Universidade do Minho
Escola de Ciências

António Carlos Pinto Oliveira

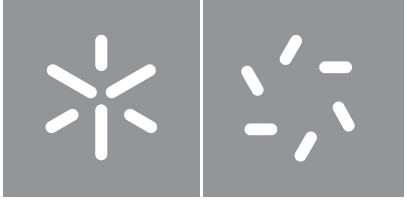
Supervised machine learning techniques in high energy physics

Supervised machine learning techniques in high energy physics

António Oliveira

UMinho | 2022

Março 2022



Universidade do Minho
Escola de Ciências

António Carlos Pinto Oliveira

Supervised machine learning techniques in high energy physics

Dissertação de Mestrado
Mestrado em Física

Trabalho efetuado sob a orientação do
Professor Doutor Nuno Filipe da Silva Fernandes de Castro
e do
Doutor Miguel Correia dos Santos Crispim Romão

Março 2022

Acknowledgements

For the successful completion of this work I relied on the support of my supervisors, Professor Nuno Castro and Doctor Miguel Romão. I am also grateful to my LIP-Minho colleagues for their help, especially to Maura and Henrique.

I would like to thank my family for their love and support.

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the Universidade do Minho.

_____, _____

(Location) (Date)

_____

(António Carlos Pinto Oliveira)

Resumo

Técnicas de aprendizagem automática supervisionada em física de altas energias

O Modelo Padrão da Física de Partículas (MP) é uma teoria extremamente bem sucedida na confrontação experimental. No entanto, a busca por fenómenos que não caibam no seu quadro explicativo é um campo ativo. Várias famílias de teorias que estendem o MP são construídas e novos fenómenos por elas previstos são objeto de pesquisa. Nomeadamente, são feitas buscas por partículas que essas teorias prevêm que se manifestem nos dados adquiridos no detetor ATLAS no LHC a uma energia do centro de massa de 13 TeV. Um tipo particular de buscas consiste em estabelecer limites para certos parâmetros da teoria.

Para levar a cabo essas pesquisas vários métodos são usados. Todos eles se baseiam em otimizar a capacidade para distinguir o que é modelizado como acontecimentos esperados assumindo o MP como explicação para as observações (fundo) e o que é modelizado como acontecimentos que deveriam estar presentes se a extensão ao MP fosse correta (sinal). Têm vindo a ser usadas técnicas de aprendizagem de máquina para esse efeito como alternativa a uma análise em que se delimita o espaço de fase da pesquisa usando regiões retangulares e se usam variáveis discriminantes motivadas pelo conhecimento da física do problema em estudo. As redes neuronais escolhem regiões do espaço de fase com formas mais gerais e constroem uma variável discriminante que é opaca no seu significado físico, mas eficaz. Neste trabalho é feita uma comparação do uso de redes neuronais profundas com a análise mais tradicional para estabelecer limites inferiores da massa dum hipotético bosão Z' usando dados públicos de ATLAS. É também estudado o efeito do uso da variável motivada fisicamente como componente de uma análise baseada em redes neuronais. Por fim, um estudo adicional é feito sobre a transferibilidade de redes neuronais treinadas para reconhecer um sinal específico para discriminar sinais diferentes.

Palavras-chave: Além do modelo padrão, Aprendizagem automática, ATLAS.

Abstract

Supervised machine learning techniques in high energy physics

The Standard Model of particle physics (SM) is an extremely successful theory in the comparison with experimental data. However, the search for phenomena that do not fit into its explanatory framework is an active field. Several families of theories extending the SM are constructed and new phenomena predicted by them are the subject of research. Particularly, searches are conducted to find particles that these theories predict will manifest themselves in data acquired at the ATLAS detector at the LHC at a center-of-mass energy of 13 TeV. One particular type of search consists of setting limits on certain parameters of the theory, namely the mass of said particles.

To carry out these searches various methods are used. They are all based on optimizing the ability to distinguish between what is modeled as expected events assuming the MP as an explanation for the observations (background) and what is modeled as events that should be present if the extension to the MP were correct (signal). Machine learning techniques have been used for this purpose as an alternative to an analysis in which one delimits the phase space of the search using rectangular regions and uses discriminant variables motivated by knowledge of the physics of the problem under study. Neural networks choose regions of the phase space with more general shapes and construct a discriminant variable that is opaque in its physical meaning, but effective. In this work a comparison is made of the use of deep neural networks with more traditional analysis to establish lower limits on the mass of a hypothetical Z' boson using ATLAS open data. The effect of using the physically motivated variable as a component of a neural network-based analysis is also studied. Finally, an additional study is done on the transferability of neural networks trained to recognize a specific signal to discriminate different signals

Keywords: ATLAS, Beyond the Standard Model, Machine learning.

Contents

List of Figures

List of Tables

Chapter 1

Theory Overview

1.1 The Standard Model

The Standard Model (SM) is our present response to the age-old question, "What is matter made of?". That answer has required a long history of intellectual inquiry, both in the formulation of concepts and in the ingenious creation of experimental devices to empirically test ideas with unparalleled levels of precision and using extremely high levels of energy to probe the inner structure of matter as finely as possible.

It tells us that matter comes in three families, called generations. In each family, there are two flavors of quarks (up and down), (charm and strange), (top and bottom), which respond to the strong interaction, being assigned a color charge, and because they also have an electric charge, also respond to the electromagnetic interaction, and there are two leptons (electron and electron neutrino), (muon and muon neutrino), (tau and tau neutrino), the neutrino being electrically neutral and the other having the same electric charge as the electron. The strong interaction does not affect the leptons, while the weak interaction does, and also the electromagnetic in the case of the charged particles. These matter particles have spin $1/2$ and are thus fermions, obeying the Pauli Exclusion Principle. On the other hand, interactions are mediated by bosons with spin 1. The weak interaction is mediated by three bosons with mass, giving it a finite range, one of which is electrically neutral and the other two charged with opposed signs. The electromagnetic interaction is mediated by photons, particles without mass, giving it an infinite range; the strong interaction is mediated by 8 gluons, which have color themselves, leading to confinement, that is, no quark exist in isolation, always showing as components of hadrons, that are bound states of quarks, existing in two varieties: baryons $(qqq, \bar{q}\bar{q}\bar{q})$ and mesons $(q\bar{q})$.

All particles with mass are subject to gravitation, described by General Relativity and not by the SM. The first generation forms the familiar matter, with quarks up and down forming protons and neutrons. The particles that belong to the other generations, being similar in many ways to those of the first family, are heavier and unstable, decaying to the lighter ones of the first generation. Moreover, each fermion has a companion anti-particle, with the same properties but with the charges reversed. There is also the Higgs

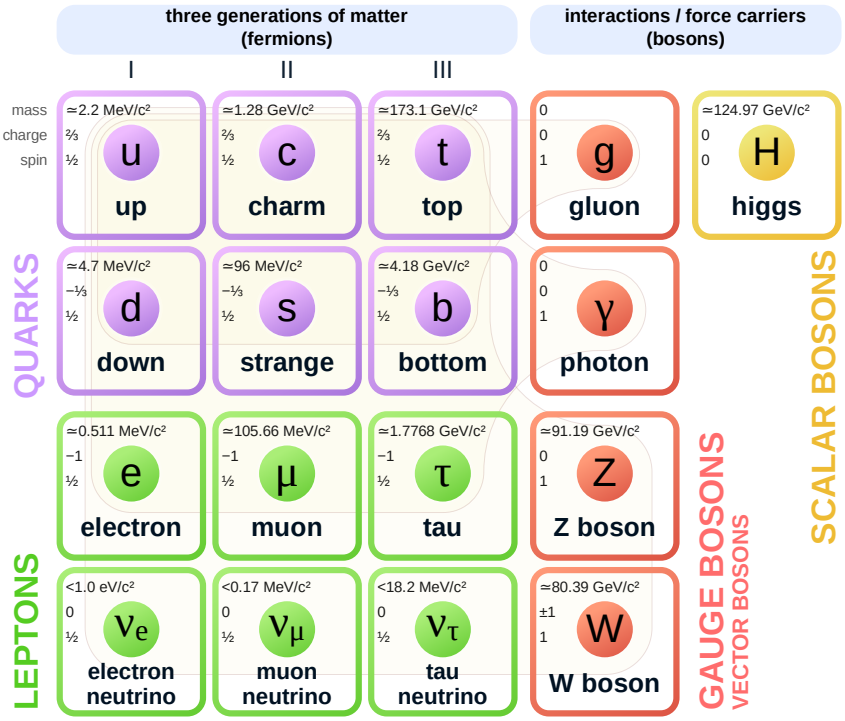


Figure 1.1: SM elementary particles and their proprieties. Adapted from [1]

boson, with spin 0, responsible for the masses of all particles via the Higgs mechanism.

1.1.1 Quantum Electrodynamics

The standard model was developed in stages. The first of them was the creation of quantum electrodynamics by Tomonaga [2], Schwinger [3] [4] and Feynman [5] [6] [7]. It came about combining relativistic quantum mechanics and the quantization of the electromagnetic field. This is the prototype of a gauge quantum field theory that guided the creation of the theories for other interactions. A free fermion with spin $\frac{1}{2}$, is described by the Dirac equation

$$i\gamma^{\mu}\partial_{\mu}\psi - m\psi = 0\,, \tag{1.1}$$

which is the solution of the Euler-Lagrange equations for the Lagrangian density (from now on, simply Lagrangian)

$$\mathcal{L} = i\,\bar{\psi}\gamma^{\mu}\partial_{\mu}\psi - m\,\bar{\psi}\psi\,, \tag{1.2}$$

where γ^{μ} are the four 4x4 Dirac matrices that obey:

$$\{\gamma^{\mu},\gamma^{\nu}\} = \gamma^{\mu}\gamma^{\nu} + \gamma^{\nu}\gamma^{\mu} = 2\eta^{\mu\nu}\,, \mu = 0\cdots 3\,, \tag{1.3}$$

and $\eta^{\mu\nu}$ is the Minkowski metric with signature (+ − − −) and a 4x4 identity matrix is implicit. The Proca Lagrangian for a field A^μ,

$$\mathcal{L} = -\frac{1}{4}(\partial^\mu A^\nu - \partial^\nu A^\mu)(\partial_\mu A_\nu - \partial_\nu A_\mu) + \frac{m_A^2}{2}A^\nu A_\nu, \tag{1.4}$$

gives rise through the Euler-Lagrange equations to

$$\partial_\mu(\partial^\mu A^\nu - \partial^\nu A^\mu) + m_A^2 A^\nu = 0, \tag{1.5}$$

which describe a particle of spin 1. For $m_A = 0$, these are Maxwell equations in free space,

$$\partial_\mu F^{\mu\nu} = j^v \quad , \quad F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu \, . \tag{1.6}$$

Adding a source term with a current that obeys the continuity equation, this Lagrangian describes the electromagnetic field.

The free fermion Lagrangian is invariant under global phase-rotation transformations, that is, the transformation $\psi \longrightarrow e^{-iq\chi}\psi$ does not alter it. If we impose the condition that invariance must be local (called local gauge invariance [1]), that is, that χ is a function of space-time coordinates, $\mathcal{L}$ must have an additional term in order to preserve the invariance, because the derivative term introduces a dependency on χ that must be cancelled out. As $\psi \longrightarrow e^{-iq\chi(x)}\psi$, then $\mathcal{L}$ must be given by

$$\mathcal{L} = [\, i\bar{\psi}\gamma^\mu\partial_\mu\psi - m\bar{\psi}\psi \,] - (q\bar{\psi}\gamma^\mu\psi)A_\mu \tag{1.7}$$

where A_μ is a vectorial field that transforms as

$$A_\mu \longrightarrow A_\mu + \partial_\mu\chi(x) \tag{1.8}$$

in order to cancel the unwanted term. So, a term that couples the Dirac field with a vectorial field is introduced . Now, one has to add the free part of the Lagrangian for that field, that is, the Proca Lagrangian (1.4). For the Proca Lagrangian to be invariant under (1.8) the mass term must be zero, that is, is necessary to set $m_A = 0$. As it turns out, imposing local gauge invariance in the Lagrangian for spin $1/2$ particles force us to introduce a field for a spin 1 massless particle, the photon, that describes the electromagnetic interaction. The existence of gauge invariance means that not all components of A_μ correspond to physical degrees of freedom because physical quantities must not depend on arbitrary choices of $\chi(x)$. It is important to note, for the discussion that will follow, that the local gauge transformation under which the Lagrangian had to remain invariant is multiplication by a unitary 1x1 matrix, an element of the $U(1)$ group. Theories for other interactions were built imposing gauge invariance having particular symmetries related to unitary groups of other degrees. It is also worth noting the concept of *covariant derivative*, which

[1] For a short account about the use of the term gauge see Reference [8].

summarizes the procedure just described to transform a global phase-rotation invariant theory into a local gauge invariant one and to express the interaction between the gauge boson and the fermion. The partial derivative in the free Lagrangian for the fermion is replaced by a covariant derivative [2]

$$\mathcal{D}_\mu = \partial_\mu + iqA_\mu. \tag{1.9}$$

1.1.2 Quantum Chromodynamics

Quantum Chromodynamics (QCD) is the gauge [3] theory of the strong interaction. It describes the interaction between carriers of color charge, that is, quarks, antiquarks, and gluons. Quarks can have three colors, namely, red (r), green (g), blue (b). Anti-quarks have they anti-colors. Color as a quantum number, label states of quarks, antiquarks, and gluons and allow to account for the absence of, for example, free quarks or (q q) hadrons, that are colorless. For a quark of a given flavor with mass m the free Lagrangian is

$$\mathcal{L} = \sum_{c\in\{r,g,b\}} [\, i\,\bar{\psi}_c\gamma^\mu\partial_\mu\psi_c - m\,\bar{\psi}_c\psi_c]\,. \tag{1.10}$$

Defining the color triplet:

$$\psi = \begin{pmatrix}\psi_r\\ \psi_g\\ \psi_b\end{pmatrix} \quad , \quad \bar{\psi} = (\psi_r\psi_g\psi_b), \tag{1.11}$$

(1.10) can be rewritten as

$$\mathcal{L} = i\,\bar{\psi}\gamma^\mu\partial_\mu\psi - mc^2\,\bar{\psi}\psi\,. \tag{1.12}$$

This Lagrangian is invariant under unitary transformations of the new ψ

$$\psi \longrightarrow U\psi\,, \tag{1.13}$$

where U is an unitary 3x3 matrix. Unitary 3x3 matrices can be written as

$$U = e^{i\mathbf{H}}, \tag{1.14}$$

where $\mathbf{H}$ is a Hermitian matrix. It can be decomposed further as

$$U = e^{i\theta}e^{i\frac{\boldsymbol{\lambda}}{2}\cdot\boldsymbol{\theta}}, \tag{1.15}$$

where $\boldsymbol{\lambda}$ are the eight Gell-Mann matrices [4]. Ignoring the scalar phase factor (the $U(1)$ symmetry already studied), which amounts to consider only unitary matrices with determinant 1 [5], we can consider then only the $SU(3)$ group that is generated by the elements of the algebra defined by

$$[t_a, t_b] = if_{abc}t_c\,, \tag{1.16}$$

[2] This procedure is called the minimal coupling rule.

[3] Finding the form of the interaction imposing a local gauge unitary symmetry is based on the 1954 work by Yang and Mills [9] related to isospin in nuclear physics.

[4] The 3x3 linearly independent traceless Hermitian matrices.

[5] Due to the relation $det(e^A) = e^{tr(A)}$.

where $\boldsymbol{t} = \frac{\boldsymbol{\lambda}}{2}$ and f_{abc} are structure constants, not all zero, differently from the $U(1)$ situation. Using the procedure described in the last subsection for imposing local gauge invariance, namely, considering the transformation

$$\psi \rightarrow e^{-ig\boldsymbol{\lambda}\cdot\boldsymbol{\phi}(x)}\psi \tag{1.17}$$

and using the minimal coupling rule

$$\mathcal{D}_\mu = \partial_\mu + ig\boldsymbol{\lambda}\cdot\boldsymbol{G_\mu} \tag{1.18}$$

for which 8 vector fields $\boldsymbol{G_\mu}$ must be introduced, that correspond to 8 different gluons, we work out other proprieties of these fields and the complete Lagrangian. Again, the fields must be massless, and due to the noncommutability of the $\boldsymbol{\lambda}$ matrices, manifested in the existence of nonzero structure constants, also the field strength tensors $G^{\mu\nu}$ must have an additional term to ensure the removing of unwanted terms,

$$G_a^{\mu\nu} = \partial^\mu G_a^\nu - \partial^\nu G_a^\mu - 2g\sum_{b,c=1}^{8} f_{abc}G_b^\mu G_c^\nu\,, \tag{1.19}$$

which results in gluons coupling with each other. The component of the Lagrangian for the quark flavor with mass m is then given by

$$\mathcal{L} = [i\bar{\psi}\gamma^\mu\partial_\mu\psi - m\bar{\psi}\psi] - \frac{1}{4}\mathbf{G}^{\mu\nu}\cdot\mathbf{G}_{\mu\nu} - (g\bar{\psi}\gamma^\mu\boldsymbol{\lambda}\psi)\cdot\mathbf{G}_\mu\,. \tag{1.20}$$

The strong interaction has two distinct properties that contrast to electromagnetic interactions. One is asymptotic freedom, which means that the strength of the interaction diminish at shorter distances and higher energies. Other is confinement, whereby the strength of the interaction increases with distance making that, as two quarks are separated, it becomes more energetically favorable to create new quark-antiquark pairs than to have free particles. Hence the observation of jets, a set of hadrons traveling together in a narrow cone, in High Energy Physics (HEP) experiments, where quarks and gluons tracks are expected as decay products. A third propriety is that physical particles must have electric charges that are integer multiples of the electron charge and that limits the possible combinations of quarks and antiquarks in composite particles.

1.1.3 Electroweak Theory

The weak interaction was recognized for the first time in nuclear β decays. It was in this context that the electron neutrino was introduced to satisfy energy conservation. It has several unique proprieties, namely, it can change the flavor of quarks, it violates parity (and also charge conjugation-parity symmetry), it is mediated by bosons with mass, actually, very significant masses around 90 GeV, which gives the interaction a very short range. Until the development by Salam [10], Weinberg [11], and Glashow [12] of a theory that unified electromagnetic and weak interactions, it was explained considering charged massive

vector intermediate bosons $W^{\pm}$ as force carriers and considering that only left-handed chiral states of leptons and right-handed chiral states of antileptons participated in the interaction. This theory predicted also a a third intermediate boson without electric charge, the Z^0. The gauge symmetry considered was $SU_L(2) \otimes U(1)_Y$, where L refers to the lef-handness of the particles interacting and Y refers to the hypercharge that is given by $Y = 2Q - 2I_3$ where Q is the electric charge, in units of the charge of the proton, and I_3 is the value of the third component of isospin, which components are the generators of SU(2), that is $I_i = 1/2\,\sigma_i$, where σ_i are the Pauli matrices. The fermions are arranged as left-handed doublets and right-handed singlets :

$$\begin{gathered} \psi_L^i = \begin{pmatrix} \nu_L^i \\ l_L^i \end{pmatrix}, \begin{pmatrix} u_L^i \\ d_L^i \end{pmatrix} \ , \\ \psi_R^i = l_R^i, u_R^i, d_R^i \end{gathered} \tag{1.21}$$

where i runs for the three generations of fermions. To build the gauge theory, the following covariant derivative must be used:

$$\mathcal{D}_\mu = \partial_\mu - ig\boldsymbol{I} \cdot \boldsymbol{W_\mu} - ig' \frac{Y}{2} B_\mu \,, \tag{1.22}$$

where g and g' are coupling constants, and were gauge fields were introduced, B_μ for the $U(1)_Y$ group and $\boldsymbol{W_\mu}$ for $SU_L(2)$. Also kinetic terms would be necessary as before, constructed once again with the field strengths tensors:

$$\begin{aligned} W_{\mu v}^i &= \partial_\mu W_\nu^i - \partial_v W_\mu^i - g\varepsilon^{ijk} W_\mu^j W_\nu^k \\ B_{\mu v} &= \partial_\mu B_\nu - \partial_\nu B_\mu \end{aligned} \tag{1.23}$$

where ε^{ijk} are the components of the totally antisymmetric Levi-Civita tensor,the structure constants of SU(2). Then the Lagrangian will result in:

$$\mathcal{L} = i\bar{\psi}\gamma^\mu \mathcal{D}_\mu \psi - \frac{1}{4}\left(W_{\mu\nu}^i W^{i,\mu\nu} + B_{\mu v} B^{\mu v}\right) \ . \tag{1.24}$$

Notice that the gauge bosons must be massless but also the fermions (in order to decouple right-handed and left-handed states). However, massive bosons are needed because that is what is observed. Another ingredient was necessary to obviate this problem. It is the Higgs mechanism based on spontaneous symmetry breaking.

1.1.4 The Brout-Englert-Higgs Mechanism

The mechanism of spontaneous symmetry breaking can be understood considering the following steps. Consider the Lagrangian [13] [14]:

$$\mathcal{L} = \frac{1}{2}(\partial^\mu \phi \partial_\mu \phi) + \frac{1}{2}\mu^2\phi^2 - \frac{1}{4}\lambda^2\phi^4 \,. \tag{1.25}$$

The Mexican hat potential

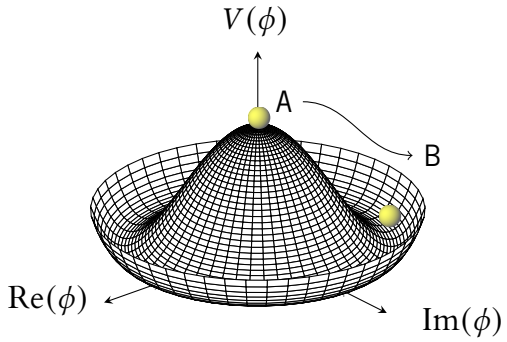


Figure 1.2: The Brout-Englert-Higgs potential.

Consider it like a sum of a kinetic part and the potential part. However, apparently, the mass term has the wrong sign. The minimum for the potential will occur, not for zero but for $\phi = \pm\frac{\mu}{\lambda}$. The Lagrangian must be rewritten in terms of perturbations around the ground state. If we write $\eta = \phi \pm \frac{\mu}{\lambda}$, the Lagrangian become,

$$\mathcal{L} = \frac{1}{2}(\partial^\mu \eta \partial_\mu \eta) - \mu^2\eta^2 \pm \mu\lambda\eta^3 - \frac{1}{4}\lambda^2\eta^4 + \frac{1}{4}\left(\frac{\mu^2}{\lambda}\right)^2 . \tag{1.26}$$

Then, the mass term is the second, corresponding to a mass $\sqrt{2}\mu$. Also, the symmetry ($\phi \longrightarrow -\phi$) that was present in 1.25 no longer holds in terms of the field η. Now, let us consider two fields,

$$\mathcal{L} = \frac{1}{2}\left(\partial_\mu\phi_1\right)\left(\partial^\mu\phi_1\right) + \frac{1}{2}\left(\partial_\mu\phi_2\right)\left(\partial^\mu\phi_2\right) + \frac{1}{2}\mu^2\left(\phi_1^2+\phi_2^2\right) - \frac{1}{4}\lambda^2\left(\phi_1^2+\phi_2^2\right)^2 . \tag{1.27}$$

This Lagrangian is invariant under rotations in the space of the linear combinations of both fields. The minimum for the potential is now given by any point in the circle:

$$\phi_1^2 + \phi_2^2 = 0\,. \tag{1.28}$$

Choosing the particular solution:

$$\phi_1 = \frac{\mu}{\lambda}, \phi_2 = 0 \tag{1.29}$$

and defining the fluctuations about that minimum:

$$\eta = \phi_1 - \frac{\mu}{\lambda}, \xi = \phi_2\,, \tag{1.30}$$

we can write the Lagrangian as

$$\begin{aligned}\mathcal{L} = &\left[\frac{1}{2}\left(\partial_\mu\pi\right)\left(\partial^\mu\pi\right) - \mu^2\eta^2\right] + \left[\frac{1}{2}\left(\partial_\mu\xi\right)\left(\partial^\mu\xi\right)\right] \\ &+ \left[\mu\lambda\left(\eta^3+\eta\xi^2\right) - \frac{\lambda^2}{4}\left(\eta^4+\xi^4+2\eta^2\xi^2\right)\right] + \mu^4/\left(4\lambda^2\right) .\end{aligned} \tag{1.31}$$

We recognize a Klein-Gordon field with mass $\sqrt{2}\mu$ and a free Lagrangian for ξ with no mass. Also, the original SO(2) symmetry is no longer to be seen. Rewriting 1.27 using:

$$\phi = \phi_1 + i\phi_2\,, \tag{1.32}$$

$$\mathcal{L} = \frac{1}{2}(\partial^\mu\phi)^*(\partial_\mu\phi) + \frac{1}{2}\mu^2(\phi^*\phi) - \frac{1}{4}\lambda^2(\phi^*\phi)^2\,. \tag{1.33}$$

We see that in this guise, the Lagrangian has the symmetry U(1) ($\phi \longrightarrow e^{i\theta}\phi$). Imposing that the Lagrangian must be invariant under local gauge transformations, and replacing partial derivatives with covariant derivatives, introducing massless vector fields for that effect, the Lagrangian becomes:

$$\begin{aligned}\mathcal{L} =& \frac{1}{2}\left[(\partial_\mu - iqA_\mu)\phi^*\right]\left[(\partial^\mu + iqA^\mu)\phi\right] \\ &+ \frac{1}{2}\mu^2(\phi^*\phi) - \frac{1}{4}\lambda^2(\phi^*\phi)^2 - \frac{1}{4}F^{\mu\nu}F_{\mu\nu}\,.\end{aligned} \tag{1.34}$$

Defining new fields, η and ξ as before, it becomes:

$$\begin{aligned}\mathcal{L} =& \left[\frac{1}{2}\left(\partial_\mu\eta\right)\left(\partial^\mu\eta\right) - \mu^2\eta^2\right] + \left[\frac{1}{2}\left(\partial_\mu\xi\right)\left(\partial^\mu\xi\right)\right] \\ &+ \left[-\frac{1}{4}F^{\mu\nu}F_{\mu\nu} + \frac{1}{2}\left(q\frac{\mu}{\lambda}\right)^2 A_\mu A^\mu\right] - 2i\left(\frac{\mu}{\lambda}q\right)\left(\partial_\mu\xi\right)A^\mu + \quad \text{interaction terms.}\end{aligned} \tag{1.35}$$

This procedure brought us a massless scalar boson ξ and the particle η with mass $\sqrt{2}\mu$, as before, but the field A^μ acquired mass. Still, some problems remain. Using the freedom given by the gauge invariance we can fix the gauge in such a way that, given

$$\phi = \eta + \frac{\mu}{\lambda} + i\xi\,, \tag{1.36}$$

we make:

$$\phi \longrightarrow e^{i\theta}\phi = (\cos\theta + i\sin\theta)(\phi_1 + i\phi_2) \tag{1.37}$$

real. This happens if we choose $\theta = -\arctan(\phi_2/\phi_1)$. Then, the unwanted terms disappear, including the massless boson. We are left with a massive vectorial particle and a scalar massive boson. This procedure was used to break U(1). In a similar manner, the electroweak symmetry can be broken.

Consider the doublet scalar field, called Higgs field:

$$\Phi = \begin{pmatrix}\phi^+ \\ \phi^0\end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix}\phi_1 + i\phi_2 \\ \phi_3 + i\phi_4\end{pmatrix} \tag{1.38}$$

containing an electrically charged and a neutral fields. A Lagrangian to govern this fields has the form:

$$\mathcal{L} = (\mathcal{D}_\mu)^\dagger(\mathcal{D}^\mu) - \left(-\mu^2\Phi^\dagger\Phi + \lambda^2(\Phi^\dagger\Phi)^2\right)\,, \tag{1.39}$$

where the covariant derivative is given in Equation 1.22, and the signs in the Higgs potential introduced after the kinetic term (and depicted in Figure 1.2) were chosen to have a stable minimum different from $|\Phi| = 0$. Actually, with this choice, there is an infinite set of degenerate minima in the ring:

$$\Phi^{\dagger}\Phi = \frac{1}{2}(\phi_1^2 + \phi_2^2 + \phi_3^2 + \phi_4^2) = \frac{v^2}{2} = \frac{\mu^2}{2\lambda^2}\,. \tag{1.40}$$

The fields can be parametrized as:

$$\Phi = \frac{1}{\sqrt{2}}\begin{pmatrix} 0 \\ v + h(x) \end{pmatrix}\,. \tag{1.41}$$

The term where to look to find the masses is:

$$(\mathcal{D}_\mu\Phi)^{\dagger}(\mathcal{D}^\mu\Phi) = |(\partial_\mu - ig\boldsymbol{I}\cdot\boldsymbol{W_\mu} - ig'\frac{Y}{2}B_\mu)\Phi|^2\,. \tag{1.42}$$

Making use of the identity $\boldsymbol{I} = \boldsymbol{\sigma}/2$ this results in:

$$\begin{aligned}(\mathcal{D}_\mu\Phi)^{\dagger}(\mathcal{D}^\mu\Phi) =& \frac{1}{2}\left(\partial_\mu h\right)\left(\partial^\mu h\right) + \frac{1}{8}g_W^2\left(W_\mu^{(1)} + iW_\mu^{(2)}\right)\left(W^{(1)\mu} - iW^{(2)\mu}\right)(v+h)^2 \\ &+ \frac{1}{8}\left(g_W W_\mu^{(3)} - g'B_\mu\right)\left(g_W W^{(3)\mu} - g'B^\mu\right)(v+h)^2\,.\end{aligned} \tag{1.43}$$

Equating the terms quadratic in the boson fields with:

$$\frac{1}{2}m_W^2 W_\mu^{(i)}W^{(i)\mu}\,, for\; i = 1,2 \tag{1.44}$$

follows the mass of the W boson:

$$m_W = \frac{1}{2}gv\,. \tag{1.45}$$

Using the relation:

$$\begin{pmatrix} Z_\mu \\ A_\mu \end{pmatrix} = \begin{pmatrix} \cos\theta_W & -\sin\theta_W \\ \sin\theta_W & \cos\theta_W \end{pmatrix}\begin{pmatrix} W_\mu^3 \\ A_\mu \end{pmatrix}\,, \tag{1.46}$$

follows

$$M_Z = \frac{1}{2}v\sqrt{g^2 + g'^2}\,, M_A = 0\,. \tag{1.47}$$

The three gauge bosons that intermediate the weak interaction gain mass and the photon rests massless. The Weinberg angle can be obtained by $\cos\theta_W = M_W/M_Z$. The fermion masses are obtained by terms like:

$$\mathcal{L}_{\mathrm{e}} = -g_{\mathrm{e}}\left[\begin{pmatrix} \bar{\nu}_{\mathrm{e}} & \bar{\mathrm{e}} \end{pmatrix}_L\begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix}\mathrm{e}_R + \bar{\mathrm{e}}_R\begin{pmatrix} \phi^{++}\phi^{0+} \end{pmatrix}\begin{pmatrix} \nu_{\mathrm{e}} \\ \mathrm{e} \end{pmatrix}_L\right] \tag{1.48}$$

Using the expansion of Equation 1.41, it follows:

$$m_{\mathrm{e}} = \frac{1}{\sqrt{2}}g_{\mathrm{e}}v\,. \tag{1.49}$$

From Equation 1.42 comes the mass of the Higgs boson:

$$m_H = \sqrt{2}\lambda v\,. \tag{1.50}$$

The values of g and g' can be obtained by $g \sin\theta_W = g' \cos\theta_W = e$. From Equation 1.45, the value of v, the vacuum expectation value, follows with $v \approx 246$ GeV [15]. Only the value of λ is left to determine the Higgs boson mass. So, this is a parameter to be determined by experiment.

1.2 Beyond the Standard Model Z' Boson

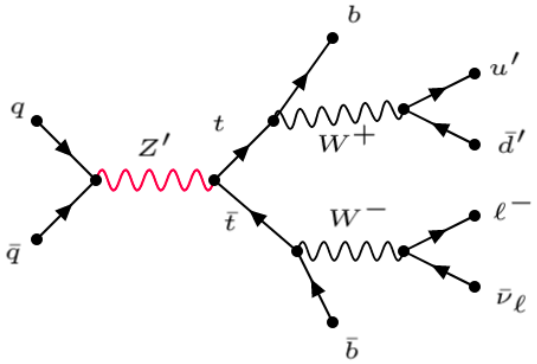


Figure 1.3: $t\bar{t}$ decaying semileptonically. u' and d' stand for up-type and down-type quarks, respectively, and ℓ for electron or muon.

Despite its accuracy and success, the SM is thought to be incomplete. The observation of neutrino oscillations indicate that neutrinos have mass, in contradiction to the Standard Model predictions. Astrophysical observations demand the existence of a different kind of matter not described by the SM, known only by its gravitational effects, called Dark Matter because it doesn't interact via the electromagnetic field. In cosmology, Dark Energy, assumed to exist to justify the acceleration of the expansion of the Universe, is also not explained by the SM. There is the matter-antimatter asymmetry, the fact that SM doesn't describe gravity, the large number of free parameters, etc.

Several extensions of the SM introduce a heavy, electrically neutral, spin-1 boson called Z', that decays into $t\bar{t}$. In this work a specific model will be used, that corresponds to a leptophobic, topophyllic Z' corresponding to a specific model of the topcolor-assisted-technicolor [16] [17] family, more concretely the Model IV [18], that couples only to first and third generation quarks. A search [19] performed by ATLAS Collaboration found no significant deviation from the Standard Model predictions but set exclusion limits on the production cross-section times branching ratio on the production of Z'. Namely, upper limits on the production cross section vary between 25 pb to 0.02 pb for masses from 0.4 TeV to 5 TeV. Masses of Z' lower than 2.6 TeV were excluded (c.f. Figure 1.4).

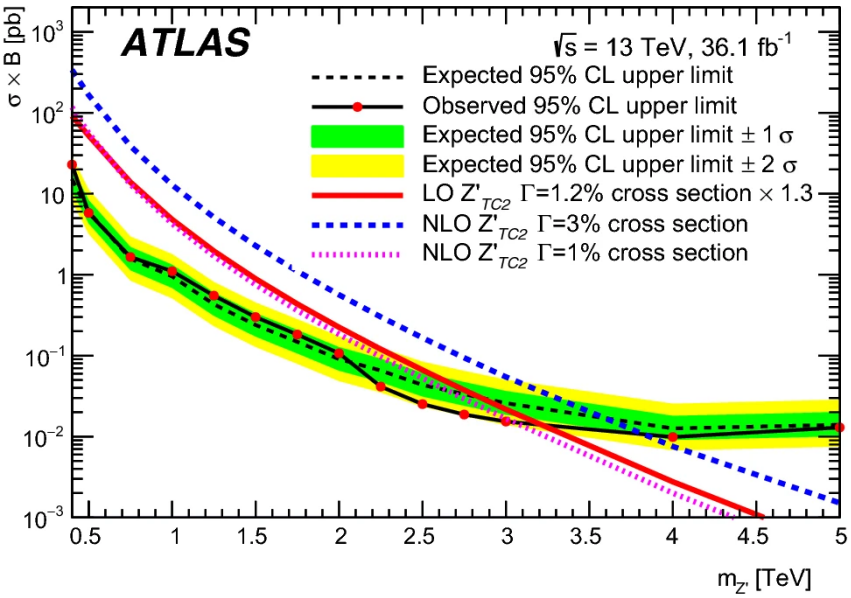


Figure 1.4: The observed and expected cross-section 95% CL upper limits on the Z' signal and the theoretical predictions for the production cross-section times branching ratio of Z' [19].

The high center-of-mass energy of LHC allows it to be a top factory. The top quark decays almost exclusively to a W^+ boson and a quark bottom. Moreover, having a very short lifetime ($\approx 4 \times 10^{-25}$ s) it do not form hadrons. The semileptonic channel of the pair top antitop (see Figure 1.3) is chosen for study because it is a good compromise between competing advantages. For one part, the all-hadronic channel, where both W bosons decay into quarks, is the dominant channel (46%), which has the advantage of providing more statistics. But the final state of this decay consists in six jets what makes it hard to distinguish from QCD multijets events. Furthermore, the large multiplicity of jets makes difficult the task of ascribing each one to the correct top quark from which they decay. The dileptonic channel where both W bosons decay to leptons has a small branching ratio (9%) which is a drawback in terms of statistics. Its advantage is that it is more easy to distinguish its final states from QCD multijets. It has the drawback, though, of having two undetected particles. Lastly, the semileptonic channel has a branching ratio comparable to the all-hadronic channel (45%) but the decay $W \longrightarrow \tau\nu_\tau$ (15%) is usually not considered in analyses because it introduces an additional neutrino (missing energy) when it decays via the weak interaction ($\tau \longrightarrow W\nu_\tau$). But it decays with more probability into quarks, giving rise to more jets and making hard the task of event reconstruction. In this way, omitting this decay, the semileptonic channel is less statistically advantageous than the all-hadronic but still good (30%) [20]. It is, although, easier to distinguish its final state from background and more suitable for event reconstruction.

Chapter 2

Experimental Setup

2.1 LHC

The Large Hadron Collider (LHC) is a particle accelerator at the European Laboratory for Particle Physics (CERN), in Geneva, Switzerland. It consists of a ring of superconducting magnets located underground, having a circumference of 26.7 km. [22] It is divided in 8 archs and 8 straight sections between them. At 4 of these sections are located four main detectors: ATLAS [23], ALICE [24], CMS [25] and LHCb [26]. Of these, CMS and ATLAS are general-purpose detectors, LHCb is dedicated to the study of B physics, and ALICE to the study of heavy-ion physics. LHC was designed to achieve a center-of-mass energy of 14 TeV, through proton-proton collisions. Lead-lead and proton-lead collisions are also carried out. The purpose is to firmly establish the validity of the SM (namely, the discovery of the Higgs boson [27] [28] announced on 4th July 2012, is a highlight of the goals accomplished so far) and to find new physics beyond it.

Protons are accelerated in opposite beams, each one inside its own pipeline (kept at ultra-high vacuum) until they reach the wanted energy. They are then focused to the point of interaction inside the detectors. Beams consist of bunches of particles, around 25 ns apart [1], guided in the pipelines by very strong magnetic fields created by the superconducting magnets operating at 1.9 K for which a cooling system based on liquid helium is necessary.

The acceleration process is done in several phases as illustrated in Figure 2.1. First, protons are collected from a container where hydrogen molecules are split into electrons and protons after application of an intense electric field. Then, they are accelerated in a linear accelerator, called LINAC 2, until they reach an energy of 50 MeV. After, they are injected into the proton synchrotron booster (PSB) where they reach 1.4 GeV before they enter the Proton Synchrotron (PS), that accelerate them until 26 GeV. After, they are injected into the Super Proton Synchrotron (SPS) and are accelerated to 450 Gev. At last the enter the LHC. In the second run of the LHC, from 2015 to 2018, the protons collide with a center-of-mass energy of $\sqrt{s}$ of 13 TeV.

[1] This refers to run 2 at which the data analyzed in this work was collected.

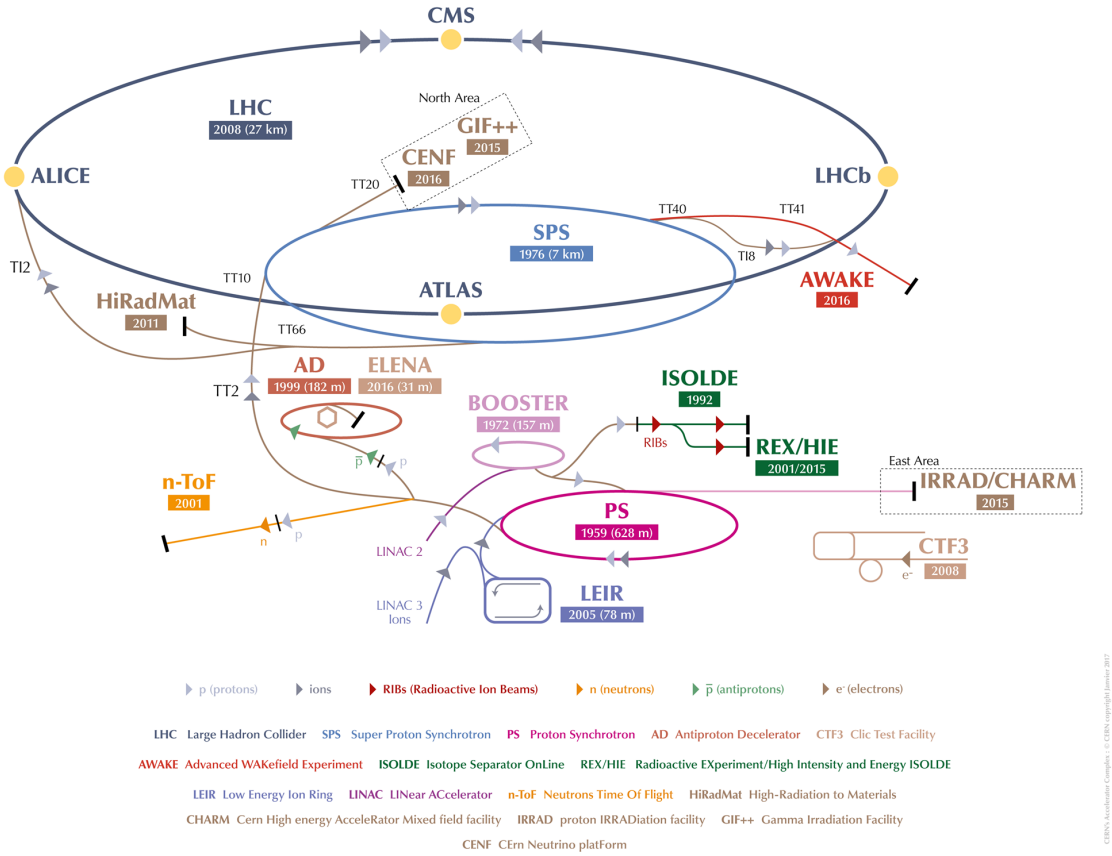


Figure 2.1: The several components of the CERN accelerator complex. [21]

In the LHC, acceleration is achieved using superconducting dipole magnets for beam bending, quadrupole magnets for focusing, and radio-frequency cavities for accelerating. In each cavity protons receive 2 MeV per pass, clumping around the synchronous particle (which is the one that is exactly synchronized with the radio-frequency), forming bunches that are separated by 25 ns (in Run 2).

For a given process the number of events generated in collisions is given by:

$$N = \sigma \int L(t)dt\,, \tag{2.1}$$

where σ is the cross-section for that process and L is the instantaneous luminosity. For two bunches colliding head-on with frequency f having n_1 and n_2 particles, the luminosity is given by:

$$L = f\frac{n_1 n_2}{4\pi\sigma_x^*\sigma_y^*}\mathcal{F}\,, \tag{2.2}$$

where σ_x^* and σ_y^* characterize the transverse dimensions of the beam, horizontally and vertically, and $\mathcal{F}$ is a factor of order 1 that takes in account several geometrical effects.

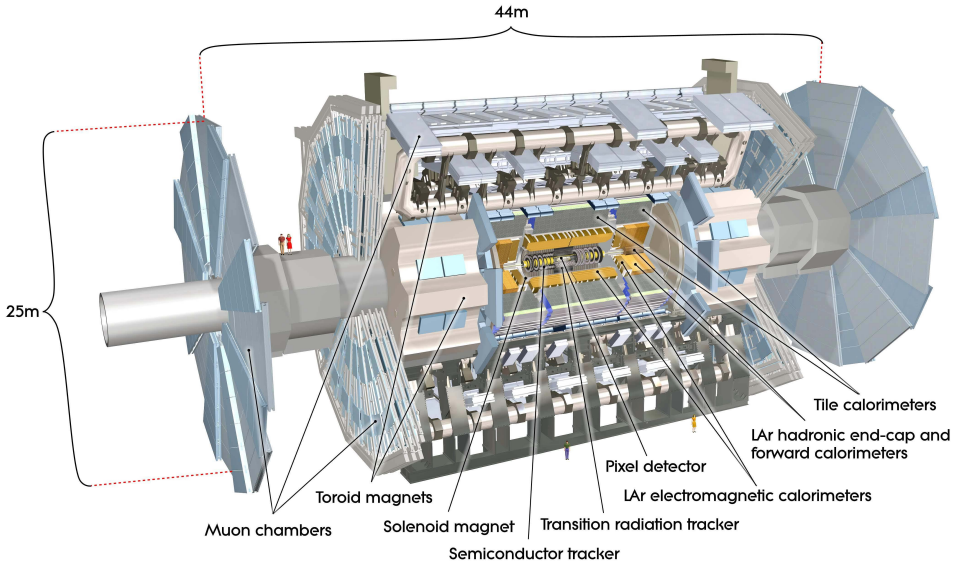


Figure 2.2: Cut-away view of the ATLAS detector. [23]

2.2 ATLAS

ATLAS (A Toroidal LHC ApparatuS) is a cylindrically symmetrical general purpose detector built for probing proton-proton and heavy-ion heavy-ion (in particular lead nuclei) collisions [23] and was optimized for the study of a broad range of processes, including Higgs boson searches and BSM. It weighs 7000 metric tons, is 44 m long, and has a diameter of 25 m. It lays at a depth of 100 m. It surrounds one of the LHC collision points. A coordinate system is defined, associated with it, with its origin in the nominal interaction point. The beam direction defines the z-axis and the plane transverse to it defines the x-y plane. The positive x-axis point to the center of the LHC ring whereas the positive y-axis points upwards. The azimuthal angle ϕ is measured around the beam direction. The polar angle θ is measured from the beam axis. Another related variable commonly used is the pseudorapidity η defined as $-\ln\tan(\theta/2)$, but for massive objects like jets, rapidity is used:

$$y = \frac{1}{2}\ln\left[\frac{E+p_z}{E-p_z}\right], \tag{2.3}$$

as η is the approximation of y when the mass is zero. Differences of rapidity are boost invariant, for boosts in the z-direction, so ΔR, an angular distance defined as $\Delta R = \sqrt{\Delta\eta^2 + \Delta\phi^2}$, is boost invariant. When protons collide at very high energy only a quark or a gluon from each proton interact, the rest of the components go along with the beam and are not observed. Only transverse energy and transverse momentum are observed and can be balanced. Neutrinos are inferred from missing transverse momentum. Transverse momentum p_T and transverse energy E_T are defined in the transverse plane. Missing transverse momentum is given by $\vec{p}_T^{\,miss} = -\sum \vec{p}_T$ and the missing transverse energy $E_T^{miss} = |\vec{p}_T^{\,miss}|$.

ATLAS is composed of several sub-detectors to track and identify different kinds of particles and

measure their energy and momentum. An inner tracking detector (ID) immersed in a 2 T magnetic field parallel to the beam axis measures the charge and momentum of electrically charged particles. The energy of electrons and photons is measured by an electromagnetic calorimeter (ECAL) that surrounds the ID. Around it, a layer of calorimeters is used to measure the energy of hadrons. It act as well as an absorber, letting pass only the energetic muons and the feebly interacting neutrinos. The outermost layer is a muon spectrometer. A two-level trigger system is in place, due to the necessity to select the interesting events from the many produced at a very high rate in the collisions.

Detector component	Required resolution	η coverage	
		Measurement	Trigger
Tracking	$\sigma_{p_T}/p_T = 0.05\% p_T \oplus 1\%$	± 2.5	
EM calorimetry	$\sigma_E/E = 10\%/\sqrt{E} \oplus 0.7\%$	± 3.2	± 2.5
Hadronic calorimetry (jets)			
barrel and end-cap	$\sigma_E/E = 50\%/\sqrt{E} \oplus 3\%$	± 3.2	± 3.2
forward	$\sigma_E/E = 100\%/\sqrt{E} \oplus 10\%$	$3.1 < \lvert\eta\rvert < 4.9$	$3.1 < \lvert\eta\rvert < 4.9$
Muon spectrometer	$\sigma_{pT}/p_T = 10\%$ at $p_T = 1TeV$	± 2.7	± 2.4

Table 2.1: The units for E and p_T are in GeV. [23]

2.2.1 Inner Detector

ID is the inner component, the closest to the interaction point. It tracks charged particles, allowing the measurement of their momenta and the sign of the charges. For that purpose, a solenoid that provides a 2 T magnetic field bends the path of charged particles. Because it is located in the area with the higher density of particle tracks, it requires a high momentum and vertex reconstruction resolution. It provides vertex reconstruction within $|\eta| < 2.5$. It is composed of four sub-detectors. The Insertable B-Layer (IBL), is the innermost component of the ID. It was added in the 2014 upgrade for a precise identification and localization of b-jets. A pixel detector (Pixel), close to the beam pipe and the Semiconductor Tracker (SCT), a silicon microstrip detector, at intermediate radii, are used to reconstruct the origin of the particles. The Transition Radiation Tracker (TRT), at outer radii, provides additional spacial measurements and information about the particle type and covers a region of $|\eta| < 2.0$.

2.2.2 Calorimeters

Calorimeters are the detectors that measure the particles energy. They consist in material that absorb the incident particles, converting the energy deposited in measurable signals. They usually are segmented transversely, to gather information about the direction, in addition, to the particles energy. At ATLAS, the calorimeter system, that covers the region $|\eta| < 4.9$, is composed of the ECAL, for precise measurements of the energy of photons and electrons, and the hadronic calorimeter (HCAL) for the measurement of the

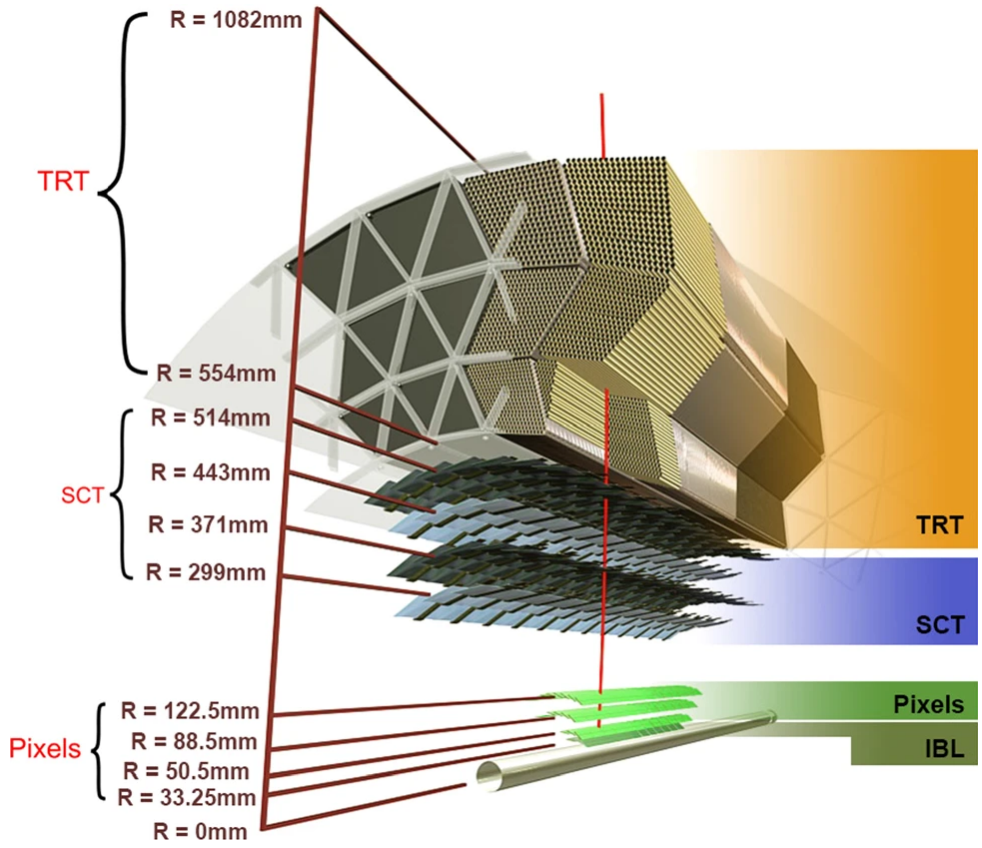


Figure 2.3: Cut-away view of the ATLAS inner detector. [29]

energy of hadrons. The ECAL has three parts: one barrel section that surrounds the ID and the other two in the opposed endcaps of the ID barrel. All use liquid argon as the active material (for energy measurement). The absorber material (for stopping the incoming particles) used by the central part is lead. In the caps, copper is used for that end. The hadronic calorimeter, also composed of three parts, use liquid argon as active material in both caps, and as absorber, copper is used in the endcap calorimeter (HEC) and tungsten in the forward calorimeter (FCal). In the central part, the Tile Hadronic Calorimeter (TileCal), scintillating tiles are used as the active material and steel as absorber material.

2.2.3 Muon Spectrometer

The Muon Spectrometer (MS) measure the momentum of muons, in a range of p_T, between 3 GeV and 1 TeV. This is done deflecting the track of muons by super-conducting toroidal magnets, within a region of $|\eta| < 2.7$. The measurement is done with four different types of muon chambers. Resistive Plate Chambers (PRCs) and Thin Gap Chambers (TGCs) are used for triggering and (η, ϕ) position measurements. Monitored Drift Tubes (MDTs) and Cathode Strip Chambers (CSCs) are used for precise muon track measurement.

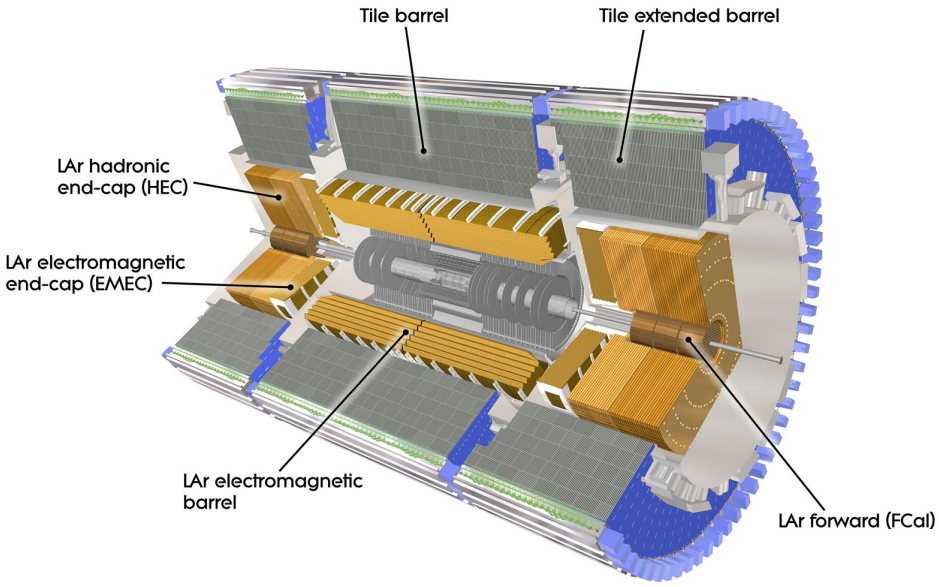


Figure 2.4: Cut-away view of the ATLAS calorimeter system. [30]

2.2.4 Trigger and Data Acquisition systems

ATLAS was designed to produce more than a billion events per second, which amounts to tens of millions of megabytes per second. Technical limitations do not allow to record all this information, making necessary to select only the interesting events that can lead to new discoveries. For this purpose a two-level trigger system [32] is used. The second level refines the decisions made by the previous level, and apply additional selection criteria when necessary.

The data acquisition system receives and buffers the data from readout electronics from the detectors and direct it to the first level, called L1 trigger, which is hardware-based. It uses just a subset of information from the calorimeters and the muon system, namely, reduced granularity information. This is combined in a Central trigger to make a decision in less than 2.5 μs, reducing the rate to 100 KHz. At this level, searches occur for high transverse momentum of muons, electrons, photons, jets, large missing transverse energy and large total energy. Also, it is at this point that regions of interest in the detector are defined. In the next level, the information from these regions are used with full granularity. The second level is software-based, called the High-Level Trigger (HLT), and operates from a farm of about 40,000 CPU cores. It takes only 200 μs to make further decisions after elaborate analysis of each event, accessing the additional data from the specific regions of the detector chosen in the first stage. That results in a reduction to about 1000 events per second, stored for offline analysis, corresponding to 1.3 megabyte per event.

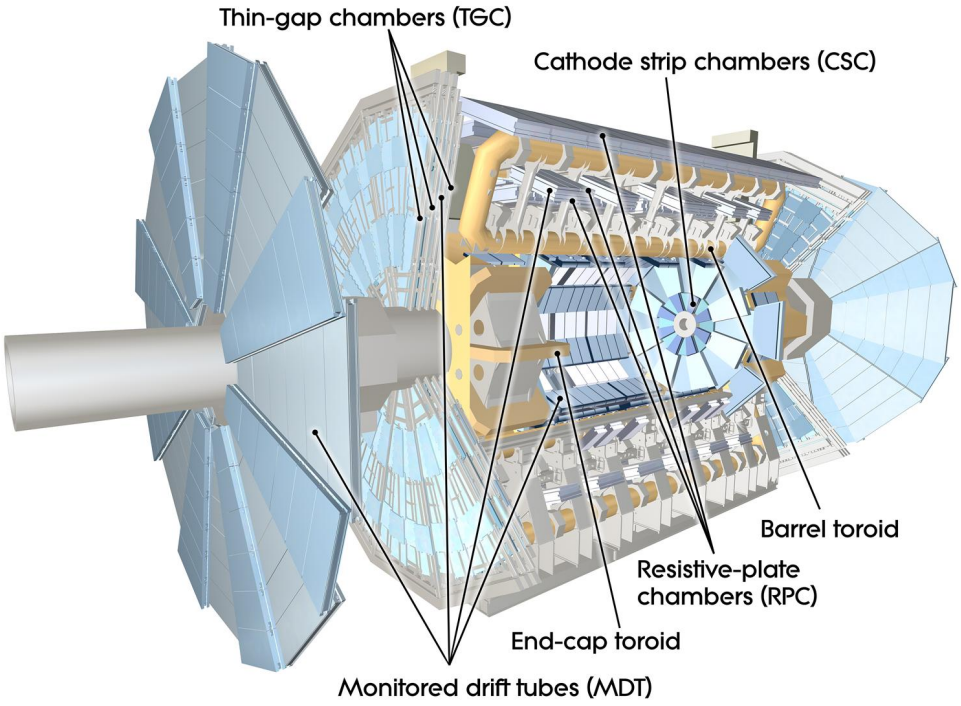


Figure 2.5: Cut-away view of the ATLAS Muon Spectrometer. [31]

Chapter 3

The 13 TeV ATLAS Open Dataset

The 13 TeV ATLAS Open Dataset [33] is a collection of data, Monte Carlo (MC) simulations and tools provided by the ATLAS experiment to be used, among other things, for educational purposes and Machine Learning challenges. It provides tools that implement examples of some physics analysis. These can be the point of departure for further analysis and development of techniques to perform them. Data is available in the ROOT [34] format. The data is comprised of events from 61 runs that were collected in 2016 by the ATLAS detector from p-p collisions at $\sqrt{s} = 13$ TeV. It consists of approximately 270 millions events. Only events recorded when all subsystems of ATLAS were working acceptably were included. Also quality criteria related to the beam and data were imposed. In total, the dataset correspond to an integrated luminosity of $10.06 \pm 0.37\,\text{fb}^{-1}$.

In addition to the data, the 13 TeV ATLAS Open Dataset includes MC simulations that describe various SM processes used to model expected signal and background. Data and MC simulations are submitted to the same quality and trigger criteria as the data. In the end of a loose preselection, performed to reduce subsequent processing time, they are grouped in collections according to type and multiplicity of reconstructed objects with high transverse momentum. For our analysis, the collection of interest is the one labeled *1largeRjet1lep*, where among the final state reconstructed objects are at least one jet with large-R with minimum p_T of 250 Gev and exactly one charged lepton with minimum p_T of 25 GeV.

3.1 Preselection and particle identification

Electron candidates are reconstructed matching isolated energy deposits (clusters) in the ECAL to tracks in the ID. It is considered only the precision region of the ATLAS detector, called fiducial, defined by $|\eta_{\text{cluster}}| < 2.47$. It is necessary, in addition, to exclude the transition region between the barrel and the endcap of the ECAL, $1.37 < |\eta_{\text{cluster}}| < 1.52$. The candidate must have $p_T > 7$ GeV and pass *loose* identification criteria [35]. Very loose, loose, medium or tight criteria refer to how tight the match

between the track and the cluster is enforced, depending on the particle identification efficiency required. The identification is made using a likelihood-based discriminant and the loose operating point correspond to 93 % efficiency for identifying a prompt electron with $E_T = 40$ GeV. For muons, the reconstruction is based in matching tracks in the ID and in the Muon Spectrometer. The muon candidate must also have $p_T > 7$ GeV and pass *loose* identification criteria [36]. These criteria aims to maximize the reconstruction efficiency while keeping good-quality muon tracks. Isolation criteria are imposed on muons and electrons to reduce contributions from unwanted sources, such as non-prompt leptons, photon conversions and hadrons. Events containing at least an electron or a muon are selected with single-lepton triggers with p_T threshold of 26 GeV and isolation requirements, or with a larger threshold of 50-60 Gev, looser identification requirements and no isolation requirement.

The reconstruction of photon candidates, like for electron candidates, is based on detecting energy clusters in the ECAL not matched by any track in the ID, and, additionally, searching for a process of photon conversion into e^- e^+ at the ID and corresponding clusters in the ECAL. To reduce hadronic background, the photon candidates must obey "loose"isolation criteria. Other criteria are shown in the Table 3.1.

Electron (e)	Muon (μ)	Photon (γ)
ID & ECAL rec.	ID & MS rec.	ID & ECAL rec.
loose identification	loose identification	tight identification
loose isolation	loose isolation	loose isolation
$p_T > 7$ GeV	$p_T > 7$ GeV	$E_T > 25$ GeV
$\|\eta\| < 2.47$	$\|\eta\| < 2.5$	$\|\eta\| < 2.37$

Hadronically decaying τ-leptons (τ_h)	Small-R jets	Large- R jets
ID & ECAL rec.	ECAL & HCAL rec.	ECAL & HCAL rec.
medium identification	anti- k_t, R =0.4	anti- k_t, R =1.0
$P_T > 20$ GeV	$P_T > 20$ GeV	$P_T > 250$ GeV
$\|\eta\| < 2.5$	$\|\eta\| < 2.5$	$\|\eta\| < 2.0$
1 or 3 associated tracks	b-tagging (MV2c10)	trimming: $R_{\text{sub}} = 0.2$, $f_{\text{cut}} = 0.05$

Table 3.1: Preselection requirements.

Jet candidates are reconstructed using three-dimensional energy clusters in ECAL and HCAL using a specific algorithm called anti-k_t[1] with radius parameter of 0.4 for the "small-R" jets. They must fulfil $|\eta| < 2.5$ and $p_T > 20$ GeV. One effect that needs to be minimized is "pile-up". It consists of low transverse momentum collisions other than the hard-scatter p-p collisions that are of interest, which would hide the rare events that we want to study. These additional collisions from the same bunch crossing are called

[1]The procedure goes as follow. The pairwise distance between objects is computed. The two closest objects are merged and the procedure is repeated until no pair of particles are closer than a distance R. Also, a distance between the object i and the beam, d_{iB} is considered. If that distance is larger than d_{ij}, the pairwise distance between two objects, those objects are merged, otherwise the object is considered a jet and no more merging is performed. In the anti-k_t algorithm, d_{iB} is given by $\frac{1}{p_{Ti}^2}$ and d_{ij} is given by $min(\frac{1}{p_{Ti}^2}, \frac{1}{p_{Tj}^2})R_{ij}^2/R^2$, where $R_{ij} = \sqrt{(\eta_i - \eta_j)^2 + (\phi_i - \phi_j)^2}$ [37].

in-time pile-up. Also, energy deposits associated with previous or following bunch crossings relative to the triggered event must be dealt with (due to the response time of subdetectors being larger than the interval between successive bunch crossings), and they are called out-of-time pile-up. To reduce this effect, a condition is imposed on the score of the jet vertex tagger (JVT) [38] discriminant, for jets with $p_T <$ 60 GeV and $|\eta| < 2.4$.

Large-R jets are built using also the anti-k_t algorithm with R = 1.0. Then, the are trimmed. [39] which reduces the effects of pile-up. This consists of recluster the components of the large-R jet into subjets with a R_{sub} parameter. Subjets with transverse momentum lesser than a fraction f_{cut} of the original jet are discarded. The parameters $R_{\text{sub}} = 0.2$ and $f_{\text{cut}} = 0.05$ are chosen based on a study of sensitivity to pile-up [40]. Large-R jets must have $|\eta| < 2.0$ and $p_T > 250$ GeV. The visible products of the τ-lepton decays are also reconstructed.

It is very important to identify jets containing B-hadrons (b-tagging) for various physical analysis as this allows a huge rejection of background processes. The present analysis demands to find jets originated from b quarks into which top quarks decay almost exclusively, plus a W boson, by way of the weak interaction. Because B-hadrons decay very close to the point of creation (around 0.5 mm, corresponding to a decay time of 10^{-12} s), leave therefore a secondary vertex. This vertex can be reconstructed by the convergence of tracks to a vertex very near the primary vertex [2] or if it cannot be resolved, the impact parameters of tracks relative to the primary vertex are used. In the pre-selection that is being presented, a multivariate discriminant, MV2c10, is used, that combines this information. For each jet, the value for the discriminant is calculated. For a required efficiency of b-tagging there is a threshold, the working point (WP), the value of which the discriminant must surpass [41].

After this pre-selection, data quality criteria are applied to guarantee that detectors were correctly functioning and tracks were not reconstructed from deposits that were due to cosmic-ray showers, or hardware problems. Also, events must contain at least one reconstructed vertex with at least two tracks with $p_T > 0.4$ GeV. The tracks associated to muons and electrons must correspond to the primary vertex of the event.

Several SM processes that can mimic the signal were simulated, namely the production of $t\bar{t}$, single-top, W plus jets, Z plus jets and diboson. In addition, simulations of some BSM processes, namely, the Z' production are included. They are listed in Table 3.2 with references to the software used to perform the MC simulations.

The 13 TeV ATLAS Open Dataset includes several SM and BSM physics analyses. For all, the reconstructed physics objects are subject to additional selection (Table 3.3) that correspond to the requirement of calorimeter (etcone20) and track (ptcone30) isolation for electrons, photons and muons. etcone20 is given by the sum of the energy of the clusters located inside a cone of $\Delta R = 0.2$ around the object considered and ptcone30 is defined as the scalar sum of the p_T of tracks within a cone of $\Delta R = 0.3$.

[2] This is the vertex that corresponds to highest sum of squared transverse momentum of the tracks associated with it.

Process	Generator, hadronisation	Additional information
Top-quark production		
$t\bar{t}$ + jets	POWHEG-BOX v2 + PYTHIA 8	only 1 ℓ and 2 ℓ decays of $t\bar{t}$-system
single (anti)top t-channel	POWHEG-BOX v1 + PYTHIA 6	
single (anti)top W t-channel	POWHEG-BOX v2 + PYTHIA 6	
single (anti)top s-channel	POWHEG-BOX v2 + PYTHIA 6	
W/Z (+ jets) production		
Z $\rightarrow ee, \mu\mu, \tau\tau$	POWHEG-BOX v2 + PYTHIA 8	LO accuracy up to N_{jets} = 1
W $\rightarrow e\nu, \mu\nu, \tau\nu$	POWHEG-BOX v2 + PYTHIA 8	LO accuracy up to N_{jets} = 1
W $\rightarrow e\nu, \mu\nu, \tau\nu$ + jets	SHERPA 2.2	LO accuracy up to 3-jets final states
Z $\rightarrow ee, \mu\mu, \tau\tau$ + jets	SHERPA 2.2	LO accuracy up to 3-jets final states
Dibosons production		
WW	SHERPA 2.2	$qq'\ell\nu$ final states
WW	SHERPA 2.2	$\ell\nu\ell'\nu'$ final states
ZZ	SHERPA 2.2	$qq'\ell^+\ell^-$ final states
ZZ	SHERPA 2.2	$\ell^+\ell^-\ell'^+l'^-$ final states
WZ	SHERPA 2.2	$qq'\ell^+\ell^-$ final states
WZ	SHERPA 2.2	$\ell\nu qq'$ final states
WZ	SHERPA 2.2	$\ell\nu\ell^+\ell^-$ final states
WZ	SHERPA 2.2	$\ell\nu\nu'$ final states
Z' production		
Z' $\rightarrow t\bar{t}$	PYTHIA 8	$m_{Z'}$ =1 TeV

Table 3.2: MC samples contained in 13 Tev ATLAS Open Dataset used in the analysis.

Large R-jets with mass lower than 50 GeV or p_T larger than 1500 GeV are excluded because they belong to a region of phase-space that is not well-calibrated.

Electrons and Muons	Small-R jets	Photons	Large-R jets	τ_h
p_T > 25 GeV lep_ptcone30 < 0.15 lep_etcone20 < 0.15	p_T > 25 GeV JVT > 0.59	photon_ptcone30 < 0.065 photon_etcone20 < 0.065	p_T < 1500 GeV mass > 50 GeV	p_T > 25 GeV

Table 3.3: Additional object selection.

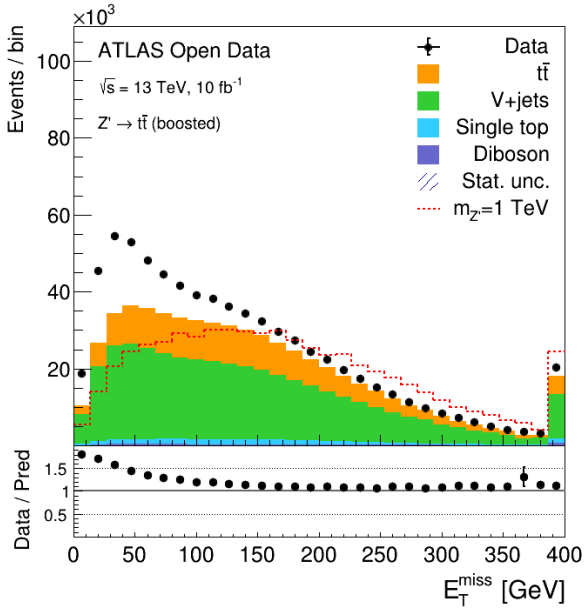
3.2 Search of the decay of Z' into top quark pairs

This work will compare a machine learning based analysis to the search of the decay of Z' into top quark pairs in events that contain a single charged lepton, large R-jets and missing momentum with the analysis included in the 13 Tev ATLAS Open Dataset that will be described below. The analysis is based in previous searches [19] for decays of heavy particles to top-quark pairs in pp collisions at $\sqrt{13}$ = 13 TeV with ATLAS. It applies a selection known as single-lepton boosted topology to the final products of a top and antitop decay. [42]. This correspond to semileptonic decays of the $t\bar{t}$ system as exemplified in Figure

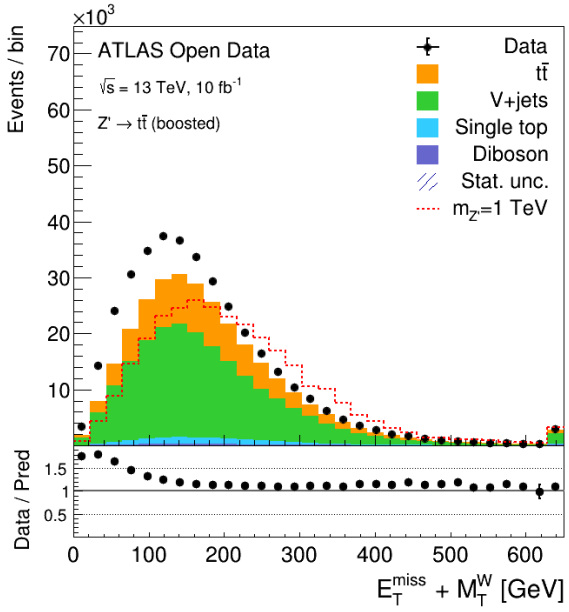
1.3, where the decay products of the top quark that decays hadronically quark are enclosed within one large-radius jet. Tops decay into a W boson and a bottom quark. One W boson decays into a an electron or a muon plus a neutrino while the other decays into quarks. An all-hadronic topology is more frequent but is more difficult to separate from non-$t\bar{t}$ background. This selection requires a single isolated electron or muon, large missing transverse momentum and hadronic jets, one of which must contain a b-hadron. More concretely, the final event selection criteria are:

1. There must be at least a preselected large-R jet.
2. The missing transverse momentum E_T^{miss}, defined as the magnitude of the negative of the vector sum of the transverse momentum of all selected physics objects, must be greater than 20 GeV.
3. Single-electron or single-muon trigger satisfied.
4. There is exactly on good lepton, that is, one for which $p_T > 30$ GeV, and the track associated with it must must match the candidate that triggered the event. Also, the identification must be tight.
5. In order to have a situation consistent with a leptonic W decay, additionally to point 2, $E_T^{\text{miss}} + M_T^W$ must be larger than 60 GeV, where M_T^W is the transverse mass of the W boson candidate (the selected lepton ℓ plus the E_T^{miss}), given by $M_T^W = \sqrt{2p_T^{\ell} E_T^{\text{miss}}(1 - \cos\Delta\phi(\ell, E_T^{\text{miss}}))}$ [3].
6. At least, one small-R jet close to the lepton, that is, ΔR(lepton, jet) < 2.0. This would correspond to the leptonic W decay.
7. There must be exactly one large-R jet that pass simplified requirements to be compatible with a hadronically top decay (top-tagged), namely, to have mass larger than 100 GeV and to have N-subjettiness ratio [43] $\tau_{32} < 0.75$. $\tau_{32} = \frac{\tau_3}{\tau_2}$, τ_N expressing how well a jet can have N or fewer subjets. This variable allows the discrimination between jets containing three subjets and jets containing two subjets. Also, in addition to the conditions already stated for a well-calibrated region of phase space, we must have $p_T > 300$ Gev and $|\eta| < 2$.
8. The large-R jet of item 7 must be well apart from the small-R jet ($\Delta R > 1.5$) and from the lepton ($\Delta\phi > 1.0$).
9. At least one b-tagged jet (with the requirement of WP corresponding to 70% of efficiency of tagging). This jet must be within the top-tagged large-R jet or to be the small-R jet close to the lepton, that is ΔR(large-R jet, b-tagged jet) < 1.0 or ΔR(small-R jet, b-tagged jet) < 0.01.

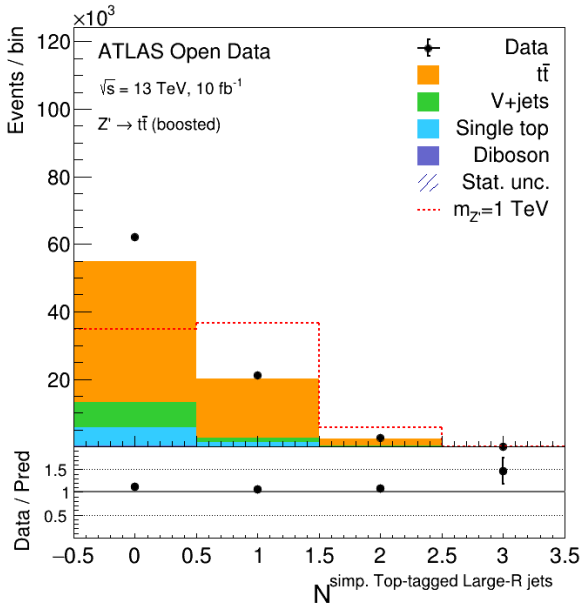
[3] This quantity is defined in a special manner when we have a particle to decay into two particles, of which one is invisible [20]. In this case, $M_T^2 = (E_T(1) + E_T(2))^2 - (\vec{p}_T(1) + \vec{p}_T(2))^2$, where $E_T^2 = m^2 + \vec{p}_T^2$ is the transverse energy. For the case where the daughter particles are massless or can be considered as such, as in the case under consideration, the result has the form stated.



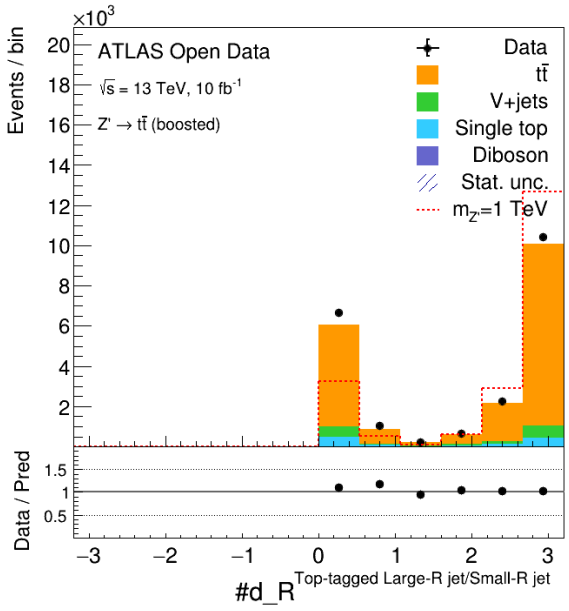
(a) Missing transverse momentum with application of cuts 1, 3 and 4.



(b) Application of the same cuts as in Figure 3.1a.

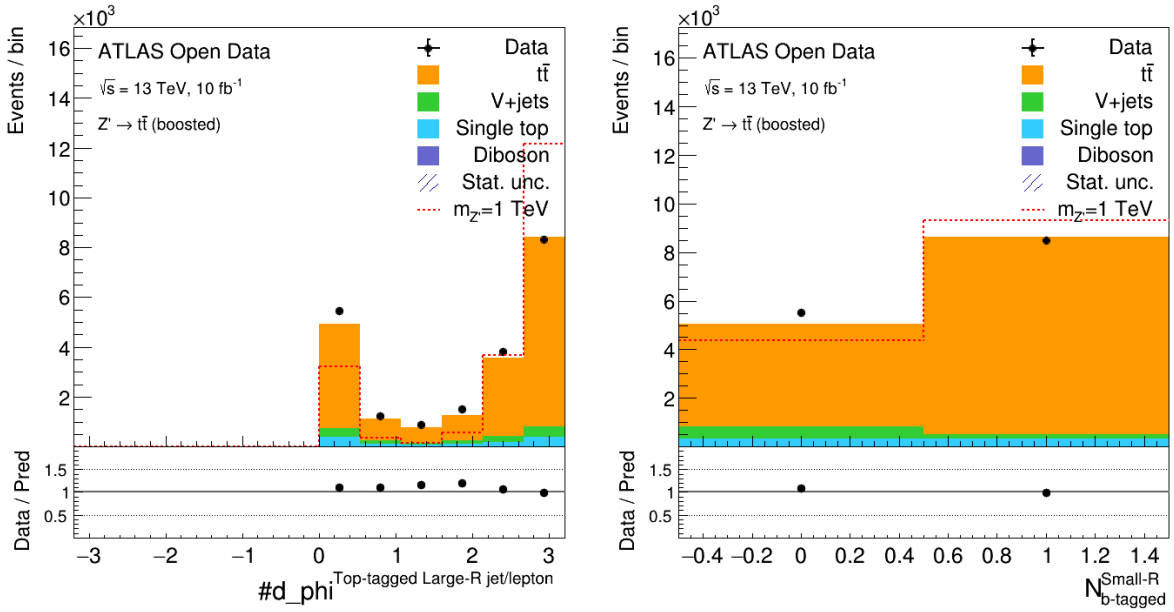


(c) Number of large-R jets that pass simplified top-tagging requirements.Application of the same cuts as in Figure 3.1a plus cuts 2, 5, 6 and at least one b-tagged jet.



(d) Application of all cuts except 8 and 9. The variable is the ΔR referred in 8.

Figure 3.1: Relevant plots to justify the cuts enumerated above.



(e) Application of all cuts except 8 and 9. The variable is the $\Delta\phi$ referred in 8.

(f) Application of all cuts except 9. This variable tells if the small-R jet described in that item is b-tagged.

Figure 3.1: Relevant plots to justify the cuts enumerated above. (continuation)

Sample	Number of Events
Single Top	610.78
Diboson	22.34
$t\bar{t}$	13600.48
V + jets	682.68
Z' 1 TeV	426.62

Table 3.4: Expected number of selected events for a luminosity of 10 fb^{-1}.

In Figures 3.1a and 3.1b we can see that it is justified to cut in the missing energy as there are disagreement between data and MC samples at low missing energy because multijet background was not simulated due to its large cross-section, which would require simulating a huge number of events. In Figure 3.1c it is possible to see that demanding exactly one large-R jet that pass simplified top-tagging requirement increases the signal to background ratio. The same reasoning applies for Figures 3.1d and 3.1e that justify the cuts 8 and Figure 3.1f to justify cut 9.

The number of selected events are shown in Table 3.4 distributed according to the type of background.

In Figure 3.2 is shown an observable that approximates [4] the mass of the top-antitop system, adding the four-momenta of the charged lepton, of the b-tagged small-R jet and of the top-tagged large-R jet. It can be seen that the Standard Model prediction is consistent with the data, from which the simulated hypothesized Z' model deviates considerably.

[4] For simplicity, the four-momentum of the neutrino is not included as it would involve the difficult reconstruction of the longitudinal component of the missing energy.

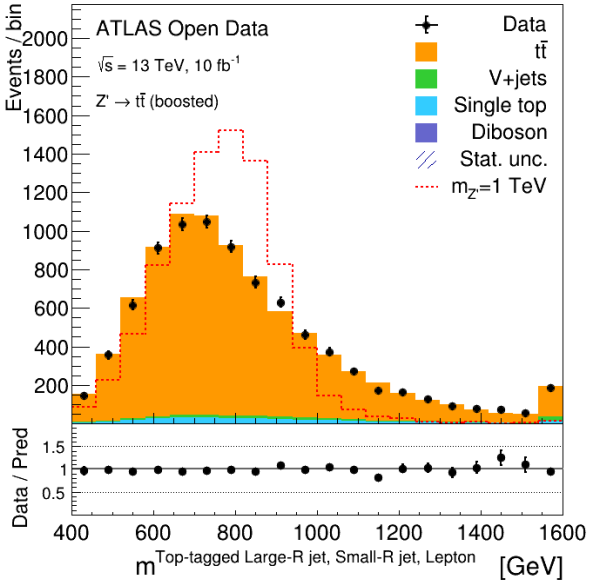


Figure 3.2: Approximate mass of the top-antitop system.

Chapter 4

Deep Neural Networks

4.1 Machine Learning

Artificial Intelligence is the area of computer science concerned with implementing the ability to perform tasks usually associated with intelligent beings. Some problems are amenable to be described by formal rules easily translatable to computer programs. The emblematic example is playing chess. Other tasks are intuitive for humans but hard to formally describe by explicit rules (e.g., driving, recognizing faces.) The way to solve them is by allowing the rules to be learned from the data. This ability is called machine learning. *Machine Learning* (ML) is the subfield of artificial that occupies itself with this problem.

Datasets consist of examples of some phenomenon under study, that is, observations, organized as vectors. Each dimension of such vectors, called a *feature*, represents a relevant aspect of the phenomenon to the task at hand. The process of selecting those features is called feature engineering. For example, if the task is to tell apart signal from background events in a High-Energy Physics experiment, a good guess would be to choose variables used in traditional, non-ML based, analysis or even more low-level variables (i.e., the ones from which the former were constructed.)

Machine learning algorithms can be *supervised*, *unsupervised* or semi-supervised. The rules to be *machine learned* are functions (models) that map observations to predictions. Training a machine learning model means finding its parameters. In supervised learning (the only type used in this work), the examples are labeled. The model parameters are found by minimizing an error function, also called a cost function, that measures the difference between predictions and their true values expressed in the labels. When labels refer to discrete categories, it is a *classification* task. When they take continuous values, it is a *regression* task, and they are often called targets.

In unsupervised learning, the examples are not labeled. The algorithm must extract patterns from the data (for example, clustering, dimensionality reduction, etc).

If the model learns noise (statistical fluctuations present in data) during training (in which case we say to have high variance), its predictive power decreases when applied to different data. In addition to what

is similar among the various examples of a given phenomenon, it also learns the peculiarities present in that specific sample. This high predictive power of the model for data on which it was trained but which decreases for different data is called *overfitting*. We say that the model does not generalize well. To avoid this, we use only part of the data to train the model. We use another part to validate the model, that is, to verify that the model is not overfitting. Finally, we use a third part to test the model (in this work we use that part to perform the analysis). In addition to this, other methods, called regularization, are used. *Regularization* consists in making the model less complex. Various techniques are used for the effect, some will be described later. If the model is less complex, it has more difficulty to fit so well the training data (we say that it has more bias). In that way, it is less able to fit to noise.

4.1.1 Example of regression: Linear Regression

We are given some set of N observations (examples) x_i and corresponding targets y_i, where $i = 1, \ldots, N$, and want to predict y for some new observation x. Each observation has D features. We want to build a model

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + w_1 x_1 + \cdots + w_D x_D, \tag{4.1}$$

where $\boldsymbol{x}$ is the D-dimensional feature vector and $\boldsymbol{w}$ are the model parameters to be learned during training. To train the model (4.1) we need to define a cost function and minimize it to determine the parameters $\boldsymbol{w}$. In linear regression, the Mean Squared Error

$$C(\boldsymbol{w}) = \frac{1}{N} \sum_{i=1}^{N} (y(\boldsymbol{x_i}, \boldsymbol{w}) - y_i)^2 \tag{4.2}$$

is used.

For linear regression, it is possible to minimize mean squared error in closed-form, but, in general, we need to use numerical methods. The *Gradient Descent* algorithm is commonly used. When minimizing a function, we need to compute its gradient, $\boldsymbol{\nabla} C(\boldsymbol{w})$, and find where it vanishes. We begin at some initial point in the parameter space. We define a positive *learning rate* χ to control the update of the parameters. We compute the gradient of the cost function C (as a function of the model parameters) at that location and move the parameters $\hat{\boldsymbol{w}}$ in the direction opposite to the gradient (that is, in the direction of the steepest descent), in steps proportional to χ, that is,

$$\hat{\boldsymbol{w}} \leftarrow \hat{\boldsymbol{w}} - \chi \boldsymbol{\nabla}_{\boldsymbol{w}} C|_{\hat{\boldsymbol{w}}}, \tag{4.3}$$

until the gradient is sufficiently small. χ is not learned but specified in advance by the analyst. Parameters of this kind, e.g., the architecture of artificial neural networks, the optimization algorithm (gradient descent being one of them), the cost function, are called hyperparameters. These are parameters that can not be learned from data. The learning rate should not be too small (learning will take longer) nor too large (the

procedure can jump over the local minimum and possibly diverge from it). Typically, a preprocessing step, called *normalization*, could be necessary. If features have values in very different ranges, in the process of updating weights, larger features can dominate the process. Normalization consists in rescaling the features in the following way:

$$x_j \leftarrow \frac{x_j - min(x_j)}{max(x_j) - min(x_j)}, \tag{4.4}$$

where $max(x_j)$ and $min(x_j)$ are, respectively, the maximum and the minimum values that feature x_j can take. It also serves to maintain values always within a certain range to avoid numerical overflow. Another similar procedure is *standardization* where features are rescaled to have mean value 0 and standard deviation 1,

$$x_j \leftarrow \frac{x_j - \mu(x_j)}{\sigma(x_j)}, \tag{4.5}$$

where $\mu(x_j)$ and $\sigma(x_j)$ are, respectively, the mean and the standard deviation of x_j.

Linear regression is an example of a linear basis function model with the basis consisting of the feature variables. The general form for a linear basis function model with M parameters is

$$y(\boldsymbol{x}, \boldsymbol{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\boldsymbol{x}), \tag{4.6}$$

where $\phi_j(\boldsymbol{x})$ are the basis functions. After defining $\phi_0(\boldsymbol{x}) = 1$, we can write

$$y(\boldsymbol{x}, \boldsymbol{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\boldsymbol{x}). \tag{4.7}$$

Several choices are available for basis functions, that do not need to be linear in the input vector x. For example,in polynomial regression we have a single input variable x and the basis functions are the powers of x, $\phi_j(x) = x^j$. We can also have localized basis functions,

$$\phi_j(x) = \sigma(\frac{x - \mu_j}{s}), \tag{4.8}$$

where the μs serve to localize the functions, s gives the spatial scale and σ is the *logistic sigmoid* function (see Figure 4.1):

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{4.9}$$

4.1.2 Example of classification: Logistic Regression

Logistic regression is a classification algorithm, despite its name. Given some observation, we want to predict to which class it belongs. As before, we have N observations and the correspondent labels in the

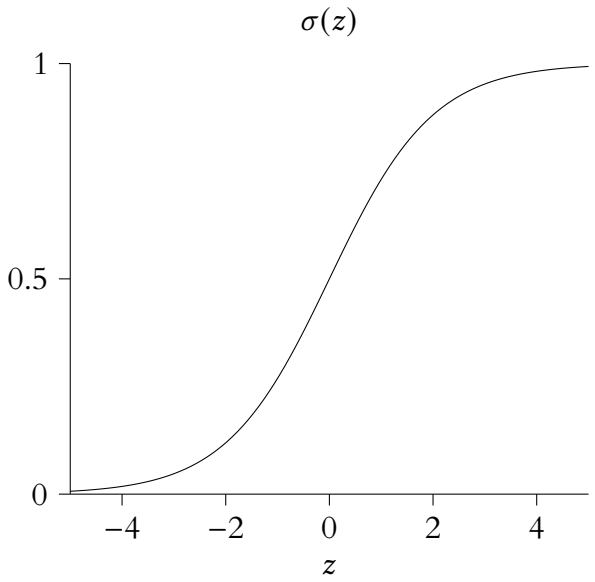


Figure 4.1: Logistic sigmoid function.

train set. Labels could be $y_i = 0, 1$, the negative and positive class, respectively, where $i = 1, \ldots, N$. The model is given by

$$\begin{aligned} y(\boldsymbol{x}, \boldsymbol{w}) &= \sigma(\boldsymbol{w} \cdot \boldsymbol{x}) \\ &= \frac{1}{1 + e^{-\boldsymbol{w} \cdot \boldsymbol{x}}}, \end{aligned} \tag{4.10}$$

where we maintain the conventions adopted early regarding x_0, and $\boldsymbol{w}$ is a D dimensional vector of features. It is apt to use this model for classification purposes, interpreting y as the probability of the label being positive. We say that an example belongs to the positive class if y is greater than some threshold, which is chosen according to the problem at hand. The training consists in adapting the weights $\boldsymbol{w}$ to minimize an appropriate cost function. Mean squared error was utilized for linear regression. It can be shown that that choice results from the *Maximum Likelihood* principle [44]. The same principle gives the *binary cross entropy* as the cost function for logistic regression. The goal is to maximize the likelihood of the data given the model. As stated before, the likelihood that some observation $\boldsymbol{x_i}$ belongs to the positive class is $y(\boldsymbol{x_i}, \boldsymbol{w})$ given by Equation 4.10. A compact way of writing the likelihood of some observation given the model is $y(\boldsymbol{x_i}, \boldsymbol{w})^{y_i}(1 - y(\boldsymbol{x_i}, \boldsymbol{w}))^{(1-y_i)}$, being that for $y_i = 1$ we get $y(\boldsymbol{x_i}, \boldsymbol{w})$, and for $y_i = 0$ we get $1 - y(\boldsymbol{x_i}, \boldsymbol{w})$, that is the likelihood of the observation to belong to the negative class. The likelihood for the entire train set is given by

$$L(\boldsymbol{w}) = \prod_{i=1}^{N} y(\boldsymbol{x_i}, \boldsymbol{w})^{y_i}(1 - y(\boldsymbol{x_i}, \boldsymbol{w}))^{(1-y_i)}. \tag{4.11}$$

To avoid dealing with huge numbers due to the exponential (numerical overflow) and the simplicity of dealing with addition rather than multiplication, *log-likelihood* is maximized instead of likelihood. We are

allowed to do it because the logarithm is a strictly growing function. The result is

$$LogL(\boldsymbol{w}) = \sum_{i=1}^{N} [y_i \log y(\boldsymbol{x_i}, \boldsymbol{w}) + (1 - y_i) \log (1 - y(\boldsymbol{x_i}, \boldsymbol{w}))]. \tag{4.12}$$

The binary cross-entropy cost function, that needs to be minimized, is simply given by the average log-likelihood multiplied by -1,

$$-\frac{1}{N} \sum_{i=1}^{N} [y_i \log y(\boldsymbol{x_i}, \boldsymbol{w}) + (1 - y_i) \log (1 - y(\boldsymbol{x_i}, \boldsymbol{w}))]. \tag{4.13}$$

4.1.3 Model Performance Metrics

After training is completed, the performance of a model must be evaluated to validate or compare it to a competing model. It is necessary to see if it generalizes well to data to which it has never been exposed, not even in the validation phase, that was still part of the training. We must note that in contrast to a simple optimization problem, where we just are interested in minimizing an error function, in training a model we have in mind to get an optimal performance metric, that only indirectly is linked to minimization of the error function. [44]

A typical performance metric is *accuracy*. Accuracy is the fraction of correctly labeled examples among the total number of them. If we divide examples used for performance evaluation into groups based on whether they were correctly classified or misclassified, as well as their true label, we get true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). Then, accuracy is given by

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{4.14}$$

This metric is employed when the consequences of misclassifying any of the classes are the same. When people are screened for cancer, for instance, it is more important to avoid a false negative than a false positive. Two other widely used metrics are *precision* and *recall*. They are defined as:

$$precision = \frac{TP}{TP + FP}, \tag{4.15}$$

$$recall = \frac{TP}{TP + FN}. \tag{4.16}$$

Precision is the fraction of true positives among the ones flagged as positives. Recall is the fraction of true positives among the ones that are actually positives, also called sensitivity or true positive rate (TPR), whereas false positive rate (FPR) is defined as

$$FPR = \frac{FP}{FP + TN}. \tag{4.17}$$

Specificity, the fraction of true negatives among all actual negatives, is given by $1 - FPR$.

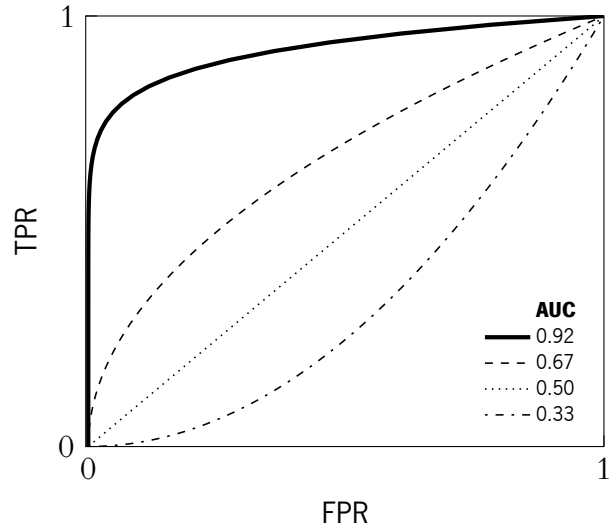


Figure 4.2: Example of ROC curves.
As the curve approaches the upper left corner, the area under the curve increases.

One metric that combines information from TPR and FPR, and that is utilized in this work, is Area Under the ROC Curve (AUC). The ROC curve [1] plots TPR against FPR, both calculated at specific thresholds ranging from 0 to 1. That is, only when a continuous probability can be assigned to each observation and the concept of a decision threshold can be specified, can this type of assessment be applied. That is what was done when likelihood was introduced in the context of logistic regression. There was a probability that some observation would fall into the positive class, but it would only be categorized as positive if it exceeded a certain threshold. It is easy to see that if TPR is always equal to FPR (see dotted line in Figure 4.2), AUC is half of its maximum possible value. That corresponds to a random classifier. The ideal curve would be one that passes through the upper left corner, as this corresponds to maximum sensitivity and maximum specificity. AUC is larger when the curve approaches that ideal.

4.2 Artificial Neural Networks

Artificial neural networks (from now on, just neural networks) are a group of machine learning algorithms that can be used for clustering, classification or regression tasks. Therefore, they are applied in computer vision, speech recognition, drug design, machine translation, etc. In this work, a neural network was used for classification. The term derives from studies that attempted to mathematically describe biological systems [45]. Only the *multilayer perceptron* architecture will be described, being the most simple type and the one used in this work.

The models for linear regression and logistic regression can both be written in the general form:

[1] ROC means receiver operating characteristic. It was introduced in the Second World War for analysis of radar signals.

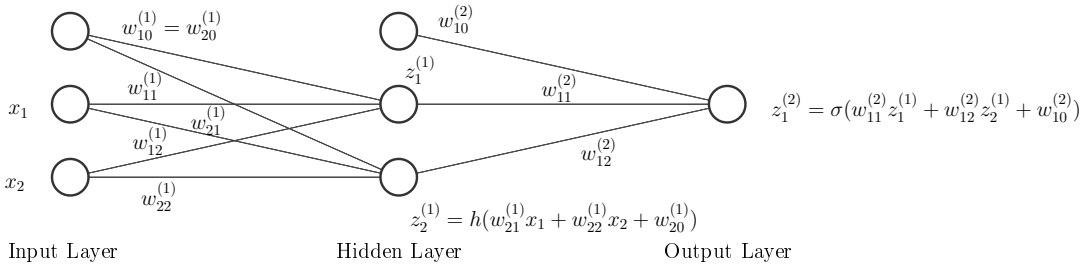


Figure 4.3: Neural network with one hidden layer.
We can say that this is a two-layer network, following [45], because its properties result from the two layers of adaptive weights.

$$y(\boldsymbol{x}, \boldsymbol{w}) = f(\sum_{j=0}^{M} w_j \phi_j(\boldsymbol{x})), \tag{4.18}$$

where M is the dimension of the function space [2], f is an *activation function* and ϕ_j are basis functions. In both cases, the basis functions are the features. The activation function for linear regression is the identity, whereas for logistical regression the sigmoid function is used. In a neural network this idea is used with basis functions that are themselves parameterized, which parameters need to be adjusted during training. More concretely, a neural network is comprised of several layers, namely, an input layer, one or more hidden layers and an output layer. The neural network model $f_{nn}(\boldsymbol{x})$ is a function of the feature vector in the form of a nested composition of functions. That is, for L layers,

$$f_{nn}(\boldsymbol{x}) = f_L(f_{L-1}(...(f_1(\boldsymbol{x})))). \tag{4.19}$$

Each layer has one or more units. Say, the unit k of layer l with N_l units has the output

$$z_k^{(l)} = h\left(\sum_{j=1}^{N_{l-1}} \left(w_{kj}^{(l-1)} z_j^{(l-1)} + w_{k0}^{(l-1)}\right)\right) \tag{4.20}$$

where h is an activation function that must be differentiable with respect to the model parameters. It is common to use the same non-linear[3] function for every unit except in the output layer. There is just one unit in the output layer for regression tasks, and identity is employed as activation function, whereas the logistic function is used for binary classification tasks, as is the case in this work, also for its sole unit. The coefficients $w_{kj}^{(l)}$ are called weights. Terms like $w_{l0}^{(l)}$ are called bias, being $z_0^{(l)} = 1$ by definition. In the

[2] ϕ_0 is 1 in order to utilize the convenient notation of the dot product between weights and inputs, and w_0 as the bias.

[3] If the activation functions were linear, the composition of linear functions would result in a linear function and the neural network would implement just a linear regression.

sum, in (4.20), $z_j^{(0)}$ are the input features discussed early. This is illustrated in Figure 4.3 for a simple case, where Equation 4.19 takes the form

$$y(\boldsymbol{x}, \boldsymbol{w}) = \sigma\left(\sum_{i=0}^{2} w_{ik}^{(2)} h\left(\sum_{j=0}^{2} w_{kj}^{(1)} x_j\right)\right), \tag{4.21}$$

where its structure of being composed by basis functions with adjustable parameters is apparent. A sufficiently complex neural network can approximate any function. [45]

The training consists, as before, in minimizing a cost function using gradient descent. But with deep neural networks (DNN), that is, networks with more than two hidden layers [4], we have nested functions, and the computation of the gradient is more complicated. In fact, it took the invention of an effective way for doing these computations, *backpropagation*, as well the development of more powerful computers, for the use of gradient descent methods be feasible. Backpropagation reduces greatly the number of computations to calculate the gradient. Without getting into details of the method, we must mention that for calculating the derivative of the cost function with respect to the weights, the closer the layer they belong is to input layer, more factors need to be multiplied. Several of the factors are the derivatives of activation functions with respect to their argument. If we look to Figure 4.1, we see that the sigmoid function have flat sections for which their derivatives are very small numbers. Then, a multiplication of several small quantities result in a vanishing gradient. As the weights are updated by terms proportional to this very small numbers, the learning becomes impossibly slow. In part, this is solved by using other activation functions. One typical activation function used in hidden layers is the Rectified Linear Unit (ReLU),

$$h(z) = \begin{cases} 0 \text{ if } z < 0 \\ z \text{ otherwise} \end{cases}, \tag{4.22}$$

leading to the derivative of the active function being one or zero. The last case, resulting in sparse networks, is actually considered advantageous [47].

The method of gradient descent described so far use all examples of the training dataset in every step. It is possible to use a suitable approximation that utilize just a different minibatch of examples in each step or even just one example per step. However, in the succession of steps, all examples of the training step will be used completing what is called an epoch. Several epochs are necessary to complete the training. This is called *stochastic gradient descent*. The approximation is acceptable and diminish drastically the number of calculations, a very important factor when using large datasets. Variants of stochastic gradient descent that are often utilized, use the ideas of momentum and adaptive learning rate. Momentum refers to an added term in the update step (see Equation 4.3) that depends of the gradients computed in previous

[4] Adding more layers was necessary because in the very beginning of neural network history, it was proved [46] that the first model studied, a perceptron (with no hidden layers) had problems, namely the impossibility to be a universal function approximator. Incorrectly, that problems were thought to be present also in multilayer perceptrons, which brought the development of the study of neural networks to a near halt. [44]

steps that help avoid oscillations, making the convergence quicker. This work was based in the use of ADAM [48], a stochastic gradient descent algorithm that uses both improvements.

In neural networks, several methods are used for regularization. Some of them were used in this work and will be briefly described. *Batch normalization* consists in standardize the output of a unit before using it as input to other units. This has a regularization effect. *Dropout* consist in randomly remove some units in each step of the optimization. This effect is controlled by a dropout rate parameter that specifies the fraction of units to drop. This is a hyperparameter that must be chosen. In this work,for tuning these and other hyperparameters the tool Optuna [49] was used. *Early stopping* consists in monitoring the performance of the model after each epoch of training in validation data and stop the training if it deteriorates.

Chapter 5

Strategy and Results

In Chapter 3, it was described a strategy for selecting a signal-enriched region of the phase space, for an hypothetical Z' signal coupling exclusively to $t\bar{t}$. It was shown in a graphical way that the observed data was better described solely by the Standard Model, excluding the BSM hypothesis, building for the effect a discriminant variable, the approximate mass of the $t\bar{t}$ system.

In Section 5.2.1, a statistical procedure will be used to interpret the obtained results. More concretely, it will be assumed that the hypothesis of discovery of new physics was rejected and an exclusion limit for the mass of Z' will be given, meaning that all hypothesized signals corresponding to masses lesser than that limit are excluded, within a certain confidence level, given the observed data. The main purpose of this work is to test if a better result, that is, a larger lower limit on the mass of Z', can be found constructing a different discriminant variable using an artificial neural network. In the first approach, rectangular regions of phase space are selected for an enhanced ratio of signal to background ($s/\sqrt{b}$) for a distribution of a variable chosen for its physical meaning. The machine learning approach will build a new variable that is a non-linear function of the feature space, which distribution better discriminate signal-rich from signal-poor regions. Two variables of this kind will be built, corresponding to two different neural networks. The difference between the two is that for one of them, in addition to low level features, one of more high level is used, namely, the approximate mass of the $t\bar{t}$ system for the same events selected in the traditional approach and that takes the value zero otherwise, that will be called tt_m. Moreover, the event selection for the machine learning approach is less strict, having in common with the cut-based approach all the cuts already described, mostly to guarantee the correct topological final state, excepting one. That is, the one that demands that there is a single large radius jet that is top-tagged and well separated from the small radius jet from the leptonic decay.

A different part of the work consists in studying how well a neural network trained to discriminate a particular signal will perform when used to separate a different signal from the same background.

5.1 Neural Network Training

The new discriminant variables used as discriminant are single-valued non-linear functions of input features of neural networks that determine their output. The output layer is therefore constituted by a single unit, which activation is the logistic function. The network is trained to recognize a given signal through the use of an appropriate cost function, the binary cross-entropy.

The event selection was made using Root but building a dataset in tabular form. For that end it was recorded for each event characteristics of a fixed number of jets (10), and of large-R jets (6). The identified jets and large-R jets for each event are sorted by descending order of their transverse momenta. When the recorded number is larger than the actual number, the proprieties of non existent jets are set to zero.

Used as features are the following:

- The multiplicity of small-R jets close to the lepton that are classified as good [1].
- The multiplicity of good b-tagged jets.
- The energy, transverse momentum, charge, type, pseudorapidity and azimuthal angle of the lepton.
- The energy, transverse momentum, pseudorapidity and azimuthal angle of the jets, and the value of MV2c10, the discriminant used to perform b-tagging.
- The energy, invariant mass, transverse momentum, pseudorapidity and azimuthal angle and two additional features of large-R jets. These two last features are τ_{32}, the weight from the algorithm for top-quark tagging, and a boolean variable that identifies that particular jet as the single large-R jet that is top-tagged and well separated from the small radius jet, for the events for which this condition is satisfied, and is set always to false otherwise.
- In one of neural networks, tt_m is also used as a feature.

To each event is associated a physical weight (here named so to distinguish them for other kind of weights, like the parameters of the DNNs or the class weights and MC weights described below). By this, it is taken in account that MC simulated events are generated with different luminosities than data and even with distorted distributions. Also, for different physical processes the efficiency of detection is different. Moreover, there are weights assigned to events by the MC generator w_g (related to specific ways of implementing integration). In the end, a final weight is calculated as follow:

$$w = F\sigma L w_g / \sum w_g, \tag{5.1}$$

where L is the integrated luminosity (a third of 10 fb^{-1}), σ is the cross-section for the process, and the fraction $w_g/\sum w_g$ gives us the inverse of the number of generated events. The sum is over the events

[1] They have to pass a pile-up exclusion filter and to have an appropriate location in the calorimeter.

corresponding to the same process [2]. F represents a combination of scale factors, including for b-tagging, pile-up, etc.

The dataset is equally divided in three parts. One part is used for training the model, other for performing validation assessment, and the last to apply the model and obtain results. During training, when computing loss, each event is given a weight that correspond to his physical weights but normalized in the sense that the sum of weights for signal events is equal to the sum of weights for background events, guaranteeing equal importance is given to both classes.

The neural network has 100 or 101 units in the input layer, corresponding to the number of features and one in the output layer. The number of hidden layers and the number of units per layer are considered hyperparameters, determined using Optuna. They can assume values between 1 and 9 and between 40 and 60, respectively. The activation function used in units belonging to hidden layers is the ReLU. The outputs of all units except the single unit of the output layer are subject to batch normalization. Instead, the input for this last unit is subject to dropout. The dropout rate, allowed to take values from 0 to 0.5 and the learning rate, allowed to vary between 10^{-5} and 10^{-2}, are, jointly with the number of hidden layers and the number of units per layer, hyperparameters.

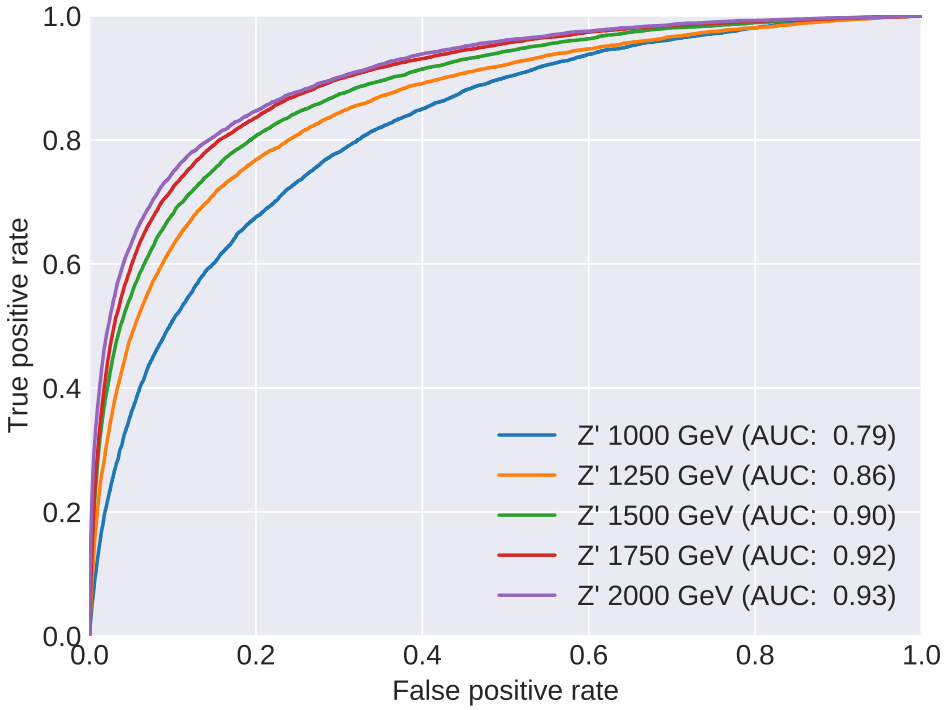
Optuna will search in the hyperparameter space, in the predefined ranges stated above, the combination that will maximize the area under the ROC curve (AUC) of the neural network (the objective function), conveniently pruning unpromising trials to speedup the process. For efficiently sampling the hyperparameter space the Tree-structured Parzen Estimator algorithm [50] is used. It begins with a random search but the history search is taken in account to suggest new values to search based in a probability model of the objective function. This extra effort put in determining the direction of the search in each step is more than compensated by the reduced number of calls to the objective function (each requiring a training session to evaluate the AUC.) The results are presented in Tables 5.1 and 5.2.

Z' mass (GeV)	number of hidden layers	number of units in each layer	learning rate	dropout rate
1000	10	56, 48, 58, 41, 54, 54, 44, 59, 60	0.008169	0.5
1250	3	51, 49	0.007106	0.4
1500	9	42, 46, 60, 52, 47, 60, 49, 56	0.009649	0.1
1750	4	50, 50, 48	0.009540	0.3
2000	7	53, 43, 50, 44, 56, 59	0.009865	0.0

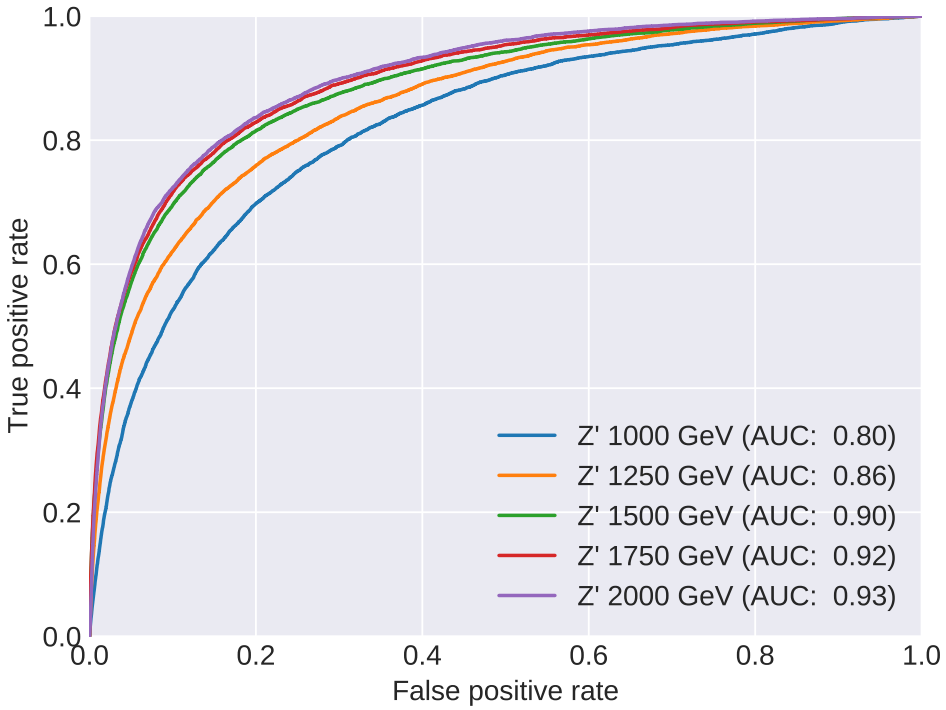
Table 5.1: Hyperparameters for NNs for which tt_m was not used as a feature.

The training was done using batches of size 1024 for 100 epochs using the early stopping method monitoring the auc with patience of 10 epochs. The performance assessment of the trained DNN is shown in Figure 5.1.

[2] This procedure is needed because there is a different weight for each event.



(a) ROC curves evaluating the performance of DNNs that don't use tt_m as feature trained with different signals.



(b) ROC curves evaluating the performance of DNNs that use tt_m as feature trained with different signals.

Figure 5.1: Performance assessment of DNNs trained to discriminate signals corresponding to Z' of different masses from background.

Z' mass (GeV)	number of hidden layers	number of units in each layer	learning rate	dropout rate
1000	5	50, 44, 57, 54, 45	0.009198	0.3
1250	6	46, 58, 54, 51, 49, 59	0.009825	0.3
1500	3	53, 55, 59	0.009621	0.2
1750	7	40, 44, 55, 50, 44, 47, 48	0.008370	0.0
2000	1	52	0.006612	0.2

Table 5.2: Hyperparameters for NNs for which tt_m was used as a feature.

5.2 Exclusion Limits

5.2.1 Exclusion Limits

The goal of a search in HEP is either to claim that a signal has been discovered or to exclude the presence of the signal. However, a search that has as result an exclusion convey information, notwithstanding. This information is presented as lower or higher limits. In our case that would be to state that if a boson Z' exists at all it would have to have a mass higher than a certain limit (a lower limit, then) because all signals with a mass smaller than that limit were excluded.

The procedure used, the so-called CL_s method [51], follows closely a standard statistical analysis based on a hypothesis test. However, it departs from it to deal with the lack of sensitivity that occurs when the expected signal is extremely low. In a standard analysis, there is a hypothesis to be rejected, called the null hypothesis (H_0), and an alternative hypothesis (H_1). A quantity q is defined, the test statistic, that is a function of the data. That quantity has a known distribution $f(q|H_0)$ if H_0 is correct. When an observation is made it is then possible to say how probable it is to get an observation that is at least so extreme, assuming that the H_0 is correct. This is made placing the test statistic observed in its distribution. The use of an alternative hypothesis allow us to define what means more extreme, in the sense that the observation is increasingly more likely to be explained by H_1 instead of H_0. The probability just described is called the p-value for that observation given H_0. H_0 is rejected at confidence level of 95% if the p-value is less than 0.05.

Let us consider the case where the null hypothesis H_{s+b} correspond to assume that we have background plus signal and H_b) is the alternative hypothesis that assumes that only background is present. Using the method just described, H_{s+b} would be rejected at confidence level 95% if

$$p_{s+b} = \int_{q_{\text{obs}}}^{\infty} f(q|H_{s+b})\mathrm{d}q < 0.05\,, \tag{5.2}$$

assuming that smaller values of q are more compatible with H_{s+b}. Likewise,

$$p_b = \int_{-\infty}^{q_{\text{obs}}} f(q|H_b)\mathrm{d}q\,, \tag{5.3}$$

is the p-value of the observation given H_b. It could happen that the distributions $f(q|H_{s+b})$ and $f(q|H_b)$ almost overlap completely [3]. In this case, the standard analysis, in some cases would lead us to reject H_{s+b} with probability close to 5 % in situations to which one has little sensitivity. In order to mitigate this effect, a new "p-value"is defined, the CL_s:

$$CL_s = \frac{p_{s+b}}{1 - p_b} \,. \tag{5.4}$$

It can be noticed that when the two distributions are well separated, the denominator is almost 1, and the standard p-value is recovered. When there is no sensitivity, the denominator makes the new "p-value"larger, allowing that less observations will reject the hypothesis of signal being present. So, in the CL_s method, the rejected models are only a part of the models that would be rejected using p_{s+b} as p-value, being therefore more conservative.

In our problem we will test several hypotheses of the type s + b, meaning, $b + \mu s$, considering different signal strengths μ, taking values from 0 to 1. An upper limit for μ_{upper} will be found, meaning that all hypotheses with $\mu > \mu_{\text{upper}}$ were rejected. This is made adjusting μ in such a way that $CL_s(\mu) = \frac{p_{\mu s+b}}{1-p_b} = 0.05$ For different masses of Z', different values of μ_{upper} are found. Then, a value for the mass of Z' will be calculated that corresponds to a full rejection of the signal hypothesis, that is, for which $\mu_{\text{upper}} = 1$. All masses lesser than that limit are rejected.

The computation of the limits was made using the Python tool pyhf [52]. The test statistic used was $\tilde{q}_\mu$ appropriated for calculate upper limits for signal strength as presented in Reference [53]. The results for masses lower limits are in Table 5.3 calculated from results from signal strength upper limits presented in Figure 5.5.

5.2.2 Results

The following results will be relative to a luminosity of 3.3 fb^{-1}. The outputs of the DNNs, the predictions for all inputs reserved in the application set (also called test set), shown in Figures 5.2 and 5.3, can be interpreted as probability distributions of an event being the signal in which the DNN was trained. In these plots and the ones in Figure 5.4, the signal is amplified in a way that its integral has the same value has the integral of the background. The signal events are extremely rare. In Figure 5.4 it is shown the distribution of the approximate mass of the $t\bar{t}$ system. There, events known to correspond to signal or to one of different backgrounds are represented separately, the backgrounds being stacked.

It can be seen in Figures 5.1 that DNNs trained with signals corresponding to higher masses have better performances. This is due to the fact that those signals are more distinct from the background as it is shown in Figures [5.2-5.4]. The use of a higher-level feature has a positive impact on the performance as can be seen comparing Figures 5.1a and 5.1b. This feature by itself already encode information that

[3] That is, not only in the tail region.

allows the process of discrimination. This discussion will be retaken in Section 5.3. The lower limits for the mass of Z' computed according to the method described in Subsection 5.2.1 are presented in Table 5.3. The same hierarchy among the methods used is obtained. The lower limit calculated based in a DNN that has a high-level feature is higher than the limit obtained using a DNN without the higher-level feature. Using a more traditional analysis, making rectangular cuts in the phase-space and using a physical observable know to have discriminant power results in a smaller lower limit. This is the result of a loss of discriminant power as we pass from one to other of these methods enumerated above. The conclusion is that a non-linear function of the features learned in an automated way can surpass the observable traditionally used based in the knowledge of basic physics, but that that knowledge can help if added to the machine learning workflow.

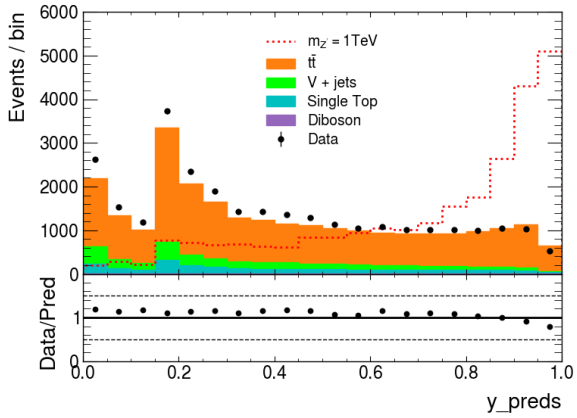
DNN based without tt_m as feature	DNN based with tt_m as feature	Based on tt_m distribution
1350.21	1364.68	1322.41

Table 5.3: 95% CL lower limits on the Z' mass (in GeV) at a luminosity of 3.3 fb^{-1}.

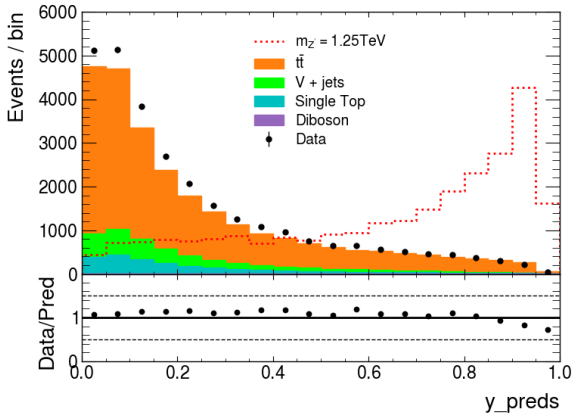
5.3 Study of transferability.

In this section is presented a study of the transferability of DNNs trained to discriminate a specific Z' signal from background to be used as a good discriminant if different signals are present amidst the same background. The study was made for DNNs that use and for those that do not use the high-level feature, each one trained for a different Z' signal. They were used Z' signals with masses: 1000, 1250, 1500 and 1750 GeV. Then, all DNNs are employed to separate all signals available from the same background. The results are presented in Figures [5.6 - 5.9].

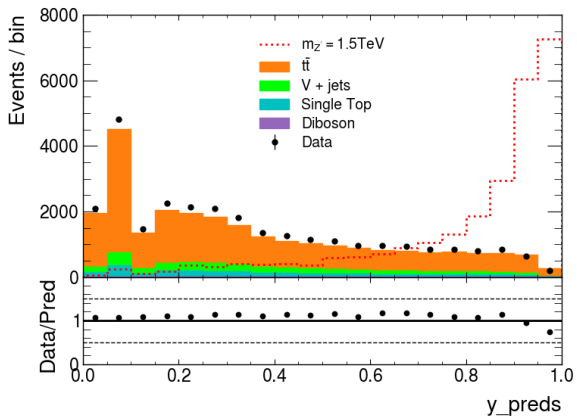
It stands out that neural networks trained with signals of larger mass retain more ability to discriminate different signals from background. This follows from the fact evidenced in Figures [5.2-5.4] that the signal is more different than background when the mass of Z' is larger. This not only explains the better AUC scores for DNNs trained on larger signal masses as shown in Figures 5.1 but can explain also the resilience of these DNNs. The neural networks learn a boundary in the features space between background and signal. As the signal is more different from background, more sharply outlined is the boundary to isolate the background. The ability of all DNNs to isolate the background explains, albeit in different degrees, their ability to discriminate what is not background. Furthermore, the larger the similarity between the signal used for training and the signal being discriminated, the better is the performance of the network. This follows from the fact that the DNNs also learn how to isolate the signal, in addition to isolate the background. When the signal is barely distinguishable from background, as it happens for the mass 1 TeV, it can occur that the DNNs trained with it loose all ability to discriminate when the signals are very different (c.f. the black squares in heatmaps 5.7 and 5.9).



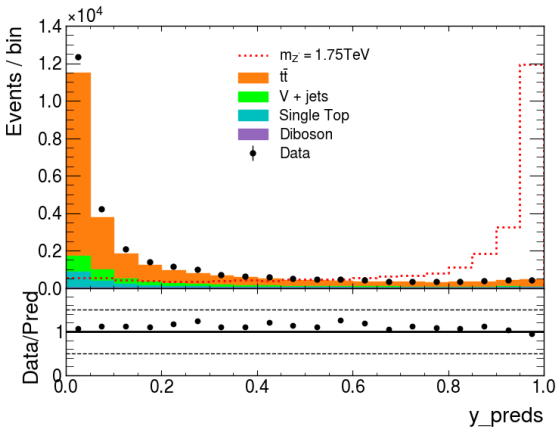
(a) DNN trained in signal Z' 1000 GeV



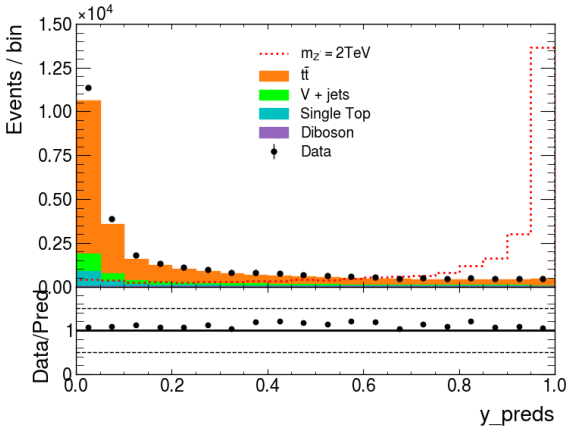
(b) DNN trained in signal Z' 1250 GeV



(c) DNN trained in signal Z' 1500 GeV

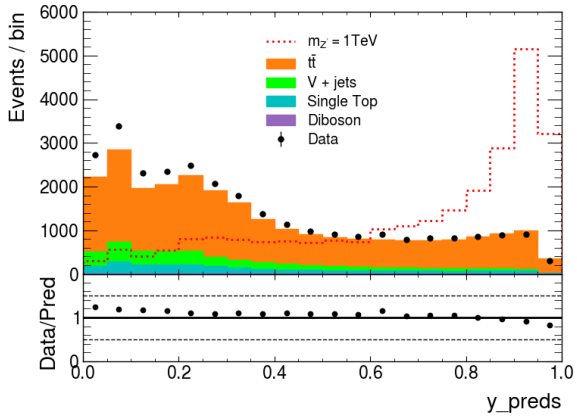


(d) DNN trained in signal Z' 1750 GeV

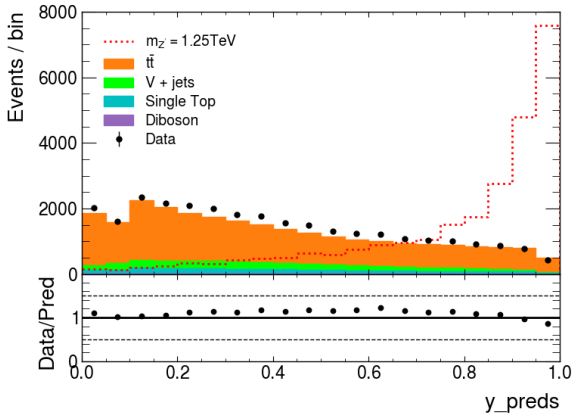


(e) DNN trained in signal Z' 2000 GeV

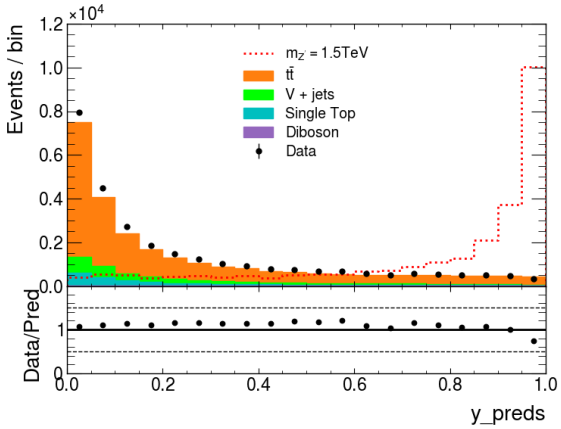
Figure 5.2: Outputs of DNNs that don't use tt_m as feature.



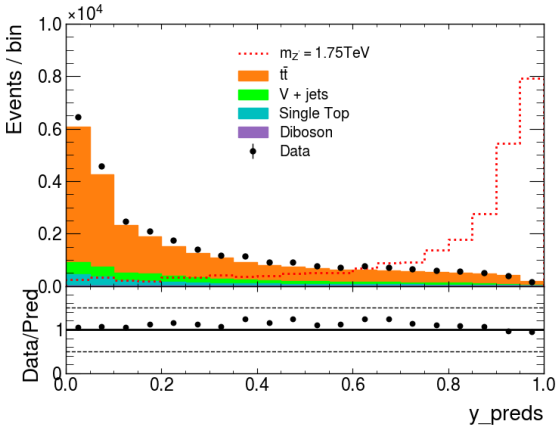
(a) DNN trained in signal Z' 1000 GeV



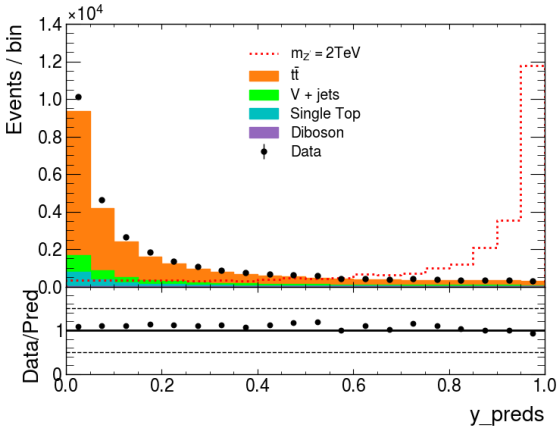
(b) DNN trained in signal Z' 1250 GeV



(c) DNN trained in signal Z' 1500 GeV

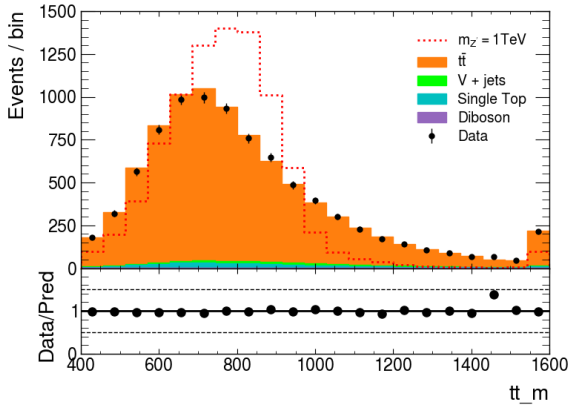


(d) DNN trained in signal Z' 1750 GeV

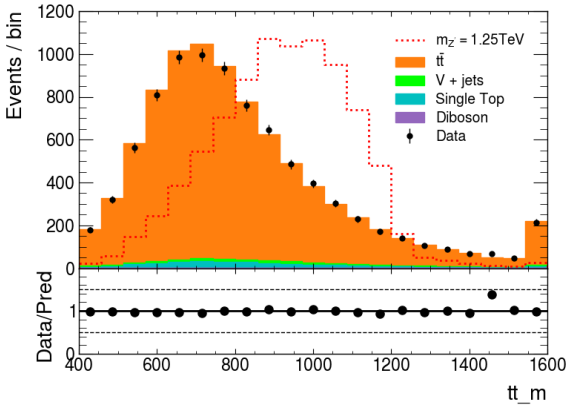


(e) DNN trained in signal Z' 2000 GeV

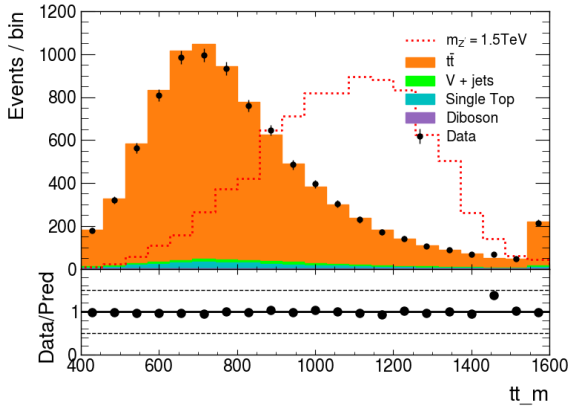
Figure 5.3: Outputs of DNNs that use tt_m as feature.



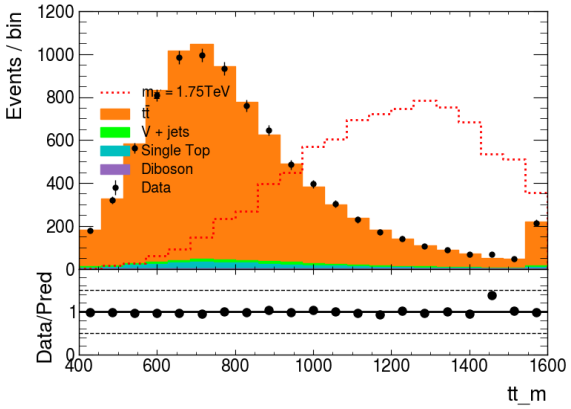
(a) Distribution of the approximate mass of $t\bar{t}$ for signal Z' 1000 GeV



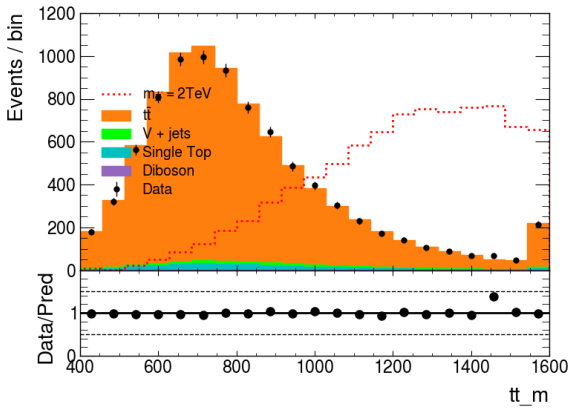
(b) Distribution of the approximate mass of $t\bar{t}$ for signal Z' 1250 GeV



(c) Distribution of the approximate mass of $t\bar{t}$ for signal Z' 1500 GeV

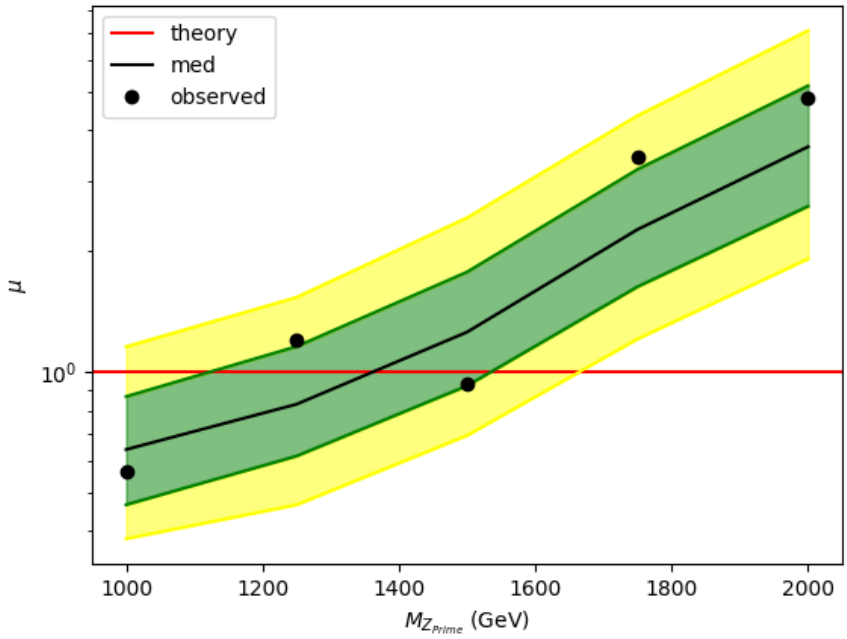


(d) Distribution of the approximate mass of $t\bar{t}$ for signal Z' 1750 GeV

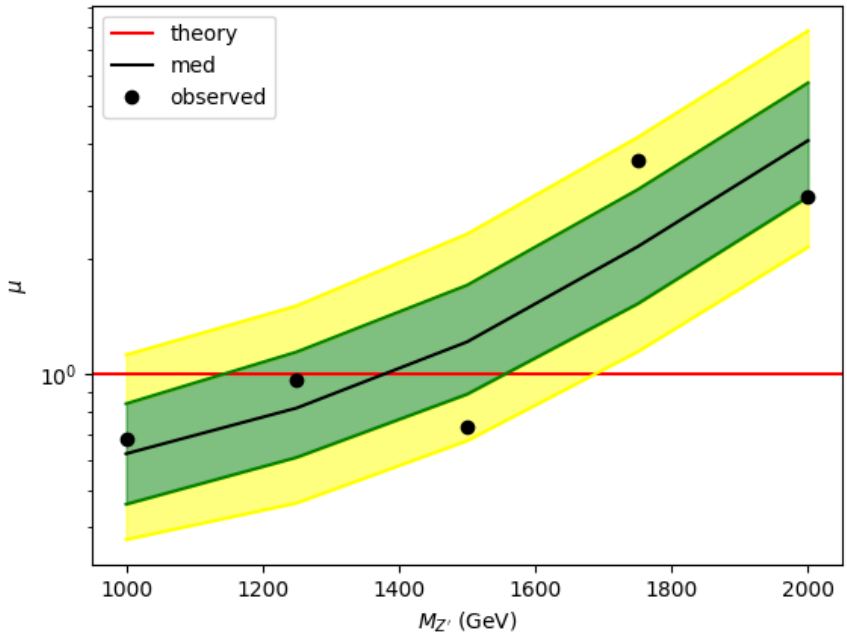


(e) Distribution of the approximate mass of $t\bar{t}$ for signal Z' 2000 GeV

Figure 5.4: Distribution of the approximate mass of $t\bar{t}$ with different signals.

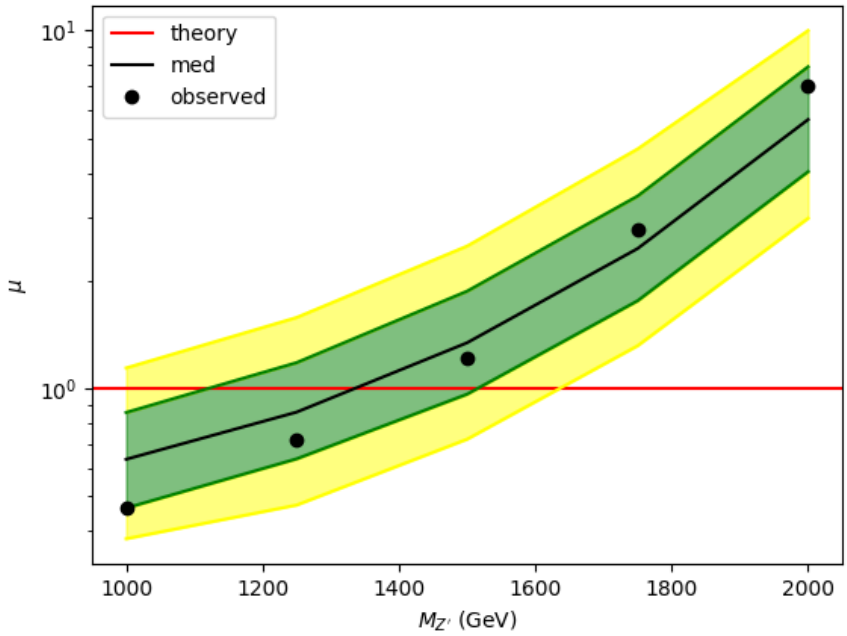


(a) Based in the output of DNNs that don't use tt_m as feature.



(b) Based in the output of DNNs that use tt_m as feature.

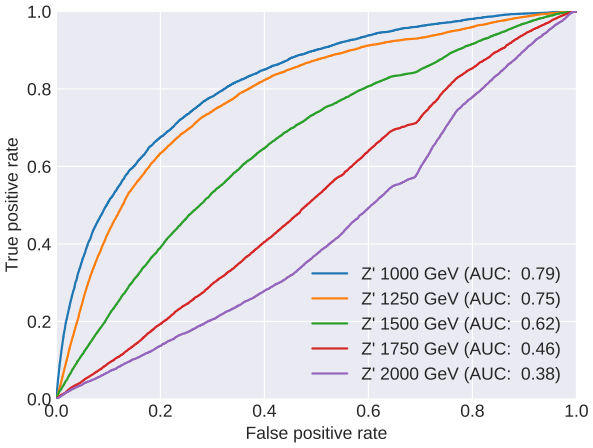
Figure 5.5: 95% CL upper limits on μ as a function of $m_{Z'}$.



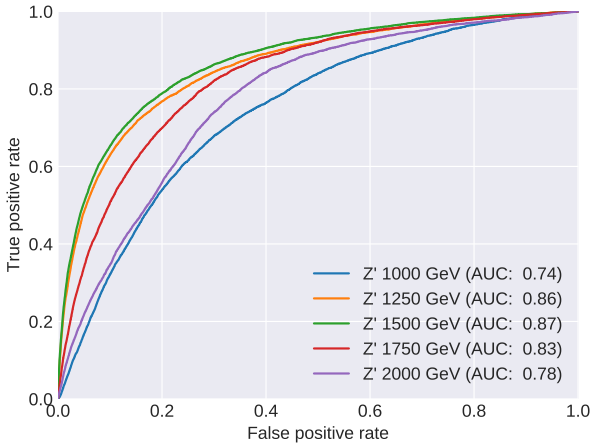
(c) Based in the distribution of the tt_m variabe.

Figure 5.5: 95% CL upper limits on μ as a function of $m_{Z'}$.

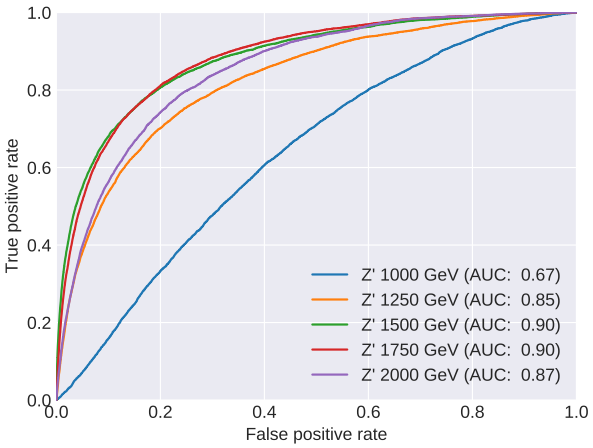
Another conclusion is that the high-level feature tt_m, by it self, carries information that helps to discriminate signal from background. This can be seen comparing Figures 5.2 to 5.3 and to 5.4. The result is that DNNs that use the high-level feature maintain more resilience in their ability to discriminate signals different from the one with which they were trained, as the ability to discriminate persists, encoded in that feature. This is shown comparing ROC curves 5.6 to 5.8.



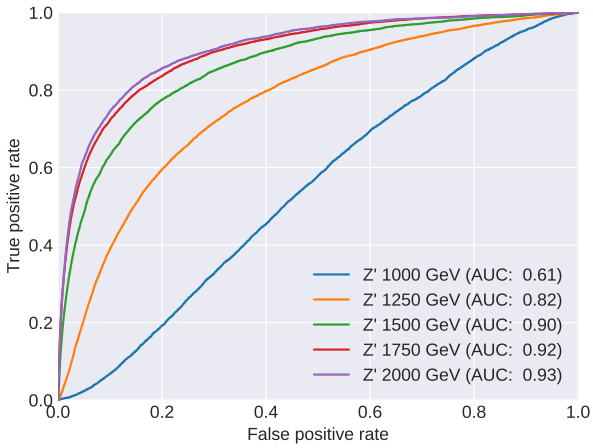
(a) DNN trained in signal Z' 1000 GeV



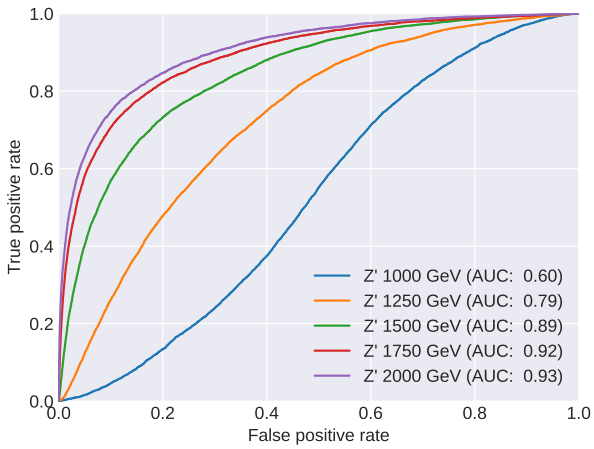
(b) DNN trained in signal Z' 1250 GeV



(c) DNN trained in signal Z' 1500 GeV



(d) DNN trained in signal Z' 1750 GeV



(e) DNN trained in signal Z' 2000 GeV

Figure 5.6: ROC curves showing the degradation of the performance of DNNs that don't use tt_m as feature as they are used to distinguish from background signals in which they were not trained.

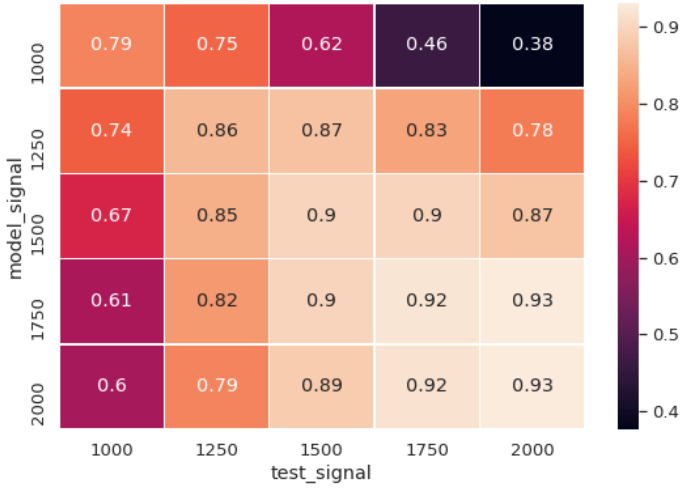
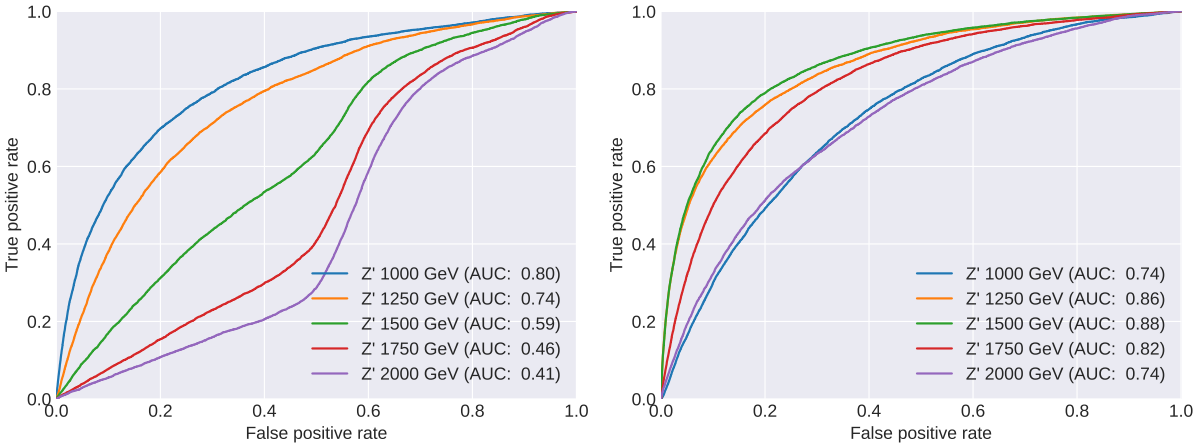
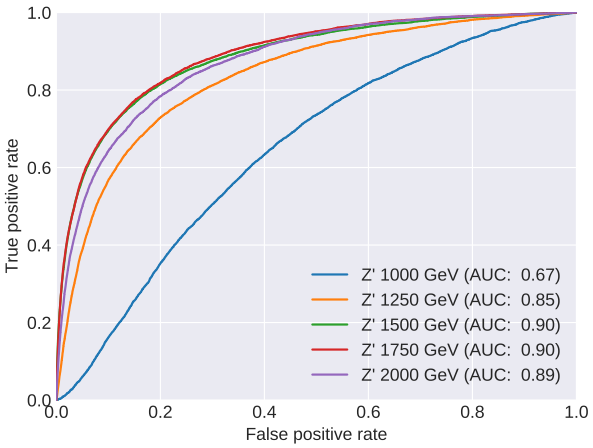


Figure 5.7: Heatmap of the AUC of the DNNs that don't use tt_m as feature. In the vertical axis is represented the signal with wich the DNN was trained. In the horizontal axis is represented the signal used for prediction.

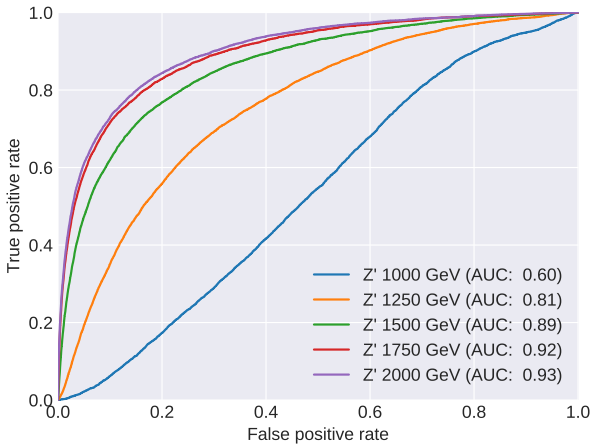


(a) DNN trained in signal Z' 1000 GeV

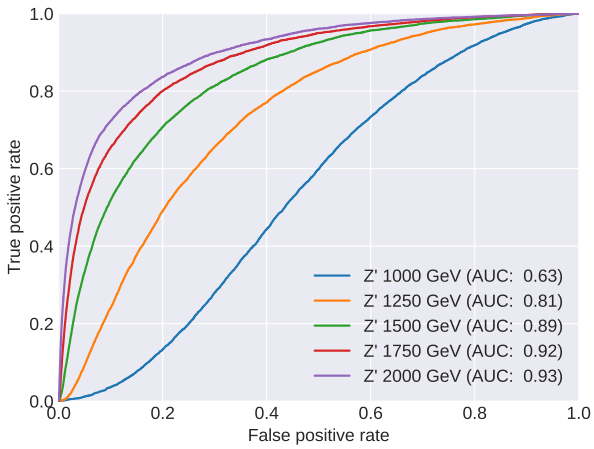
(b) DNN trained in signal Z' 1250 GeV



(c) DNN trained in signal Z' 1500 GeV



(d) DNN trained in signal Z' 1750 GeV



(e) DNN trained in signal Z' 2000 GeV

Figure 5.8: ROC curves showing the degradation of the performance of DNNs that use tt_m as feature as they are used to distinguish from background signals in which they were not trained.

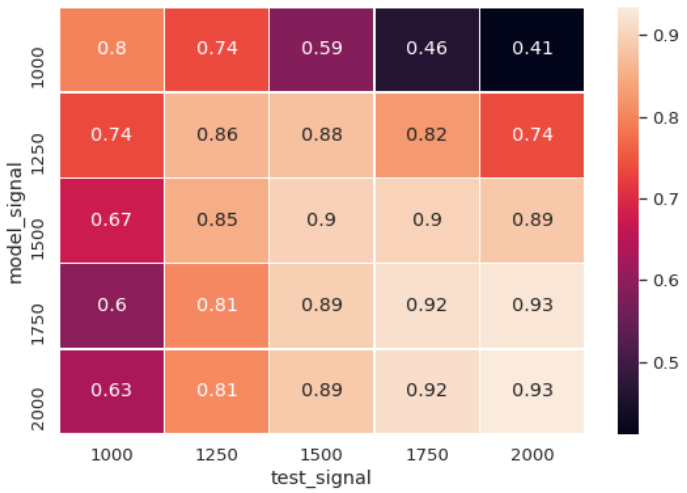


Figure 5.9: Heatmap of the AUC of the DNNs that use tt_m as feature. In the vertical axis is represented the signal with wich the DNN was trained. In the horizontal axis is represented the signal used for prediction.

Chapter 6

Conclusions and future work

It is possible to conclude from this work that the use of DNNs in establishing exclusion limits in searches for new physics is an improvement over the use of a more traditional approach. Nonetheless, the incorporation of previous knowledge in the form of physically motivated features in the building of the neural networks increases even more their performance. This, apparently, is not a novel result as this constitutes the field of feature engineering. But it also could have happened that all the information contained in the physically motivated feature had be learned by the DNN making it redundant. It is observed that transferability of DNNs occurs but degrades as signals to discriminate becomes more different from the ones used during training. Also stands out that the better situation when one is preparing a single DNN for use to discriminate different signals is to train it with the signal that is more distinct from background. In the case studied, for larger masses of Z'.

It must be noticed that the comparisons were limited to basically the same model of Z', only varying the mass. Furthermore the comparison with analyses not based in machine learning was restricted to a particular case. These limitations could be tackled in a future work. Additionally, this work could be continued in the future exploring the following points. Further comparisons could be done using different models of the boson Z'. How resilient the DNNs will be if the signal will be more different, when they depart from the background in different regions of the phase-space? Another possible study would be to compare Deep Neural Networks with other multivariate methods, for example, Boosted Decision Trees, a method commonly used in data analysis in High-Energy Physics. In addition to the comparison with other supervised methods, a comparison with unsupervised or semi-supervised methods could be performed. The absence of supervision to tell where signal is induce naturally a degradation of the performance to discriminate but these methods can be good at recognizing the background, and detecting anomalies from it. Is the degradation of the discriminative power inherent in the unsupervising nature of the method larger than the loss of ability of the DNNs studied in this work to discriminate signals different from the specific signal to which they were presented during the the supervising training? It would be also interesting to study the the systematic uncertainties related to the detectors and to the modelling of backgrounds, which

are out of the scope of this thesis.

Bibliography

[1] Wikipedia, the free encyclopedia. *Standard model of elementary particles*. [Online; accessed October 25, 2021]. 2019. url: https://commons.wikimedia.org/wiki/File:Standard_Model_of_Elementary_Particles.svg.

[2] S. Tomonaga. "On a Relativistically Invariant Formulation of the Quantum Theory of Wave Fields*." In: *Progress of Theoretical Physics* 1.2 (Aug. 1, 1946), pp. 27–42. issn: 0033-068X. doi: 10.1143/PTP.1.27.

[3] J. Schwinger. "On Quantum-Electrodynamics and the Magnetic Moment of the Electron." In: *Physical Review* 73.4 (Feb. 15, 1948), pp. 416–417. doi: 10.1103/PhysRev.73.416.

[4] J. Schwinger. "Quantum Electrodynamics. I. A Covariant Formulation." In: *Physical Review* 74.10 (Nov. 15, 1948), pp. 1439–1461. doi: 10.1103/PhysRev.74.1439.

[5] R. P. Feynman. "The Theory of Positrons." In: *Physical Review* 76.6 (Sept. 15, 1949), pp. 749–759. doi: 10.1103/PhysRev.76.749.

[6] R. P. Feynman. "Space-Time Approach to Quantum Electrodynamics." In: *Physical Review* 76.6 (Sept. 15, 1949), pp. 769–789. doi: 10.1103/PhysRev.76.769.

[7] R. P. Feynman. "Mathematical Formulation of the Quantum Theory of Electromagnetic Interaction." In: *Physical Review* 80.3 (Nov. 1, 1950), pp. 440–457. doi: 10.1103/PhysRev.80.440.

[8] G. '. Hooft. "Gauge Theories of the Forces between Elementary Particles." In: *Scientific American* 242.6 (1980), pp. 104–141. issn: 00368733, 19467087.

[9] C. N. Yang and R. L. Mills. "Conservation of Isotopic Spin and Isotopic Gauge Invariance." In: *Physical Review* 96.1 (Oct. 1, 1954), pp. 191–195. doi: 10.1103/PhysRev.96.191.

[10] A. Salam and J. C. Ward. "Electromagnetic and Weak Interactions." In: *Physics Letters* 13.2 (Nov. 15, 1964), pp. 168–171. issn: 0031-9163. doi: 10.1016/0031-9163(64)90711-5.

[11] S. Weinberg. "A Model of Leptons." In: *Physical Review Letters* 19.21 (Nov. 20, 1967), pp. 1264–1266. doi: 10.1103/PhysRevLett.19.1264.

[12] S. L. Glashow. "Partial-Symmetries of Weak Interactions." In: *Nuclear Physics* 22.4 (Feb. 1, 1961), pp. 579–588. issn: 0029-5582. doi: 10.1016/0029-5582(61)90469-2.

[13] E. Abers and B. Lee. "Gauge Theories." In: *Physics Reports* 9.1 (Nov. 1, 1973), pp. 1–2. issn: 0370-1573. doi: 10.1016/0370-1573(73)90027-6.

[14] D. Griffiths. *Introduction to Elementary Particles*. Wiley, Dec. 1987. doi: 10.1002/9783527618460.

[15] K. Olive. "Review of Particle Physics." In: *Chinese Physics C* 38.9 (Aug. 2014), p. 090001. doi: 10.1088/1674-1137/38/9/090001.

[16] C. T. Hill and S. J. Parke. "Top Quark Production: Sensitivity to New Physics." In: *Physical Review D* 49.9 (May 1, 1994), pp. 4454–4462. doi: 10.1103/PhysRevD.49.4454.

[17] C. T. Hill. "Topcolor Assisted Technicolor." In: *Physics Letters B* 345.4 (Feb. 23, 1995), pp. 483–489. issn: 0370-2693. doi: 10.1016/0370-2693(94)01660-5.

[18] R. M. Harris, C. T. Hill, and S. J. Parke. *Cross Section for Topcolor Z' Decaying to Top-Antitop*. Nov. 9, 1999. arXiv: hep-ph/9911288.

[19] The ATLAS Collaboration. "Search for Heavy Particles Decaying into Top-Quark Pairs Using Lepton-plus-Jets Events in Proton-Proton Collisions at $\sqrt{s} = 13$TeV with the ATLAS Detector." In: *The European Physical Journal C* 78.7 (July 2018), p. 565. issn: 1434-6044, 1434-6052. doi: 10.1140/epjc/s10052-018-5995-6. arXiv: 1804.10823.

[20] Particle Data Group. "Review of Particle Physics." In: *Progress of Theoretical and Experimental Physics* 2020.8 (Aug. 2020). issn: 2050-3911. doi: 10.1093/ptep/ptaa104.

[21] E. Mobs. "The CERN accelerator complex. Complexe des accélérateurs du CERN." In: (July 2016). General Photo. url: https://cds.cern.ch/record/2197559.

[22] L. Evans and P. Bryant. "LHC Machine." In: 3.08 (Aug. 2008), S08001–S08001. issn: 1748-0221. doi: 10.1088/1748-0221/3/08/S08001. url: https://doi.org/10.1088/1748-0221/3/08/s08001.

[23] The ATLAS Collaboration. "The ATLAS Experiment at the CERN Large Hadron Collider." In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08003–S08003. issn: 1748-0221. doi: 10.1088/1748-0221/3/08/S08003.

[24] The ALICE Collaboration. "The ALICE Experiment at the CERN LHC." In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08002–S08002. issn: 1748-0221. doi: 10.1088/1748-0221/3/08/S08002.

[25] The CMS Collaboration. "The CMS Experiment at the CERN LHC." In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08004–S08004. issn: 1748-0221. doi: 10.1088/1748-0221/3/08/S08004.

[26] The LHCb Collaboration. "The LHCb Detector at the LHC." In: *Journal of Instrumentation* 3.08 (Aug. 2008), S08005–S08005. issn: 1748-0221. doi: 10.1088/1748-0221/3/08/S08005.

[27] The ATLAS Collaboration. "Observation of a New Particle in the Search for the Standard Model Higgs Boson with the ATLAS Detector at the LHC." In: *Physics Letters B* 716.1 (Sept. 17, 2012), pp. 1–29. issn: 0370-2693. doi: 10.1016/j.physletb.2012.08.020.

[28] The CMS Collaboration. "Observation of a New Boson at a Mass of 125 GeV with the CMS Experiment at the LHC." In: *Physics Letters B* 716.1 (Sept. 17, 2012), pp. 30–61. issn: 0370-2693. doi: 10.1016/j.physletb.2012.08.021.

[29] The ATLAS Collaboration. "Alignment of the ATLAS Inner Detector in Run 2." In: *The European Physical Journal C* 80.12 (Dec. 24, 2020), p. 1194. issn: 1434-6052. doi: 10.1140/epjc/s10052-020-08700-6.

[30] J. Pequenão. "Computer Generated image of the ATLAS calorimeter." Mar. 2008. url: https://cds.cern.ch/record/1095927.

[31] J. Pequenão. "Computer generated image of the ATLAS Muons subsystem." Mar. 2008. url: https://cds.cern.ch/record/1095929.

[32] The ATLAS Collaboration. "Performance of the ATLAS Trigger System in 2015." In: *The European Physical Journal C* 77.5 (May 18, 2017), p. 317. issn: 1434-6052. doi: 10.1140/epjc/s10052-017-4852-3.

[33] The ATLAS Collaboration. *Review of the 13 TeV ATLAS Open Data Release*. ATL-OREACH-PUB-2020-001. Jan. 24, 2020. url: https://cds.cern.ch/record/2707171.

[34] R. Brun and F. Rademakers. "ROOT — An Object Oriented Data Analysis Framework." In: *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*. New Computing Techniques in Physics Research V 389.1 (Apr. 11, 1997), pp. 81–86. issn: 0168-9002. doi: 10.1016/S0168-9002(97)00048-X.

[35] The ATLAS Collaboration. "Electron Reconstruction and Identification in the ATLAS Experiment Using the 2015 and 2016 LHC Proton-Proton Collision Data at $\sqrt{s} = 13$TeV." In: *The European Physical Journal C* 79.8 (Aug. 3, 2019), p. 639. issn: 1434-6052. doi: 10.1140/epjc/s10052-019-7140-6.

[36] The ATLAS Collaboration. "Muon reconstruction performance of the ATLAS detector in proton-proton collision data at $\sqrt{s} = 13$ TeV." In: *The European Physical Journal C* 76.5 (May 2016). doi: 10.1140/epjc/s10052-016-4120-y. url: https://doi.org/10.1140/epjc/s10052-016-4120-y.

[37] M. Cacciari, G. P. Salam, and G. Soyez. "The anti-ktjet clustering algorithm." In: *Journal of High Energy Physics* 2008.04 (Apr. 2008), pp. 063–063. doi: 10.1088/1126-6708/2008/04/063. url: https://doi.org/10.1088/1126-6708/2008/04/063.

[38] The ATLAS Collaboration. *Tagging and Suppression of Pileup Jets with the ATLAS Detector*. ATLAS-CONF-2014-018. ATLAS-COM-CONF-2014-025, May 12, 2014. url: https://cds.cern.ch/record/1700870.

[39] D. Krohn, J. Thaler, and L.-T. Wang. "Jet Trimming." In: *Journal of High Energy Physics* 2010.2 (Feb. 24, 2010), p. 84. issn: 1029-8479. doi: 10.1007/JHEP02(2010)084.

[40] The ATLAS Collaboration. *Performance of jet substructure techniques in early $\sqrt{s} = 13$ TeV pp collisions with the ATLAS detector*. Tech. rep. Geneva: CERN, Aug. 2015. url: https://cds.cern.ch/record/2041462.

[41] The ATLAS Collaboration. "Measurements of B-Jet Tagging Efficiency with the ATLAS Detector Using $$ T\overline{t} $$events at $$ \sqrt{s}=13 $$TeV." In: *Journal of High Energy Physics* 2018.8 (Aug. 16, 2018), p. 89. issn: 1029-8479. doi: 10.1007/JHEP08(2018)089.

[42] The ATLAS Collaboration. "Measurements of Top-Quark Pair Differential Cross-Sections in the Lepton+jets Channel in Pp Collisions at $\sqrt{s} = 13$ TeV Using the ATLAS Detector." In: *Journal of High Energy Physics* 2017.11 (Nov. 28, 2017), p. 191. issn: 1029-8479. doi: 10.1007/JHEP11(2017)191.

[43] The ATLAS Collaboration. *Boosted hadronic top identification at ATLAS for early 13 TeV data*. Tech. rep. Geneva: CERN, Dec. 2015. url: https://cds.cern.ch/record/2116351.

[44] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[45] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[46] M. Minsky and S. A. Papert. *Perceptrons; an introduction to computational geometry*. MIT Press, 1969.

[47] X. Glorot, A. Bordes, and Y. Bengio. "Deep Sparse Rectifier Neural Networks." In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. Ed. by G. Gordon, D. Dunson, and M. Dudík. Vol. 15. Proceedings of Machine Learning Research. Fort Lauderdale, FL, USA: PMLR, 2011, pp. 315–323. url: https://proceedings.mlr.press/v15/glorot11a.html.

[48] D. P. Kingma and J. Ba. *Adam: A Method for Stochastic Optimization*. 2017. arXiv: 1412.6980 [cs.LG].

[49] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. "Optuna: A Next-generation Hyperparameter Optimization Framework." In: *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2019.

[50] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl. “Algorithms for Hyper-Parameter Optimization.” In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Q. Weinberger. Vol. 24. Curran Associates, Inc., 2011.

[51] A. L. Read. “Presentation of Search Results: The CL_S Technique.” In: *Journal of Physics G: Nuclear and Particle Physics* 28.10 (Oct. 1, 2002), pp. 2693–2704. issn: 0954-3899. doi: 10.1088/0954-3899/28/10/313.

[52] L. Heinrich, M. Feickert, G. Stark, and K. Cranmer. “pyhf: pure-Python implementation of HistFactory statistical models.” In: *Journal of Open Source Software* 6.58 (2021), p. 2823. doi: 10.21105/joss.02823.

[53] G. Cowan, K. Cranmer, E. Gross, and O. Vitells. “Asymptotic Formulae for Likelihood-Based Tests of New Physics.” In: *The European Physical Journal C* 71.2 (Feb. 9, 2011), p. 1554. issn: 1434-6052. doi: 10.1140/epjc/s10052-011-1554-0.