in the ATLAS experiment using the 2015 and 2016 LHC proton-proton collision data at $\sqrt{s} =$ 13 TeV. *Eur. Phys. J. C*, 79(8):639, 2019.

[54] Georges Aad et al. Muon reconstruction performance of the ATLAS detector in proton–proton collision data at $\sqrt{s}$ =13 TeV. *Eur. Phys. J. C*, 76(5):292, 2016.

[55] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. FastJet User Manual. *Eur. Phys. J. C*, 72:1896, 2012.

[56] Matteo Cacciari, Gavin P. Salam, and Gregory Soyez. The anti-k_t jet clustering algorithm. *JHEP*, 04:063, 2008.

[57] Dan Guest, Kyle Cranmer, and Daniel Whiteson. Deep Learning and its Application to LHC Physics. *Ann. Rev. Nucl. Part. Sci.*, 68:161–181, 2018.

[58] Warren S. McCulloch. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

[59] Milos Miljanovic. Comparative analysis of Recurrent and Finite Impulse Response Neural Networks in Time Series Prediction Abstract. *Indian Journal of Computer and Engineering*, 3, 2011.

[60] Q. Liao H. Mhaskar and T. Poggio. Learning Real and Boolean Functions: When Is Deep Better Than Shallow. *Center for Brains, Minds and Machines*, 2016.

[61] Michael A. Nielsen. Neural Networksand Deep Learning. *Determination Press*, 2015.

[62] G. E. Hinton Y. LeCun, Y. Bengio. Deep Learning. *Nature*, 521:436–444, 2015.

[63] Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for Exotic Particles in High-Energy Physics with Deep Learning. *Nature Commun.*, 5:4308, 2014.

[64] `https://www.educative.io/edpresso/learning-rate-in-machine-learning`. Accessed: 11/10/2021.

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

[66] F. E. Curtis L. Bottou and J. Nocedal. Optimization Methods for Large-Scale Machine Learning. 2018.

[67] Enrico Bothmann et al. Event Generation with Sherpa 2.2. *SciPost Phys.*, 7(3):034, 2019.

[68] 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. *JHEP*, 07:079, 2014.

[69] Torbjorn Sjostrand, Stephen Mrenna, and Peter Z. Skands. PYTHIA 6.4 Physics and Manual. *JHEP*, 05:026, 2006.

[70] J. de Favereau et al. DELPHES 3, A modular framework for fast simulation of a generic collider experiment. *JHEP*, 02:057, 2014.

[71] Neil D. Christensen and Claude Duhr. FeynRules - Feynman rules made easy. *Comput. Phys. Commun.*, 180:1614–1641, 2009.

[72] I. Antcheva et al. ROOT: A C++ framework for petabyte data storage, statistical analysis and visualization. *Comput. Phys. Commun.*, 180:2499–2512, 2009.

[73] Georges Aad et al. Measurements of $W^+W^-+ \geq 1$ jet production cross-sections in pp collisions at $\sqrt{s} = 13$ TeV with the ATLAS detector. *JHEP*, 06:003, 2021.

[74] Richard D. Ball et al. Parton distributions with LHC data. *Nucl. Phys. B*, 867:244–289, 2013.

[75] Celine Degrande, Nicolas Greiner, Wolfgang Kilian, Olivier Mattelaer, Harrison Mebane, Tim Stelzer, Scott Willenbrock, and Cen Zhang. Effective Field Theory: A Modern Approach to Anomalous Couplings. *Annals Phys.*, 335:21–32, 2013.

[76] Giovanni Guerrieri. Charged Triple Gauge Couplings at present and future colliders. *Master Thesis*, 2019.

[77] G. Van Rossum and F. L. Drake Jr. Python reference manual. *Centrum voor Wiskunde en Informatica Amsterdam*, 1995.

[78] F. Chollet et al. Keras. *XH Lu et al./Application of Machine Learning and Grocery Transaction Data. Available at: https://keras.io*, 2015.

[79] A. Martin et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. *Available at: http://tensorflow.org/*, 2015.

[80] High-Luminosity Large Hadron Collider (HL-LHC): Technical Design Report V. 0.1. 4/2017, 2017.