UNIVERSITÀ DEGLI STUDI DI NAPOLI
"FEDERICO II"

Scuola Politecnica e delle Scienze di Base

Area Didattica di Scienze Matematiche Fisiche e Naturali

Dipartimento di Fisica "Ettore Pancini"

Laurea Magistrale in Fisica

Search for new heavy resonances decaying into top and bottom quarks at the CMS experiment with Machine Learning techniques.

Relatori:
Prof. Luca Lista
Dott. Alberto Orso Maria Iorio

Candidata:
Cristina Giordano
Matr. *N*94000630

Anno Accademico 2020/2021

Contents

The lesson you need to learn right now
can't be taught with words,
only with action.

LEVI ACKERMAN

Introduction

The Standard Model of Particle Physics (SM) is the theory that accurately describes three out of the four known fundamental interactions occurring between particles. The predictions of this model have been experimentally verified extensively and with high precision, and all the particles included in its framework have been observed, the last one being the Higgs boson at the Large Hadron Collider (LHC) at CERN, in 2012. Despite its great level of predictivity, the SM is inherently incomplete, since it does not account for a plethora of phenomena, among others the existence of Dark Matter and Dark Energy, gravitational interactions, the presence of neutrino masses, the suppression of flavour changing neutral currents, and the matter-antimatter asymmetry. Many theories have been formulated in order to overcome these shortcomings; some of these theories either propose the existence of extra dimensions for the inclusion of General Relativity (hence gravity) or the presence of extra interactions and fields, providing an extension of the already existing SM. Almost all these theories, classified as theories Beyond Standard Model (BSM), predict the existence of heavy resonances such as the W' boson, a hypothetical heavy copy of the SM W boson of electroweak interactions, with the same quantum numbers of its SM counterpart but with much higher mass. Such particles could be produced in proton-proton collisions in the energy reach of the LHC. Experiments such as the Compact Muon Solenoid (CMS) actively search for these and many other particles that could be hint of new physics BSM. The W' boson could decay in a top and a bottom quark; this decay mode is of particular interest as it could be the main signature for many models foreseeing the W' coupling preferentially with the third generation of quarks and leptons. Besides, it has a peculiar decay chain, where the top quark further decays to a b-jet, lepton, and neutrino triplet. The main purpose of this work is the identification and reconstruction (*tagging*) of the top quark with the CMS detector via the study of its decay in the previously mentioned final leptonic state. In order to improve on the existing searches and enhance the possibility of a discovery of the W' boson at the LHC, Machine Learning (ML) techniques were used in the top-tagging process. The performance of said ML algorithm can be tested by performing an analysis on simulated samples and comparing the expected results to the existing analyses. This work is thusly structured:

- *Chapter 1:* an overview of the SM of Particle Physics;
- *Chapter 2:* a description of the LHC and the CMS experiment;

- *Chapter 3:* a presentation of BSM Physics theories, with main focus on those predicting the existence of the W' boson;
- *Chapter 4:* an account of the object selection and reconstruction, with a brief description of the functioning of the used Machine Learning algorithm;
- *Chapter 5:* the analysis developed for the reconstruction of the W' boson and the description of the fit procedure and its results.

Chapter 1

Standard Model of Particle Physics

The Standard Model of Particle Physics (SM) is a quantum field theory that describes three out of the four known fundamental interactions: the electromagnetic, strong and weak forces. In 1961 S. Glashow unified the electromagnetic and weak interactions, and in 1979 S. Weinberg and A. Salam incorporated the Higgs mechanism into Glashow's framework. The theory of strong interaction, namely Quantum Chromodynamics (QCD), was developed during the 60s and 70s, as a result of the efforts of many scientists, who contributed to shape it in its currently known form. Fundamental contributions worthy of note were, for instance, those of M. Gell-Mann and G. Zweig, for the development of the first three-quark model and that of O. W. Greenberg, who first introduced a new quantum number called *colour*. The picture of strong interactions was reached in 1973, when the theory of the asymptotic freedom of strong interactions was proposed. The SM has been proved to be incredibly precise and accurate in its predictions. However, it does not account for many observed phenomena; a dedicated discussion on some of those and their potential experimental implications is given in Chapter 3.

1.1 Overview of the Standard Model

The SM describes particles and force fields with the same formalism, for the interaction itself arises from the exchange of particles called *mediators*. It is a $SU(3)_C \times SU(2)_L \times U(1)_Y$ quantum field theory, meaning that the Lagrangian Density $\mathcal{L}$ must be invariant under the action of this group. The subscripts refer to the *colour* charge C, *left-hand chirality* L and *hypercharge* Y; these quantities are conserved under transformations belonging to the $SU(3)_C$, $SU(2)_L$ and $U(1)_Y$ groups respectively.
Each group has different generators:

- 8 Gell-Mann matrices λ_i for the Special Unitary group $SU(3)_C$;
- 3 Pauli matrices τ_i for the Special Unitary group $SU(2)_L$;

- 1 Identity matrix I for the Unitary group $U(1)_Y$.

Each generator is associated to a gauge boson, the mediators of the interaction between particles:

- From the $SU(3)_C$ symmetry, 8 gluons act as mediators of the strong interaction taking place between gluons;
- From the $SU(2)_L \times U(1)_Y$ symmetry, 3 W bosons of weak isospin and the B hypercharge boson arise.

Actually, $W_{1,2,3}$ and B are not physical fields: the real bosons are generated through a spontaneous symmetry breaking of $SU(2)_L \times U(1)_Y$. As a result of this break, 3 vector bosons, $W^{\pm}$ and Z, mediate the weak interaction, while the electromagnetic force is carried out through the exchange of a photon γ. A summary of the gauge bosons of SM and their properties is shown in Table 1.1.

Boson	Electric Charge	Mass[1]	Spin	Interaction
Gluon	0	0	1	Strong
Photon	0	0	1	Electromagnetic
$W^{\pm}$	± 1	80.379 ± 0.012 GeV	1	Weak(CC)
Z	0	91.187 ± 0.002 GeV	1	Weak(NC)

Table 1.1: Gauge bosons of the SM [1]; CC stands for Charged Current, while NC stands for Neutral Current.

In SM, particles are divided in bosons and fermions: the former have integer spin values and obey the Bose-Einstein statistics, the latter have half-integer spin values and abide by the rules of the Fermi-Dirac statistics. As already discussed above, interactions are associated to one absolutely conserved quantum number and to a boson multiplet, whose components are the actual mediators of the interaction. In the SM, elementary fermions are either quarks or leptons. Quarks take part in electromagnetic, weak, and strong interactions, therefore they are elecromagnetically, weakly, and strongly charged. They have different flavours: up, down, charm, strange, top(truth), and bottom(beauty). They are divided in three generations in accordance with the timeline of their discovery. Quarks and their properties are listed in Table 1.2.
Leptons can interact weakly and electromagnetically. They are distinguished in three generations, just like quarks. Each generation (also called family) is composed of a massive, charged particle and a massless, electrically neutral one. Table 1.3 lists their main characteristics.

1.2 Quantum Electrodynamics

Quantum Electrodynamics (QED) is the quantum field theory aimed at describing the electromagnetic interaction. The Lagrangian density for QED is obtained by imposition of gauge principles on the free Lagrangian density:

Generation	Particle	Charge	Mass[MeV]
I	u	$2/3\ e$	2.32 ± 0.10
	d	-1/3e	4.71 ± 0.09
II	s	$2/3\ e$	1280 ± 25
	c	-1/3e	92.9 ± 0.7
III	t	$2/3\ e$	$173.34 \pm 0.27 \pm 0.71 \times 10^3$
	b	-1/3 e	4180 ± 30

Table 1.2: Quarks and their properties [1]

Generation	Particle	Spin	Charge	Mass[MeV]
I	e	1/2	$-e$	0.511
	ν_e	1/2	0	0
II	μ	1/2	$-e$	105.7
	ν_μ	1/2	0	0
III	τ	1/2	$-e$	1776.86
	ν_τ	1/2	0	0

Table 1.3: Leptons and their properties [1].

$$\mathcal{L}_D = i\bar{\psi}\gamma^\mu\partial_\mu\psi - m\bar{\psi}\psi \tag{1.1}$$

where ψ is the Dirac-bispinor field, $\bar{\psi}$ is its adjoint, m is the mass of the field, and γ^μ are the Dirac matrices. In order to obtain the Lagrangian density for QED, Equation 1.1 must be locally invariant under $U(1)$ transformation:

$$\psi \longrightarrow \psi' = e^{i\theta(x)}\psi. \tag{1.2}$$

By substituting 1.2 in Equation 1.1, the Lagrangian becomes:

$$\mathcal{L'}_D = i\bar{\psi'}\gamma^\mu\partial_\mu\psi' - m\bar{\psi'}\psi'. \tag{1.3}$$

Expanding the terms one obtains:

$$\mathcal{L}' = i\bar{\psi}\gamma^\mu\partial_\mu\psi - \partial_\mu\theta(x)\bar{\psi}\gamma^\mu\psi - m\bar{\psi}\psi \tag{1.4}$$

This new Lagrangian density has an additional term that renders it not locally invariant. Gauge principles restore the invariance by adding an interaction field to the lagrangian, namely the electromagnetic four-potential A^μ, that abides by the following transformation law:

$$A^\mu \longrightarrow A'^\mu = A^\mu - \partial^\mu\Lambda \tag{1.5}$$

in which $\Lambda(\bar{x}, t)$ is a generic twice continuously differentiable function. By replacing the derivative ∂_μ in Equation 1.1 with the covariant derivative

$$D_\mu = \partial_\mu + ieA_\mu \tag{1.6}$$

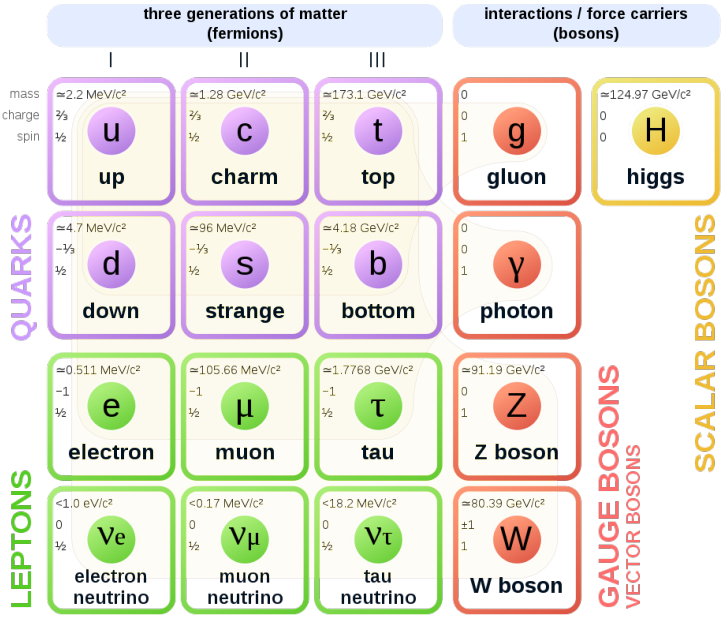


Figure 1.1: *Standard model of Particle Physics*: 12 fundamental fermions and 5 fundamental bosons. Brown loops indicate which bosons (red) couple to which fermions (purple and green).

where e is the electric charge of the bispinor field, one obtains:

$$\mathcal{L} = i\bar{\psi}\gamma^\mu\partial_\mu\psi - e\bar{\psi}\gamma^\mu A_\mu\psi - m\bar{\psi}\psi. \tag{1.7}$$

The new interaction term cancels the extra term in the previous expression for $\mathcal{L}'$ (Equation 1.4) and grants local gauge invariance; the Lagrangian density for QED is obtained by finally adding a kinetic term for the electromagnetic field. Defining the Electromagnetic Field Tensor:

$$F^{\mu\nu} = \partial^\mu A^\nu - \partial^\nu A^\mu \tag{1.8}$$

The Lagrangian density for QED is:

$$\mathcal{L}_{QED} = i\bar{\psi}\gamma^\mu\partial_\mu\psi - e\bar{\psi}\gamma^\mu A_\mu\psi - m\bar{\psi}\psi - \frac{1}{4}F^{\mu\nu}F_{\mu\nu} \tag{1.9}$$

1.3 The Electroweak Unification

At the beginning of the 20th century, the electromagnetic interaction had been the subject of extensive and thorough studies that brought to the formulation of QED, the theory that describes electromagnetic phenomena in great details. However, not all the known occurrences could be explained via the application of QED principles, for instance the β-decays. Two types of β decays were known, β^- and β^+; the former consists in a proton that converts into a neutron emitting an electron (called, in fact, β particles in Nuclear Physics), while the

latter sees a neutron becoming a proton and releasing a positron.

$$X(Z,A) \longrightarrow Y(Z+1,A) + e^- \tag{1.10}$$

$$X(Z,A) \longrightarrow Y(Z-1,A) + e^+ \tag{1.11}$$

$$X(Z,A) + e^- \longrightarrow Y(Z-1,A) \tag{1.12}$$

Equation 1.12 is a process called Electron Capture, another phenomenon discovered at the beginning of the last century that was deemed impossible to explain with the prevailing knowledge of QED. E. Fermi and W. Pauli first noticed that the energy spectrum associated with the electron was continuum; according to the kinematic laws of two-body decays, in order for the momentum an energy conservation laws to still be valid, the energy of the electron must be discrete. Furthermore, the non conservation of angular momentum posed another problem to the idea that these were closed kinematics processes. This discrepancy resulted in the discovery of *neutrinos*; the existence of this massless particles explained the continuum energy spectrum of *beta* decays:

$$X(Z,A) \longrightarrow Y(Z+1,A) + e^- + \bar{\nu_e} \tag{1.13}$$

$$X(Z,A) \longrightarrow Y(Z-1,A) + e^+ + \nu_e \tag{1.14}$$

$$X(Z,A) + e^- \longrightarrow Y(Z-1,A) + \nu_e \tag{1.15}$$

In 1933, the first theory of the weak interaction was proposed by E. Fermi, according to which β decays were described as four-fermion contact interactions, without the aid of a mediator. In the 1960s, S. Glashow, A. Salam and S. Weinberg unified the electromagnetic force and the weak interaction, naming this as the Electroweak Unification.

1.3.1 The Fermi Theory of Weak Interaction

The weak interaction was considered by Fermi as a four-fermion contact interaction; the lagrangian for this process is:

$$\mathcal{L} = \frac{G_F}{\sqrt{2}} J^\dagger_\mu(x) J^\mu(x), \tag{1.16}$$

where $G_F = 1.16638 \times 10^{-5} GeV^{-2}$ is the Fermi constant and $J^\mu(x)$ is the weak current, which can be either leptonic or hadronic. In 1957, C.S. Wu proved that the purely leptonic weak current has a Vectorial-Axial structure; in the case of a transition from an electron to its corresponding neutrino:

$$J^\mu(x) = \bar{\psi}_e(x)\gamma^\mu(1-\gamma^5)\psi_\nu(x) \tag{1.17}$$

In the previous expression, each Dirac-bispinor can be further decomposed in right-handed and left-handed chiral bispinor.

$$\psi = \psi_L + \psi_R \tag{1.18}$$

It is to be noted that the current in Equation 1.17 is charged; this caused great turmoil in the scientific community at the time, for the other noteworthy example of charged current is to be found in the theory of strong interactions, Quantum Chromodynamics (QCD), which was developed significantly later than Fermi's theory. According to this notion, only left-handed particles and right-handed antiparticles take part in weak interactions, for the left chiral projector P_L is proportional to $(1-\gamma^5)$. This theory describes weak interactions remarkably well. Iowever, it was abandoned, for experimental data proved that charged leptons couple with neutral currents, even though the coupling constants are different for the different chiralities. Furthermore, the calculated cross section is proportional to the square of the energy:

$$\sigma \propto G_F E^2 \tag{1.19}$$

so that for energies higher than $100GeV$ the unitarity of the scattering matrix is violated, therefore rendering this theory only valid for low energies.

1.3.2 The Glashow-Salam-Weinberg Model

In order to overcome the problems posed by Fermi's theory of weak interactions, S. Glashow, S. Weinberg, and A. Salam proposed a non-Abelian $SU(2)_L$ quantum field theory, where L is the left-handed chiral components of the fields. In this model, left-handed fermions are coupled in doublets with fixed eigenvalue of the weak isospin:

$$\begin{array}{llll} I = 1/2 & I_3 = +1/2 & \begin{pmatrix}\nu_l \\ l^-\end{pmatrix}_L & \begin{pmatrix}U \\ D'\end{pmatrix}_L \\ I = 1/2 & I_3 = -1/2 & & \end{array} \tag{1.20}$$

where I is the weak isospin, I_3 is the projection of I along a chosen axis, l^- represents charged leptons and ν_l their corresponding neutrinos, U represents an up-type quark ($u,\ c,\ t$ quarks), while D' stands for the corresponding down-type quarks. The weak interaction eigenstates d', s' and b' are a linear combination of the strong interaction eigenstates (mass eigenstates). The mixing matrix between mass and weak D' eigenstates is called CKM (Cabibbo-Kobayashi-Maskawa) matrix and acts as follows:

$$\begin{pmatrix}d' \\ s' \\ b'\end{pmatrix} = V_{CKM} \begin{pmatrix}d \\ s \\ b\end{pmatrix} \tag{1.21}$$

where V_{CKM} is the CKM matrix:

$$V_{CKM} = \begin{pmatrix}V_{ud} & V_{us} & V_{ub} \\ V_{cd} & V_{cs} & V_{cb} \\ V_{td} & V_{ts} & V_{tb}\end{pmatrix} \tag{1.22}$$

Taking the modulus squared of each element of this complex unitary matrix one obtains the probability of transition to one quark to another. Oftentimes, the Wolfenstein parametrization is used:

$$V_{CKM} = \begin{pmatrix} 1-\frac{\lambda^2}{2} & \lambda & A\lambda^3(\rho - i\eta) \\ -\lambda & 1-\frac{\lambda^2}{2} & A\lambda^2 \\ A\lambda^3(1-\rho-i\eta) & -A\lambda^2 & 1 \end{pmatrix} + \mathcal{O}(\lambda^4) \tag{1.23}$$

Measurements for A and λ show that the CKM matrix is quasi-diagonal:

$$\lambda = 0.22650 \pm 0.00048 \;\; ; \;\; A = 0.790^{+0.017}_{-0.012} \tag{1.24}$$

meaning that u, c and t quarks have higher probability to interact with their corresponding down-type quarks. The procedure to obtain the Lagrangian density for the electroweak interaction is the same as for QED: gauge principles are applied to the free Lagrangian, using the appropriate covariant derivative:

$$D_\mu = \partial_\mu + ig\frac{\vec{\tau}}{2} \cdot \vec{W}_\mu + ig'\frac{Y}{2}B_\mu \tag{1.25}$$

in which g and g' are the equivalent of the electrical charge in Equation 1.6, and the fields $W_\mu^{1,2,3}$ and B_μ ensure the local invariance of the lagrangian under the action of the group $SU(2)_L \times U(1)_Y$. Since an $SU(2)_L$ transformation can be espressed as:

$$\psi_L \longrightarrow \psi'_L = e^{ig\frac{\vec{\tau}}{2}\cdot\vec{\alpha}}\psi_L, \tag{1.26}$$

where $\vec{\alpha}$ is the parameter of the transformation, the gauge property for the W_μ^i fields is:

$$W_\mu^i \longrightarrow W_\mu'^i = W_\mu^i - \partial_\mu\alpha^i - g\epsilon^{ijk}\alpha^j W_\mu^k. \tag{1.27}$$

The gauge properties for the group $U(1)$ have already been discussed in Section 1.2, and its transformation law is:

$$B_\mu \longrightarrow B'_\mu = B_\mu - \partial_\mu\Lambda. \tag{1.28}$$

The interaction terms between the fermions and the fields are:

$$\mathcal{L}_{EW}^I = -g\bar{\psi}_L\gamma^\mu\frac{\vec{\tau}}{2}\cdot\vec{W}_\mu\bar{\psi}_L - g'\frac{Y}{2}\psi\gamma^\mu B_\mu\psi \tag{1.29}$$

The complete form of the Lagrangian is once again obtained by adding the kinetic terms for the interaction fields. B_μ follows the same rule as in Equation 1.9, while the one for $\vec{W}_\mu$ is:

$$\mathcal{L}_K = -\frac{1}{4}\vec{W}^{\mu\nu}\cdot\vec{W}_{\mu\nu} \tag{1.30}$$

where $\vec{W}_{\mu\nu}$ is:

$$\vec{W}_{\mu\nu} = \partial_\mu\vec{W}_\nu - \partial_\nu\vec{W}_\mu - g\vec{W}_\mu \times \vec{W}_\nu. \tag{1.31}$$

The cross product in Equation 1.31 results in triple and quadruple gauge boson vertices that represent self-interaction terms of the W boson. The GSW model does not account for all the characteristics of the weak interaction. For instance, its bosons are massive but when a mass term is added to the Lagrangian, the gauge invariance is broken. Furthermore, the interaction of right-handed charged particles via neutral weak current is not explained. These problems are solved through the aid of the Higgs mechanism.

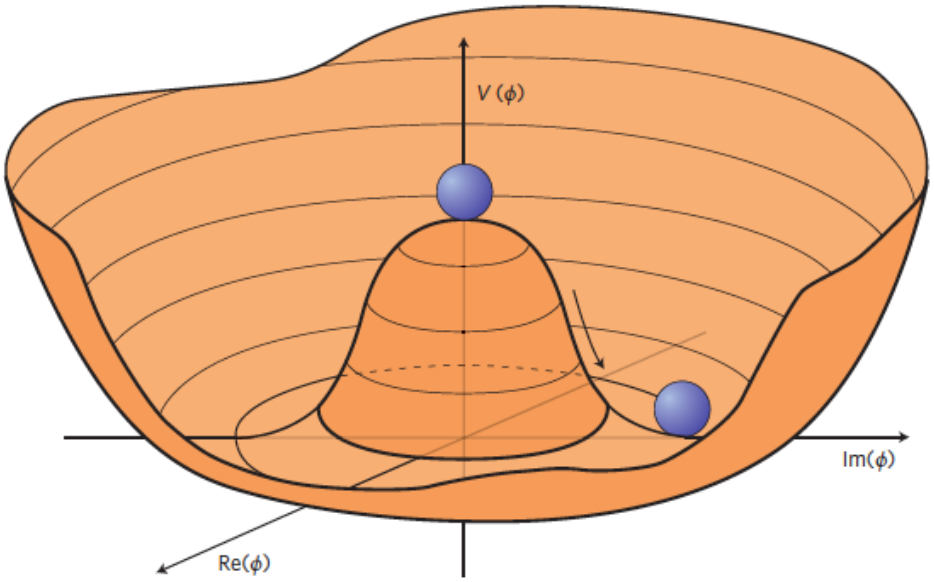


Figure 1.2: Higgs potential for $\mu^2 < 0$.

1.4 The Higgs Mechanism

The Higgs mechanism is an essential part of the Standard Model, for it represents the way in which the weak interaction gauge bosons acquire mass without breaking the local gauge symmetry of the SM. Developed by P. Higgs, R. Brout and F. Englert [2] [3] in 1964, this mechanism explains the mass of Z, $W^{\pm}$, and fermions as a result of their interaction with the Higgs boson field. Its particle counterpart was later observed in 2012 at LHC by the ATLAS and CMS experiments. [4] [5]. The minimal Higgs model consists of two complex scalar fields placed in a weak isospin doublet:

$$\phi = \begin{pmatrix} \phi^+ \\ \phi^0 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} \phi_1 + i\phi_2 \\ \phi_3 + i\phi_4 \end{pmatrix}. \tag{1.32}$$

This mechanism is required to generate the masses of the electroweak gauge bosons; therefore, one of the components of the doublet must be neutral (ϕ^0), while the other must be charged (ϕ^+). The Lagrangian for the Higgs doublet is:

$$\mathcal{L} = (\partial_\mu \phi)^\dagger (\partial^\mu \phi) - V(\phi), \tag{1.33}$$

where $V(\phi)$ is the Higgs potential:

$$[b]V(\phi) = \mu^2 \phi^\dagger \phi + \lambda(\phi^\dagger \phi)^2 \tag{1.34}$$

For $\mu^2 < 0$, this potential (shown in Figure 1.2) has an infinite set of degenerate minima that satisfy the following condition:

$$\phi^\dagger \phi = \frac{1}{2}(\phi_1^2 + \phi_2^2 + \phi_3^2 + \phi_4^2) = -\frac{\mu^2}{2\lambda} = \frac{v^2}{2} \tag{1.35}$$

After symmetry breaking, the neutral photon is required to remain massless, and therefore the minimum of the potential must correspond to a non-zero vacuum expectation value only of the neutral scalar field ϕ^0, namely:

$$\langle 0|\phi|0\rangle = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\v\end{pmatrix} \tag{1.36}$$

The fields can be expanded around this minimum by rewriting:

$$\phi(x) = \frac{1}{\sqrt{2}}\begin{pmatrix}\phi_1(x) + i\phi_2(x)\\v + \eta(x) + i\phi_4(x)\end{pmatrix} \tag{1.37}$$

After the spontaneous breaking of the symmetry, there will be a massive scalar and three massless Goldstone bosons, which will give the longitudinal degrees of freedom of the $W^{\pm}$ and Z bosons. The three massless bosons can be eliminated from the Lagrangian by making an appropriate gauge transformation, called Unitary Gauge, that consists in choosing the complex scalar field $\phi(x)$ to be entirely real:

$$\phi(x) = \frac{1}{\sqrt{2}}\begin{pmatrix}0\\v + h(x)\end{pmatrix}. \tag{1.38}$$

The Lagrangian in Equation 1.33 is required to be locally invariant under the action of $SU(2)_L \times U(1)_Y$; the resulting mass matrices are diagonalized and the physical bosons are obtained. The terms corresponding to these bosons and their mass terms depend on the interaction of the gauge bosons with the Higgs field:

$$\begin{aligned}\mathcal{L} = {} & \frac{1}{2}\partial_\mu H \partial^\mu H + \\ & - \frac{1}{4}\left(\partial_\mu W^1_\nu - \partial_\nu W^1_\mu\right)\left(\partial^\mu W^{1\nu} - \partial^\nu W^{1\mu}\right) + \frac{1}{8}v^2 g^2 W^1_\mu W^{1\mu} + \\ & - \frac{1}{4}\left(\partial_\mu W^2_\nu - \partial_\nu W^2_\mu\right)\left(\partial^\mu W^{2\nu} - \partial^\nu W^{2\mu}\right) + \frac{1}{8}v^2 g^2 W^2_\mu W^{2\mu} + \\ & - \frac{1}{4}\left(\partial_\mu Z_\nu - \partial_\nu Z_\mu\right)\left(\partial^\mu Z^\nu - \partial^\nu Z^\mu\right) + \frac{1}{8}v^2 g^2 Z_\mu Z^\mu + \\ & - \frac{1}{4}\left(F_{\mu\nu}F^{\mu\nu}\right).\end{aligned} \tag{1.39}$$

The mass of the physical bosons $W^{\pm}$ is:

$$m_W = \frac{1}{2}gv; \tag{1.40}$$

the mass of the Z boson is:

$$m_Z = \frac{1}{2}v\sqrt{g^2 + g'^2} = \frac{m_W}{\cos\theta_W} \tag{1.41}$$

in which θ_W is the Weinberg angle, that represents the rotation performed on the W^3 and B fields to obtain the real Z and γ bosons, namely:

$$\begin{pmatrix}A_\mu\\Z_\mu\end{pmatrix} = \begin{pmatrix}\cos\theta_W & \sin\theta_W\\-\sin\theta_W & \cos\theta_W\end{pmatrix}\begin{pmatrix}B_\mu\\W^3_\mu\end{pmatrix}. \tag{1.42}$$

This mechanism is able to explain why the weak neutral current couples to both left-handed and right-handed chirality: the Z boson results from a combination of two neutral bosons, one of which (B) couples to particles independently from their chirality. This minimal formulation of the Higgs mechanism is the simplest, but in no way the only one. For instante, in Supersymmetry (SUSY), a very popular extension to the SM, employs at lease two complex doublets of scalar fields, which give rise to five physical Higgs boson. Furthermore, the nature of the Higgs boson is still a conundrum: the answer to the question whether it is fundamental or composite is still unknown.

1.5 Quantum Chromodynamics

In the context of the SM, strong interactions are described by Quantum Chromodynamics (QCD), a non-Abelian quantum gauge field theory. Its symmetry group is $SU(3)_c$, whose generators are:

$$T_{1,\dots,8} = \frac{\lambda_{1,\dots,8}}{2}, \tag{1.43}$$

where the $\lambda_{1,\dots,8}$ are the Gell-Mann matrices, whose commutation rules are:

$$\left[\frac{\lambda_a}{2}, \frac{\lambda_b}{2}\right] = i f_{abc} \frac{\lambda_c}{2}, \tag{1.44}$$

where f_{abc} are the structure constants of the groups and the indices run from 1 to 8. The Gell-Mann are 3×3 matrices, therefore the ψ field includes three additional degrees of freedom: this new degree of freedom is called colour. As per usual, the gauge principle leads to the following covariant derivative:

$$D_\mu = \partial_\mu + i g_s \frac{\vec{\lambda}}{2} \cdot \vec{G}_\mu \tag{1.45}$$

where $\vec{G}_\mu$ are the eight fields representing the mediators of strong interaction, called gluons. The transformation law for these fields that ensure the local invariance of the lagrangian is:

$$G^i_\mu \longrightarrow G'^i_\mu = G^i_\mu + i g_s f_{ijk} \theta^j(x) G^k_\mu \tag{1.46}$$

Going through the same passages and adding a kinetic term for the gluon field, one obtains the full lagrangian for QCD:

$$\mathcal{L}_{\mathcal{QCD}} = \bar{\psi}\gamma^\mu \partial_\mu \psi - m\bar{\psi}\psi - g_s \bar{\psi}\gamma^\mu \frac{\vec{\lambda}}{2}_\mu \psi - \frac{1}{4}\vec{G}_{\mu\nu}\vec{G}^{\mu\nu}, \tag{1.47}$$

where:

$$G^i_{\mu\nu} = \partial_\mu G^i_\nu - \partial_\nu G^i_\mu - g_s f^{ijk} G^j_\mu \times G^k_\nu. \tag{1.48}$$

The kinetic term causes self-interaction between gluons, as it was the case for the weak interaction. The reason for the existence of charged currents is to be

found in their non-Abelian nature of the groups representing these interactions. Two important features of QCD are colour confinement and asymptotic freedom; these properties were firstly theorized and later experimentally proved. The former stems from the need to explain the absence of coloured hadrons in nature; therefore, hadrons are considered colour singlets, since they are bound states of quarks in the QCD parton model. Furthermore, only colourless, quark bound state configurations can exist. The latter was discovered in 1973 by D. Gross, F. Wilczek [6], and indipendenty by D. Politzer [7]. It consists in the fact that strong interaction becomes asymptotically weaker as the transferred momentum $|q|$ increases. This can be explained with the fact that the coupling constant of the strong interaction varies with the scale of the interaction, hence earning the name of *running* coupling constant:

$$\alpha_s(|q^2) = \frac{\alpha_s(\mu^2)}{\left[1 + \alpha_s(\mu^2)\frac{11N_C - 2N_f}{12\pi}\ln\left(\frac{q^2}{\mu^2}\right)\right]} \tag{1.49}$$

where $N_C = 3$ is the number of colours, N_f is the number of flavours available at transferred momentum q^2 and μ is a scale parameter for the strength of the coupling. For $|q|$ $200MeV$ the value of the running constant is large enough that perturbative approaches cannot be applied; this region call for another formulation of the theory, called Lattice QCD.

Chapter 2

The CMS Experiment

The Compact Muon Solenoid (CMS) is one of the four great experiments taking data produced by the Large Hadron Collider (LHC), located underground near Geneva. The LHC is the latest proton-proton accelerator realized by the European Organization for Nuclear Research (CERN), which is a supranational organization founded in 1954. Born with the goal of providing a research centre untainted by the unpleasant rise of nationalism still present in the aftermath of WWII, CERN is nowadays the site of the currently biggest particle accelerators needed for High-Energy Physics research. Several breakthrough discoveries have been made by CERN experiments, among others the discovery of neutral currents (Gargamelle in 1973), W and Z bosons (UA1 and UA2 experiments in 1983) and the Higgs boson (ATLAS and CMS at LHC in 2012). In this Chapter, an overview of the LHC and a more detailed description of the CMS experiment are given [8].

2.1 The Large Hadron Collider

The LHC at CERN near Geneva is the world's largest circular accelerator to this day. Proton beams are accelerated up to a design collision energy of 14 TeV and luminosity of $10^{34} cm^{-2} s^{-1}$, while heavy ions (Pb) have energy of 2.8 TeV per nucleon and luminosity $10^{27} cm^{-2} s^{-1}$. LHC constitutes the last stage in a series of different accelerators which have the purpose of bringing particle beams up to a certain energy threshold before injecting them into the next stage. The Pb atoms are obtained from a source of vaporised lead, the protons are obtained from a hydrogen gas tank connected to a machine called duoplasmatron; this device strips the H atoms of their electrons, producing a plasma of protons, electrons, and molecular ions. Protons are extracted, then a beam is formed and fed to the subsequent chain of machines:

- Linac 2, a linear accelerator that speeds them up to 50 Mev of energy;
- Proton Synchrotron Booster (PBS), which accelerates them to 1.4 GeV;
- Proton Synchrotron (PS), which brings them to 25 GeV;

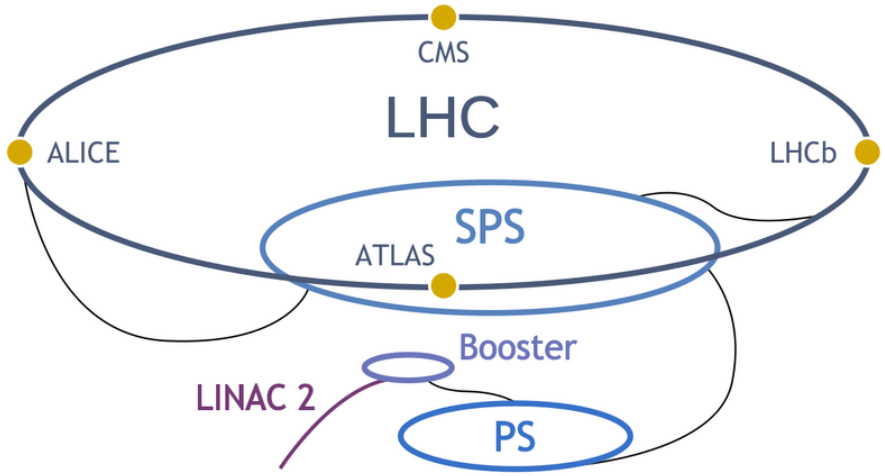


Figure 2.1: A simplified view of the LHC complex at CERN.

- Super Proton Synchrotron (SPS), which raises the energy to 450 GeV.

The beams are lastly fed to LHC, whose structure is shown in Figure 2.1; the main ring has a circumference of 27 kms on which the beams are accelerated to the maximum energy of 7 TeV via a system of superconductive radiofrequency cavities and focusing superconductive magnets that have the task of keeping the particles on a circular path. Particles are led to collision in four points of intersection, where the four main experiments are located. These experiments are: Compact Muon Solenoid (CMS), A Toroidal LHC ApparatuS (ATLAS), A Large Ion Collider Experiment (ALICE), LHC-beauty (LHCb). One of the most important parameters of accelerators is the Luminosity $\mathcal{L}$, a quantity related to the rate of events detected R and to the total interaction cross-section σ:

$$R = \mathcal{L}\sigma \tag{2.1}$$

The Luminosity is strictly dependent on the characteristics of the accelerator, and for circular accelerators it is calculated as

$$\mathcal{L} = \frac{n_b N_b^2 \gamma_r f_{rev}}{4\pi\epsilon_n \beta*} F \tag{2.2}$$

where n_b is the number of bunches per beam, N_b the number of particles per bunch, γ_r the relativistic Lorentz factor, ϵ_n the normalised transverse beam emittance, $\beta*$ the beta function at the collision point, f_{ref} the revolution frequency and F the geometric luminosity reduction factor. Another important quantity is the Integrated Luminosity L:

$$L = \int_{t_1}^{t_2} \mathcal{L}(t)dt \tag{2.3}$$

where the extremes of integration t_1 and t_2 represent the time during which the collider is functioning. This quantity L is related to the number of total

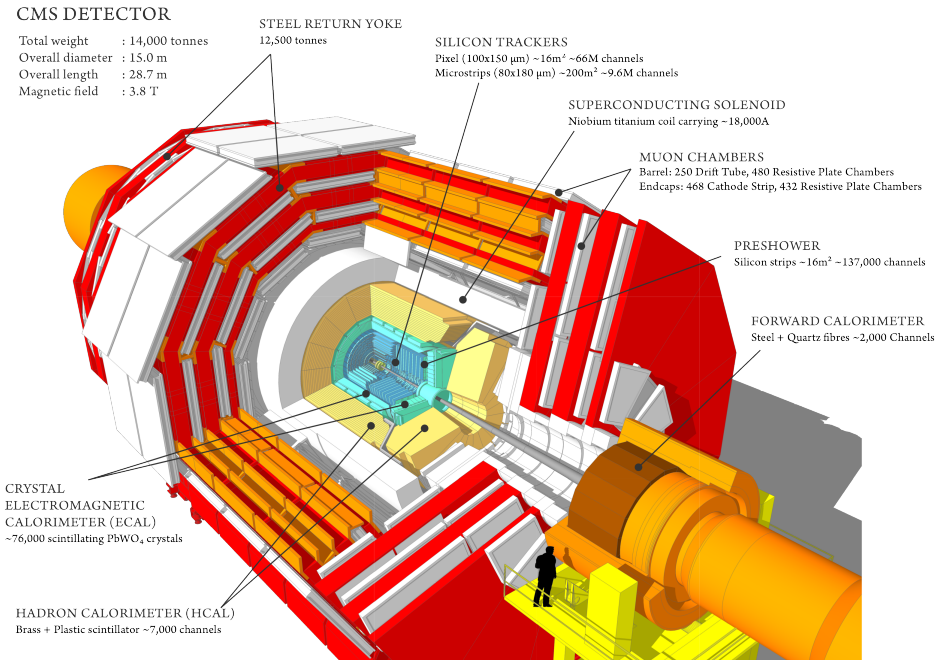


Figure 2.2: A view of the structure of CMS.

events

$$N = L\sigma. \tag{2.4}$$

2.2 The Compact Muon Solenoid

The Compact Muon Solenoid is one of the four main experiments at LHC and it is a general-purpose detector designed to observe a wide range of phenomena within the context of the Standard Model; however, it could also observe events of Physics Beyond Standard Model (BSM). Overall, the detector has the structure of a cylinder, 21 metres high and with a diameter of 15 metres, positioned horizontally so that its centre (called Barrel) surrounds the collision point; two structures (Endcaps) orthogonal to the beam axis enclose the Barrel. CMS is formed of different subdetectors adhering to precise requirements, namely high spatial and time resolution and high radiation hardness, each to different degrees depending on the specific function of the subdetector and on its positioning and distance with respect to the beam axis. High spatial resolution is needed in order to distinguish between particles that cross the detector in close positions, while high time resolution is necessary for triggering purposes and to distinguish between subsequent collisions. Furthermore, radiation hardness is fundamental, for the detector works with high levels of radiation for long periods of time. As shown in Figure 2.2, the CMS detector has different layers each with a specific purpose, namely (from inner to outer shell):

- a tracking system;

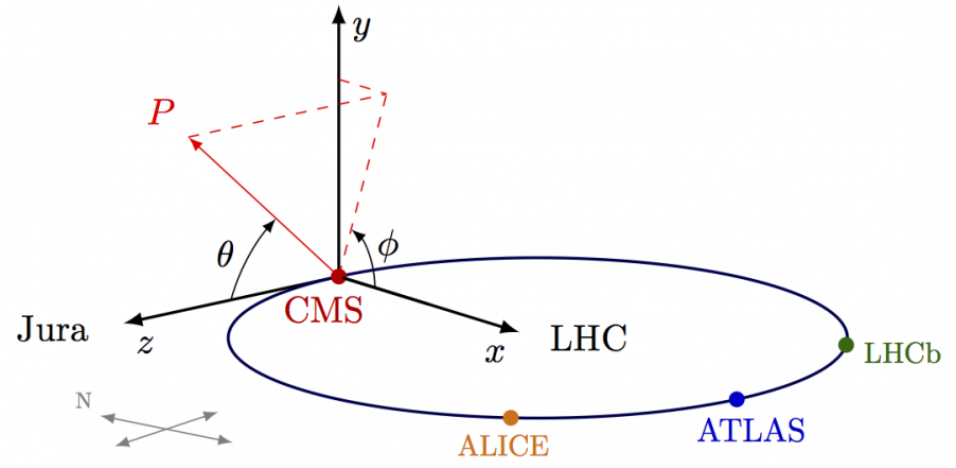


Figure 2.3: The coordinate systems used in CMS;
the cartesian system is deployed for the description of the detector, while the cylindrical coordinate system is used to describe the characteristics of the particles arising from the p-p collision.

- an electromagnetic calorimeter (ECAL);
- hadron calorimeter (HCAL);
- a superconductive solenoid, from wich the experiment takes its name;
- a muon tracking system;

A right-handed cartesian system of coordinates is used to describe the collision events, as seen in Figure 2.3: the x-axis points to the centre of LHC, the y-axis points upwards, and the z-axis is tangential to the beam in the counterclockwise direction. Furthermore, a cylindrical system is employed for describing the quantities related to the particle arising from the collision (Figure 2.3).

- the radial distance r from the z-axis;
- the azimuthal angle Φ around the z-axis;
- the polar angle θ around the x-axis;

An alternative variable is frequently used instead of θ, called pseudo-rapidity

$$\eta = -\ln\left(\tan\left(\frac{\theta}{2}\right)\right). \tag{2.5}$$

which is a proxy variable of another important quantity named rapidity:

$$y = \frac{1}{2}\ln\left(\frac{E + p_z}{E - p_z}\right) \tag{2.6}$$

Both y and η are invariant under a Lorentz boost along the z-axis. For particles in the ultrarelativistic limit, η tends to for it transforms linearly under a

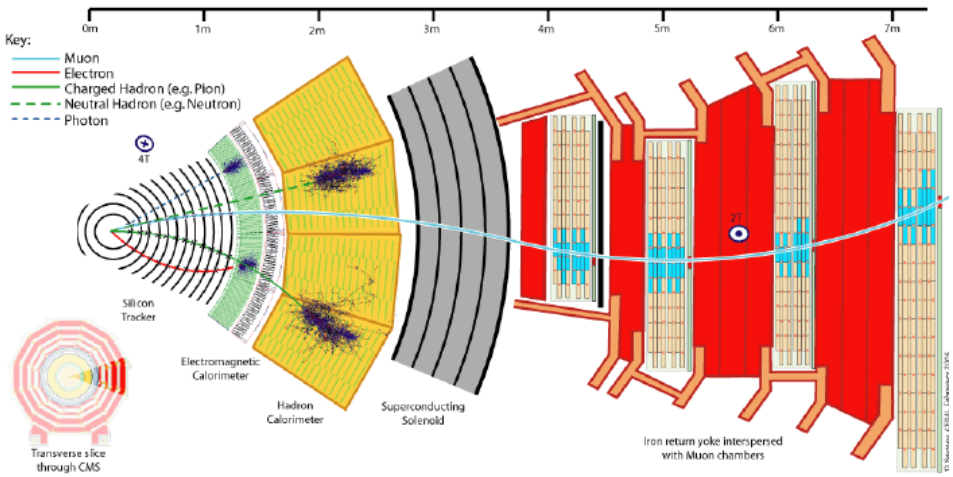


Figure 2.4: The CMS subdetector configuration.

Lorentz boost along the z-axis. This implies that the difference of the η values of two particles is also invariant for this type of boost. Furthermore, using η and Φ it is possible to define the angular distance between two particles R:

$$\Delta R = \sqrt{(\Delta\eta)^2 + (\Delta\Phi)^2}. \tag{2.7}$$

Another important quantity that is invariant under this kind of boost is called transverse momentum. In the cartesian frame of reference, the momentum is conventionally divided into two components: p_z, the longitudinal momentum, and p_T, the transverse momentum, defined as:

$$p_T = \sqrt{p_x^2 + p_y^2}. \tag{2.8}$$

2.2.1 The subdetector system of CMS

Starting from the interaction point and going outwards, the subdetector system of CMS consists of: the inner tracking system, the electromagnetic calorimeter, the hadron calorimeter, the superconducting solenoid, and lastly the iron return yoke interspersed with muon chambers. Each layer of subdetectors has a barrel-and-endcaps structure.

The tracking system

The inner tracking system is the closest subdetector to the interaction point; the CMS collaboration is currently preparing for the complete refurbishment of the CMS tracking system [9]. Like the current Tracker, the new Tracker will also be operated in a 3.8 T magnetic field. It has a diameter of 2.5 m and a length of 5.8 m with an acceptance in pseudorapidity $|\eta| < 4$. It is designed to measure the trajectory of the charged particles emerging from the p-p interaction and to provide a reconstruction of the secondary vertex with great

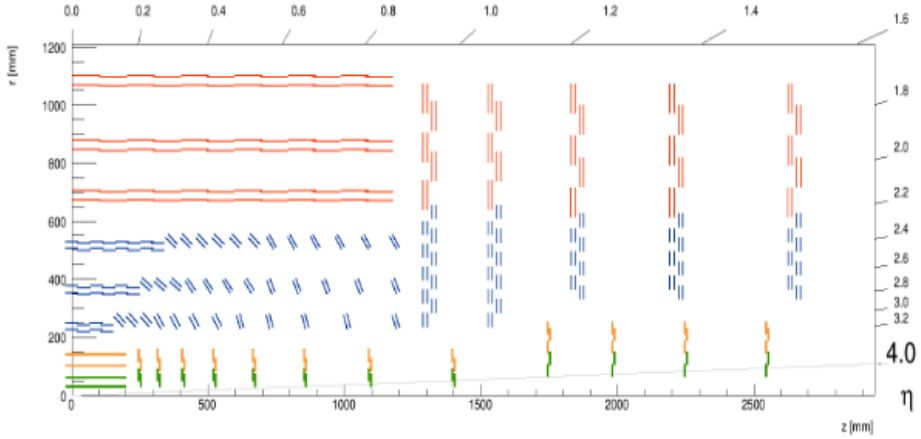


Figure 2.5: Disposition of the detectors in one quarter of the CMS tracker in r-z view.

precision. The layout of the Tracker consists of a barrel part complemented by two endcaps. The barrel part will have ten silicon detector layers as shown in Figure 2.5. The six outmost layers form the Outer Tracker (OT) and the four innermost ones the Inner Tracker (IT). In OT, the layers from eight to ten (red) consist of the strip-strip modules and the layers from five to seven (blue) of strip-pixel modules providing the p_T information in addition to the track position. The barrel part of the Tracker is complemented at larger η by endcap discs. The OT endcap holds five large discs: in Figure 2.5 the strip-pixel modules are marked with blue and the strip-strip modules with red. The barrel section of the strip-pixel modules is gradually inclined in the range of pseudo-rapidity η between 0.6 and 2.2. The Inner Tracker barrel holds layers one to four and it consists of hybrid silicon pixel modules installed into ladder like structures. The IT endcap holds eight small discs in the forward section with four rings of modules each, and four large discs in the high η extension section with five rings of modules. In Figure 2.5, green lines correspond to modules made of two readout chips and orange lines represent larger modules with four chips.

The electromagnetic calorimeter

The electromagnetic calorimeter is a hermetic homogeneous calorimeter made of lead tungstate ($PbWO_4$) crystals, and it extends at a radial distance from the centre of the detector between 1.25m and 1.8m. Avalanche photodiodes (APDs) are used as photodetectors in the barrel and vacuum phototriodes in the endcaps; APDs cannot be used in the endcaps, for that area is subject to radiation too intense for APDs to handle. The use of high density crystals is appropriate for operation at LHC: small density $\rho = 8.28 g/cm^3$, small radiation length $X_0 = 0.89 cm$ and Molière radius $R_M = 2.2 cm$ result in a fast, compact and fine granularity calorimeter. The scintillation decay time of these production crystals is of the same order of magnitude as the LHC bunch crossing time: about 80% of the light is emitted in 25 ns. The barrel part of ECAL

Properties of C26000	
Chemical composition	70% Cu, 30% Zn
Density	8.53 g/cm^3
Radiation Length	1.49 cm
Interaction Length	16.42 cm

Table 2.1: Properties of C26000 Cartridge brass.

(EB) covers a pseudorapidity range of $|\eta| < 1.479$ while the endcaps(EE) cover the range $1.479 < |\eta| < 3.0$. The subdetector layout is shown in Figure 2.6.

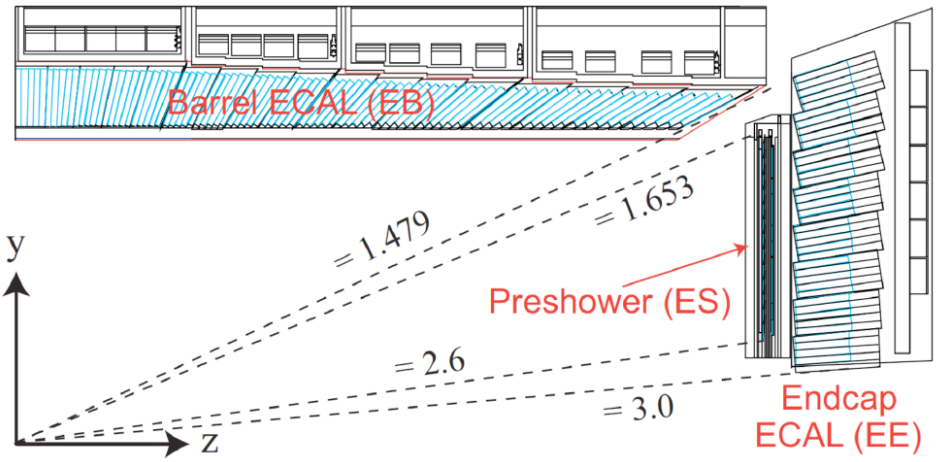


Figure 2.6: Layout of the ECAL of CMS.

The hadron calorimeter

The main purpose of the HCAL is the measurement of the energy of hadrons produced in the collision; it is also fundamental in the measurement of the neutrino contribution and of the exotic particles resulting in apparent missing transverse energy. Covering the pseudorapidity range of $|\eta| < 1.3$, the barrel part of the detector (HB) is a sampling calorimeter made of layers of active material, a fluorescent scintillator, and absorber, C26000 Cartridge brass, whose properties are listed in Table 2.1.

The endcaps (HE) cover a substantial portion of the rapidity range $1.3 < |\eta| < 3.0$, approximately 13.2% of the solid angle. The calorimeter is inserted into the ends of a 4 T superconducting magnet on the outer side, so the absorber must be made from a non magnetic material, with the further requirements of high enough interaction length (for the containment of the showers), good mechanical properties and reasonable cost; therefore C2600 brass was chosen. Since in the central pseudorapidity range the combined stopping power of HE and HB combined does not provide sufficient containment for hadron showers, the hadron calorimeter is extended outside the solenoid with a tail catcher called the outer calorimeter (HO). The HO utilises the solenoid coil as

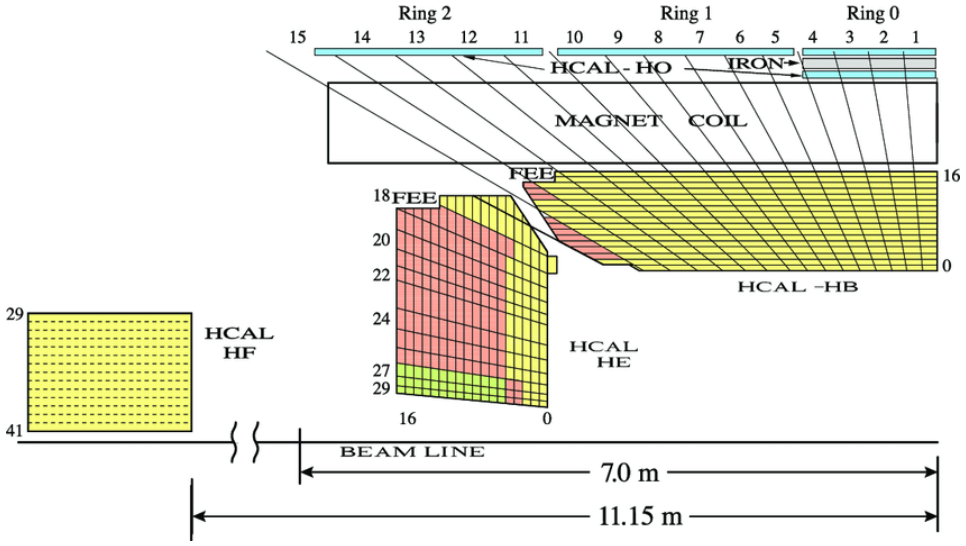


Figure 2.7: Layout of the HCAL of CMS.

an additional absorber (plus a few scintillators) equal to $1.4/\sin\theta$ interaction lengths and is used to identify late starting showers and to measure the shower energy deposited after HB; its goal is to provide adequate sampling depth for $|\eta| < 1.3$. Furthermore, a forward section (HF) is put at a radial distance of 11.2 m from the interaction point along the z-axis. It covers the $3 < |\eta| < 5.2$ range and it is made of quartz fibers embedded within a 165 cm long steel absorbers and is sensitive to Cherenkov radiation ($E > 190keV$ for electrons). A layout of the detector is shown in Figure 2.7. The main parameter used to describe the performance of both the calometers is the energy resolution, whose main contributions are:

$$\left(\frac{\sigma}{E}\right)^2 = \left(\frac{a}{\sqrt{E}}\right)^2 + \left(\frac{b}{E}\right)^2 + c^2 \tag{2.9}$$

where:

- a is a stochastic term accounting for the fluctuations in the number of primary particles and number of photons produced by charged particles;
- b is a noise term representing the electronic noise and pile-up energy;
- c is a constant term standing for calibration errors and leakage.

For CMS, these parameters, which differ between the two calometers, are reported in Table 2.2.

The Superconducting Magnet

The superconducting magnet for CMS has been designed to reach a 4 T field in a free bore of 6m diameter and 12.5m length with a stored energy of 2.6 GJ at full current. This coil surrounding the tracker and calorimeter systems

has the purpose of bending the trajectory of charged particles in the detector to measure their transverse momentum. An iron yoke surrounds the magnet in order to avoid border effects and bend the field lines so that the magnetic field outside the solenoid bore is approximately constant and equal to 1.8 T. A simplified layout of the module and lines of the magnetic field is shown in Figure 2.8

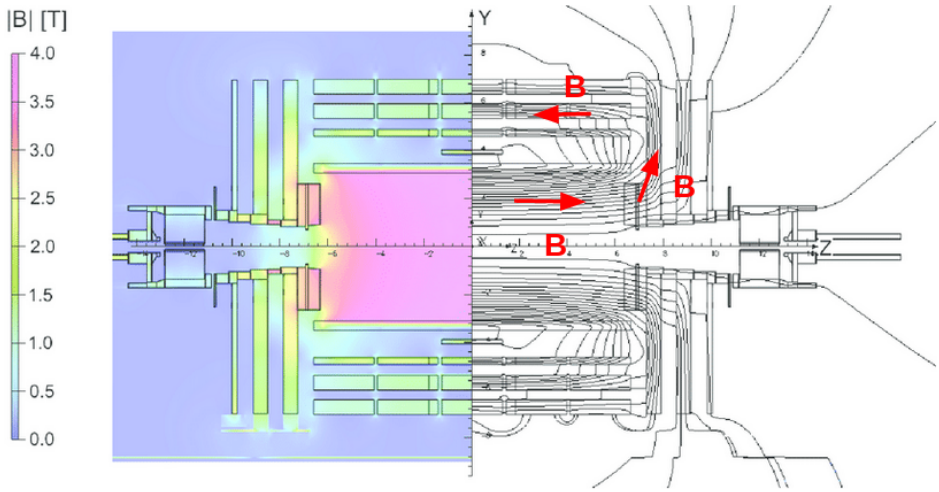


Figure 2.8: A view of the module of the magnetic field (left) and its field lines(left); is it noticeable how the field is constant inside the yoke and equal to 4 T, while it is almost constant (1.8 T) in the plates.

The muon system

The muon system has 3 main functions: muon identification, measurement of their momentum and triggering. Due to their mass, muons emit less deceleration radiation (Bremsstrahlung) than electrons, so that they are able to penetrate through many layers of materials without stopping, therefore preventing their energy measurement. CMS uses 4 types of gaseous particle detectors for muon identification: drift tubes, cathode strip chambers, resistive plate chambers(RPCs), and gas electron multipliers detectors (GEMs), arranged in cylindrical symmetry over the inner solenoid magnet. Because the muon system consists of about $25000m^2$ of detection planes, the muon chambers had to be inexpensive, reliable, and robust. The drift tubes (DTs) form the barrel region of this system and cover the $[0, 1.2]$ pseudorapidity range. This barrel consists of 4 stations forming concentric cylinders around the beamline, so

Parameter	ECAL	HCAL
a	0.0280	0.8470
b	0.12	0
c	0.003	0.074

Table 2.2: Parameters for the ECAL and the HCAL, reported in GeV.

that three of them are used for measurements in the $r - \phi$ plane, while the other is used for the z-axis. The cathode strip chambers (CSCs) form the endcaps, where the muon rate is higher. In addition to high radiation resistance, they also have good segmentation and fast response, covering a pseudorapidity range $0.9 < |\eta| < 2.4$. The cathode strips of each chamber are used for position measurements in the $r-\phi$ plane, while the anode strips grant pseudorapidity measurements and the beam-crossing time for each muon. The RPCs are placed in both the barrel and the endcaps, for they combine adequate spatial resolution with a time resolution comparable to that of scintillators. These qualities are perfect for triggering purposes. GEMs represent the latest addition to the muon system in CMS. By complementing the already existing systems in the endcaps, GEM chambers provide additional redundancy and measurement points, allowing a better muon track identification and also wider coverage in the very forward region.

The Trigger system

The LHC provides proton-proton and heavy-ion collisions at high interaction rates. The beam crossing interval for protons is 25 ns, corresponding to a 40 MHz crossing frequency. Multiple collisions occur at each crossing of the bunches (20 collisions at the nominal design luminosity $10^{34} cm^{-2} s{-}1$), so that is impossible to store and process that large of an amount of data on the high number of events; a reduction in the rate of stored events is performed by the trigger, which is composed of Level-1 trigger (L1 trigger) and High Level Trigger (HLT). This system reduces the rate of a factor of at least 10^6. For reasons of flexibility the L1 Trigger hardware is implemented in FPGA technology where possible, but ASICs and programmable memory lookup tables (LUT) are also widely used where speed, density and radiation resistance requirements are important. A software system, the Trigger Supervisor, controls the configuration and operation of the trigger components. The L1 Trigger has local, regional and global components. The local components are based on energy deposit in the calorimeters, track segments or hit patterns in the muon chambers; regional triggers combine the information provided by the local ones, sort objects as electron and muon candidates and pass this information on to the Global Muon Trigger or to the Global Calorimeter Trigger, which are connected to the Global Trigger. The main task of the Global Trigger is deciding whether rejecting an event or passing it to the HLT, which is a software online trigger that performs a further evaluation running quality reconstruction modules and filters to process and select events for storage.

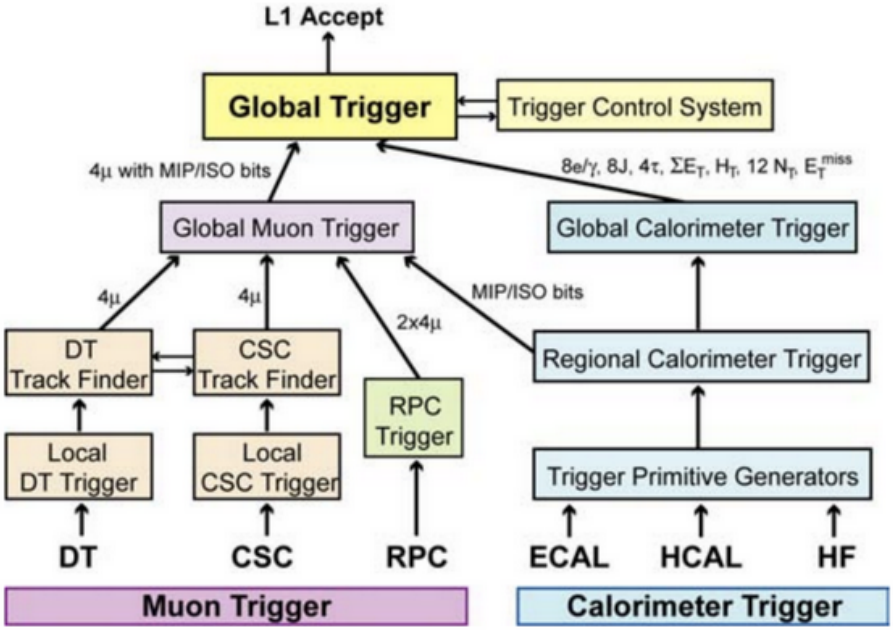


Figure 2.9: Architecture of the L1 trigger.

Chapter 3

Physics Beyond Standard Model

The SM is considered the most successful model in Particle Physics to date. It is highly accurate in the feat of describing three out of the four fundamental interactions (electromagnetic, weak, and strong) and it holds an unbeaten number of experimentally verified predictions. However, albeit being one of the most outstanding theories formulated over the course of the 20th century, it is inherently incomplete: many observed phenomena find no explanations in the context of the SM, for instance the existence of Dark Matter, Dark Energy, and the matter-antimatter asymmetry. Furthermore, as a theory, it has some unsatisfactory aspects, for example it does not explain the fine tuning of the mass of the Higgs boson nor the hierarchy of fermion masses, and why the gravitational interaction has such a smaller characteristic scale with respect to the other fundamental interactions. Numerous models have been crafted in an attempt to provide a solution to these problems; these theories are labelled as theories Beyond Standard Model(BSM). Almost all of these theories predict the existence of new spin-1 gauge bosons W' and Z', with the same quantum numbers of the SM bosons for weak interactions, W and Z. However, they could differ in terms of the coupling strength and chirality: depending on the model, they could couple either left or right-handedly or could favour couplings with a particular generation. This last instance could potentially give some insight on the anomalies regarding lepton universality raised by the BaBar, Belle, and LHCb experiments [10]. In this chapter, a brief description of some of these models is presented. The Lagrangian for the interaction of these bosons with other SM particles is then described; finally, the current state of search for the W' gauge boson is reported.

3.1 Unsolved Problems of the Standard Model

In order to understand the need for new theories beyond the scope of the SM, it is important to acknowledge its limitations, which could be classified into shortcomings and formal problems.

- **Gravity:** the SM does not account for gravitational interactions. General Relativity is widely believed to be incompatible with the SM, for

there seems to be no explanation of the great difference between the Planck and the SM interaction scales. Moreover, theories based on the addition of graviton (namely, a gravity quantum) do not predict what is experimentally observed withouth further modification to the SM framework.

- **Dark Matter:** in 1975, V.Rubin discovered that most stars in spiral galaxies orbit at approximately the same speed, meaning that the rotation curve of these galaxies remains flat instead of decreasing as the distance from the centre increases. This implies that galaxy masses grow approximately linearly with radius. This is, among others, one of the most compelling evidence of the existence of another form of matter, called Dark Matter (from the German *dunkle Materie*), not described by particles included in the SM.

- **Dark Energy:** from cosmological observations, namely red shift measurements, it is known that the Universe is accelerated. Since only 5% of the Universe is composed of ordinary matter (also called Baryonic matter) and Dark Matter amounts approximately its 23%, the rest is believed to be consisted of a form of energy. The latter is called Dark Energy, which is hypothesised to permeate all of space, tending to accelerate the expansion of the universe. This phenomenon has no explanation in the SM framework.

- **Matter-antimatter asymmetry:** SM predicts that matter and antimatter should have been created in equal amounts. However, there is a great imbalance in favour of matter over antimatter, and the sole SM CP-violation in the quark sector is not enough to justify the measured imbalance.

- **Existence of neutrino masses:** according to the predictions of the SM, neutrinos should be massless. However, neutrino oscillation observed by the Super-Kamiokande Observatory and the Sudbury Neutrino Observatories prove that neutrinos do, in fact, possess mass.

- **Flavour Changing Neutral Current:** the suppression of flavour changing neutral currents is not predicted by the SM. These processes are not present at tree level, and the unitarity of the CKM matrix creates suppression in loop processes. Theories regarded as extensions of the SM generate new flavour changing neutral current processes, leading to signals which, if observed, would be unambiguous evidence of new interactions.

Moreover, there are intrinsic theoretical problems in the SM:

- **Number of Adjustable Parameters:** SM depends on 18 numerical parameters: 6 quark masses, 3 lepton masses (since neutrino masses are zero in the SM), 1 Higgs mass, 4 mixing angles from the CKM matrix,

one mixing angle for QCD and 3 coupling constant. Their values are known from experiment, but the origin of the values is unknown.

- **Hierarchy Problems:** particle masses are introduced through spontaneous symmetry breaking, as already said in Section 1.4. The tree level Higgs mass receives corrections from fermion loop diagrams which are quadratically-divergent and that are not cancelled by the boson loop diagrams. Within the present framework, the Higgs mass should be of several orders of magnitude greater than the observed one.

3.2 Models predicting W' and Z'

The presence of new W' and Z' bosons are predicted by almost all BSM theories. They have properties similar to the W and Z bosons: they have integer spin values equal to 1, they are electrically charged and neutral respectively and they mediate the charged and neutral current processes. One of the main differences between these new resonances and their SM counterpart is that, in most models, the former have significantly larger masses. It is to be noted that models for which the masses of these new particles are very small exist, but they will not be discussed in this overview. These heavy resonances could be detected via Drell-Yan processes, that have a quite clean di-lepton or lepton-neutrino final state. These bosons also have a significant coupling with quarks, but hadronic final states are usually more difficult to identify. The study of final states including a top quark is of particular interest owing to the special properties of the top quark. Some of these models that include the W' and Z' bosons are described below.

Extra dimensions

In 1920, T. Kaluza and O. Klein [11] proposed the introduction of a fifth dimension to the four dimensional space-time. If proven true, this theory would eventually allow for the unification of the electromagnetic and gravitational interactions. The first formulation was given by Kaluza as a simple, purely classical extension of General Relativity in 5 dimensions: 10 out of the 15 components of the metric tensor are associated to the 4 dimensional space-time, 4 are interpreted as the electromagnetic vector potential and 1 component with a scalar field usually referred to as Radion of Dilation. Kalusa's hypothesis was that none of the components of the metric depended on the fifth additional dimension; this is called the Cylinder Condition. Subsequently, Klein contributed to this model, providing the classical formulation by Kaluza with a normalized 5D metric and giving it a proper quantum interpretation. According to this vision, the additional dimension is microscopic (with a radius of $10^{-30}cm$) and curled. The addition of two extra dimensions proposed by Arkani-Hamed, Dimopoulos and Dvali (ADD), the weakness of gravitational interactions could be easily explained. The idea that gravity could only propagate through these additional dimension opened new possibilities in the context

of warped geometry-based models, with effects that could be detected through the current state of experimentation. Physics BSM due to extra dimensions could potentially be detected as a deviation from the SM in experiments like ATLAS and CMS at CERN, for instance by finding evidence of new particles like W', Z', gravitons, or radions. No deviation has been found to date.

Alterative Higgs Models

Many models predicting the W' and Z' bosons are based on the addition of new fields and interactions to the SM. One interesting subcategory of these theories involves the Higgs boson, the last discovered particle of the SM. While its discovery was another fundamental prove of the validity of the SM at scale of electroweak interactions, on the other hand it presents some problems, such as Naturalness, the idea that contributions to its mass should have contributions up to the Planck scale ($1.22 \times 10^{19} GeV$). However, the way that these contributions are cancelled out is totally unknown. The models providing a solution for Naturalness are essentially divided in two categories; the first one includes theories based on Supersymmetry, the idea that every particle has a supersymmetric counterpart that could bring contributions able to balance the Higgs mass. The second category is based on the possibility that the Higgs boson may not be elementary, thus some degrees of freedom could be excluded from the mass contribution, therefore explaining why its mass is relatively small. The Little Higgs Models are based on the idea that the Higgs doublet is a Goldstone boson that arises from global symmetry breaking at the TeV scale. The gauge group is obtained from the direct product of many copies of the same group, each one living in additional spacial dimensions[12] [13].

Topflavour Model

There is a rather broad variety of models based on the idea of extending the SM using larger gauge groups; one of these theories is the Topflavour model, an extension of the electroweak theory [14] [15]. The symmetry group is $SU(2)_1 \times SU(2)_2 \times U(1)_Y$: the first and second generations of fermions only couple to $SU(2)_1$, the third one couples to $SU(2)_2$. The fermions of the first two generations have the following representation under the three groups composing the gauge group:

$$(U, D)_L \longrightarrow \left(2, 1, \frac{1}{3}\right), \; U_R \longrightarrow \left(1, 1, \frac{4}{3}\right), \; D_R \longrightarrow \left(1, 1, \frac{2}{3}\right) \tag{3.1}$$

$$(\nu_l, l)_L \longrightarrow (2, 1, 1), \; l_R \longrightarrow (1, 1, -2). \tag{3.2}$$

where $(U, D) = (u, d), \; (c, s)$ and $l = e, \mu, \tau$. The fermions of the third generations have the following representation:

$$(t, b)_L \longrightarrow \left(1, 2, \frac{1}{3}\right), \; t_R \longrightarrow \left(1, 1, \frac{4}{3}\right), \; b_R \longrightarrow \left(1, 1, \frac{2}{3}\right) \tag{3.3}$$

The appropriate form of the covariant derivative is:

$$D_\mu = \partial_\mu - i\frac{g'}{2}YB_\mu - ig_1\vec{T}\vec{W}_\mu - ig_2\vec{\tilde{T}}\vec{\tilde{W}}_\mu \tag{3.4}$$

in which g_1 and g_2 are the coupling constants, $\vec{W}_\mu$ and $\vec{\tilde{W}}_\mu$ are the interaction fields for the $SU(2)_1$ and the $SU(2)_2$ groups respectively. After having obtained the SM $SU(2)_L \times U(1)_Y$ group from the standard Higgs mechanism, the symmetry breaking proceeds as follows. A Higgs field Φ is introduced; it transforms as a doublet under $SU(2)_1$ and $SU(2)_2$ and has a vacuum expectation value (*vev*):

$$\Phi = \frac{1}{\sqrt{2}}\begin{pmatrix} u & 0 \\ 0 & u \end{pmatrix} \tag{3.5}$$

Subsequently, a doublet Higgs field with vev$=$ v is introduced. The mass matrix obtained for the neutral sector of the theory (the basis order is $W, \tilde{W}, B$) is:

$$\frac{1}{2}\begin{pmatrix} g_1^2u^2 & -g_1g_2u^2 & 0 \\ -g_1g_2u^2 & g_2^2(v^2+u^2) & -g'g_2v^2 \\ 0 & -g'g_2v^2 & g'^2v^2 \end{pmatrix} \tag{3.6}$$

Using an appropriate orthogonal matrix R, the previous matrix is diagonalized:

$$\begin{pmatrix} A \\ Z_l \\ Z_h \end{pmatrix} = R\begin{pmatrix} W_3 \\ \tilde{W}_3 \\ B \end{pmatrix} \tag{3.7}$$

with A, Z_l, Z_h being the mass eigenstates. The coupling constants are:

$$g_1 = \frac{e}{\cos\phi\sin\theta_w},\ g_2 = \frac{e}{\sin\phi\sin\theta_w},\ g' = \frac{e}{\cos\theta_w}, \tag{3.8}$$

in which θ_w is the standard weak mixing angle and ϕ is an additional mixing angle. The eigenstate A is identifiable as the photon for it has zero mass. Z_l and Z_h masses are obtained by solving the following equation:

$$M_Z^4 - \frac{1}{2}u^2(g_1^2 + g_2^2 + g'^2\epsilon + g_2^2\epsilon)M_Z^2 + \frac{1}{4}u^4\epsilon(g_1^2g'^2 + g_1^2g_2^2 + g_2^2g'^2) = 0 \tag{3.9}$$

where $\epsilon = v^2/u^2$ Z_l has lower mass and it is the eigenstate that represents the Z boson. The charged sector has mass matrix (basis order is $W, \tilde{W}$):

$$\frac{1}{2}\begin{pmatrix} g_1^2u^2 & -g_1g_2u^2 \\ -g_1g_2u^2 & g_2(v^2+u^2) \end{pmatrix} \tag{3.10}$$

Following the same procedure used for the neutral sector, performing diagonalization via an orthogonal matrix R':

$$\begin{pmatrix} W_l \\ W_h \end{pmatrix} = R'\begin{pmatrix} W \\ \tilde{W} \end{pmatrix} \tag{3.11}$$

and subsequently solving the following equation:

$$M_W^4 - \frac{1}{2}u^2[g_1^2 + g_2^2(1+\epsilon)]M_W^2 + \frac{1}{4}u^4g_1^2g_2^2\epsilon = 0. \tag{3.12}$$

W_l is identified as the SM W boson, while W_h is an example of W'.

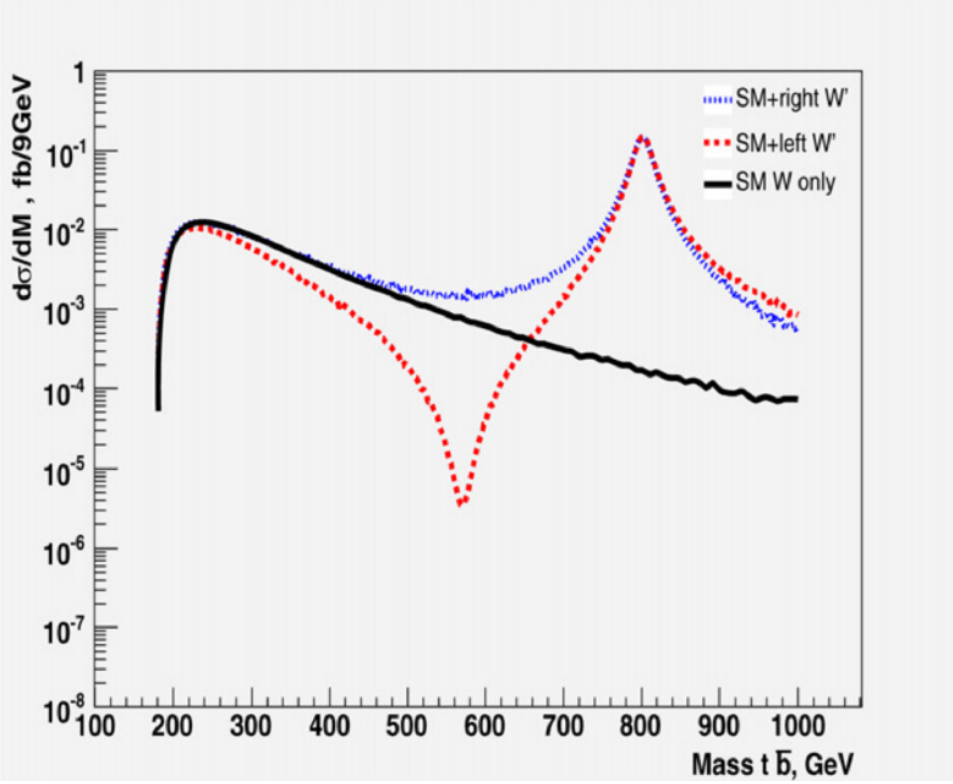


Figure 3.1: Differential cross section for the $pp \longrightarrow W/W' \longrightarrow t\bar{b}$ process [16]; this graphic is obtained considering the invariant mass of the $t\bar{b}$ couple for a simulated $M_{W'}$ of $800GeV$. The image shows three cases; the SM only (black continuous line), the SM + right interacting W' (dotted blue line) and the SM + left interacting W' (dotted red line).

3.3 The interaction Lagrangian

The effective Lagrangian describing the interaction of W' with fermions is, in its most general form:

$$\mathcal{L} = \frac{W'_\mu}{\sqrt{2}} \left[q'_i (C^R_{q_{ij}} P_R + C^L_{q_{ij}} P_L) \gamma^\mu q_j + \bar{\nu}_i (C^R_{l_{ij}} P_R + C^L_{l_{ij}} P_L) \gamma^\mu l_j \right] \tag{3.13}$$

where i and j are the generation indices, q, q', l and ν are SM fermions in mass eigenstates. For W, the coefficients are $C^R_{q_{ij}} = C^R_{l_{ij}} = 0$, $C^L_{q_{ij}} = g_w V_{CKM}$ and $C^L_{l_{ij}} = g_w$. The simplest SM extension which predicts the W' boson is $SU(2)_1 \times SU(2)_2 \times U(1)_Y$, the group already described in Section 3.2. This group includes a mixing between W and W' in case the latter couples with left-handed currents. Figure 3.1 shows the trend of the differential cross section for the $pp \longrightarrow W/W' \longrightarrow t\bar{b}$ in the case of a simulated W' mass of 800 $GeVs$. It is to be noted that the case of SM and W' with left-handed couplings, there is a local minimum due to the interference between the two bosons [16].

3.4 Decay channels of the W' Boson

As for its SM equivalent, W' counts many decay channels. The total decay width is [17]:

$$\Gamma_{tot}(W') = \Gamma(W' \longrightarrow t\bar{q}') + \Gamma(W' \longrightarrow q\bar{q}') + \Gamma(W' \longrightarrow l\bar{\nu}) \tag{3.14}$$

where the partial width containing the top quark is given separately, for this decay channel is of interest in this thesis work. The leading order partial widths are

$$\Gamma_{LO}(W' \longrightarrow t\bar{q}') = \frac{g^2\beta^2}{16\pi m_{W'}}|V'_{t\bar{q}'}|^2(m^2_{W'} + m^2_t/2), \tag{3.15}$$

$$\Gamma_{LO}(W' \longrightarrow q\bar{q}') = \frac{g^2}{16\pi}|V'_{q\bar{q}'}|^2 m_{W'}, \tag{3.16}$$

$$\Gamma_{LO}(W' \longrightarrow l\bar{\nu}) = \frac{g^2}{16\pi}|V'_{l\bar{\nu}}|^2\frac{m_{W'}}{3}. \tag{3.17}$$

where $\beta = 1 - m_t^2/m_{W'}^2$ and it is assumed that the coupling constant $g = 8m_W^2 G_F/\sqrt{2}$, as it is for the SM. Hence, the partial widths of the W' boson have the same form as the SM W boson, so that new couplings and GCKM matrix elements absorbed into the $V'_{f_i f_j}$ matrix elements. From Equation 3.15, it is evident that W' bosons tend to have a large branching fraction into top quarks; this has repercussions on the size of the single-top-quark production cross section at hadron colliders. This process happens through three channels, as shown in Figure 3.2. Cross sections for the t-channel production and the

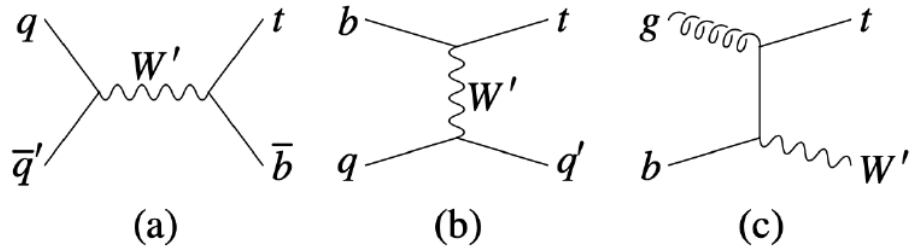


Figure 3.2: Feynman diagrams for single top production including a W' boson; (a) is the s-channel production, (b) the t-channel production and (c) the $W' - t$ associated production.

associated W' production are negligible at the LHC energy scale [17]. W' could also decay into $W - Z$ and $W - H$ pairs. Figure 3.3 displays the upper limits at 95% CL in many two-boson decay channels [18].

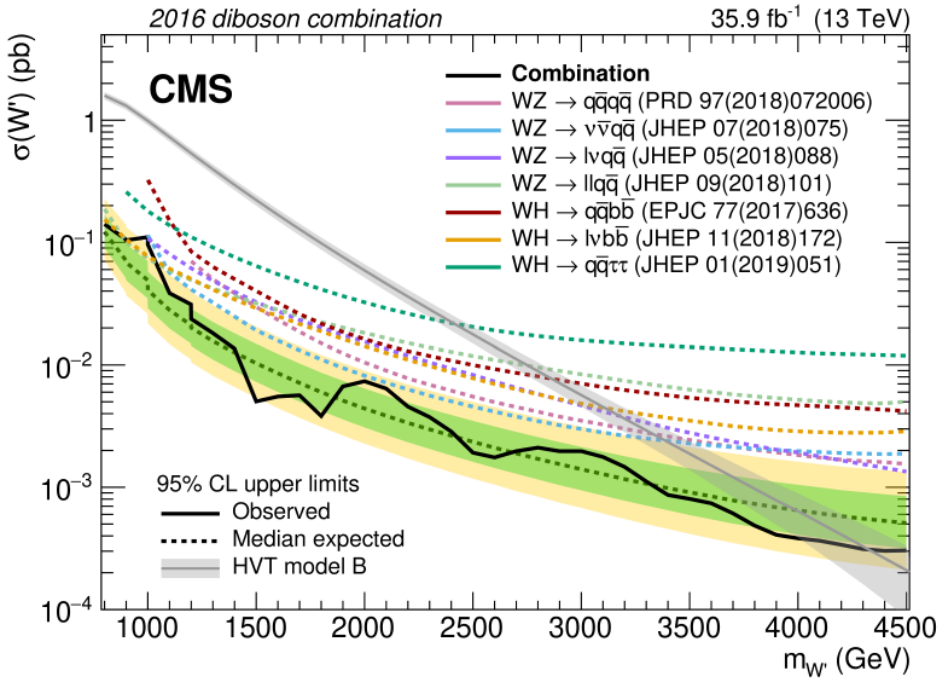


Figure 3.3: Observed and expected 95% CL upper limits on the W' cross section as a function of the W'. The inner green and outer yellow bands represent the ± 1 and ± 2 standard deviation variations on the expected limits of the statistical combination of the VV and VH channels considered (in which V represents either W or Z). The expected limits in individual channels are represented by the colored dashed lines [18].

3.5 Search for W'

The search for these bosons has been conducted both and at Fermilab's Tevatron and at LHC. Among the many processes involving the W' bosons, its decay into a top quark and a bottom quark is of particular interest: the bottom quark results in jet formation, while the top quark can decay either hadronically or leptonically, namely

$$t \longrightarrow b\, W \longrightarrow b\, q\, q' \tag{3.18}$$

$$t \longrightarrow b\, W \longrightarrow b\, l^{+}\, \nu_l \tag{3.19}$$

The branching ratios for these processes are reported in Table 3.1. In 2017, the CMS collaboration published a search for the leptonic decay of the top quark in the e-channel and the μ-channel [19]; the data used in this analysis were collected at $\sqrt{s} = 13\ TeV$ whith an Integrated Luminosity $L = 35.9 fb^{-1}$; Figure 3.4 shows that the expected events are comparable with background predictions, meaning that the search for higher mass points would be limited by the presence of background in the signal region that is not rejected by standard cuts.

Furthermore, the production of right-handed W' bosons is excluded at 95% CL for masses up to 3.6 TeV; Figure 3.5 displays two different theoretical production cross sections, seen as function of the potential sterile neutrino mass

Top Decay Process	BR
$t \longrightarrow b\ q\ q'$	$s(66.5 \pm 1.4)\%$
$t \longrightarrow b\ e^+\nu_e$	$(11.10 \pm 0.30)\%$
$t \longrightarrow b\ \mu^+\nu_\mu$	$(11.40 \pm 0.20)\%$
$t \longrightarrow b\ \tau^+\nu_\tau$	$(11.1 \pm 0.9)\%$

Table 3.1: Branching ratios for both hadronic and leptonic top quark decays.

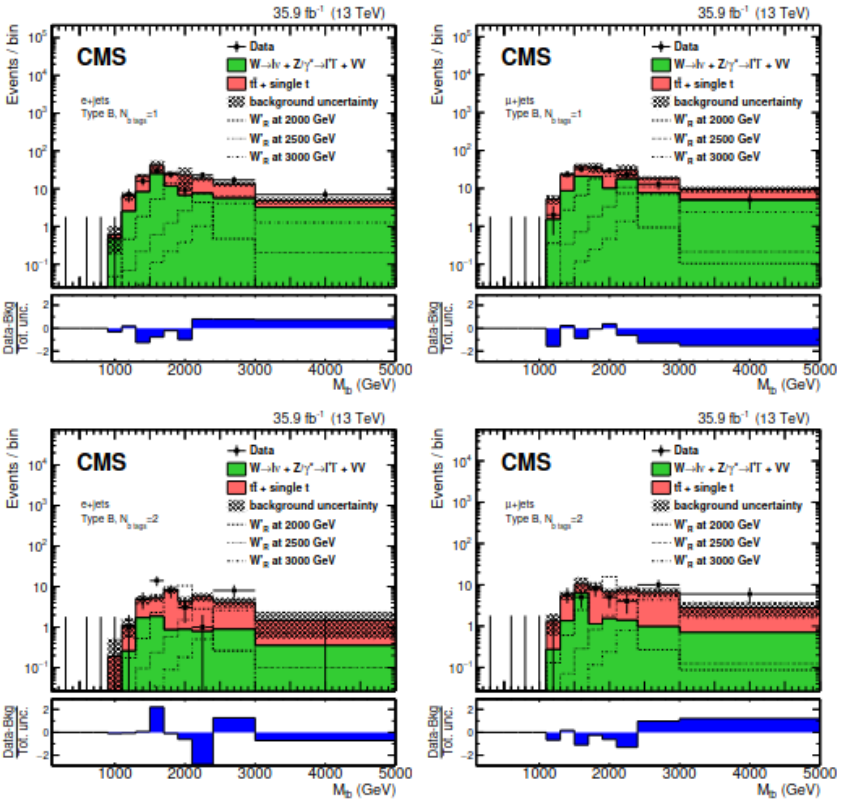


Figure 3.4: Reconstructed invariance mass of the b-jet and the top quark with 1 (upper row) or 2 (lower row) b-tagged jets for the e-channel (left) and μ-channel (right) after selection. Distributions for W'_R boson with masses of 2, 2.5 and 3 TeV are shown.

m_{ν_R}. If there is a right-handed neutrino ν_R and $M_{W'} > m_{\nu_R}$, the BR for the $W' \longrightarrow tb$ would have to decrease in order to account for the ulterior channel $W' \longrightarrow \nu_R l$. A similar search was carried out by the ATLAS experiment [20], with an integrated luminosity of $36.1 fb^{-1}$, and similar results were obtained, excluding at 95% CL the existence of W'_R for masses up to 3.15 Tev, as shown in Figure 3.6. Another search performed at CMS considered all-hadronic final states and exploited a Machine Learning algorithm to recognize the hadronic jets which originated from a top quark. An integrated luminosity of $137 fb^{-1}$ collected at $\sqrt{s} = 13TeV$ was used for this analysis, which excluded the existence of both right-handed and left-handed W' with mass below $3.4TeV$ at 95% confidence level, as shown in Figure 3.7. [21].

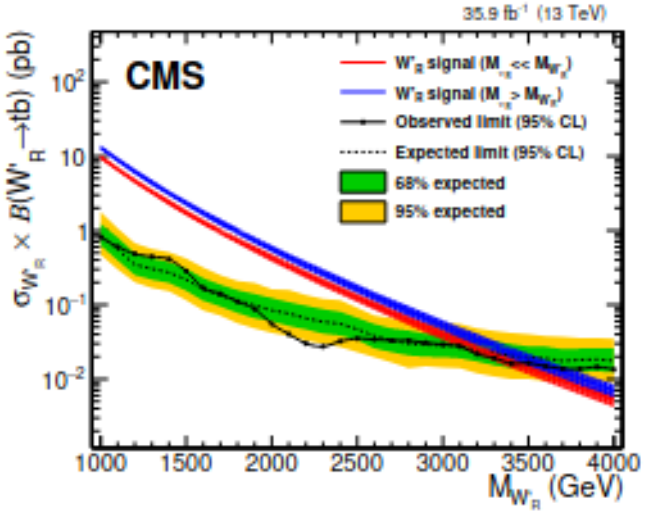


Figure 3.5: Upper limit at 95% CL on the W'_R boson production cross section for the combined electron and muon channels. Signal masses for which the theoretical cross section (red and blue) exceeds the observed upper limit (solid black) are excluded. The green and yellow bands represent the ±1 and 2 standard deviation uncertainties in the expected limit, respectively.

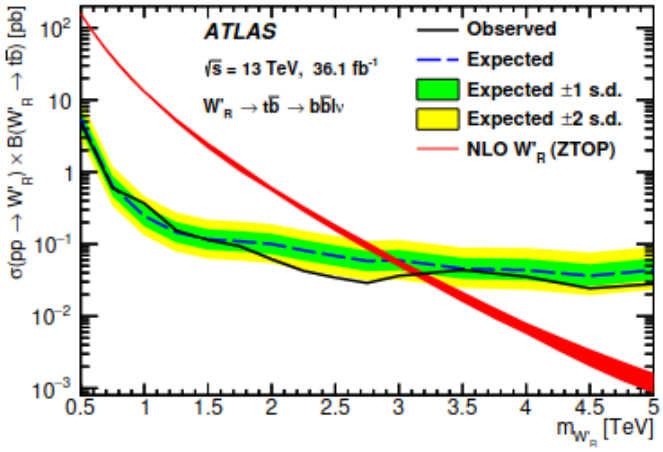


Figure 3.6: Upper limit at 95% CL on the W'_R boson production cross section times $W'_R \longrightarrow b\bar{b}$ branching fraction as a function of resonance mass. The solid curve corresponds to the observed limit, while the dashed curve and shaded bands correspond to the limit expected in the absence of signal and the regions enclosing one/two standard deviation fluctuations of the expected limit

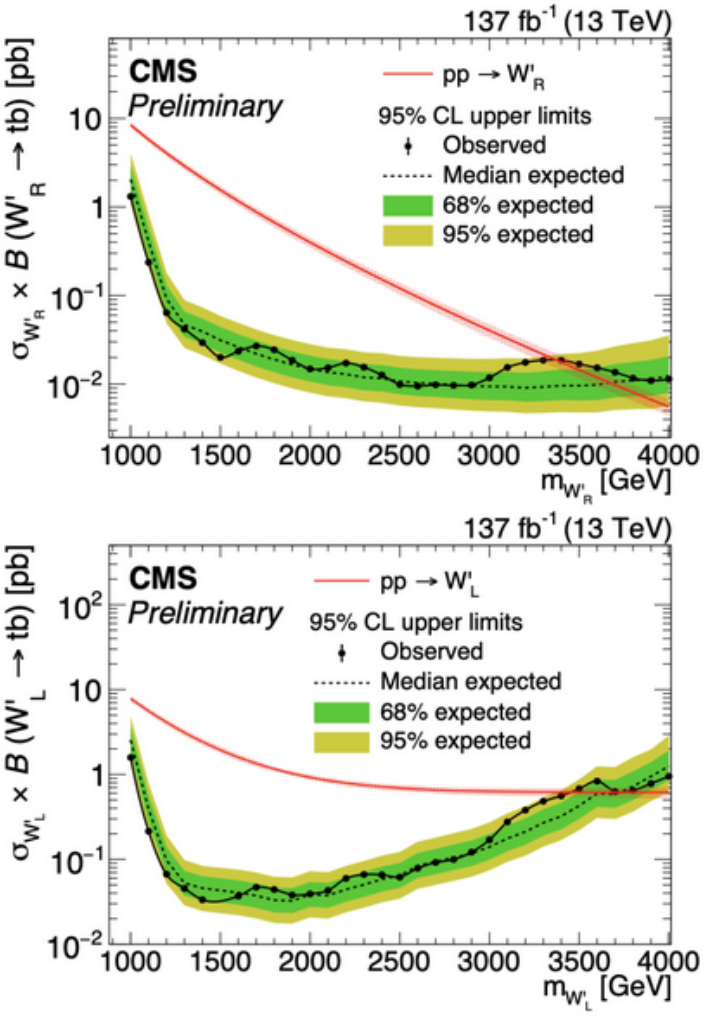


Figure 3.7: Upper limit at 95% CL for the production of W'_R boson (top) and W'_L boson (bottom). The two cross sections differ because a left-handed W' boson would undergo interference with the SM W boson.

Chapter 4

Object Selection and Reconstruction

The aim of this work is the search for the W' boson via the analysis of the $W' \longrightarrow t\ b$ channel. which is of utmost importance in the context of the models described in Chapter 3 that foresee enhanced couplings to third generation quarks. In such cases, i.e. Topflavour described in Section 3.2, the decay width for the $\Gamma(W' \longrightarrow tb)$ process can be fairly large, which has also consequences on the experimental side. The top quark decays into a b quark and a W boson, further cascading to a lepton-neutrino pair; Figure 4.1 shows the $W' \longrightarrow tb$ process with the top quark decaying into a muon. As already stated in Chapter 3, the main production channel for the W' boson is the s-channel, for the cross sections for the t-channel production are negligible at the LHC energy scale [17]. The leptonic decay channel considered in this work has a relatively small background from QCD multijet processes, so that, despite having a lower branching fraction in comparison to its hadronic counterpart (as seen in Chapter 3, Table 3.1) it is relatively easier to analyze. Since the W' boson is expected to have a large mass for the models considered, the Lorentz boost for the top quark decay products will be high in the laboratory frame of reference; therefore, the final state products will tend to collimate along the direction of the top quark momentum.

4.1 Physics Object Selection

In order to identify the W' boson, b quark hadronization products, muons or electrons, and neutrinos must be selected amidst all the final products of a collision event in the CMS detector. After passing through the subdetector system described in Chapter 2, these particles are parsed by the Particle Flow (PF) algorithm [22]; its main goal is the identification and reconstruction of each individual particle arising from the LHC $p-p$ collision. The PF algorithm operates by combining the basic information derived from all the layers of the CMS detector (namely, tracks and clusters) to obtain each final state product. The reconstruction of the physics objects follows a precise order:

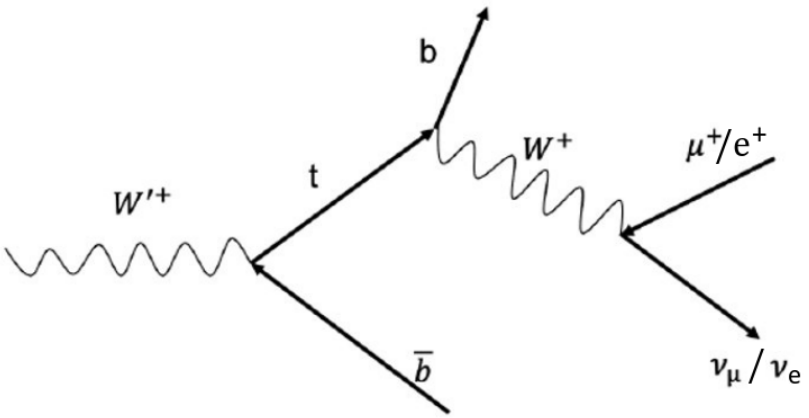


Figure 4.1: Feynman diagram for the $W' \longrightarrow tb$ process, with the final state of the top quark decay consisting of a jet-lepton-neutrino triplet.

1. *muons*, in this case muons are the first particles to be identified: tracks in the inner tracker and in muon detectors are combined with the energy releases in the calorimeter;

2. *electrons and single photons* are then reconstructed at the same time by combining the energy in the ECAL and the tracks in the inner tracking system. If clusters in the ECAL are matched to charged particle tracks, they are identified as electrons. If not, they are identified as isolated photons;

3. *hadrons and non-isolated photons* are the last objects reconstructed by the algorithm, that combines information incoming from both the ECAL and HCAL. Neutral hadrons or non-isolated photons result in clusters that do not correspond to any tracks, otherwise they are considered charged hadrons;

4. *missing transverse energy and hadronic jets* are high-level objects, obtainable only by the combination of all the previously acquired information on the already reconstructed objects.

Tracks and clusters are later taken out from further processing, for they consist of low-level information, as most of the analyses make use of the PF objects rather than their low-level components. In the case of the analysis in question, τ particles are not considered, for they mainly decay hadronically, although their leptonic final states are note vetoed in the following study.

4.1.1 Leptons

Muons

The high level muon physics objects are obtained by performing a thorough combination of information arising from different parts of the detector. For

instance, the outer muon spectrometer allows muons to be identified with high efficiency over the full detector acceptance. This is mainly due to the fact that the calorimeters absorb most of the particles (with the exception of said muons and neutrinos), granting high purity by way of energy deposits. Furthermore, the inner tracker provides a precise measurement of the momentum of these muons. The final collection consists of three different muon types:

- **Standalone Muons:** obtained by only fitting the hits in the muon detector, they have a minimum transverse momentum of approximately 3 GeV, in order to be able to cross one half of the entire detector;
- **Global Muons:** are the result of the matching of Standalone muons with tracks in the inner tracker
- **Tracker Muons:** tracks with $p_T > 0.5$ GeV and a total momentum larger than 2.5 GeV are propagated to the muon system; if there is a match with the hits of the muon system, the track qualifies as a tracker muon.

Global muons have the highest reconstruction efficiency if they have a momentum higher than 10 GeV, that corresponds to the case in which they have hits in at least two muon stations. For lower values of momentum, muons suffer from multiple scattering in the iron of the return yoke, so their efficiency is worsened. Standalone muons could be affected by contamination from the cosmic rays muon component, able to reach the detector. Therefore, tracker muons turn out to be the ones with the higher efficiency [23]. Since lower p_T leptons are unlikely to be generated from a top quark, only muons with $p_t > 10$ are considered in the subsequent analysis.

Electrons

The conventional seeding method for electrons exploits tracker and ECAL measurements. Energetic clusters within the ECAL with an $E_T > 4\ GeV$ are taken into consideration; their energies and positions are used to infer the position of the hits in the inner tracker. This method is called ECAL-based approach. Most of the electrons passing through the tracker emit a significant fraction of their energy in the form of Bremsstrahlung photons that convert in electron-positron pairs. Therefore, the performance of the seeding method relies on the ability to gather and evaluate this radiated energy; ECAL clusters with a small window in η and an extended window in ϕ are grouped into Superclusters, which are used to collect the energy of electrons and possible bremsstrahlung photons. The ECAL-based approach fails for electrons in jet and electrons with small p_T[24]. As for the case of muons, only electrons with $p_t > 10$ are considered in the subsequent analysis.

Lepton Isolation

In order to make a proper event selection, the top quark must be reconstructed from *prompt* leptons: a prompt lepton comes from the parton-parton elemen-

tary interaction vertex, therefore they have a generally lower impact parameters with regards to their counterpart, non-prompt leptons. The latter are generated in two ways: either through the decay of the jets, or a as a result of mis-identification. In the first case, tracks and hits left by non-prompt leptons are correlated to those hits coming from jets or b-tagged activity, therefore they are singled out as non-prompt and thusly excluded. In the case of mis-identification, due to a particular jet signature or a fault in a part of the detector, a jet is reconstructed as a lepton. This fake leptons will also be considered as non-prompt. In order to select prompt leptons, both electrons and muons, and to reject the leptons produced in jets through the decay of flavoured hadrons or of charged pions and kaons, a quantity called *isolation* has been defined. Due to the high number of interaction per bunch crossing and the high boost of the final state particles, the leptons can be misidentified as jet or viceversa. To prevent this from happening, the lepton track is required to be isolated in a fixed size cone around the lepton. This isolation is quantified by estimating the p_T of the particles emitted around the direction of a lepton, and is is defined as:

$$I_{PF} = \frac{1}{p_T}\left(\sum_{h^{\pm}} p_T^{h^{\pm}} + \sum_{h^{\gamma}} p_T^{h^{\gamma}} + \sum_{h^0} p_T^{h^0}\right), \tag{4.1}$$

in which the sums run over the charged hadrons ($h^{\pm}$), photons (γ), and neutral hadrons ($h^{\pm}$) with an angular distance ΔR (see definition in Chapter 2, Section 2.2) to the lepton smaller than either 0.3 or 0.4 in the (η, ϕ) plane. In the subsequent analysis, *Mini-Isolation* is used, for it allows to recover efficiency when leptons are produced in the decay chain of boosted objects. When the boost is large, standard isolation cuts fail, because the lepton overlaps with the jet produced in the same decay chain. Therefore, MiniIso helps evaluating whether or not a lepton in proximity of a hadronic jet is the direct byproduct of said jet. This cone has a variable radius $R(\eta, \phi)$, which varies between 0.2 and 0.05, for $R \propto 1/p_T^{lep}$.

4.1.2 Jets

Because of the colour confinement, quarks and gluons are not observed as free particles. As they travel through the detector, moving away from the point of interaction, they hadronize, meaning that they create jets of colour-neutral hadrons. By analyzing these jets, it is possible to extract information on the partons that generated them. The anti-k_T clustering algorithms [25] are used to reconstruct jets with a radius parameter of 0.4 ; for this very reason the reconstructed jets are called AK4 jets. This algorithm, that provides an infrared-safe and collinear-safe clustering for jets, introduces a distance d_{ij} between the PF candidates (*pseudojets*) i and j, and another distance d_{iB} between the entity i and the beam B. Their definition is:

$$d_{ij} = min\left(k_{ti}^{-2}, k_{tj}^{-2}\right) \frac{\Delta_{ij}^2}{R^2}, d_{iB} = k_{ti}^{-2}, \tag{4.2}$$

in which i is the i-th entity, $\Delta_{ij}^2 = (y_i - y_j)^2 + (\phi_i \phi_j)^2$, k_{ti} is its transverse momentum, y_i is its rapidity and ϕ_i its azimuth. The parameter R symbolizes a radius-like quantity used to module the size of the jet; for the purpose of this study, its value was set as $R = 0.4$. The clustering proceeds by finding the smallest distance for each identity i, against all others: if the minimum a d_{ij} for some j, entities i and j are recombined; if it is d_{iB}, then entity i is identified as a jet and removed from the list of the entities. The distances are then recalculated and the procedure repeated until no entities are left. The entities considered for jet clustering are all the Particle Flow candidates, including muons and electrons, thus their reconstruction inside a jet can be performed. Only jets with $p_t > 30GeV$ were considered in the subsequent analysis.

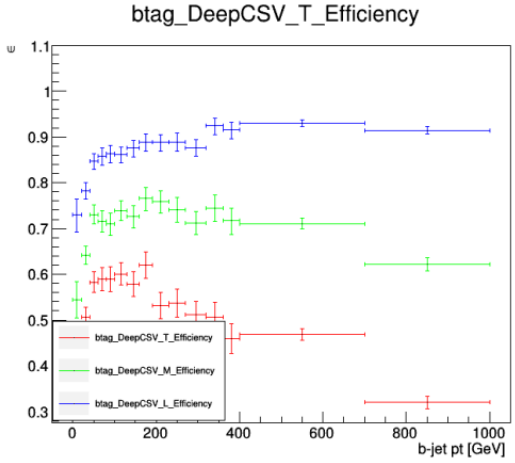
B-jets and b-tagging

In order to reconstruct the W' boson, the identification of hadronic jets originated from b-quarks, called b-jets, is fundamental in the context of this analysis. This operation is called b-tagging, for which several algorithms were developed and provided by CMS, the latest being the DeepCSV and the DeepFlavour [26] [27]. They are based on Deep Neural Networks, Machine Learning algorithms inspired by biological neural networks. These models assign to a jet its probability of being a b-jet. The DeepCSV inherits the inner workings of another algorithm called Combined Secondary Vertex (CSV), which uses the information on secondary vertices, which are the decay vertices of particles arising from the p-p collision, in addition to particle lifetime information. DeepCSV extends the prevous versions of CSV, by extending the range of the maximum considered tracks per jets. The DeepFlavour employs all the features used by DeepCSV and adds additional variables regarding charged and neutral particles in the jet. They are important in the process of identification, for they provide insight on the origin of the hadronic jets. For both algorithms multiple working points are defined at fixed background rejection and studied by the CMS working group:

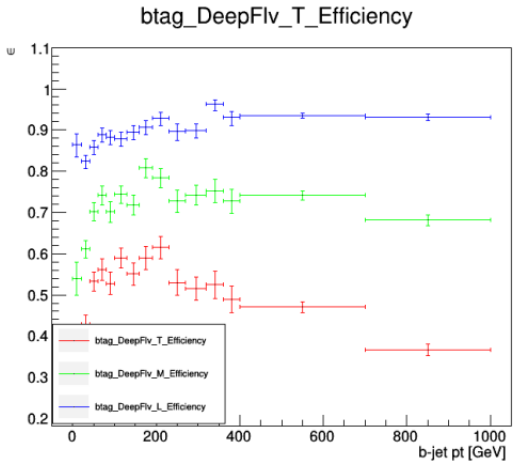
- **Loose:** for which the mistagging rate is approximately 10%;
- **Medium:** for which the mistagging rate is approximately 1%;
- **Tight:** for which the mistagging rate is approximately 0.1%;

A jet is *tagged* if the discriminator value is above some threshold value, often referred to as the cut value. The efficiency is obtained by dividing the number of jets which pass the cut selection by the total number of jets of the same flavour. The tagging efficiency is defined by selecting dependends on the values

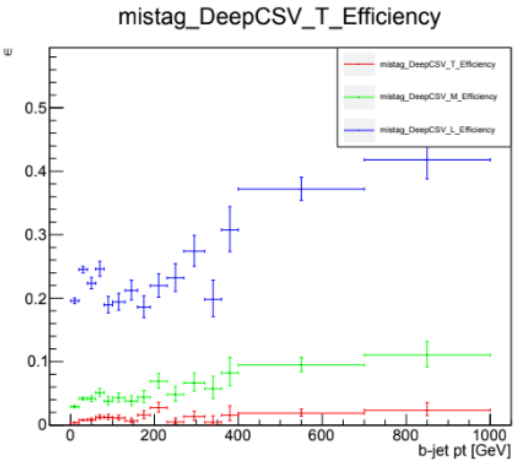
of p_T. Figures 4.2a and 4.2b show that the tagging efficiency of the DeepCSV and the DeepFlavour are similar, while the mis-tagging is better for the DeepFlavour, as show by simply comparing Figures 4.2c and 4.2d. Furthermore, the medium working point seems to be the best in terms of efficiency/mis-tagging balance, therefore, it will be the one used in the subsequent analysis



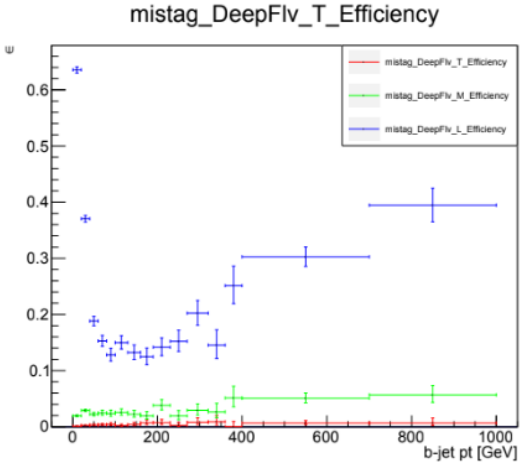
(a) Plot of the b-tag efficiency for the DeepCSV algorithm vs the b-jet p_T.



(b) Plot of the b-tag efficiency for the DeepFlavour algorithm vs the b-jet p_T.



(c) Plot of the b-tag mistagging for the DeepCSV algorithm vs the b-jet p_T.



(d) Plot of the b-tag mistagging for the DeepFlavour algorithm vs the b-jet p_T.

4.1.3 Missing Transverse Energy

The component in the (x-y) plane of the momentum of the particle beams at LHC is called the transverse component; momentum conservation in the transverse plane allows for measurement of the transverse momentum of all the undetected particles. Missing Transverse Energy (MET) is the energy corresponding to this momentum. The definition of said missing momentum (also called raw MET) is:

$$\vec{\not{p}}_t = -\sum_i \vec{p}_{ti} \tag{4.3}$$

where the sum is performed on all PF candidates. In this type of analysis, it is customary to use MET to generically refer to the missing transverse momentum. The MET actually used in analysis at CMS is further corrected and reweighted, in order to account for non-compensating calorimeters, detector misalignments, and energy corrections to reconstructed physics objects, whose the largest contribution comes from jets. In the context of this work, MET is used to gain knowledge on muonic and electronic neutrinos in the final state of the top quark decay, as seen in Figure 4.1.

4.2 Top Quark Reconstruction

Lepton, neutrino, and b-jet are used for the reconstruction of the top quark. The first step of this process is the calculation of the 4-momentum of the top quark by performing the sum of the 4-momenta of the aforementioned objects; this operation is quite complex, for MET is considered as the transverse momentum of neutrinos since there is no way of measuring the z-component of their momentum. The decay width of the W is considered negligible with respect to the experimental resolutions involved with the top quark reconstruction; by imposing $\sqrt{s}(\mu, \nu) = m_W$, the following equation is obtained:

$$p_{\nu,z} = \frac{\Lambda p_{\mu,z}}{p_{\mu,t}^2} \pm \frac{1}{\vec{p_{\mu,t}}} \sqrt{2\Lambda^2 - E_\mu^2 \vec{\not{E}}_t^{\,2}}, \tag{4.4}$$

$$\Lambda = \frac{m_W}{2} \pm \vec{p_{\mu,t}} \cdot \vec{\not{p}}_t, \tag{4.5}$$

in which $\vec{p}_\mu$ and E_μ are respectively the momentum and the energy of the lepton, $\vec{\not{p}}_\mu$ and $\vec{\not{E}}_\mu$ are the missing transverse momentum and energy. The squared root in Equation 4.4 usually has a positive argument and the solution with the smaller absolute value is chosen. In case it is not positive, the imaginary component is eliminated by imposing that the square root is equal to zero; the obtained quadratic relation between $p_{\nu,x}$ and $p_{\nu,y}$ has two solutions and one degree of freedom. The solution with the minimum vectorial distance between the two momenta is chosen. Once the neutrino momentum is obtained, top reconstruction can proceed; a top candidate 4-momentum is the result of the sum of a jet, a lepton and its corresponding neutrino. In order to identify the correct top reconstruction, Machine Learning techniques were used.

4.2.1 Top Categories

Top candidates are reconstructed by selecting the appropriate final state objects, as shown in Figure 4.3. However, Figure 4.4 shows that some objects originated from the previous steps in the decay chain could be incorrectly identified as top quarks, creating a significant source of background for the real

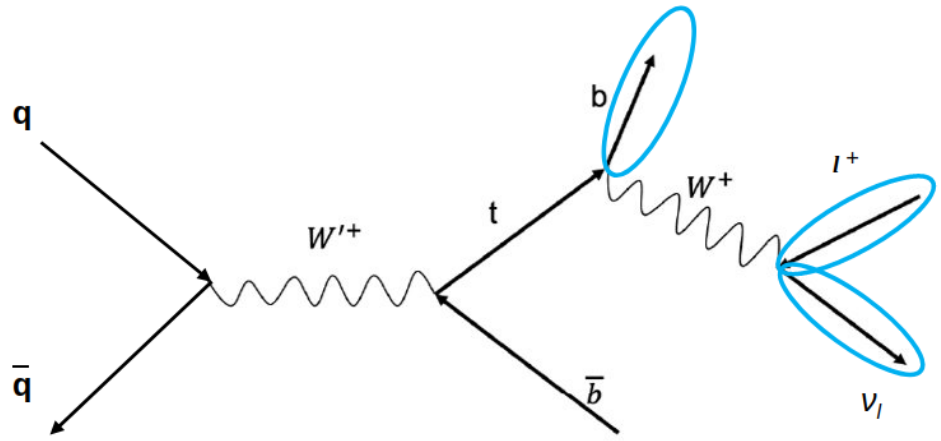


Figure 4.3: Diagram depicting the complete chain of decay from $W' \longrightarrow t\bar{b}$ to $t \longrightarrow bl\nu$. The objects circled in blue are those used for the top quark reconstruction.

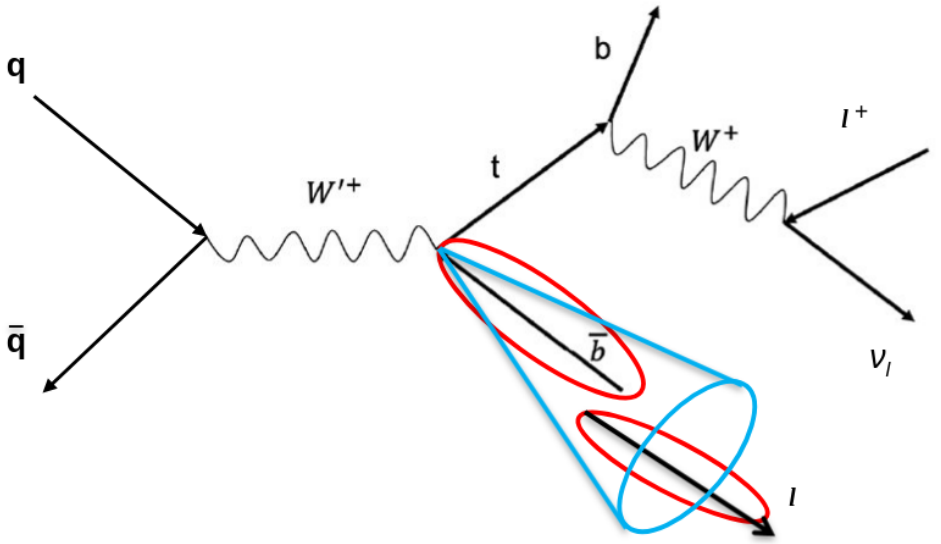


Figure 4.4: The objects circled in red are source of background to the real top candidates.

top quark category. In order to single out only the real tops, it is important to identify the correct objects. Top candidates were divided into two categories based on the angular distance ΔR between the lepton and the jet, shown in Figure 4.5:

- **Merged:** $\Delta R(j, l) < 0.4$;
- **Resolved:** $0.4 < \Delta R(j, l) < 2$;

Signal candidate triplets are constructed by selecting a reconstructed lepton within $\Delta R < 0.4$ from the true lepton, identified via Monte Carlo (MC) truth information. The reconstructed b-jet for the signal category is thusly identified thanks to the pdgId, a particle numbering scheme by the Particle Data Group

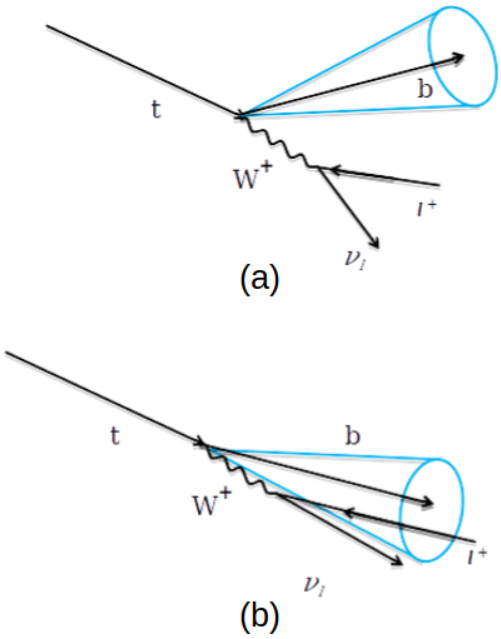
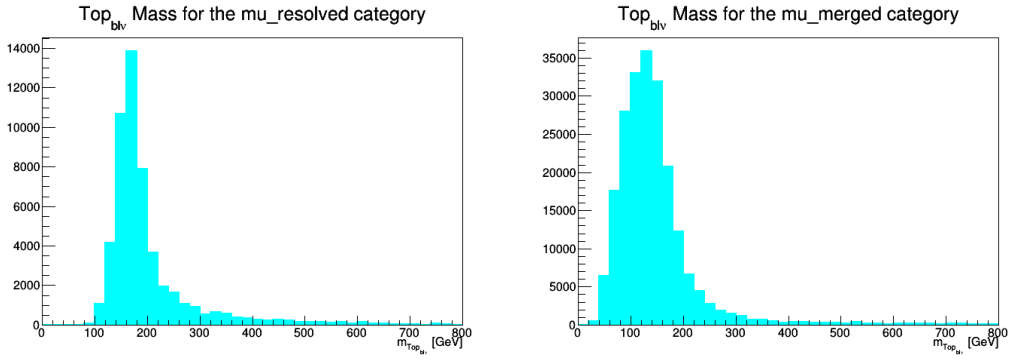


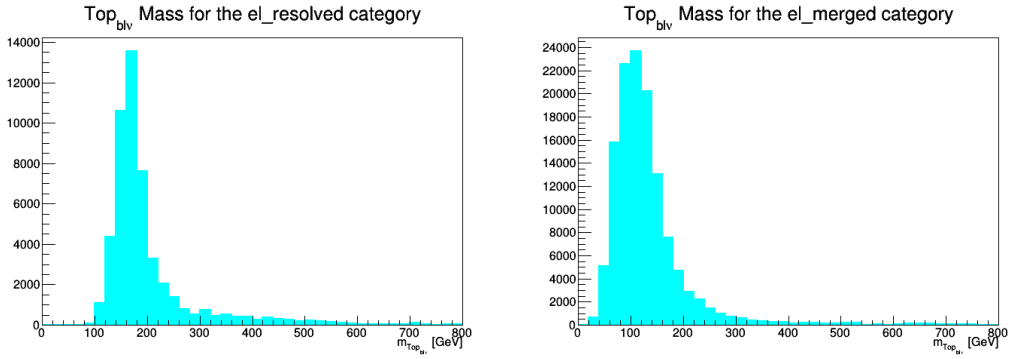
Figure 4.5: Visualization of the merged and resolved top categories. **(a)** represents the "Resolved" kind, in which the lepton and the b-jet are separate and with an angular distance $0.4 < \Delta R(j, l) < 2$; **(b)** is the "Merged" category, in which, as the name suggests, the b-jet includes the lepton in a cone with a $\Delta R(j, l) < 0.4$ cone opening.

used in all modern MC event generators. This scheme assigns a unique code to each type of particle [1]. As seen in Figure 4.4, objects circled in red have a non negligible chance to return large values of top quark mass, given that signal 4-momenta are usually larger than the top quark mass itself due to fluctuations in the energy of the jets or objects involved. The reconstruction was performed on MC simulated signal events, in particular three data files of simulated W' production at LHC. The mass of the right-handed W' boson is respectively 4000, 5000 an 6000 GeV in the three files, with a decay width of 1% of the masses. Figures 4.6 and 4.7 show the distributions of the masses of the top quarks reconstructed with both electrons and muons in the merged and resolved configurations. While the peak for the resolved category is centered around the expected mass value for the top quark (namely 170 GeV), the merged category for both muonic and electronic tops is lower. There are many factors that possibly contribute to this phenomenon, one of them being the uncertainty on the momentum for boosted particles: i.e., when a lepton is produced within a jet, its identification is made difficult by the fact that it tends to pass through the detector. As already seen in Section 4.1.2, b-tagging also reveals failings at higher values of p_T, so that some b-jets used in the reconstruction could have been misidentified. In summary, energy releases from the leptons are not always correctly identified and singled out from the ones arising from jets, resulting into object mis-identification and subsequent lowering of the most probable mass value for the top quark.



(a) (b)

Figure 4.6: Plots depicting the mass distribution of the top quarks reconstructed with the missing transverse energy with a muon in the final state. (a) is the case of the resolved configuration, while (b) is the shape obtained for the merged category.



(a) (b)

Figure 4.7: Plots depicting the mass distribution of the top quarks reconstructed with the missing transverse energy with an electron in the final state. (a) is the case of the resolved configuration, while (b) is the shape obtained for the merged category.

4.3 The Top Standalone category

Another top quark category was taken into consideration, in which the lepton is not reconstructed or it is wrongly identified as a jet or jet fraction. Furthermore, the expected branching ratio for the $W \longrightarrow t\bar{b}$ is approximately be 10.8% per lepton, while the observed efficiency in simulation of signal events was lower by a factor approximately equal to 58%, and it was observed that the number of top quarks reconstructed with electrons in the final state was lower that the one expected from theory. Such inefficiency cannot be explained with the pre-selection applied to electrons ($p_T > 10$ GeV), as if that were the case, it should have also been present for muons. Therefore, an inefficiency in the electron identification and subsequent reconstruction is present. In order to try to recreate these missing top quarks from the information readily available from data, a new top category was created. This new brand of top quark was called StandAlone(SA), because they are reconstructed with those jets (namely Stan-

dalone Jets) for which there is no reconstructed prompt lepton, even though the MC simulation states a prompt lepton was generated at matrix element level. In such cases, the prompt lepton is still within the $\Delta R < 0.4$ cone but it is not actually reconstructed, leading to its energy deposits or track mixing with the ones of the jet components. The equivalent of the electron for the Standalone category was obtained by considering the charged electromagnetic energy fraction arising from the ECAL. Each component of the Standalone jet 4-momentum was multiplied by said fraction, which is interpreted as the electron energy coming from the prompt electron. Therefore, the new electron has a momentum:

$$p_{4,el} = E_{charged}\, p_{4,jet}; \tag{4.6}$$

while the equivalent of the b-jet has:

$$p_{4,b-jet} = (1 - E_{charged}) p_{4,jet}. \tag{4.7}$$

The same procedure was repeated in the case of muons, using the muon energy fraction, although in these cases the chance of not reconstructing a muon because its track overlaps with one from the jets is slim, as they have a much cleaner signature that includes tracks in the muon system.

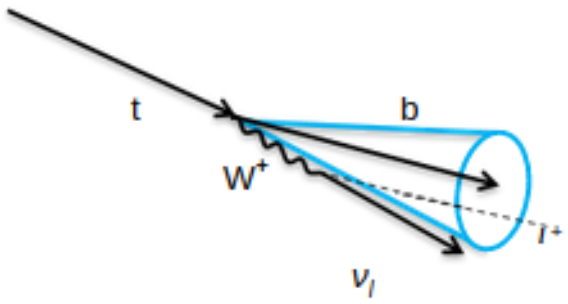
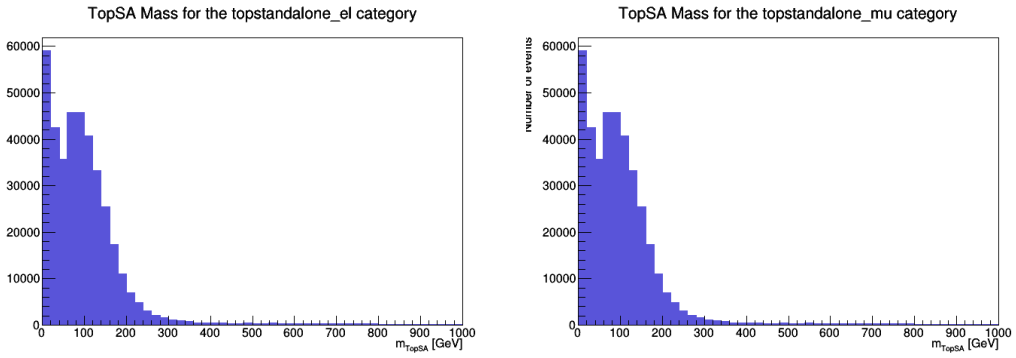


Figure 4.8: Visualization of the Top StandAlone category. Even though the ΔR cone opening is the same as its standard counterparts, the lepton is not reconstructed or it is considered as a jet/jet fraction.

Figure 4.9 shows the obtained masses for the TopSA category. Even in this case, a shift to the left in the distribution of mass values is present. This happens because the reconstruction algorithm has been thought for the resolved top quarks category, and it has not been optimized for the StandAlone (SA) category. Furthermore, the peak at low values of mass is caused by misidentification: standard jets have a very small mass, therefore, by identifying some of them as SA jets, they create a peak at low values of mass. However, for the purpose of the subsequent analysis, this level of reconstruction is satisfactory: the main focus is the reconstruction of this top category and the comparison of its performance with regards to the standard analysis.



(a) (b)

Figure 4.9: Plots depicting the mass distribution of the TopSA quark category reconstructed with the energy fractions in the calorimeter system. (a) is the case of the TopSA reconstructed with the charged electromagnetic fraction, while (b) is the shape obtained for the muonic energy fraction. The shift to the left of the distribution is due to identification problems.

4.4 Machine Learning Algorithms

Machine Learning (ML) is a branch of Artificial Intelligence; its main goal is the development of algorithms able to make predictions based on the knowledge derived from data. Developed during the course of the 20th century, ML algorithms allow the handling of huge quantities of information in a relatively simple way while at the same time finding relations between data. There are three main types of ML: supervised (SL), unsupervised (UL) and reinforcement learning (RL). Supervised Learning is the ML task of learning a function that maps a labelled input to an output based on example input-output pairs in which the labelling is already known. A SL process with discrete class labels is called classification, while, if the outcome is continuous, it is called regression. If the output is unkonwn, Unsupervised ML algorithms are used. This category of models study and generate functions in order to describe patterns found in data. However, this class of the algorithms has no way of determining whether or not these patterns make sense. This is something that requires human intervention in retrospect. Finally, RL is the category of ML in which learning happens without any human involvement, for these type of algorithms have a system (called agent) that learns a behavioural pattern of interaction with its environment by performing actions and then learn from the outcome. If said outcome satisfies a set of conditions imposed by the environment then the ML algorithm can be considered successful in its learning process. Needless to say, ML algorithms have a wide scope of possible applications. In the case of high energy Physics, their ubiquitous usage proves of utmost importance, since are currently allowing the development of new methodologies of data analysis [28]. The performance of the algorithm was trained and tested on data obtained from MC simulations of W' production at LHC.
The subsequent analysis will make use of a binary classifier, a ML algorithm able to group entries into two categories, namely true and false reconstructed

top quarks.

4.4.1 Boosted Decision Tree

A Boosted Decision Tree is a ML algorithm that performs classification. A decision tree is a sequence of requirements (cuts) applied in a specific order on a given variable dataset [29]. The cuts split the dataset into nodes, each corresponding to a certain number of samples classified as either signal or background. Other cuts may be applied in order to further split each node. Nodes in which either signal or background is dominant are classified as leafs, and no further selection is applied. Other cases in which nodes are classified as leafs is when there are too few remaining observations per node or when the number of nodes becomes too large. Each branch on a tree represents a sequence of cuts, as shown in Figure 4.10. In order to achieve the best split level in each node, the cuts can be tuned according to some metrics. Once the process of training the algorithm on an already classified dataset is done, the tree will be able to make predictions on unknown data. A possible optimization consists in maximizing for each node the gain of Gini index achieved after a splitting. The Gini index is defined as:

$$G = P(1 - P), \tag{4.8}$$

in which P is the fraction of signal samples in the node, called *purity*. This index is zero for nodes containing only signal or background. As an alternative to the Gini index, another important metric frequently used is called cross entropy:

$$E = -Plog(P) + (1 - P)log(1 - P). \tag{4.9}$$

The gain due to the splitting of a node A into two nodes B_1 and B_2 is defined as

$$\Delta I = I(A) - I(B_1) - I(B_2). \tag{4.10}$$

where I represents the chosen metric. Due to their easy interpretability, decision trees are very useful ML algorithms; however, they are fairly susceptible to overtraining, an occurrence in which the algorithms learns the fluctuations in the training dataset instead of identifying the pattern in the data. As a result of overtraining, the model does not generalize properly to new, unlabelled data. In order to improve the robustness of decision trees, they are ofter combined into ensembles. There are many methods of creating combinations of decision trees, and one of these is called *boosting*. This iterative procedure is performed as follows:

- training observations are reweighted using the previous iteration's classifier result;
- a new tree is built and optimized using the reweighted observations as a training sample;
- a score is given to each tree;

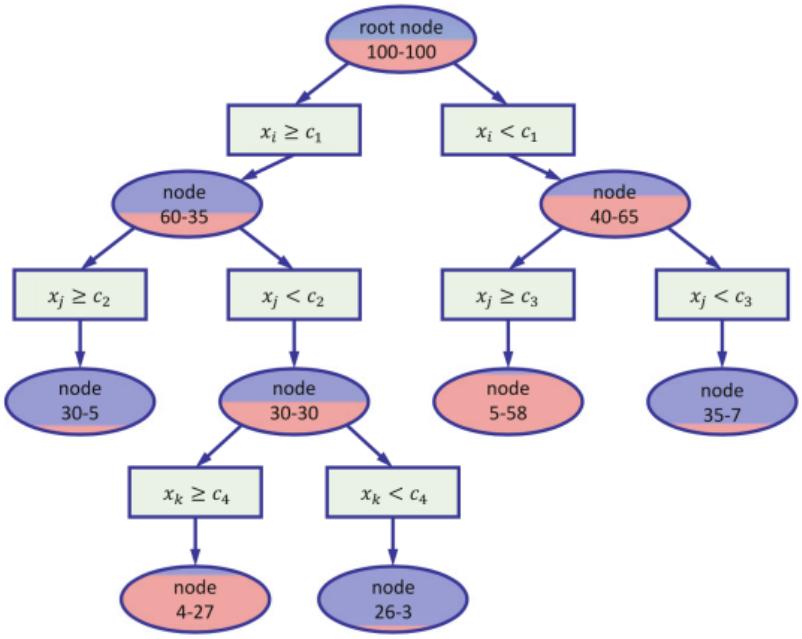


Figure 4.10: Chart depicting the branching process of a decision tree.

- the output of the final BDT classifier is the weighted average of every tree:

$$y = \sum_{i=1}^{N_{trees}} w_k y_k, \tag{4.11}$$

in which w_k are the weight (score) and y_k are the prediction of the $k-th$ tree.

4.5 Top Tagging with BDTs

For the resolved and merged top quark recostruction, a total of 12 models were developed using the XGBoost module in Python[30]. The dataset used for classification is composed of triplets of b-jets, leptons and MET in simulated collision events at LHC. The algorithm was trained on events from the MC simulations, and the performance was tested on an indipendent event set from the same MC simulation. The dataset obtained after the top quark reconstruction consists of a grid, in which each row represents a single reconstructed top quark. Each column contains different attributes of each top quark (i.e. its p_T, η, ϕ) and an additional column for the labels True and False, derived from the MC truth.

Before the actual training, the dataset is prepared by following two main steps:

- *Preselection:* a simple preselection is applied on leptons, in order to find the cuts that make the background (False Tops) comparable to the signal (True Tops):

Muon	Electron
$p_T > 10$GeV	$p_T > 10$GeV
isLoose=1	mvanoIsoL=1
MiniIso<5	MiniIso<4
|Dxy|<0.5	|Dxy|<0.05

Table 4.1: Pre-selection applied on leptons.

The meaning of some of these features will be explained later in this Section;

- *Binning:* each category is split into 3 bins based on the p_T values of the reconstructed top quark candidate :
  - **high p_T:** $p_T > 1000GeV$;
  - **medium p_T:** $p_T > 500GeV$ and $p_T < 1000GeV$;
  - **low p_T:** $p_T < 500GeV$;

Top quarks reconstructed with electrons and muons were analysed separately. After this procedure, 12 datasets are obtained, 4 for each top p_T bin, two of which belong to the Merged category (on for each lepton), and the other two to the Resolved one.
Tables 4.2 and 4.3 list all the variables used for the training. The label "Top" represents a top quark reconstructed with the sum of the lepton and b-jet 4-momenta, while "Top$_\nu$" also includes the missing transverse energy. The label *ub* (unboosted) is used for the variables calculated in the top quark candidate centre of mass frame. Beside the standard kinematic variables, some quantities of particular interest for this analysis are:

- **Iso04, Iso03:** as already described in Section 4.1, they are different degrees of isolation, respectively PF relative isolation with $\Delta R = 0.4$, $\Delta R = 0.3$;
- **mvanoIsoL:** an identification criterion on electrons based on a BDT analysis developed by the electron CMS working group;
- $\theta_{l,b}$**:** angle between the lepton momentum direction and the b-jet momentum direction;
- $p_{T,rel}$**:** quantity used to identify the p_T component of the lepton perpendicular to the b-jet, it is also used to discriminate between prompt and non-prompt leptons:
$$p_{T,rel} = \frac{|\vec{p_l} \times \vec{p}_{jet}|}{\vec{p_l}}; \tag{4.12}$$
- **Dxy & Dz:** impact parameters in the transverse plane and in the longitudinal plane respectively;

- **DeepFlavB:** bottom flavour tag discriminator, defined in the context of the DeepFlavour algorithm employed for b-tagging;
- **Over_Jet_Pt:** defined as p_l/p_{jet};
- **isLoose:** indicates that the lepton must comply to some criteria decided by the CMS Collaboration.

Jet	Muon	Jet_{ub}	Muon_{ub}	Top	Top_{ν}
M	p_T	p_T	M	M	M
p_T	Dxy	E	p_T	E	p_T
η	Dz		η	mT	E
ϕ	MiniIso		ϕ	dR	
DeepFlavB	Iso04			$cos(\theta_{l,b})$	
	Over_Jet_pt			$p_{T,rel}$	

Table 4.2: Variables employed by the BDT top tagger for candidates reconstructed with muons.

Jet	Electron	Jet_{ub}	Electron_{ub}	Top	Top_{ν}
M	pt	M	M	M	M
p_T	Dxy	p_T	p_T	E	p_T
η	Dz	E	η	mT	E
ϕ	MiniIso		ϕ	dR	
DeepFlavB	mvanoIsoL			$cos(\theta_{l,b})$	
	Iso03			$p_{T,rel}$	
	Over_Jet_pt				

Table 4.3: Variables employed by the BDT top tagger for candidates reconstructed with electrons.

Each one of the 12 datasets has been randomly split into a training and test set; as the names imply, the former is used for the training process, while the latter (also called validation set) is meant to be used for the evaluation of the performance of the algorithm. When the performance of the algorithm is the same for test and train, it means that the model is extrapolating from one dataset to the other, i.e. is not mistakenly interpreting statistical fluctuations of the training set as features of the signal. At first, 12 BDTs (one for each dataset) were trained with an identical set of input variables and hyperparameters, in order to minimize the correlation between the output of the ML algorithm and the transverse momentum of the top candidate. Then, after identifying the variables mostly used by the BDTs, the Receiver Operating Characteristic (ROC) curve was obtained. The ROC curve shows the performance of a classification model by plotting two parameters, the *True Positive*

Rate:

$$TPR = \frac{TP}{TP + FN} \tag{4.13}$$

and the *False Positive Rate*:

$$FPR = \frac{FP}{FP + TN} \tag{4.14}$$

in which TP, FN, FP, and TN respectively stand for True Positive, False Negative, False Positive, and True Negative. Therefore, in the case of the analysis in question, $T = TP + FN$, $N = FP + TN$ are respectively the total number of signal events and background samples in the training set, while TP and FP represent the number of signal and background samples selected by the ML algorithm among the instances of the training set. As seen in Figure 4.11, ROC curves plot TPR vs FPR at different classification thresholds; an important parameter is the Area under the ROC curve (AUC), a parameter accounting for the entire area underneath the ROC curve from the point (0,0) to the point (1,1). A model whose AUC is 0 has efficiency 0 on the signal, meaning each signal instance is rejected regardless of the requirement on the score; one whose AUC is 1 perfectly rejects background entries for every requirement applied to the score.

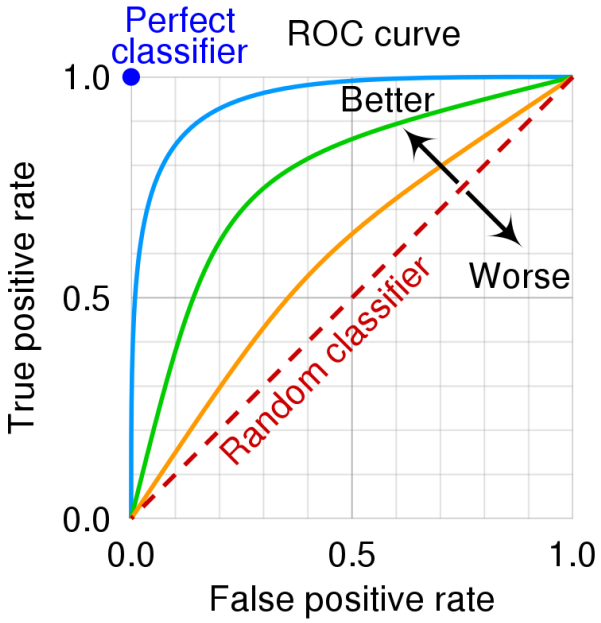
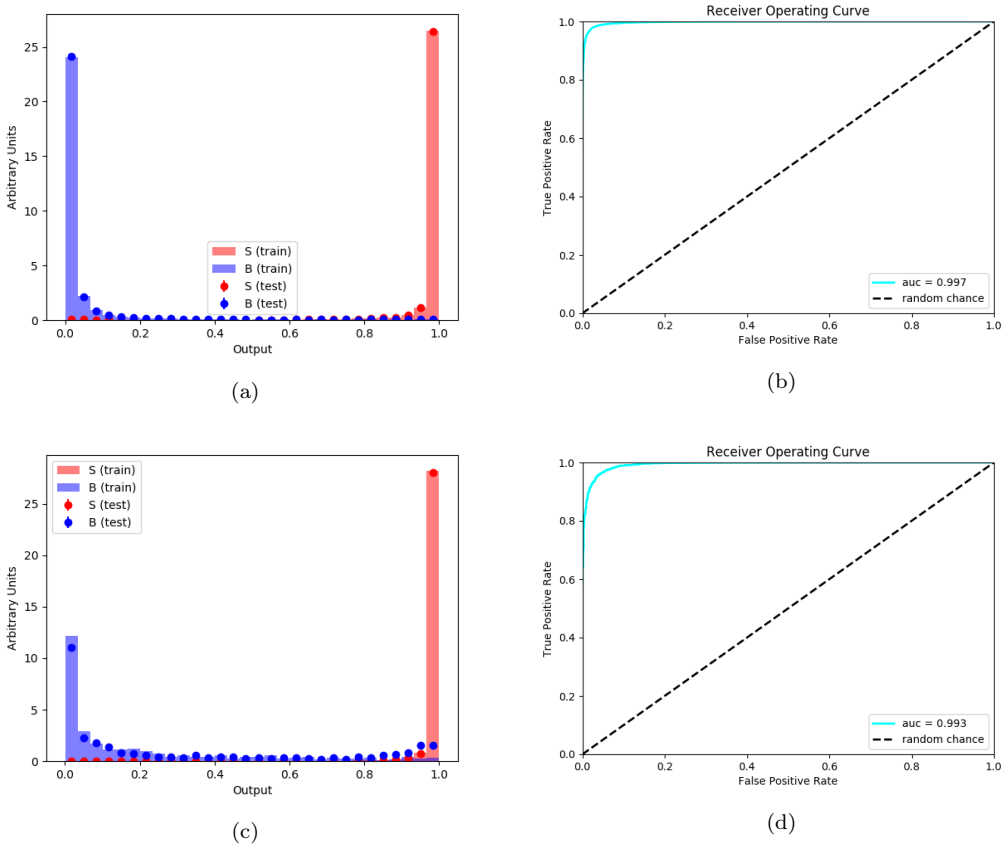


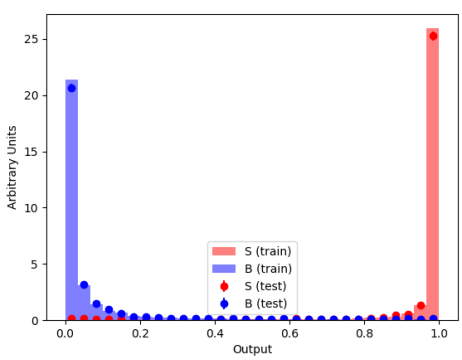
Figure 4.11: A few examples of ROC curves.

The results of the training of the BDTs and the related ROC curves are shown below. Figure 4.12a shows the output of the BDT for the resolved signal with muons in the final state for the high p_T configuration. There are no visible signs of overtraining and the signal and background are correctly identified and separated. The relative ROC curve is shown in Figure 4.12b, and the AUC is 0.997. Below, all the results of the 12 trainings are shown.

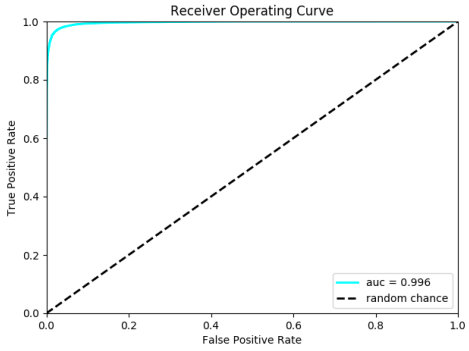


(a) (b) (c) (d)

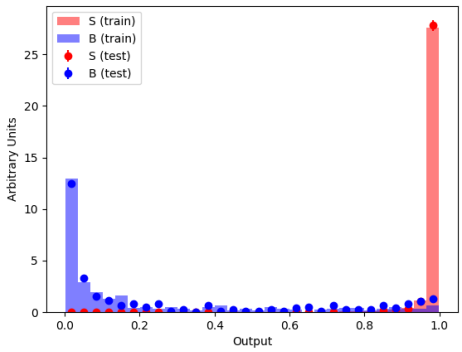
Figure 4.12: (a)BDT output for the resolved category reconstructed with a muon in the final state, in the high p_T range; the true top quarks (red) are correctly identified by the algorithm, which separates them from the false tops (blue). No sign of overtraining is present, for the performances of the model in the training (histogram bars) and the evaluation (dots) sets are similar. (b) ROC curve for the high p_T muon resolved category. As expected from the output of the BDT, the model is working properly: the AUC is very close to 1, meaning that the model is perfectly rejecting background entries for every requirement applied to the score and the distance from the random chance is maximized. (c) BDT output and ROC curve for the high p_T muon merged category. Even though a slight residual overtraining is visible in the corresponding BDT score, the AUC is pretty close to being unitary (d).



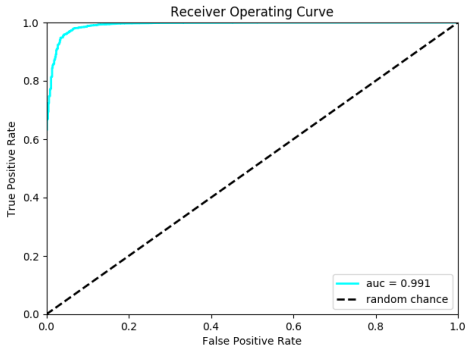
(a)



(b)



(c)



(d)

Figure 4.13: (a) BDT output for the resolved category reconstructed with an electron in the final state, in the high p_T range. Classification is correct for both the training and the validation set, no signs of overtraining are visible. (b) ROC curve for the high p_T electron resolved category. As expected by the degree of separation of the signal and background classes in the corresponding BDT score, AUC is very close to 1. (c) BDT output for the merged category reconstructed with an electron in the final state, in the high p_T range. As expected, since the statistic for this category is pretty low, this model suffers from evident overtraining and misclassification is visible in the signal region. (d) ROC curve for the high p_T electron merged category. AUC is slightly lower with respect to the categories seen up to this point, as it is expected from the output of the BDT score.

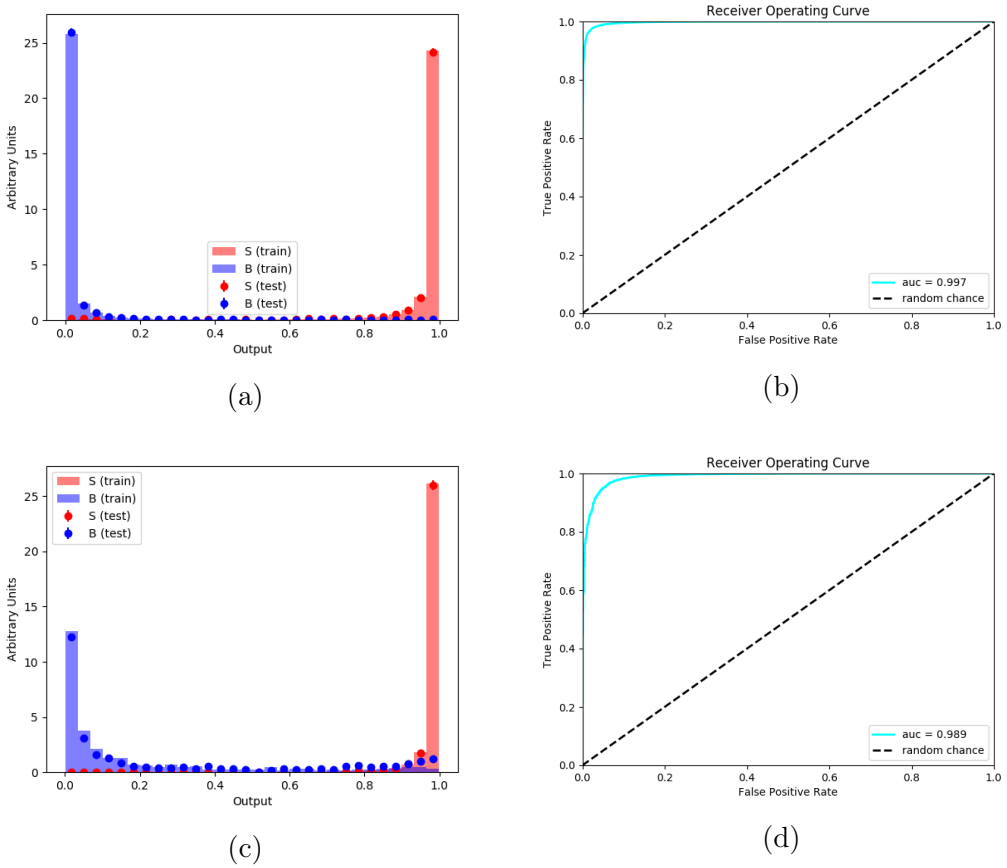


Figure 4.14: (a)BDT output for the resolved category reconstructed with a muon in the final state, in the medium p_T range. (b) ROC curve for the medium p_T muon resolved category. (c) BDT output for the medium p_T mu merged category. Again, this model suffers from overtraining and misclassification, with a clear although almost negligible background tail in the signal region. (d) BDT output and ROC curve for the high p_T muon merged category.

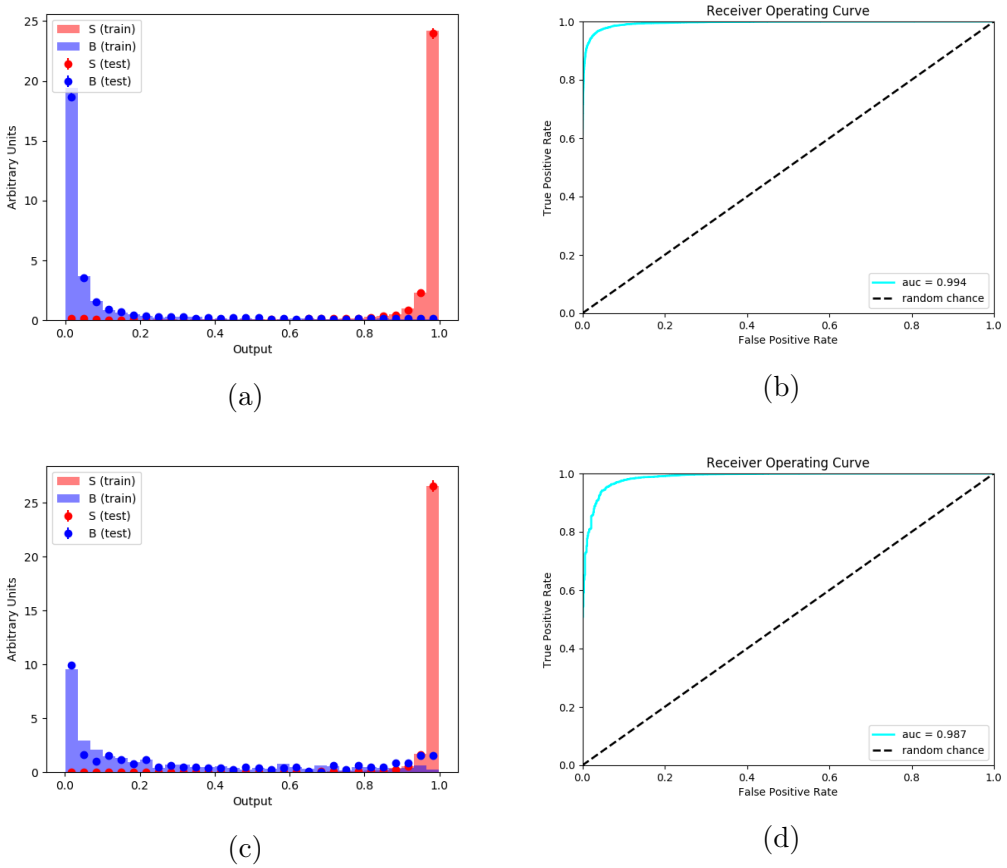


Figure 4.15: (a) BDT output for the resolved category reconstructed with an electron in the final state, in the medium p_T range. (b) ROC curve for the medium p_T electron resolved category. (c) BDT output for the merged category reconstructed with an electron in the final state, in the medium p_T range. Since the number of entries for this category is low, overtraining and misidentification are unavoidable. (d) ROC curve for the medium p_T electron merged category. Needless to say, AUC has a lower value with regards to its muonic counterpart.

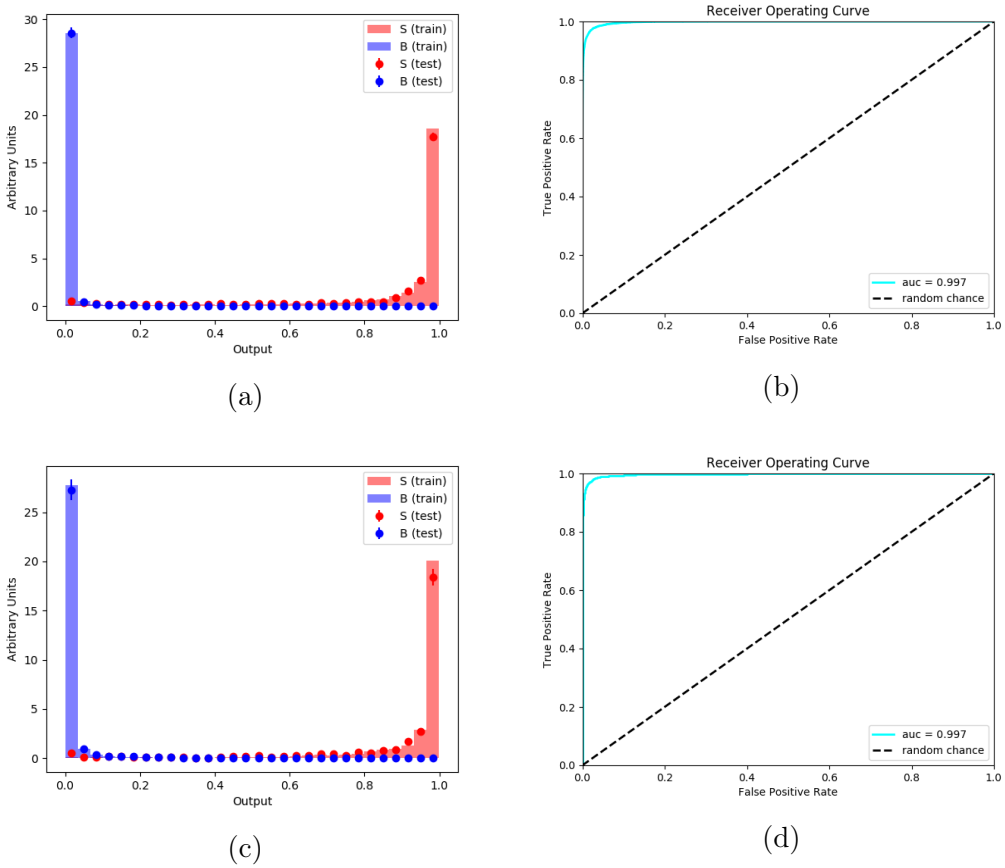


Figure 4.16: (a) BDT output for the resolved category reconstructed with a muon in the final state, in the low p_T range. (b) ROC curve for the low p_T muon resolved category. (c) BDT output for the low p_T mu merged category. (d) BDT output and ROC curve for the high p_T muon merged category.

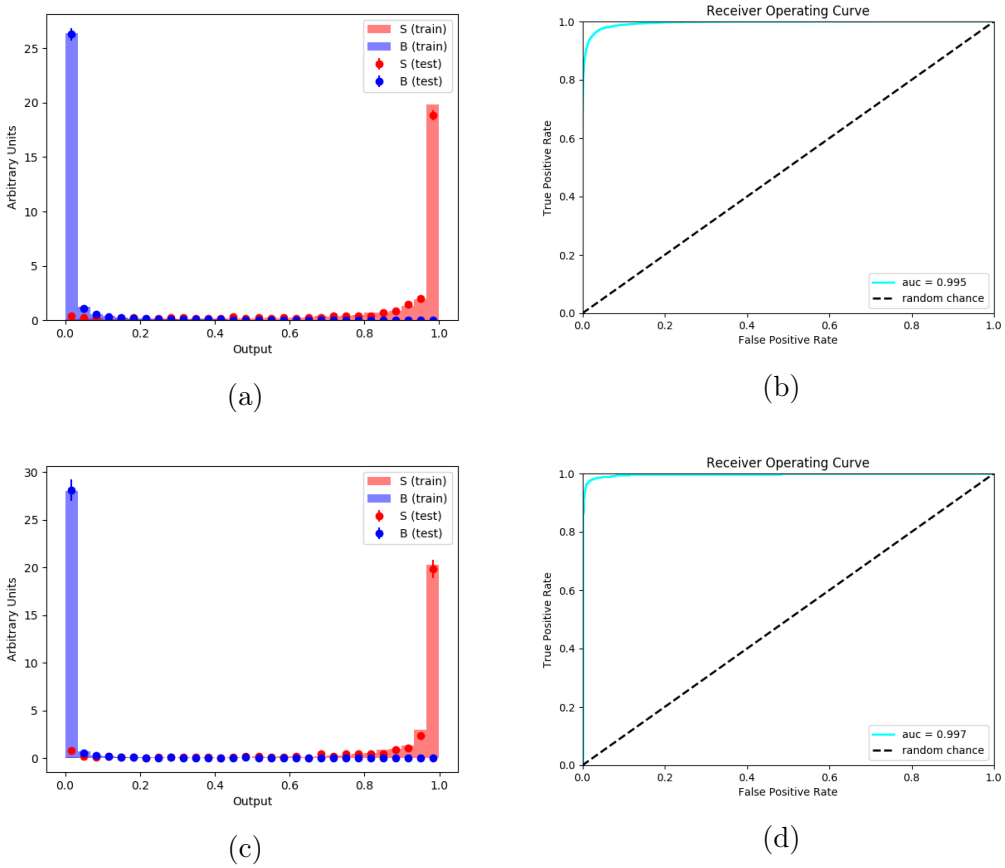


Figure 4.17: (a) BDT output for the resolved category reconstructed with an electron in the final state, in the low p_T range. (b) ROC curve for the low p_T electron resolved category. (c) BDT output for the merged category reconstructed with an electron in the final state, in the low p_T range. (d) ROC curve for the low p_T electron merged category.

4.6 TopSA Tagging with the BDTs

For the Top Standalone category, the main variables used in the training are inherited from the jets collection. The preparation of the dataset is in every way similar to the one used in the standard analysis, except for the selection on leptons, which of course is not feasible in this case. However, the requirement on the flavour of the selected jet stands, for this TopSA must have a bottom quark. The variables used in the training are listed in Table 4.4; the labelling is similar to the one used in the standard analysis. "TopSA$_\nu$" is the only real category of top candidate considered in this training, for its counterpart without the missing transverse energy would have been simply a jet, therefore inappropriate for this type of analysis. The momentum components of the new object called "LeptonSA" have been calculated in the Standalone Top centre of mass frame, therefore the pedix. The same was done for the equivalent of the jet in this new analysis.

TopSA$_\nu$	LeptonSA$_{ub}$	JetSA$_{ub}$	Class Variables
p_T	p_T	p_T	area
η	ϕ	e	bRegRes
M	e		btagCMVA
	M		btagDeepFlavB
			partonFlavour

Table 4.4: Variables employed by the BDT top tagger for candidates reconstructed with electrons.

- **area:** area of the cone opening in the $(\eta - \phi)$ plane in which tracks are detected, mainly used in the context of jet energy corrections;
- **bRegRes:** p_T resolution corrected with b-jet regression. The b-jet energy resolution is worse with respect to the light quark/gluon induced jets since in the 35% of the cases a neutrino is involved in the B hadron decay.The regression technique is a multidimensional calibration targeting the jet transverse momentum at generator level, exploiting several jet and event properties;
- **btagCMVA:** another b-tagger discriminator. The Combined Multi Variate Algorithm tagger combines the discriminator values of various taggers to improve the identification of b-jets:
- **btagDeepFlavB:** DeepFlavour algorithm tagger, already described in Section 4.1;
- **partonFlavour:** flavour from parton matching;

BDTs and ROC curves for the 6 models for TopSA are shown and commented below.

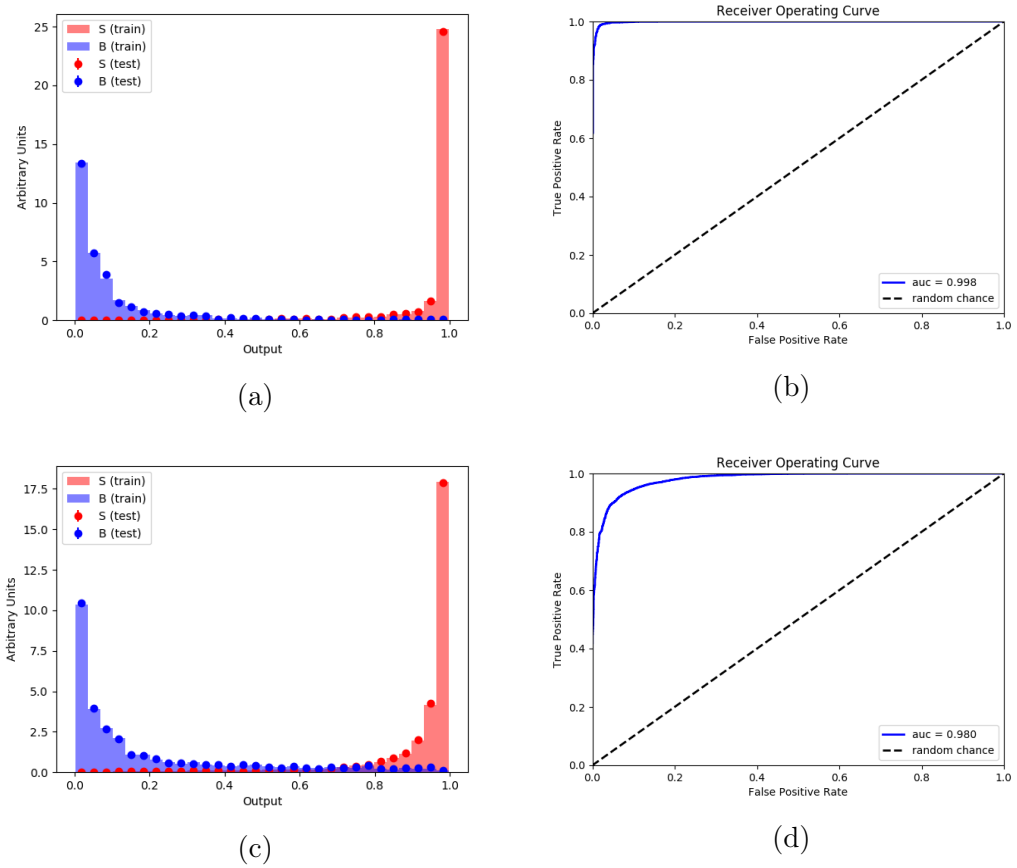


Figure 4.18: (a) BDT output for the StandAlone category reconstructed with the charged calorimeter energy fraction in the high p_T range. (b) ROC curve for the standalone category reconstructed with the charged calorimeter energy fraction in the high p_T range. (c) BDT output for the standalone category reconstructed with the muonic calorimeter energy fraction in the high p_T range. (d) ROC curve for the standalone category reconstructed with the muonic calorimeter energy fraction in the high p_T range.

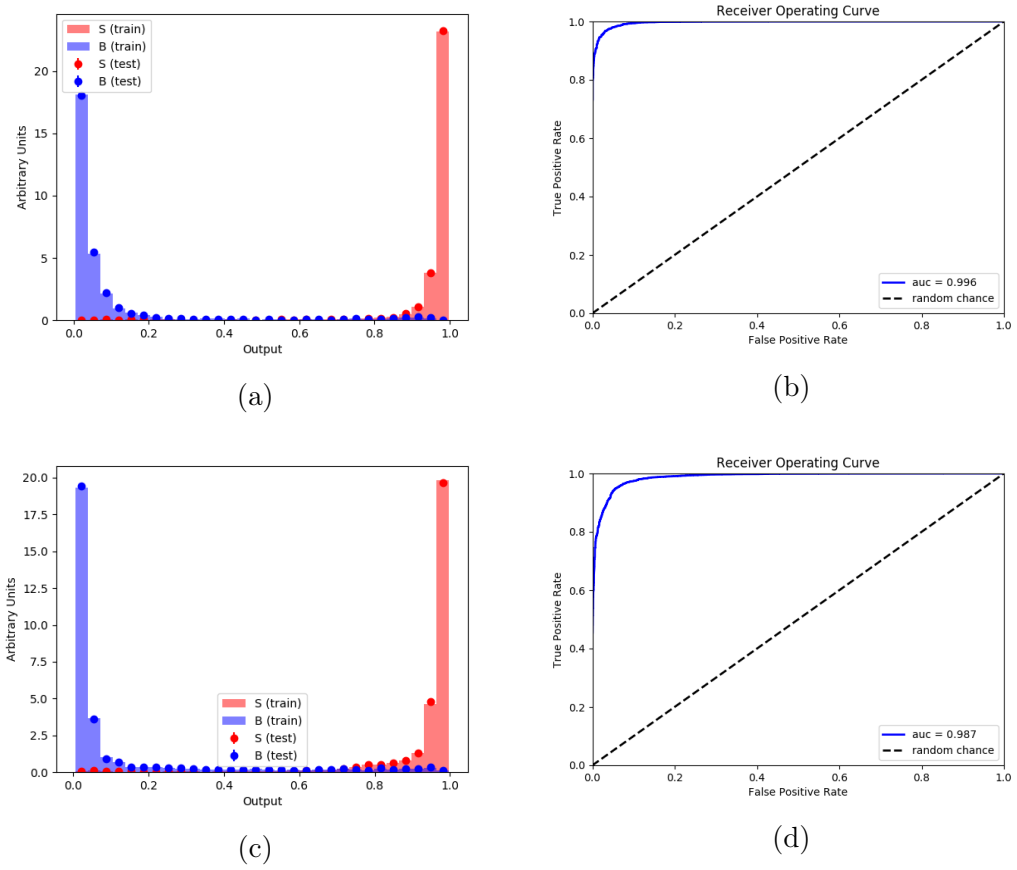


Figure 4.19: (a) BDT output for the StandAlone category reconstructed with the charged calorimeter energy fraction in the medium p_T range. (b) ROC curve for the standalone category reconstructed with the charged calorimeter energy fraction in the medium p_T range. (c) BDT output for the standalone category reconstructed with the muonic calorimeter energy fraction in the medium p_T range. (d) ROC curve for the standalone category reconstructed with the muonic calorimeter energy fraction in the medium p_T range.

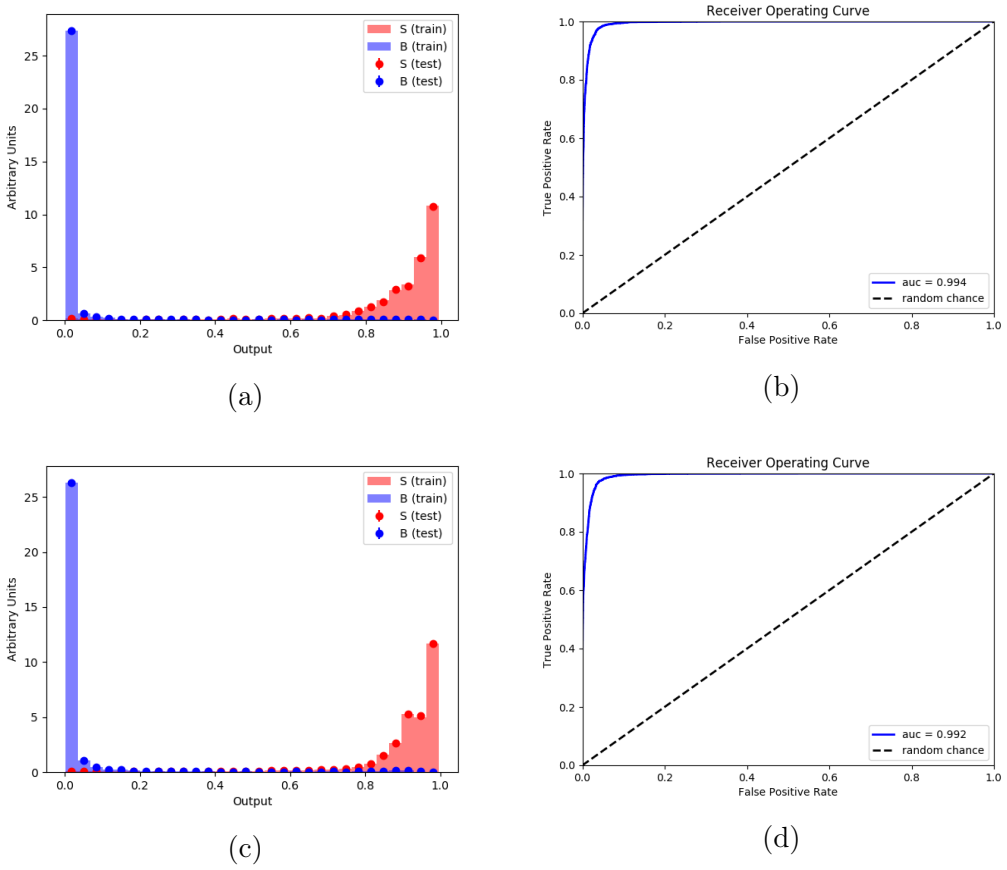


Figure 4.20: (a) BDT output for the StandAlone category reconstructed with the charged calorimeter energy fraction in the low p_T range. (b) ROC curve for the standalone category reconstructed with the charged calorimeter energy fraction in the low p_T range. (c) BDT output for the standalone category reconstructed with the muonic calorimeter energy fraction in the low p_T range. (d) ROC curve for the standalone category reconstructed with the muonic calorimeter energy fraction in the low p_T range.

Chapter 5

Application of ML algorithm and W' analysis

In this Chapter, the analysis strategy for the search of the W' is presented. As previously stated, this work focuses on the processes where the W' decays into a bottom and a top quarks, with the latter later cascading leptonically into a jet-lepton-neutrino triplet. Figure 5.1 shows the complete decay chain. The bottom quark undergoes hadronization, resulting in a b-jet. The reconstruction of the top quark candidate through its experimental signature is performed using the algorithm described in the previous Chapter, Section 4.2. For each lepton, 3 top quark categories have been identified: the Merged, the Resolved, and the StandAlone category. While the Merged and Resolved categories are not mutually exclusive, the SA is exclusive with the Merged one, but not with the Resolved one: in fact, SA top quarks are reconstructed from jets that do not overlap with reconstructed leptons. After selecting the best top quark candidate for each category, the next step in the reconstruction of the W' is the identification of the other object involved in its decay, namely the b-jet. Jets derived from bottom quarks are selected ensuring that those b-jets involved in the top quark reconstruction are singled out and, therefore, excluded from the reconstruction of the W'. For each top quark category the highest BDT score top quark candidate and the highest-p_T eligible b-jet are selected. Further selection criteria can be defined on top of this reconstruction to reduce specific background contributions, as it will be detailed later on in this Chapter. The 4-momentum of the W' boson is obtained by adding the 4-momenta of the reconstructed top quark and of the b-jet. Before proceeding in the description of the analysis, it is to be noted that the nature of the top quark and its behaviour are deeply linked to the value of the W' mass hypothesis. From the existing searches at the LHC [20] [21] [19], it is seen that the W' production has been excluded for the most common models in a mass range below 2-3 TeV. Therefore, the presented analysis is limited to a mass range above 2 TeV. The module of the top quark momentum is of order the magnitude of $m_{W'}/2$. The decay products of the top quark are produced in a cone roughly with angular opening $R \sim 2m_t/p_T$, in which m_t is the top quark mass, and p_T its transverse momentum. As the p_T gradually increases with the

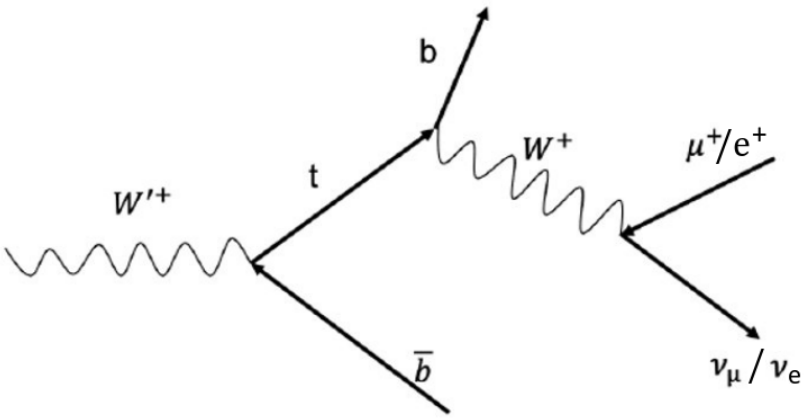


Figure 5.1: Feynman diagram for the $W' \longrightarrow t\bar{b} \longrightarrow l^{+}\nu_{l}b\bar{b}$ process. After selecting the best top quark candidate, the next step is the reconstruction of the W' boson, performed by adding the 4-momenta of the selected top quark candidate and b-jet.

W' mass hypothesis, it is expected that the prevailing top quark configuration will be the Merged one, for it is defined as the case in which the lepton and the b-jet are at an angular distance $\Delta R < 0.4$. The analyzed masses are in the range 2-6 TeV, therefore, for higher values of $m_{W'}$, one expects that the prevailing top quark configurations will be the Merged or the SA ones. In order to simplify the description of the subsequent analysis, the following notation, referring to the previously discussed top quark reconstruction categories, will be adopted:

- **mu_merged:** the top quark is reconstructed with a muon in its final state, and the angular distance between the lepton and the b-jet is $\Delta R < 0.4$;
- **mu_resolved:** the top quark is reconstructed with a muon in its final state, and the angular distance between the lepton and the b-jet is $0.4 < \Delta R < 2$;
- **el_merged:** the top quark is reconstructed with an electron in its final state, and the angular distance between the lepton and the b-jet is $\Delta R < 0.4$;
- **mu_resolved:** the top quark is reconstructed with an electron in its final state, and the angular distance between the lepton and the b-jet is $0.4 < \Delta R < 2$;
- **mu_topSA:** the top quark is reconstructed from a jet with an angular opening of the cone equal to $\Delta R = 0.4$, in which the muon is reconstructed from the muon energy fraction;
- **el_topSA:** the top quark is reconstructed from a jet with an angular opening of the cone equal to $\Delta R = 0.4$ in which the electron is reconstructed from the electromagnetic energy fraction in the ECAl.

Given the high-energetic nature of top quarks it is natural to assume that the high and medium $p_{T,top}$ range will return the best W' reconstructions. Furthermore, since muons are naturally better separated from jets in the CMS reconstruction chain and in the CMS subdetector system, it is expected that the mu_topSA category will not be of particular significance. The most abundant background processes that can mimic a W' decay are $t\bar{t}$, W+jets, and QCD, described in the following section, plus minor contributions from other backgrounds that are considered as negligible for the sake of evaluating the performance of the analysis at this stage.

5.1 Background description

The analysis was performed on MC simulated signals samples, considering right-handed W' production, with 3 different sets of masses: 2, 4, and 6 TeV, with a decay width of 1% of the respective mass.

- $t\bar{t}$**:** a top quark and anti-quark couple is produced, as Figure 5.2 shows. If at least one of the top quarks decays leptonically, it can exactly mimic the final state of the W'-originated top quark. The other top quark of the pair generates a b-jet that could be selected as the jet deriving from the W', creating a fake signal event;

- *W+Jets:* the associated production of the W boson and two jets (Figure 5.3) could recreate a final state for which the W boson, decaying into a lepton-neutrino couple, could be misidentified as a signal if accidentally paired with a bottom quark-originated jet, while the other b-jet in the event reproduces the other leg of the W' decay;

- *QCD:* a quark pair is generated from gluon-gluon interaction, and the combination of one of the quark jets with the lepton-neutrino pair in the other jet coming from the non-prompt hadronic decay chains of, for instance, b-hadrons, could amount to the top mass, therefore constituting background for the event in question;

Table 5.1 lists the samples used for this analysis, their respective cross sections, and the number of expected entries for each sample, obtained by multiplying each cross section by the nominal luminosity, which, for the case considered is the 2016 luminosity $L = 35.9 fb^{-1}$. The $t\bar{t}$, W+Jets, and QCD samples are divided into subsamples in order to increase the available MC statistics. The bins are made in $M_{t\bar{t}}$ for $t\bar{t}$, and in HT for W+Jets and QCD. $M_{t\bar{t}}$ is the invariant mass of the $t\bar{t}$ quark pair, and in HT is the hadronic transverse energy of the jets. The W+Jet cross sections are multiplied by the scale factor obtained from the ratio of the next-to leading order (NLO) over leading order (LO) cross section.

Samples		σ (pb)	$N = \sigma \cdot L$
Signals	W'(2 TeV)	1.397	$55 \cdot 10^3$
	W'(4 TeV)	.01736	685
	W'(6 TeV)	.0009153	36
Total number of signal events			55902
$M_{t\bar{t}}$	700 to 1000 GeV	80.5	$34 \cdot 10^5$
	>1000 GeV	21.3	$84 \cdot 10^4$
Total number of $t\bar{t}$ events			$42 \cdot 10^5$
W+Jets; HT ranges	70 to 100 GeV	1353.0×1.21	$64 \cdot 10^6$
	100 to 200 GeV	1345×1.21	$64 \cdot 10^6$
	200 to 400 GeV	359.7×1.21	$17 \cdot 10^6$
	400 to 600 GeV	48.91×1.21	$23 \cdot 10^5$
	600 to 800 GeV	12.05×1.21	$57 \cdot 10^4$
	800 to 1200 GeV	5.501×1.21	$26 \cdot 10^4$
	1200 to 2500 GeV	1.329×1.21	$63 \cdot 10^3$
	>2500 GeV	0.03216×1.21	1580
Total number of W+Jets events			$14 \cdot 10^7$
QCD; HT ranges	300 to 500 GeV	347700	$13 \cdot 10^9$
	500 to 700 GeV	32100	$12 \cdot 10^8$
	700 to 1000 GeV	6831	$26 \cdot 10^7$
	1000 to 1500 GeV	1207	$47 \cdot 10^6$
	1500 to 2000 GeV	119.9	$47 \cdot 10^5$
	> 2000 GeV	25.24	$9 \cdot 10^5$
Total number for QCD events			$14 \cdot 10^9$

Table 5.1: Cross sections and expected number of events per each sample used in the analysis. Each sample is split accordingly to the nature of the phenomenon in question; for each subsample, the cross sections are listed. The product of the 2016 nominal luminosity and the cross section are reported for each sample, for this quantity is used to define selection efficiencies for both signal and background categories in later steps of the analysis.

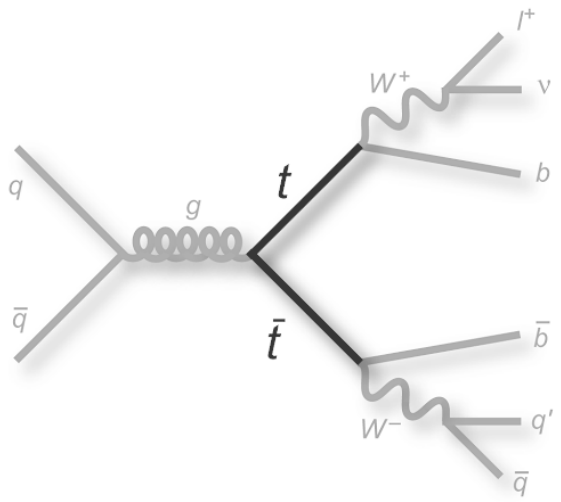


Figure 5.2: Feynman diagram of the $t\bar{t}$ quark pair creation.

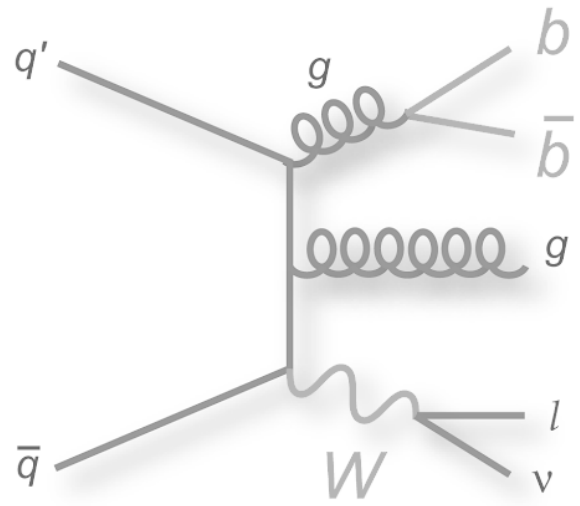


Figure 5.3: Feynman diagram of the W+Jets background events.

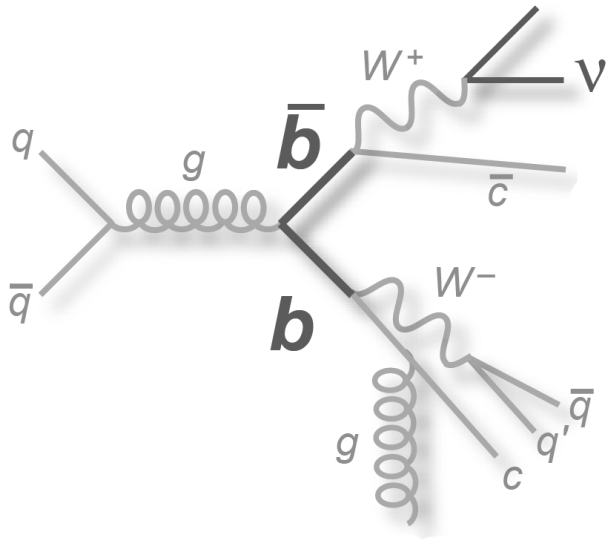


Figure 5.4: Feynman diagram of the QCD background events.

5.2 W' reconstruction and baseline selection

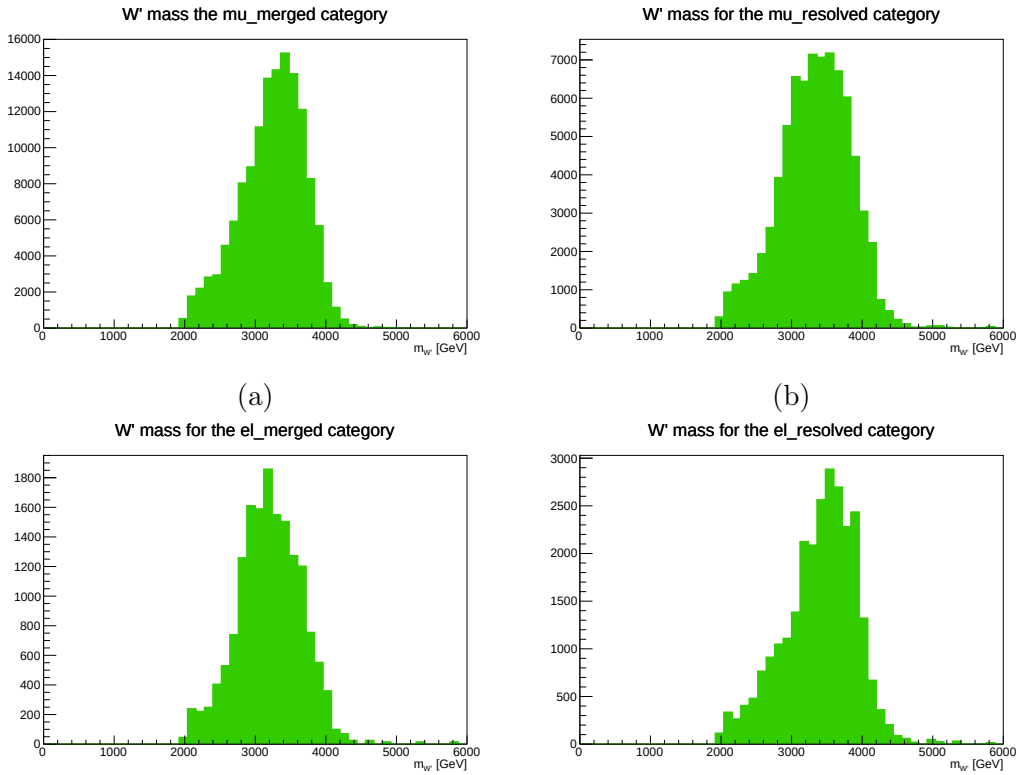
As stated in Figure 5.1, a selection of the most suitable top quark and b-jet candidate is necessary in order to avoid combinatorics background. For each category described in Chapter 4, the best top quark candidate has been selected by requiring a 90% background rejection. This selection requirement was applied to the top tagging Score, so that 90% of fake top quarks are rejected, according to the studies performed in Chapter 4. The b-jet stemming from the hadronization of the b-quark originated from the W' decay vertex was selected among the jet candidates that pass the following conditions:

- $p_T > 30$ GeV;
- value of the b-tag discriminator score (named *btagDeepFlavB*) > 0.4 which corresponds to a misidentification probability of 1%, namely the so-called Medium working point of the algorithm.

In order to reject those jets not compatible with the high p_T spectrum of the jet from a direct W' decay, an additional requirement was applied to the already flavour-tagged b-jets, namely that $p_T > 100$ GeV. Among these jets, the ones with the highest p_T are selected. Of course, these jets are also vetoed from being the ones involved in the top quark reconstruction. After this selection and the construction of the 4-momenta of these two objects, the 4-momentum of the W' boson is thusly obtained:

$$p_{4,W'} = p_{4,top} + p_{4,bjet} \tag{5.1}$$

Figures 5.5, 5.6, and 5.7 show the mass distribution of a W' boson in the 4 TeV mass hypothesis, reconstructed respectively for the high, medium, and low p_T categories. The peaks are clearly shifted to the left. This could be due to the fact that the selection of the b-jet coming from the W' decay has not been optimized, and the top quark 4-momentum reconstruction algorithm was originally tuned on top quarks in the Resolved regime. Additionally, given the kinematics of the top quark, events in the low-p_T regime for signal processes are extremely unlikely, therefore, they might stem from cases in the very tail of the kinematic distributions. For these reasons, the best results are obtained for those categories corresponding to higher values of the reconstructed top quark transverse momentum, while the low p_T categories clearly suffer from misidentification in both the top reconstruction and the b-jet selection.



(a) (b) (c) (d)

Figure 5.5: W' masses reconstructed considering the 4 TeV mass hypothesis, using the (a) mu_merged (b) mu_resolved (c) el_merged (d) el_resolved top quark categories in the high p_T range.

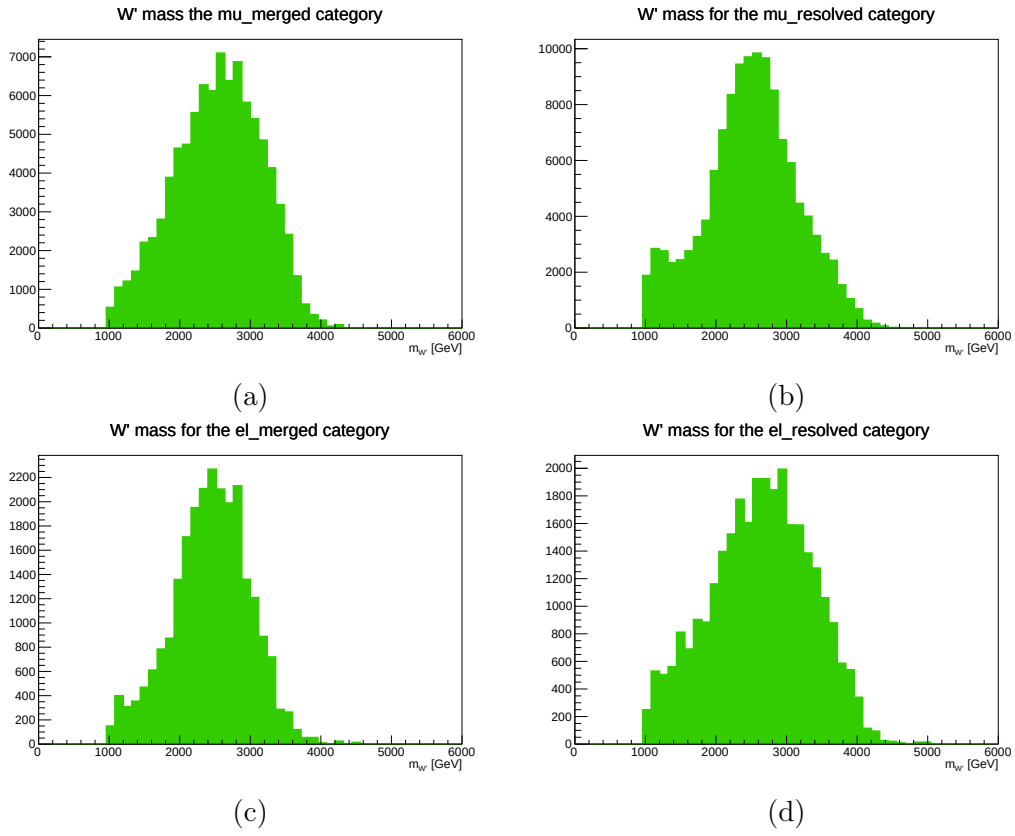
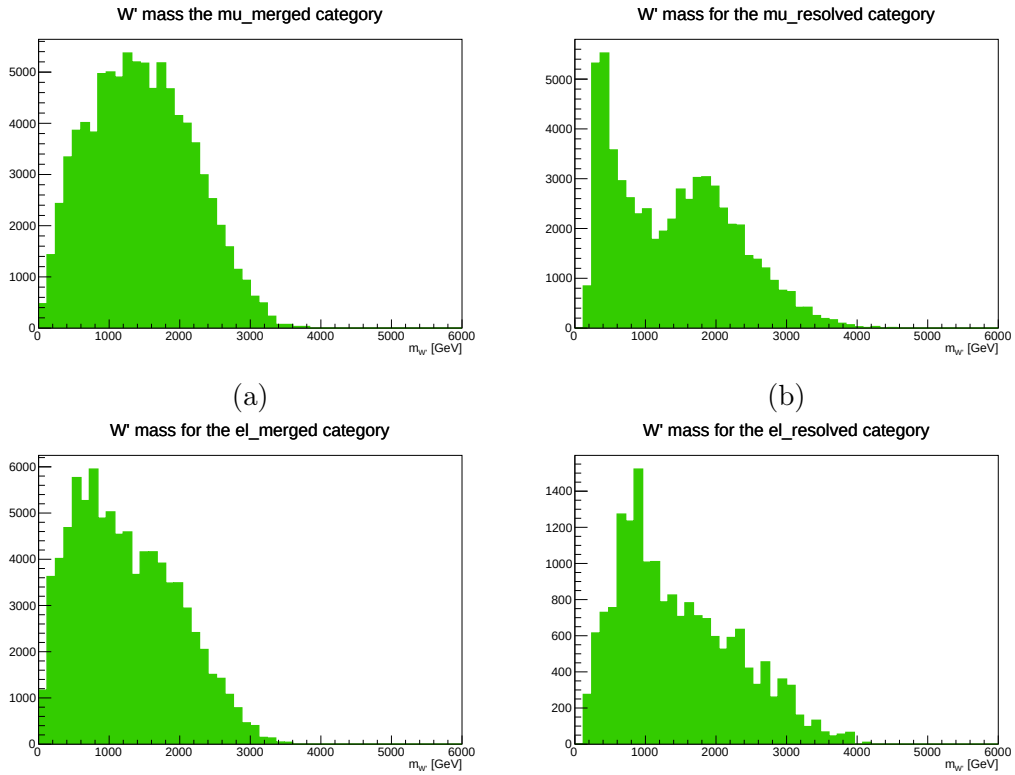
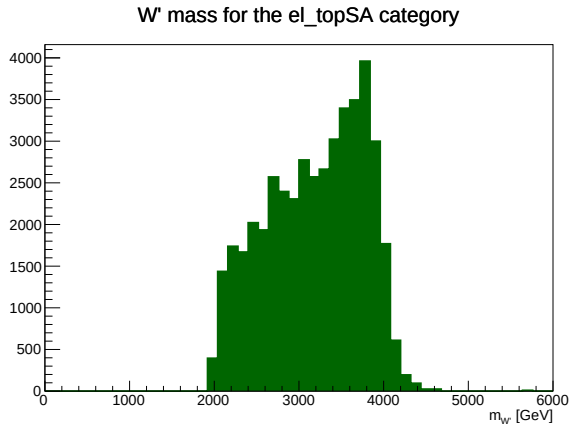


Figure 5.6: W' masses reconstructed considering the 4 TeV mass hypothesis, using the (a) mu_merged (b) mu_resolved (c) el_merged (d) el_resolved top quark categories in the medium p_T range.

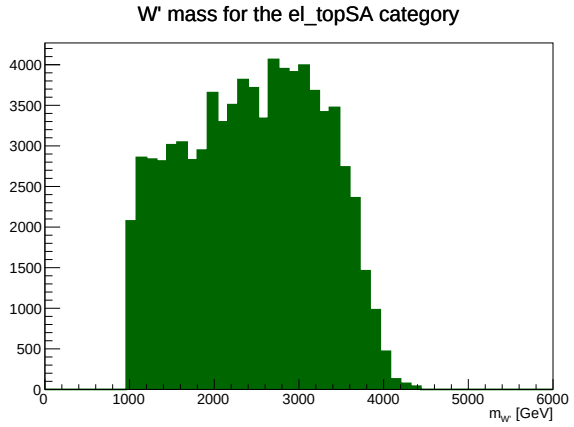


(a) (b) (c) (d)

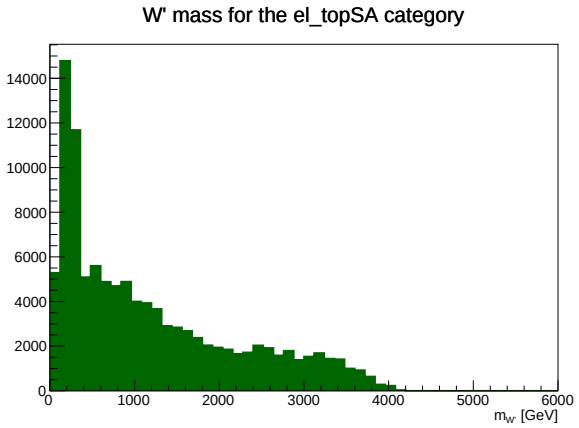
Figure 5.7: W' masses reconstructed considering the 4 TeV mass hypothesis, using the (a) mu_merged (b) mu_resolved (c) el_merged (d) el_resolved top quark categories in the low p_T range.



(a)



(b)



(c)

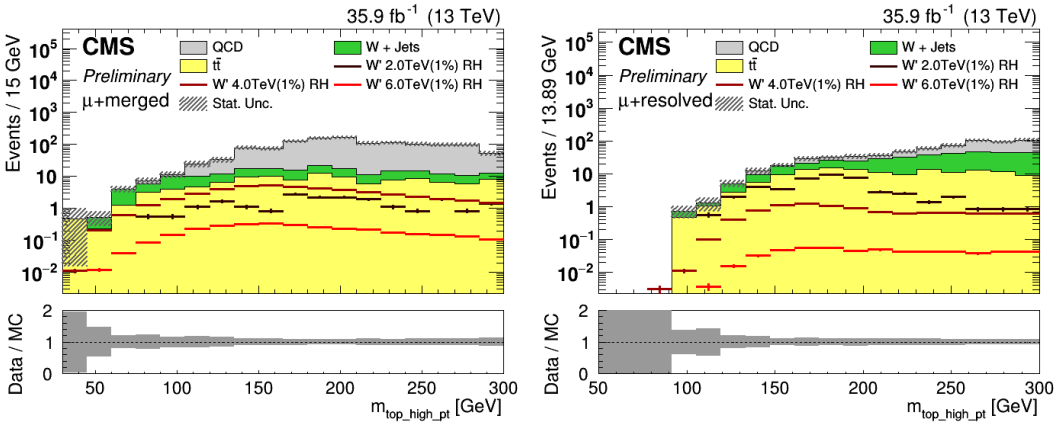
Figure 5.8: W' masses reconstructed considering the 4 TeV mass hypothesis, reconstructed with the el_topSA top quark category for the (a) high, (b) medium, and (c) low p_T range. In the first two cases the reconstruction, although imperfect, is perfunctory to the reconstruction of objects with a mass > 2 TeV, the low p_T case is clearly suffering from misidentification issues.

5.3 Analysis strategy

After the event reconstruction and categorization, additional kinematic requirements are imposed on the objects used for the W' reconstruction to reduce the background contamination. Characteristic variables for this process are the reconstructed top quark mass and the transverse momentum of the b-jet deriving from the W' decay. Figure 5.9 and Figure 5.10 show the distribution of the reconstructed top quark mass for the mu_merged and mu_resolved top quark categories in the high and medium p_T ranges. Figure 5.11 and Figure 5.16 show the b-jet p_T for each category in the high and medium p_T range. By requiring that the top quarks involved in the W' reconstruction must have mass $m_t > 50$ GeV (or alternatively $m_t > 100$ GeV), a significant part of the QCD and $t\bar{t}$ backgrounds are removed, while the requirement $m_t < 250$ GeV helps in both the cases of W+Jets and QCD. By imposing a requirement the b-jet p_T to be at least 300 GeV (200 GeV for the case where a medium p_T top quark is reconstructed), an improvement in the $t\bar{t}$ background is seen. Summarizing, the requirements applied to the W' decay products are:

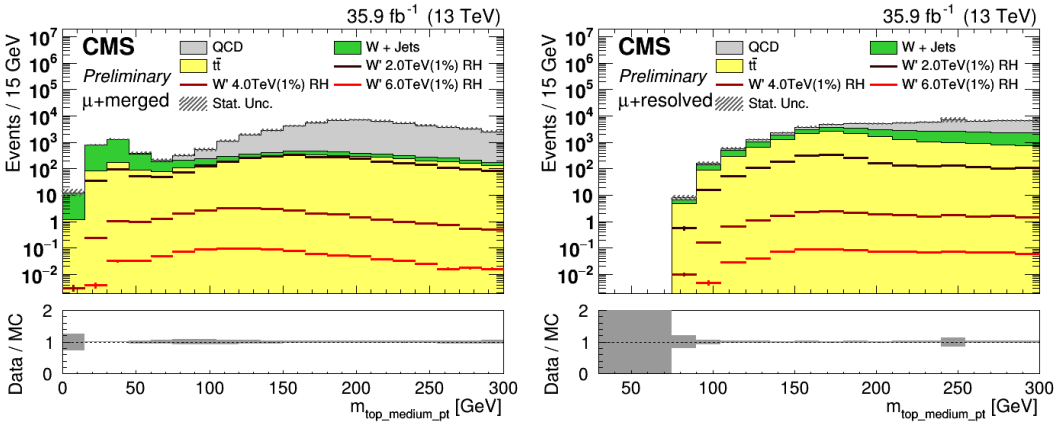
- $Score > 0.9(0.7)$ for high (medium) p_T top quarks;
- $(50 < m_t < 250)$ GeV;
- $p_{T,bjet} > 300(200)$ GeV for high (medium) p_T top quarks;
- $btagDeepFlavB > 0.4$;

The BDT score criterion for top quark selection is less efficient at lower p_T ranges, therefore a lower threshold has been chosen in order to find the best trade-off between selection efficiency and mis-identification.



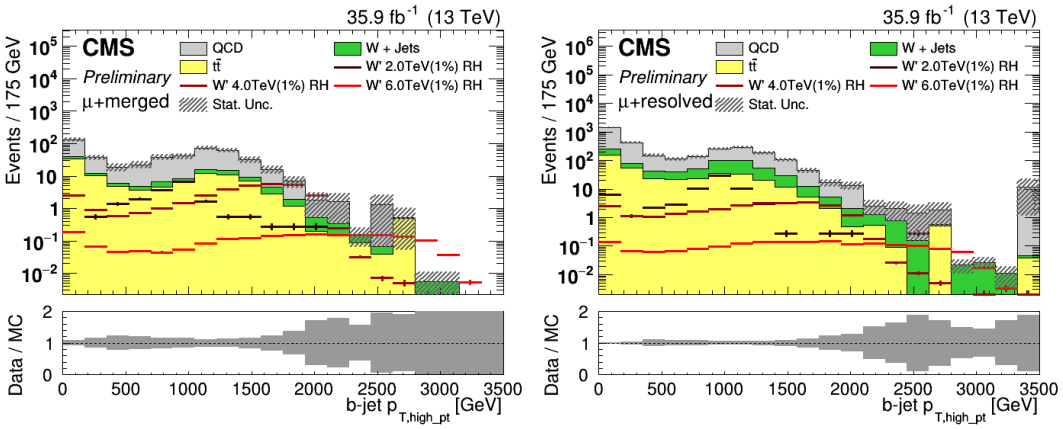
(a) (b)

Figure 5.9: Masses of the reconstructed (a) mu_merged (b) mu_resolved top quark candidates in the high p_T range. These variables were used to optimize the cut on the W' mass.



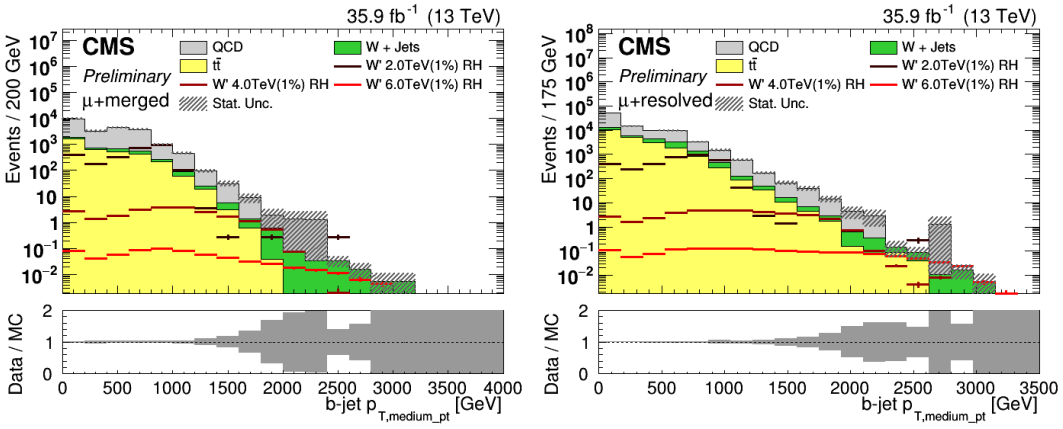
(a) (b)

Figure 5.10: Masses of the reconstructed (a) mu_merged (b) mu_resolved top quark candidates in the medium p_T range. These variables were used to optimize the cut on the W' mass.



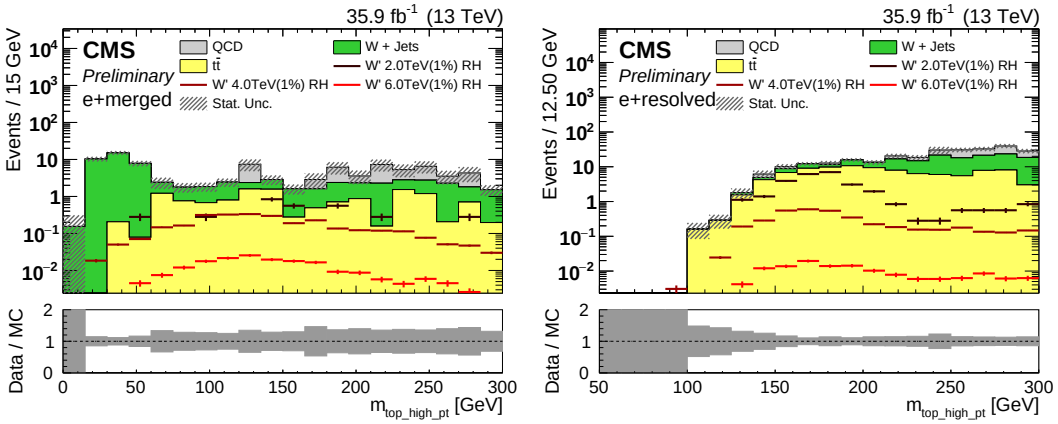
(a) (b)

Figure 5.11: Transverse momentum distributions of the b-jet involved in the W' reconstruction for the (a) mu_merged (b) mu_resolved top quark categories in the high p_T range. These variables were used to optimize the cut on the W' mass.



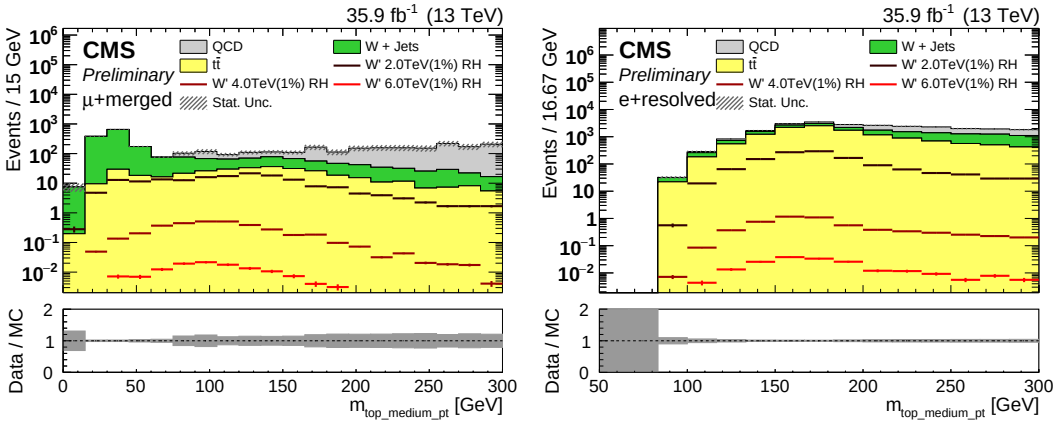
(a) (b)

Figure 5.12: Transverse momentum distributions of the b-jet involved in the W' reconstruction for the (a) mu_merged (b) mu_resolved top quark categories in the medium p_T range. These variables were used to optimize the cut on the W' mass.



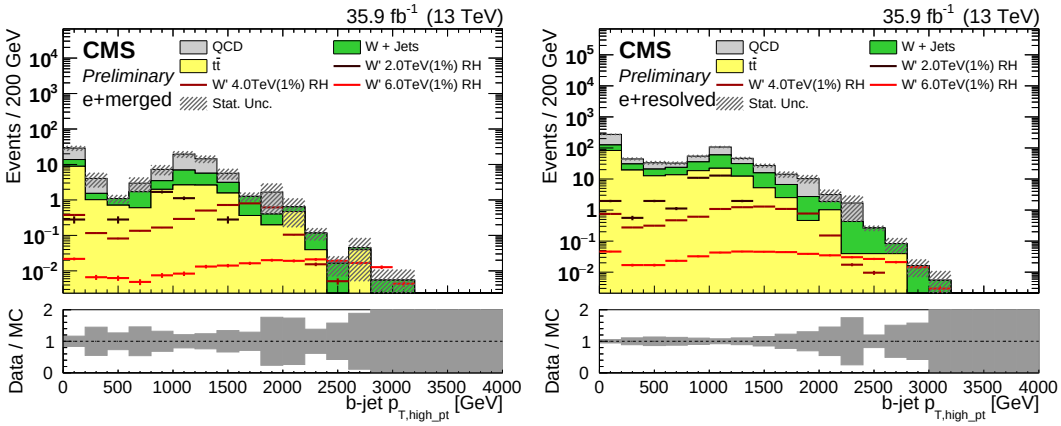
(a) (b)

Figure 5.13: Masses of the reconstructed (a) el_merged (b) el_resolved top quark categories in the high p_T range. These variables were used to optimize the cut on the W' mass.



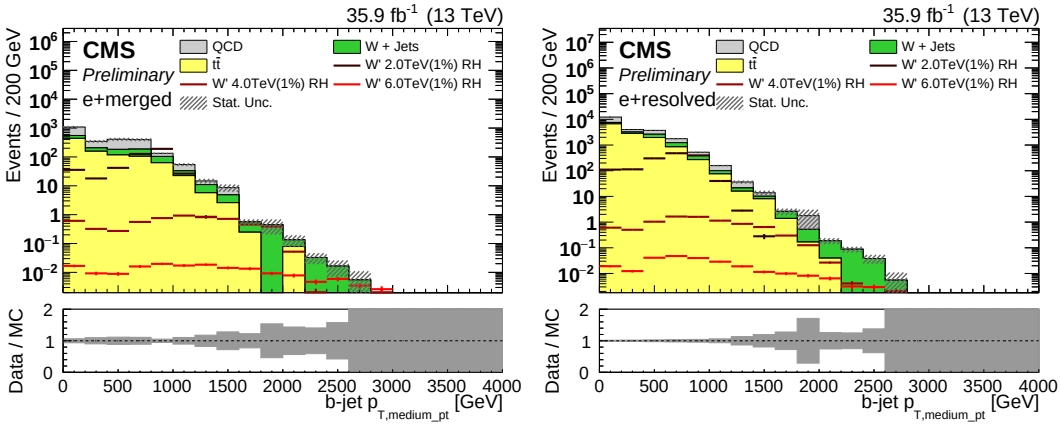
(a) (b)

Figure 5.14: Masses of the reconstructed (a) el_merged (b) el_resolved top quark categories in the medium p_T range. These variables were used to optimize the cut on the W' mass.



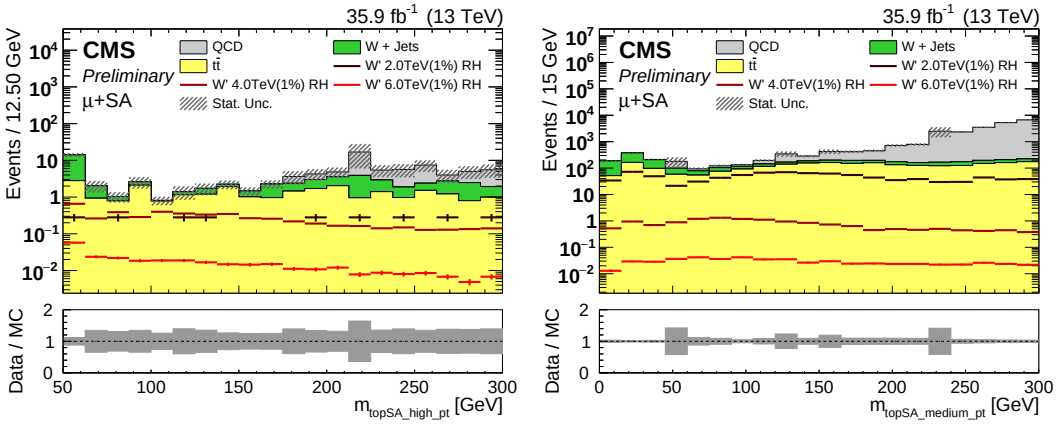
(a) (b)

Figure 5.15: Transverse momentum distributions of the b-jet involved in the W' reconstruction for the (a) el_merged (b) el_resolved top quark categories in the high p_T range. These variables were used to optimize the cut on the W' mass.



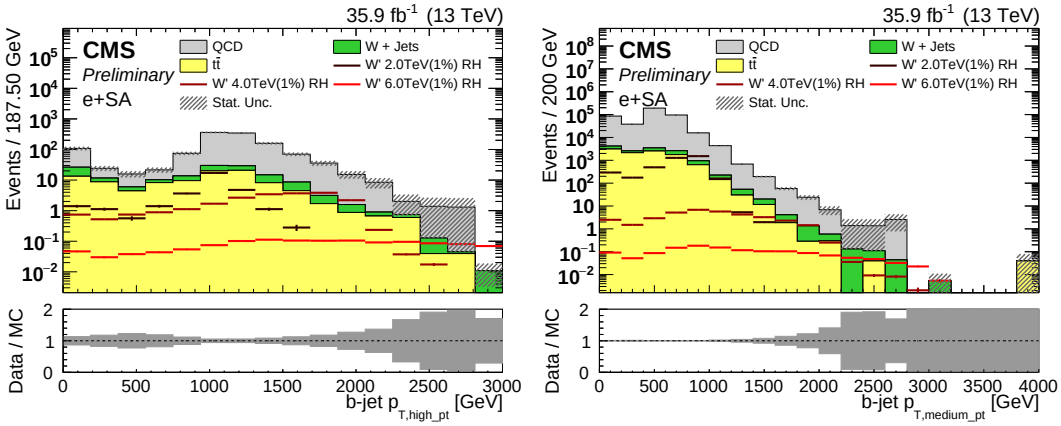
(a) (b)

Figure 5.16: Transverse momentum distributions of the b-jet involved in the W' reconstruction for the (a) el_merged (b) el_resolved top quark categories in the medium p_T range. These variables were used to optimize the cut on the W' mass.



(a) (b)

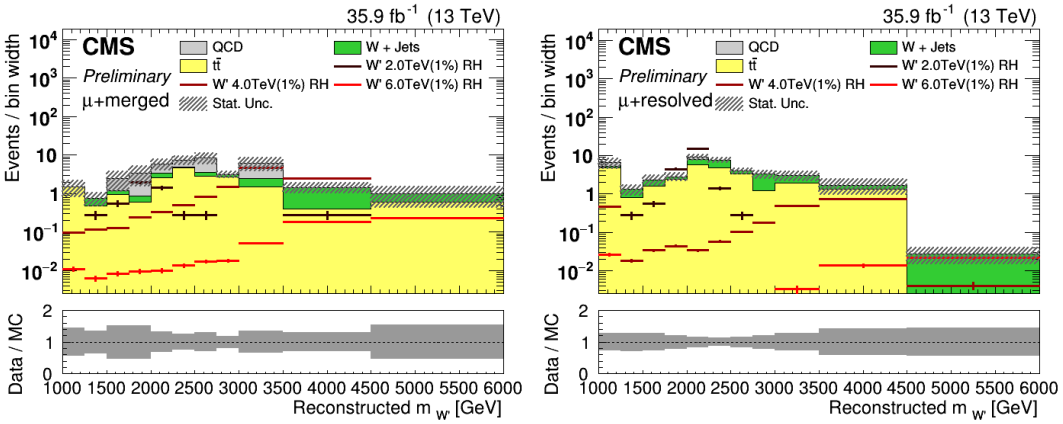
Figure 5.17: Masses of the reconstructed top candidates of the el_topSA configuration in the (a) high p_T (b) medium p_T range. These variables were used to optimize the cut on the W' mass.



(a) (b)

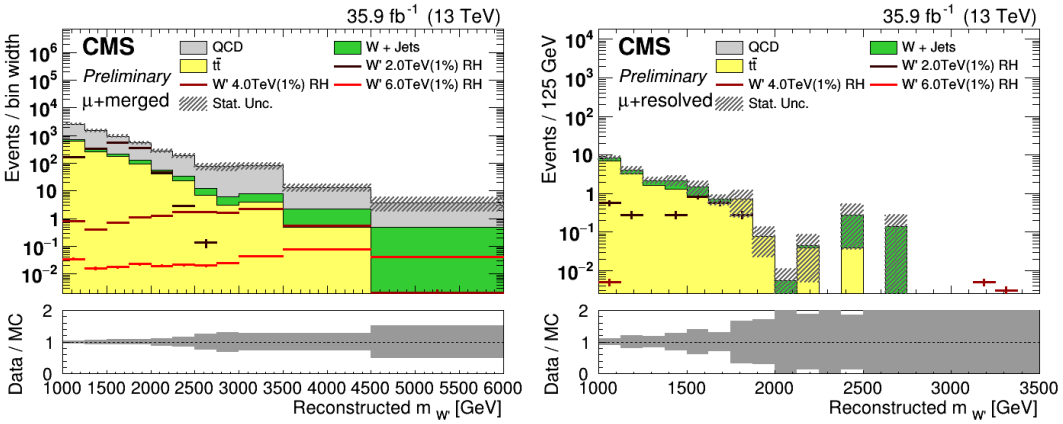
Figure 5.18: Transverse momentum distributions of the b-jet involved in the W' reconstruction for the el_topSA configuration in the (a) high p_T (b) medium p_T range. These variables were used to optimize the cut on the W' mass.

Finally, the W' distributions for the considered signals and backgrounds were obtained. The reconstructed W' mass shows a good discrimination power between signals and background, in particular for the Merged categories for both muons and electrons, and the SA category for electrons in the high top quark p_T range. In order to require the Merged and SA categories to be mutually exclusive, an additional requirement vetoing the el_merg in the presence of a el_topSA is applied. The same procedure was repeated in the case of the Resolved category, and seemingly the difference in the W' yield and shape was found to be negligible. Therefore, top quark categories result in event reconstruction categories that are almost completely exclusive, given the low likelihood of having two top quarks passing the top-tagging requirements.



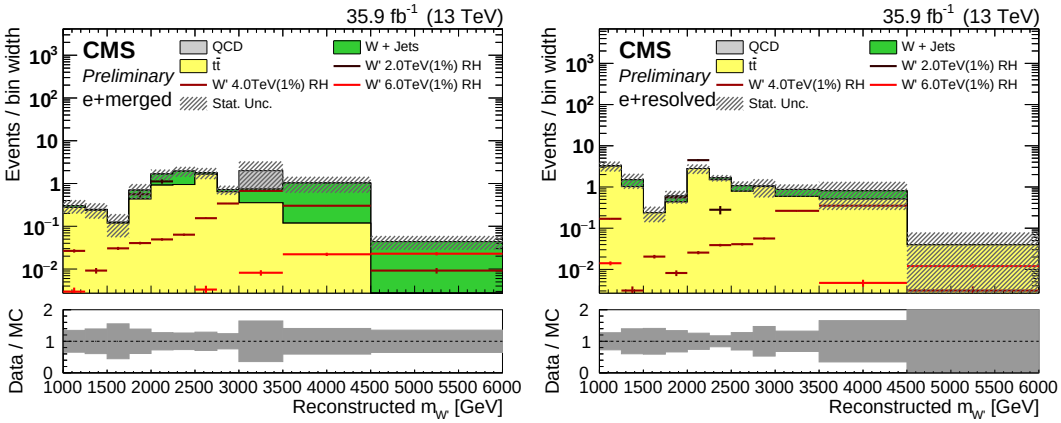
(a) (b)

Figure 5.19: W' boson mass distributions for the (a) mu_merged (b) mu_resolved top quark categories in the high p_T range.



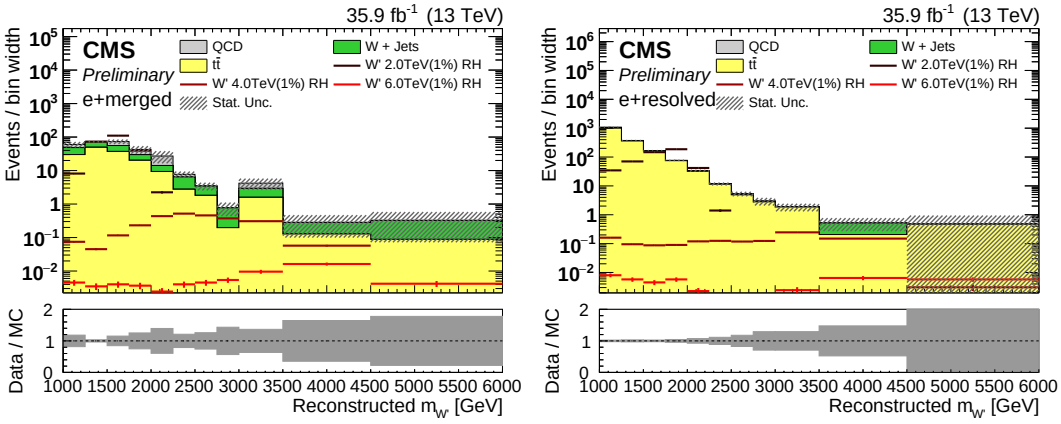
(a) (b)

Figure 5.20: W' boson mass distributions for the (a) mu_merged (b) mu_resolved top quark categories in the medium p_T range.



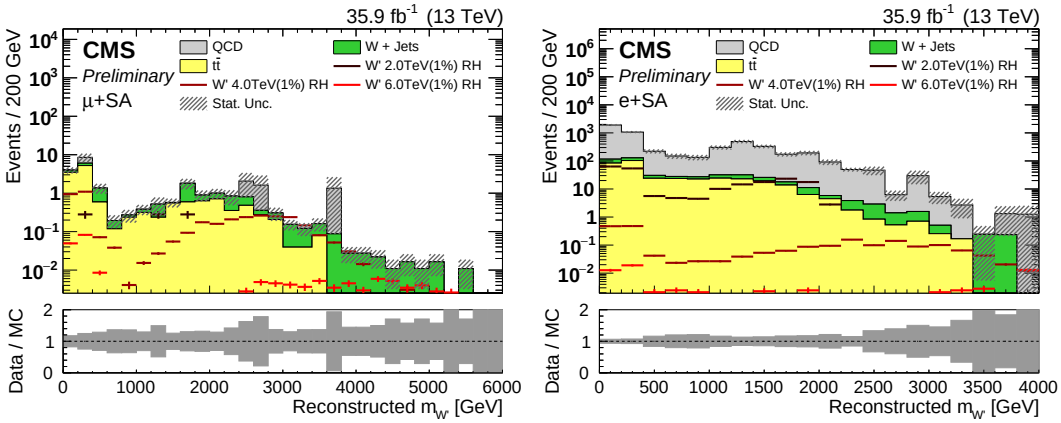
(a) (b)

Figure 5.21: W' boson mass distributions for the (a) el_merged (b) el_resolved top quark categories in the high p_T range.



(a) (b)

Figure 5.22: W' boson mass distributions for the (a) el_merged (b) el_resolved top quark categories in the medium p_T range.



(a) (b)

Figure 5.23: W' boson mass distributions for the el_topSA top quark category in the (a) high (b) medium p_T range.

After obtaining the distributions of the W' mass for the set of chosen samples, selection efficiencies were derived for all categories. The selection efficiency is defined as:

$$\epsilon = \frac{N_{sel}}{N} \tag{5.2}$$

In which N_{sel} is the number of events passing the selection for the considered category, while the $N = \sigma \cdot L$ is the total number of events, listed in Table 5.1. It is to be noted that this efficiency represents the combination of all the previous steps in the analysis process. A more accurate formula would be

$$\epsilon = \epsilon_{BR} \cdot \epsilon_{ML} \cdot \epsilon_{b-tag} \cdot \epsilon_{p_T}, \tag{5.3}$$

in which:

- ϵ_{BR}: each category is reconstructing a specific lepton, and, as already stated in Chapter 3, the BR is approximately 10% for both the electron and the muon cases. Therefore, the events of interest after the initial selection is approximately 10% of the initial number of entries;

- ϵ_{ML}: there are two steps of the ML procedure that create a natural event selection; the categorization, intrinsically designed to recognize and label those events that correspond to true signal events, and the score selection, base on the output of the training process;

- ϵ_{b-tag}: as already seen in Chapter 4 Section 4.1.2, the b-tagging efficiency is order of 50%-70% for the adopted working point (i.e. medium) in the case of very high p_T b-jets;

- ϵ_{p_T}: the binning into 3 different kinematic ranges for the reconstructed top quark applies an ulterior selection to the number of signal events;

The Tables 5.2, 5.3, 5.4, and 5.5 show that the numbers obtained are of an order of magnitude comparable with the expected values of efficiencies for this kind of process. The selection efficiencies in the case of the mu_topSA category are not particularly meaningful: the reconstructed W' mass distribution is pathological, meaning there are issues in the previous steps of reconstruction. This was somehow expected, because, as already stated, muons are naturally better separated from jets in the CMS reconstruction chain and in the CMS subdetector system, therefore they do not suffer from misidentification issues, rendering the construction of the SA muon categories from scratch non particularly meaningful or effective. Therefore, these objects will not be included in the subsequent analysis.

mu_merged			
high p_T		medium p_T	
Sample	Efficiency	Sample	Efficiency
W'(2 TeV)	0.009	W'(2 TeV)	0.02
W'(4 TeV)	0.015	W'(4 TeV)	0.017
W'(6 TeV)	0.016	W'(6 TeV)	0.009
$t\bar{t}$	$4.8 \cdot 10^{-6}$	$t\bar{t}$	$3.3 \cdot 10^{-4}$
W+Jets	$4.7 \cdot 10^{-8}$	W+Jets	$3.3 \cdot 10^{-6}$
QCD	$1.4 \cdot 10^{-9}$	QCD	$3.7 \cdot 10^{-7}$

Table 5.2: Selection efficiencies for the mu_merged category.

mu_resolved			
high p_T		medium p_T	
Sample	Efficiency	Sample	Efficiency
W'(2 TeV)	$4.8 \cdot 10^{-4}$	W'(2 TeV)	$6.0 \cdot 10^{-5}$
W'(4 TeV)	$3.6 \cdot 10^{-3}$	W'(4 TeV)	$2.4 \cdot 10^{-5}$
W'(6 TeV)	$2.0 \cdot 10^{-3}$	W'(6 TeV)	$1.4 \cdot 10^{-5}$
$t\bar{t}$	$7.2 \cdot 10^{-6}$	$t\bar{t}$	$3.9 \cdot 10^{-6}$
W+Jets	$8.8 \cdot 10^{-8}$	W+Jets	$3.7 \cdot 10^{-8}$
QCD	$1.9 \cdot 10^{-10}$	QCD	$7.8 \cdot 10^{-16}$

Table 5.3: Selection efficiencies for the mu_resolved category.

el_merged

high p_T		medium p_T	
Sample	Efficiency	Sample	Efficiency
W'(2 TeV)	$3.8 \cdot 10^{-5}$	W'(2 TeV)	$4.6 \cdot 10^{-5}$
W'(4 TeV)	$2.6 \cdot 10^{-3}$	W'(4 TeV)	$4.1 \cdot 10^{-3}$
W'(6 TeV)	$1.9 \cdot 10^{-3}$	W'(6 TeV)	$1.8 \cdot 10^{-3}$
$t\bar{t}$	$1.4 \cdot 10^{-6}$	$t\bar{t}$	$3.9 \cdot 10^{-5}$
W+Jets	$2.7 \cdot 10^{-8}$	W+Jets	$6.4 \cdot 10^{-7}$
QCD	$9.7 \cdot 10^{-11}$	QCD	$5.3 \cdot 10^{-9}$

Table 5.4: Selection efficiencies for the el_merged category.

el_resolved

high p_T		medium p_T	
Sample	Efficiency	Sample	Efficiency
W'(2 TeV)	$5.6 \cdot 10^{-5}$	W'(2 TeV)	$9.57 \cdot 10^{-3}$
W'(4 TeV)	$6.3 \cdot 10^{-4}$	W'(4 TeV)	$3.7 \cdot 10^{-3}$
W'(6 TeV)	$2.0 \cdot 10^{-7}$	W'(6 TeV)	$8.9 \cdot 10^{-7}$
$t\bar{t}$	$5.6 \cdot 10^{-7}$	$t\bar{t}$	$4.2 \cdot 10^{-4}$
W+Jets	$3.8 \cdot 10^{-12}$	W+Jets	$1.6 \cdot 10^{-7}$
QCD	$7.8 \cdot 10^{-15}$	QCD	$7.5 \cdot 10^{-7}$

Table 5.5: Selection efficiencies for the el_resolved category.

el_topSA

high p_T		medium p_T	
Sample	Efficiency	Sample	Efficiency
W'(2 TeV)	$1.6 \cdot 10^{-5}$	W'(2 TeV)	$4.2 \cdot 10^{-3}$
W'(4 TeV)	0.007	W'(4 TeV)	0.004
W'(6 TeV)	0.008	W'(6 TeV)	0.003
$t\bar{t}$	$3.4 \cdot 10^{-6}$	$t\bar{t}$	$9.1 \cdot 10^{-5}$
W+Jets	$4.5 \cdot 10^{-8}$	W+Jets	$8.6 \cdot 10^{-7}$
QCD	$4.9 \cdot 10^{-10}$	QCD	$3.7 \cdot 10^{-7}$

Table 5.6: Selection efficiencies for the el_topSA category.

mu_topSA			
high p_T		medium p_T	
Sample	Efficiency	Sample	Efficiency
W'(2 TeV)	$1.2 \cdot 10^{-4}$	W'(2 TeV)	$1.3 \cdot 10^{-3}$
W'(4 TeV)	0.05	W'(4 TeV)	$9.1 \cdot 10^{-3}$
W'(6 TeV)	0.05	W'(6 TeV)	$6.0 \cdot 10^{-3}$
$t\bar{t}$	$5.9 \cdot 10^{-5}$	$t\bar{t}$	$1.0 \cdot 10^{-7}$
W+Jets	$5.3 \cdot 10^{-7}$	W+Jets	$7.5 \cdot 10^{-8}$
QCD	$4.7 \cdot 10^{-8}$	QCD	$3.7 \cdot 10^{-9}$

Table 5.7: Selection efficiencies for the mu_topSA category

Finally, a note on the low p_T range: the reconstruction of the event is clearly pathological, as already seen in Figure 5.7 for the Merged and Resolved categories, and in Figure 5.8 for the SA one. There are clear signs of misidentification; therefore, although the selection efficiencies are relatively high with regards to the corresponding efficiencies for background samples, the algorithm is selecting the wrong events most of the times. This is probably due to the fact that the algorithm of reconstruction of the top quarks and the identification of the b-jets is not optimized in the low energy ranges, and since the nature of the considered objects is intrinsically energetic, being them a direct byproduct of a massive object such as the W', the process of analysis is more difficult and in need of a separate in-depth study in order to make a proper object selection. Therefore, the subsequent analysis will not be carried forward on these categories in the low p_T range.

low p_T			
mu_merged		mu_resolved	
Sample	Efficiency	Sample	Efficiency
W'(2 TeV)	$3.7 \cdot 10^{-5}$	W'(2 TeV)	$1.3 \cdot 10^{-3}$
W'(4 TeV)	$1.5 \cdot 10^{-4}$	W'(4 TeV)	$3.1 \cdot 10^{-4}$
W'(6 TeV)	$2.9 \cdot 10^{-4}$	W'(6 TeV)	$1.2 \cdot 10^{-4}$
$t\bar{t}$	$1.8 \cdot 10^{-7}$	$t\bar{t}$	$5.9 \cdot 10^{-4}$
W+Jets	$7.8 \cdot 10^{-10}$	W+Jets	$2.4 \cdot 10^{-7}$
QCD	$1.5 \cdot 10^{-9}$	QCD	$2.8 \cdot 10^{-8}$

Table 5.8: Selection efficiencies for the low p_T range for the categories involving muons.

low p_T			
el_merged		el_resolved	
Sample	Efficiency	Sample	Efficiency
W'(2 TeV)	$7.0 \cdot 10^{-3}$	W'(2 TeV)	$3.9 \cdot 10^{-3}$
W'(4 TeV)	$3.6 \cdot 10^{-3}$	W'(4 TeV)	$1.2 \cdot 10^{-3}$
W'(6 TeV)	$1.9 \cdot 10^{-3}$	W'(6 TeV)	$1.2 \cdot 10^{-3}$
$t\bar{t}$	$2.0 \cdot 10^{-4}$	$t\bar{t}$	$1.7 \cdot 10^{-3}$
W+Jets	$1.1 \cdot 10^{-7}$	W+Jets	$1.3 \cdot 10^{-6}$
QCD	$4.0 \cdot 10^{-8}$	QCD	$1.2 \cdot 10^{-7}$

Table 5.9: Selection efficiencies for the low p_T range for the categories involving electrons.

low p_T el_topSA	
Signal	Efficiency
W'(2 TeV)	$7.0 \cdot 10^{-3}$
W'(4 TeV)	$3.7 \cdot 10^{-3}$
W'(6 TeV)	$1.9 \cdot 10^{-3}$
$t\bar{t}$	$2.09 \cdot 10^{-4}$
W+Jets	$1.1 \cdot 10^{-7}$
QCD	$1.3 \cdot 10^{-8}$

Table 5.10: Selection efficiencies for the low p_T range for the el_topSA category.

5.4 Fit procedure

For this analysis, a maximum likelihood estimator is used to extract the signal. The method is based on the construction of the combined probability distribution of all entries in a data sample, called likelihood function:

$$L(x_1, ..., x_n; \theta_1, ..., \theta_m) = f(x_1, ..., x_n; \theta_1, ..., \theta_m), \tag{5.4}$$

in which f is the joint Probability Distribution Function (PDF) of the random variables $x_1, ..., x_n$, and $\theta_1, ..., \theta_m$ are a set of unknown parameters. The estimate of the parameters to determine is obtained by finding the parameter set that corresponds to the maximum value of the likelihood function. This approach gives the name of *maximum likelihood method* to this technique [29]. In the case of N repeated measuremets, the likelihood function is the probability density corresponding to the total sample $\vec{x} = \{(x_1^1, ..., x_n^1), ..., (x_1^N, ..., x_n^N)\}$. If the observations are indipendent of each other, the likelihood function can be written as:

$$L(\vec{x}; \vec{\theta}) = \prod_{i=1}^{N} = f(x_1^i, ..., x_n^i; \theta_1, ..., \theta_n) \tag{5.5}$$

Usually, the logarithm of the likelihood function is computed, so that:

$$-logL(\vec{x}; \vec{\theta}) = -\sum_{i=1}^{N} logf(x_1^i, ..., x_n^i; \theta_1, ..., \theta_n). \tag{5.6}$$

For this analysis, an extended binned maximum likelihood is used:

$$L(\vec{x}, \theta) = \mathcal{P}(s(\theta) + b(\theta)) \prod_{i=1}^{N} [w_s f_s(x_i, \theta) + w_b f_b(x_i, \theta)] \tag{5.7}$$

in which $P(s + b)$ is the Poisson distribution, and f_s and f_b are the PDFs taken from the MC template histogram. The quantities w_s and w_b are the relative fractions of the signal and background; $b = b_{t\bar{t}} + b_{W+Jets} + b_{QCD}$ is the number of background events, and s is the number of signal events $s_{W'}$. This parameter allows to extract information on the cross section by making use of the signal efficiencies derived from MC, and reported in Tables 5.2, 5.3, 5.4, 5.5, and 5.6. The ratio $s_{W'}/s_{theory}$ is commonly referred to as signal strength R. A test statistic t can be used to measure the degree of compatibility between data and a hypothesis to refute, named null hypothesis. In case of a search for new physics, the null hypothesis is the case in which only background is present, while a competing hypothesis is the case where both the expected signal and background are present. In order to define the test-statistic, the profile likelihood ratio of the fit to the W' reconstructed mass has been used:

$$\lambda(\theta) = \frac{L_{s+b}(\vec{x}, \theta)}{L_b(\vec{x}, \theta)}, \tag{5.8}$$

from which the test statistic is defined as

$$t = -2log\lambda(\theta) \tag{5.9}$$

The probability p that the considered test statistic assumes a value greater or equal to the one observed in data due to a statistical fluctuation of the null-hypthesis is called p-value. The upper limits to the cross section were evaluated using the Combined toolkit provided by the CMS collaboration, which uses the method of CL in order to provide a measure of the level of compatibility of data with a signal hypothesis. This method defines the quantity:

$$CL_s(\vec{\theta}) = \frac{p_{s+b}(\vec{\theta})}{1 - p_b(\vec{\theta})} \tag{5.10}$$

in which p_b is the p-value of the *null hypothesis*, namely the probability that only background is present, while p_{s+b} is the p-value of the signal hypothesis. The profile maximum likelihood fit is set up as a simultaneous binned maximum likelihood fit to the $m_{W'}$ in the signal categories. The fit is performed in multiple configurations to compare the performances of the different algorithms. The expected upper limits on the signal cross section at 95% CL were evaluated, considering a luminosity of $35.9fb^{-1}$ corresponding to the 2016 data set, on the following configurations:

- μ_{tot} : mu_merged + mu_resolved categories in the high and medium p_T range (Figure 5.24);
- $e_{merg+SA}$: el_merged + el_topSA in the the high and medium p_T range (Figure 5.26);
- e_{tot} : el_merged + el_resolved + el_topSA in the high and medium p_T range (Figure 5.25);
- $Merged_{e+\mu}$: el_merged + mu_merged in the high and medium p_T range (Figure 5.27);
- $Total_{e+\mu}$: el_merged + mu_merged + el_resolved + mu_resolved + el_topSA in the high and medium p_T range (Figure 5.28);

No systematic uncertainties were included at this stage of the analysis, with the sole exception of the one concerning luminosity, which is treated as a nuisance parameter in the construction of the maximum likelihood function. As expected, muons perform better than electrons. In regards to the electron categories, the el_topSA category provides a non-negligible contribution to the evaluation of the upper limit, which is compatible with the order of magnitude of the one provided by the el_merged category. Furthermore, the upper limit with the $Total_{e+\mu}$ is slightly improving on the existing analysis limits on the cross section, and allowing to access a higher mass range mitigating high p_T efficiency loss, ultimately raising the value of the excluded mass. [31].

Configuration	Upper Limit on R at 95% CL		
	W'(2 TeV)	W'(4 TeV)	W'(6 TeV)
μ_{tot}	0.0718	0.7852	9.9062
e_{tot}	0.0713	1.5078	22.7500
$e_{merg+SA}$	0.1255	1.6641	24.8750
$Merged_{e+\mu}$	0.0835	0.7695	9.4062
$Total_{e+\mu}$	0.0513	0.6660	8.1875

Table 5.11: Upper limit on the signal strength R = $\sigma/\sigma_{expected}$ at 95% CL for each configuration.

Configuration	Cross section (pb) Upper Limit at 95% CL		
	W'(2 TeV)	W'(4 TeV)	W'(6 TeV)
μ_{tot}	0.1003	0.0136	$9.1 \cdot 10^{-3}$
e_{tot}	0.0996	0.0261	$2.1 \cdot 10^{-2}$
$e_{merg+SA}$	0.1753	0.0288	$2.3 \cdot 10^{-2}$
$Merged_{e+\mu}$	0.1166	0.01335	$8.6 \cdot 10^{-3}$
$Total_{e+\mu}$	0.0716	0.0115	$7.4 \cdot 10^{-3}$

Table 5.12: Expected cross sections for each configuration, obtained by multiplying the signal strength with the theoretical σ value.

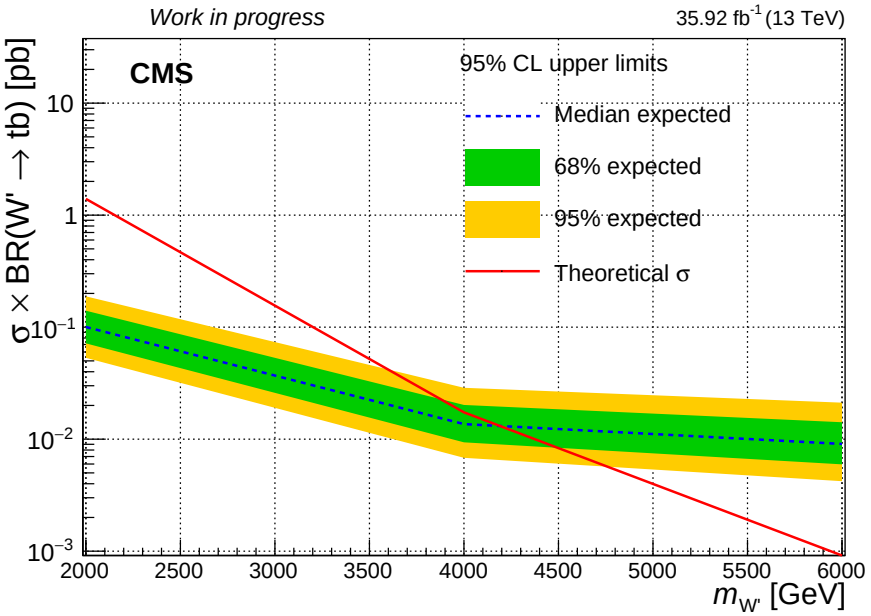


Figure 5.24: Expected 95% CL upper limit for the μ_{tot} configuration.

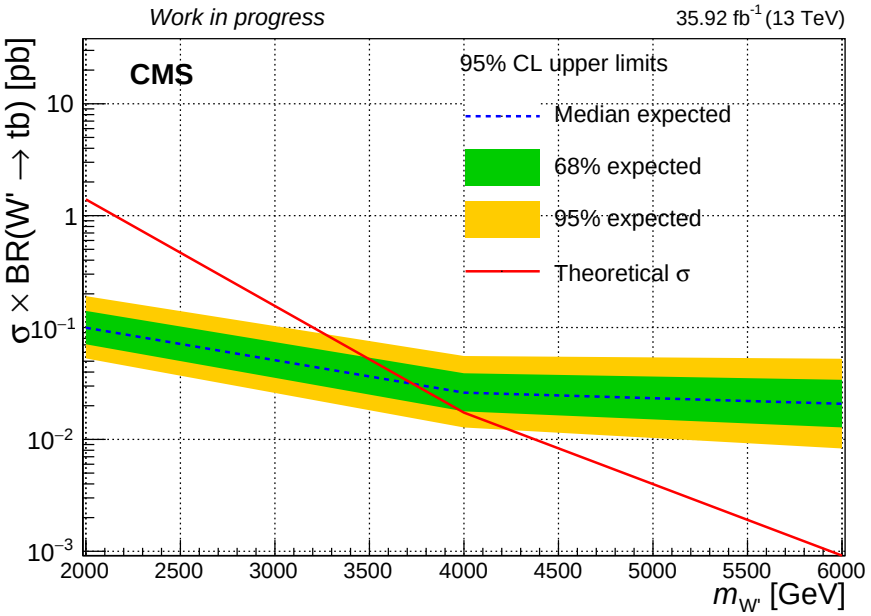


Figure 5.25: Expected 95% CL upper limit for the e_{tot} configuration.

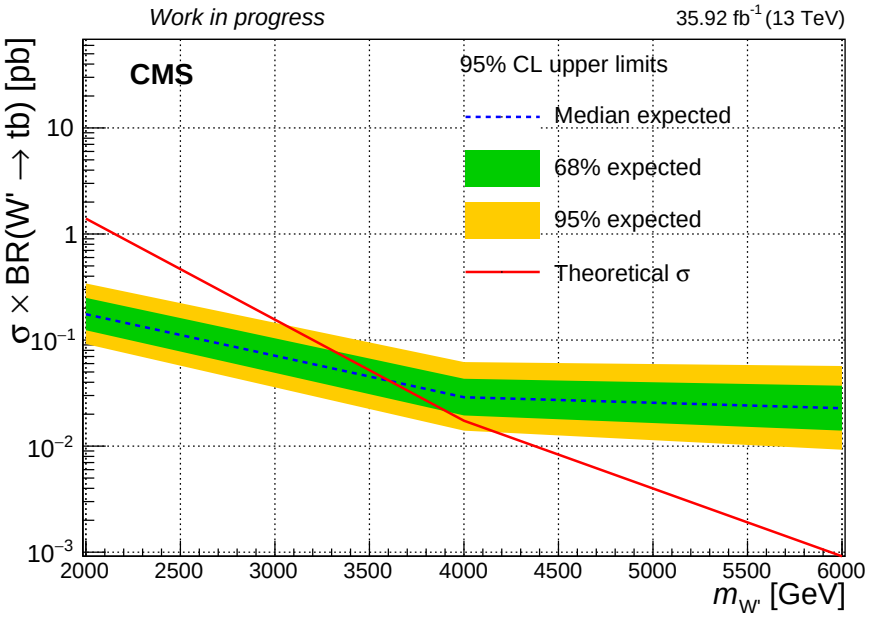


Figure 5.26: Expected 95% CL upper limit for the $e_{merg+SA}$ configuration.

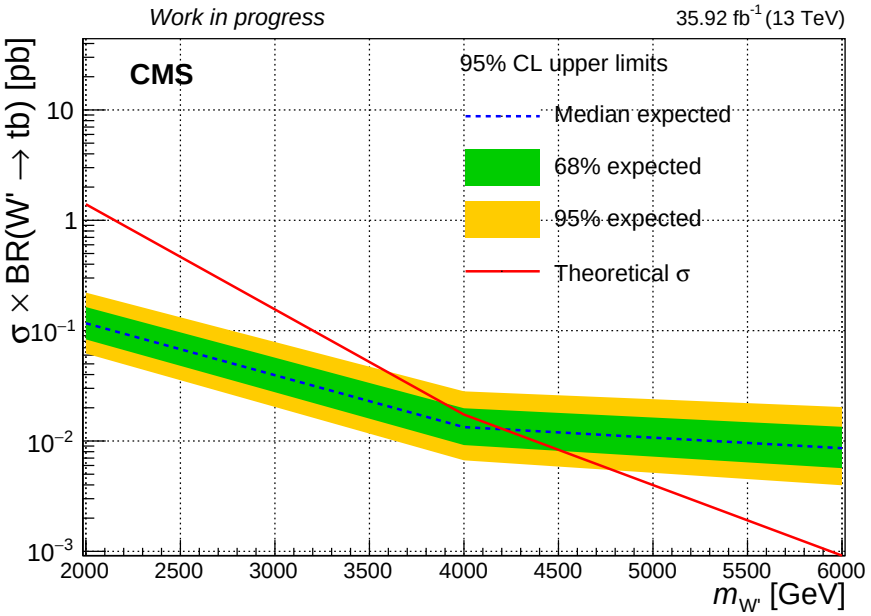


Figure 5.27: Expected 95% CL upper limit for the $Merged_{e+\mu}$ configuration.

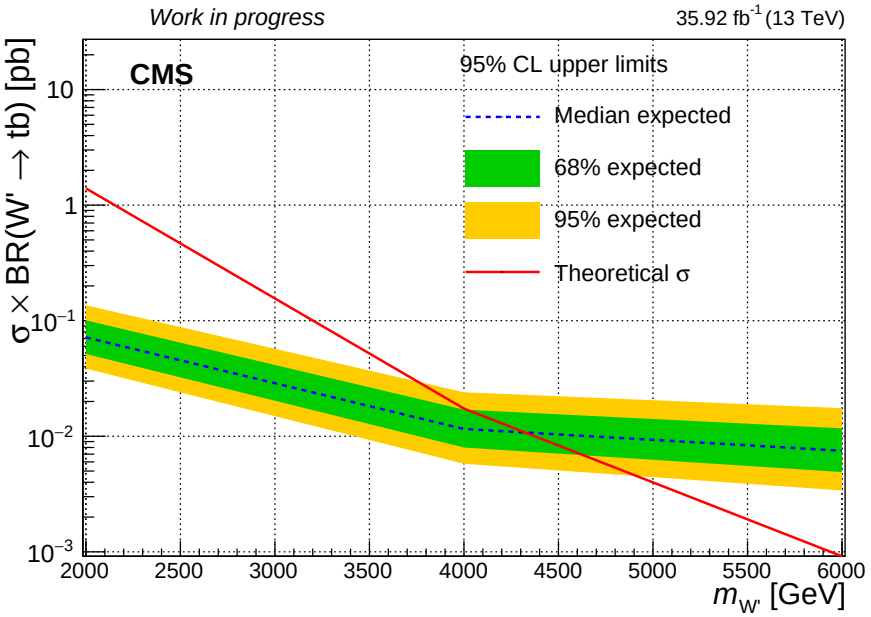


Figure 5.28: Expected 95% CL upper limit for the $Total_{e+\mu}$ configuration.

Conclusions

In this thesis, an analysis strategy for the search of the W' boson using Machine Learning techniques has been developed. The analysis was performed on Monte Carlo simulated samples of W' signal events and its most important backgrounds, produced in proton-proton collisions at LHC with a centre-of-mass energy of 13 TeV, and reconstructed with the CMS detector. The simulated data samples reproduce the data taking conditions of the detector during 2016, and they are scaled to the corresponding integrated luminosity of 35.9 fb^{-1}. The W' boson is a new resonance predicted by many theories Beyond Standard Model, and among these models, particularly interesting are the ones that predict the W' to have a preferential coupling with the third generation of quarks and leptons, and they foresee values of the W' mass in the TeV range. Such boson is a heavier counterpart of the SM W boson, therefore it has the chance to decay into a top and bottom quark. The bottom quark undergoes hadronization, creating a jet, while the top quark can decay to both hadronic and leptonic final states. The considered leptonic decay channels include either a muon or electron in the final state, in association with a bottom quark and a neutrino. The main goal of this work was to study the top quark reconstruction for its leptonic decay making use of Machine Learning techniques. Three categories of reconstructed top quarks have been defined: the Merged category, for which the lepton is reconstructed inside the b-jet, the Resolved category, for which the lepton and the b-jet are angularly separated, and the StandAlone(SA) category, considering the cases in which leptons are overlapping with the b-jet, as in the Merged category, but not reconstructed or identified by standard algorithms. This last category was defined in order to recover signal inefficiency, observed in W' signal simulation, in the electron channel, in particular for Merged events, after observing from simulation a non-negligible difference between the number of events with reconstructed and true top quarks. These three categories have been split into three kinematic ranges, based on the p_T of the reconstructed top quark, for a total of 18 models. Instead of the standard selection based on requirements on the single decay components, a *top-tagging* was performed with the aid of a Boosted Decision Tree algorithm, a binary classifier which performs a search on the variables in order to select the true top quark candidates in each category. In each event, for those categories having more than one reconstructed top quark candidate, a hierarchy was established based on the output score of the BDT. The best candidate for each category was then selected with a requirement that rejected

90% of fake top quarks. For each category, the W' mass was reconstructed; the jet stemming from the bottom quark hadronization at the W' decay vertex was selected by making requirements on its p_T and the score of an appropriate algorithm for tagging b-jets. An extended binned maximum likelihood fit to the MC simulated dataset, reproducing a realistic 2016 dataset, is performed; the fit accommodates various configurations in order to evaluate the impact on the result of various algorithms. Upper limits on the cross sections have been estimated at 95% CL. The addition of the SA categories for electrons results in a betterment on the existing analysis limits on the cross section, and allows to access a higher mass range mitigating high p_T efficiency loss, ultimately raising the value of the excluded mass. Further implementations in the analysis could be the introduction of systematic uncertainties, and the optimization of the reconstruction algorithm; the top-tagging could make use of different Machine Learning classifiers able to give a hierarchy not only to top candidates but also to reconstruction strategies. Furthermore, the implemented analysis could be performed on data from Run-II after making the proper tuning of the Machine Learning algorithm and eventually applied on the up-and-coming data from the Run-III data taking.

Ackowledgements

Since I am writing these words, I can fairly assume that this part of my life is almost over. These two years have been very tough for everyone, and this thesis is, in spite of everything, an important albeit small personal milestone that I would not have been able to reach, had it not been for the support I had from all the people in my life. I was hesitant about writing Acknowledgements for a Master Thesis, but I have so much gratitude to spare, so here we go.
First and foremost, I would like to thank Orso, for his infinite amount of patience, kindness, competence and wisdom, all bestowed unconditionally. Supervisor, teacher, occasional shrink, steady provider of memes (see Figure 1.29), I really could not have done it without him, since he believed in me more than I believed in myself.

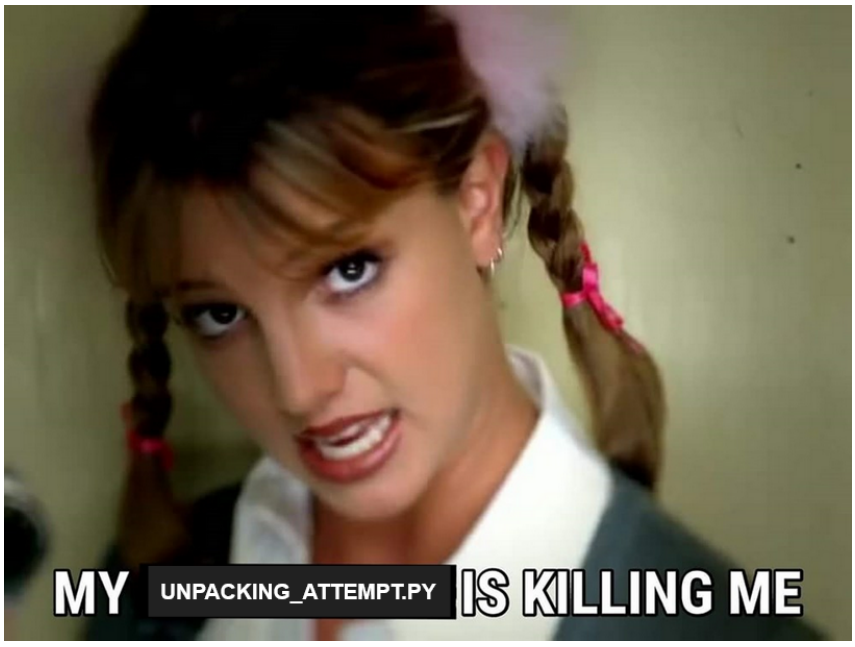


Figure 1.29

I would also like to thank my CMS-mates, Agostino, Antimo, Francesco, and Oriana; they really are a bunch of "GEMs" (pun intended) and they really know how to make everyday more cheerful.
Throughout my time at university, four people have been my constant companions, much to their chagrin. Even though we took on different career paths, we

shared a lot, from travels to mental breakdowns; their company and friendship is priceless and I feel the need to address them personally.
To Emanuele, with which I have a moral debt that goes way beyond initiating me to my current fandom fixation; your ability to suspend judgement in order to ask all the right questions and get to know all the perspectives of a topic is something that I have always admired and it's always a pleasure to converse with you.
To Fabrizio, who is currently working as an undercover agent in Bologna and therefore we see less, but still fondly consider as our own woodland elf and who we miss each day; the drive and determination you put in everything you do have always been admirable, and getting to hear about your scouting experiences is always a knee-slapper.
To Gracey (she is going to kill me), who, despite everything she might say, is the kindest, sweetest, most honest and reliable person I know; your support is invaluable, the simple fact that you have managed to put up with me for so long is a testament to the kind of person you are, and I am happy that, a least, I make you laugh, even when you don't want to.
To Manuel, my twin brother from another mother, the apple of my eye, I do not have enough words to describe what your presence in my life really means; from university to Ballroom, we share a lot, and my deepest wish is that we will always be able to. "Me & you, trading smooches at every Ball" is not a statement, but a threat.
I also thank my other classmates and friends, some of which currently dislocated all over the globe. We may not be as close (at least physically) as we once were, but I deeply cherish each and every one of you nonetheless: Alessandra, Erasmo, Matteo, Michele, Pasquale, Renato, Roberta, and Salvatore. A special mention goes to Giuseppe, my lab-5 mate and companion in misfortunes, with which I got to know the joys of distance learning while trying to keep a sliver of sanity intact. You knew the struggle, and if one good thing came out of these last few years is that now I can proudly call you a friend.
A big, fat, wholeheartedly felt thank you goes to Mike, my cynical soulmate, who gets me more than any other. You always were a cornerstone in all these years spent languishing on books, and those times spent sharing music recommendations are something I cherish deeply.
Thank you to Luca, with whom I shared some of the most amazing years of my life, I miss going to the Conservatory, attend lessons with the Maestro and snickering about everyone and everything with you. Despite my frequent absences and general unavailability, you still put up with me for all these years, and I thank you for your support and friendship.
Thanks to my chosen family, the House of Lamborghini: my Mother Yunikon, my brothers and sisters Jupiter, Masako, Parsifal (yes, Manuel, you are counted twice!) and little baby Jaguar. You made my life so much brighter than it was before and I cannot thank you enough for that.
To my family: my aunts Raffaella and Anna, and my grandma, who are probably more excited about this than I am, and then all my uncles and cousins. To my father, and to my baby brother Luca, who I miss every second of the

day ever since he moved to Rome, and who I love very much.
And finally to you, mom. Not for importance, but because I will never have enough time or space or words to convey the depth of the gratitude I have towards you, so I am intentionally limiting myself. These past five years have been particularly rough for us, no one knows that better than you. Now that they are almost over, I can say with absolute certainty than I would not be alive to tell the tale if it weren't for you. Alas I am, alive and still kicking, and all thanks to you.
Thanks to you, I carried on doing the things I love, even though most people thought I would not be able to keep up.
Thanks to you, even in the darkest days I was reminded that I had had worse, and it made everything a little easier, a little brighter.
Thanks to you, I never felt alone in facing all the obstacles thrown at me because you always took on a fair share of the burden.
Thanks to you, I learnt that struggle, discomfort, failure, and even illness are not necessarily a death sentence, but just part of the process.
We always joke about you being strict and overly critical of everything Luca and I do, but in the end we all know that it's all a façade, for you are and always have been our number one supporter. I owe you so much, and I sincerely hope that one day I will get to be just a sliver of the person you are: kind, loving, honest, compassionate and fair.
I love you and I thank you, for supporting and accepting me for everything I do and everything I am.
Ultimately, this is for you.

Bibliography

[1] P. Zyla et al. "*Review of Particle Physics*". *PTEP*, 2020:083C01. 2093 p, 2020.

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

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

[4] G. Aad et al. "*Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC*". *Physics Letters B*, 716(1):1–29, Sep 2012.

[5] S. Chatrchyan et al. "*Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC*". *Physics Letters B*, 716(1):30–61, Sep 2012.

[6] D. J. Gross and F. Wilczek. "*Ultraviolet Behavior of Non-Abelian Gauge Theories*". *Phys. Rev. Lett.*, 30:1343–1346, Jun 1973.

[7] H. D. Politzer. "*Reliable Perturbative Results for Strong Interactions?*". *Phys. Rev. Lett.*, 30:1346–1349, Jun 1973.

[8] S. Chatrchyan et al. The cms experiment at the cern lhc. *Journal of Instrumentation*, 3:S08004, 08 2008.

[9] Panja Luukka. "*CMS Inner Tracker Upgrade*". Technical report, CERN, Geneva, Jan 2020.

[10] Abdullah M. "*Probing a simplified W' model of R(D(*)) anomalies using b tags, tau leptons, and missing energy*". *Physical Review D*, 98(5), Sep 2018.

[11] T. Appelquist et al. *"Modern Kaluza-Klein theories"*. Addison-Wesley Pub. Co, 1987.

[12] N. Arkani-Hamed, A. G Cohen, E. Katz, and A. E Nelson. "*The Littlest Higgs*". *Journal of High Energy Physics*, 2002(07):034–034, Jul 2002.

[13] R. S. Chivukula, B. A. Dobrescu, H. Georgi, and C. T. Hill. "*Top quark seesaw theory of electroweak symmetry breaking*". *Physical Review D*, 59(7), Mar 1999.

[14] David J Muller and Satyanarayan Nandi. "*Topflavor: a separate SU(2) for the third family*". *Physics Letters B*, 383(3):345–350, Sep 1996.

[15] R. Calabrese, A. De Iorio, D. Fiorillo, A. O. M. Iorio, G. Miele, and S. Morisi. "*Top-flavor scheme in the context of W' searches at LHC*". *Physical Review D*, 104(5), Sep 2021.

[16] E. Boos, V. Bunichev, L. Dudko, and M. Perfilov. "*Interference between W' and W in single-top quark production processes*". *Physics Letters B*, 655(5-6):245–250, Nov 2007.

[17] Z. Sullivan. "*Fully differential W' production and decay at next-to-leading order in QCD.*". *Physical Review D*, 66(7), Oct 2002.

[18] A.M. Sirunyan et al. "*Combination of CMS searches for heavy resonances decaying to pairs of bosons or leptons*". *Physics Letters B*, 798:134952, Nov 2019.

[19] A.M. Sirunyan et al. "*Search for heavy resonances decaying to a top quark and a bottom quark in the lepton+jets final state in proton-proton collisions at 13TeV*". *Physics Letters B*, 777:39–63, Feb 2018.

[20] G. Aad et al. "*Search for vector-boson resonances decaying to a top quark and bottom quark in the lepton plus jets final state in pp collisions at s=13 TeV with the ATLAS detector*". *Physics Letters B*, 788:347–370, Jan 2019.

[21] S. Chatrchyan and V. Khachatryan et al. "*Search for a W' boson decaying to a bottom quark and a top quark in pp collisions at s=7 TeV*". *Physics Letters B*, 718(4):1229–1251, 2013.

[22] A.M. Sirunyan, A. Tumasyan, W. Adam, E. Asilar, T. Bergauer, J. Brandstetter, E. Brondolin, M. Dragicevic, J. Erö, and M. Flechl et al. "*Particle-flow reconstruction and global event description with the CMS detector*". *Journal of Instrumentation*, 12(10):P10003–P10003, Oct 2017.

[23] A.M. Sirunyan et al. "*Performance of the CMS muon detector and muon reconstruction with proton-proton collisions at* $\sqrt{s} = 13$ *TeV*". 13(06):P06015–P06015, jun 2018.

[24] CMS Collaboration. Performance of electron reconstruction and selection with the CMS detector in proton-proton collisions at $\sqrt{s} = 8$ tev. 10(06):P06005–P06005, jun 2015.

[25] M. Cacciari, G. P. Salam, and G. Soyez. "*The anti-ktjet clustering algorithm*". *Journal of High Energy Physics*, 2008(04):063–063, Apr 2008.

[26] Mauro Verzetti. "*Machine learning techniques for jet flavour identification at CMS*". *EPJ Web of Conferences*, 214:06010, 01 2019.

[27] A.M. Sirunyan et al. "*Identification of heavy-flavour jets with the CMS detector in pp collisions at 13 TeV*". 13(05):P05011–P05011, may 2018.

[28] Matthew D. Schwartz. "*Modern Machine Learning and Particle Physics*", 2021.

[29] Luca Lista. *"Statistical Methods for Data Analysis in Particle Physics"*. Lecture Notes in Physics 909. Springer, 2016 edition, 2015.

[30] T. Chen and C. Guestrin. "*XGBoost: A Scalable Tree Boosting System*". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. ACM.

[31] A.M. Sirunyan et al. "*Search for heavy resonances decaying to a top quark and a bottom quark in the lepton+jets final state in proton–proton collisions at 13 TeV*". *Physics Letters B*, 777:39–63, Feb 2018.